Heinz H. Bauschke
Regina S. Burachik
D. Russell Luke *Editors*

Splitting Algorithms, Modern Operator Theory, and Applications

Springer

Splitting Algorithms, Modern Operator Theory, and Applications

Heinz H. Bauschke • Regina S. Burachik
D. Russell Luke
Editors

Splitting Algorithms, Modern Operator Theory, and Applications

Springer

Editors
Heinz H. Bauschke
Department of Mathematics
University of British Columbia
Kelowna, BC, Canada

Regina S. Burachik
School of IT & Mathematical Sciences
University of South Australia
Mawson Lakes, SA, Australia

D. Russell Luke
Inst. Numerische & Angewandte
Mathematik
Universität Göttingen
Göttingen, Niedersachsen, Germany

ISBN 978-3-030-25938-9 ISBN 978-3-030-25939-6 (eBook)
https://doi.org/10.1007/978-3-030-25939-6

This book is dedicated to the memory of Jonathan Michael Borwein, who was a great visionary, researcher, teacher, and mentor to many of us.

Preface

This book brings together several carefully reviewed papers in the broad areas of optimization and numerical analysis, with a particular emphasis on algorithms. The volume is a compendium of topics presented at the interdisciplinary workshop on *Splitting Algorithms, Modern Operator Theory, and Applications*, held at Casa Matemática Oaxaca (CMO) in Mexico, September 17–22, 2017. The participants came from Australia, Austria, Belgium, Brazil, Canada, Chile, France, Germany, Hong Kong, Japan, the Netherlands, Poland, Russia, Spain, Sweden, and the United States. Most papers in this volume grew out of talks delivered at this workshop. We believe that the reader will find this volume to be a valuable snapshot of the state of the art. We thank Elizabeth Loew from Springer for her help guiding this volume to completion and through production. Also, thanks to Arian Berdellima who helped with the technical compilation of the book.

The editors also thank CMO and BIRS for their support in hosting the workshop, in particular Claudia Arias Cao Romero and Miguel Ricardo Altamirano Ibarra for their local support. We also thank the very dedicated referees who were instrumental in ensuring the high quality of contributions.

Last but not least, we dedicate this volume to the memory of Jonathan Michael Borwein who was an early and enthusiastic supporter of this workshop. He was a great collaborator, teacher, mentor, and friend to us and to the many attendants of the workshop. (See also https://jonborwein.org for more on Jon's incomparable contributions.)

Kelowna, BC, Canada — Heinz H. Bauschke
Adelaide, SA, Australia — Regina S. Burachik
Göttingen, Germany — D. Russell Luke
May 2019

Contents

Contributors

Masoud Ahookhosh Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Leuven, Belgium

Hedy Attouch IMAG, Univ. Montpellier, CNRS, Montpellier, France

Heinz H. Bauschke Department of Mathematics, University of British Columbia, Kelowna, BC, Canada

Radu Ioan Boţ Faculty of Mathematics, University of Vienna, Vienna, Austria

Regina S. Burachik School of IT & Mathematical Sciences, University of South Australia, Mawson Lakes, SA, Australia

Christian Clason Faculty of Mathematics, University Duisburg-Essen, Essen, Germany

Ernö Robert Csetnek Faculty of Mathematics, University of Vienna, Vienna, Austria

Aris Danillidis DIM-CMM, Universidad de Chile, Santiago, Chile

Asen L. Dontchev University of Michigan, Ann Arbor, MI, USA

Xiaoling Fu School of Economics and Management, Southeast University, Nanjing, China

Lilian E. Glaudin Laboratoire Jacques-Louis Lions, Sorbonne Université, Paris, France

Rafal Goebel Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, USA

Sorin-Mihai Grad Faculty of Mathematics, University of Vienna, Vienna, Austria

Sylvain Gretchko Mathematics, UBCO, Kelowna, BC, Canada

Bingsheng He Department of Mathematics, South University of Science and Technology of China, Shenzhen, China

Department of Mathematics, Nanjing University, Nanjing, China

Barbara Kaltenbacher Institute of Mathematics, Alpen-Adria Universität Klagenfurt, Klagenfurt, Austria

C. Yalçın Kaya School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, Adelaide, SA, Australia

D. Russell Luke Inst. Numerische & Angewandte Mathematik, Universität Göttingen, Göttingen, Niedersachsen, Germany

Juan Enrique Martínez-Legaz Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona and Barcelona Graduate School of Mathematics (BGSMath), Barcelona, Spain

Dennis Meier Faculty of Mathematics, University of Vienna, Vienna, Austria

Walaa M. Moursi Electrical Engineering, Stanford University, Stanford, CA, USA

Faculty of Science, Mathematics Department, Mansoura University, Mansoura, Egypt

Dominikus Noll Université de Toulouse, Institut de Mathématiques, Toulouse, France

Panagiotis Patrinos Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Leuven, Belgium

Juan Peypouquet Departamento de Ingeniería Matemática & Centro de Modelamiento Matemático (CNRS UMI2807), FCFM, Universidad de Chile, Santiago, Chile

Elena Resmerita Institute of Mathematics, Alpen-Adria Universität Klagenfurt, Klagenfurt, Austria

Stephen Simons Department of Mathematics, University of California, Santa Barbara, CA, USA

Wilfredo Sosa Programa de Pôs-Graduação em Economia, Universidade Católica de Brasília, Taguatinga, Brazil

Matthew Tam University of Göttingen, Göttingen, Germany

Andreas Themelis Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Leuven, Belgium

Xiangfeng Wang School of Computer Science and Technology, East China Normal University, Shanghai, China

Isao Yamada Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan

Masao Yamagishi Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan

Xiaoming Yuan Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong

Yuriy Zinchenko University of Calgary, Department of Mathematics and Statistics, Calgary, AB, Canada

Chapter 1
Convergence Rate of Proximal Inertial Algorithms Associated with Moreau Envelopes of Convex Functions

Hedy Attouch and Juan Peypouquet

Abstract In a Hilbert space setting $\mathcal{H}$, we develop new inertial proximal-based algorithms that aim to rapidly minimize a convex lower-semicontinuous proper function $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. The guiding idea is to use an accelerated proximal scheme where, at each step, Φ is replaced by its Moreau envelope, with varying approximation parameter. This leads to consider a Relaxed Inertial Proximal Algorithm (RIPA) with variable parameters which take into account the effects of inertia, relaxation, and approximation. (RIPA) was first introduced to solve general maximally monotone inclusions, in which case a judicious adjustment of the parameters makes it possible to obtain the convergence of the iterates towards the equilibria. In the case of convex minimization problems, convergence analysis of (RIPA) was initially addressed by Attouch and Cabot, based on its formulation as an inertial gradient method with varying potential functions. We propose a new approach to this algorithm, along with further developments, based on its formulation as a proximal algorithm associated with varying Moreau envelopes. For convenient choices of the parameters, we show the fast optimization property of the function values, with the order $o(k^{-2})$, and the weak convergence of the iterates. This is in line with the recent studies of Su-Boyd-Candès, Chambolle-Dossal, Attouch-Peypouquet. We study the impact of geometric assumptions on the convergence rates, and the stability of the results with respect to perturbations and errors. Finally, in the case of structured minimization problems *smooth + nonsmooth*, based on this approach, we introduce new proximal-gradient inertial algorithms for which similar convergence rates are shown.

H. Attouch (✉)
IMAG, Univ. Montpellier, CNRS, Montpellier, France
e-mail: hedy.attouch@umontpellier.fr

J. Peypouquet
Departamento de Ingeniería Matemática & Centro de Modelamiento Matemático (CNRS UMI2807), FCFM, Universidad de Chile, Santiago, Chile
e-mail: jpeypou@dim.uchile.cl

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_1

Keywords Inertial proximal algorithms · Lyapunov analysis · Maximally monotone operators · Moreau envelopes · Nesterov accelerated gradient method · Nonsmooth convex minimization · Proximal-gradient algorithms · Relaxation

AMS Subject Classification 37N40, 46N10, 49M30, 65K05, 65K10, 90B50, 90C25

1.1 Introduction, Preliminary Results

In this paper, we analyze the convergence rate of a general class of inertial proximal algorithms for convex optimization. Following [10, 16], our main motivation is to put to the fore inertial proximal algorithms that converge for general monotone inclusions, and which, in the case of convex minimization, give fast convergence rates for the values. Let $\mathcal{H}$ be a real Hilbert space endowed with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, and let $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a convex, lower-semicontinuous, proper function. We will study the convergence rate of the Relaxed Inertial Proximal Algorithm, (RIPA) for short,

$$\text{(RIPA)} \qquad \begin{cases} y_k = x_k + \alpha_k (x_k - x_{k-1}) \\ x_{k+1} = (1 - \rho_k) y_k + \rho_k \text{prox}_{\mu_k \Phi}(y_k), \end{cases}$$

where the varying parameters α_k, ρ_k, and μ_k, take into account of the inertial, relaxation, and approximation effects, respectively. Convergence analysis of (RIPA) was recently considered by Attouch-Cabot in [10], relying on its formulation as an inertial *gradient* algorithm associated with varying Moreau envelopes. We propose a new approach to (RIPA), based on its reformulation as an inertial *proximal* algorithm. In doing so, we obtain convergence rates relying on *different* types of assumptions, and enrich the analysis of (RIPA) in several aspects. We study the impact of geometric assumptions on the convergence rates, and the stability of the results with respect to perturbations and errors. In the case of minimization problems with additive *smooth* + *nonsmooth* structure, our study naturally provides new inertial proximal-gradient algorithms with fast convergence properties.

1.1.1 Introducing (RIPA) from a Dynamic Perspective

As pointed out in [9, 16], (RIPA) bears close connection with the *Regularized Inertial Gradient System*, which is the non-autonomous second-order differential equation

$$\text{(RIGS)} \quad \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla \Phi_{\lambda(t)}(x(t)) = 0. \tag{1.1}$$

It involves two varying parameters: $\gamma(\cdot)$ is a positive viscous damping parameter, and $\lambda(t)$ is the approximation parameter entering the Moreau envelope of Φ, which will be conveniently tuned. Recall that, for $\lambda > 0$, and $x \in \mathcal{H}$,

$$\Phi_\lambda(x) = \inf_{\xi \in \mathcal{H}} \left\{ \Phi(\xi) + \frac{1}{2\lambda} \|x - \xi\|^2 \right\}, \tag{1.2}$$

and that $\text{prox}_{\lambda\Phi}(x)$ is the unique point where the above minimum is achieved. Writing the optimality condition for (1.2), we get $\text{prox}_{\lambda\Phi}(x) + \lambda\partial\Phi\left(\text{prox}_{\lambda\Phi}(x)\right) \ni x$, that is $\text{prox}_{\lambda\Phi}(x) = (I + \lambda\partial\Phi)^{-1}(x)$ (see Appendix "Some Properties of the Moreau Envelope" for further results). The regularity of the function $]0, +\infty[\times \mathcal{H} \ni (\lambda, x) \mapsto \nabla\Phi_\lambda(x)$ makes (1.1) a classical ordinary differential equation, whose Cauchy problem is well posed (see Appendix, Theorem 1.6.2). This makes a big difference with the approach based on the differential inclusion $\ddot{x}(t) + \gamma(t)\dot{x}(t) + \partial\Phi(x(t)) \ni 0$, for which the study of the existence and uniqueness of trajectories raises significant difficulties, even in a finite dimensional setting, see [11]. Thus, our approach will be to replace Φ by $\Phi_{\lambda(t)}$, while keeping $\lambda(t)$ bounded below by a positive constant. This makes sense because, for $\lambda > 0$ fixed, Φ_λ completely determines Φ. This contrasts with the Yosida approximation method which is based on $\lambda(t) \to 0$.

The time discretization of (RIGS) naturally leads to (RIPA), as follows: Take a time step $h_k > 0$, and set $\tau_k = \sum_{i=1}^{k} h_i$, $x_k = x(\tau_k)$, $\lambda_k = \lambda(\tau_k)$, $\gamma_k = \gamma(\tau_k)$. An implicit finite-difference scheme for (1.1) with centered second-order variation gives

$$\frac{1}{h_k^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma_k}{h_k}(x_k - x_{k-1}) + \nabla\Phi_{\lambda_k}(x_{k+1}) = 0.$$

Equivalently, $x_{k+1} + h_k^2\nabla\Phi_{\lambda_k}(x_{k+1}) = x_k + (1 - \gamma_k h_k)(x_k - x_{k-1})$, which gives

$$x_{k+1} = \left(I + h_k^2\nabla\Phi_{\lambda_k}\right)^{-1}(x_k + (1 - \gamma_k h_k)(x_k - x_{k-1})). \tag{1.3}$$

Setting $s_k = h_k^2$ and $\alpha_k = 1 - \gamma_k h_k$, we obtain

$$\text{(RIPA)}_1 \quad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{s_k\Phi_{\lambda_k}}(y_k). \end{cases}$$

This is an inertial proximal algorithm, where the potential function Φ_{λ_k}, and the step size s_k, vary at each iteration. Let us show that this algorithm can be equivalently written as (RIPA). We have

$$x_{k+1} = \text{prox}_{s_k\Phi_{\lambda_k}}(y_k) = y_k - s_k\nabla(\Phi_{\lambda_k})_{s_k}(y_k) = y_k - s_k\nabla\Phi_{\lambda_k + s_k}(y_k),$$

where the last equality comes from the resolvent equation (or semi-group property) $(\Phi_{\lambda_k})_{s_k} = \Phi_{\lambda_k+s_k}$. This gives

$$(\text{RIPA})_2 \quad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = y_k - s_k \nabla \Phi_{\mu_k}(y_k) \end{cases}$$

with $\mu_k = \lambda_k + s_k$. Note that $\mu_k > s_k$. Developing the above relation, we obtain

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \dfrac{\lambda_k}{\lambda_k + s_k} y_k + \dfrac{s_k}{\lambda_k + s_k} \text{prox}_{(\lambda_k+s_k)\Phi}(y_k). \end{cases}$$

We recover the algorithm (RIPA), with relaxation parameter $\rho_k = \frac{s_k}{\lambda_k+s_k}$, and proximal parameter $\mu_k = \lambda_k + s_k$.

Conversely, let us show that (RIPA) can be equivalently formulated as $(\text{RIPA})_1$ or $(\text{RIPA})_2$. It is convenient to set $\Phi_0 = \Phi$, which is in accordance with the limiting behavior of Φ_λ as $\lambda \to 0$ (see Section 1.3.1 for further details).

Lemma 1.1 *Suppose that the relaxation parameter ρ_k satisfies $0 < \rho_k \leq 1$ for all $k \geq 1$. Then, the algorithm (RIPA) with parameters (ρ_k, μ_k) can be equivalently formulated either as $(\text{RIPA})_1$ with parameters $s_k = \rho_k\mu_k$ and $\lambda_k = \mu_k(1 - \rho_k)$, or as $(\text{RIPA})_2$ with parameters $s_k = \rho_k\mu_k$ and μ_k.*

Proof By definition of the Yosida approximation,

$$(1 - \rho_k)y_k + \rho_k \text{prox}_{\mu_k\Phi}(y_k) = y_k - \rho_k\mu_k \nabla \Phi_{\mu_k}(y_k).$$

Set $s_k = \rho_k\mu_k$. When $\rho_k < 1$ we have $\mu_k > s_k$. Setting $\mu_k = s_k + \lambda_k$, with $\lambda_k > 0$, similar computation as above gives

$$\begin{aligned}(1 - \rho_k)y_k + \rho_k \text{prox}_{\mu_k\Phi}(y_k) &= y_k - s_k \nabla \Phi_{s_k+\lambda_k}(y_k) = y_k - s_k \nabla (\Phi_{\lambda_k})_{s_k}(y_k) \\ &= \text{prox}_{s_k\Phi_{\lambda_k}}(y_k),\end{aligned}$$

which gives $(\text{RIPA})_1$. When $\rho_k = 1$, we have $s_k = \mu_k$ and $\lambda_k = 0$. Since $\Phi_{\lambda_k} = \Phi_0 = \Phi$, it follows

$$(1 - \rho_k)y_k + \rho_k \text{prox}_{\mu_k\Phi}(y_k) = \text{prox}_{\mu_k\Phi}(y_k) = \text{prox}_{s_k\Phi}(y_k) = \text{prox}_{s_k\Phi_{\lambda_k}}(y_k).$$

$(\text{RIPA})_2$ follows at once by a similar argument as above.

According to the formulation of the algorithm, proximal as in $(\text{RIPA})_1$, or gradient as in $(\text{RIPA})_2$, we obtain close but different results. This is because each method is based on a specific descent inequality. We will compare our approach, based on $(\text{RIPA})_1$, to the approach developed by Attouch-Cabot in [8, Part C], and which is based on $(\text{RIPA})_2$. An advantage of the proximal approach $(\text{RIPA})_1$ is that,

for $\rho_k = 0$, we find the classical form of the results concerning the inertial proximal algorithm: they express themselves directly with Φ, not with a Moreau envelope.

1.1.2 The Sequence (t_k)

In the study of (RIPA), the sequence (t_k), defined by

$$t_k := 1 + \sum_{l=k}^{+\infty} \left(\prod_{j=k}^{l} \alpha_j \right), \tag{1.4}$$

plays a central role. The sequence (t_k) is well defined provided the following condition holds:

$$\sum_{l=k}^{+\infty} \left(\prod_{j=k}^{l} \alpha_j \right) < +\infty \quad \text{for every } k \geq 1. \tag{K_0}$$

One can simply retrieve (α_k) from (t_k), thanks to the following formula:

Lemma 1.2 *Assume that the nonnegative sequence (α_k) satisfies (K_0). Then, $1 + \alpha_k t_{k+1} = t_k$ for every $k \geq 1$. Equivalently, $\alpha_k = \dfrac{t_k - 1}{t_{k+1}}$ for every $k \geq 1$.*

Therefore, (RIPA) can be written in an equivalent way

$$\begin{cases} y_k = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{s_k \Phi_{\lambda_k}}(y_k). \end{cases} \tag{1.5}$$

In line with [7, 21, 26], we introduce the following property

$$t_{k+1}^2 - t_k^2 \leq t_{k+1} \quad \text{for every } k \geq 1. \tag{K_1}$$

It is convenient to introduce the following quantity, which, under assumption (K_1), is nonnegative:

$$\delta_k := t_k^2 - t_{k+1}^2 + t_{k+1}. \tag{1.6}$$

To prove the convergence of the iterations, we will need the slightly stronger notion:

$$t_{k+1}^2 - t_k^2 \leq m\, t_{k+1} \quad \text{for some } m \in (0, 1) \quad \text{and every } k \geq 1. \tag{K_1^+}$$

So, assumption (K_1^+) gives

$$\delta_k \geq (1-m)t_{k+1}. \tag{1.7}$$

These conditions can be found in several previous works:

- In a seminal work [33], Nesterov introduced the accelerated gradient method, which corresponds to the sequence (t_k) obtained by taking equality in (K_1), that is $t_{k+1} = \frac{1}{2}\left(1+\sqrt{1+4t_k^2}\right)$, and $t_1 = 1$. This choice leads to an increasing sequence (α_k), whose asymptotic behavior is similar to that of $1 - \frac{3}{k}$ as $k \to +\infty$. It has been further extended to structured minimization problems by Beck-Teboulle in [21] (FISTA). This scheme exhibits the convergence rate of values $\mathcal{O}(\frac{1}{k^2})$, which is optimal among all methods having only information about the gradient at consecutive iterates [34].
- Recently, the case $\alpha_k = 1 - \frac{\alpha}{k}$ where $\alpha > 0$ has gained special attention, see [13, 15, 26, 38]. A rather involved but straightforward computation gives that for $\alpha > 1$, the condition (K_0) is satisfied and $t_k = \frac{k-1}{\alpha-1}$. We immediately deduce that (K_1) is equivalent to $\alpha \geq 3$, and (K_1^+) is equivalent to $\alpha > 3$. While keeping the same computational complexity as in the case $\alpha = 3$, taking $\alpha > 3$ offers many advantages. First, it ensures the convergence of the sequences (x_k), as proved by Chambolle-Dossal [26], see also [13]. The convergence of the sequences generated by (FISTA) has not been established so far. Second, as proved by Attouch-Peypouquet in [15], it provides the better convergence rate $o\left(\frac{1}{k^2}\right)$. Thus, $\alpha = 3$ appears as a critical value. The subcritical case $\alpha < 3$ has been recently considered by Apidopoulos-Aujol-Dossal [3] and Attouch-Chbani-Riahi [14].
- The case $t_k \equiv 1$ is equivalent to $\alpha_k \equiv 0$, which gives the proximal algorithm (without inertia).
- An extended study of the convergence rates of the inertial forward-backward methods with general damping coefficient (α_k) has been developed by Attouch-Cabot [7].

1.1.3 A Model Result

Consider the case $\alpha_k = 1 - \frac{\alpha}{k}$. The relaxed inertial proximal algorithm writes

$$\begin{cases} y_k = x_k + (1-\dfrac{\alpha}{k})(x_k - x_{k-1}) \\ x_{k+1} = (1-\rho_k)y_k + \rho_k \mathrm{prox}_{\mu_k \Phi}(y_k). \end{cases}$$

The following result is a consequence of Theorems 1.2.2 and 1.2.3 proved in the next section:

Theorem 1.1.1 *Suppose that* $\alpha > 3$, *and that* $((1-\rho_k)\mu_k)$ *and* $(\rho_k\mu_k)$ *are nondecreasing sequences of positive numbers. Setting* $\lambda_k = (1-\rho_k)\mu_k$, *we have* $\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2}\right)$. *Equivalently, setting* $p_k := \mathrm{prox}_{\lambda_k \Phi} x_k$, *we have*

$$\Phi(p_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2}\right) \quad \text{and} \quad \|x_k - p_k\| = o\left(\frac{\sqrt{\mu_k}}{k}\right).$$

Moreover, the sequence (x_k) *converges weakly to a minimizer of* Φ *if* $\sup_{k\geq 0} \frac{\sqrt{\mu_k}}{k} < +\infty$.

This makes a difference with the approach developed in [10], where it is assumed that (μ_k) and $(\rho_k\mu_k)$ are nondecreasing sequences of positive numbers, and where the convergence results are expressed with Φ_{μ_k} instead of Φ_{λ_k}.

1.1.4 Organization of the Paper

In Section 1.2, we analyze the convergence properties of (RIPA), thus obtaining new inertial proximal algorithms, with fast convergence properties. In Section 1.3, we compare the results obtained using the proximal and gradient approaches, and combine our results with those for general maximally monotone operators. In Section 1.4, we obtain better convergence rates and strong convergence of the iterates, provided the function has a strong minimum. In Section 1.5, we examine the stability of (RIPA) with respect to perturbations and errors. In Section 1.6, we develop regularized inertial proximal-gradient algorithms to solve structured *smooth + nonsmooth* minimization problems. The Appendix contains auxiliary technical results used along the paper.

1.2 A Regularized Inertial Proximal Algorithm

In this section, we analyze the convergence properties of (RIPA). We set $\lambda_k := (1-\rho_k)\mu_k$ and $s_k := \rho_k\mu_k$, and make the following standing assumptions on the parameters:

$$(\mathcal{A}) \begin{cases} \bullet\ \Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\} \text{ is a convex lower-semicontinuous proper function;} \\ \bullet\ S := \mathrm{argmin}\Phi \neq \emptyset; \\ \bullet\ (\alpha_k) \text{ is a sequence in } [0,1] \text{ satisfying } (K_0); \\ \bullet\ 0 < \rho_k \leq 1 \text{ for all } k \geq 1; \\ \bullet\ (\mu_k) \text{ is a sequence of positive numbers;} \\ \bullet\ (\lambda_k) = ((1-\rho_k)\mu_k) \text{ and } (s_k) = (\rho_k\mu_k) \\ \quad \text{are nondecreasing sequences of positive numbers.} \end{cases}$$

As shown in Lemma 1.1, (RIPA) takes the compact proximal formulation with varying potential functions

$$(\text{RIPA})_1 \quad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{s_k \Phi_{\lambda_k}}(y_k). \end{cases}$$

We denote by $(x_k)_{k\in\mathbb{N}}$, $(y_k)_{k\in\mathbb{N}}$ the sequences in $\mathcal{H}$ defined by (RIPA) for $k \geq 0$, and x_{-1}, $x_0 \in \mathcal{H}$. Under assumption $(\mathcal{A})$, for $x^* \in \arg\min \Phi$, we define the sequence $(\mathcal{E}_{x^*,k})_{k\in\mathbb{N}}$ of nonnegative real numbers

$$\mathcal{E}_{x^*,k} = t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{1}{2s_k} \|x^* - x_{k-1} + t_k(x_{k-1} - x_k)\|^2, \tag{1.8}$$

where the sequence (t_k) has been defined in (1.4). We also define the sequence $(E_k)_{k\in\mathbb{N}}$

$$E_k = t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_k^2}{2s_k} \|x_{k-1} - x_k\|^2. \tag{1.9}$$

We have $E_k = t_k^2 W_k$, where W_k is the global energy at stage k, namely:

$$W_k = \Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi + \frac{1}{2s_k} \|x_{k-1} - x_k\|^2. \tag{1.10}$$

Remark 1.1 Recall the equalities $\min_{\mathcal{H}} \Phi_\lambda = \min_{\mathcal{H}} \Phi$ and

$$\Phi_\lambda(x) - \min_{\mathcal{H}} \Phi = \left(\Phi(\text{prox}_{\lambda\Phi} x) - \min_{\mathcal{H}} \Phi \right) + \frac{1}{2\lambda} \|\text{prox}_{\lambda\Phi} x - x\|^2,$$

by which an upper bound of $\Phi_\lambda(x) - \min_{\mathcal{H}} \Phi$ is also an upper bound of $\Phi(\text{prox}_{\lambda\Phi} x) - \min_{\mathcal{H}} \Phi$ and of $\frac{1}{2\lambda}\|\text{prox}_{\lambda\Phi} x - x\|^2$. This argument will be used repeatedly in the sequel.

1.2.1 Preliminary Lemmas

As a classical key ingredient, we use the descent rule for the proximal method. It can be obtained as a special case of the descent rule for the proximal-gradient method, see, for example, [21, Lemma 2.3], [26, Lemma 3.1]. For the convenience of the reader, we give a short proof in the case of the proximal method. Then we will apply it to the algorithm (RIPA).

Lemma 1.3 *Let $\varphi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a convex, lower-semicontinuous, and proper function. Let $s > 0$, $y \in \mathcal{H}$, and set $p = prox_{s\varphi} y$.*

(i) *For any $u \in \mathcal{H}$ we have*

$$\varphi(p) + \frac{1}{2s}\|u - p\|^2 + \frac{1}{2s}\|y - p\|^2 \leq \varphi(u) + \frac{1}{2s}\|u - y\|^2; \quad (1.11)$$

$$\varphi(p) + \frac{1}{s}\|y - p\|^2 + \frac{1}{s}\langle y - p, u - y\rangle \leq \varphi(u). \quad (1.12)$$

(ii) *For any x, x^* in $\mathcal{H}$, and for any $t \geq 1$ we have*

$$t^2(\varphi(p) - \varphi(x^*)) + \frac{1}{2s}\|x^* - x + t(x - p)\|^2 + \frac{t^2}{2s}\|y - p\|^2$$
$$\leq (t^2 - t)(\varphi(x) - \varphi(x^*)) + \frac{1}{2s}\|x^* - x + t(x - y)\|^2.$$

Proof

(i) By the definition of p, $\frac{1}{s}(y - p)$ is a subgradient of φ at p. Hence, for all $u \in \mathcal{H}$, we have

$$\varphi(u) \geq \varphi(p) + \langle \frac{1}{s}(y - p), u - p\rangle.$$

Inequality (1.11) is a consequence of $\langle y - p, u - p\rangle = \frac{1}{2}\{\|y - p\|^2 + \|u - p\|^2 - \|u - y\|^2\}$, while (1.12) is obtained by writing $u - p = (u - y) + (y - p)$ in the above inequality.

(ii) Following the proof of [22, Lemma 1.7], we set $u = \frac{1}{t}x^* + (1 - \frac{1}{t})x$ in (1.11), and we use the convexity of φ at u:

$$\varphi(p) + \frac{1}{2s}\left\|\frac{1}{t}x^* + \left(1 - \frac{1}{t}\right)x - p\right\|^2 + \frac{1}{2s}\|y - p\|^2$$
$$\leq \frac{1}{t}\varphi(x^*) + \left(1 - \frac{1}{t}\right)\varphi(x) + \frac{1}{2s}\left\|\frac{1}{t}x^* + \left(1 - \frac{1}{t}\right)x - y\right\|^2.$$

It suffices to rearrange the terms, and to multiply by t^2 to obtain the desired inequality.

□

Lemma 1.4 *Let assumption* $(\mathcal{A})$ *hold, and let* (x_k) *be a sequence generated by the algorithm (RIPA).*

(i) *For any* $x^* \in \arg\min \Phi$ *and* $k \geq 0$*, we have*

$$\mathcal{E}_{x^*,k+1} + \delta_k \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_{k+1}^2}{2s_k} \|x_{k+1} - y_k\|^2 \leq \mathcal{E}_{x^*,k}.$$

(ii) *The energy sequence* (W_k) *satisfies, for every* $k \geq 1$*,*

$$W_{k+1} - W_k \leq -\frac{1-\alpha_k^2}{2s_k} \|x_k - x_{k-1}\|^2.$$

As a consequence, the sequence (W_k) *is nonincreasing.*

(iii) *Assume, moreover, that condition* (K_1) *is satisfied. Then, for each* $k \geq 0$*, we have*

$$E_{k+1} - t_{k+1} \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_k}{2s_k} \|x_k - x_{k-1}\|^2 \leq E_k.$$

Proof

(i) We apply Lemma 1.3(ii) with $\varphi = \Phi_{\lambda_k}$, $x^* \in \arg\min \Phi_{\lambda_k}$, $t = t_{k+1}$, $s = s_k$, $x = x_k$ and $y = y_k$. Hence $p = \operatorname{prox}_{s_k \Phi_{\lambda_k}} y_k = x_{k+1}$. We obtain, for each $k \geq 0$,

$$t_{k+1}^2(\Phi_{\lambda_k}(x_{k+1}) - \Phi_{\lambda_k}(x^*)) + \frac{1}{2s_k}\|x^* - x_k + t_{k+1}(x_k - x_{k+1})\|^2 + \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - y_k\|^2$$
$$\leq (t_{k+1}^2 - t_{k+1})(\Phi_{\lambda_k}(x_k) - \Phi_{\lambda_k}(x^*)) + \frac{1}{2s_k}\|x^* - x_k + t_{k+1}(x_k - y_k)\|^2.$$

By assumption $(\mathcal{A})$, the sequences (λ_k) and (s_k) are nondecreasing. These properties, respectively, imply that for all $k \geq 0$, $\Phi_{\lambda_{k+1}} \leq \Phi_{\lambda_k}$ and $\frac{1}{s_{k+1}} \leq \frac{1}{s_k}$. Moreover $\Phi_{\lambda_k}(x^*) = \min_{\mathcal{H}} \Phi$. Therefore, we deduce from the above inequality that

$$t_{k+1}^2(\Phi_{\lambda_{k+1}}(x_{k+1}) - \min_{\mathcal{H}} \Phi) + \frac{1}{2s_{k+1}}\|x^* - x_k + t_{k+1}(x_k - x_{k+1})\|^2$$
$$+ \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - y_k\|^2$$
$$\leq (t_{k+1}^2 - t_{k+1})(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi) + \frac{1}{2s_k}\|x^* - x_k + t_{k+1}(x_k - y_k)\|^2.$$

Let us examine the last term of this inequality. By (1.5) we have

$$x^*-x_k+t_{k+1}(x_k-y_k) = x^*-x_k-(t_k-1)(x_k-x_{k-1}) = x^*-x_{k-1}+t_k(x_{k-1}-x_k).$$

Collecting these results, we obtain

$$\begin{aligned} t_{k+1}^2(\Phi_{\lambda_{k+1}}(x_{k+1}) - \min_{\mathcal{H}} \Phi) + \frac{1}{2s_{k+1}}\|x^* - x_k + t_{k+1}(x_k - x_{k+1})\|^2 \\ + \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - y_k\|^2 \\ \leq (t_{k+1}^2 - t_{k+1})(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi) + \frac{1}{2s_k}\|x^* - x_{k-1} + t_k(x_{k-1} - x_k)\|^2. \end{aligned}$$

By definition of $\delta_k := t_k^2 - t_{k+1}^2 + t_{k+1}$, and $\mathcal{E}_{x^*,k}$ we get

$$\mathcal{E}_{x^*,k+1} + \delta_k\left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi\right) + \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - y_k\|^2 \leq \mathcal{E}_{x^*,k},$$

which is the desired inequality.

(ii) In inequality (1.11) of Lemma 1.3(i), we neglect the term $\frac{1}{2s}\|p - y\|^2$ and we set $\varphi = \Phi_{\lambda_k}$, $s = s_k$, $u = x_k$ and $y = y_k$. Hence $p = \text{prox}_{s_k\Phi_{\lambda_k}} y_k = x_{k+1}$. For each $k \geq 0$, we obtain

$$\Phi_{\lambda_k}(x_{k+1}) + \frac{1}{2s_k}\|x_k - x_{k+1}\|^2 \leq \Phi_{\lambda_k}(x_k) + \frac{1}{2s_k}\|x_k - y_k\|^2.$$

Recalling that $\Phi_{\lambda_{k+1}}(x_{k+1}) \leq \Phi_{\lambda_k}(x_{k+1})$, $\frac{1}{s_{k+1}} \leq \frac{1}{s_k}$, and $y_k - x_k = \frac{t_k-1}{t_{k+1}}(x_k - x_{k-1})$, we get

$$\Phi_{\lambda_{k+1}}(x_{k+1}) + \frac{1}{2s_{k+1}}\|x_k - x_{k+1}\|^2 \leq \Phi_{\lambda_k}(x_k) + \frac{1}{2s_k}\frac{(t_k-1)^2}{t_{k+1}^2}\|x_k - x_{k-1}\|^2. \tag{1.13}$$

By Lemma 1.2 we have $1 + \alpha_k t_{k+1} = t_k$. Hence, inequality (1.13) can be equivalently formulated as

$$\Phi_{\lambda_{k+1}}(x_{k+1}) + \frac{1}{2s_{k+1}}\|x_k - x_{k+1}\|^2 \leq \Phi_{\lambda_k}(x_k) + \frac{1}{2s_k}{\alpha_k}^2\|x_k - x_{k-1}\|^2. \tag{1.14}$$

According to the definition of W_k, we obtain

$$W_{k+1} - W_k \leq -\frac{1-\alpha_k^2}{2s_k}\|x_k - x_{k-1}\|^2,$$

which is the desired inequality.

(iii) In inequality (1.13), we subtract $\min_{\mathcal{H}} \Phi$ from each side, and multiply by t_{k+1}^2, resulting in

$$E_{k+1} = t_{k+1}^2 \left(\Phi_{\lambda_{k+1}}(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) + \frac{t_{k+1}^2}{2s_{k+1}} \|x_k - x_{k+1}\|^2$$
$$\leq t_{k+1}^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{1}{2s_k}(t_k - 1)^2 \|x_k - x_{k-1}\|^2.$$

In view of (K_1) we have $t_{k+1}^2 \leq t_{k+1} + t_k^2$. Since $t_k \geq 1$ we have $(t_k - 1)^2 = t_k^2 - t_k + (1 - t_k) \leq t_k^2 - t_k$. Collecting these results we obtain

$$E_{k+1} \leq t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_k^2}{2s_k} \|x_k - x_{k-1}\|^2$$
$$+ t_{k+1} \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) - \frac{t_k}{2s_k} \|x_k - x_{k-1}\|^2,$$

which establishes the desired inequality.

Lemma 1.5 *Let assumption $(\mathcal{A})$ hold, and let (x_k) be a sequence generated by the algorithm (RIPA). Fix $x^* \in \mathcal{H}$, and consider the anchor sequence (h_k), which is defined by $h_k = \frac{1}{2}\|x_k - x^*\|^2$. We have, for every $k \geq 1$,*

$$h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) = \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 + \langle \mathrm{prox}_{s_k \Phi_{\lambda_k}}(y_k) - y_k, y_k - x^* \rangle$$
$$+ \frac{1}{2}\|\mathrm{prox}_{s_k \Phi_{\lambda_k}}(y_k) - y_k\|^2. \tag{1.15}$$

If, moreover, $x^* \in \arg\min \Phi$, then

$$h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \leq \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 - s_k(\Phi_{\lambda_k}(x_{k+1}) - \min \Phi).$$

Proof Observe that

$$\begin{aligned}
\|y_k - x^*\|^2 &= \|x_k + \alpha_k(x_k - x_{k-1}) - x^*\|^2 \\
&= \|x_k - x^*\|^2 + \alpha_k^2\|x_k - x_{k-1}\|^2 + 2\alpha_k\langle x_k - x^*, x_k - x_{k-1}\rangle \\
&= \|x_k - x^*\|^2 + \alpha_k^2\|x_k - x_{k-1}\|^2 \\
&+ \alpha_k\|x_k - x^*\|^2 + \alpha_k\|x_k - x_{k-1}\|^2 - \alpha_k\|x_{k-1} - x^*\|^2 \\
&= \|x_k - x^*\|^2 + \alpha_k(\|x_k - x^*\|^2 - \|x_{k-1} - x^*\|^2) + (\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 \\
&= 2[h_k + \alpha_k(h_k - h_{k-1})] + (\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2.
\end{aligned}$$

Setting briefly $A_k = h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) = \frac{1}{2}\|x_{k+1} - x^*\|^2 - [h_k + \alpha_k(h_k - h_{k-1})]$, we deduce that

$$\begin{aligned} A_k &= \frac{1}{2}\|x_{k+1} - x^*\|^2 - \frac{1}{2}\|y_k - x^*\|^2 + \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 \\ &= \langle x_{k+1} - y_k, y_k - x^*\rangle + \frac{1}{2}\|x_{k+1} - y_k\|^2 + \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2. \end{aligned}$$

Using the equality $x_{k+1} = \mathrm{prox}_{s_k \Phi_{\lambda_k}}(y_k)$, we obtain (1.15).

Let us now assume that $x^* \in \arg\min \Phi$, and apply inequality (1.12) with $\phi = \Phi_{\lambda_k}$, $y = y_k$ and $u = x^*$. By definition of (RIPA), $\mathrm{prox}_{s_k \Phi_{\lambda_k}}(y_k) = x_{k+1}$. Hence,

$$\Phi_{\lambda_k}(x^*) \geq \Phi_{\lambda_k}(x_{k+1}) + \frac{1}{s_k}\|y_k - x_{k+1}\|^2 + \frac{1}{s_k}\langle x^* - y_k, y_k - x_{k+1}\rangle.$$

Since $\Phi_{\lambda_k}(x^*) = \min \Phi$, we infer that

$$\langle x_{k+1} - y_k, y_k - x^*\rangle \leq -s_k(\Phi_{\lambda_k}(x_{k+1}) - \min \Phi) - \|y_k - x_{k+1}\|^2.$$

Returning to (1.15), we obtain

$$\begin{aligned} h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \leq{}& \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 \\ &- s_k(\Phi_{\lambda_k}(x_{k+1}) - \min \Phi) - \frac{1}{2}\|\mathrm{prox}_{s_k \Phi_{\lambda_k}}(y_k) - y_k\|^2. \end{aligned}$$

Neglecting the last term of the inequality above, this completes the proof of Lemma 1.5.

1.2.2 Fast Convergence of the Values

Theorem 1.2.1 *Let us make assumption $(\mathcal{A})$, and suppose that the sequence (t_k) satisfies (K_1). Then, the following properties hold:*

(i) *For any $x^* \in \arg\min \Phi$, the sequence $(\mathcal{E}_{x^*,k})_{k\in\mathbb{N}}$ is nonincreasing and converges. Moreover, for all $k \geq 1$, we have*

$$\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \leq \frac{\mathcal{E}_{x^*,0}}{t_k^2} \tag{1.16}$$

$$\Phi(\text{prox}_{\lambda_k\Phi}x_k) - \min_{\mathcal{H}}\Phi \leq \frac{\mathcal{E}_{x^*,0}}{t_k^2} \tag{1.17}$$

$$\|\text{prox}_{\lambda_k\Phi}x_k - x_k\| \leq \sqrt{\frac{2\lambda_k\mathcal{E}_{x^*,0}}{t_k^2}} \tag{1.18}$$

$$\|\nabla\Phi_{\lambda_k}(x_k)\| \leq \sqrt{\frac{2\mathcal{E}_{x^*,0}}{\lambda_k t_k^2}}. \tag{1.19}$$

(ii) $\displaystyle\sum_{k=0}^{\infty}\delta_k\left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}}\Phi\right) < +\infty.$

(iii) $\displaystyle\sum_{k=0}^{\infty}s_k t_{k+1}^2\|\nabla\Phi_{\lambda_k+s_k}(y_k)\|^2 < +\infty$, *and* $\|\nabla\Phi_{\lambda_k+s_k}(y_k)\| = o(1/t_{k+1})$.

Proof

(i) The assertion is a consequence of Lemma 1.4(i). From the inequality $\mathcal{E}_{x^*,k} \leq \mathcal{E}_{x^*,0}$ and the definition (1.8) of $\mathcal{E}_{x^*,k}$ we then deduce (1.16). Estimations (1.17), (1.18) are consequences of (1.16) and Remark 1.1. Estimation (1.19) follows from (1.18) and $\nabla\Phi_{\lambda_k}(x_k) = \frac{1}{\lambda_k}(x_k - \text{prox}_{\lambda_k\Phi}x_k)$.

(ii) A consequence of Lemma 1.4(i) and of the existence of $\lim_{k\to+\infty}\mathcal{E}_{x^*,k}$ (notice that δ_k may be zero, and the information conveyed is then void).

(iii) The same argument shows $\sum_{k=0}^{\infty}\frac{t_{k+1}^2}{s_k}\|x_{k+1} - y_k\|^2 < +\infty$. From the formulation $(\text{RIPA})_2$ of the algorithm, we deduce $\sum_{k=0}^{\infty}s_k t_{k+1}^2\|\nabla\Phi_{\lambda_k+s_k}(y_k)\|^2 < +\infty$. Since (s_k) is nondecreasing, we deduce $\sum_{k=0}^{\infty}t_{k+1}^2\|\nabla\Phi_{\lambda_k+s_k}(y_k)\|^2 < +\infty$. As a consequence $t_{k+1}\|\nabla\Phi_{\lambda_k+s_k}(y_k)\| \to 0$, as $k \to +\infty$. If $(t_k)_{k\in\mathbb{N}}$ is nondecreasing, a fortiori, we have $t_k\|\nabla\Phi_{\lambda_k+s_k}(y_k)\| \to 0$, as $k \to +\infty$.

Remark 1.2 Recall that $\lambda_k + s_k = \mu_k$. Since $\lambda \mapsto \Phi_\lambda$ is nonincreasing, we have

$$\Phi_{\mu_k}(x_k) - \min_{\mathcal{H}}\Phi = \Phi_{\lambda_k+s_k}(x_k) - \min_{\mathcal{H}}\Phi \leq \Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}}\Phi.$$

The same arguments as in the proof of Theorem 1.2.1(i) yield $\Phi_{\mu_k}(x_k) - \min_{\mathcal{H}}\Phi \leq \frac{\mathcal{E}_{x^*,0}}{t_k^2}$, which gives for $k \geq 1$

$$\Phi(\text{prox}_{\mu_k\Phi}x_k) - \min_{\mathcal{H}}\Phi \leq \frac{\mathcal{E}_{x^*,0}}{t_k^2}, \qquad \|\text{prox}_{\mu_k\Phi}x_k - x_k\| \leq \sqrt{\frac{2\mu_k\mathcal{E}_{x^*,0}}{t_k^2}},$$

$$\text{and} \quad \|\nabla\Phi_{(\mu_k)}(x_k)\| \leq \sqrt{\frac{2\mathcal{E}_{x^*,0}}{\mu_k t_k^2}}.$$

The convergence rates are similar to [10, Theorem 2.5] (our assumptions on the parameters are slightly different).

Remark 1.3 Take $x_{-1} = x_0$. Then $\mathcal{E}_{x^*,0} = t_0^2 \left(\Phi_{\lambda_0}(x_0) - \min_{\mathcal{H}} \Phi\right) + \frac{1}{2s_0}\|x^* - x_0\|^2$. From Theorem 1.2.1 we obtain

$$\Phi(\text{prox}_{\lambda_k \Phi} x_k) - \min_{\mathcal{H}} \Phi \leq \frac{1}{t_k^2}\left(t_0^2\left(\Phi_{\lambda_0}(x_0) - \min_{\mathcal{H}} \Phi\right) + \frac{1}{2s_0}\|x^* - x_0\|^2\right).$$

Since $\Phi_{\lambda_0}(x_0) \leq \Phi(x_0)$, the constant entering the above rate of convergence of the values is at least as good as in the classical inertial algorithms based on Nesterov's acceleration.

1.2.3 Faster Convergence

Theorem 1.2.2 *Under $(\mathcal{A})$, assume that the sequence (α_k) satisfies (K_1^+). Then, for any sequence (x_k) generated by the algorithm (RIPA), the following holds true:*

$$\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right) \quad \text{and} \quad \|x_k - x_{k-1}\| = o\left(\frac{s_k}{\sum_{i=1}^k t_i}\right)^{\frac{1}{2}} \quad \text{as } k \to +\infty, \tag{1.20}$$

thus implying

$$\Phi(\text{prox}_{\lambda_k \Phi} x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right)$$

$$\text{and} \quad \|x_k - \text{prox}_{\lambda_k \Phi} x_k\| = o\left(\frac{\mu_k}{\sum_{i=1}^k t_i}\right)^{\frac{1}{2}} \quad \text{as } k \to +\infty. \tag{1.21}$$

As a consequence, $\lim_{k\to+\infty} E_k = 0$. *Moreover, we have*

$$\Phi(\text{prox}_{\lambda_k \Phi} x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{t_k^2}\right), \quad \|x_k - \text{prox}_{\lambda_k \Phi} x_k\| = o\left(\frac{\sqrt{\mu_k}}{t_k}\right)$$

$$\text{and} \quad \|x_k - x_{k-1}\| = o\left(\frac{\sqrt{s_k}}{t_k}\right) \quad \text{as } k \to +\infty.$$

Proof In (1.7), we saw that (K_1^+) gives $\delta_k \geq (1-m)t_{k+1}$. According to this property, Theorem 1.2.1(ii) implies

$$\sum_{l=0}^{\infty} t_{l+1}(\Phi_{\lambda_l}(x_l) - \min_{\mathcal{H}} \Phi) < +\infty. \tag{1.22}$$

Thus we may add $\sum_{l=k}^{\infty} t_{l+1}(\Phi_{\lambda_l}(x_l) - \min_{\mathcal{H}} \Phi)$ to either side of the inequality of Lemma 1.4(iii); which gives

$$E_{k+1} + \sum_{l=k+1}^{\infty} t_{l+1}\left(\Phi_{\lambda_l}(x_l) - \min_{\mathcal{H}} \Phi\right) + \frac{t_k}{2s_k}\|x_k - x_{k-1}\|^2 \leq E_k + \sum_{l=k}^{\infty} t_{l+1}\left(\Phi_{\lambda_l}(x_l) - \min_{\mathcal{H}} \Phi\right).$$

Therefore the positive sequence $k \to E_k + \sum_{l=k}^{\infty} t_{l+1}\left(\Phi_{\lambda_l}(x_l) - \min_{\mathcal{H}} \Phi\right)$ is decreasing. As a consequence,

$$\lim_{k\to+\infty} E_k \text{ exists} \tag{1.23}$$

$$\sum_{k=0}^{\infty} \frac{t_k}{2s_k}\|x_k - x_{k-1}\|^2 < +\infty. \tag{1.24}$$

Let us now observe that assumption (K_1) implies $t_{k+1} - t_k \leq \frac{t_{k+1}}{t_{k+1}+t_k} \leq 1$. Since $t_k \geq 1$, we deduce that $t_{k+1} \leq 2t_k$ for every $k \geq 1$. Then estimate (1.24) yields

$$\sum_{k=1}^{+\infty} \frac{t_{k+1}}{2s_k}\|x_k - x_{k-1}\|^2 < +\infty.$$

Recall from (1.22) that $\sum_{k=0}^{\infty} t_{k+1}(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi) < +\infty$. Adding the two last inequalities, and by definition of W_k, we obtain $\sum_{k=0}^{+\infty} t_{k+1} W_k < +\infty$. On the other hand, it follows from Lemma 1.4 (iii) and $0 \leq \alpha_k \leq 1$ that the sequence (W_k) is nonincreasing. Hence $\sum_{k=0}^{+\infty} t_{k+1} W_{k+1} < +\infty$, which gives

$$\sum_{k=1}^{+\infty} t_k W_k < +\infty.$$

We follow now the same argument as in [10], and apply Lemma 1.10 in the Appendix, with the sequences (t_k) and (W_k), respectively, in place of (τ_k) and (ε_k). We obtain that

$$W_k = o\left(\frac{1}{\sum_{i=1}^{k} t_i}\right) \quad \text{as } k \to +\infty.$$

The estimates in (1.20) follow immediately. From the definition of Φ_{μ_k}, we easily deduce (1.21). In view of assumption (K_1), we have $t_{i+1}^2 - t_i^2 \leq t_{i+1}$ for every $i \geq 1$, hence by summing from $i = 1$ to $k - 1$

$$t_k^2 \leq t_1^2 + \sum_{i=1}^{k-1} t_{i+1} = t_1^2 - t_1 + \sum_{i=1}^{k} t_i,$$

and the last claims follow.

Remark 1.4 In the line of Remark 1.2, we also have $\Phi_{\mu_k}(x_k) - \min_{\mathcal{H}} \Phi = o\left(t_k^{-2}\right)$, $\Phi(\text{prox}_{\mu_k \Phi} x_k) - \min_{\mathcal{H}} \Phi = o\left(t_k^{-2}\right)$, $\|\text{prox}_{\mu_k \Phi} x_k - x_k\| = o\left(\sqrt{\lambda_k} t_k^{-1}\right)$, and $\|\nabla \Phi_{\mu_k}(x_k)\| = o\left(\lambda_k^{-1/2} t_k^{-1}\right)$.

1.2.4 Convergence of the Iterates

Theorem 1.2.3 *Under* $(\mathcal{A})$*, assume that the sequence* (α_k) *satisfies* (K_1^+)*. Suppose, further, that* $\sup_{k\geq 0} \lambda_k t_k^{-2} < +\infty$ *and* $\sup_k s_k < +\infty$*. Then, the sequences* (x_k)*,* $(\text{prox}_{\lambda_k \Phi} x_k)$*,* $(\text{prox}_{(\lambda_k + s_k)\Phi} x_k)$ *and* (y_k) *converge weakly to the same minimizer of* Φ*.*

Proof We use Opial's Lemma (see Appendix Lemma 1.8). Let x^* be a minimum point of Φ. By Theorems 1.2.1(i) and 1.2.2 we know that $\mathcal{E}_{x^*,k}$ has a limit, and E_k converges to 0, as $k \to +\infty$. This last property is equivalent to $t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi\right) \to 0$ and $\frac{t_k}{\sqrt{2s_k}} \|x_{k-1} - x_k\| \to 0$. Let us rewrite $\mathcal{E}_{x^*,k}$ as

$$\mathcal{E}_{x^*,k} = t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi\right) + \|\frac{1}{\sqrt{2s_k}} \left(x^* - x_{k-1}\right) + \frac{t_k}{\sqrt{2s_k}} (x_{k-1} - x_k)\|^2.$$

From the convergence of $\mathcal{E}_{x^*,k}$ and the convergence to zero of the above-mentioned sequences, we deduce that

$$\lim_{k\to+\infty} \mathcal{E}_{x^*,k} = \lim_{k\to+\infty} \left\| \frac{1}{\sqrt{2s_k}} \left(x^* - x_{k-1}\right) \right\|^2. \tag{1.25}$$

The sequence (s_k) has been supposedly nondecreasing and bounded from above. Hence it converges to some positive real, which by (1.25) implies that $\|x^* - x_{k-1}\|$ has a limit, as $k \to +\infty$. This gives the first hypothesis in Opial's Lemma.

Let $\bar{x}$ be a weak cluster point of $(x_k)_{k\in\mathbb{N}}$. By Theorem 1.2.2 and the assumption $\sup_{k\geq 0} \lambda_k t_k^{-2} < +\infty$, $\bar{x}$ is also a weak cluster point of $(\text{prox}_{\lambda_k \Phi} x_k)_{k\in\mathbb{N}}$. Now, Theorem 1.2.2 also implies, along with the weak lower semicontinuity of Φ, that $\bar{x}$ is a minimum point of Φ. This gives the second hypothesis in Opial's Lemma, and we conclude that x_k and $\text{prox}_{\lambda_k \Phi} x_k$ converge weakly to $\bar{x}$. Remark 1.4 shows that $\text{prox}_{(\lambda_k + s)\Phi} x_k$ also converges weakly to $\bar{x}$. Lastly, we have $y_k - x_k =$

$\frac{t_k-1}{t_{k+1}t_k}t_k(x_k - x_{k-1})$, where $\left|\frac{t_k-1}{t_{k+1}t_k}\right| < 1$ (an immediate consequence of $t_{k+1} \geq 1$), and $t_k(x_k - x_{k-1}) \to 0$, as $k \to +\infty$. Hence y_k converges weakly to $\bar{x}$ as well.

1.3 Comparison of the Various Approaches

1.3.1 The Case $\rho_k \equiv 1$

The approach developed in this paper contains, in particular, the classical Inertial Proximal Algorithm (IPA). Indeed, when taking $\rho_k \equiv 1$, (RIPA) reduces to

$$\text{(IPA)} \qquad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{\mu_k \Phi}(y_k), \end{cases}$$

which is an inertial proximal algorithm without relaxation. A rich literature has been devoted to this class of algorithms in recent years, see [1, 7, 15, 18, 21, 26, 27, 30, 31, 34, 37, 38]. The new aspects of the algorithm (IPA) are the general inertial coefficients (α_k), and the general proximal parameters (μ_k), which can be interpreted as variable step sizes. With this choice of $\rho_k \equiv 1$, we have $\lambda_k = (1 - \rho_k)\mu_k = 0$. As explained in the introduction, all the calculations developed in the previous section can be conducted with $\Phi_0 = \Phi$. Thus, Theorems 1.2.2 and 1.2.3 specialize to give:

Theorem 1.3.1 *Under $(\mathcal{A})$, assume that the sequence (α_k) satisfies (K_1^+). Suppose also that (μ_k) is nondecreasing. Then for any sequence (x_k) generated by the algorithm (IPA), we have*

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right) \quad \text{and} \quad \|x_k - x_{k-1}\| = o\left(\frac{\mu_k}{\sum_{i=1}^k t_i}\right)^{\frac{1}{2}} \quad \text{as } k \to +\infty, \tag{1.26}$$

Moreover, if $\sup_k \mu_k < +\infty$, then (x_k) converges weakly to a minimizer of Φ.

The gradient approach $(\text{RIPA})_2$ gives the convergence rate $\Phi(\operatorname{prox}_{\mu_k \Phi}(x_k)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right)$. It is expressed with the Moreau envelopes, which is close but different from (1.26).

1.3.2 Proximal Versus Gradient Approach

The following table (Figure 1.1) shows the hypotheses and convergence results for the Gradient formulation of (RIPA) (left column) and Proximal formulation (right

Formulation	Gradient: $x_{k+1} = y_k - \rho_k \mu_k \nabla \Phi_{\mu_k}(y_k)$	Proximal: $x_{k+1} = \mathrm{prox}_{\rho_k \mu_k \Phi_{(1-\rho_k)\mu_k}}(y_k)$
Condition for fast minimization	(μ_k) and $(\rho_k \mu_k)$ nondecreasing	$((1-\rho_k)\mu_k)$ and $(\rho_k \mu_k)$ nondecreasing
Convergence rate I	$\Phi_{\mu_k}(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right)$	$\Phi_{(1-\rho_k)\mu_k}(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right)$
Convergence rate II	$\Phi(\mathrm{prox}_{\mu_k \Phi}(x_k)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right)$	$\Phi(\mathrm{prox}_{(1-\rho_k)\mu_k \Phi}(x_k)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\sum_{i=1}^k t_i}\right)$
Convergence rate III	$\|x_k - x_{k-1}\| = o\left(\frac{\sqrt{s_k}}{t_k}\right)$	$\|x_k - x_{k-1}\| = o\left(\frac{\sqrt{\rho_k \mu_k}}{t_k}\right)$
Condition for convergence of (x_k)	$\sup_k \mu_k t_k^{-2} < +\infty$, $\sup_k \rho_k \mu_k < +\infty$	$\sup_k (1-\rho_k)\mu_k t_k^{-2} < +\infty$, $\sup_k \rho_k \mu_k < +\infty$

Fig. 1.1 Gradient versus proximal approach to (RIPA)

column). In both cases, $\alpha_k \in [0, 1]$ for every $k \geq 1$, (α_k) is a sequence in $[0, 1]$ that satisfies (K_0)-(K_1^+), and (ρ_k) is a sequence in $(0, 1]$. Clear similarities between them suggest finding a unifying approach, an interesting topic for future research.

As we already noticed in Remark 1.2 the convergence rates of the values obtained by the proximal approach are better than those for the gradient approach, since

$$\Phi_{\mu_k}(x_k) - \min_{\mathcal{H}} \Phi = \Phi_{\lambda_k + s_k}(x_k) - \min_{\mathcal{H}} \Phi \leq \Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi = \Phi_{(1-\rho_k)\mu_k}(x_k) - \min_{\mathcal{H}} \Phi,$$

and the inequality is strict unless x_k minimizes Φ.

1.3.3 Link with the General Maximally Monotone Case

As we already stressed in the introduction, another important motivation is to put to the fore inertial proximal algorithms that converge for general monotone inclusions, and which, in the case of convex minimization, give fast convergence rates of the values in the worst case. For a general maximally monotone operator $A : \mathcal{H} \to 2^{\mathcal{H}}$, (RIPA) is defined by, for $k \geq 1$

$$\text{(RIPA)} \quad \begin{cases} y_k = x_k + \alpha_k (x_k - x_{k-1}) \\ x_{k+1} = (1 - \rho_k) y_k + \rho_k J_{\mu_k A}(y_k) \end{cases}$$

(see [16]). In the above formula, $J_{\mu A} = (I + \mu A)^{-1}$ is the *resolvent* of A with index $\mu > 0$. When $A = \partial \Phi$ where $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower-semicontinuous proper function, $J_{\mu A} = \mathrm{prox}_{\mu \Phi}$, and we recover the algorithm studied in this paper. The above algorithm was introduced by Attouch-Peypouquet [16], in the case $\alpha_k = 1 - \frac{\alpha}{k}$, and further extended by Attouch-Cabot [9] in the case of a general inertial (extrapolation) coefficient α_k.

Because of its numerical importance, let us discuss the case $\alpha_k = 1 - \frac{\alpha}{k}$. Combining the results obtained here with the ones in [16], we deduce a set of assumptions for which iterate convergence and fast minimization both hold.

Corollary 1.3.2 *For each $k \geq 1$, take $\alpha_k = 1 - \dfrac{\alpha}{k}$ with $\alpha \geq 3$, $\rho_k = \dfrac{\beta}{k^2}$ with $\beta < \alpha(\alpha - 2)$, and $\mu_k = c\, k^{r'}$ for some $r' \geq 2$ and $c > 0$. Let (x_k) be generated by (RIPA). Then,*

a) If A is a maximally monotone operator, then x_k converges weakly to a zero of A. Moreover, $\|x_{k+1} - x_k\| = \mathcal{O}\left(\frac{1}{k}\right)$.

b) Let $A = \partial\Phi$ and $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is convex, lower-semicontinuous, proper. Then, $\Phi(p_k) - \min \Phi = \mathcal{O}(\frac{1}{k^2})$, where $p_k = \operatorname{prox}_{(1-\rho_k)\mu_k \Phi}(x_k)$.

1.4 The Impact of Geometry on the Rates of Convergence

To assume only the convexity of Φ is not sufficient to guarantee the strong convergence of the iterates generated by the proximal algorithm. This is known since the seminal work of Baillon [19] on the continuous steepest descent, and the parallel study on the proximal algorithm developed by Güler [28]. However, under some additional geometric assumptions on Φ, one can obtain the strong convergence property, and improve the convergence rate compared to the worst-case estimates. An important case is when the function Φ has the strong minimum property. This has been investigated in the case of classical proximal-gradient algorithms with general damping coefficients in [7]. We will develop similar results for (RIPA).

We say Φ has a strong minimum at $x^* \in \mathcal{H}$ if there is $\beta > 0$ such that

$$\Phi(x) \geq \Phi(x^*) + \frac{\beta}{2}\|x - x^*\|^2 \tag{1.27}$$

for all $x \in \mathcal{H}$. This clearly implies that $\arg\min \Phi = \{x^*\}$. Under this condition, we will determine the decay rate of the energy sequence (W_k) associated with the algorithm (RIPA). It is easy to see that

$$\Phi_\lambda(x) \geq \min \Phi + \frac{\beta}{2(1 + \beta\lambda)}\|x - x^*\|^2. \tag{1.28}$$

for all $x \in \mathcal{H}$. Since $\min \Phi = \min \Phi_\lambda$, we deduce that Φ_λ has still a strong minimum at x^*. The first results concerning the convergence rate of inertial algorithms of (FISTA) type for strongly convex minimization problems were obtained in [13, 38]. The following result is largely inspired by the techniques developed in [7, Theorem 11] for inertial proximal-gradient algorithms with general damping coefficient. Its adaptation to our setting is not immediate, because we have to deal with several variable data, which makes the proof rather technical. For simplicity, we assume that $s_k \equiv s$.

Theorem 1.4.1 *Let Φ admit a strong minimum x^*, in the sense of (1.27). Under $(\mathcal{A})$, suppose that the sequence $(1-\alpha_k)$ is nonincreasing, converges to 0, and satisfies $\sum_{k=1}^{+\infty}(1-\alpha_k) = +\infty$. Suppose moreover that $\sup_k \mu_k > 0$ and $s_k \equiv s$.*

Let (x_k) be a sequence generated by (RIPA), and let (W_k) be the associated energy sequence. The following holds:

(i) If $\alpha_{k+1} - \alpha_k = o(1-\alpha_k)$ as $k \to +\infty$, then for any $m \in]0, 2/3[$, we have, for k large enough,

$$W_k = \mathcal{O}\left(e^{-m\sum_{i=1}^{k}(1-\alpha_i)}\right) \tag{1.29}$$

Equivalently, $\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(e^{-m\sum_{i=1}^{k}(1-\alpha_i)}\right)$ and $\|x_k - x_{k-1}\|^2 = \mathcal{O}\left(e^{-m\sum_{i=1}^{k}(1-\alpha_i)}\right)$.
Therefore, setting $p_k = \operatorname{prox}_{\lambda_k \Phi} x_k$, we have

$$\Phi(p_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(e^{-m\sum_{i=1}^{k}(1-\alpha_i)}\right), \quad \|p_k - x_k\|^2 = \mathcal{O}\left(e^{-m\sum_{i=1}^{k}(1-\alpha_i)}\right)$$
$$\text{and} \quad \|p_k - x^*\|^2 = \mathcal{O}\left(e^{-m\sum_{i=1}^{k}(1-\alpha_i)}\right)$$

As a consequence, the iterates (x_k) and (p_k) converge strongly to the unique minimizer x^*.

(ii) If $\sum_{k=1}^{+\infty}(1-\alpha_k)^2 < +\infty$, then the estimates of (i) are satisfied for $m = 2/3$.

Proof By Lemma 1.5, the sequence (h_k) defined by $h_k = \frac{1}{2}\|x_k - x^*\|^2$ satisfies for every $k \geq 1$,

$$\begin{aligned} h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) &\leq \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 - s_k(\Phi_{\lambda_k}(x_{k+1}) - \min \Phi).) \\ &\leq \|x_k - x_{k-1}\|^2 - s_k(\Phi_{\lambda_k}(x_{k+1}) - \min \Phi)) \\ &\leq \|x_k - x_{k-1}\|^2 - s_k(\Phi_{\lambda_{k+1}}(x_{k+1}) - \min \Phi)) \end{aligned}$$

where the last inequality comes from $k \mapsto \lambda_k$ nondecreasing. From the definition of (W_k), we have

$$s_k W_{k+1} = s_k\left(\Phi_{\lambda_{k+1}}(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) + \frac{s_k}{2s_{k+1}}\|x_{k+1} - x_k\|^2.$$

Combining the two above expressions we obtain

$$h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) + s_k W_{k+1} \leq \|x_k - x_{k-1}\|^2 + \frac{s_k}{2s_{k+1}}\|x_{k+1} - x_k\|^2. \tag{1.30}$$

Recalling the expression of the decay of (W_k) given in Lemma 1.4 (iii), we have

$$\|x_k - x_{k-1}\|^2 \leq \frac{2s_k}{1-\alpha_k^2}(W_k - W_{k+1}) \quad \text{and} \quad \|x_{k+1} - x_k\|^2 \leq \frac{2s_{k+1}}{1-\alpha_{k+1}^2}(W_{k+1} - W_{k+2}).$$

Multiplying inequality (1.30) by $1 - \alpha_{k+1}^2$ and using that $\alpha_k \leq \alpha_{k+1}$, we obtain

$$\begin{aligned}(1 - \alpha_{k+1}^2)[h_{k+1} - h_k - \alpha_k(h_k - h_{k-1})] + s_k(1 - \alpha_{k+1}^2)W_{k+1} \\ \leq 2s_k(W_k - W_{k+1}) + s_k(W_{k+1} - W_{k+2}).\end{aligned} \tag{1.31}$$

Let us now define the sequence $(\widehat{W}_k)$ by $\widehat{W}_k = \frac{2}{3}W_k + \frac{1}{3}W_{k+1}$. Since the sequence (W_k) is nonnegative and nonincreasing, we have

$$\frac{2}{3}W_k \leq \widehat{W}_k \leq W_k. \tag{1.32}$$

Using the definition of $(\widehat{W}_k)$, elementary computation gives

$$\begin{aligned}2(W_k - W_{k+1}) + (W_{k+1} - W_{k+2}) = 2(W_k - W_{k+1}) + W_{k+1} - (3\widehat{W}_{k+1} - 2W_{k+1}) = \\ 2W_k + W_{k+1} - 3\widehat{W}_{k+1} = 3(\widehat{W}_k - \widehat{W}_{k+1}).\end{aligned}$$

We deduce from (1.31) that for every $k \geq 1$,

$$(1-\alpha_{k+1}^2)[h_{k+1} - h_k - \alpha_k(h_k - h_{k-1})] + s_k\left[(1 - \alpha_{k+1}^2)\widehat{W}_{k+1} + 3(\widehat{W}_{k+1} - \widehat{W}_k)\right] \leq 0. \tag{1.33}$$

Now observe that

$$\begin{aligned}&(1 - \alpha_{k+1}^2)[h_{k+1} - h_k - \alpha_k(h_k - h_{k-1})] \qquad (1.34)\\ &= (1 - \alpha_{k+1}^2)(h_{k+1} - h_k) - (1 - \alpha_{k+2}^2)\alpha_{k+1}(h_{k+1} - h_k) \\ &\quad + (1 - \alpha_{k+2}^2)\alpha_{k+1}(h_{k+1} - h_k) - (1 - \alpha_{k+1}^2)\alpha_k(h_k - h_{k-1}) \\ &= \big(1 - \alpha_{k+1}^2 - (1 - \alpha_{k+2}^2)\alpha_{k+1}\big)(h_{k+1} - h_k) \qquad (1.35)\\ &\quad + (1 - \alpha_{k+2}^2)\alpha_{k+1}(h_{k+1} - h_k) - (1 - \alpha_{k+1}^2)\alpha_k(h_k - h_{k-1}).\end{aligned}$$

At this point, it is simpler to assume that $s_k \equiv s$ is constant. Let us introduce the sequence $(\widetilde{W}_k)$ given by

$$\widetilde{W}_k = \widehat{W}_k + \frac{1}{3s}(1 - \alpha_{k+1}^2)\alpha_k(h_k - h_{k-1}). \tag{1.36}$$

Dividing inequality (1.33) by $3s$ and using (1.35), we infer that

$$\frac{1}{3}(1-\alpha_{k+1}^2)\widehat{W}_{k+1}+\widetilde{W}_{k+1}-\widetilde{W}_k \le \frac{1}{3s}\left[(1-\alpha_{k+2}^2)\alpha_{k+1}-(1-\alpha_{k+1}^2)\right](h_{k+1}-h_k),$$

which can be written, only in terms of the sequence $(\widetilde{W}_k)$, as

$$\frac{1}{3}(1-\alpha_{k+1}^2)\widetilde{W}_{k+1}+\widetilde{W}_{k+1}-\widetilde{W}_k \tag{1.37}$$
$$\frac{1}{3s}\left[\frac{1}{3}(1-\alpha_{k+1}^2)(1-\alpha_{k+2}^2)\alpha_{k+1}+\le(1-\alpha_{k+2}^2)\alpha_{k+1}-(1-\alpha_{k+1}^2)\right](h_{k+1}-h_k).$$

Using that the sequence (α_k) satisfies $\alpha_k \in [0,1]$ for every $k \ge 1$, we have

$$0 \le (1-\alpha_{k+1}^2)(1-\alpha_{k+2}^2)\alpha_{k+1} \le (1-\alpha_{k+1}^2)^2\alpha_{k+1} \le 4(1-\alpha_{k+1})^2.$$

Using, moreover, that the sequence (α_k) is nondecreasing, we have

$$\begin{aligned}0 \le (1-\alpha_{k+1}^2)-(1-\alpha_{k+2}^2)\alpha_{k+1} &= (1-\alpha_{k+1}^2)(1-\alpha_{k+1})+(\alpha_{k+2}^2-\alpha_{k+1}^2)\alpha_{k+1}\\ &\le 2(1-\alpha_{k+1})^2+2(\alpha_{k+2}-\alpha_{k+1}).\end{aligned}$$

It ensures that the term between brackets in (1.37) is comprised between $-2(1-\alpha_{k+1})^2-2(\alpha_{k+2}-\alpha_{k+1})$ and $\frac{4}{3}(1-\alpha_{k+1})^2$. This implies that its absolute value is majorized by $2(1-\alpha_{k+1})^2+2(\alpha_{k+2}-\alpha_{k+1})$. We then deduce from (1.37) that

$$\frac{1}{3}(1-\alpha_{k+1}^2)\widetilde{W}_{k+1}+\widetilde{W}_{k+1}-\widetilde{W}_k \le \frac{2}{3s}\left[(1-\alpha_{k+1})^2+(\alpha_{k+2}-\alpha_{k+1})\right]|h_{k+1}-h_k|. \tag{1.38}$$

Now observe that

$$\begin{aligned}h_{k+1}-h_k &= \frac{1}{2}\|x_{k+1}-x^*\|^2-\frac{1}{2}\|x_k-x^*\|^2\\ &= \langle x_{k+1}-x_k, x_{k+1}-x^*\rangle-\frac{1}{2}\|x_{k+1}-x_k\|^2.\end{aligned}$$

Since $|\langle x_{k+1}-x_k, x_{k+1}-x^*\rangle| \le \frac{1}{2}\|x_{k+1}-x_k\|^2+\frac{1}{2}\|x_{k+1}-x^*\|^2$, we infer that

$$|h_{k+1}-h_k| \le \|x_{k+1}-x_k\|^2+\frac{1}{2}\|x_{k+1}-x^*\|^2. \tag{1.39}$$

By the strong minimum property (1.28) we have

$$\frac{\beta}{2(1+\beta\lambda_{k+1})}\|x_{k+1}-x^*\|^2 \le \Phi_{\lambda_{k+1}}(x_{k+1})-\min\Phi.$$

Using this inequality in (1.39), together with the assumption $\sup_k \lambda_k < +\infty$, we find

$$\begin{aligned} |h_{k+1} - h_k| &\leq \|x_{k+1} - x_k\|^2 + \frac{1+\beta\lambda_{k+1}}{\beta}(\Phi_{\lambda_{k+1}}(x_{k+1}) - \min\Phi) \\ &\leq 2sC\, W_{k+1} \quad \text{with } C = \max\{1, \frac{1+\beta\sup_k \lambda_k}{2\beta s}\} \\ &\leq 3sC\, \widehat{W}_{k+1} \quad \text{in view of (1.32).} \end{aligned} \tag{1.40}$$

Since $\lim_{k\to+\infty} \alpha_k = 1$, the expression (1.36) of $\widetilde{W}_k$ and inequality (1.40) show that

$$\widetilde{W}_k = \widehat{W}_k + o(\widehat{W}_k) \quad \text{as } k \to +\infty. \tag{1.41}$$

Let $C' > C$. By combining (1.38), (1.40), and (1.41), we obtain the existence of $k_0 \geq 1$ such that for every $k \geq k_0$,

$$\frac{1}{3}(1-\alpha_{k+1}^2)\widetilde{W}_{k+1} + \widetilde{W}_{k+1} - \widetilde{W}_k \leq 2C'\,\widetilde{W}_{k+1}\Big[(1-\alpha_{k+1})^2 + (\alpha_{k+2} - \alpha_{k+1})\Big].$$

Noting that $1 - \alpha_{k+1}^2 = 2(1-\alpha_{k+1}) - (1-\alpha_{k+1})^2$, the above inequality can be rewritten as

$$\widetilde{W}_{k+1}\left(1 + \frac{2}{3}(1-\alpha_{k+1}) - u_{k+1}\right) \leq \widetilde{W}_k,$$

where the sequence (u_k) is defined by

$$u_k = \left(\frac{1}{3} + 2C'\right)(1-\alpha_k)^2 + 2C'(\alpha_{k+1} - \alpha_k).$$

Let $n \geq k_0 + 1$. By multiplying the inequalities above, as k ranges from k_0 to $n-1$, we obtain

$$\widetilde{W}_n \leq \frac{\widetilde{W}_{k_0}}{\prod_{k=k_0+1}^{n}\left(1 + \frac{2}{3}(1-\alpha_k) - u_k\right)} = \widetilde{W}_{k_0}\, e^{-\left[\sum_{k=k_0+1}^{n} \ln\left(1+\frac{2}{3}(1-\alpha_k)-u_k\right)\right]}. \tag{1.42}$$

(*i*) Let us now fix $m \in]0, 2/3[$ and assume that $\alpha_{k+1} - \alpha_k = o(1-\alpha_k)$ as $k \to +\infty$. Since $\lim_{k\to+\infty} \alpha_k = 1$, the expression of u_k shows that $u_k = o(1-\alpha_k)$ as $k \to +\infty$. It ensures that

$$\ln\left(1 + \frac{2}{3}(1-\alpha_k) - u_k\right) = \frac{2}{3}(1-\alpha_k) + o(1-\alpha_k) \quad \text{as } k \to +\infty,$$

and hence for k large enough,

$$\ln\left(1+\frac{2}{3}(1-\alpha_k)-u_k\right)\geq m\,(1-\alpha_k).$$

Coming back to (1.42), we easily deduce that

$$\widetilde{W}_n=\mathcal{O}\left(e^{-m\sum_{k=1}^{n}(1-\alpha_k)}\right)\quad\text{as } n\to+\infty.$$

In view of (1.32) and (1.41), we immediately derive the estimate (1.29). The other estimates follow directly.

(ii) Let us now assume that $\sum_{k=1}^{+\infty}(1-\alpha_k)^2<+\infty$. Observe that $\sum_{k=1}^{+\infty}(\alpha_{k+1}-\alpha_k)<+\infty$, resulting from the sequence (α_k) is nondecreasing and tends to 1 as $k\to+\infty$. The expression of (u_k) then shows that $\sum_{k=1}^{+\infty}u_k<+\infty$. Using that $\lim_{k\to+\infty}\alpha_k=1$ and $\lim_{k\to+\infty}u_k=0$, we have for k large enough,

$$\ln\left(1+\frac{2}{3}(1-\alpha_k)-u_k\right)=\frac{2}{3}(1-\alpha_k)-u_k-\frac{1}{2}\left[\frac{2}{3}(1-\alpha_k)-u_k\right]^2+o\left(\left[\frac{2}{3}(1-\alpha_k)-u_k\right]^2\right).$$

Since $\sum_{k=1}^{+\infty}(1-\alpha_k)^2<+\infty$ and $\sum_{k=1}^{+\infty}u_k^2<+\infty$ (recall that (u_k) is summable from what precedes), we obtain

$$\sum_{k=1}^{+\infty}\left[\frac{2}{3}(1-\alpha_k)-u_k\right]^2<+\infty.$$

Defining the sequence (v_k) by

$$v_k=\ln\left(1+\frac{2}{3}(1-\alpha_k)-u_k\right)-\frac{2}{3}(1-\alpha_k),$$

we deduce from what precedes that the series $\sum_k v_k$ is convergent. It ensues that

$$\begin{aligned}\sum_{k=k_0+1}^{n}\ln\left(1+\frac{2}{3}(1-\alpha_k)-u_k\right)&=\frac{2}{3}\sum_{k=k_0+1}^{n}(1-\alpha_k)+\sum_{k=k_0+1}^{n}v_k\\&=\frac{2}{3}\sum_{k=k_0+1}^{n}(1-\alpha_k)\\&\quad+\sum_{k=k_0+1}^{+\infty}v_k+o(1)\quad\text{as } n\to+\infty.\end{aligned}$$

Coming back to (1.42), we easily deduce that

$$\widetilde{W}_n = \mathcal{O}\left(e^{-\frac{2}{3}\sum_{k=1}^{n}(1-\alpha_k)}\right) \quad \text{as } n \to +\infty.$$

The conclusion then follows from (1.32) and (1.41).

Corollary 1.4.2 *Under* $(\mathcal{A})$*, assume that* $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ *admits a strong minimum* $x^* \in \mathcal{H}$.

(i) *If there exists* $\alpha > 0$ *such that* $\alpha_k = 1 - \frac{\alpha}{k}$ *for every* $k \geq 1$*, then*

$$W_k = \mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right) \quad \text{as } k \to +\infty.$$

(ii) *If there exist* $\alpha > 0$ *and* $r \in]1/2, 1[$ *such that* $\alpha_k = 1 - \frac{\alpha}{k^r}$ *for every* $k \geq 1$*, then*

$$W_k = \mathcal{O}\left(e^{-\frac{2\alpha}{3(1-r)}k^{1-r}}\right) \quad \text{as } k \to +\infty.$$

(iii) *If there exist* $\alpha > 0$ *and* $r \in]0, 1[$ *such that* $\alpha_k = 1 - \frac{\alpha}{k^r}$ *for every* $k \geq 1$*, then for any* $m \in]0, 2/3[$*, we have*

$$W_k = \mathcal{O}\left(e^{-\frac{m\alpha}{1-r}k^{1-r}}\right) \quad \text{as } k \to +\infty.$$

Proof

(i) The condition $\sum_{k=1}^{+\infty}(1-\alpha_k)^2 < +\infty$ is clearly satisfied. It is then sufficient to apply Theorem 1.4.1 (ii) and to recall that

$$\sum_{i=1}^{k}\frac{1}{i} = \ln k + \gamma + o(1) \quad \text{as } k \to +\infty,$$

for some $\gamma \in \mathbb{R}$ (Euler's constant).

(ii) The condition $\sum_{k=1}^{+\infty}(1-\alpha_k)^2 < +\infty$ is guaranteed by the assumption $r \in]1/2, 1[$. Then apply Theorem 1.4.1 (ii), together with the following asymptotic expansion

$$\sum_{i=1}^{k}\frac{1}{i^r} = \frac{k^{1-r}}{1-r} + l + o(1) \quad \text{as } k \to +\infty, \tag{1.43}$$

for some $l \in \mathbb{R}$.

(iii) The condition $\alpha_{k+1} - \alpha_k = o(1-\alpha_k)$ is satisfied as $k \to +\infty$, hence the announced estimate is a consequence of Theorem 1.4.1 (i), combined with the equality (1.43).

Remark 1.5 The strong minimum assumption (1.27) is a particular case of the Hölderian error bound inequality

$$\Theta(x) - \min \Theta \geq c \operatorname{dist}(x, \arg\min \Theta)^p.$$

In our case, we have $p = 2$, and the set of solutions is reduced to a single element. In the context of convex functions satisfying the above error bound inequality, this notion is equivalent to a Kurdyka-Lojasiewicz (KL) inequality. The connection between error bounds and Kurdyka-Lojasiewicz inequality was first established by Bolte, Daniilidis, Ley, and Mazet in [23]. The above equivalence was recently obtained in [24], which also provides a recent account on the rich interaction between these concepts. Indeed, the KL property and the corresponding desingularizing function play a central role in analyzing the convergence rate of first-order methods in nonsmooth structured optimization. Its applicability goes well beyond the convex case and the situation examined here, which suggests further developments.

1.5 Stability with Respect to Perturbations, Errors

Consider the perturbed version of the evolution equation (RIGS):

$$(\text{RIGS})_{\text{pert}} \quad \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla \Phi_{\lambda(t)}(x(t)) = g(t). \tag{1.44}$$

It will serve as a guide for the introduction of perturbations, errors in the algorithm (RIPA). The second member of (1.44), denoted by $g(\cdot)$, reflects an external action on the system (source term). We can also see (1.44) as a perturbation of the initial system (RIGS), or resulting from errors in the computation of the Moreau envelopes of Φ. We follow a parallel approach to the time discretization procedure described in Section 1.1.1.

Take a time step $h_k > 0$, and set $\tau_k = \sum_{i=1}^{k} h_i$, $x_k = x(\tau_k)$, $\lambda_k = \lambda(\tau_k)$, $\gamma_k = \gamma(\tau_k)$. An implicit finite-difference scheme for $(\text{RIGS})_{\text{pert}}$ with centered second-order variation gives

$$\frac{1}{h_k^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\gamma_k}{h_k}(x_k - x_{k-1}) + \nabla \Phi_{\lambda_k}(x_{k+1}) = g_k.$$

Equivalently, $x_{k+1} + h_k^2 \nabla \Phi_{\lambda_k}(x_{k+1}) = x_k + (1 - \gamma_k h_k)(x_k - x_{k-1}) + h_k^2 g_k$, which gives

$$x_{k+1} = \left(I + h_k^2 \nabla \Phi_{\lambda_k}\right)^{-1} \left(x_k + (1 - \gamma_k h_k)(x_k - x_{k-1}) + h_k^2 g_k\right). \tag{1.45}$$

Setting $s_k = h_k^2$, $\alpha_k = 1 - \gamma_k h_k$, and $y_k = x_k + \alpha_k(x_k - x_{k-1})$ (that's Nesterov extrapolation term), we get

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{s_k \Phi_{\lambda_k}}(y_k + s_k g_k). \end{cases}$$

Using the resolvent equation (or semi-group property) $(\Phi_{\lambda_k})_{s_k} = \Phi_{\lambda_k + s_k}$, we obtain the equivalent formulation

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \dfrac{\lambda_k}{\lambda_k + s_k}(y_k + s_k g_k) + \dfrac{s_k}{\lambda_k + s_k} \operatorname{prox}_{(\lambda_k + s_k)\Phi}(y_k + s_k g_k). \end{cases}$$

The change of parametrization $s_k = \rho_k \mu_k$, $\lambda_k = \mu_k(1 - \rho_k)$ gives the following equivalent form of the algorithm:

$$(\text{RIPA})_{pert} \qquad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = (1 - \rho_k) y_k + \rho_k \operatorname{prox}_{\mu_k \Phi}(y_k + s_k g_k). \end{cases}$$

It is a relaxed inertial proximal algorithm with perturbation, hence the terminology $(\text{RIPA})_{pert}$. When it comes to the numerical implementation of (RIPA), computational errors are unavoidable. The above approach allows us to consider that x_{k+1} no longer involves the proximal image of y_k, as required by (1.5), but the proximal image of a close point to y_k, namely $y_k + s_k g_k$. For the mathematical analysis, it is convenient to formulate $(\text{RIPA})_{pert}$ with the help of the sequence (t_k), which gives

$$\begin{cases} y_k = x_k + \dfrac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \\ x_{k+1} = \operatorname{prox}_{s_k \Phi_{\lambda_k}}(y_k + s_k g_k). \end{cases} \tag{1.46}$$

Recall that the sequence (t_k) is linked to the sequence (α_k) by (1.4) (see Section 1.1.2). To express that $y_k + s_k g_k$ is close to y_k (all the more as $k \to +\infty$), we will use the condition

$$\sum_{k=0}^{\infty} \sqrt{s_{k+1}} t_{k+1} \|g_k\| < +\infty. \tag{1.47}$$

Under this hypothesis, we will see that system (1.46) can be studied in much the same way as system (1.5), and retains the convergence properties of the latter. Our study bears a natural relation to [12] and [13], which deal with the simpler situation where the potential is fixed (without Moreau's envelopes), as well as the parameter $s_k \equiv s > 0$, and without relaxation. Note that when $s_k \equiv s > 0$, we recover the summability condition $\sum_{k=0}^{\infty} t_{k+1} \|g_k\| < +\infty$ introduced in the previous papers.

In the sequel, $(x_k)_{k\in\mathbb{N}}$, $(y_k)_{k\in\mathbb{N}}$ are sequences in $\mathcal{H}$ defined by x_{-1}, x_0 given and system (1.46) for $k \geq 0$. We will use the same energy functions $(E_k)_{k\in\mathbb{N}}$ and $(\mathcal{E}_{x^*,k})_{k\in\mathbb{N}}$ as in the unperturbed case: for $k \geq 0$

$$E_k = t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_k^2}{2s_k} \|x_{k-1} - x_k\|^2, \tag{1.48}$$

and

$$\mathcal{E}_{x^*,k} = t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{1}{2s_k} \|z_k - x^*\|^2,$$

where we set

$$z_k := x_k + (t_k - 1)(x_k - x_{k-1}).$$

Unlike the unperturbed case, where the sequence $(\mathcal{E}_{x^*,k})_{k\in\mathbb{N}}$ is nonincreasing, we will only prove that it is a convergent sequence. This will suffice to obtain similar conclusions. The following lemma contains the basic ingredients for the Lyapunov analysis of the perturbed algorithm $(\text{RIPA})_{pert}$.

Lemma 1.6 *Let $(\mathcal{A})$ hold.*

(i) *For any $x^* \in \arg\min \Phi$ and $k \geq 0$ we have*

$$\mathcal{E}_{x^*,k+1} + \delta_k \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_{k+1}^2}{2s_k} \|x_{k+1} - y_k\|^2 \leq \mathcal{E}_{x^*,k} + \langle z_{k+1} - x^*, t_{k+1} g_k \rangle. \tag{1.49}$$

(ii) *Assume, moreover, that condition (K_1) is satisfied. For each $k \geq 0$, we have*

$$E_{k+1} - t_{k+1} \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) + \frac{t_k}{2s_k} \|x_k - x_{k-1}\|^2 \leq E_k + \langle x_{k+1} - x_k, t_{k+1}^2 g_k \rangle.$$

(iii) *In addition, assume the summability condition (1.47) is satisfied. Then,*

$$\sup_k \frac{1}{\sqrt{s_k}} \|z_k - x^*\| < +\infty. \tag{1.50}$$

Proof

(i) Let us apply Lemma 1.3(ii) with $\varphi = \Phi_{\lambda_k}$, $x^* \in \arg\min \Phi_{\lambda_k}$, $t = t_{k+1}$, $s = s_k$, $x = x_k$ and $y = y_k + s_k g_k$. With this choice, (1.46) gives $p = \text{prox}_{s_k \Phi_{\lambda_k}}(y_k + s_k g_k) = x_{k+1}$; we obtain for any $k \geq 0$

$$t_{k+1}^2(\Phi_{\lambda_k}(x_{k+1}) - \Phi_{\lambda_k}(x^*)) + \frac{1}{2s_k}\|x^* - x_k + t_{k+1}(x_k - x_{k+1})\|^2$$
$$+ \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - (y_k + s_k g_k)\|^2$$
$$\leq (t_{k+1}^2 - t_{k+1})(\Phi_{\lambda_k}(x_k) - \Phi_{\lambda_k}(x^*)) + \frac{1}{2s_k}\|x^* - x_k + t_{k+1}(x_k - (y_k + s_k g_k))\|^2.$$

Introducing the variable z_k, and using the same arguments as in the proof of Lemma (1.4)(i), we obtain

$$\mathcal{E}_{x^*,k+1} + \delta_k(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi) + \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - (y_k + s_k g_k)\|^2$$
$$\leq t_k^2(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi) + \frac{1}{2s_k}\|x^* - z_k - t_{k+1}s_k g_k\|^2.$$

By developing the squares involving g_k, we get

$$\mathcal{E}_{x^*,k+1} + \delta_k(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi) + \frac{t_{k+1}^2}{2s_k}\|x_{k+1} - y_k\|^2$$
$$\leq \mathcal{E}_{x^*,k} + \langle z_k + t_{k+1}(x_{k+1} - y_k) - x^*, t_{k+1} g_k\rangle.$$

Now we have

$$\begin{aligned} z_k + t_{k+1}(x_{k+1} - y_k) &= x_k + (t_k - 1)(x_k - x_{k-1}) \\ &\quad + t_{k+1}\left(x_{k+1} - x_k - \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})\right) \\ &= x_{k+1} + (t_{k+1} - 1)(x_{k+1} - x_k) = z_{k+1}, \end{aligned}$$

which yields (1.49).

(ii) In inequality (1.11) of Lemma 1.3(i), set $\varphi = \Phi_{\lambda_k}$, $s = s_k$, $u = x_k$ and $y = y_k + s_k g_k$, hence $p = x_{k+1}$; we obtain for any $k \geq 0$

$$\Phi_{\lambda_k}(x_{k+1}) + \frac{1}{2s_k}\|x_k - x_{k+1}\|^2 + \frac{1}{2s_k}\|x_{k+1} - (y_k + s_k g_k)\|^2$$
$$\leq \Phi_{\lambda_k}(x_k) + \frac{1}{2s_k}\|x_k - (y_k + s_k g_k)\|^2.$$

Recalling $\Phi_{\lambda_{k+1}}(x_{k+1}) \leq \Phi_{\lambda_k}(x_{k+1})$, and expanding the squares involving g_k, we get

$$\Phi_{\lambda_{k+1}}(x_{k+1}) + \frac{1}{2s_k}\|x_k - x_{k+1}\|^2 + \frac{1}{2s_k}\|x_{k+1} - y_k\|^2 - \langle x_{k+1} - y_k, g_k\rangle \\ \leq \Phi_{\lambda_k}(x_k) + \frac{1}{2s_k}\|x_k - y_k\|^2 - \langle x_k - y_k, g_k\rangle.$$

Replacing y_k by its value $x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$ on the right-hand side, we obtain

$$\Phi_{\lambda_{k+1}}(x_{k+1}) + \frac{1}{2s_k}\|x_k - x_{k+1}\|^2 + \frac{1}{2s_k}\|x_{k+1} - y_k\|^2 \\ \leq \Phi_{\lambda_k}(x_k) + \frac{(t_k - 1)^2}{2s_k t_{k+1}^2}\|x_k - x_{k-1}\|^2 + \langle x_{k+1} - x_k, g_k\rangle.$$

We neglect $\frac{1}{2s_k}\|x_{k+1} - y_k\|^2$ on the left-hand side, then we subtract $\min_{\mathcal{H}} \Phi$ from each side, and multiply by t_{k+1}^2

$$E_{k+1} \leq t_{k+1}^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi\right) + \frac{t_k^2}{2s_k}\|x_k - x_{k-1}\|^2 \\ - \frac{2t_k - 1}{2s_k}\|x_k - x_{k-1}\|^2 + \langle x_{k+1} - x_k, t_{k+1}^2 g_k\rangle.$$

In view of $t_{k+1}^2 \leq t_{k+1} + t_k^2$ and $2t_k - 1 \geq t_k$, we obtain further

$$E_{k+1} \leq t_k^2 \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi\right) + \frac{t_k^2}{2s_k}\|x_k - x_{k-1}\|^2 \\ + t_{k+1}\left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi\right) - \frac{t_k}{2s_k}\|x_k - x_{k-1}\|^2 + \langle x_{k+1} - x_k, t_{k+1}^2 g_k\rangle,$$

which establishes the desired inequality.

(iii) From (1.49), we deduce for $k \geq 0$

$$\mathcal{E}_{x^*,k+1} \leq \mathcal{E}_{x^*,k} + \langle z_{k+1} - x^*, t_{k+1} g_k\rangle \leq \mathcal{E}_{x^*,0} + \sum_{l=0}^{k} \langle z_{l+1} - x^*, t_{l+1} g_l\rangle. \quad (1.51)$$

According to the definition of $\mathcal{E}_{x^*,k}$ we obtain

$$\frac{1}{2s_{k+1}}\|z_{k+1} - x^*\|^2 \leq \mathcal{E}_{x^*,0} + \sum_{l=0}^{k} \|z_{l+1} - x^*\| \, \|t_{l+1} g_l\|.$$

Equivalently, and after reindexing,

$$\left(\frac{1}{\sqrt{s_k}}\|z_k - x^*\|\right)^2 \leq 2\mathcal{E}_{x^*,0} + 2\sum_{l=1}^{k}\sqrt{s_l}t_l\|g_{l-1}\|\left(\frac{1}{\sqrt{s_l}}\|z_l - x^*\|\right).$$

Let us apply the Gronwall lemma 1.9 with $a_k = \frac{1}{\sqrt{s_k}}\|z_k - x^*\|$ and $\beta_j = 2\sqrt{s_j}t_j\|g_{j-1}\|$. We obtain

$$\frac{1}{\sqrt{s_k}}\|z_k - x^*\| \leq \sqrt{2\mathcal{E}_{x^*,0}} + 2\sum_{l=1}^{\infty}\sqrt{s_l}t_l\|g_{l-1}\| = \sqrt{2\mathcal{E}_{x^*,0}} + 2\sum_{l=0}^{\infty}\sqrt{s_{l+1}}t_{l+1}\|g_l\|. \tag{1.52}$$

From assumption (1.47), we deduce that

$$\sup_k \frac{1}{\sqrt{s_k}}\|z_k - x^*\| < +\infty.$$

The preceding lemma, a counterpart of Lemma 1.4, leads to the same convergence results as in Sections 1.2.2–1.2.4, with some obvious arrangements. For instance, we have the following perturbed version of Theorem 1.2.1:

Theorem 1.5.1 *Under* $(\mathcal{A})$, *assume that the sequence* (α_k) *satisfies* (K_1), *and the perturbation terms* (g_k) *satisfy the summability assumption* (1.47). *Then, the following properties hold:*

(i) *For any* $x^* \in \arg\min \Phi$, *the sequence* $(\mathcal{E}_{x^*,k})_{k\in\mathbb{N}}$ *converges. Setting*

$$C := \mathcal{E}_{x^*,0} + \left(\sqrt{2\mathcal{E}_{x^*,0}} + 2\sum_{l=0}^{\infty}\sqrt{s_{l+1}}t_{l+1}\|g_l\|\right)\sum_{l=0}^{\infty}\sqrt{s_{l+1}}t_{l+1}\|g_l\|$$

which is finite, we have: for all $k \geq 1$,

$$\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}}\Phi \leq \frac{C}{t_k^2}, \quad \Phi(\mathrm{prox}_{\lambda_k\Phi}x_k) - \min_{\mathcal{H}}\Phi \leq \frac{C}{t_k^2},$$

$$\|\mathrm{prox}_{\lambda_k\Phi}x_k - x_k\| \leq \sqrt{\frac{2C\lambda_k}{t_k^2}},$$

$$\|\nabla\Phi_{\lambda_k}(x_k)\| \leq \sqrt{\frac{2C}{\lambda_k t_k^2}}.$$

(ii) $$\sum_{k=0}^{\infty}\delta_k\left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}}\Phi\right) < +\infty.$$

(iii) $\sum_{k=0}^{\infty} t_{k+1}^2 s_k \|g_k - s_k \nabla \Phi_{\lambda_k+s_k}(y_k + s_k g_k)\|^2 < +\infty.$

(iv) $\|\nabla \Phi_{\lambda_k+s}(y_k)\| = o\left(\frac{1}{\sqrt{s_k} t_{k+1}}\right).$

Proof From (1.51) and (1.52), we have for $k \geq 0$

$$\begin{aligned}
\mathcal{E}_{x^*,k} &\leq \mathcal{E}_{x^*,0} + \sum_{l=0}^{k-1} \langle z_{l+1} - x^*, t_{l+1} g_l \rangle \\
&\leq \mathcal{E}_{x^*,0} + \sum_{l=0}^{k} \frac{1}{\sqrt{s_{l+1}}} \|z_{l+1} - x^*\| \sqrt{s_{l+1}} t_{l+1} \|g_l\| \\
&\leq \mathcal{E}_{x^*,0} + \left(\sqrt{2\mathcal{E}_{x^*,0}} + 2 \sum_{l=0}^{\infty} \sqrt{s_{l+1}} t_{l+1} \|g_l\| \right) \sum_{l=0}^{\infty} \sqrt{s_{l+1}} t_{l+1} \|g_l\| := C < +\infty.
\end{aligned}$$

Assertion (i) follows directly from the above majorization, and the definitions of $\mathcal{E}_{x^*,k}$, $\Phi_{\lambda_k}(x_k)$, $\nabla \Phi_{\lambda_k}(x_k)$;
From (1.49) we get

$$\mathcal{E}_{x^*,k+1} + \delta_k \left(\Phi_{\lambda_k}(x_k) - \min_{\mathcal{H}} \Phi \right) \leq \mathcal{E}_{x^*,k} + \frac{1}{\sqrt{s_{k+1}}} \|z_{k+1} - x^*\| \sqrt{s_{k+1}} t_{k+1} \|g_k\|. \tag{1.53}$$

Summing the above inequalities, and using (1.47), (1.50), we obtain assertion (ii).
From (1.49)

$$\mathcal{E}_{x^*,k+1} + \frac{t_{k+1}^2}{2s_k} \|x_{k+1} - y_k\|^2 \leq \mathcal{E}_{x^*,k} + \langle z_{k+1} - x^*, t_{k+1} g_k \rangle. \tag{1.54}$$

By a similar argument as above, and by summing the corresponding inequalities, we get

$$\sum_{k=0}^{\infty} \frac{t_{k+1}^2}{s_k} \|x_{k+1} - y_k\|^2 < +\infty. \tag{1.55}$$

Then, assertion (iii) follows from the above inequality and $x_{k+1} - y_k = s_k g_k - s_k \nabla \Phi_{\lambda_k+s_k}(y_k + s_k g_k)$.
Assertion (iv) follows from

$$s_k \nabla \Phi_{\lambda_k+s_k}(y_k) = s_k (\nabla \Phi_{\lambda_k+s_k}(y_k) - \nabla \Phi_{\lambda_k+s_k}(y_k + s_k g_k)) + s_k g_k - (x_{k+1} - y_k)$$

which implies

$$\sqrt{s_k}t_{k+1}\|\nabla\Phi_{\lambda_k+s}(y_k)\| \leq \sqrt{s_k}t_{k+1}\frac{s_k}{\lambda_k+s_k}\|g_k\|+\sqrt{s_k}t_{k+1}\|g_k\|+\frac{t_{k+1}}{s_k}\|x_{k+1}-y_k\|.$$

Inequalities (1.47) and (1.55) imply that the second member of the above inequality tends to zero as k goes to infinity, which gives the claim.

1.6 A Regularized Inertial Proximal-Gradient Algorithm

In many practical situations, the function $\Theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ to minimize is the sum of two convex functions $\Theta = \Phi + \Psi$, where $\Phi : \mathcal{H} \to \mathbb{R}$ is continuously differentiable, and $\Psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a lower-semicontinuous function whose proximal mapping is easy to calculate, at least approximatively. Hence the use of proximal-gradient methods, equivalently called forward-backward methods. An abundant literature has been devoted to this efficient and relatively easy to implement splitting method, see, for example, [7, 13, 15, 26, 38]. We will give an example of adaptation of (RIPA) to this additively structured situation. The general study is a subject to be studied further.

In order to develop relaxed inertial proximal-gradient algorithms, let us start from the continuous dynamic that served us as an introduction to (RIPA) in Section 1.1, namely

$$\ddot{x}(t)+\frac{\alpha}{t}\dot{x}(t)+\nabla\Theta_{\lambda(t)}(x(t))=0, \tag{1.56}$$

and which is associated to the minimization of the function $\Theta = \Phi + \Psi$. Equivalently,

$$\ddot{x}(t)+\frac{\alpha}{t}\dot{x}(t)+\frac{1}{\lambda(t)}\left(x(t)-\text{prox}_{\lambda(t)\Theta}\,(x(t))\right)=0. \tag{1.57}$$

Let us recall that

$$\text{prox}_{\lambda(t)\Theta}\,(x(t))=\arg\min_{\xi\in\mathcal{H}}\{\Phi(\xi)+\Psi(\xi)+\frac{1}{2\lambda(t)}\|\xi-x(t)\|^2\}. \tag{1.58}$$

To find an approximation $y(t)$ of $\text{prox}_{\lambda(t)\Theta}(x(t)$, we consider the inertial differential inclusion associated with the minimization problem (1.58),

$$\ddot{y}(t)+\frac{\alpha}{t}\dot{y}(t)+\frac{1}{\lambda(t)}\,(y(t)-x(t))+\nabla\Phi(y(t))+\partial\Psi(y(t))\ni 0. \tag{1.59}$$

For simplicity, take $\lambda(t) = \lambda$ constant. This argument leads us to considering the evolution of the coupled system

$$\begin{cases} \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \frac{1}{\lambda}(x(t) - y(t)) = 0; \\ \ddot{y}(t) + \frac{\alpha}{t}\dot{y}(t) + \frac{1}{\lambda}(y(t) - x(t)) + \nabla\Phi(y(t)) + \partial\Psi(y(t)) \ni 0, \end{cases} \tag{1.60}$$

as an approximation of (1.57). The fact that the coupling forces have opposite sign in the two above equations can be interpreted as an *action-reaction* principle. Another justification for (1.60) is the following: for $Z = (x, y) \in \mathcal{H} \times \mathcal{H}$ set

$$\Gamma(Z) := \Phi(y) + \Psi(y) + \frac{1}{2\lambda}\|x - y\|^2.$$

Then, with $Z(t) = (x(t), y(t))$, (1.60) is equivalent to the inertial differential inclusion in $\mathcal{H} \times \mathcal{H}$

$$\ddot{Z}(t) + \frac{\alpha}{t}\dot{Z}(t) + \partial\Gamma(Z(t)) \ni 0. \tag{1.61}$$

As a key property, we have

$$\inf_{Z \in \mathcal{H} \times \mathcal{H}} \Gamma(Z) = \inf_{y \in \mathcal{H}} \{\Phi(y) + \Psi(y)\}.$$

For $\alpha > 3$, the convergence analysis developed in [6, 13, 32] guarantees the rate of convergence for the values:

$$\Gamma(Z(t)) - \inf_{Z \in \mathcal{H} \times \mathcal{H}} \Gamma(Z) = o(\frac{1}{t^2}).$$

Equivalently

$$\Phi(y(t)) + \Psi(y(t)) + \frac{1}{2\lambda}\|x(t) - y(t)\|^2 - \inf_{\mathcal{H}}(\Phi + \Psi) = o(\frac{1}{t^2}),$$

which implies

$$(\Phi + \Psi)(y(t)) - \inf_{\mathcal{H}}(\Phi + \Psi) = o(\frac{1}{t^2}).$$

Let us now examine the algorithmic version of the above results. One possibility is to discretize with respect to the time variable t the equation (1.60). But, it is not clear how to discretize the coupling term, and at what point to take the gradient of Φ. Moreover, such an approach would require a complete analysis of the algorithm thus obtained. We choose a different and simpler approach, which consists in applying the existing convergence theory concerning the inertial proximal-gradient

algorithms with inertial coefficient $\alpha_k = 1 - \frac{\alpha}{k}$ to the structured potential function Γ in the product space $\mathcal{H} \times \mathcal{H}$. For some recent references to these algorithms, see [7, 13, 15, 26, 38].

a) We first examine the case where the quadratic coupling term is incorporated into the non-smooth part. Then,

$$\Gamma(Z) = \tilde{\Phi}(Z) + \tilde{\Psi}(Z)$$

where

$$\tilde{\Phi}(Z) = \Phi(y), \quad \text{and} \quad \tilde{\Psi}(Z) = \Psi(y) + \frac{1}{2\lambda}\|x - y\|^2.$$

Setting $Z_k = (x_k, y_k)$, the algorithm writes

$$\begin{cases} \Xi_k = Z_k + (1 - \frac{\alpha}{k})(Z_k - Z_{k-1}); \\ Z_{k+1} = \operatorname{prox}_{s\tilde{\Psi}}\left(\Xi_k - s\nabla\tilde{\Phi}\,(\Xi_k)\right). \end{cases} \tag{1.62}$$

Classical convex subdifferential calculus gives

$$\nabla\tilde{\Phi}(Z) = (0, \nabla\Phi(y)), \quad \partial\tilde{\Psi}(Z) = (\frac{1}{\lambda}(x - y),\ \partial\Psi(y) + \frac{1}{\lambda}(y - x)).$$

By definition $\operatorname{prox}_{s\tilde{\Psi}}(Z) := Z_s$ is the solution of $Z_s + s\partial\tilde{\Psi}(Z_s) \ni Z$. Setting $Z_s = (x_s, y_s)$ this amounts to solving

$$\begin{cases} x_s + \frac{s}{\lambda}(x_s - y_s) = x; \\ y_s + s\partial\Psi(y_s) + \frac{s}{\lambda}(y_s - x_s) \ni y. \end{cases} \tag{1.63}$$

Elementary computation gives

$$\operatorname{prox}_{s\tilde{\Psi}}(Z) = \left(\frac{\lambda}{\lambda + s}x + \frac{s}{\lambda + s}\operatorname{prox}_{\frac{s(\lambda+s)}{\lambda+2s}\Psi}\left(\frac{s}{\lambda + 2s}x + \frac{\lambda + s}{\lambda + 2s}y\right),\right.$$
$$\left.\operatorname{prox}_{\frac{s(\lambda+s)}{\lambda+2s}\Psi}\left(\frac{s}{\lambda + 2s}x + \frac{\lambda + s}{\lambda + 2s}y\right)\right).$$

Set $\Xi_k = (\xi_k, \eta_k)$, and $\gamma = \frac{s(\lambda+s)}{\lambda+2s}$. The above relations and (1.62) give the Regularized Inertial Proximal-Gradient Algorithm (RIPGA-1)

$$\text{(RIPGA-1)}\quad \begin{cases} \xi_k = x_k + (1-\frac{\alpha}{k})(x_k - x_{k-1}); \\ \eta_k = y_k + (1-\frac{\alpha}{k})(y_k - y_{k-1}); \\ y_{k+1} = \operatorname{prox}_{\gamma\Psi}\left(\frac{s}{\lambda+2s}\xi_k + \frac{\lambda+s}{\lambda+2s}(\eta_k - s\nabla\Phi(\eta_k))\right); \\ x_{k+1} = \frac{\lambda}{\lambda+s}\xi_k + \frac{s}{\lambda+s}y_{k+1}. \end{cases} \tag{1.64}$$

Then notice that $\nabla\tilde{\Phi}$ is Lipschitz continuous, with the Lipschitz constant L. We can now state our main result.

Theorem 1.6.1 *Let $\Psi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower-semicontinuous, and convex function. Let $\Phi : \mathcal{H} \to \mathbb{R}$ be a convex, continuously differentiable function, whose gradient has a Lipschitz constant L satisfying $sL \leq 1$. Suppose that $\arg\min(\Phi + \Psi) \neq \emptyset$. Let (x_k) and (y_k) be sequences generated by algorithm (RIPGA-1) with $\alpha > 3$. Then,*

(i) *x_k and y_k converge weakly, as $k \to +\infty$, to the same limit $x^* \in \arg\min(\Phi + \Psi)$.*

(ii) $(\Phi + \Psi)(y_k) - \min_{\mathcal{H}}(\Phi + \Psi) = o\left(\frac{1}{k^2}\right)$.

Proof The proof follows from a direct application of the convergence results concerning the inertial proximal-gradient algorithms with inertial coefficient $\alpha_k = 1 - \frac{\alpha}{k}$, see [7, 15]. Consider the structured potential function $\Gamma = \tilde{\Phi} + \tilde{\Psi}$ in the product space $\mathcal{H} \times \mathcal{H}$. Note that

$$\inf_{Z\in\mathcal{H}\times\mathcal{H}} \Gamma(Z) = \inf_{y\in\mathcal{H}}\{\Phi(y) + \Psi(y)\}.$$

Precisely, by applying of [15, Theorem 1], or [7, Corollary 17] to algorithm (1.62), for $\alpha > 3$, we obtain the rate of convergence for the values

$$\Gamma(Z_k) - \inf_{Z\in\mathcal{H}\times\mathcal{H}} \Gamma(Z) = o(\frac{1}{k^2}).$$

Equivalently,

$$\Phi(y_k) + \Psi(y_k) + \frac{1}{2\lambda}\|x_k - y_k\|^2 - \inf_{\mathcal{H}}(\Phi + \Psi) = o(\frac{1}{k^2}), \tag{1.65}$$

which implies

$$(\Phi + \Psi)(y_k) - \inf_{\mathcal{H}}(\Phi + \Psi) = o(\frac{1}{k^2}).$$

We also obtain the weak convergence of the sequence (Z_k) with $Z_k = (x_k, y_k)$ to a point of $\arg\min \Gamma = \arg\min(\Phi + \Psi) \times \arg\min(\Phi + \Psi)$. Equivalently x_k converges weakly, as $k \to +\infty$, to a point $x^* \in \arg\min(\Phi + \Psi)$, and y_k converges weakly, as $k \to +\infty$, to a point $y^* \in \arg\min(\Phi + \Psi)$. From (1.65), we have $\|x_k - y_k\| = o(\frac{1}{k})$, which clearly implies $x^* = y^*$, and completes the proof.

b) Let us now incorporate the quadratic coupling term in the smooth part. We have

$$\Gamma(Z) = \tilde{\Phi}(Z) + \tilde{\Psi}(Z)$$

where

$$\tilde{\Phi}(Z) = \Phi(y) + \frac{1}{2\lambda}\|x - y\|^2, \quad \text{and} \quad \tilde{\Psi}(Z) = \Psi(y).$$

Then, algorithm $(\mathrm{AVD})_\alpha$ writes, with $Z_k = (x_k, y_k)$:

$$\begin{cases} \Xi_k &= Z_k + (1 - \frac{\alpha}{k})(Z_k - Z_{k-1}); \\ Z_{k+1} &= \mathrm{prox}_{s\tilde{\Psi}}\left(\Xi_k - s\nabla\tilde{\Phi}\,(\Xi_k)\right). \end{cases} \tag{1.66}$$

Elementary calculations give

$$\nabla\tilde{\Phi}(Z) = \left(\frac{1}{\lambda}(x - y), \nabla\Phi(y) + \frac{1}{\lambda}(y - x)\right), \quad \mathrm{prox}_{s\tilde{\Psi}}(Z) = (x, \mathrm{prox}_{s\Psi}(y)).$$

Set $\Xi_k = (\xi_k, \eta_k)$. Thanks to the above relations, algorithm (1.62) writes

$$\text{(RIPGA-2)} \quad \begin{cases} \xi_k = x_k + (1 - \frac{\alpha}{k})(x_k - x_{k-1}); \\ \eta_k = y_k + (1 - \frac{\alpha}{k})(y_k - y_{k-1}); \\ x_{k+1} = \xi_k - \frac{s}{\lambda}(\xi_k - \eta_k); \\ y_{k+1} = \mathrm{prox}_{s\Psi}\left(\eta_k - s\nabla\Phi\,(\eta_k) - \frac{s}{\lambda}(\eta_k - \xi_k)\right). \end{cases}$$

Then notice that $\nabla\tilde{\Phi}$ is Lipschitz continuous, with the Lipschitz constant $L + \frac{2\sqrt{2}}{\lambda}$, where L is the Lipschitz constant of $\nabla\Phi$. Assuming $s(L + \frac{2\sqrt{2}}{\lambda}) \leq 1$, for $\alpha > 3$, by the same argument as in the above paragraph, one can easily deduce the convergence results:

i) The sequences (x_k) and (y_k) converge weakly, as $k \to +\infty$, to the same limit $x^* \in \arg\min(\Phi + \Psi)$.

ii) $(\Phi + \Psi)(y_k) - \min_{\mathcal{H}}(\Phi + \Psi) = o\left(\frac{1}{k^2}\right)$.

Remark 1.6 As we have already noted in the previous sections, the variables x_k and y_k do not play symmetrical roles. The variable for which we obtained a result of fast convergence of the values is y_k.

Remark 1.7 A major interest of the above algorithm is that, when the trajectory is bounded away from the diagonal $\Delta = \{(x, y) \in \mathcal{H} \times \mathcal{H} : x = y\}$, the function of the two variables Γ, which is to minimize, is strongly convex. This results in favorable numerical features, particularly when using a restart method, see [38].

Remark 1.8 In system (1.60), the two equations involve the same parameter α. Using different values would lead to studying inertial dynamics (with respect to Z) with a nonisotropic diagonal damping matrix. Some corresponding results for the heavy ball method can be found in [1]. This is an interesting topic for further research.

Remark 1.9 As noticed in [5], the above technique, which consists in introducing an auxiliary variable, is related to the inertial dynamics with Hessian driven damping. To give an idea in a simple case, consider the continuous steepest descent associated to the potential $\Gamma(x, y) = \Theta(y) + \frac{1}{2\lambda}\|x - y\|^2$, where Θ is a smooth convex function. It writes

$$\begin{cases} \dot{x}(t) + \frac{1}{\lambda}(x(t) - y(t)) = 0; \\ \dot{y}(t) + \nabla\Theta(y(t)) + \frac{1}{\lambda}(y(t) - x(t)) = 0. \end{cases} \tag{1.67}$$

Eliminating the auxiliary variable (here x) gives

$$\ddot{y}(t) + \frac{1}{2\lambda}\dot{y}(t) + \nabla^2\Theta(y(t))\dot{y}(t) + \frac{1}{\lambda}\nabla\Theta(y(t)) = 0;$$

It is the (DIN-AVD) dynamical system [2, 17], which combines inertial features with geometrical damping. Hessian damping is naturally linked to Newton's method, and is particularly convenient for poorly conditioned problems [17].

Appendix

Some Properties of the Moreau Envelope

For a detailed presentation of the Moreau envelope, we refer the reader to [20, 25, 35, 36]. We merely point out the following properties, of constant use here:

(i) the function $\lambda \in]0, +\infty[\mapsto \Phi_\lambda(x)$ is nonincreasing for each $x \in \mathcal{H}$;
(ii) the equality $\inf_{\mathcal{H}} \Phi = \inf_{\mathcal{H}} \Phi_\lambda$ holds in $\mathbb{R} \cup \{-\infty\}$ for all $\lambda > 0$;
(iii) $\arg\min \Phi = \arg\min \Phi_\lambda$ for all $\lambda > 0$.

It will be convenient to consider the Moreau envelope as a function of the two variables $x \in \mathcal{H}$ and $\lambda \in]0, +\infty[$. Its differentiability with respect to (x, λ) plays a crucial role in our analysis.

a

Let us first recall some classical facts concerning the differentiability of the function $x \mapsto \Phi_\lambda(x)$ for fixed $\lambda > 0$. The infimum in (1.2) is attained at a unique point

$$\text{prox}_{\lambda\Phi}(x) = \text{argmin}_{\xi\in\mathcal{H}} \left\{ \Phi(\xi) + \frac{1}{2\lambda}\|x - \xi\|^2 \right\}, \tag{1.68}$$

which gives

$$\Phi_\lambda(x) = \Phi(\text{prox}_{\lambda\Phi}(x)) + \frac{1}{2\lambda}\|x - \text{prox}_{\lambda\Phi}(x)\|^2. \tag{1.69}$$

Writing the optimality condition for (1.68), we get $\text{prox}_{\lambda\Phi}(x) + \lambda\partial\Phi\left(\text{prox}_{\lambda\Phi}(x)\right) \ni x$, that is

$$\text{prox}_{\lambda\Phi}(x) = (I + \lambda\partial\Phi)^{-1}(x).$$

Thus, $\text{prox}_{\lambda\Phi}$ is the resolvent of index $\lambda > 0$ of the maximal monotone operator $\partial\Phi$. As a consequence, the mapping $\text{prox}_{\lambda\Psi} : \mathcal{H} \to \mathcal{H}$ is firmly nonexpansive. The function $x \mapsto \Phi_\lambda(x)$ is continuously differentiable, with

$$\nabla\Phi_\lambda(x) = \frac{1}{\lambda}\left(x - \text{prox}_{\lambda\Phi}(x)\right). \tag{1.70}$$

Equivalently

$$\nabla\Phi_\lambda = \frac{1}{\lambda}\left(I - (I + \lambda\partial\Phi)^{-1}\right) = (\partial\Phi)_\lambda \tag{1.71}$$

which is the Yosida approximation of the maximal monotone operator $\partial\Phi$. As such, $\nabla\Phi_\lambda$ is Lipschitz continuous, with Lipschitz constant $\frac{1}{\lambda}$, and $\Phi_\lambda \in \mathcal{C}^{1,1}$.

b

A less known result is the $\mathcal{C}^1$-regularity of the function $\lambda \mapsto \Phi_\lambda(x)$, for each $x \in \mathcal{H}$. Its derivative is given by

$$\frac{d}{d\lambda}\Phi_\lambda(x) = -\frac{1}{2}\|\nabla\Phi_\lambda(x)\|^2. \tag{1.72}$$

This result is known as the Lax-Hopf formula for the above first-order Hamilton-Jacobi equation, see [4, Remark 3.2; Lemma 3.27], [8, Lemma A.1], and [29].

Lemma 1.7 *For each $x \in \mathcal{H}$, the real-valued function $\lambda \mapsto \Phi_\lambda(x)$ is continuously differentiable on $]0, +\infty[$, with*

$$\frac{d}{d\lambda}\Phi_\lambda(x) = -\frac{1}{2}\|\nabla\Phi_\lambda(x)\|^2. \tag{1.73}$$

As a consequence, for any $x \in \mathcal{H}$, $\lambda > 0$ and $\mu > 0$,

$$(\Phi_\lambda)_\mu(x) = \Phi_{(\lambda+\mu)}(x). \tag{1.74}$$

Indeed, (1.74) is the semi-group property satisfied by the orbits of the autonomous evolution equation (1.72). Differentiating (1.74) with respect to x, and using (1.71) gives the classical resolvent equation

$$(A_\lambda)_\mu = A_{(\lambda+\mu)}, \tag{1.75}$$

where $A = \partial\Phi$. Indeed, (1.75) is valid for a general maximally monotone operator A, see, for example, [20, Proposition 23.6] or [25, Proposition 2.6].

Auxiliary Results

Theorem 1.6.2 *Let $\lambda : [t_0, +\infty[\to]0, +\infty[$ be continuous and nondecreasing. Let $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be convex, lower-semicontinuous, and proper. Then, given any x_0 and v_0 in $\mathcal{H}$, system (1.1) has a unique twice continuously differentiable global solution $x : [t_0, +\infty[\to \mathcal{H}$ verifying $x(t_0) = x_0$, $\dot{x}(t_0) = v_0$.*

Proof The assertion appeals to the most elementary form of the Cauchy-Lipschitz theorem (see any textbook) and hinges on the (t, x)-continuity of $\nabla\Phi_\lambda$ and on its Lipschitz continuity with respect to x, uniform with respect to t.

Indeed, for $t \in [t_0, +\infty[$ and $(x, x') \in \mathcal{H} \times \mathcal{H}$ we have

$$\|\nabla\Phi_{\lambda(t)}(x') - \nabla\Phi_{\lambda(t)}(x)\| \le \frac{1}{\lambda(t)}\|x' - x\| \le \frac{1}{\lambda(t_0)}\|x' - x\|.$$

Next, the continuity of $\nabla\Phi_{\lambda(t)}(x) = \frac{1}{\lambda(t)}(x - \mathrm{prox}_{\lambda(t)\Phi}x)$ boils down to the continuity of the mapping $(t, x) \in [t_0, +\infty[\times\mathcal{H} \to \mathrm{prox}_{\lambda(t)\Phi}x \in \mathcal{H}$. For (t, x) and (t', x') in $[t_0, +\infty[\times\mathcal{H}$ we have

$$\|\mathrm{prox}_{\lambda(t')\Phi}x' - \mathrm{prox}_{\lambda(t)\Phi}x\| \le \|\mathrm{prox}_{\lambda(t')\Phi}x' - \mathrm{prox}_{\lambda(t')\Phi}x\| + \|\mathrm{prox}_{\lambda(t')\Phi}x - \mathrm{prox}_{\lambda(t)\Phi}x\|.$$

But, since $\text{prox}_{\lambda\Phi}$ is nonexpansive

$$\|\text{prox}_{\lambda(t')\Phi}x' - \text{prox}_{\lambda(t')\Phi}x\| \leq \|x' - x\|;$$

and also (see [20, Prop. 23.28(iii)])

$$\|\text{prox}_{\lambda(t')\Phi}x - \text{prox}_{\lambda(t)\Phi}x\| \leq \left|\frac{\lambda(t')}{\lambda(t)} - 1\right| \|\text{prox}_{\lambda(t)\Phi}x - x\|.$$

Therefore

$$\|\text{prox}_{\lambda(t')\Phi}x' - \text{prox}_{\lambda(t)\Phi}x\| \leq \|x' - x\| + |\lambda(t') - \lambda(t)|\|\nabla\Phi_{\lambda(t)}(x)\|,$$

which proves the continuity of $\text{prox}_{\lambda\Phi}$ at point (t, x).

Let us state the discrete version of Opial's lemma.

Lemma 1.8 *Let S be a nonempty subset of $\mathcal{H}$, and (x_k) a sequence of elements of $\mathcal{H}$. Assume that*

(i) for every $z \in S$, $\lim_{k\to+\infty}\|x_k - z\|$ exists;
(ii) every weak sequential cluster point of (x_k), as $k \to \infty$, belongs to S.

Then x_k converges weakly as $k \to \infty$ to a point in S.

We shall also make use of the following discrete version of the Gronwall lemma:

Lemma 1.9 *Let (a_k) be a sequence of nonnegative numbers such that, for all $k \in \mathbb{N}$*

$$a_k^2 \leq c^2 + \sum_{j=1}^{k} \beta_j a_j,$$

where (β_j) is a summable sequence of nonnegative numbers, and $c \geq 0$. Then, $a_k \leq c + \sum_{j=1}^{\infty} \beta_j$ for all $k \in \mathbb{N}$.

Proof For $k \in \mathbb{N}$, set $A_k := \max_{1\leq m\leq k} a_m$. Then, for $1 \leq m \leq k$, we have

$$a_m^2 \leq c^2 + \sum_{j=1}^{m} \beta_j a_j \leq c^2 + A_k \sum_{j=1}^{\infty} \beta_j.$$

Taking the maximum over $1 \leq m \leq k$, we obtain

$$A_k^2 \leq c^2 + A_k \sum_{j=1}^{\infty} \beta_j.$$

Bounding by the roots of the corresponding quadratic equation, we obtain the result.

The next lemma provides an estimate of the convergence rate of a sequence that is summable with respect to weights.

Lemma 1.10 ([7, Lemma 22]) *Let (τ_k) be a nonnegative sequence such that $\sum_{k=1}^{+\infty} \tau_k = +\infty$. Assume that (ϵ_k) is a nonnegative and nonincreasing sequence satisfying $\sum_{k=1}^{+\infty} \tau_k \epsilon_k < +\infty$. Then we have $\epsilon_k = o\left(\frac{1}{\sum_{i=1}^{k} \tau_i}\right)$ as $k \to +\infty$.*

Acknowledgements This work was supported by FONDECYT Grant 1181179 and CMM-Conicyt PIA AFB170001.

References

1. Álvarez, F.: On the minimizing property of a second-order dissipative system in Hilbert spaces. SIAM J. Control Optim. **38**, 1102–1119 (2000)
2. Álvarez, F., Attouch, H., Bolte, J., Redont, P.: A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics. J. Math. Pures Appl. **81**, 747–779 (2002)
3. Apidopoulos, V., Aujol, J.-F., Dossal, Ch.: Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule. Math. Prog. (Ser. A). , 1–20 (2018)
4. Attouch, H.: Variational Analysis for Functions and Operators. Pitman (1984)
5. Attouch, H., Bolte, J., Redont, P.: Optimizing properties of an inertial dynamical system with geometric damping. Link with proximal methods. Control Cybernet. **31**, 643–657 (2002)
6. Attouch, H., Cabot, A.: Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. J. Differential Equations **263**, 5412–5458 (2017)
7. Attouch, H., Cabot, A.: Convergence rates of inertial forward-backward algorithms. SIAM J. Optim. **28**, 849–874 (2018)
8. Attouch, H., Cabot, A.: Convergence of damped inertial dynamics governed by regularized maximally monotone operators. J. Differential Equations to appear. HAL-01648383v2 (2018)
9. Attouch, H., Cabot, A.: Convergence of a relaxed inertial proximal algorithm for maximally monotone operators. HAL-01708905 (2018)
10. Attouch, H., Cabot, A.: Convergence rate of a relaxed inertial proximal algorithm for convex minimization. HAL-01807041 (2018)
11. Attouch, H., Cabot, A., Redont, P.: The dynamics of elastic shocks via epigraphical regularization of a differential inclusion. Adv. Math. Sci. Appl. **12**, 273–306 (2002)
12. Attouch, H., Cabot, A., Chbani, Z., Riahi, H.: Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization. J. Optim. Th. Appl. **179**, 1–36 (2018)
13. Attouch, H., Chbani, Z., Peypouquet, J., Redont, P.: Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. Math. Prog. (Ser. B) **168**, 123–175 (2018)
14. Attouch, H., Chbani, Z., Riahi, H.: Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$, ESAIM: COCV **25** (2019)
15. Attouch, H., Peypouquet, J.: The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. SIAM J. Optim. **26**, 1824–1834 (2016)
16. Attouch, H., Peypouquet, J.: Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators. Math. Prog. **174**, 319–432 (2019)

17. Attouch, H., Peypouquet, J., Redont, P.: Fast convex minimization via inertial dynamics with Hessian driven damping. J. Differential Equations **261**, 5734–5783 (2016)
18. Aujol, J.-F., Dossal, Ch.: Stability of over-relaxations for the Forward-Backward algorithm, application to FISTA. SIAM J. Optim. **25**, 2408–2433 (2015)
19. Baillon, J.-B.:, Un exemple concernant le comportement asymptotique de la solution du problème $\frac{du}{dt} + \partial\phi(u) \ni 0$. J. Functional Anal. **28**, 369–376 (1978)
20. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert spaces. Springer (2011)
21. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
22. Beck, A., Teboulle, M.: Gradient-Based Algorithms with Applications in Signal Recovery Problems. In Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar Eds., Cambridge University Press, 33–88 (2010)
23. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Lojasiewicz Inequalities and Applications. Trans. AMS **362**, 3319–3363 (2010)
24. Bolte, J., Nguyen, T.P., Peypouquet, J., Suter, B.: From error bounds to the complexity of first-order descent methods for convex functions. Math. Prog. **165**, 471–507 (2017)
25. Brézis, H.: Opérateurs maximaux monotones dans les espaces de Hilbert et équations d'évolution. North Holland, (1972)
26. Chambolle, A., Dossal, Ch.: On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm. J. Optim. Theory Appl. **166**, 968–982 (2015)
27. Drori, Y., Teboulle, M.: Performance of first-order methods for smooth convex minimization: a novel approach. Math. Prog. (Ser. A) **145**, 451–482 (2014)
28. Güler, O.: On the convergence of the proximal point algorithm for convex optimization. SIAM J. Control Optim. **29** 403–419 (1991)
29. Imbert, C.: Convex Analysis techniques for Hopf-Lax formulae in Hamilton-Jacobi equations. J. of Nonlinear Convex Anal. **2** 333–343 (2001)
30. Kim, D., Fessler, J.A.: Optimized first-order methods for smooth convex minimization. Math. Prog. to appear. DOI 10.1007/s10107–015-0949-3.
31. Liang, J., Fadili, J., Peyré, G.: Local linear convergence of forward-backward under partial smoothness. Advances in Neural Information Processing Systems, 1970–1978 (2014)
32. May, R.: Asymptotic for a second-order evolution equation with convex potential and vanishing damping term. Turkish J. Math. **41**, 681–685 (2017)
33. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Doklady **27**, 372–376 (1983)
34. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA (2004)
35. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends in Optimization **1**, 123–231 (2013)
36. Peypouquet, J.: Convex Optimization in Normed Spaces: Theory, Methods and Examples. Springer (2015)
37. Villa, S., Salzo, S., Baldassarres, L., Verri A.: Accelerated and inexact forward-backward. SIAM J. Optim. **23**, 1607–1633 (2013)
38. Su, W., Boyd, S., Candès, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. Neural Information Processing Systems **27**, 2510–2518 (2014)

Chapter 2
Constraint Splitting and Projection Methods for Optimal Control of Double Integrator

Heinz H. Bauschke, Regina S. Burachik, and C. Yalçın Kaya

We dedicate our contribution to the memory of our friend and mentor Jonathan Borwein

Abstract We consider the minimum-energy control of a car, which is modelled as a point mass sliding on the ground in a fixed direction, and so it can be mathematically described as the double integrator. The control variable, representing the acceleration or the deceleration, is constrained by simple bounds from above and below. Despite the simplicity of the problem, it is not possible to find an analytical solution to it because of the constrained control variable. To find a numerical solution to this problem we apply three different projection-type methods: (i) Dykstra's algorithm, (ii) the Douglas–Rachford (DR) method and (iii) the Aragón Artacho–Campoy (AAC) algorithm. To the knowledge of the authors, these kinds of (projection) methods have not previously been applied to continuous-time optimal control problems, which are infinite-dimensional optimization problems. The problem we study in this article is posed in infinite-dimensional Hilbert spaces. Behaviour of the DR and AAC algorithms are explored via numerical experiments with respect to their parameters. An error analysis is also carried out numerically for a particular instance of the problem for each of the algorithms.

H. H. Bauschke
Department of Mathematics, University of British Columbia, Kelowna, BC, Canada
e-mail: heinz.bauschke@ubc.ca

R. S. Burachik
School of IT & Mathematical Sciences, University of South Australia, Mawson Lakes, SA, Australia
e-mail: regina.burachik@unisa.edu.au

C. Y. Kaya (✉)
School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, Adelaide, SA, Australia
e-mail: yalcin.kaya@unisa.edu.au

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_2

Keywords Optimal control · Dykstra projection method · Douglas-Rachford method · Aragón Artacho–Campoy algorithm · Linear quadratic optimal control · Control constraints · Numerical methods

AMS 2010 Subject Classification 49M27, 65K10, 90C20

2.1 Introduction

In this paper, we provide (to the best of our knowledge also first) application of various best approximation algorithms to solve a continuous-time optimal control problem. Operator splitting methods were applied previously to discrete-time optimal control problems [19, 26], which are finite-dimensional problems. In [26], for example, the state difference equations comprise the constraint $\mathcal{A}$, and the box constraints on the state and control variables comprise $\mathcal{B}$. The condition of belonging to the sets $\mathcal{A}$ and $\mathcal{B}$ are then appended to the objective function via indicator functions. The original objective function that is considered in [26] is quadratic in the state and control variables. In the next step in [26], the new objective function is split into its quadratic and convex parts and the Douglas-Rachford splitting method is applied to solve the problem.

In the current paper, we deal with continuous-time optimal control problems, which are infinite-dimensional optimization problems that are set in Hilbert spaces. After splitting the constraints of the problem, we apply Dykstra's algorithm [11], the Douglas–Rachford (DR) method [6, 9, 17, 18, 25, 29], and the Aragón Artacho–Campoy (AAC) algorithm [3], all of which solve the underlying best approximation problem.

The exposure of the current paper is more in the style of a tutorial. We pose the problem of minimum-energy control of a simplified model of a car, amounting to the double integrator, where the control variable has simple lower and upper bounds and the initial and terminal state variables are specified. We split the constraints into two, $\mathcal{A}$ and $\mathcal{B}$, representing respectively the state differential equations (the double integrator) along with their boundary conditions and the constraints on the control variable. We define two subproblems, one subject to $\mathcal{A}$, and the other one subject to $\mathcal{B}$. We take advantage of the relatively simple form of the optimal control problem and derive analytical expressions for the optimality conditions and implement these in defining the projections onto $\mathcal{A}$ and $\mathcal{B}$.

The solutions of these subproblems provide the projections of a given point in the control variable space onto the constraint sets $\mathcal{A}$ and $\mathcal{B}$, respectively, in some optimal way. By performing these projections in the way prescribed by the above-listed algorithms, we can ensure convergence to a solution of the original optimal control problem,

Note that while the minimum-energy control of the double integrator without any constraints on the control variable can be solved analytically, the same problem with (even simple bound, i.e., box) constraints on the control variable can in general be solved only numerically. This problem should be considered within the framework

of control-constrained linear-quadratic optimal control problems for which new numerical methods are constantly being developed—see for example [1, 12] and the references therein.

The current paper is a prototype for future applications of projection methods to solving more general optimal control problems. Indeed, the minimum-energy control of double integrator is a special case of linear quadratic optimal control problems; so, with the reporting of the current study, an extension to more general problems will be imminent.

The paper is organized as follows. In Section 2.2, we state the control-constrained minimum-energy problem for the double integrator, and write down the optimality conditions. We provide the analytical solution for the unconstrained problem. For the control-constrained case, we briefly describe the standard numerical approach and consider an instance of the problem which we use in the numerical experiments in the rest of the paper. We define the constraint sets $\mathcal{A}$ and $\mathcal{B}$. In Section 2.3, we provide the expressions for the projections onto $\mathcal{A}$ and $\mathcal{B}$. We describe the algorithms in Section 2.4 and in the beginning of Section 2.5. In the remaining part of Section 2.5, we present numerical experiments to study parametric behaviour of the algorithms as well as the errors in the state and control variables with each algorithm. In Section 2.6, we provide concluding remarks and list some open problems.

2.2 Minimum-Energy Control of Double Integrator

We consider the minimum-energy control of a car, with a constrained control variable. Consider the car as a point unit mass, moving on a frictionless ground in a fixed line of action. Let the position of the car at time t be given by $y(t)$ and the velocity by $\dot{y}(t) := (dy/dt)(t)$. By Newton's second law of motion, $\ddot{y}(t) = u(t)$, where $u(t)$ is the summation of all the external forces applied on the car, in this case the force simply representing the acceleration and deceleration of the car. This differential equation model is referred to as the *double integrator* in system theory literature, since $y(t)$ can be obtained by integrating $u(t)$ twice.

Optimal Control Problem Suppose that the total force on the car, i.e., the acceleration or deceleration of the car, is constrained by a magnitude of $a > 0$. Let $x_1 := y$ and $x_2 := \dot{y}$. Then the problem of minimizing the energy of the car, which starts at a position $x_1(0) = s_0$ with a velocity $x_2(0) = v_0$ and finishes at some other position $x_1(1) = s_f$ with velocity $x_2(1) = v_f$, within one unit of time, can be posed as follows.

$$
\text{(P)} \quad \begin{cases} \min \quad \dfrac{1}{2}\displaystyle\int_0^1 u^2(t)\,dt \\ \text{subject to} \quad \dot{x}_1(t) = x_2(t)\,, \quad x_1(0) = s_0\,, \quad x_1(1) = s_f\,, \\ \qquad\qquad \dot{x}_2(t) = u(t)\,, \quad x_2(0) = v_0\,, \quad x_2(1) = v_f\,, \quad |u(t)| \le a\,. \end{cases}
$$

Here, the functions x_1 and x_2 are referred to as the *state variables* and u the *control variable*. As a first step in writing the conditions of optimality for this optimization problem, define the Hamiltonian function H for Problem (P) simply as

$$H(x_1, x_2, u, \lambda_1, \lambda_2) := \frac{1}{2}\, u^2 + \lambda_1\, x_2 + \lambda_2\, u\,, \tag{2.1}$$

where $\lambda(t) := (\lambda_1(t), \lambda_2(t)) \in \mathbf{R}^2$ is the *adjoint variable* (or *costate*) *vector* such that (see [21])

$$\dot{\lambda}_1 = -\partial H/\partial x_1 \quad \text{and} \quad \dot{\lambda}_2 = -\partial H/\partial x_2\,. \tag{2.2}$$

Equations in (2.2) simply reduce to

$$\lambda_1(t) = c_1 \quad \text{and} \quad \lambda_2(t) = -c_1\, t - c_2\,, \tag{2.3}$$

where c_1 and c_2 are real constants. Let $x(t) := (x_1(t), x_2(t)) \in \mathbf{R}^2$ denote the state variable vector.

Maximum Principle If u is an optimal control for Problem (P), then there exists a continuously differentiable vector of adjoint variables λ, as defined in (2.2), such that $\lambda(t) \neq 0$ for all $t \in [0, t_f]$, and that, for a.e. $t \in [0, t_f]$,

$$u(t) = \operatorname*{arg\,min}_{v \in [-a,a]} H(x, v, \lambda(t))\,, \tag{2.4}$$

i.e.,

$$u(t) = \operatorname*{arg\,min}_{v \in [-a,a]} \frac{1}{2}\, v^2 + \lambda_2(t)\, v\,; \tag{2.5}$$

see e.g. [21]. Condition (2.5) implies that the optimal control is given by

$$u(t) = \begin{cases} -\lambda_2(t)\,, & \text{if } -a \le \lambda_2(t) \le a\,, \\ a\,, & \text{if } \lambda_2(t) \le -a\,, \\ -a\,, & \text{if } \lambda_2(t) \ge a\,. \end{cases} \tag{2.6}$$

From (2.6), we can also conclude that the optimal control u for Problem (P) is continuous.

When a is large enough, the control constraint does not become active, so the optimal control is simply $-\lambda_2$, and it is a straightforward classroom exercise to find the analytical solution as

$$u(t) = c_1\, t + c_2\,,$$

$$x_1(t) = \frac{1}{6}\, c_1\, t^3 + \frac{1}{2}\, c_2\, t^2 + v_0\, t + s_0\,,$$

$$x_2(t) = \frac{1}{2}\, c_1\, t^2 + c_2\, t + v_0\,,$$

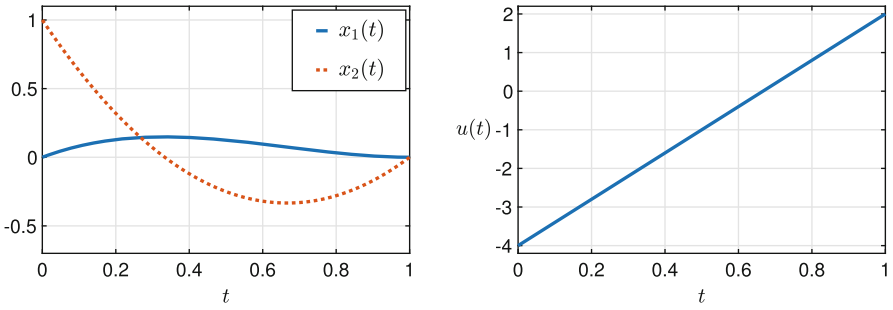


Fig. 2.1 Solution of Problem (P) with large a (so that $u(t)$ is unconstrained), $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$. (**a**) Optimal state variables. (**b**) Optimal control variable

for all $t \in [0, 1]$, where

$$c_1 = -12\,(s_f - s_0) + 6\,(v_0 + v_f)\,,$$

$$c_2 = 6\,(s_f - s_0) - 2\,(2\,v_0 + v_f)\,.$$

The solution of an instance of Problem (P), with $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$, and large a, say $a = 9$, is depicted in Figure 2.1. Note that, for all $t \in [0, 1]$, $\lambda_2(t) = -u(t) = -6t + 4$ and $\lambda_1(t) = c_1 = 6$. The graphs of λ_1 and λ_2 are not displayed for this particular instance.

When a is not so large, say $a = 2.5$, as we will consider next so that the control constraint becomes active, it is usually not possible to find an analytical solution, i.e., a solution has to be found numerically, as described below.

Numerical Approach A straightforward and popular numerical approach to solving Problem (P) is to discretize Problem (P) over a partition of the time horizon [0, 1] and then use some finite-dimensional optimization software to get a *discrete* (finite-dimensional) *solution* for the state and control variables $x(t)$ and $u(t)$. The discrete solution is an approximation of the continuous-time solution. This approach is often referred to as the *direct method* or the (*first-*)*discretize-then-optimize* approach. A survey and discussion of Euler discretization of linear-quadratic optimal control problems and convergence of their discretized solutions to their continuous-time solutions can be found in [12, Section 5].

Figure 2.2 depicts the discrete solution of Problem (P) with the instance where $a = 2.5$, $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$. The solution was obtained by pairing up the optimization modelling language AMPL [20] and the finite-dimensional

optimization software Ipopt [30]. The number of discretization nodes was taken to be 2000. The multipliers of the (Euler approximation of the) state differential equation constraints are provided by Ipopt when it finds an optimal solution to the discretized (finite-dimensional) problem. These multipliers have been plotted in Figure 2.2c. It should be noted that the graph of the adjoint variable $\lambda_2(t)$ given in Figure 2.2c verifies the graph of the optimal control $u(t)$ in Figure 2.2b via the optimal control rule in (2.6). In Figure 2.2b and c, the bounds ± 2.5 have been marked by horizontal dashed lines for ease of viewing.

Remark 2.1 If a is too small, there will obviously be no solution to Problem (P). For the particular instance of the problem considered here, the critical value of a, below which there exists no solution, is somewhere between 2.414 and 2.415, as our numerical experiments show (not reported in detail here). At this critical value, the only feasible solution is bang–bang, i.e., $u(t)$ switches once from $-a$ to a at around $t = 0.71$. It should be noted that, in this case, the optimal control in (2.6) requires the adjoint variable λ_2 to switch from a value $\alpha \geq a$ to another value $\beta \leq -a$, i.e., be discontinuous, which is not allowed by the maximum principle. In this paper, we

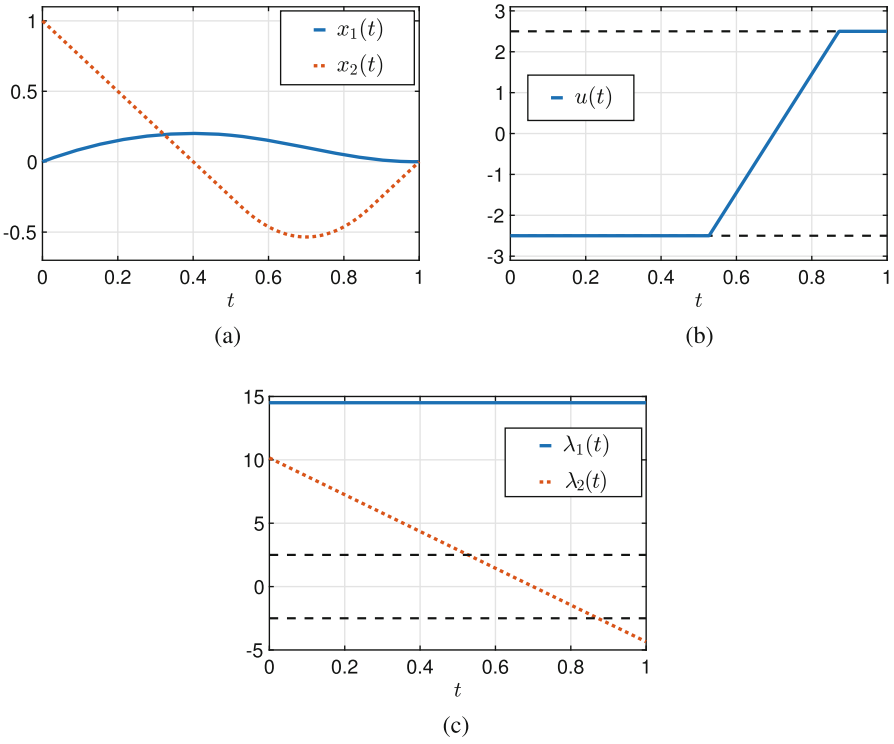


Fig. 2.2 Solution of direct discretization of Problem (P), with $a = 2.5$, $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$. (**a**) Optimal state variables. (**b**) Optimal control variable. (**c**) Adjoint variables

only consider the case when a is strictly greater than its critical value so that the maximum principle can be applied.

Function Spaces For the numerical methods, we consider projection/reflection methods in Hilbert spaces. The spaces associated with Problem (P) are set up as follows. Let $q \in \mathbf{N}$ and $L^2(0, 1; \mathbf{R}^q)$ be the Banach space of Lebesgue measurable functions

$$\begin{aligned} z : [0, 1] &\to \mathbf{R}^q \\ t &\mapsto (z_1(t), \ldots, z_q(t)), \end{aligned}$$

with finite L^2 norm. Namely, define

$$\|z\|_2 := \left(\sum_{i=1}^{q} \|z_i\|_2^2 \right)^{1/2},$$

where

$$\|z_i\|_2 := \left(\int_0^1 |z_i(t)|^2 \, \mathrm{d}t \right)^{1/2},$$

for $i = 1, \ldots, q$, with $|\cdot|$ the modulus or absolute value. In other words,

$$L^2(0, 1; \mathbf{R}^q) := \left\{ z : [0, 1] \to \mathbf{R}^q \ : \ \|z\|_2 < \infty \right\}.$$

Furthermore, $W^{1,2}(0, 1; \mathbf{R}^q)$ is the Sobolev space of absolutely continuous functions, namely

$$W^{1,2}(0, 1; \mathbf{R}^q) = \{ z \in L^2(0, 1; \mathbf{R}^q) \mid \dot{z} = dz/dt \in L^2(0, 1; \mathbf{R}^q) \},$$

endowed with the norm

$$\|z\|_{W^{1,2}} := \left(\sum_{i=1}^{q} \left[\|z_i\|_2^2 + \|\dot{z}_i\|_2^2 \right] \right)^{1/2}.$$

In Problem (P), the state variable $x \in W^{1,2}(0, 1; \mathbf{R}^2)$ and the control variable $u \in L^2(0, 1; \mathbf{R})$.

Constraint Splitting Next, we split the constraints of Problem (P) into two subsets, $\mathcal{A}$ and $\mathcal{B}$. The subset $\mathcal{A}$ collects together all the feasible control functions satisfying only the dynamics of the car. The subset $\mathcal{B}$, on the other hand, collects all the control functions whose values are constrained by $-a$ and a.

$$\mathcal{A} := \big\{u \in L^2(0,1;\mathbf{R}) \mid \exists x \in W^{1,2}(0,1;\mathbf{R}^2) \text{ which solves}$$
$$\dot{x}_1(t) = x_2(t)\,,\ x_1(0) = s_0\,,\ x_1(1) = s_f\,,$$
$$\dot{x}_2(t) = u(t)\,,\quad x_2(0) = v_0\,,\ x_2(1) = v_f\,,\ \forall t \in [0,1]\big\}\,, \tag{2.7}$$

$$\mathcal{B} := \big\{u \in L^2(0,1;\mathbf{R}) \mid -a \le u(t) \le a\,,\ \text{for all } t \in [0,1]\big\}\,. \tag{2.8}$$

The rationale behind this sort of splitting is as follows: The problem of minimizing the energy of the car subject to only $\mathcal{A}$ or only $\mathcal{B}$ is much easier to solve—in fact, the solutions can be analytically written up in each case. If, for some given u, a solution exists to the two-point boundary-value problem (TPBVP) in (2.7) then that solution is unique by the linearity of the TPBVP [5, 28]. Note that a control solution u as in (2.7) exists by the (Kalman) controllability of the double integrator—see [27]. So the set $\mathcal{A}$ is nonempty. Note that the constraint set $\mathcal{A}$ is an *affine subspace* and $\mathcal{B}$ a *box*.

2.3 Projections

All of the projection methods that we will consider involve projections onto the sets $\mathcal{A}$ and $\mathcal{B}$. The projection onto $\mathcal{A}$ from a current iterate u^- is the point u solving the following problem.

$$\text{(P1)} \quad \begin{cases} \min & \dfrac{1}{2}\displaystyle\int_0^1 (u(t) - u^-(t))^2\,dt \\ \text{subject to} & u \in \mathcal{A}\,. \end{cases}$$

In (P1), we minimize the squared L^2-norm distance between u^- and u. The projection onto $\mathcal{B}$ from a current iterate u^- is similarly the point u solving the following problem.

$$\text{(P2)} \quad \begin{cases} \min & \dfrac{1}{2}\displaystyle\int_0^1 (u(t) - u^-(t))^2\,dt \\ \text{subject to} & u \in \mathcal{B}\,. \end{cases}$$

Proposition 2.1 (Projection onto $\mathcal{A}$) *The projection $P_{\mathcal{A}}$ of $u^- \in L^2(0,1;\mathbf{R})$ onto the constraint set $\mathcal{A}$, as the solution of Problem (P1), is given by*

$$P_{\mathcal{A}}(u^-)(t) = u^-(t) + c_1\,t + c_2\,, \tag{2.9}$$

for all $t \in [0, 1]$*, where*

$$c_1 = 12\,(x_1(1) - s_f) - 6\,(x_2(1) - v_f)\ , \tag{2.10}$$

$$c_2 = -6\,(x_1(1) - s_f) + 2\,(x_2(1) - v_f)\ , \tag{2.11}$$

and $x_1(1)$ *and* $x_2(1)$ *are obtained by solving the initial value problem*

$$\dot{x}_1(t) = x_2(t)\,, \quad x_1(0) = s_0\,, \tag{2.12}$$

$$\dot{x}_2(t) = u^-(t)\,, \quad x_2(0) = v_0\,, \tag{2.13}$$

for all $t \in [0, 1]$.

Proof The Hamiltonian function for Problem (P1) is

$$H_1(x_1, x_2, u, \lambda_1, \lambda_2, t) := \frac{1}{2}\,(u - u^-)^2 + \lambda_1\, x_2 + \lambda_2\, u\,,$$

where the adjoint variables λ_1 and λ_2 are defined as in (2.2), with H replaced by H_1, and the subsequent solutions are given as in (2.3). The optimality condition for Problem (P1) is akin to that in (2.4) for Problem (P) and, owing to the fact that the control u is now unconstrained, can more simply be written as

$$\frac{\partial H_1}{\partial u}(x, u, \lambda, t) = 0\,,$$

which yields the optimal control as $u(t) = u^-(t) - \lambda_2(t)$, i.e.

$$u(t) = u^-(t) + c_1\, t + c_2\,, \tag{2.14}$$

for all $t \in [0, 1]$. We need to show that c_1 and c_2 are found as in (2.10)–(2.11). Using (2.14) in (2.7) yields the following time-varying, linear two-point boundary-value problem.

$$\dot{x}_1(t) = x_2(t)\,, \qquad x_1(0) = s_0\,, \quad x_1(1) = s_f\,, \tag{2.15}$$

$$\dot{x}_2(t) = u^-(t) + c_1\, t + c_2\,, \quad x_2(0) = v_0\,, \quad x_2(1) = v_f\,, \tag{2.16}$$

for all $t \in [0, 1]$. In other words, Problem (P1) is reduced to solving Equations (2.15)–(2.16) for the unknown parameters c_1 and c_2. Once c_1 and c_2 are found, the projected point u in (2.14) is found. Since Equations (2.15)–(2.16) are linear in x_1 and x_2, a *simple shooting technique* [5, 28] provides the solution for c_1 and c_2 in just one iteration. The essence of this technique is that the initial-value problem (IVP)

$$\frac{\partial z_1(t,c)}{\partial t} = z_2(t,c)\,, \qquad z_1(0,c) = s_0\,, \tag{2.17}$$

$$\frac{\partial z_2(t,c)}{\partial t} = u^-(t) + c_1\,t + c_2\,, \quad z_2(0,c) = v_0\,, \tag{2.18}$$

for all $t \in [0,1]$, is solved repeatedly, so as to make the *discrepancy at* $t = 1$ vanish. Namely, we seek a parameter $c := (c_1, c_2)$ such that $z_1(1,c) - s_f = 0$ and $z_2(1,c) - v_f = 0$. The procedure is as follows. For a given c, there exists a unique solution $z(t,c) := (z_1(t,c), z_2(t,c))$ of (2.17)–(2.18). Define the *near-miss* (vector) *function* $\varphi : \mathbf{R}^2 \to \mathbf{R}^2$ as follows:

$$\varphi(c) := \begin{bmatrix} z_1(1,c) - s_f \\ z_2(1,c) - v_f \end{bmatrix}. \tag{2.19}$$

The Jacobian of the near-miss function is

$$J_\varphi(c) := \begin{bmatrix} \dfrac{\partial z_1(1,c)}{\partial c_1} & \dfrac{\partial z_1(1,c)}{\partial c_2} \\ \dfrac{\partial z_2(1,c)}{\partial c_1} & \dfrac{\partial z_2(1,c)}{\partial c_2} \end{bmatrix}$$

The shooting method looks for a pair c such that $\varphi(c) := 0$ (i.e., a pair c such that the terminal boundary conditions are met). Expanding φ about, say, $\bar{c} = 0$, and discarding the terms of order 2 or higher, we obtain

$$\varphi(c) \approx \varphi(0) + J_\varphi(0)\,c\,.$$

Substituting $\varphi(c) = 0$ in the above expression, replacing "$\approx$" with "$=$", and rearranging, gives the single (Newton) iteration of the shooting method:

$$c = -[J_\varphi(0)]^{-1}\varphi(0)\,. \tag{2.20}$$

The components $(\partial z_i/\partial c_j)(1,c)$, $i, j = 1, 2$, of $J_\varphi(c)$, can be obtained by solving the variational equations for (2.15)–(2.16) with respect to c_1 and c_2, i.e., by solving the following system for $(\partial z_i/\partial c_j)(\cdot, c)$:

$$\frac{\partial}{\partial t}\left(\frac{\partial z_1}{\partial c_1}\right)(t,c) = \frac{\partial z_2}{\partial c_1}(t,c)\,, \quad \frac{\partial z_1}{\partial c_1}(0,c) = 0\,,$$

$$\frac{\partial}{\partial t}\left(\frac{\partial z_1}{\partial c_2}\right)(t,c) = \frac{\partial z_2}{\partial c_2}(t,c)\,, \quad \frac{\partial z_1}{\partial c_2}(0,c) = 0\,,$$

$$\frac{\partial}{\partial t}\left(\frac{\partial z_2}{\partial c_1}\right)(t,c) = t\,, \qquad \frac{\partial z_2}{\partial c_1}(0,c) = 0\,,$$
$$\frac{\partial}{\partial t}\left(\frac{\partial z_2}{\partial c_2}\right)(t,c) = 1\,, \qquad \frac{\partial z_2}{\partial c_2}(0,c) = 0\,.$$

Elementary calculations lead to the following solution of the above system:

$$\frac{\partial z}{\partial c}(t,c) = \begin{bmatrix} t^3/6 & t^2/2 \\ t^2/2 & t \end{bmatrix},$$

which is independent of c. Hence,

$$J_\varphi(0) = \frac{\partial z}{\partial c}(1,0) = \begin{bmatrix} 1/6 & 1/2 \\ 1/2 & 1 \end{bmatrix},$$

with inverse:

$$\left[\frac{\partial z}{\partial c}(1,0)\right]^{-1} = \left[J_\varphi(0)\right]^{-1} = \begin{bmatrix} -12 & 6 \\ 6 & -2 \end{bmatrix}. \tag{2.21}$$

Setting $(x_1(\cdot), x_2(\cdot)) := (z_1(\cdot,0), z_2(\cdot,0))$, the IVP (2.17)–(2.18) becomes (2.12)–(2.13). Then substitution of (2.19) and (2.21) with $c = 0$ into Equation (2.20), and expanding out, yield (2.10)–(2.11). The proof is complete. □

Proposition 2.2 (Projection onto $\mathcal{B}$) *The projection $P_{\mathcal{B}}$ of $u^- \in L^2(0,1;\mathbf{R})$ onto the constraint set $\mathcal{B}$, as the solution of Problem (P2), is given by*

$$P_{\mathcal{B}}(u^-)(t) = \begin{cases} u^-(t)\,, & \text{if } -a \le u^-(t) \le a\,, \\ -a\,, & \text{if } u^-(t) \le -a\,, \\ a\,, & \text{if } u^-(t) \ge a\,, \end{cases} \tag{2.22}$$

for all $t \in [0,1]$.

Proof The expression (2.22) is the straightforward solution of Problem (P2). □

2.4 Best Approximation Algorithms

In this section, we discuss best approximation algorithms. In the following,

$$X \text{ is a real Hilbert space} \tag{2.23}$$

with inner product $\langle \cdot, \cdot \rangle$, induced norm $\| \cdot \|$. We also assume that

$$A \text{ is a closed affine subspace of } X, \text{ and } B \text{ is a nonempty closed convex subset of } X. \tag{2.24}$$

Given $z \in X$, our aim is to find

$$P_{A\cap B}(z), \tag{2.25}$$

the projection of z onto the intersection $A \cap B$ which we assume to be *nonempty*. We also assume that we are able to compute the projectors P_A and P_B onto the constraints A and B, respectively.

Many algorithms are known which could be employed to find $P_{A\cap B}(z)$; here, however, we focus on three simple methods that do not require a product space set-up as some of those considered, in, e.g., [6, 7, 13, 14].

In the next section, we will numerically test these algorithms when $X = L^2(0, 1; \mathbf{R})$, $A = \mathcal{A}$, $B = \mathcal{B}$, and $z = 0$.

2.4.1 Dykstra's Algorithm

We start with Dykstra's algorithm (see [11]), which operates as follows[1]: Set $a_0 := z$ and $q_0 := 0$. Given a_n, q_n, where $n \geq 0$, update

$$b_n := P_B(a_n + q_n), \quad a_{n+1} := P_A(b_n), \quad \text{and} \quad q_{n+1} := a_n + q_n - b_n. \tag{2.26}$$

It is known that both $(a_n)_{n\in\mathbb{B}}$ and $(b_n)_{n\in\mathbb{N}}$ converge *strongly* to $P_{A\cap B}(z)$.

2.4.2 Douglas–Rachford Algorithm

Given $\beta > 0$, we specialize the Douglas–Rachford algorithm (see [17], [25] and [18]) to minimize the sum of the two functions $f(x) = \iota_B(x) + \frac{\beta}{2}\|x - z\|^2$ and $g := \iota_A$ which have respective proximal mappings (see [6, Proposition 23.29(i)]) $P_f(x) = P_B\big(\frac{1}{1+\beta}x + \frac{\beta}{1+\beta}z\big)$ and $P_g = P_A$. Set $\lambda := \frac{1}{1+\beta} \in]0, 1[$. It follows that the Douglas–Rachford operator $T := \mathrm{Id} - P_f + P_g(2P_f - \mathrm{Id})$ turns into

$$Tx = x - P_B\big(\lambda x + (1-\lambda)z\big) + P_A\Big(2P_B\big(\lambda x + (1-\lambda)z\big) - x\Big). \tag{2.27}$$

[1] In the general case, there is also an auxiliary sequence (p_n) associated with A; however, because A is an affine subspace, it is not needed in our setting.

Now let $x_0 \in X$ and given $x_n \in X$, where $n \geq 0$, update

$$b_n := P_B\big(\lambda x_n + (1-\lambda)z\big), \quad x_{n+1} := Tx_n = x_n - b_n + P_A\big(2b_n - x_n\big). \tag{2.28}$$

Then it is known (see [29] or [9]) that $(b_n)_{n\in\mathbb{N}}$ converges weakly to $P_{A\cap B}(z)$. Note that (2.28) simplifies to

$$x_{n+1} := x_n - P_B(\lambda x_n) + P_A\big(2P_B(\lambda x_n) - x_n\big) \quad \text{provided that } z = 0. \tag{2.29}$$

2.4.3 *Aragón Artacho–Campoy Algorithm*

The Aragón Artacho–Campoy (AAC) Algorithm was recently presented in [3]; see also [2, 4]. Given two fixed parameters α and β in $]0, 1[$, define

$$\begin{aligned} Tx &= (1-\alpha)x \\ &\quad + \alpha\bigg(2\beta\Big(P_A\Big(2\beta\big(P_B(x+z)-z\big)-x+z\Big)-z\Big)+x+2\beta\big(z-P_B(x+z)\big)\bigg) \\ &= x + 2\alpha\beta\Big(P_A\Big(2\beta\big(P_B(x+z)-z\big)-x+z\Big)-P_B(x+z)\Big). \end{aligned} \tag{2.30}$$

Now let $x_0 \in X$ and given $x_n \in X$, where $n \geq 0$, update

$$b_n := P_B(x_n + z), \tag{2.31}$$

and

$$x_{n+1} := Tx_n = x_n + 2\alpha\beta\Big(P_A\Big(2\beta\big(b_n - z\big) - x_n + z\Big) - b_n\Big). \tag{2.32}$$

By [3, Theorem 4.1(iii)], the sequence $(b_n)_{n\in\mathbb{N}}$ converges strongly to $P_{A\cap B}(z)$ provided that[2] $z - P_{A\cap B}(z) \in (N_A + N_B)(P_{A\cap B}z)$. Note that (2.32) simplifies to

$$x_{n+1} := Tx_n = x_n + 2\alpha\beta\Big(P_A\big(2\beta P_B x_n - x_n\big) - P_B x_n\Big) \quad \text{provided that } z = 0. \tag{2.33}$$

[2] It appears that this constraint qualification is not easy to check in our setting.

2.5 Numerical Implementation

2.5.1 The Algorithms

In this section, we gather the algorithms considered abstractly and explain how we implemented them.

We start with Dykstra's algorithm from Section 2.4.1.

Algorithm 1 (Dykstra)

Step 1 (*Initialization*) Choose the initial iterates $u^0 = 0$ and $q^0 = 0$. Choose a small parameter $\varepsilon > 0$, and set $k = 0$.
Step 2 (*Projection onto* $\mathcal{B}$) Set $u^- = u^k + q^k$. Compute $\widetilde{u} = P_{\mathcal{B}}(u^-)$ by using (2.22).
Step 3 (*Projection onto* $\mathcal{A}$) Set $u^- := \widetilde{u}$. Compute $\widehat{u} = P_{\mathcal{A}}(u^-)$ by using (2.9).
Step 4 (*Update*) Set $u^{k+1} := \widehat{u}$ and $q^{k+1} := u^k + q^k - \widetilde{u}$.
Step 5 (*Stopping criterion*) If $\|u^{k+1} - u^k\|_{L^\infty} \leq \varepsilon$, then return $\widetilde{u}$ and stop. Otherwise, set $k := k + 1$ and go to Step 2.

Next is the Douglas–Rachford method from Section 2.4.2.

Algorithm 2 (DR)

Step 1 (*Initialization*) Choose a parameter $\lambda \in]0, 1[$ and the initial iterate u^0 arbitrarily. Choose a small parameter $\varepsilon > 0$, and set $k = 0$.
Step 2 (*Projection onto* $\mathcal{B}$) Set $u^- = \lambda u^k$. Compute $\widetilde{u} = P_{\mathcal{B}}(u^-)$ by using (2.22).
Step 3 (*Projection onto* $\mathcal{A}$) Set $u^- := 2\widetilde{u} - u^k$. Compute $\widehat{u} = P_{\mathcal{A}}(u^-)$ by using (2.9).
Step 4 (*Update*) Set $u^{k+1} := u^k + \widehat{u} - \widetilde{u}$.
Step 5 (*Stopping criterion*) If $\|u^{k+1} - u^k\|_{L^\infty} \leq \varepsilon$, then return $\widetilde{u}$ and stop. Otherwise, set $k := k + 1$ and go to Step 2.

Finally, we describe the Aragón Artacho–Campoy algorithm from Section 2.4.3.

Algorithm 3 (AAC)

Step 1 (*Initialization*) Choose the initial iterate u^0 arbitrarily. Choose a small parameter $\varepsilon > 0$, two parameters[3] α and β in $]0, 1[$, and set $k = 0$.
Step 2 (*Projection onto* $\mathcal{B}$) Set $u^- = u^k$. Compute $\widetilde{u} = P_{\mathcal{B}}(u^-)$ by using (2.22).
Step 3 (*Projection onto* $\mathcal{A}$) Set $u^- = 2\beta\widetilde{u} - u^k$. Compute $\widehat{u} = P_{\mathcal{A}}(u^-)$ by using (2.9).
Step 4 (*Update*) Set $u^{k+1} := u^k + 2\alpha\beta(\widehat{u} - \widetilde{u})$.
Step 5 (*Stopping criterion*) If $\|u^{k+1} - u^k\|_{L^\infty} \leq \varepsilon$, then return $\widetilde{u}$ and stop. Otherwise, set $k := k + 1$ and go to Step 2.

We provide another version of each of Algorithms 1–3, as Algorithms 1b–3b, in Appendix A. In Algorithm 1b, we *monitor* the sequence of iterates which are

[3] Aragón Artacho and Campoy recommend $\alpha = 0.9$ and $\beta \in [0.7, 0.8]$; see [3, End of Section 7].

the projections onto set $\mathcal{A}$, instead of monitoring the projections onto set $\mathcal{B}$ in Algorithm 1. On the other hand, in Algorithms 2b–3b, the order in which the projections are done is reversed: the first projection is done onto the set $\mathcal{A}$ and the second projection onto $\mathcal{B}$.

Although the order of projections will not matter in view of the existing results stating that convergence is achieved under any order—see [8, Proposition 2.5(i)], the order does make a difference in early iterations (as well as in the number of iterations required for convergence of Algorithms 2 and 2b, as we will elaborate on later). If our intent is to stop the algorithm early so that we can use the current iterate as an initial guess in more accurate computational optimal control algorithms, which can find the junction times with a high precision (see [22–24]), then it is desirable to implement Algorithms 1–3 above, rather than Algorithms 1b–3b, because any iterate of Algorithms 1–3 will satisfy the constraints on the control variable, while that of Algorithms 1b–3b will in general not.

2.5.2 Numerical Experiments

In what follows, we study the working of Algorithms 1–3 for an instance of Problem (P). Suppose that the car is initially at a reference position 0 and has unit speed. It is desired that the car come back to the reference position and be at rest after one unit of time; namely that $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$. For these boundary conditions, no solution exists if one takes the control variable bound $a = 2.4$ or smaller but a solution does exist for $a = 2.5$. So, we use $a = 2.5$. In the ensuing discussions, we use the *stopping tolerance* $\varepsilon = 10^{-8}$ unless otherwise stated.

Discretization Algorithms 1–3, as well as 1b–3b, carry out iterations with functions. For computations, we consider discrete approximations of the functions over the partition $0 = t_0 < t_1 < \ldots < t_N = 1$ such that

$$t_{i+1} = t_i + h\,, \quad i = 0, 1, \ldots, N\,,$$

$h := 1/N$ and N is the number of subdivisions. Let u_i be an approximation of $u(t_i)$, i.e., $u_i \approx u(t_i)$, $i = 0, 1, \ldots, N-1$; similarly, $x_{1,i} \approx x_1(t_i)$ and $x_{2,i} \approx x_2(t_i)$, or $x_i := (x_{1,i}, x_{2,i}) \approx x(t_i)$, $i = 0, 1, \ldots, N$. In other words, the functions u, x_1 and x_2 are approximated by the N-dimensional array u_h, with components u_i, $i = 0, 1, \ldots, N-1$, and the $(N+1)$-dimensional arrays $x_{1,h}$ and $x_{2,h}$, with components $x_{1,i}$ and $x_{2,i}$, $i = 0, 1, \ldots, N$, respectively. We define a discretization $P^h_{\mathcal{A}}$ of the projection $P_{\mathcal{A}}$ as follows.

$$P^h_{\mathcal{A}}(u_h^-)(t) = u_h^- + c_1\, t_h + c_2\,, \tag{2.34}$$

where $t_h = (0, t_1, \ldots, t_N)$,

$$c_1 = 12\,(x_{1,N} - s_f) - 6\,(x_{2,N} - v_f)\ , \tag{2.35}$$

$$c_2 = -6\,(x_{1,N} - s_f) + 2\,(x_{2,N} - v_f)\ , \tag{2.36}$$

and $x_{1,N}$ and $x_{2,N}$ are obtained from the Euler discretization of (2.12)–(2.13): Given $x_{1,0} = s_0$ and $x_{2,0} = v_0$,

$$x_{1,i+1} = x_{1,i} + h\,x_{2,i}\ , \tag{2.37}$$

$$x_{2,i+1} = x_{2,i} + h\,u_i^-(t)\,, \tag{2.38}$$

for $i = 0, 1, \ldots, N-1$.

The discretization $P_{\mathcal{B}}^h$ of the projection $P_{\mathcal{B}}$ can be defined in a straightforward manner, by simply replacing u^- in (2.22) with the discrete components u_i^- of u_h^-.

Parametric Behaviour Obviously, the behaviour of Algorithms 2 and 2b, the Douglas–Rachford method, depend on the parameter λ, and the behaviour of Algorithms 3 and 3b on the two parameters α and β. Figure 2.3 displays the dependence of the number of iterations it takes to converge on these parameters, for various values of a. The dependence for a given value of a appears to be continuous, albeit the presence of downward *spikes*.

The graphs for Algorithms 2 and 2b, shown in parts (a) and (c) of Figure 2.3, respectively, differ significantly from one another. The bound $a = 4$ corresponds to the case when the control constraint becomes active only at $t = 0$—see Figure 2.1. In other words, when $a > 4$ the optimal control variable is truly unconstrained. When $a = 4$, the best value of λ is 1 for Algorithm 2, yielding the solution in just 6 iterations. For Algorithm 2b, the best value for λ is 0.5, as can be seen in (c), producing the solution in 30 iterations. Going back to Algorithm 2, with decreasing values of a, the values of λ minimizing the number of iterations shift to the right. For example, the minimum number of iterations is 91, with $a = 2.5$ and $\lambda = 0.7466$ (found by a refinement of the graph).

As for Algorithm 2b, the minimizer for $a = 2.5$ is $\lambda = 0.5982766$ and the corresponding minimum number of iterations is 38. This is a point where a downward spike occurs and so the number of iterations is very sensitive to changes in λ. For example, the rounded-off value of $\lambda = 0.598277$ results in 88 iterations instead of 38, and $\lambda = 0.55$ yields 444 iterations for convergence. The number of iterations is less sensitive to the local minimizer $\lambda = 0.7608$, which results in 132 iterations. It is interesting to note that the graph with $a = 4$ appears to be an envelope for the number of iterations for all $\lambda \in]0, 1[$.

The graphs for Algorithms 3 and 3b, the Aragón Artacho–Campoy algorithm, are indistinguishable to one's eye; therefore we only display the one in Figure 2.3b. Part (d) of Figure 2.3 shows surface plots of the number of iterations versus the algorithmic parameters α and β, for the same values of a as in the rest of the graphs in the figure. It is interesting to observe that the surfaces look to be cascaded with (roughly) the outermost surface being the one corresponding to $a = 4$. The

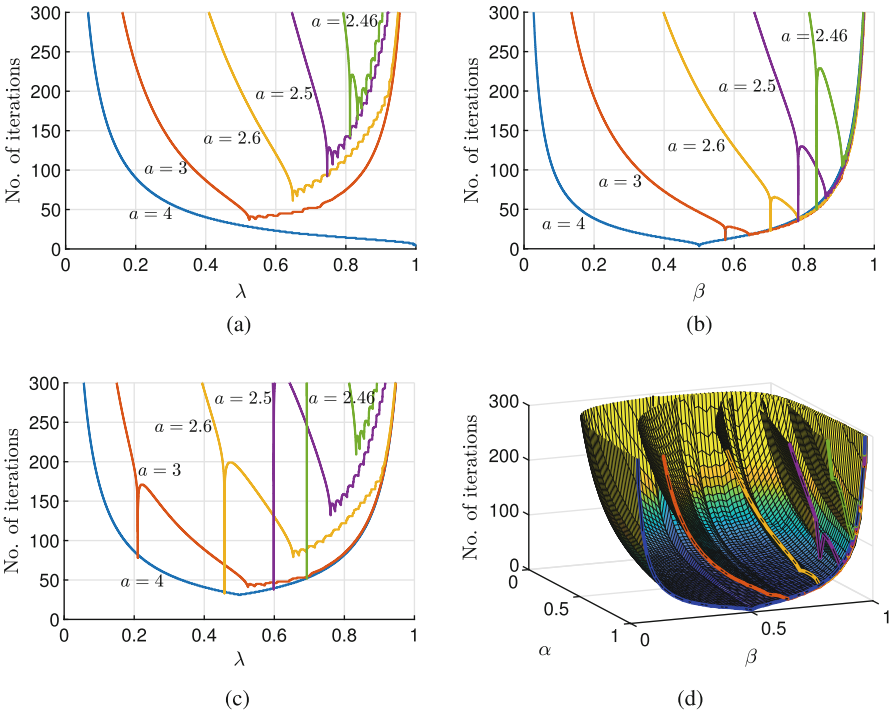


Fig. 2.3 Numerical experiments with $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$. (**a**) Algorithm 2 (DR). (**b**) Algorithms 3 (AAC) and 3b (AAC-b) for $\alpha = 1$. (**c**) Algorithm 2b (DR-b). (**d**) Algorithms 3 (AAC) and 3b (AAC-b)

surface plot suggests that for minimum number of iterations, one must have $\alpha = 1$. Although theory requires $\alpha < 1$, $\alpha = 1$ seems to cause no concerns in this particular instance; so, we set $\alpha = 1$ for the rest of the paper. The cross-sectional curves at $\alpha = 1$ are shown with much more precision in part (b) of the figure. The *spikes* that are observed in part (d) can also be seen in the graph in part (b).

In fact, the first observation one has to make here is that, for $a = 4$, convergence can be achieved in merely one iteration, with $\beta = 0.5$. This is quite remarkable, compared with Algorithms 2 and 2b. The graphs in (b) appear to be enveloped as well by the graph for $a = 4$, as in part (c). For the values of a other than 4, the globally minimum number of iterations seems to be achieved at a downward spike, which as a result is very sensitive to changes in β. For example, for $a = 2.5$, the optimal β value is 0.78249754 for a minimum 35 iterations. A rounded-off $\beta = 0.782$ results in 111 iterations, and $\beta = 0.7$ yields 243 iterations. Sensitivity at the local minimizer $\beta = 0.8617$ giving 64 iterations is far less: Choosing $\beta = 0.8$ or 0.9 results in 128 or 90 iterations, respectively. It is interesting to note that, as in the case of Algorithms 2 and 2b, the graphs in Figure 2.3b are approximately enveloped by the graph/curve drawn for $a = 4$.

Behaviour in Early Iterations Figure 2.4a–c illustrates the working of all three algorithms for the same instance. All three algorithms converge to the optimal solution, with the stopping tolerance of $\varepsilon = 10^{-8}$. The optimal values of the algorithmic parameters, $\lambda = 0.7466$ for Algorithm 2, and $\alpha = 1$ and $\beta = 0.8617$ for Algorithm 3, have been used. The third, fifth and fifteenth iterates, as well as the solution curve, are displayed for comparisons of behaviour. At least for the given instance of the problem, it is fair to say from Figure 2.4c that Algorithm 3 gets closer to the solution much more quickly than the others in the few initial iterations—see the third and fifth iterates. It also achieves convergence in a smaller number of iterations (64 as opposed to 530 and 91 iterations of the Algorithms 1 and 2, respectively).

Error Analysis via Numerical Experiments For a fixed value of N, Algorithms 1–3 converge only to some approximate solution of the original Problem. Therefore, the question as to how the algorithms behave as the time partition is refined, i.e., N is increased, needs to be investigated. For the purpose of a numerical investigation, we define, in the kth iteration, the following errors. Suppose that the pair (u^*, x^*) is the optimal solution of Problem (P) and (u_h^k, x_h^k) an approximate solution of Problem (P) in the kth iteration of a given algorithm. Define

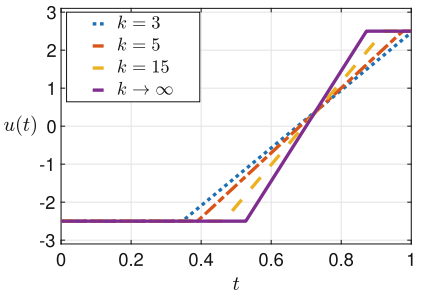
$$\sigma_u^k := \max_{0\le i\le N-1} |u_i^k - u^*(t_i)| \qquad \text{and} \qquad \sigma_x^k := \max_{0\le i\le N} ||x_i^k - x^*(t_i)||_\infty ,$$

where $||\cdot||_\infty$ is the ℓ_∞-norm in $\mathbf{R}^2$. For large N, these expressions are reminiscent of the L^∞-norm, and therefore they will be referred to as the L^∞-*error*.

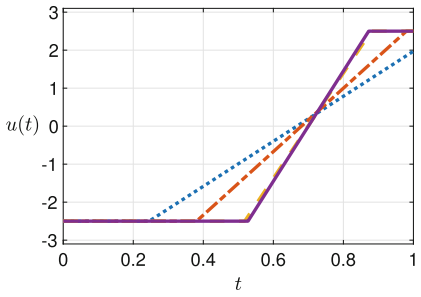
For (u^*, x^*) in the error expressions, we have used the discretized (approximate) solution obtained for the Euler-discretized Problem (P) utilizing the Ipopt–AMPL suite, with $N = 10^6$ and the tolerance set at 10^{-14}.

For $N = 2000$, these errors are depicted in Figure 2.4d and e. From the graphs it is immediately clear that no matter how much smaller the stopping tolerance is selected, the best error that is achievable with $N = 2000$ is around 10^{-2} for the control variable and around 10^{-3} for the state variable vector. In fact, the graphs also tell that perhaps a much smaller stopping threshold than 10^{-8} would have achieved the same approximation to the continuous-time solution of Problem (P). By just looking at the graphs, one can see that Algorithm 1 could have been run just for about 300 iterations instead of 530, and Algorithms 2 and 3 could have been run for about 50 iterations to achieve the best possible approximation with $N = 2000$.

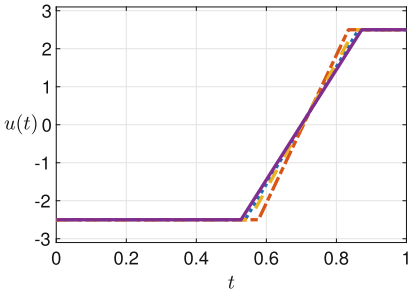
In Figure 2.5, we depict the same errors for $N = 10^3$ (parts (a) and (b)), $N = 10^4$ (in parts (c) and (d)) and $N = 10^5$ (in parts (e) and (f)). It is observed that, with a ten-fold increase in N (which is a ten-fold decrease in h) the errors in both u and x are reduced by ten-folds, implying that the error (both in x and in u) depends on the stepsize h linearly. This is in line with the theory of Euler-discretization of optimal control problems; see, for example, [15, 16]. Furthermore, even for very large values of N, it can be seen from these graphs that a stopping threshold slightly smaller than 10^{-8} would suffice to get even more stringent error levels, such as around 10^{-4} for



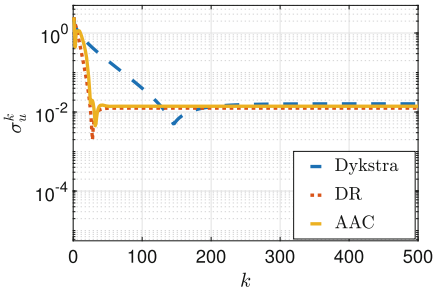
(a) Algorithm 1 (Dykstra) stops after 530 iterations.



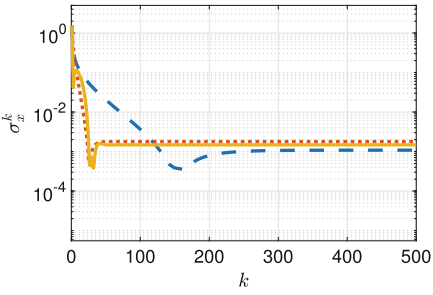
(b) Algorithm 2 (DR, $\lambda = 0.7466$) stops after 91 iterations.



(c) Algorithm 3 (AAC, $\alpha = 1$, $\beta = 0.8617$) stops after 64 iterations.



(d) L^∞-error in control by Algorithms 1–3.



(e) L^∞-error in states by Algorithms 1–3.

Fig. 2.4 Numerical experiments with $a = 2.5$, $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$, and the number of discretization subintervals $N = 2000$. The graphs in (**a**)–(**c**) show approximations of the optimal control function with Algorithms 1–3, after $k = 3, 5, 15$ iterations, with $\varepsilon = 10^{-8}$. All algorithms are observed to converge to the optimal solution indicated by $k \to \infty$, in various rates. The semi-log graphs in (**d**) and (**e**) show the L^∞ errors in the state and control variables, respectively, in each iteration of the three algorithms

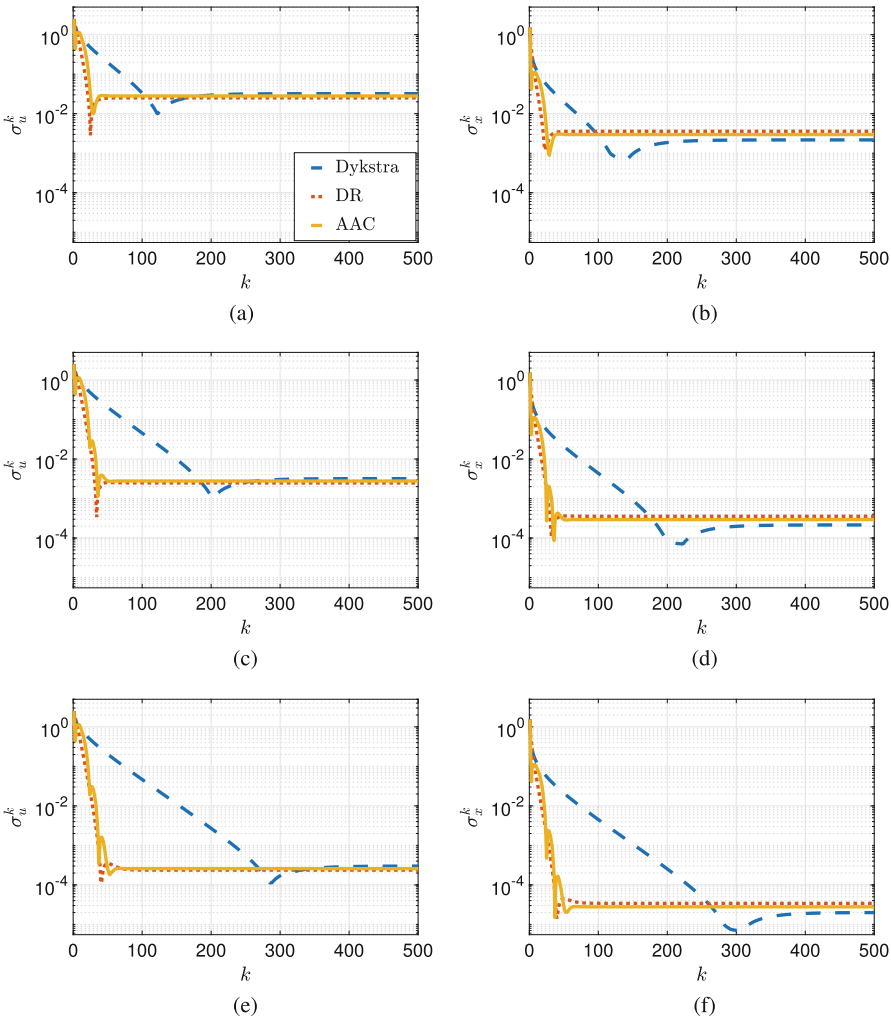


Fig. 2.5 Numerical experiments with $a = 2.5$, $s_0 = 0$, $s_f = 0$, $v_0 = 1$, $v_f = 0$. The semi-log graphs show the L^∞ errors in the state and control variables, respectively, in each iteration of the three algorithms, with various N from coarse ($N = 1000$) to fine ($N = 100000$). (**a**) L^∞-error in control with $N = 10^3$. (**b**) L^∞-error in states with $N = 10^3$. (**c**) L^∞-error in control with $N = 10^4$. (**d**) L^∞-error in states with $N = 10^4$. (**e**) L^∞-error in control with $N = 10^5$. (**f**) L^∞-error in states with $N = 10^5$

the control variable and around 10^{-5} for the state variable vector. A larger stopping threshold would obviously result in smaller number of iterations.

Table 2.1 displays the values of the errors, separately in u and x, after the stopping criteria with $\varepsilon = 10^{-8}$ was satisfied, for each of the three algorithms. A precise 10-fold reduction in error with a 10-fold increase in N can be verified with

Table 2.1 Least errors that can be achieved by Algorithms 1–3 and Ipopt, with $\varepsilon = 10^{-8}$

N	Dykstra	DR	AAC	Ipopt
(a) L^∞-error in control, σ_u^k				
10^3	3.2×10^{-2}	2.5×10^{-2}	2.8×10^{-2}	3.2×10^{-2}
10^4	3.2×10^{-3}	2.5×10^{-3}	2.8×10^{-3}	7.7×10^{-3}
10^5	3.0×10^{-4}	2.4×10^{-4}	2.6×10^{-4}	1.6×10^{-2}
(b) L^∞-error in states, σ_x^k				
10^3	2.2×10^{-3}	3.6×10^{-3}	3.0×10^{-3}	2.2×10^{-3}
10^4	2.1×10^{-4}	3.6×10^{-4}	2.9×10^{-4}	2.3×10^{-4}
10^5	2.0×10^{-5}	3.4×10^{-5}	2.8×10^{-5}	8.7×10^{-5}

Table 2.2 CPU times [sec] taken by Algorithms 1–3 and Ipopt. For $N = 10^3, 10^4, 10^5$, respectively: $\varepsilon = 10^{-6}, 10^{-6}, 10^{-7}$ for Algorithm 1, $\varepsilon = 10^{-5}, 10^{-5}, 10^{-7}$ for Algorithm 2, and $\varepsilon = 10^{-4}, 10^{-5}, 10^{-6}$ for Algorithm 3, have been used. The tolerance for Ipopt was set as 10^{-14}

N	Dykstra	DR	AAC	Ipopt
10^3	0.03	0.01	0.01	0.08
10^4	0.16	0.05	0.05	0.71
10^5	1.6	0.41	0.28	7.3

these numbers, as discussed in the previous paragraph. We have added the experiments we have carried out with Ipopt, version 3.12, an interior point optimization software [30], which solved the direct Euler-discretization of Problem (P), with the same values of N and the same tolerance 10^{-8}. Ipopt, running with linear solver MA57, was paired up with the optimization modelling language AMPL [20]. The same 10-fold decreases in error cannot be observed with Ipopt, unless one sets the tolerance for Ipopt to be much smaller than 10^{-8}, say 10^{-14} (which also means longer computational times). With the tolerance set at 10^{-14}, the error values with Ipopt becomes pretty much the same as those with Dykstra (still with $\varepsilon = 10^{-8}$), which is interesting to note.

As we pointed out earlier, the same errors listed in Table 2.1 can be achieved with bigger stopping thresholds. For $N = 10^3, 10^4, 10^5$, respectively: with $\varepsilon = 10^{-6}, 10^{-6}, 10^{-7}$, Algorithm 1 converges in 281, 359 and 454 iterations; with $\varepsilon = 10^{-5}, 10^{-5}, 10^{-7}$, Algorithm 2 converges in 65, 50 and 101 iterations; with $\varepsilon = 10^{-4}, 10^{-5}, 10^{-6}$, Algorithm 3 converges in 49, 60 and 70 iterations.

In Table 2.2, the CPU times (in seconds) each algorithm takes, with the respective ε values listed above, are tabulated. Note that Algorithms 1–3 have been coded and run on Matlab, 64-bit (maci64) version R2017b. All software, including AMPL and Ipopt, were run on MacBook Pro, with operating system macOS Sierra version 10.12.6, processor 3.3 GHz Intel Core i7 and memory 6 GB 2133 MHz LPDDR3. In Table 2.2, the CPU times for Ipopt are listed with the tolerance 10^{-14}, since with only this fine tolerance it is possible to obtain the same order of the error magnitudes as those obtained by Algorithms 1–3. With $\varepsilon = 10^{-8}$, the CPU times for Ipopt are

0.06, 0.45 and 4.4 seconds, respectively, which are significantly higher than the times taken by Algorithms 1–3, in addition to worse errors.

Numerical observations suggest two joint winners: Algorithms 2 and 3, i.e., the Douglas–Rachford method and the Aragón Artacho–Campoy algorithm, in both accuracy and speed.

2.6 Conclusion and Open Problems

We have applied three well-known projection methods to solve an optimal control problem, i.e., control-constrained minimum-energy control of double integrator. We have derived the projectors for the optimal control problem and demonstrated that they can be used in Dykstra's algorithm, the Douglas–Rachford (DR) method and the Aragón Artacho–Campoy (AAC) algorithm, effectively. We carried out extensive numerical experiments for an instance of the problem and concluded that the DR and AAC algorithms (Algorithms 2 and 3) were jointly the most successful. We also made comparisons with the standard discretization approach, only to witness the benefit of using projection methods.

It is interesting to note that when we apply alternating projections, we also seem to converge to $P_{\mathcal{A}\cap\mathcal{B}}(0)$ even though this is not supported by existing theory.

To the best of authors' knowledge, the current paper constitutes the first of its kind which involves projection methods and continuous-time optimal control problems. It can be considered as a prototype for future studies in this direction. Some of the possible directions are listed as follows.

- The setting we have introduced could be extended to general control-constrained linear-quadratic problems.
- We have used some discretization of the projector as well as the associated IVP in (2.34)–(2.38). This might be extended to problems in more general form. On the other hand, for the particular problem we have dealt with in the present paper, one might take into account the fact that if $u^-(t)$ is piecewise linear then its projection is piecewise linear. This might simplify further the expressions given in Proposition 2.1.
- Although theory for projection methods can in principle vouch convergence only for convex problems, it is well-known that the DR method can be successful for nonconvex problems, see, for example, [10]. It would be interesting to extend the formulations in the current paper to nonconvex optimal control problems.
- For a certain value of an algorithmic parameter, Figure 2.3 exhibits downward spikes. It would be interesting to see if this phenomenon is also observed in other control-constrained optimal control problems, as well as under other stopping criteria.

Appendix

Algorithm 1b (Dykstra-b)

Steps 1–4 (*Initialization*) Do as in Steps 1–4 of Algorithm 1.
Step 5 (*Stopping criterion*) If $\|u^{k+1} - u^k\|_{L^\infty} \leq \varepsilon$, then return u^{k+1} and stop. Otherwise, set $k := k + 1$ and go to Step 2.

Algorithm 2b (DR-b)

Step 1 (*Initialization*) Choose a parameter $\lambda \in]0, 1[$ and the initial iterate u^0 arbitrarily. Choose a small parameter $\varepsilon > 0$, and set $k = 0$.
Step 2 (*Projection onto* $\mathcal{A}$) Set $u^- = \lambda u^k$. Compute $\widetilde{u} = P_{\mathcal{A}}(u^-)$ by using (2.9).
Step 3 (*Projection onto* $\mathcal{B}$) Set $u^- := 2\widetilde{u} - u^k$. Compute $\widehat{u} = P_{\mathcal{B}}(u^-)$ by using (2.22).
Step 4 (*Update*) Set $u^{k+1} := u^k + \widehat{u} - \widetilde{u}$.
Step 5 (*Stopping criterion*) If $\|u^{k+1} - u^k\|_{L^\infty} \leq \varepsilon$, then return $\widetilde{u}$ and stop. Otherwise, set $k := k + 1$ and go to Step 2.

Algorithm 3b (AAC-b)

Step 1 (*Initialization*) Choose the initial iterate u^0 arbitrarily. Choose a small parameter $\varepsilon > 0$, two parameters α and β in $]0, 1[$, and set $k = 0$.
Step 2 (*Projection onto* $\mathcal{A}$) Set $u^- = u^k$. Compute $\widetilde{u} = P_{\mathcal{A}}(u^-)$ by using (2.9).
Step 3 (*Projection onto* $\mathcal{B}$) Set $u^- = 2\beta\widetilde{u} - u^k$. Compute $\widehat{u} = P_{\mathcal{B}}(u^-)$ by using (2.22).
Step 4 (*Update*) Set $u^{k+1} := u^k + 2\alpha\beta(\widehat{u} - \widetilde{u})$.
Step 5 (*Stopping criterion*) If $\|u^{k+1} - u^k\|_{L^\infty} \leq \varepsilon$, then return $\widetilde{u}$ and stop. Otherwise, set $k := k + 1$ and go to Step 2.

References

1. Alt, W., Kaya, C.Y., Schneider, C.: Dualization and discretization of linear-quadratic control problems with bang–bang solutions. EURO J. Comput. Optim. **4**, 47–77 (2016)
2. Alwadani, S., Bauschke, H.H., Moursi, W.M., Wang, X.: On the asymptotic behaviour of the Aragon Artacho-Campoy algorithm. Oper. Res. Letters **46** 585–587 (2018)
3. Aragón Artacho, F.J., Campoy, R.: A new projection method for finding the closest point in the intersection of convex sets. Comput. Optim. Appl. **69**, 99–132 (2018)
4. Aragón Artacho, F.J., Campoy, R.: Computing the resolvent of the sum of maximally monotone operators with the averaged alternating modified reflections algorithm. J. Optim. Th. Appl. **181**, 709–726 (2019)
5. Ascher, U.M., Mattheij, R.M.M., Russell, R.D.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations. SIAM Publications, Philadelphia (1995)
6. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Second edition. Springer (2017)
7. Bauschke, H.H., Koch, V.R.: Projection methods: Swiss Army knives for solving feasibility and best approximation problems with halfspaces. Infinite Products of Operators and Their Applications, 1–40 (2012)

8. Bauschke, H.H., Moursi, W.M.: On the order of the operators in the Douglas–Rachford algorithm. Optimization Letters **10**, 447–455 (2016)
9. Bauschke, H.H., Moursi, W.M.: On the Douglas–Rachford algorithm. Math. Program. (Ser. A) **164**, 263–284 (2017)
10. Borwein, J.M., Sims, B.: The Douglas-Rachford Algorithm in the absence of convexity. In: Bauschke H., Burachik R., Combettes P., Elser V., Luke D., Wolkowicz H. (eds) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer Optimization and Its Applications, vol 49, pp. 93–109. Springer, New York, NY (2011)
11. Boyle J.P., Dykstra R.L.: A method for finding projections onto the intersection of convex sets in Hilbert spaces. In: Advances in Order Restricted Statistical Inference, vol. 37, Lecture Notes in Statistics, pp. 28–47. Springer (1986)
12. Burachik, R.S., Kaya, C.Y., Majeed, S.N.: A duality approach for solving control-constrained linear-quadratic optimal control problems. SIAM J. Control Optim. **52**, 1771–1782 (2014)
13. Combettes, P.L.: A block-iterative surrogate constraint splitting method for quadratic signal recovery. IEEE Trans. Sig. Proc. **51**, 2432–2442 (2003)
14. Combettes, P.L.: Iterative construction of the resolvent of a sum of maximal monotone operators. J. Convex Anal. **16**, 727–748 (2009)
15. Dontchev, A.L., Hager, W.W.: The Euler approximation in state constrained optimal control. Math. Comp. **70**, 173–203 (2001)
16. Dontchev, A.L., Hager, W.W., Malanowski, K.: Error bound for Euler approximation of a state and control constrained optimal control problem. Numer. Funct. Anal. Optim. **21**, 653–682 (2000)
17. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Trans. Amer. Math. Soc. **82**, 421–439 (1956)
18. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Prog. (Ser. A) **55**, 293–318 (1992)
19. Eckstein, J., Ferris, M.C.: Operator-splitting methods for monotone affine variational inequalities, with a parallel application to optimal control. INFORMS J. Comput. **10**, 218–235 (1998)
20. Fourer, R., Gay, D.M., Kernighan, B.W.: AMPL: A Modeling Language for Mathematical Programming, Second Edition. Brooks/Cole Publishing Company / Cengage Learning (2003)
21. Hestenes, M.R.: Calculus of Variations and Optimal Control Theory. John Wiley & Sons, New York (1966)
22. Kaya, C.Y., Lucas, S.K. Simakov, S.T.: Computations for bang–bang constrained optimal control using a mathematical programming formulation. Optim. Contr. Appl. Meth. **25**, 295–308 (2004)
23. Kaya, C.Y., Noakes, J.L.: Computations and time-optimal controls, Opt. Cont. Appl. Meth. **17**, 171–185 (1996)
24. Kaya, C.Y., Noakes, J.L.: Computational algorithm for time-optimal switching control. J. Optim. Theory App. **117**, 69–92 (2003)
25. Lions, P.-L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
26. O'Donoghue, B., Stathopoulos, G., Boyd, S.: A splitting method for optimal control. IEEE Trans. Contr. Sys. Tech. **21**, 2432–2442 (2013)
27. Rugh, W.J.: Linear System Theory, 2nd Edition. Pearson (1995)
28. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis, 2nd Edition. Springer-Verlag, New York (1993)
29. Svaiter, B.F.: On weak convergence of the Douglas-Rachford method. SIAM J. Control Optim. **49** , 280–287 (2011)
30. Wächter, A., Biegler, L.T.: On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math. Prog. **106**, 25–57 (2006)

Chapter 3
Numerical Explorations of Feasibility Algorithms for Finding Points in the Intersection of Finite Sets

Heinz H. Bauschke, Sylvain Gretchko, and Walaa M. Moursi

Dedicated to the memory of Jonathan Borwein

Abstract Projection methods are popular algorithms for iteratively solving feasibility problems in Euclidean or even Hilbert spaces. They employ (selections of) nearest point mappings to generate sequences that are designed to approximate a point in the intersection of a collection of constraint sets. Theoretical properties of projection methods are fairly well understood when the underlying constraint sets are convex; however, convergence results for the nonconvex case are more complicated and typically only local. In this paper, we explore the perhaps simplest instance of a feasibility algorithm, namely when each constraint set consists of only finitely many points. We numerically investigate four constellations: either few or many constraint sets, with either few or many points. Each constellation is tackled by four popular projection methods each of which features a tuning parameter. We examine the behaviour for a single and for a multitude of orbits, and we also consider local and global behaviour. Our findings demonstrate the importance of the choice of the algorithm and that of the tuning parameter.

Keywords Cyclic Douglas–Rachford algorithm · Douglas–Rachford algorithm · Extrapolated parallel projection method · Method of cyclic projections · Nonconvex feasibility problem · Optimization algorithm · Projection

H. H. Bauschke (✉)
Department of Mathematics, University of British Columbia, Kelowna, BC, Canada
e-mail: heinz.bauschke@ubc.ca

S. Gretchko
Mathematics, UBCO, Kelowna, BC, Canada

W. M. Moursi
Electrical Engineering, Stanford University, Stanford, CA, USA

Faculty of Science, Mathematics Department, Mansoura University, Mansoura, Egypt
e-mail: wmoursi@stanford.edu

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_3

AMS 2010 Subject Classification 49M20, 49M27, 49M37, 65K05, 65K10, 90C25, 90C26, 90C30

3.1 Introduction

Background Let X be a Euclidean space (i.e., a finite-dimensional Hilbert space), with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. The *feasibility problem* is a common problem in science and engineering: given finitely many closed subsets $C_1, \ldots, C_m$ of X, it asks to

$$\text{Find } x \in C := C_1 \cap \cdots \cap C_m. \tag{FP}$$

We henceforth assume that the intersection C is nonempty. Algorithms for solving (3.1) exist when the constraint sets C_i allow for simple projectors P_{C_i} (i.e., nearest-point mappings). When C_i is convex, then the projector P_{C_i} is a nice (firmly nonexpansive and single-valued) operator defined on the entire space X; when C_i is not convex, then P_{C_i} is nonempty and set-valued. For notational simplicity, we will use P_{C_i} to denote an arbitrary but fixed selection of the set-valued projector. (If S is a subset of X, then $P_S(x)$ is a minimizer of the function $s \mapsto \|x - s\|$, where $s \in S$. For other notions not explicitly defined in this paper, we refer the reader to [1].)

Assuming that the operators $P_{C_1}, \ldots, P_{C_m}$ are readily available and implementable, one may try to solve (3.1) iteratively by generating a sequence $(x_k)_{k\in\mathbb{N}}$ of vectors in X that employs the projection operators P_{C_i} in some fashion to produce the next update. There are hundreds of papers dealing with algorithms for solving convex or nonconvex feasibility problems. Thus, we refrain from providing a comprehensive list of references and rather point to the following recent books and "meta" papers as starting points: [1, 2, 4, 10, 12–15]. (We note that the recent manuscript [5] deals with a feasibility problem where one set is a doubleton.) The convergence theory in the nonconvex case is much more challenging and usually of local character.

Goal of This Paper *The goal of this paper is to showcase the surprising numerical complexity of the most simple instance of* (3.1); namely, when each constraint set

$$C_i \text{ contains a } \textit{finite} \text{ number of points.}$$

In this case, the projection operator is very easy to implement—this is achieved by simply measuring the distance of the point to each point in C_i and returning the closest one. Furthermore, we will restrict ourselves to the simple case when the underlying space

$$X = \mathbb{R}^2$$

is simply the *Euclidean plane*. Even in this setting, the difficulty and richness of the dynamic behaviour is impressively illustrated.

It is our hope that the complexity revealed will spark further analytical research in feasibility algorithms with the goal to explain the observed complexity and ultimately to aid in the design of new algorithms for solving difficult feasibility problems.

Organization of the Paper The remainder of the paper is organized as follows. In Section 3.2, we present the four constellations we will use for our numerical exploration throughout the remainder of the paper. These constellations correspond to feasibility problems that we will attempt to solve using the algorithms listed in Section 3.3. Section 3.4 provides details on the implementation and execution of the numerical experiments. The "best" tuning parameter λ_{best} is determined in Section 3.5. We then track typical orbits of the algorithms in Section 3.6. Local and global behaviour is investigated in Section 3.7. Some interesting (and beautiful) behaviour outside of the main numerical experiments are collected in Section 3.8. The final Section 3.9 contains some concluding remarks.

3.2 The Four Constellations

Even though we restrict ourselves already to finitely many constraint sets with finitely many points in the Euclidean plane, the infinitely many possibilities to experiment make it a daunting task to explore this space. We opted to probe this universe as follows.

The points in each constraint set C_i are chosen randomly. We will ensure that the origin belongs to each set C_i

$$0 \in C_i \subset [-10, 10] \times [-10, 10]$$

to have a consistent feasibly problem with

$$C = C_1 \cap \cdots \cap C_m = \{0\}.$$

We will focus on two alternatives for the *number of constraint sets*, either "few" or "many". We will also consider constraint sets with a *maximum number of points* in the constraint sets, either "few" or "many". From now on, we will use the following language:

- The number of **few sets** is 3.
- The number of **many sets** is 10.
- The number of **few points** is 20.
- The number of **many points** is 100.

This will give rise to *four constellations*: few sets with few points, few sets with many points, many sets with few points, and many sets with many points. The four constellations used in our numerical experiments are shown in Figure 3.1.

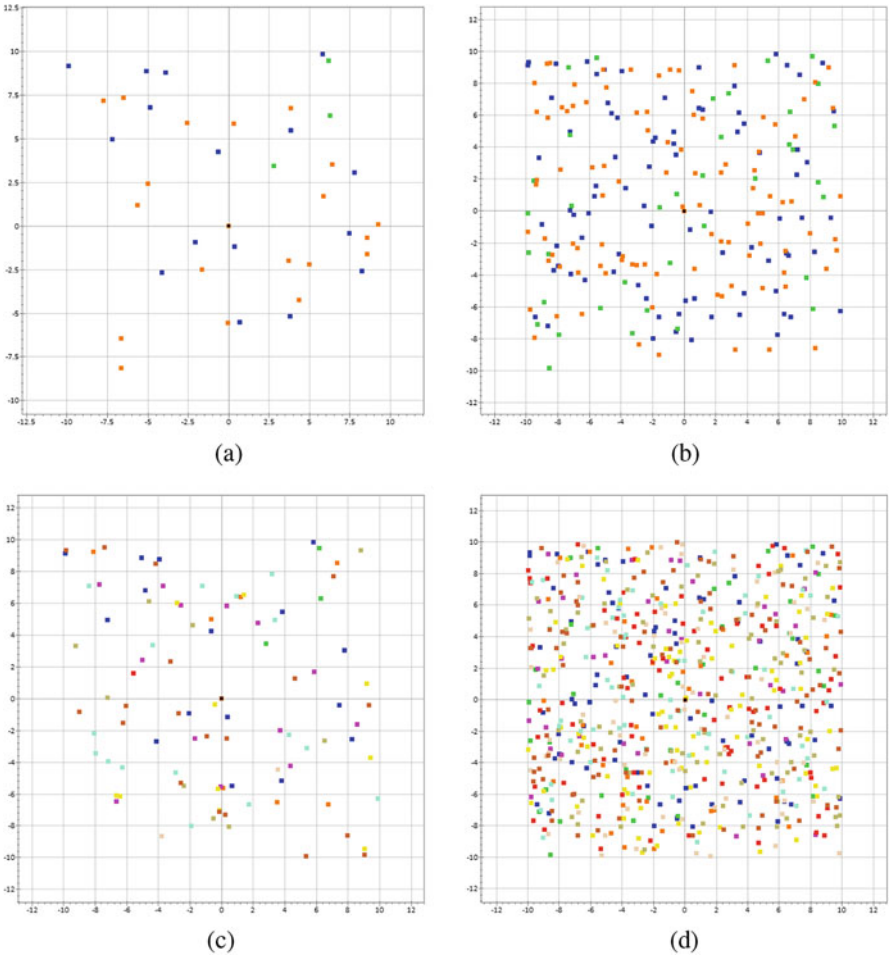


Fig. 3.1 The four constellations explored in this paper. See Section 3.2 for further information . (**a**) Few sets with few points. (**b**) Few sets with many points. (**c**) Many sets with few points. (**d**) Many sets with many points

3.3 The Four Feasibility Algorithms

We will numerically solve instances of (3.1) using four algorithms which we briefly review in this section. While there is a myriad of competing algorithms available, our selection consists of trustworthy "work horses" that have been employed elsewhere and for which the convergence theory in the *convex* case is fairly well understood. Each of these algorithms has a "tuning" parameter λ in the range $]0, 2[$. The *default value* λ_{dflt} is 1. Guided by experiments, we will also (numerically) look for the "best" value λ_{best}. We now turn to these four algorithms. Each algorithm will have a *governing sequence* driving the iteration, and a (possibly different) *monitored sequence* which is meant to find a solution of (3.1).

Cyclic Projections (CycP) Given $x_0 \in X$, the governing sequence is defined by

$$x_{k+1} := \big((1-\lambda)\mathrm{Id} + \lambda P_{C_m}\big) \circ \cdots \circ \big((1-\lambda)\mathrm{Id} + \lambda P_{C_1}\big)x_k. \tag{3.1}$$

The default parameter is $\lambda_{\text{dflt}} = 1$, from the range $]0, 2[$. The sequence monitored is $\big(\frac{1}{m}\sum_{i=1}^m P_{C_i}x_k\big)_{k\in\mathbb{N}}$. Selected references: [1, 2, 4, 10, 11, 13, 15].

Extrapolated Parallel Projections (ExParP) Given $x_0 \in X$, the governing and monitored sequence is defined by

$$x_{k+1} := x_k + \lambda \cdot \frac{\sum_{i=1}^m \|x_k - P_{C_i}x_k\|^2}{\|\sum_{i=1}^m (x_k - P_{C_i}x_k)\|^2} \sum_{i=1}^m (P_{C_i}x_k - x_k) \tag{3.2}$$

if $x_k \notin C$; $x_{k+1} = x_k$ otherwise. The default parameter is $\lambda_{\text{dflt}} = 1$, from the range $]0, 2[$. Selected references: [2, 3, 14].

Douglas–Rachford (DR) Given $x_0 \in X$, $\mathbf{x}_0 := (x_{0,1}, \ldots, x_{0,m}) = (x_0, \ldots, x_0) \in \mathbf{X} := X^m$, $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,m}) \in \mathbf{X}$, and $\bar{x}_k := \frac{1}{m}\sum_{i=1}^m x_{k,i}$, the next iterate is $\mathbf{x}_{k+1} = (x_{k+1,1}, \ldots, x_{k+1,m})$, where

$$(\forall i \in \{1, \ldots, m\}) \quad x_{k+1,i} := x_{k,i} + \lambda\big(P_{C_i}(2\bar{x}_k - x_{k,i}) - \bar{x}_k\big). \tag{3.3}$$

The default parameter is $\lambda_{\text{dflt}} = 1$, from the range $]0, 2[$. The sequence monitored is $(\bar{x}_k)_{k\in\mathbb{N}}$. Selected references: [2, 6, 17–19].

Cyclic Douglas–Rachford (CycDR) Given $x_0 \in X$, the governing sequence is defined by

$$x_{k+1} := \big((1-\tfrac{\lambda}{2})P_{C_m} + \tfrac{\lambda}{4}\big(\mathrm{Id} + R_{C_1}R_{C_m}\big)\big) \circ \cdots \circ \big((1-\tfrac{\lambda}{2})P_{C_2} + \tfrac{\lambda}{4}\big(\mathrm{Id} + R_{C_3}R_{C_2}\big)\big) \circ \big((1-\tfrac{\lambda}{2})P_{C_1} + \tfrac{\lambda}{4}\big(\mathrm{Id} + R_{C_2}R_{C_1}\big)\big)x_k. \tag{3.4}$$

The default parameter is $\lambda_{\mathrm{dflt}} = 1$, from the range $]0, 2[$. The sequence monitored is $\big(\frac{1}{m}\sum_{i=1}^{m} P_{C_i}x_k\big)_{k\in\mathbb{N}}$. Selected references: [16, 20, 21]. (For other cyclic version of DR, see [7, 9]. Also, if $m = 2$ and $C_1 = C_2 = \{0\}$, then $(x_k)_{k\in\mathbb{N}} = ((\lambda/2)^k x_0)_{k\in\mathbb{N}}$ is actually unbounded when $x_0 \neq 0$ and $\lambda > 2$.)

3.4 Setting Up the Numerical Explorations

Stopping Criteria The feasibility measure

$$d : X \to \mathbb{R}_+ : x \mapsto \sqrt{\frac{\sum_{i=1}^{m}\|x - P_{C_i}x\|^2}{\sum_{i=1}^{m}\|x_0 - P_{C_i}x_0\|^2}},$$

where $x_0 \in X \smallsetminus C$, vanishes exactly when $x \in C$. We stop each algorithm with monitored sequence $(y_k)_{k\in\mathbb{N}}$ either when

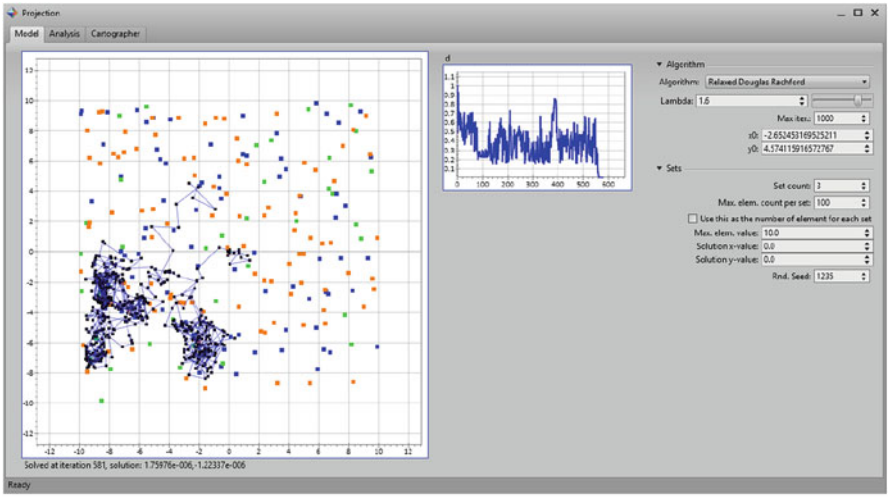
$$d(y_k) < \epsilon := 10^{-6}$$

or when the maximum number of iterations, which we set to 1000, is reached. These values were chosen to allow a reasonable exploration of the feasibility problem while maintaining computational efficiency.

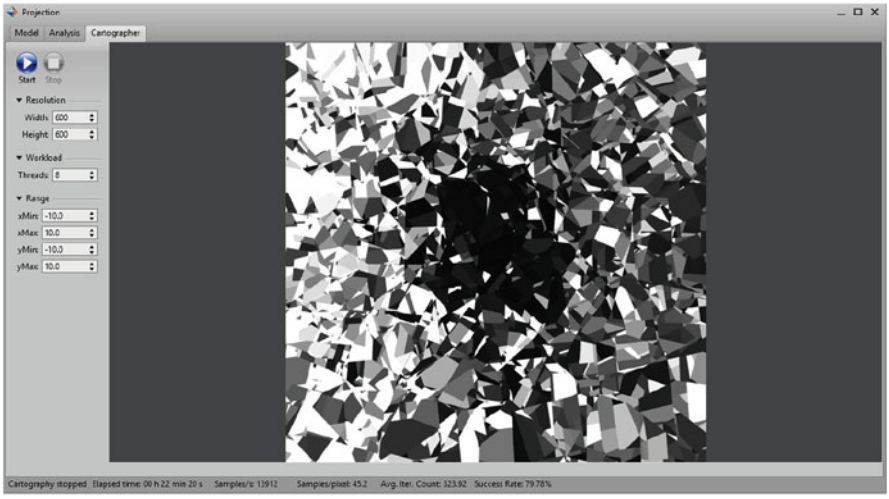
Details on Program A program was developed in C++ to run the different experiments, see Figure 3.2 for two screenshots which we describe next. In the main tab of the user interface one can select the algorithm to be used and set up the problem to be solved by choosing the number of sets and the maximum number of elements per set. By clicking on the diagram showing the current constellation of points, the user can select a starting point and immediately observe the resulting orbit being rendered over the constellation. The graph of the feasibility measure d, corresponding to the current orbit, is also displayed.

The Cartographer tab allows the exploration of a very large number of starting points to construct a picture of the performance of a given algorithm. This two-dimensional plot shows for each starting point the number of iterations required to solve the problem, ranging from zero (black) to the maximum number of iterations allowed (white). The plot is generated progressively and uses Quasi-Monte Carlo

sampling for the selection of the starting points. This is the most computationally intensive part of the software, and it is fully multi-threaded to take advantage of modern processor architectures.



(a)



(b)

Fig. 3.2 The software developed for this work. Setting the constellation of points and the algorithm to be used is done in the main tab, shown in (**a**). The generation of a performance plot is done in the cartographer tab, shown in (**b**)

3.5 Determining the "best" Parameter λ_{best}

In this section, we consider our four constellations (see Section 3.2) and run on each of them the four algorithms (see Section 3.3) with the parameter λ ranging over $]0, 2[$. The curves shown in Figures 3.3, 3.4, 3.5, and 3.6 give an estimate of the success rate of each algorithm, evaluated for 200 evenly-spaced values of λ. For each value of λ, 5000 starting points are drawn from $[-10, 10] \times [-10, 10]$ using Quasi-Monte Carlo sampling, and the success rate is estimated by dividing the number of times the algorithm is successful by this number of starting points. Thus, a "best" parameter λ_{best} is determined. It is this parameter that we will use to compare with the default parameter λ_{dflt}, which is 1 in all cases.

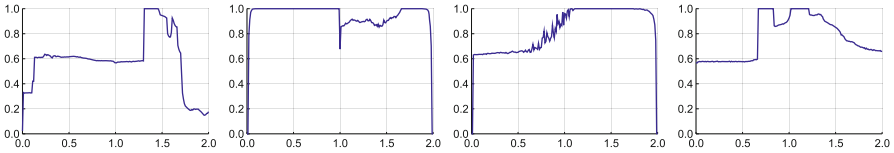


Fig. 3.3 Success rates in terms of λ for the few sets with few points constellation. (**a**) CycP. (**b**) ExParP. (**c**) DR. (**d**) CycDR

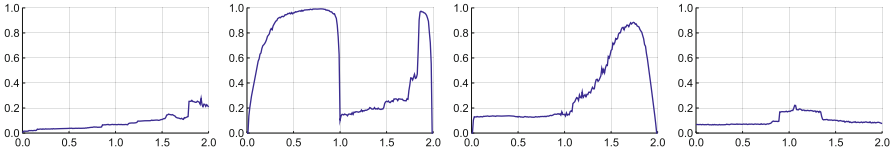


Fig. 3.4 Success rates in terms of λ for the few sets with many points constellation. (**a**) CycP. (**b**) ExParP. (**c**) DR. (**d**) CycDR

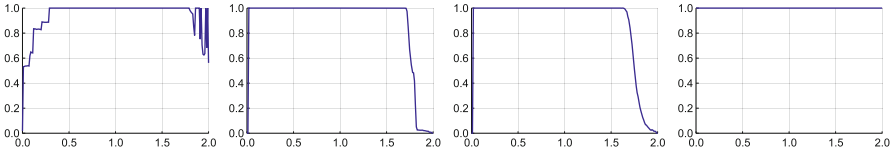


Fig. 3.5 Success rates in terms of λ for the many sets with few points constellation. (**a**) CycP. (**b**) ExParP. (**c**) DR. (**d**) CycDR

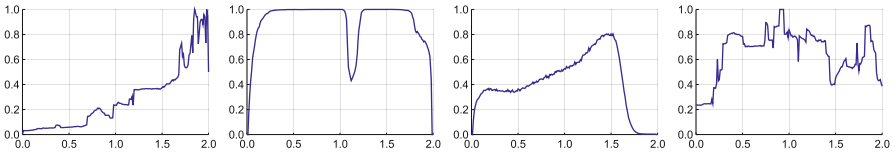


Fig. 3.6 Success rates in terms of λ for the many sets with many points constellation. (**a**) CycP. (**b**) ExParP. (**c**) DR. (**d**) CycDR

Algorithm	CycP	ExParP	DR	CycDR
λ_{dflt}	1.0	1.0	1.0	1.0
λ_{best}	1.5	0.8	1.6	1.2

Fig. 3.7 Best parameters λ_{best} chosen by inspecting the success rates curves

Discussion For each of the four constellations considered above, we visually inspected the λ-curves indicating success rates. We then picked for each algorithm a parameter called λ_{best} which improved performance over the default parameter $\lambda_{\text{dflt}} = 1$. The results are recorded in the table in Figure 3.7.

We will use the parameters λ_{best} for the experiments in subsequent sections.

3.6 Tracking Orbits

In this section, we consider our four given constellations (see Section 3.2). For each constellation, which is organized in a separate subsection, the same starting point is used. We then consider each of our four fixed algorithms (see Section 3.3) and show orbits for $\lambda_{\text{dflt}} = 1$ and for λ_{best} (see Section 3.5), and the corresponding feasibility measure d (see Section 3.4).

3.6.1 *Few Sets with Few Points*

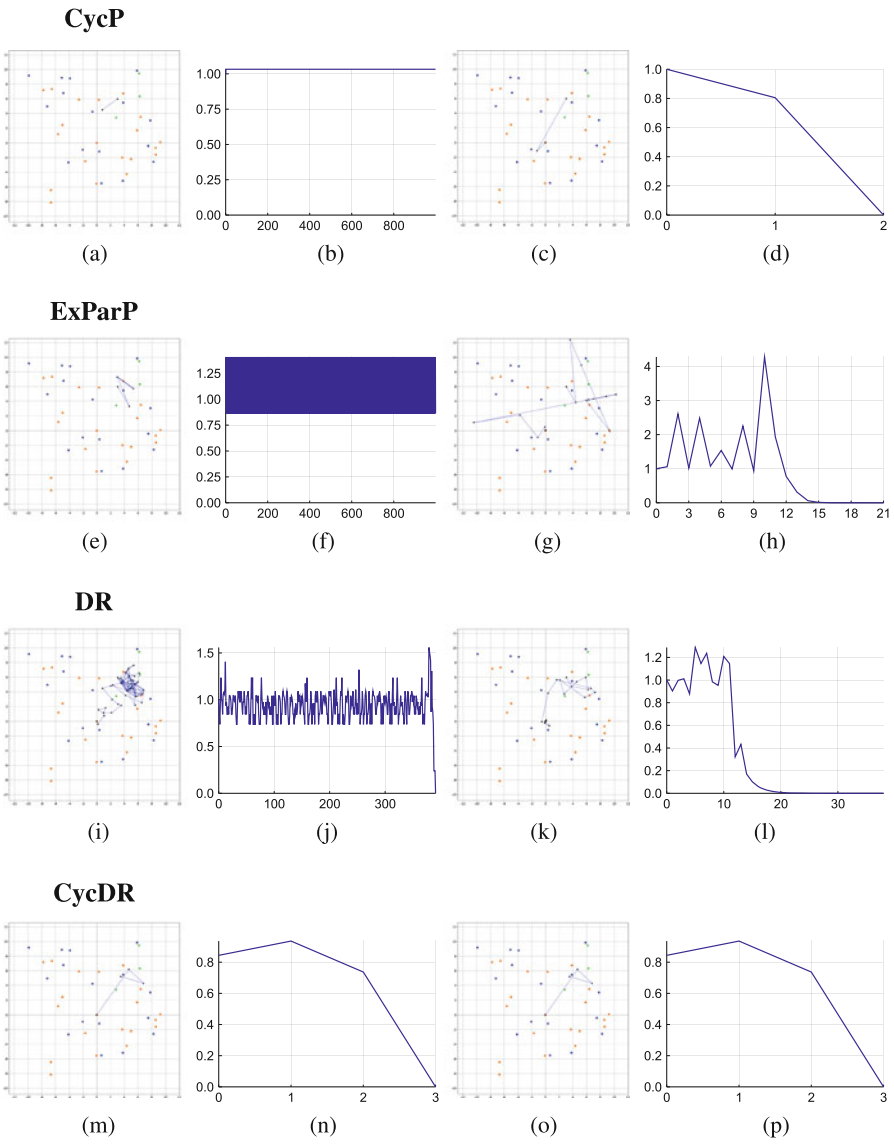


Fig. 3.8 Orbits and errors for CycP, ExParP, DR, and CycDR in the few sets with few points constellation. (**a**) λ_{dflt} orbit. (**b**) λ_{dflt} error. (**c**) λ_{best} orbit. (**d**) λ_{best} error. (**e**) λ_{dflt} orbit. (**d**) λ_{dflt} error. (**g**) λ_{best} orbit. (**h**) λ_{best} error. (**i**) λ_{dflt} orbit. (**j**) λ_{dflt} error. (**k**) λ_{best} orbit. (**l**) λ_{best} error. (**m**) λ_{dflt} orbit. (**n**) λ_{dflt} error. (**o**) λ_{best} orbit. (**p**) λ_{best} error

3.6.2 *Few Sets with Many Points*

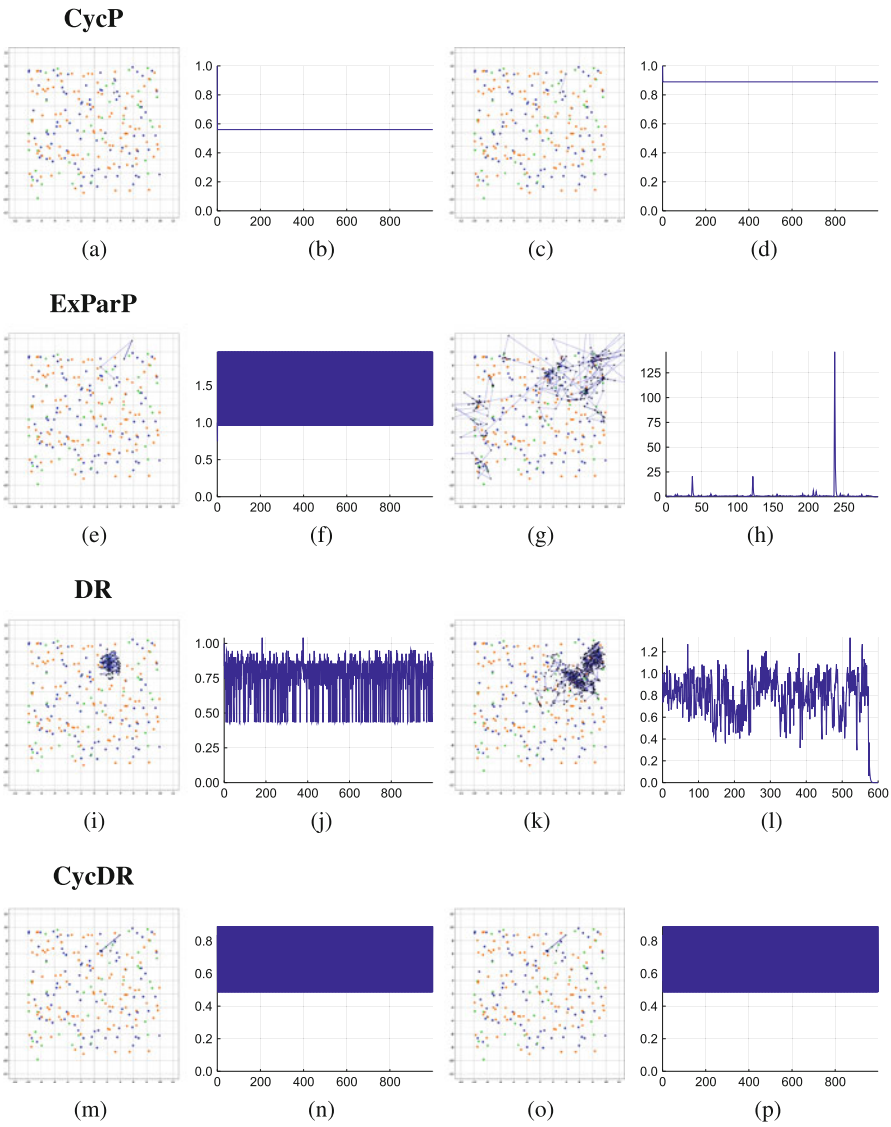


Fig. 3.9 Orbits and errors for CycP, ExParP, DR, and CycDR in the few sets with many points constellation. (**a**) λ_{dflt} orbit. (**b**) λ_{dflt} error. (**c**) λ_{best} orbit. (**d**) λ_{best} error. (**e**) λ_{dflt} orbit. (**f**) λ_{dflt} error. (**g**) λ_{best} orbit. (**h**) λ_{best} error. (**i**) λ_{dflt} orbit. (**j**) λ_{dflt} error. (**k**) λ_{best} orbit. (**l**) λ_{best} error. (**m**) λ_{dflt} orbit. (**n**) λ_{dflt} error. (**o**) λ_{best} orbit. (**p**) λ_{best} error

3.6.3 *Many Sets with Few Points*

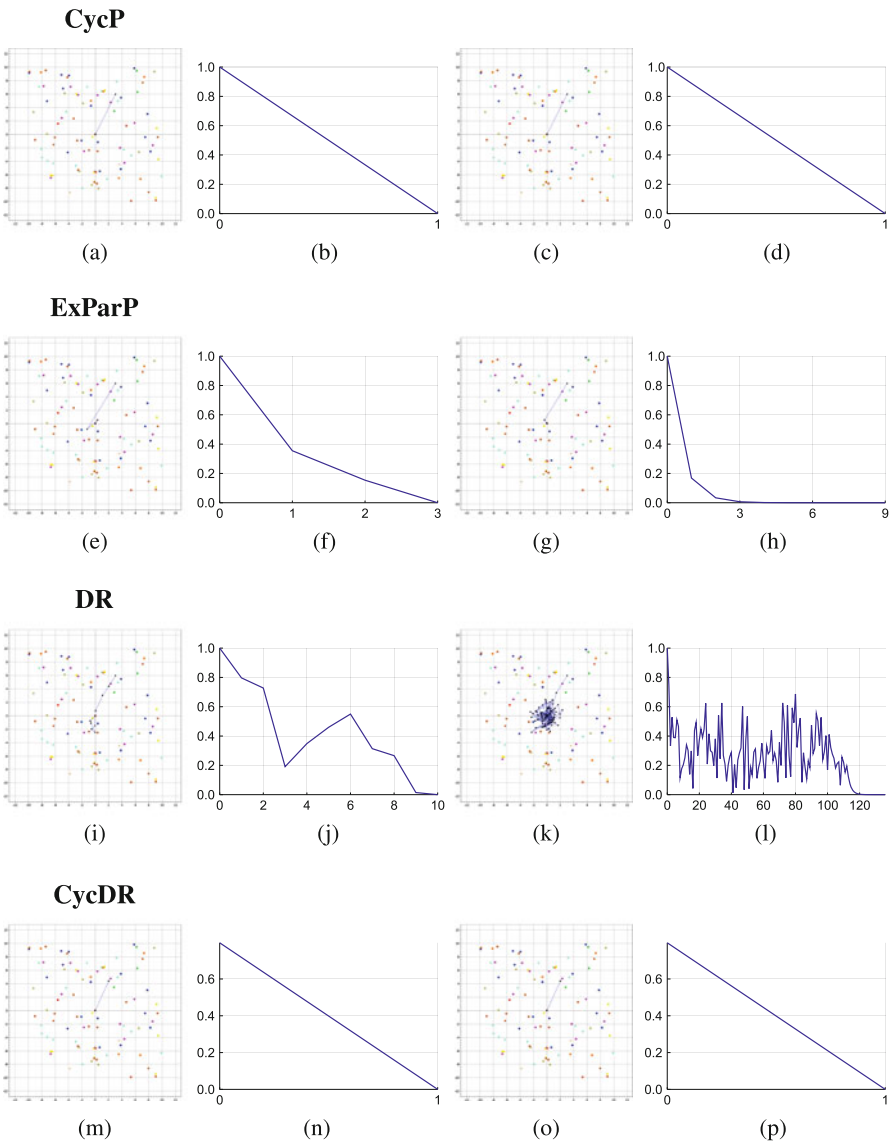


Fig. 3.10 Orbits and errors for CycP, ExParP, DR, and CycDR in the many sets with few points constellation. (**a**) λ_{dflt} orbit. (**b**) λ_{dflt} error. (**c**) λ_{best} orbit. (**d**) λ_{best} error. (**e**) λ_{dflt} orbit. (**f**) λ_{dflt} error. (**g**) λ_{best} orbit. (**h**) λ_{best} error. (**i**) λ_{dflt} orbit. (**j**) λ_{dflt} error. (**k**) λ_{best} orbit. (**l**) λ_{best} error. (**m**) λ_{dflt} orbit. (**n**) λ_{dflt} error. (**o**) λ_{best} orbit. (**p**) λ_{best} error

3.6.4 *Many Sets with Many Points*

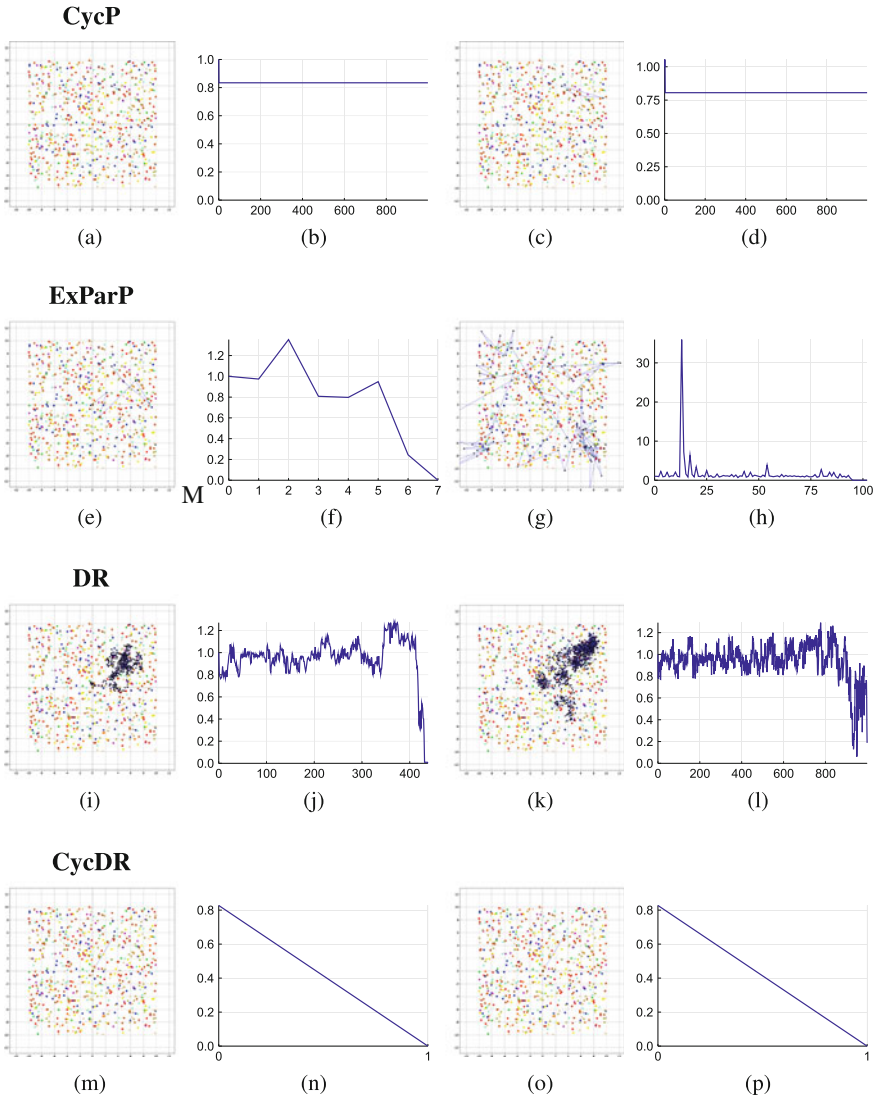


Fig. 3.11 Orbits and errors for CycP, ExParP, DR, and CycDR in the many sets with many points constellation. (**a**) λ_{dflt} orbit. (**b**) λ_{dflt} error. (**c**) λ_{best} orbit. (**d**) λ_{best} error. (**e**) λ_{dflt} orbit. (**f**) λ_{dflt} error. (**g**) λ_{best} orbit. (**h**) λ_{best} error. (**i**) λ_{dflt} orbit. (**j**) λ_{dflt} error. (**k**) λ_{best} orbit. (**l**) λ_{best} error. (**m**) λ_{dflt} orbit. (**n**) λ_{dflt} error. (**o**) λ_{best} orbit. (**p**) λ_{best} error

3.6.5 Discussion

The numerical results in this subsection suggest the following: The most challenging constellation is the one with few sets and many points. The least challenging constellation is the one with many sets and few points for which all algorithms are successful.

The worst algorithm is CycP. ExParP with λ_{best} solves all four constellations. DR solves all constellations but λ has to be chosen appropriately. CycDR works well with λ_{dflt} in terms of number of iterations required; however, it was not able to solve the constellation with few sets and many points.

The experiments in this section suggest that (i) ExParP, DR, and CycDR are algorithms worthwhile exploring and that (ii) experimenting with λ may lead to improved convergence.

Because the results in this section feature a *fixed* starting point, we will explore in the next section the four constellations for a *multitude* of starting points.

3.7 Local and Global Behaviour

In this section, we continue to consider our four constellations (see Section 3.2) which our four algorithms attempt to solve (see Section 3.3).

In contrast to Section 3.6 where we tracked a single orbit, we here illustrate *local* and *global* behaviour of the algorithms for a multitude of starting points, sampled from $[-10, 10] \times [-10, 10]$ and $[-100, 100] \times [-100, 100]$, respectively. We do this for $\lambda_{\text{dflt}} = 1$ and for λ_{best} (see the table in Figure 3.7 in Section 3.5);

For each starting point in the given range, these plots display as gray levels the number of iterations the algorithm needed in its attempt to solve the problem represented by the given constellation. Black corresponds to the minimum number of iterations (zero), and white to the maximum number of iterations (1000). The latter is obtained when the algorithm is unsuccessful. Therefore, the darker the image, the better the performance.

To quantitatively assess the performance of each algorithm, success rates are also provided. These are obtained by dividing the number of times the algorithm is successful by the number of starting points used.

Each of these images was generated using at least 15 million starting points. Depending on the constellation, the time required to generate these pictures ranged between a few minutes to about 3 hours using a quad-core computer.

3.7.1 Few Sets with Few Points

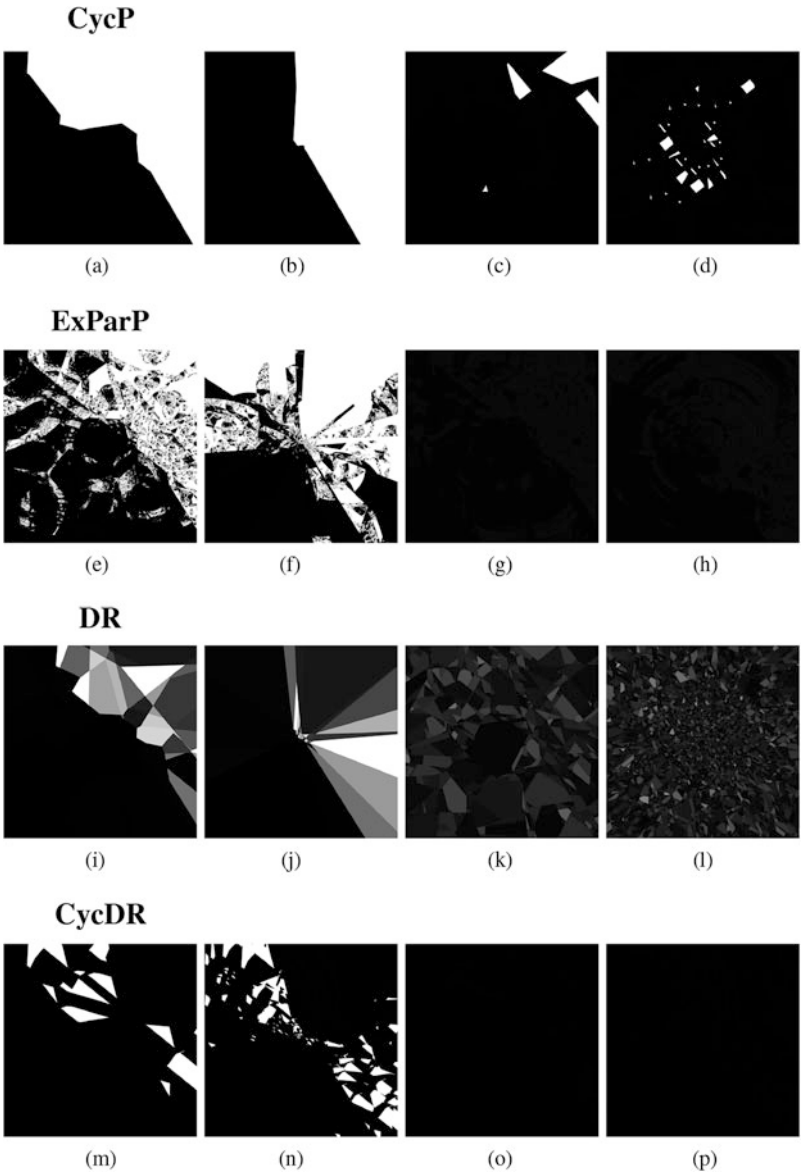


Fig. 3.12 Behaviour of CycP, ExParP, DR, and CycDR for the few sets with few points constellation (success rates indicated in parentheses). (**a**) λ_{dflt} local (57%). (**b**) λ_{dflt} global (57%). (**c**) λ_{best} local (95%). (**d**) λ_{best} global (98%). (**e**) λ_{dflt} local (68%). (**f**) λ_{dflt} global (52%). (**g**) λ_{best} local (100%). (**h**) λ_{best} global (100%). (**i**) λ_{dflt} local (96%). (**j**) λ_{dflt} global (94%). (**k**) λ_{best} local (100%). (**l**) λ_{best} global (100%). (**m**) λ_{dflt} local (93%). (**n**) λ_{dflt} global (91%). (**o**) λ_{best} local (100%). (**p**) λ_{best} global (100%)

3.7.2 Few Sets with Many Points

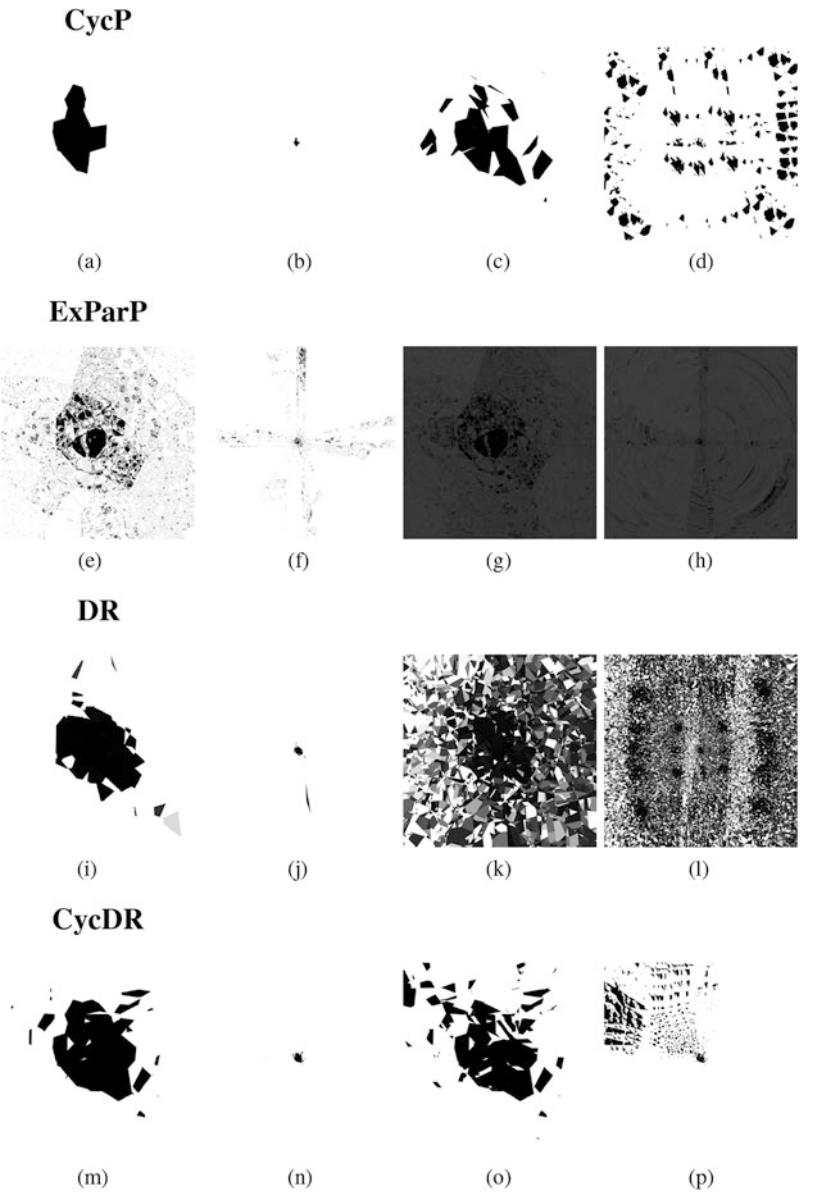


Fig. 3.13 Behaviour of CycP, ExParP, DR, and CycDR for the few sets with many points constellation (success rates indicated in parentheses). (**a**) λ_{dflt} local (6.8%). (**b**) λ_{dflt} global (0.1%). (**c**) λ_{best} local (11%). (**d**) λ_{best} global (12%). (**e**) λ_{dflt} local (10%). (**f**) λ_{dflt} global (0.9%). (**g**) λ_{best} local (99%). (**h**) λ_{best} global (99%). (**i**) λ_{dflt} local (15%). (**j**) λ_{dflt} global (0.2%). (**k**) λ_{best} local (80%). (**l**) λ_{best} global (81%). (**m**) λ_{dflt} local (17%). (**n**) λ_{dflt} global (0.2%). (**o**) λ_{best} local (18%). (**p**) λ_{best} global (4.9%)

3.7.3 Many Sets with Few Points

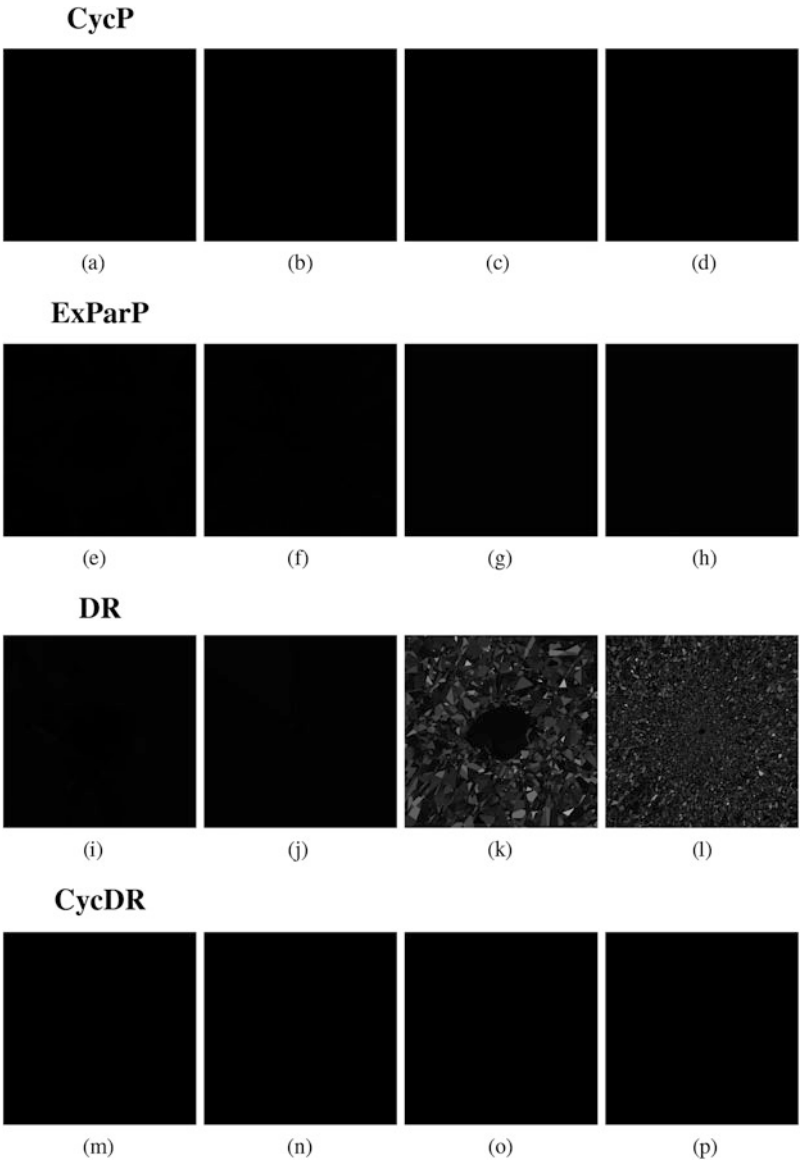


Fig. 3.14 Behaviour of CycP, ExParP, DR, and CycDR for the many sets with few points constellation (success rates indicated in parentheses). **(a)** λ_{dflt} local (100%). **(b)** λ_{dflt} global (100%). **(c)** λ_{best} local (100%). **(d)** λ_{best} global (100%). **(e)** λ_{dflt} local (100%). **(f)** λ_{dflt} global (100%). **(g)** λ_{best} local (100%). **(h)** λ_{best} global (100%). **(i)** λ_{dflt} local (100%). **(j)** λ_{dflt} global (100%). **(k)** λ_{best} local (100%). **(l)** λ_{best} global (100%). **(m)** λ_{dflt} local (100%). **(n)** λ_{dflt} global (100%). **(o)** λ_{best} local (100%). **(p)** λ_{best} global (100%)

3.7.4 Many Sets with Many Points

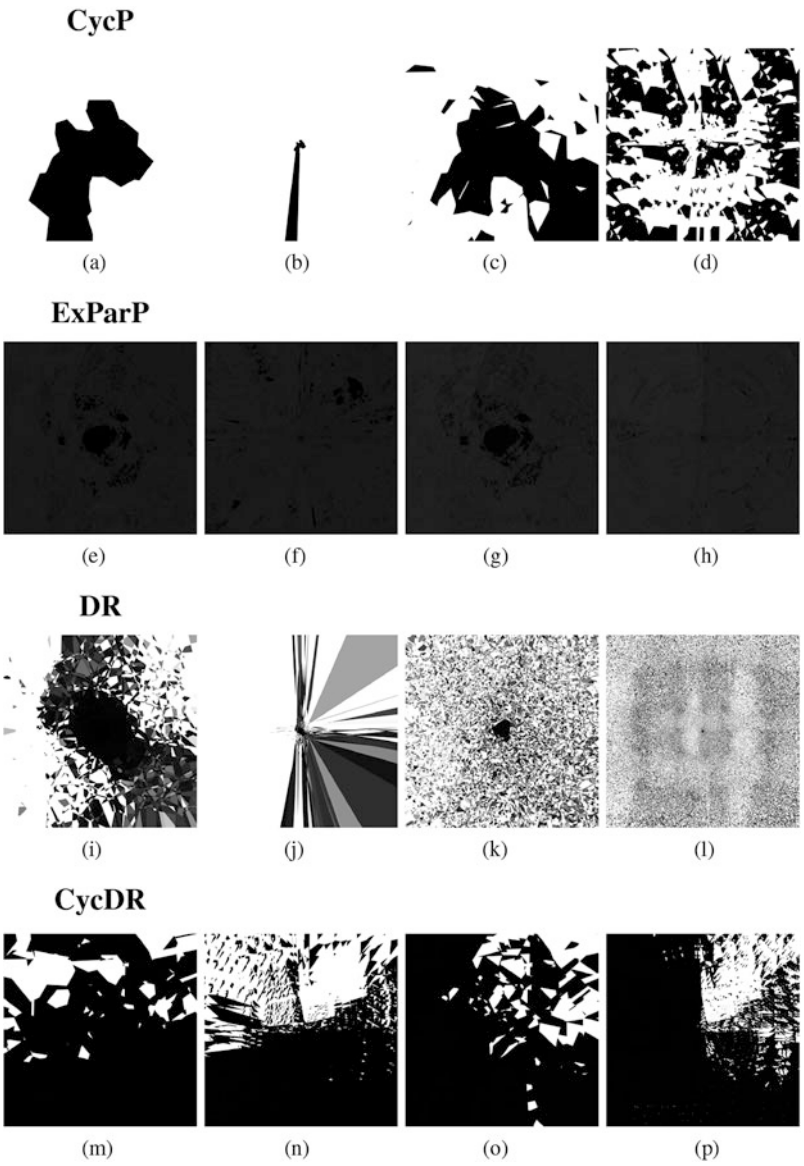


Fig. 3.15 Behaviour of CycP, ExParP, DR, and CycDR for the many sets with many points constellation (success rates indicated in parentheses). (**a**) λ_{dflt} local (24%). (**b**) λ_{dflt} global (2.2%). (**c**) λ_{best} local (38%). (**d**) λ_{best} global (47%). (**e**) λ_{dflt} local (100%). (**f**) λ_{dflt} global (100%). (**g**) λ_{best} local (100%). (**h**) λ_{best} global (100%). (**i**) λ_{dflt} local (53%). (**j**) λ_{dflt} global (40%). (**k**) λ_{best} local (56%). (**l**) λ_{best} global (57%). (**m**) λ_{dflt} local (83%). (**n**) λ_{dflt} global (66%). (**o**) λ_{best} local (84%). (**p**) λ_{best} global (82%)

3.7.5 Discussion

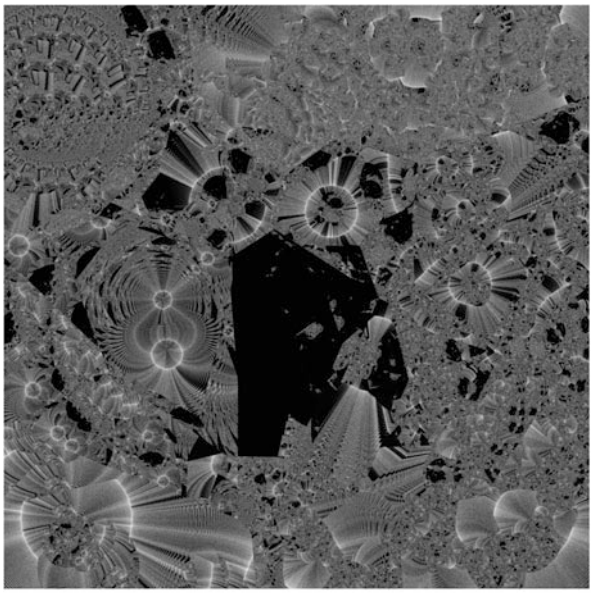
Comparing the success rates reported in the figures above, it appears that ExParP, DR, and CycDR are good choices; we recommend that CycP be not used. The effect of the tuning parameter λ is very striking for most algorithms when comparing performance of λ_{dflt} with λ_{best}.

3.8 Divertissements

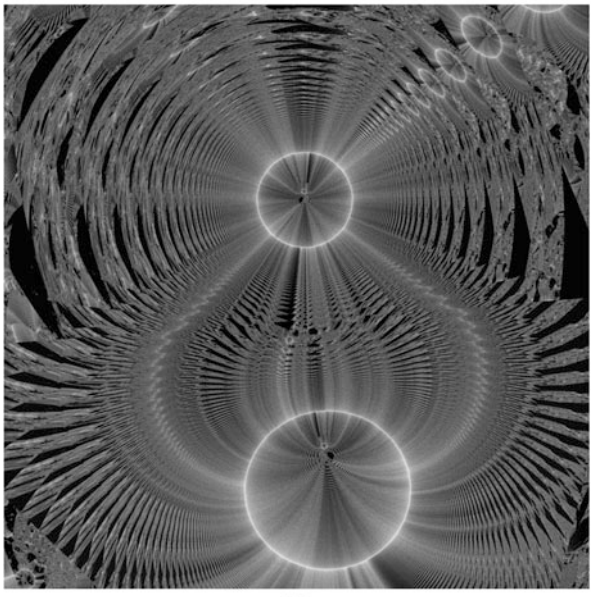
We experimented also with other constellations and encountered some interesting behaviour of ExParP. This algorithm seems to exhibit fractal-like behaviour for some constellations—whether they are created randomly or not. In the following, we present three images that we found particularly delightful in Figures 3.16 and 3.17.

3.9 Concluding Remarks

We encountered a somewhat surprising complexity in the behaviour of four algorithms for solving feasibility problems in a simple nonconvex case. The importance of the tuning parameter λ is apparent as is the proximity to solutions (local vs global behaviour). Further studies are needed to find effective guidelines for users in terms of choice of algorithms and the choice of the parameter λ. Finally, and similarly to [8], we encountered *beauty* in our numerical explorations. It is our hope that others will join us and explore theoretically and numerically this fascinating universe of constellations.



(a)



(b)

Fig. 3.16 Shown in (**a**) is the performance of ExParP on a constellation consisting of 3 sets with 20 points each, with $\lambda = 0.998$, within the region $[-10, 10] \times [-10, 10]$. A close-up of the centre-left region of (**a**) is presented in (**b**)

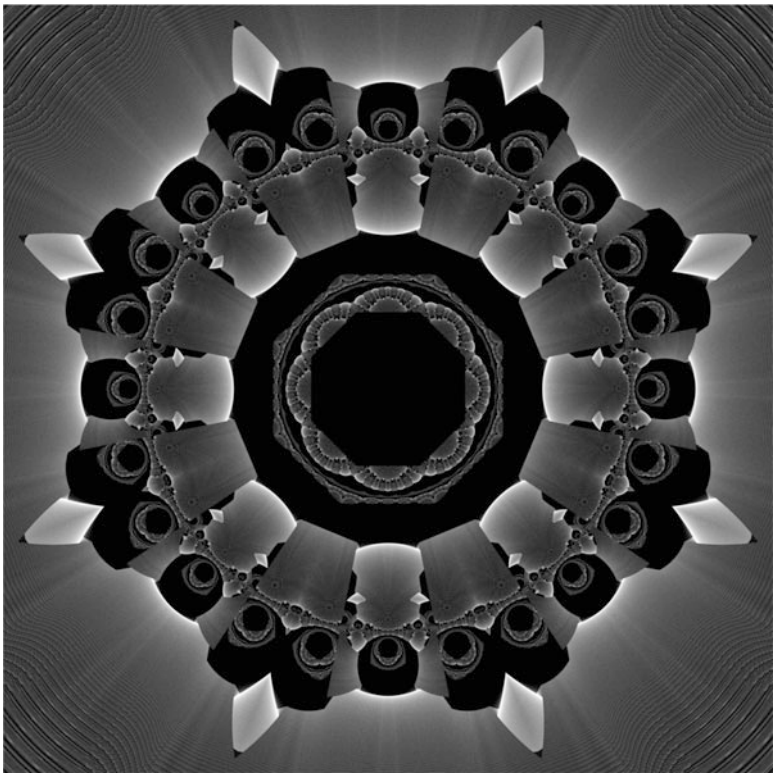


Fig. 3.17 ExParP for a constellation with $\lambda = 0.995$, consisting of 2 subsets of concentric circles centred at the origin, with radii 4 and 8, containing 8 and 16 equispaced points, respectively

Acknowledgements We thank the referee for constructive comments and suggestions. The research of HHB was partially supported by NSERC.

References

1. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. *SIAM Rev.* **38**, 367–426 (1996)
2. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, second edition. Springer, Cham (2017)
3. Bauschke, H.H., Combettes, P.L., Kruk, S.G.: Extrapolation algorithm for affine-convex feasibility problems. Numer. Algorithms **41**, 239–274 (2006)
4. Bauschke, H.H., Koch, V.R.: Projection methods: Swiss army knives for solving feasibility and best approximation problems with halfspaces. Contemp. Math. **636**, 1–40 (2015) doi: 10.1090/conm/636/12726
5. Bauschke, H.H., Dao, M.N., Lindstrom, S.B.: The Douglas–Rachford algorithm for a hyperplane and a doubleton. J. Glob. Optim. **74**, 79–93 (2019)

6. Bauschke, H.H., Moursi, W.M.: On the Douglas–Rachford algorithm. Math. Prog. (Ser. A) **164**, 263–284 (2017)
7. Bauschke, H.H., Noll, D., Phan, H.M.: Linear and strong convergence of algorithms involving averaged nonexpansive operators. J. Math. Anal. Appl. **421**, 1–20 (2015)
8. Borwein, J.M., Lindstrom, S.B., Sims, B., Schneider, A., Skerritt, M.P.: Dynamics of the Douglas–Rachford method for ellipses and p-spheres. Set-Valued Var. Anal. **26**, 385–403 (2018)
9. Borwein, J.M., Tam, M.K.: A cyclic Douglas–Rachford iteration scheme. J. Optim. Th. Appl. **160**, 1–29 (2014)
10. Cegielski, A.: Iterative methods for fixed point problems in Hilbert spaces. Springer, Heidelberg (2012)
11. Censor, Y., Chen, W., Combettes, P.L., Davidi, R., Herman, G.T.: On the effectiveness of projection methods for convex feasibility problems Extrapolation algorithm for affine-convex feasibility problems. Numer. Algorithms **41**, 239–274 (2006)
12. Censor, Y., Zaknoon, M.: Algorithms and convergence results of projection methods for inconsistent feasibility problems: a review. Pure Appl. Funct. Anal. **3**, 565–586 (2018) https://arxiv.org/abs/1802.07529 [math.OC] (2018)
13. Censor, Y., Zenios, S.A.: Parallel Optimization. Oxford University Press (1997)
14. Combettes, P.L.: Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections. IEEE Trans. Image Process. **6**, 493–506 (1997)
15. Combettes, P.L.: Hilbertian convex feasibility problems: convergence of projection methods. Appl. Math. Optim. **35**, 311–330 (1997)
16. Dao, M., Phan, H.M.: Linear convergence of the generalized Douglas–Rachford algorithm for feasibility problems. J. Glob. Optim. **72**, 443–474 (2018)
17. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Prog. (Ser. A) **55**, 293–318 (1992)
18. Elser, V., Rankenburg, I., Thibault, P.: Searching with iterated maps. Proc. Natl. Acad. Sci. U.S.A. **104**, 418–423 (2007)
19. Lions, P.-L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
20. Luke, D.R.: Finding best approximation pairs relative to a convex and a prox-regular set in a Hilbert space. SIAM J. Optim. **19**, 714–739 (2008)
21. Luke, D.R., Sabach, S., Teboulle, M.: Optimization on spheres: models and proximal algorithms with computational performance comparisons. https://arxiv.org/abs/1810.02893 [math.OC] (2018)

Chapter 4
Variable Metric ADMM for Solving Variational Inequalities with Monotone Operators over Affine Sets

Radu Ioan Boţ, Ernö Robert Csetnek, and Dennis Meier

Abstract We propose an iterative scheme for solving variational inequalities with monotone operators over affine sets in an infinite dimensional Hilbert space setting. We show that several primal-dual algorithms in the literature as well as the classical ADMM algorithm for convex optimization problems, together with some of its variants, are encompassed by the proposed numerical scheme. Furthermore, we carry out a convergence analysis of the generated iterates and provide convergence rates by using suitable dynamical step sizes together with variable metric techniques.

Keywords ADMM algorithm · Primal-dual algorithm · Monotone operators · Convex optimization

AMS 2010 Subject Classification 47H05, 65K05, 90C25

4.1 Introduction

Many problems in fields like signal and image processing, portfolio optimization, cluster analysis, location theory, network communication and machine learning as well as inverse problems can be formulated as a convex optimization problem of the form

$$\inf_{x\in\mathcal{H},z\in\mathcal{G}}\{f(x)+h(x)+g(z)\},\quad \text{s.t. } L_1x+L_2z=d \tag{4.1}$$

where $\mathcal{H}$, $\mathcal{G}$ and $\mathcal{Z}$ are real Hilbert spaces, $f:\mathcal{H}\to\overline{\mathbb{R}}:=\mathbb{R}\cup\{\pm\infty\}$ and $g:\mathcal{G}\to\overline{\mathbb{R}}$ are proper, convex and lower semicontinuous functions, $h:\mathcal{H}\to\mathbb{R}$ is

R. I. Boţ (✉) · E. R. Csetnek · D. Meier
Faculty of Mathematics, University of Vienna, Vienna, Austria
e-mail: radu.bot@univie.ac.at; ernoe.robert.csetnek@univie.ac.at; meierd61@univie.ac.at

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_4

a convex and Fréchet differentiable function with Lipschitz continuous gradient, $L_1 : \mathcal{H} \to \mathcal{Z}$, $L_2 : \mathcal{G} \to \mathcal{Z}$ are linear continuous operators and $d \in \mathcal{Z}$.

One of the most prominent numerical algorithms one can find in the literature for solving optimization problems of the form (4.1) is the *alternating direction method of multipliers (ADMM)*. In the case $h = 0$, which represents the standard setting in the literature addressing ADMM methods, the augmented Lagrangian associated with problem (4.1) is given for a fixed real number $c > 0$ as

$$L_c : \mathcal{H} \times \mathcal{G} \times \mathcal{Z} \to \overline{\mathbb{R}},$$

$$L_c(x, z, y) = f(x) + g(z) + \langle y, L_1 x + L_2 z - d \rangle + \frac{c}{2}\|L_1 x + L_2 z - d\|^2.$$

The ADMM algorithm generates a sequence $(x^k, z^k, y^k)_{k\geq 0} \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$ by iterating for every $k \geq 0$

$$x^{k+1} \in \operatorname*{arg\,min}_{x\in\mathcal{H}} L_c(x, z^k, y^k) = \operatorname*{arg\,min}_{x\in\mathcal{H}} \left\{ f(x) + \frac{c}{2}\|L_1 x + L_2 z^k - d + c^{-1} y^k\|^2 \right\}$$

$$z^{k+1} \in \operatorname*{arg\,min}_{z\in\mathcal{G}} L_c(x^{k+1}, z, y^k) = \operatorname*{arg\,min}_{x\in\mathcal{G}} \left\{ g(z) + \frac{c}{2}\|L_1 x^{k+1} + L_2 z - d + c^{-1} y^k\|^2 \right\}$$

$$y^{k+1} = y^k + c(L_1 x^{k+1} + L_2 z^{k+1} - d).$$

Since the function f and the operator L_1 are not evaluated independently in the first line of the algorithm, the minimization with respect to the variable x does not lead to a proximal step (the same is true for the second minimization). This results in less attractiveness for implementations than for primal-dual splitting algorithms, which represent the second class of prominent iterative methods for solving (4.1). This drawback has been overcome in the literature by introducing a suitable regularizer equipped with a (semi-)metric, see, for example, [10] for a finite dimensional approach (in case $\mathcal{G} = \mathcal{Z}$, $L_2 = -\mathrm{Id}$ and $d = 0$ see also [15], and also [1] for an extension of the ADMM algorithm by involving also smooth parts in the objective, by employing variable metrics and by working in an infinite dimensional Hilbert setting). This so-called alternating direction *proximal* method of multipliers (AD-PMM) reveals a bridge which connects the classical ADMM algorithm with primal-dual methods. This observation served as the starting point for the investigations made in [4], where a generalization of the AD-PMM algorithm to monotone inclusions was proposed and investigated from the point of view of its convergence properties.

In this paper we propose an iterative algorithm for solving variational inequalities with monotone operators of the type

$$\text{find } (x, z) \in \mathcal{H} \times \mathcal{G} \text{ such that } 0 \in (Ax + Cx) \times Bz + N_S(x, z),$$

where $A : \mathcal{H} \rightrightarrows \mathcal{H}$ and $B : \mathcal{G} \rightrightarrows \mathcal{G}$ are maximally monotone operators, $C : \mathcal{H} \to \mathcal{H}$ is an η-cocoercive operator, for $\eta \geq 0$, $S := \{(x, z) \in \mathcal{H} \times \mathcal{G} :$

$L_1x + L_2z = d\}$, and N_S denotes the normal cone operator to the set S. This delivers a unifying framework for solving monotone inclusions in Hilbert spaces which encompasses in particular the ADMM algorithm in [4], several primal-dual iterative methods [5, 7, 9, 16] as well as the classical ADMM algorithm designed to solve problems of type (4.1) (and its variants from [10, 15], see also [6, 11, 12]). After giving the necessary preliminaries, we formulate the ADMM iterative scheme for variational inequalities and carry out a convergence analysis. Furthermore, under additional strong monotonicity assumptions, we derive convergence rates for the primal iterates by using a dynamic step size strategy combined with variable metric techniques.

4.2 Notation and Preliminaries

Throughout, $\mathcal{H}$, $\mathcal{G}$ and $\mathcal{Z}$ denote real Hilbert spaces with scalar products $\langle \cdot, \cdot \rangle$ and associated norms $\| \cdot \|$ (since there is no risk of confusion, they are denoted in the same way). Let $M : \mathcal{H} \rightrightarrows \mathcal{H}$ be an arbitrary set-valued operator. We denote by $\operatorname{gra} M := \{(x, u) \in \mathcal{H} \times \mathcal{H} : u \in Mx\}$ its graph, by $\operatorname{dom} M := \{x \in \mathcal{H} : Mx \neq \emptyset\}$ its domain and by $M^{-1} : \mathcal{H} \rightrightarrows \mathcal{H}$ its inverse operator, defined by $(u, x) \in \operatorname{gra} M^{-1}$ if and only if $(x, u) \in \operatorname{gra} M$. M is said to be monotone, if $\langle x - y, u - v \rangle \geq 0$ for all $(x, u), (y, v) \in \operatorname{gra} M$. A monotone operator M is said to be maximally monotone, if there exists no proper monotone extension of the graph of M on $\mathcal{H} \times \mathcal{H}$. For an arbitrary $\gamma > 0$, the operator M is called γ-strongly monotone, if $\langle x - y, u - v \rangle \geq \gamma \|x - y\|^2$ for all $(x, u), (y, v) \in \operatorname{gra} M$.

The resolvent of M is the mapping $J_M : \mathcal{H} \rightrightarrows \mathcal{H}$, defined by $J_A := (\mathrm{Id} + M)^{-1}$. If M is maximally monotone, then $J_M : \mathcal{H} \to \mathcal{H}$ is single-valued and maximally monotone (see [2, Proposition 23.7 and Corollary 23.10]). Furthermore, for an arbitrary $\gamma > 0$ we have (see [2, Proposition 23.18])

$$J_{\gamma M} + \gamma J_{\gamma^{-1} M^{-1}} \circ \gamma^{-1} \mathrm{Id} = \mathrm{Id}, \tag{4.2}$$

where Id denotes the identity operator on $\mathcal{H}$.

For a linear continuous operator $L : \mathcal{H} \to \mathcal{G}$, its adjoint operator $L^* : \mathcal{G} \to \mathcal{H}$ is defined by $\langle L^* y, x \rangle = \langle y, Lx \rangle$ for all $(x, y) \in \mathcal{H} \times \mathcal{G}$. The norm of L is defined by $\|L\| := \sup\{\|Lx\| : x \in \mathcal{H}, \|x\| \leq 1\}$. The linear operator L is said to be skew, if $\langle x, Lx \rangle = 0$ for all $x \in \mathcal{H}$. A single-valued operator $M : \mathcal{H} \to \mathcal{H}$ is said to be β-cocoercive, for $\beta \geq 0$, if $\beta \langle x - y, Mx - My \rangle \geq \|Mx - My\|^2$ for all $(x, y) \in \mathcal{H} \times \mathcal{H}$. Moreover, M is β-Lipschitz continuous, if $\|Mx - My\| \leq \beta \|x - y\|$ for all $(x, y) \in \mathcal{H} \times \mathcal{H}$.

We write

$$\mathcal{S}_+(\mathcal{H}) := \{L : \mathcal{H} \to \mathcal{H} : L \text{ is linear, bounded, positive semidefinite and } L = L^*\}.$$

The Loewner partial ordering on $\mathcal{S}_+(\mathcal{H})$ is defined by

$$(\forall U \in \mathcal{S}_+(\mathcal{H}))(\forall V \in \mathcal{S}_+(\mathcal{H})) \qquad U \succeq V :\Leftrightarrow (\forall x \in \mathcal{H}) \ \ \langle x, Ux\rangle \geq \langle x, Vx\rangle.$$

Further, for every $U \in \mathcal{S}_+(\mathcal{H})$, we define a semi-scalar product and a semi-norm by

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \qquad \langle x, y\rangle_U := \langle x, Uy\rangle \quad \text{and} \quad \|x\|_U := \sqrt{\langle x, Ux\rangle},$$

respectively. For $\alpha > 0$ we set

$$\mathcal{P}_\alpha(\mathcal{H}) := \{U \in \mathcal{S}_+(\mathcal{H}) \ \mid \ U \succeq \alpha \,\mathrm{Id}\}.$$

Since we will also address convex optimization problems, we recall some elements of convex analysis. For a function $f : \mathcal{H} \to \overline{\mathbb{R}}$ we denote by $\mathrm{dom}\, f := \{x \in \mathcal{H} : f(x) < +\infty\}$ its effective domain and say that f is proper, if $\mathrm{dom}\, f \neq \emptyset$ and $f(x) \neq -\infty$ for all $x \in \mathcal{H}$. The (convex) conjugate function $f^* : \mathcal{H} \to \overline{\mathbb{R}}$ of f is defined by $f^*(u) := \sup_{x\in\mathcal{H}}\{\langle u, x\rangle - f(x)\}$ for all $u \in \mathcal{H}$. The (convex) subdifferential $\partial f : \mathcal{H} \rightrightarrows \mathcal{H}$ of f is given by

$$\partial f(x) := \{p \in \mathcal{H} : f(y) - f(x) \geq \langle p, y - x\rangle \ \ \forall y \in \mathcal{H}\},$$

for $x \in \mathcal{H}$ with $f(x) \in \mathbb{R}$ and $\partial f(x) = \emptyset$, otherwise. In case f is a proper, convex and lower semi-continuous function, $\partial f : \mathcal{H} \rightrightarrows \mathcal{H}$ is a maximally monotone operator [14].

For $f, g : \mathcal{H} \to \overline{\mathbb{R}}$ two proper functions, the infimal convolution $f\square g : \mathcal{H} \to \overline{\mathbb{R}}$ is defined by $(f\square g)(x) = \inf_{u\in\mathcal{H}}\{f(u) + g(x - u)\}$ for all $x \in \mathcal{H}$.

For a proper, convex and lower semi-continuous function $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $\gamma > 0$, for every $x \in \mathcal{H}$ we denote by $\mathrm{prox}_{\gamma f}(x)$ the proximal point of parameter γ of f at x, which is defined by

$$\mathrm{prox}_{\gamma f}(x) := \mathrm{argmin}_{y\in\mathcal{H}}\left\{f(y) + \frac{1}{2\gamma}\|y - x\|^2\right\}.$$

Since $J_{\gamma\partial f} = (\mathrm{Id} + \gamma\partial f)^{-1} = \mathrm{prox}_{\gamma f}$, this gives a single-valued operator $\mathrm{prox}_{\gamma f} : \mathcal{H} \to \mathcal{H}$ fulfilling the extended Moreau's decomposition formula

$$\mathrm{prox}_{\gamma f} + \gamma\, \mathrm{prox}_{\gamma^{-1} f^*} \circ \gamma^{-1}\,\mathrm{Id} = \mathrm{Id}\,.$$

Last, for a nonempty convex subset S of $\mathcal{H}$ and for $x \in \mathcal{H}$, the normal cone to S at x is

$$N_S(x) = \begin{cases} \{u \in \mathcal{H} \mid \sup\langle s - x, u\rangle \leq 0 \ \forall s \in S\}, & \text{if } x \in S \\ \emptyset, & \text{if } x \notin S \end{cases}.$$

4.3 A Variable Metric ADMM for Monotone Operators

In this section we present the variational inequality problem to solve, formulate the iterative numerical scheme and prove convergence for the sequence of generated iterates.

4.3.1 Problem Formulation and Algorithm

We start by describing the problem under investigation.

Problem 4.1 Let $\mathcal{H}$, $\mathcal{G}$ and $\mathcal{Z}$ be real Hilbert spaces, $A : \mathcal{H} \rightrightarrows \mathcal{H}$ and $B : \mathcal{G} \rightrightarrows \mathcal{G}$ be maximally monotone operators and $C : \mathcal{H} \to \mathcal{H}$ an η-cocoercive operator, for $\eta \geq 0$. Further, let $L_1 : \mathcal{H} \to \mathcal{Z}$ and $L_2 : \mathcal{G} \to \mathcal{Z}$ be linear continuous operators and $S := \{(x, z) \in \mathcal{H} \times \mathcal{G} : L_1 x + L_2 z = d\}$. To solve is the variational inequality with monotone operators over the set S

$$\text{find } (x, z) \in \mathcal{H} \times \mathcal{G} \text{ such that } 0 \in (Ax + Cx) \times Bz + N_S(x, z), \tag{4.3}$$

which can be reformulated as

$$\text{find } (x, z) \in S \text{ such that } \exists (p, q) \in -(Ax + Cx) \times (-Bz) \text{ with the property}$$
$$\langle (p, q), (u, v) - (x, z) \rangle \leq 0 \; \forall (u, v) \in S.$$

We will propose an algorithm for determining the KKT points associated to the variational inequality (4.3), namely, those $(x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$ which fulfill

$$-L_1^* y \in Ax + Cx, \;\; -L_2^* y \in Bz \text{ and } L_1 x + L_2 z = d. \tag{4.4}$$

Remark 4.1 If $(x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$ is a KKT point of (4.3), then, obviously,

$$(-L_1^* y, -L_2^* y) \in (Ax + Cx) \times Bz + N_S(x, z),$$

which means that (x, z) is a solution of (4.3).

On the other hand, if $(x, z) \in \mathcal{H} \times \mathcal{G}$ is a solution of (4.3), then there exists $(p, q) \in -(Ax + Cx) \times (-Bz)$ such that

$$L_1 x + L_2 z = d \text{ and } (x, z) \in \arg\min_{(u,v) \in S} \langle (-p, -q), (u, v) \rangle.$$

Using duality theory, we obtain under suitable constraint qualifications the existence of $y \in \mathcal{Z}$ such that

$$\langle(-p,-q),(x,z)\rangle = \inf_{(u,v)\in\mathcal{H}\times\mathcal{G}}\{\langle(-p,-q),(u,v)\rangle + \langle y, L_1u + L_2v - d\rangle\}$$
$$= \inf_{u\in\mathcal{H}}\langle u, -p + L_1^*y\rangle + \inf_{v\in\mathcal{G}}\langle z, -q + L_2^*y\rangle - \langle y, d\rangle.$$

Since the term on the left-hand side is finite, this holds only when $p = L_1^*y$ and $q = L_2^*y$. In other words, (x, z, y) is a KKT point of (4.3).

Next, we relate Problem 4.1 to a particular convex optimization problem with affine constraints.

Problem 4.2 Let $\mathcal{H}$, $\mathcal{G}$ and $\mathcal{Z}$ real Hilbert spaces, $f : \mathcal{H} \to \overline{\mathbb{R}}$, $g : \mathcal{G} \to \overline{\mathbb{R}}$ be proper, convex and lower semicontinuous functions, $h : \mathcal{H} \to \mathbb{R}$ a convex and Fréchet differentiable function with η-Lipschitz continuous gradient, for $\eta \geq 0$, $L_1 : \mathcal{H} \to \mathcal{Z}$, $L_2 : \mathcal{G} \to \mathcal{Z}$ linear continuous operators and $d \in \mathcal{Z}$. We consider the convex optimization problem

$$\inf_{x\in\mathcal{H},z\in\mathcal{G}}\{f(x) + h(x) + g(z)\}. \tag{4.5}$$
$$\text{s.t. } L_1x + L_2z = d$$

The system of KKT optimality conditions associated to this optimization problem is given by

$$-L_1^*y \in \partial f(x) + \nabla h(x), \quad -L_2^*y \in \partial g(z) \text{ and } L_1x + L_2z = d. \tag{4.6}$$

If (x, z, y) is a solution of (4.6), then (x, z) is an optimal solution of (4.5) and y is an optimal solution of its dual problem

$$\sup_{y\in\mathcal{G}}\left\{-(f^*\square h^*)(-L_1^*y) - g^*(-L_2^*y) - \langle d, y\rangle\right\}, \tag{4.7}$$

The system of KKT optimality conditions (4.6) is for

$$A := \partial f, B := \partial g, C := \nabla h \tag{4.8}$$

nothing else than (4.4). Notice that ∂f and ∂g are maximally monotone operators, while, by the Baillon-Haddad Theorem (see [2, Corollary 18.16]), the gradient of h is η-cocoercive.

Remark 4.2 Consider the optimization problem

$$\inf_{y\in\mathcal{Z}}\{f(L_1y) + g(L_2y)\}, \tag{4.9}$$

where $f : \mathcal{H} \to \overline{\mathbb{R}}$ and $g : \mathcal{G} \to \overline{\mathbb{R}}$ are proper, convex and lower semicontinuous functions, and $L_1 : \mathcal{Z} \to \mathcal{H}$ and $L_2 : \mathcal{Z} \to \mathcal{G}$ are linear continuous operators. The associated dual problem can be written as

$$\inf_{(p,q)\in\mathcal{H}\times\mathcal{G}}\{f^*(p)+g^*(q)\}, \quad \text{s.t. } L_1^*p+L_2^*q=0 \tag{4.10}$$

while (4.9) is on its turn the dual problem of (4.10). Finding a solution (p,q,y) of the system of KKT optimality conditions associated to (4.10)

$$-L_1y\in\partial f^*(p),\ -L_2y\in\partial g^*(y) \text{ and } L_1^*p+L_2^*q=0$$

provides an optimal solution y of the problem (4.9) and an optimal solution (p,q) of the problem (4.10).

We propose the following iterative scheme for determining the KKT points of the variational inequality (4.3).

Algorithm 4.11 Let $M_1^k\in\mathcal{S}_+(\mathcal{H})$, $M_2^k\in\mathcal{S}_+(\mathcal{G})$ and $c>0$ be such that $cL_1^*L_1+M_1^k\in\mathcal{P}_{\alpha_k}(\mathcal{H})$ and $cL_2^*L_2+M_2^k\in\mathcal{P}_{\beta_k}(\mathcal{G})$, with $\alpha_k,\beta_k>0$, for all $k\geq 0$. Choose $(x^0,z^0,y^0)\in\mathcal{H}\times\mathcal{G}\times\mathcal{Z}$. For all $k\geq 0$ generate the sequence $(x^k,z^k,y^k)_{k\geq 0}$ as follows:

$$x^{k+1}:=(cL_1^*L_1+M_1^k+A)^{-1}\left[cL_1^*(-L_2z^k+d-c^{-1}y^k)+M_1^kx^k-Cx^k\right] \tag{4.11a}$$

$$z^{k+1}:=(cL_2^*L_2+M_2^k+B)^{-1}\left[cL_2^*(-L_1x^{k+1}+d-c^{-1}y^k)+M_2^kz^k\right] \tag{4.11b}$$

$$y^{k+1}:=y^k+c(L_1x^{k+1}+L_2z^{k+1}-d). \tag{4.11c}$$

Remark 4.3 For the choice $\mathcal{G}:=\mathcal{Z}$, $L_2:=-\mathrm{Id}$ and $d:=0$, the variational inequality to solve simplifies to the following monotone inclusion problem

$$\text{find } x\in\mathcal{H} \text{ such that } 0\in Ax+Cx+(L_1^*\circ B\circ L_1)(x),$$

while Algorithm 4.11 becomes the iterative scheme proposed in [4] for solving it.

We show in the following that the numerical scheme above encompasses several other algorithms from the literature. For all $k\geq 0$, the equations (4.11a) and (4.11b) are equivalent to

$$-cL_1^*(L_1x^{k+1}+L_2z^k-d+c^{-1}y^k)+M_1^k(x^k-x^{k+1})-Cx^k\in Ax^{k+1}, \tag{4.12}$$

and, respectively,

$$-cL_2^*(L_1x^{k+1}+L_2z^{k+1}-d+c^{-1}y^k)+M_2^k(z^k-z^{k+1})\in Bz^{k+1}. \tag{4.13}$$

In the variational setting of Problem 4.2, i.e. considering the particular choice (4.8), the inclusion (4.12) becomes

$$0 \in \partial f(x^{k+1}) + cL_1^*(L_1x^{k+1} + L_2z^k - d + c^{-1}y^k) + M_1^k(x^{k+1} - x^k) + \nabla h(x^k),$$

which is equivalent to

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathcal{H}} \left\{ f(x) + \langle x - x^k, \nabla h(x^k)\rangle + \frac{c}{2}\|L_1x + L_2z^k - d + c^{-1}y^k\|^2 + \frac{1}{2}\|x - x^k\|^2_{M_1^k} \right\}.$$

On the other hand, (4.13) becomes

$$-cL_2^*(L_1x^{k+1} + L_2z^{k+1} - d + c^{-1}y^k) + M_2^k(z^k - z^{k+1}) \in \partial g(z^{k+1}),$$

which is equivalent to

$$z^{k+1} = \operatorname*{arg\,min}_{x\in\mathcal{G}} \left\{ g(z) + \frac{c}{2}\|L_1x^{k+1} + L_2z - d + c^{-1}y^k\|^2 + \frac{1}{2}\|z - z^k\|^2_{M_2^k} \right\}.$$

In conclusion, the iterative scheme (4.11a)–(4.11c) applied to the variational setting of Problem 4.2 reads

$$x^{k+1} = \operatorname*{arg\,min}_{x\in\mathcal{H}} \left\{ f(x) + \langle x - x^k, \nabla h(x^k)\rangle + \frac{c}{2}\|L_1x + L_2z^k - d + c^{-1}y^k\|^2 + \frac{1}{2}\|x - x^k\|^2_{M_1^k} \right\} \tag{4.14}$$

$$z^{k+1} = \operatorname*{arg\,min}_{x\in\mathcal{G}} \left\{ g(z) + \frac{c}{2}\|L_1x^{k+1} + L_2z - d + c^{-1}y^k\|^2 + \frac{1}{2}\|z - z^k\|^2_{M_2^k} \right\} \tag{4.15}$$

$$y^{k+1} = y^k + c(L_1x^{k+1} + L_2z^{k+1} - d). \tag{4.16}$$

The situation when $h = 0$ and the sequences $(M_1^k)_{k\geq 0}$, $(M_2^k)_{k\geq 0}$ are constant has been considered, for example, in [10]. The case $\mathcal{G} = \mathcal{Z}$, $L_2 = -\mathrm{Id}$ and $d = 0$ delivers the algorithm formulated and investigated by Banert, Boţ and Csetnek in [1]. The latter is a generalization of the iterative scheme proposed by Shefi and Teboulle [15], which addresses the case when $h = 0$ and the sequences $(M_1^k)_{k\geq 0}$, $(M_2^k)_{k\geq 0}$ are constant in the setting of finite dimensional Hilbert spaces. Finally, when $h = 0$ and $M_1^k = M_2^k = 0$ for all $k \geq 0$, the iterative scheme (4.14)–(4.16) collapses into the classical version of the ADMM algorithm (see, for example, [11, 12]).

We refer the reader to [4, Remark 5], where it is shown that several primal-dual-type algorithms from the literature [5, 7, 9, 16] can be embedded in the algorithm designed for the situation where $\mathcal{G} = \mathcal{Z}$, $L_2 = -\mathrm{Id}$ and $d = 0$ and, consequently, in the general algorithm considered above.

4.3.2 Convergence Analysis

An important ingredient for our convergence analysis will be the following version of the Opial Lemma (see [8, Theorem 3.3]).

Lemma 4.1 *Let C be a nonempty subset of $\mathcal{H}$ and $(x^k)_{k\geq 0}$ be a sequence in $\mathcal{H}$. Let $\alpha > 0$ and $W^k \in \mathcal{P}_\alpha(\mathcal{H})$ be such that $W^k \succeq W^{k+1}$ for all $k \geq 0$. Assume that:*

(i) for all $z \in C$ and for all $k \geq 0$: $\|x^{k+1} - z\|_{W^{k+1}} \leq \|x^k - z\|_{W^k}$.
(ii) every weak sequential cluster point of $(x^k)_{k\geq 0}$ belongs to C.

Then $(x^k)_{k\geq 0}$ converges weakly to an element in C.

The following theorem is the main result of this section.

Theorem 4.1 *In the context of Problem 4.1, assume that the set of KKT points of the variational inequality with monotone operators (4.3) is nonempty and that $M_1^k - \frac{\eta}{2}\mathrm{Id} \in \mathcal{S}_+(\mathcal{H})$, $M_1^k \succeq M_1^{k+1}$, $M_2^k \in \mathcal{S}_+(\mathcal{G})$, $M_2^k \succeq M_2^{k+1}$, and $M_2^k + cL_2^*L_2 \in \mathcal{S}_+(\mathcal{G})$ for all $k \geq 0$. Let $(x^k, z^k, y^k)_{k\geq 0}$ be the sequence generated by Algorithm 4.11. Suppose that one of the following assumptions is fulfilled:*

(I) there exist $\alpha_1, \beta_1 > 0$ such that $M_1^k - \frac{\eta}{2}\mathrm{Id} \in \mathcal{P}_{\alpha_1}(\mathcal{H})$ and $M_2^k + cL_2^*L_2 \in \mathcal{P}_{\beta_1}(\mathcal{G})$ for all $k \geq 0$;
(II) there exist $\alpha_2, \beta_2 > 0$ such that $L_1^*L_1 \in \mathcal{P}_{\alpha_2}(\mathcal{H})$ and $M_2^k \in \mathcal{P}_{\beta_2}(\mathcal{G})$ for all $k \geq 0$;
(III) there exist $\alpha_3, \beta_3 > 0$ such that $M_1^k - \frac{\eta}{2}\mathrm{Id} + cL_1^*L_1 \in \mathcal{P}_{\alpha_3}(\mathcal{H})$, $L_2^*L_2 \in \mathcal{P}_{\beta_3}(\mathcal{G})$ and $2M_2^{k+1} \succeq M_2^k \succeq M_2^{k+1}$ for all $k \geq 0$.

Then $(x^k, z^k, y^k)_{k\geq 0}$ converges weakly to a KKT point of the variational inequality (4.3).

Proof Let (x^*, z^*, y^*) be a KKT point of the variational inequality with monotone operators (4.3). Then

$$-L_1^*y^* - Cx^* \in Ax^*, \quad -L_2^*y^* \in Bz^* \text{ and } L_1x^* + L_2z^* = d.$$

Let $k \geq 0$ be fixed. By (4.12), (4.13) and the monotonicity of A and B, we obtain the inequalities

$$\langle -cL_1^*(L_1x^{k+1} + L_2z^k - d + c^{-1}y^k) + M_1^k(x^k - x^{k+1}) - Cx^k + L_1^*y^* + Cx^*, x^{k+1} - x^*\rangle \geq 0$$

and

$$\langle -cL_2^*(L_1x^{k+1} + L_2z^{k+1} - d + c^{-1}y^k) + M_2^k(z^k - z^{k+1}) + L_2^*y^*, z^{k+1} - z^*\rangle \geq 0.$$

Since C is η-cocoercive, we have

$$\eta\langle Cx^* - Cx^k, x^* - x^k\rangle \geq \|Cx^* - Cx^k\|^2.$$

We consider first the case when $\eta > 0$. Summing up the three inequalities from above we get

$$\begin{aligned}
&c\langle -L_1x^{k+1} - L_2z^k + d, L_1x^{k+1} - L_1x^*\rangle + \langle y^* - y^k, L_1x^{k+1} - L_1x^*\rangle \\
&\quad + \langle Cx^* - Cx^k, x^{k+1} - x^*\rangle + \langle M_1^k(x^k - x^{k+1}), x^{k+1} - x^*\rangle \\
&\quad + c\langle L_2^*(-L_1x^{k+1} - L_2z^{k+1} + d), z^{k+1} - z^*\rangle + \langle -L_2^*y^k + L_2^*y^*, z^{k+1} - z^*\rangle \\
&\quad + \langle M_2^k(z^k - z^{k+1}), z^{k+1} - z^*\rangle + \langle Cx^* - Cx^k, x^* - x^k\rangle - \eta^{-1}\|Cx^* - Cx^k\|^2 \geq 0.
\end{aligned}$$

By taking into account (4.11c) we also obtain

$$\begin{aligned}
&\langle y^* - y^k, L_1x^{k+1} - L_1x^*\rangle + \langle -L_2^*y^k + L_2^*y^*, z^{k+1} - z^*\rangle \\
&\quad = \langle y^* - y^k, L_1x^{k+1} - L_1x^*\rangle + \langle y^* - y^k, L_2(z^{k+1} - z^*)\rangle \\
&\quad = \langle y^* - y^k, L_1x^{k+1} + L_2z^{k+1} - \underbrace{(L_1x^* + L_2z^*)}_{=d}\rangle \\
&\quad = c^{-1}\langle y^* - y^k, y^{k+1} - y^k\rangle.
\end{aligned}$$

Hence the above inequality reads as

$$\begin{aligned}
&c\langle (d - L_2z^k) - L_1x^{k+1}, L_1x^{k+1} - L_1x^*\rangle + c^{-1}\langle y^* - y^k, y^{k+1} - y^k\rangle \\
&\quad + \langle Cx^* - Cx^k, x^{k+1} - x^k\rangle + \langle M_1^k(x^k - x^{k+1}), x^{k+1} - x^*\rangle \\
&\quad + c\langle (d - L_1x^{k+1}) - L_2z^{k+1}, L_2z^{k+1} - L_2z^*\rangle + \langle M_2^k(z^k - z^{k+1}), z^{k+1} - z^*\rangle \\
&\quad - \eta^{-1}\|Cx^* - Cx^k\|^2 \geq 0.
\end{aligned}$$

By expressing the inner products through norms the above inequality becomes

$$\frac{c}{2}\left(\|(d-L_2z^k)-L_1x^*\|^2-\|(d-L_2z^k)-L_1x^{k+1}\|^2-\|L_1x^{k+1}-L_1x^*\|^2\right)$$
$$+\frac{c}{2}\left(\|(d-L_1x^{k+1})-L_2z^*\|^2-\|(d-L_1x^{k+1})-L_2z^{k+1}\|^2-\|L_2z^{k+1}-L_2z^*\|^2\right)$$
$$+\frac{1}{2c}\left(\|y^*-y^k\|^2+\|y^{k+1}-y^k\|^2-\|y^{k+1}-y^*\|^2\right)$$
$$+\frac{1}{2}\left(\|x^k-x^*\|^2_{M_1^k}-\|x^k-x^{k+1}\|^2_{M_1^k}-\|x^{k+1}-x^*\|^2_{M_1^k}\right)$$
$$+\frac{1}{2}\left(\|z^k-z^*\|^2_{M_2^k}-\|z^k-z^{k+1}\|^2_{M_2^k}-\|z^{k+1}-z^*\|^2_{M_2^k}\right)$$
$$+\langle Cx^*-Cx^k,x^{k+1}-x^k\rangle-\eta^{-1}\|Cx^*-Cx^k\|^2\geq 0.$$

By using again that $y^{k+1}=y^k+c(L_1x^{k+1}+L_2z^{k+1}-d)$ and by taking into account that

$$\langle Cx^*-Cx^k,x^{k+1}-x^k\rangle-\eta^{-1}\|Cx^*-Cx^k\|^2$$
$$=-\eta\left\|\eta^{-1}(Cx^*-Cx^k)+\frac{1}{2}(x^k-x^{k+1})\right\|^2+\frac{\eta}{4}\|x^k-x^{k+1}\|^2,$$

we obtain

$$\frac{1}{2}\|x^{k+1}-x^*\|^2_{M_1^k}+\frac{1}{2}\|z^{k+1}-z^*\|^2_{M_2^k}+\frac{1}{2}\|L_2z^{k+1}-L_2z^*\|^2_{c\,\mathrm{Id}}+\frac{1}{2c}\|y^{k+1}-y^*\|^2\leq$$
$$\frac{1}{2}\|x^k-x^*\|^2_{M_1^k}+\frac{1}{2}\|z^k-z^*\|^2_{M_2^k}+\frac{1}{2}\|(d-L_2z^k)-L_1x^*\|^2_{c\,\mathrm{Id}}+\frac{1}{2c}\|y^*-y^k\|^2$$
$$-\frac{c}{2}\|(d-L_2z^k)-L_1x^{k+1}\|^2-\frac{1}{2}\|x^k-x^{k+1}\|^2_{M_1^k-\frac{\eta}{2}\mathrm{Id}}-\frac{1}{2}\|z^k-z^{k+1}\|^2_{M_2^k}$$
$$-\eta\left\|\eta^{-1}(Cx^*-Cx^k)+\frac{1}{2}(x^k-x^{k+1})\right\|^2.$$

Since $(d-L_2z^k)-L_1x^*=-L_2z^k+L_2z^*$ and by using the monotonicity assumptions on $(M_1^k)_{k\geq 0}$ and $(M_2^k)_{k\geq 0}$ it yields

$$\frac{1}{2}\|x^{k+1}-x^*\|^2_{M_1^{k+1}}+\frac{1}{2}\|z^{k+1}-z^*\|^2_{M_2^{k+1}+cL_2^*L_2}+\frac{1}{2c}\|y^{k+1}-y^*\|^2\leq$$
$$\frac{1}{2}\|x^k-x^*\|^2_{M_1^k}+\frac{1}{2}\|z^k-z^*\|^2_{M_2^k+cL_2^*L_2}+\frac{1}{2c}\|y^*-y^k\|^2$$
$$-\frac{c}{2}\|L_1x^{k+1}+L_2z^k-d\|^2-\frac{1}{2}\|x^k-x^{k+1}\|^2_{M_1^k-\frac{\eta}{2}\mathrm{Id}}-\frac{1}{2}\|z^k-z^{k+1}\|^2_{M_2^k}$$
$$-\eta^{-1}\left\|\eta(Cx^*-Cx^k)+\frac{1}{2}(x^k-x^{k+1})\right\|^2. \tag{4.17}$$

In the case when $\eta = 0$, by repeating the above calculations, we obtain

$$\frac{1}{2}\|x^{k+1} - x^*\|^2_{M_1^{k+1}} + \frac{1}{2}\|z^{k+1} - z^*\|^2_{M_2^{k+1}+cL_2^*L_2} + \frac{1}{2c}\|y^{k+1} - y^*\|^2 \leq$$
$$\frac{1}{2}\|x^k - x^*\|^2_{M_1^k} + \frac{1}{2}\|z^k - z^*\|^2_{M_2^k+cL_2^*L_2} + \frac{1}{2c}\|y^* - y^k\|^2$$
$$-\frac{c}{2}\|L_1x^{k+1} + L_2z^k - d\|^2 - \frac{1}{2}\|x^k - x^{k+1}\|^2_{M_1^k} - \frac{1}{2}\|z^k - z^{k+1}\|^2_{M_2^k}. \tag{4.18}$$

By using arguments involving telescoping sums, each of the inequalities (4.17) and (4.18) yields

$$\sum_{k\geq 0}\|L_1x^{k+1} + L_2z^k - d\|^2 < +\infty, \ \sum_{k\geq 0}\|x^k - x^{k+1}\|^2_{M_1^k - \frac{\eta}{2}\mathrm{Id}} < +\infty,$$
$$\sum_{k\geq 0}\|z^k - z^{k+1}\|^2_{M_2^k} < +\infty. \tag{4.19}$$

Assume that condition (I) holds. By neglecting the negative terms (notice that $M_1^k - \frac{\eta}{2}\mathrm{Id} \in \mathcal{S}_+(\mathcal{H})$ for all $k \geq 0$), from each of the inequalities (4.17) and (4.18) it follows that assumption (i) the Opial Lemma holds, when applied in the product space $\mathcal{H} \times \mathcal{G} \times \mathcal{Z}$, for the sequence $(x^k, z^k, y^k)_{k\geq 0}$, for $W^k := (M_1^k, M_2^k + cL_2^*L_2, c^{-1}\,\mathrm{Id})$ for $k \geq 0$, and for $C \subseteq \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$ the set of KKT points of the variational inequality (4.3).

Since $M_1^k - \frac{\eta}{2}\mathrm{Id} \in \mathcal{P}_{\alpha_1}(\mathcal{H})$ for all $k \geq 0$ with $\alpha_1 > 0$, we get from (4.19)

$$x^k - x^{k+1} \to 0 \ (k \to +\infty) \tag{4.20}$$

and

$$L_1x^{k+1} + L_2z^k - d \to 0 \ (k \to +\infty). \tag{4.21}$$

Therefore

$$\begin{aligned}\|z^{k+1} - z^k\|_{L_2^*L_2} =& \|L_2z^{k+1} - L_2z^k\| \\ \leq& \|L_1x^{k+2} + L_2z^{k+1} - d\| + \|L_1x^{k+1} + L_2z^k - d\| \\ &+ \|L_1x^{k+1} - L_1x^{k+2}\|,\end{aligned}$$

which means that

$$\|z^{k+1} - z^k\|_{L_2^*L_2} \to 0 \ (k \to +\infty).$$

Using the third condition in (4.19) and the fact that $M_2^k + cL_2^*L_2 \in \mathcal{P}_{\beta_1}(\mathcal{G})$ we conclude

$$z^k - z^{k+1} \to 0 \ (k \to +\infty). \tag{4.22}$$

From (4.11c) we derive

$$\begin{aligned} \|y^k - y^{k+1}\| &= c\|L_1x^{k+1} + L_2z^{k+1} - d\| \\ &\leq c\left(\|L_1x^{k+1} + L_2z^k - d\| + \|L_2z^{k+1} - L_2z^k\|\right), \end{aligned}$$

hence, by (4.21) and (4.22)

$$y^k - y^{k+1} \to 0 \ (k \to +\infty). \tag{4.23}$$

Now we are able to verify the second assumption in the Opial Lemma for C taken as the set of KKT points of (4.3). Let $(\bar{x}, \bar{z}, \bar{y}) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$ be such that there exists $(k_n)_{n\geq 0}$, $k_n \to +\infty$ (as $n \to +\infty$), and $(x^{k_n}, z^{k_n}, y^{k_n})$ converges weakly to $(\bar{x}, \bar{z}, \bar{y})$ (as $n \to +\infty$). From (4.20) and the linearity of L_1 we obtain that $(L_1x^{k_n+1} + L_2z^{k_n})_{n\geq 0}$ converges weakly to $L_1\bar{x} + L_2\bar{z}$ (as $n \to +\infty$), which combined with (4.21) yields $L_1\bar{x} + L_2\bar{z} = d$. For $n \geq 0$, we use now the following notations

$$\begin{aligned} a_n^* &:= cL_1^*(-L_1x^{k_n+1} - L_2z^{k_n} + d) + L_1^*(y^{k_n+1} - y^{k_n}) \\ &\quad + M_1^{k_n}(x^{k_n} - x^{k_n+1}) + Cx^{k_n+1} - Cx^{k_n} \\ a_n &:= x^{k_n+1} \\ b_n^* &:= M_2^{k_n}(z^{k_n} - z^{k_n+1}) \\ b_n &:= z^{k_n+1} \\ c_n^* &:= -L_1x^{k_n+1} - L_2z^{k_n+1} + d \\ c_n &:= y^{k_n+1}. \end{aligned}$$

From (4.12) we have

$$a_n^* \in (A + C)a_n + L_1^*c_n \tag{4.24}$$

and by combining (4.13) with (4.11c) we obtain

$$b_n^* \in Bb_n + L_2^*c_n \tag{4.25}$$

for all $n \geq 0$. From (4.20), (4.22) and (4.23) we have that

$$(a_n, b_n, c_n) \text{ converges weakly to } (\bar{x}, \bar{z}, \bar{y}) \ \ (\text{as } n \to +\infty). \tag{4.26}$$

Moreover, by (4.20)–(4.23) and the Lipschitz continuity of C we obtain

$$(a_n^*, b_n^*, c_n^*) \text{ converges strongly to } (0, 0, 0) \ \ (\text{as } n \to +\infty). \tag{4.27}$$

Next we define the maximally monotone operator

$$T : \mathcal{H} \times \mathcal{G} \times \mathcal{Z} \rightrightarrows \mathcal{H} \times \mathcal{G} \times \mathcal{Z}, (x, z, y) \mapsto ((A + C)x, Bz, 0),$$

and the linear continuous operator

$$\tilde{K} : \mathcal{H} \times \mathcal{G} \times \mathcal{Z} \to \mathcal{H} \times \mathcal{G} \times \mathcal{Z}, (x, z, y) \mapsto (L_1^* y, L_2^* y, -L_1 x - L_2 z).$$

For all $(x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$ we have

$$\begin{aligned} \langle \tilde{K}(x, z, y), (x, z, y) \rangle &= \langle L_1^* y, x \rangle + \langle L_2^* y, z \rangle + \langle -L_1 x - L_2 z, y \rangle \\ &= \langle y, L_1 x \rangle + \langle y, L_2 z \rangle - \langle L_1 x, y \rangle - \langle L_2 z, y \rangle = 0, \end{aligned}$$

hence $\tilde{K}$ is maximally monotone and therefore the shifted operator

$$K : \mathcal{H} \times \mathcal{G} \times \mathcal{Z} \to \mathcal{H} \times \mathcal{G} \times \mathcal{Z}, K(x, y, z) := \tilde{K}(x, y, z) + (0, 0, d),$$

is maximally monotone, as well. Since K has full domain we obtain that

$$T + K \text{ is a maximally monotone operator.} \tag{4.28}$$

On the other hand, from (4.24) and (4.25) we have that

$$\big((a_n, b_n, c_n), (a_n^*, b_n^*, c_n^*)\big) \in \text{gra}(T + K) \ \forall n \geq 0. \tag{4.29}$$

Since the graph of a maximally monotone operator is sequentially closed with respect to the weak×strong topology (see [2, Proposition 20.33]), from (4.26), (4.27), (4.28) and (4.29) we derive that

$$((\bar{x}, \bar{z}, \bar{y}), (0, 0, 0)) \in \text{gra}(T + K),$$

which is equivalent to

$$(0, 0, 0) \in ((A + C)\bar{x} + L_1^* \bar{y}, B\bar{z} + L_2^* \bar{y}, -L_1 \bar{x} - L_2 \bar{z} + d).$$

The latter means nothing else than saying that $(\bar{x}, \bar{z}, \bar{y})$ is a KKT point of (4.3), thus assumption (ii) in the Opial Lemma is verified, too. In conclusion, $(x^k, z^k, y^k)_{k\geq 0}$ converges weakly to a KKT point of (4.3).

Consider now the situation in assumption (II). From (4.17) and (4.18) it follows that (4.21) and (4.22) hold. From (4.11c), (4.21) and (4.22) we obtain (4.23). Finally, by using that $L_1^* L_1 \in \mathcal{P}_{\alpha_2}(\mathcal{H})$ for $\alpha_2 > 0$, relation (4.20) holds, too.

On the other hand, (4.17) and (4.18) yield that

$$\exists \lim_{k\to+\infty} \left(\frac{1}{2}\|x^k - x^*\|^2_{M_1^k} + \frac{1}{2}\|z^k - z^*\|^2_{M_2^k + cL_2^* L_2} + \frac{1}{2c}\|y^k - y^*\|^2 \right), \tag{4.30}$$

hence $(y^k)_{k\geq 0}$ and $(z^k)_{k\geq 0}$ are bounded. Combining this with

$$\begin{aligned} \alpha_2 \|x^k - x^*\|^2 &\leq \|L_1 x^k - L_1 x^*\|^2 \\ &\leq \frac{1}{3}\|L_1 x^k + L_2 z^k - d\|^2 \\ &\quad + \frac{1}{3}\|L_1 x^* + L_2 z^* - d\|^2 + \frac{1}{3}\|L_2 z^* - L_2 z^k\|^2, \end{aligned}$$

which holds for all $k \geq 0$, and using (4.11c), we derive that $(x^k)_{k\geq 0}$ is bounded, too. Hence there exists a weakly convergent subsequence of $(x^k, z^k, y^k)_{k\geq 0}$. By using the same arguments as in the second part of the proof of (I) it follows that every weak sequential cluster point of $(x^k, z^k, y^k)_{k\geq 0}$ is a KKT point of (4.3).

Now we show that the set of weak sequential cluster points of $(x^k, z^k, y^k)_{k\geq 0}$ is a singleton. Let (x_1, z_1, y_1), (x_2, z_2, y_2) be two such weak sequential cluster points. Then there exist $(k_p)_{p\geq 0}$, $(k_q)_{q\geq 0}$, $k_p \to +\infty$ (as $p \to +\infty$), $k_q \to +\infty$ (as $q \to +\infty$), a subsequence $(x^{k_p}, z^{k_p}, y^{k_p})_{p\geq 0}$ which converges weakly to (x_1, z_1, y_1) (as $p \to +\infty$), and a subsequence $(x^{k_q}, z^{k_q}, y^{k_q})_{q\geq 0}$ which converges weakly to (x_2, z_2, y_2) (as $q \to +\infty$). As seen above, (x_1, z_1, y_1) and (x_2, z_2, y_2) are KKT points of (4.3), thus $L_1 x_i + L_2 z_i = d$ for $i \in \{1, 2\}$. From (4.30), which is true for every KKT point of (4.3), we derive

$$\exists \lim_{k\to+\infty} \left(E(x^k, z^k, y^k; x_1, z_1, y_1) - E(x^k, z^k, y^k; x_2, z_2, y_2) \right), \tag{4.31}$$

where

$$E(x^k, z^k, y^k; x, z, y) := \frac{1}{2}\|x^k - x\|^2_{M_1^k} + \frac{1}{2}\|z^k - z\|^2_{M_2^k + cL_2^* L_2} + \frac{1}{2c}\|y^k - y\|^2.$$

We have for all $k \geq 0$

$$\frac{1}{2}\|x^k - x_1\|^2_{M_1^k} - \frac{1}{2}\|x^k - x_2\|^2_{M_1^k} = \frac{1}{2}\|x_2 - x_1\|^2_{M_1^k} + \langle x^k - x_2, M_1^k(x_2 - x_1)\rangle,$$

$$\begin{aligned}&\frac{1}{2}\|z^k - z_1\|^2_{M_2^k + cL_2^*L_2} - \frac{1}{2}\|z^k - z_2\|^2_{M_2^k + cL_2^*L_2}\\ &= \frac{1}{2}\|z_2 - z_1\|^2_{M_2^k + cL_2^*L_2} + \langle z^k - z_2, (M_2^k + cL_2^*L_2)(z_2 - z_1)\rangle\end{aligned}$$

and

$$\frac{1}{2c}\|y^k - y_1\|^2 - \frac{1}{2c}\|y^k - y_2\|^2 = \frac{1}{2c}\|y_2 - y_1\|^2 + \frac{1}{c}\langle y^k - y_2, y_2 - y_1\rangle.$$

According to[13, Théorème 104.1] there exist $M_1 \in \mathcal{S}_+(\mathcal{H})$ such that $(M_1^k)_{k\geq 0}$ converges pointwise to M_1 in the strong topology (as $k \to +\infty$). Similarly, the monotonicity condition imposed on $(M_2^k)_{k\geq 0}$ implies that $\sup_{k\geq 0}\|M_2^k + cL_2^*L_2\| < +\infty$. Thus, according to [8, Lemma 2.3], there exists $\alpha' > 0$ and $M_2 \in \mathcal{P}_{\alpha'}(\mathcal{G})$ such that $(M_k^2 + cL_2^*L_2)_{k\geq 0}$ converges pointwise to M_2 in the strong topology (as $k \to +\infty$).
Taking the limit in (4.31) along the subsequences $(k_p)_{p\geq 0}$ and $(k_q)_{q\geq 0}$ and using the last three identities above, we obtain

$$\begin{aligned}&\frac{1}{2}\|x_1 - x_2\|^2_{M_1} + \langle x_1 - x_2, M_1(x_2 - x_1)\rangle + \frac{1}{2}\|z_1 - z_2\|^2_{M_2} + \langle z_1 - z_2, M_2(z_2 - z_1)\rangle\\ &+ \frac{1}{2c}\|y_1 - y_2\|^2 + \frac{1}{c}\langle y_1 - y_2, y_2 - y_1\rangle\\ =&\frac{1}{2}\|x_1 - x_2\|^2_{M_1} + \frac{1}{2}\|z_1 - z_2\|^2_{M_2} + \frac{1}{2c}\|y_1 - y_2\|^2,\end{aligned}$$

hence

$$-\|x_1 - x_2\|^2_{M_1} - \|z_1 - z_2\|^2_{M_2} - \frac{1}{c}\|y_1 - y_2\|^2 = 0,$$

thus $z_1 = z_2$ and $y_1 = y_2$. Further, since $L_1x_i + L_2z_i = d$ for $i \in \{1, 2\}$,

$$\begin{aligned}\alpha_2\|x_1 - x_2\|^2 &\leq \|L_1x_1 - L_1x_2\|^2\\ &\leq \frac{1}{3}\|L_1x_1 + L_2z_1 - d\|^2 + \frac{1}{3}\|L_1x_2 + L_2z_2 - d\|^2 + \frac{1}{3}\|L_2z_1 - L_2z_2\|^2\\ &= 0,\end{aligned}$$

thus $x_1 = x_2$. In conclusion, $(x^k, z^k, y^k)_{k\geq 0}$ converges weakly to a KKT point of (4.3).

Finally, we consider the situation when the hypotheses in assumption (III) hold. Let $k \geq 1$ be fixed. Combining (4.13) with (4.11c) gives

$$-L_2^* y^{k+1} + M_2^k(z^k - z^{k+1}) \in Bz^{k+1}.$$

Considering this monotone inclusion for consecutive iterates and by taking into account the monotonicity of B, we obtain

$$\langle z^{k+1} - z^k, -L_2^*(y^{k+1} - y^k) + M_2^k(z^k - z^{k+1}) - M_2^{k-1}(z^{k-1} - z^k)\rangle \geq 0,$$

hence

$$\begin{aligned}
&\langle z^{k+1} - z^k, -L_2^*(y^{k+1} - y^k)\rangle \\
&\geq \|z^{k+1} - z^k\|^2_{M_2^k} + \langle z^{k+1} - z^k, M_2^{k-1}(z^{k-1} - z^k)\rangle \\
&\geq \|z^{k+1} - z^k\|^2_{M_2^k} - \frac{1}{2}\|z^{k+1} - z^k\|^2_{M_2^{k-1}} - \frac{1}{2}\|z^k - z^{k-1}\|^2_{M_2^{k-1}}.
\end{aligned}$$

Using that $y^{k+1} - y^k = c(L_1x^{k+1} + L_2z^{k+1} - d)$, the last inequality yields

$$\begin{aligned}
&\|z^{k+1} - z^k\|^2_{M_2^k} - \frac{1}{2}\|z^{k+1} - z^k\|^2_{M_2^{k-1}} - \frac{1}{2}\|z^k - z^{k-1}\|^2_{M_2^{k-1}} \\
&\leq c\langle L_2z^{k+1} - L_2z^k, -L_1x^{k+1} - L_2z^{k+1} + d\rangle \\
&= \frac{c}{2}\left(\|L_1x^{k+1} + L_2z^k - d\|^2 - \|L_2z^{k+1} - L_2z^k\|^2 - \|L_1x^{k+1} + L_2z^{k+1} - d\|^2\right),
\end{aligned}$$

which, after adding it with (4.17) and using (4.11c), leads to

$$\begin{aligned}
&\frac{1}{2}\|x^{k+1} - x^*\|^2_{M_1^{k+1}} + \frac{1}{2}\|z^{k+1} - z^*\|^2_{M_2^{k+1} + cL_2^*L_2} + \frac{1}{2c}\|y^{k+1} - y^*\|^2 + \\
&\frac{1}{2}\|z^{k+1} - z^k\|^2_{3M_2^k - M_2^{k-1}} \\
&\leq \frac{1}{2}\|x^k - x^*\|^2_{M_1^k} + \frac{1}{2}\|z^k - z^*\|^2_{M_2^k + cL_2^*L_2} + \frac{1}{2c}\|y^* - y^k\|^2 + \frac{1}{2}\|z^k - z^{k-1}\|^2_{M_2^{k-1}} - \\
&\frac{1}{2}\|x^k - x^{k+1}\|^2_{M_1^k - \frac{\eta}{2}\mathrm{Id}} - \frac{c}{2}\|L_2z^{k+1} - L_2z^k\|^2 - \frac{1}{2c}\|y^{k+1} - y^k\|^2 - \\
&\eta\|\eta^{-1}(Cx^* - Cx^k) + \frac{1}{2}(x^k - x^{k+1})\|^2.
\end{aligned}$$

Taking into account that according to (III) we have $3M_2^k - M_2^{k-1} \succeq M_2^k$, we can conclude that for all $k \geq 1$ it holds

$$\begin{aligned}
&\frac{1}{2}\|x^{k+1}-x^*\|^2_{M_1^{k+1}}+\frac{1}{2}\|z^{k+1}-z^*\|^2_{M_2^{k+1}+cL_2^*L_2}+\frac{1}{2c}\|y^{k+1}-y^*\|^2+\frac{1}{2}\|z^{k+1}-z^k\|^2_{M_2^k}\\
&\leq \frac{1}{2}\|x^k-x^*\|^2_{M_1^k}+\frac{1}{2}\|z^k-z^*\|^2_{M_2^k+cL_2^*L_2}+\frac{1}{2c}\|y^*-y^k\|^2-\frac{1}{2}\|x^k-x^{k+1}\|^2_{M_1^k-\frac{\eta}{2}\mathrm{Id}}-\\
&\frac{1}{2}\|z^{k+1}-z^k\|^2_{cL_2^*L_2}-\frac{1}{2c}\|y^{k+1}-y^k\|^2+\frac{1}{2}\|z^k-z^{k-1}\|^2_{M_2^{k-1}}-\\
&\eta^{-1}\|\eta(Cx^*-Cx^k)+\frac{1}{2}(x^k-x^{k+1})\|^2, \qquad (4.32)
\end{aligned}$$

while, by using when $\eta = 0$ (4.18) instead of (4.17), it yields

$$\begin{aligned}
&\frac{1}{2}\|x^{k+1}-x^*\|^2_{M_1^{k+1}}+\frac{1}{2}\|z^{k+1}-z^*\|^2_{M_2^{k+1}+cL_2^*L_2}+\frac{1}{2c}\|y^{k+1}-y^*\|^2+\frac{1}{2}\|z^{k+1}-z^k\|^2_{M_2^k}\\
&\leq \frac{1}{2}\|x^k-x^*\|^2_{M_1^k}+\frac{1}{2}\|z^k-z^*\|^2_{M_2^k+cL_2^*L_2}+\frac{1}{2c}\|y^*-y^k\|^2+\frac{1}{2}\|z^k-z^{k-1}\|^2_{M_2^{k-1}}-\\
&\frac{1}{2}\|x^k-x^{k+1}\|^2_{M_1^k}-\frac{1}{2}\|z^{k+1}-z^k\|^2_{cL_2^*L_2}-\frac{1}{2c}\|y^{k+1}-y^k\|^2. \qquad (4.33)
\end{aligned}$$

Using telescoping sum arguments, we obtain that $\|x^k - x^{k+1}\|^2_{M_1^k-\frac{\eta}{2}\mathrm{Id}} \to 0$, $y^k - y^{k+1} \to 0$ and $z^k - z^{k+1} \to 0$ as $k \to +\infty$. Using (4.11c), it follows that $L_1(x^k - x^{k+1}) \to 0$ as $k \to +\infty$, which, combined with the hypotheses imposed on $M_1^k - \frac{\eta}{2}\mathrm{Id} + cL_1^*L_1$, implies that $x^k - x^{k+1} \to 0$ as $k \to +\infty$. Consequently, $L_1x^{k+1} + L_2z^k - d \to 0$ as $k \to +\infty$. Hence the relations (4.20)–(4.23) are fulfilled. On the other hand, from (4.32) and (4.33) if follows that the limit

$$\begin{aligned}
\lim_{k\to+\infty}\Bigg(&\frac{1}{2}\|x^k-x^*\|^2_{M_1^k}+\frac{1}{2}\|z^k-z^*\|^2_{M_2^k+cL_2^*L_2}\\
&+\frac{1}{2c}\|y^k-y^*\|^2+\frac{1}{2}\|z^k-z^{k-1}\|^2_{M_2^{k-1}}\Bigg).
\end{aligned}$$

exists. By using that

$$\|z^k - z^{k-1}\|^2_{M_2^{k-1}} \leq \|z^k - z^{k-1}\|^2_{M_2^0} \leq \|M_2^0\|\|z^k - z^{k-1}\|^2 \quad \forall k \geq 1,$$

it follows that $\lim_{k\to+\infty}\|z^k - z^{k-1}\|^2_{M_2^{k-1}} = 0$, which further implies that (4.30) holds. From here the conclusion follows by arguing as in the second part of the proof provided in the setting of assumption (II). □

4.4 Convergence Rates in the Case When $A + C$ Is Strongly Monotone

In this section we address the following modification of the Problem 4.1.

Problem 4.3 In the context of Problem 4.1, we replace the cocoercivity of C by the assumptions that C is monotone and μ-Lipschitz continuous, for $\mu \geq 0$. Further, we assume that the sum $A + C$ is γ-strongly monotone for $\gamma > 0$, and that $d = 0$.

We have the following characterization for a KKT point of (4.3):

$$\exists (x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}: \begin{cases} -L_1^* y \in Ax + Cx \\ -L_2^* y \in Bz \\ L_1 x = -L_2 z \end{cases}$$

$$\Leftrightarrow \exists (x, z, y) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}: \begin{cases} -L_1^* y \in Ax + Cx \\ z \in B^{-1} \circ (-L_2^*) y \\ L_1 x = -L_2 z \end{cases}$$

$$\Leftrightarrow \exists (x, y) \in \mathcal{H} \times \mathcal{Z}: \begin{cases} -L_1^* y \in Ax + Cx \\ L_1 x \in (-L_2) \circ B^{-1} \circ (-L_2^*) y. \end{cases}$$

The latter means that (x, y) is a so-called primal-dual solution associated to the monotone inclusion problem

$$\text{find } x \in \mathcal{H} \text{ such that } 0 \in Ax + Cx + (L_1^* \circ \bar{B} \circ L_1)(x),$$

and its Attouch-Thera dual inclusion problem, where $\bar{B} := [(-L_2) \circ B^{-1} \circ (-L_2^*)]^{-1}$. Algorithm 14 in [4] designed to determine these primal-dual solutions in a setting which is similar to the one in Problem 4.3 gives rise to Algorithm 4.34.

Algorithm 4.34 For all $k \geq 0$, let $M_2^k : \mathcal{Z} \to \mathcal{Z}$ be a linear, continuous and self-adjoint operator such that $\tau_k L_1 L_1^* + M_2^k \in \mathcal{P}_{\alpha_k}(\mathcal{Z})$, for $\alpha_k > 0$. Choose $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{Z}$. For all $k \geq 0$ generate the sequence $(x^k, z^k, y^k)_{k \geq 0}$ as follows:

$$y^{k+1} = \left(\tau_k L_1 L_1^* + M_2^k + (-L_2) \circ B^{-1} \circ (-L_2^*)\right)^{-1} [-\tau_k L_1 (z^k - \tau_k^{-1} x^k) + M_2^k y^k] \tag{4.34a}$$

$$z^{k+1} = \left(\frac{\theta_k}{\lambda} - 1\right) L_1^* y^{k+1} + \frac{\theta_k}{\lambda} C x^k + \frac{\theta_k}{\lambda} (\mathrm{Id} + \lambda \tau_{k+1}^{-1} A^{-1})^{-1} (-L_1^* y^{k+1} + \lambda \tau_{k+1}^{-1} x^k - C x^k) \tag{4.34b}$$

$$x^{k+1} = x^k + \frac{\tau_{k+1}}{\theta_k} (-L_1^* y^{k+1} - z^{k+1}), \tag{4.34c}$$

where $\lambda, \tau_k, \theta_k > 0$.

Algorithm 4.35 For all $k \geq 0$, let $M_2^k : \mathcal{G} \to \mathcal{G}$ be a linear, continuous and self-adjoint operator such that $\tau_k L_2^{-1} L_1 (L_2^{-1} L_1)^* + L_2^{-1} M_2^k (L_2^*)^{-1} \in \mathcal{P}_{\alpha_k}(\mathcal{Z})$ for $\alpha_k > 0$. Choose $(x^0, z^0, y^0) \in \mathcal{H} \times \mathcal{G} \times \mathcal{G}$. For all $k \geq 0$ generate the sequence $(x^k, z^k, y^k)_{k \geq 0}$ as follows:

$$y^{k+1} = (-L_2^*)^{-1} \circ \left(\tau_k L_2^{-1} L_1 (L_2^{-1} L_1)^* + L_2^{-1} M_2^k (L_2^*)^{-1} + B^{-1}\right)^{-1} \circ (-L_2)^{-1}[-\tau_k L_1 (z^k - \tau_k^{-1} x^k) + M_2^k y^k] \tag{4.35a}$$

$$z^{k+1} = \left(\frac{\theta_k}{\lambda} - 1\right) L_1^* y^{k+1} + \frac{\theta_k}{\lambda} C x^k + \frac{\theta_k}{\lambda} (\mathrm{Id} + \lambda \tau_{k+1}^{-1} A^{-1})^{-1} (-L_1^* y^{k+1} + \lambda \tau_{k+1}^{-1} x^k - C x^k) \tag{4.35b}$$

$$x^{k+1} = x^k + \frac{\tau_{k+1}}{\theta_k} (-L_1^* y^{k+1} - z^{k+1}), \tag{4.35c}$$

where $\lambda, \tau_k, \theta_k > 0$.

In case $\mathcal{G} = \mathcal{Z}$ and the linear continuous operator $L_2 : \mathcal{G} \to \mathcal{G}$ is invertible, we obtain Algorithm 4.35, which is a full splitting formulation for Algorithm 4.34.

Concerning the parameters involved in Algorithm 4.34, we assume that

$$\mu \tau_1 < 2\gamma, \ \lambda \geq \mu + 1, \tag{4.36}$$

that there exists $\sigma_0 > 0$ such that

$$\sigma_0 \tau_1 \|L_1\|^2 \leq 1, \tag{4.37}$$

and that for all $k \geq 0$

$$\theta_k = \frac{1}{\sqrt{1 + \tau_{k+1} \lambda^{-1} (2\gamma - \mu \tau_{k+1})}} \tag{4.38}$$

$$\tau_{k+2} = \theta_k \tau_{k+1} \tag{4.39}$$

$$\sigma_{k+1} = \theta_k^{-1} \sigma_k \tag{4.40}$$

$$\tau_k L_1 L_1^* + M_2^k \succeq \sigma_k^{-1} \mathrm{Id} \tag{4.41}$$

$$\frac{\tau_k}{\tau_{k+1}} L_1 L_1^* + \frac{1}{\tau_{k+1}} M_2^k \succeq \frac{\tau_{k+1}}{\tau_{k+2}} L_1 L_1^* + \frac{1}{\tau_{k+2}} M_2^{k+1}. \tag{4.42}$$

The following convergence rate result follows from [4, Theorem 19].

Theorem 4.2 *Consider the setting of Problem 4.3 in the hypothesis $(-L_2) \circ B^{-1} \circ (-L_2^*)$ is maximally monotone. Let (x, z, y) be a KKT point of the variational inequality (4.3). Let $(x^k, z^k, y^k)_{k \geq 0}$ be the sequence generated by Algorithm 4.34 and assume that the relations (4.36)–(4.42) are fulfilled. Then we have for all $n \geq 2$*

$$\frac{\lambda\|x^n - x\|^2}{\tau_{n+1}^2} + \frac{1 - \sigma_0\tau_1\|L_1\|^2}{\sigma_0\tau_1}\|y^n - y\|^2 \leq$$

$$\frac{\lambda\|x^1 - x\|^2}{\tau_2^2} + \frac{\|y^1 - y\|^2_{\tau_1 L_1 L_1^* + M_2^1}}{\tau_2} + \frac{\|x^1 - x^0\|^2}{\tau_1^2} + \frac{2}{\tau_1}\langle L_1(x^1 - x^0), y^1 - y\rangle.$$

Moreover, $\lim_{n\to+\infty} n\tau_n = \frac{\lambda}{\gamma}$, *hence one obtains for* $(x^n)_{n\geq 0}$ *an order of convergence of* $\mathcal{O}(\frac{1}{n})$.

Remark 4.4 Conditions guaranteeing the maximal monotonicity of compositions of a maximally monotone operator with a linear continuous operator have been intensively studied in the Hilbert space setting; for more insights we refer the reader to [2] and [3] and to the references therein.

Acknowledgements The first author has been partially supported by FWF (Austrian Science Fund), project I 2419-N32. The second author has been supported by FWF (Austrian Science Fund), project P 29809-N32. The third author has been partially supported by FWF (Austrian Science Fund), project I 2419-N32, and by the Doctoral Programme Vienna Graduate School on Computational Optimization (VGSCO), project W1260-N35.

References

1. Banert, S., Boţ, R.I., Csetnek, E.R.: Fixing and extending some recent results on the ADMM algorithm (2017). Available via arXiv. https://arxiv.org/abs/1612.05057
2. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics, Springer, New York (2011)
3. Boţ, R.I.: Conjugate Duality in Convex Optimization. Lecture Notes in Economics and Mathematical Systems **637**, Springer-Verlag Berlin Heidelberg (2010)
4. Boţ, R.I., Csetnek, E.R.: ADMM for monotone operators: convergence analysis and rates. Adv. Comput. Math. **45**(1), 327–359 (2019). Available via arXiv. https://arxiv.org/abs/1705.01913
5. Boţ, R.I., Csetnek, E.R., Heinrich, A.: A primal-dual splitting algorithm for finding zeros of sums of maximal monotone operators. SIAM J. Optim. **23**, 2011–2036 (2013)
6. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**, 1–12 (2010)
7. Chambolle, P.L., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vision **40**, 120–145 (2011)
8. Combettes, P.L., Vũ, B.C.: Variable metric quasi-Feyér monotonicity. Nonlinear Anal. **78**, 17–31 (2013)
9. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**, 460–479 (2013)
10. Fazel, M., Pong, T.K., Sun, D., Tseng, P.: Hankel matrix rank minimization with applications in system identification and realization. SIAM J. Matrix Anal. **34**, 946–977 (2013)
11. Fortin, M., Glowinski, R.: On decomposition-coordination methods using an augmented Lagrangian. In: Fortin, M. and Glowinski, R. (eds.), Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, North-Holland, Amsterdam (1983)

12. Gabay, D.: Applications of the method of multipliers to variational inequalities. In: Fortin, M. and Glowinski, R. (eds.), Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems, North-Holland, Amsterdam (1983)
13. Riesz, F., Nagy, B.Sz.: Leçons d'Analyse Fonctionelle. Fifth ed., Gauthier-Villars, Paris (1968)
14. Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pacific J. Math. **33**, 209–216 (1970)
15. Shefi, R., Teboulle, M.: Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization. SIAM J. Optim. **24**, 269–297 (2014)
16. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**, 667–681 (2013)

Chapter 5
Regularization of Ill-Posed Problems with Non-negative Solutions

Christian Clason, Barbara Kaltenbacher, and Elena Resmerita

Dedicated to the memory of Jonathan M. Borwein

Abstract This survey reviews variational and iterative methods for reconstructing non-negative solutions of ill-posed problems in infinite-dimensional spaces. We focus on two classes of methods: variational methods based on entropy-minimization or constraints, and iterative methods involving projections or non-negativity-preserving multiplicative updates. We summarize known results and point out some open problems.

Keywords Convex optimization · Fenchel duality · Entropy · Regularization · Sparsity · Signal processing

AMS 2010 Subject Classification 49M20, 65K10, 90C30

5.1 Introduction

Many inverse problems are concerned with the reconstruction of parameters that are a priori known to be non-negative, such as material properties or densities (in particular, probability densities). Non-negative solutions also frequently occur in astronomy and optical tomography, in particular in Poisson models for positron emission tomography (PET), see [57, 61]. Note that the literature on the finite-dimensional setting is very rich, and quite comprehensive surveys are already available; see, e.g., [10–12].

C. Clason (✉)
Faculty of Mathematics, University Duisburg-Essen, Essen, Germany
e-mail: christian.clason@uni-due.de

B. Kaltenbacher · E. Resmerita
Institute of Mathematics, Alpen-Adria Universität Klagenfurt, Klagenfurt, Austria
e-mail: barbara.kaltenbacher@aau.at; elena.resmerita@aau.at

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_5

Borwein and collaborators dealt in a series of papers with the case involving operators with finite-dimensional range; see, e.g., [4–6]. Concerned with reconstructing a density function from a finite number n of density moments, they have approached the problem from a few perspectives. For instance, it has been shown in [5] that the best (Boltzmann–Shannon) entropy estimates converge in the L^1-norm to the best entropy estimate of the limiting problem as $n \to \infty$. Along with this result, strong properties of the Boltzmann–Shannon entropy such as a Kadec–Klee property have been derived. Note that the dual problem of the maximum entropy estimates problem has been quite instrumental in showing further results such as error bounds. From a computational point of view, choosing one entropy (e.g., Dirac–Fermi or Burg) over the other has been the main topic in [4]. The work [3] studies the case of operators with infinite-dimensional range by proposing relaxed problems in the spirit of Morozov and Tikhonov regularization (cf. Section 5.3).

The context of infinite-dimensional function spaces for reconstructing non-negative solutions of ill-posed operator equations has been much less investigated in the literature. Therefore, this work focuses on methods for problems in such spaces from a deterministic point of view.

We will primarily consider linear operator equations

$$Au = y \tag{5.1}$$

with the operator $A : X \to Y$ mapping between suitable infinite-dimensional function spaces X and Y. We assume that (5.1) admits a non-negative solution $u^\dagger \geq 0$ (which we will make precise below) and that it is ill-posed in the sense that small perturbations of y can lead to arbitrarily large perturbations on u (or even lead to non-existence of a solution). Besides enforcing non-negativity of the solution for given data, a solution approach therefore also needs to have regularizing properties, i.e., be stable even for noisy data y^δ with

$$\|y^\delta - y\|_Y \leq \delta, \tag{5.2}$$

in place of y and yield reconstructions u^δ that converge to $u^\dagger$ as $\delta \to 0$. Two approaches are widespread in the literature:

(i) *Variational methods* are based on minimizing a weighted sum of a discrepancy term and a suitable regularization term (*Tikhonov regularization*) or on minimizing one of these terms under a constraint on the other (*Ivanov* or *Morozov regularization*, respectively).
(ii) *Iterative methods* construct a sequence of iterates approximating—for exact data—the solution $u^\dagger$; regularization is introduced by stopping the iteration based on a suitable discrepancy principle.

Regarding the regularization theory for ill-posed problems, we refer, e.g., to the classical work [22]; of particular relevance in the context of non-negative solutions

are regularization terms or iterations based on the Boltzmann–Shannon entropy and the associated Kullback–Leibler divergence, and we will focus especially on such methods.

This chapter is organized as follows. Section 5.2 recalls useful algebraic and topological properties of the mentioned entropy functionals. Section 5.3 reviews several variational entropy-based regularization methods (Morozov, Tikhonov, Ivanov), while Section 5.4 is dedicated to iterative methods for general linear ill-posed equations, both ones involving projections onto the non-negative cone and ones based on multiplicative updates preserving non-negativity.

5.2 Preliminaries

Let Ω be an open and bounded subset of $\mathbb{R}^d$. The *negative of the Boltzmann–Shannon entropy* is the function $f : L^1(\Omega) \to (-\infty, +\infty]$, given by[1]

$$f(u) = \begin{cases} \int_\Omega u(t) \log u(t)\, dt & \text{if } u \in L^1_+(\Omega) \text{ and } u \log u \in L^1(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

Here and in what follows, we set for $p \in [1, \infty]$

$$L^p_+(\Omega) := \left\{ u \in L^p(\Omega) : u(x) \geq 0 \text{ for almost every } x \in \Omega \right\},$$

while $\| \cdot \|_p$ denotes, as usual, the norm of the space $L^p(\Omega)$.

We recall some useful properties of the negative Boltzmann–Shannon entropy from, e.g., [1, proof of Thm. 1], [18, Lem. 2.1, 2.3], [52, § 3.4].

Lemma 5.1 *The followings properties hold:*

(i) *The function f is convex.*
(ii) *The function f is weakly lower semicontinuous in $L^1(\Omega)$.*
(iii) *For any $c > 0$, the sublevel set*

$$\left\{ v \in L^1_+(\Omega) : f(v) \leq c \right\}$$

is convex, weakly closed, and weakly compact in $L^1(\Omega)$.
(iv) *The domain of the function f is strictly included in $L^1_+(\Omega)$.*
(v) *The interior of the domain of the function f is empty.*
(vi) *The set $\partial f(u)$ is nonempty if and only if u belongs to $L^\infty_+(\Omega)$ and is bounded away from zero. In this case, $\partial f(u) = \{1 + \log u\}$.*

[1] We use the convention $0 \log 0 = 0$.

(vii) The directional derivative of the function f is given by

$$f'(u; v) = \int_\Omega v(t)[1 + \log u(t)]\, dt,$$

whenever it is finite.

Based on Lemma 5.1 (vi), we define in the following

$$\operatorname{dom} \partial f = \left\{ u \in L^\infty_+(\Omega) : u \text{ bounded away from zero a.e.} \right\}.$$

The *Kullback–Leibler divergence*, which coincides with the *Bregman distance* with respect to the Bolzmann–Shannon entropy, can be defined as $d : \operatorname{dom} f \times \operatorname{dom} f \to [0, +\infty]$ by

$$d(v, u) = f(v) - f(u) - f'(u; v - u),$$

where $f'(u; \cdot)$ is the directional derivative at u. One can also write

$$d(v, u) = \int \left[v(t) \log \frac{v(t)}{u(t)} - v(t) + u(t) \right] dt,$$

when $d(v, u)$ is finite. We list below several properties of the Kullback–Leibler divergence.

Lemma 5.2 *The followings properties hold:*

(i) The function $(v, u) \mapsto d(v, u)$ is convex.
(ii) The function $d(\cdot, u^*)$ is weakly lower semicontinuous in $L^1(\Omega)$ whenever $u^* \in \operatorname{dom} f$.
(iii) For any $c > 0$ and any non-negative $u \in L^1(\Omega)$, the sublevel set

$$\left\{ v \in L^1_+(\Omega) : d(v, u) \le c \right\}$$

is convex, weakly closed, and weakly compact in $L^1(\Omega)$.
(iv) The set $\partial d(\cdot, u^*)(u)$ is nonempty for $u^* \in \operatorname{dom} f$ if and only if u belongs to $L^\infty_+(\Omega)$ and is bounded away from zero. Moreover, $\partial d(\cdot, u^*)(u) = \{\log u - \log u^*\}$.

Finally, the Kullback–Leibler divergence provides a bound on the L^1 distance.

Lemma 5.3 *For any $u, v \in \operatorname{dom} f$, one has*

$$\|u - v\|_1^2 \le \left(\frac{2}{3}\|v\|_1 + \frac{4}{3}\|u\|_1 \right) d(v, u). \tag{5.3}$$

5.3 Variational Methods

Tikhonov regularization with additional convex constraints such as the non-negative cone is now classical; we refer, e.g., to [47] for linear and [13, 48] for nonlinear inverse problems in Hilbert spaces and to [27] for the Banach space setting. We therefore focus in this section on methods that are based on minimizing some combination of the regularization functional $\mathcal{R} \in \{f, d(\cdot, u_0)\}$, where $u_0 \in \operatorname{dom} f \subseteq L^1_+(\Omega)$ is an a priori guess, with the residual norm either as a penalty, i.e., as Tikhonov regularization

$$\min_{u \in L^1_+(\Omega)} \tfrac{1}{2}\|Au - y^\delta\|_Y^2 + \alpha \mathcal{R}(u)$$

for some regularization parameter $\alpha > 0$, or as a constraint, i.e., as Morozov regularization

$$\min_{u \in L^1_+(\Omega)} \mathcal{R}(u) \quad \text{s.t.} \quad \|Au - y^\delta\|_Y \leq \delta,$$

where δ is the noise level according to (5.2). Throughout this section we will set $X = L^1(\Omega)$ and assume $A : X \to Y$ to be a bounded linear operator mapping into some Banach space Y. Moreover, we will assume existence of a solution $u^\dagger$ to (5.1) with finite entropy $\mathcal{R}(u^\dagger) < \infty$ (which is therefore in particular non-negative).

5.3.1 Morozov-Entropy Regularization

The historically first study of regularizing properties of such methods can be found in [1] for the Morozov-entropy method

$$\min_{u \in L^1_+(\Omega)} f(u) \quad \text{s.t.} \quad \|Au - y^\delta\|_Y \leq \delta. \tag{5.4}$$

The reader is referred also to [3, Theorem 3.1], where a version of Morozov regularization is discussed.

We first of all state existence of a solution to (5.1) that maximizes the entropy, i.e., minimizes f.

Theorem 5.1 (Existence of Maximum Entropy Solution for Exact Data [1, Thm. 4]) *There exists a minimizer $u^\dagger \in L^1_+(\Omega)$ of*

$$\min_{u \in L^1_+(\Omega)} f(u) \quad \text{s.t.} \quad Au = y.$$

Theorem 5.2 (Existence of Regularizer [1, Thm. 1]) *For every $\delta > 0$ and $y^\delta \in Y$ satisfying (5.2), there exists a minimizer u^δ of (5.4).*

Both theorems follow from weak compactness of sublevel sets and weak lower semicontinuity of f (Lemma 5.1 (ii), (iii)) together with weak closedness and nonemptiness of $\{u \in L^1_+(\Omega) : Au = y\}$ and $\{u \in L^1_+(\Omega) : \|Au - y^\delta\|_Y \le \delta\}$.

Finally, it can be shown that (5.4) indeed defines a regularization method.

Theorem 5.3 (Convergence as $\delta \to 0$ [1, Thm. 5]) *Let $(y^\delta)_{\delta>0}$ be a family of data satisfying (5.2). Then*

$$\|u^\delta - u^\dagger\|_1 \to 0 \text{ as } \delta \to 0. \tag{5.5}$$

Stability of (5.4) in the sense that small perturbations in y^δ lead to small perturbations in u^δ has not been shown in [1]. Since one is actually interested in approaching $u^\dagger$ rather than u^δ, such stability results might be considered as of lower importance than the convergence in (5.5). We will therefore not state such stability results for the remaining variational methods either.

Convergence rates are not stated in [1], but they could be proved as in Theorem 5.6 or Theorem 5.10 below under a similar source condition.

5.3.2 Tikhonov-Entropy Regularization

The work [1] also mentions the Tikhonov-entropy regularization

$$\min_{u \in L^1_+(\Omega)} \tfrac{1}{2}\|Au - y^\delta\|_Y^2 + \alpha f(u), \tag{5.6}$$

pointing out from [60] that a regularization parameter choice $\alpha = \alpha(\delta)$ exists such that minimizers of (5.6) coincide with minimizers of (5.4). Hence, Theorem 5.3 also yields convergence of solutions of (5.6); however, this does not provide a concrete rule for choosing α.

A more detailed analysis of the constrained Tikhonov-entropy regularization

$$\min_{u \in \mathcal{D}} \tfrac{1}{2}\|Au - y^\delta\|_Y^2 + \alpha f(u) \tag{5.7}$$

with an appropriate subset $\mathcal{D}$ of $L^1_+(\Omega)$—including a priori regularization parameter choice rules—can be found in [24]. The analysis relies on a nonlinear transformation $T : L^2_{-e^{-1/2}}(\Omega) \to L^1(\Omega)$ for $L^2_{-e^{-1/2}}(\Omega) = \{v \in L^2(\Omega) : v \ge -e^{-1/2} \text{ a.e.}\}$ satisfying

$$f(T(v)) = \|v\|_2^2 - e^{-1/2}.$$

This leads to replacing (5.1) with the nonlinear problem $F(v) = y$ for $F := A \circ T : L^2_{-e^{-1/2}}(\Omega) \supseteq \mathcal{B} \to Y$. The theory on Tikhonov regularization for nonlinear problems in Hilbert spaces, in combination with a proof of weak sequential closedness of F, allows the authors of [24] to prove well-definedness and convergence of minimizers of (5.7) under the assumption that $\mathcal{B} = T^{-1}(\mathcal{D})$ is compact in measure.

Theorem 5.4 (Existence of Minimizers) *For every $\alpha > 0$, $\delta > 0$ and $y^\delta \in Y$ satisfying (5.2), there exists a minimizer u_α^δ of (5.7).*

This result also follows from Lemma 5.1 (ii), (iii) together with [24, Lem. 3.1,3.2].

Theorem 5.5 (Convergence as $\delta \to 0$ [24, Thm. 3.7]) *Let $(y^\delta)_{\delta>0}$ be a family of data satisfying (5.2), and let $\alpha = \alpha(\delta)$ be chosen such that*

$$\alpha \to 0 \quad \text{and} \quad \frac{\delta^2}{\alpha} \to 0 \quad \text{as } \delta \to 0. \tag{5.8}$$

Then

$$\|u_\alpha^\delta - u^\dagger\|_1 \to 0 \quad \text{as } \delta \to 0.$$

The reader is referred also to [3, Theorem 3.3] for convergence in the case of exact data.

Furthermore, a classical result on convergence rates for nonlinear Tikhonov regularization in Hilbert spaces from [23] can be employed to yield the following statement.

Theorem 5.6 (Convergence Rates [24, Thm. 3.8]) *Assume that $u^\dagger$ satisfies the source condition*

$$1 + \log u^\dagger = A^* w \tag{5.9}$$

for some sufficiently small $w \in Y^*$. Let $(y^\delta)_{\delta>0}$ be a family of data satisfying (5.2), and let $\alpha = \alpha(\delta)$ be chosen such that

$$\alpha \sim \delta \quad \text{as } \delta \to 0. \tag{5.10}$$

Then

$$\|u_\alpha^\delta - u^\dagger\|_1 = \mathcal{O}(\sqrt{\delta}) \quad \text{as } \delta \to 0.$$

The work [24] also treats the generalized negative Boltzmann–Shannon entropy

$$f(u) = \begin{cases} \int_\Omega u(t) \log \frac{u(t)}{u_0(t)}\, dt & \text{if } u \geq 0 \text{ a.e. and } u \log \frac{u}{u_0} \in L^1(\Omega), \\ +\infty & \text{otherwise,} \end{cases}$$

for some non-negative function u_0 carrying a priori information on u. In this case, the source condition (5.9) becomes

$$1 + \log \frac{u^\dagger}{u_0} = A^* w.$$

Finally, we point out that the convergence analysis in [24] includes the practically relevant case of inexact minimization of the Tikhonov functional.

5.3.3 Tikhonov–Kullback–Leibler Regularization

Replacing the negative Boltzmann–Shannon entropy in (5.7) by the Kullback–Leibler divergence results in the problem

$$\min_{u \in L^1_+(\Omega)} \tfrac{1}{2} \|Au - y^\delta\|_Y^2 + \alpha d(u, u_0), \tag{5.11}$$

which was investigated in [18]. In contrast to the analysis of the similar-looking problem (5.7) in [24], the analysis in [18] treats (5.11) directly by convexity arguments and does not require a nonlinear transformation. More precisely, the fact that both parts of the Tikhonov functional can be written as Bregman distances yields the estimate

$$\begin{aligned} &\tfrac{1}{2} \|A(u - u_\alpha^\delta)\|_Y^2 + \alpha d(u, u_\alpha^\delta) \\ &\qquad \leq \tfrac{1}{2} \|Au - y^\delta\|_Y^2 + \alpha d(u, u_0) - \tfrac{1}{2} \|Au_\alpha^\delta - y^\delta\|_Y^2 - \alpha d(u_\alpha^\delta, u_0) \end{aligned}$$

for a minimizer u_α^δ of (5.11) and all $u \in L^1_+(\Omega)$; cf. [9] and [18, Thm. 3.1]. In addition, the estimate

$$\tfrac{1}{2} \|A(\tilde{u}_\alpha^\delta - u_\alpha^\delta)\|_Y^2 + \alpha d(\tilde{u}_\alpha^\delta, u_\alpha^\delta) \leq 2 \|\tilde{y}^\delta - y^\delta\|_Y^2$$

holds for a minimizer $\tilde{u}_\alpha^\delta$ of (5.11) with y^δ replaced by $\tilde{y}^\delta$; cf. [18, Thm. 3.3]. These two inequalities are the basis for the following results.

We first consider existence of minimizers. Assuming the existence of a solution $u^\dagger \in L^1_+(\Omega)$ to (5.1) of finite Kullback–Leibler divergence, one obtains (as in Theorem 5.1) an existence result for exact data from the weak compactness of sublevel sets and the weak lower semicontinuity of $d(\cdot, u_0)$ (Lemma 5.2 (ii), (iii)) as well as the weak continuity of A.

Theorem 5.7 (Existence of Minimum-KL Solution for Exact Data) *For every $u_0 \in \operatorname{dom} f$, there exists a minimizer $u^\dagger$ of*

$$\min_{u \in L^1_+(\Omega)} d(\cdot, u_0) \quad s.t. \quad Au = y.$$

Next we consider the well-definedness of (5.11) for noisy data, where one also obtains uniqueness as well as a uniform positivity property of the minimizer.

Theorem 5.8 (Existence, Uniqueness, and Uniform Positivity of Regularizer [18, Thm. 2.4]) *For every $\alpha > 0$, $\delta > 0$, and $y^\delta \in Y$ satisfying (5.2), there exists a unique minimizer $u^\delta_\alpha \in L^1_+(\Omega)$ of (5.11). Moreover, $\frac{u^\delta_\alpha}{u_0}$ is bounded away from zero.*

From this, the following convergence and convergence rate results can be stated.

Theorem 5.9 (Convergence as $\delta \to 0$ [18, Thm. 4.1]) *Let $(y^\delta)_{\delta>0}$ be a family of data satisfying (5.2), and let $\alpha = \alpha(\delta)$ be chosen according to (5.8). Then*

$$\|u^\delta_\alpha - u^\dagger\|_1 \to 0 \quad as\ \delta \to 0.$$

Theorem 5.10 (Convergence Rates; [18, Thm. 4.2]) *Assume that $u^\dagger$ satisfies the source condition*

$$\log \frac{u^\dagger}{u_0} = A^* w. \tag{5.12}$$

Let $(y^\delta)_{\delta>0}$ be a family of data satisfying (5.2), and let $\alpha = \alpha(\delta)$ be chosen according to (5.10). Then

$$\|u^\delta - u^\dagger\|_1 = \mathcal{O}(\sqrt{\delta}) \quad as\ \delta \to 0. \tag{5.13}$$

We remark that in [18], compactness of A is assumed, but an inspection of the proofs shows that actually boundedness of A suffices, since this implies weak lower semicontinuity of the mapping $u \mapsto \frac{1}{2}\|Au - y^\delta\|_Y^2$ and therefore (by Lemma 5.2 (ii) and the elementary inequality $\liminf(a_n) + \liminf(b_n) \le \liminf(a_n + b_n)$) of the Tikhonov functional.

5.3.4 Nonquadratic Data Misfit

In the remainder of this section, we remark on some possible extensions and open problems.

First, the quadratic term $\frac{1}{2}\|Au - y^\delta\|_Y^2$ can be replaced by some other convex data misfit functional to take into account special features of the data or of the measurement noise. In particular, we mention the case of non-negative data resulting

from a positivity preserving operator $A : L^1_+(\Omega) \to L^1_+(\Omega)$ as in, e.g., [52] or Poisson noise as in, e.g., [63].

Indeed, it was shown in [52] that Theorems 5.8, 5.9, 5.10 also hold for the entropy–entropy regularization

$$\min_{u \in L^1_+(\Omega)} d(y^\delta, Au) + \alpha d(u, u_0)$$

with $Y = L^1(\Omega)$ and (5.2) replaced by

$$d(y^\delta, y) \leq \delta^2, \tag{5.14}$$

provided A is positivity preserving in the sense that $x > 0$ almost everywhere implies $Ax > 0$ almost everywhere.

The key estimates for proving convergence and convergence rates in this case are

$$d(y^\delta, Au_\alpha^\delta) + \alpha d(u_\alpha^\delta, u_0) \leq d(y^\delta, Au^\dagger) + \alpha d(u^\dagger, u_0) \leq \delta^2 + \alpha d(u^\dagger, u_0),$$

which follows from minimality and (5.14), and

$$\begin{aligned}
&d(y^\delta, Au_\alpha^\delta) + \alpha d(u_\alpha^\delta, u^\dagger) \\
&= d(y^\delta, Au_\alpha^\delta) + \alpha \left(d(u_\alpha^\delta, u^\dagger) - d(u^\dagger, u_0) - \int_\Omega \log \frac{u^\dagger(t)}{u_0(t)} (u_\alpha^\delta(t) - u^\dagger(t))\, dt \right) \\
&\leq \delta^2 - \alpha \langle w, A(u_\alpha^\delta - u^\dagger) \rangle_{Y^*,Y} \\
&= \delta^2 + \alpha \langle w, y - y^\delta \rangle_{Y^*,Y} - \alpha \langle w, Au_\alpha^\delta - y^\delta \rangle_{Y^*,Y} \\
&\leq \delta^2 + \tfrac{4}{3} \alpha \|w\|_{Y^*} (\|y^\delta\|_1 + \|y\|_1 + \|Au_\alpha^\delta\|_1)^{\frac{1}{2}} (\delta + d(y^\delta, Au_\alpha^\delta)^{\frac{1}{2}}).
\end{aligned}$$

The last two inequalities hold due to the source condition (5.12) and to (5.3). By using the a priori choice $\alpha \sim \delta$, the latter estimate implies $d(y^\delta, Au_\alpha^\delta) = \mathcal{O}(\delta^2)$ and $d(u_\alpha^\delta, u^\dagger) = \mathcal{O}(\delta)$. The inequality (5.3) now yields the rate (5.13).

Finally, as mentioned in [52], convergence can also be extended to the symmetric Kullback-Leibler functional as a regularizing term in

$$\min_{u \in L^1_+(\Omega)} d(y^\delta, Au) + \alpha (d(u, u_0) + d(u_0, u)).$$

5.3.5 *Measure Space Solutions*

In particular in the context of probability densities, it could be appropriate to look for solutions in the space of positive measures $\mathcal{M}_+(\Omega)$ instead of $L^1_+(\Omega)$, i.e., consider

$$\min_{u \in \mathcal{M}_+(\Omega)} \tfrac{1}{2}\|Au - y^\delta\|_Y^2 + \alpha \mathcal{R}(u). \tag{5.15}$$

For a definition of entropy functionals on measure spaces we refer, e.g., to [59]. Minimization of some data misfit with a norm penalty but without imposing non-negativity as in [8, 14] or with non-negativity constraints but without adding a penalty as in [15] has been shown to yield sparse solutions in the sense that the support of the minimizers will typically have zero Lebesgue measure. Here it would be interesting to investigate whether an entropy penalty $\mathcal{R}$ could overcome singularity of the optimality conditions (cf., e.g., [15]) for attainable data that is present also for the measure space norm as penalty. Other relevant possibilities include the Wasserstein-1 and Kantorovich–Rubinstein norms considered in [41].

5.3.6 Nonlinear Problems

A natural extension would be to consider Tikhonov regularization for a nonlinear operator $F : \mathrm{dom}(F) \to Y$, i.e.,

$$\min_{u \in \mathcal{D}} \tfrac{1}{2}\|F(u) - y^\delta\|_Y^2 + \alpha \mathcal{R}(u),$$

for $\mathcal{R} \in \{f, d(\cdot, u_0)\}$ and $\mathcal{D} \subseteq \mathrm{dom}(F) \cap L^1_+(\Omega)$. For Tikhonov-entropy regularization, the analysis of [24] by way of nonlinear transformation was extended to nonlinear operators in [25]. In addition, recent analysis for Tikhonov regularization with abstract regularization functionals from, e.g., [26, 31, 51, 62] together with Lemmas 5.1, 5.2 shows that the existence and convergence results from Theorems 5.4, 5.5, 5.8, 5.9 remain valid if A is replaced by a nonlinear operator which is weakly continuous on $L^1(\Omega)$ and Y and if $\mathcal{D}$ is weakly closed in $L^1(\Omega)$. Furthermore, Theorems 5.6, 5.10 can be recovered in the nonlinear case by replacing A in the source conditions (5.9), (5.12) by the Fréchet derivative $F'(u^\dagger)$. However, uniform positivity results like those from Theorem 5.8 and [52, Section 4.2] do not follow from this theory and would have to be subject of additional investigations.

5.3.7 Ivanov Regularization

Ivanov regularization (also called method of quasi-solutions), cf., e.g., [17, 35–37, 42, 49, 56], defines u_ρ^δ as a solution to

$$\min_{u \in L^1_+(\Omega)} \|Au - y^\delta\|_Y \quad \text{s.t.} \quad \mathcal{R}(u) \le \rho$$

with $\mathcal{R} \in \{f, d(\cdot, u_0)\}$. Here, regularization is controlled by the parameter $\rho > 0$, with larger parameters corresponding to weaker regularization. If the respective minimizers are unique, all three variational regularization methods (Tikhonov, Morozov, and Ivanov) are equivalent for a certain choice of the regularization parameters α and ρ, cf. [42, Thm. 2.3]. Nevertheless, a practically relevant regularization parameter choice might lead to different solutions; the three formulations also entail different numerical approaches, some of which might be better suited than others in concrete applications. Furthermore, as a counterexample in [42] shows, the methods are no longer equivalent in the non-convex case arising from a nonlinear forward operator F.

Concerning well-definedness and convergence for (5.3.7), one can again rely on general results on the entropy functionals as stated in Section 5.2. Indeed, the properties of sublevel sets according to Lemma 5.1 (iii) or Lemma 5.2 (iii) together with weak sequential lower semicontinuity of the mapping $u \mapsto \frac{1}{2}\|Au - y^\delta\|^2$ guarantee existence of a minimizer for any $\rho > 0$. In case the maximal entropy or the minimal Kullback–Leibler divergence of a solution to (5.1) is known, the ideal parameter choice is of course $\rho = \mathcal{R}(u^\dagger)$; note that this choice of ρ is independent of δ. In this case, we obtain from minimality of u_ρ^δ that

$$\|Au_\rho^\delta - y^\delta\|_Y \leq \|Au^\dagger - y^\delta\|_Y \leq \delta.$$

We can thus argue similarly to the convergence proof for the Morozov formulation (5.4) to obtain convergence $\|u_\rho^\delta - u^\dagger\|_1 \to 0$ as $\delta \to 0$. Convergence rates under source conditions of the type (5.9) can also be derived. To see this in case $\mathcal{R} = f$, observe that minimality of u_ρ^δ and admissibility of $u^\dagger$ together with Lemma 5.1 (vii) and (5.9) yields

$$d(u_\rho^\delta, u^\dagger) = f(u_\rho^\delta) - f(u^\dagger) - f'(u^\dagger, u_\rho^\delta - u^\dagger) \leq -\langle w, A(u_\rho^\delta - u^\dagger)\rangle_{Y^*,Y} \leq 2\|w\|_{Y^*}\delta.$$

Hence, (5.3) implies the rate (5.13). A practical choice of ρ can be carried out, e.g., by Morozov's discrepancy principle, see [39].

5.4 Iterative Methods

In practice, solutions to the approaches given in Section 5.3 cannot be computed directly but require iterative methods. This makes applying iterative regularization methods directly to (5.1) attractive. For the particular case of non-negativity constraints, there are two general approaches: The constraints can be imposed during the iteration by projection, or the iteration can be constructed such that non-negativity of the starting value is preserved. In this section, we will review examples of both methods.

5.4.1 Projected Landweber Method for Non-negative Solutions of Linear Ill-Posed Equations

The classical Landweber method for the solution of (5.1) in Hilbert spaces consists in choosing $u_0 = 0$, $\tau \in (0, 2\|A\|^{-2})$, and setting

$$u_{k+1} = u_k + \tau A^*(y - Au_k), \qquad k = 0, \dots.$$

By spectral methods, one can show that the iterates converge strongly to the minimum norm solution $u^\dagger$ for exact data $y \in \operatorname{ran} A$. For noisy data $y = y^\delta \in \overline{\operatorname{ran} A} \setminus \operatorname{ran} A$, one also initially observes convergence, but at some point the iterates start to diverge from the solution; this behavior is often referred to as *semi-convergence*. It is therefore necessary to choose an appropriate stopping index $k_* := k_*(\delta, y^\delta) < \infty$ such that $u^\delta_{k_*} \to u^\dagger$ as $\delta \to 0$; a frequent choice is a discrepancy principle, e.g., of Morozov.

This method was generalized in [21] to constrained inverse problems of the form

$$Au = y \quad \text{s.t.} \quad u \in C$$

for a convex and closed set $C \subset X$; in our context, the obvious choice is $X = L^2(\Omega)$ and

$$C = \{u \in X : u(x) \geq 0 \text{ for almost every } x \in \Omega\}.$$

The corresponding *projected Landweber method* then consists in the iteration

$$u_{k+1} = P_C\left[u_k + \tau A^*(y - Au_k)\right], \qquad k = 0, \dots, \tag{5.16}$$

where P_C denotes the metric (in our case pointwise almost everywhere) projection onto C. This coincides with a *forward–backward splitting* or *proximal gradient descent* applied to $\|Au - y\|_Y^2 + \delta_C(u)$, where δ_C denotes the indicator function of C in the sense of convex analysis; see, e.g., [16]. Thus, a standard proof yields weak convergence of the iterates in the case of exact data $y \in A(C) := \{Au : u \in C\}$.

Theorem 5.11 ([21, Thm. 3.2]) *Let $u_0 = 0$ and $\tau \in (0, 2\|A\|^{-2})$. If $y \in A(C)$, then the sequence of iterates $\{u_k\}_{k\in\mathbb{N}} \subset C$ of (5.16) converges weakly to a solution $u^\dagger \in C$ of $Au = y$.*

In contrast to the unconstrained Landweber iteration, strong convergence can only be shown under additional restrictive conditions or by including additional terms in the iteration; in the setting considered here, this holds if $\mathrm{Id} - \tau A^*A$ is compact, see [21, Thm. 3.3].

Regarding noisy data $y^\delta \notin \operatorname{ran} A$, the following stability estimate holds.

Theorem 5.12 ([21, Thm. 3.4]) *Let $u_0^\delta = u_0 = 0$ and $\tau \in (0, 2\|A\|^{-2})$. If $y \in A(C)$ and $y^\delta \in Y$ with $\|y^\delta - y\|_Y \leq \delta$, then the sequences of iterates $\{u_k\}_{k\in\mathbb{N}} \subset C$ and $\{u_k^\delta\}_{k\in\mathbb{N}} \subset C$ of (5.16) with y and y^δ, respectively, satisfy*

$$\|u_k^\delta - u_k\|_X \leq \tau \|A\| \delta k, \qquad k = 0, \ldots. \tag{5.17}$$

By usual arguments, this can be used—together with a monotonicity property for noisy data—to derive stopping rules and thus regularization properties. However, to the best of our knowledge, this has not been done in the literature so far. It should also be pointed out that the estimate (5.17) is weaker than in the unconstrained case, where an $\mathcal{O}(\sqrt{k})$ estimate can be shown.

In addition, [21] proposes a "dual" projected Landweber iteration: Setting $w_0 = 0$, compute for $k = 0, \ldots,$ the iterates

$$\begin{cases} u_k = P_C[A^* w_k], \\ w_{k+1} = w_k + \tau(y - A u_k). \end{cases}$$

(This can be interpreted as a *backward–forward splitting*.) Under the same assumptions as above, one obtains strong convergence $u_k \to u^\dagger$ (without assuming compactness of $\mathrm{Id} - \tau A^* A$), see [21, Thm. 3.5], and the stability estimate (5.17), see [21, Thm. 3.6]. Numerical examples for integral equations with non-negative solutions in $L^2(\Omega)$ using both methods can be found in [21]. Acceleration by stationary preconditioning—i.e., replacing the scalar τ by a fixed self-adjoint, positive definite, linear operator D—was considered in [50]. It is an open problem whether further acceleration by Nesterov-type extrapolation or a more general inertial approach is possible.

If X and Y are Banach spaces, the above iterations are not applicable. A version of Landweber iteration in Banach spaces that can treat convex constraints has been proposed in [2]. The iteration can be formulated as

$$\begin{cases} \xi_{k+1} = \xi_k - \tau A^* J_Y(y - A x_k), \\ x_{k+1} \in \arg\min\limits_{x \in X} G(x) - \langle \xi_{k+1}, x \rangle_X, \end{cases}$$

for $x_0 = 0$ and $\xi_0 = 0$, where J_Y denotes the so-called duality mapping between Y^* and Y, and $G : X \to \mathbb{R} \cup \{\infty\}$ is proper, convex, and lower semicontinuous. Here the choice $G = \delta_C$ yields the projected Landweber iteration in Banach spaces, while $G = \delta_C + f$ would be the choice if a non-negative minimum-entropy solution is searched for. However, convergence in [2] could only be shown under the assumption that the interior of C is non-empty in X (which is not the case for $X = L^p(\Omega)$, $p < \infty$; see, e.g., [7]), G is p-convex with $p \geq 2$ (in particular, excluding both $G = \delta_C$ and $G = \delta_C + f$), and Y is uniformly smooth (requiring Y to be reflexive). The first assumption was removed in [38], allowing application to the case $X = L^p(\Omega)$ for $2 \leq p < \infty$ with $G = \delta_C + f + \frac{1}{p}\|\cdot\|_X^p$ and $Y = L^q(\Omega)$ for $2 \leq q < \infty$. Another open issue is the practical realization for $p, q > 2$, in particular of the second step, which requires computing a generalized metric projection in a Banach space.

Alternatively, a natural first step of moving from Tikhonov-type variational regularization towards iterative methods is the so-called non-stationary Tikhonov regularization [30] (which is a proximal point method), whose entropy-based version can be formulated as

$$u_k = \arg \min_{u \in L^1_+(\Omega)} \tfrac{1}{2} \|Au - y^\delta\|_Y^2 + \alpha_k d(u, u_{k-1}),$$

where $\{\alpha_k\}_{k \in \mathbb{N}}$ is a bounded sequence of positive numbers. This has been shown to converge in finite dimension; see, e.g., [34] and the references therein. We expect that an analysis of the infinite-dimensional counterpart can be carried out using the tools presented in this review.

5.4.2 EM Method for Integral Equations with Non-negative Data and Kernel

We now consider (5.1) in the special case that A is a Fredholm integral operator of the first kind, i.e.,

$$A : L^1(\Omega) \to L^1(\Sigma), \qquad (Au)(s) = \int_\Omega a(s,t)u(t)\,dt\,, \tag{5.18}$$

where $\Omega, \Sigma \subset \mathbb{R}^d$, $d \geq 1$, are compact, and the kernel a and the data y are positive pointwise almost everywhere. In this case, the following multiplicative iteration can be seen to preserve non-negativity for $u_0 \geq 0$:

$$u_{k+1}(t) = u_k(t) \int_\Sigma \frac{a(s,t)y(s)}{(Au_k)(s)}\,ds, \quad t \in \Omega, \qquad k = 0, \ldots. \tag{5.19}$$

This method was introduced in [40] as the method of convergent weights, motivated by some problems arising in nuclear physics. Writing this concisely as

$$u_{k+1} = u_k\, A^* \frac{y}{Au_k}, \qquad k = 0, \ldots,$$

where the multiplication and division are to be understood pointwise almost everywhere, relates (5.19) to the popular method known in the finite-dimensional setting as the expectation-maximization (EM) algorithm for Poisson models for PET, cf. [57, 61], and as the Lucy–Richardson algorithm in astronomical imaging, see [43, 55].

The study of (5.19) was initiated by the series of papers [44–46] primarily for the setting $A : C([0,1]) \to C([0,1])$. More precisely, some monotonicity features have been derived, while convergence has not been shown yet. Modified

EM algorithms allowing for better convergence properties have been investigated in infinite-dimensional settings; see [19, 20].

In the following, we summarize, based on [19, 33], the convergence properties for the case $A : L^1(\Omega) \to L^1(\Sigma)$ using a similar notation as in [53, 54]. Specifically, we make the following assumptions:

(A1) The kernel a is a positive and measurable function satisfying

$$\int_\Sigma a(s,t)\,ds = 1 \qquad \text{for almost all } t \in \Omega.$$

(A2) There exist $m, M > 0$ such that

$$m \le a(s,t) \le M \qquad \text{a.e. on } \Sigma \times \Omega.$$

(A3) The exact data y in (5.1) satisfies $\int_\Sigma y(s)\,ds = 1$ and

$$y(s) \le M' \qquad \text{a.e. on } \Sigma$$

for some $M' > 0$.

(A4) Equation (5.1) admits a solution $u^\dagger \in L^1_+(\Omega) \setminus \{0\}$.

Furthermore, let

$$\Delta = \left\{ u \in L^1_+(\Omega) : \int_\Omega u(t)\,dt = 1 \right\}.$$

By noticing that a positive solution of (5.1) is also a minimizer of the function $u \mapsto d(y, Au)$ subject to $u \ge 0$ (which is related to a maximum likelihood problem in statistical setting), we formulate a classical monotonicity result for (5.19) for exact data.

Proposition 5.1 ([53, Prop. 3.3]) *Let* (A1) *and* (A3) *be satisfied and let* $u_0 \in \Delta$ *such that* $d(u^\dagger, u_0) < \infty$. *Then, for any* $k \ge 0$, *the iterates* u_k *generated by* (5.19) *satisfy*

$$\begin{aligned} d(u^\dagger, u_k) &< \infty, \\ d(u_{k+1}, u_k) &\le d(y, Au_k) - d(y, Au_{k+1}), \\ d(y, Au_k) - d(y, Au^\dagger) &\le d(u^\dagger, u_k) - d(u^\dagger, u_{k+1}). \end{aligned}$$

Therefore, the sequences $\{d(u^\dagger, u_k)\}_{k\in\mathbb{N}}$ *and* $\{d(y, Au_k)\}_{k\in\mathbb{N}}$ *are nonincreasing. Moreover,*

$$\lim_{k\to\infty} d(y, Au_k) = d(y, Au^\dagger),$$
$$\lim_{k\to\infty} d(u_{k+1}, u_k) = 0.$$

We point out that from this result, one also obtains that $\lim_{k\to\infty} \|Au_k - y\|_1 = 0$ and $\lim_{k\to\infty} \|u_{k+1} - u_k\|_1 = 0$.

Since ill-posedness is an infinite-dimensional phenomenon, and the EM algorithm has been shown to be highly unstable, we recall also the approach in [53] which investigates the noise influence on the iterations. Similar properties of the iterates (as in the noise free data case) are derived there by stopping the procedure according to a discrepancy rule, as one can see below.

For the noisy data y^δ, we make the following assumptions.

(A5) $y^\delta \in L^\infty(\Sigma)$ satisfies $\int_\Sigma y^\delta(s)\, ds = 1$ and

$$\|y^\delta - y\|_1 \le \delta, \qquad \delta > 0.$$

(A6) There exist $m_1, M_1 > 0$ such that for all $\delta > 0$,

$$m_1 \le y^\delta(s) \le M_1, \qquad \text{a.e. on } \Sigma.$$

For further use, we define

$$\gamma := \max\left\{ \left|\ln \tfrac{m_1}{M}\right|, \left|\ln \tfrac{M_1}{m}\right| \right\}, \tag{5.20}$$

where $m, M > 0$ are the constants from (A2).

Let now u_k^δ denote the iterates generated by (5.19) with y^δ in place of y and $u_0^\delta = u_0 \in \Delta$. In this case, the iterates get closer and closer to the solution as long as the residual lies above the noise level.

Theorem 5.13 ([53, Thm. 6.3]) *Fix $\delta > 0$. If assumptions (A1)–(A6) are satisfied, then*

$$d(u^\dagger, u_{k+1}^\delta) \le d(u^\dagger, u_k^\delta)$$

for all $k \ge 0$ such that

$$d(y^\delta, Au_k^\delta) \ge \delta\gamma.$$

The above result indicates a possible choice of the stopping index for the algorithm (5.19) as

$$k_*(\delta) = \min\left\{k \in \mathbb{N} : d(y^\delta, Au_k^\delta) \le \tau\delta\gamma\right\} \tag{5.21}$$

for some fixed $\tau > 1$ and γ as given by (5.20). The next statement guarantees existence of such a stopping index.

Theorem 5.14 ([53, Thm. 6.4]) *Let assumptions (A1)–(A6) be satisfied and choose $u_0 \in \Delta$ such that $d(u^\dagger, u_0) < \infty$. Then:*

(i) For all $\delta > 0$, there exists a $k_*(\delta)$ satisfying (5.21) and

$$k_*(\delta)\tau\delta\gamma \le k_*(\delta) d(y^\delta, Au^\delta_{k_*(\delta)-1}) \le d(u^\dagger, u_0) + k_*(\delta)\delta\gamma.$$

(ii) The stopping index $k_*(\delta)$ is finite with $k_*(\delta) = O\left(\delta^{-1}\right)$ and

$$\lim_{\delta \to 0^+} \|Au^\delta_{k_*(\delta)} - y\|_p = 0$$

for any $p \in [1, +\infty)$.

An interesting open problem would be to investigate the behavior of (5.19) in conjunction with other stopping rules, e.g., a monotone error-type rule defined by means of the KL divergence in a way similar to the one dealt with in [29].

5.4.3 Modified EM Algorithms

In this section, we present some modifications of algorithm (5.19) which improve its stability or its performance.

5.4.3.1 EM Algorithms with Smoothing Steps

In order to stabilize (5.19), the work [58] proposed the so-called EMS algorithm

$$u_{k+1} = S\left(u_k\, A^* \frac{y}{Au_k}\right), \qquad k = 0, \ldots,$$

with $u_0 \equiv 1$ and the smoothing operator

$$Su(s) = \int_\Omega b(s,t)u(t)\,dt,$$

where $b : \Omega \times \Omega \to \mathbb{R}$ is continuous, positive and obeys a normalization condition similar to (A1). Note that [58] presents the EMS method from a statistical perspective, while the continuous formulation mentioned above can be found in [19]. Although this yields faster convergence in practice, more information on limit points (or the unique limit point, as the numerical experiments strongly suggest) has not been provided.

The work [19] also proposes the following nonlinear smoothing procedure, called NEMS:

$$u_{k+1} = S\left(\mathcal{N}(u_k)\, A^* \frac{y}{Au_k}\right), \qquad k = 0, \ldots,$$

with $u_0 \equiv 1$ and

$$\mathcal{N}u(t) = \exp\left([S^*(\log u)](t)\right) \qquad \text{for all } t \in \Omega.$$

Properties typical to EM iterations are shown in [19]. In particular, [19, Thm. 4.1] proves that the iterates produced by the NEMS algorithm converge to solutions of

$$\min_{u \in L^1_+(\Omega)} d(y, A\mathcal{N}u) - \mathcal{N}u + \int_\Omega u,$$

in analogy with (5.19) which is designed for approximating minimizers in $L^1_+(\Omega)$ of $d(y, Au)$.

5.4.3.2 EM-Kaczmarz Type Algorithms

The ordered-subsets expectation maximization algorithm (OS-EM) proposed in [32] is a variation of the EM iteration that has proved to be quite efficient in computed tomography. It resembles a Kacmarz-type method, being conceptually based on grouping the data y into an ordered sequence of N subsets y_j. A single outer iteration step then is composed of N EM-steps, where in each step j one updates the current estimate by working only with the corresponding data subset y_j. An extension of the OS-EM to the infinite-dimensional setting was introduced in [28], and numerical experiments reported there indicate this to be at least as efficient as the classical discrete OS-EM algorithm.

Let Σ_j be (not necessarily disjoint) subsets of Σ with $\Sigma_0 \cup \cdots \cup \Sigma_{N-1} = \Sigma$, and denote $y_j := y|_{\Sigma_j}$. Set $a_j := a|_{\Omega \times \Sigma_j}$.

Thus, one can rewrite (5.18) as

$$A_j : L^1(\Omega) \to L^1(\Sigma_j), \quad (A_j u)(s) := \int_\Omega a_j(s,t)\, u(t)\, dt, \quad j = 0, \ldots, N-1. \tag{5.22}$$

Then (5.18) can be formulated as a system of integral equations of the first kind

$$A_j u = y_j, \qquad j = 0, \ldots, N-1. \tag{5.23}$$

Clearly, u is a solution of (5.23) if and only if u solves (5.18). Without loss of generality, one can work with a common domain Σ instead of Σ_j, thus considering $A_j : L^1(\Omega) \to L^1(\Sigma)$ and $y_j \in L^1(\Sigma)$. Therefore, the system (5.23) can be solved by simultaneously minimizing

$$d(y_j, A_j u), \qquad j = 0, \dots, N-1.$$

The corresponding OS-EM algorithm for solving system (5.23) can be written in the form

$$u_{k+1} = u_k \int_\Sigma \frac{a_j(s,\cdot)\, y_j(s)}{(A_j u_k)(s)}\, ds, \qquad k = 0, \dots, \tag{5.24}$$

where $j = [k] := (k \bmod N)$.

Under assumptions similar to the ones in Section 5.4.2, the following results hold for exact data.

Theorem 5.15 ([28, Thm. 3.3]) *Let the sequence $\{u_k\}_{k\in\mathbb{N}}$ be defined by iteration (5.24), and let $u^\dagger \in \Delta \setminus \{0\}$ be a solution of (5.22) with $d(u^\dagger, u_0) < \infty$. Then one has*

(i) $f_{[k]}(u_{k+1}) \le f_{[k]}(u_k)$, for every $k = 0, \dots$;
(ii) the sequence $\{d(u^\dagger, u_k)\}_{k\in\mathbb{N}}$ is nonincreasing;
(iii) $\lim_{k\to\infty} f_{[k]}(u_k) = 0$;
(iv) $\lim_{k\to\infty} d(u_{k+1}, u_k) = 0$;
(v) for each $0 \le j \le N-1$ and $p \in [1, \infty)$ we have

$$\lim_{m\to\infty} \|A_j u_{j+mN} - y_j\|_p = 0;$$

(vi) if $u_0 \in L^\infty(\Omega)$ and $\{u_k\}_{k\in\mathbb{N}}$ is bounded in $L^p(\Omega)$ for some $p \in (1, \infty)$, then there exists a subsequence converging weakly in $L^p(\Omega)$ to a solution of (5.23).

Finally, we address the case of noisy data, where the noise is allowed to be different for each group of data. Specifically, we consider $y_j^\delta \in L^1(\Sigma)$ with

$$\|y_j - y_j^\delta\|_1 \le \delta_j, \qquad j = 0, \dots, N-1,$$

and set $\delta := (\delta_0, \dots, \delta_{N-1})$. Correspondingly, the stopping rules are independently defined for each group of data. This leads to the following loping OS-EM iteration for (5.23) with noisy data:

$$u_{k+1}^\delta = \begin{cases} u_k^\delta \int_\Sigma \frac{a_{[k]}(s,\cdot)\, y_{[k]}^\delta(\cdot)}{(A_{[k]} u_k^\delta)(s)} ds & d(y_j^\delta, A_j u_k^\delta) > \tau\gamma\delta_{[k]}, \\ u_k^\delta & \text{else}, \end{cases}$$

for $\tau > 1$ and γ as defined in (5.20).

Under similar assumptions on kernels and data as in Section 5.4.2, one can show analogously to Theorem 5.14 that there exists a finite stopping index $k_*(\delta)$, after which the iteration for all groups is terminated, and that (under additional assumptions) $A_j u_{k^*(\delta)}^\delta \to y_j$ for all j and $u_{k^*(\delta)}^\delta \rightharpoonup u^\dagger$ in L^p as $\delta \to 0$, see [28, Thm. 4.4].

References

1. Amato, U., Hughes, W.: Maximum entropy regularization of Fredholm integral equations of the first kind. Inverse Problems **7**, 793–803 (1991)
2. Boţ, R.I., Hein, T.: Iterative regularization with a general penalty term—theory and application to L^1 and TV regularization. Inverse Problems **28**(10), 104010, 19 (2012)
3. Borwein, J.: On the failure of maximum entropy reconstruction for Fredholm equations and other infinite systems. Math Program **61**, 251–261 (1993)
4. Borwein, J., Goodrich, R., Limber, M.: A comparison of entropies in the underdetermined moment problem (1993). URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.56.7938. Technical report
5. Borwein, J., Lewis, A.: Convergence of best entropy estimates. SIAM Journal on Optimization **1**, 191–205 (1991)
6. Borwein, J., Lewis, A.: Duality relationships for entropy-like minimization problems. SIAM Journal on Control and Optimization **29**, 325–338 (1991)
7. Borwein, J., Limber, M.: On entropy maximization via convex programming (1996). URL wayback.cecm.sfu.ca/projects/MomEnt+/ent_max.ps.gz. Technical report
8. Bredies, K., Pikkarainen, H.K.: Inverse problems in spaces of measures. ESAIM: Control, Optimisation and Calculus of Variations **19**(1), 190–218 (2013)
9. Brègman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Comput. Math. and Math. Phys. **7**, 200–217 (1967)
10. Byrne, C.: Applied Iterative Methods. A K Peters, Ltd., Wellesley, MA (2008)
11. Byrne, C.: EM algorithms from a non-stochastic perspective. In: O. Scherzer (ed.) Handbook of Mathematical Methods in Imaging, second edn. Springer New York (2015)
12. Byrne, C., Eggermont, P.P.B.: EM algorithms. In: O. Scherzer (ed.) Handbook of Mathematical Methods in Imaging, second edn., pp. 305–388. Springer New York (2015)
13. Chavent, G., Kunisch, K.: Convergence of Tikhonov regularization for constrained ill-posed inverse problems. Inverse Problems **10**(1), 63 (1994)
14. Clason, C., Kunisch, K.: A measure space approach to optimal source placement. Computational Optimization and Applications **53**(1), 155–171 (2012)
15. Clason, C., Schiela, A.: Optimal control of elliptic equations with positive measures. Control, Optimisation and Calculus of Variations (ESAIM-COCV) **23**, 217–240 (2017)
16. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Modeling & Simulation **4**(4), 1168–1200 (2005)
17. Dombrovskaja, I., Ivanov, V.K.: On the theory of certain linear equations in abstract spaces. Sibirsk. Mat. Z. **6**, 499–508 (1965)
18. Eggermont, P.P.B.: Maximum entropy regularization for Fredholm integral equations of the first kind. SIAM Journal of Mathematical Analysis **24**, 1557–1576 (1993)
19. Eggermont, P.P.B.: Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. Applied Mathematics and Optimization **39**(1), 75–91 (1999)
20. Eggermont, P.P.B., LaRiccia, V.N.: Maximum penalized likelihood estimation and smoothed EM algorithms for positive integral equations of the first kind. Numer. Funct. Anal. Optim. **17**, 737–754 (1996)
21. Eicke, B.: Iteration methods for convexly constrained ill-posed problems in Hilbert space. Numerical Functional Analysis and Optimization **13**(5–6), 413–429 (1992)
22. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems, *Mathematics and its Applications*, vol. 375. Kluwer Academic Publishers Group, Dordrecht (1996)
23. Engl, H.W., Kunisch, K., Neubauer, A.: Convergence rates for Tikhonov regularisation of nonlinear ill-posed problems. Inverse Problems **5**, 523–540 (1989)
24. Engl, H.W., Landl, G.: Convergence rates for maximum entropy regularization. SIAM J. Num. Anal. **30**, 1509–1536 (1993)

25. Engl, H.W., Landl, G.: Maximum entropy regularization of nonlinear ill-posed problems. In: World Congress of Nonlinear Analysts '92, Vol. I–IV (Tampa, FL, 1992), pp. 513–525. de Gruyter, Berlin (1996)
26. Flemming, J.: Generalized Tikhonov regularization: Basic theory and comprehensive results on convergence rates. Ph.D. thesis, TU Chemnitz (2011). URL http://nbn-resolving.de/urn:nbn:de:bsz:ch1-qucosa-78152
27. Flemming, J., Hofmann, B.: Convergence rates in constrained Tikhonov regularization: equivalence of projected source conditions and variational inequalities. Inverse Problems **27**(8), 085001 (2011)
28. Haltmeier, M., Leitão, A., Resmerita, E.: On regularization methods of EM-Kaczmarz type. Inverse Problems **25**, 075008 (2009)
29. Hämarik, U., Kaltenbacher, B., Kangro, U., Resmerita, E.: Regularization by discretization in Banach spaces. Inverse Problems **32**, 035004 (2016)
30. Hanke, M., Groetsch, C.: Nonstationary iterated Tikhonov regularization. Journal of Optimization Theory and Applications **98**, 37–53 (1998)
31. Hofmann, B., Kaltenbacher, B., Pöschl, C., Scherzer, O.: A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Problems **23**(3), 987–1010 (2007)
32. Hudson, H.M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. IEEE Trans. Med. Imaging **13**, 601–609 (1994)
33. Iusem, A.: A short convergence proof of the EM algorithm for a specific Poisson model. Brazilian Journal of Probability and Statistics **6**, 57–67 (1992). URL http://www.jstor.org/stable/43601445
34. Iusem, A.: Metodos de Pontos Proximal EM Optimizacao. IMPA, Rio de Janeiro (1995)
35. Ivanov, V.K.: On linear problems which are not well-posed. Dokl. Akad. Nauk SSSR **145**, 270–272 (1962)
36. Ivanov, V.K.: On ill-posed problems. Mat. Sb. (N.S.) **61 (103)**, 211–223 (1963)
37. Ivanov, V.K., Vasin, V.V., Tanana, V.P.: Theory of Linear Ill-posed Problems and Its Applications. Inverse and Ill-posed Problems Series. VSP, Utrecht (2002)
38. Jin, Q., Wang, W.: Landweber iteration of Kaczmarz type with general non-smooth convex penalty functionals. Inverse Problems **29**(8), 085011, 22 (2013)
39. Kaltenbacher, B., Klassen, A.: On convergence and convergence rates for Ivanov and Morozov regularization and application to some parameter identification problems in elliptic PDEs. Inverse Problems **34**(5), 055008 (2018)
40. Kondor, A.: Method of convergent weights – an iterative procedure for solving Fredholm's integral equations of the first kind. Nuclear Instruments and Methods in Physics Research **216**, 177–181 (1983)
41. Lellmann, J., Lorenz, D.A., Schönlieb , C., Valkonen, T.: Imaging with Kantorovich–Rubinstein discrepancy. SIAM Journal on Imaging Sciences **7**(4), 2833–2859 (2014)
42. Lorenz, D., Worliczek, N.: Necessary conditions for variational regularization schemes. Inverse Problems **29**(7), 075016 (2013)
43. Lucy, L.: An iterative technique for the rectification of observed distributions. Astron. J. **7**, 81–92 (1975)
44. Mülthei, H.N.: Iterative continuous maximum-likelihood reconstruction methods. Math. Methods Appl. Sci. **15**, 275–286 (1992)
45. Mülthei, H.N., Schorr, B., Törnig, W.: On an iterative method for a class of integral equations of the first kind. Math. Methods Appl. Sci. **9**, 137–168 (1987)
46. Mülthei, H.N., Schorr, B., Törnig, W.: On properties of the iterative maximum likelihood reconstruction method. Math. Methods Appl. Sci. **11**, 331–342 (1989)
47. Neubauer, A.: Tikhonov-regularization of ill-posed linear operator equations on closed convex sets. Journal of Approximation Theory **53**(3), 304–320 (1988)
48. Neubauer, A.: Tikhonov regularisation for non-linear ill-posed problems: optimal convergence rates and finite-dimensional approximation. Inverse Problems **5**(4), 541 (1989)

49. Neubauer, A., Ramlau, R.: On convergence rates for quasi-solutions of ill-posed problems. Electron. Trans. Numer. Anal. **41**, 81–92 (2014). URL http://etna.math.kent.edu/volumes/2011-2020/vol41/abstract.php?vol=41&pages=81-92
50. Piana, M., Bertero, M.: Projected Landweber method and preconditioning. Inverse Problems **13**(2), 441–463 (1997)
51. Pöschl, C.: Tikhonov regularization with general residual term. Ph.D. thesis, University of Innsbruck (2008)
52. Resmerita, E., Anderssen, R.S.: A joint additive Kullback–Leibler residual minimization and regularization for linear inverse problems. Math. Methods Appl. Sci. **30**, 1527–1544 (2007)
53. Resmerita, E., Engl, H.W., Iusem, A.N.: The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. Inverse Problems **23**(6), 2575 (2007)
54. Resmerita, E., Engl, H.W., Iusem, A.N.: Corrigendum. The expectation-maximization algorithm for ill-posed integral equations: a convergence analysis. Inverse Problems **24**(5), 059801 (2008)
55. Richardson, W.H.: Bayesian-based iterative method of image restoration. J. Opt. Soc. Am. **62**, 55–59 (1972)
56. Seidman, T.I., Vogel, C.R.: Well posedness and convergence of some regularisation methods for non-linear ill posed problems. Inverse Problems **5**(2), 227 (1989)
57. Shepp, L.A., Vardi, Y.: Maximum likelihood reconstruction in positron emission tomography. IEEE Trans. Medical Imaging **1**, 113–122 (1982)
58. Silverman, B.W., Jones, M.C., Nychka, D.W., Wilson, J.D.: A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. J. Roy. Statist. Soc. B **52**, 271–324 (1990). URL http://www.jstor.org/stable/2345438
59. Stummer, W., Vajda, I.: On Bregman distances and divergences of probability measures. IEEE Trans. Information Theory **58**(3), 1277–1288 (2012)
60. Tikhonov, A.N., Arsenin, V.Y.: Solutions of Ill-Posed Problems. Wiley, New York (1977)
61. Vardi, Y., Shepp, L.A., Kaufmann, L.: A statistical model for positron emission tomography. J. Am. Stat. Assoc **80**, 8–37 (1985)
62. Werner, F.: Inverse problems with Poisson data: Tikhonov-type regularization and iteratively regularized Newton methods. Ph.D. thesis, University of Göttingen (2012). URL http://num.math.uni-goettingen.de/~f.werner/files/diss_frank_werner.pdf
63. Werner, F., Hohage, T.: Convergence rates in expectation for Tikhonov-type regularization of inverse problems with Poisson data. Inverse Problems **28**(10), 104004 (2012)

Chapter 6
Characterizations of Super-Regularity and Its Variants

Aris Danillidis, D. Russell Luke, and Matthew Tam

In fond remembrance of Jonathan M. Borwein

Abstract Convergence of projection-based methods for nonconvex set feasibility problems has been established for sets with ever weaker regularity assumptions. What has not kept pace with these developments is analogous results for convergence of optimization problems with correspondingly weak assumptions on the value functions. Indeed, one of the earliest classes of nonconvex sets for which convergence results were obtainable, the class of so-called *super-regular sets* (Lewis et al., Comput. Math. **9**(4), 485–513, 2009), has no functional counterpart. In this work, we amend this gap in the theory by establishing the equivalence between a property slightly stronger than super-regularity, which we call *Clarke super-regularity*, and *subsmoothness* of sets as introduced by Aussel, Daniilidis and Thibault (Amer. Math. Soc. **357**, 1275–1301, 2004). The bridge to functions shows that approximately convex functions studied by Ngai, Luc and Thera (J. Nonlinear Convex Anal. **1**, 155–176, 2000) are those which have Clarke super-regular epigraphs. Further classes of regularity of functions based on the corresponding regularity of their epigraph are also discussed.

Keywords Super-regularity · Subsmoothness · Approximately convex

AMS 2010 Subject Classification 49J53, 26B25, 49J52, 65K10

A. Danillidis
DIM-CMM, Universidad de Chile, Santiago, Chile
e-mail: arisd@dim.uchile.cl

D. R. Luke (✉)
Inst. Numerische & Angewandte Mathematik, Universität Göttingen, Göttingen, Niedersachsen, Germany
e-mail: r.luke@math.uni-goettingen.de

M. Tam
University of Göttingen, Göttingen, Germany
e-mail: m.tam@math.uni-goettingen.de

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_6

6.1 Introduction

The notion of a *super-regular* set was introduced by Lewis, Luke, and Malick [10] who recognized the property as an important ingredient for proving convergence of the method of alternating projections without convexity. This was generalized in subsequent publications [3, 6, 7, 11], with the weakest known assumptions guaranteeing local linear convergence of the alternating projections algorithm for two-set, consistent feasibility problems to date found in [15, Theorem 3.3.5]. The regularity assumptions on the individual sets in these subsequent works are vastly weaker than super-regularity, but what has not kept pace with these generalizations is their functional analogs. Indeed, it appears that the notion of a *super-regular function* has not yet been articulated. In this note, we bridge this gap between super-regularity of sets and functions as well as establishing connections to other known function-regularities in the literature. A missing link is yet another type of set regularity, what we call *Clarke super-regularity*, which is a slightly stronger version of super-regularity and, as we show, this is equivalent to other existing notions of regularity. For a general set that is not necessarily the epigraph of a function, we establish an equivalence between *subsmoothness* as introduced by Aussel, Daniilidis, and Thibault [1] and Clarke super-regularity.

To begin, in Section 6.2 we recall different concepts of the normal cones to a set as well as notions of set regularity, including Clarke regularity (Definition 6.3) and (limiting) super-regularity (Definition 6.4). Next, in Section 6.3 we introduce the notion of Clarke super-regularity (Definition 6.5) and relate it to the notion of subsmoothness (Theorem 6.1). We also provide an example illustrating that Clarke super-regularity at a point is a strictly weaker condition than Clarke regularity around the point (Example 6.2). Finally, in Section 6.4, we provide analogous statements for Lipschitz continuous functions, relating the class of approximately convex functions to super-regularity of the epigraph. After completing this work we received a preprint [16] which contains results of this flavor, including a characterization of (limiting) super-regularity in terms of (metric) subsmoothness.

6.2 Normal Cones and Clarke Regularity

The notation used throughout this work is standard for the field of variational analysis, as can be found in [14]. The *closed ball* of radius $r > 0$ centered at $x \in \mathbb{R}^n$ is denoted $\mathbb{B}_r(x)$ and the closed unit ball is denoted $\mathbb{B} := \mathbb{B}_1(0)$. *(metric) projector* onto a set $\Omega \subset \mathbb{R}^n$, denoted by $P_\Omega : \mathbb{R}^n \rightrightarrows \Omega$, is the multi-valued mapping defined by

$$P_\Omega(x) := \{\omega \in \Omega : \|x - \omega\| = d(x, \Omega)\},$$

where $d(x, \Omega)$ denotes the distance of the point $x \in \mathbb{R}^n$ to the set Ω. When Ω is nonempty and closed, its projector P_Ω is everywhere nonempty. A selection from the projector is called a *projection*.

Given a set Ω, we denote its *closure* by $\operatorname{cl}\Omega$, its *convex hull* by $\operatorname{conv}\Omega$, and its *conic hull* by $\operatorname{cone}\Omega$. In this work we shall deal with two fundamental tools in nonsmooth analysis; *normal cones* to sets and *subdifferentials* of functions (Section 6.4).

Definition 6.1 (Normal Cones) Let $\Omega \subseteq \mathbb{R}^n$ and let $\bar{\omega} \in \Omega$.

(i) The *proximal normal cone* of Ω at $\bar{\omega} \in \Omega$ is defined by

$$N_\Omega^{\mathrm{P}}(\bar{\omega}) = \operatorname{cone}\left(P_\Omega^{-1}\bar{\omega} - \bar{\omega}\right).$$

Equivalently, $\bar{\omega}^* \in N_\Omega^{\mathrm{P}}(\bar{\omega})$ whenever there exists $\sigma \geq 0$ such that

$$\langle \bar{\omega}^*, \omega - \bar{\omega}\rangle \leq \sigma\|\omega - \bar{\omega}\|^2, \quad \forall \omega \in \Omega.$$

(ii) The *Fréchet normal cone* of Ω at $\bar{\omega}$ is defined by

$$\hat{N}_\Omega(\bar{\omega}) = \left\{\bar{\omega}^* \in \mathbb{R}^n : \langle \bar{\omega}^*, \omega - \bar{\omega}\rangle \leq \mathrm{o}(\|\omega - \bar{\omega}\|),\ \forall \omega \in \Omega\right\},$$

Equivalently, $\bar{\omega}^* \in \hat{N}_\Omega(\bar{\omega})$, if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\langle \bar{\omega}^*, \omega - \bar{\omega}\rangle \leq \varepsilon\|\omega - \bar{\omega}\|, \quad \text{for all } \omega \in \Omega \cap \mathbb{B}_\delta(\bar{\omega}). \tag{6.1}$$

(iii) The *limiting normal cone* of Ω at $\bar{\omega}$ is defined by

$$N_\Omega(\bar{\omega}) = \operatorname*{Lim\,sup}_{\omega \to \bar{\omega}} \hat{N}_\Omega(\bar{\omega}),$$

where the limit superior denotes the *Painlevé–Kuratowski outer limit*.

(iv) The *Clarke normal cone* of Ω at $\bar{\omega}$ is defined by

$$N_\Omega^{\mathrm{C}}(\bar{\omega}) = \operatorname{cl}\operatorname{conv} N_\Omega(\bar{\omega}).$$

When $\bar{\omega} \notin \Omega$, all of the aforementioned normal cones at $\bar{\omega}$ are defined to be empty.

Central to our subsequent analysis is the notion of a *truncation* of a normal cone. Given $r > 0$, one defines the *r-truncated* version of each of the above cones to be its intersection with a ball centered at the origin of radius r. For instance, the *r-truncated proximal normal cone* of Ω at $\bar{\omega} \in \Omega$ is defined by

$$N_\Omega^{\mathrm{rP}}(\bar{\omega}) = \operatorname{cone}\left(P_\Omega^{-1}\bar{\omega} - \bar{\omega}\right) \cap \mathbb{B}_r,$$

that is, $\bar{\omega}^* \in N_\Omega^{\mathrm{rP}}(\bar{\omega})$ whenever $\|\bar{\omega}^*\| \leq r$ and for some $\sigma \geq 0$ we have

$$\langle \bar{\omega}^*, \omega - \bar{\omega} \rangle \leq \sigma \|\omega - \bar{\omega}\|^2, \quad \forall \omega \in \Omega.$$

In general, the following inclusions between the normal cones can deduce straightforwardly from their respective definitions:

$$N_\Omega^{\mathrm{P}}(\bar{\omega}) \subseteq \hat{N}_\Omega(\bar{\omega}) \subseteq N_\Omega(\bar{\omega}) \subseteq N_\Omega^{\mathrm{C}}(\bar{\omega}). \tag{6.2}$$

The regularity of sets is characterized by the relation between elements in the graph of the normal cones to the sets and directions constructable from points in the sets. The weakest kind of regularity of sets that has been shown to guarantee convergence of the alternating projections algorithm is *elemental subregularity* (see [7, Cor.3.13(a)] and [15, Theorem 3.3.5]). It was called *elemental* (sub)regularity in [8, Definition 5] and [11, Definition 3.1] to distinguish regularity of sets from regularity of collections of sets. Since we are only considering the regularity of sets, and later functions, we can drop the "elemental" qualifier in the present setting. We also streamline the terminology and variations on elemental subregularity used in [8, 11], replacing *uniform* elemental subregularity with a more versatile and easily distinguishable variant.

Definition 6.2 (Subregularity [8, Definition 5]) Let $\Omega \subseteq \mathbb{R}^n$ and $\bar{\omega} \in \Omega$. The set Ω is said to be *ε-subregular relative to Λ at $\bar{\omega}$ for* $(\hat{\omega}, \hat{\omega}^*) \in \operatorname{gph} N_\Omega$ if it is locally closed at $\bar{\omega}$ and there exists an $\varepsilon > 0$ together with a neighborhood U of $\bar{\omega}$ such that

$$\left\langle \hat{\omega}^* - (\omega' - \omega), \omega - \hat{\omega} \right\rangle \leq \varepsilon ||\hat{\omega}^* - (\omega' - \omega)|| \|\omega - \hat{\omega}\|, \quad \forall \omega' \in \Lambda \cap U, \ \forall \omega \in P_\Omega(\omega'). \tag{6.3}$$

If *for every* $\varepsilon > 0$ there is a neighborhood (depending on ε) such that (6.3) holds, then Ω is said to be *subregular relative to Λ at $\bar{\omega}$ for* $(\hat{\omega}, \hat{\omega}^*) \in \operatorname{gph} N_\Omega$.

The property that distinguishes the degree of regularity of sets is the diversity of vectors $(\hat{\omega}, \hat{\omega}^*) \in \operatorname{gph} N_\Omega$ for which (6.3) holds, as well as the choice of the set Λ. Of particular interest to us are *Clarke regular* sets, which satisfy (6.3) for all $\varepsilon > 0$ and for all Clarke normal vectors at $\bar{\omega}$.

Definition 6.3 (Clarke Regularity) The set Ω is said to be *Clarke regular* at $\bar{\omega} \in \Omega$ if it is locally closed at $\bar{\omega}$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $(\bar{\omega}, \bar{\omega}^*) \in \operatorname{gph} N_\Omega^{\mathrm{C}}$

$$\left\langle \bar{\omega}^*, \omega - \bar{\omega} \right\rangle \leq \varepsilon ||\bar{\omega}^*|| \|\omega - \bar{\omega}\|, \quad \forall \omega \in \Omega \cap \mathbb{B}_\delta(\bar{\omega}). \tag{6.4}$$

Note that (6.4) is (6.3) with $\Lambda = \Omega$ and $U = \mathbb{B}_\delta(\bar{\omega})$, which in the case of Clarke regularity holds for all $(\bar{\omega}, \bar{\omega}^*) \in \operatorname{gph} N_\Omega^{\mathrm{C}}$. A short argument shows that, for Ω Clarke regular at $\bar{\omega}$, the Clarke and Fréchet normal cones coincide at $\bar{\omega}$. Indeed, this property is used to *define* Clarke regularity in [14, Definition 6.4]. It is also

immediately clear from the definitions that if Ω is Clarke regular at $\bar{\omega}$, then it is subregular relative to $\Lambda = \Omega$ at $\bar{\omega}$ for all $\bar{\omega}^* \in N_\Omega(\bar{\omega})$.

By setting $\Lambda = \mathbb{R}^n$, letting $\hat{\omega} \in \Omega$ be in a neighborhood of $\bar{\omega}$ and fixing $\hat{\omega}^* = 0$ in the context of Definition 6.2, we arrive at super-regularity which, when stated explicitly, takes the following form.

Definition 6.4 (Super-Regularity [10, Definition 4.3]) Let $\Omega \subseteq \mathbb{R}^n$ and $\bar{\omega} \in \Omega$. The set Ω is said to be *super-regular at* $\bar{\omega}$ if it is locally closed at $\bar{\omega}$ and for every $\varepsilon > 0$ there is a $\delta > 0$ such that for all $(\hat{\omega}, 0) \in \operatorname{gph} N_\Omega \cap \{(\mathbb{B}_\delta(\bar{\omega}), 0)\}$

$$\langle \omega' - \omega, \hat{\omega} - \omega \rangle \leq \varepsilon \, ||\omega' - \omega|| \|\hat{\omega} - \omega\|, \quad \forall \omega' \in \mathbb{B}_\delta(\bar{\omega}), \ \forall \omega \in P_\Omega(\omega'). \tag{6.5}$$

Rewriting the above leads the following equivalent characterization of super-regularity, which is more useful for our purposes.

Proposition 6.1 ([10, Proposition 4.4]) *The set $\Omega \subseteq \mathbb{R}^n$ is* super-regular *at $\bar{\omega} \in \Omega$ if and only if it is locally closed at $\bar{\omega}$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that*

$$\langle \omega_1^*, \omega_2 - \omega_1 \rangle \leq \varepsilon \, ||\omega_1^*|| \|\omega_2 - \omega_1\|, \\ \forall (\omega_1, \omega_1^*) \in \operatorname{gph} N_\Omega \cap \left(\mathbb{B}_\delta(\bar{\omega}) \times \mathbb{R}^n\right), \forall \omega_2 \in \Omega \cap \mathbb{B}_\delta(\bar{\omega}) \tag{6.6}$$

It is immediately clear from this characterization that super-regularity implies Clarke regularity. By continuing our development of increasingly nicer regularity properties to convexity, we have the following relationships involving stronger notions of regularity.

Proposition 6.2 *Let $\Omega \subseteq \mathbb{R}^n$ be locally closed at $\bar{\omega} \in \Omega$.*

*(i) If Ω is prox-regular at $\bar{\omega}$ (*i.e., *there exists a neighborhood of $\overline{x}$ on which the projector is single-valued), then there is a constant $\gamma > 0$ such that for all $\varepsilon > 0$*

$$\langle \omega_1^*, \omega_2 - \omega_1 \rangle \leq \varepsilon ||\omega_1^*|| \|\omega_2 - \omega_1\|, \\ \forall (\omega_1, \omega_1^*) \in \operatorname{gph} N_\Omega \cap \left(\mathbb{B}_{\gamma\varepsilon}(\bar{\omega}) \times \mathbb{R}^n\right), \ \forall \omega_2 \in \Omega \cap \mathbb{B}_{\gamma\varepsilon}(\bar{\omega}). \tag{6.7}$$

(ii) If Ω is convex, then

$$\langle \omega_1^*, \omega_2 - \omega_1 \rangle \leq 0, \qquad \forall (\omega_1, \omega_1^*) \in \operatorname{gph} N_\Omega, \quad \forall \omega_2 \in \Omega. \tag{6.8}$$

Proof The proof of (i) can be found in [8, Proposition 4(vi)]. Part (ii) is classical. □

Example 6.1 (Pac-Man) Let $\overline{x} = 0 \in \mathbb{R}^2$ and consider two subsets of $\mathbb{R}^2$ given by

$$A = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 \leq 1, \ -(1/2)x_1 \leq x_2 \leq x_1, x_1 \geq 0\},$$

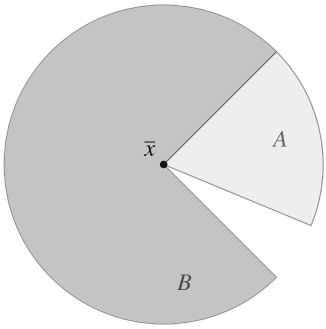


Fig. 6.1 An illustration of the sets in Example 6.1

$$B = \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1^2 + x_2^2 \leq 1,\ x_1 \leq |x_2|\}.$$

The set B looks like a "Pac-Man"' with mouth opened to the right and the set A, if you like, a piece of pizza. For an illustration, see Fig. 6.1. The set B is subregular relative to A at $\overline{x} = 0$ for all $(b, v) \in \operatorname{gph}(N_B \cap A)$ for $\varepsilon = 0$ on all neighborhoods since, for all $a \in A$, $a_B \in P_B(a)$ and $v \in N_B(b) \cap A$. To see this, we simply note that

$$\langle v - (a - a_B),\ a_B - b\rangle = \langle v,\ a_B - b\rangle - \langle a - a_B,\ a_B - b\rangle = 0.$$

In other words, from the perspective of the piece of pizza, Pac-Man looks *convex*. The set B, however, is only ε-subregular at $\overline{x} = 0$ relative to $\mathbb{R}^2$ for any $v \in N_B(0)$ for $\varepsilon = 1$ since, by choosing $x = tv \in B$ (where $0 \neq v \in B \cap N_B(0)$, $t \downarrow 0$), we have $\langle v,\ x\rangle = \|v\|\|x\| > 0$. Clearly, this also means that Pac-Man is not Clarke regular.

6.3 Super-Regularity and Subsmoothness

In the context of the definitions surveyed in the previous section, we introduce an even stronger type of regularity that we identify, in Theorem 6.1, with *subsmoothness* as studied in [1]. This will provide a crucial link to the analogous characterizations of regularity for *functions* considered in Theorem 6.2, in particular, to *approximately convex* functions studied in [12].

Definition 6.5 (Clarke Super-Regularity) Let $\Omega \subseteq \mathbb{R}^n$ and $\bar{\omega} \in \Omega$. The set Ω is said to be *Clarke super-regular* at $\bar{\omega}$ if it is locally closed at $\bar{\omega}$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $(\hat{\omega}, \hat{\omega}^*) \in \operatorname{gph} N_\Omega^{\mathrm{C}} \cap (\mathbb{B}_\delta(\bar{\omega}) \times \mathbb{R}^n)$, the following inequality holds

$$\left\langle \hat{\omega}^*, \omega - \hat{\omega} \right\rangle \leq \varepsilon\, ||\hat{\omega}^*||\|\omega - \hat{\omega}\|, \quad \forall \omega \in \Omega \cap \mathbb{B}_\delta(\bar{\omega}). \tag{6.9}$$

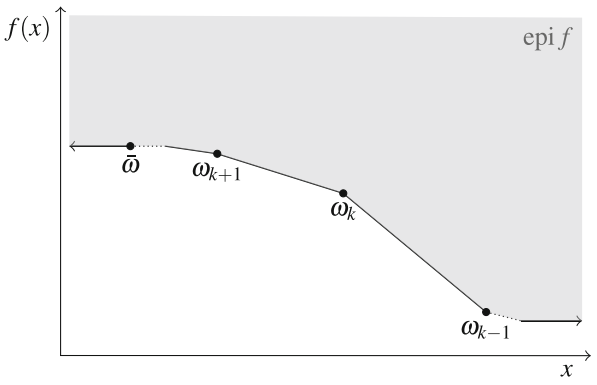


Fig. 6.2 A sketch of the function f and the sequence (ω_k) given in Example 6.2

The only difference between Clarke super-regularity and super-regularity is that, in the case of Clarke super-regularity, the key inequality above holds for all nonzero Clarke normals in a neighborhood instead holding only for limiting normals (compare (6.6) with (6.9)). It therefore follows that Clarke super-regularity at a point implies Clarke regularity there. Nevertheless, even this stronger notion of regularity does not imply Clarke regularity around $\bar{\omega}$, as the following counterexample shows.

Example 6.2 (Regularity Only At a Point) Let $f : \mathbb{R}^2 \to \mathbb{R}$ be the continuous, piecewise linear function (see Figure 6.2) defined by

$$f(x) := \begin{cases} 0, & \text{if } x \le 0 \\ -\frac{1}{2^k}(x - \frac{1}{2^k}) - \frac{1}{3.4^k}, & \text{if } \frac{1}{2^{k+1}} \le x \le \frac{1}{2^k} \\ -\frac{1}{12}, & \text{if } x \ge \frac{1}{2}. \end{cases} \quad (\text{for } k = 1, 2, \dots)$$

Notice that

$$-\frac{4}{3}x^2 \le f(x) \le -\frac{1}{3}x^2, \quad \forall x \in \left[0, \frac{1}{2}\right]. \tag{6.10}$$

Let $\Omega = \operatorname{epi} f$ denote the epigraph of f. Thanks to (6.10) it is easily seen that Ω is Clarke regular at $\bar{\omega} = (0, 0)$ in the sense of Definition 6.3. However, Ω is not Clarke regular at the sequence of points $\omega_k = (\frac{1}{2^{k+1}}, \frac{1}{2^k})$ converging to $\bar{\omega}$. Indeed, the Fréchet normal cones $\hat{N}_\Omega(\omega_k)$ are reduced to $\{0\}$ for all $k \ge 1$, while the corresponding limiting normal cones are given by

$$N_\Omega(\omega_k) = \mathbb{R}_+ \left\{ \left(-\frac{1}{2^k}, -1\right), \left(-\frac{1}{2^{k+1}}, -1\right) \right\}, \quad \forall k \in \mathbb{N}.$$

A missing link in the cascade of set regularity is subsmooth and semi-subsmooth sets, introduced and studied by Aussel, Daniilidis and Thibault in [1, Definitions 3.1 & 3.2].

Definition 6.6 ((Semi-)subsmooth Sets) Let $\Omega \subset \mathbb{R}^n$ be closed and let $\bar{\omega} \in \Omega$.

(i) The set Ω is *subsmooth* at $\bar{\omega} \in \Omega$ if, for every $r > 0$ and $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\omega_1, \omega_2 \in \mathbb{B}_\delta(\bar{\omega}) \cap \Omega$, all $\omega_1^* \in N_\Omega^{\mathrm{rC}}(\omega_1)$ and all $\omega_2^* \in N_\Omega^{\mathrm{rC}}(\omega_2)$ we have:

$$\langle \omega_1^* - \omega_2^*, \omega_1 - \omega_2 \rangle \geq -\varepsilon \|\omega_1 - \omega_2\|. \tag{6.11}$$

(ii) The set Ω is *semi-subsmooth* at $\bar{\omega}$ if, for every $r > 0$ and $\varepsilon > 0$, there exists $\delta > 0$ such that for all $\omega \in \mathbb{B}_\delta(\bar{\omega}) \cap \Omega$, all $\omega^* \in N_\Omega^{\mathrm{rC}}(\omega)$ and all $\bar{\omega}^* \in N_\Omega^{\mathrm{rC}}(\bar{\omega})$

$$\langle \omega^* - \bar{\omega}^*, \omega - \bar{\omega} \rangle \geq -\varepsilon \|\omega - \bar{\omega}\|. \tag{6.12}$$

It is clear from the definitions that subsmoothness at a point implies semi-subsmoothness at the same point. The next theorem establishes the precise connection between subsmoothness and Clarke super-regularity (Definition 6.5).

Theorem 6.1 (Characterization of Subsmoothness) *Let $\Omega \subseteq \mathbb{R}^n$ be closed and nonempty.*

(i) The set Ω is subsmooth at $\bar{\omega} \in \Omega$ if and only if Ω is Clarke super-regular at $\bar{\omega}$.
(ii) The set Ω is semi-subsmooth at $\bar{\omega} \in \Omega$ if and only if for each constant $\varepsilon > 0$ there is a $\delta > 0$ such that for every $(\bar{\omega}, \bar{\omega}^*) \in \operatorname{gph} N_\Omega^{\mathrm{C}}$

$$\left\langle \bar{\omega}^*, \omega - \bar{\omega} \right\rangle \leq \varepsilon \, ||\bar{\omega}^*|| \|\omega - \bar{\omega}\|, \quad \forall \omega \in \Omega \cap \mathbb{B}_\delta(\bar{\omega})$$

and for all $(\omega, \omega^*) \in \operatorname{gph} N_\Omega^{\mathrm{C}} \cap (\mathbb{B}_\delta(\bar{\omega}) \times \mathbb{R}^n)$,

$$\left\langle \omega^*, \bar{\omega} - \omega \right\rangle \leq \varepsilon \, ||\omega^*|| \|\bar{\omega} - \omega\|.$$

Proof (i) Assume Ω is subsmooth at $\bar{\omega} \in \Omega$ and fix an $\varepsilon > 0$. Set $r = 1$ and let $\delta > 0$ be given by the definition of subsmoothness. Then for every $\omega_1, \omega_2 \in \Omega \cap \mathbb{B}_\delta(\bar{\omega})$ and $\omega_2^* \in N_\Omega^{\mathrm{C}}(\omega_2) \setminus \{0\}$, applying (6.11) for $\omega_1^* = \{0\} \in N_\Omega^{(\mathrm{r}=1)\mathrm{C}}(\omega_1)$ and $||\omega_2^*||^{-1}\omega_2^* \in N_\Omega^{(\mathrm{r}=1)\mathrm{C}}(\omega_2)$ we deduce (6.9). The same argument applies in the case that $\omega_2^* = 0$ and $\omega_1^* \neq 0$. If both $\omega_1^* = \omega_2^* = 0$, then the required inequality holds trivially.

Let us now assume that Ω is Clarke super-regular at $\bar{\omega}$ and fix $r > 0$ and $\varepsilon > 0$. Let $\delta > 0$ be given by the definition of Clarke super-regularity corresponding to $\varepsilon' = \varepsilon/2r$ and let $\omega_1, \omega_2 \in \mathbb{B}_\delta(\bar{\omega}) \cap \Omega$, $\omega_1^* \in N_\Omega^{\mathrm{rC}}(\omega_1)$ and $\omega_2^* \in N_\Omega^{\mathrm{rC}}(\omega_2)$. It follows from (6.9) that

$$\langle \omega_1^*, \omega_1 - \omega_2 \rangle \geq -\frac{\varepsilon}{2r} ||\omega_1^*|| \|\omega_1 - \omega_2\| \geq -\frac{\varepsilon}{2} \|\omega_1 - \omega_2\|$$

and

$$\langle -\omega_2^*, \omega_1 - \omega_2 \rangle \geq -\frac{\varepsilon}{2r} ||\omega_2^*|| \|\omega_1 - \omega_2\| \geq -\frac{\varepsilon}{2} \|\omega_1 - \omega_2\|.$$

We conclude by adding the above inequalities.

Part (ii) is nearly identical and the proof is omitted. □

The following corollary utilizes Theorem 6.1 to summarize the relations between various notions of regularity for sets, the weakest of these being the weakest known regularity assumption under which local convergence of alternating projections has been established [15, Theorem 3.3.5].

Corollary 6.1 *Let $\Omega \subseteq \mathbb{R}^n$ be closed, let $\bar{\omega} \in \Omega$ and consider the following assertions.*

(i) *Ω is prox-regular at $\bar{\omega}$.*
(ii) *Ω is subsmooth at $\bar{\omega}$.*
(iii) *Ω is Clarke super-regular at ω.*
(iv) *Ω is (limiting) super-regular at ω.*
(v) *Ω is Clarke regular at ω.*
(vi) *Ω is subregular at ω relative to some nonempty $\Lambda \subset \mathbb{R}^n$ for all $(\omega, \omega^*) \in V \subset \operatorname{gph} N_\Omega^{\mathrm{P}}$.*

Then (i) $\Longrightarrow$ (ii) $\Longleftrightarrow$ (iii) $\Longrightarrow$ (iv) $\Longrightarrow$ (v) $\Longrightarrow$ (vi).

Proof (i) $\Longrightarrow$ (ii): This was shown in [1, Proposition 3.4(ii)]. (ii) $\Longleftrightarrow$ (iii): This is Theorem 6.1(i). (iii) $\Longrightarrow$ (iv) $\Longrightarrow$ (v) $\Longrightarrow$ (vi): These implications follow from the definitions. □

Remark 6.1 (Amenablility) A further regularity between convexity and prox-regularity is *amenability* [14, Definition 10.23]. This was shown in [13, Corollary 2.12] to imply prox-regularity. Amenability plays a larger role in the analysis of functions and is defined precisely in this context below.

6.4 Regularity of Functions

The extension of the above notions of set regularity to analogous notions for functions typically passes through the epigraphs. Given a function $f : \mathbb{R}^n \to [-\infty, +\infty]$, recall that its *domain* is $\operatorname{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$ and its *epigraph* is

$$\operatorname{epi} f := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq \alpha\}.$$

The *subdifferential* of a function at a point $\bar{x}$ can be defined in terms of the normal cone to its epigraph at that point. Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ and let $\bar{x} \in \operatorname{dom} f$. The *proximal subdifferential* of f at $\bar{x}$ is defined by

$$\partial^{\mathrm{P}} f(\bar{x}) = \{v \in \mathbb{R}^n : (v, -1) \in N^{\mathrm{P}}_{\operatorname{epi} f}((\bar{x}, f(\bar{x}))\}. \tag{6.13}$$

The *Fréchet (resp. limiting, Clarke) subdifferential*, denoted $\hat{\partial} f(\bar{x})$ (resp. $\partial f(\bar{x})$, $\partial^{\mathrm{C}} f(\bar{x})$), is defined analogously by replacing normal cone $N^{\mathrm{P}}_{\operatorname{epi} f}(\bar{\omega})$ with $\hat{N}_{\operatorname{epi} f}(\bar{\omega})$ (resp. $N_{\operatorname{epi} f}(\bar{\omega})$, $N^{\mathrm{C}}_{\operatorname{epi} f}(\bar{\omega})$) in (6.13) where $\bar{\omega} = (\bar{x}, f(\bar{x}))$. The *horizon and Clarke horizon subdifferentials* at $\overline{x}$ are defined, respectively, by

$$\partial_\infty f(\bar{x}) = \{v \in \mathbb{R}^n : (v, 0) \in N_{\operatorname{epi} f}((\bar{x}, f(\bar{x}))\},$$
$$\partial^{\mathrm{C}}_\infty f(\bar{x}) = \{v \in \mathbb{R}^n : (v, 0) \in N^{\mathrm{C}}_{\operatorname{epi} f}((\bar{x}, f(\bar{x}))\}.$$

In what follows, we define regularity of functions in terms of the regularity of their epigraphs. We refer to a regularity defined in such a way as *epi-regularity*.

Definition 6.7 (Epi-Regular Functions) Let $f : \mathbb{R}^n \to (-\infty, +\infty]$, $\bar{x} \in \operatorname{dom} f$, $\Lambda \subseteq \operatorname{dom} f$, and $(\overline{y}, \overline{v}) \in \operatorname{gph} \partial f \cup \operatorname{gph} \partial_\infty f$.

(i) f is said to be *ε-epi-subregular* at $\bar{x} \in \operatorname{dom} f$ relative to $\Lambda \subseteq \operatorname{dom} f$ for $(\overline{y}, \overline{v})$ whenever epi f is *ε-subregular* at $\bar{x} \in \operatorname{dom} f$ relative to $\{(x, \alpha) \in \operatorname{epi} f \mid x \in \Lambda\}$ for $(\overline{y}, (\overline{v}, e))$ with fixed $e \in \{-1, 0\}$.

(ii) f is said to be *epi-subregular* at $\bar{x}$ relative to $\Lambda \subseteq \operatorname{dom} f$ for $(\overline{y}, \overline{v})$ whenever epi f is *subregular* at $(\bar{x}, f(\overline{x}))$ relative to $\{(x, \alpha) \in \operatorname{epi} f \mid x \in \Lambda\}$ for $(\overline{y}, (\overline{v}, e))$ with fixed $e \in \{-1, 0\}$.

(iii) f is said to be *epi-Clarke regular* at $\bar{x}$ whenever epi f is Clarke regular at $(\bar{x}, f(\overline{x}))$. Similarly, the function is said to be *epi-Clarke super-regular* (resp. epi-super-regular, epi-prox-regular) at $\bar{x}$ whenever its epigraph is Clarke super-regular (resp. super-regular, or prox-regular) at $(\bar{x}, f(\overline{x}))$.

Recent work [2, 4] makes use of the *directional* regularity (in particular Lipschitz regularity) of functions or their gradients. The next example illustrates how this fits naturally into our framework.

Example 6.3 The negative absolute value function $f(x) = -|x|$ is the classroom example of a function that is not Clarke regular at $x = 0$. It is, however, ε-epi-subregular relative to $\mathbb{R}$ at $x = 0$ for all limiting subdifferentials there for the same reason that the Pac-Man of Example 6.1 is ε-subregular relative to $\mathbb{R}^2$ at the origin for $\varepsilon = 1$. Indeed, $\partial f(0) = \{-1, +1\}$ and at any point (x, y) below epi f the vector $(x, y) - P_{\operatorname{epi} f}(x, y) \in \{\alpha(-1, -1), \alpha(1, -1)\}$ with $\alpha \geq 0$. So by the Cauchy-Schwarz inequality

$$\begin{aligned}&\big\langle (\pm 1, -1) - \alpha(\pm 1, -1), P_{\operatorname{epi} f}(x, y)\big\rangle \\ &\quad \leq \|(\pm 1, -1) - \alpha(\pm 1, -1)\| \|P_{\operatorname{epi} f}(x, y)\|, \quad \forall (x, y) \in \mathbb{R}^2. \end{aligned} \tag{6.14}$$

In particular, any point $(x, x) \in \operatorname{gph} f$ we have

$$P_{\operatorname{epi} f}(x, x) = (x, x) \quad \text{and} \quad (x, x) - P_{\operatorname{epi} f}(x, x) = (0, 0),$$

so the inequality is tight for the subgradient $-1 \in \partial f(0)$. Following (6.3), this shows that epi f is ε-subregular at the origin relative to $\mathbb{R}^2$ for all limiting normals (in fact, for all Clarke normals) at $(0, 0)$ for $\varepsilon = 1$. In contrast, the function f is not epi-subregular at $x = 0$ relative to $\mathbb{R}$ since the inequality above is tight on all balls around the origin, just as with the Pac-Man of Example 6.1. If one employs the restriction $\Lambda = \{x \mid x < 0\}$, then epi-subregularity of f is recovered at the origin relative to the negative orthant for the subgradient $v = 1$ for $\varepsilon = 0$ on the neighborhood $U = \mathbb{R}$, that is, $-|x|$ looks *convex* from this direction!

In a subsequent section, we develop an equivalent, though more elementary, characterizations of these regularities of functions defined in Definition 6.7.

6.4.1 Lipschitz Continuous Functions

In this section, we consider the class of locally Lipschitz functions, which allows us to avoid the horizon subdifferential (since this is always $\{0\}$ for Lipschitz functions). Recall that a set Ω is called *epi-Lipschitz* at $\bar{\omega} \in \Omega$ if it can be represented near $\bar{\omega}$ as the epigraph of a Lipschitz continuous function. Such a function is called a *locally Lipschitz representation* of Ω at $\bar{\omega}$.

The following notion of *approximately convex* functions was introduced by Ngai, Luc and Thera [12] and turns out to fit naturally within our framework.

Definition 6.8 (Approximate Convexity) A function $f : \mathbb{R}^n \to (-\infty, +\infty]$ is said to be *approximately convex* at $\bar{x} \in \mathbb{R}^n$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$(\forall x, y \in \mathbb{B}_\delta(\bar{x}))(\forall t \in]0, 1[\,) :$$
$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) + \varepsilon t(1-t)\|x - y\|.$$

Daniilidis and Georgiev [5] and subsequently Daniilidis and Thibault [1, Theorem 4.14] showed the connection between approximately convex functions and subsmooth sets. Using our results in the previous section, we are able to provide the following extension of their characterization. In what follows, set $\omega = (x, t) \in \mathbb{R}^n \times \mathbb{R}$ and denote by $\pi(\omega) = x$ its projection onto $\mathbb{R}^n$.

Proposition 6.3 (Subsmoothness of Lipschitz Epigraphs) *Let Ω be an epi-Lipschitz subset of $\mathbb{R}^n$ and let $\bar{\omega} \in \operatorname{bdry}\Omega$. Then the following statements are equivalent:*

(i) Ω is Clarke super regular at $\bar{\omega}$.
(ii) Ω is subsmooth at $\bar{\omega}$.

(iii) every locally Lipschitz representation f of Ω at $\bar{\omega}$ is approximately convex at $\pi(\bar{\omega})$.
(iv) some locally Lipschitz representation f of Ω at $\bar{\omega}$ is approximately convex at $\pi(\bar{\omega})$.

Proof The equivalence of (i) and (ii) follows from Theorem 6.1(i), and does not require Ω to be epi-Lipschitz. The equivalence of (ii), (iii), and (iv) by [1, Theorem 4.14]. □

Remark 6.2 The equivalences in Theorem 6.3 actually hold in the Hilbert space setting without any changes. In fact, the equivalence of (ii)-(iv) remains true in Banach spaces [1, Theorem 4.14].

The following characterization extends [5, Theorem 2].

Theorem 6.2 (Characterizations of Approximate Convexity) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz on $\mathbb{R}^n$ and let $\bar{x} \in \mathbb{R}^n$. Then the following are equivalent:*

(i) epi f *is Clarke super-regular at $\bar{x}$.*
(ii) f is approximately convex at $\bar{x}$.
(iii) For every $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$(\forall x, y \in \mathbb{B}_\delta(\bar{x}))(\forall v \in \partial^{\mathrm{C}} f(x)) \qquad f(y) - f(x) \geq \langle v, y - x\rangle - \varepsilon\|y - x\|.$$

(iv) ∂f is submonotone [5, Definition 7] at x_0, that is, for every ε there is a δ such that for all $x_1, x_2 \in \mathbb{B}_\delta(x_0) \cap \operatorname{dom} \partial f$, and all $x_i^* \in \partial f(x_i)$ $(i = 1, 2)$, one has

$$\langle x_1^* - x_2^*,\ x_1 - x_2\rangle \geq -\varepsilon\|x_1 - x_2\|. \tag{6.15}$$

Proof (i) $\iff$ (ii): Since f is locally Lipschitz at $\bar{x}$, it is trivially a local Lipschitz representation of $\Omega = \operatorname{epi} f$ at $\bar{\omega} = (\bar{x}, f(\bar{x})) \in \Omega$. The result thus follows from Proposition 6.3. (ii) $\iff$ (iii) $\iff$ (iv): This is [5, Theorem 2]. □

6.4.2 Non-Lipschitzian Functions

In this section, we collect results which hold true without assuming local Lipschitz continuity.

Proposition 6.4 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be lower semicontinuous (lsc) and approximately convex. Then* epi f *is Clarke-super regular.*

Proof As a proper, lsc, approximately convex function is locally Lipschitz at each point in the interior of its domain [12, Proposition 3.2] and dom $f = \mathbb{R}^n$, the result follows from Theorem 6.2. □

Example 6.4 (Clarke Super-Regularity Does Not Imply Approximate Convexity) Consider the counting function $f : \mathbb{R}^n \to \{0, 1, \ldots, n\}$ defined by

$$f(x) = \|x\|_0 := \sum_{j=1}^{n} |\mathrm{sign}(x_j)|, \quad \text{where} \quad \mathrm{sign}(t) := \begin{cases} -1 \text{ for } t < 0 \\ 0 \quad \text{for } t = 0 \\ +1 \text{ for } t > 0\,. \end{cases}$$

This function is lower-semicontinuous and Clarke epi-super-regular almost everywhere, but not locally Lipschitz at x whenever $\|x\|_0 < n$; *a fortiori*, f it is actually discontinuous at all such points. Indeed, the epigraph of f is *locally convex* almost everywhere and, in particular, at any point (x, α) with $\alpha > f(x)$. At the point $(x, f(x))$ however, the epigraph is not even Clarke regular when $\|x\|_0 < n$. Nevertheless, it is ε-subregular, for the limiting subgradient 0 with $\varepsilon = 1$. Conversely, if x is any point with $\|x\|_0 = n$, then the counting function is locally constant and so in fact locally convex. These observations agree nicely with those in [9], namely, that the rank function (a generalizaton of the counting function) is subdifferentially regular everywhere (*i.e.,* all the various subdifferentials coincide) with $0 \in \partial \|x\|_0$ for all $x \in \mathbb{R}^n$.

In order to state the following corollary, recall that an extended real-valued function f is strongly amenable at $\bar{x}$ is if $f(\bar{x})$ is finite and there exists an open neighborhood U of $\bar{x}$ on which f has a representation as a composite $g \circ F$ with F of class $\mathcal{C}^2$ and g a proper, lsc, convex function on $\mathbb{R}^n$.

Proposition 6.5 *Let $f : \mathbb{R}^n \to (-\infty, +\infty]$ and consider the following assertions:*

(i) *f is strongly amenable at $\bar{x}$.*
(ii) *f is prox-regular at $\bar{x}$.*
(iii) epi f *is Clarke super-regular at $(\bar{x}, f(\bar{x}))$.*

Then: (i) $\Longrightarrow$ (ii) $\Longrightarrow$ (iii).

Proof The fact that strong amenability implies prox-regularity is discussed in [13, Proposition 2.5]. To see that (ii) implies (iii), suppose f is prox-regular at $\bar{x}$. Then epi f is prox-regular at $(\bar{x}, f(\bar{x}))$ by [13, Theorem 3.5] and hence Clarke super-regular at $(\bar{x}, f(\bar{x}))$ by Theorem 6.1. □

To conclude, we establish a *primal* characterization of epi-subregularity analogous to the characterization of Clarke epi-super-regularity in Theorem 6.2. It is worth noting that, unlike the results in Section 6.4.1, this characterization includes the possibility of horizon normals. In what follows, we denote the epigraph of a function f restricted to a subset $\Lambda \subset \mathrm{dom}\, f$ by $\mathrm{epi}(f_\Lambda) := \{(x, \alpha) \in \mathrm{epi}\, f \mid x \in \Lambda\}$.

Proposition 6.6 *Consider a function $f : \mathbb{R}^n \to (-\infty, +\infty]$, let $\overline{x} \in \mathrm{dom}\, f$ and let $(\overline{x}, \overline{v}) \in (\mathrm{gph}\, \partial^{\mathrm{C}} f \cup \mathrm{gph}\, \partial^{\mathrm{C}}{}_{\infty} f)$. Then the following assertions hold:*

(i) *f has an ε-subregular epigraph at $\bar{x} \in \mathrm{dom}\, f$ relative to $\Lambda \subseteq \mathrm{dom}\, f$ for $(\overline{x}, \overline{v})$ if and only if for some constant $\varepsilon > 0$ there is a neighborhood U of $(\overline{x}, f(\overline{x}))$*

such that, for all $(x, \alpha) \in \operatorname{epi}(f_\Lambda) \cap U$*, one of the following two inequalities holds:*

$$f(\overline{x}) + \langle \overline{v}, x - \overline{x} \rangle \leq \alpha + \varepsilon \|\overline{v}\| \|x - \overline{x}\| \left(\left(1 + \|\overline{v}\|^{-2} \right) \left(1 + |\alpha - f(\overline{x})|^2 \|x - \overline{x}\|^{-2} \right) \right)^{\frac{1}{2}} \tag{6.16a}$$

$$\langle \overline{v}, x - \overline{x} \rangle \leq \varepsilon \|\overline{v}\| \|x - \overline{x}\| \left(1 + |\alpha - f(\overline{x})|^2 \|x - \overline{x}\|^{-2} \right)^{\frac{1}{2}}. \tag{6.16b}$$

(ii) *f is epi-subregular at $\bar{x} \in \operatorname{dom} f$ for $(\overline{x}, \overline{v})$ relative to $\Lambda \subseteq \operatorname{dom} f$ if and only if for all $\varepsilon > 0$ there is a neighborhood (depending on ε) of $(\overline{x}, f(\overline{x}))$ such that, for all $(x, \alpha) \in \operatorname{epi}(f_\Lambda) \cap U$, either (6.16a) or (6.16b) holds.*

Proof

(i): First observe that since

$$N_{\operatorname{epi} f}(\overline{x}) \supseteq \{(v, -1) \mid v \in \partial f(\overline{x})\} \cup \{(v, 0) \mid v \in \partial_\infty f(\overline{x})\}, \text{ and}$$
$$N^{\mathrm{C}}_{\operatorname{epi} f}(\overline{x}) \supseteq \left\{(v, -1) \mid v \in \partial^{\mathrm{C}} f(\overline{x})\right\} \cup \left\{(v, 0) \mid v \in \partial^{\mathrm{C}}_\infty f(\overline{x})\right\},$$

any point $(\overline{x}, \overline{v}) \in (\operatorname{gph} \partial f \cup \operatorname{gph} \partial_\infty f)$ corresponds to either a normal vector of the form $(\overline{v}, -1)$ or a horizon normal of the form $(\overline{v}, 0)$. Suppose first that f is ε-epi-subregular at $\overline{x}$ relative to $\Lambda \subset \operatorname{dom} f$ for $\overline{v} \in \partial^{\mathrm{C}} f(\overline{x})$ with constant ε and neighborhood U' of $\overline{x}$. Then epi f is ε-subregular at $(\overline{x}, f(\overline{x}))$ relative to $\operatorname{epi}(f_\Lambda)$ for $(\overline{v}, -1) \in N^{\mathrm{C}}_{\operatorname{epi} f}(\overline{x}, f(\overline{x}))$ with constant ε and neighborhood U of $(\overline{x}, f(\overline{x}))$ in (6.3). Thus, for all $(x, \alpha) \in \operatorname{epi}(f_\Lambda) \cap U$, we have

$$\begin{aligned} &\langle (\overline{v}, -1), (x, \alpha) - (\overline{x}, f(\overline{x})) \rangle \leq \varepsilon \|(\overline{v}, -1)\| \|(x, \alpha) - (\overline{x}, f(\overline{x}))\| \\ \iff &\langle \overline{v}, x - \overline{x} \rangle - \alpha + f(\overline{x}) \leq \varepsilon \left(\|\overline{v}\|^2 + 1 \right)^{\frac{1}{2}} \left(\|x - \overline{x}\|^2 + (\alpha - f(\overline{x}))^2 \right)^{\frac{1}{2}} \\ &= \varepsilon \|\overline{v}\| \|x - \overline{x}\| \left(1 + \|\overline{v}\|^{-2} \right)^{\frac{1}{2}} \\ &\left(1 + (\alpha - f(\overline{x}))^2 \|x - \overline{x}\|^{-2} \right)^{\frac{1}{2}} \end{aligned}$$

which from the claim follows.

The only other case to consider is that f is ε-epi-subregular at $\overline{x}$ relative to $\Lambda \subset \operatorname{dom} f$ for $\overline{v} \in \partial^{\mathrm{C}}_\infty f(\overline{x})$ with constant ε and neighborhood U' of $\overline{x}$. In this case, epi f is ε-subregular at $(\overline{x}, f(\overline{x}))$ relative to $\operatorname{epi}(f_\Lambda)$ for $(\overline{v}, 0) \in N^{\mathrm{C}}_{\operatorname{epi} f}(\overline{x}, f(\overline{x}))$ with constant ε and neighborhood U of $(\overline{x}, f(\overline{x}))$ in (6.3). Thus, for all $(x, \alpha) \in \operatorname{epi}(f_\Lambda) \cap U$, we have

$$\begin{aligned} & \langle(\overline{v},0),(x,\alpha)-(\overline{x},f(\overline{x}))\rangle \le \varepsilon\|(\overline{v},0)\|\|(x,\alpha)-(\overline{x},f(\overline{x}))\| \\ \iff \quad & \langle\overline{v},x-\overline{x}\rangle \le \varepsilon\|\overline{v}\|\left(\|x-\overline{x}\|^2+(\alpha-f(\overline{x}))^2\right)^{1/2} \\ \iff \quad & \langle\overline{v},x-\overline{x}\rangle \le \varepsilon\|\overline{v}\|\|x-\overline{x}\|\left(1+(\alpha-f(\overline{x}))^2\|x-\overline{x}\|^{-2}\right)^{1/2}, \end{aligned}$$

which completes the proof of (i).

(ii): Follows immediately from the definition.

□

Remark 6.3 (Indicator Functions of Subregular Sets) When $f = \iota_\Omega$ for a closed set Ω the various subdifferentials coincide with the respective normal cones to Ω. In this case, inequality (6.16b) subsumes (6.16a) since all subgradients are also horizon subgradients and (6.16b) reduces to (6.3) in agreement with the corresponding notions of regularity of sets.

Acknowledgements The research of AD has been supported by the grants AFB170001 (CMM) & FONDECYT 1171854 (Chile) and MTM2014-59179-C2-1-P (MINECO of Spain). The research of DRL was supported in part by DFG Grant SFB755 and DFG Grant GRK2088. The research of MKT was supported in part by a post-doctoral fellowship from the Alexander von Humboldt Foundation.

References

1. Aussel, D., Daniilidis, A., Thibault, L.: Subsmooth sets: functional characterizations and related concepts. Trans. Amer. Math. Soc. **357**, 1275–1301 (2004)
2. Bauschke, H.H., Bolte, J., Teboulle, M.: A descent lemma beyond Lipschitz gradient continuity: first order methods revisited and applications. Math. Oper. Res. **42**(2), 330–348 (2016)
3. Bauschke, H.H., Luke, D.R., Phan, H.M., Wang, X.: Restricted Normal Cones and the Method of Alternating Projections: Theory. Set-Valued Var. Anal. **21**, 431–473 (2013). DOI 10.1007/s11228-013-0239-2. URL http://dx.doi.org/10.1007/s11228-013-0239-2
4. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. **28**, 2131–2151 (2018)
5. Daniilidis, A., Georgiev, P.: Approximate convexity and submonotonicity. J. Math. Anal. Appl. **291**, 292–301 (2004)
6. Dao, M.N., Phan, H.M.: Linear convergence of projection algorithms. Math. Oper. Res. **44**(2), 377–766 (2018). DOI 10.1287/moor.2018.0942. URL https://doi.org/10.1287/moor.2018.0942
7. Hesse, R., Luke, D.R.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. SIAM J. Optim. **23**(4), 2397–2419 (2013)
8. Kruger, A.Y., Luke, D.R., Thao, N.H.: Set regularities and feasibility problems. Math. Program. **168**, 279–311 (2018). DOI 10.1007/s10107-016-1039-x.
9. Le, H.Y.: Generalized subdifferentials of the rank function. Optimization Letters pp. 1–13 (2012). DOI 10.1007/s11590-012-0456-x. URL http://dx.doi.org/10.1007/s11590-012-0456-x
10. Lewis, A.S., Luke, D.R., Malick, J.: Local linear convergence of alternating and averaged projections. Found. Comput. Math. **9**(4), 485–513 (2009)

11. Luke, D.R., Thao, N.H., Tam, M.K.: Quantitative convergence analysis of iterated expansive, set-valued mappings. Math. Oper. Res. **43**(4), 1143–1176 (2018). DOI 10.1287/moor.2017.0898. URL https://doi.org/10.1287/moor.2017.0898
12. Ngai, H.V., Luc, D.T., Théra, M.: Approximate convex functions. J. Nonlinear Convex Anal. **1**, 155–176 (2000)
13. Poliquin, R.A., Rockafellar, R.T.: Prox-regular functions in variational analysis. Trans. Amer. Math. Soc. **348**, 1805–1838 (1996)
14. Rockafellar, R.T., Wets, R.J.: Variational Analysis. Grundlehren Math. Wiss. Springer-Verlag, Berlin (1998)
15. Thao, N.H.: Algorithms for structured nonconvex optimization: Theory and practice. Ph.D. thesis, Georg-August Universität Göttingen, Göttingen (2017)
16. Thibault, L.: Subsmooth functions and sets. Linear and Nonlinear Analysis (to appear)

Chapter 7
The Inverse Function Theorems of L. M. Graves

Asen L. Dontchev

Abstract The classical inverse/implicit function theorem revolves around solving an equation involving a differentiable function in terms of a parameter and tells us when the solution mapping associated with this equation is a differentiable function. Already in 1927 Hildebrand and Graves observed that one can put aside differentiability using instead Lipschitz continuity. Subsequently, Graves developed various extensions of this idea, most known of which are the Lyusternik-Graves theorem, where the inverse of a function is a set-valued mapping with certain Lipschitz type properties, and the Bartle-Graves theorem which establishes the existence of a continuous and calm selection of the inverse. In the last several decades more sophisticated results have been obtained by employing various concepts of regularity of mappings acting in metric spaces, mainly aiming at applications to numerical analysis and optimization. This paper presents a unified view to the inverse function theorems that originate from the works of Graves. It has a historical flavor, but not entirely, tracing the development of ideas from a personal perspective rather than surveying the literature.

Keywords Inverse function theorem · Nonsmooth analysis · Set-valued mapping · Metric regularity · Calmness · Continuous selection

AMS 2010 Subject Classification 47J07, 49J53, 49K40, 90C31

7.1 Introduction

Lawrence Murry Graves (1896–1973) earned his Ph.D. in calculus of variations in 1924 from the University of Chicago; his advisor was G. A. Bliss. Thanks to Mathematics Genealogy Project, one can trace the mathematical roots of Graves:

A. L. Dontchev (✉)
University of Michigan, Ann Arbor, MI, USA
e-mail: dontchev@umich.edu

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_7

Bliss was a student of O. Bolza, who was a student of C. Felix Klein, who had two advisers: J. Plüker and R. O. S. Lipschitz. Lipschitz also had two advisers, G. Dirichlet and M. Ohm, and Dirichlet's advisers were S. Poisson and J.-B. Fourier. Poisson also had two advisers, J. L. Lagrange and Laplace, while Fourier's adviser was Lagrange, and the advisor of Lagrange was L. Euler. So everything goes right to the origins of the calculus of variations.

After staying at Harvard for two years on a National Research Fellowship, Graves returned to the University of Chicago in 1926 and spent the rest of his career there, retiring in 1961. From 1930 to 1952 he supervised at least 18 PhD students, among them L. Alaoglu, E. McShane, H. Goldstine and Robert G. Bartle. Robert (Bob) Bartle recruited me, the author of the present paper, to Mathematical Reviews in 1990, where he was Executive Editor at that time. Around 1992 I read the paper [11] by Graves and discovered that what is written about it in [6] is not correct; this is explained in detail in the paper [8]. Around that time and later I had the opportunity to spend a number of hours with Bob Bartle, talking also about Graves and his mathematics. According to Bob Bartle, Professor Graves, as he always called him, was a private man, who did not like meetings and preferred to communicate with his students by mail (not by e-mail!). He was very rigorous and even meticulous in his mathematical writings, always looking for the precise statement under minimal assumption, and this could be seen in his papers.

H. Hildebrand (who apparently met Graves in Harvard but spent most of his professional life at the University of Michigan) and L. M. Graves are most likely the authors of the first nonsmooth implicit function theorem, published in [12], about 50 years before the beginning of the era of nonsmooth analysis. Section 2 of this paper is devoted to the Hildebrand–Graves theorem and extensions, reaching out to recent developments in that direction.

Section 7.3 is devoted to the most known nowadays theorem of Graves, which is now commonly called the Lyusternik–Graves theorem but other names are used as well. It plays a major role in modern variational analysis and optimization, and not only. According to the author's knowledge, it is the first inverse function theorem where the inverse is not a function but a set-valued mapping having a certain Lipschitz-type property. A number of papers and several monographs have been written about this theorem and around, the most recent being [13], which I enthusiastically recommend. In Section 7.3 a nonsmooth Lyusternik–Graves theorem is presented, which is proved in the recent paper [4].

In the late 90s there was a Mathematical Reviews Executive Committee meeting and Jon Borwein came to Ann Arbor as a member of that committee. We met with Jon already in the early 80s when we tried to solve an open problem in approximation theory (which was solved about 20 years later by L. Qi, H. Qi and me). In Ann Arbor we with Jon had had extensive conversations about mathematics but not only, and I mentioned to him one of the theorems (Theorem 4) in the paper [1] by Bartle and Graves and a generalization of it by Páles [16]. Jon jumped on that and, as a result, we published the paper [2]. I gave the paper to Bob who was

already diagnosed with cancer[1]; shortly after he sent us a letter, in which he wrote the following[2]:

> Your results are, indeed, an impressive and far-reaching extension of the theorem that Professor Graves and I published over a half-century ago. I was a student in a class of Graves in which he presented the theorem in the case that the parameter domain is the interval [0, 1]. He expressed the hope that it could be generalized to a more general domain, but said that he didn't see how to do so. By a stroke of luck, I had attended a seminar a few months before given by André Weil, which he titled "On a theorem by Stone." I (mis)understood that he was referring to M. H. Stone, rather than A. H. Stone, and attended. Fortunately, I listened carefully enough to learn about paracompactness and continuous partition of unity[3] (which were totally new to me) and which I found to be useful in extending Graves' proof. So the original theorem was entirely due to Graves; I only provided an extension of his proof, using methods that were not known to him. However, despite the fact that I am merely a 'middleman,' I am pleased that this result has been found to be useful.

In the Bartle-Graves paper [1] there are several theorems, and together with Jon we generalized one of them, Theorem 4. In the late 1990s Hector Sussmann wrote to me about a problem he bumped into when doing his constructions in geometric control. I got lucky to connect his problem to another theorem (Theorem 6) in the paper by Bartle and Graves [1]. This theorem is about a continuous and calm selection of the inverse of a smooth function whose derivative at the reference point is surjective. On the way to solving Sussmann's problem, I generalized in [9] that theorem for set-valued mapping. Section 7.4 of this paper discusses that Bartle-Graves theorem (Theorem 6 in [1]) and its generalization, and states a conjecture concerning a possible extension of that theorem to nonsmooth mappings.

There is another basic theorem in optimization which is not connected with the name of Graves but perhaps should be: this is the Karush-Kuhn-Tucker theorem. As is well known now, Kuhn and Tucker published this theorem in 1948 but in the 60s it was discovered that the result is contained in the master thesis of W. Karush from 1939 written under the supervision of Graves. Specifically, in his vita at the end of his master thesis of 25 pages, published in [15], Karush wrote: "...particular thanks are due to Professor Graves for his guidance as a teacher and in the writing of this dissertation." The work of Karush, guided by Graves, had remained unknown for quite a while, but the computers were still in the future and the finite-dimensional optimization was not regarded as important as it is now.

This paper presents a unified view to the inverse/implicit function theorems that originated from the works of Graves. To keep things simple, the focus is on inverse function theorems; their implicit function versions are not discussed. The terminology and notations are from the book [10], where the reader can find a broader coverage of some of the results given here. This paper also presents some more recent developments and poses open problems.

[1] Bob passed away September 18, 2002.

[2] This letter is also published in [9] and [10].

[3] Michael's theorem was not known at that time.

7.2 Hildebrand–Graves Theorem

Recall that the Lipschitz modulus of a function f acting between Banach spaces X and Y and having $\bar{x}$ in the interior of its domain, is defined as follows:

$$\operatorname{lip}(f;\bar{x}) := \limsup_{\substack{x',x\to\bar{x}\\ x\neq x'}} \frac{\|f(x')-f(x)\|}{\|x'-x\|}.$$

Also recall that, given a set-valued mapping $F : X \rightrightarrows Y$ and a point $(\bar{x},\bar{y})$ in its graph $\operatorname{gph} F$, the mapping F is said to have a single-valued graphical localization around $\bar{x}$ for $\bar{y}$ whenever there exist neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that the truncated mapping

$$U \ni x \mapsto F(x) \cap V$$

is single-valued, a function. If this function is Lipschitz continuous on U, we say that the mapping F has a Lipschitz localization around $\bar{x}$ for $\bar{y}$. The following definition goes back to Robinson [18]: A mapping $F : X \rightrightarrows Y$ with $(\bar{x},\bar{y}) \in \operatorname{gph} F$ is said to be strongly regular at $\bar{x}$ for $\bar{y}$ whenever F^{-1} has a Lipschitz localization around $\bar{y}$ for $\bar{x}$. For a function f we simply say that f is strongly regular at $\bar{x}$ (dropping for $f(\bar{x})$).

We start with the following slightly extended inverse function version of the Hildebrand–Graves theorem [12, Theorem 4]. Since this result is central in this paper, and also for pedagogical purposes, we supply it with a short proof which is close to the original proof.[4]

Theorem 7.1 *Let X be a Banach space and consider a function $f : X \to X$ and a linear bounded mapping $A : X \to X$ which is invertible. Suppose that*

$$\operatorname{lip}(f-A;\bar{x})\cdot\|A^{-1}\| < 1. \tag{7.1}$$

Then f is strongly regular at $\bar{x}$.

Proof Without loss of generality, let $f(\bar{x}) = 0$. Choose a positive κ such that

$$\operatorname{lip}(f-A;\bar{x}) < \kappa < 1/\|A^{-1}\|$$

and let $a > 0$ be such that the function $f - A$ is Lipschitz continuous on $\mathbb{B}_a(\bar{x})$ with Lipschitz constant κ. Let $y \in \mathbb{B}_{\kappa a}(0)$ and consider the function

[4]Interestingly enough, Hildebrand and Graves cite the 1922 paper by Banach published in Fundamenta Mathematicae, where Banach presented his contraction mapping theorem, but they prove it independently in their Theorem 1. Apparently, the contracting mapping iteration was known to Picard and Goursat long before Banach.

$$\mathbb{B}_a(\bar{x}) \ni x \mapsto \Phi_y(x) := -A^{-1}[(f - A)(x) - y].$$

We have

$$\|\bar{x} - \Phi_y(\bar{x})\| = \|\bar{x} + A^{-1}(-A\bar{x} - y)\| = \|A^{-1}y\| \le \kappa a/\kappa = a.$$

Further, for any $x, x' \in \mathbb{B}_a(\bar{x})$ we obtain

$$\|\Phi_y(x) - \Phi_y(x')\| \le \|A^{-1}\|\|f(x) - f(x') - A(x - x')\| \le \|A^{-1}\|\kappa\|x - x'\|;$$

that is, Φ_y is Lipschitz continuous on $\mathbb{B}_a(\bar{x})$ with Lipschitz constant $\|A^{-1}\|\kappa < 1$. By the contraction mapping theorem, see, e.g., [10, Theorem 1A.2], we obtain that for each $y \in \mathbb{B}_{\kappa a}(0)$ there exists only one $x(y) \in f^{-1}(y) \cap \mathbb{B}_a(\bar{x})$, that is, f^{-1} has a single-valued graphical localization around $\bar{x}$. Let $y, y' \in \mathbb{B}_{\kappa a}(0)$. Then

$$\begin{aligned}\|x(y) - x(y')\| &= \|\Phi_y(x(y)) - \Phi_{y'}(x(y'))\| \\ &\le \|\Phi_y(x(y)) - \Phi_y(x(y'))\| + \|\Phi_y(x(y')) - \Phi_{y'}(x(y'))\| \\ &\le \|A^{-1}\|\kappa\|x(y) - x(y')\| + \|A^{-1}\|\|y - y'\|.\end{aligned}$$

Hence, the graphical localization $y \mapsto x(y)$ is Lipschitz continuous on $\mathbb{B}_a(\bar{x})$ with Lipschitz constant $\|A^{-1}\|/(1 - \kappa\|A^{-1}\|)$. □

Recall that f is strictly differentiable at $\bar{x}$ with derivative $Df(\bar{x})$ if and only if $\operatorname{lip}(f - Df(\bar{x}); \bar{x}) = 0$. Thus, the Hildebrand–Graves theorem implies the basic (Dini) inverse function theorem.

The main novelty in the Hildebrand–Graves theorem is that *differentiability is replaced by Lipschitz continuity*. This is not spelled out clearly in their paper [12] but can be gleaned from their proof.

To the best of the author's knowledge, the next nonsmooth implicit function theorem came almost 50 years after the theorem of Hildebrand and Graves and is due to F. H. Clarke [5]. Here we adopt the following inverse function version of it:

Theorem 7.2 *Consider a function $f : \mathbb{R}^n \to \mathbb{R}^n$ which is Lipschitz continuous around $\bar{x}$ and suppose that all matrices in the Clarke generalized Jacobian $\partial f(\bar{x})$ are nonsingular. Then f is strongly regular at $\bar{x}$.*

Note that the Hildebrand–Graves theorem is stated in Banach spaces, while the Clarke theorem is finite dimensional.[5] These two theorems are of different nature, and do not follow from each other. If it were true, however, that

$$\inf_{A \in \partial f(\bar{x})} \operatorname{lip}(f - A; \bar{x}) = 0, \tag{7.2}$$

[5]There are some partial extensions to infinite dimensions but we shall not go into that here.

then the assumption (7.1) would hold automatically and Hildebrand–Graves theorem would imply Clarke's theorem. The relation (7.2) is actually not true; for a simple counterexample,[6] take $X = \mathbb{R}$, $f(x) = |x|$ whose generalized Jacobian is the interval $[-1, 1]$, and $\bar{x} = 0$; then the infimum on the left of (7.2) is 1.

Clarke's implicit function theorem has found important applications in optimization and beyond, most notably for solving nonsmooth equations, e.g., by the semismooth Newton methods. This came after another seminal contribution to variational analysis—the implicit function theorem of S. M. Robinson published in [18]. We will state this theorem in the following inverse function form:

Theorem 7.3 *Let X be a Banach space and consider a function $f : X \to X$ which is strictly differentiable at $\bar{x}$ and any set-valued mapping $F : X \rightrightarrows X$. Let $\bar{y} \in f(\bar{x}) + F(\bar{x})$. Then $f + F$ is strongly regular at $\bar{x}$ for $\bar{y}$ if and only if the mapping*

$$x \mapsto f(\bar{x}) + Df(\bar{x})(x - \bar{x}) + F(x) \tag{7.3}$$

has the same property.

Robinson stated his theorem for variational inequalities, i.e., when the map F is the normal cone mapping. To the best of my knowledge, it was first noted in [7] that the theorem remains valid for *any* set-valued mapping F. Moreover, in the same paper [7] it was shown that that if the graphical localization of the inverse of the linearized mapping (7.3) is not only Lipschitz continuous but also differentiable (in the sense of Fréchet, Gateaux, Bouligand, or directionally) at the reference point $\bar{y}$, then the graphical localization of $(f + F)^{-1}$ has the same property; for an extended version of this result, see [10, Theorem 2B.9].

In the paper [14] A. F. Izmailov made a significant step ahead by merging the Robinson and Clarke inverse function theorems. Specifically, he obtained[7] the following result:

Theorem 7.4 *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be Lipschitz continuous around $\bar{x}$, let $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, and let $\bar{y} \in f(\bar{x}) + F(\bar{x})$. Suppose that for every $A \in \partial f(\bar{x})$ the mapping $f(\bar{x}) + A(\cdot - \bar{x}) + F(\cdot)$ is strongly regular at $\bar{x}$ for $\bar{y}$. Then $f + F$ has the same property.*

If f is strictly differentiable at $\bar{x}$, Theorem 7.4 implies Robinson's theorem in finite dimensions; if F is the zero mapping, it implies Clarke's theorem.

In [3], we generalized Izmailov's theorem to Banach spaces, as follows:

[6] This counterexample was communicated to the author by Radek Cibulka.

[7] The original proof in [14] has a gap which was later fixed by Radek Cibulka and me in [3]. Subsequently, Izmailov sent us a nice letter saying that, yes, a student of his found the gap, and yes, it is now fixed.

Theorem 7.5 *Let $(\bar{x}, \bar{y}) \in X \times Y$, let $f : X \to Y$ and $F : X \rightrightarrows Y$ be such that $\bar{y} \in f(\bar{x}) + F(\bar{x})$. Suppose that there exist a convex subset $\mathcal{A}$ of $\mathcal{L}(X, Y)$ and a constant $c > 0$ such that*

(i) *there exists $r > 0$ such that for each u and v in $\mathbb{B}_r(\bar{x})$ one can find $A \in \mathcal{A}$ such that*

$$\|f(v) - f(u) - A(v - u)\| \leq c\|v - u\|;$$

(ii) *for every $A \in \mathcal{A}$ the mapping*

$$G_A : x \mapsto f(\bar{x}) + A(x - \bar{x}) + F(x) \tag{7.4}$$

is strongly regular at $\bar{x}$ for $\bar{y}$; moreover, if s_A is any single-valued graphical localization of G_A^{-1} around $\bar{y}$ for $\bar{x}$, then

$$(c + \chi(\mathcal{A})) \sup_{A \in \mathcal{A}} \operatorname{lip}(s_A; \bar{y}) < 1,$$

where $\chi(\mathcal{A})$ is the measure of non-compactness of the set $\mathcal{A}$. Then the mapping $f + F$ is strongly regular at $\bar{x}$ for $\bar{y}$.

When $X = Y = \mathbb{R}^n$, Theorem 7.5 reduces to Theorem 7.4.

7.3 The Lyusternik-Graves Theorem

Recall that, for a positively homogeneous mapping H acting between Banach spaces X and Y, the inner norm is defined as

$$\|H\|^- = \sup_{\|x\| \leq 1} \inf_{y \in H(x)} \|y\|. \tag{8}$$

When H is linear and bounded singe-valued mapping, this gives us the operator norm. The following theorem is a restatement of the Banach open mapping principle:

Theorem 7.6 *Let A be a linear and bounded mapping acting between Banach spaces X and Y. Then the following are equivalent:*

(a) *A is surjective;*
(b) *A is open (at every point);*
(c) $\|A^{-1}\|^- < \infty$.

In 1950 L. M. Graves published a theorem[8] in [11], which we present next in a form similar to the Hildebrand–Graves Theorem 7.1:

Theorem 7.7 *Let X, Y be Banach spaces and consider a function $f : X \to Y$ and a point $\bar{x} \in \operatorname{int} \operatorname{dom} f$ along with a bounded linear mapping $A : X \to Y$ which is surjective and such that*

$$\operatorname{lip}(f - A; \bar{x}) \cdot \|A^{-1}\|^{-} < 1. \tag{7.5}$$

Then there exist a positive constant κ and neighborhoods U of $\bar{x}$ and V of $f(\bar{x})$ such that

$$\sup_{x \in f^{-1}(y) \cap U} d\big(x, f^{-1}(y')\big) \le \kappa \|y - y'\| \quad \text{for every} \quad y, y' \in V. \tag{7.6}$$

If A is not just surjective but also invertible, then condition (7.5) becomes the same as (7.1). If f^{-1} has a single-valued graphical localization around $f(\bar{x})$, then condition (7.6) implies that this localization is Lipschitz continuous around $f(\bar{x})$; that is, f is strongly regular at $\bar{x}$. But it is possible that any graphical localization of the inverse f^{-1} is multi-valued. Still, the inverse has a Lipschitz-type property, which is now called the Aubin property, and the theorem itself can be regarded as an inverse mapping theorem. It turned out that the Aubin property of the inverse is equivalent to a property of the mapping itself, which was called *metric regularity* by Jon Borwein in 1986, and the name remained, even after so much recent re-naming in the area.

Let X and Y be metric spaces. A mapping $F : X \rightrightarrows Y$ is said to be metrically regular at $\bar{x}$ for $\bar{y}$ when $\bar{y} \in F(\bar{x})$, $\operatorname{gph} F$ is locally closed at $(\bar{x}, \bar{y})$ and there is a constant $\tau \ge 0$ together with neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that

$$d\big(x, F^{-1}(y)\big) \le \tau d\big(y, F(x)\big) \quad \text{for every} \quad (x, y) \in U \times V.$$

The infimum of all constants $\tau \ge 0$ for which this inequality holds is the regularity modulus of F at $\bar{x}$ for $\bar{y}$ denoted by $\operatorname{reg}(F; \bar{x} \,|\, \bar{y})$.

The following theorem is extracted from [10, Chapter 5] and combines extended versions of both Robinson's theorem and the Lyusternik-Graves theorem in metric spaces.

Theorem 7.8 *Let X be a complete metric space, Y be a linear metric space with shift-invariant metric. Suppose that*

1) κ and μ are positive constants with $\kappa\mu < 1$.
2) $F : X \rightrightarrows Y$ is [strongly] metrically regular at $\bar{x}$ for $\bar{y}$ with $\operatorname{reg}(F; \bar{x} \,|\, \bar{y}) \le \kappa$.
3) $g : X \to Y$ and $\operatorname{lip}(g; \bar{x}) \le \mu$.

[8] A predecessor of that theorem was given by Lyusternik, for a statement and a comparison, see [10, Section 5D].

Then $g + F$ is [strongly] metrically regular at $\bar{x}$ for $\bar{y} + g(\bar{x})$ with

$$\operatorname{reg}(g + F; \bar{x} \,|\, g(\bar{x}) + \bar{y}) \leq (\kappa^{-1} - \mu)^{-1}.$$

In 1977 B. H. Pourciau [17] proved that a function $f : \mathbb{R}^n \to \mathbb{R}^d$, with $d \leq n$, which is Lipschitz continuous around $\bar{x}$, is metrically regular at $\bar{x}$ if every element of $\partial f(\bar{x})$ has full rank. This result is very much in line with Clarke's Theorem 7.2 which gives a sufficient condition for Lipschitz invertibility. A generalization of the theorem of Pourciau in the spirit of Theorem 7.5 was obtained recently in [4], as follows:

Theorem 7.9 *Let X and Y be Banach spaces. Consider a function $f : X \to Y$, a set-valued mapping $F : X \rightrightarrows Y$, and a point $(\bar{x}, \bar{y}) \in \operatorname{gph}(f + F)$ with $\bar{x} \in \operatorname{int} \operatorname{dom} f$. Consider also a convex subset $\mathcal{A}$ of $\mathcal{L}(X, Y)$ and a constant $\mu \geq 0$, and assume that condition (i) in Theorem 7.5 is satisfied and, moreover, the following conditions hold:*

- *for every $A \in \mathcal{A}$ the mapping G_A defined in (7.4) is metrically regular at $\bar{x}$ for $\bar{y}$ and, in addition,*

$$\beta\big(c + \chi(\mathcal{A})\big) < 1,$$

  where $\beta := \sup_{A \in \mathcal{A}} \operatorname{reg}(G_A; \bar{x} \,|\, \bar{y})$ and $\chi(\mathcal{A})$ is the measure of non-compactness of the set $\mathcal{A}$;
- *there are neighborhoods U of $\bar{x}$ and V of $\bar{y}$ such that the set $G_A^{-1}(v) \cap U$ is convex whenever $v \in V$ and $A \in \mathcal{A}$.*

Then the mapping $f + F$ is metrically regular at $\bar{x}$ for $\bar{y}$; moreover,

$$\operatorname{reg}(f + F; \bar{x} \,|\, \bar{y}) \leq (\beta^{-1} - (c + \chi(\mathcal{A}))^{-1}.$$

The convexity requirement in the second assumption of the theorem comes from the Michael selection theorem which we use in the proof; it is an open question whether it could be relaxed.

7.4 Bartle–Graves Theorem

In this section X and Y are Banach spaces unless stated differently. The Theorem 7.7 published in by Graves in [11] implies that, for a function $f : X \to Y$ which is strictly differentiable at $\bar{x}$ and such that the strict derivative $Df(\bar{x})$ is surjective, then f^{-1} has the Aubin property at $f(\bar{x})$ for $\bar{x}$. Only two years after the publication of [11], Bartle and Graves published in [1] a theorem claiming that under the *same* assumptions, the inverse f^{-1} has a continuous local selection which is calm. The original statement of that Bartle–Graves theorem is as follows:

Theorem 7.10 *Let X and Y be Banach spaces and let $f : X \to Y$ be a function which is strictly differentiable at $\bar{x}$ and such that the derivative $Df(\bar{x})$ is surjective. Then there is a neighborhood V of $f(\bar{x})$ along with a constant $\gamma > 0$ such that f^{-1} has a continuous selection s on V with the property*

$$\|s(y) - \bar{x}\| \leq \gamma \|y - f(\bar{x})\| \quad \text{for every } y \in V.$$

It should be noted that if we restrict our attention to Hilbert spaces, then it is easy to obtain the existence of a local selection which is even differentiable. Specifically, we have the following (see [10, Exercise 5J.2]):

Theorem 7.11 *Let X and Y be Hilbert spaces and let $f : X \to Y$ be a function which is strictly differentiable at $\bar{x}$ and such that the derivative $A := Df(\bar{x})$ is surjective. Then the inverse f^{-1} has a local selection s around $\bar{y} := f(\bar{x})$ for $\bar{x}$ which is strictly differentiable at $\bar{y}$ with derivative $Ds(\bar{y}) = A^*(AA^*)^{-1}$, where A^* is the adjoint of A.*

The following generalization of Theorem 7.10 for set-valued mapping was published in [9][9]:

Theorem 7.12 *Consider a mapping $G : X \rightrightarrows Y$ and any $(\bar{x}, \bar{y}) \in \operatorname{gph} G$ and suppose that for some $c > 0$ the mapping $I\!B_c(\bar{y}) \ni y \mapsto G^{-1}(y) \cap I\!B_c(\bar{x})$ is closed-convex-valued. Consider also a function $g : X \to Y$ with $\bar{x} \in \operatorname{int} \operatorname{dom} g$.[10] Let κ and μ be nonnegative constants such that*

$$\kappa\mu < 1, \quad \operatorname{reg}(G; \bar{x} \,|\, \bar{y}) \leq \kappa \quad \text{and} \quad \operatorname{lip}(g; \bar{x}) \leq \mu.$$

Then for every $\gamma > \kappa/(1 - \kappa\mu)$ the mapping $(g + G)^{-1}$ has a continuous local selection s around $g(\bar{x}) + \bar{y}$ for $\bar{x}$ with the property

$$\|s(y) - \bar{x}\| \leq \gamma \|y - \bar{y}\| \quad \text{for every } y \in V.$$

Here again, the convexity requirement comes from Michael's selection theorem—see the open question at the end of the preceding section.

Is there a Bartle-Graves theorem of the kind of Theorem 7.10 for the case when the function g is merely Lipschitz continuous, as in the theorems of Clarke and Pourciau? Specifically, putting aside the set-valued part, is the following statement true?

Conjecture (A Nonsmooth Bartle-Graves Theorem) Consider a function $f : I\!R^n \to I\!R^m$ which is Lipschitz continuous around $\bar{x}$ and suppose that all matrices in

[9] A more general version of Theorem 7.12 will be presented in author's paper *Bartle-Graves theorem revisited*, submitted to Set-Valued and Variational Analysis, July 2019.

[10] As noted by the referee of this paper, here it is sufficient to assume that $\bar{x}$ is in the core of the domain of g.

the generalized Jacobian $\partial f(\bar{x})$ are surjective. Then f^{-1} has a continuous local selection around $f(\bar{x})$ for $\bar{x}$ which is calm at $\bar{y}$.

Acknowledgements The author wishes to thank Radek Cibulka for his help when preparing this paper. This work was supported by the National Science Foundation (NSF) Grant 156229, the Austrian Science Foundation (FWF) Grant P31400-N32, and the Australian Research Council (ARC) Project DP160100854.

References

1. Bartle, R.G., Graves, L.M.: Mappings between function spaces. Trans. Amer. Math. Soc. **72**, 400–413 (1952)
2. Borwein, J.M., Dontchev, A.L.: On the Bartle-Graves theorem. Proc. Amer. Math. Soc. **131**, 2553–2560 (2003)
3. Cibulka, R., Dontchev, A.L.: A nonsmooth Robinson's inverse function theorem in Banach spaces. Math. Program. A **156**, 257–270 (2016)
4. Cibulka, R., Dontchev, A.L., Veliov, V.M.: Lyusternik-Graves theorems for the sum of a Lipschitz function and a set-valued mapping. SIAM J. Control Optim. **54**, 3273–3296 (2016)
5. Clarke, F.H.: On the inverse function theorem. Pacific Journal of Mathematics **64**, 97–102 (1976)
6. Dmitruk, A.V., Milyutin, A.A., Osmolovskiĭ, N.P.: Ljusternik's theorem and the theory of the extremum. Uspekhi Mat. Nauk **35**, 11–46 (1980)
7. Dontchev, A.L.: Implicit function theorems for generalized equations. Math. Program. A **70**, 91–106 (1995)
8. Dontchev, A.L.: The Graves theorem revisited. J. Convex Anal. **3**, 45–53 (1996)
9. Dontchev, A.L.: A local selection theorem for metrically regular mappings. J. Convex Anal. **11**, 81–94 (2004)
10. Dontchev, A.L., Rockafellar, R.T.: Implicit Functions and Solution Mappings. A view from variational analysis., 2nd edn. Springer (2014)
11. Graves, L.M.: Some mapping theorems. Duke Math. Journal **17**, 111–114 (1950)
12. Hildebrand, H., Graves, L.M.: Implicit functions and their differentials in general analysis. Trans. AMS **29**, 127–153 (1927)
13. Ioffe, A.: Variational Analysis of Regular Mappings. Theory and Applications. Springer (2017)
14. Izmailov, A.: Strongly regular nonsmooth generalized equations. Math. Program. Ser. A **147**, 581–590 (2014)
15. Karush, W.: Minima of functions of several variables with inequalities as side conditions. Traces and emergence of nonlinear programming. Birkhäuser/Springer Basel AG, Basel (2014)
16. Páles, Z.: Inverse and implicit function theorems for nonsmooth maps in Banach spaces. J. Math. Anal. Appl. **209**, 202–220 (1997)
17. Pourciau, B.: Analysis and optimization of Lipschitz continuous mappings. J. Opt. Theory Appl. **22**, 311–351 (1977)
18. Robinson, S.M.: Strongly regular generalized equations. Math. of Oper. Res. **5**, 43–62 (1980)

Chapter 8
Block-Wise Alternating Direction Method of Multipliers with Gaussian Back Substitution for Multiple-Block Convex Programming

Xiaoling Fu, Bingsheng He, Xiangfeng Wang, and Xiaoming Yuan

Abstract We consider the linearly constrained convex minimization model with a separable objective function which is the sum of m functions without coupled variables, and discuss how to design an efficient algorithm based on the fundamental technique of splitting the augmented Lagrangian method (ALM). Our focus is the specific big-data scenario where m is huge. A pretreatment on the original data is to regroup the m functions in the objective and the corresponding m variables as t subgroups, where t is a handleable number (usually $t \geq 3$ but much smaller than m). To tackle the regrouped model with t blocks of functions and variables, some existing splitting methods in the literature are applicable. We concentrate on the application of the alternating direction method of multiplier with Gaussian back substitution (ADMM-GBS) whose efficiency and scalability have been well verified in the literature. The block-wise ADMM-GBS is thus resulted and named when the ADMM-GBS is applied to solve the t-block regrouped model. To alleviate the difficulty of the resulting ADMM-GBS subproblems, each of which may still require minimizing more than one function with coupled variables, we suggest further decomposing these subproblems but regularizing these further

X. Fu
School of Economics and Management, Southeast University, Nanjing, China

B. He
Department of Mathematics, South University of Science and Technology of China, Shenzhen, China

Department of Mathematics, Nanjing University, Nanjing, China
e-mail: hebma@nju.edu.cn

X. Wang
School of Computer Science and Technology, East China Normal University, Shanghai, China
e-mail: xfwang@sei.ecnu.edu.cn

X. Yuan (✉)
Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong
e-mail: xmyuan@hku.hk

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_8

decomposed subproblems with proximal terms to ensure the convergence. With this further decomposition, each of the resulting subproblems only requires handling one function in the original objective plus a simple quadratic term; it thus may be very easy for many concrete applications where the functions in the objective have some specific properties. Moreover, these further decomposed subproblems can be solved in parallel, making it possible to handle big-data by highly capable computing infrastructures. Consequently, a splitting version of the block-wise ADMM-GBS is proposed for the particular big-data scenario. The implementation of this new algorithm is suitable for a centralized-distributed computing system, where the decomposed subproblems of each block can be computed in parallel by a distributed-computing infrastructure and the blocks are updated by a centralized-computing station. For the new algorithm, we prove its convergence and establish its worst-case convergence rate measured by the iteration complexity. Two refined versions of this new algorithm with iteratively calculated step sizes and linearized subproblems are also proposed, respectively.

Keywords Convex programming · Alternating direction method of multipliers · Operator splitting · Convergence rate

AMS 2010 Subject Classification 49M20, 65K10, 90C30

8.1 Introduction

We consider a separable convex minimization problem with linear constraints and its objective function is the sum of more than one function without coupled variables:

$$\min\left\{\sum_{i=1}^{m}\theta_i(x_i)\;\Big|\;\sum_{i=1}^{m}A_ix_i=b,\;x_i\in X_i,i=1,\ldots,m\right\},\tag{8.1.1}$$

where $\theta_i:\mathbb{R}^{n_i}\to\mathbb{R}$ $(i=1,\ldots,m)$ are convex (not necessarily smooth) and continuous, n_i $(i=1,\ldots,m)$ are the dimensions of variables x_i and $\sum_{i=1}^{m}n_i=n$; $A_i\in\mathbb{R}^{\ell\times n_i}$, $b\in\mathbb{R}^{\ell}$, and $X_i\subseteq\mathbb{R}^{n_i}$ $(i=1,\cdots,m)$ are closed convex sets. The solution set of (8.1.1) is assumed to be nonempty throughout our discussions in this paper. We also assume that matrices $A_i^TA_i$ $(i=1,\ldots,m)$ are all nonsingular.

Our discussion is under the assumption that each function θ_i in the objective of (8.1.1) has some specific properties and it is worthwhile to take advantage of them in algorithmic design. One representative case, which has wide applications in many sparse- and/or low-rank-related fields, is when the following problem

$$\arg\min_{x_i\in\mathbb{R}^{n_i}}\left\{\theta_i(x_i)+\frac{\tau}{2}\|x_i-a_i\|^2\right\}\tag{8.1.2}$$

has a closed-form solution for any given vector $a_i \in \mathbb{R}^{n_i}$ and scalar $\tau > 0$. In (8.1.2), $\|\cdot\|$ denotes the standard ℓ_2 norm. Note that solving (8.1.2) is equivalent to estimating the proximal operator of θ_i. Thus, we do not discuss the case where the model (8.1.1) is treated as a whole and its separable structures are ignored in algorithmic design. Instead, we are interested in such an algorithm whose x_i-subproblems at each iteration are all of the difficulty of solving

$$\arg\min_{x_i \in \mathcal{X}_i} \left\{\theta_i(x_i) + \frac{\tau}{2}\|x_i - a_i\|^2\right\}, \quad a_i \in \mathbb{R}^{n_i}, \tag{8.1.3}$$

or, at most, of

$$\arg\min_{x_i \in X_i} \left\{\theta_i(x_i) + \frac{\tau}{2}\|A_i x_i - a_i\|^2\right\}, \quad a_i \in \mathbb{R}^{\ell}. \tag{8.1.4}$$

Note that when the problem (8.1.2) has a closed-form solution, solving (8.1.3) or (8.1.4) could be generally easy, especially for the case where $X_i = \mathbb{R}^{n_i}$. For instance, if $X_i = \mathbb{R}^{n_i}$, the problem (8.1.4) can be iteratively solved by linearizing the quadratic term in (8.1.4) because the linearized subproblem reduces to a problem in form of (8.1.2). This is indeed an implementation of the forward–backward splitting method which was originated in [27]. Therefore, to expose our main idea of algorithmic design with easier notation, we mainly focus on the discussion of designing an algorithm with subproblems in form of (8.1.4) and only briefly mention its advanced version with subproblems in form of (8.1.3).

The augmented Lagrangian method (ALM) in [24, 29] is the basis for a number of splitting methods in the literature for solving the model (8.1.1). Let the Lagrangian function of (8.1.1) be

$$L^m(x_1, x_2, \ldots, x_m, \lambda) = \sum_{i=1}^{m} \theta_i(x_i) - \lambda^T\Big(\sum_{i=1}^{m} A_i x_i - b\Big), \tag{8.1.5}$$

with $\lambda \in \mathbb{R}^{\ell}$ the Lagrange multiplier and it be defined on $\Omega = X_1 \times X_2 \times \cdots \times X_m \times \mathbb{R}^{\ell}$. The augmented Lagrangian function is

$$\mathcal{L}_\beta^m(x_1, \ldots, x_m, \lambda) = L^m(x_1, \ldots, x_m, \lambda) + \frac{\beta}{2}\Big\|\sum_{i=1}^{m} A_i x_i - b\Big\|^2, \tag{8.1.6}$$

where $L^m(x_1, x_2, \ldots, x_m, \lambda)$ is given by (8.1.5) and $\beta > 0$ is a penalty parameter with respect to the violation of the linear constraints in (8.1.1). If we treat the primal variables in model (8.1.1) as a whole and apply directly the ALM, then the resulting scheme is

$$\begin{cases} (x_1^{k+1}, x_2^{k+1}, \ldots, x_m^{k+1}) = \arg\min\left\{\mathcal{L}_\beta^m(x_1, x_2, \ldots, x_m, \lambda^k) \mid x_i \in X_i,\ i = 1, \ldots, m\right\}, \\ \lambda^{k+1} = \lambda^k - \beta(\sum_{i=1}^m A_i x_i^{k+1} - b). \end{cases} \tag{8.1.7}$$

The minimization subproblem in (8.1.7) is clearly not efficient under the mentioned assumption that each θ_i has specific properties. Thus, when considering the model (8.1.1), the scheme (8.1.7) is only of conceptual sense. But it is the basis of a number of efficient methods in the literature whose common feature is decomposing the minimization subproblem in (8.1.7) appropriately and then to ensure the convergence with some additional steps if necessary. The most successful case is decomposing the minimization subproblem in (8.1.7) in Gauss–Seidel order for the special case of (8.1.1) with $m = 2$:

$$\begin{cases} x_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta^2(x_1, x_2^k, \lambda^k) \mid x_1 \in X_1\right\}, \\ x_2^{k+1} = \arg\min\left\{\mathcal{L}_\beta^2(x_1^{k+1}, x_2, \lambda^k) \mid x_2 \in X_2\right\}, \\ \lambda^{k+1} = \lambda^k - \beta(A_1 x_1^{k+1} + A_2 x_2^{k+1} - b). \end{cases} \tag{8.1.8}$$

This is the so-called alternating direction method of multiplier (ADMM) in [11] and it has found many efficient applications in a broad spectrum of application domains such as image processing, statistical learning, computer vision, network optimization, and so on. We refer to [3, 8, 10] for some review papers on the ADMM.

With the efficiency of ADMM, it is natural to consider directly extending the scheme (8.1.8) to the case of (8.1.1) with $m > 2$. The resulting direct extension of ADMM reads as

$$\begin{cases} x_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1, x_2^k, \ldots, x_m^k, \lambda^k) \mid x_1 \in X_1\right\}, \\ \qquad\vdots \\ x_i^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \mid x_i \in X_i\right\}, \\ \qquad\vdots \\ x_m^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1^{k+1}, \ldots, x_{m-1}^{k+1}, x_m, \lambda^k) \mid x_m \in X_m\right\}, \\ \lambda^{k+1} = \lambda^k - \beta(\sum_{i=1}^m A_i x_i^{k+1} - b). \end{cases} \tag{8.1.9}$$

Empirically, the direct extension of ADMM scheme (8.1.9) indeed works well for some applications, as shown in, e.g., [28, 30]. However, it was shown in [6] that theoretically the scheme (8.1.9) is not necessarily convergent. Hence, like the extreme case of treating (8.1.1) as a whole and applying no splitting at all to the ALM (8.1.7), this scheme (8.1.9) resulted by applying a full splitting to the ALM (8.1.7) does not work either.

In the literature, some surrogates with provable convergence and numerical performance competitive to (8.1.9) have been well studied. For example, the schemes in [15, 16] treat the output of (8.1.9) as a predictor and suggest correcting

it appropriately to ensure the convergence. These schemes are all in the prediction-correction framework. The scheme in [17] requires no correction step, but it slightly changes the order of updating the Lagrange multiplier and twists some of the subproblems appropriately to obtain the convergence. Accordingly, the $(x_2, \cdots, x_m)$-subproblems can be solved in parallel but they should be regularized by appropriate proximal terms with sufficiently large proximal coefficients. Moreover, the scheme in [25] suggests attaching a shrinking factor to the Lagrange multiplier updating step in (8.1.9). In [6], it was shown that it could be very difficult to find such a factor to guarantee the convergence of the direct extension of the ADMM scheme (8.1.9). Let us recall the ADMM with a Gaussian back substitution (ADMM-GBS for short) proposed in [15] whose iterative scheme reads as

$$\begin{cases} \begin{cases} \bar{x}_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1, x_2^k, \ldots, x_m^k, \lambda^k) \;\middle|\; x_1 \in X_1\right\}, \\ \qquad\vdots \\ \bar{x}_i^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(\bar{x}_1^{k+1}, \ldots, \bar{x}_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \;\middle|\; x_i \in X_i\right\}, \\ \qquad\vdots \\ \bar{x}_m^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(\bar{x}_1^{k+1}, \ldots, \bar{x}_{m-1}^{k+1}, x_m, \lambda^k) \;\middle|\; x_m \in X_m\right\}, \\ \bar{\lambda}^{k+1} = \lambda^k - \beta(\sum_{i=1}^m A_i \bar{x}_i^{k+1} - b), \end{cases} \\ x_1^{k+1} = \bar{x}_1^{k+1}, \\ \boldsymbol{v}^{k+1} = \boldsymbol{v}^k - \alpha P^{-1}(\boldsymbol{v}^k - \bar{\boldsymbol{v}}^{k+1}), \quad \alpha \in (0, 1), \end{cases} \tag{8.1.10}$$

where $\mathcal{L}_\beta^m$ is defined in (8.1.6) and the matrix P is a block-wise upper triangular matrix defined as

$$P = \begin{pmatrix} I_{n_2}(A_2^T A_2)^{-1} A_2^T A_3 & \cdots & (A_2^T A_2)^{-1} A_2^T A_m & 0 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & (A_{m-1}^T A_{m-1})^{-1} A_{m-1}^T A_m & 0 \\ 0 & \cdots & I_{n_m} & 0 \\ 0 & \cdots & 0 & I_\ell \end{pmatrix}. \tag{8.1.11}$$

Here, the matrix $P \in \mathbb{R}^{(n-n_1+\ell)\times(n-n_1+\ell)}$. Note that in (8.1.10), $\boldsymbol{v}$ represents the collection of variables $(x_2^T, \cdots, x_m^T, \lambda^T)^T \in \mathbb{R}^{(n-n_1+\ell)}$ which are essentially required in the iteration, and we have

$$\boldsymbol{v}^k = \begin{pmatrix} x_2^k \\ \vdots \\ x_m^k \\ \lambda^k \end{pmatrix}, \quad \bar{\boldsymbol{v}}^k = \begin{pmatrix} \bar{x}_2^k \\ \vdots \\ \bar{x}_m^k \\ \bar{\lambda}^k \end{pmatrix}. \tag{8.1.12}$$

As mentioned in [3], the first variable x_1 is not required to execute the iteration; it is "intermediate" in the iteration. This is why in the scheme (8.1.10), the back substitution procedure is only implemented to $\boldsymbol{v}$, without x_1. Clearly, the last step in (8.1.10) can be written as

$$P(\boldsymbol{v}^{k+1} - \boldsymbol{v}^k) = \alpha(\bar{\boldsymbol{v}}^{k+1} - \boldsymbol{v}^k).$$

Thus, with the block-wise upper triangular matrix P defined in (8.1.11), the entries of $\boldsymbol{v}^{k+1}$ can be updated in the order of $\lambda \to x_m \to \cdots \to x_2$, just like the standard Gaussian back substitution procedure for solving a system of liner equations.

For the ADMM-GBS (8.1.10), the ADMM splitting step (i.e., the x_i-subproblems in (8.1.9)) is mainly for yielding easier subproblems so that it becomes possible to exploit the properties of θ_i's individually. However, yielding these easier subproblems means that the individual m x_i-subproblems in (8.1.10) is only an approximation of the ALM subproblem in (8.1.7) and thus the decomposed subproblems, even if all are solved exactly, are not necessarily accurate enough to provide a qualified input to update the Lagrange multiplier such that the convergence can be still ensured. This is an explanation of the failure of convergence for the direct extension of ADMM (8.1.9) for $m > 2$, see the counter example given in [6] showing the divergence of the direct extension of ADMM (8.1.9). The Gaussian back substitution step in (8.1.10) can thus be regarded as a correction step to compensate the inaccuracy resulted by the decomposition on the ALM and so as to ensure the contraction property for the iterative sequence to the solution set. With this contraction, the convergence of (8.1.10) can be established from the contraction method perspective.

In this paper, we focus on the particular case of (8.1.1) which arises from a big-data scenario; thus, m is assumed to be huge. Under this big-data scenario with a huge m, a pretreatment on the original model (data) is usually implemented. For example, we can classify the original functions and the corresponding variables into t classes by identifying some common features or data-processing in particular applications. A more specific case is that t represents the number of features in a data-mining application of the abstract model (8.1.1). In general, t is a handleable number but it is much smaller than m. The general model (8.1.1) is thus treated as a separable model with t blocks of functions and variables. For the r-th block ($r = 1, 2, \ldots, t$), let m_r be the number of variables in the r-th block and thus $\sum_{r=1}^{t} m_r = m$. That is, we consider regrouping the variables $x = (x_1, x_2, \ldots, x_m)$ and functions $(\theta_1, \theta_2, \ldots, \theta_m)$ in (8.1.1) as $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t)$ with $\boldsymbol{x}_r = (x_{r_1}, x_{r_2}, \ldots, x_{r_{m_r}})$ and $(\vartheta_1(\boldsymbol{x}_1), \vartheta_2(\boldsymbol{x}_2), \ldots, \vartheta_t(\boldsymbol{x}_t))$ with $\vartheta_r(\boldsymbol{x}_r) = \sum_{j=1}^{m_r} \theta_{r_j}(x_{r_j})$, respectively; and furthermore, we define

$$\mathcal{A}_r = (A_{r_1}, \ldots, A_{r_{m_r}}), \qquad \mathcal{X}_r = \prod_{j=1}^{m_r} X_{r_j},\ r = 1, \ldots, t. \tag{8.1.13}$$

Then, the model (8.1.1) can be reformulated as the block-wise form

$$\min \left\{ \sum_{r=1}^{t} \vartheta_r(\boldsymbol{x}_r) \,\Big|\, \sum_{r=1}^{t} \mathcal{A}_r \boldsymbol{x}_r = b,\ \boldsymbol{x}_r \in \mathcal{X}_r,\ r = 1, \ldots, t \right\}. \tag{8.1.14}$$

Note that the block-wise reformulation (8.1.14) may account for the application where each block of variables and functions represents a specific set of decision variables and cost functions in the same classification. Accordingly, the Lagrangian function (8.1.5) can be written as the block-wise

$$L^t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \lambda) = \sum_{r=1}^{t} \vartheta_r(\boldsymbol{x}_r) - \lambda^T (\textstyle\sum_{r=1}^{t} \mathcal{A}_r \boldsymbol{x}_r - b), \tag{8.1.15}$$

and thus the augmented Lagrangian function (8.1.6) as

$$\mathcal{L}_\beta^t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \lambda) = L^t(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \lambda) + \frac{\beta}{2} \|\textstyle\sum_{r=1}^{t} \mathcal{A}_r \boldsymbol{x}_r - b\|^2. \tag{8.1.16}$$

When $t = 2$, the original ADMM scheme (8.1.8) can be applicable to the block-wise reformulation (8.1.14) and its iterative scheme reads as

$$\begin{cases} \boldsymbol{x}_1^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^2(\boldsymbol{x}_1, \boldsymbol{x}_2^k, \lambda^k) \,\big|\, \boldsymbol{x}_1 \in \mathcal{X}_1 \right\}, \\ \boldsymbol{x}_2^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^2(\boldsymbol{x}_1^{k+1}, \boldsymbol{x}_2, \lambda^k) \,\big|\, \boldsymbol{x}_2 \in \mathcal{X}_2 \right\}, \\ \lambda^{k+1} = \lambda^k - \beta(\mathcal{A}_1 \boldsymbol{x}_1^{k+1} + \mathcal{A}_2 \boldsymbol{x}_2^{k+1} - b). \end{cases} \tag{8.1.17}$$

We refer to [19, 23] for the discussion of how to further decompose the subproblems in (8.1.17) and obtain solvable subproblems in form of (8.1.4).

We focus on the case of $t \geq 3$ and discuss how to design implementable algorithms for the block-wise reformulation (8.1.14). Recall that the scheme (8.1.9) is not necessarily convergent. Thus, its block-wise extension to (8.1.14), which reads as

$$\begin{cases} \boldsymbol{x}_1^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\boldsymbol{x}_1, \boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \,\big|\, \boldsymbol{x}_1 \in \mathcal{X}_1 \right\}, \\ \qquad \vdots \\ \boldsymbol{x}_r^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\boldsymbol{x}_1^{k+1}, \ldots, \boldsymbol{x}_{r-1}^{k+1}, \boldsymbol{x}_r, \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \,\big|\, \boldsymbol{x}_r \in \mathcal{X}_r \right\}, \\ \qquad \vdots \\ \boldsymbol{x}_t^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\boldsymbol{x}_1^{k+1}, \ldots, \boldsymbol{x}_{t-1}^{k+1}, \boldsymbol{x}_t, \lambda^k) \,\big|\, \boldsymbol{x}_t \in \mathcal{X}_t \right\}, \\ \lambda^{k+1} = \lambda^k - \beta(\textstyle\sum_{r=1}^{t} \mathcal{A}_r \boldsymbol{x}_r^{k+1} - b), \end{cases} \tag{8.1.18}$$

is not necessarily convergent, either; and it is important to investigate how to design implementable algorithms for (8.1.14) based on the scheme (8.1.18). Because of the well-verified efficiency and stability of the ADMM-GBS (8.1.10) in some areas such as image processing, statistical learning, and SDP, it is natural to consider extending it to a block-wise form. The resulting block-wise version of the ADMM-GBS (8.1.10) for the regrouped model (8.1.14) reads as

$$\begin{cases} \begin{cases} \bar{\boldsymbol{x}}_1^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\boldsymbol{x}_1, \boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \mid \boldsymbol{x}_1 \in \mathcal{X}_1 \right\}, \\ \qquad \vdots \\ \bar{\boldsymbol{x}}_r^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\bar{\boldsymbol{x}}_1^{k+1}, \ldots, \bar{\boldsymbol{x}}_{r-1}^{k+1}, \boldsymbol{x}_r, \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \mid \boldsymbol{x}_r \in \mathcal{X}_r \right\}, \\ \qquad \vdots \\ \bar{\boldsymbol{x}}_t^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\bar{\boldsymbol{x}}_1^{k+1}, \ldots, \bar{\boldsymbol{x}}_{t-1}^{k+1}, \boldsymbol{x}_t, \lambda^k) \mid \boldsymbol{x}_t \in \mathcal{X}_t \right\}, \\ \bar{\lambda}^{k+1} = \lambda^k - \beta(\sum_{r=1}^t \mathcal{A}_r \bar{\boldsymbol{x}}_r^{k+1} - b), \end{cases} \\ \boldsymbol{x}_1^{k+1} = \bar{\boldsymbol{x}}_1^{k+1}, \\ \boldsymbol{v}^{k+1} = \boldsymbol{v}^k - \alpha \mathcal{P}^{-1}(\boldsymbol{v}^k - \bar{\boldsymbol{v}}^{k+1}), \quad \alpha \in (0, 1), \end{cases} \tag{8.1.19}$$

where $\mathcal{L}_\beta^t$ is defined in (8.1.16) and the matrix $\mathcal{P}$ in (8.1.19) is a block-wise upper triangular matrix similar as in (8.1.11), see (8.3.2) for details. Note that this block-matrix $\mathcal{P}$ makes the output of (8.1.18) updated via a Gaussian back substitution procedure in block-wise form in the scheme (8.1.19).

The convergence of the block-wise scheme (8.1.19) is certainly ensured provided that all the resulting subproblems are solved exactly. For a general case, however, similar as (8.1.18), each of the minimization subproblems in (8.1.19) involves more than one function in its objective and the m_r variables are coupled by the quadratic term in (8.1.16). This may make it difficult to solve these subproblems exactly unless the special case $m_r = 1$. Recall that we only consider the case where each subproblem to be solved is in the form of (8.1.3) or (8.1.4). Thus, we suggest further decomposing the $\boldsymbol{x}_r$-subproblem in (8.1.19) as m_r smaller subproblems so that each function θ_i is treated individually. More specifically, the block-wise $\boldsymbol{x}_r$-subproblem in (8.1.19) is decomposed as the following m_r smaller subproblems:

$$\begin{cases} \bar{x}_{r_1}^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\bar{\boldsymbol{x}}_1^{k+1}, \ldots, \bar{\boldsymbol{x}}_{r-1}^{k+1}, x_{r_1}, x_{r_2}^k, \ldots, x_{r_{m_r}}^k, \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \mid x_{r_1} \in X_{r_1} \right\}, \\ \qquad \vdots \\ \bar{x}_{r_j}^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\bar{\boldsymbol{x}}_1^{k+1}, \ldots, \bar{\boldsymbol{x}}_{r-1}^{k+1}, x_{r_1}^k, \cdots, x_{r_{j-1}}^k, x_{r_j}, x_{r_{j+1}}^k, \ldots, x_{r_{m_r}}^k, \right. \\ \qquad\qquad \left. \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \mid x_{r_j} \in X_{r_j} \right\}, \\ \qquad \vdots \\ \bar{x}_{r_{m_r}}^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t(\bar{\boldsymbol{x}}_1^{k+1}, \ldots, \bar{\boldsymbol{x}}_{r-1}^{k+1}, x_{r_1}^k, \ldots, x_{r_{m_r-1}}^k, x_{r_{m_r}}, \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \mid x_{r_{m_r}} \in X_{r_{m_r}} \right\}. \end{cases} \tag{8.1.20}$$

Note that we only consider implementing the parallel decomposition to the $\boldsymbol{x}_r$-subproblem in (8.1.19). This makes it possible to implement parallel computation to tackle each block of subproblems by, e.g., a distributed-computing system. To summarize, the implementation of the new algorithm can be ordered as t main phases which are proceeded sequentially according to the block-wise ADMM-GBS scheme (8.1.19); and for the r-th phase, there are m_r subtasks in form of (8.1.4) which can be proceeded in parallel. This feature is useful for big-data scenarios where parallel computation is necessary.

The rest of this paper is organized as follows: In Section 8.2, we review some known results and prove some preliminary propositions which are useful for further analysis. The new algorithm is presented in Section 8.3, followed by some remarks. Then, we prove the convergence for the new algorithm in Section 8.4, and establish its worst-case convergence rate in Section 8.5. In Section 8.6, we elucidate some special cases of the new algorithm and see its relationship to some existing schemes in the literature. We present a refined version for the new algorithm with an iteratively calculated step size in Section 8.7, and briefly mention its convergence analysis. In Section 8.8, we present a linearized version of the new algorithm proposed in Section 8.3, whose subproblems are in form of (8.1.3) rather than (8.1.4). In addition, two key results which essentially guarantee its convergence are established for this linearized version. In Section 8.9, we repost some numerical results to verify the convergence of the new algorithms and the fact that different grouping strategies may result in different numerical performance. Finally, we make some conclusions in Section 8.10.

8.2 Preliminaries

In this section, we summarize some results known in the literature and introduce some additional notations for the convenience of analysis later.

8.2.1 Variational Inequality Characterization

Let $\left(x_1^*, x_2^*, \ldots, x_m^*, \lambda^*\right)$ be a saddle point of the Lagrangian function (8.1.5), it follows that

$$\begin{aligned}\sup_{\lambda\in\mathbb{R}^\ell} L^m(x_1^*, x_2^*, \ldots, x_m^*, \lambda) &\le L^m(x_1^*, x_2^*, \ldots, x_m^*, \lambda^*)\\ &\le \inf_{x_i\in X_i,\ i=1,\ldots,m} L^m(x_1, x_2, \ldots, x_m, \lambda^*).\end{aligned}$$

Then, finding a saddle point of $L^m(x_1, x_2, \ldots, x_m, \lambda)$ is equivalent to finding $(x_1^*, x_2^*, \ldots, x_m^*, \lambda^*) \in \Omega = X_1 \times X_2 \times \cdots \times X_m \times \mathbb{R}^\ell$ such that

$$\begin{cases} x_1^* \in X_1,\ \theta_1(x_1) - \theta_1(x_1^*) + (x_1 - x_1^*)^T(-A_1^T\lambda^*) \geq 0, & \forall x_1 \in X_1, \\ x_2^* \in X_2,\ \theta_2(x_2) - \theta_2(x_2^*) + (x_2 - x_2^*)^T(-A_2^T\lambda^*) \geq 0, & \forall x_2 \in X_2, \\ \qquad\qquad \vdots & \\ x_m^* \in X_m,\ \theta_m(x_m) - \theta_m(x_m^*) + (x_m - x_m^*)^T(-A_m^T\lambda^*) \geq 0, & \forall x_m \in X_m, \\ \lambda^* \in \mathbb{R}^\ell, \qquad (\lambda - \lambda^*)^T(\sum_{i=1}^m A_i x_i^* - b) \geq 0, & \forall \lambda \in \mathbb{R}^\ell. \end{cases} \tag{8.2.1}$$

We denote by Ω^* the set of all saddle points of $L^m(x_1, x_2, \ldots, x_m, \lambda)$. More compactly, (8.2.1) can be written as the following variational inequality:

$$\mathrm{VI}(\Omega, F, \theta) \quad \boldsymbol{w}^* \in \Omega, \quad \vartheta(\boldsymbol{x}) - \vartheta(\boldsymbol{x}^*) + (\boldsymbol{w} - \boldsymbol{w}^*)^T F(\boldsymbol{w}^*) \geq 0, \quad \forall\, \boldsymbol{w} \in \Omega, \tag{8.2.2a}$$

where

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \quad \boldsymbol{w} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \\ \lambda \end{pmatrix}, \quad \vartheta(\boldsymbol{x}) = \sum_{i=1}^m \theta_i(x_i), \quad F(\boldsymbol{w}) = \begin{pmatrix} -A_1^T\lambda \\ \vdots \\ -A_m^T\lambda \\ \sum_{i=1}^m A_i x_i - b \end{pmatrix}. \tag{8.2.2b}$$

Here, F is monotone. Using the mentioned block-wise notation, we can rewrite (8.2.1)–(8.2.2), respectively, as

$$\begin{cases} \boldsymbol{x}_1^* \in \mathcal{X}_1,\ \vartheta_1(\boldsymbol{x}_1) - \vartheta_1(\boldsymbol{x}_1^*) + (\boldsymbol{x}_1 - \boldsymbol{x}_1^*)^T(-\mathcal{A}_1^T\lambda^*) \geq 0, & \forall \boldsymbol{x}_1 \in \mathcal{X}_1, \\ \boldsymbol{x}_2^* \in \mathcal{X}_2,\ \vartheta_2(\boldsymbol{x}_2) - \vartheta_2(\boldsymbol{x}_2^*) + (\boldsymbol{x}_2 - \boldsymbol{x}_2^*)^T(-\mathcal{A}_2^T\lambda^*) \geq 0, & \forall \boldsymbol{x}_2 \in \mathcal{X}_2, \\ \qquad\qquad \vdots & \\ \boldsymbol{x}_t^* \in \mathcal{X}_t,\ \vartheta_t(\boldsymbol{x}_t) - \vartheta_t(\boldsymbol{x}_t^*) + (\boldsymbol{x}_t - \boldsymbol{x}_t^*)^T(-\mathcal{A}_m^T\lambda^*) \geq 0, & \forall \boldsymbol{x}_t \in \mathcal{X}_t, \\ \lambda^* \in \mathbb{R}^\ell, \qquad (\lambda - \lambda^*)^T(\sum_{r=1}^t \mathcal{A}_r \boldsymbol{x}_r^* - b) \geq 0, & \forall \lambda \in \mathbb{R}^\ell, \end{cases} \tag{8.2.3}$$

and

$$\mathrm{VI}(\Omega, F, \theta) \quad \boldsymbol{w}^* \in \Omega, \quad \vartheta(\boldsymbol{x}) - \vartheta(\boldsymbol{x}^*) + (\boldsymbol{w} - \boldsymbol{w}^*)^T F(\boldsymbol{w}^*) \geq 0, \quad \forall\, \boldsymbol{w} \in \Omega, \tag{8.2.4a}$$

where

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_t \end{pmatrix}, \quad \boldsymbol{w} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_t \\ \lambda \end{pmatrix}, \quad \vartheta(\boldsymbol{x}) = \sum_{r=1}^t \vartheta_r(\boldsymbol{x}_r), \quad F(\boldsymbol{w}) = \begin{pmatrix} -\mathcal{A}_1^T\lambda \\ \vdots \\ -\mathcal{A}_t^T\lambda \\ \sum_{r=1}^t \mathcal{A}_r \boldsymbol{x}_r - b \end{pmatrix}. \tag{8.2.4b}$$

8.2.2 Some Properties

Recall the matrices $\mathcal{A}_r$'s defined in (8.1.13). Then, for $\mathcal{A}_r$ and $\mathcal{A}_s$, we have

$$\mathcal{A}_r^T\mathcal{A}_s = \begin{pmatrix} A_{r_1}^T A_{s_1} & \cdots\cdots & A_{r_1}^T A_{s_{m_s}} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ A_{r_{m_r}}^T A_{s_1} & \cdots\cdots & A_{r_{m_r}}^T A_{s_{m_s}} \end{pmatrix}.$$

The matrices $\mathcal{A}_r$'s have a useful property for further analysis. We summarize it in the following lemma and omit its trivial proof.

Lemma 8.2.1 *For the matrix $\mathcal{A}_r$ defined in (8.1.13), if we define*

$$\operatorname{diag}(\mathcal{A}_r^T\mathcal{A}_r) := \begin{pmatrix} A_{r_1}^T A_{r_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_{r_{m_r}}^T A_{r_{m_r}} \end{pmatrix}, \tag{8.2.5}$$

then we have

$$m_r \cdot \operatorname{diag}(\mathcal{A}_r^T\mathcal{A}_r) \succeq \mathcal{A}_r^T\mathcal{A}_r, \quad r = 1, \ldots, t. \tag{8.2.6}$$

Proof Clearly, we have

$$\begin{aligned} & m_r \cdot \operatorname{diag}(\mathcal{A}_r^T\mathcal{A}_r) - \mathcal{A}_r^T\mathcal{A}_r \\ = & \begin{pmatrix} m_r A_{r_1}^T A_{r_1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & m_r A_{r_{m_r}}^T A_{r_{m_r}} \end{pmatrix} - \begin{pmatrix} A_{r_1}^T A_{r_1} & \cdots\cdots & A_{r_1}^T A_{r_{m_r}} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ A_{r_{m_r}}^T A_{r_1} & \cdots\cdots & A_{r_{m_r}}^T A_{r_{m_r}} \end{pmatrix} \\ = & \begin{pmatrix} (m_r-1)A_{r_1}^T A_{r_1} & \cdots\cdots & -A_{r_1}^T A_{r_{m_r}} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ -A_{r_{m_r}}^T A_{r_1} & \cdots\cdots & (m_r-1)A_{r_{m_r}}^T A_{r_{m_r}} \end{pmatrix} \end{aligned}$$

$$= \mathcal{A}_r^T \begin{pmatrix} (m_r-1)I_\ell & \cdots\cdots & -I_\ell \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ -I_\ell & \cdots\cdots & (m_r-1)I_\ell \end{pmatrix} \mathcal{A}_r \succeq 0. \tag{8.2.7}$$

□

Furthermore, we define

$$\tau_r \geq m_r - 1, \qquad \text{and} \qquad D_r = (\tau_r + 1)\text{diag}(\mathcal{A}_r^T \mathcal{A}_r), \quad r = 1, \ldots, t. \tag{8.2.8}$$

8.3 The Block-Wise ADMM with Gaussian Back Substitution

In this section, we propose an implementable version of the block-wise ADMM-GBS with solvable subproblems in form of (8.1.4). In particular, this block-wise ADMM-GBS turns out to be a unified scheme including the existing algorithms in [15, 17] as special cases. Some remarks are also given.

8.3.1 The New Algorithm

Based on the previous discussion, we now propose the new algorithm which embeds the parallel computation (8.1.4) into the block-wise ADMM-GBS (8.1.19). As analyzed in [14, 23], if we replace the $\boldsymbol{x}_r$-subproblems in (8.1.19) directly by the further decomposed subproblems in (8.1.20), the convergence is not guaranteed. In fact, the proximity to the last iterate should be controlled when solving the further subproblems in (8.1.20). Therefore, we should embed not the subproblems in (8.1.20), but their regularized counterparts:

$$\begin{aligned} & x_{r_j}^{k+1} \\ & = \arg\min \left\{ \begin{array}{c} \mathcal{L}_\beta^t(\boldsymbol{x}_1^{k+1}, \ldots, \boldsymbol{x}_{r-1}^{k+1}, x_{r_1}^k, \ldots, x_{r_j-1}^k, x_{r_j}, x_{r_j+1}^k, \ldots, x_{r_{m_r}}^k, \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) \\ + \frac{\tau_r \beta}{2} \| A_{r_j}(x_{r_j} - x_{r_j}^k) \|^2 \end{array} \Big| x_{r_j} \in X_{r_j} \right\} \end{aligned} \tag{8.3.1}$$

with τ_r $(r = 1, \cdots, t)$ into the block-wise ADMM-GBS (8.1.19). By defining a matrix

$$\mathcal{P} = \begin{pmatrix} I & 0 & 0 & \cdots & 0 & 0 \\ 0 & I & D_2^{-1}\mathcal{A}_2^T\mathcal{A}_3 & \cdots & D_2^{-1}\mathcal{A}_2^T\mathcal{A}_t & 0 \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & D_{t-1}^{-1}\mathcal{A}_{t-1}^T\mathcal{A}_t & 0 \\ 0 & 0 & \cdots & 0 & I & 0 \\ 0 & 0 & \cdots & 0 & 0 & I_\ell \end{pmatrix}, \tag{8.3.2}$$

where $\mathcal{P} \in \mathbb{R}^{(n+\ell)\times(n+\ell)}$ and D_r is defined in (8.2.8), we summarize the resulting algorithm as follows.

Algorithm 1: A splitting version of the block-wise ADMM-GBS (8.1.19) for (8.1.1)

Initialization: Specify a regrouping for the model (8.1.1) with determined values of t and m_r for $r = 1, 2, \ldots, t$. Choose constants τ_r such that $\tau_r \geq m_r - 1$ for $r = 1, \ldots, t$ and $\beta > 0$. Let $\mathcal{P}$ be defined in (8.3.2). Choose $\boldsymbol{w}^0 = (\boldsymbol{x}_1^0, \boldsymbol{x}_2^0, \cdots, \boldsymbol{x}_t^0, \lambda^0) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_t \times \mathbb{R}^\ell$, for every $k \geq 0$,

$$\begin{cases} \text{for } r = 1, 2, \ldots t, \text{ do:} \\ \quad \text{for } j = 1, \ldots m_r, \text{ parallel do:} \\ \quad\quad \bar{x}_{r_j}^{k+1} = \arg\min \left\{ \mathcal{L}_\beta^t\left(\bar{\boldsymbol{x}}_1^{k+1}, \ldots, \bar{\boldsymbol{x}}_{r-1}^{k+1}, x_{r1}^k, \ldots, x_{r_{j-1}}^k, x_{r_j}, x_{r_{j+1}}^k, \ldots, x_{rm_r}^k, \right. \right. \\ \quad\quad\quad \left. \left. \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k\right) + \frac{\tau_r\beta}{2}\|A_{r_j}(x_{r_j} - x_{r_j}^k)\|^2 \;\middle|\; x_{r_j} \in X_{r_j} \right\}; \\ \quad \text{end.} \\ \text{end.} \\ \bar{\lambda}^{k+1} = \lambda^k - \beta\left(\sum_{r=1}^t \mathcal{A}_r\bar{\boldsymbol{x}}_r^{k+1} - b\right). \end{cases}$$

$$\mathcal{P}(\boldsymbol{w}^{k+1} - \boldsymbol{w}^k) = \alpha(\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k), \quad \alpha \in (0, 1). \tag{8.3.3}$$

Remark 8.3.1 To implement the proposed algorithm (8.3.3), at most $\max\{m_1, \ldots, m_t\}$ work stations are needed. Also, the proximal parameters τ_r is only dependent on the number of variables m_r of the r-th group; they thus can be significantly smaller than $m - 1$ as required in (8.3.5). This feature thus can avoid slow convergence due to too large proximal coefficients. Certainly, when a specific application of the abstract model (8.1.1) is considered, the user can optimally determine the values of t and m_r for $r = 1, 2, \ldots, t$, so that the balance among the sequential and parallel computation is achieved and the optimal overall performance is achieved. But in this paper, we focus on the general methodology of algorithmic design for the generic

case of (8.1.1), and do not discuss the specific regrouping strategies among variables which are case-dependent.

8.3.2 Some Remarks

It is easy to see that at each iteration, the new algorithm (8.3.3) mainly requires solving m subproblems in form of (8.1.4). We use the proximal terms $\frac{\tau_r\beta}{2}\|A_{r_j}(x_{r_j} - x_{r_j}^k)\|^2$ to regularize the further decomposed subproblems in (8.3.3). But, just like the analysis in [22], we can instead use the terms $\frac{\tau_r\beta}{2}\|x_{r_j} - x_{r_j}^k\|^2$, or more generally $\frac{\tau_r\beta}{2}\|x_{r_j} - x_{r_j}^k\|_G^2$ with a positive definite matrix G. Therefore, for the case where A_i is not the identity matrix while θ_i is simple in the sense that its proximal operator defined in (8.1.3) has a closed-form representation, then we can easily further consider linearizing the quadratic term in its corresponding subproblem in (8.3.3) and thus propose a linearized version of the algorithm (8.3.3). The corresponding analysis is not much different from our analysis to be presented. We thus will only briefly discuss the linearized version in Section 8.8, and mainly focus on the discussion for the scheme (8.3.3) for the purpose of exposing our main idea with easier notation.

It is also worthwhile to mention that if the alternating decomposition is implemented to the $\boldsymbol{x}_r$-subproblem in (8.1.19), then the resulting scheme reduces to the original ADMM-GBS (8.1.10). Recall that the ADMM-GBS (8.1.10) requires solving all the decomposed subproblems in a completely sequential way. Hence, when the big-data scenario is considered where m is huge in (8.1.1), the waiting time resulted by the sequential computing might be too expensive if the ADMM-GBS (8.1.10) is directly used. We are thus interested in implementing the ADMM-GBS in the block-wise form (8.1.19) but further decomposing the block-wise subproblems in the parallel way of (8.1.20). In this way, the advantage of the ADMM-GBS such as its efficiency and stability is preserved among blocks while the parallel computation to tackle big-data scenarios is applicable within each block. This is the main motivation of the new algorithm to be proposed.

We have emphasized the importance of parallel computation to tackle the big-data scenarios of the model (8.1.1). One may ask why not just implement the full parallel decomposition directly to the ALM (8.1.7) and thus obtain the following scheme whose m x_i-subproblems can be solved fully in parallel:

$$\begin{cases} x_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1, x_2^k, \ldots, x_m^k, \lambda^k) \mid x_1 \in X_1\right\}, \\ \quad\vdots \\ x_i^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \mid x_i \in X_i\right\}, \\ \quad\vdots \\ x_m^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) \mid x_m \in X_m\right\}, \\ \lambda^{k+1} = \lambda^k - \beta(\sum_{i=1}^m A_i x_i^{k+1} - b). \end{cases} \tag{8.3.4}$$

In fact, the scheme (8.3.4) requires m work stations to realize the parallel computation. When m is huge for a big-data scenario, it might be too expensive to be practical. Moreover, from methodological point of view, as shown in [13], the scheme (8.3.4) is not necessarily convergent even for $m = 2$. Later, it was shown in [18] that the convergence of (8.3.4) can be guaranteed if all the x_i-subproblems are proximally regularized by certain proximal term

$$\begin{cases} x_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1, x_2^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_1(x_1 - x_1^k)\|^2 \mid x_1 \in X_1\right\}, \\ \quad\vdots \\ x_i^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1^k, \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) + \frac{s\beta}{2}\|A_i(x_i - x_i^k)\|^2 \mid x_i \in X_i\right\}, \\ \quad\vdots \\ x_m^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1^k, \ldots, x_{m-1}^k, x_m, \lambda^k) + \frac{s\beta}{2}\|A_m(x_m - x_m^k)\|^2 \mid x_m \in X_m\right\}, \\ \lambda^{k+1} = \lambda^k - \beta(\sum_{i=1}^m A_i x_i^{k+1} - b), \end{cases} \tag{8.3.5}$$

where the proximal parameter s is required to be greater than $m - 1$. The x_i-subproblems in the scheme (8.3.5) are also eligible for parallel computation. But recall that we are considering the big-data scenarios where m is huge. Thus, the proximal terms in (8.3.5) with $s \geq m - 1$ play a dominate role in the objective functions and the convergence is doomed to be extremely slow due to the huge value of $m - 1$, though the convergence can be guaranteed asymptotically. Therefore, we do not expect that the existing schemes based on the technique of directly decomposing the ALM (8.1.7) in a parallel way are applicable for the big-data scenarios of (8.1.1) with a huge m. Note that in [12, 13], it was also suggested to correct the output of (8.3.4) by certain correction steps and the proximal terms are not needed to regularize the decomposed subproblems. But these schemes also require m work stations to realize the parallel computation.

8.4 Convergence

In this section, we prove the global convergence for the proposed algorithm (8.3.3).

8.4.1 Some Matrices

First of all, for the convenience of analysis, let us define some matrices and prove some useful properties for these matrices. Let

$$Q = \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1) & 0 & \cdots & \cdots & 0 & 0 \\ 0 & \beta D_2 & \ddots & & \vdots & \vdots \\ 0 & \beta\mathcal{A}_3^T\mathcal{A}_2 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \beta\mathcal{A}_t^T\mathcal{A}_2 & \cdots & \beta\mathcal{A}_t^T\mathcal{A}_{t-1} & \beta D_t & 0 \\ 0 & -\mathcal{A}_2 & \cdots & -\mathcal{A}_{t-1} & -\mathcal{A}_t & \frac{1}{\beta}I \end{pmatrix}, \tag{8.4.1}$$

where $\mathcal{A}_r$ and D_r are defined in (8.1.13) and (8.2.8), respectively.

In fact, the matrix Q in (8.4.1) can be written as the block-wise form

$$Q = \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta\mathcal{Q}_e & 0 \\ 0 & -\mathcal{A} & \frac{1}{\beta}I \end{pmatrix}, \tag{8.4.2}$$

with

$$\mathcal{A} = (\mathcal{A}_2, \ldots, \mathcal{A}_t) \tag{8.4.3}$$

and

$$\mathcal{Q}_e = \begin{pmatrix} D_2 & 0 & \cdots & 0 \\ \mathcal{A}_3^T\mathcal{A}_2 & D_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \mathcal{A}_t^T\mathcal{A}_2 & \cdots & \mathcal{A}_t^T\mathcal{A}_{t-1} & D_t \end{pmatrix}. \tag{8.4.4}$$

Moreover, we use $\mathcal{D}_e$ to denote the diagonal part of $\mathcal{Q}_e$, i.e.,

$$\mathcal{D}_e = \begin{pmatrix} D_2 & 0 & \cdots & 0 \\ 0 & D_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & D_t \end{pmatrix}. \tag{8.4.5}$$

With the just defined matrices $\mathcal{A}$, $\mathcal{Q}_e$, and $\mathcal{D}_e$, we further define

$$M = \begin{pmatrix} I & 0 & 0 \\ 0 & \mathcal{Q}_e^{-T}\mathcal{D}_e & 0 \\ 0 & -\beta\mathcal{A} & I \end{pmatrix}, \tag{8.4.6}$$

where $M \in \mathbb{R}^{(n+\ell)\times(n+\ell)}$. These matrices will help us present the upcoming analysis more succinctly.

Indeed, proving the convergence for the proposed algorithm (8.3.3) crucially depends on some important properties of the just defined matrices. We summarize them in the following two lemmas.

Lemma 8.4.1 *For the matrices $\mathcal{A}$, $\mathcal{Q}_e$, and $\mathcal{D}_e$ which are defined in* (8.4.3), (8.4.4), *and* (8.4.5), *respectively, we have*

$$\mathcal{Q}_e^T + \mathcal{Q}_e \begin{cases} \succeq \mathcal{D}_e + \mathcal{A}^T\mathcal{A}, \ \tau_r \geq m_r - 1, \ r = 1, \ldots, t; \\ \succ \mathcal{D}_e + \mathcal{A}^T\mathcal{A}, \ \tau_r > m_r - 1, \ r = 1, \ldots, t. \end{cases} \tag{8.4.7}$$

Proof Using the structure of the matrices $\mathcal{Q}_e$ and $\mathcal{D}_e$ (see (8.4.4) and (8.4.5)), we obtain

$$\mathcal{Q}_e^T + \mathcal{Q}_e = \mathcal{D}_e + \begin{pmatrix} D_2 & \mathcal{A}_2^T\mathcal{A}_3 & \cdots & \mathcal{A}_2^T\mathcal{A}_t \\ \mathcal{A}_3^T\mathcal{A}_2 & D_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathcal{A}_{t-1}^T\mathcal{A}_t \\ \mathcal{A}_t^T\mathcal{A}_2 & \cdots & \mathcal{A}_t^T\mathcal{A}_{t-1} & D_t \end{pmatrix}.$$

Since we choose $\tau_r \geq$ (resp. $>$) $m_r - 1$, it follows that

$$D_r = (\tau_r + 1)\mathrm{diag}(\mathcal{A}_r^T\mathcal{A}_r) \succeq (\text{Resp.}, \succ)\mathcal{A}_r^T\mathcal{A}_r, \quad r = 1, \ldots, t,$$

and consequently,

$$\begin{pmatrix} D_2 & \mathcal{A}_2^T\mathcal{A}_3 & \cdots & \mathcal{A}_2^T\mathcal{A}_t \\ \mathcal{A}_3^T\mathcal{A}_2 & D_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathcal{A}_{t-1}^T\mathcal{A}_t \\ \mathcal{A}_t^T\mathcal{A}_2 & \cdots & \mathcal{A}_t^T\mathcal{A}_{t-1} & D_t \end{pmatrix} \succeq (\text{Resp.}, \succ)\mathcal{A}^T\mathcal{A}.$$

The assertions (8.4.7) hold immediately. □

Lemma 8.4.2 *For the matrices Q and M defined in* (8.4.1) *and* (8.4.6), *respectively, let*

$$H := QM^{-1} \tag{8.4.8a}$$

and

$$G := Q^T + Q - \alpha M^T H M. \tag{8.4.8b}$$

Then, we have the following conclusions:

(i) The matrix H defined in (8.4.8a) is symmetric and positive definite.
(ii) For the matrix G defined in (8.4.8b), we have

$$G=Q^T+Q-\alpha M^T H M \begin{cases} \begin{cases} \succ 0, \ \forall \alpha \in (0,1), \\ \succeq 0, \quad \alpha = 1, \end{cases} & \text{if} \quad \tau_r \geq m_r - 1, \ r = 1, \ldots, t; \\ \\ \succ 0, \quad \forall\, \alpha \in (0,1], \ \text{if} \quad \tau_r > m_r - 1, \ r = 1, \ldots, t. \end{cases} \tag{8.4.9}$$

Proof First, we check the positive definiteness of the matrix H. For the matrix M defined in (8.4.6), we have

$$M^{-1} = \begin{pmatrix} I & 0 & 0 \\ 0 & \mathcal{D}_e^{-1}\mathcal{Q}_e^T & 0 \\ 0 & \beta\mathcal{A}\mathcal{D}_e^{-1}\mathcal{Q}_e^T & I \end{pmatrix}.$$

Thus, according to the definition of the matrix H (see (8.4.8a)), we conclude that

$$\begin{aligned} H = QM^{-1} &= \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta\mathcal{Q}_e & 0 \\ 0 & -\mathcal{A} & \frac{1}{\beta}I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & \mathcal{D}_e^{-1}\mathcal{Q}_e^T & 0 \\ 0 & \beta\mathcal{A}\mathcal{D}_e^{-1}\mathcal{Q}_e^T & I \end{pmatrix} \\ &= \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta\mathcal{Q}_e\mathcal{D}_e^{-1}\mathcal{Q}_e^T & 0 \\ 0 & 0 & \frac{1}{\beta}I \end{pmatrix} \end{aligned}$$

is symmetric and positive definite.

Now, we turn to check the positive definiteness of the matrix G. Note that

$$\begin{aligned} Q^T + Q &= \begin{pmatrix} 2\beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{Q}_e^T + \mathcal{Q}_e) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{2}{\beta}I \end{pmatrix} \\ &\overset{(8.4.7)}{\succeq} \begin{pmatrix} 2\beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{D}_e + \mathcal{A}^T\mathcal{A}) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{2}{\beta}I \end{pmatrix} \end{aligned}$$

and

$$M^T HM = M^T Q = Q^T M = \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T \mathcal{A}_1) & 0 & 0 \\ 0 & \beta \mathcal{Q}_e^T & -\mathcal{A}^T \\ 0 & 0 & \frac{1}{\beta} I \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & \mathcal{Q}_e^{-T} \mathcal{D}_e & 0 \\ 0 & -\beta \mathcal{A} & I \end{pmatrix}$$

$$= \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T \mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{D}_e + \mathcal{A}^T \mathcal{A}) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{1}{\beta} I \end{pmatrix}. \tag{8.4.10}$$

From the definition of G (see (8.4.8b) and the two different cases of (8.4.7)), it follows that

$$\begin{aligned} G &= Q^T + Q - \alpha M^T HM \\ &\begin{pmatrix} \succeq \\ \succ \end{pmatrix} \begin{pmatrix} (2-\alpha)\beta(D_1 - \mathcal{A}_1^T \mathcal{A}_1) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} + (1-\alpha) \begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta(\mathcal{D}_e + \mathcal{A}^T \mathcal{A}) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{1}{\beta} I \end{pmatrix} \\ &\succeq 0. \end{aligned}$$

The assertion (8.4.9) is proved. □

As we shall see, Lemma 8.4.2 actually play a very important role in proving the convergence for the proposed algorithm (8.3.3).

8.4.2 *A Prediction-Correction Reformulation of (8.3.3)*

Now, with the matrices introduced in the last subsection, we can rewrite the proposed algorithm (8.3.3) as the following prediction-correction form.

Prediction. For the given $\boldsymbol{w}^k = (x_1^k, x_2^k \ldots, x_m^k, \lambda^k) = (\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_t^k, \lambda^k)$, generate the predictor $\tilde{\boldsymbol{w}}^k = (\tilde{x}_1^k, \tilde{x}_2^k \ldots, \tilde{x}_m^k, \tilde{\lambda}^k) = (\tilde{\boldsymbol{x}}_1^k, \ldots, \tilde{\boldsymbol{x}}_t^k, \tilde{\lambda}^k)$ by the following steps:

$$\left\{\begin{array}{l} \text{for } r = 1, 2, \ldots t, \text{ do:} \\ \quad \text{for } j = 1, \ldots m_r, \text{ parallel do:} \\ \qquad \tilde{x}_{r_j}^k = \arg\min \left\{ \begin{array}{l} \mathcal{L}_\beta^t(\tilde{\boldsymbol{x}}_1^k, \ldots, \tilde{\boldsymbol{x}}_{r-1}^k, x_{r_1}^k, \ldots, x_{r_{j-1}}^k, x_{r_j}, x_{r_{j+1}}^k, \ldots, x_{r_{mr}}^k, \\ \quad \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k) + \frac{\tau_r \beta}{2} \|A_{r_j}(x_{r_j} - x_{r_j}^k)\|^2 \end{array} \middle| x_{r_j} \in X_{r_j} \right\}; \\ \quad \text{end.} \\ \text{end.} \end{array}\right. \tag{8.4.11a}$$

Additionally, we define

$$\tilde{\lambda}^k = \lambda^k - \beta\big(\mathcal{A}_1 \tilde{\boldsymbol{x}}_1^k + \textstyle\sum_{j=2}^t \mathcal{A}_j \boldsymbol{x}_j^k - b\big). \tag{8.4.11b}$$

Correction. The new iterate $\boldsymbol{w}^{k+1}$ is given by

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \alpha M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{8.4.12a}$$

where $\tilde{\boldsymbol{w}}^k$ is the predictor generated by (8.4.11), the matrix M is defined in (8.4.6) and

$$\alpha \in \left\{ \begin{array}{ll} (0, 1), \text{ if } & \tau_r \geq m_r - 1, \ r = 1, \ldots, t; \\ (0, 1], \text{ if } & \tau_r > m_r - 1, \ r = 1, \ldots, t. \end{array} \right. \tag{8.4.12b}$$

As mentioned in [23], we conduct the convergence analysis in the context of the prediction-correction form (8.4.11)–(8.4.12) because the proof of the convergence is essentially to prove the Féjer monotonicity property with respect to the solution set, while the progress of the proximity to the solution set is measured by the quantity $\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2$, where G is defined in (8.4.8b). Thus, it is convenient to explicitly analyze the predictor $\tilde{\boldsymbol{w}}^k$ and accordingly revisit the algorithm (8.3.3) from the prediction-correction perspective. The other reason is that this prediction-correction reformulation enables us to investigate the relationship between the proposed algorithm (8.3.3) and some existing schemes in the literature by a unified framework, as elaborated in Sections 8.6.1 and 8.6.2.

Let us take a closer look at the correction step (8.4.12). Recall that the matrix M defined in (8.4.6) and the matrices $\mathcal{Q}_e$, $\mathcal{D}_e$ in M are defined in (8.4.4) and (8.4.5),

respectively. Moreover, using (8.4.3) and (8.4.11b), we can see that the correction step (8.4.12) consists of the following computations:

$$\begin{cases} \boldsymbol{x}_1^{k+1} - \boldsymbol{x}_1^k = \alpha(\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k), \\ \mathcal{D}_e^{-1}\mathcal{Q}_e^T \begin{pmatrix} \boldsymbol{x}_2^{k+1} - \boldsymbol{x}_2^k \\ \vdots \\ \boldsymbol{x}_t^{k+1} - \boldsymbol{x}_t^k \end{pmatrix} = \alpha \begin{pmatrix} \tilde{\boldsymbol{x}}_2^k - \boldsymbol{x}_2^k \\ \vdots \\ \tilde{\boldsymbol{x}}_t^k - \boldsymbol{x}_t^k \end{pmatrix}, \\ \lambda^{k+1} = \lambda^k - \alpha\beta\big(\sum_{s=1}^t \mathcal{A}_s\tilde{\boldsymbol{x}}_s^k - b\big). \end{cases} \tag{8.4.13}$$

Notice that $\mathcal{D}_e^{-1}\mathcal{Q}_e^T$ is a block-wise upper triangular matrix whose diagonal parts are identities. Thus, the block-wise variables $(\boldsymbol{x}_2, \boldsymbol{x}_3, \ldots, \boldsymbol{x}_t)$ are updated consecutively in the back substitution order: $\boldsymbol{x}_t^{k+1} \to \boldsymbol{x}_{t-1}^{k+1} \to \cdots\cdots \to \boldsymbol{x}_2^{k+1}$. Recall that within each block variable, the further decomposed subproblems are eligible for parallel computation. Thus, the correction step (8.4.12) can be viewed as a Gaussian back substitution procedure to correct the output of (8.4.11).

Now, let us come back to the prediction step (8.4.11). In the following lemma, we analyze the optimality conditions of the $\tilde{x}_{r_j}$-subproblems in (8.4.11) and represent the predictor generated by (8.4.11) as a VI reformulation. This VI reformulation helps us better discern its difference from the VI characterization (8.2.4) of the original model (8.1.1), and thus clearly see how far the predictor $\tilde{w}^k$ is from a solution point. It also inspires the correction step (8.4.12).

Lemma 8.4.3 *Let $\tilde{\boldsymbol{x}}^k$ be generated by (8.4.11a) from the given vector $\boldsymbol{w}^k$ and $\tilde{\lambda}^k$ be defined by (8.4.11b). Then, the predictor $\tilde{\boldsymbol{w}}^k \in \Omega$ satisfies*

$$\tilde{\boldsymbol{w}}^k \in \Omega, \quad \vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \geq (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \quad \forall \boldsymbol{w} \in \Omega, \tag{8.4.14}$$

where Q is defined in (8.4.1).

Proof Using the notation of the augmented Lagrangian function (see (8.1.16)), we observe the optimality condition of the x_{r_j}-subproblem in the r-th group of (8.4.11a) for $r = 1, \ldots, t$. Ignoring some constant terms in the objective function of the subproblems, we have

$$\begin{aligned} \tilde{x}_{r_j}^k &= \arg\min \left\{ \begin{array}{c} \mathcal{L}_\beta^t\big(\tilde{\boldsymbol{x}}_1^k, \ldots, \tilde{\boldsymbol{x}}_{r-1}^k, x_{r_1}^k, \ldots, x_{r_{j-1}}^k, x_{r_j}, x_{r_{j+1}}^k, \ldots, x_{r_{m_r}}^k, \\ \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k\big) + \frac{\tau_r\beta}{2}\|A_{r_j}(x_{r_j} - x_{r_j}^k)\|^2 \end{array} \middle| x_{r_j} \in X_{r_j} \right\} \\ &\overset{(8.1.6)}{=} \arg\min \left\{ \begin{array}{c} \theta_{r_j}(x_{r_j}) - (\lambda^k)^T A_{r_j} x_{r_j} + \frac{\beta}{2}\|A_{r_j}(x_{r_j} - x_{r_j}^k) + \sum_{s=1}^{r-1} \mathcal{A}_s\tilde{\boldsymbol{x}}_s^k \\ + \sum_{s=r}^t \mathcal{A}_s\boldsymbol{x}_s^k - b\|^2 + \frac{\tau_r\beta}{2}\|A_{r_j}(x_{r_j} - x_{r_j}^k)\|^2 \end{array} \middle| x_{r_j} \in X_{r_j} \right\}. \end{aligned}$$

The optimality condition of the above convex minimization problem is

$$\begin{aligned}\tilde{x}_{r_j}^k \in X_{r_j},\ \theta_{r_j}(x_{r_j}) - \theta_{r_j}(\tilde{x}_{r_j}^k) + (x_{r_j} - \tilde{x}_{r_j}^k)^T \big\{ - A_{r_j}^T \lambda^k \\ \beta A_{r_j}^T \big[\textstyle\sum_{s=1}^{r-1} \mathcal{A}_s \tilde{\boldsymbol{x}}_s^k + \sum_{s=r}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b \big] \\ + (\tau_r + 1)\beta A_{r_j}^T A_{r_j} (\tilde{x}_{r_j}^k - x_{r_j}^k) \big\} \geq 0, \quad \forall x_{r_j} \in X_{r_j}.\end{aligned}$$

For $r = 2, \ldots, m$, by using the definition of $\tilde{\lambda}^k$ (see (8.4.11b)), we have

$$\lambda^k = \tilde{\lambda}^k + \beta\big(\mathcal{A}_1 \tilde{\boldsymbol{x}}_1^k + \textstyle\sum_{s=2}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b\big).$$

Substituting it into the last inequality, we obtain

$$\begin{aligned}\tilde{x}_{r_j}^k \in X_{r_j}, \theta_{r_j}(x_{r_j}) - \theta_{r_j}(\tilde{x}_{r_j}^k) + (x_{r_j} - \tilde{x}_{r_j}^k)^T \big\{ - A_{r_j}^T \tilde{\lambda}^k \\ + \beta A_{r_j}^T \big[\textstyle\sum_{s=2}^{r-1} \mathcal{A}_s (\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) \big] + (\tau_r + 1)\beta A_{r_j}^T A_{r_j} (\tilde{x}_{r_j}^k - x_{r_j}^k) \big\} \geq 0,\ \forall x_{r_j} \in X_{r_j}.\end{aligned}$$

Applying this inequality for the cases of $j = 1, \ldots, m_r$, and summarizing the resulting inequalities, we get

$$\begin{aligned}\tilde{\boldsymbol{x}}_r^k \in \mathcal{X}_r,\ \vartheta_r(\boldsymbol{x}_r) - \vartheta_r(\tilde{\boldsymbol{x}}_r^k) + (\boldsymbol{x}_r - \tilde{\boldsymbol{x}}_r^k)^T \big\{ - \mathcal{A}_r^T \tilde{\lambda}^k + \beta \mathcal{A}_r^T \big[\textstyle\sum_{s=2}^{r-1} \mathcal{A}_s (\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) \big] \\ + (\tau_r + 1)\beta \mathrm{diag}(\mathcal{A}_r^T \mathcal{A}_r)(\tilde{\boldsymbol{x}}_r^k - \boldsymbol{x}_r^k) \big\} \geq 0,\ \forall \boldsymbol{x}_r \in \mathcal{X}_r.\end{aligned} \tag{8.4.15}$$

For $r = 1$, recall the optimality condition and combine the definition of $\tilde{\lambda}^k$. We obtain that

$$\begin{aligned}\tilde{\boldsymbol{x}}_1^k \in \mathcal{X}_1, \quad \vartheta_1(\boldsymbol{x}_1) - \vartheta_1(\tilde{\boldsymbol{x}}_1^k) + (\boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1^k)^T \big\{ - \mathcal{A}_1^T \tilde{\lambda}^k \\ - \beta \mathcal{A}_1^T \mathcal{A}_1 (\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k) + (\tau_1 + 1)\beta \mathrm{diag}(\mathcal{A}_1^T \mathcal{A}_1)(\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k) \big\} \geq 0, \quad \forall \boldsymbol{x}_1 \in \mathcal{X}_1.\end{aligned}$$

Using the notation of matrix D_1 (see (8.2.8)), it can be written as

$$\begin{aligned}\tilde{\boldsymbol{x}}_1^k \in \mathcal{X}_1, \quad \vartheta_1(\boldsymbol{x}_1) - \vartheta_1(\tilde{\boldsymbol{x}}_1^k) + (\boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1^k)^T \big\{ - \mathcal{A}_1^T \tilde{\lambda}^k + \beta (D_1 - \mathcal{A}_1^T \mathcal{A}_1)(\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k) \big\} \\ \geq 0, \quad \forall \boldsymbol{x}_1 \in \mathcal{X}_1.\end{aligned} \tag{8.4.16}$$

In addition, by using (8.4.11b), we have

$$(\sum_{r=1}^{t} \mathcal{A}_r \tilde{\boldsymbol{x}}_r^k - b) - \sum_{s=2}^{t} \mathcal{A}_s (\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k) = 0,$$

and it can be rewritten as

$$\tilde{\lambda}^k \in \mathbb{R}^\ell, \quad (\lambda - \tilde{\lambda}^k)^T \big\{ (\sum_{r=1}^{t} \mathcal{A}_r \tilde{\boldsymbol{x}}_r^k - b) - \sum_{s=2}^{t} \mathcal{A}_s (\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k) \big\} \geq 0, \quad \forall\, \lambda \in \mathbb{R}^\ell. \tag{8.4.17}$$

Combining (8.4.16), (8.4.15) ($r = 2, \ldots, t$), and (8.4.17) together and using the notations $F(\boldsymbol{w})$, Q, and D_r (see (8.2.2), (8.4.1), and (8.2.8)), we have the result of this lemma. □

Recall the VI reformulation (8.2.4a)–(8.2.4b) of the model (8.1.1). Lemma 8.4.3 thus indicates that the accuracy of the predictor $\tilde{\boldsymbol{w}}^k$ to a solution point $\boldsymbol{w}^*$ is measured by the quantity $\max\{(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \mid \boldsymbol{w} \in \Omega\}$. This is also the reason we search for a better iterate at the correct step (8.4.12) along the direction $-(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)$ to further reduce the proximity and to guarantee that the whole sequence is monotonically closer to the solution set. With this strict contraction property, it becomes standard to prove the convergence from the contraction method perspective in [2].

8.4.3 An Illustrative Example of Lemma 8.4.3

For better understanding the proposed algorithm (8.3.3) and seeing the assertion in Lemma 8.4.3 more specifically, we consider the special case of (8.1.1) with $m = 6$:

$$\min\Big\{\sum_{i=1}^{6}\theta_i(x_i)\ \Big|\ \sum_{i=1}^{6}A_i x_i = b,\ x_i \in X_i,\ i = 1, 2, \cdots, 6\Big\};$$

and regroup the variables as

$$\boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \boldsymbol{x}_3\end{pmatrix} \quad \text{with} \quad \boldsymbol{x}_1 = \begin{pmatrix}x_1\\ x_2\end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix}x_3\\ x_4\end{pmatrix} \quad \text{and} \quad \boldsymbol{x}_3 = \begin{pmatrix}x_5\\ x_6\end{pmatrix}. \tag{8.4.18a}$$

Therefore, $m_1 = m_2 = m_3 = 2$. Accordingly, we regroup

$$\mathcal{A}_1 = (A_1, A_2), \qquad \mathcal{A}_2 = (A_3, A_4), \qquad \mathcal{A}_3 = (A_5, A_6), \tag{8.4.18b}$$

and

$$\mathcal{X}_1 = X_1 \times X_2, \qquad \mathcal{X}_2 = X_3 \times X_4, \qquad \mathcal{X}_3 = X_5 \times X_6. \tag{8.4.18c}$$

The corresponding augmented Lagrangian function is

$$\mathcal{L}_\beta^6(x_1, x_2, x_3, x_4, x_5, x_6, \lambda) = \sum_{i=1}^{6}\theta_i(x_i) - \lambda^T\Big(\sum_{i=1}^{6}A_i x_i - b\Big) + \frac{\beta}{2}\Big\|\sum_{i=1}^{6}A_i x_i - b\Big\|^2. \tag{8.4.19}$$

With the given $\boldsymbol{w}^k = (x_1^k, x_2^k, x_3^k, x_4^k, x_5^k, x_6^k, \lambda^k)$, the prediction step (8.4.11) at the k-th iteration can be specified as

$$\begin{cases} \tilde{x}_1^k = \arg\min\left\{\mathcal{L}_\beta^6(x_1, x_2^k, x_3^k, x_4^k, x_5^k, x_6^k, \lambda^k) + \frac{\tau_1\beta}{2}\|A_1(x_1 - x_1^k)\|^2 \;\middle|\; x_1 \in X_1\right\}, \\ \tilde{x}_2^k = \arg\min\left\{\mathcal{L}_\beta^6(x_1^k, x_2, x_3^k, x_4^k, x_5^k, x_6^k, \lambda^k) + \frac{\tau_1\beta}{2}\|A_2(x_2 - x_2^k)\|^2 \;\middle|\; x_2 \in X_2\right\}; \end{cases} \tag{8.4.20a}$$

$$\begin{cases} \tilde{x}_3^k = \arg\min\left\{\mathcal{L}_\beta^6(\tilde{x}_1^k, \tilde{x}_2^k, x_3, x_4^k, x_5^k, x_6^k, \lambda^k) + \frac{\tau_2\beta}{2}\|A_3(x_3 - x_3^k)\|^2 \;\middle|\; x_3 \in X_3\right\}, \\ \tilde{x}_4^k = \arg\min\left\{\mathcal{L}_\beta^6(\tilde{x}_1^k, \tilde{x}_2^k, x_3^k, x_4, x_5^k, x_6^k, \lambda^k) + \frac{\tau_2\beta}{2}\|A_4(x_4 - x_4^k)\|^2 \middle| \; x_4 \in X_4\right\}; \end{cases} \tag{8.4.20b}$$

$$\begin{cases} \tilde{x}_5^k = \arg\min\left\{\mathcal{L}_\beta^6(\tilde{x}_1^k, \tilde{x}_2^k, \tilde{x}_3^k, \tilde{x}_4^k, x_5, x_6^k, \lambda^k) + \frac{\tau_3\beta}{2}\|A_5(x_5 - x_5^k)\|^2 \middle| \; x_5 \in X_5\right\}, \\ \tilde{x}_6^k = \arg\min\left\{\mathcal{L}_\beta^6(\tilde{x}_1^k, \tilde{x}_2^k, \tilde{x}_3^k, \tilde{x}_4^k, x_5^k, x_6, \lambda^k) + \frac{\tau_3\beta}{2}\|A_6(x_6 - x_6^k)\|^2 \middle| \; x_6 \in X_6\right\}; \end{cases} \tag{8.4.20c}$$

$$\tilde{\lambda}^k = \lambda^k - \beta\big(A_1\tilde{x}_1^k + A_2\tilde{x}_2^k + \textstyle\sum_{j=3}^6 A_j x_j^k - b\big). \tag{8.4.20d}$$

Using (8.4.19) and combining the notations in (8.4.18), we obtain

$$\begin{aligned} &\vartheta(\boldsymbol{x}_1) - \vartheta(\tilde{\boldsymbol{x}}_1^k) + (\boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1^k)^T\{-\mathcal{A}_1^T\tilde{\lambda}^k - \beta\mathcal{A}_1^T\mathcal{A}_1(\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k) \\ &\qquad + (\tau_1+1)\beta\mathrm{diag}(\mathcal{A}_1^T\mathcal{A}_1)(\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k)\} \geq 0, \ \forall\, \boldsymbol{x}_1 \in \mathcal{X}_1. \end{aligned} \tag{8.4.21}$$

$$\vartheta(\boldsymbol{x}_2) - \vartheta(\tilde{\boldsymbol{x}}_2^k) + (\boldsymbol{x}_2 - \tilde{\boldsymbol{x}}_2^k)^T\{-\mathcal{A}_2^T\tilde{\lambda}^k + (\tau_2+1)\beta\mathrm{diag}(\mathcal{A}_2^T\mathcal{A}_2)(\tilde{\boldsymbol{x}}_2^k - \boldsymbol{x}_2^k)\} \geq 0, \ \forall\, \boldsymbol{x}_2 \in \mathcal{X}_2. \tag{8.4.22}$$

$$\begin{aligned} &\vartheta(\boldsymbol{x}_3) - \vartheta(\tilde{\boldsymbol{x}}_3^k) + (\boldsymbol{x}_3 - \tilde{\boldsymbol{x}}_3^k)^T\{-\mathcal{A}_3^T\tilde{\lambda}^k \\ &\quad + \beta\mathcal{A}_3^T\mathcal{A}_2(\tilde{\boldsymbol{x}}_2^k - \boldsymbol{x}_2^k) + (\tau_3+1)\beta\mathrm{diag}(\mathcal{A}_3^T\mathcal{A}_3)(\tilde{\boldsymbol{x}}_3^k - \boldsymbol{x}_3^k)\} \geq 0, \ \forall\, \boldsymbol{x}_3 \in \mathcal{X}_3. \end{aligned} \tag{8.4.23}$$

Using the notations in (8.4.18), we rewrite (8.4.20d) as

$$\begin{aligned} \tilde{\lambda}^k \in \mathbb{R}^\ell, \quad (\lambda - \tilde{\lambda}^k)^T\Big\{&(\sum_{r=1}^3 \mathcal{A}_r\tilde{\boldsymbol{x}}_r^k - b) - \mathcal{A}_2(\tilde{\boldsymbol{x}}_2^k - \boldsymbol{x}_2^k) - \mathcal{A}_3(\tilde{\boldsymbol{x}}_3^k - \boldsymbol{x}_3^k) \\ &+ \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k)\Big\} \geq 0, \quad \forall\, \lambda \in \mathbb{R}^\ell. \end{aligned} \tag{8.4.24}$$

Combining (8.4.21), (8.4.22), (8.4.23), and (8.4.24) together, and using the VI (8.2.2), the predictor $\tilde{\boldsymbol{w}}^k \in \Omega$ satisfies (8.4.14) with the concrete matrix Q defined as

$$Q = \begin{pmatrix} (\tau_1+1)\beta\mathrm{diag}(\mathcal{A}_1^T\mathcal{A}_1) - \beta\mathcal{A}_1^T\mathcal{A}_1 & 0 & 0 & 0 \\ 0 & (\tau_2+1)\beta\mathrm{diag}(\mathcal{A}_2^T\mathcal{A}_2) & 0 & 0 \\ 0 & \beta\mathcal{A}_3^T\mathcal{A}_2 & (\tau_3+1)\beta\mathrm{diag}(\mathcal{A}_3^T\mathcal{A}_3) & 0 \\ 0 & -\mathcal{A}_2 & -\mathcal{A}_3 & \frac{1}{\beta}I \end{pmatrix}.$$

Therefore, for a given scenario of the abstract model (8.1.1) and when the regrouping strategy is determined, the matrix Q in (8.4.14) can be easily specified.

8.4.4 Convergence Proof

With the proved propositions, we are now ready to prove the convergence for the proposed algorithm (8.3.3). First of all, let us further analyze the term $(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)$ in the right-hand side of (8.4.14), which will help us show the strict contraction for the sequence $\{w^k\}$ generated by (8.3.3) with respect to the solution set Ω^*.

Theorem 8.4.4 *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the proposed algorithm (8.3.3). We have*

$$\begin{aligned}&\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k)\\ &\geq \frac{1}{2\alpha}\big(\|\boldsymbol{w} - \boldsymbol{w}^{k+1}\|_H^2 - \|\boldsymbol{w} - \boldsymbol{w}^k\|_H^2\big) + \frac{1}{2}\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2, \quad \forall \boldsymbol{w} \in \Omega\end{aligned} \tag{8.4.25}$$

Proof First, it follows from (8.4.8a) that $Q = HM$. We thus have

$$(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) = \frac{1}{\alpha}(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}).$$

Together with (8.4.14), this identity means

$$\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \geq \frac{1}{\alpha}(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}), \quad \forall \boldsymbol{w} \in \Omega. \tag{8.4.26}$$

Applying the identity

$$(a-b)^T H(c-d) = \frac{1}{2}\big(\|a-d\|_H^2 - \|a-c\|_H^2\big) + \frac{1}{2}\big(\|c-b\|_H^2 - \|d-b\|_H^2\big),$$

to the term $(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k - \boldsymbol{w}^{k+1})$ in the right-hand side of (8.4.26) with

$$a = \boldsymbol{w}, \quad b = \tilde{\boldsymbol{w}}^k, \quad c = \boldsymbol{w}^k, \quad \text{and} \quad d = \boldsymbol{w}^{k+1},$$

we thus obtain

$$\begin{aligned}(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T H(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) = &\frac{1}{2}\big(\|\boldsymbol{w} - \boldsymbol{w}^{k+1}\|_H^2 - \|\boldsymbol{w} - \boldsymbol{w}^k\|_H^2\big)\\ &+ \frac{1}{2}(\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^k\|_H^2).\end{aligned} \tag{8.4.27}$$

For the last group term of (8.4.27), we have

$$\begin{aligned}
&\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^k\|_H^2 \\
&= \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^k - \boldsymbol{w}^{k+1})\|_H^2 \\
&\overset{(8.4.8a)}{=} \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_H^2 - \|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - \alpha M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2 \\
&= 2\alpha(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T H M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - \alpha^2(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T M^T H M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \\
&= \alpha(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T [Q^T + Q - \alpha M^T H M](\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \\
&\overset{(8.4.8b)}{=} \alpha\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2. \qquad (8.4.28)
\end{aligned}$$

Substituting (8.4.27), (8.4.28) in (8.4.26), the assertion of this theorem is proved. □

Now, we are ready to show the strict contraction property of the sequence $\{\boldsymbol{w}^k\}$ generated by the proposed scheme (8.3.3).

Theorem 8.4.5 *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the proposed algorithm (8.3.3). Then we have*

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \le \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \alpha\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2, \quad \forall \boldsymbol{w}^* \in \Omega^*. \qquad (8.4.29)$$

Proof Setting $\boldsymbol{w} = \boldsymbol{w}^*$ in (8.4.25), we get

$$\begin{aligned}
\|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \ge \alpha\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2 + 2\alpha\{\vartheta(\tilde{\boldsymbol{x}}^k) - \vartheta(\boldsymbol{x}^*) \\
+ (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\tilde{\boldsymbol{w}}^k)\}.
\end{aligned}$$

Using the optimality of $\boldsymbol{w}^*$ and the monotonicity of $F(\boldsymbol{w})$, we have

$$\vartheta(\tilde{\boldsymbol{x}}^k) - \vartheta(\boldsymbol{x}^*) + (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\tilde{\boldsymbol{w}}^k) \ge \vartheta(\tilde{\boldsymbol{x}}^k) - \vartheta(\boldsymbol{x}^*) + (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\boldsymbol{w}^*) \ge 0,$$

and thus

$$\|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \ge \alpha\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G^2.$$

The assertion (8.4.29) follows directly. □

Finally, the convergence of $\{\boldsymbol{w}^k\}$ generated by the algorithm (8.3.3) can be proved easily. We summarize it in the following theorem.

Theorem 8.4.6 *The sequence $\{\boldsymbol{w}^k\}$ generated by the proposed algorithm (8.3.3) converges to a solution point of VI(Ω, F, θ).*

Proof First, according to (8.4.29), it holds that $\{\boldsymbol{w}^k\}$ is bounded and

$$\lim_{k\to\infty} \|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|_G = 0. \qquad (8.4.30)$$

Thus, $\{\boldsymbol{w}^k\}$ (and $\{\tilde{\boldsymbol{w}}^k\}$) has a cluster point $\boldsymbol{w}^\infty$. Using the continuity of ϑ and F and (8.4.30), then (8.4.14) becomes

$$\tilde{\boldsymbol{w}}^\infty \in \Omega, \quad \vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^\infty) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^\infty)^T F(\tilde{\boldsymbol{w}}^\infty) \geq 0, \quad \forall \boldsymbol{w} \in \Omega,$$

and thus $\tilde{\boldsymbol{w}}^\infty$ is a solution point of VI(Ω, F, θ). According to (8.4.29), the sequence $\{\boldsymbol{w}^k\}$ cannot have another cluster point and it converges to $\tilde{\boldsymbol{w}}^\infty$. The proof is complete. □

8.5 Convergence Rate

In this section, we establish the $O(1/T)$ worst-case convergence rates measured by the iteration complexity in both the ergodic and nonergodic senses for the new algorithm (8.3.3), where t denotes the iteration counter. Recall the prediction-correction algorithm (8.4.11)–(8.4.12) is a reformulation of (8.3.3).

8.5.1 Convergence Rate in the Ergodic Sense

We first establish a worst-case $O(1/T)$ convergence rate for the scheme (8.3.3) in the ergodic sense. The proof is inspired by our earlier work in [20] for the ADMM (8.1.8).

For this convergence rate analysis, we need to recall a characterization of the solution set Ω^*, which is described in the following theorem. Its proof can be found in [9] (Theorem 2.3.5) or [20] (Theorem 2.1).

Theorem 8.5.1 *The solution set of VI(Ω, F, θ) is convex and it can be characterized as*

$$\Omega^* = \bigcap_{\boldsymbol{w} \in \Omega} \big\{ \tilde{\boldsymbol{w}} \in \Omega : \big(\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}})\big) + (\boldsymbol{w} - \tilde{\boldsymbol{w}})^T F(\boldsymbol{w}) \geq 0 \big\}. \tag{8.5.1}$$

For given $\epsilon > 0$, $\tilde{\boldsymbol{w}} \in \Omega$ is called an ϵ-approximate solution of VI(Ω, F, θ) if it satisfies

$$\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}) + (\boldsymbol{w} - \tilde{\boldsymbol{w}})^T F(\boldsymbol{w}) \geq -\epsilon, \quad \forall\, \boldsymbol{w} \in \mathcal{D}_{(\tilde{\boldsymbol{w}})},$$

where

$$\mathcal{D}_{(\tilde{\boldsymbol{w}})} = \{\boldsymbol{w} \in \Omega \,|\, \|\boldsymbol{w} - \tilde{\boldsymbol{w}}\| \leq 1\}.$$

We refer the reader to [26] (Definition 1 therein) for the definition of an ϵ-approximate solution using the above set.

In the following, we shall show that based on T iterations generated by the proposed algorithm (8.3.3), we can find $\tilde{\boldsymbol{w}} \in \Omega$ such that

$$\tilde{\boldsymbol{w}} \in \Omega \quad \text{and} \quad \sup_{\boldsymbol{w} \in \mathcal{D}_{(\tilde{\boldsymbol{w}})}} \left\{ \vartheta(\tilde{\boldsymbol{x}}) - \vartheta(\boldsymbol{x}) + (\tilde{\boldsymbol{w}} - \boldsymbol{w})^T F(\boldsymbol{w}) \right\} \leq \epsilon, \tag{8.5.2}$$

with $\epsilon = O(1/T)$. Theorem 8.4.4 is also the basis for the upcoming analysis about the worst-case convergence rate.

Note that it follows from the monotonicity of F that

$$(\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\boldsymbol{w}) \geq (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k).$$

Substituting it into (8.4.25), we obtain

$$\vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\boldsymbol{w}) + \frac{1}{2\alpha}\|\boldsymbol{w} - \boldsymbol{w}^k\|_H^2 \geq \frac{1}{2\alpha}\|\boldsymbol{w} - \boldsymbol{w}^{k+1}\|_H^2, \ \forall \boldsymbol{w} \in \Omega. \tag{8.5.3}$$

Note that the above assertion holds whenever $G \succeq 0$.

Theorem 8.5.2 *Let $\{\boldsymbol{w}^k\}$ be generated by the proposed algorithm (8.3.3) and $\{\tilde{\boldsymbol{w}}^k\}$ be defined in (8.4.11). For any integer $T > 0$, let $\tilde{\boldsymbol{w}}_T$ be defined as*

$$\tilde{\boldsymbol{w}}_T = \frac{1}{T+1} \sum_{k=0}^{T} \tilde{\boldsymbol{w}}^k. \tag{8.5.4}$$

Then, we have $\tilde{\boldsymbol{w}}_T \in \Omega$ and

$$\vartheta(\tilde{\boldsymbol{x}}_T) - \vartheta(\boldsymbol{x}) + (\tilde{\boldsymbol{w}}_T - \boldsymbol{w})^T F(\boldsymbol{w}) \leq \frac{1}{2\alpha(T+1)}\|\boldsymbol{w} - \boldsymbol{w}^0\|_H^2, \quad \forall \boldsymbol{w} \in \Omega. \tag{8.5.5}$$

Proof First, it holds that $\tilde{\boldsymbol{w}}^k \in \Omega$ for all $k \geq 0$. Together with the convexity of $\mathcal{X}$ and $\mathbb{R}^\ell$, (8.5.4) implies that $\tilde{\boldsymbol{w}}_T \in \Omega$. Applying (8.5.3) to the cases with $k = 0, 1, \ldots, T$, and adding all the resulting inequalities together, we obtain

$$\begin{aligned} (T+1)\vartheta(\boldsymbol{x}) - \sum_{k=0}^{T} \vartheta(\tilde{\boldsymbol{x}}^k) + \Big((T+1)\boldsymbol{w} - \sum_{k=0}^{T} \tilde{\boldsymbol{w}}^k\Big)^T F(\boldsymbol{w}) \\ + \frac{1}{2\alpha}\|\boldsymbol{w} - \boldsymbol{w}^0\|_H^2 \geq 0, \quad \forall \boldsymbol{w} \in \Omega. \end{aligned}$$

Using the notation of $\tilde{\boldsymbol{w}}_T$, it can be written as

$$\frac{1}{T+1} \sum_{k=0}^{T} \vartheta(\tilde{\boldsymbol{x}}^k) - \vartheta(\boldsymbol{x}) + (\tilde{\boldsymbol{w}}_T - \boldsymbol{w})^T F(\boldsymbol{w}) \leq \frac{1}{2\alpha(T+1)}\|\boldsymbol{w} - \boldsymbol{w}^0\|_H^2, \quad \forall \boldsymbol{w} \in \Omega. \tag{8.5.6}$$

Since $\vartheta(\boldsymbol{x})$ is convex and

$$\tilde{\boldsymbol{x}}_T = \frac{1}{T+1}\sum_{k=0}^{T}\tilde{\boldsymbol{x}}^k,$$

we have that

$$\vartheta(\tilde{\boldsymbol{x}}_t) \le \frac{1}{T+1}\sum_{k=0}^{T}\vartheta(\tilde{\boldsymbol{x}}^k).$$

Substituting it in (8.5.6), the assertion of this theorem follows directly. □

Recall (8.5.2) and the conclusion (8.5.5) thus indicates that based on T iterations of the proposed algorithm (8.3.3), we can find an approximate solution of $\mathrm{VI}(\Omega, F, \theta)$ (i.e., $\tilde{\boldsymbol{w}}_T$ defined in (8.5.4)) with an accuracy of $O(1/T)$. That is, a worst-case $O(1/T)$ convergence rate is established for the proposed algorithm (8.3.3) in the ergodic sense.

8.5.2 Convergence Rate in a Nonergodic Sense

In this subsection, we establish a worst-case $O(1/T)$ convergence rate in a nonergodic sense for the proposed Algorithm 1. The proof is inspired by our earlier work in [21] for the ADMM (8.1.8). We first need to prove the following lemma.

Lemma 8.5.3 *For the sequences* $\{\boldsymbol{w}^k\}$ *and* $\{\tilde{\boldsymbol{w}}^k\}$ *generated by the proposed prediction-correction algorithm* (8.4.11)–(8.4.12)*, we have*

$$\begin{aligned}&(\boldsymbol{w}^k-\tilde{\boldsymbol{w}}^k)^T M^T H M\{(\boldsymbol{w}^k-\tilde{\boldsymbol{w}}^k)-(\boldsymbol{w}^{k+1}-\tilde{\boldsymbol{w}}^{k+1})\}\\ &\ge \frac{1}{2\alpha}\|(\boldsymbol{w}^k-\tilde{\boldsymbol{w}}^k)-(\boldsymbol{w}^{k+1}-\tilde{\boldsymbol{w}}^{k+1})\|^2_{(Q^T+Q)}.\end{aligned} \tag{8.5.7}$$

Proof First, set $\boldsymbol{w} = \tilde{\boldsymbol{w}}^{k+1}$ in (8.4.14), we have

$$\vartheta(\tilde{\boldsymbol{x}}^{k+1})-\vartheta(\tilde{\boldsymbol{x}}^k)+(\tilde{\boldsymbol{w}}^{k+1}-\tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \ge (\tilde{\boldsymbol{w}}^{k+1}-\tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k-\tilde{\boldsymbol{w}}^k). \tag{8.5.8}$$

Note that (8.4.14) is also true for $k := k+1$. Thus, we have

$$\vartheta(\boldsymbol{x})-\vartheta(\tilde{\boldsymbol{x}}^{k+1})+(\boldsymbol{w}-\tilde{\boldsymbol{w}}^{k+1})^T F(\tilde{\boldsymbol{w}}^{k+1}) \ge (\boldsymbol{w}-\tilde{\boldsymbol{w}}^{k+1})^T Q(\boldsymbol{w}^{k+1}-\tilde{\boldsymbol{w}}^{k+1}),\ \forall\, \boldsymbol{w}\in\Omega.$$

Setting $\boldsymbol{w} = \tilde{\boldsymbol{w}}^k$ in the above inequality, we obtain

$$\vartheta(\tilde{\boldsymbol{x}}^k) - \vartheta(\tilde{\boldsymbol{x}}^{k+1}) + (\tilde{\boldsymbol{w}}^k - \tilde{\boldsymbol{w}}^{k+1})^T F(\tilde{\boldsymbol{w}}^{k+1}) \geq (\tilde{\boldsymbol{w}}^k - \tilde{\boldsymbol{w}}^{k+1}) Q(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1}). \tag{8.5.9}$$

Adding (8.5.8) and (8.5.9), and using the monotonicity of F, we get

$$(\tilde{\boldsymbol{w}}^k - \tilde{\boldsymbol{w}}^{k+1})^T Q\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\} \geq 0. \tag{8.5.10}$$

Further, adding the term

$$\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\}^T Q\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\}$$

to both sides of (8.5.10), and using $\boldsymbol{w}^T Q \boldsymbol{w} = \frac{1}{2}\boldsymbol{w}^T(Q^T + Q)\boldsymbol{w}$, we obtain

$$(\boldsymbol{w}^k - \boldsymbol{w}^{k+1})^T Q\{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\} \geq \frac{1}{2}\|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|^2_{(Q^T+Q)}.$$

Substituting $(\boldsymbol{w}^k - \boldsymbol{w}^{k+1}) = \alpha M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)$ into the left-hand side of the last inequality and using $Q = HM$, we obtain (8.5.7) and the lemma is proved. □

In the following theorem, we prove a key inequality for establishing the worst-case $O(1/T)$ convergence rate in a nonergodic sense for the proposed algorithm (8.3.3).

Theorem 8.5.4 *For the sequences* $\{\boldsymbol{w}^k\}$ *and* $\{\tilde{\boldsymbol{w}}^k\}$ *generated by the proposed prediction-correction algorithm* (8.4.11)–(8.4.12)*, we have*

$$\|M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_H \leq \|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H, \quad \forall\, k > 0, \tag{8.5.11}$$

where M *and* H *are defined in (8.4.6) and (8.4.8a), respectively.*

Proof Setting $a = M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)$ and $b = M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})$ in the identity

$$\|a\|_H^2 - \|b\|_H^2 = 2a^T H(a - b) - \|a - b\|_H^2,$$

we obtain

$$\begin{aligned}
&\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2 - \|M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_H^2 \\
&\quad = 2(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T M^T H M[(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \\
&\qquad - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})] - \|M[(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})]\|_H^2.
\end{aligned}$$

Inserting (8.5.7) into the first term of the right-hand side of the last equality, we obtain

$$\begin{aligned}
&\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2 - \|M(\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_H^2 \\
&\quad \geq \frac{1}{\alpha}\|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_{(Q^T+Q)}^2 \\
&\qquad - \|M[(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})]\|_H^2 \\
&\quad \overset{(8.4.8b)}{=} \frac{1}{\alpha}\|(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\boldsymbol{w}^{k+1} - \tilde{\boldsymbol{w}}^{k+1})\|_G^2 \geq 0,
\end{aligned}$$

where the last inequality is because of the positive definiteness of the matrix $(Q^T + Q) - \alpha M^T H M \succeq 0$. The assertion (8.5.11) follows immediately. □

Note that it follows from $G \succ 0$ and Theorem 8.4.5 that there exists a constant $c_0 > 0$ such that

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \leq \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - c_0\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2, \quad \forall \boldsymbol{w}^* \in \Omega^*.$$

Since $\alpha M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) = (\boldsymbol{w}^k - \boldsymbol{w}^{k+1})$, we have a constant $c > 0$ such that

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \leq \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - c\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2, \quad \forall \boldsymbol{w}^* \in \Omega^*. \tag{8.5.12}$$

Now, with (8.5.12) and (8.5.11), we are ready to establish a worst-case $O(1/T)$ convergence rate in a nonergodic sense for the proposed algorithm (8.3.3).

Theorem 8.5.5 *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the proposed algorithm (8.3.3). For any integer $T > 0$, we have*

$$\|\boldsymbol{w}^T - \boldsymbol{w}^{T+1}\|_H^2 \leq \frac{1}{(T+1)c}\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|_H^2, \quad \forall \boldsymbol{w}^* \in \Omega^*, \tag{8.5.13}$$

with a constant $c > 0$.

Proof First, it follows from (8.5.12) that

$$\sum_{k=0}^{\infty} c\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2 \leq \|\boldsymbol{w}^0 - \boldsymbol{w}^*\|_H^2, \quad \forall\, \boldsymbol{w}^* \in \Omega^*. \tag{8.5.14}$$

According to Theorem 8.5.4, the sequence $\{\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2\}$ is monotonically nonincreasing. Therefore, we have

$$(T+1)\|\boldsymbol{w}^T - \boldsymbol{w}^{T+1}\|_H^2 \leq \sum_{k=0}^{T} \|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2. \tag{8.5.15}$$

The assertion (8.5.13) follows from (8.5.14) and (8.5.15) immediately. □

Let $d := \inf\{\|\boldsymbol{w}^0 - \boldsymbol{w}^*\|_H \mid \boldsymbol{w}^* \in \Omega^*\}$. Then, for any given $\epsilon > 0$, Theorem 8.5.5 shows that the proposed algorithm (8.3.3) needs at most $\lfloor d^2/c\epsilon \rfloor$ iterations to ensure that $\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2 \leq \epsilon$. Recall (8.4.26) and $\alpha > 0$ is a constant. It indicates that $\boldsymbol{w}^k$ is a solution point of VI(Ω, F, θ) if $\|\boldsymbol{w}^k - \boldsymbol{w}^{k+1}\|_H^2 = 0$. A worst-case $O(1/T)$ convergence rate in a nonergodic sense is thus established for the proposed algorithm (8.3.3).

8.6 Some Special Cases

In this section, we discuss some special cases when a regrouping strategy for (8.1.1) is specified and demonstrate the new algorithm in some more specific contexts. In particular, we show that the existing algorithms in [15, 17] can both be recovered by regrouping the variables in (8.1.1) appropriately. Therefore, the convergence rate results established in Sections 8.5.1 and 8.5.2 are applicable to the methods in [15, 17]. This is a by-product of this paper.

In such special cases, we always consider the first group as x_1, thus we have

$$\boldsymbol{x}_1 = x_1, \qquad \text{and} \qquad m_1 = 1. \tag{8.6.1}$$

In addition, we take

$$\tau_1 = m_1 - 1 = 0.$$

Because $\boldsymbol{x}_1 = x_1$ and $\tau_1 = 0$, the first subproblem in the prediction step (8.4.11a) becomes

$$\tilde{\boldsymbol{x}}_1^k = \arg\min\left\{\mathcal{L}_\beta^t[\boldsymbol{x}_1, \boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_t^k, \lambda^k] \;\middle|\; \boldsymbol{x}_1 \in \mathcal{X}_1\right\}.$$

For this case, the prediction step (8.4.11) can be specified as follows.

Prediction. For given $\boldsymbol{v}^k = (x_2^k, \ldots, x_m^k, \lambda^k) = (\boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_t^k, \lambda^k)$,

$$\begin{cases} \tilde{\boldsymbol{x}}_1^k = \arg\min\left\{\mathcal{L}_\beta^t[\boldsymbol{x}_1, \boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_t^k, \lambda^k] \;\middle|\; \boldsymbol{x}_1 \in \mathcal{X}_1\right\}; \\ \text{for } r = 2, \ldots t, \text{ do:} \\ \quad \text{for } j = 1, \ldots m_r, \text{ do:} \\ \qquad \tilde{x}_{r_j}^k = \arg\min\left\{\begin{array}{c}\mathcal{L}_\beta^t\big(\tilde{\boldsymbol{x}}_1^k, \ldots, \tilde{\boldsymbol{x}}_{r-1}^k, x_{r_1}^k, \ldots, x_{r_{j-1}}^k, x_{r_j}, x_{r_{j+1}}^k, \ldots, x_{r_{mr}}^k, \\ \boldsymbol{x}_{r+1}^k, \ldots, \boldsymbol{x}_t^k, \lambda^k\big) + \frac{\tau_r\beta}{2}\|A_{r_j}(x_{r_j} - x_{r_j}^k)\|^2\end{array}\;\middle|\; x_{r_j} \in X_{r_j}\right\}; \\ \quad \text{end.} \\ \text{end.} \end{cases} \tag{8.6.2a}$$

Additionally, we define

$$\tilde{\lambda}^k = \lambda^k - \beta\big(A_1\tilde{\boldsymbol{x}}_1^k + \sum_{r=2}^{t}\mathcal{A}_r\boldsymbol{x}_r^k - b\big). \tag{8.6.2b}$$

According to (8.6.2), because we choose $\boldsymbol{x}_1 = x_1$ and $\tau_1 = 0$, then $\boldsymbol{x}_1 = x_1$ is an intermediate variable and it is not needed in the iteration. In other words, to implement the proposed algorithm (8.3.3) with $\boldsymbol{x}_1 = x_1$, we only need $\boldsymbol{v}^k = (\boldsymbol{x}_2^k, \ldots, \boldsymbol{x}_t^k, \lambda^k)$. Moreover, note that $\beta(D_1 - \mathcal{A}_1^T\mathcal{A}_1)$ is the unique non-zero elements in the first row and first column of the matrix Q (see (8.4.1)). In this case,

$$D_1 = (\tau_1 + 1)\mathrm{diag}(\mathcal{A}_1^T\mathcal{A}_1) = \mathrm{diag}(\mathcal{A}_1^T\mathcal{A}_1)$$

and $(D_1 - \mathcal{A}_1^T\mathcal{A}_1)$ becomes the zero matrix. Accordingly, Lemma 8.4.3 is reduced to the following lemma (for convenience, we still use the same letters to denote the matrices).

Lemma 8.6.1 *Let $\tilde{\boldsymbol{x}}^k$ be generated by (8.6.2a) from the given vector $\boldsymbol{v}^k$ and $\tilde{\lambda}^k$ be defined by (8.6.2b). Then, the predictor $\tilde{\boldsymbol{w}}^k \in \Omega$ satisfies*

$$\tilde{\boldsymbol{w}}^k \in \Omega, \quad \vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \geq (\boldsymbol{v} - \tilde{\boldsymbol{v}}^k)^T Q(\boldsymbol{v}^k - \tilde{\boldsymbol{v}}^k), \quad \forall \boldsymbol{w} \in \Omega, \tag{8.6.3}$$

where Q is defined by

$$Q = \begin{pmatrix} \beta\mathcal{Q}_e & 0 \\ -\mathcal{A} & \frac{1}{\beta}I \end{pmatrix}, \tag{8.6.4}$$

$\mathcal{A}$ and $\mathcal{Q}_e$ are defined by (8.4.3) and (8.4.4), respectively.

Note that the matrix Q in (8.6.4) can be generated by cutting off the first row and column of the matrix Q given in (8.4.1). This is because the first block-wise

variable $\boldsymbol{x}_1$ is just an intermediate variable for the special case under our current consideration. Thus, the matrix Q originally given in (8.4.1) for the generic case of Algorithm 1 has all zeros in its first row and column for the special case where $\boldsymbol{x}_1 = x_1$; hence, we only consider (8.6.4) for this special case. Likewise, with the specific Q in (8.6.4), we can also define the corresponding matrix H and G as in (8.4.8a) and (8.4.8b), respectively. Moreover, as shown in Lemma 8.4.2, the positive definiteness of these two matrices is crucial for proving the convergence of Algorithm 1 for the special case where $\boldsymbol{x}_1 = x_1$.

Moreover, the correction step (8.4.12) in the generic setting can be specified as follows.

Correction. The new iterate $\boldsymbol{v}^{k+1}$ is given by

$$\boldsymbol{v}^{k+1} = \boldsymbol{v}^k - \alpha M(\boldsymbol{v}^k - \tilde{\boldsymbol{v}}^k), \tag{8.6.5a}$$

where

$$M = \begin{pmatrix} \mathcal{Q}_e^{-T}\mathcal{D}_e & 0 \\ -\beta\mathcal{A} & I \end{pmatrix}, \qquad \alpha \in \begin{cases} (0,1), \text{ if } & \tau_r \geq m_r - 1,\ r = 2, \ldots, t; \\ (0,1], \text{ if } & \tau_r > m_r - 1,\ r = 2, \ldots, t, \end{cases} \tag{8.6.5b}$$

and $\tilde{\boldsymbol{v}}^k$ is the related sub-vector of the predictor $\tilde{\boldsymbol{w}}^k$ generated by (8.6.2).

The matrices $\mathcal{Q}_e$, $\mathcal{D}_e$ in (8.6.5b) are defined in (8.4.4) and (8.4.5), respectively. It follows from (8.6.5b) that

$$M = \begin{pmatrix} \mathcal{Q}_e^{-T}\mathcal{D}_e & 0 \\ -\beta\mathcal{A} & I \end{pmatrix} \quad \text{and} \quad \mathcal{A} = (A_2, A_3, \ldots, A_m).$$

Also, because of (8.6.2b), we have

$$\lambda^{k+1} = \lambda^k - \alpha\beta\big(\textstyle\sum_{j=1}^m A_j\tilde{x}_j^k - b\big). \tag{8.6.6a}$$

In addition, the variables $x_2, \cdots, x_m$ are updated by the back substitution procedure:

$$\mathcal{D}_e^{-1}\mathcal{Q}_e^T \begin{pmatrix} \boldsymbol{x}_2^{k+1} - \boldsymbol{x}_2^k \\ \vdots \\ \boldsymbol{x}_t^{k+1} - \boldsymbol{x}_t^k \end{pmatrix} = \alpha \begin{pmatrix} \tilde{\boldsymbol{x}}_2^k - \boldsymbol{x}_2^k \\ \vdots \\ \tilde{\boldsymbol{x}}_t^k - \boldsymbol{x}_t^k \end{pmatrix}. \tag{8.6.6b}$$

In the following, we show that both the methods in [15, 17] are special cases of the proposed algorithm (8.3.3) with $\boldsymbol{x}_1 = x_1$.

8.6.1 The ADMM-GBS in [15]

Let us consider the special regrouping strategy with $\boldsymbol{x}_i = x_i$ for $i = 1, \cdots, m$ for (8.1.1). That is, each block of variables only consists of one variable. For this case, we have

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix}, \quad \text{where} \quad \boldsymbol{x}_i = x_i, \quad i = 1, \ldots, m. \tag{8.6.7}$$

Clearly, for this regrouping strategy, in the implementation of the proposed algorithm (8.3.3), we have

$$\tau_i = 0, \quad i = 1, 2, \ldots, m,$$

and thus the matrix $\mathcal{Q}_e$ (8.4.4) and $\mathcal{D}_e$ can be specified as

$$\mathcal{Q}_e = \begin{pmatrix} A_2^T A_2 & 0 & \cdots & 0 \\ A_3^T A_2 & A_3^T A_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ A_m^T A_2 & \cdots & A_m^T A_{m-1} & A_m^T A_m \end{pmatrix} \quad \text{and} \quad \mathcal{D}_e = \begin{pmatrix} A_2^T A_2 & 0 & \cdots & 0 \\ 0 & A_3^T A_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & A_m^T A_m \end{pmatrix},$$

respectively. According to (8.6.2), the prediction step (8.4.11) is reduced to

$$\begin{cases} \tilde{x}_1^k = \arg\min \left\{ \mathcal{L}_\beta^m(x_1, x_2^k, x_3^k, \ldots, x_m^k, \lambda^k) \mid x_1 \in X_1 \right\}; \\ \tilde{x}_2^k = \arg\min \left\{ \mathcal{L}_\beta^m(\tilde{x}_1^k, x_2, x_3^k, \ldots, x_m^k, \lambda^k) \mid x_2 \in X_2 \right\}; \\ \qquad \vdots \\ \tilde{x}_i^k = \arg\min \left\{ \mathcal{L}_\beta^m(\tilde{x}_1^k, \ldots, \tilde{x}_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \mid x_i \in X_i \right\}; \\ \qquad \vdots \\ \tilde{x}_m^k = \arg\min \left\{ \mathcal{L}_\beta^m(\tilde{x}_1^k, \ldots, \tilde{x}_{m-1}^k, x_m, \lambda^k) \mid x_m \in X_m \right\}, \end{cases} \tag{8.6.8a}$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta\big(A_1 \tilde{x}_1^k + \sum_{j=2}^m A_j x_j^k - b\big). \tag{8.6.8b}$$

The new iterate $\boldsymbol{v}^{k+1}$ is given by (8.6.5). Since $\tau_i = m_i - 1 = 0$, the step size $\alpha \in (0, 1)$.

If we denote the output (8.6.8a) by $\bar{x}_1^{k+1}, \bar{x}_2^{k+1}, \ldots, \bar{x}_m^{k+1}$, namely

$$\begin{cases} \bar{x}_1^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(x_1, x_2^k, x_3^k, \ldots, x_m^k, \lambda^k) \mid x_1 \in X_1\right\}; \\ \bar{x}_2^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(\bar{x}_1^{k+1}, x_2, x_3^k, \ldots, x_m^k, \lambda^k) \mid x_2 \in X_2\right\}; \\ \qquad\vdots \\ \bar{x}_i^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(\bar{x}_1^{k+1}, \ldots, \bar{x}_{i-1}^{k+1}, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \mid x_i \in X_i\right\}; \\ \qquad\vdots \\ \bar{x}_m^{k+1} = \arg\min\left\{\mathcal{L}_\beta^m(\bar{x}_1^{k+1}, \ldots, \bar{x}_{m-1}^{k+1}, x_m, \lambda^k) \mid x_m \in X_m\right\}, \end{cases} \tag{8.6.9a}$$

and set

$$\bar{\lambda}^{k+1} = \lambda^k - \beta\big(\sum_{j=1}^m A_j \bar{x}_j^{k+1} - b\big). \tag{8.6.9b}$$

The implementation of (8.6.6) becomes

$$\begin{cases} \mathcal{D}_e^{-1}\mathcal{Q}_e^T \begin{pmatrix} x_2^{k+1} - x_2^k \\ \vdots \\ x_m^{k+1} - x_m^k \end{pmatrix} = \alpha \begin{pmatrix} \bar{x}_2^{k+1} - x_2^k \\ \vdots \\ \bar{x}_m^{k+1} - x_m^k \end{pmatrix}, \\ \lambda^{k+1} - \lambda^k \quad = \quad \alpha(\bar{\lambda}^{k+1} - \lambda^k). \end{cases} \tag{8.6.10}$$

Note that for this special case, we have

$$\mathcal{D}_e^{-1}\mathcal{Q}_e^T = \begin{pmatrix} I_{n_2}(A_2^TA_2)^{-1}A_2^TA_3 & \cdots & (A_2^TA_2)^{-1}A_2^TA_m \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & (A_{m-1}^TA_{m-1})^{-1}A_{m-1}^TA_m \\ 0 & \cdots & 0 & I_{n_m} \end{pmatrix}.$$

It is just the left-upper part of the matrix P (see (8.1.11)). Thus, the method (8.6.9)–(8.6.10) reduces to the ADMM-GBS in [15].

8.6.2 *The Splitting Method in [17]*

Then, we consider another regrouping for (8.1.1):

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}, \quad \text{where} \quad \boldsymbol{x}_1 = x_1 \quad \text{and} \quad \boldsymbol{x}_2 = \begin{pmatrix} x_2 \\ \vdots \\ x_m \end{pmatrix}. \tag{8.6.11}$$

For this regrouping, we have

$$m_1 = 1 \qquad \text{and} \qquad m_2 = m - 1.$$

Besides (8.6.1), for the implementation of the new algorithm (8.3.3), we have

$$\tau_2 = \tau > m - 2 = m_2 - 1$$

and thus the matrix $\mathcal{Q}_e$ given in (8.4.4) is specified as

$$\mathcal{Q}_e = \mathcal{D}_e = \begin{pmatrix} (\tau+1)A_2^T A_2 & 0 & \cdots & 0 \\ 0 & (\tau+1)A_3^T A_3 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & (\tau+1)A_m^T A_m \end{pmatrix}. \tag{8.6.12}$$

Thus, according to (8.6.2), the prediction step (8.4.11) is reduced to

$$\begin{cases} \tilde{x}_1^k = \arg\min \left\{ \mathcal{L}_\beta^2(x_1, x_2^k, x_3^k, \ldots, x_m^k, \lambda^k) \mid x_1 \in X_1 \right\}; \\ \tilde{x}_i^k = \arg\min \left\{ \mathcal{L}_\beta^2(\tilde{x}_1^k, x_2^k \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \right. \\ \qquad\qquad \left. + \frac{\tau\beta}{2}\|A_i(x_i - x_i^k)\|^2 \;\middle|\; x_i \in X_i \right\}, \quad i = 2, \ldots, m, \end{cases} \tag{8.6.13a}$$

and

$$\tilde{\lambda}^k = \lambda^k - \beta\big(A_1\tilde{x}_1^k + \sum_{j=2}^{m} A_j x_j^k - b\big). \tag{8.6.13b}$$

Since $\tau_2 = \tau > m_2 - 1 = m - 2$, we take the step size $\alpha = 1$ in the correction step (8.6.5). The new iterate is given by

$$\boldsymbol{v}^{k+1} = \boldsymbol{v}^k - M(\boldsymbol{v}^k - \tilde{\boldsymbol{v}}^k).$$

Because $\boldsymbol{x}_2 = (x_2, \ldots, x_m)$, we have $\mathcal{Q}_e = \mathcal{D}_e$ (see (8.4.4) and (8.4.5)). Thus the matrix M in (8.6.5b) becomes

$$M = \begin{pmatrix} I & 0 \\ -\beta\mathcal{A} & I \end{pmatrix}.$$

Using $\mathcal{Q}_e = \mathcal{D}_e$ and $\alpha = 1$, the implementation of correction (8.6.5) is reduced to

$$\lambda^{k+1} = \lambda^k - \beta\big(\textstyle\sum_{j=1}^m A_j \tilde{x}_j^k - b\big), \tag{8.6.14a}$$

and

$$\begin{pmatrix} x_2^{k+1} \\ \vdots \\ x_m^{k+1} \end{pmatrix} = \begin{pmatrix} \tilde{x}_2^k \\ \vdots \\ \tilde{x}_m^k \end{pmatrix}. \tag{8.6.14b}$$

Therefore, (8.6.13) and (8.6.14) can be represented by

$$\begin{cases} x_1^{k+1} = \arg\min \big\{\mathcal{L}_\beta^2(x_1, x_2^k, x_3^k, \ldots, x_m^k, \lambda^k) \;\big|\; x_1 \in X_1\big\}; \\ x_i^{k+1} = \arg\min \big\{\mathcal{L}_\beta^2(x_1^{k+1}, x_2^k \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \\ \qquad\qquad + \frac{\tau\beta}{2}\|A_i(x_i - x_i^k)\|^2 \big|\; x_i \in X_i\big\}, i = 2, \ldots, m. \\ \lambda^{k+1} = \lambda^k - \beta\big(\sum_{j=1}^m A_j x_j^{k+1} - b\big). \end{cases} \tag{8.6.15}$$

To clearly see the relationship between (8.6.15) and the method in [17], let us summarize a conclusion in the following lemma.

Lemma 8.6.2 *Let the augmented Lagrangian function* $\mathcal{L}_\beta^m(x_1, \ldots, x_m, \lambda)$ *be defined in* (8.1.6)*. Then we have*

$$\begin{aligned} &\arg\min \Big\{\mathcal{L}_\beta^m(x_1^{k+1}, x_2^k \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) + \frac{\tau\beta}{2}\|A_i(x_i - x_i^k)\|^2 \big|\; x_i \in X_i\Big\} \\ &\quad = \arg\min \Big\{\theta_i(x_i) - (\lambda^{k+\frac{1}{2}})^T A_i x_i + \frac{(\tau+1)\beta}{2}\|A_i(x_i - x_i^k)\|^2 \big|\; x_i \in X_i\Big\}, \end{aligned} \tag{8.6.16}$$

where

$$\lambda^{k+\frac{1}{2}} = \lambda^k - \beta\big(A_1 x_1^{k+1} + \sum_{i=2}^m A_i x_i^k - b\big). \tag{8.6.17}$$

Proof Let us observe the x_i-subproblems in the left-hand side of (8.6.16). Notice that

$$\begin{aligned} &\mathcal{L}_\beta^m(x_1^{k+1}, x_2^k \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) \\ &= \theta_i(x_i) - \theta_i(x_i^k) + \textstyle\sum_{j=1}^m \theta_j(x_j^k) - (\lambda^k)^T[A_1 x_1^{k+1} + A_i x_i + \sum_{j=2, j\neq i}^m A_j x_j^k - b] \\ &\quad + \frac{\beta}{2}\|A_i(x_i - x_i^k) + A_1 x_1^{k+1} + \textstyle\sum_{j=2}^m A_j x_j^k - b\|^2. \end{aligned}$$

Ignoring some constant terms in the objective function of the minimization problem, we have

$$\begin{aligned}&\arg\min\{\mathcal{L}_\beta^m(x_1^{k+1}, x_2^k \ldots, x_{i-1}^k, x_i, x_{i+1}^k, \ldots, x_m^k, \lambda^k) + \tfrac{\tau\beta}{2}\|A_i(x_i - x_i^k)\|^2 | x_i \in X_i\}\\ &= \arg\min\left\{\begin{matrix}\theta_i(x_i) - (\lambda^k)^T A_i x_i + \frac{\beta}{2}\|A_i(x_i - x_i^k) + A_1 x_1^{k+1} + \sum_{j=2}^m A_j x_j^k - b\|^2\\ +\frac{\tau\beta}{2}\|A_i(x_i - x_i^k)\|^2\end{matrix}\middle| x_i \in X_i\right\}.\end{aligned}$$

Thus, the optimality condition of the x_i-subproblem is

$$\begin{aligned}x_i^{k+1} \in X_i,\ \theta_i(x_i) - \theta_i(x_i^{k+1}) + (x_i - x_i^{k+1})^T\big\{ &- A_i^T\lambda^k +\\ &+\beta A_i^T\big[A_i(x_i^{k+1} - x_i^k) + (A_1 x_1^{k+1} + \textstyle\sum_{j=2}^m A_j x_j^k - b)\big]\\ &+\tau\beta A_i^T A_i(x_i^{k+1} - x_i^k)\big\} \geq 0, \qquad \forall x_i \in X_i.\end{aligned}$$

It follows from (8.6.17) that

$$\lambda^k = \lambda^{k+\frac{1}{2}} + \beta\big(A_1 x_1^{k+1} + \textstyle\sum_{j=2}^m A_j x_j^k - b\big).$$

Substituting this identity into the last inequality, we obtain

$$\begin{aligned}&x_i^{k+1} \in X_i,\quad \theta_i(x_i) - \theta_i(x_i^{k+1})\\ &\quad + (x_i - x_i^{k+1})^T\big\{ - A_i^T\lambda^{k+\frac{1}{2}} + (1+\tau)\beta A_i^T A_i(x_i^{k+1} - x_i^k)\big\} \geq 0,\quad \forall\, x_i \in X_i.\end{aligned}$$

This is just the optimality condition of the x_i-subproblem of the right-hand side of (8.6.16). □

Thus, by setting $\mu = \tau + 1$, the scheme (8.6.15) can be represented as the following scheme:

$$\left\{\begin{aligned}&x_1^{k+1} = \arg\min\big\{\mathcal{L}_\beta^m(x_1, x_2^k, x_3^k, \ldots, x_m^k, \lambda^k)\ \big|\ x_1 \in X_1\big\};\\ &\lambda^{k+\frac{1}{2}} = \lambda^k - \beta\big(A_1 x_1^{k+1} + \textstyle\sum_{i=2}^m A_i x_i^k - b\big);\\ &x_i^{k+1} = \arg\min\Big\{\theta_i(x_i) - (\lambda^{k+\frac{1}{2}})^T A_i x_i + \tfrac{\mu\beta}{2}\|A_i(x_i - x_i^k)\|^2\big|\ x_i \in X_i\Big\},\\ &\qquad\qquad i = 2, \ldots, m.\\ &\lambda^{k+1} = \lambda^k - \beta\big(\textstyle\sum_{j=1}^m A_j x_j^{k+1} - b\big).\end{aligned}\right. \tag{8.6.18}$$

This is just the method proposed in [17]. Recall that $\mu > m - 1$ (since $\tau > m - 2$) is the condition to ensure the convergence of the method in [17].

8.7 A Refined Version of Algorithm 1 with Calculated Step Sizes

Instead of taking the constant step size α in the correction step (8.4.12), we can refine the algorithm (8.3.3) by choosing a calculated step size α_k at each iteration. Recall the role of the correction step in the algorithm (8.3.3) is to ensure the strict contraction property of the sequence (see (8.4.29)). The main idea of refining the algorithm (8.3.3) is that we can find a better step size, which is iteration-dependent, for each iteration such that the proximity to the solution set can be further reduced. For the case where calculating the step size is not computationally expensive, this refined version can accelerate the convergence and the number of iteration can be reduced, while the computation per iteration is just slightly increased. However, if the step size itself is computationally expensive, we still recommend the scheme (8.3.3) with a constant step size because for this case, the computation per iteration might be significantly increased; thus, the overall convergence might be slower even though the number of iteration might be smaller.

To see how to find a better step size to further reduce the proximity to the solution set, let us revisit Lemma 8.4.3. Indeed, setting $\boldsymbol{w} = \boldsymbol{w}^*$ in (8.4.14), we get

$$(\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \geq \vartheta(\tilde{\boldsymbol{x}}^k) - \vartheta(\boldsymbol{x}^*) - (\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T F(\tilde{\boldsymbol{w}}^k), \quad \forall \boldsymbol{w}^* \in \Omega^*.$$

Using the monotonicity of F and (8.2.4), it follows that

$$(\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^*)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \geq 0 \tag{8.7.1}$$

and consequently

$$(\boldsymbol{w}^k - \boldsymbol{w}^*)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) \geq (\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \quad \forall\, \boldsymbol{w}^* \in \Omega^*. \tag{8.7.2}$$

Because $Q = HM$, it follows that

$$\langle H(\boldsymbol{w}^k - \boldsymbol{w}^*), M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\rangle \geq \frac{1}{2}\|\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k\|^2_{(Q^T+Q)}, \quad \forall \boldsymbol{w}^* \in \Omega^*.$$

This means that $M(\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^k)$ is a descent direction of the distance function $\|\boldsymbol{w} - \boldsymbol{w}^*\|^2_H$ at the point $\boldsymbol{w}^k$, even if $\boldsymbol{w}^*$ is unknown. Along the direction $M(\tilde{\boldsymbol{w}}^k - \boldsymbol{w}^k)$ with well-chosen step size α, we can reduce the unknown distance function $\|\boldsymbol{w} - \boldsymbol{w}^*\|^2_H$. We define the step-size-dependent new iterate by

$$\boldsymbol{w}^{k+1}(\alpha) = \boldsymbol{w}^k - \alpha M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{8.7.3}$$

and

$$p(\alpha) = \|\boldsymbol{w}^k - \boldsymbol{w}^*\|^2_H - \|\boldsymbol{w}^{k+1}(\alpha) - \boldsymbol{w}^*\|^2_H. \tag{8.7.4}$$

By using $HM = Q$, we have

$$\begin{aligned} p(\alpha) &= \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|\boldsymbol{w}^{k+1}(\alpha) - \boldsymbol{w}^*\|_H^2 \\ &= \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \|(\boldsymbol{w}^k - \boldsymbol{w}^*) - \alpha M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2 \\ &= 2\alpha(\boldsymbol{w}^k - \boldsymbol{w}^*)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - \alpha^2 \|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2. \end{aligned}$$

Ideally we want to maximize the quadratic function $p(\alpha)$. However, it is impossible due to the lack of the unknown solution point $\boldsymbol{w}^*$. By using (8.7.2), we obtain

$$p(\alpha) \geq q(\alpha), \tag{8.7.5}$$

where

$$q(\alpha) = 2\alpha(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - \alpha^2 \|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2. \tag{8.7.6}$$

We thus turn to the second best choice: Maximizing the quadratic function $q(\alpha)$ which is a lower bound of $p(\alpha)$. This promotes us to take the value of α as

$$\alpha_k^* = \frac{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)}{\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2} = \frac{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)}{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T (M^T H M)(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)}. \tag{8.7.7}$$

We take $\alpha = \gamma \alpha_k^*$ with $\gamma \in (0, 2)$. According to (8.4.9), we have

$$Q^T + Q - M^T H M \succeq 0$$

and thus

$$\alpha_k^* \geq \frac{1}{2}. \tag{8.7.8}$$

Therefore, the iteration-dependent step size calculated by (8.7.7) is bounded away from 0.

Moreover, it is worth to mention that it follows from (8.4.10) that

$$M^T H M = \begin{pmatrix} \beta(D_1 - \mathcal{A}_1^T \mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{D}_e + \mathcal{A}^T \mathcal{A}) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{1}{\beta} I \end{pmatrix}.$$

Therefore, the denominator in (8.7.7) can be calculated directly based on the matrix defined above before implementing the Gaussian back substitution procedure and there is no need to calculate the inverse of any matrix for determining α_k.

So, the proposed algorithm (8.3.3) can be altered to a refined version where the constant step size α in (8.4.12b) is iteratively calculated by (8.7.7). The resulting

refined version differs from the proposed algorithm (8.3.3) only in its correction step as shown below.

Correction step: The new iterate $\boldsymbol{w}^{k+1}$ is given by

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - \alpha_k M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{8.7.9a}$$

where $\tilde{\boldsymbol{w}}^k$ is generated by the prediction step (8.4.11) and M is given by (8.4.6). The step size α_k is given by

$$\alpha_k = \gamma\alpha_k^*, \quad \gamma \in (0, 2) \quad \text{and} \quad \alpha_k^* = \frac{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)}{(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T (M^T H M)(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)}. \tag{8.7.9b}$$

Note that it follows from (8.7.6) and (8.7.7) that

$$\begin{aligned} q(\gamma\alpha_k^*) &= 2\gamma\alpha_k^*(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k) - (\gamma\alpha_k^*)^2 \|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2 \\ &= \gamma(2-\gamma)(\alpha_k^*)^2 \|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2. \end{aligned} \tag{8.7.10}$$

The following theorem shows the strict contraction property of the sequence generated by the refined algorithm with the iteratively calculated step size (8.7.7). Its proof is similar as Theorem 8.4.5 and thus omitted.

Theorem 8.7.1 *Let $\{\boldsymbol{w}^k\}$ be the sequence generated by the refined algorithm of (8.3.3) with the iteratively calculated step size (8.7.7). Then, it holds*

$$\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^*\|_H^2 \le \|\boldsymbol{w}^k - \boldsymbol{w}^*\|_H^2 - \frac{\gamma(2-\gamma)}{4}\|M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k)\|_H^2, \quad \forall\, \boldsymbol{w}^* \in \Omega^*. \tag{8.7.11}$$

Based on Theorem 8.7.1, the convergence and the convergence rates can all be established similar as the analysis in Sections 8.4 and 8.5. We omit them for succinctness.

8.8 A Linearized Splitting Block-Wise ADMM with Gaussian Back Substitution

As analyzed, the x_{r_j}-subproblems (see (8.3.1)) in the proposed splitting version of block-wise ADMM-GBS (8.3.3) are in form of (8.1.4) and we can further alleviate them by linearizing their quadratic terms if these subproblems are still too hard

for a particular application of the model (8.1.1). More specifically, recall the x_{r_j}-subproblem (8.3.1) in (8.3.3) and ignore some constant terms in its objective. Then, if its quadratic term is linearized, the resulting linearized subproblem becomes

$$\bar{x}_{r_j}^{k+1} = \arg\min \left\{ \begin{array}{c} \theta_{r_j}(x_{r_j}) - (\lambda^k)^T A_{r_j} x_{r_j} + (x_{r_j} - x_{r_j}^k)^T \\ \beta A_{r_j}^T \left(\sum_{s=1}^{r-1} \mathcal{A}_s \bar{\boldsymbol{x}}_s^{k+1} + \sum_{s=r}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b \right) + \frac{\nu_r \beta}{2} \|x_{r_j} - x_{r_j}^k\|^2, \end{array} \right\} \tag{8.8.1}$$

which is indeed in form of (8.1.3). Note that in (8.8.1), the constant $\nu_r > 0$ plays the role of controlling the proximity of the linearization, and it should be sufficiently large to ensure the accuracy of this linearized subproblem and finally the convergence. As well studied in the literature such as [20, 22, 31–33], we require

$$\nu_r > \rho(\mathcal{A}_r^T \mathcal{A}_r), \quad r = 1, \ldots, t, \tag{8.8.2}$$

where $\rho(\cdot)$ denotes the spectrum radius of a matrix.

Therefore, replacing the x_{r_j}-subproblems in (8.3.3) by their linearized counterparts given in (8.8.1), we can obtain a linearized version of the proposed splitting block-wise ADMM-GBS (8.3.3) whose x_{r_j}-subproblems are in form of (8.1.3).

8.8.1 Algorithm

For the algorithm in this section, we define the matrix

$$\mathcal{P}_L = \begin{pmatrix} I & 0 & 0 & 0 & 0 & 0 \\ 0 & I & \frac{1}{\nu_2}\mathcal{A}_2^T\mathcal{A}_3 & \cdots & \frac{1}{\nu_2}\mathcal{A}_2^T\mathcal{A}_t & 0 \\ 0 & 0 & \ddots & \ddots & \vdots & \vdots \\ & \vdots & \ddots & I & \frac{1}{\nu_{t-1}}\mathcal{A}_{t-1}^T\mathcal{A}_t & 0 \\ 0 & 0 & \cdots & 0 & I & 0 \\ 0 & 0 & \cdots & 0 & 0 & I_\ell \end{pmatrix}, \tag{8.8.3}$$

where $\mathcal{P}_L \in \mathbb{R}^{(n+\ell)\times(n+\ell)}$, and we summarize the linearized version of the scheme (8.3.3) as follows.

Algorithm 2: A linearized version of the splitting block-wise ADMM-GBS (8.3.3) for (8.1.1)

Initialization: Specify a regrouping for the model (8.1.1) with determined values of t and m_r for $r = 1, 2, \ldots, t$. Choose constants ν_r such that $\nu_r > \rho(\mathcal{A}_r^T \mathcal{A}_r)$ for $r = 1, \ldots, t$. Let $\mathcal{P}_L$ be defined in (8.8.3). Choose $\boldsymbol{w}^0 = (\boldsymbol{x}_1^0, \boldsymbol{x}_2^0, \ldots, \boldsymbol{x}_t^0, \lambda^0) \in \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_t \times \mathbb{R}^\ell$, for every $k \geq 0$,

$$
\begin{cases}
\text{for } r = 1, 2, \ldots t, \text{ do:} \\
\quad \text{for } j = 1, 2, \ldots, m_r, \text{ parallel do:} \\
\quad\quad \bar{x}_{r_j}^{k+1} \\
\quad\quad = \arg\min \left\{ \begin{array}{c} \theta_{r_j}(x_{r_j}) - (\lambda^k)^T A_{r_j} x_{r_j} + (x_{r_j} - x_{r_j}^k)^T \\ \beta A_{r_j}^T \left(\sum_{s=1}^{r-1} \mathcal{A}_s \bar{\boldsymbol{x}}_s^{k+1} + \sum_{s=r}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b \right) + \frac{\nu_r \beta}{2} \| x_{r_j} - x_{r_j}^k \|^2 \end{array} \middle| \, x_{r_j} \in X_{r_j} \right\}; \\
\quad \text{end.} \\
\text{end.} \\
\bar{\lambda}^{k+1} = \lambda^k - \beta \left(\sum_{r=1}^{t} \mathcal{A}_r \bar{\boldsymbol{x}}_r^{k+1} - b \right). \\
\\
\mathcal{P}_L(\boldsymbol{w}^{k+1} - \boldsymbol{w}^k) = (\bar{\boldsymbol{w}}^{k+1} - \boldsymbol{w}^k).
\end{cases}
\tag{8.8.4}
$$

Remark 8.8.1 Just like the scheme (8.3.3), with the block-wise upper triangular matrix $\mathcal{P}_L$ defined in (8.1.11), the entries of $\boldsymbol{v}^{k+1}$ can be updated in the order of $\lambda \to x_m \to \cdots x_2$ by the Gaussian back substitution procedure when implementing (8.8.4). Moreover, the matrix $\mathcal{P}_L$ does not require computing any inverse of matrix, not like the matrix $\mathcal{P}$ defined in (8.3.2). Therefore, it is an easier substitution procedure compared with the one in (8.3.3). Meanwhile, theoretically it is required to estimate $\rho(\mathcal{A}_r^T \mathcal{A}_r)$ for $r = 1, 2, \cdots, t$, which might not be easy. This is the cost of alleviating the difficulty levels of subproblems from (8.1.4) to (8.1.3) for (8.8.4). Finally, it is worth to mention that the requirements $\nu_r > \rho(\mathcal{A}_r^T \mathcal{A}_r)$ for $r = 1, 2, \cdots, t$ are sufficient conditions to ensure the convergence of the linearized version (8.8.4) and they represent conservative estimates on the parameters ν_r's. In implementation, usually we can choose smaller values for ν_r's which might not satisfy these sufficient conditions while can lead to better numerical performance.

Remark 8.8.2 In the scheme (8.8.4), we take the step size as 1 constantly in the Gaussian back substitution procedure. As in Section 8.7, we can analogously discuss how to choose an iteratively calculated step size for the Gaussian back substitution step in (8.8.4). For succinctness, we omit it and refer to [22] for some useful analysis.

8.8.2 Convergence Analysis

In this subsection, we prove two important results for the proposed linearized version (8.8.4). Based on them, the convergence analysis including both the global

convergence and the worst-case convergence rates can be established analogously as the analysis in Sections 8.4 and 8.5. As in Section 8.4, we need to rewrite the scheme (8.8.4) as a prediction-correction form for analysis. For this purpose, similarly as in Section 8.4.1, we first write the matrix Q as the block-wise form

$$Q = \begin{pmatrix} \beta(\nu_1 I - \mathcal{A}_1^T \mathcal{A}_1) & 0 & 0 \\ 0 & \beta \mathcal{Q}_e & 0 \\ 0 & -\mathcal{A} & \frac{1}{\beta} I \end{pmatrix}, \tag{8.8.5}$$

with $\mathcal{A}$ defined in (8.4.3) and

$$\mathcal{Q}_e = \begin{pmatrix} \nu_2 I & 0 & \cdots & 0 \\ \mathcal{A}_3^T \mathcal{A}_2 & \nu_3 I & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ \mathcal{A}_t^T \mathcal{A}_2 & \cdots & \mathcal{A}_t^T \mathcal{A}_{t-1} & \nu_t I \end{pmatrix}. \tag{8.8.6}$$

Moreover, we use $\mathcal{D}_e = \mathrm{diag}(\nu_2 I, \nu_3 I, \ldots, \nu_t I)$ to denote the diagonal part of $\mathcal{Q}_e$. Using (8.8.2), we have

$$\mathcal{Q}_e^T + \mathcal{Q}_e \succ \mathcal{D}_e + \mathcal{A}^T \mathcal{A}. \tag{8.8.7}$$

With these matrices, we can rewrite the scheme (8.8.4) as follows.

Prediction. For the given $\boldsymbol{w}^k = (x_1^k, x_2^k \ldots, x_m^k, \lambda^k) = (\boldsymbol{x}_1^k, \ldots, \boldsymbol{x}_t^k, \lambda^k)$, generate the predictor $\tilde{\boldsymbol{w}}^k = (\tilde{x}_1^k, \tilde{x}_2^k \ldots, \tilde{x}_m^k, \tilde{\lambda}^k) = (\tilde{\boldsymbol{x}}_1^k, \ldots, \tilde{\boldsymbol{x}}_t^k, \tilde{\lambda}^k)$ by the following steps:

$$\left\{ \begin{array}{l} \text{for } r = 1, 2, \ldots t, \text{ do:} \\ \quad \text{for } j = 1, \ldots m_r, \text{ parallel do:} \\ \quad\quad \tilde{x}_{r_j}^k = \arg\min \left\{ \begin{array}{c} \theta_{r_j}(x_{r_j}) - (\lambda^k)^T A_{r_j} x_{r_j} + (x_{r_j} - x_{r_j}^k)^T \\ \beta A_{r_j}^T \big(\sum_{s=1}^{r-1} \mathcal{A}_s \tilde{\boldsymbol{x}}_s^k + \sum_{s=r}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b \big) + \frac{\nu_r \beta}{2} \| x_{r_j} - x_{r_j}^k \|^2 \end{array} \middle| x_{r_j} \in X_{r_j} \right\}; \\ \quad \text{end.} \\ \text{end.} \end{array} \right. \tag{8.8.8a}$$

Additionally, we define

$$\tilde{\lambda}^k = \lambda^k - \beta \big(\mathcal{A}_1 \tilde{\boldsymbol{x}}_1^k + \textstyle\sum_{j=2}^{t} \mathcal{A}_j \boldsymbol{x}_j^k - b \big). \tag{8.8.8b}$$

Correction. The new iterate $\boldsymbol{w}^{k+1}$ is given by

$$\boldsymbol{w}^{k+1} = \boldsymbol{w}^k - M(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \tag{8.8.9a}$$

where $\tilde{\boldsymbol{w}}^k$ is the predictor generated by (8.8.8) and

$$M = \begin{pmatrix} I & 0 & 0 \\ 0 & \mathcal{Q}_e^{-T}\mathcal{D}_e & 0 \\ 0 & -\beta\mathcal{A} & I \end{pmatrix}. \tag{8.8.9b}$$

Note that the matrix M in (8.8.9b) is the same form as the matrix defined in (8.4.6). In the following, we prove a result similar as Lemma 8.4.3. This assertion enables us to discern the difference between the predictor $\tilde{\boldsymbol{w}}^k$ and a solution point $\boldsymbol{w}^*$.

Lemma 8.8.3 *Let $\tilde{\boldsymbol{x}}^k$ be generated by (8.8.8a) from the given vector $\boldsymbol{w}^k$ and $\tilde{\lambda}^k$ be defined by (8.8.8b). Then, the predictor $\tilde{\boldsymbol{w}}^k \in \Omega$ satisfies*

$$\tilde{\boldsymbol{w}}^k \in \Omega, \quad \vartheta(\boldsymbol{x}) - \vartheta(\tilde{\boldsymbol{x}}^k) + (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T F(\tilde{\boldsymbol{w}}^k) \geq (\boldsymbol{w} - \tilde{\boldsymbol{w}}^k)^T Q(\boldsymbol{w}^k - \tilde{\boldsymbol{w}}^k), \quad \forall \boldsymbol{w} \in \Omega, \tag{8.8.10}$$

where

$$Q = \begin{pmatrix} \beta(\nu_1 I - \mathcal{A}_1^T\mathcal{A}_1) & 0 & \cdots & \cdots & 0 & 0 \\ 0 & \beta\nu_2 I & \ddots & & \vdots & \vdots \\ 0 & \beta\mathcal{A}_3^T\mathcal{A}_2 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \beta\mathcal{A}_t^T\mathcal{A}_2 & \cdots & \beta\mathcal{A}_t^T\mathcal{A}_{t-1} & \beta\nu_t I & 0 \\ 0 & -\mathcal{A}_2 & \cdots & -\mathcal{A}_{t-1} & -\mathcal{A}_t & \frac{1}{\beta}I \end{pmatrix}. \tag{8.8.11}$$

Proof The optimality condition of the convex minimization problem (8.8.8a) is

$$\begin{aligned} \tilde{x}_{r_j}^k \in X_{r_j}, \quad \theta_{r_j}(x_{r_j}) - \theta_{r_j}(\tilde{x}_{r_j}^k) + (x_{r_j} - \tilde{x}_{r_j}^k)^T \big\{ & - A_{r_j}^T \lambda^k \\ & + \beta A_{r_j}^T \big[\textstyle\sum_{s=1}^{r-1} \mathcal{A}_s \tilde{\boldsymbol{x}}_s^k + \sum_{s=r}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b \big] \\ & + \nu_r \beta (\tilde{x}_{r_j}^k - x_{r_j}^k) \big\} \geq 0, \quad \forall x_{r_j} \in X_{r_j}. \end{aligned}$$

For $r = 2, \ldots, m$, using the definition of $\tilde{\lambda}^k$ (see (8.8.8b)), we have

$$\lambda^k = \tilde{\lambda}^k + \beta\big(\mathcal{A}_1 \tilde{\boldsymbol{x}}_1^k + \textstyle\sum_{s=2}^{t} \mathcal{A}_s \boldsymbol{x}_s^k - b\big).$$

Substituting it into the last inequality, we obtain

$$\tilde{x}_{r_j}^k \in X_{r_j}, \quad \theta_{r_j}(x_{r_j}) - \theta_{r_j}(\tilde{x}_{r_j}^k) + (x_{r_j} - \tilde{x}_{r_j}^k)^T \big\{ - A_{r_j}^T \tilde{\lambda}^k + \beta A_{r_j}^T \big[\sum_{s=2}^{r-1} \mathcal{A}_s(\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) \big] + \nu_r \beta (\tilde{x}_{r_j}^k - x_{r_j}^k) \big\} \geq 0, \quad \forall x_{r_j} \in X_{r_j}.$$

Applying this inequality for the cases of $j = 1, \ldots, m_r$, and summarizing the resulting inequalities, we get

$$\tilde{\boldsymbol{x}}_r^k \in \mathcal{X}_r, \quad \vartheta_r(\boldsymbol{x}_r) - \vartheta_r(\tilde{\boldsymbol{x}}_r^k) + (\boldsymbol{x}_r - \tilde{\boldsymbol{x}}_r^k)^T \big\{ - \mathcal{A}_r^T \tilde{\lambda}^k + \beta \mathcal{A}_r^T \big[\sum_{s=2}^{r-1} \mathcal{A}_s(\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) \big] + \nu_r \beta (\tilde{\boldsymbol{x}}_r^k - \boldsymbol{x}_r^k) \big\} \geq 0, \ \forall \boldsymbol{x}_r \in \mathcal{X}_r. \tag{8.8.12}$$

For $r = 1$, recall the optimality condition and the definition of $\tilde{\lambda}^k$. We obtain

$$\tilde{\boldsymbol{x}}_1^k \in \mathcal{X}_1, \quad \vartheta_1(\boldsymbol{x}_1) - \vartheta_1(\tilde{\boldsymbol{x}}_1^k) + (\boldsymbol{x}_1 - \tilde{\boldsymbol{x}}_1^k)^T \big\{ - \mathcal{A}_1^T \tilde{\lambda}^k + \beta(\nu_1 I - \mathcal{A}_1^T \mathcal{A}_1)(\tilde{\boldsymbol{x}}_1^k - \boldsymbol{x}_1^k) \big\} \geq 0, \forall \boldsymbol{x}_1 \in \mathcal{X}_1. \tag{8.8.13}$$

In addition, by using (8.8.8b), we have

$$(\sum_{r=1}^{t} \mathcal{A}_r \tilde{\boldsymbol{x}}_r^k - b) - \sum_{s=2}^{t} \mathcal{A}_s(\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k) = 0,$$

and it can be rewritten as

$$\tilde{\lambda}^k \in \mathbb{R}^\ell, \quad (\lambda - \tilde{\lambda}^k)^T \big\{ (\sum_{r=1}^{t} \mathcal{A}_r \tilde{\boldsymbol{x}}_r^k - b) - \sum_{s=2}^{t} \mathcal{A}_s(\tilde{\boldsymbol{x}}_s^k - \boldsymbol{x}_s^k) + \frac{1}{\beta}(\tilde{\lambda}^k - \lambda^k) \big\} \geq 0, \quad \forall \lambda \in \mathbb{R}^\ell. \tag{8.8.14}$$

Combining (8.8.13), (8.8.12) ($r = 2, \ldots, t$), and (8.8.14) together and using the notations $F(\boldsymbol{w})$, Q (see (8.2.2) and (8.8.11)), the assertion of this lemma is followed directly. □

Then, in the following lemma, we prove some assertions with respect to the matrices defined before.

Lemma 8.8.4 *For the matrices Q and M defined in (8.8.11) and (8.8.9b), respectively, let*

$$H := QM^{-1} \tag{8.8.15a}$$

and

$$G := Q^T + Q - M^T H M. \tag{8.8.15b}$$

Then, both the matrices H and G are symmetric and positive definite.

Proof First, we check the positive definiteness of the matrix H. For the matrix M defined in (8.8.9b), we have

$$M^{-1} = \begin{pmatrix} I & 0 & 0 \\ 0 & \mathcal{D}_e^{-1}\mathcal{Q}_e^T & 0 \\ 0 & \beta\mathcal{A}\mathcal{D}_e^{-1}\mathcal{Q}_e^T & I \end{pmatrix}.$$

Thus, according to the definition of the matrix H (see (8.8.15a)), we conclude that

$$H = QM^{-1} = \begin{pmatrix} \beta(\nu_1 I - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta\mathcal{Q}_e\mathcal{D}_e^{-1}\mathcal{Q}_e^T & 0 \\ 0 & 0 & \frac{1}{\beta}I \end{pmatrix}$$

is symmetric and positive definite.

Now, we turn to check the positive definiteness of the matrix G. Note that

$$Q^T + Q = \begin{pmatrix} 2\beta(\nu_1 I - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{Q}_e^T + \mathcal{Q}_e) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{2}{\beta}I \end{pmatrix} \tag{8.8.16}$$

and

$$M^T H M = Q^T M = \begin{pmatrix} \beta(\nu_1 I - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{D}_e + \mathcal{A}^T\mathcal{A}) & -\mathcal{A}^T \\ 0 & -\mathcal{A} & \frac{1}{\beta}I \end{pmatrix}. \tag{8.8.17}$$

Then, it follows from (8.8.16), (8.8.17), and (8.8.7)) that

$$\begin{aligned} G &= Q^T + Q - M^T H M \\ &= \begin{pmatrix} \beta(\nu_1 I - \mathcal{A}_1^T\mathcal{A}_1) & 0 & 0 \\ 0 & \beta(\mathcal{Q}_e^T + \mathcal{Q}_e) - \beta(\mathcal{D}_e + \mathcal{A}^T\mathcal{A}) & 0 \\ 0 & 0 & \frac{1}{\beta}I \end{pmatrix} \succ 0. \end{aligned}$$

The assertion of this lemma is proved. □

Based on Lemmas 8.8.3 and 8.8.4, and following the analysis in Sections 8.4.4 and 8.5, we can easily establish the convergence and worst-case convergence rate for the linearized version (8.8.4). We omit the detail for succinctness.

8.9 Numerical Experiments

In this section, we provide some numerical results to verify the convergence of the proposed Algorithm 1, and its refined versions with iteratively calculated step sizes and linearized subproblems, respectively. Three subsections are thus organized accordingly.

Recall that our emphasis is discussing the big-data scenario of (8.1.1) with a huge value of m, and we pay particular attention to the case where each function θ_i is simple in sense of that the subproblems (8.1.3) or (8.1.4) can be easily solved. Moreover, we want to test the affection of different grouping strategies in the block-wise reformulation (8.1.14), i.e., the difference in numerical performance for different values of the group number t. With these considerations, the basis pursuit problem is a good choice to generate various synthetic datasets in this desired setting and thus verify the theoretical assertions.

The basis pursuit (BP) problem can be mathematically modeled as

$$\min \ \|\boldsymbol{x}\|_1 \quad \text{s.t.} \quad A\boldsymbol{x} = b, \ \boldsymbol{x} \in \mathbb{R}^m, \tag{8.9.1}$$

where $A \in \mathbb{R}^{\ell \times m}$ with $\ell \ll m$ and $b \in \mathbb{R}^{\ell}$, see, e.g., [4, 5, 7]. This model captures a wide range of applications in areas such as signal processing, sparse optimization, variable selection, dimension reduction, and so on. Certainly, the model (8.9.1) is a special case of (8.1.1) with $\theta_i(x_i) = |x_i|$, $n_i = 1$ and $X_i = \mathbb{R}$ for $i = 1, \ldots, m$.

Our code was written by Matlab R2014b and all our experiments were performed on a desktop with Windows 7 system and 4 Intel(R)-i7 CPU processor (3.10 GHz) and 8.00 GB memory.

8.9.1 Convergence of Algorithm 1

We first verify the convergence of Algorithm 1 by the BP model (8.9.1). We generate a matrix $A \in \mathbb{R}^{\ell \times m}$ randomly using the standard Gaussian distribution; then a sparse vector $x \in \mathbb{R}^m$ using the standard Gaussian distribution with a 6% sparsity level (meaning 94% components of x being zero); and finally a vector $b \in \mathbb{R}^{\ell}$ by $b = Ax$. For completeness, we test both the cases of $\ell \ll m$ (the case with sparse-solution-seeking explanation) and $\ell = m$ (the case of a general ℓ_1 minimization model with linear constraints).

Then, we consider different grouping strategies for the model (8.1.1), i.e., choosing different values for t in (8.1.14) and regrouping all the functions and variables equally as t groups, each having $\frac{m}{t}$ components. For simplicity, we assume that t can be divided by m without remainder. Algorithm 1 with t groups is denoted by "ADMM-GBS-t." Note that when t is chosen as m, then the BP model (8.9.1) is considered as a m-block convex minimization model whose function is simply $|x_i|$ and variable is x_i for each block. It is easy to see that for "ADMM-GBS-t," each subproblem arising in Algorithm 1 can be computed easily by

$$\tilde{x}_{r_j}^k = \arg\min_{x_{r_j}} |x_{r_j}| + \frac{\beta}{2}\left\| A_{r_j}x_{r_j} + \sum_{j=1}^{r-1}\mathcal{A}_j\tilde{\boldsymbol{x}}_j^k + \sum_{j=r_1,j\neq r_j}^{r_{m_r}} A_j x_j^k + \sum_{j=r+1}^{t}\mathcal{A}_j\boldsymbol{x}_j^k - b - \frac{\lambda^k}{\beta}\right\|^2$$
$$+\frac{\tau_r\beta}{2}\left\|A_{r_j}(x_{r_j} - r_{r_j}^k)\right\|^2$$
$$= \textbf{shrinkage}\left(\frac{1}{A_{r_j}^T A_{r_j}}A_{r_j}^T\left[\boldsymbol{a}_{r_j} + \tau_r\boldsymbol{b}_{r_j}\right], \frac{1}{(\tau_r+1)\beta(A_{r_j}^T A_{r_j})}\right), \tag{8.9.2}$$

where

$$\boldsymbol{a}_{r_j} = b + \frac{\lambda^k}{\beta} - \sum_{j=1}^{r-1}\mathcal{A}_j\tilde{\boldsymbol{x}}_j^k - \sum_{j=r_1,j\neq r_j}^{r_{m_r}} A_j x_j^k - \sum_{j=r+1}^{t}\mathcal{A}_j\boldsymbol{x}_j^k,$$

$\boldsymbol{b}_{r_j} = A_{r_j}x_{r_j}^k$, and the operator **shrinkage** : $\mathbb{R}^\ell \times \mathbb{R} \to \mathbb{R}^\ell$ is defined componentwisely as

$$\textbf{shrinkage}(x,\tau) = \text{sign}(x).*\max(|x|-\tau, 0). \tag{8.9.3}$$

For comparison, we also apply the direct extension of ADMM (8.1.9) to this model because of its empirical efficiency even without provable convergence, and denote it by "ADMM-Direct." In our experiments, the penalty parameter β is set to be 0.01 for "ADMM-Direct" and all cases of "ADMM-GBS-t." For all cases of "ADMM-GBS-t," the parameter α is set to be 0.99 and $\tau_r := m/t - 1$. Before presenting the stopping criterion, we calculate a nearly optimal solution by implementing the code available on the website "http://stanford.edu/boyd/papers/admm/" and denote it by $\boldsymbol{x}_{optimal}$. For all the tested cases, the stopping criterion is set to be

$$\max\left\{\frac{\left\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\right\|}{\left\|\boldsymbol{x}^k\right\|}, \|A\boldsymbol{x}-b\|\right\} \le 10^{-4}, \quad \text{and} \quad \left\|\boldsymbol{x}^{k+1}\right\|_1 \le \left\|\boldsymbol{x}_{optimal}\right\|_1.$$

We report the numerical results in terms of number of iterations ("Iter"), objective function values ("Obj"), computing time in seconds ("Time"), the difference of $\boldsymbol{x}^{k+1}$ and $\boldsymbol{x}^k$ measured by $\|\boldsymbol{x}^{k+1}-\boldsymbol{x}^k\|_2/\left\|\boldsymbol{x}^k\right\|_2$ ("Error"), and the constraint violation of $\|A\boldsymbol{x}-b\|_2$ ("Constraint"). Note that for the case $t \le m$ where the $\boldsymbol{x}_r$-subproblem is further splitted into m/t subproblems in parallel in Algorithm 1, we only count the computing time of the subproblem with the longest computing time because we can only use a sequential computation to simulate the parallel computation for these further splitted subproblems.

From Tables 8.1 and 8.2, we see that all the tested cases of Algorithm 1 with different group numbers perform favorably with a convergence; thus, the convergence of Algorithm 1 is well verified. Moreover, it is shown that extreme values of t, meaning too small or too close to m, perform less efficiently than

Table 8.1 Comparison of Algorithm 1 and ADMM-Direct for (8.9.1) with $\ell = m$

Dimension	Method	Iter	Obj	Time	Error	Constraint
$\ell = m =$ 1000	ADMM-GBS-20	3537	162.178	5.187096	5.074078e−05	9.665843e−05
	ADMM-GBS-50	1944	162.178	3.900150	6.764180e−05	9.911781e−05
	ADMM-GBS-100	805	162.178	1.975037	8.561838e−05	9.101627e−05
	ADMM-GBS-200	426	162.178	1.785631	6.736172e−05	9.795935e−05
	ADMM-GBS-250	247	162.178	1.784326	7.288522e−05	9.086058e−05
	ADMM-GBS-500	272	162.178	2.621949	6.958933e−05	9.192190e−05
	ADMM-GBS-1000	289	162.178	5.112578	8.596721e−05	9.789457e−05
	ADMM-Direct	321	162.178	11.365211	6.601057e−05	9.894851e−05
$\ell = m =$ 5000	ADMM-GBS-20	16,405	768.328	492.757187	8.994438e−06	9.949918e−05
	ADMM-GBS-50	8759	768.328	290.486375	1.389909e−05	9.802019e−05
	ADMM-GBS-100	4317	768.328	186.969837	1.532337e−05	9.828724e−05
	ADMM-GBS-200	2084	768.328	123.789375	2.148270e−05	9.986633e−05
	ADMM-GBS-250	1059	768.328	101.603750	2.932857e−05	9.741456e−05
	ADMM-GBS-500	849	768.328	102.531844	3.641548e−05	9.525081e−05
	ADMM-GBS-1000	446	768.328	100.180726	7.609816e−05	9.191748e−05
	ADMM-GBS-1250	282	768.328	115.009375	8.196895e−05	9.305580e−05
	ADMM-GBS-2500	281	768.328	127.253906	8.850048e−05	9.065167e−05
	ADMM-GBS-5000	263	768.328	211.710938	7.253421e−05	9.617933e−05
	ADMM-Direct	327	768.328	234.125578	9.766126e−05	8.707594e−05
$\ell = m =$ 10000	ADMM-GBS-20	33,592	1608.342	3567.999828	3.530865e−06	9.998092e−05
	ADMM-GBS-50	18,050	1608.342	2001.936219	4.834372e−06	9.973279e−05
	ADMM-GBS-100	7836	1608.342	1008.372188	7.528407e−06	9.977020e−05
	ADMM-GBS-200	4192	1608.342	649.298438	8.119648e−05	9.879046e−05
	ADMM-GBS-250	2215	1608.342	461.073125	7.393204e−05	9.983498e−05
	ADMM-GBS-500	1741	1608.342	411.725000	2.153406e−05	9.712915e−05
	ADMM-GBS-1000	930	1608.342	356.833594	2.118703e−05	9.946718e−05
	ADMM-GBS-1250	552	1608.342	379.593750	5.508611e−05	8.041246e−05
	ADMM-GBS-2500	919	1608.342	813.880859	5.741985e−05	4.547159e−05
	ADMM-GBS-5000	523	1608.342	853.058594	7.560450e−05	6.447262e−05
	ADMM-GBS-10,000	341	1608.342	1011.812500	7.357012e−05	9.878123e−05
	ADMM-Direct	234	1608.342	849.640625	8.980877e−05	3.806441e−05

moderate values. For example, for the case where $\ell = 100$ and $m = 1000$, the choices of $t = 100, 200$ or 250 are much better than the extreme cases of $t = 20, 50, 1000$. As mentioned, this is because a larger t means a higher extent of splitting on the augmented Lagrangian function outside but easier subproblems with smaller proximal parameters and lower extent of parallelism inside, while the opposite for a smaller t. Thus, an appropriate value of t can balance the loss of accuracy to the augmented Lagrangian function outside and the solvability of the subproblems inside; extreme values of t cannot achieve this balance and thus lead to less favorable performance.

Table 8.2 Comparison of Algorithm 1 and ADMM-Direct for (8.9.1) with $\ell \ll m$

Dimension	Method	Iter	Obj	Time	Error	Constratints
$\ell = 100$, $m = 1000$	ADMM-GBS-20	24,712	73.112	19.723581	6.084257e−05	9.973273e−05
	ADMM-GBS-50	24,768	73.112	27.347500	7.910947e−05	9.915442e−05
	ADMM-GBS-100	8932	73.112	18.612819	7.650294e−05	9.830939e−05
	ADMM-GBS-200	6731	73.112	25.433852	8.547038e−05	9.688150e−05
	ADMM-GBS-250	3483	73.112	23.978025	8.159635e−05	9.781603e−05
	ADMM-GBS-500	4304	73.113	34.582416	7.826775e−05	9.961987e−05
	ADMM-GBS-1000	3462	73.113	57.752663	8.844657e−05	9.821201e−05
	ADMM-Direct	3316	73.113	45.125785	8.606919e−05	9.869295e−05
$\ell = 500$, $m = 5000$	ADMM-GBS-20	23,342	127.522	72.566731	7.132412e−05	9.992486e−05
	ADMM-GBS-50	19,221	127.522	70.582162	9.455009e−05	9.963828e−05
	ADMM-GBS-100	15,224	127.522	62.960977	7.305311e−05	9.886987e−05
	ADMM-GBS-200	9123	127.522	55.292716	6.437095e−05	9.978086e−05
	ADMM-GBS-250	4916	127.522	50.874821	8.685067e−05	9.949931e−05
	ADMM-GBS-500	4672	127.522	42.775000	7.349265e−05	9.955053e−05
	ADMM-GBS-1000	3696	127.522	50.490625	8.261306e−05	9.971985e−05
	ADMM-GBS-1250	3873	127.522	87.723192	8.530503e−05	9.774368e−05
	ADMM-GBS-2500	3326	127.522	94.406066	6.515167e−05	9.855248e−05
	ADMM-GBS-5000	7338	127.522	453.875000	7.647595e−05	9.729608e−05
	ADMM-Direct	3218	127.522	135.281250	8.525945e−05	9.852127e−05
$\ell = 1000$, $m = 10{,}000$	ADMM-GBS-20	50,000	184.680	496.394914	6.673783e−04	3.149458e−03
	ADMM-GBS-50	50,000	184.680	376.299071	6.437142e−04	1.058711e−03
	ADMM-GBS-100	25,636	184.660	182.541629	7.201667e−05	9.992378e−05
	ADMM-GBS-200	17,266	184.660	145.324531	7.371996e−05	9.989235e−05
	ADMM-GBS-250	11,274	184.660	120.511632	6.823286e−05	9.975338e-05
	ADMM-GBS-500	8813	184.660	105.782215	7.657899e−05	9.519138e−05
	ADMM-GBS-1000	6221	184.660	101.566047	8.601931e−05	9.782919e−05
	ADMM-GBS-1250	5532	184.660	182.300125	6.413814e−05	9.743909e−05
	ADMM-GBS-2500	4232	184.660	162.531762	7.124923e−05	9.879833e−05
	ADMM-GBS-5000	3621	184.660	256.775816	8.551503e−05	9.947986e−05
	ADMM-GBS-10,000	2206	184.660	293.561335	7.182545e−05	9.109035e−05
	ADMM-Direct	1406	184.660	215.156502	9.184770e−05	7.911372e−05

To further discern the numerical difference of different tested cases, in Figures 8.1 and 8.2, we choose the particular two cases, where $\ell = m = 5000$ and $(\ell = 500, m = 5000)$, to plot the evaluations of the objective function values and constraint violations with respect to iterations. For each row, the right-hand side figure is a zoom-in counterpart of the left-hand side one, to show the difference more clearly.

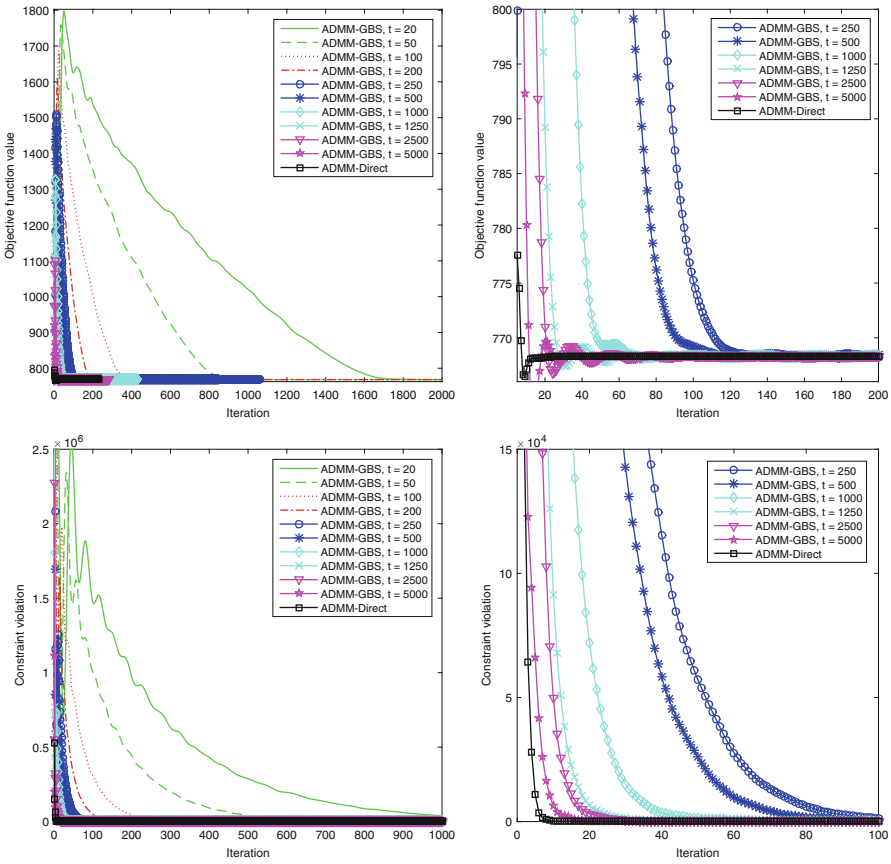


Fig. 8.1 Evolutions of Algorithm 1 and ADMM-Direct for (8.9.1) with $\ell = m = 5000$

8.9.2 Convergence of Algorithm 1 with Iteratively Calculated Step Sizes

In this subsection, we verify the convergence of the refined version of Algorithm 1 (denoted by "R-ADMM-GBS") whose step size is iteratively calculated by the scheme (8.7.9), and the acceleration with the refined step sizes. We still use the BP model (8.9.1). For succinctness, we only report the results for the cases where $\ell = m = 5000$ and $(\ell = 500, m = 5000)$, and report their results in Tables 8.3 and 8.4, respectively. To compare, the results of "ADMM-Direct" are also reported. The parameters are set exactly as those in the last subsection. The additional parameter γ in (8.7.9) is set as 1.8.

According to Tables 8.3 and 8.4, we see that the R-ADMM-GBS requires less iterations than the ADMM-GBS, it is faster for most of the tested cases. This is because the refined step size is chosen with the purpose of reducing the proximity

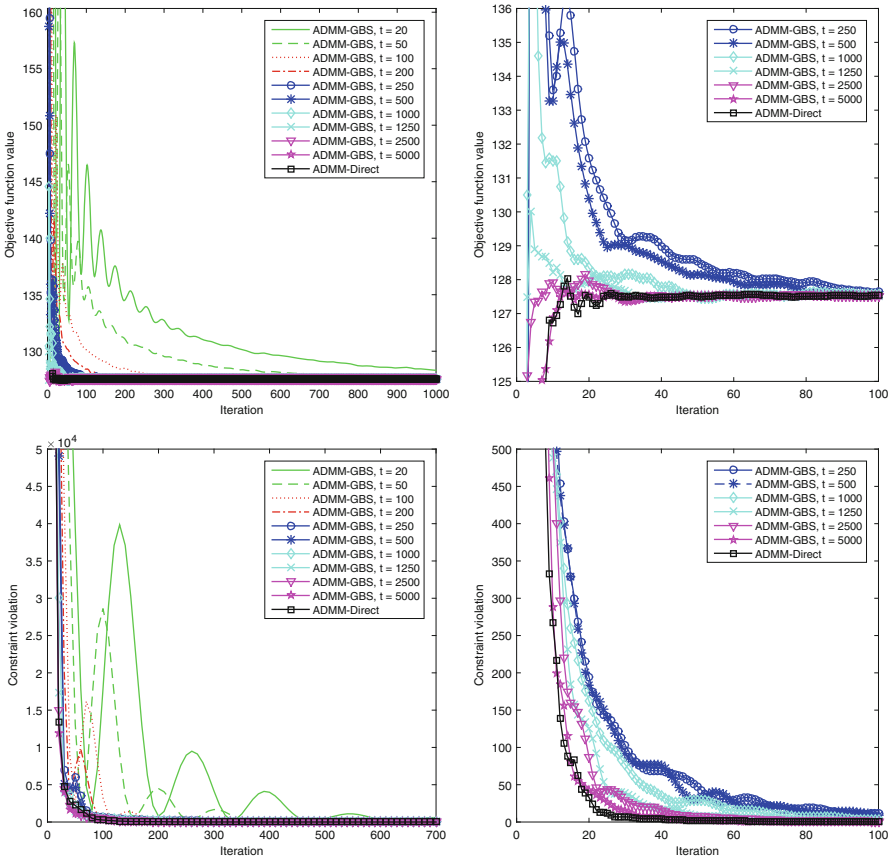


Fig. 8.2 Evolutions of Algorithm 1 and ADMM-Direct for (8.9.1) with $\ell = 500$, $m = 5000$

to the solution set, or improving the contraction of the sequence. Thus, the R-ADMM-GBS requires less iterations to achieve the same level of accuracy. But the calculation of the step sizes itself needs time; thus, the overall computing time of R-ADMM-GBS is not necessary to be less. We recommend the R-ADMM-GBS especially when the computation of step sizes is cheap. Moreover, the data in Tables 8.3 and 8.4 still reflects the fact that a non-extreme value of t can lead to better numerical results.

8.9.3 Convergence of Algorithm 2

Finally, we verify the convergence of the linearized version of the splitting block-wise ADMM-GBS (denoted by "Linearized-ADMM-GBS"), i.e., Algorithm 2,

Table 8.3 Comparison of R-ADMM-GBS and ADMM-Direct for (8.9.1) with $\ell = m$

Dimension	Block number t	Method	Iter	Obj	Time	Error	Constraint
$\ell = 5000$	100	ADMM-GBS	3912	768.329746	163.837969	1.532337e−05	9.828724e−05
$m = 5000$		R-ADMM-GBS	1723	768.329715	97.412262	3.468721e−05	9.691743e−05
	250	ADMM-GBS	1059	768.328434	101.603750	2.932857e−05	9.741456e−05
		R-ADMM-GBS	679	768.327641	82.039320	1.271965e−05	9.913485e−05
	500	ADMM-GBS	849	768.328581	93.844531	3.641548e−05	9.525081e−05
		R-ADMM-GBS	517	768.327719	76.453887	4.353316e−05	9.647576e−05
	1000	ADMM-GBS	431	768.329373	84.750000	5.816609e−05	9.748191e−05
		R-ADMM-GBS	406	768.328233	107.776044	4.419142e−05	9.866019e−05
	2500	ADMM-GBS	281	768.328477	127.253906	8.850048e−05	9.065167e−05
		R-ADMM-GBS	252	768.329172	153.755282	8.850048e−05	9.124582e−05
	5000	ADMM-GBS	263	768.328355	211.710938	9.253421e−05	7.617933e−05
		R-ADMM-GBS	247	768.328721	234.094445	5.778205e−05	9.983320e−05
		ADMM-Direct	231	768.329118	191.578125	9.660766e−05	4.594707e−05

Table 8.4 Comparison of R-ADMM-GBS and ADMM-Direct for (8.9.1) with $\ell \ll m$

Dimension	Block number t	Method	Iter	Obj	Time	Error	Constraint
$\ell = 500$	100	ADMM-GBS	13571	127.522848	54.077969	9.305311e−05	9.886987e−05
$m = 5000$		R-ADMM-GBS	9160	127.522771	46.502629	9.715218e−05	9.901157e−05
	250	ADMM-GBS	4601	127.522489	36.002500	9.065768e−05	9.949931e−05
		R-ADMM-GBS	3105	127.522532	31.332517	9.178332e−05	9.868076e−05
	500	ADMM-GBS	4212	127.522491	36.775000	9.654329e−05	9.955053e−05
		R-ADMM-GBS	3143	127.522447	36.270515	9.960288e−05	9.982264e−05
	1000	ADMM-GBS	3069	127.522656	46.490625	9.321660e−05	9.971985e−05
		R-ADMM-GBS	2771	127.522490	49.557654	9.763329e−05	9.885742e−05
	2500	ADMM-GBS	2462	127.522919	84.066406	9.515155e−05	9.855248e−05
		R-ADMM-GBS	1661	127.522238	76.505248	9.438861e−05	9.910689e−05
	5000	ADMM-GB	6783	127.522237	453.875000	9.576549e−05	9.729608e−05
		R-ADMM-GBS	3278	127.522179	185.701152	9.741995e−05	9.832210e−05
		ADMM-Direct	2292	127.522284	135.281250	9.555249e−05	9.852127e−05

whose subproblems are in form of (8.1.4) and a linearization is used to make the subproblems easier.

We still use the BP model (8.9.1) and regroup it as an equally grouped t-block reformulation in form of (8.1.14), where $t > 3$. We compare the proposed Algorithm 2 with the block-wise versions of both the direct extension of ADMM (8.1.18) and the ADMM-GBS (8.1.19). Note that for this grouped case of (8.9.1), the subproblems arising in (8.1.18) and (8.1.19) are all in dimensionality of $\frac{m}{t}$ as

$$\tilde{x}_{r_j}^{k+1} = \arg\min_{x_{r_j}} \|x_{r_j}\|_1 + \frac{\beta}{2}\left\|A_{r_j}x_{r_j} + \sum_{i=1}^{r-1}\mathcal{A}_i\tilde{\boldsymbol{x}}_i^{k+1} + \sum_{i=1,i\neq j}^{m_r} A_i x_i^k + \sum_{i=r+1}^{t}\mathcal{A}_i\boldsymbol{x}_i^k - \frac{\lambda^k}{\beta}\right\|^2 + \frac{\tau_r\beta}{2}\left\|A_{r_j}(x_{r_j} - x_{r_j}^k)\right\|^2. \tag{8.9.4}$$

In general, $A_{r_j} \in \mathbb{R}^{\ell\times\frac{m}{t}}$ is in generally not an identity matrix. Thus, the subproblem (8.9.4) generally has no closed-form solution and must be solved iteratively. In our experiments, we apply the FISTA in [1] to solve these subproblems iteratively, and denote the combination of FISTA with (8.1.18) and (8.1.19) by "ADMM-Direct-FISTA" and "ADMM-GBS-FISTA," respectively. Recall that when Algorithm 2 is implemented to this particular regrouped case of (8.9.1), the subproblems (8.9.4) are solved by linearizing their quadratic terms, not further splitting in parallel as Algorithm 1. Accordingly, a subproblem of Algorithm 2 has the closed-form solution given by

$$\begin{aligned}\tilde{x}_{r_j}^{k+1} &= \arg\min_{x_{r_j}} \|x_{r_j}\|_1 + \frac{\nu_r\beta}{2}\left\|x_{r_j} - x_{r_j}^k + \frac{1}{\nu_r\beta}\left(\sum_{i=1}^{r-1}\mathcal{A}_i\tilde{\boldsymbol{x}}_i^{k+1} + \sum_{i=r}^{t}\mathcal{A}_i\boldsymbol{x}_i^k - \frac{\lambda^k}{\beta}\right)\right\|^2 \\ &= \mathbf{shrinkage}\left(x_{r_j}^k - \frac{1}{\nu_r\beta}\left(\sum_{i=1}^{r-1}\mathcal{A}_i\tilde{\boldsymbol{x}}_i^{k+1} + \sum_{i=r}^{t}\mathcal{A}_i\boldsymbol{x}_i^k - \frac{\lambda^k}{\beta}\right), \frac{1}{\nu_r\beta}\right),\end{aligned} \tag{8.9.5}$$

where the operator **shrinkage** is also defined in (8.9.3).

For all the algorithms, the parameter β is set as 0.01 for all methods. For Algorithm 2, ν_r is set as $1.01 \times \rho(\mathcal{A}_r^T\mathcal{A}_r)$. For ADMM-GBS, $\alpha = 0.99$ and $\tau_r = m/t - 1$. The stopping criterion for all the algorithms is set as

$$\max\left\{\frac{\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|}{\|\boldsymbol{x}^k\|}, \|A\boldsymbol{x} - b\|\right\} \leq \epsilon. \tag{8.9.6}$$

To implement FISTA to solve the subproblems for ADMM-Direct-FISTA and ADMM-GBS-FISTA, we calculate the maximal eigenvalues for the matrices $A_{r_j}^T A_{r_j}$ and estimate the Lipschitz constants for the objective function in (8.9.4). For succinctness, we only report the results for (8.9.1) with $\ell = m = 1000$ and $t = 100$. We choose different levels of tolerance in (8.9.6) as $\epsilon = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$; and for each case of ϵ in (8.9.6), we test different levels of tolerance internally

Table 8.5 Comparison of Algorithm 2 with ADMM-GBS-FISTA and ADMM-Direct-FISTA for (8.9.1) with $\ell = m = 1000$

Method	ϵ			
	10^{-3}	10^{-4}	10^{-5}	10^{-6}
ADMM-GBS-FISTA($100 \cdot \epsilon$)	3.6962	4.6703	9.2634	17.6497
ADMM-GBS-FISTA($10 \cdot \epsilon$)	3.0765	5.5917	12.3913	24.7783
ADMM-GBS-FISTA(ϵ)	3.3162	7.1126	15.2145	32.4819
ADMM-GBS-FISTA($0.1 \cdot \epsilon$)	4.7231	9.3329	22.1796	47.0625
ADMM-Direct-FISTA($100 \cdot \epsilon$)	3.1294	4.1179	6.2247	12.3386
ADMM-Direct-FISTA($10 \cdot \epsilon$)	2.8815	4.6812	7.7629	19.3764
ADMM-Direct-FISTA(ϵ)	3.0047	5.8647	10.0718	25.2881
ADMM-Direct-FISTA($0.1 \cdot \epsilon$)	4.0562	7.2977	18.6328	33.7715
Linearized-ADMM-GBS	**1.2754**	**2.8862**	**4.3315**	**10.6922**

Bold values indicate best result

for FISTA as $100 \cdot \epsilon$, $10 \cdot \epsilon$, ϵ, and $0.1 \cdot \epsilon$. The computing times in seconds are reported in Table 8.5, in which the numbers in parentheses are the tolerance levels for implementing FISTA for a given ϵ in (8.9.6).

The efficiency of Algorithm 2 is supported by the results reported in Table 8.5. Indeed, the advantage of Algorithm 2 is because of the fact that when the quadratic term in (8.9.4) is linearized, then the resulting subproblem has a closed-form solution. This specific linearization technique thus can fully take advantage of the remaining $\|\cdot\|_1$ function, and it is generally better than a generic strategy of solving it iteratively. It is also worthwhile to mention that for solving the internal subproblems (8.9.4) iteratively, the tolerance generally should be less accurate than that for the outside iterations and it is not beneficial to pursue too accurate solutions for the subproblems, especially at the first phase of the iteration process. Finally, in Figure 8.3, we plot the evolutions of objective function values and constraint violations with respect to the first 500 iterations for Linearized-ADMM-GBS, and some cases of ADMM-GBS-FISTA and ADMM-Direct-FISTA.

8.10 Conclusions

In this paper, we discuss how to develop an algorithm for the separable multiple-block convex minimization models with linear constraints and an objective function which is in the sum of m functions without coupled variables. We focus on the big-data scenario with a huge m, to which the existing splitting schemes in the literature seem not to be directly applicable. With the assumption that the variables and functions are regrouped as more than two blocks, we investigate how to apply the alternating direction method of multiplier with a Gaussian back substitution (ADMM-GBS) in [15] to the regrouped model which is still in a multiple-block form. The resulting block-wise ADMM-GBS, however, may involve hard subproblems. To yield solvable easier subproblems, we suggest embedding

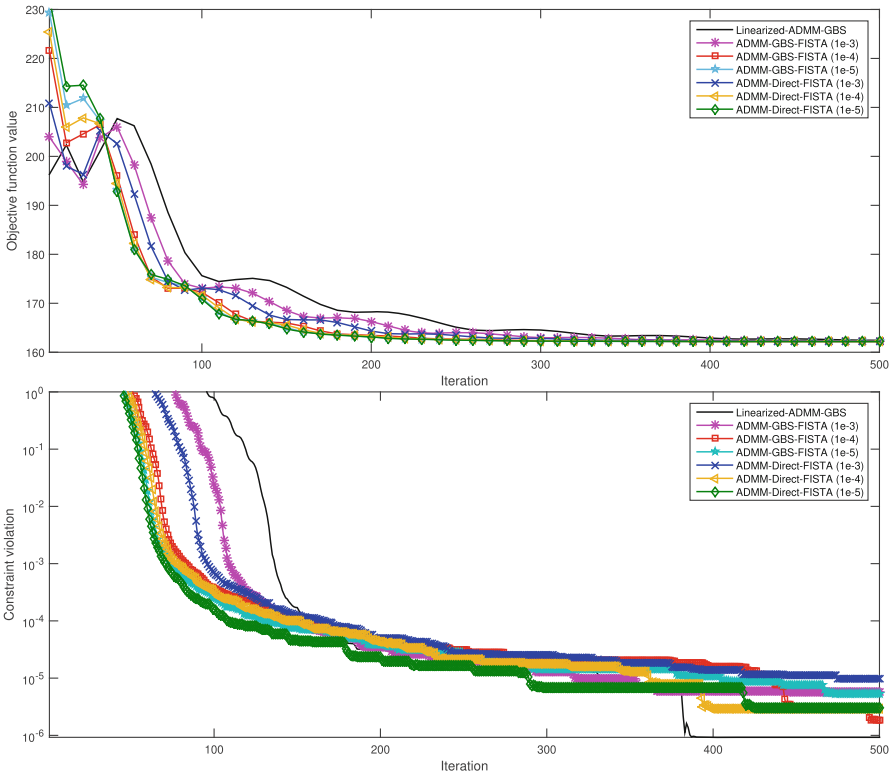


Fig. 8.3 Evolutions of linearized-ADMM-GBS, ADMM-GBS-FISTA, and ADMM-Direct, for (8.9.1) with $\ell = m = 1000$

a parallel computation into the block-wise ADMM-GBS, and consequently propose a splitting version of the block-wise ADMM-GBS which is suitable for a distributed-centralized computing system. The global convergence and the worst-case convergence rates measured by the iteration complexity in both the ergodic and nonergodic senses are established for the new algorithm. Moreover, the new algorithm turns to include some existing schemes as special cases; thus, a by-product of this paper is that the convergence rates for these existing schemes are also established. We also discuss how to refine the new scheme by choosing an iteratively calculated step size and further alleviating the resulting subproblems. Thus, two advanced versions with refined step sizes and linearized subproblems are proposed, respectively.

We verify the convergence of the proposed algorithms and the importance of an appropriate grouping strategy by the basis pursuit problem in several settings. Particularly, we show that the grouping strategy adopted in the block-wise reformulation (8.1.14), or more concretely, how to determine an appropriate value for the group number t, is very important because different strategies may result in

different numerical performance. For the synthetic datasets tested in our numerical experiments, there is no obvious difference in different functions and variables. We thus group them "blindly" with the only consideration of function and variable numbers. For a particular application of (8.1.14), based on some known information or features (e.g., by a learning process), we may group the functions and variables more smartly so that the subproblems could be easier while the group number may be smaller. We would emphasize that how to group the variables and functions smartly or even optimally for a particular application really depends on the particular structure and features of a given application itself. In this paper, we just provide the methodology and theoretical analysis to guarantee the convergence for the most general setting in form of (8.1.1).

The proposed scheme is a basic scheme which can easily inspire specific algorithms when concrete applications of the abstract model under consideration are specified. For example, as mentioned, we can consider further linearizing the subproblems such that each subproblem is of the difficulty level of estimating a function's proximal operator. Also, in addition to the Gaussian back substitution, other correction steps in the literature (e.g., [12, 13, 16]) can be used. In [23], we focused on the case where the model (8.1.1) is regrouped as two groups and thus a block-wise version of the original ADMM (8.1.8) is applied. In this paper, we consider the case where the model (8.1.1) is regrouped as at least three groups and thus the direct extension of ADMM (8.1.9) is not necessarily convergent. Because of the significant difference between the two- and three-block cases in ADMM-oriented schemes (see [6]), we regard this paper complementary to the most recent one [23] for using block-wise ADMM-based schemes for the multiple-block separable convex minimization model (8.1.1).

Acknowledgements The author "Xiaoling Fu" was supported by the Fundamental Research Funds for the Central Universities 2242019K40168 and partly supported by Natural Science Foundation of Jiangsu Province Grant BK20181258. The author "Bingsheng He" was supported by the NSFC Grant 11871029 and 11471156. The author "Xiangfeng Wang" was supported by the NSFC Grant 61672231, 11871279 and 11971090. The author "Xiaoming Yuan" was supported by the General Research Fund from Hong Kong Research Grants Council: 12313516.

References

1. A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imag. Sci., 2 (2009), pp. 183–202.
2. E. Blum and W. Oettli, Mathematische Optimierung. Grundlagen und Verfahren. Ökonometrie und Unternehmensforschung, Springer-Verlag, Berlin-Heidelberg-New York, 1975.
3. S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foun. Trends Mach. Learn., 3 (2010), pp. 1–122.
4. E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inform. Theory, 51 (2004), pp. 4203–4215.

5. E. J. Candès and T. Tao, *Reflections on compressed sensing*, IEEE Inform. Theory Soc. News., 58(4) (2008), pp. 14–17.
6. C. H. Chen, B. S. He, Y. Y. Ye and X. M. Yuan, *The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent*, Math. Program., Ser. A, 155(2016), pp. 57–79.
7. S. S. Chen, D. Donoho and M. A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Rev., 43(1) (2006), pp. 129–159.
8. J. Eckstein and W. Yao, *Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results*, Pacific J. Optim., 11(4) (2015), pp. 619–644.
9. F. Facchinei and J. S. Pang, Finite-Dimensional Variational Inequalities and Complementarity problems, Volume I, Springer Series in Operations Research, Springer-Verlag, 2003.
10. R. Glowinski, *On alternating direction methods of multipliers: A historical perspective*, Modeling, Simulation and Optimization for Science and Technology, W. Fitzgibbon, Y.A. Kuznetsov, P. Neittaanmäki and O. Pironneau, eds., Computational Methods in Applied Sciences, 34 (2014).
11. R. Glowinski and A. Marrocco, *Approximation par éléments finis d'ordre un et résolution par pénalisation-dualité d'une classe de problèmes non linéaires*, R.A.I.R.O., R2 (1975), pp. 41–76.
12. D. R. Han, X. M. Yuan and W. X. Zhang, *An augmented-Lagrangian-based parallel splitting method for separable convex programming with applications to image processing*, Math. Comput., 83 (2014), pp. 2263–2291.
13. B. S. He, L. S. Hou and X. M. Yuan, *On full Jacobian decomposition of the augmented Lagrangian method for separable convex programming*, SIAM J. Optim., 25(4) (2015), pp. 2274–2312.
14. B. S. He, H. Liu, J. Lu, and X. M. Yuan, *Application of the strictly contractive Peaceman-Rachford splitting method to multi-block convex programming*, in *Operator Splitting Methods and Applications*, edited by R. Glowinski, S. Osher and W. Yin, Springer, 2016.
15. B. S. He, M. Tao and X. M. Yuan, *Alternating direction method with Gaussian back substitution for separable convex programming*, SIAM J. Optim., 22 (2012), pp. 313–340.
16. B. S. He, M. Tao and X. M. Yuan, *Convergence rate and iteration complexity on the alternating direction method of multipliers with a substitution procedure for separable convex programming*, Math. Oper. Res., 42 (3) (2017), pp. 662–691.
17. B. S. He, M. Tao and X. M. Yuan, *A splitting method for separable convex programming*, IMA J. Numer. Anal., 35(2015), pp. 394–426.
18. B. S. He, H. K. Xu and X. M. Yuan, *On the proximal Jacobian decomposition of ALM for multiple-block separable convex minimization problems and its relationship to ADMM*, J. Sci. Comput., 66 (2016), 1204–1217.
19. B. S. He, M. H. Xu and X. M. Yuan, *Block-wise ADMM with a relaxation factor for multiple-block convex programming*, Journal of the Operational Research Society of China, 6(4) (2018), pp. 485–505.
20. B. S. He and X. M. Yuan, *On the $O(1/n)$ convergence rate of the alternating direction method*, SIAM J. Numer. Anal., 50 (2012), pp. 700–709.
21. B. S. He and X. M. Yuan, *On nonergodic convergence rate of Douglas-Rachford alternating direction method of multipliers*, Numer. Math., 130 (3)(2015), pp. 567–577.
22. B. S. He and X. M. Yuan, *Linearized alternating direction method with Gaussian back substitution for separable convex programming*, Numer. Alge., Cont. and Opt., 3(2)(2013), pp. 247–260.
23. B. S. He and X. M. Yuan, *Block-wise alternating direction method of multipliers for multiple-block convex programming and beyond*, SMAI J. Comp. Math., 1(2015), pp. 145–174.
24. M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appli., 4(1969), pp. 303–320.
25. M. Hong and Z. Q. Luo, *On the linear convergence of the alternating direction method of multipliers*, Math. Program., 162(1–2)(2017), pp. 165–199.

26. Y. E. Nesterov, *Gradient methods for minimizing composite objective function*, Math. Prog., Ser. B, 140 (2013), pp. 125–161.
27. G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Analy. Applic. 72 (1979), pp. 383–390.
28. Y. G. Peng, A. Ganesh, J. Wright, W. L. Xu and Y. Ma, *Robust alignment by sparse and low-rank decomposition for linearly correlated images*, IEEE Tran. Pattern Anal. Mach. Intel., 34 (2012), pp. 2233–2246.
29. M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, In Optimization edited by R. Fletcher, pp. 283–298, Academic Press, New York, 1969.
30. M. Tao and X. M. Yuan, *Recovering low-rank and sparse components of matrices from incomplete and noisy observations*, SIAM J. Optim., 21 (2011), pp. 57–81.
31. X. F. Wang and X. M. Yuan, *The linearized alternating direction method for Dantzig Selector*, SIAM J. Sci. Comput., 34 (5) (2012), pp. A2792 - A2811.
32. J. F. Yang and X. M. Yuan, Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization, Math. Comput., 82 (281) (2013), pp. 301–329.
33. X. Q. Zhang, M. Burger and S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*. J. Sci. Comput., 46 (2011), pp. 20–46.

Chapter 9
Variable Metric Algorithms Driven by Averaged Operators

Lilian E. Glaudin

Abstract The convergence of a new general variable metric algorithm based on compositions of averaged operators is established. Applications to monotone operator splitting are presented.

Keywords Averaged operator · Composite algorithm · Convex optimization · Fixed point iteration · Monotone operator splitting · Primal-dual algorithm · Variable metric

AMS 2010 Subject Classification 47H05, 49M27, 49M29, 90C25

9.1 Introduction

Iterations of averaged nonexpansive operators provide a synthetic framework for the analysis of many algorithms in nonlinear analysis, e.g., [3, 4, 7, 9, 18]. We establish the convergence of a new general variable metric algorithm based on compositions of averaged operators. These results are applied to the analysis of the convergence of a new forward-backward algorithm for solving the inclusion

$$0 \in Ax + Bx, \tag{9.1}$$

where A and B are maximally monotone operators on a real Hilbert space. The theory of monotone operators is used in many applied mathematical fields, including optimization [10], partial differential equations and evolution inclusions [5, 21, 23], signal processing [13, 17], and statistics and machine learning [12, 19, 20]. In recent years, variants of the forward-backward algorithm with variable metric have been proposed in [15, 16, 22, 24], as well as variants involving overrelaxations [18].

L. E. Glaudin (✉)
Laboratoire Jacques-Louis Lions, Sorbonne Université, Paris, France
e-mail: glaudin@ljll.math.upmc.fr

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_9

The goal of the present paper is to unify these two approaches in the general context of iterations of compositions of averaged operators. In turn, this provides new methods to solve the problems studied in [1, 4, 6, 8, 9, 14].

The paper is organized as follows: Section 9.2 presents the background and notation. We establish the proof of the convergence of the general algorithm in Section 9.3. Special cases are provided in Section 9.4. Finally, by recasting these results in certain product spaces, we present and solve a general monotone inclusion in Section 9.5.

9.2 Notation and Background

Throughout this paper, $\mathcal{H}$, $\mathcal{G}$, and $(\mathcal{G}_i)_{1\leqslant i\leqslant m}$ are real Hilbert spaces. We use $\langle\cdot\mid\cdot\rangle$ to denote the scalar product of a Hilbert space and $\|\cdot\|$ for the associated norm. Weak and strong convergence are, respectively, denoted by $\rightharpoonup$ and $\to$. We denote by $\mathcal{B}(\mathcal{H},\mathcal{G})$ the space of bounded linear operators from $\mathcal{H}$ to $\mathcal{G}$, and set $\mathcal{B}(\mathcal{H})=\mathcal{B}(\mathcal{H},\mathcal{H})$ and $\mathcal{S}(\mathcal{H})=\big\{L\in\mathcal{B}(\mathcal{H})\;\big|\;L=L^*\big\}$, where L^* denotes the adjoint of L, and Id denotes the identity operator. The Loewner partial ordering on $\mathcal{S}(\mathcal{H})$ is defined by

$$(\forall U\in\mathcal{S}(\mathcal{H}))(\forall V\in\mathcal{S}(\mathcal{H}))\quad U\succcurlyeq V\quad\Leftrightarrow\quad(\forall x\in\mathcal{H})\quad\langle Ux\mid x\rangle\geqslant\langle Vx\mid x\rangle. \tag{9.2}$$

Let $\alpha\in\left]0,+\infty\right[$. We set

$$\mathcal{P}_\alpha(\mathcal{H})=\big\{U\in\mathcal{S}(\mathcal{H})\;\big|\;U\succcurlyeq\alpha\,\mathrm{Id}\big\}, \tag{9.3}$$

and we denote by $\sqrt{U}$ the square root of $U\in\mathcal{P}_\alpha(\mathcal{H})$. Moreover, for every $U\in\mathcal{P}_\alpha(\mathcal{H})$, we define a scalar product and a norm by

$$(\forall x\in\mathcal{H})(\forall y\in\mathcal{H})\quad\langle x\mid y\rangle_U=\langle Ux\mid y\rangle\quad\text{and}\quad\|x\|_U=\sqrt{\langle Ux\mid x\rangle}, \tag{9.4}$$

and we denote this Hilbert space by $(\mathcal{H},U)$. Let $A\colon\mathcal{H}\to 2^{\mathcal{H}}$ be a set-valued operator. We denote by $\operatorname{dom}A=\big\{x\in\mathcal{H}\;\big|\;Ax\neq\varnothing\big\}$ the domain of A, by $\operatorname{gra}A=\big\{(x,u)\in\mathcal{H}\times\mathcal{H}\;\big|\;u\in Ax\big\}$ the graph of A, by $\operatorname{ran}A=\big\{u\in\mathcal{H}\;\big|\;(\exists\,x\in\mathcal{H})\;u\in Ax\big\}$ the range of A, by $\operatorname{zer}A=\big\{x\in\mathcal{H}\;\big|\;0\in Ax\big\}$ the set of zeros of A, and by A^{-1} the inverse of A which is the operator with graph $\big\{(u,x)\in\mathcal{H}\times\mathcal{H}\;\big|\;u\in Ax\big\}$. The resolvent of A is $J_A=(\mathrm{Id}+A)^{-1}$. Moreover, A is monotone if

$$(\forall(x,u)\in\operatorname{gra}A)(\forall(y,v)\in\operatorname{gra}A)\quad\langle x-y\mid u-v\rangle\geqslant 0, \tag{9.5}$$

and maximally monotone if there exists no monotone operator $B\colon\mathcal{H}\to 2^{\mathcal{H}}$ such that $\operatorname{gra}A\subset\operatorname{gra}B\neq\operatorname{gra}A$. The parallel sum of $A\colon\mathcal{H}\to 2^{\mathcal{H}}$ and $B\colon\mathcal{H}\to 2^{\mathcal{H}}$ is

$$A\,\square\,B=(A^{-1}+B^{-1})^{-1}. \tag{9.6}$$

An operator $B\colon \mathcal{H} \to 2^{\mathcal{H}}$ is cocoercive with constant $\beta \in]0, +\infty[$ if

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \quad \langle x - y \mid Bx - By\rangle \geqslant \beta\|Bx - By\|^2. \tag{9.7}$$

Let C be a nonempty subset of $\mathcal{H}$. The interior of C is $\operatorname{int}C$. Finally, the set of summable sequences in $[0, +\infty[$ is denoted by $\ell^1_+(\mathbb{N})$.

Definition 9.1 Let $\mu \in]0, +\infty[$, let $U \in \mathcal{P}_\mu(\mathcal{H})$, let $\alpha \in]0, 1]$, and let $T\colon \mathcal{H} \to \mathcal{H}$ be an operator. Then T is an α-averaged operator on $(\mathcal{H}, U)$ if

$$(\forall x \in \mathcal{H})(\forall y \in \mathcal{H}) \quad \|Tx - Ty\|_U^2 \leqslant \|x - y\|_U^2 - \frac{1-\alpha}{\alpha}\|Tx - x\|_U^2. \tag{9.8}$$

If $\alpha = 1$, T is nonexpansive on $(\mathcal{H}, U)$.

Lemma 9.1 ([4, Proposition 4.46]) *Let $m \geqslant 1$ be an integer. For every $i \in \{1, \ldots, m\}$, let $T_i\colon \mathcal{H} \to \mathcal{H}$ be averaged. Then $T_1 \cdots T_m$ is averaged.*

Lemma 9.2 ([4, Proposition 4.35]) *Let $\mu \in]0, +\infty[$, let $U \in \mathcal{P}_\mu(\mathcal{H})$, let $\alpha \in]0, 1]$, and let T be an α-averaged operator on $(\mathcal{H}, U)$. Then the operator $R = (1 - 1/\alpha)\operatorname{Id} + (1/\alpha)T$ is nonexpansive on $(\mathcal{H}, U)$.*

Lemma 9.3 ([4, Lemma 5.31]) *Let $(\alpha_n)_{n\in\mathbb{N}}$ and $(\beta_n)_{n\in\mathbb{N}}$ be sequences in $[0, +\infty[$, let $(\eta_n)_{n\in\mathbb{N}}$ and $(\varepsilon_n)_{n\in\mathbb{N}}$ be sequences in $\in \ell^1_+(\mathbb{N})$ such that*

$$(\forall n \in \mathbb{N}) \quad \alpha_{n+1} \leqslant (1 + \eta_n)\alpha_n - \beta_n + \varepsilon_n. \tag{9.9}$$

Then $(\beta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$.

Lemma 9.4 ([15, Proposition 4.1]) *Let $\alpha \in]0, +\infty[$, let $(W_n)_{n\in\mathbb{N}}$ be in $\mathcal{P}_\alpha(\mathcal{H})$, let C be a nonempty subset of $\mathcal{H}$, and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ such that*

$$\big(\exists\, (\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})\big)\big(\forall z \in C\big)\big(\exists\, (\varepsilon_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})\big)(\forall n \in \mathbb{N})$$
$$\|x_{n+1} - z\|^2_{W_{n+1}} \leqslant (1 + \eta_n)\|x_n - z\|^2_{W_n} + \varepsilon_n. \tag{9.10}$$

Then $(x_n)_{n\in\mathbb{N}}$ is bounded and, for every $z \in C$, $(\|x_n - z\|_{W_n})_{n\in\mathbb{N}}$ converges.

Proposition 9.1 ([15, Theorem 3.3]) *Let $\alpha \in]0, +\infty[$, and let $(W_n)_{n\in\mathbb{N}}$ and W be operators in $\mathcal{P}_\alpha(\mathcal{H})$ such that $W_n \to W$ pointwise, as is the case when*

$$\sup_{n\in\mathbb{N}} \|W_n\| < +\infty \quad \text{and} \quad (\exists\, (\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N}))(\forall n \in \mathbb{N}) \quad (1 + \eta_n)W_n \succcurlyeq W_{n+1}. \tag{9.11}$$

Let C be a nonempty subset of $\mathcal{H}$, and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ such that (9.10) is satisfied. Then $(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in C if and only if every weak sequential cluster point of $(x_n)_{n\in\mathbb{N}}$ is in C.

Proposition 9.2 ([16, Proposition 3.6]) *Let $\alpha \in]0, +\infty[$, let $(\nu_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, and let $(W_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that $\sup_{n\in\mathbb{N}} \|W_n\| < +\infty$ and $(\forall n \in \mathbb{N})$ $(1+\nu_n)W_{n+1} \succcurlyeq W_n$. Furthermore, let C be a subset of $\mathcal{H}$ such that $\operatorname{int} C \neq \varnothing$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ such that*

$$\big(\exists\, (\varepsilon_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})\big)\big(\exists\, (\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})\big)(\forall x \in \mathcal{H})(\forall n \in \mathbb{N})$$
$$\|x_{n+1} - x\|^2_{W_{n+1}} \leqslant (1+\eta_n)\|x_n - x\|^2_{W_n} + \varepsilon_n. \tag{9.12}$$

Then $(x_n)_{n\in\mathbb{N}}$ converges strongly.

Proposition 9.3 ([15, Proposition 3.4]) *Let $\alpha \in]0, +\infty[$, let $(W_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that $\sup_{n\in\mathbb{N}} \|W_n\| < +\infty$, let C be a nonempty closed subset of $\mathcal{H}$, and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{H}$ such that*

$$\big(\exists\, (\varepsilon_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})\big)\big(\exists\, (\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})\big)(\forall z \in C)(\forall n \in \mathbb{N})$$
$$\|x_{n+1} - z\|^2_{W_{n+1}} \leqslant (1+\eta_n)\|x_n - z\|^2_{W_n} + \varepsilon_n. \tag{9.13}$$

Then $(x_n)_{n\in\mathbb{N}}$ converges strongly to a point in C if and only if $\varliminf d_C(x_n) = 0$.

Lemma 9.5 ([16, Lemma 3.1]) *Let $\alpha \in]0, +\infty[$, let $\mu \in]0, +\infty[$, and let A and B be operators in $\mathcal{S}(\mathcal{H})$ such that $\mu \operatorname{Id} \succcurlyeq A \succcurlyeq B \succcurlyeq \alpha \operatorname{Id}$. Then the following hold:*

(i) $\alpha^{-1} \operatorname{Id} \succcurlyeq B^{-1} \succcurlyeq A^{-1} \succcurlyeq \mu^{-1} \operatorname{Id}$.
(ii) $(\forall x \in \mathcal{H})$ $\langle A^{-1}x \mid x\rangle \geqslant \|A\|^{-1}\|x\|^2$.
(iii) $\|A^{-1}\| \leqslant \alpha^{-1}$.

9.3 Main Convergence Result

We present our main result.

Theorem 9.1 *Let $\alpha \in]0, +\infty[$, let $(\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, and let $(U_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that*

$$\mu = \sup_{n\in\mathbb{N}} \|U_n\| < +\infty \quad \text{and} \quad (\forall n \in \mathbb{N}) \quad (1+\eta_n)U_{n+1} \succcurlyeq U_n. \tag{9.14}$$

Let $\varepsilon \in]0, 1[$, let $m \geqslant 1$ be an integer, and let $x_0 \in \mathcal{H}$. For every $i \in \{1, \ldots, m\}$ and every $n \in \mathbb{N}$, let $\alpha_{i,n} \in]0, 1[$, let $T_{i,n} \colon \mathcal{H} \to \mathcal{H}$ be $\alpha_{i,n}$-averaged on $(\mathcal{H}, U_n^{-1})$, let ϕ_n be an averagedness constant of $T_{1,n} \cdots T_{m,n}$, let $\lambda_n \in]0, \phi_n]$, and let $e_{i,n} \in \mathcal{H}$. Iterate

$$\begin{aligned}
&\text{for } n = 0, 1, \ldots \\
&\quad \left\lfloor \begin{aligned}
y_n &= T_{1,n}\Big(T_{2,n}\big(\cdots T_{m-1,n}(T_{m,n}x_n + e_{m,n}) + e_{m-1,n}\cdots\big) + e_{2,n}\Big) + e_{1,n} \\
x_{n+1} &= x_n + \lambda_n(y_n - x_n).
\end{aligned} \right.
\end{aligned} \tag{9.15}$$

Suppose that

$$S = \bigcap_{n\in\mathbb{N}} \operatorname{Fix}(T_{1,n}\cdots T_{m,n}) \neq \varnothing \tag{9.16}$$

and

$$(\forall i \in \{1,\ldots,m\}) \quad \sum_{n\in\mathbb{N}} \lambda_n \|e_{i,n}\|_{U_n^{-1}} < +\infty, \tag{9.17}$$

and define

$$(\forall i \in \{1,\ldots,m\})(\forall n \in \mathbb{N}) \quad T_{i+,n} = \begin{cases} T_{i+1,n}\cdots T_{m,n}, & \text{if } i \neq m;\\ \mathrm{Id}, & \text{if } i = m. \end{cases} \tag{9.18}$$

Then the following hold:

(i) $\sum_{n\in\mathbb{N}} \lambda_n(1/\phi_n - \lambda_n)\|T_{1,n}\cdots T_{m,n}x_n - x_n\|^2_{U_n^{-1}} < +\infty$.

(ii) Suppose that $(\forall n \in \mathbb{N})\ \lambda_n \in\,]0, \varepsilon + (1-\varepsilon)/\phi_n]$. *Then* $(\forall x \in S)$

$$\max_{1\leqslant i\leqslant m} \sum_{n\in\mathbb{N}} \frac{\lambda_n(1-\alpha_{i,n})}{\alpha_{i,n}} \left\|(\mathrm{Id}-T_{i,n})T_{i+,n}x_n - (\mathrm{Id}-T_{i,n})T_{i+,n}x\right\|^2_{U_n^{-1}} < +\infty. \tag{9.19}$$

(iii) $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point in* S *if and only if every weak sequential cluster point of* $(x_n)_{n\in\mathbb{N}}$ *is in* S. *In this case, the convergence is strong if* $\operatorname{int} S \neq \varnothing$.

(iv) $(x_n)_{n\in\mathbb{N}}$ *converges strongly to a point in* S *if and only if* $\varliminf d_S(x_n) = 0$.

Proof Let $n \in \mathbb{N}$ and let $x \in S$. Set

$$T_n = T_{1,n}\cdots T_{m,n} \tag{9.20}$$

and

$$e_n = y_n - T_n x_n. \tag{9.21}$$

Using the nonexpansiveness on $(\mathcal{H}, U_n^{-1})$ of the operators $(T_{i,n})_{1\leqslant i\leqslant m}$, we first derive from (9.21) that

$$\|e_n\|_{U_n^{-1}} \leqslant \sum_{i=1}^{m} \|e_{i,n}\|_{U_n^{-1}}. \tag{9.22}$$

Let us rewrite (9.15) as

$$x_{n+1} = x_n + \lambda_n\big(T_n x_n + e_n - x_n\big), \tag{9.23}$$

and set

$$R_n = (1-1/\phi_n)\,\mathrm{Id} + (1/\phi_n)T_n \quad \text{and} \quad \mu_n = \phi_n\lambda_n. \tag{9.24}$$

Then $\operatorname{Fix} R_n = \operatorname{Fix} T_n$ and, by Lemmas 9.1 and 9.2, R_n is nonexpansive on $(\mathcal{H}, U_n^{-1})$. Furthermore, (9.23) can be written as

$$x_{n+1} = x_n + \mu_n\big(R_n x_n - x_n\big) + \lambda_n e_n, \quad \text{where} \quad \mu_n \in \,]0, 1[\,. \tag{9.25}$$

Now set $z_n = x_n + \mu_n(R_n x_n - x_n)$. Since $x \in \operatorname{Fix} R_n$, we derive from [4, Corollary 2.14] that

$$\begin{aligned}\|z_n - x\|_{U_n^{-1}}^2 &= (1-\mu_n)\|x_n - x\|_{U_n^{-1}}^2 + \mu_n\|R_n x_n - R_n x\|_{U_n^{-1}}^2 \\ &\quad - \mu_n(1-\mu_n)\|R_n x_n - x_n\|_{U_n^{-1}}^2 \qquad (9.26)\\ &\leqslant \|x_n - x\|_{U_n^{-1}}^2 - \lambda_n(1/\phi_n - \lambda_n)\|T_n x_n - x_n\|_{U_n^{-1}}^2. \qquad (9.27)\end{aligned}$$

Hence, (9.25), (9.14), and (9.27) yield

$$\begin{aligned}\|x_{n+1} - x\|_{U_{n+1}^{-1}} &\leqslant \sqrt{1+\eta_n}\|z_n - x\|_{U_n^{-1}} + \lambda_n\sqrt{1+\eta_n}\|e_n\|_{U_n^{-1}} \qquad (9.28)\\ &\leqslant \sqrt{1+\eta_n}\|x_n - x\|_{U_n^{-1}} + \lambda_n\sqrt{1+\eta_n}\|e_n\|_{U_n^{-1}} \qquad (9.29)\end{aligned}$$

and, since $\sum_{k\in\mathbb{N}} \lambda_k\|e_k\|_{U_k} < +\infty$, it follows from Lemma 9.4 that

$$\nu = \sum_{k\in\mathbb{N}} \lambda_k\|e_k\|_{U_k^{-1}} + 2\sup_{k\in\mathbb{N}} \|x_k - x\|_{U_k^{-1}} < +\infty. \tag{9.30}$$

Moreover, using (9.28) and (9.27) we write

$$\begin{aligned}(1+\eta_n)^{-1}\|x_{n+1} - x\|_{U_{n+1}^{-1}}^2 &\leqslant \|z_n - x\|_{U_n^{-1}}^2 \\ &\quad + (2\|z_n - x\|_{U_n^{-1}} + \lambda_n\|e_n\|_{U_n^{-1}})\lambda_n\|e_n\|_{U_n^{-1}} \qquad (9.31)\\ &\leqslant \|x_n - x\|_{U_n^{-1}}^2 - \lambda_n(1/\phi_n - \lambda_n)\|T_n x_n - x_n\|_{U_n^{-1}}^2 \\ &\quad + \nu\lambda_n\|e_n\|_{U_n^{-1}}. \qquad (9.32)\end{aligned}$$

(i): This follows from (9.32), (9.20), (9.16), (9.30), and Lemma 9.3.
(ii): We apply the definition of averagedness of the operators $(T_{i,n})_{1\leqslant i\leqslant m}$ to obtain

$$\begin{aligned}\|T_n x_n - x\|_{U_n^{-1}}^2 &= \big\|T_{1,n}\cdots T_{m,n}x_n - T_{1,n}\cdots T_{m,n}x\big\|_{U_n^{-1}}^2 \\ &\leqslant \big\|T_{2,n}\cdots T_{m,n}x_n - T_{2,n}\cdots T_{m,n}x\big\|_{U_n^{-1}}^2 \\ &\quad - \frac{1-\alpha_{1,n}}{\alpha_{1,n}}\big\|(\mathrm{Id} - T_{1,n})T_{2,n}\cdots T_{m,n}x_n \\ &\quad - (\mathrm{Id} - T_{1,n})T_{2,n}\cdots T_{m,n}x\big\|_{U_n^{-1}}^2\end{aligned}$$

$$\begin{aligned}
&\vdots \\
&\leqslant \|x_n - x\|^2_{U_n^{-1}} \\
&\quad - \sum_{i=1}^{m} \frac{1-\alpha_{i,n}}{\alpha_{i,n}} \left\| (\mathrm{Id} - T_{i,n}) T_{i+,n} x_n - (\mathrm{Id} - T_{i,n}) T_{i+,n} x \right\|^2_{U_n^{-1}}.
\end{aligned} \tag{9.33}$$

Note also that

$$\begin{aligned}
\lambda_n \leqslant \varepsilon + \frac{1-\varepsilon}{\phi_n} &\Rightarrow \frac{1}{\varepsilon}\lambda_n \leqslant \left(\frac{1}{\varepsilon} - 1\right)\frac{1}{\phi_n} \\
&\Leftrightarrow \lambda_n - 1 \leqslant \left(\frac{1}{\varepsilon} - 1\right)\left(\frac{1}{\phi_n} - \lambda_n\right).
\end{aligned} \tag{9.34}$$

Thus (9.31), the definition of z_n, and [4, Corollary 2.14] yield

$$\begin{aligned}
(1+\eta_n)^{-1}\|x_{n+1} - x\|^2_{U_{n+1}^{-1}} &\leqslant \|(1-\lambda_n)(x_n - x) + \lambda_n(T_n x_n - x)\|^2_{U_n^{-1}} + \nu\lambda_n \|e_n\|_{U_n^{-1}} \\
&= (1-\lambda_n)\|x_n - x\|^2_{U_n^{-1}} + \lambda_n \|T_n x_n - x\|^2_{U_n^{-1}} \\
&\quad + \lambda_n(\lambda_n - 1)\|T_n x_n - x_n\|^2_{U_n^{-1}} + \nu\lambda_n \|e_n\|_{U_n^{-1}} \\
&\leqslant (1-\lambda_n)\|x_n - x\|^2_{U_n^{-1}} + \lambda_n \|T_n x_n - x\|^2_{U_n^{-1}} + \varepsilon_n,
\end{aligned} \tag{9.35}$$

where

$$\varepsilon_n = \lambda_n \left(\frac{1}{\varepsilon} - 1\right)\left(\frac{1}{\alpha_n} - \lambda_n\right)\|T_n x_n - x_n\|^2_{U_n^{-1}} + \nu\lambda_n \|e_n\|_{U_n^{-1}}. \tag{9.36}$$

Now set

$$\beta_n = \lambda_n \max_{1 \leqslant i \leqslant m} \left(\frac{1-\alpha_{i,n}}{\alpha_{i,n}} \left\| (\mathrm{Id} - T_{i,n}) T_{i+,n} x_n - (\mathrm{Id} - T_{i,n}) T_{i+,n} x \right\|^2_{U_n^{-1}} \right). \tag{9.37}$$

On the one hand, it follows from (i), (9.30), and (9.16) that

$$\sum_{k \in \mathbb{N}} \varepsilon_k < +\infty. \tag{9.38}$$

On the other hand, combining (9.33) and (9.35), we obtain

$$(1+\eta_n)^{-1}\|x_{n+1} - x\|^2_{U_{n+1}^{-1}} \leqslant \|x_n - x\|^2_{U_n^{-1}} - \beta_n + \varepsilon_n. \tag{9.39}$$

Consequently, Lemma 9.3 implies that $\sum_{k\in\mathbb{N}} \beta_k < +\infty$.

(iii)–(iv): The results follow from (9.39), (9.38), and Proposition 9.1 for the weak convergence, and Propositions 9.2 and 9.3 for the strong convergence.

□

Remark 9.1 Suppose that $(\forall n \in \mathbb{N})\ U_n = \mathrm{Id}$ and $\lambda_n \leqslant (1-\varepsilon)(1/\phi_n + \varepsilon)$. Then Theorem 9.1 reduces to [18, Theorem 3.5] which itself extends [9, Section 3] in the case $(\forall n \in \mathbb{N})\ \lambda_n \leqslant 1$. As far as we know, it is the first inexact overrelaxed variable metric algorithm based on averaged operators.

9.4 Applications to the Forward-Backward Algorithm

A special case of Theorem 9.1 of interest is the following.

Corollary 9.1 *Let $\alpha \in]0, +\infty[$, let $(\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, and let $(U_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that*

$$\mu = \sup_{n\in\mathbb{N}} \|U_n\| < +\infty \quad \text{and} \quad (\forall n \in \mathbb{N}) \quad (1+\eta_n)U_{n+1} \succcurlyeq U_n. \tag{9.40}$$

Let $\varepsilon \in]0, 1[$ and let $x_0 \in \mathcal{H}$. For every $n \in \mathbb{N}$, let $\alpha_{1,n} \in]0, 1/(1+\varepsilon)]$, let $\alpha_{2,n} \in]0, 1/(1+\varepsilon)]$, let $T_{1,n} \colon \mathcal{H} \to \mathcal{H}$ be $\alpha_{1,n}$-averaged on $(\mathcal{H}, U_n^{-1})$, let $T_{2,n} \colon \mathcal{H} \to \mathcal{H}$ be $\alpha_{2,n}$-averaged on $(\mathcal{H}, U_n^{-1})$, let $e_{1,n} \in \mathcal{H}$, and let $e_{2,n} \in \mathcal{H}$. In addition, for every $n \in \mathbb{N}$, let

$$\lambda_n \in \left[\varepsilon, \varepsilon + \frac{1-\varepsilon}{\phi_n}\right], \quad \text{where} \quad \phi_n = \frac{\alpha_{1,n} + \alpha_{2,n} - 2\alpha_{1,n}\alpha_{2,n}}{1 - \alpha_{1,n}\alpha_{2,n}}, \tag{9.41}$$

and iterate

$$x_{n+1} = x_n + \lambda_n\Big(T_{1,n}\big(T_{2,n}x_n + e_{2,n}\big) + e_{1,n} - x_n\Big). \tag{9.42}$$

Suppose that

$$S = \bigcap_{n\in\mathbb{N}} \mathrm{Fix}\,(T_{1,n}T_{2,n}) \neq \varnothing, \quad \sum_{n\in\mathbb{N}} \lambda_n\|e_{1,n}\| < +\infty, \quad \text{and} \quad \sum_{n\in\mathbb{N}} \lambda_n\|e_{2,n}\| < +\infty. \tag{9.43}$$

Then the following hold:

(i) $\sum_{n\in\mathbb{N}} \|T_{1,n}T_{2,n}x_n - x_n\|^2 < +\infty$.
(ii) $(\forall x \in S)\ \sum_{n\in\mathbb{N}} \|T_{1,n}T_{2,n}x_n - T_{2,n}x_n + T_{2,n}x - x\|^2 < +\infty$.
(iii) $(\forall x \in S)\ \sum_{n\in\mathbb{N}} \|T_{2,n}x_n - x_n - T_{2,n}x + x\|^2 < +\infty$.

(iv) *Suppose that every weak sequential cluster point of* $(x_n)_{n\in\mathbb{N}}$ *is in* S. *Then* $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point in* S, *and the convergence is strong if* $\operatorname{int} S \neq \varnothing$.

(v) $(x_n)_{n\in\mathbb{N}}$ *converges strongly to a point in* S *if and only if* $\underline{\lim}\, d_S(x_n) = 0$.

Proof For every $n \in \mathbb{N}$,

$$\frac{1}{\sqrt{\mu}}\|e_{1,n}\| \leqslant \|e_{1,n}\|_{U_n^{-1}} \quad \text{and} \quad \frac{1}{\sqrt{\mu}}\|e_{2,n}\| \leqslant \|e_{2,n}\|_{U_n^{-1}}, \tag{9.44}$$

and $T_{1,n}T_{2,n}$ is ϕ_n-averaged by [4, Proposition 4.44]. Thus, we apply Theorem 9.1 with $m = 2$.

(i)–(iii): This follows from Theorem 9.1(i) that

$$(\forall x \in S) \quad \begin{cases} \displaystyle\sum_{n\in\mathbb{N}} \frac{\lambda_n(1-\alpha_{1,n})}{\alpha_{1,n}} \left\|(\mathrm{Id}-T_{1,n})T_{2,n}x_n-(\mathrm{Id}-T_{1,n})T_{2,n}x\right\|_{U_n^{-1}}^2 < +\infty \\ \displaystyle\sum_{n\in\mathbb{N}} \frac{\lambda_n(1-\alpha_{2,n})}{\alpha_{2,n}} \left\|(\mathrm{Id}-T_{2,n})x_n-(\mathrm{Id}-T_{2,n})x\right\|_{U_n^{-1}}^2 < +\infty \\ \displaystyle\sum_{n\in\mathbb{N}} \lambda_n\Big(\frac{1}{\phi_n}-\lambda_n\Big)\left\|T_{1,n}T_{2,n}x_n - x_n\right\|_{U_n^{-1}}^2 < +\infty. \end{cases} \tag{9.45}$$

However, we derive from the assumptions that

$$(\forall x \in S)(\forall n \in \mathbb{N}) \quad \begin{cases} T_{1,n}T_{2,n}x = x \\ \dfrac{\lambda_n(1-\alpha_{1,n})}{\alpha_{1,n}} \geqslant \varepsilon^2 \\ \dfrac{\lambda_n(1-\alpha_{2,n})}{\alpha_{2,n}} \geqslant \varepsilon^2 \\ \lambda_n\Big(\dfrac{1}{\phi_n}-\lambda_n\Big) \geqslant \varepsilon\dfrac{1-\phi_n}{\phi_n} \geqslant \dfrac{2\varepsilon^2}{2\varepsilon+1}. \end{cases} \tag{9.46}$$

Combining (9.40), (9.45), and (9.46) completes the proof.

(iv)–(v): This follows from Theorem 9.1(iii)–(iv).

□

Remark 9.2 This corollary is a variable metric version of [18, Corollary 4.1] where $(\forall n \in \mathbb{N})\ U_n = \mathrm{Id}$ and $\lambda_n \leqslant (1-\varepsilon)(1/\phi_n + \varepsilon)$.

We recall the definition of a demiregular operator. See [2] for examples of demiregular operators.

Definition 9.2 ([2, Definition 2.3]) An operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is *demiregular* at $x \in \operatorname{dom} A$ if, for every sequence $((x_n, u_n))_{n\in\mathbb{N}}$ in $\operatorname{gra} A$ and every $u \in Ax$ such that $x_n \rightharpoonup x$ and $u_n \to u$, we have $x_n \to x$.

Proposition 9.4 *Let $\alpha \in]0, +\infty[$, let $U \in \mathcal{P}_\alpha(\mathcal{H})$, let $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ be a maximally monotone operator, let $\beta \in]0, +\infty[$, let $\gamma \in]0, 2\beta/\|U\|]$, and let B a β-cocoercive operator. Then the following hold:*

(i) $J_{\gamma UA}$ *is a* $1/2$*-averaged operator on* $(\mathcal{H}, U^{-1})$.
(ii) $\mathrm{Id} - \gamma UB$ *is a* $\gamma\|U\|/(2\beta)$*-averaged operator on* $(\mathcal{H}, U^{-1})$.

Proof

(i): [16, Lemma 3.7].
(ii): We derive from (9.7) and Lemma 9.5(iii) that for every $x \in \mathcal{H}$ and for every $y \in \mathcal{H}$

$$\begin{aligned}\langle x - y \mid UBx - UBy\rangle_{U^{-1}} &= \langle x - y \mid Bx - By\rangle \\ &\geqslant \beta\langle Bx - By \mid Bx - By\rangle \\ &= \beta\langle U^{-1}(UBx - UBy) \mid UBx - UBy\rangle_{U^{-1}} \\ &\geqslant \|U\|^{-1}\beta\|UBx - UBy\|_{U^{-1}}^2. \end{aligned} \tag{9.47}$$

Thus, for every $x \in \mathcal{H}$ and for every $y \in \mathcal{H}$

$$\|(x - \gamma UBx) - (y - \gamma UBy)\|_{U^{-1}}^2 = \|x - y\|_{U^{-1}}^2 + \|\gamma UBx - \gamma UBy\|_{U^{-1}}^2 - 2\gamma\langle x - y \mid UBx - UBy\rangle_{U^{-1}} \tag{9.48}$$

$$\leqslant \|x - y\|_{U^{-1}}^2 - \gamma(2\beta/\|U\| - \gamma)\|UBx - UBy\|_{U^{-1}}^2, \tag{9.49}$$

which concludes the proof. □

Next, we introduce a new variable metric forward-backward splitting algorithm.

Proposition 9.5 *Let $\beta \in]0, +\infty[$, let $\varepsilon \in]0, \min\{1/2, \beta\}[$, let $\alpha \in]0, +\infty[$, let $(\eta_n)_{n\in\mathbb{N}} \in \ell_+^1(\mathbb{N})$, and let $(U_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that*

$$\mu = \sup_{n\in\mathbb{N}} \|U_n\| < +\infty \quad \text{and} \quad (\forall n \in \mathbb{N}) \quad (1 + \eta_n)U_{n+1} \succcurlyeq U_n. \tag{9.50}$$

Let $x_0 \in \mathcal{H}$, let $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ be maximally monotone, and let $B\colon \mathcal{H} \to \mathcal{H}$ be β-cocoercive. Furthermore, let $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ be sequences in $\mathcal{H}$ such that $\sum_{n\in\mathbb{N}} \|a_n\| < +\infty$ and $\sum_{n\in\mathbb{N}} \|b_n\| < +\infty$. Suppose that $\operatorname{zer}(A + B) \neq \varnothing$ and, for every $n \in \mathbb{N}$, let

$$\gamma_n \in \left[\varepsilon, \frac{2\beta}{(1+\varepsilon)\|U_n\|}\right] \quad \text{and} \quad \lambda_n \in \left[\varepsilon, 1 + (1-\varepsilon)\left(1 - \frac{\gamma_n\|U_n\|}{2\beta}\right)\right], \tag{9.51}$$

and iterate

$$x_{n+1} = x_n + \lambda_n\Big(J_{\gamma_n U_n A}\big(x_n - \gamma_n U_n(Bx_n + b_n)\big) + a_n - x_n\Big). \tag{9.52}$$

Then the following hold:

(i) $\sum_{n\in\mathbb{N}} \|J_{\gamma_n U_n A}(x_n - \gamma_n U_n B x_n) - x_n\|^2 < +\infty$.
(ii) *Let* $x \in \operatorname{zer}(A+B)$. *Then* $\sum_{n\in\mathbb{N}} \|Bx_n - Bx\|^2 < +\infty$.
(iii) $(x_n)_{n\in\mathbb{N}}$ *converges weakly to a point in* $\operatorname{zer}(A+B)$.
(iv) *Suppose that one of the following holds:*

- *(i)* A *is demiregular at every point in* $\operatorname{zer}(A+B)$.
- *(ii)* B *is demiregular at every point in* $\operatorname{zer}(A+B)$.
- *(iii)* $\operatorname{int} S \neq \varnothing$.

Then $(x_n)_{n\in\mathbb{N}}$ *converges strongly to a point in* $\operatorname{zer}(A+B)$.

Proof We apply Corollary 9.1. Set

$$(\forall n \in \mathbb{N})\quad T_{1,n} = J_{\gamma_n U_n A},\quad T_{2,n} = \mathrm{Id} - \gamma_n U_n B,\quad e_{1,n} = a_n,\quad \text{and}\quad e_{2,n} = -\gamma_n U_n b_n. \tag{9.53}$$

Then, for every $n \in \mathbb{N}$, $T_{1,n}$ is $\alpha_{1,n}$-averaged on $(\mathcal{H}, U_n^{-1})$ with $\alpha_{1,n} = 1/2$ and $T_{2,n}$ is $\alpha_{2,n}$-averaged on $(\mathcal{H}, U_n^{-1})$ with $\alpha_{2,n} = \gamma_n\|U_n\|/(2\beta)$ by Proposition 9.4. Moreover, for every $n \in \mathbb{N}$,

$$\phi_n = \frac{\alpha_{1,n} + \alpha_{2,n} - 2\alpha_{1,n}\alpha_{2,n}}{1 - \alpha_{1,n}\alpha_{2,n}} = \frac{2\beta}{4\beta - \gamma_n\|U_n\|} \tag{9.54}$$

and, therefore, (9.51) yields

$$\lambda_n \in \left[\varepsilon, \varepsilon + \frac{1-\varepsilon}{\phi_n}\right]. \tag{9.55}$$

Hence, we derive from (9.54) and (9.55) that $(\forall n \in \mathbb{N})\ \lambda_n \leqslant 2 + \varepsilon$. Consequently,

$$\begin{cases} \sum_{n\in\mathbb{N}} \lambda_n\|e_{1,n}\| = (2+\varepsilon)\sum_{n\in\mathbb{N}} \|a_n\| < +\infty \\ \sum_{n\in\mathbb{N}} \lambda_n\|e_{2,n}\| \leqslant 2\beta(2+\varepsilon)\mu\alpha^{-1}\sum_{n\in\mathbb{N}} \|b_n\| < +\infty. \end{cases} \tag{9.56}$$

Furthermore, it follows from [4, Proposition 26.1(iv)] that

$$(\forall n \in \mathbb{N})\quad S = \operatorname{zer}(A+B) = \operatorname{Fix}(T_{1,n}T_{2,n}) \neq \varnothing. \tag{9.57}$$

Hence, the assumptions of Corollary 9.1 are satisfied.

(i): This is a consequence of Corollary 9.1(i) and (9.53).
(ii): Corollary 9.1(ii), (9.53), and Lemma 9.5(iii) yield

$$\sum_{n\in\mathbb{N}} \|Bx_n - Bx\|^2 = \sum_{n\in\mathbb{N}} \gamma_n^{-2}\|U_n^{-1}(T_{2,n}x_n - x_n - T_{2,n}x + x)\|^2$$

$$\leqslant \frac{1}{\varepsilon^2\alpha^2}\sum_{n\in\mathbb{N}}\|T_{2,n}x_n - x_n - T_{2,n}x + x\|^2$$
$$< +\infty. \tag{9.58}$$

(iii): Let $(k_n)_{n\in\mathbb{N}}$ be a strictly increasing sequence in $\mathbb{N}$ and let $y \in \mathcal{H}$ be such that $x_{k_n} \rightharpoonup y$. In view of Corollary 9.1(iv), it remains to show that $y \in \operatorname{zer}(A+B)$. Set

$$(\forall n \in \mathbb{N}) \quad \begin{cases} y_n = J_{\gamma_n U_n A}(x_n - \gamma_n U_n B x_n) \\ u_n = \gamma_n^{-1}U_n^{-1}(x_n - y_n) - Bx_n \\ v_n = Bx_n \end{cases} \tag{9.59}$$

and let $z \in \operatorname{zer}(A+B)$. Hence, we derive from (i) that $y_n - x_n \to 0$. Then $y_{k_n} \rightharpoonup y$ and by (ii) $Bx_n \to Bz$. Altogether, $y_{k_n} \rightharpoonup y$, $v_{k_n} \rightharpoonup Bz$, $y_{k_n} - x_{k_n} \to 0$, $u_{k_n} + v_{k_n} \to 0$, and, for every $n \in \mathbb{N}$, $u_{k_n} \in Ay_{k_n}$ and $v_{k_n} \in Bx_{k_n}$. It therefore follows from [11, Lemma 4.5(ii)] that $y \in \operatorname{zer}(A+B)$.

(iv): The proof is the same that in [18, Proposition 4.4(iv)].

□

Remark 9.3 Suppose that $(\forall n \in \mathbb{N})$ $U_n = \mathrm{Id}$ and $\lambda_n \leqslant (1-\varepsilon)(1/\phi_n + \varepsilon)$. Then Proposition 9.5 captures [18, Proposition 4.4]. Now suppose that $(\forall n \in \mathbb{N})$ $\lambda_n \leqslant 1$. Then Proposition 9.5 captures [16, Theorem 4.1].

Using the averaged operators framework allows us to obtain an extended forward-backward splitting algorithm in Euclidean spaces.

Example 9.1 Let $\alpha \in]0, +\infty[$, let $(\eta_n)_{n\in\mathbb{N}} \in \ell^1_+(\mathbb{N})$, and let $(U_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that

$$\mu = \sup_{n\in\mathbb{N}}\|U_n\| < +\infty \quad \text{and} \quad (\forall n \in \mathbb{N}) \quad (1+\eta_n)U_{n+1} \succcurlyeq U_n. \tag{9.60}$$

Let $\varepsilon \in]0, 1/2[$, let $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ be a maximally monotone operator, let $\beta \in]0, +\infty[$, let B a β-cocoercive operator, and let $(\gamma_n)_{n\in\mathbb{N}}$ and $(\mu_n)_{n\in\mathbb{N}}$ be sequences in $[\varepsilon, +\infty[$ such that

$$\phi_n = \frac{2\mu_n\beta}{4\beta - \|U_n\|\gamma_n} \leqslant 1-\varepsilon. \tag{9.61}$$

Let $x_0 \in \mathcal{H}$ and iterate

$$(\forall n \in \mathbb{N}) \quad x_{n+1} = x_n + \mu_n\Big(J_{\gamma_n U_n A}(x_n - \gamma_n U_n B) - x_n\Big). \tag{9.62}$$

Suppose that $\mathcal{H}$ is finite-dimensional and that $\operatorname{zer}(A+B) \neq \varnothing$. Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in $\operatorname{zer}(A+B)$.

Proof Set $(\forall n \in \mathbb{N})\ T_n = \mathrm{Id} + \mu_n(J_{\gamma_n U_n A}(\mathrm{Id} - \gamma_n U_n B) - \mathrm{Id})$. Remark that, for every $n \in \mathbb{N}$, T_n is ϕ_n-averaged. Hence we apply Theorem 9.1 with $m = 1$ and $\lambda \equiv 1$. It follows from Theorem 9.1(i) and (9.61) that $T_n x_n - x_n \to 0$. Since $\mathcal{H}$ is finite-dimensional, the claim follows from Theorem 9.1(iii). □

Remark 9.4 An underrelaxation or an appropriate choice of the metric of the algorithm allows us to exceed the classical bound $2/\beta$ for $(\gamma_n)_{n\in\mathbb{N}}$. For instance, the parameters $\gamma_n \equiv 2.99/\beta$, $\mu_n \equiv 1/2$, and $U_n \equiv \mathrm{Id}$ satisfy the assumptions.

9.5 A Composite Monotone Inclusion Problem

We study the composite monotone inclusion presented in [14].

Problem 9.1 Let $z \in \mathcal{H}$, let $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ be maximally monotone, let $\mu \in]0, +\infty[$, let $C\colon \mathcal{H} \to \mathcal{H}$ be μ-cocoercive, and let m be a strictly positive integer. For every $i \in \{1, \ldots, m\}$, let $r_i \in \mathcal{G}_i$, let $B_i\colon \mathcal{G}_i \to 2^{\mathcal{G}_i}$ be maximally monotone, let $\nu_i \in]0, +\infty[$, let $D_i\colon \mathcal{G}_i \to 2^{\mathcal{G}_i}$ be maximally monotone and ν_i-strongly monotone, and suppose that $0 \neq L_i \in \mathcal{B}(\mathcal{H}, \mathcal{G}_i)$. The problem is to find $\overline{x} \in \mathcal{H}$ such that

$$z \in A\overline{x} + \sum_{i=1}^{m} L_i^*\big((B_i \,\square\, D_i)(L_i\overline{x} - r_i)\big) + C\overline{x}, \tag{9.63}$$

the dual problem of which is to find $\overline{v}_1 \in \mathcal{G}_1, \ldots, \overline{v}_m \in \mathcal{G}_m$ such that

$$(\exists\, x \in \mathcal{H}) \quad \begin{cases} z - \sum_{i=1}^{m} L_i^*\overline{v}_i \in Ax + Cx \\ (\forall i \in \{1, \ldots, m\})\ \overline{v}_i \in (B_i \,\square\, D_i)(L_i x - r_i). \end{cases} \tag{9.64}$$

The following corollary is an overrelaxed version of [16, Corollary 6.2].

Corollary 9.2 *In Problem 9.1, suppose that*

$$z \in \operatorname{ran}\left(A + \sum_{i=1}^{m} L_i^*\big((B_i \,\square\, D_i)(L_i \cdot - r_i)\big) + C\right), \tag{9.65}$$

and set

$$\beta = \min\{\mu, \nu_1, \ldots, \nu_m\}. \tag{9.66}$$

Let $\varepsilon \in]0, \min\{1, \beta\}[$, let $\alpha \in]0, +\infty[$, let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $]0, +\infty[$, let $x_0 \in \mathcal{H}$, let $(a_n)_{n\in\mathbb{N}}$ and $(c_n)_{n\in\mathbb{N}}$ be absolutely summable sequences in $\mathcal{H}$, and let $(U_n)_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{H})$ such that $(\forall n \in \mathbb{N})\ U_{n+1} \succcurlyeq U_n$. For every $i \in \{1, \ldots, m\}$, let $v_{i,0} \in \mathcal{G}_i$, and let $(b_{i,n})_{n\in\mathbb{N}}$ and $(d_{i,n})_{n\in\mathbb{N}}$ be absolutely summable

sequences in $\mathcal{G}_i$, and let $(U_{i,n})_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}_\alpha(\mathcal{G}_i)$ such that $(\forall n \in \mathbb{N})$ $U_{i,n+1} \succcurlyeq U_{i,n}$. For every $n \in \mathbb{N}$, set

$$\delta_n = \left(\sqrt{\sum_{i=1}^{m} \|\sqrt{U_{i,n}} L_i \sqrt{U_n}\|^2}\right)^{-1} - 1, \tag{9.67}$$

suppose that

$$\zeta_n = \frac{1+\delta_n}{(1+\delta_n)\max\{\|U_n\|, \|U_{1,n}\|, \ldots, \|U_{m,n}\|\}} \geqslant \frac{1}{2\beta-\varepsilon}, \tag{9.68}$$

and let

$$\lambda_n \in \left[\varepsilon, 1+(1-\varepsilon)\left(1-\frac{1}{2\zeta_n\beta}\right)\right]. \tag{9.69}$$

Iterate

$$\begin{array}{l}
\text{for } n = 0, 1, \ldots \\
\quad \left\lfloor \begin{array}{l}
p_n = J_{U_n A}\Big(x_n - U_n\big(\sum_{i=1}^{m} L_i^* v_{i,n} + Cx_n + c_n - z\big)\Big) + a_n \\
y_n = 2p_n - x_n \\
x_{n+1} = x_n + \lambda_n(p_n - x_n) \\
\text{for } i = 1, \ldots, m \\
\quad \left\lfloor \begin{array}{l}
q_{i,n} = J_{U_{i,n}B_i^{-1}}\Big(v_{i,n} + U_{i,n}\big(L_i y_n - D_i^{-1} v_{i,n} - d_{i,n} - r_i\big)\Big) + b_{i,n} \\
v_{i,n+1} = v_{i,n} + \lambda_n(q_{i,n} - v_{i,n}).
\end{array} \right.
\end{array} \right.
\end{array} \tag{9.70}$$

Then the following hold for some solution $\overline{x}$ to (9.63) and some solution $(\overline{v}_1, \ldots, \overline{v}_m)$ to (9.64):

(i) $x_n \rightharpoonup \overline{x}$.
(ii) $(\forall i \in \{1, \ldots, m\})$ $v_{i,n} \rightharpoonup \overline{v}_i$.
(iii) Suppose that C is demiregular at $\overline{x}$. Then $x_n \to \overline{x}$.
(iv) Suppose that, for some $j \in \{1, \ldots, m\}$, D_j^{-1} is demiregular at $\overline{v}_j$. Then $v_{j,n} \to \overline{v}_j$.

Proof Set $\mathcal{G} = \mathcal{G}_1 \oplus \cdots \oplus \mathcal{G}_m$, $\mathcal{K} = \mathcal{H} \oplus \mathcal{G}$, and

$$\begin{cases}
\widetilde{A}\colon \mathcal{K} \to 2^{\mathcal{K}}\colon (x, v_1, \ldots, v_m) \mapsto \big(\sum_{i=1}^{m} L_i^* v_i - z + Ax\big) \\
\qquad\qquad \times (r_1 - L_1 x + B_1^{-1} v_1) \times \cdots \times (r_m - L_m x + B_m^{-1} v_m) \\
\widetilde{B}\colon \mathcal{K} \to \mathcal{K}\colon (x, v_1, \ldots, v_m) \mapsto \big(Cx, D_1^{-1} v_1, \ldots, D_m^{-1} v_m\big) \\
\widetilde{S}\colon \mathcal{K} \to \mathcal{K}\colon (x, v_1, \ldots, v_m) \mapsto \Big(\sum_{i=1}^{m} L_i^* v_i, -L_1 x, \ldots, -L_m x\Big).
\end{cases} \tag{9.71}$$

Now, for every $n \in \mathbb{N}$, define

$$\begin{cases} \widetilde{U}_n \colon \mathcal{K} \to \mathcal{K} \colon (x, v_1, \ldots, v_m) \mapsto \Big(U_n x, U_{1,n} v_1, \ldots, U_{m,n} v_m\Big) \\ \widetilde{V}_n \colon \mathcal{K} \to \mathcal{K} \colon \\ \qquad (x, v_1, \ldots, v_m) \mapsto \Big(U_n^{-1} x - \sum_{i=1}^m L_i^* v_i, \big(-L_i x + U_{i,n}^{-1} v_i\big)_{1 \leqslant i \leqslant m}\Big) \end{cases} \tag{9.72}$$

and

$$\begin{cases} \widetilde{x}_n = (x_n, v_{1,n}, \ldots, v_{m,n}) \\ \widetilde{y}_n = (p_n, q_{1,n}, \ldots, q_{m,n}) \\ \widetilde{a}_n = (a_n, b_{1,n}, \ldots, b_{m,n}) \\ \widetilde{c}_n = (c_n, d_{1,n}, \ldots, d_{m,n}) \\ \widetilde{d}_n = (U_n^{-1} a_n, U_{1,n}^{-1} b_{1,n}, \ldots, U_{m,n}^{-1} b_{m,n}) \end{cases} \quad \text{and} \quad \widetilde{b}_n = (\widetilde{S} + \widetilde{V}_n)\widetilde{a}_n + \widetilde{c}_n - \widetilde{d}_n. \tag{9.73}$$

It follows from the proof of [16, Corollary 6.2] that (9.70) is equivalent to

$$(\forall n \in \mathbb{N}) \quad \widetilde{x}_{n+1} = \widetilde{x}_n + \lambda_n \Big(J_{\widetilde{V}_n^{-1} \widetilde{A}} \big(\widetilde{x}_n - \widetilde{V}_n^{-1} (\widetilde{B} \widetilde{x}_n + \widetilde{b}_n) \big) + \widetilde{a}_n - \widetilde{x}_n \Big), \tag{9.74}$$

that the operators $\widetilde{A}$ and $\widetilde{B}$ are maximally monotone, and $\widetilde{B}$ is β-cocoercive on $\mathcal{H}$. Furthermore, for every $(\overline{x}, \overline{v}) \in \operatorname{zer}(\widetilde{A} + \widetilde{B})$, $\overline{x}$ solves (9.63) and $\overline{v}$ solves (9.64). Now set $\rho = 1/\alpha + \sqrt{\sum_{i=1}^m \|L_i\|^2}$. We deduce from the proof of [16, Corollary 6.2] that $(\forall n \in \mathbb{N})$ $\|\widetilde{V}_n^{-1}\| \leqslant \zeta_n^{-1} \leqslant 2\beta - \varepsilon$ and $\widetilde{V}_{n+1}^{-1} \succcurlyeq \widetilde{V}_n^{-1} \in \mathcal{P}_{1/\rho}(\mathcal{K})$. We observe that (9.74) has the structure of the variable metric forward-backward splitting algorithm (9.52) and that all the conditions of Proposition 9.5 are satisfied.

(i)&(ii): Proposition 9.5(iii) asserts that there exists

$$\overline{\overline{x}} = (\overline{x}, \overline{v}_1, \ldots, \overline{v}_m) \in \operatorname{zer}(\widetilde{A} + \widetilde{B}) \tag{9.75}$$

such that $\widetilde{x}_n \rightharpoonup \overline{\overline{x}}$.

(iii)&(iv): It follows from Proposition 9.5(ii) that $\widetilde{B}\widetilde{x}_n \to \widetilde{B}\overline{\overline{x}}$. Hence, (9.71), (9.73), and (9.75) yield

$$C x_n \to C\overline{x} \quad \text{and} \quad \big(\forall i \in \{1, \ldots, m\}\big) \quad D_i^{-1} v_{i,n} \to D_i^{-1} \overline{v}_i. \tag{9.76}$$

We derive the results from Definition 9.2 and (i)–(ii) above.

□

Remark 9.5 Suppose that $(\forall n \in \mathbb{N})$ $\lambda_n \leqslant 1$. Then Corollary 9.2 captures [15, Corollary 6.2].

Acknowledgements The author thanks his Ph.D. advisor P. L. Combettes for his guidance during this work, which is part of his Ph.D. dissertation.

References

1. Alotaibi, A., Combettes, P.L., and Shahzad, N.: Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn-Tucker set. SIAM J. Optim. **24**, 2076–2095 (2014)
2. Attouch, H., Briceño-Arias, L.M., and Combettes, P.L.: A parallel splitting method for coupled monotone inclusions. SIAM J. Control Optim. **48**, 3246–3270 (2010)
3. Baillon, J.-B., Bruck, R.E., and Reich, S.: On the asymptotic behavior of nonexpansive mappings and semigroups, Houston J. Math., **4**, 1–9 (1978)
4. Bauschke, H.H. and Combettes, P.L.: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* 2nd ed. Springer, New York (2017)
5. Brézis, H.: *Opérateurs Maximaux Monotones et Semi-groupes de Contractions dans les Espaces de Hilbert.* North-Holland/Elsevier, New York (1973)
6. Briceño-Arias, L.M. and Combettes, P.L.: A monotone + skew splitting model for composite monotone inclusions in duality. SIAM J. Optim. **21**, 1230–1250 (2011)
7. Cegielski, A., *Iterative Methods for Fixed Point Problems in Hilbert Spaces,* Lecture Notes in Mathematics, **2057**. Springer, Heidelberg (2012)
8. Chambolle, A. and Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vision **40**, 120–145 (2011)
9. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization **53**, 475–504 (2004)
10. Combettes, P.L.: Monotoned operator theory in convex optimization. Math. Program. **170**, 177–206 (2018)
11. Combettes, P.L. and Glaudin, L.E.: Quasi-nonexpansive iterations on the affine hull of orbits: from Mann's mean value algorithm to inertial methods. SIAM J. Optim. **27**, 2356–2380 (2017)
12. Combettes, P.L. and Müller, C.L.: Perspective functions: proximal calculus and applications in high-dimensional statistics. J. Math. Anal. Appl. **457**, 1283–1306 (2018)
13. Combettes, P.L. and Pesquet, J.-C.: Proximal splitting methods in signal processing, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, 185–212. Springer, New York (2011)
14. Combettes, P.L. and Pesquet, J.-C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Valued Var. Anal. **20**, 307–330 (2012)
15. Combettes, P.L. and Vũ, B.C: Variable metric quasi-Fejér monotonicity. Nonlinear Anal. **78**, 17–31 (2013)
16. Combettes, P.L. and Vũ, B.C.: Variable metric forward-backward splitting with applications to monotone inclusions in duality. Optimization **63**, 1289–1318 (2014)
17. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting., Multiscale Model. Simul. **4**, 1168–1200 (2005)
18. Combettes, P.L. and Yamada, I.: Compositions and convex combinations of averaged nonexpansive operators. J. Math. Anal. Appl. **425**, 55–70 (2015)
19. Duchi, J. and Singer, Y.: Efficient online and batch learning using forward backward splitting. J. Mach. Learn. Res. **10**, 2899–2934 (2009)
20. Jenatton, R., Mairal, J., Obozinski, G., and Bach, F.: Proximal methods for hierarchical sparse coding. J. Mach. Learn. Res. **12**, 2297–2334 (2011)
21. Peypouquet, J. and Sorin, S.: Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time. J. Convex Anal. **17**, 1113–1163 (2010)
22. Salzo, S.: The variable metric forward-backward splitting algorithm under mild differentiability assumptions. SIAM J. Optim. **27**, 2153–2181 (2017)
23. Showalter, R.E.: *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations.* American Mathematical Society, Providence, RI (1997)
24. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**, 667–681 (2013)

Chapter 10
A Glimpse at Pointwise Asymptotic Stability for Continuous-Time and Discrete-Time Dynamics

Rafal Goebel

Abstract Given a dynamical system, pointwise asymptotic stability, also called semistability, of a set requires that every point in the set be a Lyapunov stable equilibrium, and that every solution converge to one of the equilibria in the set. This note provides examples of pointwise asymptotic stability related to optimization and states select results from the literature, focusing on necessary and sufficient Lyapunov and Lyapunov-like conditions for and robustness of this stability property. Background on the classical asymptotic stability is included.

Keywords Pointwise asymptotic stability · Differential inclusion · Difference inclusion · Monotone operator · Set-valued Lyapunov function

AMS 2010 Subject Classification 93D05, 49J53, 90C25, 34D20, 47H05

10.1 Introduction

Asymptotic stability is an important concept in dynamical systems and control theory. It is often the goal of control engineering design. Given a dynamical system, asymptotic stability of a set requires that the set be Lyapunov stable, i.e., solutions originating near that set remain near that set, and that the distance of every solution to the set decrease asymptotically to 0. Pointwise asymptotic stability is a related property, which requires that every point in the set be a Lyapunov stable equilibrium, and that every solution converge to one of the equilibria in the set. The property appears naturally in algorithms in convex optimization, when the algorithms are viewed as discrete-time dynamical systems; in continuous-time dynamics generated by monotone operators, including the steepest descent for a convex function and the so-called saddle dynamics; in control algorithms for multi-agent systems when

R. Goebel (✉)
Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, USA
e-mail: rgoebel1@luc.edu

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_10

consensus is an objective; and in biological, chemical, physiological, and engineered systems not related to optimization.

This note reviews the notion of asymptotic stability; defines and provides examples of pointwise asymptotic stability; and collects some results characterizing these stability properties, with focus on necessary and sufficient Lyapunov and Lyapunov-like conditions, which feature set-valued Lyapunov mappings for the pointwise property, and on robustness of the properties. Continuous-time and discrete-time dynamics are considered and are modeled, respectively, by differential inclusions and difference inclusions.

10.2 Dynamics

Continuous-time dynamics in this note are modeled by differential inclusions

$$\dot{x} \in F(x), \tag{10.1}$$

where $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued mapping.[1] For a general theory of differential inclusions, see [7] or [60]. A *solution* to (10.1) is a function $\phi : I \to \mathbb{R}^n$, where I is an interval containing and beginning at 0, such that ϕ is absolutely continuous on every compact subinterval of I and $\dot{\phi}(t) = \frac{d\phi}{dt}(t) \in F(\phi(t))$ for almost all $t \in I$. A solution is *maximal* if it cannot be extended, and *complete* if its domain is unbounded.

The differential inclusion (10.1), or the mapping F, is said to satisfy *basic assumptions* if

- F is locally bounded[2] and outer semicontinuous (osc)[3] and for every $x \in \mathbb{R}^n$, $F(x)$ is nonempty and convex.

The basic assumptions are sufficient for existence of solutions to (10.1), and also ensure some regularity of the space of solutions to (10.1); see [7, Chapter 2]. Further assumptions on F, like linear growth, or the knowledge that all solutions are bounded, may ensure that maximal solutions are complete.

[1] The set-valued terminology in this note follows [56]. In particular, a set-valued mapping $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ associates to each $x \in \mathbb{R}^n$, a subset $F(x) \subset \mathbb{R}^n$.

[2] $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is locally bounded if for every bounded set $C \subset \mathbb{R}^n$, $F(C) := \bigcup_{x \in C} F(x)$ is bounded.

[3] $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *outer semicontinuous at* $x \in \mathbb{R}^n$ if for every $x_i \to x$ and every convergent $y_i \in F(x_i)$, $\lim_{i \to \infty} y_i \in F(x)$. F is *outer semicontinuous* if it is outer semicontinuous at every $x \in \mathbb{R}^n$. If F is locally bounded and has closed (hence compact) values, outer semicontinuity at x is equivalent to a property of a set-valued F often called upper semicontinuity at x: for every $\varepsilon > 0$, there exists $\delta > 0$ such that $F(x + \delta\mathbb{B}) \subset F(x) + \varepsilon\mathbb{B}$. Here, and in the remainder of this note, $\mathbb{B} \subset \mathbb{R}^n$ is a closed unit ball centered at 0; $x + \delta\mathbb{B}$ is the closed ball of radius δ centered at x; and $F(x) + \varepsilon\mathbb{B}$ is the Minkowski sum of $F(x)$ and $\varepsilon\mathbb{B}$, i.e., $\{y + z \mid y \in F(x),\ z \in \varepsilon\mathbb{B}\}$.

Discrete-time dynamics in this note are modeled by difference inclusions

$$x^+ \in G(x), \tag{10.2}$$

where $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued mapping. A *solution* to (10.2) is a function $\phi : \mathbb{N}_0 \to \mathbb{R}^n$ such that $\phi(j+1) \in G(\phi(j))$ for all $j \in \mathbb{N}_0 := \{0, 1, 2, \dots\}$. The difference inclusion (10.2), or the mapping G, satisfies *basic assumptions* if

- G is locally bounded and outer semicontinuous and for every $x \in \mathbb{R}^n$, $G(x)$ is nonempty.

Nonemptiness of values of G ensures existence and completeness of maximal solutions for (10.2); further conditions in the basic assumptions ensure that the sets of solutions to (10.2) depend outer semicontinuously on initial conditions.

Several concepts, definitions, and results recalled in this note stated for (10.1) have parallel results for (10.2) and vice versa. In fact, they often extend to the setting where constraints are present, i.e., to

$$\dot{x} \in F(x), \qquad x \in C, \tag{10.3}$$

where $C \subset \mathbb{R}^n$, and

$$x^+ \in G(x), \qquad x \in D, \tag{10.4}$$

where $D \subset \mathbb{R}^n$. In such cases, basic assumptions require that C, or D, be closed. Furthermore, some results carry over to the so-called hybrid dynamical systems, which combine (10.3) and (10.4); see [33].

To quickly suggest what hybrid systems are and to illustrate some concepts that follow, an example from [31] is recalled. Let $x_1, x_2, \dots, x_K \in \mathbb{R}^m$ represent the positions of K agents. Agents move towards an agreed-upon target a, according to $\dot{x}_k = a - x_k$, and every $T > 0$ amount of time update the target a by picking $a \in \Gamma(x_1, x_2, \dots, x_K)$, where Γ is some set-valued mapping. The resulting dynamical system, with τ serving as a timer variable, is

$$\begin{aligned} \dot{x}_k = a - x_k,\ \dot{a} = 0,\ \dot{\tau} = -1 \qquad & \text{if } \tau \in [0, T], \\ x_k^+ = x_k,\ a^+ \in \Gamma(x_1, x_2, \dots, x_K),\ \tau^+ = T \ \text{ if } \quad & \tau = 0, \end{aligned}$$

which fits the combination of (10.3) and (10.4) by taking $x = (x_1, \dots, x_K, a, \tau) \in \mathbb{R}^{(K+1)m+1}$, $C = \mathbb{R}^{(K+1)m} \times (0, T]$, $D = \mathbb{R}^{(K+1)m} \times \{0\}$, and F and G as determined by the dynamics above. If $\Gamma(x_1, x_2, \dots, x_K)$ is the convex hull of x_k's, namely, the smallest convex set containing $x_1, x_2, \dots, x_K$, then, as time and the number of updates of a and τ go to ∞, x_k's converge to a common limit. In fact, the set $\{x \in \mathbb{R}^{(K+1)m+1} \mid x_1 = \cdots = x_K = a\}$ has a kind of pointwise asymptotic property. The property can be shown by noting that the convex hull of x_k's and a (a is included to account for a bad initial condition for a, i.e., when a is initially not in the convex hull

of x_k's) is not increasing along solutions and does not remain constant forever unless it consists of a single point. This approach can be made formal, and generalized, using set-valued Lyapunov functions.

10.2.1 Why Basic Assumptions?

Basic assumptions lead to desirable structure of the set of solutions to (10.1) and (10.2). For (10.1), they also have an interesting control engineering motivation. A control system is, in rough terms, a differential equation where the right-hand side depends not just on the state x of the system but also on the input, or control, u. A control system can be thus represented by $\dot{x} = c(x, u)$. Feedback control of a control system is about letting the control u be a function of the state x, $u = k(x)$. In some cases, control objectives—including asymptotic stability—cannot be achieved through continuous feedback k, even if c is continuous, but can be achieved using discontinuous k. See [4] for an example. Discontinuous k may lead to discontinuous $x \mapsto c(x, k(x))$, and thus to a differential equation $\dot{x} = f(x)$ with a discontinuous right-hand side. Solutions to such differential equations may be quite sensitive to initial conditions. When such differential equations result from a discontinuous feedback $u = k(x)$ applied to a control system, the sensitivity may be to measurement error as well. That is, small errors represented here by e and the application of $u = k(x + e)$ can result in behaviors quite different from those resulting from $u = k(x)$.

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a function. An absolutely continuous $\phi : [0, T] \to \mathbb{R}^n$ is a *Hermes solution* to the differential equation $\dot{x} = f(x)$ if there exist sequences $\phi_i : [0, T] \to \mathbb{R}^n$ of absolutely continuous functions converging uniformly to ϕ and $e_i : [0, T] \to \mathbb{R}^n$ of measurable functions converging uniformly to the 0 function such that

$$\dot{\phi}_i(t) = f\left(\phi_i(t) + e_i(t)\right) \qquad \text{for almost all } t \in [0, T].$$

An absolutely continuous $\phi : [0, T] \to \mathbb{R}^n$ is a *Krasovskii solution* to $\dot{x} = f(x)$ if it is a solution to the differential inclusion (10.1) with $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by[4]

$$F(x) = \bigcap_{\delta > 0} \overline{\text{con}}\, f(x + \delta \mathbb{B}) \qquad \forall x \in \mathbb{R}^n.$$

The set-valued regularization above, of the possibly discontinuous f, comes from [43] and differs from the one used by Filippov [26] in that the latter excludes from $x + \delta\mathbb{B}$ sets of 0 measure. Both regularizations lead to F that satisfies the basic

[4] $\overline{\text{con}}\, f(x + \delta\mathbb{B})$ is the closure of the convex hull of $f(x + \delta\mathbb{B})$, i.e., of the smallest convex set containing $f(x + \delta\mathbb{B})$.

assumptions. The connection between the two notions of generalized solutions to $\dot{x} = f(x)$, stated below, was observed by [37] and formally proved in [35]; see also an extension to hybrid systems in [58].

Theorem 10.1 *If $f : \mathbb{R}^n \to \mathbb{R}^n$ is locally bounded, then $\phi : [0, T] \to \mathbb{R}^n$ is a Hermes solution to $\dot{x} = f(x)$ if and only if it is a Krasovskii solution.*

In summary, the effect of small perturbations, including measurement error in a control system, on a differential equation with a discontinuous right-hand side can be studied by passing to a well-behaved differential inclusion.

To illustrate the role of basic assumptions in the study of convergence of solutions to the difference inclusion (10.2), a result of [64] is recalled below. In the nonlinear programming setting of [64], (10.2) was understood as an algorithm with a set of solutions (minimizers), denoted below by A. In that sense, the conclusion of the result is that an algorithm either terminates in finite number of steps at a solution or generates a sequence, the convergent subsequences of which converge to solutions.

Theorem 10.2 *Let $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map. Let $A \subset \mathbb{R}^n$ be a set. Suppose that:*

- *There exists a continuous $V : \mathbb{R}^n \to \mathbb{R}$ such that*

  $$\text{if } x \notin A \text{ then for every } x^+ \in G(x),\ V(x^+) < V(x);$$
  $$\text{if } x \in A \text{ then for every } x^+ \in G(x),\ V(x^+) \leq V(x).$$

- *For every $x \notin A$, $G(x)$ is nonempty and G is outer semicontinuous at x.*

Let ϕ be a bounded maximal solution to (10.2). Then either the domain of ϕ is bounded, given by $\{0, 1, \ldots, J\}$, and $x(J) \in A$; or ϕ is complete and the limit of any convergent subsequence of $\{\phi(j)\}_{j=0}^{\infty}$ is in A.

The gist of the proof is that if ϕ is a complete solution to (10.2), then a limit of any convergent subsequence, denoted $\overline{x}$, satisfies $\overline{x} \in A$. Indeed, otherwise $G(\overline{x}) \neq \emptyset$ and for some $y \in G(\overline{x})$, $V(y) = V(\overline{x})$, which cannot hold if $\overline{x} \notin A$. The result and the method of proof resemble what in control theory is known as the Invariance Principle [10, 44] where, usually, one of the assumptions requires that level sets of V that are invariant under the dynamics be in A.

10.3 Asymptotic Stability

A set $A \subset \mathbb{R}^n$ is *asymptotically stable* for the differential inclusion (10.1) if every maximal solution to (10.1) is complete; if

- A is *Lyapunov stable*: for every $\varepsilon > 0$ there exists $\delta > 0$ so that, for every complete solution ϕ to (10.1), if $d_A(\phi(0)) < \delta$ then $d_A(\phi(t)) < \varepsilon$ for all $t \geq 0$; and if
- A is *attractive*: for every complete solution ϕ to (10.1), $\lim_{t\to\infty} d_A(\phi(t)) = 0$.

Above, d_A is the distance to A, i.e., $d_A(x) = \inf_{a\in A} \|x - a\|$. The property just defined is usually referred to as global asymptotic stability or asymptotic stability in the large, as opposed to local asymptotic stability that requires Lyapunov stability and local attractivity of A: only solutions from a neighborhood of A are required to satisfy $\lim_{t\to\infty} d_A(\phi(t)) = 0$. For simplicity of presentation, only the global concept is discussed here and the adjective is skipped.

Well-known sufficient conditions for asymptotic stability, dating back to Lyapunov [46], involve a function that decreases along solutions to (10.1). A variety of conditions involving generalized derivatives of a Lyapunov functions have been used to describe that decrease; see [8]. Roughly, they take the form

$$\partial V(x) \cdot f \leq -W(x) \quad \forall f \in F(x), \tag{10.5}$$

where V is a Lyapunov function, ∂V is an appropriate generalized gradient of V, and the nonnegative W represents the rate of decrease of V along solutions to (10.1): the inequality (10.5) should ensure that $V(\phi(t))$ decreases along a solution ϕ with rate $W(\phi(t))$. Here, for simplicity—and also because existence of such functions can be shown if the dynamics or the asymptotic stability is somewhat regular, only smooth Lyapunov functions are precisely defined. A function $V : \mathbb{R}^n \to [0, \infty)$ is a *smooth Lyapunov function* for (10.1) and a compact set A if it is C^∞, $V(x) = 0$ if and only if $x \in A$, $\lim_{|x|\to\infty} V(x) = \infty$, and

$$\nabla V(x) \cdot f \leq -V(x) \quad \forall x \in \mathbb{R}^n,\ f \in F(x). \tag{10.6}$$

Theorem 10.3 *If every maximal solution to* (10.1) *is complete and there exists a smooth Lyapunov function for* (10.1) *and a compact set A, then A is asymptotically stable for* (10.1)*.*

Less known and usually harder to prove are converse Lyapunov results, guaranteeing the existence of Lyapunov functions for asymptotically stable differential and difference inclusions. For linear dynamics with an asymptotically stable origin, Lyapunov showed that there exist quadratic Lyapunov functions, but for nonlinear dynamics, many converse results produce irregular Lyapunov functions. For a historical accounting, see the survey [40]. A strong converse result, guaranteeing the existence of a smooth Lyapunov function for (10.1), was given by [23, Theorem 1.2] for $A = \{0\}$ and extended to more general A, and more general stability concepts, in [61]. The result below follows from Theorems 1, 3 and Proposition 3 in [61].

Theorem 10.4 *If F satisfies the basic assumptions and a compact set A is asymptotically stable for* (10.1)*, then there exists a smooth Lyapunov function for* (10.1) *and A.*

The approach in [23, 61] to the converse result is to first establish robustness of asymptotic stability and then rely on robustness to smooth out a possibly discontinuous and nonsmooth Lyapunov function candidate. The set A is *robustly asymptotically stable* for (10.1) if it is asymptotically stable and there exists a

continuous function $\rho : \mathbb{R} \to [0, \infty)$, with $\rho(x) = 0$ if and only if $x \in A$, such that A is asymptotically stable for

$$\dot{x} \in F_\rho(x), \tag{10.7}$$

where the set-valued mapping $F_\rho : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is

$$F_\rho(x) = \overline{\text{con}} F\,(x + \rho(x)\mathbb{B}) + \rho(x)\mathbb{B} \quad \forall x \in \mathbb{R}^n. \tag{10.8}$$

In common words, robustness of asymptotic stability of A requires that the property remain true if the dynamics are enlarged, or inflated, away from A. The result below follows from Theorem 3 and Proposition 3 in [61].

Theorem 10.5 *If F satisfies the basic assumptions and a compact set A is asymptotically stable for* (10.1)*, then A is robustly asymptotically stable for* (10.1).

In fact, robust asymptotic stability of a compact set A is, essentially, equivalent to the existence of a smooth Lyapunov function. This is visible in [23, Theorem 1.2, Proposition 3.1] for the case of $A = \{0\}$, and explicitly stated in greater generality in [61, Theorem 1]. Subject to adding some conditions on the completeness of maximal solutions, one can conclude that the following are equivalent:

- A is robustly asymptotically stable for (10.1);
- there exists a smooth Lyapunov function for (10.1) and A.

This and the other ideas and results above carry over to discrete dynamics [41, 42] and generalize to hybrid dynamical systems [33]. For completeness, the definition parallel to (10.6) and results parallel to Theorem 10.3 and Theorem 10.4, in the setting of (10.2) are given below, summarizing [41, 42]. Note that only continuity is required below; this property, for difference inclusions, plays the role that smoothness (though continuity of the gradient is what matters the most) plays for differential inclusions. However, converse results in [41, 42] do yield smooth V.

A function $V : \mathbb{R}^n \to [0, \infty)$ is a *continuous Lyapunov function* for (10.2) and a compact set A if it is continuous, $V(x) = 0$ if and only if $x \in A$, $\lim_{|x|\to\infty} V(x) = \infty$, and

$$V(g) \leq \frac{1}{e} V(x) \quad \forall x \in \mathbb{R}^n,\ g \in G(x).$$

The constant $1/e$ above is picked to mimic the exponential decay of V in (10.5); any positive constant less than 1 leads to asymptotic stability.

Theorem 10.6 *If there exists a continuous Lyapunov function for* (10.2) *and a compact set A, then A is asymptotically stable for* (10.2)*. If* (10.2) *satisfies the basic assumptions and a compact set A is asymptotically stable for* (10.2)*, then there exists a continuous Lyapunov function for* (10.2).

10.4 Pointwise Asymptotic Stability: Some Examples

A set $A \subset \mathbb{R}^n$ is *pointwise asymptotically stable* for the differential inclusion (10.1) if every maximal solution to (10.1) is complete; if

- every point $a \in A$ is Lyapunov stable: for every $\varepsilon > 0$ there exists $\delta > 0$ so that, for every complete solution ϕ to (10.1), if $\|\phi(0) - a\| < \delta$ then $\|\phi(t) - a\| < \varepsilon$ for all $t \geq 0$; and if
- every complete solution ϕ to (10.1) is convergent and $\lim_{t\to\infty} \phi(t) \in A$.

As it was the case for asymptotic stability, the concept above is global/in the large and the adjective is skipped. If A is a singleton, the pointwise asymptotic stability is the same as asymptotic stability. If A is compact, pointwise asymptotic stability implies asymptotic stability. Indeed, given an $\varepsilon > 0$ and, for each $a \in A$, a $\delta_a > 0$ that verifies Lyapunov stability of a, compactness of A leads to a single $\delta > 0$ verifying Lyapunov stability of A. If A is unbounded, this is no longer the case, and Lyapunov stability of every $a \in A$ need not imply Lyapunov stability of A in the sense of Section 10.3. Thus, for unbounded A, asymptotic and pointwise asymptotic stability properties are not comparable. Furthermore, an asymptotically stable compact set A consisting of equilibria of (10.1) only, i.e., points with $F(x) = 0$, need not be pointwise asymptotically stable, as pointed out in a nice example [12, Example 1.1]: for a system on $\mathbb{R}^2$ given by

$$\begin{aligned}\dot{x} = f(x) := \operatorname{sign}(x_1^2 + x_2^2 - 1)\left|x_1^2 + x_2^2 - 1\right|^{\alpha}\begin{pmatrix} -x_1 \\ -x_2 \end{pmatrix} \\ + \operatorname{sign}(x_1^2 + x_2^2 - 1)\left|x_1^2 + x_2^2 - 1\right|^{\beta}\begin{pmatrix} x_2 \\ -x_1 \end{pmatrix},\end{aligned}$$

the unit circle consists of equilibria, but depending on the parameters α, β, solutions can converge, in distance, to the unit circle while winding themselves around it infinitely many times, violating Lyapunov stability of each point on the circle; or may have limits in the unit circle, leading to pointwise asymptotic stability.

A linear differential equation $\dot{x} = Mx$ has a pointwise asymptotically stable subspace if and only if M has index 0 or 1 and its nonzero eigenvalues have negative real parts; equivalently, if $\lim_{t\to\infty} e^{Mt}$ exists. Such matrices are called semistable [19]. The name *semistability* was adopted to study pointwise asymptotic stability in general nonlinear differential equations and more general settings by [12, 38] and others. Pointwise asymptotic stability is the name adopted by the author, beginning in [27], but it has appeared before, including in [63] in the setting of saddle-point dynamics. The reason for a different name was that while, for a matrix M, semistability is indeed weaker than stability,[5] for nonlinear dynamics and compact

[5] A square matrix M is stable, or Hurwitz, if all of its eigenvalues have negative real parts. For such a matrix and a linear differential equation $\dot{x} = Mx$, the origin is not just (Lyapunov) stable but also attractive, and hence asymptotically stable.

sets A, pointwise asymptotic stability is in fact a stronger property than asymptotic stability.

Examples of pointwise asymptotic stability presented below are related to optimization of convex functions, dynamics are generated by monotone mappings and have a nonexpansive property, and, consequently, the pointwise asymptotic stability is particularly nice: Lyapunov stability of equilibria is verified with $\delta = \varepsilon$. The example from [12], recalled above, does not have this feature. Further examples where pointwise asymptotic stability appears include modeling of hysteresis [51], biological and physiological systems [34], and mass-action kinetics in chemistry [21].

10.4.1 Fejér Monotonicity

A sequence $\{x_j\}_{j=0}^{\infty}$ is *Fejér monotone* with respect to a set A if, for every $n \in \mathbb{N}_0$,

$$\|x_{j+1} - a\| \leq \|x_j - a\| \quad \forall a \in A.$$

Clearly, if G is nonexpansive, i.e., Lipschitz continuous with constant 1, then solutions to (10.2) are Fejér monotone with respect to the set A of equilibria/fixed points of G, i.e., points a with $G(a) = a$. Several algorithms in convex optimization, whose purpose is to find a minimum of a function f, generate sequences that are Fejér monotone with respect to the set of minimizers of f; see [25] for an exposition. In the terminology of this note, if every solution to (10.2) is Fejér monotone with respect to the set of equilibria of (10.2), then every equilibrium is Lyapunov stable. If, additionally, every solution to (10.2) converges to an equilibrium, then the set of equilibria is pointwise asymptotically stable.

10.4.2 Steepest Descent for Convex Functions, and Beyond

A prototype for the differential inclusions considered in this section is the steepest descent or gradient flow $\dot{x} = -\nabla f(x)$ for a convex and differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. A natural generalization is to consider convex, but nondifferentiable f, leading to the subdifferential inclusion $\dot{x} \in -\partial f(x)$. A further generalization, motivated by the consideration of constraints, is to consider extended-valued convex functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. A different extension is to begin with a function convex in one variable, concave in the other variable—for example, the Lagrangian for a convex optimization problem—and consider the steepest descent in the first variable, steepest ascent in the second.

These generalizations fit under the umbrella of differential inclusions defined by maximal monotone mappings, discussed in the next section. The cases of convex and convex-concave functions follow.

To account for constraints and so to only consider solutions to the dynamics (10.1) or (10.2) that remain in a particular subset of the state space, in the remainder of Section 10.4 the following variant of pointwise asymptotic stability is considered: Given a set $C \subset \mathbb{R}^n$, a set $A \subset \mathbb{R}^n$ is *pointwise asymptotically stable relative to* C for the differential inclusion (10.1) if every maximal solution ϕ to (10.1) with $\phi(0) \in C$ is complete; if every point $a \in A$ is Lyapunov stable relative to C: for every $\varepsilon > 0$ there exists $\delta > 0$ so that, for every complete solution ϕ to (10.1) with $\phi(0) \in C$, if $\|\phi(0) - a\| < \delta$ then $\|\phi(t) - a\| < \varepsilon$ for all $t \geq 0$; and if every complete solution ϕ to (10.1) with $\phi(0) \in C$ is convergent and $\lim_{t\to\infty} \phi(t) \in A$.

10.4.2.1 Differential Inclusions Given by Maximal Monotone Mappings

Let $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximal monotone mapping.[6] Consider the differential inclusion

$$\dot{x} \in -M(x), \tag{10.9}$$

i.e., consider (10.1) with $F = -M$. Then [15], [7, Chapter 3, Section 2]:

- For every $x_0 \in \mathrm{dom} M$[7] there exists a unique maximal solution to (10.9) with $\phi(0) = x_0$ and this solution is complete.
- For any two complete solutions ϕ, ψ to (10.9), $t \mapsto \|\phi(t) - \psi(t)\|$ is nonincreasing. Indeed,

$$\frac{d}{dt}\frac{1}{2}\|\phi(t) - \psi(t)\|^2 = (\phi(t) - \psi(t)) \cdot \left(\dot{\phi}(t) - \dot{\psi}(t)\right) \leq 0,$$

  where the inequality follows directly from monotonicity of M. In particular, the solutions to (10.9) depend continuously on initial conditions.

Furthermore, for every solution ϕ to (10.9), $\|\dot{\phi}(t)\|$ is nonincreasing, and, for almost all $t \geq 0$,

$$\dot{\phi}(t) = m\left(-M(\phi(t))\right),$$

where $m(S)$ is the element of the closed set S with minimum norm. In other words, every solution is "slow," "heavy," or "lazy."

[6] A set-valued mapping $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *monotone* if for every $x, x' \in \mathbb{R}^n$, every $v \in M(x)$, $v' \in M(x')$, one has $(x - x') \cdot (v - v') \geq 0$. It is *maximal monotone* if it is monotone and its graph, $\{(x, v) \in \mathbb{R}^{2n} \mid v \in M(x)\}$, cannot be enlarged without violating monotonicity. In particular, a linear M given by $M(x) = Lx$ is monotone if and only if L is positive semidefinite, and if such M is monotone then it is maximal monotone.

[7] For a set-valued mapping $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, its effective domain, denoted $\mathrm{dom} M$, is the set $\{x \in \mathbb{R}^n \mid M(x) \neq \emptyset\}$.

Let A be the set of equilibria of (10.9). Clearly, every $a \in A$ is Lyapunov stable, and if $A \neq \emptyset$, then every solution to (10.9) is bounded. If, additionally, every solution converges to a point in A, the set A is pointwise asymptotically stable relative to $\mathrm{dom}M$. A natural sufficient condition for such convergence, proposed by [17], is demipositivity. Let x be a complete solution to (10.9) and let a be an equilibrium of (10.9), i.e., $0 \in M(a)$. Then

$$\frac{d}{dt}\frac{1}{2}\|\phi(t) - a\|^2 = (\phi(t) - a) \cdot \dot{\phi}(t) \leq 0,$$

and there must exist $t_i \nearrow \infty$ such that $(\phi(t_i) - a) \cdot \dot{\phi}(t_i) \to 0$. Without loss of generality, one can assume that $\phi(t_i)$ and $\dot{\phi}(t_i)$ converge. If $\overline{x} = \lim_{i\to\infty} \phi(t_i)$ is an equilibrium of (10.9), then it is Lyapunov stable, and consequently, $\lim_{t\to\infty} \phi(t) = \overline{x}$. This leads to the definition: M is *demipositive* if there exists $a \in A$ such that, for every convergent sequence x_i and every bounded sequence $v_i \in M(x_i)$, if $(x_i - a) \cdot v_i \to 0$ then $\lim_{i\to\infty} x_i \in A$. The original definition, given in an infinite-dimensional Hilbert space setting, considered weak convergence of x_i. Here, since the graph of a maximally monotone $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is closed, demipositivity furthermore reduces to a property sometimes referred to as *firm positivity*: there exists $a \in A$ such that, if $v \cdot (x - a) = 0$ for some $v \in M(x)$ then $x \in A$.

Sufficient conditions for demipositivity include strict and strong monotonicity of M,[8] or the interior of A being nonempty. For more, see [54, Proposition 6.2]. In case of strict monotonicity, A reduces to a singleton $\{a\}$ and $\|\phi(t) - \psi(t)\|$ is strictly decreasing for any two solutions ϕ, ψ whenever $\phi(t) \neq \psi(t)$. In case of strong monotonicity, A is also a singleton, the function $V(x) = \|x - a\|^2$ turns out to be a smooth Lyapunov function, and the convergence is exponential.[9]

10.4.2.2 Steepest Descent for a Convex Function

An important case of a maximal monotone mapping is the subdifferential of a proper, lower semicontinuous (lsc), and convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$.[10] The subdifferential mapping $\partial f : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ of such a function is defined, at every $x \in \mathbb{R}^n$, where $f(x) < \infty$, by

[8] A monotone $M : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is *strictly monotone* if for every $x, x' \in \mathbb{R}^n$ with $x \neq x'$, every $v \in M(x)$, $v' \in M(x')$, one has $(x - x') \cdot (v - v') > 0$, and *strongly monotone* if there exists $\rho > 0$ such that, for every $x, x' \in \mathbb{R}^n$, every $v \in M(x)$, $v' \in M(x')$, one has $(x - x') \cdot (v - v') \geq \rho \|x - x'\|^2$.

[9] In systems theory, a system where $\|\phi(t) - \psi(t)\|$ is eventually decreasing to 0, for all solutions, often with appropriately understood uniform decrease rate over $\|\phi(0) - \psi(0)\|$ is called *incrementally stable*, see [3] and the references therein, and *contractive* if $\|\phi(t) - \psi(t)\|$ is decreasing, often at an exponential rate, see [2]. For applications of the contractive property, not related to monotonicity of the dynamics, see the survey [2] and the references therein.

[10] A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is proper if it is not identically equal to ∞ and lsc if, for every $x \in \mathbb{R}^n$ and every $x_i \to x$, $\liminf_{i\to\infty} f(x_i) \geq f(x)$. A useful condition, equivalent to f being proper, lsc, and convex is that the *epigraph* of f, namely the set $\{(x, r) \in \mathbb{R}^n \,|\, r = f(x)\}$ be nonempty, closed, and convex.

$$\partial f(x) = \{v \in \mathbb{R}^n \mid f(x') \geq f(x) + v \cdot (x' - x)\ \forall x' \in \mathbb{R}^n\}. \tag{10.10}$$

The subdifferential mapping is maximal monotone, [56, Theorem 12.17], and when f is a differentiable convex function, $\partial f = \nabla f$. Strong convexity of f is equivalent to strong monotonicity of ∂f, see [56, Exercise 12.59], and thus implies demipositivity of ∂f. A similar equivalence and implication holds for strict convexity, see [56, Theorem 12.17].

To illustrate the reason to consider convex functions that are not necessarily differentiable or even finite-valued, and the utility of the subdifferential, consider the question of minimizing a convex and differentiable function $g : \mathbb{R}^n \to \mathbb{R}$ over a nonempty, closed, and convex $C \subset \mathbb{R}^n$. The question is equivalent to that of minimizing, over $\mathbb{R}^n$, a proper, lsc, and convex $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ given by

$$f(x) = \begin{cases} g(x) \text{ if } x \in C, \\ \infty \quad \text{if } x \notin C. \end{cases} \tag{10.11}$$

The subdifferential of this function is given by

$$\partial f(x) = \begin{cases} \nabla g(x) + N_C(x) \text{ if } x \in C, \\ \emptyset \qquad\qquad\quad \text{if } x \notin C. \end{cases}$$

Here, $N_C(x)$ is the normal cone to C at x.[11] At points x in the interior of C, $\partial f(x) = \nabla g(x)$, while at points on the boundary of C, $\partial f(x)$ is naturally unbounded, and so ∂f does not satisfy the basic assumptions. To ensure that solutions to $\dot{x} = -\nabla g(x)$ remain in the set C, projected dynamics are often considered, i.e., at the boundary points of C, $-\nabla g(x)$ is projected onto the tangent cone to C at x. It turns out that the projection is the same as the minimum norm element of $-\partial f(x)$, with f given by (10.11), and so solutions to the projected gradient dynamics are the same as those to $\dot{x} \in -\partial f(x)$; see [36] or the recent [16, Corollary 2]. These solutions converge to minimizers, if the minimizers exist.

Indeed, consider any proper, lsc, and convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ with a nonempty set of minimizers. Then ∂f is demipositive/firmly positive. To see this, note that $a \in A := \arg\min f$ if and only if $0 \in \partial f(a)$, which follows from (10.10), and then $v \cdot (x - a) = 0$ with $v \in \partial f(x)$ ensures, via (10.10), that $f(x) = f(a)$. Additionally, in the case of M being the subdifferential of a proper, lsc, and convex function, the existence of solutions to (10.9) holds not just for $x_0 \in \text{dom}\partial f$ but for every $x_0 \in \overline{\text{dom} f}$; see [14, Theorem 22].[12]

[11] The normal cone to a closed and convex set $C \subset \mathbb{R}^n$ at $x \in C$ is $N_C(x) = \{v \in \mathbb{R}^n \mid v \cdot (x' - x) \leq 0\ \forall x' \in C\}$.

[12] For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, $\text{dom} f$ is the *effective domain* of f, i.e., the set $\{x \in \mathbb{R}^n \mid f(x) \in \mathbb{R}\}$.

Theorem 10.7 *Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be lower semicontinuous and convex and suppose $A := \arg\min f \neq \emptyset$. Then A is pointwise asymptotically stable, relative to* $\overline{\operatorname{dom} f}$, *for*

$$\dot{x} \in -\partial f(x).$$

The case of steepest descent/steepest ascent for a convex-concave function is more interesting.

10.4.2.3 Saddle-Point Dynamics for a Convex-Concave Function

Let $H : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{-\infty, \infty\}$ be a proper and closed, in the sense of [55], convex-concave function. Closedness plays here a role similar to the role lower semicontinuity plays for convex functions, but the definition is more technical. See [55, Chapter 33] for details and Theorem 10.8 below for an example of a proper and closed function with ∞ and $-\infty$ as values. The convex-concave subdifferential of H, namely the mapping

$$(x, y) \mapsto \partial_x H(x, y) \times \left(-\tilde{\partial}_y H(x, y)\right), \tag{10.12}$$

where $\partial_x H(x, y)$ is the convex analysis subdifferential of the convex function $x \mapsto H(x, y)$, and $\tilde{\partial}_y H(x, y)$ is the negative of the subdifferential of the convex function $y \mapsto -H(x, y)$, is maximal monotone; see [55, Corollary 37.5.2]. Consequently, saddle-point dynamics

$$\dot{x} \in -\partial_x H(x, y), \quad \dot{y} \in \tilde{\partial}_y H(x, y), \tag{10.13}$$

are a special case of (10.9). Equilibria of (10.13) are saddle points of H, namely, points (x^*, y^*) such that $H(x^*, y) \leq H(x^*, y^*) \leq H(x, y^*)$ for all $x \in \mathbb{R}^n$ and all $y \in \mathbb{R}^m$, and the set of all saddle points of H is a closed convex product set, denoted $X^* \times Y^*$. Every saddle point of H is thus Lyapunov stable for (10.13), but convergence of solutions to (10.13) is more delicate than what was the case for steepest descent. In particular, the simple convex-concave $H : \mathbb{R}^2 \to \mathbb{R}$ given by $H(x, y) = xy$ has a unique saddle point $(0, 0)$ and every solution to (10.13) is periodic. This example also shows that, in general, (10.12) is not demipositive.

Sufficient conditions for convergence of every solution to (10.13) to a saddle point of H include, of course, strong convexity in x and strong concavity in y, as then the convex-concave subdifferential is strongly monotone; and strict convexity in x and strict concavity in y, as then the subdifferential is strictly monotone. A weaker sufficient condition, leading to demipositivity of the convex-concave subdifferential, is the existence of a saddle point (x^*, y^*) such that, if (x, y) satisfies $H(x^*, y) = H(x^*, y^*) = H(x, y^*)$, then (x, y) is a saddle point too; see [20].

A similar, but weaker sufficient condition, requiring a kind of partial demipositivity was noted in [62] for differentiable H extended to include constraints $x \in X$ and $y \in Y$ for closed convex sets X and Y in [63]; and was recently revisited to allow nondifferentiable H in [29, Theorem 3.3].

Theorem 10.8 *Let $h : \mathbb{R}^{n+m} \to \mathbb{R}$ be a convex-concave function, let $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$ be nonempty, closed, and convex sets, and let $H : \mathbb{R}^{n+m} \to \mathbb{R} \cup \{-\infty, \infty\}$ be given by*

$$H(x, y) = \begin{cases} h(x, y) & \text{if } x \in X, y \in Y, \\ \infty & \text{if } \quad x \notin X, \\ -\infty & \text{if } x \in X, y \notin Y. \end{cases}$$

If the set of saddle points of H, $X^* \times Y^*$, is nonempty and either $H(x^*, y^*) < H(x, y^*)$ for all $x^* \in X^*$, $y^* \in Y^*$, $x \notin X^*$ or $H(x^*, y) < H(x^*, y^*)$ for all $x^* \in X^*$, $y^* \in Y^*$, $y \notin Y^*$, then $X^* \times Y^*$ is pointwise asymptotically stable, relative to $X \times Y$, for (10.13).

An example where the assumptions of the theorem hold but (10.12) is not demipositive is $H(x, y) = x^2 + xy$. For further discussion and a proof of the result above, and several references with applications of saddle-point dynamics to control engineering problems, see [29].

10.4.3 Consensus Algorithms

For $x = (x_1, x_2, \ldots, x_K) \in \mathbb{R}^{Km}$, where $x_k \in \mathbb{R}^m$ and which could model positions of K agents, consider

$$\dot{x}_i = \sum_{k=1}^{K} a_{ik}(x_k - x_i), \quad i = 1, 2, \ldots, K \tag{10.14}$$

for some constants $a_{ik} \in \mathbb{R}$, $i, k = 1, 2, \ldots, K$. If these constants are nonnegative and $a_{ik} = a_{ki}$, then (10.14) is the steepest descent $\dot{x} = -\nabla f(x)$ for the convex function $f : \mathbb{R}^{Km} \to \mathbb{R}$ given by

$$f(x) = \frac{1}{4} \sum_{i,k=1}^{K} a_{ik}(x_i - x_k)^2.$$

From Theorem 10.7, the set of equilibria in (10.14) is pointwise asymptotically stable. Of particular interest in control literature is the case when the set of equilibria of (10.14) is

$$A = \{x \in \mathbb{R}^{mK} \mid x_1 = x_2 = \cdots = x_K\}, \tag{10.15}$$

in which case convergence to A represents the agents reaching consensus. This occurs in particular when nonnegative $a_{ik} = a_{ki}$'s represent weights in an undirected communication graph between the agents—for example, it may be that $a_{ik} = a_{ki} = 1$ if the agents communicate, 0 otherwise—and the communication graph is connected. See the surveys [52, 53] for a broader exposition, [59] and the literature therein for some generalizations of (10.14) that come from more general convex functions or that allow for rapid changes in the communication graph, and [38] for links between consensus and pointwise asymptotic stability. Many of the mentioned generalizations fit under a broader umbrella of a switching system, as noted in [32], leading to the result below. For a general exposition of switching systems in control, see [45]. In common words, a switching system is a differential (or difference) equation (or inclusion) where the right-hand side switches, according to usually a time-dependent but sometimes a state-dependent rule, between several given mappings.

Consider the switching system

$$\dot{x} \in -\partial f_q(x), \tag{10.16}$$

where Q is a set and for each $q \in Q$, $f_q : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex function. A *switching signal* is a function $\sigma : [0, \infty) \to Q$ such that there exist $0 = t_0 < t_1 < t_2 < \ldots$ with $t_j \nearrow \infty$ such that σ is constant on each $[t_j, t_{j+1})$. Given a switching signal σ, a *solution* to (10.16) is a locally absolutely continuous function $\phi : [0, T] \to \mathbb{R}^n$ or $\phi : [0, \infty) \to \mathbb{R}^n$ such that $\dot{\phi}(t) \in -\partial f_{\sigma(t)}(\phi(t))$ for almost every t in the domain of ϕ. The following is shown in [32]. The proof relies on picking $a \in A_\infty$ and using $\|x - a\|^2$ as a kind of Lyapunov function, which is nondecreasing. (Below, μ stands for the Lebesgue measure, and reduces to a sum of lengths of intervals.)

Theorem 10.9 *Suppose that $Q = \{1, 2, \ldots, p\}$; for every $q \in Q$, $f_q : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a proper, lsc, and convex function; and that $\bigcap_{q \in Q} \arg\min f_q \neq \emptyset$. Let $\sigma : [0, \infty) \to Q$ be a switching signal and let*

$$T_q(\sigma) := \mu\left(\{t \geq 0 \mid \sigma(t) = q\}\right), \quad Q_\infty(\sigma) := \{q \in Q \mid T_q(\sigma) = \infty\}.$$

Then every complete and uniformly continuous solution to (10.16) converges to A_∞, where

$$A_\infty(\sigma) := \bigcap_{q \in Q_\infty(\sigma)} \arg\min f_q.$$

In the setting of Theorem 10.9, the distance of every solution to (10.16) to the set $\bigcap_{q \in Q} \arg\min f_q$ and to $A_\infty(\sigma)$ is nondecreasing, and so, subject to further existence and uniform continuity assumptions, pointwise asymptotic stability of A_∞

can be concluded. In this setting, the issue of agents reaching consensus reduces to the question whether $A_\infty(\sigma)$, the set of common minimizers of f_q for $q \in Q_\infty(\sigma)$, correspond to points representing consensus (10.15).

The switching system (10.16) is a particular kind of time-varying dynamics generalizing (10.9). Different time dependence is considered in, for example, [9] and [5]. When arbitrarily fast switching in (10.16) is allowed, solutions approximate those to $\dot{x} \in -F(x)$ where $F(x)$ is the convex hull of the union of $\partial f_q(x)$. Arbitrary solutions to this inclusion are not expected to converge to common minimizers. However, slow solutions to the inclusion, i.e., solutions to $\dot{x} = m(-F(x))$, are expected to converge to common minimizers if they exist, and to Pareto optimal points in general; see [6, 48] and earlier works referenced therein. Ideas similar to what is behind Theorem 10.9 are related to the ideas behind the alternating or cyclic projection method and other methods of finding common zeros of monotone mappings; see [11, 57], and the numerous references therein. In the discrete-time consensus algorithm setting, this relationship is discussed in [50].

10.5 Pointwise Asymptotic Stability: Some Results

Pointwise asymptotic stability theory has seen contributions in [12, 13, 19, 38, 39], and more, under the name "semistability," and in the work by the author [27, 28], and [31]. Selected results are recalled below. Focus is on set-valued Lyapunov functions which enable necessary and sufficient conditions for pointwise asymptotic stability and characterizations of its robustness, inspired by what was previously done for asymptotic stability, as recalled in Section 10.3.

10.5.1 Sufficient Conditions

The usual Lyapunov conditions for asymptotic stability of a set A don't imply pointwise asymptotic stability, unless A consists of a single point. Additional conditions can be posed ensure pointwise asymptotic stability. For example, [12] considered Lyapunov conditions and appropriately understood nontangent to A behavior of solutions, in the setting of differential equations. A condition of a different nature, which is sufficient for Lyapunov stability and can be combined with other conditions to yield pointwise asymptotic stability, is based on the length of solutions. In the setting of a difference inclusion (10.2), let Length : $\mathbb{R}^n \to \mathbb{R} \cup \{-\infty, \infty\}$ be defined by

$$\text{Length}(x_0) = \sup \{\text{length}(\phi) \mid \phi \text{ is a solution to (10.2)}, \ \phi(0) = x_0\},$$

where

$$\text{length}(\phi) = \sum_{j=0}^{\infty} \|\phi(j+1) - \phi(j)\|.$$

If $\text{Length}(x) = 0$ and the function Length is upper semicontinuous at x, equivalently, continuous at x, then x is Lyapunov stable. See [47], where the length was considered in a general metric space and for a difference inclusion; [13] for related results for differential equations, where the length of a complete solution $x : [0, \infty) \to \mathbb{R}^n$ is $\int_0^\infty \|\dot{\phi}(t)\|\, dt$; and [31] for hybrid systems.

Consensus issues for multiagent systems modeled by difference equations led [49] to introduce a set-valued Lyapunov function, to its use in a sufficient condition for pointwise asymptotic stability, and to applications of the condition to particular cases where the convex hull of the positions of agents can serve as a set-valued Lyapunov function.

Let $A \subset \mathbb{R}^n$ be a closed set. A set-valued mapping $W : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a *set-valued Lyapunov function* for A and the difference inclusion (10.2) if:

- $x \in W(x)$ for every $x \in \mathbb{R}^n$, $W(x) = \{x\}$ if and only if $x \in A$, W is outer semicontinuous at every $x \in A$, and W is locally bounded;
- there exists a continuous $\alpha : \mathbb{R}^n \to [0, \infty)$ such that $\alpha(x) = 0$ if and only if $x \in A$ and

$$W(G(x)) + \alpha(x)\mathbb{B} \subset W(x) \quad \forall x \in \mathbb{R}^n. \tag{10.17}$$

In [49], an inequality involving some measure of the size of W was used in place of (10.17); the inclusion (10.17) is from [27], and is meant to resemble a version of a Lyapunov inequality: $V(G(x)) \leq V(x) - \alpha(x)$, i.e., $V(G(x)) + \alpha(x) \leq V(x)$. The result below is thus a minor variation of [49, Theorem 4].

Theorem 10.10 *If there exists a set-valued Lyapunov function for a closed set $A \subset \mathbb{R}^n$ and the difference inclusion (10.2), then A is pointwise asymptotically stable for (10.2).*

The proof relies on the fact that, for every solution ϕ to (10.2),

$$\phi(j) + \sum_{i=0}^{j-1} \alpha(\phi(i))\mathbb{B} \subset W(\phi(0)).$$

Then, boundedness of solutions follows from $\phi(j) \in W(\phi(0))$; Lyapunov stability of every $a \in A$ follows from $\phi(j) \in W(\phi(0))$, $W(a) = \{a\}$, and outer semicontinuity of W at a; while convergence, for complete solutions, follows from summability of the series of $\alpha(\phi(i))$. Details can be found in [31, Theorem 3.3, Theorem 3.7], in a hybrid system setting. In absence of strict decrease of W, i.e., if (10.17) is replaced by $W(G(x)) \subset W(x)$, invariance-based arguments can lead to similar conclusions.

For completeness, a set-valued mapping $W : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a *set-valued Lyapunov function* for A and the constrained differential inclusion (10.3) if:

- $x \in W(x)$ for every $x \in \overline{C}$, $W(x) = \{x\}$ if and only if $x \in A$, W is outer semicontinuous at every $x \in A$, and W is locally bounded;
- there exists a continuous $\alpha : \mathbb{R}^n \to [0, \infty)$ such that $\alpha(x) = 0$ if and only if $x \in A$ and, for every solution $\phi : [0, T] \to \mathbb{R}^n$ to (10.3),

$$W(\phi(t)) + \left(\int_0^t \alpha(\phi(s))\, ds \right) \mathbb{B} \subset W(\phi(0)) \quad \forall t \in [0, T].$$

Existence of such a W is sufficient for pointwise asymptotic stability; see [31, Theorem 3.3, Theorem 3.7].

10.5.2 Reachable Sets and Limits of Solutions

Consider the difference inclusion (10.2). Let $\mathcal{R}_{\leq J} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be the finite-horizon reachable set, i.e., the set-valued mapping given by

$$\mathcal{R}_{\leq J}(x_0) = \{\phi(j) \,|\, \phi \text{ is a solution to (10.2) with } \phi(0) = x_0,\ j = 0, 1, \ldots, J\},$$

let $\mathcal{R}_\infty : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be the infinite-horizon reachable set, given by

$$\mathcal{R}_\infty(x_0) = \{\phi(j) \,|\, \phi \text{ is a solution to (10.2) with } \phi(0) = x_0,\ j \in \mathbb{N}_0\},$$

and let $\overline{\mathcal{R}}_\infty : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be its closure, i.e., $\overline{\mathcal{R}}_\infty(x_0) = \overline{\mathcal{R}_\infty(x_0)}$. When all complete solutions to (10.2) converge, the limit set-valued mapping $\mathcal{L} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ can be defined by

$$\mathcal{L}(x_0) = \left\{ \lim_{j \to \infty} \phi(j) \,|\, \phi \text{ is a complete solution to (10.2) with } \phi(0) = x_0 \right\}.$$

Parallel definitions can be stated for the differential inclusion (10.1).

Simple examples show that neither the infinite-horizon reachable sets nor limits of solutions, if they exist in the first place, need to depend regularly on initial conditions. The presence of an asymptotically stable compact set does not fix this. For example, for $\dot{x} = x^2(1 - x)$, $[0, 1]$ is asymptotically stable (in fact, it is the smallest asymptotically stable set), while the infinite horizon reachable set from $x < 0$ is $\mathcal{R}_\infty(x) = [x, 0)$; for $x = 0$, $\mathcal{R}_\infty(x) = \{0\}$; and for $0 < x < 1$, $\mathcal{R}_\infty(x) = [x, 1)$. Both $\mathcal{R}_\infty$ and $\overline{\mathcal{R}}_\infty$ fail to be outer semicontinuous at $x = 0$. Similarly, $\mathcal{L}(x) = \{0\}$ for $x \leq 0$ and $\mathcal{L}(x) = \{1\}$ for $x > 0$, and both inner and outer semicontinuity of $\mathcal{L}$ fail at $x = 0$. On the other hand, the existence and continuous dependence of limits of solutions on initial conditions does not imply

Lyapunov stability of equilibria. For $\dot{r} = 0$, $\dot{\theta} = \theta(2\pi - \theta)$ in polar coordinates, where $\theta \in [0, 2\pi)$, the limit of a solution from (r, θ) is $(r, 0)$, and each such limit is an equilibrium, but only $(0, 0)$ is Lyapunov stable.

The presence of a closed and pointwise asymptotically stable set does lead to regularity of $\overline{\mathcal{R}}_\infty$ and $\mathcal{L}$.

Theorem 10.11 *Suppose that $A \subset \mathbb{R}^n$ is a nonempty, closed, and pointwise asymptotically stable for (10.2) set and that G satisfies the basic assumptions. Then*

(a) the set-valued mappings $\overline{\mathcal{R}}_\infty$ and $\mathcal{L}$ are locally bounded and outer semicontinuous, and for every $x_0 \in \mathbb{R}^n$, $\overline{\mathcal{R}}_\infty(x_0) = \mathcal{R}_\infty(x_0) \cup \mathcal{L}(x_0)$;
(b) for every $\varepsilon > 0$ there exists $J \in \mathbb{N}_0$ such that $\overline{\mathcal{R}}_\infty(x_0) \subset \mathcal{R}_{\leq J}(x_0) + \varepsilon\mathbb{B}$.

If, additionally, G is continuous, then $\overline{\mathcal{R}}_\infty$ and $\mathcal{L}$ are continuous.

The result (a) is in [27, Proposition 4.1], (b) is in [28, Lemma 2.12], and the conclusion on continuity—which follows from (b) and continuity of the finite-horizon reachable set for continuous dynamics—is in [28, Proposition 2.13]. Parallel results for a differential inclusion (10.1) hold, but for the conclusions about continuity of $\overline{\mathcal{R}}_\infty$ and $\mathcal{L}$, an assumption of local Lipschitz continuity of F is needed.

10.5.3 Converse Set-Valued Lyapunov Results and Robustness

The set-valued Lyapunov function concept allows for converse results and characterization of robustness of pointwise asymptotic stability for the difference inclusion (10.2). A converse of Theorem 10.10, first given in [27], is below.

Theorem 12 *If the difference inclusion (10.2) satisfies the basic assumptions and a compact set $A \subset \mathbb{R}^n$ is pointwise asymptotically stable for (10.2), then there exists a set-valued Lyapunov function for A and (10.2).*

One approach to proving this is as follows. By Theorem 10.11, the closure of the reachable set $\overline{\mathcal{R}}_\infty : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is locally bounded and osc. Also, by the definition of the reachable set, for every $x \in \mathbb{R}^n$, $x \in \overline{\mathcal{R}}_\infty(x)$ and $\overline{\mathcal{R}}_\infty(G(x)) \subset \overline{\mathcal{R}}_\infty(x)$. By Lyapunov stability of every $a \in A$, for each such a one has $\overline{\mathcal{R}}_\infty(a) = \{a\}$. What is missing is the strict decrease along solutions, as required by (10.17). Since A is asymptotically stable for (10.2), by Theorem 10.6 there exists a continuous Lyapunov function $V : \mathbb{R}^n \to [0, \infty)$, so that $V(g) \leq \frac{1}{e}V(x)$ for every $x \in \mathbb{R}^n$, $g \in G(x)$. Then

$$W(x) = \overline{\mathcal{R}}_\infty(x) + V(x)\mathbb{B}$$

satisfies (10.17) with $\alpha(x) = (1 - 1/e)V(x)$ and thus $W : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as defined above is a set-valued Lyapunov function for A and (10.2).

Under further assumptions on the data in (10.2), $\overline{\mathcal{R}}_\infty$ is a continuous set-valued mapping, and then so is the W constructed above.

Corollary 10.1 *If the difference inclusion* (10.2) *satisfies the basic assumptions, G is continuous, and a compact set $A \subset \mathbb{R}^n$ is pointwise asymptotically stable for* (10.2)*, then there exists a continuous set-valued Lyapunov function for A and* (10.2).

Continuity of a set-valued Lyapunov function is important, since it relates to robustness of pointwise asymptotic stability. The set A is *robustly pointwise asymptotically stable* for (10.2) if it is pointwise asymptotically stable and there exists a continuous function $\rho : \mathbb{R} \to [0, \infty)$, with $\rho(x) = 0$ if and only if $x \in A$, such that A is pointwise asymptotically stable for $x^+ \in G_\rho(x)$, where the set-valued mapping $G_\rho : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is

$$G_\rho(x) = \bigcup_{y \in G(x+\rho(x)\mathbb{B})} y + \rho(y)\mathbb{B} \quad \forall x \in \mathbb{R}^n. \tag{10.18}$$

The result below is [28, Theorem 4.3].

Theorem 10.13 *Suppose that $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is locally bounded. Let the compact set $A \subset \mathbb{R}^n$ be pointwise asymptotically stable for* (10.2)*. Then, the following are equivalent:*

(a) A is robustly pointwise asymptotically stable for (10.2).
(b) There exists a continuous set-valued Lyapunov function for A and (10.2).

The more involved part of the proof, of the implication (a) $\Longrightarrow$ (b), takes an outer semicontinuous set-valued Lyapunov function W, the existence of which is guaranteed by Theorem 10.12, and using the robustness margin ρ, constructs from W a continuous (in fact locally Lipschitz) set-valued Lyapunov function. The construction uses the technique that was used in [23, Proposition 3.5] to smooth a Lyapunov function candidate when proving Theorem 10.4, but applies it to W and not to the dynamics.

It is not immediate how robustness of pointwise asymptotic stability for discrete dynamics, as defined by the perturbation (10.18) of (10.2), and characterized by Corollary 10.1 and Theorem 10.13, relates to sensitivity of optimization algorithms to computational errors [65] and, in particular, to quasi-Fejér monotonicity [24]. Clearly, robustness considered here deals with stability and convergence, while most considerations in the optimization literature focus on convergence only.

Results for continuous-time dynamics (10.1), parallel to Theorem 10.12, Corollary 10.1, and Theorem 10.13 in particular the robustness of pointwise asymptotic stability for continuous (in the set-valued sense) dynamics are expected to hold but have not been published. What is not clear is whether robustness of pointwise asymptotic stability of a compact set automatically holds for outer semicontinuous dynamics, as it is the case for asymptotic stability (recall Theorem 10.5). Partial results in this direction, in the setting of differential inclusions (10.1), have been shown in [29], for the maximal monotone case, and announced in [30] for constrained dynamics with a local Fejér property. A special case of the latter result is below.

Note that even if the dynamics (10.1) are given by a monotone mapping, which ensures the Fejér property in Theorem 10.14, the inflated dynamics (10.8) in the definition of robustness are different from the enlargements of monotone mappings used in analysis of discrete-time optimization algorithms [18] and (10.8) is not monotone. Similarly, the inflated dynamics (10.8) are not single-valued and so robustness considered below is different from what is considered, for example, in [1, 22]. Relating these approaches to robustness properties to one another may be of interest.

Theorem 10.14 *Suppose that* (10.1) *satisfies the basic assumptions; a nonempty, compact* $A \subset \mathbb{R}^n$ *is asymptotically stable for* (10.1)*; and* (10.1) *is locally Fejér monotone with respect to A, in the sense that there exists a neighborhood U of A such that, for every solution* ϕ *to* (10.1) *with* $\phi(0) \in U$, $\|\phi(t) - a\| \leq \|\phi(0) - a\|$ *for every* $a \in A$, *every* $t \in \mathrm{dom}\phi$. *Then A is robustly pointwise asymptotically stable for* (10.1).

An outline of the proof is given, to illustrate the utility of the asymptotic stability results of Section 10.3. By Theorem 10.5, there exists a continuous $\rho_0 : \mathbb{R}^n \to [0, \infty)$, with $\rho_0(x) = 0$ if and only if $x \in A$, such that A is asymptotically stable for

$$\dot{x} \in F_{\rho_0}(x),$$

where F_{ρ_0} is given by (10.8). By Theorem 10.4, there exists a smooth Lyapunov function for this inclusion and A. Without loss of generality suppose that $\{x \in \mathbb{R}^n \mid V(x) \leq 1\} \subset U$, where U comes from the definition of local Fejér monotonicity of (10.1) with respect to A. Let

$$R_i := \left\{ x \in \mathbb{R}^n \mid 2^{-i} \leq V(x) \leq 2^{-i+1} \right\}, \quad i = 1, 2, \ldots$$

which are nonempty compact sets, and let

$$c_i := \min_{x \in R_i} \min_{a \in A} \|x - a\|, \quad d_i = c_i - c_{i+1}, \quad r_i = \min_{x \in R_i} \rho_0(x)$$

which are positive. Compactness-based arguments show that, for each $i = 1, 2, \ldots$, there exists a positive $\rho_i \leq r_i$ such that, for every solution ϕ to (10.8) where ρ is given by $\rho(x) = \rho_i$ for all $x \in \mathbb{R}^n$ and where $\phi(0) \in R_i$; for every t such that $\phi(t) \in R_i$; and for every $a \in A$, one has

$$\|\phi(t) - a\| \leq \|\phi(0) - a\| + d_i. \tag{10.19}$$

Let $\rho : \mathbb{R}^n \to [0, \infty)$ be continuous, with $\rho(x) = 0$ if and only if $x \in A$, and such that $\rho(x) \leq \rho_i$ for all $x \in R_i$. Then $\rho \leq \rho_0$ and so A is asymptotically stable for (10.8). It is left to show that every $a \in A$ is Lyapunov stable for (10.8). Indeed, for a compact set, this property combined with asymptotic stability implies pointwise asymptotic stability.

Let $a \in A$. Let ϕ be a solution to (10.8) with $\phi(0) \in R_i$ for some i. Let t_j, $j = 1, 2, \ldots$ be such that $V(x(t_j)) = 2^{-i-j+1}$. By (10.19), for $t \in [0, t_1]$, $\|x(t) - a\| \leq \|x(0) - a\| + d_i$. For $t \in [t_1, t_2]$, $\|x(t) - a\| \leq \|x(t_1) - a\| + d_{i+1} \leq \|x(0) - a\| + d_i + d_{i+1}$. In general, for $t \in [t_j, t_{j+1}]$, $j = 1, 2, \ldots$, $\|x(t) - a\| \leq \|x(0) - a\| + \sum_{k=i}^{i+j} d_k$. Hence, for every $t \in [0, \infty)$,

$$\|x(t)-a\| \leq \|x(0)-a\| + \sum_{k=i}^{\infty} d_k = \|x(0) - a\| + c_i \leq \|x(0) - a\| + \|x(0) - a\|.$$

Lyapunov stability of a follows, by taking $\delta = \min\{\varepsilon/2, 1\}$ for any given $\varepsilon > 0$.

A minor variation of the proof above shows that, in the setting of Theorem 10.14, given any $\varepsilon > 0$, the robustness margin $\rho : \mathbb{R} \to [0, \infty)$ can be picked so that not only is A pointwise asymptotically stable for (10.7), but also there exists a neighborhood U of A such that, for every solution ϕ to (10.7) with $\phi(0) \in U$, $\|\phi(t) - a\| \leq (1 + \varepsilon)\|\phi(0) - a\|$ for every $a \in A$, every $t \in \text{dom}\phi$. That is, Fejér monotonicity is almost preserved locally by the inflated dynamics (10.8).

Acknowledgements This work was partially supported by the Simons Foundation Grant 315326.

References

1. Aizicovici, S., Reich, S., Zaslavski, A.: Minimizing convex functions by continuous descent methods. Electron. J. Differential Equations (19) (2010)
2. Aminzare, Z., Sontag, E.: Contraction methods for nonlinear systems: a brief introduction and some open problems. In: Proc. 53rd IEEE Conference on Decision and Control (2014)
3. Angeli, D.: A Lyapunov approach to the incremental stability properties. IEEE Trans. Automat. Control **47**(3), 410–421 (2002)
4. Artstein, Z.: Stabilization with relaxed controls. Nonlinear Anal. **7**(11), 1163–1173 (1983)
5. Attouch, H., Cabot, A., Czarnecki, M.O.: Asymptotic behavior of nonautonomous monotone and subgradient evolution equations. Trans. Amer. Math. Soc. **370**(2), 755–790 (2018)
6. Attouch, H., Garrigos, G., Goudou, X.: A dynamic gradient approach to Pareto optimization with nonsmooth convex objective functions. J. Math. Anal. Appl. **422**(1), 741–771 (2015)
7. Aubin, J.P., Cellina, A.: Differential Inclusions. Springer-Verlag (1984)
8. Bacciotti, A., Rosier, L.: Liapunov Functions and Stability in Control Theory, *Lecture Notes in Control and Information Sciences*, vol. 267. Springer Verlag (2001)
9. Baillon, J., Cominetti, R.: A convergence result for nonautonomous subgradient evolution equations and its application to the steepest descent exponential penalty trajectory in linear programming. J. Funct. Anal. **187**(2), 263–273 (2001)
10. Barbašin, E., Krasovskiĭ, N.: On stability of motion in the large. Doklady Akad. Nauk SSSR (N.S.) **86**, 453–456 (1952)
11. Bauschke, H., Borwein, J., Lewis, A.: The method of cyclic projections for closed convex sets in Hilbert space. In: Recent developments in optimization theory and nonlinear analysis (Jerusalem, 1995), *Contemp. Math.*, vol. 204, pp. 1–38. Amer. Math. Soc., Providence, RI (1997)
12. Bhat, S., Bernstein, D.: Nontangency-based Lyapunov tests for convergence and stability in systems having a continuum of equilibria. SIAM J. Control Optim. **42**(5), 1745–1775 (2003)

13. Bhat, S., Bernstein, D.: Arc-length-based Lyapunov tests for convergence and stability with applications to systems having a continuum of equilibria. Math. Control Signals Systems **22**(2), 155–184 (2010)
14. Brézis, H.: Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations. In: Contributions to nonlinear functional analysis (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1971), pp. 101–156. Academic Press, New York (1971)
15. Brézis, H.: Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert. North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York (1973)
16. Brogliato, B., Daniilidis, A., Lemaréchal, C., Acary, V.: On the equivalence between complementarity systems, projected systems and differential inclusions. Systems Control Lett. **55**(1), 45–51 (2006)
17. Bruck, R.: Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. J. Funct. Anal. **18**, 15–26 (1975)
18. Burachik, R., Iusem, A.: Set-valued mappings and enlargements of monotone operators, *Springer Optimization and Its Applications*, vol. 8. Springer, New York (2008)
19. Campbell, S., Rose, N.: Singular perturbation of autonomous linear systems. SIAM J. Math. Anal. **10**(3), 542–551 (1979)
20. Chbani, Z., Riahi, H.: Existence and asymptotic behaviour for solutions of dynamical equilibrium systems. Evol. Equ. Control Theory **3**(1), 1–14 (2014)
21. Chellaboina, V., Bhat, S., Haddad, W., Bernstein, D.: Modeling and analysis of mass-action kinetics: nonnegativity, realizability, reducibility, and semistability. IEEE Control Syst. Mag. **29**(4), 60–78 (2009)
22. Choudhary, R.: Generic convergence of a convex Lyapounov function along trajectories of nonexpansive semigroups in Hilbert space. J. Nonlinear Convex Anal. **7**(2), 245–268 (2006)
23. Clarke, F., Ledyaev, Y., Stern, R.: Asymptotic stability and smooth Lyapunov functions. J. Diff. Eq. **149**(1), 69–114 (1998)
24. Combettes, P.: Quasi-Fejérian analysis of some optimization algorithms. In: Inherently parallel algorithms in feasibility and optimization and their applications (Haifa, 2000), *Stud. Comput. Math.*, vol. 8, pp. 115–152. North-Holland, Amsterdam (2001)
25. Combettes., P.: Fejér monotonicity in convex optimization. In: Encyclopedia of Optimization, second edn., pp. 1016–1024. Springer, New York (2009)
26. Filippov, A.: Differential Equations with Discontinuous Righthand Sides. Kluwer (1988)
27. Goebel, R.: Set-valued Lyapunov functions for difference inclusions. Automatica **47**(1), 127–132 (2011)
28. Goebel, R.: Robustness of stability through necessary and sufficient Lyapunov-like conditions for systems with a continuum of equilibria. Systems Control Lett. **65**, 81–88 (2014)
29. Goebel, R.: Stability and robustness for saddle-point dynamics through monotone mappings. Systems Control Lett. **108**, 16–22 (2017)
30. Goebel, R., Sanfelice, R.: Applications of convex analysis to consensus algorithms, pointwise asymptotic stability, and its robustness. In: Proc. 57th IEEE Conference on Decision and Control (2018). Accepted
31. Goebel, R., Sanfelice, R.: Pointwise Asymptotic Stability in a Hybrid System and Well-Posed Behavior Beyond Zeno. SIAM J. Control Optim. **56**(2), 1358–1385 (2018)
32. Goebel, R., Sanfelice, R.: A unifying convex analysis and switching system approach to consensus with undirected communication graphs (2018). Submitted, https://arxiv.org/abs/1808.00989
33. Goebel, R., Sanfelice, R., Teel, A.: Hybrid Dynamical Systems: Modeling, Stability, and Robustness. Princeton University Press (2012)

34. Haddad, W., Chellaboina, V.: Stability and dissipativity theory for nonnegative dynamical systems: a unified analysis framework for biological and physiological systems. Nonlinear Anal. Real World Appl. **6**(1), 35–65 (2005)
35. Hájek, O.: Discontinuous differential equations, I. J. Diff. Eq. **32**, 149–170 (1979)
36. Henry, C.: An existence theorem for a class of differential equations with multivalued right-hand side. J. Math. Anal. Appl. **41**, 179–186 (1973)
37. Hermes, H.: Discontinuous vector fields and feedback control. In: Differential Equations and Dynamical Systems, pp. 155–165. Academic Press (1967)
38. Hui, Q., Haddad, W., Bhat, S.: Finite-time semistability and consensus for nonlinear dynamical networks. IEEE Trans. Automat. Control **53**(8), 1887–1900 (2008)
39. Hui, Q., Haddad, W., Bhat, S.: Semistability, finite-time stability, differential inclusions, and discontinuous dynamical systems having a continuum of equilibria. IEEE Trans. Automat. Control **54**(10), 2465–2470 (2009)
40. Kellett, C.: Classical converse theorems in Lyapunov's second method. Discrete Contin. Dyn. Syst. Ser. B **20**(8), 2333–2360 (2015)
41. Kellett, C., Teel, A.: Smooth Lyapunov functions and robustness of stability for difference inclusions. Systems & Control Letters **52**, 395–405 (2004)
42. Kellett, C., Teel, A.: On the robustness of $\mathcal{KL}$-stability for difference inclusions: Smooth discrete-time Lyapunov functions. SIAM J. Control Optim. **44**(3), 777–800 (2005)
43. Krasovskii, N., Subbotin, A.: Game-theoretical control problems. Springer-Verlag, New York (1988)
44. LaSalle, J.P.: An invariance principle in the theory of stability. In: Differential equations and dynamical systems. New York: Academic Press (1967)
45. Liberzon, D.: Switching in Systems and Control. Systems and Control: Foundations and Applications. Birkhauser (2003)
46. Lyapunov, A.M.: The general problem of the stability of motion. Internat. J. Control **55**(3), 521–790 (1992). Translated by A. T. Fuller from Édouard Davaux's French translation (1907) of the 1892 Russian original.
47. Maschler, M., Peleg, B.: Stable sets and stable points of set-valued dynamic systems with applications to game theory. SIAM J. Control Optimization **14**(6), 985–995 (1976)
48. Miglierina, E.: Slow solutions of a differential inclusion and vector optimization. Set-Valued Anal. **12**(3), 345–356 (2004)
49. Moreau, L.: Stability of multiagent systems with time-dependent communication links. IEEE Trans. Automat. Control **50**(2), 169–182 (2005)
50. Nedić, A., Ozdaglar, A., Parrilo, P.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Automat. Control **55**(4), 922–938 (2010)
51. Oh, J., Drincic, B., Bernstein, D.: Nonlinear feedback models of hysteresis. IEEE Control Syst. Mag. **29**(1), 100–119 (2009)
52. Oh, K.K., Park, M.C., Ahn, H.S.: A survey of multi-agent formation control. Automatica J. IFAC **53**, 424–440 (2015)
53. Olfati-Saber, R., Fax, J., Murray, R.: Consensus and cooperation in networked multi-agent systems. Proceedings of the IEEE **95**(1), 215–233 (2007)
54. Peypouquet, J., Sorin, S.: Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time. J. Convex Anal. **17**(3–4), 1113–1163 (2010)
55. Rockafellar, R.: Convex Analysis. Princeton University Press (1970)
56. Rockafellar, R., Wets, R.J.B.: Variational Analysis. Springer (1998)
57. Sabach, S.: Products of finitely many resolvents of maximal monotone mappings in reflexive Banach spaces. SIAM J. Optim. **21**(4), 1289–1308 (2011)
58. Sanfelice, R., Goebel, R., Teel, A.: Generalized solutions to hybrid dynamical systems. ESAIM Control Optim. Calc. Var. **14**, 699–724 (2008)
59. Shi, G., Proutiere, A., Johansson, K.: Network synchronization with convexity. SIAM J. Control Optim. **53**(6), 3562–3583 (2015)
60. Smirnov, G.: Introduction to the theory of differential inclusions, *Graduate Studies in Mathematics*, vol. 41. American Mathematical Society, Providence, RI (2002)

61. Teel, A., Praly, L.: A smooth Lyapunov function from a class-$\mathcal{KL}$ estimate involving two positive semidefinite functions. ESAIM Control Optim. Calc. Var. **5**, 313–367 (2000)
62. Venets, V.: A continuous algorithm for finding the saddle points of convex-concave functions. Avtomat. i Telemekh. (1), 42–47 (1984)
63. Venets, V.: Continuous algorithms for solution of convex optimization problems and finding saddle points of convex-concave functions with the use of projection operations. Optimization **16**(4), 519–533 (1985).
64. Zangwill, W.: Nonlinear programming: a unified approach. Prentice-Hall, Inc., Englewood Cliffs, N.J. (1969)
65. Zaslavski, A.: Convergence of a proximal point method in the presence of computational errors in Hilbert spaces. SIAM J. Optim. **20**(5), 2413–2421 (2010).

Chapter 11
A Survey on Proximal Point Type Algorithms for Solving Vector Optimization Problems

Sorin-Mihai Grad

Dedicated to the memory of J.M. Borwein

Abstract In this survey paper we present the existing generalizations of the proximal point method from scalar to vector optimization problems, discussing some of their advantages and drawbacks, respectively, presenting some open challenges and sketching some possible directions for future research.

Keywords Vector optimization problem · Proximal point algorithm · Weakly efficient solution · Efficient solution · Splitting method · Multiobjective optimization problem · Vector function · Scalarization function

AMS 2010 Subject Classification 90C25, 90C29, 90C46

11.1 Introduction

The usual way to solve a vector optimization problem is by scalarizing it, i.e. by attaching to it a scalar optimization problem whose optimal solutions are also optimal in some sense to the original problem. However, this approach can often lead to unbounded scalar optimization problems, hence the necessity to address the vector optimization problems directly, especially when it comes to numerically solving them. One can find some results on the choice of scalarizing parameters in order to guarantee the existence of optimal solutions of the scalarized problems in the literature, but the imposed conditions are quite restrictive (see [47, 62]) and their verification, when possible, may prove to be too expensive from a computational point of view. There are some scalarization methods (for instance the one with the

S.-M. Grad (✉)
Faculty of Mathematics, University of Vienna, Vienna, Austria
e-mail: sorin-mihai.grad@univie.ac.at

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_11

scalarization function introduced by Tammer (Gerstewitz) in [43] or by means of a (semi-)norm, see also [49, Chapter 4]) that lead to scalar optimization problems that are bounded from below, however the objective functions of the latter consist of compositions of functions that are often unsuitable for the existing algorithms. This situation has motivated research on iterative methods for directly solving multiobjective or vector optimization problems consisting in vector-minimizing a vector function, sometimes subject to (geometric) constraints, that are more or less immediate extensions of scalar algorithms. Some of the first contributions to this direction can be found in [54–56, 64] and the interest towards such algorithms remained active during the next decades (see, for instance, [11, 60, 61]), several other methods being adapted or developed. More recently, one can find generalizations from the scalar case to the vector one of several classical methods for solving both smooth optimization problems, such as the Newton's method (cf. [42]), the projected gradient method (cf. [46]) or the steepest descent method (cf. [48]), and nonsmooth ones, for instance the proximal point method (cf. [24, 75]), the proximal bundle one (cf. [58]) or the subgradient method (cf. [40]). Moreover, one can find even methods for solving vector optimization problems that rely on dynamical systems, such as the ones proposed in [5–7].

In this survey we focus on the existing generalizations of the proximal point method from scalar to vector optimization problems, briefly presenting them and discussing about their advantages and drawbacks, respectively, mentioning some open problems and sketching some possible directions for future research. The (already classical) proximal point algorithm was first proposed by Martinet in [59] and shortly afterwards developed and extended by Rockafellar for solving monotone inclusions, in particular convex (scalar) optimization problems. Since then there were many contributions to this area of research, proximal point type algorithms being now available for various complexly structured convex optimization problems as well as for some classes of nonconvex optimization problems. We refer the reader to [9] for more on the state of the art on this topic.

The first major contribution to extending the proximal point method from scalar to vector optimization problems is the paper [24] due to Bonnel, Iusem and Svaiter. One could argue that the earlier contributions [61] and [45, Section 4.2] contain some proximal point type algorithms for solving vector optimization problems, too, however, in the introduction of [24] the authors explicitly address this issue, stressing that in these works the proximal point steps are actually applied on scalar optimization problems. The mentioned work, written roughly fifteen years ago and cited over one hundred times (according to google scholar) is still the (gold) standard in the field and basically every further paper containing a proximal point type method for solving vector optimization problems builds on it. In the following we discuss around thirty such subsequent contributions (see [4, 12–17, 19, 27–39, 41, 51–53, 67–75]) where one finds algorithms for solving vector optimization problems of various types, where the objective functions are cone-convex, cone-quasiconvex, differences of cone-convex functions or sums of such, have a special structure or are arbitrary and map from Euclidean, Hilbert, Banach or even Hadamard spaces to Euclidean or Banach ones, being minimized over

the whole space or only over some sets or even subject to some other explicitly formulated constraints. In some of these papers the proximality paradigm is present in a classical manner, in some the proximal terms contain Bregman distances, quasi-distances, or are formulated via viscosity functions. There are also two papers where inertial/memory effects or hybrid constructions are added to the algorithm, and we mention, too, some contributions where the regularization is performed by means of a Tikhonov type function instead of the Moreau-Yosida one from the proximal point algorithms.

The method proposed in [24] and extended and refined in subsequent contributions does not directly scalarize the original vector optimization problem. The proximal step of the algorithm consists in choosing as the next iterate a weakly efficient solution of the intermediate vector optimization problem corresponding to the current iteration. The intermediate vector optimization problems are constructed by means of a Moreau-Yosida type regularization and, consequently, they always have weakly efficient solutions. Moreover, the scalarized optimization problems attached to them by any nonzero linear continuous functional from the dual cone of the ordering cone have optimal solutions and, thus, deliver weakly efficient solutions to the intermediate vector optimization problems. This observation is used in the convergence proof, however it does not mean that the original vector optimization problem or the intermediate ones have to be actually scalarized when the algorithm is running. Note also that most of the proximal point type algorithms for vector optimization problems deliver weakly efficient solutions or even, in the nonconvex case, critically efficient solutions to them. The convergence to efficient solutions can be then guaranteed under additional hypotheses.

Note that we have used in this survey the vector optimization problems as they were considered in the original works, i.e. in Euclidean, Hilbert or Banach spaces and with or without constraints, respectively. In order to maintain a reasonable length of the work, we gave for each algorithm only a convergence statement, not other related results such as well definiteness of the iterations.

Most of the proximal point type algorithms for solving vector optimization problems are formulated as theoretical schemes and some of them are accompanied by inexact versions that should be more suitable for implementation. While the papers presenting algorithms for solving scalar optimization problems usually contain applications and computational results, this is rarely the case for the ones dealing with methods for solving vector optimization problems. The algorithm introduced in [24] is explicitly presented as a theoretical scheme meant to be implemented someday and many of its followers are introduced in a similar manner. We discuss later the difficulties encountered while trying to actually numerically test such algorithms. However, making use of additional methods in order to solve the intermediate problems, various authors managed to provide viable implementations of the proximal point type algorithms they introduced for solving vector optimization problems and hence to deliver concrete computational results. We mention in the following, where applicable, on which classes of problems were the considered algorithms tested.

11.2 Preliminaries

In the following we present the general framework we consider within this study, following [24]. Where necessary we mention the changes to this setting. Note, however, that this work is not completely self-contained and the reader is referred to the original sources for some definitions and more properties of notions that are only briefly employed or mentioned within this study.

Let X be a Hilbert space and $(Y, \|\cdot\|)$ a separable Banach space that is partially ordered by a pointed closed convex cone $C \subseteq Y$. Recall that $C \subseteq Y$ is said to be a *cone* when $tC \subseteq C$ for all $t \geq 0$, that is called *pointed* when $-C \cap C = \{0\}$. The partial ordering induced by C on Y is denoted by "$\leqq_C$" (i.e. it holds $x \leqq_C y$ when $y - x \in C$, where $x, y \in Y$) and we write $x \leq_C y$ if $x \leqq_C y$ and $x \neq y$. A greatest element with respect to "$\leqq_C$" denoted by ∞_C which does not belong to Y is attached to this space, and let $Y^\bullet = Y \cup \{\infty_C\}$. Then for any $y \in Y$ one has $y \leq_C \infty_C$ and we consider on $Y^\bullet$ the operations $y + \infty_C = \infty_C + y = \infty_C$ for all $y \in Y^\bullet$ and $t \cdot \infty_C = \infty_C$ for all $t \geq 0$. By $\langle y^*, y\rangle$ we denote the value at $y \in Y$ of the linear continuous functional $y^* \in Y^*$, where $(Y^*, \|\cdot\|_*)$ is the dual space of Y, and by convention we take $\langle y^*, \infty_C\rangle = +\infty$ for all $y^* \in C^*$, where $C^* = \{y^* \in Y^* : \langle y^*, y\rangle \geq 0\ \forall y \in C\}$ is the *dual cone* to C. The *restricted polar* to the cone C is $K_\delta = \{z^* \in Y^* : \langle z^*, y\rangle \geq \delta\|y\|\|z^*\| \text{ for all } y \in C\}$ for some $\delta > 0$. Given a subset U of X, by $\operatorname{cl} U$, $\operatorname{cone} U$, $\operatorname{int} U$ and δ_U we denote its *closure*, *conical hull*, *interior* and *indicator function*, respectively. As $\operatorname{int} C \cup \{0\}$ is a convex cone, too, we also write $x <_C y$ when $y - x \in \operatorname{int} C$. A set $W \subseteq Y$ is said to have the *domination property* with respect to C, if there exists $w \in Y$ such that $W \subseteq w + C$. The closed unit ball of Y is denoted by $\mathcal{B}_Y$ and its unit sphere by $\mathcal{S}_Y$. The convergence in the (corresponding) weak topology is denoted by "$\rightharpoonup$", $\mathrm{id}_X : X \to X$ is the *identity operator* on X and by Pr_U we denote the *projection* onto the (closed convex) set $U \subseteq X$. When Y is finitely dimensional we consider it endowed with the Euclidean norm, unless otherwise specified. Denote also $e = (1, \ldots, 1)^\top \in \mathbb{R}^m$.

A Banach space $(Z, \|\cdot\|)$ is said to be *strictly convex* if $\|(1/2)(x + y)\| < 1$ for all $x, y \in Z$ with $\|x\| = \|y\| = 1$ and $x \neq y$, and *uniformly convex* if $\lim_{n\to+\infty} \|x_n - y_n\| = 0$ for any two sequences $(x_n)_n, (y_n)_n \subseteq \mathcal{S}_Z$ such that $\lim_{n\to+\infty}(\|x_n + y_n\|)/2 = 1$. One says that Z is *(uniformly) smooth* if the limit $\lim_{t\to 0}(\|x+ty\|x\|)/t$ exists (and is attained uniformly) for all $x, y \in \mathcal{S}_Z$. The *normalized duality mapping* of Z is $J_Z : Z \to 2^{Z^*}$ defined by $J_Z(x) = \{x^* \in Z^* : \langle x^*, x\rangle = \|x\|^2 = \|x^*\|_*^2\}$.

When $f : X \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is *proper* (i.e. is nowhere equal to $-\infty$ and has at least a real value) and $\varepsilon \geq 0$, if $f(x) \in \mathbb{R}$ the *(convex) ε-subdifferential* of f at x is $\partial_\varepsilon f(x) = \{x^* \in X^* : f(y) - f(x) \geq \langle x^*, y - x\rangle - \varepsilon\ \forall y \in X\}$, while if $f(x) = +\infty$ we take by convention $\partial_\varepsilon f(x) = \emptyset$. The ε-subdifferential of f becomes in case $\varepsilon = 0$ its classical *(convex) subdifferential* denoted by ∂f. Then $\bar{x} \in X$ is a minimum of f if and only if $0 \in \partial f(\bar{x})$. Denote also by $[t]_+ = \max\{t, 0\}$ for any $t \in \mathbb{R}$.

A vector function $F : X \to Y^\bullet = Y \cup \{\infty_C\}$ is said to be *proper* if its *domain* $\operatorname{dom} F = \{x \in X : F(x) \in Y\}$ is nonempty, *(strictly) C-convex* if $F(tx + (1-t)y) \leqq_C (<_C) tF(x) + (1-t)F(y)$ for all $x, y \in X$ and all $t \in (0,1)$ and *positively C-lower semicontinuous* (in the literature also *star C-lower semicontinuous*) when the function $x \mapsto \langle z^*, F(x)\rangle$, further denoted by $(z^*F) : X \to \overline{\mathbb{R}}$, is lower semicontinuous for all $z^* \in C^* \setminus \{0\}$. A slightly stronger generalization of the classical lower semicontinuity to vector functions is the one due to Penot and Théra (cf. [66]) who called F to be *C-lower semicontinuous* $x \in X$ if for any neighborhood V of 0 and for any $b \in Y$ satisfying $b \leqq_C F(x)$, there exists a neighborhood U of x in X such that $F(U) \subseteq b + V + C \cup \{\infty_C\}$. Last but not least, F is called (cf. [36, 37, 39]) *positively partially continuous* if (z^*F) is continuous on every closed convex subset of $\operatorname{dom} F$ for every $z^* \in C^*$, and *C^*-asymptotically uniformly continuous* when for every bounded sequences $(x_n)_n, (y_n)_n \subseteq X$ such that $\lim_{n\to+\infty} \|x_n - y_n\| = 0$ and each sequence $(z_n^*)_n \subseteq C^*$ weakly*-converging to some $z^* \in C^*$ there holds $\lim_{n\to+\infty} \langle F(x_n) - F(y_n), z_n^* - z^*\rangle = 0$. Related to this one can also define F to be *C^*-uniformly semicontinuous* (on some closed convex set $S \subseteq X$) when for every weakly convergent sequence $(x_n)_n \subseteq S$ to some $x \in S$ and each sequence $(z_n^*)_n \subseteq C^*$ weakly*-converging to some $z^* \in C^*$, one has $\lim_{n\to+\infty} |\langle z_n^*, F(x_n) - F(y_n)\rangle - \langle z^*, F(x) - F(y_n)\rangle| = 0$ for any sequence $(y_n)_n \subseteq S$ for which $\lim_{n\to+\infty} \|x_n - y_n\| = 0$. A generalization of the (convex) ε-subdifferential for vector functions is necessary for our presentation, too. When $F : \mathbb{R}^n \to \mathbb{R}^m$, $K \subseteq \mathbb{R}^m$ is a convex cone and $\varepsilon \geq 0$, the *(vector) ε-subdifferential* of F at $x \in \mathbb{R}^n$ is $\partial_\varepsilon F(x) = \{V \in \mathbb{R}^{m\times n} : F(x) + V^T(y - x) \leqq_K F(y) + \varepsilon e \ \forall y \in X\}$ and it becomes the *(vector) subdifferential* of F denoted by ∂F when $\varepsilon = 0$.

Some notions of nonconvex nonsmooth analysis are necessary as well. Let $f : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz at $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$. The *Clarke directional derivative* of f at x in the direction d is defined as $f^C(x; d) = \lim_{t\downarrow 0} \sup_{y\to x} (f(y + td) - f(y))/t$, while the *Clarke subdifferential* of f at x is $\partial^C f(x) = \{w \in \mathbb{R}^n : w^\top d \leq f^C(x; d) \ \forall d \in \mathbb{R}^n\}$.

In order to introduce some generalized distances, the following notions are necessary. A function $d : \mathbb{R}^n \times \mathbb{R}^n \to R_+ \cup \{+\infty\}$ is said to be a *proximal distance* with respect to a nonempty open convex set $S \subseteq \mathbb{R}^n$ (cf. [8]) if for each $y \in S$ it satisfies the following properties

(P1) $d(\cdot, y)$ is proper, convex and continuously differentiable on S;
(P2) $\operatorname{dom} d(\cdot, y) \subseteq \operatorname{cl} S$ and $\operatorname{dom} \nabla_1 d(\cdot, y) = S$, where ∇_1 denotes the gradient map with respect to the first variable;
(P3) $d(\cdot, y)$ is level bounded on $\mathbb{R}^n$, i.e., $\lim_{\|x\|\to+\infty} d(x, y) = +\infty$;
(P4) $d(y, y) = 0$.

Moreover, a function $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+ \cup \{+\infty\}$ is called the *induced proximal distance* to a given proximal distance d if H is finitely valued on $\operatorname{cl} S \times S$ and for each $y, z \in S$ it satisfies $H(y, y) = 0$, $\nabla_1 d(z, y)^\top (x - z) \leq H(x, y) - H(x, z)$ for all $x \in \operatorname{cl} S$ and $H(x, \cdot)$ is level bounded on S, for all $x \in \operatorname{cl} S$. One denotes by

$\mathcal{F}^*(\mathrm{cl}\, S)$ the set of pairs (d, H) as introduced above that satisfy the following two additional properties

(P5) if $(y_n)_n \subseteq S$ is a bounded sequence in S and $\bar{y} \in \mathrm{cl}\, S$ such that $\lim_{n\to+\infty} H(\bar{y}, y_n) = 0$, then $\lim_{n\to+\infty} y_n = \bar{y}$;
(P6) if $(y_n)_n \subseteq S$ converges to y, then at least one of the relations $\lim_{n\to+\infty} H(y, y_n) = 0$ and $\lim_{n\to+\infty} H(\bar{y}, y_n) = +\infty$ for all $\bar{y} \in S$ such that $\bar{y} \neq y$ holds true.

On the other hand, one says that $q : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ is a *quasi-distance* (cf. [72]) when for any $x, y, z \in \mathbb{R}^n$ it holds

(Q1) $q(x, y) = q(y, x) = 0 \Leftrightarrow x = y$;
(Q2) $q(x, z) \leq q(x, y) + q(y, z)$.

In [33] a vector-valued Bregman distance function was introduced. Let the proper, strictly C-convex and C-lower semicontinuous vector function $G : \mathbb{R}^n \to (\mathbb{R}^m)^\bullet$ whose domain is closed, convex and has a nonempty interior, on which it is Gâteaux differentiable with the Gâteaux derivative $DG(\cdot)$. The *vector-valued Bregman distance with respect to* G is the map $B_G : \mathrm{dom}\, G \times \mathrm{int}(\mathrm{dom}\, G) \to \mathbb{R}^m$, defined by $B_G(z, x) = G(z) - G(x) - DG(x)(z - x)$. Moreover, G is said to be a *vector-valued Bregman distance function* if it satisfies the following hypotheses

(A1) for any $x, y, z \in \mathrm{int}(\mathrm{dom}\, G)$, if $(DG(x) - DG(y))^\top (z - x) \notin -\mathrm{int}\, C$, then $(DG(x) - DG(y))^\top (z - x) \in C$;
(A2) for any $x \in \mathrm{dom}\, G$, $\lambda \in \{a \in \mathbb{R}^m_+ : \|a\| = 1\}$, bounded sequences $(x_n)_n, (y_n)_n \subseteq \mathrm{int}(\mathrm{dom}\, G)$ such that $\lim_{n\to+\infty} \|x_n - y_n\| = 0$, it holds $\lim_{n\to+\infty} (B_G(x, x_n) - B_G(x, y_n))^\top \lambda = 0$;
(A3) for any bounded sequences $(x_n)_n, (y_n)_n \subseteq \mathrm{int}(\mathrm{dom}\, G)$ such that $\lim_{n\to+\infty} y_n = y$ and, for any $\lambda \in \{a \in \mathbb{R}^m_+ : \|a\| = 1\}$, $\lim_{n\to+\infty} B_G(x_n, y_n)^\top \lambda = 0$, one has $\lim_{n\to+\infty} x_n = y$.

A *vector-valued Bregman distance function* $G : \mathbb{R}^n \to (\mathbb{R}^m)^\bullet$ that satisfies also the condition

(A4) for every $y \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m_+ \cap \mathcal{S}_{\mathbb{R}^m}$, there exists $x \in \mathrm{int}(\mathrm{dom}\, G)$ such that $DG(x)^\top \lambda = y$;

is said to be a *strengthened vector-valued Bregman distance function*.

On the other hand, a *vector-valued coercive viscosity function* $G : \mathbb{R}^n \to (\mathbb{R}^m)^\bullet$ is a proper, strictly C-convex and C-lower semicontinuous vector function with a closed convex domain with a nonempty interior, that is Gâteaux differentiable on the interior of its domain and whose Gâteaux derivative $DG(\cdot)$ is nonexpansive on $\mathrm{dom}\, G$ such that there exists an $x \in \mathrm{dom}\, G$ with $\|x\| < +\infty$ such that $DG(x) = 0$.

When Z is a smooth Banach space, one defines the *Lyapunov functional* as $L : Z \times Z \to \mathbb{R}_+$, defined by $L(x, y) = \|x\|^2 - 2\langle J_Z(y), x\rangle + \|y\|^2$, $x, y \in Z$. For further properties of this function that are relevant for the algorithms discussed in this paper the reader is referred to [35, Section 2].

Let $S \subseteq X$. A mapping $A : S \to X$ is said to be *monotone* when $\langle Ax - Ay, x - y\rangle \geq 0$ for all $x, y \in S$. Using it, one can define a *monotone variational inequality*

problem that consists of determining an $x \in S$ such that $\langle Ax, y - x\rangle \geq 0$ for all $y \in S$ and whose set of solutions is denoted by $VI(A, S)$.

Assume further, unless explicitly stated otherwise, that $\operatorname{int} C \neq \emptyset$. Consider the vector optimization problem

$$(VP) \qquad \operatorname*{Min}_{x \in X} F(x),$$

where $F : X \to Y^\bullet$ is a proper vector function. When $Y = \mathbb{R}^m$ we usually write $F = (F_1, \ldots, F_m)^\top$.

In case the vector minimization of F is considered subject to some nonempty subset S of X such that $\operatorname{dom} F \cap S \neq \emptyset$, we consider the vector optimization problem

$$(VPG) \qquad \operatorname*{Min}_{x \in S} F(x).$$

Later in Remark 11.33 a vector optimization problem with both geometric and equality constraints is briefly discussed, while in Section 11.5 we consider other vector optimization problems whose objective functions consist of sums or differences of (C-convex) vector functions.

In the literature one can find different solution notions for vector optimization problems. We present here the ones needed for our presentation. An element $\bar{x} \in \operatorname{dom} F$ is said to be an *efficient solution* to (VP) if there is no $x \in X$ such that $F(x) \leq_C F(\bar{x})$ and a *weakly efficient solution* to (VP) if $(F(\bar{x}) - \operatorname{int} C) \cap F(\operatorname{dom} F) = \emptyset$, respectively. We denote by $\mathcal{E}(VP)$ the *efficiency set* to (VP), i.e. set of all efficient solutions to (VP), and by $\mathcal{WE}(VP)$ the one of all weakly efficient ones, i.e. the *weak efficiency set*. Moreover, $\bar{x} \in \operatorname{dom} F$ is a *properly efficient solution (in the sense of Henig and Lampe)* to (VP) if there is a pointed closed convex cone $K \subseteq Y$ such that $C \setminus \{0\} \subseteq \operatorname{int} K$ and $(F(\operatorname{dom} F) - F(\bar{x})) \cap (-K) = \{0\}$, and we denote this by $\bar{x} \in \mathcal{PE}(VP)$. Another proper efficiency notion considered in this presentation is the following (for other types of properly efficient solutions to (VP), such as the ones due to Borwein from [25, 26], we refer to [21, Section 2.4]). We say, for $\delta \in (0, 1]$, that $\bar{x} \in \operatorname{dom} F$ is a *properly efficient solution (with respect to K_δ)* to (VP) when there exists some $z^* \in K_\delta \setminus \{0\}$ such that $\langle z^*, F(\bar{x})\rangle \leq \langle z^*, F(x)\rangle$ for all $x \in X$ and we write this $\bar{x} \in \mathcal{PE}_\delta(VP)$. The corresponding efficiency notions for (VPG) are defined analogously, by replacing X by S and $\operatorname{dom} F$ by $\operatorname{dom} F \cap S$ in the definitions. Further, when $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ and $F : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz, an element $\bar{x} \in \mathbb{R}^n$ is said to be a *(Pareto-Clarke) critically efficient solution* to (VP) if, for any direction $d \in \mathbb{R}^n$, there exists a $\bar{j} \in \{1, \ldots, m\}$, such that $F_{\bar{j}}^C(\bar{x}; d) \geq 0$. When F can be written as the difference of two C-convex functions $F_1, F_2 : \mathbb{R}^n \to \mathbb{R}^m$ the last definition collapses to the existence of a $u \in C^* \cap \mathcal{S}_{\mathbb{R}^m}$ such that $0 \in (\partial F_1(x) - \partial F_2(x))^\top u$. Theoretically nice, but without much practical relevance are the so-called *ideally efficient solutions* to (VP) in case $Y = \mathbb{R}^m$, defined as those $\bar{x} \in \operatorname{dom} F$ for which $F_j(\bar{x}) \leq F_j(x)$ for all $x \in X$, where $F = (F_1, \ldots, F_m)^\top$.

From [21, Corollary 2.4.26] one has the following characterization of the weakly efficient solutions to (VP) in the convex setting.

Lemma 11.1 *If F is also C-convex, then $\bar{x} \in \mathcal{WE}(VP)$ if and only if*

$$\exists z^* \in C^* \setminus \{0\} : \ \langle z^*, F(\bar{x})\rangle \leq \langle z^*, F(x)\rangle \ \forall x \in X.$$

Remark 11.1 As noted above, there are quite simple vector optimization problems where an unfortunate choice of the scalarizing function can often lead to unbounded scalar optimization problems. Take, for instance, the example presented in [24, Remark 1], where $X = Y = \mathbb{R}^2$, $C = \mathbb{R}^2_+$ and $F(x_1, x_2) = (x_1^2 - x_2, x_2)^\top$. For every $z^* = (z_1^*, z_2^*)^\top \in C^* = \mathbb{R}^2_+$ with $z_1^* \neq z_2^*$ the scalarized optimization problem $\inf_{x\in X}(z^*F)(x)$ is unbounded from below, hence it has no optimal solutions and is of little use in identifying the weakly efficient solutions to the original vector optimization problem. Only the scalarization functionals based on $z^* = (z_1^*, z_2^*)^\top \in \mathbb{R}^2_+ \setminus \{0\}$ with $z_1^* = z_2^*$ generate scalarized optimization problems that have optimal solutions, delivering hence weakly efficient solutions to the original vector optimization problem.

For guaranteeing the convergence of many of the algorithms that are presented in this work the following notion is necessary. Note that it is defined in a more general framework in [57, Definition 3.2]. It is followed by a weaker version needed only for the convergence of the method from [28, 38] (see Theorem 11.15 and Theorem 11.16).

Definition 11.1 (cf. [24]) Given $x \in X$, the set $F(X) \cap (F(x) - C)$ is said to be *C-complete* when for all sequences $(a^n)_n \subseteq X$ with $a^0 = x$ such that $F(a^{n+1}) \leqq_C F(a^n)$ for all $n \geq 1$ there exists an $a \in X$ such that $F(a) \leqq_C F(a^n)$ for all $n \geq 1$.

Definition 11.2 (cf. [28]) Given $x \in S$, where $S \subseteq X$ is a closed convex set, the set $F(S) \cap (F(x) - C)$ is said to be *C-quasicomplete* for S when for all sequences $(a^n)_n \subseteq X$ with $a^0 = x$ such that $F(a^{n+1}) \leqq_C F(a^n)$ for all $n \geq 1$ one has $F(a) \leqq_C F(a^n)$ for all $n \geq 1$ and all $a \in \mathcal{WE}(VP) \cap VI(S, A)$, where $A : S \to X$ is a monotone mapping.

11.3 The Original Proximal Point Type Method for Vector Optimization Problems

As mentioned above, the first work where the classical proximal point method was actually extended from scalar optimization to vector optimization problems is [24]. The algorithm introduced there, on which basically all the future contributions to this field rely on, is the following one.

Algorithm 1 *Choose the starting point $x_1 \in \operatorname{dom} F$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, \alpha]$, with $\alpha > 0$, and $(e_n)_n \subseteq \operatorname{int} C$ such that $\|e_n\| = 1$ for all $n \geq 1$. Consider the following iterative steps*

1 *let* $n = 1$*;*
2 *if* $x_n \in \mathcal{WE}(VP)$*, then* $x_{n+p} = x_n$ *for all* $p \geq 1$*;*
3 *otherwise find* $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n : x \in \Omega_n\right\}$*, where* $\Omega_n = \{x \in X : F(x) \leqq_C F(x_n)\}$*;*
4 *take* $n := n + 1$ *and go to Step 2.*

Under usual convexity and topological hypotheses applied on F one can prove the following weak convergence result.

Proposition 11.1 (cf. [24, Proposition 3.3]) *Let F be C-convex and positively C-lower semicontinuous. If the sequence $(x_n)_n$ generated by Algorithm 1 has a weak cluster point, then it is weakly convergent towards a weakly efficient solution to (VP).*

However, in order to prove the weak convergence of this algorithm towards a weakly efficient solution to (VP) regardless of the knowledge available only after running it, an additional hypothesis is necessary.

Theorem 11.1 (cf. [24, Theorem 3.1]) *Let F be C-convex and positively C-lower semicontinuous and assume that $F(X) \cap (F(x_1) - C)$ is C-complete. Then any sequence $(x_n)_n$ generated by Algorithm 1 converges weakly towards a weakly efficient solution to (VP).*

Some comments are in order.

Remark 11.2 When $Y = \mathbb{R}$ and $C = \mathbb{R}_+$ (i.e. in the scalar case), Algorithm 1 collapses to the classical proximal point method for scalar optimization problems, supporting thus the fact that it is a direct extension of the latter.

Remark 11.3 In every iteration of Algorithm 1 a different intermediate vector optimization problem is addressed, each of them having a smaller feasible set than its predecessor.

Remark 11.4 The operation that takes place in Step 3 of Algorithm 1 can be considered as a vector counterpart of determining the proximal point of a scalar function at a given point, i.e. one could call the set-valued mapping

$$v \mapsto \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2}\|x - v\|^2 e_n : x \in \Omega_v\right\}, \tag{11.1}$$

where $\Omega_v = \{x \in X : F(x) \leqq_C F(v)\}$, *vector proximal point operator*. Of course the analogy is not perfect, since the scalar (Moreau) proximal point operator is single-valued if the involved scalar function is proper, convex and lower-semicontinuous. Note also that the arg min operation within is unconstrained. However, (11.1) is at the moment the closest construction to the scalar (Moreau) proximal point operator one has in the vector case and when particularized to the scalar framework it coincides with the latter.

Remark 11.5 The construction of x_{n+1} in Algorithm 1 guarantees the decreasing monotonicity of the sequence $(F(x_n))_n$ with respect to the cone C. However, this is not enough to guarantee its convergence.

Remark 11.6 Alternate stopping rules to the one used in the formulation of Algorithm 1 can be found in [24, Remark 2 and Proposition 3.2]. Since it is usually not an easy task to verify whether $x_n \in \mathcal{WE}(VP)$, one can instead check if $x_{n+1} = x_n$.

Remark 11.7 At a first look the construction of the new iterate in Algorithm 1 contradicts the basic fact that the subproblems that are employed in an iterative process have to be simpler and more easily solvable than the original optimization problem one aims to solve with the method in discussion, as the intermediate problems have more complicated objective functions than (VP) and, on the top of it, they are constrained (in a world where the proximal point methods still lack the ability to solve general constrained optimization problems even in the scalar setting). However, any $z^* \in C^* \setminus \{0\}$ provides a suitable scalarization functional (whose existence is guaranteed by Lemma 11.1 under the hypotheses of Theorem 11.1) for the vector optimization problems in Step 3 of Algorithm 1. This endows the method with additional flexibility properties that may prove to be useful when implementing it. Moreover, even if the function

$$x \mapsto \left\langle z^*, F(x) + \frac{\alpha_n}{2}\|x - x_n\|^2 e_n \right\rangle + \delta_{\Omega_n}(x)$$

has, because it is lower semicontinuous and strongly convex, exactly one minimum that is x_{n+1}, the sequence $(x_n)_n$ is not uniquely determined because for each choice of $z^* \in C^* \setminus \{0\}$ one deals with a different such function. This does not mean that the vector optimization problem (VP) is a priori scalarized by means of a linear continuous functional, because this scalarization is applied to the intermediate vector optimization problems not to (VP).

Remark 11.8 Different to the classical proximal point method, in the convergence statement Theorem 11.1 it is not necessary to assume the existence of a solution of the considered optimization problem, i.e. a weakly efficient solution to (VP), in order to prove the convergence of Algorithm 1. The role of such a hypothesis in showing the convergence of the method has been fully covered in the proof of Theorem 11.1 (see [24, Theorem 3.1]) by the assumed C-completeness hypothesis. Considering the former instead of the latter, the role of $\cap_{n\geq 1}\Omega_n$ would be taken by $\mathcal{WE}(VP)$ in the proof of Theorem 11.1 (i.e. [24, Theorem 3.1]). However, then is the inclusion $\mathcal{WE}(VP) \subseteq \Omega_n$ for all $n \geq 1$ not obviously guaranteed by construction and should be separately investigated (or imposed). Note moreover that assuming the existence of some $\bar{x} \in \mathcal{WE}(VP)$ does not automatically deliver the corresponding scalarizing parameter $\bar{z}^*$ that exists according to Lemma 11.1, which would probably be needed in formulating the algorithm under the mentioned hypothesis.

Remark 11.9 In the scalar setting (i.e. when $Y = \mathbb{R}$ and $C = \mathbb{R}_+$), when the set of minimizers of a function $f : X \to \overline{\mathbb{R}}$ is nonempty, for every $x_1 \in \operatorname{dom} F$ the intersection $(f(x_1) - \mathbb{R}_+) \cap f(X)$ is $\mathbb{R}_+$-complete, i.e. the C-completeness hypothesis of Theorem 11.1 is always satisfied. However, in even slightly more complex frameworks this hypothesis is no longer automatically valid and its eventual fulfillment is not always easy to verify. Sufficient conditions for guaranteeing that $F(X) \cap (F(x_1) - C)$ is C-complete were proposed in [57, Lemma 3.5], namely

- the set $(F(x_1) - C) \cap F(X)$ is compact;
- the set $(F(x_1) - C) \cap F(X)$ is weakly compact;
- the set $(F(x_1) - C) \cap F(X)$ is closed and has a lower bound and the cone C has the *Daniell property* (i.e., any decreasing net having a lower bound converges to its infimum).

On the other hand, it could be interesting to investigate whether the weaker hypothesis of C-quasicompleteness imposed on $F(X) \cap (F(x_1) - C)$ in Theorem 11.15 and Theorem 11.16 could prove to be sufficient for convergence for other algorithms as well.

Remark 11.10 For determining the optimal solutions of the scalarized optimization problems attached to the vector optimization problems in Step 3 of Algorithm 1 one can try to employ a splitting type algorithm designed for finding the optimal solutions of optimization problems consisting in minimizing sums of convex functions, like the ones proposed in [9, 18, 20, 22]. However, the processing of the functions δ_{Ω_n}, $n \geq 1$, may prove to be quite difficult, due to the special structure of the sets Ω_n, $n \geq 1$. A way to go round this nuisance is, as seen in [4, 19, 52, 53, 67–69], by employing some other algorithms for solving the intermediate scalar optimization problems, for instance one based on interior point methods.

In [24, Section 4] it is discussed about the additional hypotheses needed by Algorithm 1 in order to deliver efficient solutions to (VP) instead of weakly efficient ones. In this case it is not necessary to have $\operatorname{int} C \neq \emptyset$.

Theorem 11.2 (cf. [24, Theorem 4.1]) *Let F be C-convex and positively C-lower semicontinuous and assume that $F(X) \cap (F(x_1) - C)$ is C-complete and that there exists some $\delta > 0$ such that the set K_δ is nonempty. Then any sequence $(x_n)_n$ generated by Algorithm 1 with the selection*

$$x_{n+1} = \operatorname*{arg\,min}_{x \in \Omega_n} \langle z_n^*, F(x) + \frac{\alpha_n}{2} \|x - x_n\|^2 e_n \rangle,$$

where $(z_n^*)_n \subseteq K_\delta$, converges weakly towards an efficient solution to (VP).

Last but not least, an inexact version of Algorithm 1 is proposed in [24, Section 5] for the purpose of providing an implementable iterative scheme, that, however, was not concretely tested on an example or an application. As the authors put it, the Step 3 in Algorithm 2 is formulated in a scalar manner in order to avoid

unnecessary complications, but x_{n+1} could be as well taken as some approximate weakly efficient solution to the corresponding intermediate vector optimization problem in Algorithm 1.

Algorithm 2 *Choose the starting point* $x_1 \in \operatorname{dom} F$*, the relative error tolerance* $\sigma \in [0, 1)$ *and the exogenous sequences* $(\alpha_n)_n \subseteq (0, \alpha]$*, with* $\alpha > 0$*,* $(e_n)_n \subseteq \operatorname{int} C$ *such that* $\|e_n\| = 1$ *for all* $n \geq 1$ *and* $(z_n^*)_n \in C^*$ *such that* $\|z_n^*\| = 1$ *for all* $n \geq 1$*. Consider the following iterative steps*

1 *let* $n = 1$*;*
2 *if* $x_n \in \mathcal{E}(VP)$*, then* $x_{n+p} = x_n$ *for all* $p \geq 1$*;*
3 *otherwise find* $x_{n+1} \in X$ *as a solution of the inclusion* $0 \in \partial_{\varepsilon_n} \langle z_n^*, F(x) + \delta_{\Omega_n} \rangle + \alpha_n \langle z_n^*, e_n \rangle (x - x_n)$*, where* $\varepsilon_n \leq \sigma \frac{\alpha_n}{2} \langle z_n^*, e_n \rangle \|x - x_n\|^2$*;*
4 *take* $n := n + 1$ *and go to Step 2.*

The convergence of Algorithm 2 is obtained in a similar manner to the one of Algorithm 1 and in [24, Remark 6] it is stated that one can also obtain efficient solutions to (VP) under the additional hypotheses from Theorem 11.2.

Theorem 11.3 (cf. [24, Theorem 5.1]) *Let* F *be* C*-convex and positively* C*-lower semicontinuous and assume that* $F(X) \cap (F(x_1) - C)$ *is* C*-complete. Then any sequence* $(x_n)_n$ *generated by Algorithm 2 converges weakly towards a weakly efficient solution to* (VP)*.*

Remark 11.11 As noted in [19, Remark 11], vector optimization problems with the ordering cones of the image spaces having empty interiors, but nonempty generalized interiors can be found in finance mathematics (see, for instance, [1, 44]) and other research fields. This has motivated the weakening of the definition of the weakly efficient solutions to (VP) (cf. [44, 49, 50]) for the case when $\operatorname{int} C = \emptyset$ by replacing this hypothesis with the nonemptiness of the *quasi interior* of C (i.e. the set of all $y \in Y$ such that $\operatorname{cl}(\operatorname{cone}(V - y)) = Y$). In order to characterize these more general weakly efficient solutions to (VP) one can use [50, Corollary 9] instead of Lemma 11.1. However, since the key result [23, Lemma 2.2] does not hold in case $\operatorname{int} C = \emptyset$, the proof of the algorithm convergence statement Theorem 11.1 has to be modified, for instance, by scalarizing all the subproblems with the same $\bar{z}^* \in C^* \setminus \{0\}$. On the other hand, in finitely dimensional spaces so-called *relatively weakly efficient solutions* to (VP) can be defined when C has an empty interior but a nonempty relative interior and they can be characterized by means of linear scalarization (cf. [49]) while the impediment mentioned above does not occur due to the equivalence of the corresponding weak and strong topologies.

Remark 11.12 Algorithm 1 is applied in [31] for vector-minimizing in the finitely dimensional setting where $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ the composition of a vector function with a linear continuous mapping subject to a geometric constraint. We have not included this method in Section 11.5 because the author merely replaced in Algorithm 1 the vector function $F : \mathbb{R}^p \to (\mathbb{R}^m)^\bullet$ with its composition with the considered linear continuous mapping $A : \mathbb{R}^n \to \mathbb{R}^p$ and did not apply some splitting method in order to process F and A separately during the iterative

process. In order to guarantee the convergence of the method towards a weakly efficient solution to the considered vector optimization problem no C-completeness hypothesis is employed, however the domination property of the image of the objective vector function is imposed and the weak efficiency set is taken to be compact.

Remark 11.13 In the literature one can find contributions where the convergence of Algorithm 1 (or of its inexact version) is proven under different hypotheses than the ones of Theorem 11.1, in the sense that the objective function F of the considered vector optimization problem needs not be C-convex in order to achieve the desired result. In the following we discuss briefly the results of [4, 12, 14, 15, 69]. Consider the finitely dimensional setting with $X = \mathbb{R}^n$, $Y = \mathbb{R}^m$ and, moreover, $C = \mathbb{R}^m_+$. In [12] Algorithm 1 (with the difference that the elements of the sequence $(x_n)_n$ are additionally asked to lie in S) is employed for solving (VPG) (the geometrically constrained counterpart of (VP)). The components of the vector function F are asked to be locally Lipschitz and, under some additional hypotheses imposed on the involved sequences it is shown in [12, Theorem 3] that every cluster point of $(x_n)_n$ is a critically efficient solution to (VPG). Taking the components of F to be quasiconvex and imposing the $\mathbb{R}^m_+$-completeness hypothesis, the convergence of the generated sequence $(x_n)_n$ towards a critically efficient solution to (VPG) is achieved. An application to the compromise solution problem is discussed, however no numerical results are provided. The same algorithm is employed in [15] for vector minimizing a vector function $F : \mathbb{R}^n \to \mathbb{R}^m$ whose components are maxima of continuously differentiable functions over some given subset of $\mathbb{R}^n$. In [15, Theorem 1] the convergence of the iterative method is investigated and it is shown that each cluster point of the generated sequence is a critically efficient solution to (VPG). Under additional hypotheses the convergence of the generated sequence towards a weakly efficient solution to (VPG) is achieved. Further, in [14] a proximal point type algorithm that solves at each step a scalarized version (by means of a special case of the Tammer scalarization function - the so-called maximum scalarization, see also Remark 11.31 and Remark 11.32) of the corresponding intermediate vector optimization problem from Algorithm 1 is proposed, with the convergence of the generated sequence towards a critically efficient solution to (VP) shown under quasiconvexity hypotheses imposed on the components of F in [14, Theorem 4.1]. It is also shown that under stronger assumptions the sequence can converge towards a weakly efficient solution to (VP) and even an efficient one in [14, Theorem 4.3], while an application to behavioural science is also discussed. Last but not least in [4, 69] one finds proximal point type algorithms for solving (VP) with the components of F taken quasiconvex, where the iterations are defined as zeros of subdifferential inclusions, in the vein of Algorithm 2 but without the error sequence $(\varepsilon_n)_n$ and with the Clarke and Fréchet subdifferential, respectively, instead of the convex one. The convergence of the generated sequence is guaranteed towards critically efficient solutions (and, under additional convexity hypotheses imposed on the components of F, also towards weakly efficient solutions) to (VP) in [4] and towards a minimizing set (and, under additional continuity hypotheses imposed on the components of F, also towards weakly efficient, and, if these are

taken also convex, even efficient solutions) to (VP). Worth noticing, however, is that in both these papers computational results obtained in MATLAB are presented, too. Note also that in [4] one can find an application to location theory, while in [69] an inexact version of the considered algorithm and an application to consumer demand theory are given as well.

11.4 Modifications and Extensions of the Original Method

Several authors have proposed various modifications and extensions of the proximal point type method introduced in [24]. Besides the extensions towards nonconvex vector optimization problems mentioned in Remark 11.13, there are algorithms where the norm distance from the classical proximal point iteration is replaced by a quasi-distance, a Bregman distance, a Lyapunov distance or is formulated by means of a viscosity function and proximal point type algorithms with inertial/memory effects or hybrid constructions. Generalizations of the algorithm from [24] towards Hadamard and Banach spaces were proposed, too. We also briefly mention some contributions where the regularization is performed by means of a Tikhonov type function instead of the Moreau-Yosida type one from the proximal point algorithms.

We begin with algorithms where the classical distance expressed via a norm is replaced by a generalization of it.

11.4.1 Algorithms with Bregman-Type Distances

In [75], a so-called interior proximal method is proposed for solving a geometrically constrained version of (VP) in finitely-dimensional spaces, i.e. (VPG), where $F : \mathbb{R}^n \to (\mathbb{R}^m)^\bullet$ is a proper vector function and $S \subseteq \mathbb{R}^n$ is a closed convex set with nonempty interior. The algorithm employs a proximal distance d with respect to $\operatorname{int} S$ and before formulating it one assumes that $S \subseteq \operatorname{dom} F$.

Algorithm 3 *Choose the starting point $x_1 \in \operatorname{int} S$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, \alpha]$, with $\alpha > 0$, and $(e_n)_n \subseteq \operatorname{int} C$ such that $\|e_n\| = 1$ for all $n \geq 1$. Consider the following iterative steps*

1. *let $n = 1$;*
2. *if $x_n \in \mathcal{WE}(VPG)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3. *otherwise find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \alpha_n d(x, x_n)e_n : x \in \Omega_n\right\}$;*
4. *take $n := n + 1$ and go to Step 2.*

For the statement on convergence towards a weakly efficient solution to (VPG), the following additional hypotheses are required

(B1) $\exists \tilde{z} \in C^* \setminus \{0\}$ such that $\tilde{z}^\top F(x) > -\infty$ for all $x \in S$;
(B2) $d(\cdot, u)$ is coercive for any $u \in S$ in $\mathbb{R}^n$.

Theorem 11.4 (cf. [75, Theorem 4.1]) *Let F be C-convex and positively C-lower semicontinuous and assume that $F(S) \cap (F(x_1) - C)$ is C-complete, the conditions $(B1) - (B2)$ hold and $(d, H) \in \mathcal{F}^*(\operatorname{cl} S)$. Then any sequence $(x_n)_n$ generated by Algorithm 3 converges towards a weakly efficient solution to (VPG).*

Remark 11.14 In [75, Section 5] an inexact version of Algorithm 3 was proposed, too, with the corresponding convergence towards a weakly efficient solution to (VPG) obtained in [75, Theorem 5.1] under some additional hypotheses.

Remark 11.15 Another inexact algorithm based on a proximal distance was proposed in [17] for vector minimizing a vector function $F : \mathbb{R}^n \to \mathbb{R}^m$ whose components are maxima of continuously differentiable functions over some given subset of $\mathbb{R}^n$, like the one treated in [15]. In [17, Theorem 4.1] the convergence of the iterative method is investigated and it is shown that each cluster point of the generated sequence is a critically efficient solution to (VPG). An application to a problem of distributive justice is also presented together with some ideas for future research.

A somehow similar algorithm was proposed in [33] for solving (VPG), with the difference that the authors take $F : S \to (\mathbb{R}^m)^\bullet$ and assume that $\operatorname{int} S \cap \operatorname{dom} F \neq \emptyset$. In order to introduce it, one needs to consider a strengthened vector-valued Bregman distance function $G : \mathbb{R}^n \to (\mathbb{R}^m)^\bullet$.

Algorithm 4 *Assume that a starting point $x_1 \in \operatorname{int} S \cap \operatorname{dom} F$ such that $\{x \in \mathbb{R}^n : F(x) \leqq_C F(x_1)\} \subseteq \operatorname{dom} F \cap \operatorname{int}(\operatorname{dom} G)$ exists and choose the exogenous sequence $(\alpha_n)_n \subseteq (0, \alpha]$, with $\alpha > 0$. Consider the following iterative steps*

1 let $n = 1$;
2 if $x_n \in \mathcal{WE}(VPG)$, then $x_{n+p} = x_n$ for all $p \geq 1$;
3 otherwise find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2} B_G(x, x_n) : x \in \Omega_n\right\}$;
4 take $n := n + 1$ and go to Step 2.

The statement on convergence of the sequence generated via Algorithm 4 towards a weakly efficient solution to (VPG) follows. Note that, unlike its counterparts presented above, it requires no C-completeness in order to achieve the convergence, however the domination property of the image of the objective vector function is imposed.

Theorem 11.5 (cf. [33, Theorem 3.3]) *Let F be C-convex and C-lower semicontinuous such that $F(\mathbb{R}^n)$ has the domination property, $\mathcal{WE}(VPG)$ is nonempty and compact and DG is norm-to-norm continuous. Then any sequence $(x_n)_n$ generated by Algorithm 4 converges towards a weakly efficient solution to (VPG).*

Further generalizations of the classical distance function can be found in the class of the so-called quasi-distances. Successfully employed in proximal point type algorithms, they ended up being used in iterative methods for determining weakly efficient solutions to vector optimization problems, too, for instance in [72], as

follows, where $F = (F_1, \ldots, F_m)^\top : \mathbb{R}^n \to \mathbb{R}^m$, $C = \mathbb{R}^m_+$ and $q : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ is a quasi-distance.

Algorithm 5 *Choose the starting point $x_1 \in \mathbb{R}^n$ and the exogenous sequence $(\alpha_n)_n \subseteq (\eta, \alpha]$, with $0 < \eta < \alpha$. Consider the following iterative steps*

1. *let $n = 1$;*
2. *if $x_n \in \mathcal{WE}(VP)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3. *otherwise find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2} q^2(x, x_n) : x \in \Omega_n\right\}$;*
4. *take $n := n + 1$ and go to Step 2.*

The convergence statement of Algorithm 5 follows, noting that the uniqueness of the cluster point of the sequence generated by the method is not guaranteed. Worth noticing is also that the considered hypotheses imply the usual C-completeness assumption needed in the rest of the paper for guaranteeing the convergence of the considered algorithms.

Theorem 11.6 (cf. [72, Theorem 3.1]) *Let F_i be convex, $i = 1, \ldots, m$, and assume that there is some $j \in \{1, \ldots, m\}$ such that $\lim_{\|x\| \to +\infty} F_j(x) = +\infty$. Suppose there are positive constants a and b such that $a\|x - y\| \leq q(x, y) \leq b\|x - y\|$ for any $x, y \in \mathbb{R}^n$. Then any cluster point of any sequence $(x_n)_n$ generated by Algorithm 5 is a weakly efficient solution to (VP).*

Remark 11.16 Computational results obtained by implementing Algorithm 5 in MATLAB are presented in [72, Section 4].

Remark 11.17 An inexact version of Algorithm 5 was proposed in [70] by means of the limiting subdifferential. The usage of this nonsmooth subdifferential instead of the classical convex one is justified by the fact that even if F is taken $\mathbb{R}^m_+$-convex, the quasi-distance needs not be convex.

11.4.2 Algorithms with Viscosity Functions and Tikhonov Type Regularizations

In order to present the proximal point type algorithms for solving vector optimization problems where the regularization is done by means of viscosity functions that were introduced in [30, 39] one needs to take $X = \mathbb{R}^n$ and $Y = \mathbb{R}^m$ within this subsection. Moreover, let $G : \mathbb{R}^n \to \mathbb{R}^m$ be a vector-valued coercive viscosity function such that $\operatorname{dom} F \cap \operatorname{int}(\operatorname{dom} G) \neq \emptyset$.

We begin with the method proposed in [30] for determining weakly efficient solutions to (VP).

Algorithm 6 *Choose the starting point $x_1 \in \operatorname{dom} F \cap \operatorname{int}(\operatorname{dom} G)$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, +\infty)$, with $\lim_{n \to +\infty} \alpha_n = +0$, and $(\beta_n)_n \subseteq [0, 1]$ with $\lim_{n \to +\infty} \beta_n = 0$. Consider the following iterative steps*

1 let $n = 1$;
2 if $x_n \in \mathcal{WE}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 otherwise find $y_{n+1} \in \mathcal{WE}\left\{F(x) + \alpha_n G(x) : x \in \Omega_n\right\}$;
4 take $x_{n+1} = (1 - \beta_n)y_n + \beta_n x_n$;
5 take $n := n + 1$ *and go to Step 2.*

The convergence statement of Algorithm 6 follows. Note that the constructions considered in this subsection do not require the usual C-completeness hypothesis needed in the rest of the paper for guaranteeing the convergence of the considered algorithms, its role being covered by asking $\mathcal{WE}(VP)$ to be nonempty and compact. Connected to this issue see also the discussion in Remark 11.8.

Theorem 11.7 (cf. [30, Theorem 3.3]) *Let F be C-convex, C-lower semicontinuous and C^*-asymptotically uniformly continuous such that $\mathcal{WE}(VP)$ is nonempty and compact. Further suppose that there exist a sequence $(\eta_n)_n \in \mathbb{R}$ such that $\|y_n - x_n\| \leq \eta_n$ for all $n \geq 1$ and $\sum_{n\geq 1} \eta_n < +\infty$, where $(x_n)_n$ and $(y_n)_n$ are sequences generated by Algorithm 6. Then $(x_n)_n$ converges towards a weakly efficient solution to (VP).*

Remark 11.18 An legitimate question regarding Theorem 11.7 concerns the necessity of both topological assumptions imposed on F, the C-lower semicontinuity and the C^*-asymptotically uniform continuity in order to achieve the statement. However, at the moment we are unaware of any result connecting or comparing these two notions in some way.

A modification of Algorithm 6 proposed in [39] guarantees the convergence of the generated sequence towards an efficient solution to (VP). More precisely, instead of determining weakly efficient solutions of the intermediate problems, one looks in Step 3 for properly efficient ones with respect to K_δ. In this case it is not necessary to have $\operatorname{int} C \neq \emptyset$.

Algorithm 7 *Choose the starting point $x_1 \in \operatorname{dom} F \cap \operatorname{int}(\operatorname{dom} G)$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, +\infty)$, with $\lim_{n\to+\infty} \alpha_n = +0$, and $(\beta_n)_n \subseteq [0, 1]$ with $\lim_{n\to+\infty} \beta_n = 0$. Assume moreover that there exists some $\delta \in (0, 1]$ such that the set K_δ is nonempty. Consider the following iterative steps*

1 let $n = 1$;
2 if $x_n \in \mathcal{E}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 otherwise find $y_{n+1} \in \mathcal{PE}_\delta\left\{F(x) + \alpha_n G(x) : x \in \Omega_n\right\}$;
4 take $x_{n+1} = (1 - \beta_n)y_n + \beta_n x_n$;
5 take $n := n + 1$ *and go to Step 2.*

The convergence statement [39, Theorem 3.4] is different to the others presented so far since it does not guarantee the convergence of the sequence generated by Algorithm 7 (towards an efficient solution to (VP)). Adding to it an additional hypothesis from Theorem 11.7 guarantees the unicity of the cluster point of this sequence.

Theorem 11.8 (cf. [39, Theorem 3.4]) *Let F be C-convex, C-lower semicontinuous, positively partially continuous and C^*-asymptotically uniformly continuous such that $\mathcal{WE}(VP)$ is nonempty and compact. Further suppose that there exist a sequence $(\alpha_n)_n \in \mathbb{R}$ such that $\|y_n - x_n\| \leq \alpha_n$ for all $n \geq 1$ and $\sum_{n\geq 1} \alpha_n < +\infty$, where $(x_n)_n$ and $(y_n)_n$ are sequences generated by Algorithm 7. Then $(x_n)_n$ converges towards an efficient solution to (VP).*

Remark 11.19 Examining the proof of [39, Theorem 3.4] (and that of [37, Theorem 3.2], cited below as Theorem 11.13), one can note that the usage of the property of positive C-lower semicontinuity of F can be covered by its positive partial continuity, so the firstly mentioned hypothesis seems to be redundant. On the other hand, at least at a first look, the role of the second hypothesis in the proof is not crucial and it can be replaced by the first one as well. Connected to this issue, a discussion on whether a positively partially continuous vector function is in general also positively C-lower semicontinuous could prove to be interesting as well, taking also in consideration that positively partially continuous vector functions (at least under this name) can be found only in [36, 37, 39].

Related to Algorithm 6 and Algorithm 7 from which special cases can be derived when $\beta_n = 0$ for all $n \geq 1$ and $G = \|\cdot\|_{\tilde{e}}$, for some suitable $\tilde{e} \in \mathbb{R}^m_+$ (respectively $\tilde{e} \in C$) are the methods proposed in [34, 36] where the regularization in the iterative steps is performed by means of a Tikhonov type function instead of the Moreau-Yosida type one from the proximal point algorithms.

The algorithm proposed in [34] for determining weakly efficient solutions to (VPG) is the following, where $\operatorname{dom} F \cap \operatorname{int} S \neq \emptyset$.

Algorithm 8 *Choose the starting point $x_1 \in \operatorname{dom} F \cap \operatorname{int} S$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, +\infty)$, with $\lim_{n\to+\infty} \alpha_n = +0$, and $(e_n)_n \subseteq \mathbb{R}^m_+$ with $\|e_n\| = 1$ for all $n \geq 1$. Consider the following iterative steps*

1. *let $n = 1$;*
2. *if $x_n \in \mathcal{WE}(VGP)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3. *otherwise find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \alpha_n \|x\|^2 e_n : x \in \Omega_n\right\}$;*
4. *take $n := n + 1$ and go to Step 2.*

Theorem 11.9 (cf. [34, Theorem 3.2]) *Let F be C-convex and C-lower semicontinuous, S be closed and convex, and $\mathcal{WE}(VPG)$ nonempty and compact. Then any sequence $(x_n)_n$ generated by Algorithm 8 converges towards a weakly efficient solution to (VPG).*

Finally, we also present the algorithm proposed in [36] for determining efficient solutions to (VP). In this case it is not necessary to have $\operatorname{int} C \neq \emptyset$.

Algorithm 9 *Choose the starting point $x_1 \in \operatorname{dom} F$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, +\infty)$, with $\lim_{n\to+\infty} \alpha_n = 0$, and $(e_n)_n \subseteq C$ with $\|e_n\| = 1$ for all $n \geq 1$. Assume moreover that there exists some $\delta \in (0, 1]$ such that the set K_δ is nonempty. Consider the following iterative steps*

1 let $n = 1$;
2 if $x_n \in \mathcal{E}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 otherwise find $x_{n+1} \in \mathcal{PE}_\delta \left\{ F(x) + \alpha_n \|x\|^2 e_n : x \in \Omega_n \right\}$;
4 take $n := n + 1$ *and go to Step 2.*

The corresponding convergence statement follows, although in a different manner than many of the other ones presented in this survey, as the unicity of the cluster point of the sequence generated by Algorithm 9 is not guaranteed. The idea presented in Remark 11.19 is valid for this result as well.

Theorem 11.10 (cf. [36, Theorem 3.1]) *Let F be C-convex, C-lower semicontinuous and positively partially continuous, and $\mathcal{E}(VP)$ is nonempty and compact. Then any sequence $(x_n)_n$ generated by Algorithm 9 converges towards an efficient solution to (VP).*

11.4.3 Algorithms with Lyapunov-Type Distances

A more general framework, even than the one in [24], is considered in the papers [32, 35, 37], where algorithms of proximal point type for solving (VP) are formulated by means of the Lyapunov functional. In this subsection let $(X, \|\cdot\|)$ be a uniformly convex and uniformly smooth Banach space. We begin with the algorithm proposed in [35].

Algorithm 10 *Choose the starting point $x_1 \in \operatorname{dom} F$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, \alpha]$, with $\alpha > 0$, and $(e_n)_n \subseteq \operatorname{int} C$ such that $\|e_n\| = 1$ for all $n \geq 1$. Consider the following iterative steps*

1 let $n = 1$;
2 if $x_n \in \mathcal{WE}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 otherwise find $x_{n+1} \in \mathcal{WE} \left\{ F(x) + \frac{\alpha_n}{2} L(x, x_n) e_n : x \in \Omega_n \right\}$;
4 take $n := n + 1$ *and go to Step 2.*

Because of the more general framework, the convergence of the method towards a weakly efficient solution to (VP) can be guaranteed under some additional hypotheses to the ones in Theorem 11.1.

Theorem 11.11 (cf. [35, Theorem 3.5]) *Let F be C-convex and positively C-lower semicontinuous and assume that $F(S) \cap (F(x_1) - C)$ is C-complete, $\mathcal{WE}(VP)$ is nonempty and J_X is weak-to-weak continuous. Then any sequence $(x_n)_n$ generated by Algorithm 10 converges weakly towards a weakly efficient solution to (VP).*

Remark 11.20 In the proof of [35, Theorem 3.5] it is claimed that the weak limit of the sequence generated by Algorithm 10 under the hypotheses of Theorem 11.11 was the only weakly efficient solution to (VP). However, nothing supports this fact, as there it is shown only that any such sequence has a unique weak cluster point.

Moreover, the hypothesis of nonemptiness of $\mathcal{WE}(VP)$ in Theorem 11.11 seems, in the light of Remark 11.8, superfluous.

Closely related to this method is the approximate one proposed in [32].

Algorithm 11 *Choose the starting point $x_1 \in \operatorname{dom} F$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, \alpha]$, with $\alpha > 0$, and $(e_n)_n \subseteq \operatorname{int} C$ such that $\|e_n\| = 1$ for all $n \geq 1$, as well as the error sequence $(\epsilon_n)_n \subseteq X^*$ satisfying $\sum_{n\geq 1} \|\epsilon_n\|_* < +\infty$. Consider the following iterative steps*

1 *let $n = 1$;*
2 *if $x_n \in \mathcal{WE}(VP)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3 *otherwise find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2}(L(x, x_n) - \langle \epsilon_{n+1}, x\rangle)e_n : x \in \Omega_n\right\}$;*
4 *take $n := n + 1$ and go to Step 2.*

The corresponding convergence statement is similar to Theorem 11.11 to which a hypothesis regarding the error sequence and involving the generated sequence is added. Note that Remark 11.20 applies for the following statement, too.

Theorem 11.12 (cf. [32, Theorem 3.5]) *Let F be C-convex and positively C-lower semicontinuous and assume that $F(S) \cap (F(x_1) - C)$ is C-complete, $\mathcal{WE}(VPG)$ is nonempty and J_X is weak-to-weak continuous. Then any sequence $(x_n)_n$ generated by Algorithm 11 converges weakly towards a weakly efficient solution to (VP) when $\sum_{n\geq 1}\langle \epsilon_n, x_n\rangle$ exists and is finite.*

A modification of Algorithm 11 proposed in [37] delivers efficient solutions to (VP). In this case it is not necessary to have $\operatorname{int} C \neq \emptyset$.

Algorithm 12 *Choose the starting point $x_1 \in \operatorname{dom} F$ and the exogenous sequences $(\alpha_n)_n \subseteq (0, \alpha]$, with $\alpha > 0$, and $(e_n)_n \subseteq \operatorname{int} C$ such that $\|e_n\| = 1$ for all $n \geq 1$, as well as the error sequence $(\epsilon_n)_n \subseteq X^*$ satisfying $\sum_{n\geq 1} \|\epsilon_n\|_* < +\infty$. Assume moreover that there exists some $\delta > 0$ such that the set K_δ is nonempty. Consider the following iterative steps*

1 *let $n = 1$;*
2 *if $x_n \in \mathcal{E}(VP)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3 *otherwise find $x_{n+1} \in \mathcal{PE}_\delta\left\{F(x) + \frac{\alpha_n}{2}(L(x, x_n) - \langle \epsilon_{n+1}, x\rangle)e_n : x \in \Omega_n\right\}$;*
4 *take $n := n + 1$ and go to Step 2.*

The convergence statement contains both the hypotheses of Theorem 11.12 and the additional requirement imposed on F to be positively partially continuous.

Theorem 11.13 (cf. [37, Theorem 3.2]) *Let F be C-convex, positively C-lower semicontinuous and positively partially continuous, and assume that $F(S) \cap (F(x_1) - C)$ is C-complete, and J_X is weak-to-weak continuous. Then any sequence $(x_n)_n$ generated by Algorithm 12 converges weakly towards an efficient solution to (VP) when $\sum_{n\geq 1}\langle \epsilon_n, x_n\rangle$ exists and is finite.*

11.4.4 Algorithms with Hybrid and Inertial Steps

In this subsection we have gathered some algorithms for solving vector optimization problems where the proximal point steps are combined with other ideas, leading to so-called hybrid methods and inertial type algorithms.

In [29] one can find the following method for determining weakly efficient solutions to vector optimization problems.

Algorithm 13 *Choose the starting point* $x_1 \in \operatorname{dom} F$ *and the sequences* $(\alpha_n)_n \subseteq (0, \alpha)$, *where* $\alpha > 0$, $(\beta_n)_n \subseteq [0, 1]$, $(\theta_n)_n \subseteq X$ *and* $(e_n)_n \subseteq \operatorname{int} C$ *such that* $\|e_n\| = 1$ *for all* $n \geq 1$. *Consider the following iterative steps*

1 *let* $n = 1$;
2 *if* $x_n \in \mathcal{WE}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 *find* $y_n \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2}\|x - x_n - \theta_n\|^2 e_n : x \in \Omega_n\right\}$;
4 *take* $x_{n+1} = \beta_n x_n + (1 - \beta_n) y_n$;
5 *take* $n := n + 1$ *and go to Step 2.*

Different to Algorithm 1 is not only the fact that an additional iterative sequence $(y_n)_n$ was employed in order to define the one that will converge towards a weakly efficient solution to (VP) (as seen below), but also the usage of another sequence $(\theta_n)_n$ in the proximal step. The corresponding convergence statement follows, with a dynamic condition that cannot be verified before running the algorithm. Note also that the comment from Remark 11.18 applicable here, too.

Theorem 11.14 (cf. [29, Theorem 3.1]) *Let* F *be* C*-convex, positively* C*-lower semicontinuous and* C^**-asymptotically uniformly continuous and assume that* $F(S) \cap (F(x_1) - C)$ *is* C*-complete. Let the sequences* $(x_n)_n$ *and* $(y_n)_n$ *be generated by Algorithm 13. If there is a bounded sequence* $(\eta_n)_n \subseteq (0, +\infty)$ *such that* $\sum_{n\geq 1} \eta_n^2 < +\infty$ *and* $\|\theta_n\| \leq \eta_n \|x_n - y_n\|$ *and a constant* $\delta \in (0, 1)$ *such that* $0 \leq \beta_n \leq 1 - \delta$ *for all* $n \geq 1$, *and it holds* $\lim_{n\to+\infty} \beta_n = 0$, *then* $(x_n)_n$ *converges weakly towards a weakly efficient solution to* (VP).

In [29] one can find also a modification of Algorithm 13 where the usual distance is replaced by a Bregman type one as follows, where $h : X \to \mathbb{R}$ is a strictly convex function that is Gâteaux differentiable with the Gâteaux derivative $Dh(\cdot)$ and $B_h : X \times X \to \mathbb{R}$ is the corresponding Bregman distance with respect to h, namely $B_h(x, y) = h(x) - h(y) - Dh(y)(x - y)$.

Algorithm 14 *Choose the starting point* $x_1 \in \operatorname{dom} F$ *and the sequences* $(\alpha_n)_n \subseteq (0, \alpha)$, *where* $\alpha > 0$, $(\beta_n)_n \subseteq [0, 1]$, $(\theta_n)_n \subseteq X$ *and* $(e_n)_n \subseteq \operatorname{int} C$ *such that* $\|e_n\| = 1$ *for all* $n \geq 1$. *Consider the following iterative steps*

1 *let* $n = 1$;
2 *if* $x_n \in \mathcal{WE}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 *find* $x_n \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2}(2h(x) + \|x - Dh(x_n) - \theta_n\|^2) e_n : x \in \Omega_n\right\}$;
4 *take* $n := n + 1$ *and go to Step 2.*

Remark 11.21 The weak convergence of Algorithm 14 towards a weakly efficient solution to (VP) is achieved in [29, Theorem 4.1]. Different to Theorem 11.14, the hypothesis of C^*-asymptotically uniform continuity of F seems no longer necessary, however Dh is required to be weak-to-weak sequentially continuous and the boundedness condition imposed on the sequence $(\theta_n)_n$ is replaced with $\|\theta_n\| \leq \eta_n D_h^{1/2}(x_{n+1}, x_n)$.

Remark 11.22 In [29] one can also find an inexact version of Algorithm 13 called relative approximate proximal algorithm, whose weak convergence towards a weakly efficient solution to (VP) is obtained in [29, Theorem 5.1]. Interestingly, the hypothesis of C^*-asymptotically uniform continuity of F seems no longer necessary for this statement either.

A further development of Algorithm 13 can be found in [28] in the form of a hybrid proximal point type algorithm for finding weakly efficient solutions to (VPG), where the iterative steps contain projections and involve monotone mappings, while a variational inequality is involved in the convergence statement. Note that in this case actually four sequences are generated during the iterative process in order to construct the one that actually converges towards a weakly efficient solutions to (VPG). Let S be closed and convex and $A : S \to X$ be monotone.

Algorithm 15 *Choose the starting point* $x_1 \in \operatorname{dom} F \cap S$ *and the sequences* $(\alpha_n)_n \subseteq (0, \alpha)$, *where* $\alpha > 0$, $(\lambda_n)_n \subseteq (0, 1)$, $(\beta_n)_n \subseteq [0, 1]$, $(\eta_n)_n \subseteq [0, 1]$, $(\theta_n)_n \subseteq X$ *and* $(e_n)_n \subseteq \operatorname{int} C$ *such that* $\|e_n\| = 1$ *for all* $n \geq 1$. *Consider the following iterative steps*

1 *let* $n = 1$;
2 *if* $x_n \in \mathcal{WE}(VPG)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 *take* $y_n = \operatorname{Pr}_S(x_n - \lambda_n A x_n)$;
4 *take* $z_n = \eta_n x_n + (1 - \eta_n) \operatorname{Pr}_S(x_n - \lambda_n A y_n)$;
5 *find* $w_n \in \mathcal{WE}\left\{F(x) + \frac{\alpha_n}{2}\|x - z_n - \theta_n\|^2 e_n : x \in \Omega_n\right\}$;
6 *take* $x_{n+1} = \beta_n x_n + (1 - \beta_n) w_n$;
7 *take* $n := n + 1$ *and go to Step 2.*

The convergence statement is apparently more complicated than the others, as it solves not only the considered vector optimization problem but also an attached variational inequality.

Theorem 11.15 (cf. [28, Theorem 3.1]) *Let* F *be* C*-convex, positively* C*-lower semicontinuous and* C^**-uniformly semicontinuous and assume that* $F(S) \cap (F(x_1) - C)$ *is* C*-quasicomplete. Assume that* A *is Lipschitz continuous with the Lipschitz constant* $\kappa > 0$, *such that* $\mathcal{WE}(VPG) \cap VI(S, A) \neq \emptyset$ *and that there is a bounded sequence* $(\eta_n)_n \subseteq (0, +\infty)$ *such that* $\sum_{n \geq 1} \eta_n^2 < +\infty$ *and* $\|\theta_n\| \leq \eta_n \|x_n - y_n\|$, *where the sequences* $(x_n)_n$ *and* $(y_n)_n$ *are generated by Algorithm 15. When there are some* $\epsilon, \delta \in (0, 1)$ *such that* $(\beta_n)_n \subseteq (\epsilon, \delta)$, $c \in [0, 1)$ *such that* $(\eta_n)_n \subseteq [0, c)$

and $a, b \in (0, 1/\kappa)$ *such that* $(\lambda_n)_n \subseteq [a, b]$, *then* $(x_n)_n$ *converges weakly towards a weakly efficient solution to* (VPG) *that also lies in* $VI(S, A)$.

Remark 11.23 In [28] one can also find a modification of Algorithm 15 where the usual distance is replaced by a Bregman type one in the spirit of Algorithm 14 as well as an inexact version of Algorithm 15 called relative hybrid approximate proximal algorithm, whose weak convergences towards weakly efficient solutions to (VPG) (that also lie in $VI(S, A)$) are obtained in [28, Theorem 4.3 & Theorem 5.1], respectively, under some additional hypotheses to the ones from Theorem 11.15. Like above (see Remark 11.22), the hypothesis of C^*-asymptotically uniform semicontinuity of F seems no longer necessary for these statements.

Another modification of Algorithm 15 is available in [38], where it is shown to deliver efficient solutions to (VPG). In this case it is not necessary to have $\operatorname{int} C \neq \emptyset$.

Algorithm 16 *Choose the starting point* $x_1 \in \operatorname{dom} F \cap S$ *and the sequences* $(\alpha_n)_n \subseteq (0, \alpha)$, *where* $\alpha > 0$, $(\lambda_n)_n \subseteq (0, 1)$, $(\beta_n)_n \subseteq [0, 1]$, $(\eta_n)_n \subseteq [0, 1]$, $(\theta_n)_n \subseteq X$ *and* $(e_n)_n \subseteq \operatorname{int} C$ *such that* $\|e_n\| = 1$ *for all* $n \geq 1$. *Assume moreover that there exists some* $\delta \in (0, 1]$ *such that the set* K_δ *is nonempty. Consider the following iterative steps*

1 *let* $n = 1$;
2 *if* $x_n \in \mathcal{E}(VPG)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 *take* $y_n = \operatorname{Pr}_S(x_n - \lambda_n A x_n)$;
4 *take* $z_n = \eta_n x_n + (1 - \eta_n) \operatorname{Pr}_S(x_n - \lambda_n A y_n)$;
5 *find* $w_n \in \mathcal{PE}_\delta \left\{ F(x) + \frac{\alpha_n}{2}\|x - z_n - \theta_n\|^2 e_n : x \in \Omega_n \right\}$;
6 *take* $x_{n+1} = \beta_n x_n + (1 - \beta_n) w_n$;
7 *take* $n := n + 1$ *and go to Step 2.*

For the convergence one basically needs the hypotheses of Theorem 11.15.

Theorem 11.16 (cf. [38, Theorem 3.1]) *Let* F *be* C*-convex, positively* C*-lower semicontinuous and* C^**-uniformly semicontinuous and assume that* $F(S) \cap (F(x_1) - C)$ *is* C*-quasicomplete. Assume that* A *is Lipschitz continuous with the Lipschitz constant* $\kappa > 0$ *and that there is a bounded sequence* $(\eta_n)_n \subseteq (0, +\infty)$ *such that* $\sum_{n \geq 1} \eta_n^2 < +\infty$ *and* $\|\theta_n\| \leq \eta_n \|x_n - y_n\|$, *where the sequences* $(x_n)_n$ *and* $(y_n)_n$ *are generated by Algorithm 15. When there are some* $\epsilon, \delta \in (0, 1)$ *such that* $(\beta_n)_n \subseteq (\epsilon, \delta)$, $c \in [0, 1)$ *such that* $(\eta_n)_n \subseteq [0, c)$ *and* $a, b \in (0, 1/\kappa)$ *such that* $(\lambda_n)_n \subseteq [a, b]$, *then* $(x_n)_n$ *converges weakly towards a weakly efficient solution to* (VPG) *that, when* $x_n \notin \mathcal{E}(VPG)$ *for all* $n \geq 1$, *also lies in* $VI(S, A)$.

Remark 11.24 In [38] one can also find a modification of Algorithm 16 where the usual distance is replaced by a Bregman type one in the spirit of Algorithm 14 as well as an inexact version of Algorithm 16 called relative hybrid approximate proximal algorithm, whose weak convergences towards efficient solutions to (VPG) (that also lie in $VI(S, A)$ provided that $x_n \notin \mathcal{E}(VPG)$ for all $n \geq 1$) are obtained in [38, Theorem 4.1 & Theorem 5.1], respectively, under some additional hypotheses

to the ones from Theorem 11.16. Like above (see Remark 11.22), the hypothesis of C^*-asymptotically uniform semicontinuity of F seems no longer necessary for these statements.

Further we also present an inertial proximal point algorithms with memory effects for solving vector optimization problems. The inertial proximal point methods with memory effects, first proposed by Alvarez and Attouch (cf. [2, 3]), were inspired from heavy ball with friction dynamical systems and have as a characteristic feature the fact that an iteration variable depends on the previous two elements of the same sequence, not only on its predecessor as it is usually the case for many algorithmic approaches. This modification accelerates the original proximal point method and makes it more robust. We propose an inertial version of Algorithm 1 that is slightly more general than the special case of Algorithm 19 (introduced later in Section 11.5) that can be employed for solving (VP). For completeness sake, since it cannot be found in the published literature, the proof of the convergence of this algorithm towards a weakly efficient solution to (VP) is provided in an Appendix at the end of the paper.

Algorithm 17 *Choose the starting points* $x_0, x_1 \in \operatorname{dom} F$ *and the sequences* $(\lambda_n)_n \subseteq (0, +\infty)$, $(\alpha_n)_n \subseteq [\alpha, +\infty)$, *where* $\alpha > 0$, $(\beta_n)_n \subseteq [0, \beta)$, *where* $0 < \beta < 1/3$, *and* $(e_n)_n \subseteq \operatorname{int} C$ *such that* $(\alpha_n)_n$ *is bounded,* $(\beta_n)_n$ *is nondecreasing and* $\|e_n\| = 1$ *for all* $n \geq 1$. *Consider the following iterative steps*

1 *let* $n = 1$;
2 *if* $x_n \in \mathcal{WE}(VP)$, *then* $x_{n+p} = x_n$ *for all* $p \geq 1$;
3 *find* $x_{n+1} \in \mathcal{WE}\left\{\lambda_n F(x) + \frac{\alpha_n}{2}\|x - x_n - \beta_n(x_n - x_{n-1})\|^2 e_n : x \in \Omega_n\right\}$;
4 *take* $n := n + 1$ *and go to Step 2.*

Remark 11.25 When $\beta_n = 0$ and $\lambda_n = 1$ for all $n \geq 1$, Algorithm 17 collapses to Algorithm 1. On the other hand, when $Y = \mathbb{R}$ and $C = \mathbb{R}_+$, Algorithm 17 becomes the inertial proximal point method for scalar optimization problems that can be derived from the algorithm for finding zeros of maximally monotone operators proposed in [3].

Theorem 11.17 *Let* F *be* C*-convex and positively* C*-lower semicontinuous and* $F(X) \cap (F(x_1) - C)$ *be* C*-complete. Then any sequence* $(x_n)_n$ *generated by Algorithm 17 converges weakly towards a weakly efficient solution to* (VP).

Remark 11.26 The conclusion of Theorem 11.17 remains valid when the sequence $(x_n)_n$ generated by Algorithm 17 fulfills the condition $\sum_{n=1}^{+\infty} \beta_n \|x_n - x_{n-1}\|^2 < +\infty$, in which case $(\beta_n)_n$ needs not necessarily be nondecreasing and one can take $\beta \in [0, 1)$. However, this dynamic condition might be more difficult to verify since it involves the generated sequence $(x_n)_n$, while the static hypotheses considered above can simply be imposed while defining the parameters β and $(\beta_n)_n$, respectively. Different to the inertial proximal methods proposed in the literature for solving scalar optimization problems or monotone inclusions (see, for instance, [3]), in Theorem 11.17 it is not necessary to assume the existence of a weakly efficient

solution to (VP) in order to prove the convergence of Algorithm 17. Analogous to Remark 11.8, in this case the role of such a hypothesis in showing the convergence of the method has been fully covered by the assumed C-completeness hypothesis.

Remark 11.27 Motivated by Remark 11.6, one can consider also in Algorithm 17 a stopping rule that is easier to check than the original one. It can be shown in a similar manner to [24, Proposition 3.2] that if three consecutive iterations of the sequence $(x_n)_n$ generated by Algorithm 17 coincide, they represent a weakly efficient solution to (VP).

One can provide an inexact version of Algorithm 17 inspired by Algorithm 2 as follows.

Algorithm 18 *Choose the starting points $x_0, x_1 \in \operatorname{dom} F$, the sequences $(\lambda_n)_n \subseteq (0, +\infty)$, $(\alpha_n)_n \subseteq [\alpha, +\infty)$, where $\alpha > 0$, $(\beta_n)_n \subseteq [0, \beta)$, where $0 < \beta < 1/4$, $(z_n^*)_n \subseteq C^* \setminus \{0\}$, and $(e_n)_n \subseteq \operatorname{int} C$ such that $(\alpha_n)_n$ is bounded, $(\beta_n)_n$ is nondecreasing, $\|z_n^*\| = 1$ and $\|e_n\| = 1$ for all $n \geq 1$, as well as the constant $\sigma \in [0, 1 - 4\beta)$. Consider the following iterative steps*

1 let $n = 1$;
2 if $x_n \in \mathcal{WE}(VP)$, then $x_{n+p} = x_n$ for all $p \geq 1$;
3 find $x_{n+1} \in \operatorname{dom} F$ such that

$$0 \in \partial_{\varepsilon_n}(\langle z_n^*, F(\cdot)\rangle + \delta_{\Omega_n})(x_{n+1}) + \alpha_n \langle z_n^*, e_n\rangle (x_{n+1} - x_n - \beta_n(x_n - x_{n-1}))$$

for some $0 \leq \varepsilon_n \leq \sigma \frac{\alpha_n}{2} \langle z_n^*, e_n\rangle \|x_{n+1} - x_n - \beta_n(x_n - x_{n-1})\|^2$;
4 take $n := n + 1$ and go to Step 2.

Remark 11.28 When $\beta_n = 0$ and $\lambda_n = 1$ for all $n \geq 1$ and $\beta \downarrow 0$, Algorithm 18 collapses to Algorithm 2.

The convergence of Algorithm 18 towards a weakly efficient solution to (VP) can be guaranteed under the same hypotheses as the one of its exact version Algorithm 17, the proof relying on the ones of Theorem 11.17 and [24, Theorem 5.1].

Theorem 11.18 *Let F be C-convex and positively C-lower semicontinuous and $F(X) \cap (F(x_1) - C)$ be C-complete. Then any sequence $(x_n)_n$ generated by Algorithm 17 converges weakly towards a weakly efficient solution to (VP).*

Remark 11.29 Another possible way to provide an inexact version of Algorithm 17 may be investigated by making use of the approximative inertial type proximal scheme proposed in [2, Section 3.2].

Remark 11.30 Another inertial type proximal point method was proposed in [27] for determining ideally efficient solutions to (VP) in case $Y = \mathbb{R}^m$. We opted not to present it here because of the limited significance of the ideally efficient solutions to (VP) and also due to the way it is constructed that required introducing maximally monotone operators. Note however that this algorithm is accompanied by applications to convex feasibility problems and to the problem of common fixed

points for nonexpansive potential mappings and also that some convergence rates are derived for the method.

Remark 11.31 In the literature there are also some proximal point type algorithms for solving vector optimization problems whose objective functions map from Hadamard manifolds to Euclidean spaces. The algorithmic schemes are not much different from the ones presented above, however one needs to define the whole Hadamard manifolds setting in order to give the corresponding convergence statements. In [13] a proximal point type scheme is proposed for determining weakly efficient solutions to the Hadamard version of (VPG), where F is taken as mentioned above. Weakly efficient elements to the Hadamard version of (VP) are obtained in [16] by means of a proximal point type algorithm where the intermediate optimization problems are scalarized versions of the ones in [13] by means of a special case of the scalarization function introduced by Tammer (then Gerstewitz) in [43]. An inexact version of the method that converges towards weakly efficient elements to the considered vector optimization problem is proposed as well. Another special case of Tammer's scalarization function, the so-called maximum scalarization, is employed in [74] for proposing an inexact proximal point method for determining efficient solutions to the Hadamard version of (VP). Because of the special structure of its objective function, the scalarized optimization problem should apparently be not so difficult to solve and a rate of convergence of the proposed method is provided, too.

Remark 11.32 One could include in this section also the algorithms in [41, 51, 71, 73], as they are proximal point type methods for solving vector optimization problems as well. However, in all of them the new iteration x_{n+1} is calculated as an optimal solution to a scalar optimization problem that is a scalarization of a vector one, not as some sort of an efficient solution to some vector optimization problem, failing thus to satisfy the criteria stated in [24]. The algorithm proposed in [41] employs the Tammer scalarization function (see also Remark 11.31). This function leads to scalar optimization problems that are bounded from below by 0, excluding thus the possibility to have to deal with unbounded scalar optimization problems that may occur when working with the linear scalarization. However, the objective functions of the scalar optimization problems derived by means of this scalarization contain compositions of functions that are unfortunately still unsuitable for the existing proximal point type algorithms and, moreover, this scalarization does not guarantee a descent property for the values of the objective function either. On the other hand, worth noticing is that the intermediate optimization problems in the algorithm designed in [41] for delivering weakly efficient solutions to (VP) are unconstrained, i.e. they do not require defining the sets Ω_n, $n \geq 1$. An inexact method that converges towards weakly efficient elements to the considered vector optimization problem is proposed in the Hadamard manifolds framework discussed in Remark 11.31 in [73] by means of a generalized proximal distance, where the original problem is scalarized by a special case of the Tammer scalarization

function. In [51, 71] there are proposed some logarithmic proximal point type algorithms for solving (VP) in Euclidean spaces. However, these are essentially methods for minimizing some scalar functions called strict scalar representations of the objective function of the original vector optimization problem. Note that the method proposed in [51] can be modelled, to some extent, to deal also with vector optimization problems with convex inequality constraints. On the other hand, the one from [71] employs a quasi-distance instead of the classical distance induced by a norm in the proximal step and is employed for numerically solving some test problems.

Remark 11.33 There are some papers where proximal point type algorithms for solving vector optimization problems consisting in vector-minimizing a vector function subject to both geometric and linear equality constraints are proposed, in the finitely dimensional framework where $X = \mathbb{R}^n$, $S \subseteq X$ is convex and compact, $Y = \mathbb{R}^m$, $C \subseteq \mathbb{R}^m$ is a convex cone, $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$ and $F : \mathbb{R}^n \to \mathbb{R}^m$. The constrained vector optimization problem considered in [67, 68] is

$$(VPC) \qquad \underset{\substack{Ax=b,\\ x\in S}}{\mathrm{WMin}}\, F(x).$$

However, the methods proposed in the mentioned papers for determining weakly efficient and efficient solutions to (VPC) fail to satisfy the criteria stated in [24] (see also Remark 11.32), as they consist of solving some minmax scalar optimization problems that contain scalarizations of the objective function of (VPC) and some other terms based on the equality constraint. The algorithms are shown to converge towards weakly efficient solutions to (VPC) and, under additional assumptions, towards efficient ones (called strongly efficient in [68]) to it. Applications to supply chain network risk management and computational results obtained in MATLAB are provided, too in both these works, despite the fact that usually minmax optimization problems are not easy to solve numerically. Note also that in [72, Remark 3.3] a vector optimization problem with a linear objective function and both geometric and linear inequality constraints is mentioned and a numerical scheme for solving it is sketched, however without actually exploiting the structure of the constraint set.

11.5 Proximal Point Type Algorithms for Other Vector Optimization Problems

In this section we present proximal point type algorithms for solving vector optimization problems with more complicated structure than (VP) (or its constrained counterparts (VPG) and (VPC)), namely ones with a sum or difference of vector functions as an objective function. Many of the remarks we gave in Section 11.3 remain valid for some of these classes of vector optimization problems as well and we will not mention them again here.

11.5.1 Vector-Minimization of Sums of Vector Functions

Consider next the so-called composite vector optimization problem consisting in vector-minimizing the sum of two vector functions, whose weakly efficient solutions were obtained in [19] by means of a forward-backward proximal point method with inertial/memory effects

$$(VPS) \qquad \operatorname*{WMin}_{x\in X}\big[F(x)+G(x)\big],$$

where $F : X \to Y^\bullet$ is a proper vector function and $G : X \to Y$ is a Fréchet differentiable vector function with an L-Lipschitz continuous gradient ∇G.

The exact proximal inertial forward-backward iterative method proposed in [19] for determining the weakly efficient solutions to (VPS) is the following.

Algorithm 19 *Choose the starting points $x_0, x_1 \in X$ and the sequences $(\beta_n)_n \subseteq [0,\beta)$, $(z_n^*)_n \subseteq C^* \setminus \{0\}$ and $(e_n)_n \subseteq \operatorname{int} C$ such that $(\beta_n)_n$ is nondecreasing, $\beta < 1/9$, $\|z_n^*\| = 1$ and $\langle z_n^*, e_n\rangle = 1$ for all $n \geq 1$. Consider the following iterative steps*

1. *let $n = 1$;*
2. *if $x_n \in \mathcal{WE}(VPS)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3. *find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{L}{2}\left\|x - \left(x_n + \beta_n(x_n - x_{n-1}) - \frac{1}{L}\nabla(z_n^* G)(x_n)\right)\right\|^2 e_n : x \in \Omega_n^S\right\}$, where $\Omega_n^S = \{x \in X : (F+G)(x) \leqq_C (F+G)(x_n)\}$;*
4. *take $n := n+1$ and go to Step 2.*

Remark 11.34 When $G \equiv 0$, Algorithm 19 becomes an inertial proximal point method for solving vector optimization problems that is a special case of Algorithm 17, and by additionally taking $\beta_n = 0$ for all $n \geq 1$ it collapses into the proximal point method for vector-minimizing a nonsmooth vector function introduced in [24] and presented above as Algorithm 1. On the other hand, when $Y = \mathbb{R}$ and $C = \mathbb{R}_+$ (i.e. in the scalar case), Algorithm 19 becomes the inertial proximal-gradient method for scalar optimization problems, that can be derived from the algorithm for finding zeros of maximally monotone operators proposed in [63]. When, furthermore, $G \equiv 0$, it collapses into the one from [3], while when $\beta_n = 0$ for all $n \geq 1$ it becomes the celebrated ISTA method, however in a more general framework.

The convergence of Algorithm 19 is achieved in a similar setting to the one of Algorithm 1.

Theorem 11.19 (cf. [19, Theorem 2.1]) *Let F be C-convex and positively C-lower semicontinuous, G be C-convex and assume that $(F+G)(X) \cap (F(x_1) + G(x_1) - C)$ is C-complete. Then any sequence $(x_n)_n$ generated by Algorithm 19 converges weakly towards a weakly efficient solution to (VPS).*

Remark 11.35 The conclusion of Theorem 11.19 remains valid when one takes only $F+G$ to be C-convex instead of both F and G and Remark 11.26 applies here as well. The additional hypotheses of Theorem 11.2 guarantee the weak convergence of any sequence $(x_n)_n$ generated by Algorithm 19 towards an efficient solution

to (VPS), too. Note also that the intermediate vector optimization problems can be solved as such or can be scalarized for this, in which case an obvious choice for the scalarizing functional are the corresponding z_n^*'s. Of course, the sequence $(z_n^*)_n$ can be taken constant, situation in which the intermediate vector optimization problems differ despite having the same objective vector function because their feasible sets become smaller at each iteration. This does not mean that the vector optimization problem (VPS) is a priori scalarized by means of a linear continuous functional, because this scalarization is applied to the intermediate vector optimization problems not to (VPS).

Remark 11.36 For implementation purposes one can provide an inexact version of Algorithm 19 as well, where Step 3 is replaced by

3' find $x_{n+1} \in X$ such that

$$0 \in \partial_{\varepsilon_n}(\langle z_n^*, F(\cdot) + \frac{L}{2}\|\cdot - x_n - \beta_n(x_n - x_{n-1}) + \frac{1}{L}\nabla(z_n^* G)(x_n)\|^2 e_n\rangle + \delta_{\Omega_n^S}(\cdot))(x_{n+1}),$$

where the additional sequence of tolerable nonnegative errors $(\varepsilon_n)_n$ fulfills some hypotheses, such as the ones considered in [24] or those from [2, 63]. Employing the later, i.e. $\sum_{n\geq 1}\varepsilon_n < +\infty$, the convergence statement obtained by correspondingly modifying Theorem 11.19 remains valid. Moreover, as an alternative stopping rule that is easier to check than Step 2 of Algorithm 19 one can verify whether three consecutive iterations of the sequence $(x_n)_n$ generated by the method coincide, in which case they represent a weakly efficient solution to (VPS). Note also that $x_{n-1} = x_n$ does not necessarily imply that x_{n+1} coincides with them, too. This can prove to be useful when starting the algorithm because one can begin with $x_0 = x_1$ without affecting the convergence of the method.

In [19] also the following (Nesterov type) modification of Algorithm 19 was proposed, the difference between them residing in the point where the value of $\nabla(z_n^* G)$ is calculated. The above remarks remain basically valid for it as well.

Algorithm 20 *Choose the starting points $x_0, x_1 \in X$ and the sequences $(\beta_n)_n \subseteq [0, \beta)$, $(z_n^*)_n \subseteq C^* \setminus \{0\}$ and $(e_n)_n \subseteq \operatorname{int} C$ such that $(\beta_n)_n$ is nondecreasing, $\beta < 1/9$, $\|z_n^*\| = 1$ and $\langle z_n^*, e_n\rangle = 1$ for all $n \geq 1$. Consider the following iterative steps*

1 *let $n = 1$;*
2 *if $x_n \in \mathcal{WE}(VPS)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3 *find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{L}{2}\left\|x - \left(x_n + \beta_n(x_n - x_{n-1}) - \frac{1}{L}\nabla(z_n^* G)(x_n + \beta_n(x_n - x_{n-1}))\right)\right\|^2 e_n : x \in \Omega_n^S\right\}$;*
4 *take $n := n + 1$ and go to Step 2.*

Remark 11.37 In the scalar case, when $Y = \mathbb{R}$ and $C = \mathbb{R}_+$, Algorithm 20 becomes a more general version of the celebrated FISTA method from [10], that can be

recovered by taking $\beta_n = (t_n - 1)/t_{n+1}$, where $t_{n+1} = (1 + \sqrt{1 + 4t_n^2})/2$, $n \geq 1$, with $t_1 = 1$, and restricting the framework to finitely dimensional spaces.

The convergence statement concerning Algorithm 20 is similar to Theorem 11.19.

Theorem 11.20 (cf. [19, Theorem 3.1]) *Let F be C-convex and positively C-lower semicontinuous, G be C-convex and assume that $(F + G)(X) \cap (F(x_1) + G(x_1) - C)$ is C-complete. Then any sequence $(x_n)_n$ generated by Algorithm 20 converges weakly towards a weakly efficient solution to (VPS).*

Remark 11.38 One can additionally provide, following [10, Theorem 4.4], when the sequence $(z_n^*)_n$ is constant and for the choice of the parameters β_n, $n \geq 1$, mentioned in Remark 11.37, a convergence rate statement for the values of the objective functions of the scalarized intermediate problems in Algorithm 20. Moreover, when taking the sequence $(z_n^*)_n$ constant it is no longer necessary to take $\|z_n^*\| = 1$ for all $n \geq 1$.

Stripping any of Algorithm 19 or Algorithm 20 of its inertial terms, it collapses into a forward-backward method, whose convergence is derivable from Theorem 11.19 or Theorem 11.20, as follows.

Algorithm 21 *Choose the starting point $x_1 \in X$ and the sequences $(z_n^*)_n \subseteq C^* \backslash \{0\}$ and $(e_n)_n \subseteq \operatorname{int} C$ such that $\|z_n^*\| = 1$ and $\langle z_n^*, e_n \rangle = 1$ for all $n \geq 1$. Consider the following iterative steps*

1. *let $n = 1$;*
2. *if $x_n \in \mathcal{WE}(VPS)$, then $x_{n+p} = x_n$ for all $p \geq 1$;*
3. *find $x_{n+1} \in \mathcal{WE}\left\{F(x) + \frac{L}{2}\left\|x - \left(x_n - \frac{1}{L}\nabla(z_n^* G)(x_n)\right)\right\|^2 e_n : x \in \Omega_n^S\right\}$;*
4. *take $n := n + 1$ and go to Step 2.*

A convergence rate statement for the values of the objective functions of the scalarized intermediate problems in Algorithm 21 can be deduced analogously to [10, Theorem 3.1].

Theorem 11.21 (cf. [19, Theorem 2.2]) *Let F be C-convex and positively C-lower semicontinuous, G be C-convex and assume that $(F+G)(X) \cap (F(x_1)+G(x_1)-C)$ is C-complete. Consider the sequence $(x_n)_n$ generated by Algorithm 21, where one takes $z_n^* = z^* \in C^* \setminus \{0\}$, $n \geq 1$. Then for any $n \geq 1$ and $\tilde{x} \in \cap_{n \geq 1} \Omega_n^S$ one has*

$$\langle z^*, F(x_n) + G(x_n) - F(\tilde{x}) - G(\tilde{x}) \rangle \leq \frac{L\|\tilde{x} - x_1\|^2}{2n}.$$

Remark 11.39 Unlike most of the mentioned papers where iterative methods for solving vector optimization problems were proposed, but their implementation was left for later due to the difficulty of solving the employed subproblems, in [19, Section 4] a concrete application in finance mathematics was solved in MATLAB via the inexact versions of both Algorithm 19 and Algorithm 21, whose performances

are then compared, showing that a good choice of the inertial parameters can considerably reduce the resources needed for identifying a weakly efficient solution to (VPS).

11.5.2 Vector-Minimization of Differences of Cone-Convex Vector Functions

The last class of vector optimization problems considered in this survey consists in vector-minimizing the difference of two cone-convex (DC) vector functions subject to a geometric constraint in finitely dimensional spaces. One can find in [52, 53] proximal point type algorithms for solving such problems, however, due to their structure, only critically efficient solutions are determined. In the convex case the critically efficient solutions turn out to be weakly efficient, however this cannot happen in this setting. Consider the DC vector optimization problem

$$(VPD) \qquad \underset{x\in S}{\mathrm{WMin}} \left[F(x) - G(x)\right],$$

where $F, G : \mathbb{R}^n \to \mathbb{R}^m$ are proper C-convex vector functions, $C \subseteq \mathbb{R}^m$ being a convex cone, and $S \subseteq \mathbb{R}^n$ is a closed convex set.

The proximal point type algorithm for solving (VPD) proposed in [53] is the following.

Algorithm 22 *Choose the starting point $x_1 \in \mathbb{R}^n$, the exogenous sequence $(\alpha_n)_n \subseteq (0, +\infty)$ and $\varepsilon > 0$. Consider the following iterative steps*

1 let $n = 1$;
2 find $V_n \in \partial G(x_n)$;
3 find $x_{n+1} = \arg\min_{x\in S} \max_{u\in C^*\cap \mathcal{S}_{\mathbb{R}^m}} \left\{u^\top F(x) - u^\top V_n(x - x_n) + \frac{\alpha_n}{2}\|x - x_n\|^2\right\}$;
4 if $\|x_{n+1} - x_n\| \leq \varepsilon$: STOP;
5 take $n := n + 1$ and go to Step 2.

The convergence statement regarding Algorithm 22 follows. Note however that only the fact that every cluster point of $(x_n)_n$ is a critically efficient solution to (VPD) is guaranteed and that even for this a condition involving the structure of the sequence $(x_n)_n$ is imposed.

Theorem 11.22 (cf. [53, Theorem 7]) *Let S be bounded and assume that for $n \geq 1$ large enough there is some $u \in C^* \cap \mathcal{S}_{\mathbb{R}^m}$ such that $x_{n+1} \in S \cap \left(\partial G^\top u + \alpha_n \mathrm{id}_{\mathbb{R}^n}\right)^{-1}\left(V_n^\top u + \alpha_n x_n\right)$, where the sequence $(x_n)_n$ is generated by Algorithm 22 and it has infinitely many iterations. Then any cluster point of $(x_n)_n$ is a critically efficient solution to (VPD).*

Remark 11.40 In [53] one can find also a second proximal point type algorithm for solving (VPD) where the role of α_n is taken by $r_n^\top u$, where $(r_n)_n \subseteq \mathbb{R}^m$ is some iterative sequence that satisfies $r_n^\top u > 0$. Moreover, an inexact version of Algo-

rithm 22 called ε-proximal algorithm is proposed, where the vector subdifferential of G is replaced (also in the hypotheses of the convergence statement) by its vector ε-subdifferential.

Remark 11.41 A special case of Algorithm 22 obtained when $C = \mathbb{R}^m_+$ can be found in the earlier paper [52] together with an inexact version and an application to portfolio optimization that is also numerically solved in MATLAB, some computational results being provided.

Remark 11.42 Note that in Algorithm 22 the sets Ω_n play no role and the intermediate problems are only geometrically constrained. On the other hand, at the first look they do not satisfy the criterion mentioned in [24] (see Remark 11.32), as they are scalar minmax optimization problems. However we opted to include Algorithm 22 here and not only to mention it in a remark because it employs the vector subdifferential of G and not one of some scalarization of it and, on the other hand, since the method can be seen as a splitting type one where the involved functions are processed separately, as one determines an element of the vector subdifferential of G and then uses it in a sort of a backward step. Another argument for the inclusion of this algorithm in this work is the fact that in [53] it is applied for numerically solving in MATLAB a problem of probabilistic lot sizing with service levels and computational results are provided, too. Note also that the hypotheses of the convergence statement do not include the usual C-completeness assumption considered in most of its counterparts gathered in this survey.

11.6 Conclusions and Further Research Directions

At the moment one can find more than thirty papers whose authors claim to introduce new proximal point type algorithms for solving vector optimization problems or to refine some existing ones. In most of these the method proposed in the seminal contribution due to Bonel, Iusem and Svaiter is extended in some way, usually by replacing the norm distance in the iterative step by some other (quasi-)distance function or by relaxing the hypotheses that are necessary for ensuring the convergence of the algorithm. Moreover, there are some contributions where one also has constraints or the objective function has a more complicated structure, being a sum or difference of vector functions or a composition of such a function with a linear operator.

The first conclusion one can draw is that in the decade and a half since the mentioned paper was published the interest around iteratively solving vector (and multiobjective) optimization problems by means of proximal point methods has steadily increased. Various techniques were extended from scalar to vector optimization and it seems that others would follow soon. There are proximal point type algorithms for approaching nonconvex vector optimization problems and some that deliver efficient (and not only weakly efficient) solutions to the considered vector optimization problem. The authors of the original paper have

actually admitted that their algorithm was merely a theoretical scheme and many of the ones who followed it contain no numerical results either, however in some recent contributions one can find concrete applications with computational results.

On the other hand, one can notice that the algorithmic scheme of Bonel, Iusem and Svaiter is still the standard in this research area, as many of the subsequent contributions are mostly theoretical variations or improvements. Despite the progress in relaxing the hypotheses of the convergence statement, there are still no real alternatives to the cone-completeness hypothesis, as the compactness of the (weak) efficiency set cannot usually be a priori verified. However, the most important issue relies in the construction of the intermediate vector optimization problems that have to be solved in the iterative steps. The fact that these are constrained makes the existing splitting proximal point methods not really useful in approaching them and they have to be solved via other algorithms or solvers. The role of these constraints is to ensure that the values of the objective function decrease with respect to the ordering cone, thus a possible idea could be here to find another way to guarantee this descending property of the generated sequence without making the intermediate problems constrained. Moreover, there are almost no results on convergence rates for such algorithms and the existing ones require quite restrictive hypotheses. On the other hand, with or without constraints, there is still room for improvements with respect to the implementations of this class of algorithms, as only a few contributions contain actual computational results.

Of course, the difficulties encountered while trying to adapt techniques from scalar optimization into the vector optimization framework are far from being trivial. For instance, at each iteration one has to deal with a different intermediate vector optimization problem, while in the scalar case the objective function of the considered problem is usually not modified as the algorithm advances. Moreover, at the moment there is still no characterization via a monotone inclusion of the efficient solutions of a vector optimization problem, so directly adapting a method from that area without going through the scalar case is still out of the question. And, as mentioned above, the constraints of the intermediate vector optimization problems are not making the life easier.

Besides these, there are many other challenges regarding proximal point methods for solving vector optimization problems that are more or less solvable. We list in the following some of them. A first one would be to identify weaker hypotheses or the necessary modifications of the existing algorithms in order to guarantee the identification of (properly) efficient solutions to the considered vector optimization problems instead of weakly efficient or even critically efficient ones. Some the methods known at the moment to function only in Euclidean spaces are expected to work, under additional assumptions, in infinitely dimensional settings such as Hilbert spaces, too. As suggested in a paper by Bento, da Cruz Neto and Soubeyran, and, on the other hand, in one due to Rocha, Oliveira, Gregório and Souza, in the nonconvex case one can try to employ functions having the Kurdyka-Łojasiewicz property, too. Things are at the moment only at the beginning with respect to vector optimization problems with objective functions consisting of sums or differences of vector functions and/or compositions with linear continuous mappings. Other

splitting schemes besides the forward-backward one could be applied. Speaking of the later, as mentioned in the cited paper of Boţ and the author, it would be interesting to identify a way to modify the proposed forward-backward algorithms in order to encompass as a special case also the projected gradient method proposed by Graña Drummond and Iusem for vector-minimizing a smooth cone-convex vector function. Adding constraints to the considered vector optimization problems obviously complicates things, however since the intermediate problems are already constrained it would be interesting to find some ways to approach these both, too, maybe by means of duality. Last but not least, as the ordering cones that occur in vector optimization often have empty interiors, modifications of the existing algorithms in order to maintain their convergence towards weakly efficient solutions defined by means of generalized interiors should be taken into consideration as well.

Appendix: Proof of Theorem 11.17

In the following we provide an example of a convergence proof for a proximal point algorithm for determininig weakly efficient solutions to a vector optimization problem. It originates from an earlier version of [19] and incorporates some ideas from the proofs of [24, Theorem 3.1] and [3, Theorem 2.1 and Proposition 2.1]. Before formulating it, we recall the celebrated Opial's Lemma (cf. [65]).

Lemma 11.2 *Let $(x_n)_n \subseteq X$ a sequence such that there exists a nonempty set $S \subseteq X$ such that*

(a) $\lim_{n\to+\infty} \|x_n - x\|$ *exists for every $x \in S$;*
(b) if $x_{n_j} \rightharpoonup \hat{x}$ for a subsequence $n_j \to +\infty$, then $\hat{x} \in S$.

Then, there exists an $\bar{x} \in S$ such that $x_k \rightharpoonup \bar{x}$ when $k \to +\infty$.

Theorem 11.17 *Let F be C-convex and positively C-lower semicontinuous and $F(X) \cap (F(x_1) - C)$ be C-complete. Then any sequence $(x_n)_n$ generated by Algorithm 17 converges weakly towards a weakly efficient solution to (VP).*

Proof We show first that the algorithm is well-defined. Assuming that we have obtained an x_n, where $n \geq 1$, we have to secure the existence of x_{n+1}. Take a $z_n^* \in C^* \setminus \{0\}$ and without loss of generality assume that $\|z_n^*\| = 1$ for all $n \geq 1$. Then $\langle z_n^*, e_n\rangle > 0$ and the function

$$x \mapsto \langle z_n^*, \lambda_n F(x) + \frac{\alpha_n}{2}\|x - x_n - \beta_n(x_n - x_{n-1})\|^2 e_n\rangle + \delta_{\Omega_n}(x)$$

is lower semicontinuous, being a sum of continuous and lower semicontinuous functions, respectively, and strongly convex, as the sum of some convex functions and a squared norm, having thus exactly one minimum. By Lemma 11.1 this minimum is a weakly efficient solution to the vector optimization problem in Step 3 of Algorithm 17 and we denote it by x_{n+1}.

The next step is to show the Fejér monotonicity of the sequence $(x_n)_n$ with respect to the set $\Omega = \{x \in X : F(x) \leqq_C F(x_k)\ \forall k \geq 0\}$, that is nonempty because of the C-completeness hypothesis. Let $n \geq 1$. The function $x \mapsto \langle z_n^*, \lambda_n F(x) + (\alpha_n/2)\|x - x_n - \beta_n(x_n - x_{n-1})\|^2 e_n\rangle + \delta_{\Omega_n}(x)$ attains its only minimum at x_{n+1}, and this fact can be equivalently written as

$$0 \in \partial\big(\langle z_n^*, \lambda_n F(\cdot) + \frac{\alpha_n}{2}\|\cdot - x_n - \beta_n(x_n - x_{n-1})\|^2 e_n\rangle + \delta_{\Omega_n}(\cdot)\big)(x_{n+1}).$$

Using the continuity of the norm, this yields (e.g. via [21, Theorem 3.5.6]) $0 \in \partial\big(\langle z_n^*, \lambda_n F(\cdot)\rangle + \delta_{\Omega_n}(\cdot)\big)(x_{n+1}) + \partial\big((\alpha_n/2)\langle z_n^*, e_n\rangle\|\cdot - x_n - \beta_n(x_n - x_{n-1})\|^2\big)(x_{n+1}) = \partial\big(\langle z_n^*, \lambda_n F(\cdot)\rangle + \delta_{\Omega_n}(\cdot)\big)(x_{n+1}) + \alpha_n\langle z_n^*, e_n\rangle(x_{n+1} - x_n - \beta_n(x_n - x_{n-1}))$. Then, since $x_{n+1} \in \Omega_n$, for any $x \in \Omega_n$ it holds

$$\lambda_n\langle z_n^*, F(x) - F(x_{n+1})\rangle \geq \alpha_n\langle z_n^*, e_n\rangle\langle x_{n+1} - x_n - \beta_n(x_n - x_{n-1}), x_{n+1} - x\rangle. \tag{11.2}$$

Let us take an element $\tilde{x} \in \Omega$. By construction $\tilde{x} \in \Omega_n$, thus (11.2) yields, after taking into consideration that $F(\tilde{x}) \leqq_C F(x_{n+1})$, $\lambda_n > 0$ and $z_n^* \in C^* \setminus \{0\}$, that $\alpha_n\langle z_n^*, e_n\rangle\langle x_{n+1} - x_n - \beta_n(x_n - x_{n-1}), \tilde{x} - x_{n+1}\rangle \geq 0$.

For each $k \geq 0$ denote $\varphi_k = (1/2)\|x_k - \tilde{x}\|^2$. The previous inequality, after dividing with the positive number $\alpha_n\langle z_n^*, e_n\rangle$, can be rewritten as

$$\varphi_{n+1} - \varphi_n + \frac{1}{2}\|x_{n+1} - x_n\|^2 - \beta_n\langle x_n - x_{n-1}, x_{n+1} - \tilde{x}\rangle \leq 0,$$

and, since $\langle x_n - x_{n-1}, x_{n+1} - \tilde{x}\rangle = \varphi_n - \varphi_{n-1} + (1/2)\|x_n - x_{n-1}\|^2 + \langle x_n - x_{n-1}, x_{n+1} - x_n\rangle$, it turns into

$$\varphi_{n+1} - \varphi_n - \beta_n(\varphi_n - \varphi_{n-1}) \leq \frac{\beta_n}{2}\|x_n - x_{n-1}\|^2 +$$

$$\beta_n\langle x_n - x_{n-1}, x_{n+1} - x_n\rangle - \frac{1}{2}\|x_{n+1} - x_n\|^2. \tag{11.3}$$

Since the right-hand side of (11.3) is less than or equal to $((\beta_n - 1)/2)\|x_{n+1} - x_n\|^2 + \beta_n\|x_n - x_{n-1}\|^2$, denoting $\mu_k = \varphi_k - \beta_k\varphi_{k-1} + \beta_k\|x_k - x_{k-1}\|^2$, $k \geq 1$, it follows that

$$\mu_{n+1} - \mu_n \leq \frac{3\beta - 1}{2}\|x_{n+1} - x_n\|^2 \leq 0, \tag{11.4}$$

thus the sequence $(\mu_k)_k$ is nonincreasing, as $n \geq 1$ was arbitrarily chosen. Then $\varphi_n \leq \beta^n\varphi_0 + \mu_1/(1-\beta)$ and one also gets $\|x_{n+1} - x_n\|^2 \leq (2/(1-3\beta))(\mu_n - \mu_{n+1})$. Employing (11.4), one obtains then

$$\sum_{k=1}^{n} \|x_{k+1} - x_k\|^2 \leq \frac{2}{1-3\beta}(\mu_1 - \mu_{n+1}) \leq \frac{2}{1-3\beta}\left(\beta^{n+1}\varphi_0 + \frac{\mu_1}{1-\beta}\right) < +\infty,$$

in particular

$$\sum_{k=1}^{+\infty} \|x_{k+1} - x_k\|^2 \leq \frac{2\mu_1}{(1-\beta)(1-3\beta)} < +\infty. \tag{11.5}$$

The right-hand side of (11.3) can be rewritten as $(1/2)(\beta_n(\beta_n+1)\|x_n - x_{n-1}\|^2 - \|x_{n+1} - x_n - \beta_n(x_n - x_{n-1})\|^2)$. Denoting $\tau_{k+1} = x_{k+1} - x_k - \beta_k(x_k - x_{k-1})$, $\theta_k = \varphi_k - \varphi_{k-1}$ and $\delta_k = \beta_k\|x_k - x_{k-1}\|^2$ for $k \geq 0$ and taking into consideration that $\beta_n \in [0, 1/3)$, (11.3) yields

$$\theta_{n+1} - \beta_n\theta_n \leq \delta_n - \frac{1}{2}\|\tau_{n+1}\|^2. \tag{11.6}$$

Then $[\theta_{n+1}]_+ \leq (1/3)[\theta_n]_+ + \delta_n$, followed by $[\theta_{n+1}]_+ \leq (1/3^n)[\theta_1]_+ + \sum_{k=0}^{n-1} \delta_{n-k}/3^k$. Hence $\sum_{k=0}^{+\infty}[\theta_{k+1}]_+ \leq 3/2([\theta_1]_+ + \sum_{k=0}^{+\infty}\delta_k)$ and, as the right-hand side of this inequality is finite due to (11.5), so is $\sum_{k=1}^{+\infty}[\theta_k]_+$, too. This yields that the sequence $(w_k)_k$ defined as $w_k = \varphi_k - \sum_{j=1}^{k}[\theta_j]_+$, $k \geq 0$, is bounded. Moreover, $w_{k+1} - w_k = \varphi_{k+1} - \varphi_k - [\varphi_{k+1} - \varphi_k]_+ = \varphi_{k+1} - \varphi_k + \min\{0, \varphi_k - \varphi_{k+1}\} \leq 0$ for all $k \geq 1$, thus $(w_k)_k$ is convergent. Consequently, $\lim_{k\to+\infty}\varphi_k = \lim_{k\to+\infty} w_k + \sum_{j=1}^{+\infty}[\theta_{j+1}]_+$, therefore $(\varphi_k)_k$ is convergent. Finally, $(\|x_k - \tilde{x}\|^2)_k$ is convergent, too, i.e. (a) in Lemma 11.2 with $S = \Omega$ is fulfilled.

We show now that $(x_k)_k$ is weakly convergent. The convergence of $(\varphi_k)_k$ implies that $(x_k)_k$ is bounded, so it has weak cluster points. Let $\hat{x} \in X$ be one of them and $(x_{k_j})_j$ the subsequence that converges towards it. Then, as F is positively C-lower semicontinuous and C-convex, it follows that for any $z^* \in C^*$ the function $\langle z^*, F(\cdot)\rangle$ is lower semicontinuous and convex, thus

$$\langle z^*, F(\hat{x})\rangle \leq \lim_{j\to+\infty}\langle z^*, F(x_{k_j})\rangle = \inf_{k\geq 0}\langle z^*, F(x_k)\rangle, \tag{11.7}$$

with the last equality following from the fact that the sequence $(F(x_k))_k$ is by construction nonincreasing. Assuming that there exists a $k \geq 0$ such that $F(\hat{x}) \not\leqq_C F(x_k)$, there exists a $\tilde{z} \in C^*\setminus\{0\}$ such that $\langle\tilde{z}, F(\hat{x}) - F(x_k)\rangle > 0$, which contradicts (11.7), consequently $F(\hat{x}) \leqq_C F(x_k)$ for all $k \geq 0$, i.e. $\hat{x} \in \Omega$, therefore one can employ Lemma 11.2 with $S = \Omega$ since its hypothesis (b) is fulfilled as well. This guarantees then the weak convergence of $(x_k)_k$ to a point $\bar{x} \in \Omega$.

The last step is to show that $\bar{x} \in \mathcal{WE}(VP)$. Assuming that $\bar{x} \notin \mathcal{WE}(VP)$, there exists an $x' \in X$ such that $F(x') <_C F(\bar{x})$. This yields $x' \in \Omega$. As $\|z_k^*\| = 1$ for all $k \geq 0$, the sequence $(z_k^*)_k$ has a weak* cluster point, say $\bar{z}^*$, that is the limit of a subsequence $(z_{k_j}^*)_j$. Because $z_k^* \in C^*$ for all $k \geq 0$ and C^* is weakly* closed, it follows that $\bar{z}^* \in C^*$. Moreover, $\bar{z}^* \neq 0$, since it can be shown via [23, Lemma 2.2]

that $\langle \bar{z}^*, c\rangle > 0$ for any $c \in \operatorname{int} C$. Consequently, $\langle \bar{z}^*, F(x') - F(\bar{x})\rangle < 0$. For any $j \geq 0$ it holds by (11.2)

$$\lambda_{k_j}\langle z^*_{k_j}, F(x') - F(x_{k_j+1})\rangle \geq -\langle \alpha_{k_j}\langle z^*_{k_j}, e_{k_j}\rangle (x_{k_j+1} - x_{k_j} - \beta_{k_j}(x_{k_j} - x_{k_j-1}), x' -$$

$$x_{k_j+1}\rangle \geq -\alpha_{k_j}\langle z^*_{k_j}, e_{k_j}\rangle \|x' - x_{k_j+1}\|\big(\|x_{k_j+1} - x_{k_j}\| + \beta_{k_j}\|x_{k_j} - x_{k_j-1}\|\big). \quad (11.8)$$

Because of (11.5), $(\|x_k - x_{k-1}\|)_k$ converges towards 0 for $k \to +\infty$, therefore so does the last expression in the inequality chain (11.8) when $j \to +\infty$ as well. Letting j converge towards $+\infty$, (11.8) yields $\langle \bar{z}^*, F(x') - F(\bar{x})\rangle \geq 0$, contradicting the inequality obtained above. Consequently, $\bar{x} \in \mathcal{WE}(VP)$. □

Remark 11.43 In order to guarantee the lower semicontinuity of the functions δ_{Ω_n}, $n \geq 1$, it is enough to have the vector function F only *C-level closed* (i.e. the set $\{x \in X : F(x) \leqq_C y\}$ is closed for any $y \in Y$), a hypotheses weaker than the positive C-lower semicontinuity imposed on F in Theorem 11.17 and Theorem 11.18. However, the latter is also necessary in the proofs of these statements in order to guarantee the lower semicontinuity of the functions $(z^*_n F)$, $n \geq 1$.

Acknowledgements This work was partially supported by FWF (Austrian Science Fund), project M-2045 and DFG (German Research Foundation), project $GR3367/4 - 1$ The author is grateful to an anonymous reviewer for making him aware of the paper [73] and for carefully reading this survey, and to the editors of this volume for the invitation to the CMO-BIRS Workshop on Splitting Algorithms, Modern Operator Theory, and Applications (17w5030) in Oaxaca.

References

1. Aliprantis, C., Florenzano, M., da Rocha, V.M., Tourky, R.: Equilibrium analysis in financial markets with countably many securities. Journal of Mathematical Economics **40**, 683–699 (2004)
2. Alvarez, F.: On the minimizing property of a second order dissipative system in Hilbert spaces. SIAM Journal on Control and Optimization **38**, 1102–1119 (2000)
3. Alvarez, F., Attouch, H.: An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping. Set-Valued Analysis **9**, 3–11 (2001)
4. Apolinário, H., Quiroz, E.P., Oliveira, P.: A scalarization proximal point method for quasiconvex multiobjective minimization. Journal of Global Optimization **64**, 79–96 (2016)
5. Attouch, H., Garrigos, G.: Multiobjective optimization - an inertial dynamical approach to Pareto optima. arXiv **1506.02823** (2015)
6. Attouch, H., Garrigos, G., Goudou, X.: A dynamic gradient approach to Pareto optimization with nonsmooth convex objective functions. Journal of Mathematical Analysis and Applications **422**, 741–771 (2015)
7. Attouch, H., Goudou, X.: A continuous gradient-like dynamical approach to Pareto-optimization in Hilbert spaces. Set-Valued and Variational Analysis **22**, 189–219 (2014)
8. Auslender, A., Teboulle, M.: Interior gradient and proximal methods for convex and conic optimization. SIAM Journal on Optimization **16**, 697–725 (2006)

9. Bauschke, H., Combettes, P.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. CMS Books in Mathematics / Ouvrages de mathématiques de la SMC. Springer-Verlag, New York (2011)
10. Beck, A., Teboulle, M.: A fast iterative shrinkage-tresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**, 183–202 (2009)
11. Benker, H., Hamel, A.H., Tammer, C.: An algorithm for vectorial control approximation problems. In: Multiple Criteria Decision Making (Hagen, 1995), *Lecture Notes in Economics and Mathematical Systems*, vol. 448, pp. 3–12. Springer-Verlag, Berlin (1997)
12. Bento, G.C., da Cruz Neto, J.X., López, G., Soubeyran, A., Souza, J.C.O.: The proximal point method for locally Lipschitz functions in multiobjective optimization with application to the compromise problem. SIAM Journal on Optimization **28**, 1104–1120 (2018)
13. Bento, G.C., da Cruz Neto, J.X., de Meireles, L.V.: Proximal point method for locally Lipschitz functions in multiobjective optimization of Hadamard manifolds. Journal of Optimization Theory and Applications **179**, 37–52 (2018)
14. Bento, G.C., da Cruz Neto, J.X., Soubeyran, A.: A proximal point-type method for multicriteria optimization. Set-Valued and Variational Analysis **22**, 557–573 (2014)
15. Bento, G.C., Ferreira, O.P., Junior, V.L.S.: Proximal point method for a special class of nonconvex multiobjective optimization functions. Optimization Letters **12**, 311–320 (2018)
16. Bento, G.C., Ferreira, O.P., Pereira, Y.R.L.: Proximal point method for vector optimization on Hadamard manifolds. Operations Research Letters **46**, 13–18 (2018)
17. Bento, G.C., Ferreira, O.P., Soubeyran, A., de Sousa Júnior, V.L., Valdinês, L.: Inexact multi-objective local search proximal algorithms: application to group dynamic and distributive justice problems. Journal of Optimization Theory and Applications **177**, 181–200 (2018)
18. Boţ, R.I., Csetnek, E.R., Heinrich, A.: A primal-dual splitting algorithm for finding zeros of sums of maximal monotone operators. SIAM Journal on Optimization **23**, 2011–2036 (2013)
19. Boţ, R.I., Grad, S.M.: Inertial forward-backward methods for solving vector optimization problems. Optimization **67**, 959–974 (2018)
20. Boţ, R.I., Hendrich, C.: A variable smoothing algorithm for solving convex optimization problems. TOP **23**(1), 124–150 (2015)
21. Boţ, R.I., Grad, S.M., Wanka, G.: Duality in Vector Optimization. Vector Optimization. Springer-Verlag, Berlin (2009)
22. Boţ, R.I., Hendrich, C.: A Douglas-Rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators. SIAM Journal on Optimization **23**, 2541–2565 (2013)
23. Bolintineanu, Ş.: Approximate efficiency and scalar stationarity in unbounded nonsmooth convex vector optimization problems. Journal of Optimization Theory and Applications **106**, 265–296 (2000)
24. Bonnel, H., Iusem, A.N., Svaiter, B.F.: Proximal methods in vector optimization. SIAM Journal on Optimization **15**, 953–970 (2005)
25. Borwein, J.M.: Proper efficient points for maximizations with respect to cones. SIAM Journal on Control and Optimization **15**, 57–63 (1977)
26. Borwein, J.M.: The geometry of Pareto efficiency over cones. Mathematische Operationsforschung und Statistik Series Optimization **11**, 235–248 (1980)
27. Buong, N.: Inertial proximal point regularization algorithm for unconstrained vector convex optimization problems. Ukrainian Mathematical Journal **60**, 1483–1491 (2008)
28. Ceng, L.C., Mordukhovich, B.S., Yao, J.C.: Hybrid approximate proximal method with auxiliary variational inequality for vector optimization. Journal of Optimization Theory and Applications **146**, 267–303 (2010)
29. Ceng, L.C., Yao, J.C.: Approximate proximal methods in vector optimization. European Journal of Operational Research **183**, 1–19 (2007)
30. Chen, Z.: Generalized viscosity approximation methods in multiobjective optimization problems. Computational Optimization and Applications **49**, 179–192 (2011)
31. Chen, Z.: Asymptotic analysis in convex composite multiobjective optimization problems. Journal of Global Optimization **55**, 507–520 (2013)

32. Chen, Z., Huang, H., Zhao, K.: Approximate generalized proximal-type method for convex vector optimization problem in Banach spaces. Computers & Mathematics with Applications **57**, 1196–1203 (2009)
33. Chen, Z., Huang, X.X., Yang, X.Q.: Generalized proximal point algorithms for multiobjective optimization problems. Applicable Analysis **90**, 935–949 (2011)
34. Chen, Z., Xiang, C., Zhao, K., Liu, X.: Convergence analysis of Tikhonov-type regularization algorithms for multiobjective optimization problems. Applied Mathematics and Computation **211**, 167–172 (2009)
35. Chen, Z., Zhao, K.: A proximal-type method for convex vector optimization problem in Banach spaces. Numerical Functional Analysis and Optimization **30**, 70–81 (2009)
36. Chuong, T.D.: Tikhonov-type regularization method for efficient solutions in vector optimization. Journal of Computational and Applied Mathematics **234**, 761–766 (2010)
37. Chuong, T.D.: Generalized proximal method for efficient solutions in vector optimization. Numerical Functional Analysis and Optimization **32**, 843–857 (2011)
38. Chuong, T.D., Mordukhovich, B.S., Yao, J.C.: Hybrid approximate proximal algorithms for efficient solutions in vector optimization. Journal of Nonlinear and Convex Analysis **12**, 257–286 (2011)
39. Chuong, T.D., Yao, J.C.: Viscosity-type approximation method for efficient solutions in vector optimization. Taiwanese Journal of Mathematics **14**, 2329–2342 (2010)
40. Cruz, J.Y.B.: A subgradient method for vector optimization problems. SIAM Journal on Optimization **23**, 2169–2182 (2013)
41. Durea, M., Strugariu, R.: Some remarks on proximal point algorithm in scalar and vectorial cases. Nonlinear Functional Analysis and Applications **15**, 307–319 (2010)
42. Fliege, J., Graña Drummond, L.M., Svaiter, B.F.: Newton's method for multiobjective optimization. SIAM Journal on Optimization **20**, 602–626 (2009)
43. Gerstewitz, C.: Nichtkonvexe Dualität in der Vektoroptimierung. Wissenschaftliche Zeitschrift der Technischen Hochschule Carl Schorlemmer Leuna-Merseburg **25**, 357–364 (1983)
44. Gong, X.H.: Optimality conditions for Henig and globally proper efficient solutions with ordering cone has empty interior. Journal of Mathematical Analysis and Applications **307**, 12–31 (2005)
45. Göpfert, A., Riahi, H., Tammer, C., Zălinescu, C.: Variational Methods in Partially Ordered Spaces. CMS Books in Mathematics / Ouvrages de mathématiques de la SMC. Springer-Verlag, New York, New York (2003)
46. Graña Drummond, L.M., Iusem, A.N.: A projected gradient method for vector optimization problems. Computational Optimization and Applications **28**, 5–29 (2004)
47. Graña Drummond, L.M., Maculan, N., Svaiter, B.F.: On the choice of parameters for the weighting method in vector optimization. Mathematical Programming **111**, 201–216 (2008)
48. Graña Drummond, L.M., Svaiter, B.F.: A steepest descent method for vector optimization. Journal of Computational and Applied Mathematics **175**, 395–414 (2005)
49. Grad, S.M.: Vector Optimization and Monotone Operators via Convex Duality. Vector Optimization. Springer-Verlag, Cham (2015)
50. Grad, S.M., Pop, E.L.: Vector duality for convex vector optimization problems by means of the quasi interior of the ordering cone. Optimization **63**, 21–37 (2014)
51. Gregório, R.M., Oliveira, P.R.: A logarithmic-quadratic proximal point scalarization method for multiobjective programming. Journal of Global Optimization **49**, 281–291 (2011)
52. Ji, Y., Goh, M., de Souza, R.: Proximal point algorithms for multi-criteria optimization with the difference of convex objective functions. Journal of Optimization Theory and Applications **169**, 280–289 (2016)
53. Ji, Y., Qu, S.: Proximal point algorithms for vector DC programming with applications to probabilistic lot sizing with service levels. Discrete Dynamics in Nature and Society **- Article ID 5675183** (2017)
54. Kiwiel, K.C.: An aggregate subgradient descent method for solving large convex nonsmooth multiobjective minimization problems. In: A. Straszak (ed.) Large Scale Systems: Theory and Applications 1983, *International Federation of Automatic Control Proceedings Series*, vol. 10, pp. 283–288. Pergamon Press, Oxford (1984)

55. Kiwiel, K.C.: An algorithm for linearly constrained nonsmooth convex multiobjective minimization. In: A. Sydow, S.M. Thoma, R. Vichnevetsky (eds.) Systems Analysis and Simulation 1985 Part I: Theory and Foundations, pp. 236–238. Akademie-Verlag, Berlin (1985)
56. Kiwiel, K.C.: A descent method for nonsmooth convex multiobjective minimization. Large Scale Systems **8**, 119–129 (1985)
57. Luc, D.T.: Theory of Vector Optimization, *Lecture Notes in Economics and Mathematical Systems*, vol. 319. Springer-Verlag, Berlin (1989)
58. Mäkelä, M.M., Karmitsa, N., Wilppu, O.: Proximal bundle method for nonsmooth and nonconvex multiobjective optimization. In: Mathematical Modeling and Optimization of Complex Structures, *Computational Methods in Applied Sciences*, vol. 40, pp. 191–204. Springer-Verlag, Cham (2016)
59. Martinet, B.: Régularisation d'inéquations variationelles par approximations succesives. Revue Française de d'Informatique et de Recherche Opérationnelle **4**, 154–159 (1970)
60. Miettinen, K., Mäkelä, M.M.: An interactive method for nonsmooth multiobjective optimization with an application to optimal control. Optimization Methods and Software **2**, 31–44 (1993)
61. Miettinen, K., Mäkelä, M.M.: Interactive bundle-based method for nondifferentiable multiobjective optimization: NIMBUS. Optimization **34**, 231–246 (1995)
62. Miglierina, E., Molho, E., Recchioni, M.C.: Box-constrained multi-objective optimization: a gradient-like method without "a priori" scalarization. European Journal of Operational Research **188**, 662–682 (2008)
63. Moudafi, A., Oliny, M.: Convergence of a splitting inertial proximal method for monotone operators. Journal of Computational and Applied Mathematics **155**, 447–454 (2003)
64. Mukai, H.: Algorithms for multicriterion optimization. IEEE Transactions on Automatic Control **25**, 177–186 (1980)
65. Opial, Z.: Weak convergence of the sequence of successive approximations for nonexpansive mappings. Bulletin of the American Mathematical Society **73**, 591–597 (1967)
66. Penot, J.P., Théra, M.: Semi-continuous mappings in general topology. Archiv der Mathematik (Basel) **38**, 158–166 (1982)
67. Qu, S., Goh, M., Ji, Y., de Souza, R.: A new algorithm for linearly constrained *c*-convex vector optimization with a supply chain network risk application. European Journal of Operational Research **247**, 359–365 (2015)
68. Qu, S.J., Goh, M., de Souza, R., Wang, T.N.: Proximal point algorithms for convex multi-criteria optimization with applications to supply chain risk management. Journal of Optimization Theory and Applications **163**, 949–956 (2014)
69. Quiroz, E.A.P., Apolinário, H.C.F., Villacorta, K.D.V., Oliveira, P.R.: A linear scalarization proximal point method for quasiconvex multiobjective minimization. arXiv **1510.00461** (2015)
70. Rocha, R.A., Gregório, R.M.: Um algoritmo de ponto proximal inexato para programaçao multiobjetivo. In: Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, vol. 6 (2018)
71. Rocha, R.A., Oliveira, P.R., Gregório, R.M., Souza, M.: Logarithmic quasi-distance proximal point scalarization method for multi-objective programming. Applied Mathematics and Computation **273**, 856–867 (2016)
72. Rocha, R.A., Oliveira, P.R., Gregório, R.M., Souza, M.: A proximal point algorithm with quasi-distance in multi-objective optimization. Journal of Optimization Theory and Applications **171**, 964–979 (2016)
73. Souza, J.C.O.: Proximal point methods for Lipschitz functions on Hadamard manifolds: scalar and vectorial cases. Journal of Optimization Theory and Applications **179**, 745–760 (2018)
74. Tang, F.M., Huang, P.L.: On the convergence rate of a proximal point algorithm for vector function on Hadamard manifolds. Journal of the Operations Research Society of China **5**, 405–417 (2017)
75. Villacorta, K.D.V., Oliveira, P.R.: An interior proximal method in vector optimization. European Journal of Operational Research **214**, 485–492 (2011)

Chapter 12
Non-polyhedral Extensions of the Frank and Wolfe Theorem

Juan Enrique Martínez-Legaz, Dominikus Noll, and Wilfredo Sosa

Abstract In 1956 Marguerite Frank and Paul Wolfe proved that a quadratic function which is bounded below on a polyhedron P attains its infimum on P. In this work we search for larger classes of sets F with this Frank-and-Wolfe property. We establish the existence of non-polyhedral Frank-and-Wolfe sets, obtain internal characterizations by way of asymptotic properties, and investigate stability of the Frank-and-Wolfe class under various operations.

Keywords Quadratic optimization · Asymptotes · Motzkin-sets · Frank-and-Wolfe theorem

AMS 2010 Subject Classification 49M20, 65K10, 90C30

12.1 Introduction

In this paper we investigate extensions of the famous Frank and Wolfe theorem [1, 5–8], which states that a quadratic function f which is bounded below on a closed convex polyhedron P attains its infimum on P. This has applications to linear complementarity problems, and a natural question is whether this property is shared by larger classes of non-polyhedral convex sets F.

J. E. Martínez-Legaz
Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, and Barcelona Graduate School of Mathematics (BGSMath), Barcelona, Spain
e-mail: JuanEnrique.Martinez.Legaz@uab.es

D. Noll (✉)
Université de Toulouse, Institut de Mathématiques, Toulouse, France
e-mail: Dominikus.Noll@math.univ-toulouse.fr

W. Sosa
Programa de Pôs-Graduação em Economia, Universidade Católica de Brasília, Taguatinga, Brazil
e-mail: sosa@ucb.br

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_12

The present work expands on [15], where the Frank-and-Wolfe property was successfully related to asymptotic properties of a set F. Following this line, we presently obtain a complete characterization of the Frank-and-Wolfe property within the class of Motzkin decomposable sets. In particular, the converse of a result of Kummer [12] is obtained.

A second theme addresses versions of the Frank-and-Wolfe theorem where the class of quadratic functions is further restricted. One may, for instance, ask for sets F on which convex or quasi-convex quadratics attain their finite infima. It turns out that this class has a complete characterization as those sets which have no flat asymptotes in the sense of Klee. As a consequence we obtain a version of the Frank-and-Wolfe theorem which extends a result of Rockafellar [17, Sect. 27] and Belousov and Klatte [4] on convex polynomials.

Invariance of Frank-and-Wolfe type sets under various operations such as finite intersections, unions, cross-products, sums, and under affine images and pre-images are also investigated.

The structure of the chapter is as follows. In section 12.2 we give the definition and collect basic information on *FW*-sets. In section 12.3 we consider quasi-Frank-and-Wolfe sets, where a version of the Frank and Wolfe theorem for quasi-convex quadratics is discussed. It turns out that the same class allows many more applications, as it basically suffices to have polynomial functions which have at least one convex sub-level set. In section 12.4 we consider sets with a generalized Motzkin decomposition of the form $F = K + D$ with K compact and D a closed convex cone. This class was used by Kummer [12], who proved a version of the Frank and Wolfe theorem in this class when D is polyhedral. We give a new proof of this result and also establish its converse, that is, if a Motzkin set satisfies the Frank and Wolfe theorem, then the cone D must be polyhedral. Section 12.5 discusses invariance properties of the class of Motzkin sets with the Frank and Wolfe property.

Notations

We generally follow Rockafellar's book [17]. The closure of a set F is $\overline{F}$. The Euclidean norm in $\mathbb{R}^n$ is $\|\cdot\|$, and the Euclidean distance is $\operatorname{dist}(x, y) = \|x - y\|$. For subsets M, N of $\mathbb{R}^n$ we write $\operatorname{dist}(M, N) = \inf\{\|x - y\| : x \in M, y \in N\}$. A direction d with $x + td \in F$ for every $x \in F$ and every $t \geq 0$ is called a direction of recession of F, and the cone of all directions of recession is denoted as 0^+F.

A function $f(x) = \frac{1}{2}x^{\mathsf{T}}Ax + b^{\mathsf{T}}x + c$ with $A = A^{\mathsf{T}} \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$ is called quadratic. The quadratic $f : \mathbb{R}^n \to \mathbb{R}$ is quasi-convex on a convex set $F \subset \mathbb{R}^n$ if the sub-level sets of $f_{|F} : F \to \mathbb{R}$ are convex. Similarly, f is convex on the set F if $f_{|F}$ is convex.

12.2 Frank-and-Wolfe Sets

The following definition is the basis for our investigation:

Definition 12.1 A set $F \subset \mathbb{R}^n$ is called a *Frank-and-Wolfe set*, for short a *FW*-set, if every quadratic function f which is bounded below on F attains its infimum on F.

In [15] this notion was introduced for convex sets F, but in the present note we extend it to arbitrary sets, as this property is not really related to convexity. The classical Frank-and-Wolfe theorem says that every closed convex polyhedron is a *FW*-set, cf. [5–8]. Here we are interested in identifying and characterizing more general classes of sets with this property. We start by collecting some basic information about *FW*-sets.

Proposition 12.1 *Affine images of FW-sets are again FW-sets.*

Proof Let F be a *FW*-set in $\mathbb{R}^n$ and $T : \mathbb{R}^n \to \mathbb{R}^m$ an affine mapping. We have to show that $T(F)$ is a *FW*-set. Let f be a quadratic on $\mathbb{R}^m$ which is bounded below on $T(F)$, then $f \circ T$ is a quadratic on $\mathbb{R}^n$, which is bounded below on F, hence attains its infimum at some $x \in F$. Then f attains its infimum at $Tx \in T(F)$. □

It is equally easy to see that every *FW*-set is closed, because if $x \in \overline{F}$, then the quadratic function $f = \| \cdot - x\|^2$ has infimum 0 on F, and if this infimum is to be attained, then $x \in F$. As a consequence, a bounded set F is *FW* iff it is closed, so there is nothing interesting to report on bounded *FW*-sets, and the property is clearly aimed at the analysis of unbounded sets.

One can go a little further than just proving closedness of *FW*-sets and get first information about their asymptotic behavior. We need the following:

Definition 12.2 An affine manifold M in $\mathbb{R}^n$ is called an *f-asymptote* of the set $F \subset \mathbb{R}^n$ if $F \cap M = \emptyset$ and $\operatorname{dist}(F, M) = 0$.

This expands on Klee [11], who introduced this notion for convex sets F. The symbol f stands for *flat* asymptote. This allows us now to propose the following:

Proposition 12.2 *Let F be a FW-set. Then F has no f-asymptotes.*

Proof Let M be an affine subspace such that $\operatorname{dist}(F, M) = 0$. We have to show that $M \cap F \neq \emptyset$. Let $M = \{x \in \mathbb{R}^n : Ax - b = 0\}$ for a suitable matrix A and vector b. Put $f(x) = \|Ax - b\|^2$, then f is quadratic, and $\gamma = \inf\{f(x) : x \in F\} \geq 0$. Now there exist $x_k \in F$ and $y_k \in M$ with $\operatorname{dist}(x_k, y_k) \to 0$. But $Ay_k = b$, and $\|A(x_k - y_k)\| \leq \|A\|\|x_k - y_k\| \to 0$, hence $Ax_k \to b$, which implies $\gamma = 0$. Now since F is a FW-set, this infimum is attained, hence there exists $x \in F$ with $f(x) = 0$, which means $Ax = b$, hence $x \in M$. That shows $F \cap M \neq \emptyset$, so M is not an f-asymptote of F. □

Remark 12.1 An immediate consequence of Propositions 12.1, 12.2 is that affine images of *FW*-sets, and in particular, projections of *FW*-sets, are always closed.

Yet another trivial fact is the following:

Proposition 12.3 *Finite unions of FW-sets are FW.*

We conclude this preparatory section by looking at invariance of the FW-class under affine pre-images. First we need the following:

Proposition 12.4 *If $F \subset \mathbb{R}^n$ is a FW-set and $M \subset \mathbb{R}^m$ is an affine manifold, then $F \times M$ is a FW-set in $\mathbb{R}^n \times \mathbb{R}^m$.*

Proof Since translates of FW-sets are FW-sets, we may assume that M is a linear subspace, and then there is no loss of generality in assuming that $M = \mathbb{R}^m$. Moreover, by an easy induction argument, we only need to consider the case when $m = 1$, because $F \times \mathbb{R}^m = (F \times \mathbb{R}) \times \mathbb{R}^{m-1}$.

Let q be a quadratic function on $\mathbb{R}^n \times \mathbb{R}$ bounded below on $F \times \mathbb{R}$. We can write $q(x,t) = \frac{1}{2}x^{\mathsf{T}}Ax + \frac{1}{2}bt^2 + tc^{\mathsf{T}}x + d^{\mathsf{T}}x + et + f$ for suitable A, b, c, d, e and f. Clearly, $b \geq 0$, as otherwise q could not be bounded below on $F \times \mathbb{R}$. Now we have $\inf_{(x,t)\in F\times\mathbb{R}} q(x,t) = \inf_{x\in F} \inf_{t\in\mathbb{R}} q(x,t)$.

First consider the case $b > 0$. Then the inner infimum in the preceding expression is attained at $t = -\frac{c^{\mathsf{T}}x+e}{b}$. Hence we have $\inf_{(x,t)\in F\times\mathbb{R}} q(x,t) = \inf_{x\in F} q\left(x, -\frac{c^{\mathsf{T}}x+e}{b}\right)$. Given that $q\left(x, -\frac{c^{\mathsf{T}}x+e}{b}\right)$ is a quadratic function of x and is obviously bounded below on F, it attains its infimum over F at some $\overline{x} \in F$. Therefore q attains its infimum over $F \times \mathbb{R}$ at $\left(\overline{x}, -\frac{c^{\mathsf{T}}\overline{x}+e}{b}\right)$.

Now consider the case $b = 0$, $c \neq 0$. Here F must be contained in the hyperplane $c^{\mathsf{T}}x + e = 0$. Substituting this, we get $\inf_{(x,t)\in F\times\mathbb{R}} q(x,t) = \inf_{x\in F}\left\{\frac{1}{2}x^{\mathsf{T}}Ax + d^{\mathsf{T}}x\right\} + f$. Hence, the quadratic function given by $\frac{1}{2}x^{\mathsf{T}}Ax + d^{\mathsf{T}}x$ is bounded below on F and, for every minimizer $\overline{x} \in F$ and every $t \in \mathbb{R}$, the point $(\overline{x}, t)$ is a minimizer of q over $F \times \mathbb{R}$.

Finally, when $b = 0$, $c = 0$ it follows that we must also have $e = 0$, so q no longer depends on t, and we argue as in the previous case. □

Remark 12.2 As we shall see in the next section (example 1), the cross product $F_1 \times F_2$ of two FW-sets F_i is in general no longer a FW-set, so Proposition 12.4 exploits the very particular situation.

We have the following consequence:

Corollary 12.1 *Let F be a FW-set in $\mathbb{R}^n$ and M an affine subspace of $\mathbb{R}^n$. Then $F + M$ is a FW-set.*

Proof $F \times M$ is a FW-set by Proposition 12.4, and its image under the mapping $(x, y) \to x + y$ is a FW-set by Proposition 12.1, and that set is $F + M$. □

Concerning pre-images, we have the following consequence of Proposition 12.4:

Proposition 12.5 *Let T be an affine operator and suppose the FW-set F is contained in the range of T. Then $T^{-1}(F)$ is a FW-set.*

Proof Since the notion of a FW-set is invariant under translations and under coordinate changes, we can assume that T is a surjective linear operator $T : \mathbb{R}^n \to \mathbb{R}^m$ and $F \subset \mathbb{R}^m$. Now $\widetilde{F} = (T_{|\ker(T)^\perp})^{-1}(F)$ is an affine image of the FW-set F, hence by Proposition 12.1 is a FW-subset of $\mathbb{R}^n$. By Corollary 12.1 the set $\widetilde{F} + \ker(T)$ is a FW-set, but this set is just $T^{-1}(F)$. □

Remark 12.3 It is not clear whether this result remains true when F is not entirely contained in the range of T, i.e., when only $F \cap \text{range}(T) \neq \emptyset$. In contrast, see Corollary 12.3 and Proposition 12.9.

More sophisticated invariance properties of the class of FW-sets will be investigated later. For instance, one may ask whether or under which conditions finite intersections, cartesian products, or closed subsets of FW-sets are again FW.

12.3 Frank-and-Wolfe Theorems for Restricted Classes of Quadratic Functions

Following [15] it is also of interest to investigate versions of the Frank and Wolfe theorem, where the class of quadratic functions is further restricted. The following notion is from [15]:

Definition 12.3 A convex set $F \subset \mathbb{R}^n$ is called a *quasi-Frank-and-Wolfe set*, for short a qFW-set, if every quadratic function f which is quasi-convex on F and bounded below on F attains its infimum on F.

Note that for the class of qFW-sets we have to maintain convexity as part of the definition, because quasi-convex functions have to be defined on convex sets. (Otherwise, for instance with $F = \{(x, y) \in \mathbb{R}^2 : y = x^2\}$, the norm squared $f(x, y) = x^2 + y^2$ would not be quasi-convex on F). Hence the notion is precisely as introduced in [15].

Remark 12.4 Every convex FW-set is clearly a qFW-set. The converse is not true, i.e., qFW-sets need not be FW, as will be seen in Example 12.1. It is again clear that qFW-sets are closed, and that affine images of qFW-sets are qFW.

It turns out that f-asymptotes are the key to understanding the quasi-Frank-and-Wolfe property. We have the following:

Theorem 12.1 *Let F be a convex set in $\mathbb{R}^n$. Then the following statements are equivalent:*

(1) *Every polynomial f which has at least one nonempty convex sub-level set on F and which is bounded below on F attains its infimum on F.*
(2) *F is a qFW-set.*

(3) *Every quadratic function q which is convex on F and bounded below on F attains its infimum on F.*
(4) *F has no f-asymptotes.*
(5) *$T(F)$ is closed for every affine mapping T.*
(6) *$P(F)$ is closed for every orthogonal projection P.*

Proof The implication (1) $\implies$ (2) is clear, because for a quasi-convex function on F *every* sub-level set on F is convex. The implication (2) $\implies$ (3) is also evident. Implication (3) $\implies$ (4) follows immediately with the same proof as Proposition 12.2, because the quadratic $f(x) = \|Ax - b\|^2$ used there is convex.

Let us prove (4) $\implies$ (5). We may without loss of generality assume that T is linear, as properties (4) and (5) are invariant under translations. Suppose $T(F)$ is not closed and pick $y \in \overline{T(F)} \setminus T(F)$. Put $M = T^{-1}(y)$, then M is an affine manifold. Note that $M \cap F = \emptyset$, because $T(M) = \{y\}$. Now pick $y_k \in T(F)$ such that $y_k \to y$, and choose $x_k \in T^{-1}(y_k) \cap F$. Since T is linear, and hence an isomorphism from $\ker(T)^\perp$ onto $T(\mathbb{R}^n)$, there exist $x'_k \in T^{-1}(y_k)$ such that $x'_k \to x' \in T^{-1}(y)$. (Take $x'_k = (T_{|\ker(T)^\perp})^{-1}(y_k)$). We have $\|x_k - (x' - x'_k + x_k)\| \to 0$, with $x_k \in F$, and since $x_k - x'_k \in \ker(T)$, we have $x' - x'_k + x_k \in x' + \ker(T) = M$. That proves $\operatorname{dist}(F, M) = 0$, and so F has M as an f-asymptote, a contradiction.

The implication (5) $\implies$ (6) is clear. Let us prove (6) $\implies$ (1). We will prove this by induction on n. For $n = 1$ the implication is clearly true, because any polynomial $f : \mathbb{R} \to \mathbb{R}$ which is bounded below on a convex set $F \subset \mathbb{R}$ satisfying (6) attains its infimum on F, as (6) implies that F is closed. Suppose therefore that the result is true for dimension $n - 1$, and consider a polynomial $f : \mathbb{R}^n \to \mathbb{R}$ which is bounded below on a set $F \subset \mathbb{R}^n$ with property (6) such that $S_\alpha := \{x \in F : f(x) \leq \alpha\}$ is nonempty and convex for some $\alpha \in \mathbb{R}$. We may without loss of generality assume that the dimension of F is n, i.e., that F has nonempty interior, as otherwise F is contained in a hyperplane, and then the result follows directly from the induction hypothesis. If $\alpha = \gamma := \inf\{f(x) : x \in F\}$, then f clearly attains α, so we assume from now on that $\alpha > \gamma$. If $S_\alpha := \{x \in F : f(x) \leq \alpha\}$ is bounded, then by the Weierstrass extreme value theorem the infimum of f over S_α is attained, because by hypothesis (6) the set F is closed. But this infimum is also the infimum of f over F, so in this case we are done. Assume therefore that S_α is unbounded. Since S_α is a closed convex set, it has a direction of recession d, that is, $x + td \in S_\alpha$ for every $t \geq 0$ and every $x \in S_\alpha$. Fix $x \in S_\alpha$. This means

$$\gamma \leq f(x + td) \leq \alpha \tag{12.1}$$

for every $t \geq 0$. Since $t \mapsto f(x + td)$ is a polynomial on the real line, which is now bounded on $[0, \infty)$, it must be constant as a function of t, so that $f(x) = f(x + td)$ for all $t \geq 0$, and then clearly also $f(x + td) = f(x)$ for every $t \in \mathbb{R}$. But the argument is valid for every $x \in S_\alpha$. By assumption F has dimension n, so S_α has nonempty interior. That shows $f(x + td) = f(x)$ for all x in a nonempty open set contained in S_α and all $t \in \mathbb{R}$. Altogether, since f is a polynomial, we obtain

$$f(x + td) = f(x) \text{ for every } x \in \mathbb{R}^n \text{ and every } t \in \mathbb{R}. \tag{12.2}$$

Now let P be the orthogonal projection onto the hyperplane $H = d^{\perp}$. Then $\widetilde{f} := f_{|H}$ is a polynomial on the $(n-1)$-dimensional space H and takes the same values as f due to (12.2). In particular, $\widetilde{f} = f_{|H}$ is bounded below on the set $\widetilde{F} = P(F)$.

We argue that the induction hypothesis applies to $\widetilde{F}$. Indeed, $\widetilde{F}$ being the image of F under a projection, is closed by condition (6). Its dimension is $n-1$, and moreover, every projection of $\widetilde{F}$ is closed, because any such projection is also a projection of F.

It remains to prove that the restriction of $\widetilde{f}$ to $\widetilde{F}$ has a nonempty convex sub-level set. To this end it will suffice to prove that, for $\widetilde{S}_\alpha := \{x \in \widetilde{F} : \widetilde{f}(x) \leq \alpha\}$, one has $\widetilde{S}_\alpha = P(S_\alpha)$. This will easily follow from the observation that $\widetilde{f} \circ P = f$, which is an immediate consequence of (12.2). Let $x \in \widetilde{S}_\alpha$. Since $x \in \widetilde{F}$, we have $P(x') = x$ for some $x' \in F$, and hence $f(x') = (\widetilde{f} \circ P)(x') = \widetilde{f}(P(x')) = \widetilde{f}(x) \leq \alpha$, which proves that $x' \in S_\alpha$. Therefore $x \in P(S_\alpha)$, which shows $\widetilde{S}_\alpha \subset P(S_\alpha)$. To prove the opposite inclusion, let $x \in P(S_\alpha)$. We then have $x = P(x')$ for some $x' \in S_\alpha$. From the inclusion $S_\alpha \subset F$, it follows that $x \in P(F) = \widetilde{F}$. On the other hand, $\widetilde{f}(x) = f(x') \leq \alpha$. This shows $x \in \widetilde{S}_\alpha$ and proves the inclusion $P(S_\alpha) \subset \widetilde{S}_\alpha$ and hence our claim $\widetilde{S}_\alpha = P(S_\alpha)$.

Altogether, $\widetilde{f}$ now attains its infimum on $\widetilde{F}$ by the induction hypothesis, and then f, having the same values, also attains its infimum on F. This proves the validity of (1). □

Remark 12.5 The equivalence of (4) and (6) can already be found in [11].

Remark 12.6 All that matters in condition (1) is the *rigidity* of polynomials. Any class $\mathcal{F}(L)$ of continuous functions defined on affine subspaces L of $\mathbb{R}^n$ with the following properties would work as well:

(i) $\mathcal{F}(L)$ is defined for every $L \subset \mathbb{R}^n$ and every n.
(ii) If $f \in \mathcal{F}(\mathbb{R})$ is bounded below on a closed interval on $\mathbb{R}$, then f attains its infimum.
(iii) If $f \in \mathcal{F}(\mathbb{R}^n)$ and H is a hyperplane in $\mathbb{R}^n$, then $f_{|H} \in \mathcal{F}(H)$.
(iv) If $f \in \mathcal{F}(\mathbb{R}^n)$ is bounded (above and below) on some ray $x + \mathbb{R}^+ d \subset \mathbb{R}^n$, then f does not depend on d, i.e., $f(x) = f(x + td)$ for all $t \in \mathbb{R}$.

We had seen in section 12.2 that *FW*-sets have no f-asymptotes. Moreover, from the results of this section we see that if F is convex and has no f-asymptotes, then it is already a *qFW*-set. This raises the question whether the absence of f-asymptotes also serves to characterize *FW*-sets, or if not, whether it does so at least for convex F. We indicate by way of two examples that this is not the case, i.e., the absence of f-asymptotes does *not* characterize Frank-and-Wolfe sets. Or put differently, there exist quasi-Frank-and-Wolfe sets which are not Frank-and-Wolfe.

Example 12.1 We construct a closed convex set F without f-asymptotes which is not Frank-and-Wolfe. We use Example 2 of [14], which we reproduce here for convenience. Consider the optimization program

$$\begin{aligned} \text{minimize } & q(x) = x_1^2 - 2x_1x_2 + x_3x_4 + 1 \\ \text{subject to } & c_1(x) = x_1^2 - x_3 \leq 0 \\ & c_2(x) = x_2^2 - x_4 \leq 0 \\ & x \in \mathbb{R}^4 \end{aligned}$$

then as Lou and Zhang [14] show the constraint set $F = \{x \in \mathbb{R}^4 : c_1(x) \leq 0, c_2(x) \leq 0\}$ is closed and convex, and the quadratic function q has infimum $\gamma = 0$ on F, but this infimum is not attained.

Let us show that F has no f-asymptotes. Note that $F = F_1 \times F_2$, where $F_1 = \{(x_1, x_3) \in \mathbb{R}^2 : x_1^2 - x_3 \leq 0\}$, $F_2 = \{(x_2, x_4) \in \mathbb{R}^2 : x_2^2 - x_4 \leq 0\}$. Observe that $F_1 \cong F_2$, and that F_1 does not have asymptotes, being a parabola. Therefore, F does not have f-asymptotes either. This can be seen from the following:

Proposition 12.6 *Any nonempty finite intersection of qFW-sets is again a qFW-set.*

Proof By Theorem 12.1 the result follows immediately from a theorem of Klee [11, Thm. 4], which says that finite intersections of sets without f-asymptotes have no f-asymptotes. □

Corollary 12.2 *If $F_1, \ldots, F_m$ are qFW-sets, then the cartesian product $F_1 \times \cdots \times F_m$ is again a qFW-set.*

Proof Consider for the ease of notation the case of two sets $F_i \subset \mathbb{R}^{d_i}$, $i = 1, 2$. Then write

$$F_1 \times F_2 = \left(F_1 \times \mathbb{R}^{d_2}\right) \cap \left(\mathbb{R}^{d_1} \times F_2\right).$$

Now $F_1 \times \mathbb{R}^{d_2}$ is also qFW, and so is $\mathbb{R}^{d_1} \times F_2$, and hence the result follows from Proposition 12.6. The fact that $F_1 \times \mathbb{R}^{d_2}$ is qFW is easily seen as follows: If M is a f-asymptote of $F_1 \times \mathbb{R}^{d_2}$, then $L = \{x : (x, y) \in M \text{ for some } y\}$ is a f-asymptote of F_1. □

Remark 12.7 Example 1 also tells us that the sum of FW-sets need not be a FW-set even when closed, as follows from the identity $F_1 \times F_2 = (F_1 \times \{0\}) + (\{0\} \times F_2)$. Note that even though $F_1 \times F_2$ fails to be FW, it remains qFW due to Corollary 12.2.

Example 12.2 Let F be the epigraph of $f(x) = x^2 + \exp(-x^2)$ in $\mathbb{R}^2$. Then $q(x, y) = y - x^2$ is bounded below on F, but does not attain its infimum, so F is not FW. However, F has no f-asymptotes, so it is qFW.

Remark 12.8 In [15] it is shown explicitly that the ice-cream cone is not qFW. Here is a simple synthetic argument. The ice cream cone $D \subset \mathbb{R}^3$ can be cut by a plane L in such a way that $F = D \cap L$ has a hyperbola as boundary curve. Since F has asymptotes, it is not qFW, hence neither is the cone D.

The method of proof in implication (6) $\Longrightarrow$ (1) in Theorem 12.1 can be used to show that sub-level sets of convex polynomials are *qFW*-sets, see [3, Chap. II, §4, Thm. 13]. We obtain the following extension of [4, Thm. 3]:

Corollary 12.3 *Let F_0 be a qFW-set and let $f_1, \ldots, f_m$ be convex polynomials on F_0 such that the set $F = \{x \in F_0 : f_i(x) \leq 0, i = 1, \ldots, m\}$ is non-empty. Let f be a polynomial which is bounded below on F and has at least one nonempty convex sub-level set on F. Then f attains its infimum on F.*

Remark 12.9 From Corollary 12.2 and Proposition 12.6 we learn that the class of *qFW*-sets is closed under finite intersections and cross products, while Example 12.1 tells us that this is no longer true for *FW*-sets. Yet another invariance property of *qFW*-sets is the following:

Corollary 12.4 *Let $T : \mathbb{R}^n \to \mathbb{R}^m$ be an affine operator, and let $F \subset \mathbb{R}^m$ be a qFW-set. If $T^{-1}(F)$ is nonempty, then it is a qFW-set, too.*

Proof We use property (4) of Theorem 12.1. Suppose $T^{-1}(F)$ had an f-asymptote M, then $T(M)$ would be an f-asymptote of F. □

Corollary 12.5 (See [4], [17, Cor. 27.3.1]) *Let f be a polynomial which is convex and bounded below on a qFW-set F. Then f attains its infimum on F.*

The following consequence of Theorem 12.1 is surprising.

Corollary 12.6 *Let F be a convex cone. Then the following are equivalent:*

(1) *F is a FW-set;*
(2) *F is a qFW-set;*
(3) *F is polyhedral.*

Proof (1) $\Longrightarrow$ (2) is clear, because F is convex. (2) $\Longrightarrow$ (3): Let $F \subset \mathbb{R}^n$ be *qFW*, then by condition (iv) of Theorem 12.1 every orthogonal projection $P(F)$ on any two-dimensional subspace of $\mathbb{R}^n$ is closed. Therefore, by Mirkil's theorem, which we give as Lemma 12.1 below, F is polyhedral. (3) $\Longrightarrow$ (1): By the classical Frank-and-Wolfe theorem every polyhedral convex cone is *FW*. □

Lemma 12.1 (Mirkil's Theorem [16]) *Let D be a convex cone in $\mathbb{R}^n$ such that every orthogonal projection on any two-dimensional subspace is closed. Then D is polyhedral.*

Remark 12.10 This result puts an end to hopes to get new results for the linear complementarity problem by investigating *FW*-cones.

We end this section with a nice consequence of Mirkil's theorem. First we need the following characterization of f-asymptotes:

Proposition 12.7 *For a closed convex set F and a linear subspace L, the following statements are equivalent:*

1) *No translate of L is an f-asymptote of F.*
2) *The orthogonal projection of F onto the orthogonal complement $L^\perp$ is closed.*
3) *$F + L$ is closed.*

Proof 1)⇒2) Let $x \in \overline{P_{L^\perp}(F)}$. Since $P_{L^\perp}^{-1}(x) = x + L$, we can easily prove that $\operatorname{dist}(F, x + L) = \operatorname{dist}\left(P_{L^\perp}(F), x\right) = 0$. Since $x + L$ is not an f-asymptote of F, we have $F \cap (x + L) \neq \emptyset$, which amounts to saying that $x \in P_{L^\perp}(F)$.

2)⇒3) Let $x_k \in F$ and $y_k \in L$ $(k = 1, 2, \ldots)$ be such that the sequence $x_k + y_k$ converges to some point z. Then $P_{L^\perp}(z) = \lim P_{L^\perp}(x_k + y_k) = \lim P_{L^\perp}(x_k) \in P_{L^\perp}(F)$ due to closedness of $P_{L^\perp}(F)$. But $P_{L^\perp}(F) = (F + L) \cap L^\perp \subset F + L$, hence $P_{L^\perp}(z) \in F + L$. Now $z = P_{L^\perp}(z) + P_L(z) \in F + L + L = F + L$.

3)⇒1) Let us assume that $x + L$ is an f-asymptote of F for some x. Then $0 \leq \operatorname{dist}(x, F + L) \leq \operatorname{dist}\left(x, (F + L) \cap L^\perp\right) = \operatorname{dist}\left(x, P_{L^\perp}(F)\right) = \operatorname{dist}(F, x + L) = 0$, hence $\operatorname{dist}(x, F + L) = 0$. Since $F + L$ is closed, this implies $x \in F + L$. This is equivalent to saying that $F \cap (x + L) \neq \emptyset$, a contradiction to the assumption that $x + L$ is an f-asymptote of F. □

Remark 12.11 An immediate consequence is that a convex set F is a *qFW*-set if and only if $F + L$ is closed for every linear subspace L.

The consequence of Mirkil's Theorem we have in mind is the following:

Proposition 12.8 *For a closed convex cone D in $\mathbb{R}^n$ (with $n > 2$), the following statements are equivalent:*

1) *D is polyhedral.*
2) *$C + D$ is a convex polyhedron for every convex polyhedron C.*
3) *$L + D$ is closed for every $(n - 2)$-dimensional subspace L.*
4) *D has no $(n - 2)$-dimensional f-asymptotes.*

Proof Implications 1) ⇒ 2) ⇒ 3) are immediate. Implication 3) ⟹ 1) is a consequence of 3) ⇒ 2) of Proposition 12.7 combined with Mirkil's Theorem. Implication 3) ⟹ 4) follows from 3) ⟹ 1) of Proposition 12.7. Finally, implication 4) ⟹ 3) can be easily derived from implication 1) ⟹ 3) of Proposition 12.7. □

12.4 Motzkin Type Sets

Following [9, 10], a convex set F is called Motzkin decomposable if it may be written as the Minkowski sum of a compact convex set C and a closed convex cone D, that is, $F = C + D$. Motzkin's classical result states that every convex polyhedron has such a decomposition. We extend this definition as follows:

Definition 12.4 A closed set $F \subset \mathbb{R}^n$ is called a *Motzkin set*, for short an M-set, if it can be written as $F = K + D$, where K is a compact set and D is a closed convex cone.

We shall continue to reserve the term Motzkin decomposable for the case where the set F is convex. A Motzkin set F which is convex is then clearly Motzkin decomposable.

Remark 12.12 Let $F = K + D$ be a Motzkin set, then similarly to the convex case D is uniquely determined by F. Indeed, taking convex hulls, we have $\mathrm{co}(F) = \mathrm{co}(K) + \mathrm{co}(D) = \mathrm{co}(K) + D$, hence $\mathrm{co}(F)$ is a convex Motzkin set, i.e., a Motzkin decomposable set. Then from known results on Motzkin decomposable sets [9, 10], $D = 0^+\mathrm{co}(F)$, the recession cone of $\mathrm{co}(F)$. Now if we define the recession cone of F in the same way as in the convex case, i.e., $0^+F = \{u \in \mathbb{R}^n : x + tu \in F \text{ for all } x \in F \text{ and all } t \geq 0\}$, then $0^+F \subset 0^+\mathrm{co}(F) = D \subset 0^+F$, proving $D = 0^+F$. In particular, F and $\mathrm{co}(F)$ have the same recession cone.

Theorem 12.2 *Let F be a Motzkin set in $\mathbb{R}^n$, represented as $F = K + D = K + 0^+F$. Then the following are equivalent:*

(1) *F is a FW-set.*
(2) *The recession cone 0^+F of F is polyhedral.*
(3) *F has no f-asymptotes.*

Proof We prove (1) $\Longrightarrow$ (2). Let P be an orthogonal projection of $\mathbb{R}^n$ onto a subspace L of $\mathbb{R}^n$. Since $F = K + D$ is a FW-set, $P(F)$ is closed. Since $P(F) = P(K) + P(D)$ and $\overline{P(F)} = P(K) + \overline{P(D)}$, this means $P(K) + P(D) = P(K) + \overline{P(D)}$. We have to show that this implies $P(D) = \overline{P(D)}$. This follows from the so-called *order cancellation law*, which we give as Lemma 12.2 below. It is applied to the convex sets $A = \overline{P(D)}$, $B = P(D)$, and for the compact set $P(K)$. This shows indeed $\overline{P(D)} = P(D)$. This means every projection of D is closed, hence by Mirkil's theorem (Lemma 12.1), the cone D is polyhedral.

Lemma 12.2 (Order Cancellation Law, see [10]) *Let $A, B \subset \mathbb{R}^n$ be convex sets, $K \subset \mathbb{R}^n$ a compact set. If $A + K \subset B + K$, then $A \subset B$.*

Let us now prove (2) $\Longrightarrow$ (1). Write $F = K + D$ for K compact and D a polyhedral convex cone. Now consider a quadratic function $q(x) = \frac{1}{2}x^{\mathsf{T}}Ax + b^{\mathsf{T}}x$ bounded below by γ on F. Hence

$$\inf_{x\in F} q(x) = \inf_{y\in K}\inf_{z\in D} q(y+z) = \inf_{y\in K}\left(q(y) + \inf_{z\in D}\left[y^{\mathsf{T}}Az + q(z)\right]\right) \geq \gamma. \tag{12.3}$$

Observe that for fixed $y \in K$ the function $q_y : z \mapsto y^{\mathsf{T}}Az + q(z)$ is bounded below on D by $\eta = \gamma - \max_{y'\in K} q(y')$. Indeed, for $z \in D$ we have

$$\begin{aligned} y^{\mathsf{T}}Az + q(z) &\geq \left(q(y) + \inf_{z' \in D} \left[y^{\mathsf{T}}Az' + q(z') \right] \right) - q(y) \\ &\geq \inf_{y \in K} \left(q(y) + \inf_{z' \in D} \left[y^{\mathsf{T}}Az' + q(z') \right] \right) - \max_{y' \in K} q(y') \\ &\geq \gamma - \max_{y' \in K} q(y') = \eta. \end{aligned}$$

Since q_y is a quadratic function bounded below on the polyhedral cone D, the inner infimum is attained at some $z = z(y)$. This is in fact the classical Frank and Wolfe theorem on a polyhedral cone. In consequence the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ defined as

$$f(y) = \inf_{z \in D} \left(y^{\mathsf{T}}Az + q(z) \right),$$

satisfies $f(y) = y^{\mathsf{T}}Az(y) + q(z(y)) > -\infty$ for every $y \in K$, so the compact set K is contained in the domain of f. But now a stronger result holds, which one could call a parametric Frank and Wolfe theorem, and which we shall prove in Lemma 12.3 below. We show that f is continuous relative to its domain. Once this is proved, the infimum (12.3) can then be written as

$$\inf_{x \in F} q(x) = \inf_{y \in K} \left(q(y) + f(y) \right),$$

and this is now attained by the Weierstrass extreme value theorem due to the continuity of $q + f$ on the compact K. Continuity of f on K is now a consequence of the following □

Lemma 12.3 *Let D be a polyhedral convex cone and define*

$$f(c) = \inf_{x \in D} \left(c^{\mathsf{T}}x + \tfrac{1}{2}x^{\mathsf{T}}Gx \right),$$

where $G = G^{\mathsf{T}}$. Then $\mathrm{dom}(f)$ *is a polyhedral convex cone, and f is continuous relative to* $\mathrm{dom}(f)$.

Proof If $x^{\mathsf{T}}Gx < 0$ for some $x \in D$, then $\mathrm{dom}(f) = \emptyset$, so we may assume for the remainder of the proof that $x^{\mathsf{T}}Gx \geq 0$ for every $x \in D$. The proof is now divided into three parts. In part 1) we establish a formula for the domain $\mathrm{dom}(f)$. In part 2) we use this formula to show that $\mathrm{dom}(f)$ is polyhedral, and in part 3) we show that the latter implies continuity of f relative to $\mathrm{dom}(f)$.

(1) We start by proving that

$$\mathrm{dom}(f) = \left\{ c : c^{\mathsf{T}}x \geq 0 \text{ for every } x \in D \text{ such that } x^{\mathsf{T}}Gx = 0 \right\}. \tag{12.4}$$

The inclusion $\subseteq$ being obvious, we have to prove the following implication:

$$c^{\mathsf{T}}x \geq 0 \text{ for every } x \in D \text{ such that } x^{\mathsf{T}}Gx = 0 \Longrightarrow \inf_{x\in D}\left(c^{\mathsf{T}}x + \tfrac{1}{2}x^{\mathsf{T}}Gx\right) > -\infty.$$

We establish this by induction on the number l of generators of D. The case $l = 1$ being clear, let $l > 1$, and suppose the implication is correct for every polyhedral convex cone D' with $l' < l$ generators. Let c be such that $c^{\mathsf{T}}x \geq 0$ for every $x \in D$ having $x^{\mathsf{T}}Gx = 0$. We have to show that $c \in \mathrm{dom}(f)$. Assume on the contrary that

$$\inf_{x\in D}\left(c^{\mathsf{T}}x + \tfrac{1}{2}x^{\mathsf{T}}Gx\right) = -\infty, \tag{12.5}$$

and choose a sequence $x_k \in D$ with $\|x_k\| \to \infty$ such that

$$c^{\mathsf{T}}x_k + \tfrac{1}{2}x_k^{\mathsf{T}}Gx_k \longrightarrow -\infty. \tag{12.6}$$

Passing to a subsequence, we can assume that the sequence $y_k = x_k/\|x_k\|$ converges to some $y \in D$. We must have $y^{\mathsf{T}}Gy = 0$, as otherwise we would have $c^{\mathsf{T}}x_k + \frac{1}{2}x_k^{\mathsf{T}}Gx_k = \|x_k\|\, c^{\mathsf{T}}y_k + \frac{1}{2}\|x_k\|^2\, y_k^{\mathsf{T}}Gy_k \longrightarrow +\infty$, a contradiction. Hence, by our assumption, $c^{\mathsf{T}}y \geq 0$. We cannot have $c^{\mathsf{T}}y > 0$, as otherwise for large enough k we would have $c^{\mathsf{T}}x_k = \|x_k\|\, c^{\mathsf{T}}y_k > 0$ and thus $c^{\mathsf{T}}x_k + \frac{1}{2}x_k^{\mathsf{T}}Gx_k > 0$ due to $x_k^{\mathsf{T}}Gx_k \geq 0$, which is impossible because of (12.6). Therefore $c^{\mathsf{T}}y = 0$. This will be used later.

Collecting more facts about y, note that as a consequence of our standing assumption $x^{\mathsf{T}}Gx \geq 0$ for $x \in D$, y is a minimizer of the quadratic form $\frac{1}{2}x^{\mathsf{T}}Gx$ over D, which implies that Gy belongs to the positive polar cone of D, that is, $x^{\mathsf{T}}Gy \geq 0$ for every $x \in D$. This property will also be used below.

Let $E = \{e_1, \ldots, e_l\}$ be the set of generating rays of D, and for $i = 1, \ldots, l$ denote by D_i and $\widehat{D}_i$ the cones generated by $E \setminus \{e_i\}$ and $(E \setminus \{e_i\}) \cup \{y\}$, respectively. As the induction hypothesis applies to each D_i, we have $\inf_{x\in D_i}\left(c^{\mathsf{T}}x + \frac{1}{2}x^{\mathsf{T}}Gx\right) > -\infty$ for every i, so the infimum m of $c^{\mathsf{T}}x + \frac{1}{2}x^{\mathsf{T}}Gx$ over $\bigcup_{i=1}^{l} D_i$ is finite.

Now observe that

$$D = \bigcup_{i=1}^{l} \widehat{D}_i. \tag{12.7}$$

Indeed, the inclusion $\supseteq$ being clear, take $x \in D$ and write it as $x = \sum_{i=1}^{l} \lambda_i e^i$ for certain $\lambda_i \geq 0$. Since $y \in D \setminus \{0\}$, we have $y = \sum_{i\in I} \mu_i e^i$ for some $\emptyset \neq I \subset \{1, \ldots, l\}$ and $\mu_i > 0$. Put $\nu = \min\{\lambda_i/\mu_i : i \in I\} =: \lambda_{i_0}/\mu_{i_0}$, then

$$x = \sum_{i\in I} \lambda_i e^i + \sum_{j\notin I} \lambda_j e^j + \nu \left(y - \sum_{i\in I} \mu_i e^i \right) = \sum_{i\in I} (\lambda_i - \nu\mu_i)\, e^i + \sum_{j\notin I} \lambda_j e^j + \nu y.$$

Since $\lambda_i - \nu\mu_i \geq 0$ for every $i \in I$, and $\lambda_{i_0} - \nu\mu_{i_0} = 0$, we have shown $x \in \widehat{D}_{i_0}$. That proves (12.7).

Now, using (12.7), for every $x \in D$ there exist $i \in \{1, \ldots, l\}$, $z \in D_i$, and $\lambda \geq 0$ such that $x = z + \lambda y$. We then have $c^\mathsf{T} x + \frac{1}{2} x^\mathsf{T} G x = c^\mathsf{T} z + \lambda c^\mathsf{T} y + \frac{1}{2} z^\mathsf{T} G z + \lambda z^\mathsf{T} G y + \frac{1}{2}\lambda^2 y^\mathsf{T} G y = c^\mathsf{T} z + \frac{1}{2} z^\mathsf{T} G z + \lambda z^\mathsf{T} G y \geq c^\mathsf{T} z + \frac{1}{2} z^\mathsf{T} G z \geq m$, which gives $\inf_{x\in D} \left(c^\mathsf{T} x + \frac{1}{2} x^\mathsf{T} G x \right) = m$, contradicting (12.5). This shows that our claim (12.4) was correct.

(2) Now by the Farkas-Minkowski-Weyl theorem (cf. [17, Thm. 19.1] or [18, Cor. 7.1a]) the polyhedral cone D is the linear image of the positive orthant of a space $\mathbb{R}^p$ of appropriate dimension, i.e. $D = \{Zu : u \in \mathbb{R}^p, u \geq 0\}$. Using (12.4), this implies

$$\mathrm{dom}(f) = \{c : c^\mathsf{T} Z u \geq 0 \text{ for every } u \geq 0 \text{ such that } u^\mathsf{T} Z^\mathsf{T} G Z u = 0\}.$$

Now observe that if $u \geq 0$ satisfies $u^\mathsf{T} Z^\mathsf{T} G Z u = 0$, then it is a minimizer of the quadratic function $u^\mathsf{T} Z^\mathsf{T} G Z u$ on the cone $u \geq 0$, hence $Z^\mathsf{T} G Z u \geq 0$ by the Kuhn-Tucker conditions. Therefore we can write the set $P = \{u \in \mathbb{R}^p : u \geq 0, u^\mathsf{T} Z^\mathsf{T} G Z u = 0\}$ as

$$P = \bigcup_{I\subset\{1,\ldots,p\}} P_I,$$

where the P_I are the polyhedral convex cones

$$P_I = \{u \geq 0 : Z^\mathsf{T} G Z u \geq 0, u_i = 0 \text{ for all } i \in I, (Z^\mathsf{T} G Z u)_j = 0 \text{ for all } j \notin I\}.$$

For every $I \subset \{1, \ldots, p\}$ choose m_I generators $u_{I1}, \ldots, u_{Im_I}$ of P_I. Then,

$$\begin{aligned} \mathrm{dom}(f) &= \left\{ c : c^\mathsf{T} Z u \geq 0 \text{ for every } u \in P \right\} \\ &= \left\{ c : c^\mathsf{T} Z u \geq 0 \text{ for every } u \in \textstyle\bigcup_{I\subset\{1,\ldots,p\}} P_I \right\} \\ &= \textstyle\bigcap_{I\subset\{1,\ldots,p\}} \left\{ c : c^\mathsf{T} Z u \geq 0 \text{ for every } u \in P_I \right\} \\ &= \textstyle\bigcap_{I\subset\{1,\ldots,p\}} \left\{ c : c^\mathsf{T} Z u_{Ij} \geq 0 \text{ for all } j = 1, \ldots, m_I \right\}. \end{aligned} \tag{12.8}$$

Since a finite intersection of polyhedral cones is polyhedral, this proves that $\mathrm{dom}(f)$ is a polyhedral convex cone.

(3) To conclude, continuity of f relative to its domain now follows from polyhedrality of $\mathrm{dom}(f)$, and using [17, Thm. 10.2], since f is clearly concave and upper semicontinuous. This completes the proof of (2) $\Longrightarrow$ (1).

(1) $\Longrightarrow$ (3) was proved in Proposition 12.2. Let us prove (3) $\Longrightarrow$ (2). By Mirkil's theorem (Lemma 12.1) it suffices to show that every orthogonal projection $P(F)$ is closed. Suppose this is not the case, and let $y \in \overline{P(F)} \setminus P(F)$. Let $L = y + \ker(P)$, then $F \cap L = \emptyset$. Now choose $y_k \in F$ such that $P(y_k) \to P(y) = y$. Then $y_k = P(y_k) + z_k$ with $z_k \in \ker(P)$. Hence $P(y) + z_k \in L$, but $\|(P(y_k) + z_k) - (P(y) + z_k)\| \to 0$, which shows $\mathrm{dist}(F, L) = 0$. That means that F has an f-asymptote, a contradiction. □

Remark 12.13 The main implication (2) $\Longrightarrow$ (1) in Theorem 12.2 was first proved by Kummer [12]. Our proof of (2) $\Longrightarrow$ (1) is slightly stronger in so far as it gives additional information on the polyhedrality of the domain of f in Lemma 12.3.

Remark 12.14 We refer to Bank *et al.* [2, Thm. 5.5.1 (4)] for a result related to Lemma 12.3 in the case where $G \succeq 0$. For the indefinite case see also Tam [19]. For further comments on this result, see also Klatte [4].

Remark 12.15 The statement of Theorem 12.2 is no longer correct if one drops the hypothesis that F is a Motzkin set. We take the convex $F = \{(x, y) \in \mathbb{R}^2 : x > 0, y > 0, xy \geq 1\}$, then F, being limited by a hyperbola, has f-asymptotes, hence is not qFW, but 0^+F is the positive orthant, which is polyhedral.

Corollary 12.7 *A Motzkin decomposable set F without f-asymptotes is Frank-and-Wolfe.*

Proof Since F has no f-asymptotes and is convex, it is a qFW-set by Theorem 12.1. But then by Theorem 12.2, F is even a FW-set. □

Remark 12.16 Let $F = K + 0^+F = K + D$ be a Motzkin set, then as our analysis shows, we need 0^+F to be polyhedral if we want *all* affine images $T(F)$ of F closed. Naturally, this leaves still room to investigate when or whether for a *fixed* affine mapping T the image $T(F)$ is closed. The latter reduces to the question whether $T(D)$ is closed, and in [13, Sect. 5] the authors relate this to infeasibility of suitable conic linear programs. For cones D their notion of asymptote coincides with Klee's f-asymptotes, which is at the basis of our Definition 12.2.

12.5 Invariance Properties of Motzkin *FW*-Sets

We have seen in Example 12.1 that finite intersections of FW-sets need no longer be FW-sets, not even when convexity is assumed. In contrast, the class of qFW-sets turned out closed under finite intersections. This raises the question whether more amenable sub-classes of the class of FW-sets with better invariance properties may

be identified. In response we show in this chapter that the class of Motzkin *FW*-sets, for short *FWM*-sets, is better behaved with regard to invariance properties.

Lemma 12.4 *Consider a set of the form $K + D$, where K is compact and D is a polyhedral closed convex cone in $\mathbb{R}^n$. Let L be a linear subspace of $\mathbb{R}^n$. Then there exists a compact set K_0 such that $(K + D) \cap L = K_0 + (D \cap L)$.*

Proof 1) We assume for the time being that the cone $D \cap L$ is pointed. For fixed $x \in K$ consider the polyhedron $P_x := (x + D) \cap L$. Define $M(P_x) = \{x' \in P_x : (x' - (D \cap L)) \cap P_x = \{x'\}\}$, and let $K(P_x)$ be the closed convex hull of $M(P_x)$. Then according to [9, Thm. 19] the set $K(P_x)$ is compact, and we have the minimal Motzkin decomposition $P_x = K(P_x) + (D \cap L)$. This uses the fact that $D \cap L$ is the recession cone of P_x. It follows that

$$(K + D) \cap L = \bigcup_{x \in K} (x + D) \cap L = \bigcup_{x \in K} K(P_x) + (D \cap L),$$

so all we have to do is show that the set $\bigcup_{x \in K} K(P_x)$ is bounded, as then its closure K_0 is the compact set announced in the statement of the Lemma. To prove boundedness of $\bigcup_{x \in K} K(P_x)$ it clearly suffices to show that $\bigcup_{x \in K} M(P_x)$ is bounded.

Let $\mathcal{F}$ be the finite set of faces of D, where we assume that D itself is a face. Let $x' \in M(P_x)$, then x' is in the relative interior of one of the faces $x + F$, $F \in \mathcal{F}$, of the shifted cone $x + D$.

We divide the faces $F \in \mathcal{F}$ of the cone D into two types: $\mathcal{F}_1$ is the class of those faces $F \in \mathcal{F}$ for which there exists $d \in L$, $d \neq 0$, such that d is a direction of recession of F, i.e., those where $F \cap L$ does not reduce to $\{0\}$. The class $\mathcal{F}_2$ gathers the remaining faces of D which are not in the class $\mathcal{F}_1$.

Now suppose the set $\bigcup_{x \in K} M(P_x)$ is unbounded. Then there exists a sequence $x_k \in K$ and $x'_k \in M(P_{x_k})$ with $\|x'_k\| \to \infty$. From the above we know that each x'_k is in the relative interior of $x_k + F_k$ for some $F_k \in \mathcal{F}$. Since there are only finitely many faces, we can extract a subsequence, also denoted x_k and satisfying $\|x'_k\| \to \infty$, such that the x'_k are relative interior points of $x_k + F$ for the same fixed face $F \in \mathcal{F}$. Due to compactness of K we may, in addition, assume that $x_k \to x \in K$. Using the definition of $M(P_{x_k})$ write $x'_k = x_k + t_k d_k \in L$ with $d_k \in F \subset D$, $\|d_k\| = 1$, $t_k > 0$, $t_k \to \infty$. Passing to yet another subsequence, assume that $d_k \to d$, where $\|d\| = 1$. It follows that $d \in L$, because in the expression $x'_k/t_k = x_k/t_k + d_k$ the middle term tends to 0 due to compactness of K and $t_k \to \infty$, while the left hand term is in L because x'_k belongs to L. Since F is a cone, it also follows that $x + \mathbb{R}_+ d \subset x + F$, hence $d \in F$. This shows that the face F is in the class $\mathcal{F}_1$.

2) So far we have shown that $\bigcup_{F \in \mathcal{F}_2} \{x' \in M(P_x) : x \in K, x' \in \mathrm{ri}(x + F)\}$ is a bounded set. It remains to prove that this set contains already all points $x' \in M(P_x)$, $x \in K$, i.e., that $x' \in M(P_x)$ cannot be a relative interior point of any of the faces $x + F$ with $F \in \mathcal{F}_1$.

3) Contrary to what is claimed, consider $x \in K \setminus L$ such that $x' \in M(P_x)$ satisfies $x' \in \mathrm{ri}(x+F)$ for some $F \in \mathcal{F}_1$. By definition of the class $\mathcal{F}_1$ there exists $d \in L \cap F$, $d \neq 0$. Since $x' \in L$ by the definition of $M(P_x)$, we have $x' + \mathbb{R}d \subset L$. But this line is also contained in $x + \mathrm{span}(F)$, because we have $d \in \mathrm{span}(F)$ and $x' = x + d'$ for some $d' \in F$, hence $x' + \mathbb{R}d \subset x + \mathrm{span}(F)$.

Since x' is a relative interior point of $x + F$, there exists $\epsilon > 0$ such that $N_\epsilon = \{x' + sd : |s| < \epsilon\}$ is contained in $x + F$. Since $d \in F \cap L \subset D \cap L$, we have arrived at a contradiction with the fact that $x' \in M(P_x)$. Namely, moving in N_ϵ we can stay in P_x while going from x' slightly in the direction of $-d \in -(D \cap L)$. This contradiction shows that what was claimed in 2) is true. The Lemma is therefore proved for pointed $D \cap L$.

4) Suppose now D is allowed to contain lines. With a change of coordinates we may arrange that $\mathbb{R}^n = \mathbb{R}^m \times \mathbb{R}^p$ and $D \subset \mathbb{R}^m \times \{0\}$, where the possibility $p = 0$ is not excluded and corresponds to the case where $D - D = \mathbb{R}^n$. Now consider the space $\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^p$ and define the cone $\widetilde{D} \subset \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^p$ as $\widetilde{D} = \{(x^+, x^-, 0) : x^\pm \in \mathbb{R}^m, x^\pm \geq 0, x^+ - x^- \in D\}$. Then $\widetilde{D}$ is polyhedral and pointed. Let T be the mapping $(x^+, x^-, y) \mapsto (x^+ - x^-, y)$, then $T(\widetilde{D}) = D$. Since T maps $\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^p$ onto $\mathbb{R}^m \times \mathbb{R}^p$, there exists a compact set $\widetilde{K} \subset \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^p$ such that $T(\widetilde{K}) = K$. Put $\widetilde{L} = T^{-1}(L)$. Now since $\widetilde{D}$ is pointed, the first part of the proof gives a compact $\widetilde{K}_0 \subset \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^p$ such that $(\widetilde{K} + \widetilde{D}) \cap \widetilde{L} = \widetilde{K}_0 + (\widetilde{D} \cap \widetilde{L})$. Applying T on both sides, and using the fact that $\widetilde{L}$ is a pre-image, we deduce $(K + D) \cap L = T(\widetilde{K}_0) + (D \cap L)$. On putting $K_0 = T(\widetilde{K}_0)$ which is compact, we get the desired statement $(K + D) \cap L = K_0 + (D \cap L)$. That completes the proof of the Lemma. □

Corollary 12.8 *Any finite intersection of sets of the form $K + D$ with K compact and D a polyhedral convex cone is again a set of this form.*

Proof It suffices to consider the case of two sets $F_i = K_i + D_i$ in $\mathbb{R}^n$, $i = 1, 2$, with compact K_i and D_i polyhedral convex cones. We build the set $F = F_1 \times F_2$ in $\mathbb{R}^n \times \mathbb{R}^n$, which is of the same form, because trivially $(K_1 + D_1) \times (K_2 + D_2) = (K_1 \times K_2) + (D_1 \times D_2)$, and since the product of two polyhedral cones is a polyhedral cone.

Now by Lemma 12.4 the intersection of $F_1 \times F_2$ with the diagonal $\Delta = \{(x, x) : x \in \mathbb{R}^n\}$ is a set of the form $\mathcal{K} + \mathcal{D}$ with $\mathcal{K}$ compact and $\mathcal{D}$ a polyhedral convex cone, because the diagonal is a linear subspace. Finally, $F_1 \cap F_2$ is the image of $\mathcal{K} + \mathcal{D}$ under the projection $p : (x, y) \to x$ onto the first coordinate, hence is of the form $p(\mathcal{K}) + p(\mathcal{D})$, and since $p(\mathcal{D})$ is a polyhedral convex cone, we are done. □

We conclude with the following invariance property of the class FWM:

Proposition 12.9 *If the pre-image of a FWM-set under an affine mapping is nonempty, then it is a FWM-set.*

Proof Let T be an affine mapping and F be a FWM-set such that $T^{-1}(F) \neq \emptyset$. Since translates of FWM-sets are FWM, there is no loss of generality in assuming

that T is linear. Then the restriction of T to $\ker(T)^{\perp}$ is a bijection from $\ker(T)^{\perp}$ onto $R(T)$, and one has

$$T^{-1}(F) = \left(T_{|\ker(T)^{\perp}}\right)^{-1}(F \cap R(T)) + \ker(T).$$

Since $R(T)$ is a subspace, hence a convex polyhedron, and $T^{-1}(F) \neq \emptyset$, the set $F \cap R(T)$ is FWM by Corollary 12.8. Since $\left(T_{|\ker(T)^{\perp}}\right)^{-1}$ is an isomorphism from $R(T)$ onto $\ker(T)^{\perp}$, the set $\left(T_{|\ker(T)^{\perp}}\right)^{-1}(F \cap R(T))$ is FWM. Hence it suffices to observe that $\ker(T)$, being a subspace, is FWM, and that the class of FWM-sets is closed under taking sums. □

Remark 12.17 It is worth mentioning that in general the affine pre-image of a Motzkin decomposable set need not be Motzkin decomposable. To wit, consider the ice cream cone F in $\mathbb{R}^3$ and the mapping $T : (x_1, x_2, x_3) \mapsto (1, x_2, x_3)$, then the linear function $x_3 - x_2$ does not attain its infimum on $T^{-1}(F)$, which proves that $T^{-1}(F)$ is not Motzkin decomposable.

Remark 12.18 In Proposition 12.5 we had proved that the affine pre-image $T^{-1}(F)$ of a FW-set is FW if F is contained in the range of T. A priori this additional range condition cannot be removed, because we have no result which guarantees that $F \cap \text{range}(T)$ is still a FW-set (if nonempty). As we just saw, this range condition *can* be removed for FWM-sets, and also for qFW-sets, so these two classes are invariant under affine pre-images without further range restriction.

Open Question Let F be a FW-set and L a linear subspace, is $F \cap L$ a FW-set?

Remark 12.19 Altogether we have found the class of FWM-sets to be closed under finite products, finite intersections, images and pre-images under affine maps. If we call a set $FWMU$ if it is a finite union of FWM-sets, then sets in this class are still FW-sets. By De Morgan's law the class $FWMU$ remains closed under finite intersections. The class $FWMU$ remains also closed under affine pre-images, because the pre-image of a union coincides with the union of the pre-images. Similarly the class $FWMU$ remains closed under affine images.

12.6 Parabolic Sets

As we have seen in Theorem 12.2, the search for new FW-sets does not lead very far beyond polyhedrality within the Motzkin class, because if a Motzkin set $F = K + D$ is to be FW, then its recession cone $D = 0^+F$ must already be polyhedral. The question is therefore whether one can find FW-sets which exhibit non-polyhedral asymptotic behavior, those then being necessarily outside the Motzkin class. The following result, with the terminology slightly adapted, shows that such FW-sets do indeed exist.

Theorem 12.3 (Luo and Zhang [14]) *Let P be a closed convex polyhedron and define $F = \{x \in P : x^\mathsf{T} Q x + q^\top x + c \leq 0\}$, where $Q = Q^\mathsf{T} \succeq 0$. Then F is a FW-set.*

The result generalizes the Frank and Wolfe theorem in the following sense: if we add just one convex quadratic constraint $x^\mathsf{T} Q x + q^\mathsf{T} x + c \leq 0$ to a linearly constrained quadratic program, then finite infima of quadratics are still attained. As Example 12.1 shows, adding a second convex quadratic constraint already fails.

The question is now: can the Luo-and-Zhang theorem, just like the Frank-and-Wolf theorem, be extended from polyhedra P to *FWM*-sets $F = K + D$? That means, if $F = K + D$ is a *FWM*-set, and if $Q = Q^\mathsf{T} \succeq 0$, will the set $\mathcal{F} = \{x \in F : x^\mathsf{T} Q x + q^\mathsf{T} x + c \leq 0\}$ still be a *FW*-set ? We show by way of a counterexample that the answer is in the negative.

Example 12.3 We consider the cylinder $F = \{(x_1, x_2, x_3, x_4) \in \mathbb{R}^4 : (x_1 - 1)^2 + x_2^2 \leq 1\}$. Note that F is a *FWM*-set, because it can be represented as $F = K + L$ for the compact convex set $K = \{(x_1, x_2, 0, 0) \in \mathbb{R}^4 : (x_1 - 1)^2 + x_2^2 \leq 1\}$ and the subspace $L = \{0\} \times \{0\} \times \mathbb{R} \times \mathbb{R}$.

Now we add the convex quadratic constraint $x_3^2 \leq x_4$ to the constraint set F, which leads to the set

$$\mathcal{F} = \{x \in F : x_3^2 \leq x_4\} = \{x \in \mathbb{R}^4 : (x_1 - 1)^2 + x_2^2 \leq 1,\ x_3^2 \leq x_4\}.$$

We will show that $\mathcal{F}$ is no longer a *FW*-set. This means that the extension of Theorem 12.3 from polyhedra P to *FWM*-sets F fails.

Consider the quadratic function $q(x) = x_4 x_1 - 2x_2 x_3 + 2$. We claim that q is bounded below on $\mathcal{F}$ by 0. Indeed, since $x_1 \geq 0$ on the feasible domain $\mathcal{F}$, we have $x_4 x_1 \geq x_3^2 x_1$ on the feasible domain, hence $q(x) \geq x_3^2 x_1 - 2x_2 x_3 + 2 = q(x_1, x_2, x_3, x_3^2)$, the expression on the right no longer depending on x_4. Let us compute the infimum of that expression on $\mathcal{F}$. This comes down to globally solving the program

$$(P) \qquad \begin{array}{ll} \text{minimize} & x_3^2 x_1 - 2x_2 x_3 + 2 \\ \text{subject to} & (x_1 - 1)^2 + x_2^2 \leq 1 \end{array}$$

and it is not hard to see that (P) has infimum 0, but that this infimum is not attained. (Solve for x_3 with fixed x_1, x_2 and show that the value at $(x_1, x_2, x_2/x_1)$ goes to 0 as $x_1 \to 0^+$, $(x_1 - 1)^2 + x_2^2 = 1$, but that 0 is not attained).

Now if $x^k \in \mathcal{F}$ is a minimizing sequence for q, then $\xi^k := (x_1^k, x_2^k, x_3^k, (x_3^k)^2) \in \mathcal{F}$ is also feasible and gives $q(x^k) \geq q(\xi^k)$, so the sequence ξ^k is also minimizing, showing that the infimum of q on $\mathcal{F}$ is the same as the infimum of (P), which is zero. But then the infimum of q on $\mathcal{F}$ could not be attained, as otherwise the infimum of (P) would also be attained. Indeed, if the infimum of q on $\mathcal{F}$ is attained at $\bar{x} \in \mathcal{F}$, then it must also be attained at $\bar{\xi} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_3^2) \in \mathcal{F}$ because $q(\bar{x}) \geq q(\bar{\xi})$, and then the infimum of (P) is attained at $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$, contrary to what was shown.

Remark 12.20 We can write the set $\mathcal{F}$ as $\mathcal{F} = K' \times F'$, where $K' = \{(x_1, x_2) : (x_1 - 1)^2 + x_2^2 \leq 1\}$ is compact convex, and where F' is the Luo-Zhang set $F' = \{(x_3, x_4) : x_3^2 \leq x_4\}$, which by Theorem 12.3 is a *FW*-set. This shows that the cross product of a convex *FW*-set (which is not *FWM*) and a compact convex set need no longer be a *FW*-set.

Remark 12.21 We can also write $\mathcal{F} = (K + L) \cap (F + M)$, where L, M are linear subspaces of $\mathbb{R}^4$. Indeed, K, L are as in Example 12.3, while $F = \{(0, 0, x_3, x_4) : x_3^2 \leq x_4\}$ and $M = \mathbb{R} \times \mathbb{R} \times \{0\} \times \{0\}$. Here $K + L$ is *FWM*, while $F + M$ is a *FW*-set by Theorem 12.3. Hence the intersection of a *FWM*-set and a *FW*-set (which is not *FWM*) need not be *FW*.

Remark 12.22 Note that $\mathcal{F}$ is a qFW-set by Proposition 12.6, see also [14, Cor. 2].

Acknowledgements Helpful discussions with B. Kummer (HU Berlin) and D. Klatte (Zürich) are gratefully acknowledged. We are indebted to Vera Roshchina (Australia) for having pointed out reference [16]. J.E. Martínez-Legaz was supported by the MINECO of Spain, Grant MTM2014-59179-C2-2-P, and by the Severo Ochoa Programme for Centres of Excellence in R&D [SEV-2015-0563]. He is affiliated with MOVE (Markets, Organizations and Votes in Economics). D. Noll was supported by Fondation Mathématiques Jacques-Hadamard (FMJH) under PGMO Grant *Robust Optimization for Control*.

References

1. Andronov, V., Belousov, E., Shironin, V.: On solvability of the problem of polynomial programming (in Russian). Izvestija Akadem. Nauk SSSR, Tekhnicheskaja Kibernetika **4**, 194–197 (1982). Translated as News of the Academy of Science of USSR, Dept. of Technical Sciences, Technical Cybernetics.
2. Bank, B., Guddat, J., Klatte, D., Kummer, B., Tammer, K.: Non-linear parametric optimization. Birkhäuser, Basel-Boston-Stuttgart (1983)
3. Belousov, E.: Introduction to Convex Analysis and Integer Programming (in Russian). Moscow University Publisher (1977)
4. Belousov, E., Klatte, D.: A Frank-Wolfe theorem for convex polynomial programs. Comput. Optim. Appl. **22**(1), 37–48 (2002)
5. Blum, E., Oettli, W.: Direct proof of the existence theorem in quadratic programming. Operations Research **20**, 165–167 (1972)
6. Collatz, L., Wetterling, W.: Optimization Problems. Springer Verlag (1975)
7. Eaves, B.: On quadratic programming. Management Sci. **17**(11), 698–711 (1971)
8. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Research Logistics Quarterly **3**, 95–110 (1956)
9. Goberna, M.A., González, E., Martínez-Legaz, J.E., Todorov, M.I.: Motzkin decomposition of closed convex sets. J Math. Anal. Appl. **364**, 209–221 (2010)
10. Iusem, A.N., Martínez-Legaz, J.E., Todorov, M.I.: Motzkin predecomposable sets. J. Global Optim. **60**(4), 635–647 (2014)
11. Klee, V.: Asymptotes and projections of convex sets. Math. Scand. **8**, 356–362 (1960)
12. Kummer., B.: Globale Stabilität quadratischer Optimierungsprobleme. Wissenschaftliche Zeitschrift der Humboldt-Universität zu Berlin, Math.-Nat. R. **XXVI**(5), 565–569 (1977)
13. Liu, M., Pataki, G.: Exact duals and short certificates of infeasibility and weak infeasibility in conic linear programming. Math. Prog. **167**, 435–480 (2018)

14. Luo, Z.Q., Zhang, S.: On extensions of the Frank-Wolfe theorems. Comput. Optim. Appl. **13**, 87–110 (1999)
15. Martínez-Legaz, J.E., Noll, D., Sosa, W.: Minimization of quadratic functions on convex sets without asymptotes. Journal of Convex Analysis **25**, 623–641 (2018)
16. Mirkil, H.: New characterizations of polyhedral cones. Can. J. Math. **9**, 1–4 (1957)
17. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1970)
18. Schrijver, A.: Theory of Linear and Integer Programming. John Wiley & Sons (1986)
19. Tam, N.N.: Continuity of the optimal value function in indefinite quadratic programming. J. Global Optim. **23**(1), 43–61 (2002)

Chapter 13
A Note on the Equivalence of Operator Splitting Methods

Walaa M. Moursi and Yuriy Zinchenko

Abstract This note provides a comprehensive discussion of the equivalences between some splitting methods. We survey known results concerning these equivalences which have been studied over the past few decades. In particular, we provide simplified proofs of the equivalence of the ADMM and the Douglas–Rachford method and the equivalence of the ADMM with intermediate update of multipliers and the Peaceman–Rachford method.

Keywords Alternating Direction Method of Multipliers (ADMM) · Chambolle–Pock method · Douglas–Rachford algorithm · Dykstra method · Equivalence of splitting methods · Fenchel–Rockafellar duality · Peaceman–Rachford algorithm

AMS 2010 Subject Classification 47H05, 47H09, 49M27, 49M29, 49N15, 90C25

13.1 Introduction

Splitting methods have become popular in solving convex optimization problems that involve finding a minimizer of the sum of two proper lower semicontinuous convex functions. Among these methods are the Douglas–Rachford and

W. M. Moursi (✉)
Electrical Engineering, Stanford University, Stanford, CA, USA

Faculty of Science, Mathematics Department, Mansoura University, Mansoura, Egypt
e-mail: wmoursi@stanford.edu

Y. Zinchenko
University of Calgary, Department of Mathematics and Statistics, Calgary, AB, Canada
e-mail: yzinchenko@ucalgary.ca

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_13

the Peaceman–Rachford methods introduced in the seminal work of Lions and Mercier [24], the forward-backward method (see, e.g., [12, 17] and [29]), Dykstra's method (see, e.g., [3] and [10]), and the Method of Alternating Projections (MAP) (see, e.g., [19]).

When the optimization problem features the composition of one of the functions with a bounded linear operator, a popular technique is the Alternating-Direction Method of Multipliers (ADMM) (see [22, Section 4], [16, Section 10.6.4] and also [7, Chapter 15]). The method has a wide range of applications including large-scale optimization, machine learning, image processing and portfolio optimization, see, e.g., [9, 15] and [20]. A powerful framework to use ADMM in the more general setting of *monotone operators* is developed in the work of Briceño-Arias and Combettes [13] (see also [8] and [14]). Another relatively recent method is the Chambolle–Pock method introduced in [11].

Equivalences between splitting methods have been studied over the past four decades. For instance, it is known that ADMM is equivalent to the Douglas–Rachford method [24] (see, also [21]) in the sense that with a careful choice of the starting point, one can prove that the sequences generated by both algorithms coincide. (See, e.g., [22, Section 5.1] and [6, Remark 3.14].) A similar equivalence holds between ADMM (with intermediate update of multiplier) and Peaceman–Rachford method [24] (see [22, Section 5.2]). In [25], the authors proved the correspondence of Douglas–Rachford and Chambolle–Pock methods.

The rest of this paper is organized as follows: Section 13.2 provides a brief literature review of ADMM, Douglas–Rachford and Peaceman–Rachford methods. In Sections 13.3 and 13.4 we explicitly describe the equivalence of ADMM (respectively ADMM with intermediate update of multipliers) and Douglas–Rachford (respectively Peaceman–Rachford) method introduced by Gabay in [22, Sections 5.1&5.2]. We provide simplified proofs of these equivalences. Section 13.5 focuses on the recent work of O'Connor and Vandenberghe concerning the equivalence of Douglas–Rachford and Chambolle–Pock methods (see [25]). Our notation is standard and follows largely, e.g., [5].

13.2 Three Techniques

In this paper, we assume that

$$X \text{ and } Y \text{ are real Hilbert spaces,}$$

and that

$f: X \to]-\infty, +\infty]$, $g: Y \to]-\infty, +\infty]$ are convex lower semicontinuous and proper.

Alternating-Direction Method of Multipliers (ADMM) In the following we assume that[1]

$L: Y \to X$ is linear such that L^*L is invertible,

that

$$\operatorname{argmin}(f \circ L + g) \neq \varnothing, \tag{13.1}$$

and that

$$0 \in \operatorname{sri}(\operatorname{dom} f - L(\operatorname{dom} g)), \tag{13.2}$$

where sri S denotes the *strong relative interior* of a subset S of X with respect to the closed affine hull of S. When X is finite-dimensional we have $\operatorname{sri} S = \operatorname{ri} S$, where ri S is the *relative interior* of S defined as the interior of S with respect to the affine hull of S.

Consider the problem

$$\underset{y \in Y}{\text{minimize}} \;\; f(Ly) + g(y). \tag{13.3}$$

Note that (13.1) and (13.2) imply that (see, e.g., [5, Proposition 27.5(iii)(a)1])

$$\operatorname{argmin}(f \circ L + g) = \operatorname{zer}(\partial(f \circ L) + \partial g) = \operatorname{zer}(L^* \circ (\partial f) \circ L + \partial g) \neq \varnothing. \tag{13.4}$$

In view of (13.4), solving (13.3) is equivalent to solving the inclusion:

$$\text{Find } y \in Y \text{ such that } 0 \in L^*(\partial f(Ly)) + \partial g(y). \tag{13.5}$$

The augmented Lagrangian associated with (13.3) is the function

[1]The adjoint of L is the unique operator $L^*: X \to Y$ that satisfies $\langle Ly, x\rangle = \langle y, L^*x\rangle$ $(\forall (x, y) \in X \times Y)$.

$$\mathcal{L}\colon X\times Y\times X \to \left]-\infty,+\infty\right] : (a,b,u)\mapsto f(a)+g(b)+\langle u, Lb-a\rangle+\frac{1}{2}\|Lb-a\|^2. \tag{13.6}$$

The ADMM (see [22, Section 4] and also [16, Section 10.6.4]) applied to solve (13.3) consists in minimizing $\mathcal{L}$ over b then over a and then applying a proximal minimization step with respect to the Lagrange multiplier u. The method applied with a starting point $(a_0, u_0) \in X \times X$ generates three sequences $(a_n)_{n\in\mathbb{N}}$; $(b_n)_{n\geq 1}$ and $(u_n)_{n\in\mathbb{N}}$ via $(\forall n \in \mathbb{N})$:

$$b_{n+1} \coloneqq (L^*L + \partial g)^{-1}(L^*a_n - L^*u_n), \tag{13.7a}$$

$$a_{n+1} \coloneqq \operatorname{Prox}_f(Lb_{n+1} + u_n), \tag{13.7b}$$

$$u_{n+1} \coloneqq u_n + Lb_{n+1} - a_{n+1}, \tag{13.7c}$$

where $\operatorname{Prox}_f\colon X \to X\colon x \mapsto \operatorname{argmin}_{y\in X}\left(f(y) + \frac{1}{2}\|x-y\|^2\right)$.

Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in X and let $\overline{x} \in X$. In the following we shall use $x_n \to \overline{x}$ (respectively $x_n \rightharpoonup \overline{x}$) to indicate that $(x_n)_{n\in\mathbb{N}}$ converges strongly (respectively weakly) to $\overline{x}$.

Fact 13.1 (Convergence of ADMM (See [22, Theorem 4.1])) *Let $(a_0, u_0) \in X \times X$, and let $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\geq 1}$ and $(u_n)_{n\in\mathbb{N}}$ be defined as in (13.7). Then, there exists $\overline{b} \in Y$ such that $b_n \rightharpoonup \overline{b} \in \operatorname{argmin}(f \circ L + g)$.*

The Douglas–Rachford Method Suppose that $Y = X$ and that $L = \mathrm{Id}$. In this case Problem (13.3) becomes

$$\underset{x\in X}{\text{minimize}}\ f(x) + g(x). \tag{13.8}$$

The Douglas–Rachford (DR) method, introduced in [24], applied to the ordered pair (f, g) with a starting point $x_0 \in X$ to solve (13.8) generates two sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ via:

$$y_n \coloneqq \operatorname{Prox}_f x_n, \tag{13.9a}$$

$$x_{n+1} \coloneqq T_{\mathrm{DR}} x_n, \tag{13.9b}$$

where

$$T_{\mathrm{DR}} \coloneqq T_{\mathrm{DR}}(f,g) = \tfrac{1}{2}(\mathrm{Id} + R_g R_f) = \mathrm{Id} - \operatorname{Prox}_f + \operatorname{Prox}_g(2\operatorname{Prox}_f - \mathrm{Id}), \tag{13.10}$$

and where $R_f \coloneqq 2\operatorname{Prox}_f - \mathrm{Id}$.

To lighten the notation, in the sequel we shall use T_{DR} to denote $T_{\mathrm{DR}}(f, g)$. Let $T\colon X \to X$. Recall that the set of fixed points of T, denoted by $\operatorname{Fix} T$, is defined as $\operatorname{Fix} T \coloneqq \{x \in X \mid x = Tx\}$.

Fact 13.2 (Convergence of Douglas–Rachford Method (See, e.g., [24, Theorem 1] or [5, Corollary 28.3])) *Let $x_0 \in X$ and let $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ be defined as in (13.9). Then, there exists $\overline{x} \in \operatorname{Fix} T_{DR}$ such that $x_n \rightharpoonup \overline{x}$ and $y_n \rightharpoonup \operatorname{Prox}_f \overline{x} \in \operatorname{argmin}(f+g)$.*

The Peaceman–Rachford Method Let $h\colon X \to]-\infty,+\infty]$ be proper and let $\beta > 0$. We say that h is *strongly convex* if $f - \frac{\beta}{2}\|\cdot\|^2$ is convex, i.e., $(\forall (x, y) \in \operatorname{dom} f \times \operatorname{dom} f)$ $(\forall \alpha \in]0, 1[)$ we have $f(\alpha x + (1-\alpha)y) + \alpha(1-\alpha)\phi(\|x-y\|) + \frac{\beta}{2}\|x-y\|^2 \leq \alpha f(x) + (1-\alpha) f(y)$.

When g is strongly convex, the Peaceman–Rachford (PR) method, introduced in [24], can be used to solve (13.8). In this case, given $x_0 \in X$, PR method generates the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ via:

$$y_n := \operatorname{Prox}_f x_n, \tag{13.11a}$$

$$x_{n+1} := T_{\mathrm{PR}} x_n, \tag{13.11b}$$

where

$$T_{\mathrm{PR}} = T_{\mathrm{PR}}(f, g) = R_g R_f = (2\operatorname{Prox}_g - \mathrm{Id})(2\operatorname{Prox}_f - \mathrm{Id}). \tag{13.12}$$

To lighten the notation, in the sequel we shall use T_{PR} to denote $T_{\mathrm{PR}}(f, g)$.

Fact 13.3 (Convergence of Peaceman–Rachford Method (See, e.g., [24, Proposition 1] or [5, Proposition 28.8])) *Suppose that g is strongly convex. Let $\overline{y}$ be the unique minimizer of $f+g$, let $x_0 \in X$ and let $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ be defined as in (13.11). Then $y_n \to \overline{y}$.*

In the sequel we use the notation

$$g^{\vee}\colon X \to]-\infty,+\infty] : x \mapsto g(-x). \tag{13.13}$$

Recall that the Fenchel–Rockafellar dual of (13.3) is

$$\operatorname*{minimize}_{x\in X} \; f^*(x) + g^*(-L^*x). \tag{13.14}$$

Remark 13.1

(i) One can readily verify that $\partial g^{\vee} = (-\mathrm{Id}) \circ \partial g \circ (-\mathrm{Id})$. Therefore, in view of [27, Theorem A] and [20, Lemma 3.5 on page 125 and Lemma 3.6 on page 133] (see also [4, Corollaries 4.2 and 4.3]) we have[2]

$$T_{\mathrm{DR}}(f, g) = T_{\mathrm{DR}}(f^*, g^{*\vee}), \tag{13.15}$$

[2]It is straightforward to verify that $g^{\vee *} = (g^*)^{\vee}$ (see, e.g., [5, Proposition 13.23(v)]).

and

$$T_{\mathrm{PR}}(f, g) = T_{\mathrm{PR}}(f^*, g^{*\vee}). \tag{13.16}$$

(ii) When $(L, Y) = (\mathrm{Id}, X)$, inclusion (13.5) reduces to: Find $y \in X$ such that $0 \in \partial f(y) + \partial g(y)$ and the *dual* inclusion (corresponding to the Fenchel–Rockafellar dual (13.14)) is: Find $y \in X$ such that $0 \in \partial f^*(y) - \partial g^*(-y) = (\partial f)^{-1}y - (\partial g)^{-1}(-y)$, which in this case coincide with the Attouch–Thera dual of (13.5) (see [2]).

One can use DR method to solve (13.14) where (f, g) in Fact 13.2 is replaced by $(f^*, g^* \circ (-L^*))$. Recalling (13.15) we learn that $T_{\mathrm{DR}} = T_{\mathrm{DR}}(f^*, g^* \circ (-L^*)) = T_{\mathrm{DR}}(f^{**}, (g^* \circ (-L^*))^{\vee *}) = T_{\mathrm{DR}}(f, (g^* \circ L^*)^*)$, where the last identity follows from [5, Proposition 13.44].

In view of (13.10) (13.15), and [5, Proposition 15.23(v)] we have

$$\begin{aligned} T_{\mathrm{DR}} &= T_{\mathrm{DR}}(f^*, g^* \circ (-L^*)) = T_{\mathrm{DR}}(f, (g^* \circ L^*)^*) \\ &= \mathrm{Id} - \mathrm{Prox}_f + \mathrm{Prox}_{(g^* \circ L^*)^*}(2\,\mathrm{Prox}_f - \mathrm{Id}). \end{aligned} \tag{13.17}$$

Similarly, under additional assumptions (see Fact 13.3), one can use PR method to solve (13.14) where (f, g) in Theorem 13.3 is replaced by $(f^*, g^* \circ (-L^*))$. In this case (13.12), (13.16) and [5, Proposition 15.23(v)] imply that

$$\begin{aligned} T_{\mathrm{PR}} &= T_{\mathrm{PR}}(f^*, g^* \circ (-L^*)) = T_{\mathrm{PR}}(f, (g^* \circ L^*)^*) \\ &= (2\,\mathrm{Prox}_{(g^* \circ L^*)^*} - \mathrm{Id})(2\,\mathrm{Prox}_f - \mathrm{Id}). \end{aligned} \tag{13.18}$$

For completeness, we provide a concrete proof of the formula for $\mathrm{Prox}_{(g^* \circ L^*)^*}$ in Appendix 1 (see Proposition 13.3(viii) below). We point out that the formula for $\mathrm{Prox}_{(g^* \circ L^*)^*}$ in a more general setting is given in [22, Proposition 4.1] (see also [16, Section 10.6.4]).

13.3 ADMM and Douglas–Rachford Method

In this section we discuss the equivalence of ADMM and DR method. This equivalence was first established by Gabay in [22, Section 5.1] (see also [6, Remark 3.14]). Let $(x_0, a_0, u_0) \in X^3$. Throughout the rest of this section, we assume that the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are as defined in (13.9), i.e.,

$$(x_{n+1}, y_n)_{n\in\mathbb{N}} = (T_{\mathrm{DR}} x_n, \mathrm{Prox}_f\, x_n), \tag{13.19}$$

$$T_{\mathrm{DR}} \coloneqq T_{\mathrm{DR}}(f,(g^* \circ L^*)^*) = \mathrm{Id} - \mathrm{Prox}_f + L(L^*L + \partial g)^{-1}L^*(2\,\mathrm{Prox}_f - \mathrm{Id}). \tag{13.20}$$

Note that the second identity in (13.20) follows from (13.17) and Proposition 13.3(viii). We also assume that

$(a_n, u_n, b_{n+1})_{n\in\mathbb{N}}$ is defined as in (13.7).

The following lemma will be used later to clarify the equivalence of DR and ADMM.

Lemma 13.1 *Let* $(b_-, a_-, u_-) \in Y \times X \times X$ *and set*

$$(b, a, u) \coloneqq ((L^*L + \partial g)^{-1}(L^*a_- - L^*u_-), \mathrm{Prox}_f(Lb + u_-), u_- + Lb - a), \tag{13.21a}$$

$$(b_+, a_+, u_+) \coloneqq ((L^*L + \partial g)^{-1}(L^*a - L^*u), \mathrm{Prox}_f(Lb_+ + u), u + Lb_+ - a_+). \tag{13.21b}$$

Then

$$T_{DR}(Lb + u_-) = Lb_+ + u, \tag{13.22a}$$

$$\mathrm{Prox}_f\, T_{DR}(Lb + u_-) = a_+. \tag{13.22b}$$

Proof Indeed, it follows from (13.20), (13.21), (13.22a) and (13.22b) that

$$\begin{aligned} T_{\mathrm{DR}}(Lb + u_-) &= (Lb + u_-) - \mathrm{Prox}_f(Lb + u_-) \\ &\quad + L(L^*L + \partial g)^{-1}L^*(2\,\mathrm{Prox}_f(Lb + u_-) - (Lb + u_-)) && \text{(13.23a)} \\ &= (Lb + u_-) - a + L(L^*L + \partial g)^{-1}L^*(2a - (Lb + u_-)) && \text{(13.23b)} \\ &= (Lb + u_-) - a + L(L^*L + \partial g)^{-1}L^*(a - (Lb + u_- - a)) && \text{(13.23c)} \\ &= u + L(L^*L + \partial g)^{-1}(L^*a - L^*u) && \text{(13.23d)} \\ &= Lb_+ + u, && \text{(13.23e)} \end{aligned}$$

where (13.23a) follows from (13.20), (13.23b) and (13.23d) follow from (13.21a), and (13.23e) follow from (13.21b). This proves (13.22a). Now (13.22b) follows from combining (13.22a) and (13.21b). □

We now prove the main result in this section by induction.

Theorem 13.4 *The following hold:*

(i) **(DR as ADMM Iteration)** *Using DR method with a starting point $x_0 \in X$ to solve (13.14) is equivalent to using ADMM with a starting point $(a_0, u_0) := (\operatorname{Prox}_f x_0, x_0 - \operatorname{Prox}_f x_0)$ to solve (13.3), in the sense that $(x_n)_{n\geq 1} = (Lb_n + u_{n-1})_{n\geq 1}$ and $(y_n)_{n\in\mathbb{N}} = (a_n)_{n\in\mathbb{N}}$.*

(ii) **(ADMM as DR Iteration)** *Using ADMM with a starting point $(a_0, u_0) \in X \times X$ to solve (13.3) is equivalent to using DR method with a starting point $x_0 = Lb_1 + u_0$ to solve (13.14), in the sense that $(x_n)_{n\in\mathbb{N}} = (Lb_{n+1} + u_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}} = (a_{n+1})_{n\in\mathbb{N}}$.*

Proof For simplicity, set $T = T_{\mathrm{DR}}$. (i): Note that (13.19) implies that $y_0 = a_0$. Now, when $n = 1$ we have

$$\begin{aligned} x_1 = Tx_0 &= x_0 - \operatorname{Prox}_f x_0 + L(L^*L + \partial g)^{-1}L^*(2\operatorname{Prox}_f x_0 - x_0) && \text{(by (13.17))} \\ &= x_0 - a_0 + L(L^*L + \partial g)^{-1}L^*(2a_0 - x_0) && \text{(by (13.19))} \\ &= (x_0 - a_0) + L(L^*L + \partial g)^{-1}L^*(a_0 - (x_0 - a_0)) \\ &= u_0 + L(L^*L + \partial g)^{-1}L^*(a_0 - u_0) \\ &= u_0 + Lb_1. && \text{(by (13.7a))} \end{aligned}$$

Combining with (13.19) and (13.7b) we get $y_1 = \operatorname{Prox}_f Tx_0 = \operatorname{Prox}_f x_1 = \operatorname{Prox}_f(u_0 + Lb_1) = a_1$, which verifies the base case. Now suppose for some $n \geq 1$ we have $x_n = Lb_n + u_{n-1}$ and $y_n = a_n$ and use Lemma 13.1 with (b_-, a_-, u_-) replaced by $(b_{n-1}, a_{n-1}, u_{n-1})$ to learn that $x_{n+1} = Lb_{n+1} + u_n$ and $y_{n+1} = a_{n+1}$. Consequently, $(x_n)_{n\geq 1} = (Lb_n + u_{n-1})_{n\geq 1}$ and $(y_n)_{n\in\mathbb{N}} = (a_n)_{n\in\mathbb{N}}$ as claimed.

(ii): At $n = 0$, $x_0 = Lb_1 + u_0 = Lb_{0+1} + u_0$, and therefore (13.9a) implies that $y_0 = \operatorname{Prox}_f x_0 = \operatorname{Prox}_f(Lb_1 + u_0) = a_1$ by (13.7b). Now suppose that for some $n \geq 0$ we have $x_n = Lb_{n+1} + u_n$ and $y_n = a_{n+1}$. The conclusion follows by applying Lemma 13.1 with (b_-, a_-, u_-) replaced by (b_n, a_n, u_n). □

13.4 ADMM and Peaceman–Rachford Method

We now turn to the equivalence of ADMM with intermediate update of multiplier and PR method. This equivalence was established in [22, Section 5.2]. Given $(a_0, u_0) \in X \times X$, the ADMM with an *intermediate* update of multiplier applied to solve (13.3) generates four sequences $(a_n)_{n\in\mathbb{N}}$, $(u_n)_{n\in\mathbb{N}}$, $(b_n)_{n\geq 1}$ and $(w_n)_{n\geq 1}$ via $(\forall n \in \mathbb{N})$:

$$b_{n+1} := (L^*L + \partial g)^{-1}(L^*a_n - L^*u_n), \tag{13.24a}$$

$$w_{n+1} := u_n + Lb_{n+1} - a_n, \tag{13.24b}$$

$$a_{n+1} := \operatorname{Prox}_f(Lb_{n+1} + w_{n+1}), \tag{13.24c}$$

$$u_{n+1} := w_{n+1} + Lb_{n+1} - a_{n+1}. \tag{13.24d}$$

Fact 13.5 (Convergence of ADMM with Intermediate Update of Multipliers (See [22, Theorem 5.4])) *Suppose that g is strongly convex. Let $(a_0, u_0) \in X \times X$, and let $(b_n)_{n\geq 1}$, $(w_n)_{n\geq 1}$, $(a_n)_{n\in\mathbb{N}}$ and $(u_n)_{n\in\mathbb{N}}$ be defined as in (13.24). Then, there exists $\overline{b} \in Y$ such that $b_n \to \overline{b} \in \operatorname{argmin}(f \circ L + g)$.*

In this section we work under the additional assumption that

$$g \text{ is strongly convex.}$$

Let $(x_0, a_0, u_0) \in X^3$. Throughout the rest of this section we assume that the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ are as defined in (13.11), i.e.,

$$(x_{n+1}, y_n)_{n\in\mathbb{N}} = (T_{\mathrm{PR}} x_n, \operatorname{Prox}_f x_n)_{n\in\mathbb{N}} \tag{13.25}$$

where

$$T_{\mathrm{PR}} := T_{\mathrm{PR}}(f, (g^* \circ L^*)^*) = 2L(L^*L + \partial g)^{-1} L^*(2\operatorname{Prox}_f - \mathrm{Id}) - 2\operatorname{Prox}_f + \mathrm{Id}\,. \tag{13.26}$$

Note that the second identity in (13.26) follows from (13.18) and Proposition 13.3(viii). We also assume that

$$(a_n, u_n, b_{n+1}, w_{n+1})_{n\in\mathbb{N}} \text{ is defined as in (13.24).}$$

Before we proceed further, we prove the following useful lemma.

Lemma 13.2 *Let $(b_-, w_-, a_-, u_-) \in Y \times X \times X \times X$ and set*

$$(b, w, a, u) = ((L^*L + \partial g)^{-1}(L^*a_- - L^*u_-), u_- + Lb - a_-, \operatorname{Prox}_f(Lb + w), w + Lb - a), \tag{13.27a}$$

$$(b_+, w_+, a_+, u_+) = ((L^*L + \partial g)^{-1}(L^*a - L^*u), u + Lb_+ - a, \operatorname{Prox}_f(Lb_+ + w_+), w_+ + Lb_+ - a_+). \tag{13.27b}$$

Then

$$T_{PR}(Lb + w) = Lb_+ + w_+, \tag{13.28a}$$

$$\operatorname{Prox}_f T_{PR}(Lb + w) = a_+. \tag{13.28b}$$

Proof Indeed, by (13.26), (13.27a) and (13.27b) we have

$$\begin{aligned}
T_{\mathrm{PR}}(Lb + w) &= Lb + w - 2\operatorname{Prox}_f(Lb + w) \\
&\quad + 2L(L^*L + \partial g)^{-1}L^*(2\operatorname{Prox}_f(Lb + w) - (Lb + w)) \\
&\qquad\qquad \text{(by (13.26))} \\
&= Lb + w - a - a + 2L(L^*L + \partial g)^{-1}L^*(a - (Lb + w - a)) \\
&\qquad\qquad \text{(by (13.27a))} \\
&= u - a + 2L(L^*L + \partial g)^{-1}L^*(a - u) = u - a + 2Lb_+ \\
&\qquad\qquad \text{(by (13.27b))} \\
&= Lb_+ + w_+, \qquad \text{(by (13.27b))}
\end{aligned}$$

which proves (13.28a). Now (13.28b) is a direct consequence of (13.28a) in view of (13.27b). □

We are now ready for the main result in this section.

Theorem 13.6 *Suppose that g is strongly convex. Then the following hold:*

(i) **(PR as ADMM Iteration)** *Using PR method with a starting point $x_0 \in X$ to solve (13.14) is equivalent to using ADMM with intermediate update of multipliers with starting points $(a_0, u_0) := (\operatorname{Prox}_f x_0, x_0 - \operatorname{Prox}_f x_0)$ to solve (13.3), in the sense that $(x_n)_{n\geq 1} = (Lb_n + w_n)_{n\geq 1}$ and $(y_n)_{n\in\mathbb{N}} = (a_n)_{n\in\mathbb{N}}$.*

(ii) **(ADMM as PR Iteration)** *Using ADMM with intermediate update of multipliers with a starting point $(a_0, u_0) \in X \times X$ to solve (13.3) is equivalent to using PR method with starting point $x_0 = Lb_1 + w_1$ to solve (13.14), in the sense that $(x_n)_{n\in\mathbb{N}} = (Lb_{n+1} + w_{n+1})_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}} = (a_{n+1})_{n\in\mathbb{N}}$.*

Proof We proceed by induction. (i): We have

$$\begin{aligned}
x_1 = T_{\mathrm{PR}}x_0 &= x_0 - 2\operatorname{Prox}_f x_0 + L(L^*L + \partial g)^{-1}L^*(2\operatorname{Prox}_f x_0 - x_0) \\
&\qquad\qquad \text{(by (13.26))} \\
&= x_0 - 2a_0 + 2L(L^*L + \partial g)^{-1}L^*(2a_0 - x_0) \\
&= (x_0 - a_0) - a_0 + 2L(L^*L + \partial g)^{-1}L^*(a_0 - (x_0 - a_0)) \\
&= u_0 - a_0 + 2Lb_1 = u_0 - a_0 + Lb_1 + Lb_1 \\
&= Lb_1 + w_1, \qquad \text{(by (13.24b))}
\end{aligned}$$

which verifies the base case. Now suppose for some $n \geq 1$ we have $x_n = Lb_n + w_n$. The conclusion follows from applying Lemma 13.2 with (b_-, w_-, a_-, u_-) replaced by $(b_{n-1}, w_{n-1}, a_{n-1}, u_{n-1})$ in view of (13.24).

(ii): At $n = 0$, the base case clearly holds in view of (13.25) and (13.24c). Now suppose that for some $n \geq 0$ we have $x_n = Lb_{n+1} + w_{n+1}$ and $y_n = a_{n+1}$ and use Lemma 13.2 with (b_-, w_-, a_-, u_-) replaced by (b_n, w_n, a_n, u_n) in view of (13.24). □

13.5 Chambolle–Pock and Douglas–Rachford Methods

In this section we survey the recent work by O'Connor and Vandenberghe [25] concerning the equivalence of Douglas–Rachford method and Chambolle–Pock method. (For a detailed study of this correspondence in the more general framework of the primal-dual hybrid gradient method and DR method with relaxation, as well as connection to linearized ADMM, we refer the reader to [25].) We work under the assumption that[3]

$$A\colon X \to Y \text{ is linear, } \sigma > 0, \tau > 0, \text{ and } \sigma\tau\|A\|^2 < 1. \tag{13.31}$$

Consider the problem

$$\underset{x\in X}{\text{minimize}} \;\; f(x) + g(Ax) \tag{13.32}$$

and its Fenchel–Rockafellar dual given by

$$\underset{x\in X}{\text{minimize}} \;\; f^*(-Ax) + g^*(x). \tag{13.33}$$

To proceed further, in the following we assume that

$$\operatorname{argmin}(f + g \circ A) \neq \varnothing \text{ and } \; 0 \in \operatorname{sri}(\operatorname{dom} g - A(\operatorname{dom} f)). \tag{13.34}$$

Note that (13.34) implies that (see, e.g., [5, Proposition 27.5(iii)(a)1])

$$\operatorname{argmin}(f+g\circ A) = \operatorname{zer}(\partial f+\partial(g\circ A)) = \operatorname{zer}(\partial f+A^*\circ(\partial g)\circ A) \neq \varnothing. \tag{13.35}$$

[3] In passing, we point out that, when X is a finite-dimensional Hilbert space, the condition $\tau\sigma\|A\|^2 < 1$ can be relaxed to $\tau\sigma\|A\|^2 \leq 1$. The convergence in this case is proved in [18, Theorem 3.3].

In view of (13.35), solving (13.32) is equivalent to solving the inclusion:

$$\text{Find } x \in X \text{ such that } 0 \in \partial f(x) + A^*(\partial g(Ax)). \tag{13.36}$$

The Chambolle–Pock (CP) method applied with a staring point $(u_0, v_0) \in X \times Y$ to solve (13.32) generates the sequences $(u_n)_{n\in\mathbb{N}}$, and $(v_n)_{n\in\mathbb{N}}$ via:

$$u_n = \operatorname{Prox}_{\tau f}(u_{n-1} - \tau A^* v_{n-1}), \tag{13.37a}$$

$$v_n = \operatorname{Prox}_{\sigma g^*}(v_{n-1} + \sigma A(2u_n - u_{n-1})), \tag{13.37b}$$

where τ and σ are as defined in (13.31).

Fact 13.7 (Convergence of Chambolle–Pock Method (See [11, Theorem 1], [30, Theorem 3.1] and Also [18, Theorem 3.1])) *Let $(u_0, v_0) \in X \times Y$ and let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be defined as in (13.37). Then, there exists $(\overline{u}, \overline{v}) \in X \times Y$ such that $(u_n, v_n)_{n\in\mathbb{N}} \rightharpoonup (\overline{u}, \overline{v})$, $\overline{u} \in \operatorname{argmin}(f + g \circ A)$ and $\overline{v} \in \operatorname{argmin}(f^* \circ (-A^*) + g^*)$.*

It is known that the method in (13.37) reduces to DR method (see, e.g., [11, Section 4.2]) when $A = \mathrm{Id}$. We state this equivalence in Proposition 13.1 below.

Proposition 13.1 (DR as a CP Iteration) *Suppose that $X = Y$, and that $A = \mathrm{Id}$. Then, using DR method, defined as in (13.9), with a starting point $x_0 \in X$ to solve (13.32) is equivalent to using CP method with a starting point $(u_0, v_0) \in \{(u, v) \mid u - v = x_0\} \subseteq X \times X$ to solve (13.32) in the sense that $(x_n)_{n\in\mathbb{N}} = (u_n - v_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}} = (u_n)_{n\in\mathbb{N}}$.*

Proof We use induction. When $n = 0$, the base case is obviously true. Now suppose that for some $n \geq 0$ we have $x_n = u_n - v_n$ and $y_n = u_n$. Then,

$$x_{n+1} = \operatorname{Prox}_f x_n - \operatorname{Prox}_{g^*}(2\operatorname{Prox}_f x_n - x_n) \tag{13.38a}$$

$$= \operatorname{Prox}_f(u_n - v_n) - \operatorname{Prox}_{g^*}(2\operatorname{Prox}_f(u_n - v_n) - (u_n - v_n)) \tag{13.38b}$$

$$= u_{n+1} - \operatorname{Prox}_{g^*}(v_n + 2u_{n+1} - u_n) = u_{n+1} - v_{n+1}. \tag{13.38c}$$

Here (13.38a) follows from Lemma 13.5 below (applied with $\gamma = 1$), (13.38b) follows from the inductive hypothesis, and (13.38c) follows from (13.37) applied with (τ, σ, A) replaced by $(1, 1, \mathrm{Id})$. The claim about y_{n+1} follows directly and the proof is complete. □

Chambolle–Pock as a DR Iteration: The O'Connor–Vandenberghe Technique Let Z be a real Hilbert space. In the following, we assume that $C\colon Z \to Y$ is linear and that

$$B\colon X \times Z \to Y\colon (x, z) \mapsto Ax + Cz \text{ satisfies that } \sigma\tau BB^* = \mathrm{Id}. \tag{13.39}$$

Note that one possible choice of C is to set $C^2 \coloneqq \mathrm{Id} - \sigma\tau AA^*$, where the existence of C follows from, e.g., [26, Theorem on page 265]. Now consider the problem

$$\underset{(x,z)\in X\times Z}{\text{minimize}}\ \widetilde{f}(x,z) + g(B(x,z)), \tag{13.40}$$

where

$$\widetilde{f}\colon X \times Z \to \left]-\infty,+\infty\right] : (x,z) \mapsto f(x) + \iota_{\{0\}}(z). \tag{13.41}$$

The following result, proved in [25, Section 4] in the more general framework of primal-dual hybrid gradient method, provides an elegant way to construct the correspondence between the DR sequence when applied to solve (13.40) and the CP sequence when applied to solve (13.32). We restate the proof for the sake of completeness.

Proposition 13.2 (CP Corresponds to a DR Iteration) *Using CP method with starting point $(u_0, v_0) \in X \times Z$ to solve (13.32) corresponds to using DR with starting point $\mathbf{x}_0 \coloneqq (u_0, 0) - \tau B^* v_0 \in X \times Z$ to solve (13.40), in the sense that $(\mathbf{x}_n)_{n\in\mathbb{N}} = ((u_n, 0) - \tau B^* v_n)_{n\in\mathbb{N}}$ and $(\mathbf{y}_n)_{n\in\mathbb{N}} = (u_{n+1}, 0)_{n\in\mathbb{N}}$.*

Proof We apply DR to solve (13.40) with $(\widetilde{f}, g)$ replaced by $(\tau\widetilde{f}, \tau g)$. The proof proceeds by induction. When $n = 0$, by assumption we have $\mathbf{x}_0 = (u_0, 0) - \tau B^* v_0$. It follows from Proposition 13.4(i)&(vii) below applied with $\widetilde{f}$ replaced by $\tau\widetilde{f}$ that $\mathbf{y}_0 = \mathrm{Prox}_{\tau\widetilde{f}}\,\mathbf{x}_0 = \mathrm{Prox}_{\tau\widetilde{f}}((u_0, 0) - \tau(A^* v_0, C^* v_0)) = \mathrm{Prox}_{\tau\widetilde{f}}(u_0 - \tau A^* v_0, -\tau C^* v_0) = (\mathrm{Prox}_{\tau f}(u_0 - \tau A^* v_0), 0)$. Now suppose that for some $n \geq 0$ we have

$$\mathbf{x}_n = (u_n, 0) - \tau B^* v_n, \tag{13.42a}$$

$$\mathbf{y}_n = (u_{n+1}, 0). \tag{13.42b}$$

Then

$$(u_{n+1}, 0) - \tau B^* v_{n+1} = (u_{n+1}, 0) - \tau B^*(\mathrm{Prox}_{\sigma g^*}(v_n + \sigma A(2u_{n+1} - u_n))) \tag{13.43a}$$

$$= \mathbf{y}_n - \tau B^*(\mathrm{Prox}_{\sigma g^*}(v_n + \sigma B(2(u_{n+1}, 0) - (u_n, 0)))) \tag{13.43b}$$

$$= \mathbf{y}_n - \tau B^*(\mathrm{Prox}_{\sigma g^*}(\sigma\tau BB^* v_n + \sigma B(2(u_{n+1}, 0) - (u_n, 0)))) \tag{13.43c}$$

$$= \mathbf{y}_n - \tau B^* \mathrm{Prox}_{\sigma g^*}(\sigma B(2(u_{n+1}, 0) - ((u_n, 0) - \tau B^* v_n))) \tag{13.43d}$$

$$= \mathbf{y}_n - \mathrm{Prox}_{(\tau g\circ B)^*}(2\mathbf{y}_n - \mathbf{x}_n)) = \mathbf{x}_{n+1}, \tag{13.43e}$$

where (13.43a) follows from (13.37b), (13.43b) follows from (13.42b) and Proposition 13.4(iii) below, (13.43c) follows from (13.39), and (13.43e) follows from (13.42a), Proposition 13.4(viii) and (13.49b) below applied with (γ, g) replaced by $(\tau, g \circ B)$.

Now by (13.37a) and Proposition 13.4(vii) below we have

$$(u_{n+2}, 0) = (\mathrm{Prox}_{\tau f}(u_{n+1} - \tau A^* v_{n+1}), 0) \tag{13.44a}$$

$$= \mathrm{Prox}_{\tau \tilde{f}}(u_{n+1} - \tau A^* v_{n+1}, -\tau C^* v_{n+1}) \tag{13.44b}$$

$$= \mathrm{Prox}_{\tau \tilde{f}}((u_{n+1}, 0) - \tau (A^* v_{n+1}, C^* v_{n+1})) \tag{13.44c}$$

$$= \mathrm{Prox}_{\tau \tilde{f}}((u_{n+1}, 0) - \tau B^* v_{n+1}) \tag{13.44d}$$

$$= \mathrm{Prox}_{\tau \tilde{f}}\, \mathbf{x}_{n+1} = \mathbf{y}_{n+1}. \tag{13.44e}$$

□

Acknowledgements WMM was supported by the Pacific Institute of Mathematics Postdoctoral Fellowship and the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant # CCF-1740425.

Appendices

Appendix 1

Let $A\colon X \to X$ be linear. Define

$$q_A\colon X \to \mathbb{R}\colon x \mapsto \tfrac{1}{2}\langle x, Ax\rangle. \tag{13.45}$$

Recall that a linear operator $A\colon X \to X$ is *monotone* if $(\forall x \in X)\ \langle x, Ax\rangle \geq 0$, and is *strictly monotone* if $(\forall x \in X \smallsetminus \{0\})\ \langle x, Ax\rangle > 0$. Let $h\colon X \to \mathbb{R}$ and let $x \in X$. We say that h is *Fréchet differentiable at* x if there exists a linear operator $Dh(x)\colon X \to \mathbb{R}$, called the *Fréchet derivative* of h at x, such that $\lim_{0\neq\|y\|\to 0} \frac{h(x+y)-h(x)-Dh(x)y}{\|y\|} = 0$; and h is *Fréchet differentiable on* X if it is Fréchet differentiable at every point in X.

The following lemma is a special case of [5, Proposition 17.36].

Lemma 13.3 *Let $A\colon X \to X$ be linear, strictly monotone, self-adjoint and invertible. Then the following hold:*

(i) *q_A and $q_{A^{-1}}$ are strictly convex, continuous, Fréchet differentiable. Moreover, $(\nabla q_A, \nabla q_{A^{-1}}) = (A, A^{-1})$.*

(ii) *$q_A^* = q_{A^{-1}}$.*

Proof Note that, likewise A, A^{-1} is linear, strictly monotone, self-adjoint (since $(A^{-1})^* = (A^*)^{-1} = A^{-1}$) and invertible. Moreover, $\operatorname{ran} A = \operatorname{ran} A^{-1} = X$. (i): This follows from [5, Example 17.11 and Proposition 17.36(i)] applied to A and A^{-1} respectively. (ii): It follows from [5, Proposition 17.36(iii)], [28, Theorem 4.8.5.4] and the invertibility of A that $q_A^* = q_{A^{-1}} + \iota_{\operatorname{ran} A} = q_{A^{-1}} + \iota_X = q_{A^{-1}}$. □

Proposition 13.3 *Let $L\colon Y \to X$ be linear. Suppose that L^*L is invertible. $g\colon Y \to]-\infty,+\infty]$ be convex, lower semicontinuous, and proper. Then the following hold:*

(i) $\ker L = \{0\}$.
(ii) L^*L *is strictly monotone.*
(iii) $\operatorname{dom}(q_{L^*L} + g)^* = X$.
(iv) $\partial(q_{L^*L} + g) = \nabla q_{L^*L} + \partial g = L^*L + \partial g$.
(v) $(q_{L^*L} + g^*)^*$ *is Fréchet differentiable on* X.
(vi) $(L^*L + \partial g^*)^{-1}$ *is single-valued and* $\operatorname{dom}(L^*L + \partial g^*)^{-1} = X$.
(vii) $\operatorname{Prox}_{g^*\circ L^*} = \operatorname{Id} - L(L^*L + \partial g)^{-1}L^*$.
(viii) $\operatorname{Prox}_{(g^*\circ L^*)^*} = L(L^*L + \partial g)^{-1}L^*$.

Proof (i): Using [5, Fact 2.25(vi)] and the assumption that L^*L is invertible we have $\ker L = \ker L^*L = \{0\}$. (ii): Using (i) we have $(\forall x \in X \smallsetminus \{0\})\ \langle L^*Lx, x\rangle = \langle Lx, Lx\rangle = \|Lx\|^2 > 0$, hence L^*L is strictly monotone. (iii): By (ii) and Lemma 13.3(i) applied with A replaced by L^*L we have $\operatorname{dom} q_{L^*L} = \operatorname{dom} q_{L^*L}^* = X$, hence

$$\operatorname{dom} q_{L^*L} - \operatorname{dom} g = X - \operatorname{dom} g = X. \tag{13.46}$$

It follows from (13.46), [1, Corollary 2.1] and Lemma 13.3(ii)&(i) that $\operatorname{dom}(q_{L^*L} + g)^* = \operatorname{dom} q_{L^*L}^* + \operatorname{dom} g^* = \operatorname{dom} q_{(L^*L)^{-1}} + \operatorname{dom} g^* = X + \operatorname{dom} g^* = X$. (iv): Combine (13.46), [1, Corollary 2.1] and Lemma 13.3(i). (v): Since q_{L^*L} is strictly convex, so is $q_{L^*L} + g$, which in view of [5, Proposition 18.9] and (iii) implies that $(q_{L^*L} + g)^*$ is Fréchet differentiable on $X = \operatorname{int}\operatorname{dom}(q_{L^*L} + g)^*$. (vi): Using (iv), Fact 13.8(i) applied with f replaced by $q_{L^*L} + g$, (v) and [5, Proposition 17.31(i)] we have $(L^*L + \partial g)^{-1} = (\partial(q_{L^*L} + g))^{-1} = \partial(q_{L^*L} + g)^* = \{\nabla(q_{L^*L} + g)^*\}$ is single-valued with $\operatorname{dom}(L^*L + \partial g)^{-1} = X$.

(vii): Let $x \in X = \operatorname{dom}(L^*L + \partial g)^{-1}$ and let $y \in X$ such that $y = x - L(L^*L + \partial g)^{-1}L^*x$. Then using (vi) we have

$$x = y + Lu \quad \text{where} \quad u = (L^*L + \partial g)^{-1}L^*x. \tag{13.47}$$

Consequently, $L^*y + L^*Lu = L^*x \in L^*Lu + \partial g(u)$, hence $L^*y \in \partial g(u)$, equivalently, in view of Fact 13.8(i) applied with f replaced by g, $u \in (\partial g)^{-1}(L^*y) = \partial g^*(L^*y)$. Combining with (13.47) we learn that

$$x \in y + L \circ (\partial g^*) \circ L^*(y). \tag{13.48}$$

Note that [5, Fact 2.25(vi) and Fact 2.26] implies that $\operatorname{ran} L^* = \operatorname{ran} L^*L = X$, hence $0 \in \operatorname{sri}(\operatorname{dom} g^* - \operatorname{ran} L^*)$. Therefore one can apply [5, Corollary 16.53(i)] to re-write (13.48) as $x \in (\operatorname{Id} + \partial(g^* \circ L^*))y$. Therefore, $y = \operatorname{Prox}_{g^* \circ L^*} x$ by [5, Proposition 16.44]. (viii): Apply Fact 13.8(ii) with f replaced by $g^* \circ L^*$. □

Appendix 2

Lemma 13.4 *Let $g\colon Y \to]-\infty, +\infty]$ be convex, lower semicontinuous, and proper. Consider the following statements:*

(i) *g is strongly convex.*
(ii) *g^* is Fréchet differentiable and ∇g^* is Lipschitz continuous.*
(iii) *$g^* \circ L^*$ is Fréchet differentiable and $\nabla(g^* \circ L^*) = L \circ (\nabla g^*) \circ L^*$ is Lipschitz continuous.*
(iv) *$(g^* \circ L^*)^*$ is strongly convex.*

Then (i)⇔(ii)⇒(iii)⇔(iv).

Proof (i)⇔(ii): See [5, Theorem 18.15]. (ii)⇒(iii): Clearly $g^* \circ L^*$ is Fréchet differentiable. Now let $(x, y) \in X \times X$ and suppose that $\beta > 0$ is a Lipschitz constant of ∇g^*. It follows from [5, Corollary 16.53] that $\|\nabla(g^* \circ L^*)x - \nabla(g^* \circ L^*)y\| = \|L \circ (\nabla g^*) \circ L^*x - L \circ (\nabla g^*) \circ L^*y\| = \|L((\nabla g^* \circ L^*)x - (\nabla g^* \circ L^*)y)\| \leq \|L\|\|(\nabla g^* \circ L^*)x - (\nabla g^* \circ L^*)y\| \leq \beta\|L\|\|L^*x - L^*y\| \leq \beta\|L\|\|L^*\|\|x - y\|$. (iii)⇔(iv): Use the equivalence of (i) and (ii) applied with g replaced by $(g^* \circ L^*)^*$. □

Appendix 3

We start by recalling the following well-known fact.

Fact 13.8 *Let $f\colon X \to]-\infty, +\infty]$ be convex, lower semicontinuous and proper and let $\gamma > 0$. Then the following hold:*

(i) $(\partial f)^{-1} = \partial f^*$.
(ii) $\operatorname{Prox}_{\gamma f} + \operatorname{Prox}_{(\gamma f)^*} = \operatorname{Id}$.

Proof (i): See, e.g., [27, Remark on page 216] or [23, Théorème 3.1].
(ii): See, e.g., [5, Theorem 14.3(iii)]. □

Lemma 13.5 *Let $\gamma > 0$. The Douglas–Rachford method given in (13.9) applied to the ordered pair $(\gamma f, \gamma g)$ with a starting point $x_0 \in X$ to solve (13.8) can be rewritten as:*

$$y_n = \operatorname{Prox}_{\gamma f}\, x_n \tag{13.49a}$$

$$x_{n+1} = y_n - \operatorname{Prox}_{(\gamma g)^*}(2y_n - x_n). \tag{13.49b}$$

Proof Using (13.9a), (13.10), and Fact 13.8(ii) applied with f replaced by g we have

$$\begin{aligned} x_{n+1} &= x_n - \operatorname{Prox}_{\gamma f}\, x_n + \operatorname{Prox}_{\gamma g}(2\operatorname{Prox}_{\gamma f}\, x_n - x_n) = x_n - y_n + \operatorname{Prox}_{\gamma g}(2y_n - x_n) \\ &= x_n - y_n + 2y_n - x_n - \operatorname{Prox}_{(\gamma g)^*}(2y_n - x_n) = y_n - \operatorname{Prox}_{(\gamma g)^*}(2y_n - x_n), \end{aligned} \tag{13.50}$$

and the conclusion follows. □

Appendix 4

Proposition 13.4 *Let $(x, y, z) \in X \times Y \times Z$ and let B and $\widetilde{f}$ be defined as in (13.39) and (13.41). Then the following hold:*

(i) $B^* y = (A^* y, C^* y)$.
(ii) $\operatorname{dom} \widetilde{f} = \operatorname{dom} f \times \{0\}$.
(iii) $(\forall (x, z) \in \operatorname{dom} \widetilde{f})$ *we have* $z = 0$ *and* $B(x, z) = Ax$.
(iv) $B(\operatorname{dom} \widetilde{f}) = A(\operatorname{dom} f)$.
(v) $0 \in \operatorname{sri}(\operatorname{dom} g - B(\operatorname{dom} \widetilde{f}))$.
(vi) $\operatorname{argmin}(\widetilde{f} + g \circ B) = \operatorname{argmin}(f + g \circ A) \times \{0\} \neq \varnothing$.
(vii) $\operatorname{Prox}_{\widetilde{f}}(x, z) = (\operatorname{Prox}_f x, 0)$.
(viii) $\operatorname{Prox}_{(\tau g \circ B)^*} = \tau B^* \operatorname{Prox}_{\sigma g^*}(\sigma B)$.

Proof (i): This clearly follows from (13.39). (ii): It follows from (13.41) that $\operatorname{dom} \widetilde{f} = \operatorname{dom} f \times \operatorname{dom} \iota_{\{0\}} = \operatorname{dom} f \times \{0\}$. (iii): The claim that $z = 0$ follows from (ii). Now combine with (13.39). (iv): Combine (ii) and (iii). (v): Combine (iv) and (13.34). (vi): We have

$$\operatorname{argmin}(\widetilde{f} + g \circ B) = \operatorname{zer}(\partial \widetilde{f} + B^* \circ \partial g \circ B) \tag{13.51a}$$

$$= \operatorname{zer}(\partial f \times N_{\{0\}} + ((A^* \circ \partial g \circ A) \times (C^* \circ \partial g \circ C))) \tag{13.51b}$$

$$= (\operatorname{zer}(\partial f + A^* \circ \partial g \circ A)) \times (\operatorname{zer}(N_{\{0\}} + C^* \circ \partial g \circ C)), \tag{13.51c}$$

where (13.51a) follows from (v) and (13.4) applied with (f, g, L) replaced by $(g, \widetilde{f}, B)$, and (13.51b) follows from (13.39) and (13.41). Therefore, $(x, z) \in \operatorname{argmin}(\widetilde{f} + g \circ B) \Leftrightarrow [z = 0 \text{ and } x \in \operatorname{zer}(\partial f + A^* \circ \partial g \circ A)] \Leftrightarrow (x, z) \in$

$\operatorname{argmin}(f + g \circ A) \times \{0\}$. Now combine with (13.4). (vii): Combine (13.41) and [5, Proposition 23.18]. (viii): Indeed, Proposition 13.3(viii) implies

$$\operatorname{Prox}_{(\tau g \circ B)^*} = B^*(BB^* + (\tau \partial g)^*)^{-1} B = B^*(\sigma^{-1}\tau^{-1}\operatorname{Id} + \partial g^* \circ \tau^{-1}\operatorname{Id})^{-1} B \tag{13.52a}$$

$$= B^*(\sigma^{-1}(\operatorname{Id} + \sigma \partial g^*)\tau^{-1} \circ \operatorname{Id})^{-1} B = \tau B^* \operatorname{Prox}_{\sigma g^*}(\sigma B). \tag{13.52b}$$

□

References

1. Attouch, H., Brézis, H.: Duality for the sum of convex functions in general Banach spaces. In: Aspects of Mathematics and Its Applications **34**, pp. 125–133. North-Holland, Amsterdam (1986)
2. Attouch, H., Théra, M.: A general duality principle for the sum of two operators. J. Convex Anal. **3**, 1–24 (1996)
3. Bauschke, H.H., Combettes, P.L.: A Dykstra-like algorithm for two monotone operators. Pacific J. Optim. **4**, 383–391 (2008)
4. Bauschke, H.H., Boţ, R.I., Hare, W.L., Moursi, W.M.: Attouch–Théra duality revisited: paramonotonicity and operator splitting. J. Approx. Th. **164**, 1065–1084 (2012)
5. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Second Edition. Springer (2017)
6. Bauschke, H.H., Koch, V.R.: Projection methods: Swiss Army knives for solving feasibility and best approximation problems with halfspaces. In: Infinite Products of Operators and Their Applications, Contemp. Math. vol. 636, pp. 1–40 (2012)
7. Beck, A.: First-Order Methods in Optimization, SIAM (2017)
8. Boţ, R.I, Csetnek, E.R.: ADMM for monotone operators: convergence analysis and rates. Advances Comp. Math. **45**, 327–359 (2019)
9. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**, 1–122 (2011)
10. Boyle, J.P., Dykstra, R.L.: A method for finding projections onto the intersection of convex sets in Hilbert spaces. Lecture Notes in Statistics **37**, 28–47 (1986)
11. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging, J. Math. Imaging Vis. **40**, 120–145 (2011)
12. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization **53**, 475–504 (2004)
13. Briceño-Arias, L.M., Combettes, P.L.: A monotone + skew splitting model for composite monotone inclusions in duality. SIAM J. Optim. **21**, 1230–1250 (2011)
14. Combettes, P.L., Pesquet, J.-C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Valued Var. Anal. **20**, 307–330 (2012)
15. Combettes, P.L., Pesquet, J.-C.: A proximal decomposition method for solving convex variational inverse problems. Inverse Problems **24**, article 065014 (2008)
16. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering, vol. 49, pp. 185–212. Springer, New York (2011)

17. Combettes, P.L., Vũ, B.-C.: Variable metric forward–backward splitting with applications to monotone inclusions in duality. Optimization **63**, 1289–1318 (2014)
18. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Th. Appl. **158**, 460–479 (2013)
19. Deutsch, F.: Best Approximation in Inner Product Spaces, Springer (2001)
20. Eckstein, J.: Splitting Methods for Monotone Operators with Applications to Parallel Optimization, Ph.D. thesis, MIT (1989)
21. Eckstein, J., Bertsekas, D.P.: On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Prog. **55**, 293–318 (1992)
22. Gabay, D.: Applications of the method of multipliers to variational inequalities. In: Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, vol. 15, pp. 299–331. North-Holland, Amsterdam (1983)
23. Gossez, J.-P.: Opérateurs monotones non linéaires dans les espaces de Banach non réflexifs. J. Math. Anal. Appl. **34**, 371–395 (1971)
24. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
25. O'Connor, D., Vandenberghe, L.: On the equivalence of the primal-dual hybrid gradient method and Douglas–Rachford splitting. Math. Prog. (Ser. A) (2018), https://doi.org/10.1007/s10107-018-1321-1.
26. Riesz, F., Sz.-Nagy, B.: Functional Analysis. Dover paperback (1990)
27. Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pacific J. Math. **33**, 209–216 (1970)
28. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. Third Edition. Springer-Verlag (2002)
29. Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Con. Optim. **29**, 119–138 (1991)
30. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**, 667–681 (2013)

Chapter 14
Quasidensity: A Survey and Some Examples

Stephen Simons

Abstract In three previous papers, we discussed quasidense multifunctions from a Banach space into its dual, or, equivalently, quasidense subsets of the product of a Banach space and its dual. In this paper, we survey (without proofs) some of the main results about quasidensity, and give some simple limiting examples in Hilbert spaces, reflexive Banach spaces, and nonreflexive Banach spaces.

Keywords Multifunction · Maximal monotonicity · Quasidensity · Sum theorem · Subdifferential · Strong maximality · Type (FPV) · Type (FP)

AMS 2010 Subject Classification 47H05, 47N10, 52A41, 46A20

14.1 Introduction

This is a sequel to the papers [17] and [18], in which we discussed *quasidense* multifunctions from a Banach space into its dual. A number of the results in [17] depend on the somewhat more abstract analysis that appears in [16].

In Section 14.2, we give some Banach space notation and definitions.

Let S be a multifunction (not assumed to be monotone) from a Banach space into its dual. We define the *quasidensity* of S in Definition 14.1. In Theorem 14.4, we establish that the (appropriately defined) subdifferential of a proper (not necessarily convex) lower semicontinuous function is quasidense, and we show in the simple Example 14.6 that the condition (14.2), which is sufficient for the quasidensity, is not necessary.

In Section 14.4, we start our investigation of *monotone* multifunctions and collect together some of the results that were proved in [17], with references to the original proofs in [17] or [16], as the case may be. We point out in Theorem 14.7 and Example 14.8 that every closed, monotone quasidense multi-

S. Simons (✉)
Department of Mathematics, University of California, Santa Barbara, CA, USA

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_14

function is maximally monotone, but that there exist maximally monotone linear operators that are not quasidense. We point out in Theorem 14.9 that the subdifferential of any proper, convex lower semicontinuous function is quasidense. By virtue of Theorem 14.7, Theorem 14.9 generalizes Rockafellar's result that such subdifferentials are maximally monotone. In Theorem 14.10 we prove that the sum of a pair of closed, monotone quasidense multifunctions that satisfy the Rockafellar constraint qualification is closed, monotone and quasidense. We note that it is apparently not known whether the sum of a pair of maximally monotone multifunctions that satisfy the Rockafellar constraint qualification is necessarily maximally monotone. (This is known as the *sum problem*.) In Theorem 14.11 we give a "parallel" sum theorem for a pair of closed, monotone quasidense multifunctions that satisfy the "dual" of the Rockafellar constraint qualification. In the process of doing this we introduce the *Fitzpatrick function* and *Fitzpatrick extension* of a closed, monotone, quasidense multifunction. In Problems 14.13 and 14.16 we give two questions that merit further study.

Quasidensity has connections with many of the subclasses of the maximally monotone multifunctions that have been investigated over the years. We explore just three of these in Section 14.5: *type (FPV)*, *type (FP)* and *strongly maximal*. Problems 14.20, 14.23 and 14.26 contain open questions about these three subclasses of the maximally monotone multifunctions. Other related subclasses are discussed in [17, Theorem 8.1], [17, Theorem 8.2], [16, Theorem 11.6, p. 1045] and [16, Theorem 11.9, pp. 1045–1046].

In the final three sections, we show how quasidensity behaves in three special cases: Hilbert spaces in Section 14.6, reflexive Banach spaces in Section 14.7 and nonreflexive Banach spaces in Section 14.8.

The author would like to express his thanks to Hedy Attouch and Heinz Bauschke for constructive discussions about the topics discussed in this paper. He would also like to thank Xianfu Wang for constructive comments about an earlier version of this paper.

14.2 Banach Space Notation and Definitions

If X is a nonzero real Banach space and $f\colon\ X \to]-\infty, \infty]$, we write $\operatorname{dom} f$ for the set $\{x \in X\colon\ f(x) \in \mathbb{R}\}$. $\operatorname{dom} f$ is the *effective domain* of f. We say that f is *proper* if $\operatorname{dom} f \neq \emptyset$. We write $\mathcal{PCLSC}(X)$ for the set of all proper convex lower semicontinuous functions from X into $]-\infty, \infty]$.

We write X^* for the dual space of X (with the pairing $\langle \cdot, \cdot \rangle\colon X \times X^* \to \mathbb{R}$). If $f \in \mathcal{PCLSC}(X)$ then, as usual, we define the *Fenchel conjugate*, f^*, of f to be the function on X^* given by $x^* \mapsto \sup_X [x^* - f]$.

We write X^{**} for the bidual of X (with the pairing $\langle \cdot, \cdot \rangle\colon X^* \times X^{**} \to \mathbb{R}$). If $f \in \mathcal{PCLSC}(X)$ and $f^* \in \mathcal{PCLSC}(X^*)$, we define $f^{**}\colon X^{**} \to]-\infty, \infty]$ by $f^{**}(x^{**}) := \sup_{X^*} [x^{**} - f^*]$. If $x \in X$, we write $\widehat{x}$ for the canonical image of x in X^{**}, that is to say, for all $(x, x^*) \in X \times X^*$, $\langle x^*, \widehat{x} \rangle = \langle x, x^* \rangle$.

If $f \in \mathcal{PCLSC}(X)$, then the *convex subdifferential of* f is the multifunction $\partial f\colon\ E \rightrightarrows E^*$ that satisfies

$$x^* \in \partial f(x) \iff f(x) + f^*(x^*) = \langle x, x^*\rangle.$$

We suppose that E is a nonzero real Banach space with dual E^*. For all $(x, x^*) \in E \times E^*$, we write $\|(x, x^*)\| := \sqrt{\|x\|^2 + \|x^*\|^2}$. We represent $(E \times E^*)^*$ by $E^* \times E^{**}$, under the pairing

$$\big\langle (x, x^*), (y^*, y^{**})\big\rangle := \langle x, y^*\rangle + \langle x^*, y^{**}\rangle.$$

The dual norm on $E^* \times E^{**}$ is given by $\|(y^*, y^{**})\| := \sqrt{\|y^*\|^2 + \|y^{**}\|^2}$.

Now let $S\colon\ E \rightrightarrows E^*$. We write $G(S)$ for the graph of S, $D(S)$ for the domain of S and $R(S)$ for the range of S. We will always suppose that $G(S) \neq \emptyset$ (equivalently, $D(S) \neq \emptyset$ or $R(S) \neq \emptyset$). We say that S is *closed* if $G(S)$ is closed. If $x \in E$, we define the multifunction ${}_xS\colon\ E \rightrightarrows E^*$ by ${}_xS = (S^{-1} - x)^{-1}$. Then ${}_xS(t) = S(t+x)$. We write $J : E \rightrightarrows E^*$ for the *duality map*. We recall that J is maximally monotone and

$$x^* \in Jx \iff \tfrac{1}{2}\|x\|^2 + \tfrac{1}{2}\|x^*\|^2 = \langle x, x^*\rangle \iff \|x\|^2 = \|x^*\|^2 = \langle x, x^*\rangle. \tag{14.1}$$

14.3 Quasidensity

Definition 14.1 We say that S is *quasidense* if, for all $(x, x^*) \in E \times E^*$,

$$\inf_{(s,s^*)\in G(S)} \left[\tfrac{1}{2}\|s - x\|^2 + \tfrac{1}{2}\|s^* - x^*\|^2 + \langle s - x, s^* - x^*\rangle\right] \leq 0.$$

See [17, Definition 3.1] and [16, Example 7.1, eqn. (28), p. 1031)].

We have the following simple result connecting J and quasidensity:

Lemma 14.2 *Let $S\colon\ E \rightrightarrows E^*$ and, for all $x \in E$, ${}_xS + J$ be surjective. Then S is quasidense.*

Proof Let $(x, x^*) \in E \times E^*$. Choose $t \in D({}_xS)$ such that $({}_xS + J)t = x^*$. So there exists $s^* \in S(t + x)$ such that $(t, x^* - s^*) \in G(J)$. Thus, writing $s := t + x$, $(s, s^*) \in G(S)$ and $(s - x, x^* - s^*) = (t, x^* - s^*) \in G(J)$, that is to say

$$\tfrac{1}{2}\|s - x\|^2 + \tfrac{1}{2}\|x^* - s^*\|^2 = \langle s - x, x^* - s^*\rangle.$$

Equivalently,

$$\tfrac{1}{2}\|s - x\|^2 + \tfrac{1}{2}\|s^* - x^*\|^2 + \langle s - x, s^* - x^*\rangle = 0.$$

This obviously implies that S is quasidense. □

We now discuss a significant example of quasidensity. The following definition was made in [18, Definition 2.1, p. 633].

Definition 14.3 A *ubiquitous subdifferential*, ∂_u, is a rule that associates with each proper lower semicontinuous function $f: E \to]-\infty, \infty]$ a multifunction $\partial_u f : E \rightrightarrows E^*$ such that

- $\partial_u f(x) = \emptyset$ if $x \notin \operatorname{dom} f$,
- $0 \in \partial_u f(x)$ if f attains a strict global minimum at x,
- $\partial_u(f + h)(x) \subseteq \partial_u f(x) + \partial h(x)$ whenever $x \in \operatorname{dom} f$ and h is a continuous convex real function on E (here ∂h is the convex subdifferential of h).

There is a list of abstract subdifferentials that satisfy these conditions in the remarks following [18, Definition 2.1]. Now suppose that ∂_u is a ubiquitous subdifferential. We have the following result:

Theorem 14.4 *Let $f : E \to \mathbb{R}$ be proper and lower semicontinuous. Let $a_0, b_0, c_0 \in \mathbb{R}$ with $a_0 < \frac{1}{2}$ and,*

$$\text{for all } x \in E, \ f(x) \geq -a_0\|x\|^2 - b_0\|x\| - c_0. \tag{14.2}$$

Then $\partial_u f$ is quasidense.

Proof See [18, Theorem 3.2, pp. 634–635]. The proof of this is based on the "elementary" proof of Theorem 14.9, that is [17, Theorem 4.6]. □

Corollary 14.5 *Let $f : E \to]-\infty, \infty]$ be proper, lower semicontinuous and dominate a continuous affine function. Then $\partial_u f$ is quasidense.*

Proof This is immediate from Theorem 14.4. □

Example 14.6 In this example, we suppose that $E = \mathbb{R}$ and that ∂_u has the special property that, whenever f is a polynomial, $\partial_u f(x) = \{f'(x)\}$. For instance, ∂_u could be the Clarke–Rockafellar subdifferential. Let f be a polynomial. Then the statement that $\partial_u f$ is quasidense can be rewritten:

$$\text{for all } z \in \mathbb{R}, \ \inf_{s\in\mathbb{R}} \tfrac{1}{2}(s + f'(s) - z)^2 \leq 0.$$

Let $\lambda \in \mathbb{R}$ and $f(x) := -\lambda x^2$. So $\partial_u f$ is quasidense if, and only if,

$$\text{for all } z \in \mathbb{R} \ \inf_{s\in\mathbb{R}} \tfrac{1}{2}(s - 2\lambda s - z)^2 \leq 0.$$

- If $\lambda \neq \frac{1}{2}$, then taking $s := z/(1 - 2\lambda)$ shows that $\partial_u f$ is quasidense.

- If $\lambda = \frac{1}{2}$, then taking $z \neq 0$ shows that $\partial_u f$ is not quasidense.
  Thus the condition (14.2) is sufficient but not necessary for the quasidensity of $\partial_u f$.
  This example (with different justification) is taken from [18, Example 3.5, p. 636].

14.4 Monotone Multifunctions: Basic Results

For the rest of this paper, we will discuss the very rich theory of the quasidensity of *monotone* multifunctions.

Theorem 14.7 (Quasidensity and Maximality) *Let $S\colon E \rightrightarrows E^*$ be closed, monotone and quasidense. Then S is maximally monotone.*

Proof See [17, Theorem 3.2], [16, Theorem 7.4(a), pp. 1032–1033] or [16, Lemma 4.7, p. 1027]. □

Example 14.8 (The Tail Operator) Let $E = \ell_1$, and define the linear map $T\colon \ell_1 \mapsto \ell_\infty = E^*$ by $(Tx)_n = \sum_{k \geq n} x_k$. T is maximally monotone, but not quasidense. See [16, Example 7.10, pp. 1034–1035].

Theorem 14.9 below is a very important result. By virtue of Theorem 14.7, it generalizes Rockafellar's result [11] that subdifferentials of proper, convex, lower semicontinuous functions are maximally monotone. The first proof of this result mentioned below was the source of Theorem 14.4.

Theorem 14.9 *Let $f \in \mathcal{PCLSC}(E)$. Then ∂f is closed, monotone and quasidense.*

Proof The more elementary proof of this result (see [17, Theorem 4.6]) uses the Brøndsted–Rockafellar theorem [4] and Rockafellar's formula [10] for the subdifferential of a sum. There is a slicker but more sophisticated proof using the properties of *Fitzpatrick functions* (see below) in [16, Theorem 7.5, p. 1033]. □

As we noted in the introduction, it is apparently not known whether the result corresponding to Theorem 14.10 with "closed, monotone and quasidense" replaced by "maximally monotone" is true.

Theorem 14.10 (Sum Theorem with Domain Cconstraints) *Let $S, T\colon E \rightrightarrows E^*$ be closed, monotone and quasidense and $D(S) \cap \operatorname{int} D(T) \neq \emptyset$. Then $S + T$ is closed, monotone and quasidense.*

Proof This is a special case of [16, Theorem 8.4(a)$\Longrightarrow$(d), pp. 1036–1037]. □

There is a "dual" version of Theorem 14.10 that we will state in Theorem 14.11. Before discussing this, we introduce the *Fitzpatrick function*, $\varphi_S\colon E \times E^* \rightarrow$

$]-\infty, \infty]$, and the *Fitzpatrick extension*, $S^{\mathbb{F}}\colon\ E^* \rightrightarrows E^{**}$, of a closed, monotone, quasidense multifunction $S\colon\ E \rightrightarrows E^*$. The function φ_S is defined by

$$\varphi_S(x, x^*) := \sup_{(s,s^*)\in G(S)} \big[\langle s, x^*\rangle + \langle x, s^*\rangle - \langle s, s^*\rangle\big].$$

See [5], [17, Definition 3.4] and many other places. The multifunction $S^{\mathbb{F}}$ was defined in [17, Definition 5.1] by

$$(y^*, y^{**}) \in G(S^{\mathbb{F}}) \text{ exactly when } \varphi_S{}^*(y^*, y^{**}) = \langle y^*, y^{**}\rangle.$$

(There is a more abstract version of this in [16, Definition 8.5, p. 1037].) The word *extension* is justified by the easily verifiable fact that

$$(x, x^*) \in G(S) \iff (x^*, \widehat{x}) \in G(S^{\mathbb{F}}).$$

It was shown in [17, Section 11] that $(y^*, y^{**}) \in G(S^{\mathbb{F}})$ exactly when (y^{**}, y^*) is in the *Gossez extension* of $G(S)$ (see [8, Lemma 2.1, p. 275]).

Theorem 14.11 (Sum Theorem with Range Constraints) *Let $S, T\colon\ E \rightrightarrows E^*$ be closed, monotone and quasidense and $R(S) \cap \operatorname{int} R(T) \neq \emptyset$. Then the multifunction $y \mapsto (S^{\mathbb{F}} + T^{\mathbb{F}})^{-1}(\widehat{y})$ is closed, monotone and quasidense. Under certain additional technical conditions, the* parallel sum *$(S^{-1} + T^{-1})^{-1}$ is closed, monotone and quasidense.*

Proof This is a special case of [16, Theorem 8.8, p. 1039]. □

If $S\colon\ E \rightrightarrows E^*$ is closed, monotone and quasidense, then it is easily seen that $S^{\mathbb{F}}$ is monotone. In fact, we have the following stronger nontrivial result:

Theorem 14.12 *Let $S\colon\ E \rightrightarrows E^*$ be closed, monotone and quasidense. Then the multifunction $S^{\mathbb{F}}\colon E^* \rightrightarrows E^{**}$ is maximally monotone.*

Proof See [16, Lemma 12.5, p. 1047]. There is also a sketch of a proof in [17, Section 11]. □

This leads to the following problem:

Problem 14.13 Let $S\colon\ E \rightrightarrows E^*$ be closed, monotone and quasidense. Then is the multifunction $S^{\mathbb{F}}\colon E^* \rightrightarrows E^{**}$ necessarily quasidense?

Theorem 14.14 *Let $f \in \mathcal{PCLSC}(E)$. Then $(\partial f)^{\mathbb{F}} = \partial(f^*)$.*

Proof See [17, Theorem 5.7]. □

Remark 14.15 Theorem 14.14 is equivalent to [8, Théorème 3.1, pp. 376–378].

Problem 14.16 The proof of [17, Theorem 5.7] (invoked in Theorem 14.14) is quite convoluted. Is there a simple direct proof of this result?

Remark 14.17 Theorems 14.14 and 14.9 show that *if* $f \in \mathcal{PCLSC}(E)$ *then* $(\partial f)^{\mathbb{F}}$ *is quasidense*, in other words, in this restricted situation we have a positive solution to Problem 14.13.

14.5 Quasidensity and the Classification of Maximally Monotone Multifunctions

The closed, monotone, quasidense multifunctions have relationships with many other subclasses of the maximally monotone multifunctions. We shall discuss just three of these. Four others are mentioned in the introduction.

Definition 14.18 Let $S\colon E \rightrightarrows E^*$ be monotone. We say that S is *of type (FPV)* or *maximally monotone locally* if, whenever U is an open convex subset of E, $U \cap D(S) \neq \emptyset$, $(w, w^*) \in U \times E^*$ and

$$(s, s^*) \in G(S) \quad \text{and} \quad s \in U \quad \Longrightarrow \quad \langle s - w, s^* - w^* \rangle \geq 0,$$

then $(w, w^*) \in G(S)$. (If we take $U = E$, we see that every monotone multifunction of type (FPV) is maximally monotone.) See [15, pp. 150–151].

Theorem 14.19 *Any closed, monotone, quasidense multifunction is maximally monotone of type (FPV).*

Proof See [17, Theorem 7.2]. □

Problem 14.20 Is every maximally monotone multifunction of type (FPV)? The tail operator (see Example 14.8) does not provide a negative example because it was proved in Fitzpatrick–Phelps, [6, Theorem 3.10, p. 68] that if $S\colon E \rightrightarrows E^*$ is maximally monotone and $D(S) = E$ then S is of type (FPV). Also, it was proved in [15, Theorem 46.1, pp. 180–182] that *if* $S\colon E \rightrightarrows E^*$ *is maximally monotone and* $G(S)$ *is convex then* S *is of type (FPV).* A negative example would lead to a negative solution for the sum problem. See [15, Theorem 44.1, p. 170]

Definition 14.21 Let $S\colon E \rightrightarrows E^*$ be monotone. We say that S is *of type (FP)* or *locally maximally monotone* if, whenever $\widetilde{U}$ is a convex open subset of E^*, $\widetilde{U} \cap R(S) \neq \emptyset$, $(w, w^*) \in E \times \widetilde{U}$ and

$$(s, s^*) \in G(S) \text{ and } s^* \in \widetilde{U} \quad \Longrightarrow \quad \langle s - w, s^* - w^* \rangle \geq 0,$$

then $(w, w^*) \in G(S)$. (If we take $\widetilde{U} = E^*$, we see that every monotone multifunction of type (FP) is maximally monotone.) See [15, pp. 149–150].

Theorem 14.22 *A maximally monotone multifunction is quasidense* $\Longleftrightarrow$ *it is of type (FP).*

Proof See [17, Theorem 10.3]. This result is related to results of Marques Alves and Svaiter, [1, Theorem 1.2(1$\Longleftrightarrow$5), p. 885], Voisei and Zălinescu, [19, Theorem 4.1, pp. 1027–1028] and Bauschke, Borwein, Wang and Yao, [2, Theorem 3.1, pp. 1878–1879]. □

Problem 14.23 The proof of ($\Longrightarrow$) in Theorem 14.22 relies on [17, Lemma 10.1]. Is there a *simple direct* proof of this result? In this connection, see also the proof of [17, Lemma 12.2], which is hardly simple and direct.

Definition 14.24 Let $S\colon\ E \rightrightarrows E^*$ be monotone. We say that S is *strongly maximal* (see [13, Theorems 6.1-2, pp. 1386–1387]) if, whenever $w \in E$ and $\widetilde{W}$ is a nonempty $w(E^*, E)$–compact convex subset of E^* such that

$$\text{for all } (s, s^*) \in G(S),\ \max\langle s - w, s^* - \widetilde{W}\rangle \geq 0,$$

then $Sw \cap \widetilde{W} \neq \emptyset$ and, further, whenever W is a nonempty $w(E, E^*)$–compact convex subset of E, $w^* \in E^*$ and,

$$\text{for all } (s, s^*) \in G(S),\ \max\langle s - W, s^* - w^*\rangle \geq 0,$$

then $w^* \in S(W)$. This property was originally proved for convex subdifferentials. If S is strongly maximal, then clearly S is maximal.

Theorem 14.25 *Let $S\colon\ E \rightrightarrows E^*$ be closed, monotone and quasidense. Then S is strongly maximal.*

Proof See [17, Theorem 8.5]. □

Problem 14.26 Is every maximally monotone multifunction strongly maximal? The tail operator (see Example 14.8) does not provide a negative example because it was proved in Bauschke–Simons, [3, Theorem 1.1, pp. 166–167] that if the function $S\colon\ D(S) \subset E \to E^*$ is linear and maximally monotone then S is strongly maximal. More generally, it was proved in [15, Theorem 46.1, pp. 180–182] that *if $S\colon E \rightrightarrows E^*$ is maximally monotone and $G(S)$ is convex then S is strongly maximal.*

14.6 The Hilbert Space Case

Let H be a real Hilbert space and $I\colon\ H \to H$ be the identity map. As usual, we identify H^* with H. Let $S\colon\ H \rightrightarrows H$. From Definition 14.1 and the properties of Hilbert spaces, S is quasidense exactly when, for all $(x, x^*) \in H \times H$,

$$\inf_{(s,s^*)\in G(S)} \tfrac{1}{2}\|s + s^* - x - x^*\|^2 \leq 0,$$

that is to say, for all $z^* \in H$, $\inf_{(s,s^*)\in G(S)} \frac{1}{2}\|s + s^* - z^*\|^2 \leq 0$. This is equivalent to the statement that $\{s + s^* \colon (s, s^*) \in G(S)\}$ is dense in H, that is to say $R(S + I)$ is dense in H. This leads to the following result:

Theorem 14.27 *Let $S \colon H \rightrightarrows H$ be closed and monotone. Then S is quasidense if, and only if, $S + I$ is surjective.*

Proof "If" is obvious from the comments above. Suppose, conversely, that S is quasidense. Then, from Theorem 14.7, S is maximally monotone, and the surjectivity of $S + I$ follows from Minty's theorem. □

Monotonicity plays a mysterious role in Theorem 14.27. This is shown by the following example.

Example 14.28 Define $S \colon \mathbb{R} \rightrightarrows \mathbb{R}$ by

$$S(x) := \begin{cases} \{1/x - x\} & (x \neq 0); \\ \emptyset & (x = 0). \end{cases}$$

Clearly, S is closed. Then

$$(S + I)(x) = \begin{cases} \{1/x\} & (x \neq 0); \\ \emptyset & (x = 0). \end{cases}$$

Thus $R(S + I) = \mathbb{R} \setminus \{0\}$. Since this is dense in $\mathbb{R}$, S is quasidense. But $S + I$ is manifestly not surjective.

14.7 The Reflexive Banach Space Case

Let E be a real reflexive Banach space.

Theorem 14.29 *Let $S \colon E \rightrightarrows E^*$ be closed and monotone. Then S is quasidense if, and only if, for all $x \in E$, ${}_xS + J$ is surjective.*

Proof "If" was established in Lemma 14.2. Suppose, conversely, that S is quasidense and $x \in E$. Since $G({}_xS) = G(S) - (x, 0)$, ${}_xS$ is closed, monotone and quasidense. Theorem 14.7 implies that ${}_xS$ is maximally monotone, and the surjectivity of ${}_xS + J$ follows from [14, Theorem 10.7, p. 24]. □

Remark 14.30 If the norm of E is produced by an Asplund renorming, one can use Rockafellar's generalization [12] of Minty's theorem instead of the result cited from [14] to prove that ${}_xS + J$ is surjective in Theorem 14.29($\Longrightarrow$).

We shall see in Example 14.31 below that the surjectivity of $S + J$ alone is not enough to ensure the quasidensity of S in Theorem 14.29($\Longleftarrow$).

Example 14.31 Define the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ on $\mathbb{R}^2$ by $\|(x_1, x_2)\|_1 = |x_1|+|x_2|$ and $\|(y_1, y_2)\|_\infty = |y_1| \vee |y_2|$. Let $E := (\mathbb{R}^2, \|\cdot\|_1)$. Then $E^* = (\mathbb{R}^2, \|\cdot\|_\infty)$. Let A be the union of the two axes in E, that is to say, $A = (\mathbb{R}, 0) \cup (0, \mathbb{R})$. Define $S\colon E \rightrightarrows E^*$ by

$$S(x) = \begin{cases} J(x) & (x \in A); \\ \emptyset & (x \notin A). \end{cases}$$

Since $G(S) = G(J) \cap (A \times E^*)$, S is closed and monotone. We shall prove that

$$S + J \text{ is surjective} \tag{14.3}$$

but

$$S \text{ is not quasidense.} \tag{14.4}$$

Let P be the square $\{y \in E^*\colon \|y\|_\infty = 1\}$, P_E be the line segment $\{1\}\times[-1, 1]$, P_N be the line segment $[-1, 1] \times \{1\}$, P_W be the line segment $-P_E$, and P_S be the line segment $-P_N$. (E, N, W and S stand for East, North, West and South, respectively.) Clearly, $P = P_E \cup P_N \cup P_W \cup P_S$. Let $e_1 = (1, 0)$ and $e_2 = (0, 1)$.

(a) If $y \in P_E$, then $\frac{1}{2}\|e_1\|_1^2 + \frac{1}{2}\|y\|_\infty^2 = \frac{1}{2} + \frac{1}{2} = 1 = \langle e_1, y\rangle$. Thus $y \in J(e_1)$.
(b) If $y \in P_N$, then interchanging the indices 1 and 2 in (a), $y \in J(e_2)$.
(c) If $y \in P_W$, then $-y \in P_E$. From (a), $-y \in J(e_1)$, and so $y \in J(-e_1)$.
(d) If $y \in P_S$, then $-y \in P_N$. From (b), $-y \in J(e_2)$, and so $y \in J(-e_2)$.
(e) Let V be the set consisting of the four points $\pm e_1$ and $\pm e_2$. It follows from (a)–(d) that $P \subset J(V)$.
(f) Let $\lambda > 0$. From (e), $\lambda P \subset \lambda J(V) = J(\lambda V) \subset J(A)$. Furthermore, $(0, 0) \in J(0, 0) \subset J(A)$. Thus $\mathbb{R}^2 = \bigcup_{\lambda>0} \lambda P \cup \{(0, 0)\} \subset J(A)$. Since J is monotone, so is S and, since A is closed, so is S.
(f) shows that S is surjective. Now $R(S + J) \supset R(S + S) \supset R(2S) = 2R(S)$, and so $S + J$ is surjective, giving (14.3). However, since $G(S)$ is a proper subset of $G(J)$, S is not maximally monotone thus, from Theorem 14.7 not quasidense, giving (14.4).

This example is patterned after two examples (one due to S. Fitzpatrick and the other due to H. Bauschke) that appear in [14, p. 25] in which S is monotone, $S + J$ is surjective but S is not maximally monotone. However, in both of these examples, S is not closed.

14.8 The Nonreflexive Banach Space Case

We now suppose that E is a nonreflexive Banach space, and we discuss a possible analog of Theorem 14.29. Theorem 14.29($\Longleftarrow$) is true in this case too, since the proof does not depend on the reflexivity of E (or even the monotonicity or

closedness of S). We shall show in Example 14.32 below that Theorem 14.29($\Longrightarrow$) fails in the most spectacular way.

Example 14.32 Since E is not reflexive, from James's theorem (see Pryce [9] or Ruiz Galán–Simons [7]), there exists $x^* \in E^* \setminus \{0\}$ such that $x \in E$ and $\|x\| = 1 \Longrightarrow \langle x, x^* \rangle < \|x^*\|$. It follows that $x \in E \setminus \{0\} \Longrightarrow \langle x, x^* \rangle < \|x\|\|x^*\|$. We now prove that $x^* \notin R(J)$. Indeed, if there existed $x \in E$ such that $x^* \in Jx$ then, from (14.1), $\frac{1}{2}\|x\|^2 + \frac{1}{2}\|x^*\|^2 = \langle x, x^* \rangle < \|x\|\|x^*\|$, which is manifestly impossible. Thus $x^* \notin R(J)$, and so J is not surjective. Now let $S = 0$. Then, for all $x \in E$, ${}_xS = 0$ and so ${}_xS + J$ is not surjective. On the other hand, S is (closed, monotone and) quasidense. The fastest way of seeing this is to note that S is a convex subdifferential and use Theorem 14.9. However, for the benefit of the reader, we will now give a direct proof of the quasidensity of S.

Let $(x, x^*) \in E \times E^*$ and $\varepsilon > 0$. The definition of $\|x^*\|$ provides an element t of E such that $\|t\| \leq \|x^*\|$ and $\langle t, x^* \rangle \geq \|x^*\|^2 - \varepsilon$. (If $x^* = 0$, we take $t = 0$). Thus, writing $s = t + x$,

$$\begin{aligned} \tfrac{1}{2}\|s - x\|^2 + \tfrac{1}{2}\|0 - x^*\|^2 + \langle s - x, 0 - x^* \rangle &= \tfrac{1}{2}\|t\|^2 + \tfrac{1}{2}\|x^*\|^2 - \langle t, x^* \rangle \\ &\leq \|x^*\|^2 - \langle t, x^* \rangle < \varepsilon. \end{aligned}$$

Since $(s, 0) \in G(S)$, this establishes the quasidensity of S.

References

1. Alves, M.M., Svaiter, B.F.: A new old class of maximal monotone operators. J. of Convex Anal. **16**, 881–890 (2009)
2. Bauschke, H., Borwein, J.M., Wang, X., Yao, L.: Every maximally monotone operator of Fitzpatrick-Phelps type is actually of dense type. Optim. Lett. **6**, 1875–1881 (2012)
3. Bauschke, H.H., Simons, S.: Stronger maximal monotonicity properties of linear operators. Bull. Austral. Math. Soc. **60**, 163–174 (1999)
4. Brøndsted, A., Rockafellar, R.: On the subdifferentiability of convex functions. Proc. Amer. Math. Soc. **16**, 605–611 (1965)
5. Fitzpatrick, S.: Representing monotone operators by convex functions. In: Functional Analysis and Optimization, vol. 20, pp. 59–65. Austral. Nat. Univ., Canberra (1988)
6. Fitzpatrick, S.P., Phelps, R.R.: Some properties of maximal monotone operators on nonreflexive Banach spaces. Set–Valued Analysis **3**, 51–69 (1995)
7. Galán, M.R., Simons, S.: A new minimax theorem and a perturbed James's theorem. Bull. Austral. Math. Soc. **66**, 43–56 (2002)
8. Gossez, J.P.: Opérateurs monotones non linéaires dans les espaces de banach non réflexifs. J. Math. Anal. Appl. **34**, 371–395 (1971)
9. Pryce, J.D.: Weak compactness in locally convex spaces. Proc. Amer. Math. Soc. **17**, 148–155 (1966)
10. Rockafellar, R.T.: Extension of Fenchel's duality theorem for convex functions. Duke Math. J. **33**, 81–89 (1966)
11. Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pac. J. Math **33**, 209–216 (1970)

12. Rockafellar, R.T.: On the maximal monotonicity of sums of nonlinear monotone operators. Trans. Amer. Math. Soc. **149**, 75–88 (1970)
13. Simons, S.: Subtangents with controlled slope. Nonlinear Analysis **22**, 1373–1389 (1994)
14. Simons, S.: Minimax and monotonicity. Lecture Notes in Mathematics., vol. 1693. Springer–Verlag (1998)
15. Simons, S.: From Hahn–Banach to monotonicity. Lecture Notes in Mathematics., vol. 1693, 2nd edn. Springer–Verlag (2008)
16. Simons, S.: "densities" and maximal monotonicity. J. Convex Anal. **23**, 1017–1050 (2016)
17. Simons, S.: Quasidense monotone multifunctions. Set–Valued Var. Anal. **26**, 5–26 (2018)
18. Simons, S., Wang, X.: Ubiquitous subdifferentials, r_l-density and maximal monotonicity. Set-Valued Var. Anal. **23**, 631–642 (2015)
19. Voisei, M.D., Zălinescu, C.: Strongly–representable monotone operators. J. of Convex Anal. **16**, 1011–1033 (2009)

Chapter 15
On the Acceleration of Forward-Backward Splitting via an Inexact Newton Method

Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos

Abstract We propose a Forward-Backward Truncated-Newton method (FBTN) for minimizing the sum of two convex functions, one of which smooth. Unlike other proximal Newton methods, our approach does not involve the employment of variable metrics, but is rather based on a reformulation of the original problem as the unconstrained minimization of a continuously differentiable function, the *forward-backward envelope (FBE)*. We introduce a generalized Hessian for the FBE that *symmetrizes* the generalized Jacobian of the nonlinear system of equations representing the optimality conditions for the problem. This enables the employment of conjugate gradient method (CG) for efficiently solving the resulting (regularized) linear systems, which can be done inexactly. The employment of CG prevents the computation of full (generalized) Jacobians, as it requires only (generalized) directional derivatives. The resulting algorithm is globally (subsequentially) convergent, Q-linearly under an error bound condition, and up to Q-superlinearly and Q-quadratically under regularity assumptions at the possibly non-isolated limit point.

Keywords Forward-backward splitting · Linear Newton approximation · Truncated-Newton method · Backtracking linesearch · Error bound · Superlinear convergence

AMS 2010 Subject Classification 49J52, 49M15, 90C06, 90C25, 90C30

15.1 Introduction

In this work we focus on convex composite optimization problems of the form

$$\text{minimize}_{x\in\mathbb{R}^n}\ \varphi(x) \equiv f(x) + g(x), \tag{15.1}$$

A. Themelis · M. Ahookhosh · P. Patrinos (✉)
Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Leuven, Belgium
e-mail: andreas.themelis@esat.kuleuven.be; masoud.ahookhosh@esat.kuleuven.be; panos.patrinos@esat.kuleuven.be

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_15

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, twice continuously differentiable and with L_f-Lipschitz-continuous gradient, and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ has a cheaply computable proximal mapping [51]. To ease the notation, throughout the chapter we indicate

$$\varphi_\star \coloneqq \inf \varphi \quad \text{and} \quad \mathcal{X}_\star \coloneqq \operatorname{argmin} \varphi. \tag{15.2}$$

Problems of the form (15.1) are abundant in many scientific areas such as control, signal processing, system identification, machine learning, and image analysis, to name a few. For example, when g is the indicator of a convex set then (15.1) becomes a constrained optimization problem, while for $f(x) = \frac{1}{2}\|Ax - b\|^2$ and $g(x) = \lambda\|x\|_1$ it becomes the ℓ_1-regularized least-squares problem (lasso) which is the main building block of compressed sensing. When g is equal to the nuclear norm, then (15.1) models low-rank matrix recovery problems. Finally, conic optimization problems such as linear, second-order cone, and semidefinite programs can be brought into the form of (15.1), see [31].

Perhaps the most well-known algorithm for problems in the form (15.1) is the forward-backward splitting (FBS) or proximal gradient method [16, 40], that interleaves gradient descent steps on the smooth function and *proximal* steps on the nonsmooth one, see Section 15.3.1. Accelerated versions of FBS, based on the work of Nesterov [5, 54, 77], have also gained popularity. Although these algorithms share favorable global convergence rate estimates of order $O(\varepsilon^{-1})$ or $O(\varepsilon^{-1/2})$ (where ε is the solution accuracy), they are first-order methods and therefore usually effective at computing solutions of low or medium accuracy only. An evident remedy is to include second-order information by replacing the Euclidean norm in the proximal mapping with that induced by the Hessian of f at x or some approximation of it, mimicking Newton or quasi-Newton methods for unconstrained problems [6, 32, 42]. However, a severe limitation of the approach is that, unless Q has a special structure, the computation of the proximal mapping becomes very hard. For example, if φ models a lasso problem, the corresponding subproblem is as hard as the original problem.

In this work we follow a different approach by reformulating the nonsmooth constrained problem (15.1) into the smooth unconstrained minimization of the *forward-backward envelope (FBE)* [57], a real-valued, continuously differentiable, exact penalty function for φ. Although the FBE might fail to be twice continuously differentiable, by using tools from nonsmooth analysis we show that one can design Newton-like methods to address its minimization, that achieve Q-superlinear asymptotic rates of convergence under nondegeneracy and (generalized) smoothness conditions on the set of solutions. Furthermore, by suitably interleaving FBS and Newton-like iterations the proposed algorithm also enjoys good complexity guarantees provided by a global (non-asymptotic) convergence rate. Unlike the approaches of [6, 32], where the corresponding subproblems are expensive to solve, our algorithm only requires the inexact solution of a linear system to compute the Newton-type direction, which can be done efficiently with a memory-free CG method.

Our approach combines and extends ideas stemming from the literature on merit functions for *variational inequalities* (VIs) and *complementarity problems* (CPs), specifically the reformulation of a VI as a constrained continuously differentiable optimization problem via the regularized gap function [23] and as an unconstrained continuously differentiable optimization problem via the D-gap function [79] (see [19, §10] for a survey and [38, 58] for applications to constrained optimization and model predictive control of dynamical systems).

15.1.1 Contributions

We propose an algorithm that addresses problem (15.1) by means of a Newton-like method on the FBE. Differently from a direct application of the classical Newton method, our approach does not require twice differentiability of the FBE (which would impose additional properties on f and g), but merely twice differentiability of f. This is possible thanks to the introduction of an *approximate generalized Hessian* which only requires access to $\nabla^2 f$ and to the generalized (Clarke) Jacobian of the proximal mapping of g, as opposed to third-order derivatives and classical Jacobian, respectively. Moreover, it allows for inexact solutions of linear systems to compute the update direction, which can be done efficiently with a truncated CG method; in particular, no computation of full (generalized) Hessian matrices is necessary, as only (generalized) directional derivatives are needed. The method is thus particularly appealing when the Clarke Jacobians are sparse and/or well structured, so that the implementation of CG becomes extremely efficient. Under an error bound condition and a (semi)smoothness assumption at the limit point, which is not required to be isolated, the algorithm exhibits asymptotic Q-superlinear rates. For the reader's convenience we collect explicit formulas of the needed Jacobians of the proximal mapping for a wide range of frequently encountered functions, and discuss when they satisfy the needed semismoothness requirements that enable superlinear rates.

15.1.2 Related Work

This work is a revised version of the unpublished manuscript [59] and extends ideas proposed in [57], where the FBE is first introduced. Other FBE-based algorithms are proposed in [69, 71, 75]; differently from the truncated-CG type of approximation proposed here, they all employ quasi-Newton directions to mimic second-order information. The underlying ideas can also be extended to enhance other popular proximal splitting algorithms: the Douglas Rachford splitting (DRS) and the alternating direction method of multipliers (ADMM) [74], and for strongly convex problems also the alternating minimization algorithm (AMA) [70].

The algorithm proposed in this chapter adopts the recent techniques investigated in [71, 75] to enhance and greatly simplify the scheme in [59]. In particular, Q-linear and Q-superlinear rates of convergence are established under an error bound condition, as opposed to uniqueness of the solution. The proofs of superlinear convergence with an error bound pattern the arguments in [82, 83], although with less conservative requirements.

15.1.3 Organization

The work is structured as follows. In Section 15.2 we introduce the adopted notation and list some known facts on generalized differentiability needed in the sequel. Section 15.3 offers an overview on the connections between FBS and the proximal point algorithm, and serves as a prelude to Section 15.4 where the forward-backward envelope function is introduced and analyzed. Section 15.5 deals with the proposed truncated-Newton algorithm and its convergence analysis. In Section 15.6 we collect explicit formulas for the generalized Jacobian of the proximal mapping of a rich list of nonsmooth functions, needed for computing the update directions in the proposed algorithm. Finally, Section 15.7 draws some conclusions.

15.2 Preliminaries

15.2.1 Notation and Known Facts

Our notation is standard and follows that of convex analysis textbooks [2, 8, 28, 63]. For the sake of clarity we now properly specify the adopted conventions, and briefly recap known definitions and facts in convex analysis. The interested reader is referred to the above-mentioned textbooks for the details.

Matrices and Vectors The $n \times n$ identity matrix is denoted as I_n, and the $\mathbb{R}^n$ vector with all elements equal to 1 is as $\mathbf{1}_n$; whenever n is clear from context we simply write I or $\mathbf{1}$, respectively. We use the Kronecker symbol $\delta_{i,j}$ for the (i, j)-th entry of I. Given $v \in \mathbb{R}^n$, with $\operatorname{diag} v$ we indicate the $n \times n$ diagonal matrix whose i-th diagonal entry is v_i. With $\mathrm{S}(\mathbb{R}^n)$, $\mathrm{S}_+(\mathbb{R}^n)$, and $\mathrm{S}_{++}(\mathbb{R}^n)$ we denote respectively the set of symmetric, symmetric positive semidefinite, and symmetric positive definite matrices in $\mathbb{R}^{n\times n}$.

The minimum and maximum eigenvalues of $H \in \mathrm{S}(\mathbb{R}^n)$ are denoted as $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$, respectively. For $Q, R \in \mathrm{S}(R^n)$ we write $Q \succeq R$ to indicate that $Q - R \in \mathrm{S}_+(\mathbb{R}^n)$, and similarly $Q \succ R$ indicates that $Q - R \in \mathrm{S}_{++}(\mathbb{R}^n)$. Any matrix $Q \in \mathrm{S}_+(\mathbb{R}^n)$ induces the semi-norm $\|\cdot\|_Q$ on $\mathbb{R}^n$, where $\|x\|_Q^2 \coloneqq \langle x, Qx\rangle$; in case $Q = \mathrm{I}$, that is, for the Euclidean norm, we omit the subscript and simply

write $\|\cdot\|$. No ambiguity occurs in adopting the same notation for the induced matrix norm, namely $\|M\| := \max\{\|Mx\| | x \in \mathbb{R}^n,\ \|x\| = 1|\}$ for $M \in \mathbb{R}^{n\times n}$.

Topology The *convex hull* of a set $E \subseteq \mathbb{R}^n$, denoted as $\operatorname{conv} E$, is the smallest convex set that contains E (the intersection of convex sets is still convex). The *affine hull* $\operatorname{aff} E$ and the *conic hull* $\operatorname{cone} E$ are defined accordingly. Specifically,

$$\begin{aligned}
\operatorname{conv} E &:= \left\{\textstyle\sum_{i=1}^k \alpha_i x_i \mid k \in \mathbb{N},\ x_i \in E,\ \alpha_i \geq 0,\ \sum_{i=1}^k \alpha_i = 1\right\},\\
\operatorname{cone} E &:= \left\{\textstyle\sum_{i=1}^k \alpha_i x_i \mid k \in \mathbb{N},\ x_i \in E,\ \alpha_i \geq 0\right\},\\
\operatorname{aff} E &:= \left\{\textstyle\sum_{i=1}^k \alpha_i x_i \mid k \in \mathbb{N},\ x_i \in E,\ \alpha_i \in \mathbb{R},\ \sum_{i=1}^k \alpha_i = 1\right\}.
\end{aligned}$$

The *closure* and *interior* of E are denoted as $\operatorname{cl} E$ and $\operatorname{int} E$, respectively, whereas its *relative interior*, namely the interior of E as a subspace of $\operatorname{aff} E$, is denoted as $\operatorname{relint} E$. With $\mathrm{B}(x; r)$ and $\overline{\mathrm{B}}(x; r)$ we indicate, respectively, the open and closed balls centered at x with radius r.

Sequences The notation $(a^k)_{k\in K}$ represents a sequence indexed by elements of the set K, and given a set E we write $(a^k)_{k\in K} \subset E$ to indicate that $a^k \in E$ for all indices $k \in K$. We say that $(a^k)_{k\in K} \subset \mathbb{R}^n$ is *summable* if $\sum_{k\in K} \|a^k\|$ is finite, and *square-summable* if $(\|a^k\|^2)_{k\in K}$ is summable. We say that the sequence converges to a point $a \in \mathbb{R}^n$ *superlinearly* if either $a^k = a$ for some $k \in \mathbb{N}$, or $\|a^{k+1}-a\|/\|a^k-a\| \to 0$; if $\|a^{k+1}-a\|/\|a^k-a\|^q$ is bounded for some $q > 1$, then we say that the sequence converges *superlinearly with order* q, and in case $q = 2$ we say that the convergence is *quadratic*.

Extended-Real Valued Functions The extended-real line is $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$. Given a function $h : \mathbb{R}^n \to [-\infty, \infty]$, its *epigraph* is the set

$$\operatorname{epi} h := \left\{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \leq \alpha\right\},$$

while its *domain* is

$$\operatorname{dom} h := \left\{x \in \mathbb{R}^n \mid h(x) < \infty\right\},$$

and for $\alpha \in \mathbb{R}$ its α-*level set* is

$$\operatorname{lev}_{\leq\alpha} h := \left\{x \in \mathbb{R}^n \mid h(x) \leq \alpha\right\}.$$

Function h is said to be *lower semicontinuous (lsc)* if $\operatorname{epi} h$ is a closed set in $\mathbb{R}^{n+1}$ (equivalently, h is said to be *closed*); in particular, all level sets of an lsc function are closed. We say that h is *proper* if $h > -\infty$ and $\operatorname{dom} h \neq \emptyset$, and that it is *level bounded* if for all $\alpha \in \mathbb{R}$ the level set $\operatorname{lev}_{\leq\alpha} h$ is a bounded subset of $\mathbb{R}^n$.

Continuity and Smoothness A function $G : \mathbb{R}^n \to \mathbb{R}^m$ is *ϑ-Hölder continuous* for some $\vartheta > 0$ if there exists $L \geq 0$ such that

$$\|G(x) - G(x')\| \leq L\|x - x'\|^{\vartheta} \tag{15.3}$$

for all x, x'. In case $\vartheta = 1$ we say that G is (L-)Lipschitz continuous. G is *strictly differentiable* at $\bar{x} \in \mathbb{R}^n$ if the Jacobian matrix $JG(\bar{x}) := \left(\frac{\partial G_i}{\partial x_j}(\bar{x})\right)_{i,j}$ exists and

$$\lim_{\substack{x,x' \to \bar{x} \\ x \neq x'}} \frac{\|G(x') - JG(\bar{x})(x' - x) - G(x)\|}{\|x' - x\|} = 0. \tag{15.4}$$

The class of functions $h : \mathbb{R}^n \to \mathbb{R}$ that are k times continuously differentiable is denoted as $C^k(\mathbb{R}^n)$. We write $h \in C^{1,1}(\mathbb{R}^n)$ to indicate that $h \in C^1(\mathbb{R}^n)$ and that ∇h is Lipschitz continuous with modulus L_h. To simplify the terminology, we will say that such an h is *L_h-smooth*. If h is L_h-smooth and convex, then for any $u, v \in \mathbb{R}^n$

$$0 \leq h(v) - \big[h(u) + \langle \nabla h(u), v - u\rangle\big] \leq \tfrac{L_h}{2}\|v - u\|^2. \tag{15.5}$$

Moreover, having $h \in C^{1,1}(\mathbb{R}^n)$ and μ_h-strongly convex is equivalent to having

$$\mu_h\|v - u\|^2 \leq \langle \nabla h(v) - \nabla h(u), v - u\rangle \leq L_h\|v - u\|^2 \tag{15.6}$$

for all $u, v \in \mathbb{R}^n$.

Set-Valued Mappings We use the notation $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ to indicate a point-to-set function $H : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^m)$, where $\mathcal{P}(\mathbb{R}^m)$ is the power set of $\mathbb{R}^m$ (the set of all subsets of $\mathbb{R}^m$). The *graph* of H is the set

$$\operatorname{gph} H := \big\{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in H(x)\big\},$$

while its *domain* is

$$\operatorname{dom} H := \big\{x \in \mathbb{R}^n \mid H(x) \neq \emptyset\big\}.$$

We say that H is *outer semicontinuous (osc)* at $\bar{x} \in \operatorname{dom} H$ if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $H(x) \subseteq H(\bar{x}) + \mathrm{B}(0; \varepsilon)$ for all $x \in \mathrm{B}(\bar{x}; \delta)$. In particular, this implies that whenever $(x^k)_{k\in\mathbb{N}} \subseteq \operatorname{dom} H$ converges to x and $(y^k)_{k\in\mathbb{N}}$ converges to y with $y^k \in H(x^k)$ for all k, it holds that $y \in H(x)$. We say that H is osc (without mention of a point) if H is osc at every point of its domain or, equivalently, if $\operatorname{gph} H$ is a closed subset of $\mathbb{R}^n \times \mathbb{R}^m$ (notice that this notion does *not* reduce to lower semicontinuity for a single-valued function H).

Convex Analysis The *indicator function* of a set $S \subseteq \mathbb{R}^n$ is denoted as $\delta_S : \mathbb{R}^n \to \overline{\mathbb{R}}$, namely

$$\delta_S(x) = \begin{cases} 0 & \text{if } x \in S, \\ \infty & \text{otherwise.} \end{cases} \tag{15.7}$$

If S is nonempty closed and convex, then δ_S is proper convex and lsc, and both the projection $\mathscr{P}_S : \mathbb{R}^n \to \mathbb{R}^n$ and the distance $\operatorname{dist}(\cdot, S) : \mathbb{R}^n \to [0, \infty)$ are well-defined functions, given by $\mathscr{P}_S(x) = \operatorname{argmin}_{z \in S} \|z - x\|$ and $\operatorname{dist}(x, S) = \min_{z \in S} \|z - x\|$, respectively.

The *subdifferential* of h is the set-valued mapping $\partial h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as

$$\partial h(x) := \left\{ v \in \mathbb{R}^n \mid h(z) \geq h(x) + \langle v, z - x \rangle \ \ \forall z \in \mathbb{R}^n \right\}. \tag{15.8}$$

A vector $v \in \partial h(x)$ is called a *subgradient* of h at x. It holds that $\operatorname{dom} \partial h \subseteq \operatorname{dom} h$, and if h is proper and convex, then $\operatorname{dom} \partial h$ is a nonempty convex set containing $\operatorname{relint} \operatorname{dom} h$, and $\partial h(x)$ is convex and closed for all $x \in \mathbb{R}^n$.

A function h is said to be μ-strongly convex for some $\mu \geq 0$ if $h - \frac{\mu}{2} \|\cdot\|^2$ is convex. Unless differently specified, we allow for $\mu = 0$ which corresponds to h being convex but not strongly so. If $\mu > 0$, then h has a unique (global) minimizer.

15.2.2 Generalized Differentiability

Due to its inherent nonsmooth nature, classical notions of differentiability may not be directly applicable in problem (15.1). This subsection contains some definitions and known facts on generalized differentiability that will be needed later on in the chapter. The interested reader is referred to the textbooks [15, 19, 65] for the details.

Definition 15.2.1 (Bouligand and Clarke Subdifferentials) Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous, and let $C_G \subseteq \mathbb{R}^n$ be the set of points at which G is differentiable (in particular $\mathbb{R}^n \setminus C_G$ has measure zero). The B-subdifferential (also known as *Bouligand* or *limiting Jacobian*) of G at $\bar{x}$ is the set-valued mapping $\partial_B G : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ defined as

$$\partial_B G(\bar{x}) := \left\{ H \in \mathbb{R}^{m \times n} \mid \exists (x^k)_{k \in \mathbb{N}} \subset C_G \text{ with } x^k \to \bar{x},\, JG(x^k) \to H \right\}, \tag{15.9}$$

whereas the (*Clarke*) *generalized Jacobian* of G at $\bar{x}$ is $\partial_C G : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m \times n}$ given by

$$\partial_C G(\bar{x}) := \operatorname{conv}(\partial_B G(\bar{x})). \tag{15.10}$$

If $G : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz on $\mathbb{R}^n$, then $\partial_C G(x)$ is a nonempty, convex, and compact subset of $\mathbb{R}^{m \times n}$ matrices, and as a set-valued mapping it is osc at every

$x \in \mathbb{R}^n$. *Semismooth* functions [60] are precisely Lipschitz-continuous mappings for which the generalized Jacobian (and consequently the B-subdifferential) furnishes a first-order approximation.

Definition 15.2.2 (Semismooth Mappings) Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz continuous at $\bar{x}$. We say that G is *semismooth* at $\bar{x}$ if

$$\limsup_{\substack{x \to \bar{x} \\ H \in \partial_C G(x)}} \frac{\|G(x) + H(\bar{x} - x) - G(\bar{x})\|}{\|x - \bar{x}\|} = 0. \tag{15.11a}$$

We say that G is ϑ*-order semismooth* for some $\vartheta > 0$ if the condition can be strengthened to

$$\limsup_{\substack{x \to \bar{x} \\ H \in \partial_C G(x)}} \frac{\|G(x) + H(\bar{x} - x) - G(\bar{x})\|}{\|x - \bar{x}\|^{1+\vartheta}} < \infty, \tag{15.11b}$$

and in case $\vartheta = 1$ we say that G is *strongly semismooth*.

To simplify the notation, we adopt the small-o and big-O convention to write expressions as (15.11a) in the compact form $G(x) + H(\bar{x} - x) - G(\bar{x}) = o(\|x - \bar{x}\|)$, and similarly (15.11b) as $G(x) + H(\bar{x} - x) - G(\bar{x}) = O(\|x - \bar{x}\|^{1+\vartheta})$. We remark that the original definition of semismoothness given by [49] requires G to be directionally differentiable at x. The definition given here is the one employed by [25]. It is also worth remarking that $\partial_C G(x)$ can be replaced with the smaller set $\partial_B G(x)$ in Definition 15.2.2. Fortunately, the class of semismooth mappings is rich enough to include many functions arising in interesting applications. For example, *piecewise smooth (PC^1) mappings* are semismooth everywhere. Recall that a continuous mapping $G : \mathbb{R}^n \to \mathbb{R}^m$ is PC^1 if there exists a finite collection of smooth mappings $G_i : \mathbb{R}^n \to \mathbb{R}^m$, $i = 1, \dots, N$, such that

$$G(x) \in \{G_1(x), \dots, G_N(x)\} \quad \forall x \in \mathbb{R}^n. \tag{15.12}$$

The definition of PC^1 mapping given here is less general than the one of, e.g., [66, §4] but it suffices for our purposes. For every $x \in \mathbb{R}^n$ we introduce the set of essentially active indices

$$I_G^e(x) := \left\{ i \mid x \in \mathrm{cl}\left(\mathrm{int}\{w \mid G(w) = G_i(w)\} \right) \right\}. \tag{15.13}$$

In other words, $I_G^e(x)$ contains only indices of the pieces G_i for which there exists a full-dimensional set on which G agrees with G_i. In accordance with Definition 15.2.1, the generalized Jacobian of G at x is the convex hull of the Jacobians of the essentially active pieces, *i.e.*, [66, Prop. 4.3.1]

$$\partial_C G(x) = \mathrm{conv}\left\{ JG_i(x) \mid i \in I_G^e(x) \right\}. \tag{15.14}$$

The following definition is taken from [19, Def. 7.5.13].

Definition 15.2.3 (Linear Newton Approximation) Let $G : \mathbb{R}^n \to \mathbb{R}^m$ be continuous on $\mathbb{R}^n$. We say that G admits a *linear Newton approximation (LNA)* at $\bar{x} \in \mathbb{R}^n$ if there exists a set-valued mapping $\mathcal{H} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{m\times n}$ that has nonempty compact images, is outer semicontinuous at $\bar{x}$, and

$$\limsup_{\substack{x\to\bar{x}\\ H\in\mathcal{H}(x)}} \frac{\|G(x)+H(\bar{x}-x)-G(\bar{x})\|}{\|x-\bar{x}\|} = 0.$$

If for some $\vartheta > 0$ the condition can be strengthened to

$$\limsup_{\substack{x\to\bar{x}\\ H\in\mathcal{H}(x)}} \frac{\|G(x)+H(\bar{x}-x)-G(\bar{x})\|}{\|x-\bar{x}\|^{1+\vartheta}} < \infty,$$

then we say that $\mathcal{H}$ is a *ϑ-order LNA*, and if $\vartheta = 1$ we say that $\mathcal{H}$ is a *strong LNA*.

Functions G as in Definition 15.2.3 are also referred to as $\mathcal{H}$-semismooth in the literature, see, *e.g.*, [78], however we prefer to stick to the terminology of [19] and rather say that $\mathcal{H}$ is a LNA for G. Arguably the most notable example of a LNA for semismooth mappings is the generalized Jacobian, cf. Definition 15.2.1. However, semismooth mappings can admit LNAs different from the generalized Jacobian. More importantly, mappings that are not semismooth may also admit a LNA.

Lemma 15.2.4 ([19, Prop. 7.4.10]) *Let $h \in C^1(\mathbb{R}^n)$ and suppose that $\mathcal{H} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ is a LNA for ∇h at $\bar{x}$. Then,*

$$\lim_{\substack{x\to\bar{x}\\ H\in\mathcal{H}(x)}} \frac{h(x)-h(\bar{x})-\langle\nabla h(\bar{x}),x-\bar{x}\rangle-\frac{1}{2}\langle H(x-\bar{x}),x-\bar{x}\rangle}{\|x-\bar{x}\|^2} = 0. \tag{15.15}$$

We remark that although [19, Prop. 7.4.10] assumes semismoothness of ∇h at $\bar{x}$ and uses $\partial_C(\nabla h)$ in place of $\mathcal{H}$; however, exactly the same arguments apply for any LNA of ∇h at $\bar{x}$ even without the semismoothness assumption.

15.3 Proximal Algorithms

15.3.1 Proximal Point and Moreau Envelope

The *proximal mapping* of a proper closed and convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ with parameter $\gamma > 0$ is $\mathrm{prox}_{\gamma h} : \mathbb{R}^n \to \mathbb{R}^n$, given by

$$\mathrm{prox}_{\gamma h}(x) := \mathrm{argmin}_{w\in\mathbb{R}^n} \overbrace{\left\{h(w)+\tfrac{1}{2\gamma}\|w-x\|^2\right\}}^{\mathcal{M}^h_\gamma(w;x)}. \tag{15.16}$$

The *majorization model* $\mathcal{M}_\gamma^h(x;\,\cdot\,)$ is a proper and strongly convex function, and therefore has a unique minimizer. The value function, as opposed to the minimizer, defines the *Moreau envelope* $h^\gamma : \mathbb{R}^n \to \mathbb{R}$, namely

$$h^\gamma(x) \coloneqq \min_{w\in\mathbb{R}^n} \left\{ h(w) + \tfrac{1}{2\gamma}\|w - x\|^2 \right\}, \tag{15.17}$$

which is real valued and Lipschitz differentiable, despite the fact that h might be extended-real valued. Properties of the Moreau envelope and the proximal mapping are well documented in the literature, see, *e.g.*, [2, §24]. For example, $\operatorname{prox}_{\gamma h}$ is nonexpansive (Lipschitz continuous with modulus 1) and is characterized by the implicit inclusion

$$\hat{x} = \operatorname{prox}_{\gamma h}(x) \quad \Leftrightarrow \quad \tfrac{1}{\gamma}(x - \hat{x}) \in \partial h(\hat{x}). \tag{15.18}$$

For the sake of a brief recap, we now list some other important known properties. Theorem 15.3.1 provides some relations between h and its Moreau envelope h^γ, which we informally refer to as *sandwich property* for apparent reasons, cf. Figure 15.1. Theorem 15.3.2 highlights that the minimization of a (proper, lsc and) convex function can be expressed as the convex smooth minimization of its Moreau envelope.

Theorem 15.3.1 (Moreau Envelope: Sandwich Property [2, 12]) *For all $\gamma > 0$ the following hold for the cost function φ:*

(i) $\varphi^\gamma(x) \le \varphi(x) - \frac{1}{2\gamma}\|x - \hat{x}\|^2$ *for all* $x \in \mathbb{R}^n$ *where* $\hat{x} \coloneqq \operatorname{prox}_{\gamma\varphi}(x)$;
(ii) $\varphi(\hat{x}) = \varphi^\gamma(x) - \frac{1}{2\gamma}\|x - \hat{x}\|^2$ *for all* $x \in \mathbb{R}^n$ *where* $\hat{x} \coloneqq \operatorname{prox}_{\gamma\varphi}(x)$;
(iii) $\varphi^\gamma(x) = \varphi(x)$ *iff* $x \in \operatorname{argmin}\varphi$.

Proof

- 15.3.1*(i)*. This fact is shown in [12, Lem. 3.2] for a more general notion of proximal point operator; namely, the square Euclidean norm appearing in (15.16) and (15.17) can be replaced by arbitrary Bregman divergences. In this simpler case, since $\frac{1}{\gamma}(x - \hat{x})$ is a subgradient of φ at $\hat{x}$, cf. (15.18), we have

$$\varphi(x) \ge \varphi(\hat{x}) + \langle \tfrac{1}{\gamma}(x - \hat{x}), x - \hat{x}\rangle = \varphi(\hat{x}) + \tfrac{1}{\gamma}\|x - \hat{x}\|^2. \tag{15.19}$$

  The claim now follows by subtracting $\frac{1}{2\gamma}\|x - \hat{x}\|^2$ from both sides.
- 15.3.1*(ii)*. Follows by definition, cf. (15.16) and (15.17).
- 15.3.1*(iii)*. See [2, Prop. 17.5]. □

□

Theorem 15.3.2 (Moreau Envelope: Convex Smooth Minimization Equivalence [2]) *For all $\gamma > 0$ the following hold for the cost function φ:*

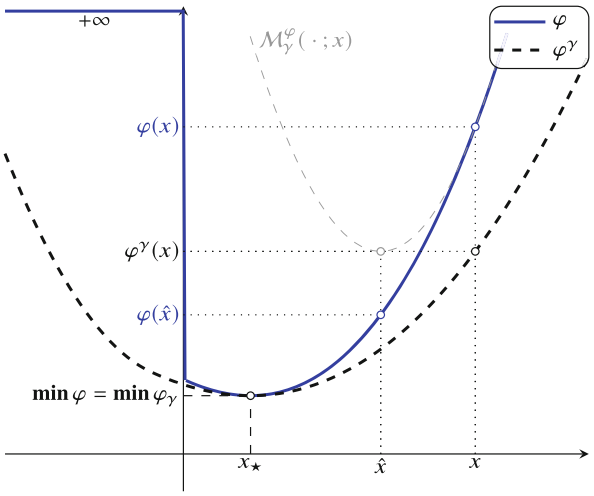


Fig. 15.1 *Moreau envelope of the function* $\varphi(x) = \frac{1}{3}x^3 + x^2 - x + 1 + \delta_{[0,\infty)}(x)$ *with parameter* $\gamma = 0.2$.
At each point x, the Moreau envelope φ^γ is the minimum of the quadratic majorization model $\mathcal{M}_\gamma^\varphi = \varphi + \frac{1}{2\gamma}(\,\cdot - x)^2$, the unique minimizer being, by definition, the proximal point $\hat{x} := \operatorname{prox}_{\gamma\varphi}(x)$. It is a convex smooth lower bound to φ, despite the fact that φ might be extended-real valued. Function φ and its Moreau envelope φ^γ have same inf and argmin; in fact, the two functions agree (only) on the set of minimizers. In general, φ^γ is *sandwiched* as $\varphi \circ \operatorname{prox}_{\gamma\varphi} \le \varphi^\gamma \le \varphi$

(i) φ^γ *is convex and smooth with* $L_{\varphi^\gamma} = \gamma^{-1}$ *and* $\nabla\varphi^\gamma(x) = \gamma^{-1}\big(x - \operatorname{prox}_{\gamma\varphi}(x)\big)$*;*
(ii) $\inf\varphi = \inf\varphi^\gamma$*;*
(iii) $x_\star \in \operatorname{argmin}\varphi$ *iff* $x_\star \in \operatorname{argmin}\varphi^\gamma$ *iff* $\nabla\varphi^\gamma(x_\star) = 0$.

Proof

- 15.3.2*(i)*. See [2, Prop.s 12.15 and 12.30].
- 15.3.2*(ii)*. See [2, Prop. 12.9(iii)].
- 15.3.2*(iii)*. See [2, Prop. 17.5]. □

□

As a consequence of Theorem 15.3.2, one can address the minimization of the convex but possibly nonsmooth and extended-real-valued function φ by means of gradient descent on the smooth envelope function φ^γ with stepsize $0 < \tau < 2/L_{\varphi^\gamma} = 2\gamma$. As first noticed by Rockafellar [64], this simply amounts to (relaxed) fixed-point iterations of the proximal point operator, namely

$$x^+ = (1-\lambda)x + \lambda \operatorname{prox}_{\gamma\varphi}(x), \tag{15.20}$$

where $\lambda = \tau/\gamma \in (0, 2)$ is a possible relaxation parameter. The scheme, known as *proximal point algorithm* (PPA) and first introduced by Martinet [45], is well covered by the broad theory of monotone operators, where convergence properties can be easily derived with simple tools of Fejérian monotonicity, see, *e.g.*, [2, Thm.s 23.41 and 27.1]. Nevertheless, not only does the interpretation as gradient method provide a beautiful theoretical link, but it also enables the employment of acceleration techniques exclusively stemming from smooth unconstrained optimization, such as Nesterov's extrapolation [26] or quasi-Newton schemes [13], see also [7] for extensions to the dual formulation.

15.3.2 Forward-Backward Splitting

While it is true that every convex minimization problem can be smoothened by means of the Moreau envelope, unfortunately it is often the case that the computation of the proximal operator (which is needed to evaluate the envelope) is as hard as solving the original problem. For instance, evaluating the Moreau envelope of the cost of modeling a convex QP at one point amounts to solving another QP with same constraints and augmented cost. To overcome this limitation there comes the idea of *splitting schemes*, which decompose a complex problem in small components which are easier to operate onto. A popular such scheme is the *forward-backward splitting* (FBS), which addresses minimization problems of the form (15.1).

Given a point $x \in \mathbb{R}^n$, one iteration of *forward-backward splitting* (FBS) for problem (15.1) with stepsize $\gamma > 0$ and relaxation $\lambda > 0$ consists in

$$x^+ = (1 - \lambda)x + \lambda T_\gamma(x), \tag{15.21}$$

where

$$T_\gamma(x) := \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \tag{15.22}$$

is the *forward-backward operator*, characterized as

$$\bar{x} = T_\gamma(x) \quad \Leftrightarrow \quad \tfrac{1}{\gamma}(x - \bar{x}) - (\nabla f(x) - \nabla f(\bar{x})) \in \partial\varphi(\bar{x}), \tag{15.23}$$

as it follows from (15.18). FBS interleaves a gradient descent step on f and a proximal point step on g, and as such it is also known as *proximal gradient method*. If both f and g are (lsc, proper and) convex, then the solutions to (15.1) are exactly the fixed points of the forward-backward operator T_γ. In other words,

$$x_\star \in \operatorname{argmin} \varphi \quad \text{iff} \quad R_\gamma(x_\star) = 0, \tag{15.24}$$

where

$$R_\gamma(x) \coloneqq \tfrac{1}{\gamma}\big(x - \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x))\big) \tag{15.25}$$

is the *forward-backward residual*.[1] FBS iterations (15.21) are well known to converge to a solution to (15.1) provided that f is smooth and that the parameters are chosen as $\gamma \in (0, 2/L_f)$ and $\lambda \in (0, 2 - \gamma L_f/2)$ [2, Cor. 28.9] ($\lambda = 1$, which is always feasible, is the typical choice).

15.3.3 Error Bounds and Quadratic Growth

We conclude the section with some inequalities that will be useful in the sequel.

Lemma 15.3.3 *Suppose that $\mathfrak{X}_\star$ is nonempty. Then,*

$$\varphi(x) - \varphi_\star \leq \operatorname{dist}(0, \partial\varphi(x)) \operatorname{dist}(x, \mathfrak{X}_\star) \quad \forall x \in \mathbb{R}^n. \tag{15.26}$$

Proof From the subgradient inequality it follows that for all $x_\star \in \mathfrak{X}_\star$ and $v \in \partial\varphi(x)$ we have

$$\varphi(x) - \varphi_\star = \varphi(x) - \varphi(x_\star) \leq \langle v, x - x_\star \rangle \leq \|v\| \|x - x_\star\| \tag{15.27}$$

and the claimed inequality follows from the arbitrarity of $x_\star$ and v. □

Lemma 15.3.4 *Suppose that $\mathfrak{X}_\star$ is nonempty. For all $x \in \mathbb{R}^n$ and $\gamma > 0$ the following holds:*

$$\|R_\gamma(x)\| \geq \tfrac{1}{1+\gamma L_f} \operatorname{dist}\big(0, \partial\varphi(T_\gamma(x))\big) \tag{15.28}$$

Proof Let $\bar{x} \coloneqq T_\gamma(x)$. The characterization (15.23) of T_γ implies that

$$\|R_\gamma(x)\| \geq \operatorname{dist}\big(0, \partial\varphi(\bar{x})\big) - \|\nabla f(x) - \nabla f(\bar{x})\| \geq \operatorname{dist}\big(0, \partial\varphi(\bar{x})\big) - \gamma L_f \|R_\gamma(x)\|. \tag{15.29}$$

After trivial rearrangements the sought inequality follows. □

Further interesting inequalities can be derived if the cost function φ satisfies an *error bound*, which can be regarded as a generalization of strong convexity that does not require uniqueness of the minimizer. The interested reader is referred to [3, 17, 43, 55] and the references therein for extensive discussions.

[1] Due to apparent similarities with gradient descent iterations, having $x^+ = x - \gamma R_\gamma(x)$ in FBS, R_γ is also referred to as (generalized) *gradient mapping*, see, *e.g.*, [17]. In particular, if $g = 0$, then $R_\gamma = \nabla f$ whereas if $f = 0$ then $R_\gamma = \nabla g^\gamma$. The analogy will be supported by further evidence in the next section where we will see that, up to a change of metric, indeed R_γ is the gradient of the *forward-backward envelope function*.

Definition 15.3.5 (Quadratic Growth and Error Bound) Suppose that $\mathcal{X}_\star \neq \emptyset$. Given $\mu, \nu > 0$, we say that

(a) φ satisfies the *quadratic growth* with constants (μ, ν) if

$$\varphi(x) - \varphi_\star \geq \tfrac{\mu}{2} \operatorname{dist}(x, \mathcal{X}_\star)^2 \quad \forall x \in \operatorname{lev}_{\leq \varphi_\star + \nu} \varphi; \tag{15.30}$$

(b) φ satisfies the *error bound* with constants (μ, ν) if

$$\operatorname{dist}(0, \partial\varphi(x)) \geq \tfrac{\mu}{2} \operatorname{dist}(x, \mathcal{X}_\star) \quad \forall x \in \operatorname{lev}_{\leq \varphi_\star + \nu} \varphi. \tag{15.31}$$

In case $\nu = \infty$ we say that the properties are satisfied *globally*.

Theorem 15.3.6 ([17, Thm. 3.3]) *For a proper convex and lsc function, the quadratic growth with constants (μ, ν) is equivalent to the error bound with same constants.*

Lemma 15.3.7 (Globality of Quadratic Growth) *Suppose that φ satisfies the quadratic growth with constants (μ, ν). Then, for every $\nu' > \nu$ it satisfies the quadratic growth with constants (μ', ν'), where*

$$\mu' \coloneqq \tfrac{\mu}{2} \min\left\{1, \tfrac{\nu}{\nu' - \nu}\right\}. \tag{15.32}$$

Proof Let $\nu' > \nu$ be fixed, and let $x \in \operatorname{lev}_{\leq \varphi_\star + \nu'}$ be arbitrary. Since $\mu' \leq \mu$, the claim is trivial if $\varphi(x) \leq \varphi_\star + \nu$; we may thus suppose that $\varphi(x) > \varphi_\star + \nu$. Let z be the projection of x onto the (nonempty closed and convex) level set $\operatorname{lev}_{\leq \varphi_\star + \nu}$, and observe that $\varphi(z) = \varphi_\star + \nu$. With Lemma 15.3.3 and Theorem 15.3.6 we can upper bound ν as

$$\nu = \varphi(z) - \varphi_\star \leq \operatorname{dist}(0, \partial\varphi(z)) \operatorname{dist}(z, \mathcal{X}_\star) \leq \tfrac{2}{\mu} \operatorname{dist}(0, \partial\varphi(z))^2. \tag{15.33}$$

Moreover, it follows from [28, Thm. 1.3.5] that there exists a subgradient $v \in \partial\varphi(z)$ such that $\langle v, x - z\rangle = \|v\|\|x - z\|$. Then,

$$\begin{aligned}\varphi(x) &\geq \varphi(z) + \langle v, x - z\rangle = \varphi(z) + \|v\|\|x - z\| \geq \varphi(z) + \operatorname{dist}(0, \partial\varphi(z))\|x - z\| \\ &\overset{(15.33)}{\geq} \varphi(z) + \sqrt{\tfrac{\mu\nu}{2}}\|x - z\|.\end{aligned} \tag{15.34}$$

By subtracting $\varphi(z)$ from the first and last terms we obtain

$$\|x - z\| \leq \sqrt{\tfrac{2}{\mu\nu}}\big(\varphi(x) - \varphi(z)\big) \leq \sqrt{\tfrac{2}{\mu\nu}}(\nu' - \nu), \tag{15.35}$$

which implies

$$\|x - z\| \geq \sqrt{\tfrac{\mu\nu}{2}} \tfrac{1}{\nu' - \nu}\|x - z\|^2. \tag{15.36}$$

Thus,

$$\varphi(x) - \varphi_\star \overset{(15.34)}{\geq} \varphi(z) - \varphi_\star + \sqrt{\tfrac{\mu\nu}{2}}\|x - z\|$$

using the quadratic growth at z and the inequality (15.36)

$$\begin{aligned} &\geq \tfrac{\mu}{2}\operatorname{dist}(z, \mathcal{X}_\star)^2 + \tfrac{\mu\nu}{2(\nu'-\nu)}\|x - z\|^2 \\ &\geq \tfrac{\mu}{2}\min\left\{1, \tfrac{\nu}{\nu'-\nu}\right\}\left[\operatorname{dist}(z, \mathcal{X}_\star)^2 + \|x - z\|^2\right]. \end{aligned}$$

By using the fact that $a^2 + b^2 \geq \frac{1}{2}(a+b)^2$ for any $a, b \in \mathbb{R}$ together with the triangular inequality $\operatorname{dist}(x, \mathcal{X}_\star) \leq \|x - z\| + \operatorname{dist}(z, \mathcal{X}_\star)$, we conclude that $\varphi(x) - \varphi_\star \geq \frac{\mu'}{2}\operatorname{dist}(x, \mathcal{X}_\star)^2$, with μ' as in the statement. Since μ' depends only on μ, ν, and ν', from the arbitrarity of $x \in \operatorname{lev}_{\leq \varphi_\star + \nu'}$ the claim follows. □

Theorem 15.3.8 ([17, Cor. 3.6]) *Suppose that φ satisfies the quadratic growth with constants (μ, ν). Then, for all $\gamma \in (0, 1/L_f)$ and $x \in \operatorname{lev}_{\leq \varphi_\star + \nu} \varphi$ we have*

$$\operatorname{dist}(x, \mathcal{X}_\star) \leq (\gamma + 2/\mu)(1 + \gamma L_f)\|R_\gamma(x)\|. \tag{15.37}$$

15.4 Forward-Backward Envelope

There are clearly infinite ways of representing the (proper, lsc and) convex function φ in (15.1) as the sum of two convex functions f and g with f smooth, and each of these choices leads to a different FBS operator T_γ. If $f = 0$, for instance, then T_γ reduces to $\operatorname{prox}_{\gamma\varphi}$, and consequently FBS (15.21) to the PPA (15.20). A natural question then arises, whether a function exists that serves as "envelope" for FBS in the same way that φ_γ does for $\operatorname{prox}_{\gamma\varphi}$. We will now provide a positive answer to this question by reformulating the nonsmooth problem (15.1) as the minimization of a differentiable function. To this end, the following requirements on f and g will be assumed throughout the chapter without further mention.

Assumption I (Basic Requirements) *In problem (15.1),*

(i) $f : \mathbb{R}^n \to \mathbb{R}$ *is convex, twice continuously differentiable and L_f-smooth;*
(ii) $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *is lsc, proper, and convex.*

Compared to the classical FBS assumptions, the only additional requirement is twice differentiability of f. This ensures that the *forward operator* $x \mapsto x - \gamma\nabla f(x)$ is differentiable; we denote its Jacobian as $Q_\gamma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$, namely

$$Q_\gamma(x) \coloneqq \mathrm{I} - \gamma\nabla^2 f(x). \tag{15.38}$$

Notice that, due to the bound $\nabla^2 f(x) \preceq L_f \mathrm{I}$ (which follows from L_f-smoothness of f, see [53, Lem. 1.2.2]) $Q_\gamma(x)$ is invertible (in fact, positive definite) whenever $\gamma < 1/L_f$. Moreover, due to the chain rule and Theorem 15.3.2*(i)* we have that

$$\begin{aligned}\nabla\big[g^\gamma \circ (\mathrm{id} - \gamma\nabla f)\big](x) &= \gamma^{-1} Q_\gamma(x)\big[x - \gamma\nabla f(x) - \mathrm{prox}_{\gamma g}\left(x - \gamma\nabla f(x)\right)\big] \\ &= Q_\gamma(x)\big[R_\gamma(x) - \nabla f(x)\big].\end{aligned}$$

Rearranging,

$$\begin{aligned}Q_\gamma(x)R_\gamma(x) &= \nabla f(x) - \gamma\nabla^2 f(x)\nabla f(x) + \nabla\big[g^\gamma \circ (\mathrm{id} - \gamma\nabla f)\big](x) \\ &= \nabla f(x) - \nabla\left[\tfrac{\gamma}{2}\|\nabla f\|^2\right](x) + \nabla\big[g^\gamma \circ (\mathrm{id} - \gamma\nabla f)\big](x) \\ &= \nabla\left[f - \tfrac{\gamma}{2}\|\nabla f\|^2 + g^\gamma \circ (\mathrm{id} - \gamma\nabla f)\right](x)\end{aligned}$$

we obtain the gradient of a real-valued function, which we define as follows.

Definition 15.4.1 (Forward-Backward Envelope) The *forward-backward envelope* (FBE) for the composite minimization problem (15.1) is the function $\varphi_\gamma : \mathbb{R}^n \to \mathbb{R}$ defined as

$$\varphi_\gamma(x) \coloneqq f(x) - \tfrac{\gamma}{2}\|\nabla f(x)\|^2 + g^\gamma(x - \gamma\nabla f(x)). \tag{15.39}$$

In the next section we discuss some of the favorable properties enjoyed by the FBE.

15.4.1 Basic Properties

We already verified that the FBE is differentiable with gradient

$$\nabla\varphi_\gamma(x) = Q_\gamma(x)R_\gamma(x). \tag{15.40}$$

In particular, for $\gamma < 1/L_f$ one obtains that a FBS step is a (scaled) gradient descent step on the FBE, similarly as the relation between Moreau envelope and PPA; namely,

$$T_\gamma(x) = x - \gamma\, Q_\gamma(x)^{-1} \nabla \varphi_\gamma(x). \tag{15.41}$$

To take the analysis of the FBE one step further, let us consider the equivalent expression of the operator T_γ as

$$T_\gamma(x) = \operatorname{argmin}_{w\in\mathbb{R}^n} \Big\{ \overbrace{f(x) + \langle \nabla f(x), w - x\rangle + \tfrac{1}{2\gamma}\|w-x\|^2 + g(w)}^{\mathcal{M}_\gamma^{f,g}(w;x)} \Big\}. \tag{15.42}$$

Differently from the quadratic model $\mathcal{M}_\gamma^{\varphi}$ in (15.16), $\mathcal{M}_\gamma^{f,g}$ replaces the differentiable component f with a linear approximation. Building upon the idea of the Moreau envelope, instead of the minimizer $T_\gamma(x)$ we consider the value attained in the subproblem (15.42), and with simple algebra one can easily verify that this gives rise once again to the FBE:

$$\varphi_\gamma(x) = \min_{w\in\mathbb{R}^n} \Big\{ f(x) + \langle \nabla f(x), w - x\rangle + \tfrac{1}{2\gamma}\|w-x\|^2 + g(w) \Big\}. \tag{15.43}$$

Starting from this expression we can easily mirror the properties of the Moreau envelope stated in Theorems 15.3.1 and 15.3.2. These results appeared in the independent works [54] and [57], although the former makes no mention of an "envelope" function and simply analyzes the *majorization-minimization* model $\mathcal{M}_\gamma^{f,g}$.

Theorem 15.4.2 (FBE: Sandwich Property) *Let $\gamma > 0$ and $x \in \mathbb{R}^n$ be fixed, and denote $\bar{x} = T_\gamma(x)$. The following hold:*

(i) $\varphi_\gamma(x) \le \varphi(x) - \frac{1}{2\gamma}\|x-\bar{x}\|^2$;

(ii) $\varphi_\gamma(x) - \frac{1}{2\gamma}\|x-\bar{x}\|^2 \le \varphi(\bar{x}) \le \varphi_\gamma(x) - \frac{1-\gamma L_f}{2\gamma}\|x-\bar{x}\|^2$.

In particular,

(iii) $\varphi_\gamma(x_\star) = \varphi(x_\star)$ *iff* $x_\star \in \operatorname{argmin}\varphi$.

In fact, the assumption of twice continuous differentiability of f can be dropped.

Proof

- 15.4.2*(i)* Since the minimum in (15.43) is attained at $w = \bar{x}$, cf. (15.42), we have

$$\varphi_\gamma(x) = f(x) + \langle \nabla f(x), \bar{x} - x\rangle + \tfrac{1}{2\gamma}\|\bar{x}-x\|^2 + g(\bar{x}) \tag{15.44}$$

$$\begin{aligned} &\leq f(x) + \langle \nabla f(x), \bar{x} - x \rangle + \tfrac{1}{2\gamma}\|\bar{x} - x\|^2 + g(x) \\ &\quad + \langle \tfrac{1}{\gamma}(x - \bar{x}) - \nabla f(x), \bar{x} - x \rangle \\ &= f(x) + g(x) - \tfrac{1}{2\gamma}\|x - \bar{x}\|^2 \end{aligned}$$

where in the inequality we used the fact that $\frac{1}{\gamma}(x - \bar{x}) - \nabla f(x) \in \partial g(\bar{x})$, cf. (15.23).

- 15.4.2*(ii)* Follows by using (15.5) (with $h = f$, $u = x$ and $v = \bar{x}$) in (15.44).
- 15.4.2*(iii)* Follows by 15.4.2*(i)* and the optimality condition (15.24). □

□

Notice that by combining Theorems 15.4.2*(i)* and 15.4.2*(ii)* we recover the "sufficient decrease" condition of (convex) FBS [54, Thm. 1], that is

$$\varphi(\bar{x}) \leq \varphi(x) - \tfrac{2-\gamma L_f}{2\gamma}\|x - \bar{x}\|^2 \tag{15.45}$$

holding for all $x \in \mathbb{R}^n$ with $\bar{x} = T_\gamma(x)$.

Theorem 15.4.3 (FBE: Smooth Minimization Equivalence) *For all $\gamma > 0$*

(i) $\varphi_\gamma \in C^1(\mathbb{R}^n)$ with $\nabla\varphi_\gamma = Q_\gamma R_\gamma$.

Moreover, if $\gamma \in (0, 1/L_f)$ then the following also hold:

(ii) $\inf\varphi = \inf\varphi_\gamma$;
(iii) $x_\star \in \operatorname{argmin}\varphi$ *iff* $x_\star \in \operatorname{argmin}\varphi_\gamma$ *iff* $\nabla\varphi_\gamma(x_\star) = 0$.

Proof

- 15.4.3*(i)*. Since $f \in C^2(\mathbb{R}^n)$ and $g^\gamma \in C^1(\mathbb{R}^n)$ (cf. Theorem 15.3.2*(i)*), from the definition (15.39) it is apparent that φ_γ is continuously differentiable for all $\gamma > 0$. The formula for the gradient was already shown in (15.40).

Suppose now that $\gamma < 1/L_f$.

- 15.4.3*(ii)*. $\inf\varphi \leq \inf_{x\in\mathbb{R}^n} \varphi(T_\gamma(x)) \overset{15.4.2(ii)}{\leq} \inf_{x\in\mathbb{R}^n} \varphi_\gamma(x) = \inf\varphi_\gamma \overset{15.4.2(i)}{\leq} \inf\varphi$.
- 15.4.3*(iii)*. We have

$$x_\star \in \operatorname{argmin}\varphi \overset{(15.24)}{\Leftrightarrow} R_\gamma(x_\star) = 0 \;\Leftrightarrow\; Q_\gamma(x_\star)R_\gamma(x_\star) = 0 \overset{15.4.3(i)}{\Leftrightarrow} \nabla\varphi_\gamma(x_\star) = 0, \tag{15.46}$$

where the second equivalence follows from the invertibility of Q_γ.
Suppose now that $x_\star \in \operatorname{argmin}\varphi_\gamma$. Since $\varphi_\gamma \in C^1(\mathbb{R}^n)$ the first-order necessary condition reads $\nabla\varphi_\gamma = 0$, and from the equivalence proven above we conclude that $\operatorname{argmin}\varphi_\gamma \subseteq \operatorname{argmin}\varphi$. Conversely, if $x_\star \in \operatorname{argmin}\varphi$, then

$$\varphi_\gamma(x_\star) \overset{15.4.2(iii)}{=} \varphi(x_\star) = \inf\varphi \overset{15.4.3(ii)}{=} \inf\varphi_\gamma, \tag{15.47}$$

proving $x_\star \in \operatorname{argmin}\varphi_\gamma$, hence the inclusion $\operatorname{argmin}\varphi_\gamma \supseteq \operatorname{argmin}\varphi$. □

□

Proposition 15.4.4 (FBE and Moreau Envelope [54, Thm. 2]) *For any* $\gamma \in (0, 1/L_f)$*, it holds that*

$$\varphi^{\frac{\gamma}{1-\gamma L_f}} \leq \varphi_\gamma \leq \varphi^\gamma. \tag{15.48}$$

Proof We have

$$\begin{aligned}
\varphi_\gamma(x) &= \min_{w\in\mathbb{R}^n} \left\{ \underbrace{f(x) + \langle \nabla f(x), w - x\rangle} + \tfrac{1}{2\gamma}\|w - x\|^2 + g(w) \right\} \\
&\overset{(15.5)}{\leq} \min_{w\in\mathbb{R}^n} \left\{ \overbrace{f(w) - \tfrac{L_f}{2}\|w - x\|^2} + \tfrac{1}{2\gamma}\|w - x\|^2 + g(w) \right\} \\
&= \min_{w\in\mathbb{R}^n} \left\{ f(w) + g(w) + \tfrac{1-\gamma L_f}{2\gamma}\|w - x\|^2 \right\} = \varphi^{\frac{\gamma}{1-\gamma L_f}}(x).
\end{aligned}$$

Using the upper bound in (15.5) instead yields the other inequality. □

Since φ_γ is upper bounded by the γ^{-1}-smooth function φ^γ with which it shares the set of minimizers $\mathcal{X}_\star$, from (15.5) we easily infer the following quadratic upper bound.

Corollary 15.4.5 (Global Quadratic Upper Bound) *If* $\mathcal{X}_\star \neq \emptyset$*, then*

$$\varphi_\gamma(x) - \varphi_\star \leq \tfrac{1}{2\gamma}\operatorname{dist}(x, \mathcal{X}_\star)^2 \quad \forall x \in \mathbb{R}^n. \tag{15.49}$$

Although the FBE may fail to be convex, for $\gamma < 1/L_f$ its stationary points and minimizers coincide and are the same as those of the original function φ (Figure 15.2). That is, the minimization of φ is equivalent to the minimization of the differentiable function φ_γ. This is a clear analogy with the Moreau envelope, which in fact is the special case of the FBE corresponding to $f \equiv 0$ in the decomposition of φ. In the next result we tighten the claims of Theorem 15.4.3*(i)* when f is a convex quadratic function, showing that in this case the FBE is convex and smooth and thus recover all the properties of the Moreau envelope.

Theorem 15.4.6 (FBE: Convexity & Smoothness for Quadratic f [24, Prop. 4.4]) *Suppose that f is convex quadratic, namely $f(x) = \frac{1}{2}\langle x, Hx\rangle + \langle h, x\rangle$ for some $H \in \mathrm{S}_+(\mathbb{R}^n)$ and $h \in \mathbb{R}^n$. Then, for all $\gamma \in (0, 1/L_f]$ the FBE φ_γ is convex and smooth, with*

$$L_{\varphi_\gamma} = \tfrac{1-\gamma\mu_f}{\gamma} \quad \text{and} \quad \mu_{\varphi_\gamma} = \min\left\{\mu_f(1-\gamma\mu_f), L_f(1-\gamma L_f)\right\}, \tag{15.50}$$

where $L_f = \lambda_{\max}(H)$ and $\mu_f = \lambda_{\min}(H)$. In particular, when f is μ_f-strongly convex the strong convexity of φ_γ is maximized for $\gamma = \frac{1}{\mu_f + L_f}$, in which case

$$L_{\varphi_\gamma} = L_f \quad \text{and} \quad \mu_{\varphi_\gamma} = \tfrac{L_f\mu_f}{\mu_f + L_f}. \tag{15.51}$$

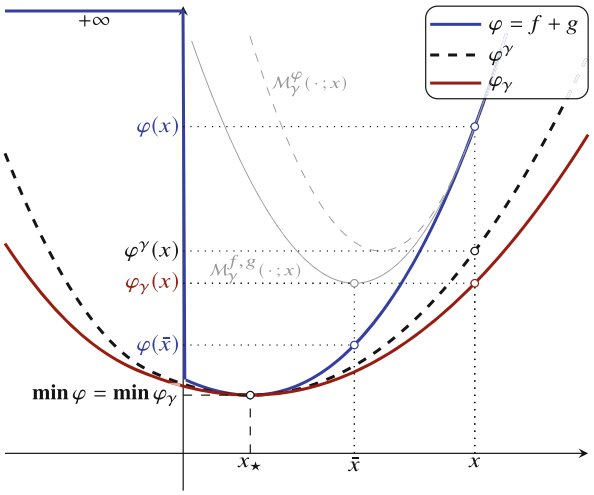


Fig. 15.2 *FBE of the function φ as in Figure 15.1 with same parameter $\gamma = 0.2$, relative to the decomposition as the sum of $f(x) = x^2 + x - 1$ and $g(x) = \frac{1}{3}x^3 + \delta_{[0,\infty)}(x)$.*
For $\gamma < 1/L_f$ ($L_f = 2$ in this example) at each point x the FBE φ_γ is the minimum of the quadratic majorization model $\mathcal{M}_\gamma^{f,g}(\cdot, x)$ for φ, the unique minimizer being the proximal gradient point $\bar{x} = T_\gamma(x)$. The FBE is a differentiable lower bound to φ and since f is quadratic in this example, it is also smooth and convex (cf. Theorem 15.4.6). In any case, its stationary points and minimizers coincide, and are equivalent to the minimizers of φ

Proof Letting $Q \coloneqq \mathrm{I} - \gamma H$, we have that $Q_\gamma \equiv Q$ and $x - \gamma\nabla f(x) = Qx - \gamma h$. Therefore,

$$\begin{aligned}\gamma\langle \nabla\varphi_\gamma(x) - \nabla\varphi_\gamma(y), x - y\rangle &\overset{(15.40)}{=} \langle Q(R_\gamma(x) - R_\gamma(y)), x - y\rangle \\ &= \langle Q(x-y), x-y\rangle - \langle Q(T_\gamma(x) - T_\gamma(y)), x - y\rangle \\ &= \|x - y\|_Q^2 \\ &\quad - \langle \operatorname{prox}_{\gamma g}(Qx - \gamma h) \\ &\quad - \operatorname{prox}_{\gamma g}(Qy - \gamma h), Q(x - y)\rangle.\end{aligned}$$

From the *firm* nonexpansiveness of $\operatorname{prox}_{\gamma g}$ (see [2, Prop.s 4.35(iii) and 12.28]) it follows that

$$0 \le \langle \operatorname{prox}_{\gamma g}(Qx-\gamma h) - \operatorname{prox}_{\gamma g}(Qy-\gamma h), Q(x-y)\rangle \le \|Q(x-y)\|^2. \tag{15.52}$$

By combining with the previous inequality, we obtain

$$\tfrac{1}{\gamma}\|x - y\|_{Q-Q^2}^2 \le \langle \nabla\varphi_\gamma(x) - \nabla\varphi_\gamma(y), x - y\rangle \le \tfrac{1}{\gamma}\|x - y\|_Q^2. \tag{15.53}$$

Since $\lambda_{\min}(Q) = 1 - \gamma L_f$ and $\lambda_{\max}(Q) = 1 - \gamma \mu_f$, from Lemma 2 we conclude that

$$\mu_{\varphi_\gamma} \|x - y\|^2 \leq \langle \nabla \varphi_\gamma(x) - \nabla \varphi_\gamma(y), x - y \rangle \leq L_{\varphi_\gamma} \|x - y\|^2 \tag{15.54}$$

with μ_{φ_γ} and L_{φ_γ} as in the statement, hence the claim, cf. (15.6). □

Lemma 15.4.7 *Suppose that φ has the quadratic growth with constants (μ, ν), and let $\varphi_\star := \min \varphi$. Then, for all $\gamma \in (0, 1/L_f]$ and $x \in \operatorname{lev}_{\leq \varphi_\star + \nu} \varphi_\gamma$ it holds that*

$$\varphi_\gamma(x) - \varphi_\star \leq \gamma \left[\tfrac{1}{2} + (1 + 2/\gamma\mu)(1 + \gamma L_f)^2 \right] \|R_\gamma(x)\|^2. \tag{15.55}$$

Proof Fix $x \in \operatorname{lev}_{\leq \varphi_\star + \nu} \varphi_\gamma$ and let $\bar{x} := T_\gamma(x)$. We have

$$\begin{aligned} \varphi_\gamma(x) - \varphi_\star &\overset{15.4.2(ii)}{\leq} \tfrac{\gamma}{2} \|R_\gamma(x)\|^2 + \varphi(\bar{x}) - \varphi_\star \\ &\overset{15.3.3}{\leq} \tfrac{\gamma}{2} \|R_\gamma(x)\|^2 + \operatorname{dist}(\bar{x}, \mathcal{X}_\star) \operatorname{dist}(0, \partial\varphi(\bar{x})) \\ &\overset{15.3.4}{\leq} \left[\tfrac{\gamma}{2} \|R_\gamma(x)\| + (1 + \gamma L_f) \operatorname{dist}(\bar{x}, \mathcal{X}_\star) \right] \|R_\gamma(x)\| \end{aligned}$$

and since $\bar{x} \in \operatorname{lev}_{\leq \varphi_\star + \nu} \varphi$ (cf. Theorem 15.4.2*(ii)*), from Theorem 15.3.8 we can bound the quantity $\operatorname{dist}(\bar{x}, \mathcal{X}_\star)$ in terms of the residual as

$$\leq \left[\tfrac{\gamma}{2} \|R_\gamma(x)\| + (\gamma + 2/\mu)(1 + \gamma L_f)^2 \|R_\gamma(\bar{x})\| \right] \|R_\gamma(x)\|.$$

The proof now follows from the inequality $\|R_\gamma(\bar{x})\| \leq \|R_\gamma(x)\|$, see [4, Thm. 10.12], after easy algebraic manipulations. □

15.4.2 Further Equivalence Properties

Proposition 15.4.8 (Equivalence of Level Boundedness) *For any $\gamma \in (0, 1/L_f)$, φ has bounded level sets iff φ_γ does.*

Proof Theorem 15.4.2 implies that $\operatorname{lev}_{\leq \alpha} \varphi \subseteq \operatorname{lev}_{\leq \alpha} \varphi_\gamma$ for all $\alpha \in \mathbb{R}$, therefore level boundedness of φ_γ implies that of φ. Conversely, suppose that φ_γ is not level bounded, and consider $(x_k)_{k \in \mathbb{N}} \subseteq \operatorname{lev}_{\leq \alpha} \varphi_\gamma$ with $\|x_k\| \to \infty$. Then from Theorem 15.4.2 it follows that $\varphi(\bar{x}_k) \leq \varphi_\gamma(x_k) - \frac{1}{2\gamma} \|x_k - \bar{x}_k\|^2 \leq \alpha - \frac{1}{2\gamma} \|x_k - \bar{x}_k\|^2$, where $\bar{x}_k = T_\gamma(x_k)$. In particular, $(\bar{x}_k)_{k \in \mathbb{N}} \subseteq \operatorname{lev}_{\leq \alpha} \varphi$. If $(\bar{x}_k)_{k \in \mathbb{N}}$ is bounded, then $\inf \varphi = -\infty$; otherwise, $\operatorname{lev}_{\leq \alpha} \varphi$ contains the unbounded sequence $(\bar{x}_k)_{k \in \mathbb{N}}$. Either way, φ cannot be level bounded. □

Proposition 15.4.9 (Equivalence of Quadratic Growth) *Let $\gamma \in (0, 1/L_f)$ be fixed. Then,*

(i) if φ satisfies the quadratic growth condition with constants (μ, ν), then so does φ_γ with constants (μ', ν), where $\mu' := \frac{1-\gamma L_f}{(1+\gamma L_f)^2} \frac{\mu\gamma}{(2+\gamma\mu)^2} \mu$;

(ii) conversely, if φ_γ satisfies the quadratic growth condition, then so does φ with same constants.

Proof Since φ and φ_γ have same infimum and minimizers (cf. Theorem 15.4.3), 15.4.9*(ii)* is a straightforward consequence of the fact that $\varphi_\gamma \leq \varphi$ (cf. Theorem 15.4.2*(i)*).

Conversely, suppose that φ satisfies the quadratic growth with constants (μ, ν). Then, for all $x \in \mathrm{lev}_{\leq \varphi_\star + \nu} \varphi_\gamma$ we have that $\bar{x} := T_\gamma(x) \in \mathrm{lev}_{\leq \varphi_\star + \nu} \varphi$, therefore

$$\varphi_\gamma(x) - \varphi_\star \overset{15.4.2(ii)}{\geq} \varphi(\bar{x}) - \varphi_\star + \gamma \tfrac{1-\gamma L_f}{2} \|R_\gamma(x)\|^2 \geq \tfrac{\mu'}{2} \operatorname{dist}(x, \mathcal{X}_\star), \tag{15.56}$$

where in the last inequality we discarded the term $\varphi(\bar{x}) - \varphi_\star \geq 0$ and used Theorem 15.3.8 to lower bound $\|R_\gamma(x)\|^2$. □

Corollary 15.4.10 (Equivalence of Strong Minimality) *For all $\gamma \in (0, 1/L_f)$, a point $x_\star$ is a (locally) strong minimizer for φ iff it is a (locally) strong minimizer for φ_γ.*

Lastly, having showed that for convex functions the quadratic growth can be extended to arbitrary level sets (cf. Lemma 15.3.7), an interesting consequence of Proposition 15.4.9 is that, although φ_γ may fail to be convex, it enjoys the same property.

Corollary 15.4.11 (FBE: Globality of Quadratic Growth) *Let $\gamma \in (0, 1/L_f)$ and suppose that φ_γ satisfies the quadratic growth with constants (μ, ν). Then, for every $\nu' > \nu$ there exists $\mu' > 0$ such that φ_γ satisfies the quadratic growth with constants (μ', ν').*

15.4.3 Second-Order Properties

Although φ_γ is continuously differentiable over $\mathbb{R}^n$, it fails to be C^2 in most cases; since g is nonsmooth, its Moreau envelope g^γ is hardly ever C^2. For example, if g is real valued, then g^γ is C^2 (and $\mathrm{prox}_{\gamma g}$ is C^1) if and only if g is C^2 [33]. Therefore, we hardly ever have the luxury of assuming continuous differentiability of $\nabla\varphi_\gamma$ and we must resort to generalized notions of differentiability stemming from nonsmooth analysis. Specifically, our analysis is largely based on generalized differentiability properties of $\mathrm{prox}_{\gamma g}$ which we study next.

Theorem 15.4.12 *For all $x \in \mathbb{R}^n$, $\partial_C(\mathrm{prox}_{\gamma g})(x) \neq \emptyset$ and any $P \in \partial_C(\mathrm{prox}_{\gamma g})(x)$ is a symmetric positive semidefinite matrix that satisfies $\|P\| \leq 1$.*

Proof Nonempty-valuedness of $\partial_C(\mathrm{prox}_{\gamma g})$ is due to Lipschitz continuity of $\mathrm{prox}_{\gamma g}$. Moreover, since g is convex, its Moreau envelope is a convex function as well, therefore every element of $\partial_C(\nabla g^\gamma)(x)$ is a symmetric positive semidefinite matrix (see, *e.g.*, [19, §8.3.3]). Due to Theorem 15.3.2*(i)*, we have that $\mathrm{prox}_{\gamma g}(x) = x - \gamma \nabla g^\gamma(x)$, therefore

$$\partial_C(\mathrm{prox}_{\gamma g})(x) = \mathrm{I} - \gamma\, \partial_C(\nabla g^\gamma)(x). \tag{15.57}$$

The last relation holds with equality (as opposed to inclusion in the general case) due to the fact that one of the summands is continuously differentiable. Now, from (15.57) we easily infer that every element of $\partial_C(\mathrm{prox}_{\gamma g})(x)$ is a symmetric matrix. Since $\nabla g^\gamma(x)$ is Lipschitz continuous with Lipschitz constant γ^{-1}, using [15, Prop. 2.6.2(d)], we infer that every $H \in \partial_C(\nabla g^\gamma)(x)$ satisfies $\|H\| \leq \gamma^{-1}$. Now, according to (15.57) it holds that

$$P \in \partial_C(\mathrm{prox}_{\gamma g})(x) \quad \Leftrightarrow \quad P = I - \gamma H, \quad H \in \partial_C(\nabla g^\gamma)(x). \tag{15.58}$$

Therefore, for every $d \in \mathbb{R}^n$ and $P \in \partial_C(\mathrm{prox}_{\gamma g})(x)$,

$$\langle d, Pd\rangle = \|d\|^2 - \gamma \langle d, Hd\rangle \geq \|d\|^2 - \gamma\gamma^{-1}\|d\|^2 = 0. \tag{15.59}$$

On the other hand, since $\mathrm{prox}_{\gamma g}$ is Lipschitz continuous with Lipschitz constant 1, using [15, Prop. 2.6.2(d)] we obtain that $\|P\| \leq 1$ for all $P \in \partial_C(\mathrm{prox}_{\gamma g})(x)$. □

We are now in a position to construct a generalized Hessian for φ_γ that will allow the development of Newton-like methods with fast asymptotic convergence rates. An obvious route to follow would be to assume that $\nabla\varphi_\gamma$ is semismooth and employ $\partial_C(\nabla\varphi_\gamma)$ as a generalized Hessian for φ_γ. However, this approach would require extra assumptions on f and involve complicated operations to evaluate elements of $\partial_C(\nabla\varphi_\gamma)$. On the other hand, what is really needed to devise Newton-like algorithms with fast local convergence rates is a *linear Newton approximation (LNA)*, cf. Definition 15.2.3, at some stationary point of φ_γ, which by Theorem 15.4.3*(iii)* is also a minimizer of φ, provided that $\gamma \in (0, 1/L_f)$.

The approach we follow is largely based on [72], [19, Prop. 10.4.4]. Without any additional assumptions we can define a set-valued mapping $\hat{\partial}^2\varphi_\gamma : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n\times n}$ with full domain and whose elements have a simpler form than those of $\partial_C(\nabla\varphi_\gamma)$, which serves as a LNA for $\nabla\varphi_\gamma$ at any stationary point $x_\star$ provided $\mathrm{prox}_{\gamma g}$ is semismooth at $x_\star - \gamma\nabla f(x_\star)$. We call it *approximate generalized Hessian* of φ_γ and it is given by

$$\hat{\partial}^2\varphi_\gamma(x) := \left\{\gamma^{-1}Q_\gamma(x)(\mathrm{I} - PQ_\gamma(x)) \mid P \in \partial_C(\mathrm{prox}_{\gamma g})(x - \gamma\nabla f(x))\right\}. \tag{15.60}$$

Notice that if f is quadratic, then $\hat{\partial}^2\varphi_\gamma \equiv \partial_C \nabla\varphi_\gamma$; more generally, the key idea in the definition of $\hat{\partial}^2\varphi_\gamma$, reminiscent of the Gauss-Newton method for nonlinear least-squares problems, is to omit terms vanishing at $x_\star$ that contain third-order derivatives of f.

Proposition 15.4.13 *Let $\bar{x} \in \mathbb{R}^n$ and $\gamma > 0$ be fixed. If $\operatorname{prox}_{\gamma g}$ is (ϑ-order) semismooth at $\bar{x} - \gamma\nabla f(\bar{x})$ (and $\nabla^2 f$ is ϑ-Hölder continuous around $\bar{x}$), then*

$$\mathcal{R}_\gamma(x) := \left\{\gamma^{-1}(\mathrm{I} - P\,Q_\gamma(x)) \mid P \in \partial_C \operatorname{prox}_{\gamma g}(x - \gamma\nabla f(x))\right\} \tag{15.61}$$

is a (ϑ-order) LNA for R_γ at $\bar{x}$.

Proof We shall prove only the ϑ-order semismooth case, as the other one is shown by simply replacing all occurrences of $O(\|\cdot\|^{1+\vartheta})$ with $o(\|\cdot\|)$ in the proof. Let $q_\gamma = \mathrm{id} - \gamma\nabla f$ be the forward operator, so that the forward-backward operator T_γ can be expressed as $T_\gamma = \operatorname{prox}_{\gamma g} \circ q_\gamma$. With a straightforward adaptation of the proof of [19, Prop. 7.2.9] to include the ϑ-Hölderian case, it can be shown that

$$q_\gamma(x) - q_\gamma(\bar{x}) - Q_\gamma(x)(x - \bar{x}) = O(\|x - \bar{x}\|^{1+\vartheta}). \tag{15.62}$$

Moreover, since ∇f is Lipschitz continuous and thus so is q_γ, we also have

$$q_\gamma(x) - q_\gamma(\bar{x}) = O(\|x - \bar{x}\|). \tag{15.63}$$

Let $U_x \in \mathcal{R}_\gamma(x)$ be arbitrary; then, there exists $P_x \in \partial_C \operatorname{prox}_{\gamma g}(x - \gamma\nabla f(x))$ such that $U_x = \gamma^{-1}(\mathrm{I} - P_x Q_\gamma(x))(\bar{x} - x)$. We have

$$\begin{aligned}
&R_\gamma(x) + U_x(\bar{x} - x) - R_\gamma(\bar{x}) \\
&= R_\gamma(x) + \gamma^{-1}(\mathrm{I} - P_x Q_\gamma(x))(\bar{x} - x) - R_\gamma(\bar{x}) \\
&= \gamma^{-1}\big[\operatorname{prox}_{\gamma g}(q_\gamma(\bar{x})) - \operatorname{prox}_{\gamma g}(q_\gamma(x)) - P_x Q_\gamma(x)(\bar{x} - x)\big]
\end{aligned}$$

due to ϑ-order semismoothness of $\operatorname{prox}_{\gamma g}$ at $q_\gamma(\bar{x})$,

$$\begin{aligned}
&= \gamma^{-1} P_x\big[q_\gamma(\bar{x}) - q_\gamma(x) + O(\|q_\gamma(\bar{x}) - q_\gamma(x)\|^{1+\vartheta}) - Q_\gamma(x)(\bar{x} - x)\big] \\
&\overset{(15.63)}{=} \gamma^{-1} P_x\big[q_\gamma(\bar{x}) - q_\gamma(x) - Q_\gamma(\bar{x})(\bar{x} - x) + O(\|\bar{x} - x\|^{1+\vartheta})\big] \\
&\overset{(15.62)}{=} \gamma^{-1} P_x\, O(\|\bar{x} - x\|^{1+\vartheta}) = O(\|\bar{x} - x\|^{1+\vartheta}),
\end{aligned}$$

where in the last equality we used the fact that $\|P_x\| \le 1$, cf. Theorem 15.4.12. □

Corollary 15.4.14 *Let $\gamma \in (0, 1/L_f)$ and $x_\star \in \mathcal{X}_\star$. If $\operatorname{prox}_{\gamma g}$ is (ϑ-order) semismooth at $x_\star - \gamma\nabla f(x_\star)$ (and $\nabla^2 f$ is locally ϑ-Hölder continuous around $x_\star$), then $\hat{\partial}^2\varphi_\gamma$ is a (ϑ-order) LNA for $\nabla\varphi_\gamma$ at $x_\star$.*

Proof Let $H_x \in \hat{\partial}^2\varphi_\gamma(x) = \left\{Q_\gamma(x)U \mid U \in \mathcal{R}_\gamma(x)\right\}$, so that $H_x = Q_\gamma(x)U_x$ for some $U_x \in \mathcal{R}_\gamma(x)$. Then,

$$\begin{aligned}\|\nabla\varphi_\gamma(x) + H_x(x_\star - x) - \nabla\varphi_\gamma(x_\star)\| &= \|Q_\gamma(x)R_\gamma(x) + Q_\gamma(x)U_x(x - x_\star))\| \\ &= \|Q_\gamma(x)[R_\gamma(x) + U_x(x - x_\star) - R_\gamma(x_\star)]\| \\ &\leq \|R_\gamma(x) + U_x(x - x_\star) - R_\gamma(x_\star)\|,\end{aligned}$$

where in the equalities we used the fact that $\nabla\varphi_\gamma(x_\star) = R_\gamma(x_\star) = 0$, and in the inequality the fact that $\|Q_\gamma\| \leq 1$. Since $\mathcal{R}_\gamma$ is a (ϑ-order) LNA of R_γ at $x_\star$, the last term is $o(\|x - x_\star\|)$ (resp. $O(\|x - x_\star\|^{1+\vartheta})$). □

As shown in the next result, although the FBE is in general not convex, for γ small enough every element of $\hat{\partial}^2\varphi_\gamma(x)$ is a (symmetric and) positive semidefinite matrix. Moreover, the eigenvalues are lower and upper bounded uniformly over all $x \in \mathbb{R}^n$.

Proposition 15.4.15 *Let $\gamma \leq 1/L_f$ and $H \in \hat{\partial}^2\varphi_\gamma(x)$ be fixed. Then, $H \in \mathrm{S}_+(\mathbb{R}^n)$ with*

$$\lambda_{\min}(H) = \min\left\{(1 - \gamma\mu_f)\mu_f, (1 - \gamma L_f)L_f\right\} \text{ and } \lambda_{\max}(H) = \gamma^{-1}(1 - \gamma\mu_f), \tag{15.64}$$

where $\mu_f \geq 0$ is the modulus of strong convexity of f.

Proof Fix $x \in \mathbb{R}^n$ and let $Q \coloneqq Q_\gamma(x)$. Any $H \in \hat{\partial}^2\varphi_\gamma(x)$ can be expressed as $H = \gamma^{-1}Q(\mathrm{I} - PQ)$ for some $P \in \partial_C(\operatorname{prox}_{\gamma g})(x - \gamma\nabla f(x))$. Since both Q and P are symmetric (cf. Theorem 15.4.12), it follows that so is H. Moreover, for all $d \in \mathbb{R}^n$

$$\begin{aligned}\langle Hd, d\rangle &= \gamma^{-1}\langle Qd, d\rangle - \gamma^{-1}\langle PQd, Qd\rangle \\ &\overset{15.4.12}{\geq} \gamma^{-1}\langle Qd, d\rangle - \gamma^{-1}\|Qd\|^2 \\ &= \langle(\mathrm{I} - \gamma\nabla^2 f(x))\nabla^2 f(x)d, d\rangle \\ &\overset{2}{\geq} \min\left\{(1 - \gamma\mu_f)\mu_f,\ (1 - \gamma L_f)L_f\right\}\|d\|^2.\end{aligned} \tag{15.65}$$

On the other hand, since $P \succeq 0$ (cf. Theorem 15.4.12) and thus $QPQ \succeq 0$, we can upper bound (15.65) as

$$\langle Hd, d\rangle \leq \gamma^{-1}\langle Qd, d\rangle \leq \|Q\|\|d\|^2 \leq \gamma^{-1}(1 - \gamma\mu_f)\|d\|^2.$$

□

The next lemma links the behavior of the FBE close to a solution of (15.1) and a nonsingularity assumption on the elements of $\hat{\partial}^2\varphi_\gamma(x_\star)$. Part of the statement is similar to [19, Lem. 7.2.10]; however, here $\nabla\varphi_\gamma$ is not required to be locally Lipschitz around $x_\star$.

Lemma 15.4.16 *Let $x_\star \in \operatorname{argmin}\varphi$ and $\gamma \in (0, 1/L_f)$. If $\operatorname{prox}_{\gamma g}$ is semismooth at $x_\star - \gamma\nabla f(x_\star)$, then the following conditions are equivalent:*

(a) $x_\star$ is a locally strong minimum for φ (or, equivalently, for φ_γ);

(b) every element of $\hat{\partial}^2\varphi_\gamma(x_\star)$ is nonsingular.

In any such case, there exist $\delta, \kappa > 0$ such that

$$\|x - x_\star\| \le \kappa\|R_\gamma(x)\| \quad \textit{and} \quad \max\left\{\|H\|, \|H^{-1}\|\right\} \le \kappa, \tag{15.66}$$

for any $x \in \mathrm{B}(x_\star; \delta)$ and $H \in \hat{\partial}^2\varphi_\gamma(x)$.

Proof Observe first that Corollary 15.4.14 ensures that $\hat{\partial}^2\varphi_\gamma$ is a LNA of $\nabla\varphi_\gamma$ at $x_\star$, thus semicontinuous and compact valued (by definition). In particular, the last claim follows from [19, Lem. 7.5.2].

- 15.4.16*(a)* $\Rightarrow$ 15.4.16*(b)* It follows from Corollary 15.4.10 that there exists $\mu, \delta > 0$ such that $\varphi_\gamma(x) - \varphi_\star \ge \frac{\mu}{2}\|x - x_\star\|^2$ for all $x \in \mathrm{B}(x_\star; \delta)$. In particular, for all $H \in \hat{\partial}^2\varphi_\gamma(x_\star)$ and $x \in \mathrm{B}(x_\star; \delta)$ we have

$$\tfrac{\mu}{2}\|x - x_\star\|^2 \le \varphi_\gamma(x) - \varphi_\star = \tfrac{1}{2}\langle H(x - x_\star), x - x_\star\rangle + o(\|x - x_\star\|^2). \tag{15.67}$$

  Let $v_{\min}$ be a unitary eigenvector of H corresponding to the minimum eigenvalue $\lambda_{\min}(H)$. Then, for all $\varepsilon \in (-\delta, \delta)$ the point $x_\varepsilon = x_\star + \varepsilon v_{\min}$ is δ-close to $x_\star$ and thus

$$\tfrac{1}{2}\lambda_{\min}(H)\varepsilon^2 \ge \tfrac{\mu}{2}\varepsilon^2 + o(\varepsilon^2) \ge \tfrac{\mu}{4}\varepsilon^2, \tag{15.68}$$

  where the last inequality holds up to possibly restricting δ (and thus ε). The claim now follows from the arbitrarity of $H \in \hat{\partial}^2\varphi_\gamma(x_\star)$.
- 15.4.16*(a)* $\Leftarrow$ 15.4.16*(b)* Easily follows by reversing the arguments of the other implication. □

□

15.5 Forward-Backward Truncated-Newton Algorithm (FBTN)

Having established the equivalence between minimizing φ and φ_γ, we may recast problem (15.1) into the smooth unconstrained minimization of the FBE. Under some assumptions the elements of $\hat{\partial}^2\varphi_\gamma$ mimic second-order derivatives of φ_γ, suggesting

Algorithm 15.1 (FBTN) Forward-Backward Truncated-Newton method

REQUIRE $\gamma \in (0, 1/L_f)$; $\sigma \in (0, \frac{\gamma(1-\gamma L_f)}{2})$; $\bar{\eta}, \zeta \in (0, 1)$; $\rho, \nu \in (0, 1]$
initial point $x_0 \in \mathbb{R}^n$; accuracy $\varepsilon > 0$
PROVIDE ε-suboptimal solution x^k (*i.e.*, such that $\|R_\gamma(x^k)\| \le \varepsilon$)
INITIALIZE $k \leftarrow 0$
1: **while** $\|R_\gamma(x^k)\| > \varepsilon$ **do**
2: $\delta_k \leftarrow \zeta \|\nabla\varphi_\gamma(x^k)\|^\nu$, $\eta_k \leftarrow \min\{\bar{\eta}, \|\nabla\varphi_\gamma(x^k)\|^\rho\}$, $\varepsilon_k \leftarrow \eta_k \|\nabla\varphi_\gamma(x^k)\|$
3: Apply CG(Alg. 15.2) to find an ε_k-approximate solution d^k to

$$[H_k + \delta_k \mathrm{I}]d^k \approx -\nabla\varphi_\gamma(x^k) \tag{15.69}$$

for some $H_k \in \hat{\partial}^2\varphi_\gamma(x^k)$
4: Let τ_k be the maximum in $\{2^{-i} \mid i \in \mathbb{N}\}$ such that

$$\varphi_\gamma(x^{k+1}) \le \varphi_\gamma(x^k) - \sigma\|R_\gamma(x^k)\|^2 \tag{15.70}$$

where $x^{k+1} \leftarrow (1-\tau_k)T_\gamma(x^k) + \tau_k[x^k + d^k]$
5: $k \leftarrow k+1$ and go to step 1
6: **end while**

the employment of Newton-like update directions $d = -(H + \delta\mathrm{I})^{-1}\nabla\varphi_\gamma(x)$ with $H \in \hat{\partial}^2\varphi_\gamma(x)$ and $\delta > 0$ (the regularization term $\delta\mathrm{I}$ ensures the well definedness of d, as H is positive semidefinite, see Proposition 15.4.15). If δ and ε are suitably selected, under some nondegeneracy assumptions updates $x^+ = x + d$ are locally superlinearly convergent. Since such d's are directions of descent for φ_γ, a possible globalization strategy is an Armijo-type linesearch. Here, however, we follow the simpler approach proposed in [71, 75] that exploits the basic properties of the FBE investigated in Section 15.4.1. As we will discuss shortly after, this is also advantageous from a computational point of view, as it allows an arbitrary warm starting for solving the underlying linear system.

Let us elaborate on the linesearch. To this end, let x be the current iterate; then, Theorem 15.4.2 ensures that $\varphi_\gamma(T_\gamma(x)) \le \varphi_\gamma(x) - \gamma\frac{1-\gamma L_f}{2}\|R_\gamma(x)\|^2$. Therefore, unless $R_\gamma(x) = 0$, in which case x would be a solution, for any $\sigma \in (0, \gamma\frac{1-\gamma L_f}{2})$ the strict inequality $\varphi_\gamma(T_\gamma(x)) < \varphi_\gamma(x) - \sigma\|R_\gamma(x)\|^2$ is satisfied. Due to the continuity of φ_γ, all points sufficiently close to $T_\gamma(x)$ will also satisfy the inequality, thus so will the point $x^+ = (1-\tau)T_\gamma(x) + \tau(x+d)$ for small enough stepsizes τ. This fact can be used to enforce the iterates to *sufficiently* decrease the value of the FBE, cf. (15.70), which straightforwardly implies optimality of all accumulation points of the generated sequence. We defer the details to the proof of Theorem 15.5.1. In Theorems 15.5.4 and 15.5.5 we will provide conditions ensuring acceptance of unit stepsizes so that the scheme reduces to a regularized version of the (undamped) linear Newton method [19, Alg. 7.5.14] for solving $\nabla\varphi_\gamma(x) = 0$, which, under due assumptions, converges superlinearly.

In order to ease the computation of d^k, we allow for inexact solutions of the linear system by introducing a tolerance $\varepsilon_k > 0$ and requiring $\|(H_k+\delta_k \mathrm{I})d^k+\nabla\varphi_\gamma(x^k)\| \leq \varepsilon_k$. Since $H_k + \delta_k \mathrm{I}$ is positive definite, inexact solutions of the linear system can be efficiently retrieved by means of CG(Alg. 15.2) , which only requires matrix-vector products and thus only (generalized) directional derivatives, namely, (generalized) derivatives (denoted as $\frac{\partial}{\partial\lambda}$) of the single-variable functions $t \mapsto \mathrm{prox}_{\gamma g}(x + t\lambda)$ and $t \mapsto \nabla f(x + t\lambda)$, as opposed to computing the full (generalized) Hessian matrix. To further enhance computational efficiency, we may warm start the CG method with the previously computed direction, as eventually subsequent update directions are expected to have a small difference. Notice that this warm starting does not ensure that the provided (inexact) solution d^k is a direction of descent for φ_γ; either way, this property is not required by the adopted linesearch, showing a considerable advantage over classical Armijo-type rules. Putting all these facts together we obtain the proposed FBE-based truncated-Newton algorithm FBTN(Alg. 15.1) for convex composite minimization.

Remark 15.1 (Adaptive Variant When L_f Is Unknown) In practice, no prior knowledge of the global Lipschitz constant L_f is required for FBTN. In fact, replacing L_f with an initial estimate $L > 0$, the following instruction can be added at the beginning of each iteration, before step 1:

0: $\bar{x}^k \leftarrow T_\gamma(x^k)$
while $f(\bar{x}^k) > f(x^k) + \langle \nabla f(x^k), \bar{x}^k - x^k\rangle + \frac{L}{2}\|\bar{x}^k - x^k\|^2$ **do**
$\gamma \leftarrow \gamma/2,\ L \leftarrow 2L,\ \bar{x}^k \leftarrow T_\gamma(x^k)$

Algorithm 15.2 (CG) Conjugate Gradient for computing the update direction

REQUIRE $\nabla\varphi_\gamma(x^k)$; δ_k; ε_k; d^{k-1} (set to 0 if $k = 0$)
(generalized) directional derivatives $\lambda \mapsto \frac{\partial^{\cdot 0}\,\mathrm{prox}_{\gamma g}}{\partial\lambda}(x^k - \gamma\nabla f(x^k))$ and $\lambda \mapsto \frac{\partial^{\cdot 0}\nabla f}{\partial\lambda}(x^k)$
PROVIDE update direction d^k
INITIALIZE $e, p \leftarrow -\nabla\varphi_\gamma(x^k)$; warm start $d^k \leftarrow d^{k-1}$
1: **while** $\|e\| > \varepsilon_k$ **do**
2: $u \leftarrow \frac{\partial\nabla f}{\partial p}(x^k)$
3: $v \leftarrow p - \gamma u$ ▷ $v = Q_\gamma(x^k)p$
4: $w \leftarrow p - \frac{\partial\,\mathrm{prox}_{\gamma g}}{\partial v}(x^k - \gamma\nabla f(x^k))$
5: $z \leftarrow \delta_k p + w - \gamma\frac{\partial\nabla f}{\partial w}(x^k)$ ▷ $z = H_k p$
6: $\alpha \leftarrow \|e\|^2/\langle p, z\rangle$
7: $d^k \leftarrow d^k + \alpha p,\ e^+ \leftarrow e - \alpha z$
8: $p \leftarrow e^+ + \left(\|e^+\|/\|e\|\right)^2 p$
9: $e \leftarrow e^+$
10: **end while**

Moreover, since positive definiteness of $H_k + \delta_k \mathrm{I}$ is ensured only for $\gamma \leq 1/L_f$ where L_f is the true Lipschitz constant of $\nabla\varphi_\gamma$ (cf. Proposition 15.4.15), special care should be taken when applying CG in order to find the update direction d^k. Specifically, CG should be stopped prematurely whenever $\langle p, z\rangle \leq 0$ in step 6, in which case $\gamma \leftarrow \gamma/2$, $L \leftarrow 2L$ and the iteration should start again from step 1.

Whenever the quadratic bound (15.5) is violated with L in place of L_f, the estimated Lipschitz constant L is increased, γ is decreased accordingly, and the proximal gradient point $\bar{x}^k$ with the new stepsize γ is evaluated. Since replacing L_f with any $L \geq L_f$ still satisfies (15.5), it follows that L is incremented only a finite number of times. Therefore, there exists an iteration k_0 starting from which γ and L are constant; in particular, all the convergence results here presented remain valid starting from iteration k_0, at latest. Moreover, notice that this step does not increase the complexity of the algorithm, since both $\bar{x}^k$ and $\nabla f(x^k)$ are needed for the evaluation of $\varphi_\gamma(x^k)$.

15.5.1 Subsequential and Linear Convergence

Before going through the convergence proofs let us spend a few lines to emphasize that FBTN is a well-defined scheme. First, that a matrix H_k as in line 1 exists is due to the nonemptyness of $\hat{\partial}^2\varphi_\gamma(x^k)$ (cf. Section 15.4.3). Second, since $\delta_k > 0$ and $H_k \succeq 0$ (cf. Proposition 15.4.15) it follows that $H_k + \delta_k \mathrm{I}$ is (symmetric and) positive definite, and thus CG is indeed applicable at line 3.

Having clarified this, the proof of the next result falls as a simplified version of [75, Lem. 5.1 and Thm. 5.6]; we elaborate on the details for the sake of self-inclusiveness. To rule out trivialities, in the rest of the chapter we consider the limiting case of infinite accuracy, that is $\varepsilon = 0$, and assume that the termination criterion $\|R_\gamma(x^k)\| = 0$ is never met. We shall also work under the assumption that a solution to the investigated problem (15.1) exists, thus in particular that the cost function φ is lower bounded.

Theorem 15.5.1 (Subsequential Convergence) *Every accumulation point of the sequence $(x^k)_{k\in\mathbb{N}}$ generated by FBTN (Alg. 15.1) is optimal.*

Proof Observe that

$$\varphi_\gamma\big(x^k - \gamma R_\gamma(x^k)\big) \overset{15.4.2}{\leq} \varphi_\gamma(x^k) - \gamma\tfrac{1-\gamma L_f}{2}\|R_\gamma(x^k)\|^2 < \varphi_\gamma(x^k) - \sigma\|R_\gamma(x^k)\|^2 \tag{15.71}$$

and that $x^{k+1} \to T_\gamma(x^k)$ as $\tau_k \to 0$. Continuity of φ_γ ensures that for small enough τ_k the linesearch condition (15.70) is satisfied, in fact, regardless of what d^k is. Therefore, for each k the stepsize τ_k is decreased only a finite number of times. By telescoping the linesearch inequality (15.70) we obtain

$$\sigma\sum_{k\in\mathbb{N}}\|R_\gamma(x^k)\|^2 \leq \sum_{k\in\mathbb{N}}\big[\varphi_\gamma(x^k) - \varphi_\gamma(x^{k+1})\big] \leq \varphi_\gamma(x^0) - \varphi_\star < \infty \tag{15.72}$$

and in particular $R_\gamma(x^k) \to 0$. Since R_γ is continuous we infer that every accumulation point $x_\star$ of $(x^k)_{k\in\mathbb{N}}$ satisfies $R_\gamma(x_\star) = 0$, hence $x_\star \in \operatorname{argmin}\varphi$, cf. (15.24). □

Remark 15.2 Since FBTN is a descent method on φ_γ, as ensured by the linesearch condition (15.70), from Proposition 15.4.8 it follows that a sufficient condition for the existence of cluster points is having φ with bounded level sets or, equivalently, having $\operatorname{argmin}\varphi$ bounded (cf. Lemma 1).

As a straightforward consequence of Lemma 15.4.7, from the linesearch condition (15.70) we infer Q-linear decrease of the FBE along the iterates of FBTN provided that the original function φ has the quadratic growth property. In particular, although the quadratic growth is a local property, Q-linear convergence holds globally, as described in the following result.

Theorem 15.5.2 (Q-Linear Convergence of FBTN Under Quadratic Growth) *Suppose that φ satisfies the quadratic growth with constants (μ, ν). Then, the iterates of FBTN(Alg. 15.1) decrease Q-linearly the value of φ_γ as*

$$\varphi_\gamma(x^{k+1}) - \varphi_\star \le \left(1 - \frac{2\sigma\mu'}{\gamma\mu + 2(2+\gamma\mu')(1+\gamma L_f)^2}\right)(\varphi_\gamma(x^k) - \varphi_\star) \quad \forall k \in \mathbb{N}, \tag{15.73}$$

where

$$\mu' := \begin{cases} \mu & \text{if } \varphi_\gamma(x_0) \le \varphi_\star + \nu, \\ \frac{\mu}{2}\min\left\{1, \frac{\nu}{\varphi_\gamma(x_0) - \varphi_\star - \nu}\right\} & \text{otherwise.} \end{cases} \tag{15.74}$$

Proof Since FBTN is a descent method on φ_γ, it holds that $(x^k)_{k\in\mathbb{N}} \subseteq \operatorname{lev}_{\le\alpha}\varphi_\gamma$ with $\alpha = \varphi_\gamma(x^0)$. It follows from Lemma 15.3.7 that φ satisfies the quadratic growth condition with constants $(\mu', \varphi_\gamma(x^0))$, with μ' is as in the statement. The claim now follows from the inequality ensured by linesearch condition (15.70) combined with Lemma 15.4.7. □

15.5.2 Superlinear Convergence

In this section we provide sufficient conditions that enable superlinear convergence of FBTN. In the sequel, we will make use of the notion of *superlinear directions* that we define next.

Definition 15.5.3 (Superlinear Directions) Suppose that $\mathcal{X}_\star \neq \emptyset$ and consider the iterates generated by FBTN(Alg. 15.1). We say that $(d^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ are *superlinearly convergent directions* if

$$\lim_{k\to\infty} \frac{\operatorname{dist}(x^k + d^k, \mathcal{X}_\star)}{\operatorname{dist}(x^k, \mathcal{X}_\star)} = 0.$$

If for some $q > 1$ the condition can be strengthened to

$$\limsup_{k\to\infty} \frac{\operatorname{dist}(x^k + d^k, \mathcal{X}_\star)}{\operatorname{dist}(x^k, \mathcal{X}_\star)^q} < \infty$$

then we say that $(d^k)_{k\in\mathbb{N}}$ are *superlinearly convergent directions with order q*.

We remark that our definition of superlinear directions extends the one given in [19, §7.5] to cases in which $\mathcal{X}_\star$ is not a singleton. The next result constitutes a key component of the proposed methodology, as it shows that the proposed algorithm does not suffer from the *Maratos' effect* [44], a well-known obstacle for fast local methods that inhibits the acceptance of the unit stepsize. On the contrary, we will show that whenever the directions $(d^k)_{k\in\mathbb{N}}$ computed in FBTN are superlinear, then indeed the unit stepsize is eventually always accepted, and the algorithm reduces to a regularized version of the (undamped) linear Newton method [19, Alg. 7.5.14] for solving $\nabla\varphi_\gamma(x) = 0$ or, equivalently, $R_\gamma(x) = 0$, and $\operatorname{dist}(x^k, \mathcal{X}_\star)$ converges superlinearly.

Theorem 15.5.4 (Acceptance of the Unit Stepsize and Superlinear Convergence) *Consider the iterates generated by FBTN(Alg. 15.1). Suppose that φ satisfies the quadratic growth (locally) and that $(d^k)_{k\in\mathbb{N}}$ are superlinearly convergent directions (with order q). Then, there exists $\bar{k} \in \mathbb{N}$ such that*

$$\varphi_\gamma(x^k + d^k) \le \varphi_\gamma(x^k) - \sigma\|R_\gamma(x^k)\|^2 \qquad \forall k \ge \bar{k}. \tag{15.75}$$

In particular, eventually the iterates reduce to $x^{k+1} = x^k + d^k$, and $\operatorname{dist}(x^k, \mathcal{X}_\star)$ converges superlinearly (with order q).

Proof Without loss of generality we may assume that $(x^k)_{k\in\mathbb{N}}$ and $(x^k + d^k)_{k\in\mathbb{N}}$ belong to a region in which quadratic growth holds. Denoting $\varphi_\star := \min\varphi$, since φ_γ also satisfies the quadratic growth (cf. Proposition 15.4.9*(i)*) it follows that

$$\varphi_\gamma(x^k) - \varphi_\star \ge \tfrac{\mu'}{2}\operatorname{dist}(x^k, \mathcal{X}_\star)^2 \tag{15.76}$$

for some constant $\mu' > 0$. Moreover, we know from Lemma 15.4.7 that

$$\varphi_\gamma(x^k + d^k) - \varphi_\star \le c\|R_\gamma(x^k + d^k)\|^2 \le c'\operatorname{dist}(x^k + d^k, \mathcal{X}_\star)^2 \tag{15.77}$$

for some constants $c, c' > 0$, where in the second inequality we used Lipschitz continuity of R_γ (Lemma 3) together with the fact that $R_\gamma(x_\star) = 0$ for all points $x_\star \in \mathcal{X}_\star$. By combining the last two inequalities, we obtain

$$t_k := \frac{\varphi_\gamma(x^k + d^k) - \varphi_\star}{\varphi_\gamma(x^k) - \varphi_\star} \le \frac{2c'\operatorname{dist}(x^k + d^k, \mathcal{X}_\star)^2}{\mu'\operatorname{dist}(x^k, \mathcal{X}_\star)^2} \to 0 \quad \text{as } k \to \infty. \tag{15.78}$$

Moreover,

$$\varphi_\gamma(x^k) - \varphi_\star \geq \varphi_\gamma(x^k) - \varphi(T_\gamma(x^k)) \overset{15.4.2(ii)}{\geq} \gamma \tfrac{1-\gamma L_f}{2} \|R_\gamma(x^k)\|^2. \tag{15.79}$$

Thus,

$$\begin{aligned}\varphi_\gamma(x^k + d^k) - \varphi_\gamma(x^k) &= \big[\varphi_\gamma(x^k + d^k) - \varphi_\star\big] - \big[\varphi_\gamma(x^k) - \varphi_\star\big] \\ &= (t_k - 1)\big[\varphi_\gamma(x^k) - \varphi_\star\big]\end{aligned}$$

and since $t_k \to 0$, eventually it holds that $t_k \leq 1 - \frac{2\sigma}{\gamma(1-\gamma L_f)} \in (0, 1)$, resulting in

$$\leq -\sigma \|R_\gamma(x^k)\|^2.$$

□

Theorem 15.5.5 *Consider the iterates generated by FBTN(Alg. 15.1). Suppose that φ satisfies the quadratic growth (locally), and let $x_\star$ be the limit point of $(x^k)_{k\in\mathbb{N}}$.*[2] *Then, $(d^k)_{k\in\mathbb{N}}$ are superlinearly convergent directions provided that*

(i) either R_γ is strictly differentiable at $x_\star$[3] *and there exists $D > 0$ such that $\|d^k\| \leq D\|\nabla\varphi_\gamma(x^k)\|$ for all k's,*

(ii) or $\mathcal{X}_\star = \{x_\star\}$ and $\operatorname{prox}_{\gamma g}$ is semismooth at $x_\star - \gamma\nabla f(x_\star)$. In this case, if $\operatorname{prox}_{\gamma g}$ is ϑ-order semismooth at $x_\star - \gamma\nabla f(x_\star)$ and $\nabla^2 f$ is ϑ-Hölder continuous close to $x_\star$, then the order of superlinear convergence is at least $1 + \min\{\rho, \vartheta, \nu\}$.

Proof Due to Proposition 15.4.9 and Theorem 15.5.2, if $\mathcal{X}_\star = \{x_\star\}$ then the sequence $(x^k)_{k\in\mathbb{N}}$ converges to $x_\star$. Otherwise, the hypothesis ensure that

$$\|x^{k+1} - x^k\| = \tau_k\|d^k\| \leq D\|\nabla\varphi_\gamma(x^k)\| \leq D\|R_\gamma(x^k)\|, \tag{15.80}$$

from which we infer that $(\|x^{k+1} - x^k\|)_{k\in\mathbb{N}}$ is R-linearly convergent, hence that $(x^k)_{k\in\mathbb{N}}$ is a Cauchy sequence, and again we conclude that the limit point $x_\star$ indeed exists. Moreover, in light of Proposition 15.4.9 we have that $(x^k)_{k\in\mathbb{N}}$ is contained in a level set of φ_γ where φ_γ has quadratic growth. To establish a notation, let $e^k := [H_k + \delta_k \mathrm{I}]d^k + \nabla\varphi_\gamma(x^k)$ be the error in solving the linear system at line 3, so that

$$\|e^k\| \leq \varepsilon_k \leq \|\nabla\varphi_\gamma(x^k)\|^{1+\rho}, \tag{15.81}$$

(cf. line 1), and let $H_k = Q_\gamma(x^k)U_k$ for some $U_k \in \mathcal{R}_\gamma(x^k)$, see (15.61). Let us now analyze the two cases separately.

[2] As detailed in the proof, under the assumptions the limit point indeed exists.

[3] From the chain rule of differentiation it follows that R_γ is strictly differentiable at $x_\star$ if $\operatorname{prox}_{\gamma g}$ is strictly differentiable at $x_\star - \gamma\nabla f(x_\star)$ (strict differentiability is closed under composition).

- 15.5.5*(i)* Let $x_\star^k := \mathscr{P}_{\mathcal{X}_\star} x^k$, so that $\operatorname{dist}(x^k, \mathcal{X}_\star) = \|x^k - x_\star^k\|$. Recall that $\nabla\varphi_\gamma = Q_\gamma R_\gamma$ and that $(1-\gamma L_f)\mathrm{I} \preceq Q_\gamma \preceq \mathrm{I}$. Since $R_\gamma(x_\star^k) = 0$, from Lemma 3 and Theorem 15.3.8 we infer that there exist $r_1, r_2 > 0$ such that

$$\|R_\gamma(x^k)\| \geq r_1 \operatorname{dist}(x^k, \mathcal{X}_\star) \quad \text{and} \quad \|\nabla\varphi_\gamma(x^k)\| \leq r_2 \operatorname{dist}(x^k, \mathcal{X}_\star). \tag{15.82}$$

  In particular, the assumption on d^k ensures that $\|d^k\| = O\big(\operatorname{dist}(x^k, \mathcal{X}_\star)\big)$. We have

$$\begin{aligned} r_1 \operatorname{dist}(x^k + d^k, \mathcal{X}_\star) &\overset{(15.82)}{\leq} \|R_\gamma(x^k + d^k)\| \\ &\leq \underbrace{\|R_\gamma(x^k + d^k) - R_\gamma(x^k) - U_k d^k\|}_{\text{(a)}} + \underbrace{\|R_\gamma(x^k) + U_k d^k\|}_{\text{(b)}}. \end{aligned}$$

  As to quantity (a), we have

$$\begin{aligned} \text{(a)} &\leq \|R_\gamma(x^k + d^k) - R_\gamma(x^k) - JR_\gamma(x_\star)d^k\| + \|U_k - JR(x_\star)\|\|d^k\| \\ &= o\big(\operatorname{dist}(x^k, \mathcal{X}_\star)\big), \end{aligned}$$

  where we used strict differentiability and the fact that $\partial_C R_\gamma(x_\star) = \{JR_\gamma(x_\star)\}$ [15, Prop. 2.2.4] which implies $U_k \to JR(x_\star)$. In order to bound (b), recall that $\delta_k = \zeta\|\nabla\varphi_\gamma(x^k)\|^\nu$ (cf. line 1). Then,

$$\begin{aligned} \text{(b)} &= \|Q_\gamma(x^k)^{-1}(e^k - \delta_k d^k)\| \\ &\overset{(15.81)}{\leq} \tfrac{1}{1-\gamma L_f}\|\nabla\varphi_\gamma(x^k)\| \left(\|\nabla\varphi_\gamma(x^k)\|^\rho + \zeta\|\nabla\varphi_\gamma(x^k)\|^{\nu-1}\|d^k\|\right). \\ &\overset{(15.82)}{\leq} \tfrac{r_2}{1-\gamma L_f} \operatorname{dist}(x^k, \mathcal{X}_\star) \left(r_2^\rho \operatorname{dist}(x^k, \mathcal{X}_\star)^\rho + \zeta\|\nabla\varphi_\gamma(x^k)\|^{\nu-1}\|d^k\|\right) \\ &= O\big(\operatorname{dist}(x^k, \mathcal{X}_\star)^{1+\min\{\rho,\nu\}}\big), \end{aligned}$$

  and we conclude that $\operatorname{dist}(x^k + d^k, \mathcal{X}_\star) \leq \text{(a)} + \text{(b)} \leq o\big(\operatorname{dist}(x^k, \mathcal{X}_\star)\big)$.
- 15.5.5*(ii)* In this case $\operatorname{dist}(x^k, \mathcal{X}_\star) = \|x^k - x_\star\|$ and the assumption of (ϑ-order) semismoothness ensures through Proposition 15.4.13 that $\mathcal{R}_\gamma$ is a (ϑ-order) LNA for R_γ at $x_\star$. Moreover, due to Lemma 15.4.16 there exists $c > 0$ such that $\|[H_k + \delta_k\mathrm{I}]^{-1}\| \leq c$ for all k's. We have

$$\begin{aligned} \|x^k + d^k - x_\star\| &= \left\|x^k + [H_k + \delta_k\mathrm{I}]^{-1}(e^k - \nabla\varphi_\gamma(x^k)) - x_\star\right\| \\ &\leq \left\|[H_k + \delta_k\mathrm{I}]^{-1}\right\| \left\|[H_k + \delta_k\mathrm{I}](x^k - x_\star) + e^k - \nabla\varphi_\gamma(x^k)\right\| \\ &\leq c\left\|H_k(x^k - x_\star) - \nabla\varphi_\gamma(x^k)\right\| + c\delta_k\|x^k - x_\star\| + c\|e^k\| \\ &= c\left\|Q_\gamma(x^k)\underbrace{\big(U_k(x^k - x_\star) - R_\gamma(x^k)\big)}\right\| + c\delta_k\|x^k - x_\star\| + c\|e^k\|. \end{aligned}$$

Since $\mathcal{R}_\gamma$ is a LNA at $x_\star$, it follows that the quantity emphasized in the bracket is a $o(\|x^k - x_\star\|)$, whereas in case of a (ϑ-order) LNA the tighter estimate $O(\|x^k - x_\star\|^{1+\vartheta})$ holds. Combined with the fact that $\delta_k = O(\|x^k - x_\star\|^\nu)$ and $\|e^k\| = O(\|x^k - x_\star\|^{1+\rho})$, we conclude that $(d^k)_{k\in\mathbb{N}}$ are superlinearly convergent directions, and with order at least $1 + \min\{\rho, \vartheta, \nu\}$ in case of ϑ-order semismoothness. □

□

Problems where the residual is (ϑ-order) semismooth are quite common. For instance, piecewise affine functions are everywhere strongly semismooth, as it is the case for the residual in lasso problems [67]. On the contrary, when the solution is not unique the condition $\|d^k\| \le D\|\nabla\varphi_\gamma(x^k)\|$ (or, equivalently, $\|d^k\| \le D'\|R_\gamma(x^k)\|$) is trickier. As detailed in [82, 83], this bound on the directions is ensured if $\rho = 1$ and for all iterates x^k and points x close enough to the limit point the following smoothness condition holds:

$$\|R_\gamma(x^k) + U_k(x - x^k)\| \le c\|x - x^k\|^2 \tag{15.83}$$

for some constant $c > 0$. This condition is implied by and closely related to local Lipschitz differentiability of R_γ and thus conservative. We remark that, however, this can be weakened by requiring $\rho \ge \nu$, and a notion of ϑ-order semismoothness at the limit point with some degree of uniformity on the set of solutions $\mathcal{X}_\star$, namely

$$\limsup_{\substack{x,x'\to x_\star \\ x'\in\mathcal{X}_\star,\, x\neq x' \\ U\in\mathcal{R}_\gamma(x)}} \frac{\|R_\gamma(x) + U(x' - x)\|}{\|x' - x\|^{1+\vartheta}} < \infty \tag{15.84}$$

for some $\vartheta \in [\nu, 1]$. This weakened requirement comes from the observation that point x in (15.83) is in fact $x^k_\star$, the projection of x^k onto $\mathcal{X}_\star$, set onto which R_γ is constant (equal to 0). To see this, notice that (15.84) implies that $\|R_\gamma(x^k) + U_k(x^k_\star - x^k)\| \le c\|x^k_\star - x^k\|^{1+\vartheta}$ for some $c > 0$. In particular, mimicking the arguments in the cited references, since $H_k \succeq 0$ and $\|Q_\gamma\| \le 1$, observe that

$$\|[H_k + \delta_k \mathrm{I}]^{-1} Q_\gamma(x^k)\| \le \|[H_k + \delta_k \mathrm{I}]^{-1}\| \le \delta_k^{-1} \tag{15.85a}$$

and

$$\|[H_k + \delta_k \mathrm{I}]^{-1} H_k\| = \|\mathrm{I} - \delta_k[H_k + \delta_k \mathrm{I}]^{-1}\| \le 2. \tag{15.85b}$$

Therefore,

$$\begin{aligned}
\|d^k\| &= \big\|[H_k + \delta_k \mathrm{I}]^{-1}\big(e^k - \nabla\varphi_\gamma(x^k)\big)\big\| \\
&\le \big\|[H_k + \delta_k \mathrm{I}]^{-1}\big\|\|e^k\| + \big\|[H_k + \delta_k \mathrm{I}]^{-1} Q_\gamma(x^k)\big(R_\gamma(x^k) + U_k(x^k_\star - x^k)\big)\big\| \\
&\quad + \big\|[H_k + \delta_k \mathrm{I}]^{-1} H_k(x^k_\star - x^k)\big\|
\end{aligned}$$

$$\begin{aligned}
&\overset{(15.85)}{\leq} \delta_k^{-1}\|\nabla\varphi_\gamma(x^k)\|^{1+\rho} + c\delta_k^{-1}\|x_\star^k - x^k\|^{1+\vartheta} + 2\|x_\star^k - x^k\| \\
&= \zeta^{-1}\|\nabla\varphi_\gamma(x^k)\|^{1+\rho-\nu} + c\zeta^{-1}\|x_\star^k - x^k\|^{1+\vartheta-\nu} + 2\|x_\star^k - x^k\| \\
&= O\big(\operatorname{dist}(x^k, \mathcal{X}_\star)^{1+\min\{0,\rho-\nu,\vartheta-\nu\}}\big),
\end{aligned}$$

which is indeed $O(\|\nabla\varphi_\gamma(x^k)\|)$ whenever $\nu \leq \min\{\vartheta, \rho\}$. Some comments are in order to expand on condition (15.84).

(i) If $\mathcal{X}_\star = \{\bar{x}\}$ is a singleton, then x' is fixed to $x_\star$ and the requirement reduces to ϑ-order semismoothness at $x_\star$.
(ii) This notion of uniformity is a *local property*: for any $\varepsilon > 0$ the set $\mathcal{X}_\star$ can be replaced by $\mathcal{X}_\star \cap \mathrm{B}(x_\star; \varepsilon)$.
(iii) The condition $U \in \mathcal{R}_\gamma(x)$ in the limit can be replaced by $U \in \hat{\mathcal{R}}_\gamma(x) := \big\{\gamma^{-1}(\mathrm{I} - PQ_\gamma(x)) \mid P \in \partial_B \operatorname{prox}_{\gamma g}(x - \gamma\nabla f(x))\big\}$, since $\mathcal{R}_\gamma(x) = \operatorname{conv}\big(\hat{\mathcal{R}}_\gamma(x)\big)$.

In particular, by exploiting this last condition it can be easily verified that if R_γ is piecewise ϑ-Hölder differentiable around $x_\star$, then (15.84) holds, yet the stronger requirement (15.83) in [82, 83] does not.

15.6 Generalized Jacobians of Proximal Mappings

In many interesting cases $\operatorname{prox}_{\gamma g}$ is PC^1 and thus semismooth. *Piecewise quadratic* (PWQ) functions comprise a special but important class of convex functions whose proximal mapping is PC^1. A convex function g is called PWQ if $\operatorname{dom} g$ can be represented as the union of finitely many polyhedral sets, relative to each of which $g(x)$ is given by an expression of the form $\frac{1}{2}\langle x, Hx\rangle + \langle q, x\rangle + c$ ($H \in \mathbb{R}^{n\times n}$ must necessarily be symmetric positive semidefinite) [65, Def. 10.20]. The class of PWQ functions is quite general since it includes *e.g.*, polyhedral norms, indicators and support functions of polyhedral sets, and it is closed under addition, composition with affine mappings, conjugation, inf-convolution and inf-projection [65, Prop.s 10.22 and 11.32]. It turns out that the proximal mapping of a PWQ function is *piecewise affine* (PWA) [65, 12.30] ($\mathbb{R}^n$ is partitioned in polyhedral sets relative to each of which $\operatorname{prox}_{\gamma g}$ is an affine mapping), hence strongly semismooth [19, Prop. 7.4.7]. Another example of a proximal mapping that is strongly semismooth is the projection operator over symmetric cones [73].

A big class with semismooth proximal mapping is formed by the semi-algebraic functions. We remind that a set $A \subseteq \mathbb{R}^n$ is *semi-algebraic* if it can be expressed as

$$A = \bigcup_{i=1}^{p} \bigcap_{j=1}^{q} \big\{x \in \mathbb{R}^m \mid P_{ij}(x) = 0,\ Q_{ij}(x) < 0\big\} \tag{15.86}$$

for some polynomial functions $P_{ij}, Q_{ij} : \mathbb{R}^n \to \mathbb{R}$, and that a function $h : \mathbb{R}^n \to \overline{\mathbb{R}}^m$ is *semi-algebraic* if $\operatorname{gph} h$ is a semi-algebraic subset of $\mathbb{R}^{n+m}$.

Proposition 15.6.1 *If $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ is semi-algebraic, then so are g^γ and $\operatorname{prox}_{\gamma g}$. In particular, g^γ and $\operatorname{prox}_{\gamma g}$ are semismooth.*

Proof Since g^γ and $\operatorname{prox}_{\gamma g}$ are both Lipschitz continuous, semismoothness will follow once we show that they are semi-algebraic [11, Rem. 4]. Every polynomial is clearly semi-algebraic, and since the property is preserved under addition [10, Prop. 2.2.6(ii)], the function $(x, w) \mapsto g(w) + \frac{1}{2\gamma}\|w - x\|^2$ is semi-algebraic. Moreover, since parametric minimization of a semi-algebraic function is still semi-algebraic (see, *e.g.*, [1, §2]), it follows that the Moreau envelope g^γ is semi-algebraic and therefore so is $h(x, w) := g(w) + \frac{1}{2\gamma}\|w - x\|^2 - g^\gamma(x)$. Notice that $\operatorname{prox}_{\gamma g}(x) = \{w \in \mathbb{R}^n \mid h(x, w) \le 0\}$, therefore

$$\begin{aligned}\operatorname{gph} \operatorname{prox}_{\gamma g} &= \left\{(x, \bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \operatorname{prox}_{\gamma g}(x) = \bar{x}\right\} \\ &= \left\{(x, \bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n \mid h(x, \bar{x}) \le 0\right\} \\ &= h^{-1}((-\infty, 0])\end{aligned}$$

is a semi-algebraic set, since the interval $(-\infty, 0]$ is clearly semi-algebraic and thus so is $h^{-1}((-\infty, 0])$ [10, Prop. 2.2.7]. □

In fact, with the same arguments it can be shown that the result still holds if "semi-algebraic" is replaced with the broader notion of "*tame*", see [11]. Other conditions that guarantee semismoothness of the proximal mapping can be found in [46–48, 50]. The rest of the section is devoted to collecting explicit formulas of $\partial_C \operatorname{prox}_{\gamma g}$ for many known useful instances of convex functions g.

15.6.1 Properties

a. Separable functions.
Whenever g is (block) separable, *i.e.*, $g(x) = \sum_{i=1}^N g_i(x_i)$, $x_i \in \mathbb{R}^{n_i}$, $\sum_{i=1}^N n_i = n$, then every $P \in \partial_C(\operatorname{prox}_{\gamma g})(x)$ is a (block) diagonal matrix. This has favorable computational implications especially for large-scale problems. For example, if g is the ℓ_1 norm or the indicator function of a box, then the elements of $\partial_C \operatorname{prox}_{\gamma g}(x)$ (or $\partial_B \operatorname{prox}_{\gamma g}(x)$) are diagonal matrices with diagonal elements in $[0, 1]$ (or in $\{0, 1\}$).

b. Convex conjugate.
With a simple application of the Moreau's decomposition [2, Thm. 14.3(ii)], all elements of $\partial_C \operatorname{prox}_{\gamma g^*}$ are readily available as long as one can compute $\partial_C \operatorname{prox}_{g/\gamma}$. Specifically,

$$\partial_C(\operatorname{prox}_{\gamma g^*})(x) = \mathrm{I} - \partial_C(\operatorname{prox}_{g/\gamma})(x/\gamma). \tag{15.87}$$

c. Support function.
The *support function* of a nonempty closed and convex set D is the proper convex and lsc function $\sigma_D(x) := \sup_{y\in D}\langle x, y\rangle$. Alternatively, σ_D can be expressed as the convex conjugate of the indicator function δ_D, and one can use the results of Section §15.6.1b to find that

$$\partial_C(\mathrm{prox}_{\gamma g})(x) = \mathrm{I} - \partial_C(\mathscr{P}_D)(x/\gamma). \tag{15.88}$$

Section 15.6.2 offers a rich list of sets D for which a closed form expression exists.

d. Spectral functions.
The eigenvalue function $\lambda : \mathrm{S}(\mathbb{R}^{n\times n}) \to \mathbb{R}^n$ returns the vector of eigenvalues of a symmetric matrix in nonincreasing order. *Spectral functions* are of the form

$$G := h \circ \lambda : \mathrm{S}(\mathbb{R}^{n\times n}) \to \overline{\mathbb{R}}. \tag{15.89}$$

where $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, convex, and *symmetric*, *i.e.*, invariant under coordinate permutations [35]. Such G inherits most of the properties of h [36, 37]; in particular, its proximal mapping is [56, §6.7]

$$\mathrm{prox}_{\gamma G}(X) = Q\,\mathrm{diag}(\mathrm{prox}_{\gamma h}(\lambda(X)))Q^\top, \tag{15.90}$$

where $X = Q\,\mathrm{diag}(\lambda(X))Q^\top$ is the spectral decomposition of X (Q is an orthogonal matrix). If, additionally,

$$h(x) = g(x_1) + \cdots + g(x_N) \tag{15.91}$$

for some $g : \mathbb{R} \to \overline{\mathbb{R}}$, then

$$\mathrm{prox}_{\gamma h}(x) = (\mathrm{prox}_{\gamma g}(x_1), \ldots, \mathrm{prox}_{\gamma g}(x_N)), \tag{15.92}$$

and therefore the proximal mapping of G can be expressed as

$$\mathrm{prox}_{\gamma G}(X) = Q\,\mathrm{diag}(\mathrm{prox}_{\gamma g}(\lambda_1(X)), \ldots, \mathrm{prox}_{\gamma g}(\lambda_n(X)))Q^\top, \tag{15.93}$$

[9, Chap. V], [29, Sec. 6.2]. Now we can use the theory of nonsmooth symmetric matrix-valued functions developed in [14] to analyze differentiability properties of $\mathrm{prox}_{\gamma G}$. In particular, $\mathrm{prox}_{\gamma G}$ is (strongly) semismooth at X iff $\mathrm{prox}_{\gamma g}$ is (strongly) semismooth at the eigenvalues of X [14, Prop. 4.10]. Moreover, for any $X \in \mathrm{S}(\mathbb{R}^{n\times n})$ and $P \in \partial_B(\mathrm{prox}_{\gamma G})(X)$ we have [14, Lem. 4.7]

$$P(X) = Q\left(\Omega^{\gamma g}_{\lambda,\lambda} \odot (Q^\top X Q)\right)Q^\top, \tag{15.94}$$

where $\odot$ denotes the Hadamard product and for vectors $u, v \in \mathbb{R}^n$ we defined $\Omega^{\gamma g}_{u,v}$ as the $n \times n$ matrix

$$(\Omega^{\gamma g}_{u,v})_{ij} := \begin{cases} \partial_B \operatorname{prox}_{\gamma g}(u_i) & \text{if } u_i = v_j, \\ \left\{ \frac{\operatorname{prox}_{\gamma g}(u_i) - \operatorname{prox}_{\gamma g}(v_j)}{u_i - v_j} \right\} & \text{otherwise.} \end{cases} \tag{15.95}$$

e. Orthogonally invariant functions. A function $G : \mathbb{R}^{m\times n} \to \overline{\mathbb{R}}$ is called *orthogonally invariant* if $G(UXV^\top) = G(X)$ for all $X \in \mathbb{R}^{m\times n}$ and orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$.[4]

A function $h : \mathbb{R}^q \to \overline{\mathbb{R}}$ is *absolutely symmetric* if $h(Qx) = h(x)$ for all $x \in \mathbb{R}^q$ and any generalized permutation matrix Q, *i.e.*, a matrix $Q \in \mathbb{R}^{q\times q}$ that has exactly one nonzero entry in each row and each column, that entry being ± 1 [34]. There is a one-to-one correspondence between orthogonally invariant functions on $\mathbb{R}^{m\times n}$ and absolutely symmetric functions on $\mathbb{R}^q$. Specifically, if G is orthogonally invariant, then

$$G(X) = h(\sigma(X)) \tag{15.96}$$

for the absolutely symmetric function $h(x) = G(\operatorname{diag}(x))$. Here, for $X \in \mathbb{R}^{m\times n}$ and $q := \min\{m, n\}$ the spectral function $\sigma : \mathbb{R}^{m\times n} \to \mathbb{R}^q$ returns the vector of its singular values in nonincreasing order. Conversely, if h is absolutely symmetric, then $G(X) = h(\sigma(X))$ is orthogonally invariant. Therefore, convex analytic and generalized differentiability properties of orthogonally invariant functions can be easily derived from those of the corresponding absolutely symmetric functions [34]. For example, assuming for simplicity that $m \leq n$, the proximal mapping of G is given by [56, Sec. 6.7]

$$\operatorname{prox}_{\gamma G}(X) = U \operatorname{diag}(\operatorname{prox}_{\gamma h}(\sigma(X)))V_1^\top, \tag{15.97}$$

where $X = U \begin{bmatrix} \operatorname{diag}(\sigma(X)) & 0 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^\top$ is the singular value decomposition of X. If we further assume that h has a separable form as in (15.91), then

$$\operatorname{prox}_{\gamma G}(X) = U \Sigma_g(X) V_1^\top, \tag{15.98}$$

where $\Sigma_g(X) = \operatorname{diag}(\operatorname{prox}_{\gamma g}(\sigma_1(X)), \ldots, \operatorname{prox}_{\gamma g}(\sigma_n(X)))$. Functions of this form are called *nonsymmetric matrix-valued functions*. We also assume that g is a non-negative function such that $g(0) = 0$. This implies that $\operatorname{prox}_{\gamma g}(0) = 0$ and guarantees that the nonsymmetric matrix-valued function (15.98) is well defined [80, Prop. 2.1.1]. Now we can use the results of [80, §2] to draw conclusions about generalized differentiability properties of $\operatorname{prox}_{\gamma G}$.

[4] In case of complex-valued matrices, functions of this form are known as *unitarily invariant* [34].

For example, through [80, Thm. 2.27] we have that $\mathrm{prox}_{\gamma G}$ is continuously differentiable at X if and only if $\mathrm{prox}_{\gamma g}$ is continuously differentiable at the singular values of X. Furthermore, $\mathrm{prox}_{\gamma G}$ is (strongly) semismooth at X if $\mathrm{prox}_{\gamma g}$ is (strongly) semismooth at the singular values of X [80, Thm. 2.3.11]. For any $X \in \mathbb{R}^{m\times n}$ the generalized Jacobian $\partial_B(\mathrm{prox}_{\gamma G})(X)$ is well defined and nonempty, and any $P \in \partial_B(\mathrm{prox}_{\gamma G})(X)$ acts on $H \in \mathbb{R}^{m\times n}$ as [80, Prop. 2.3.7]

$$P(H) = U\left[\left(\Omega^{\gamma g}_{\sigma,\sigma} \odot \left(\frac{H_1+H_1^\top}{2}\right) + \Omega^{\gamma g}_{\sigma,-\sigma} \odot \left(\frac{H_1-H_1^\top}{2}\right)\right),\ (\Omega^{\gamma g}_{\sigma,0} \odot H_2)\right] V^\top, \tag{15.99}$$

where $V = [V_1\ V_2]$, $H_1 = U^\top H V_1 \in \mathbb{R}^{m\times m}$, $H_2 = U^\top H V_2 \in \mathbb{R}^{m\times(n-m)}$ and matrices Ω are as in (15.95).

15.6.2 Indicator Functions

Smooth constrained convex problems

$$\text{minimize}_{x\in\mathbb{R}^n}\ f(x) \quad \text{subject to } x \in D \tag{15.100}$$

can be cast in the composite form (15.1) by encoding the feasible set D with the indicator function $g = \delta_D$. Whenever $\mathscr{P}_D$ is efficiently computable, then algorithms like the forward-backward splitting (15.21) can be conveniently considered. In the following we analyze the generalized Jacobian of some of such projections.

a. Affine sets. $D = \{x \in \mathbb{R}^n \mid Ax = b\}$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$.
In this case, $\mathscr{P}_D(x) = x - A^\dagger(Ax - b)$ where $A^\dagger$ is the Moore-Penrose pseudoinverse of A. For example, if A is surjective (*i.e.*, it has full row rank and thus $m \le n$), then $A^\dagger = A^\top(AA^\top)^{-1}$, whereas if it is injective (*i.e.*, it has full column rank and thus $m \ge n$), then $A^\dagger = (A^\top A)^{-1}A^\top$. Obviously $\mathscr{P}_D$ is an affine mapping, thus everywhere differentiable with

$$\partial_C(\mathscr{P}_D)(x) = \partial_B(\mathscr{P}_D)(x) = \{\nabla\mathscr{P}_D(x)\} = \left\{\mathrm{I} - A^\dagger A\right\}. \tag{15.101}$$

b. Polyhedral sets. $D = \{x \in \mathbb{R}^n \mid Ax = b,\ Cx \le d\}$, for some $A \in \mathbb{R}^{p\times n}$, $b \in \mathbb{R}^p$, $C \in \mathbb{R}^{q\times n}$ and $d \in \mathbb{R}^q$.
It is well known that $\mathscr{P}_D$ is piecewise affine. In particular, let

$$\mathcal{I}_D \coloneqq \left\{I \subseteq \{1\dots q\} \mid \exists x \in \mathbb{R}^n : Ax = b,\ C_i.x = d_i\ i \in I,\ C_j.x < d_j\ j \notin I\right\}. \tag{15.102}$$

Then, the *faces* of D can be indexed with the elements of $\mathcal{I}$ [66, Prop. 2.1.3]: for each $I \in \mathcal{I}_D$ let

$$F_I := \{x \in D \mid C_i.x = d_i,\ i \in I\}$$

be the I-th face of D,

$$S_I := \operatorname{aff} F_I = \left\{x \in \mathbb{R}^n \mid Ax = b,\ C_i.x = d_i,\ i \in I\right\}$$

be the hyperplane containing the I-th face of D,

$$N_I := \operatorname{ran} A^\top + \operatorname{cone}\left\{C_I^\top\right\}$$

be the normal cone to any point in the relative interior of F_I [66, Eq. (2.44)],[5] and

$$R_I := F_I + N_I.$$

We then have $\mathscr{P}_D(x) \in \left\{\mathscr{P}_{S_I}(x) \mid I \in \mathcal{I}_D\right\}$, *i.e.*, $\mathscr{P}_D$ is a piecewise affine function. The affine pieces of $\mathscr{P}_D$ are the projections on the corresponding affine subspaces S_I (cf. Section §15.6.2a). In fact, for each $x \in R_I$ we have $\mathscr{P}_D(x) = \mathscr{P}_{S_I}(x)$, each R_I is full dimensional and $\mathbb{R}^n = \bigcup_{I \in \mathcal{I}_D} R_I$ [66, Prop.s 2.4.4 and 2.4.5]. For each $I \in \mathcal{I}_D$ let

$$P_I := \nabla \mathscr{P}_{S_I} = \mathrm{I} - \begin{pmatrix} A \\ C_I \end{pmatrix}^\dagger \begin{pmatrix} A \\ C_I \end{pmatrix}, \tag{15.103}$$

and for each $x \in \mathbb{R}^n$ let

$$\mathcal{I}_D(x) := \{I \in \mathcal{I}_D \mid x \in R_I\}. \tag{15.104}$$

Then,

$$\partial_C(\mathscr{P}_D)(x) = \operatorname{conv} \partial_B(\mathscr{P}_D)(x) = \operatorname{conv}\{P_I \mid I \in \mathcal{I}_D(x)\}. \tag{15.105}$$

Therefore, an element of $\partial_B \mathscr{P}_D(x)$ is P_I as in (15.103) where $I = \{i \mid C_i \bar{x} = d_i\}$ is the set of active constraints of $\bar{x} = \mathscr{P}_D(x)$. For a more general analysis we refer the reader to [27, 39].

[5] Consistently with the definition in [66], the polyhedron P can equivalently be expressed by means of only inequalities as $P = \{x \in \mathbb{R}^n \mid Ax \le b,\ -Ax \le -b,\ Cx \le b\}$, resulting indeed in $\operatorname{cone}[A^\top,\ -A^\top,\ C^\top] = \operatorname{ran} A^\top + \operatorname{cone} C^\top$.

c. Halfspaces. $H = \{x \in \mathbb{R}^n \mid \langle a, x\rangle \leq b\}$ for some $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then, denoting the *positive part* of $r \in \mathbb{R}$ as $[r]_+ := \max\{0, r\}$,

$$\mathscr{P}_H(x) = x - \tfrac{[\langle a,x\rangle - b]_+}{\|a\|^2} a$$

and

$$\partial_C(\mathscr{P}_H)(x) = \begin{cases} \left\{\mathrm{I} - \|a\|^{-2} aa^\top\right\} & \text{if } x \notin H, \\ \{\mathrm{I}\} & \text{if } \langle a, x\rangle < b, \\ \operatorname{conv}\left\{\mathrm{I}, \mathrm{I} - \|a\|^{-2} aa^\top\right\} & \text{if } \langle a, x\rangle = b. \end{cases}$$

d. Boxes. $D = \{x \in \mathbb{R}^n \mid \ell \leq x \leq u\}$ for some $\ell, u \in [-\infty, \infty]^n$. We have

$$\mathscr{P}_D(x) = \min\{\max\{x, \ell\}, u\},$$

and since the corresponding indicator function δ_D is separable, every element of $\partial_C(\mathscr{P}_D)(x)$ is diagonal with (cf. Section §15.6.1a)

$$\partial_C(\mathscr{P}_D)(x)_{ii} = \begin{cases} [0,1] & \text{if } x_i \in \{\ell_i, u_i\}, \\ \{1\} & \text{if } \ell_i < x_i < u_i, \\ \{0\} & \text{otherwise,} \end{cases}$$

e. Unit simplex. $D = \left\{x \in \mathbb{R}^n \mid x \geq 0,\ \sum_{i=1}^n x_i = 1\right\}$. By writing down the optimality conditions for the corresponding projection problem, one can easily see that

$$\mathscr{P}_D(x) = [x - \lambda \mathbf{1}]_+, \tag{15.106}$$

where λ solves $\langle \mathbf{1}, [x - \lambda\mathbf{1}]_+\rangle = 1$. Since the unit simplex is a polyhedral set, we are dealing with a special case of Section §15.6.2b, where $A = \mathbf{1}^\top$, $b = 1$, $C = -\mathrm{I}$, and $d = 0$. Therefore, in order to calculate an element of the generalized Jacobian of the projection, we first compute $\mathscr{P}_D(x)$ and then determine the set of active indices $J := \{i \mid \mathscr{P}_D(x)_i = 0\}$. An element $P \in \partial_B(\mathscr{P}_D)(x)$ is given by

$$P_{ij} = \begin{cases} \delta_{i,j} - \frac{1}{n - |J|} & \text{if } i, j \notin J, \\ 0 & \text{otherwise,} \end{cases} \tag{15.107}$$

where $|J|$ denotes the cardinality of the set J. Notice that P is block diagonal after a permutation of rows and columns.

f. Euclidean unit ball. $B = \overline{\mathrm{B}}(0; 1)$.
We have

$$\mathscr{P}_B(x) = \begin{cases} x & \text{if } x \in B, \\ x/\|x\| & \text{otherwise,} \end{cases}$$

and

$$\partial_C(\mathscr{P}_B)(x) = \begin{cases} \{\mathrm{I}\} & \text{if } \|x\| < 1, \\ \operatorname{conv}\left\{\|x\|^{-1}(\mathrm{I} - ww^\top),\ \mathrm{I}\right\} & \text{if } \|x\| = 1, \\ \left\{\|x\|^{-1}(\mathrm{I} - ww^\top)\right\} & \text{if } x \notin B, \end{cases}$$

where $w \coloneqq x/\|x\|$.

g. Second-order cone. $\mathcal{K} = \left\{(x_0, \bar{x}) \in \mathbb{R} \times \mathbb{R}^{n-1} \mid x_0 \geq \|\bar{x}\|\right\}$.
Let $x \coloneqq (x_0, \bar{x})$, and for $w \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ define

$$M_{w,\alpha} \coloneqq \tfrac{1}{2}\begin{bmatrix} 1 & w^\top \\ w & (1-\alpha)\mathrm{I}_{n-1} + \alpha ww^\top \end{bmatrix}. \tag{15.108}$$

Then, $\partial_C(\mathscr{P}_{\mathcal{K}})(x) = \operatorname{conv}(\partial_B(\mathscr{P}_{\mathcal{K}})(x))$ where, for $\bar{w} \coloneqq \bar{x}/\|\bar{x}\|$ and $\bar{\alpha} \coloneqq -x_0/\|\bar{x}\|$, we have [30, Lem. 2.6]

$$\partial_B(\mathscr{P}_{\mathcal{K}})(x) = \begin{cases} \{0\} & \text{if } x_0 < -\|\bar{x}\|, \\ \{\mathrm{I}_n\} & \text{if } x_0 > \|\bar{x}\|, \\ \left\{M_{\bar{w},\bar{\alpha}}\right\} & \text{if } -\|\bar{x}\| < x_0 < \|\bar{x}\|, \\ \left\{\mathrm{I}_n,\ M_{\bar{w},\bar{\alpha}}\right\} & \text{if } x_0 = \|\bar{x}\| \neq 0, \\ \left\{0,\ M_{\bar{w},\bar{\alpha}}\right\} & \text{if } x_0 = -\|\bar{x}\| \neq 0, \\ \{0, \mathrm{I}_n\} \cup \left\{M_{w,\alpha} \mid |\alpha| \leq 1,\ \|w\| \leq 1\right\} & \text{if } x_0 = \bar{x} = 0. \end{cases} \tag{15.109}$$

h. Positive semidefinite cone. $\mathcal{S}_+ = \mathrm{S}_+(\mathbb{R}^{n\times n})$.
For any symmetric matrix M it holds that

$$\mathscr{P}_{\mathcal{S}_+}(M) = Q[\operatorname{diag}(\lambda)]_+ Q^\top, \tag{15.110}$$

where $M = Q\operatorname{diag}(\lambda)Q^\top$ is any spectral decomposition of M. This coincides with (15.93), as $\delta_{\mathcal{S}_+}$ can be expressed as in (15.89), where h has the separable form (15.91) with $g = \delta_{\mathbb{R}_+}$, so that for $r \in \mathbb{R}$ we have

$$\operatorname{prox}_{\gamma g}(r) = [r]_+ \quad \text{and} \quad \partial_B(\operatorname{prox}_{\gamma g})(r) = \begin{cases} \{0\} & \text{if } r < 0, \\ \{0, 1\} & \text{if } r = 0, \\ \{1\} & \text{if } r > 0. \end{cases} \tag{15.111}$$

An element of $\partial_B\, \mathscr{P}_{\mathrm{S}_+(\mathbb{R}^{n\times n})}(X)$ is thus given by (15.94).

15.6.3 Norms

a. ℓ_1 **norm.** $g(x) = \|x\|_1$.
The proximal mapping is the well-known soft-thresholding operator

$$(\text{prox}_{\gamma g}(x))_i = \text{sign}(x_i)[|x_i| - \gamma]_+, \quad i = 1, \ldots, n. \tag{15.112}$$

Function g is separable, and thus every element of $\partial_B(\text{prox}_{\gamma g})$ is a diagonal matrix, cf. Section §15.6.1a. Specifically, the nonzero elements are

$$\partial_C(\text{prox}_{\gamma g})(x)_{ii} = \begin{cases} \{1\} & \text{if } |x_i| > \gamma, \\ [0, 1] & \text{if } |x_i| = \gamma, \\ \{0\} & \text{if } |x_i| < \gamma. \end{cases} \tag{15.113}$$

We could also arrive to the same conclusion by applying the Moreau decomposition of Section §15.6.1b to the function of Section §15.6.2d with $u = -\ell = \mathbf{1}_n$, since the ℓ_1 norm is the conjugate of the indicator of the ℓ_∞-norm ball.

b. ℓ_∞ **norm.** $g(x) = \|x\|_\infty$.
Function g is the convex conjugate of the indicator of the unit simplex D analyzed in Section §15.6.2e. From the Moreau decomposition, see Section §15.6.1b, we obtain

$$\partial_C(\text{prox}_{\gamma g})(x) = \text{I} - \partial_C(\mathscr{P}_D)(x/\gamma). \tag{15.114}$$

Then, $\mathscr{P}_D(x/\gamma) = [x/\gamma - \lambda\mathbf{1}]_+$ where $\lambda \in \mathbb{R}$ solves $\langle \mathbf{1}, [x/\gamma - \lambda\mathbf{1}]_+ \rangle = 1$. Let $J = \{i \mid \mathscr{P}_D(x/\gamma)_i = 0\}$, then an element of $\partial_B(\text{prox}_{\gamma g})(x)$ is given by

$$P_{ij} = \begin{cases} \frac{1}{n-|J|} & \text{if } i, j \notin J, \\ \delta_{i,j} & \text{otherwise.} \end{cases} \tag{15.115}$$

c. **Euclidean norm.** $g(x) = \|x\|$.
The proximal mapping is given by

$$\text{prox}_{\gamma g}(x) = \begin{cases} (1 - \gamma\|x\|^{-1})x & \text{if } \|x\| \geq \gamma, \\ 0 & \text{otherwise.} \end{cases} \tag{15.116}$$

Since $\text{prox}_{\gamma g}$ is a PC^1 mapping, its B-subdifferential can be computed by simply computing the Jacobians of its smooth pieces. Specifically, denoting $w = x/\|x\|$ we have

$$\partial_C(\text{prox}_{\gamma g})(x) = \begin{cases} \left\{\text{I} - \gamma\|x\|^{-1}(\text{I} - ww^\top)\right\} & \text{if } \|x\| > \gamma, \\ \{0\} & \text{if } \|x\| < \gamma, \\ \text{conv}\left\{\text{I} - \gamma\|x\|^{-1}(\text{I} - ww^\top), 0\right\} & \text{otherwise.} \end{cases} \tag{15.117}$$

d. Sum of Euclidean norms. $g(x) = \sum_{s\in\mathcal{S}} \|x_s\|$, where $\mathcal{S}$ is a partition of $\{1,\dots,n\}$.
Differently from the ℓ_1-norm which induces sparsity on the whole vector, this function serves as regularizer to induce group sparsity [81]. For $s \in \mathcal{S}$, the components of the proximal mapping indexed by s are

$$(\mathrm{prox}_{\gamma g}(x))_s = (1 - \gamma\|x_s\|^{-1})_+ x_s. \tag{15.118}$$

Any $P \in \partial_B(\mathrm{prox}_{\gamma g})(x)$ is block diagonal with the s-block equal to

$$P_s = \begin{cases} \mathrm{I} - \gamma\|x_s\|^{-1}\big(\mathrm{I} - \|x_s\|^{-2}x_s x_s^\top\big) & \text{if } \|x_s\| > \gamma, \\ \mathrm{I} & \text{if } \|x_s\| < \gamma, \\ \text{any of these two matrices} & \text{if } \|x_s\| = \gamma. \end{cases} \tag{15.119}$$

e. Matrix nuclear norm. $G(X) = \|X\|_\star$ for $X \in \mathbb{R}^{m\times n}$.
The *nuclear norm* returns the sum of the singular values of a matrix $X \in \mathbb{R}^{m\times n}$, *i.e.*, $G(X) = \sum_{i=1}^m \sigma_i(X)$ (for simplicity we are assuming that $m \le n$). It serves as a convex surrogate for the rank, and has found many applications in systems and control theory, including system identification and model reduction [20–22, 41, 61]. Other fields of application include *matrix completion problems* arising in machine learning [62, 68] and computer vision [52, 76], and *nonnegative matrix factorization problems* arising in data mining [18].
The nuclear norm can be expressed as $G(X) = h(\sigma(X))$, where $h(x) = \|x\|_1$ is absolutely symmetric and separable. Specifically, it takes the form (15.91) with $g = |\cdot|$, for which $g(0) = 0$ and $0 \in \partial g(0)$, and whose proximal mapping is the soft-thresholding operator. In fact, since the case of interest here is $x \ge 0$ (because $\sigma_i(X) \ge 0$), we have $\mathrm{prox}_{\gamma g}(x) = [x - \gamma]_+$, cf. (15.116). Consequently, the proximal mapping of $\|X\|_\star$ is given by (15.98) with

$$\Sigma_g(X) = \mathrm{diag}([\sigma_1(X) - \gamma]_+, \dots, [\sigma_m(X) - \gamma]_+). \tag{15.120}$$

For $x \in \mathbb{R}_+$ we have that

$$\partial_C(\mathrm{prox}_{\gamma g})(x) = \begin{cases} 0 & \text{if } 0 \le x < \gamma, \\ [0,1] & \text{if } x = \gamma, \\ 1 & \text{if } x > \gamma, \end{cases} \tag{15.121}$$

then $\partial_B(\mathrm{prox}_{\gamma G})(X)$ takes the form as in (15.99).

15.7 Conclusions

A forward-backward truncated-Newton method (FBTN) is proposed that minimizes the sum of two convex functions one of which Lipschitz continuous and twice continuously differentiable. Our approach is based on the forward-backward

envelope (FBE), a continuously differentiable tight lower bound to the original (nonsmooth and extended-real valued) cost function sharing minima and minimizers. The method requires forward-backward steps, Hessian evaluations of the smooth function and Clarke Jacobians of the proximal map of the nonsmooth term. Explicit formulas of Clarke Jacobians of a wide variety of useful nonsmooth functions are collected from the literature for the reader's convenience. The higher-order operations are needed for the computation of symmetric and positive semidefinite matrices that serve as surrogate for the Hessian of the FBE, allowing for a generalized (regularized, truncated-) Newton method for its minimization. The algorithm exhibits global Q-linear convergence under an error bound condition, and Q-superlinear or even Q-quadratic if an additional semismoothness assumption at the limit point is satisfied.

Acknowledgements This work was supported by the *Research Foundation Flanders (FWO)* research projects G086518N and G086318N; *KU Leuven internal funding* StG/15/043; *Fonds de la Recherche Scientifique—FNRS and the Fonds Wetenschappelijk Onderzoek—Vlaanderen* under EOS Project no 30468160 (SeLMA).

Appendix: Auxiliary Results

Lemma 1 *Any proper lsc convex function with nonempty and bounded set of minimizers is level bounded.*

Proof Let h be such function; to avoid trivialities we assume that $\operatorname{dom} h$ is unbounded. Fix $x_\star \in \operatorname{argmin} h$ and let $R > 0$ be such that $\operatorname{argmin} h \subseteq B := \mathrm{B}(x_\star; R)$. Since $\operatorname{dom} h$ is closed, convex, and unbounded, it holds that h attains a minimum on the compact set $\operatorname{bdry} B$, be it m, which is strictly larger than $h(x_\star)$ (since $\operatorname{dist}(\operatorname{argmin} h, \operatorname{bdry} B) > 0$ due to compactness of $\operatorname{argmin} h$ and openness of B). For $x \notin B$, let $s_x = x_\star + R\frac{x-x_\star}{\|x-x_\star\|}$ denote its projection onto $\operatorname{bdry} B$, and let $t_x := \frac{\|x-x_\star\|}{R} \geq 1$. Then,

$$h(x) = h\big(x_\star + t_x(s_x - x_\star)\big) \geq h(x_\star) + t_x\big(h(s_x) - h(x_\star)\big) \geq h(x_\star) + t_x\big(m - h(x_\star)\big)$$

where in the first inequality we used the fact that $t_x \geq 1$. Since $m - h(x_\star) > 0$ and $t_x \to \infty$ as $\|x\| \to \infty$, we conclude that h is coercive, and thus level bounded. □

Lemma 2 *Let $H \in \mathrm{S}_+(\mathbb{R}^n)$ with $\lambda_{\max}(H) \leq 1$. Then $H - H^2 \in \mathrm{S}_+(\mathbb{R}^n)$ with*

$$\lambda_{\min}(H - H^2) = \min\left\{\lambda_{\min}(H)(1 - \lambda_{\min}(H)),\ \lambda_{\max}(H)(1 - \lambda_{\max}(H))\right\}. \tag{15.122}$$

Proof Consider the spectral decomposition $H = S^\top D S$ for some orthogonal matrix S and diagonal D. Then, $H - H^2 = S^\top \tilde{D} S$ where $\tilde{D} = D - D^2$. Apparently, $\tilde{D}$ is diagonal, hence the eigenvalues of $H - H^2$ are exactly $\big\{\lambda - \lambda^2 \mid \lambda \in \operatorname{eigs}(H)\big\}$.

The function $\lambda \mapsto \lambda - \lambda^2$ is concave, hence the minimum in $\operatorname{eigs}(\tilde{H})$ is attained at one extremum, that is, either at $\lambda = \lambda_{\min}(H)$ or $\lambda = \lambda_{\max}(H)$, which proves the claim. □

Lemma 3 *For any $\gamma \in (0, 2/L_f)$ the forward-backward operator T_γ (15.22) is nonexpansive (in fact, $\frac{2}{4-\gamma L_f}$-averaged), and the residual R_γ is Lipschitz continuous with modulus $\frac{4}{\gamma(4-\gamma L_f)}$.*

Proof By combining [2, Prop. 4.39 and Cor. 18.17] it follows that the *gradient descent operator* $x \mapsto x - \gamma \nabla f(x)$ is $\gamma L_f/2$-averaged. Moreover, since the proximal mapping is $1/2$-averaged [2, Prop. 12.28] we conclude from [2, Prop. 4.44] that the forward-backward operator T_γ is α-averaged with $\alpha = \frac{2}{4-\gamma L_f}$, thus nonexpansive [2, Rem. 4.34(i)]. Therefore, by definition of α-averagedness there exists a 1-Lipschitz continuous operator S such that $T_\gamma = (1-\alpha)\operatorname{id} + \alpha S$ and consequently the residual $R_\gamma = \frac{1}{\gamma}\big(\operatorname{id} - T_\gamma\big) = \frac{\alpha}{\gamma}(\operatorname{id} - S)$ is $(2\alpha/\gamma)$-Lipschitz continuous. □

References

1. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Mathematical Programming **137**(1), 91–129 (2013). DOI 10.1007/s10107-011-0484-9
2. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. CMS Books in Mathematics. Springer (2017). DOI 10.1007/978-3-319-48311-5
3. Bauschke, H.H., Noll, D., Phan, H.M.: Linear and strong convergence of algorithms involving averaged nonexpansive operators. Journal of Mathematical Analysis and Applications **421**(1), 1–20 (2015)
4. Beck, A.: First-Order Methods in Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA (2017). DOI 10.1137/1.9781611974997
5. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009). DOI 10.1137/080716542
6. Becker, S., Fadili, J.: A quasi-Newton proximal splitting method. In: Advances in Neural Information Processing Systems, pp. 2618–2626 (2012)
7. Bertsekas, D.P.: Constrained optimization and lagrange multiplier methods. Computer Science and Applied Mathematics, Boston: Academic Press, 1982 (1982)
8. Bertsekas, D.P.: Convex Optimization Algorithms. Athena Scientific (2015)
9. Bhatia, R.: Matrix Analysis. Graduate Texts in Mathematics. Springer New York (1997)
10. Bochnak, J., Coste, M., Roy, M.F.: Real Algebraic Geometry. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg (2013)
11. Bolte, J., Daniilidis, A., Lewis, A.: Tame functions are semismooth. Mathematical Programming **117**(1), 5–19 (2009). DOI 10.1007/s10107-007-0166-9
12. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM Journal on Optimization **3**(3), 538–543 (1993). DOI 10.1137/0803026
13. Chen, X., Fukushima, M.: Proximal quasi-Newton methods for nondifferentiable convex optimization. Mathematical Programming **85**(2), 313–334 (1999). DOI 10.1007/s101070050059

14. Chen, X., Qi, H., Tseng, P.: Analysis of nonsmooth symmetric-matrix-valued functions with applications to semidefinite complementarity problems. SIAM Journal on Optimization **13**(4), 960–985 (2003). DOI 10.1137/S1052623400380584
15. Clarke, F.H.: Optimization and Nonsmooth Analysis. Society for Industrial and Applied Mathematics (1990). DOI 10.1137/1.9781611971309
16. Combettes, P.L., Pesquet, J.C.: Proximal Splitting Methods in Signal Processing, pp. 185–212. Springer New York, New York, NY (2011). DOI 10.1007/978-1-4419-9569-8_10
17. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Mathematics of Operations Research (2018)
18. Eldén, L.: Matrix Methods in Data Mining and Pattern Recognition. Society for Industrial and Applied Mathematics (2007). DOI 10.1137/1.9780898718867
19. Facchinei, F., Pang, J.S.: Finite-dimensional variational inequalities and complementarity problems, vol. II. Springer (2003)
20. Fazel, M.: Matrix rank minimization with applications. Ph.D. thesis, Stanford University (2002)
21. Fazel, M., Hindi, H., Boyd, S.P.: A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the 2001 American Control Conference, vol. 6, pp. 4734–4739 (2001). DOI 10.1109/ACC.2001.945730
22. Fazel, M., Hindi, H., Boyd, S.P.: Rank minimization and applications in system theory. In: Proceedings of the 2004 American Control Conference, vol. 4, pp. 3273–3278 vol.4 (2004). DOI 10.23919/ACC.2004.1384521
23. Fukushima, M.: Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems. Mathematical Programming **53**(1), 99–110 (1992). DOI 10.1007/BF01585696
24. Giselsson, P., Fält, M.: Envelope functions: Unifications and further properties. Journal of Optimization Theory and Applications (2018). DOI 10.1007/s10957-018-1328-z
25. Gowda, M.S.: Inverse and implicit function theorems for H-differentiable and semismooth functions. Optimization Methods and Software **19**(5), 443–461 (2004). DOI 10.1080/10556780410001697668
26. Güler, O.: New proximal point algorithms for convex minimization. SIAM Journal on Optimization **2**(4), 649–664 (1992). DOI 10.1137/0802032
27. Han, J., Sun, D.: Newton and quasi-Newton methods for normal maps with polyhedral sets. Journal of Optimization Theory and Applications **94**(3), 659–676 (1997). DOI 10.1023/A:1022653001160
28. Hiriart-Urruty, J.B., Lemaréchal, C.: Fundamentals of Convex Analysis. Grundlehren Text Editions. Springer Berlin Heidelberg (2004)
29. Horn, R.A., Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press (1994)
30. Kanzow, C., Ferenczi, I., Fukushima, M.: On the local convergence of semismooth Newton methods for linear and nonlinear second-order cone programs without strict complementarity. SIAM Journal on Optimization **20**(1), 297–320 (2009). DOI 10.1137/060657662
31. Lan, G., Lu, Z., Monteiro, R.D.C.: Primal-dual first-order methods with $O(1/\varepsilon)$ iteration-complexity for cone programming. Mathematical Programming **126**(1), 1–29 (2011). DOI 10.1007/s10107-008-0261-6
32. Lee, J.D., Sun, Y., Saunders, M.: Proximal Newton-type methods for minimizing composite functions. SIAM Journal on Optimization **24**(3), 1420–1443 (2014). DOI 10.1137/130921428
33. Lemaréchal, C., Sagastizábal, C.: Practical aspects of the Moreau-Yosida regularization: Theoretical preliminaries. SIAM Journal on Optimization **7**(2), 367–385 (1997). DOI 10.1137/S1052623494267127
34. Lewis, A.S.: The convex analysis of unitarily invariant matrix functions. Journal of Convex Analysis **2**(1), 173–183 (1995)
35. Lewis, A.S.: Convex analysis on the Hermitian matrices. SIAM Journal on Optimization **6**(1), 164–177 (1996). DOI 10.1137/0806009

36. Lewis, A.S.: Derivatives of spectral functions. Mathematics of Operations Research **21**(3), 576–588 (1996)
37. Lewis, A.S., Sendov, H.S.: Twice differentiable spectral functions. SIAM Journal on Matrix Analysis and Applications **23**(2), 368–386 (2001). DOI 10.1137/S089547980036838X
38. Li, W., Peng, J.: Exact penalty functions for constrained minimization problems via regularized gap function for variational inequalities. Journal of Global Optimization **37**(1), 85–94 (2007). DOI 10.1007/s10898-006-9038-8
39. Li, X., Sun, D., Toh, K.C.: On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope. ArXiv e-prints (2017)
40. Lions Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM Journal on Numerical Analysis **16**(6), 964–979 (1979). DOI 10.1137/0716071
41. Liu, Z., Vandenberghe, L.: Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications **31**(3), 1235–1256 (2010). DOI 10.1137/090755436
42. Lu, Z.: Randomized block proximal damped Newton method for composite self-concordant minimization. SIAM Journal on Optimization **27**(3), 1910–1942 (2017). DOI 10.1137/16M1082767
43. Luo, Z.Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Annals of Operations Research **46**(1), 157–178 (1993). DOI 10.1007/BF02096261
44. Maratos, N.: Exact penalty function algorithms for finite dimensional and control optimization problems (1978)
45. Martinet, B.: Brève communication. Régularisation d'inéquations variationnelles par approximations successives. Revue française d'informatique et de recherche opérationnelle. Série rouge **4**(R3), 154–158 (1970)
46. Meng, F.: Moreau-Yosida regularization of Lagrangian-dual functions for a class of convex optimization problems. Journal of Global Optimization **44**(3), 375 (2008). DOI 10.1007/s10898-008-9333-7
47. Meng, F., Sun, D., Zhao, G.: Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization. Mathematical Programming **104**(2), 561–581 (2005). DOI 10.1007/s10107-005-0629-9
48. Meng, F., Zhao, G., Goh, M., De Souza, R.: Lagrangian-dual functions and Moreau-Yosida regularization. SIAM Journal on Optimization **19**(1), 39–61 (2008). DOI 10.1137/060673746
49. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM Journal on Control and Optimization **15**(6), 959–972 (1977). DOI 10.1137/0315061
50. Mifflin, R., Qi, L., Sun, D.: Properties of the Moreau-Yosida regularization of a piecewise C^2 convex function. Mathematical Programming **84**(2), 269–281 (1999). DOI 10.1007/s10107980029a
51. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France **93**, 273–299 (1965)
52. Morita, T., Kanade, T.: A sequential factorization method for recovering shape and motion from image streams. IEEE Transactions on Pattern Analysis and Machine Intelligence **19**(8), 858–867 (1997). DOI 10.1109/34.608289
53. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer (2003)
54. Nesterov, Y.: Gradient methods for minimizing composite functions. Mathematical Programming **140**(1), 125–161 (2013). DOI 10.1007/s10107-012-0629-5
55. Pang, J.S.: Error bounds in mathematical programming. Mathematical Programming **79**(1), 299–332 (1997). DOI 10.1007/BF02614322
56. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **1**(3), 127–239 (2014). DOI 10.1561/2400000003
57. Patrinos, P., Bemporad, A.: Proximal Newton methods for convex composite optimization. In: IEEE Conference on Decision and Control, pp. 2358–2363 (2013)

58. Patrinos, P., Sopasakis, P., Sarimveis, H.: A global piecewise smooth Newton method for fast large-scale model predictive control. Automatica **47**(9), 2016–2022 (2011)
59. Patrinos, P., Stella, L., Bemporad, A.: Forward-backward truncated Newton methods for convex composite optimization. ArXiv e-prints (2014)
60. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Mathematical Programming **58**(1), 353–367 (1993). DOI 10.1007/BF01581275
61. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Review **52**(3), 471–501 (2010). DOI 10.1137/070697835
62. Rennie, J.D.M., Srebro, N.: Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the 22Nd International Conference on Machine Learning, ICML '05, pp. 713–719. ACM, New York, NY, USA (2005). DOI 10.1145/1102351.1102441
63. Rockafellar, R.: Convex analysis (1970)
64. Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM Journal on Control and Optimization **14**(5), 877–898 (1976). DOI 10.1137/0314056
65. Rockafellar, R.T., Wets, R.J.B.: Variational analysis, vol. 317. Springer Science & Business Media (2011)
66. Scholtes, S.: Piecewise Differentiable Functions, pp. 91–111. Springer New York, New York, NY (2012). DOI 10.1007/978-1-4614-4340-7_4
67. Sopasakis, P., Freris, N., Patrinos, P.: Accelerated reconstruction of a compressively sampled data stream. In: 2016 24th European Signal Processing Conference (EUSIPCO), pp. 1078–1082 (2016). DOI 10.1109/EUSIPCO.2016.7760414
68. Srebro, N.: Learning with matrix factorizations. Ph.D. thesis, Cambridge, MA, USA (2004)
69. Stella, L., Themelis, A., Patrinos, P.: Forward-backward quasi-Newton methods for nonsmooth optimization problems. Computational Optimization and Applications **67**(3), 443–487 (2017). DOI 10.1007/s10589-017-9912-y
70. Stella, L., Themelis, A., Patrinos, P.: Newton-type alternating minimization algorithm for convex optimization. IEEE Transactions on Automatic Control (2018). DOI 10.1109/TAC.2018.2872203
71. Stella, L., Themelis, A., Sopasakis, P., Patrinos, P.: A simple and efficient algorithm for nonlinear model predictive control. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pp. 1939–1944 (2017). DOI 10.1109/CDC.2017.8263933
72. Sun, D., Fukushima, M., Qi, L.: A computable generalized Hessian of the D-gap function and Newton-type methods for variational inequality problems. Complementarity and Variational Problems: State of the Art, MC Ferris and JS Pang (eds.), SIAM, Philadelphia, PA pp. 452–472 (1997)
73. Sun, D., Sun, J.: Semismooth matrix-valued functions. Mathematics of Operations Research **27**(1), 150–169 (2002). DOI 10.1287/moor.27.1.150.342
74. Themelis, A., Patrinos, P.: Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results. ArXiv e-prints (2017)
75. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. SIAM Journal on Optimization **28**(3), 2274–2303 (2018). DOI 10.1137/16M1080240
76. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision **9**(2), 137–154 (1992). DOI 10.1007/BF00129684
77. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. Tech. rep. (2008)
78. Ulbrich, M.: Optimization Methods in Banach Spaces, pp. 97–156. Springer Netherlands, Dordrecht (2009). DOI 10.1007/978-1-4020-8839-1_2
79. Yamashita, N., Taji, K., Fukushima, M.: Unconstrained optimization reformulations of variational inequality problems. Journal of Optimization Theory and Applications **92**(3), 439–456 (1997). DOI 10.1023/A:1022660704427

80. Yang, Z.: A study on nonsymmetric matrix-valued functions. Master's thesis, National University of Singapore (2009)
81. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **68**(1), 49–67 (2006)
82. Zhou, G., Qi, L.: On the convergence of an inexact Newton-type method. Oper. Res. Lett. **34**(6), 647–652 (2006). DOI 10.1016/j.orl.2005.11.001
83. Zhou, G., Toh, K.C.: Superlinear convergence of a Newton-type algorithm for monotone equations. Journal of Optimization Theory and Applications **125**(1), 205–221 (2005). DOI 10.1007/s10957-004-1721-7

Chapter 16
Hierarchical Convex Optimization by the Hybrid Steepest Descent Method with Proximal Splitting Operators—Enhancements of SVM and Lasso

Isao Yamada and Masao Yamagishi

Abstract The breakthrough ideas in the modern proximal splitting methodologies allow us to express the set of all minimizers of a superposition of multiple nonsmooth convex functions as the fixed point set of computable nonexpansive operators. In this paper, we present practical algorithmic strategies for the hierarchical convex optimization problems which require further strategic selection of a most desirable vector from the solution set of the standard convex optimization. The proposed algorithms are established by applying the hybrid steepest descent method to special nonexpansive operators designed through the art of proximal splitting. We also present applications of the proposed strategies to certain unexplored hierarchical enhancements of the support vector machine and the Lasso estimator.

Keywords Convex optimization · Proximal splitting algorithms · Hybrid steepest descent method · Support Vector Machine (SVM) · Lasso · TREX · Signal processing · Machine learning · Statistical estimation

AMS 2010 Subject Classification 49M20, 65K10, 90C30

16.1 Introduction

Convex optimization has been playing a central role in a broad range of mathematical sciences and engineering. Many optimization tasks in such applications can be interpreted as special instances of the following simple model:

I. Yamada (✉) · M. Yamagishi
Department of Information and Communications Engineering, Tokyo Institute of Technology, Tokyo, Japan
e-mail: isao@ict.e.titech.ac.jp; myamagi@ict.e.titech.ac.jp

H. H. Bauschke et al. (eds.), *Splitting Algorithms, Modern Operator Theory, and Applications*, https://doi.org/10.1007/978-3-030-25939-6_16

$$\text{minimize } f(x) + g(Ax) \text{ subject to } x \in \mathcal{X}, \tag{16.1}$$

where $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}})$, $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}}, \|\cdot\|_{\mathcal{K}})$ are real Hilbert spaces, $f : \mathcal{X} \to (-\infty, \infty]$ and $g : \mathcal{K} \to (-\infty, \infty]$ are proper lower semicontinuous convex functions, i.e., $f \in \Gamma_0(\mathcal{X})$ and $g \in \Gamma_0(\mathcal{K})$, and $A : \mathcal{X} \to \mathcal{K}$ is a bounded linear operator, i.e., $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$. Such a unified simplification is indebted entirely to the remarkable expressive ability of the abstract Hilbert space. For example, a seemingly much more general model:

$$\text{find } x^\star \in \mathcal{S} := \operatorname*{argmin}_{x\in\mathcal{X}} \left[\Phi(x) := f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \tag{16.2}$$

where $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{K}_i, \langle\cdot,\cdot\rangle_{\mathcal{K}_i}, \|\cdot\|_{\mathcal{K}_i})$ $(i = 1, 2, \ldots, m)$ are real Hilbert spaces, $f \in \Gamma_0(\mathcal{X})$, $g_i \in \Gamma_0(\mathcal{K}_i)$ $(i = 1, 2, \ldots, m)$, and $A_i \in \mathcal{B}(\mathcal{X}, \mathcal{K}_i)$ $(i = 1, 2, \ldots, m)$, can also be translated into the problem in (16.1) by redefining a new Hilbert space

$$\mathcal{K} := \mathcal{K}_1 \times \cdots \times \mathcal{K}_m = \{\mathbf{x} = (x_1, \ldots, x_m) \mid x_i \in \mathcal{K}_i \ (i = 1, \ldots, m)\} \tag{16.3}$$

equipped with the addition $(\mathbf{x}, \mathbf{y}) \mapsto (x_1 + y_1, \ldots, x_m + y_m)$, the scalar multiplication $(\alpha, \mathbf{x}) \mapsto (\alpha x_1, \ldots, \alpha x_m)$, and the inner product $(\mathbf{x}, \mathbf{y}) \mapsto \langle \mathbf{x}, \mathbf{y}\rangle_{\mathcal{K}} := \sum_{i=1}^{m} \langle x_i, y_i\rangle_{\mathcal{K}_i}$, a new convex function

$$g := \bigoplus_{i=1}^{m} g_i : \mathcal{K} \to (-\infty, \infty] : (x_1, \ldots, x_m) \mapsto \sum_{i=1}^{m} g_i(x_i), \tag{16.4}$$

and a new bounded linear operator

$$A : \mathcal{X} \to \mathcal{K} : x \mapsto (A_1 x, \ldots, A_m x). \tag{16.5}$$

Indeed, for many years, the model (16.2) has been accepted widely as a standard, where all players, $f, g_i \circ A_i \in \Gamma_0(\mathcal{X})$ $(i = 1, \ldots, m)$ in (16.2) are designed strategically by users in order to achieve, after optimization, a valuable vector satisfying their requirements.

The so-called *proximal splitting methodology* has been built, on the rich mathematical foundations of convex analysis, monotone operator theory and fixed point theory of nonexpansive operators (see, e.g., [9, 45, 47, 139]), in order to broaden the applicability of the proximity operators of convex functions [101], e.g., to the model (16.2). It is well-known that the solution set $\mathcal{S}$ in (16.2) can be characterized completely as the zero of the set-valued operator $\partial\Phi : \mathcal{X} \to 2^{\mathcal{X}} : x \mapsto \{u \in \mathcal{X} \mid \Phi(y) \geq \Phi(x) + \langle y - x, u\rangle_{\mathcal{X}} \ (\forall y \in \mathcal{X})\}$, called the *subdifferential* of Φ. The maximal monotonicity of $\partial\Phi$ provides us with further equivalent fixed point characterization in terms of a single valued operator $\mathrm{prox}_{\Phi} := (\mathrm{I} + \partial\Phi)^{-1} : \mathcal{X} \to \mathcal{X} : x \mapsto \operatorname{argmin}_{y\in\mathcal{X}} \Phi(y) + \frac{1}{2}\|x - y\|_{\mathcal{X}}^2$ called the proximity operator of Φ (see Section 16.2.2) [Note: The identity operator is denoted by $\mathrm{I}: \mathcal{X} \to \mathcal{X}$ but the

common notation I is going to be used for the identity operator on any real Hilbert space in this paper]. This fact is simply stated as

$$(\forall z \in \mathcal{X}) \quad z \in \mathcal{S} \Leftrightarrow 0 \in \partial \Phi(z) \Leftrightarrow z = \mathrm{prox}_{\Phi}(z) \tag{16.6}$$

and some algorithms can generate, from any $x_0 \in \mathcal{X}$, a weakly convergent sequence $x_n \in \mathcal{X}$ $(n \in \mathbb{N})$ to a point in $\mathcal{S}$ in (16.2) if prox_{Φ} is available as a computational tool (see, e.g., [98, 99, 121]). A simplest example of such algorithms generates a sequence $(x_n)_{n\in\mathbb{N}}$ by

$$x_{n+1} = \mathrm{prox}_{\Phi}(x_n) \quad (n = 0, 1, 2, \ldots). \tag{16.7}$$

The algorithm (16.7) can be interpreted as a straightforward application of *Krasnosel'skiĭ-Mann Iterative Process* (see Fact 16.6 in Section 16.2.2) because prox_{Φ} is known to be firmly nonexpansive, i.e., $2\mathrm{prox}_{\Phi} - \mathrm{I} : \mathcal{X} \to \mathcal{X}$ is a nonexpansive operator (see (16.27)). Although the above strategy in (16.7) is conceptually simple and elegant, its applicability has been very limited because the computation of $\mathrm{prox}_{\Phi}(x)$ requires to solve a regularized convex optimization problem $\min \Phi(\cdot) + \frac{1}{2}\|x - \cdot\|_{\mathcal{X}}^2$ whose unique solution is still hard to be computed for most scenarios of type (16.2) in many application areas.

On the other hand, there are many scenarios that fall in the model (16.2) where the proximity operators of the all players, i.e., $\mathrm{prox}_{f} : \mathcal{X} \to \mathcal{X}$ and $\mathrm{prox}_{g_i} : \mathcal{K}_i \to \mathcal{K}_i$ $(i = 1, \ldots, m)$, are available as computational tools while prox_{Φ} is not practically available (see, e.g., [42, 45]). A major goal of recent active studies (see, e.g., [40, 44, 47, 139, 150]) on the *proximal splitting methodology* has been the creation of more applicable iterative algorithms, for (16.2) and its variations, than (16.7) by utilizing computable tools prox_{f} and prox_{g_i} $(i = 1, \ldots, m)$ simultaneously. Such effort has culminated in many powerful algorithms which have been applied successfully to the broader classes of optimizations including the standard model (16.2).

Usually, the standard model (16.2) is formulated in the form of a weighted average of multiple convex functions and the weights are designed in accordance with the level of importance of each convex function. However quantification of the level of importance is often challenging as well as influential to the final results of optimizations (see Section 16.5.1 for a recent advanced strategy of such a parameter tuning for the Lasso estimator which is a standard sparsity aware statistical estimation method). By keeping in mind (i) the remarkable flexibility of the standard model (16.2) proven extensively in many successful applications of the modern proximal splitting methodology, as well as (ii) the inherent difficulty in the weight design of multiple convex functions in (16.2), a question arises: *Is there any alternative model of (16.2) which can also serve as a natural optimization strategy for multiple convex criteria* ? To see the light of the tunnel regarding this primitive question, let us start to imagine important elements for us to consider in *finding residence*. We may consider the house rent, the residential environment including living space and housing equipment, the neighborhood environment, the

accessibility to public transportation systems, and the commuting time, etc. We would prioritize the elements, e.g., firstly by narrowing down the candidates to the set S_1 of all residents of which the rents and commuting times are in your acceptable range. Next we may try to narrow down the candidates to the set $S_2(\subset S_1)$ of all residents whose living spaces achieve maximum level among all in S_1. Further, we may probably like to select residents in S_2 as final choices by choosing the best ones, e.g., in the sense of the neighborhood environment or the housing equipment. This simple example suggests that we often optimize multiple criteria one by one hierarchically rather than optimize the sum of different criteria at once certainly because there exists no universal justification for adding different criteria. In fact, many mathematicians and scientists have been challenging to pave the way for the so-called hierarchical convex optimization problems (see, e.g., [3, 8, 18, 32, 33, 36, 46, 55, 95, 107, 114, 133, 142, 146–149]). Landmark theories toward $\mathcal{M}$-stage hierarchical convex optimization are found, e.g., in [3, 18] where, for given $\Phi_i \in \Gamma_0(\mathcal{X})$ $(i = 0, 1, \ldots, \mathcal{M})$ satisfying $S_i := \operatorname*{argmin}_{x \in S_{i-1}} \Phi_i(x) \neq \varnothing$ $(i = 0, 1, \ldots, \mathcal{M})$ with $S_{-1} := \mathcal{X}$, their major goals are set to establish computational strategies for iterative approximation of a point in $S_{\mathcal{M}}$. (Note: Every point in S_i of the hierarchical convex optimization is called a *viscosity solution* of S_{i-1} $(i = 1, 2, \ldots, \mathcal{M})$. To avoid confusion with "bilevel optimization" in the sense of [19, 30, 138], we do not use the designation *bilevel optimization* for S_1 in our hierarchical convex optimization). Under the assumptions that $\dim(\mathcal{X}) < \infty$ and that $\Phi_i \in \Gamma_0(\mathcal{X})$ $(i = 1, 2, \ldots, \mathcal{M})$ are real valued, Cabot [18] showed that the sequence $(x_n)_{n\in\mathbb{N}}$ defined by

$$\begin{aligned} x_{n+1} &:= \operatorname{prox}_{\left(\Phi_0+\varepsilon_n^{(1)}\Phi_1+\varepsilon_n^{(2)}\Phi_2+\cdots+\varepsilon_n^{(\mathcal{M})}\Phi_{\mathcal{M}}\right)}(x_n) \\ &= \left(\mathrm{I} + \partial\left(\Phi_0 + \varepsilon_n^{(1)}\Phi_1 + \varepsilon_n^{(2)}\Phi_2 + \cdots + \varepsilon_n^{(\mathcal{M})}\Phi_{\mathcal{M}}\right)\right)^{-1}(x_n) \quad (16.8) \end{aligned}$$

satisfies (i) $\lim_{n\to\infty} d(x_n, S_{\mathcal{M}}) = 0$ and (ii) $(\forall i \in \{0, 1, \ldots, \mathcal{M}\})$ $\lim_{n\to\infty} \Phi_i(x_n) = \min_{x\in S_{i-1}} \Phi_i(x)$ if positive number sequences $(\varepsilon_n^{(0)} := 1)_{n\in\mathbb{N}}$ and $(\varepsilon_n^{(i)})_{n\in\mathbb{N}}$ $(i \in \{1, \ldots, \mathcal{M}\})$ satisfy certain technical conditions including $\lim_{n\to\infty} \varepsilon_n^{(i)} = 0$, $\lim_{n\to\infty} \dfrac{\varepsilon_n^{(i)}}{\varepsilon_n^{(i-1)}} = 0$ $(i = 1, 2, \ldots, \mathcal{M})$, and $\sum_{n=0}^{\infty} \varepsilon_n^{(\mathcal{M})} = \infty$ [Note: The scheme (16.8) is a simplified version of the original scheme in [18] by restricting to the case $\lambda_n = 1$ and $\eta_n = 0$ $(n \in \mathbb{N})$].

Clearly, the algorithms (16.8) and (16.7) have essentially a common limitation in their practical applicabilities because (16.8) requires $\operatorname{prox}\left(\Phi_0 + \varepsilon_n^{(1)}\Phi_1 + \varepsilon_n^{(2)}\Phi_2 + \cdots + \varepsilon_n^{(\mathcal{M})}\Phi_{\mathcal{M}}\right)$, or its very good approximation, for every update in generation of $(x_n)_{n\in\mathbb{N}}$. By recalling the breakthrough ideas developed in the recent *proximal splitting methodology* for resolution of the inherent limitation in (16.7), an ideal as well as possibly realistic assumption to be imposed

upon each player $\Phi_i : \mathcal{X} \to (-\infty, \infty]$ $(i = 0, 1, \ldots, \mathcal{M})$ in the above $\mathcal{M}$-stage hierarchical convex optimization seems to be certain differentiability assumptions or proximal decomposability assumptions, e.g.

$$\Phi_i(x) := f_i(x) + \sum_{\iota_{(i)}=1}^{M_i} g_{\iota_{(i)}}(A_{\iota_{(i)}} x), \tag{16.9}$$

with real Hilbert spaces $(\mathcal{K}_{\iota_{(i)}}, \langle \cdot, \cdot \rangle_{\mathcal{K}_{\iota_{(i)}}}, \| \cdot \|_{\mathcal{K}_{\iota_{(i)}}})$ $(\iota_{(i)} = 1, 2, \ldots, M_i)$, $f_i \in \Gamma_0(\mathcal{X})$, $g_{\iota_{(i)}} \in \Gamma_0(\mathcal{K}_i)$ $(i = 0, 1, \ldots, \mathcal{M})$, and bounded linear operators $A_{\iota_{(i)}} : \mathcal{X} \to \mathcal{K}_{\iota_{(i)}}$ $(\iota_{(i)} = 1, 2, \ldots, M_i)$, where $\mathrm{prox}_{f_i} : \mathcal{X} \to \mathcal{X}$ and $\mathrm{prox}_{g_{\iota_{(i)}}} : \mathcal{K}_{\iota_{(i)}} \to \mathcal{K}_{\iota_{(i)}}$ $(\iota_{(i)} = 1, \ldots, M_i)$, are available as computational tools while prox_{Φ_i} is not necessarily available.

In this paper, we choose to cast our primary target in the iterative approximation of a solution of

$$\text{minimize } \Psi(x^\star) \text{ subject to } x^\star \in \operatorname*{argmin}_{x \in \mathcal{X}} \left[\Phi(x) := f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \tag{16.10}$$

i.e., a viscosity solution of the convex optimization problem (16.2), where we assume that $\Psi \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi : \mathcal{X} \to \mathcal{X}$, i.e.,

$$(\exists \kappa > 0,\ \forall x, y \in \mathcal{X}) \quad \|\nabla\Psi(x) - \nabla\Psi(y)\| \leq \kappa \|x - y\|, \tag{16.11}$$

and that $\mathrm{prox}_f : \mathcal{X} \to \mathcal{X}$ and $\mathrm{prox}_{g_i} : \mathcal{K}_i \to \mathcal{K}_i$ $(i = 1, \ldots, m)$ are available as computational tools.

Although the application of such iterative algorithms is certainly restrictive compared to the overwhelming potential of the general hierarchical convex optimization, our target is realistic and still allows us to cover many applications of interest to practitioners who are searching for a step ahead optimization strategy and yet to be able to exploit maximally the central ideas in the modern proximal splitting methodologies. Especially for practitioners, we remark that if the suppression of $\sum_{k=1}^{L} \psi_k \circ B_k \in \Gamma_0(\mathcal{X})$ over $\mathrm{argmin}_{x \in \mathcal{X}} \Phi(x)$ is required, where, for each $k \in \{1, 2, \ldots, L\}$, $\mathcal{Y}_k$ is a real Hilbert space, $\psi_k \in \Gamma_0(\mathcal{Y}_k)$, $\mathrm{prox}_{\gamma\psi_k} : \mathcal{Y}_k \to \mathcal{Y}_k$ $(\gamma > 0)$ is available as computational tools, and $B_k \in \mathcal{B}(\mathcal{X}, \mathcal{Y}_k)$, such a mission could be achieved satisfactorily by considering an alternative problem below of type (16.10):

$$\begin{aligned} &\text{minimize } \Psi(x^\star) := \sum_{k=1}^{L}{}^{\gamma} \psi_k(B_k x^\star) \\ &\text{subject to } x^\star \in \operatorname*{argmin}_{x \in \mathcal{X}} \left[\Phi(x) := f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \end{aligned} \tag{16.12}$$

where (i) ${}^{\gamma}\psi_k : \mathcal{Y}_k \to \mathbb{R} : y_k \mapsto \min_{y\in\mathcal{Y}_k} \psi_k(y) + \frac{1}{2\gamma}\|y - y_k\|^2_{\mathcal{Y}_k}$ $(k = 1, 2, \ldots, L)$ are *the Moreau envelopes (or the Moreau-Yosida regularizations)* of a sufficiently small index $\gamma > 0$ (see Fact 16.8 in Section 16.2.2 and [149]). This is because $\lim_{\gamma\downarrow 0} {}^{\gamma}\psi_k(y_k) = \psi_k(y_k)$ $(\forall y_k \in \mathrm{dom}\psi_k := \{y \in \mathcal{Y}_k \mid \psi_k(y) < \infty\})$ and ${}^{\gamma}\psi_k$ is Gâteaux differentiable with $\frac{1}{\gamma}$-Lipschitzian $\nabla {}^{\gamma}\psi_k : \mathcal{Y}_k \to \mathcal{Y}_k : y_k \mapsto \frac{y_k - \mathrm{prox}_{\gamma\psi_k}(y_k)}{\gamma}$ and therefore

$$(\forall x_1, x_2 \in \mathcal{X}) \quad \left\| \nabla \sum_{k=1}^{L} \left({}^{\gamma}\psi_k \circ B_k\right)(x_1) - \nabla \sum_{k=1}^{L} \left({}^{\gamma}\psi_k \circ B_k\right)(x_2) \right\|_{\mathcal{X}}$$

$$= \left\| \sum_{k=1}^{L} B_k^* \nabla {}^{\gamma}\psi_k\left(B_k x_1\right) - \sum_{k=1}^{L} B_k^* \nabla {}^{\gamma}\psi_k\left(B_k x_2\right) \right\|_{\mathcal{X}} \leq \sum_{k=1}^{L} \frac{\|B_k\|^2_{\mathrm{op}}}{\gamma} \|x_1 - x_2\|_{\mathcal{X}},$$

where $B_k^* \in \mathcal{B}(\mathcal{Y}_k, \mathcal{X})$ is the conjugate of $B_k \in \mathcal{B}_k(\mathcal{X}, \mathcal{Y}_k)$ and $\|\cdot\|_{\mathrm{op}}$ stands for the operator norm.

Fortunately, by introducing the exactly same translation used in the reformulation of Problem (16.2) as an instance of Problem (16.1), our problem (16.10) can also be simplified as

$$\text{minimize } \Psi(x^\star) \text{ subject to } x^\star \in \mathcal{S}_p := \operatorname*{argmin}_{x\in\mathcal{X}} [f(x) + g(Ax)] \neq \varnothing, \tag{16.13}$$

where $\mathcal{K}$, $g : \mathcal{K} \to (-\infty, \infty]$ and $A : \mathcal{X} \to \mathcal{K}$ are defined, respectively,[1] by (16.3), (16.4), and (16.5), and we can assume that (i) $\Psi \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi : \mathcal{X} \to \mathcal{X}$, and that (ii) $\mathrm{prox}_f : \mathcal{X} \to \mathcal{X}$ and $\mathrm{prox}_g : \mathcal{K} \to \mathcal{K}$ are available as computational tools because

$$\begin{aligned} \mathrm{prox}_g(\mathbf{x}) &:= \operatorname*{argmin}_{\mathbf{y}\in\mathcal{K}} \left[g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2_{\mathcal{K}} \right] \\ &= \operatorname*{argmin}_{(y_1,\ldots,y_m)\in\mathcal{K}_1\times\cdots\times\mathcal{K}_m} \sum_{i=1}^{m} \left[g_i(y_i) + \frac{1}{2}\|y_i - x_i\|^2_{\mathcal{K}_i} \right] \\ &= \left(\mathrm{prox}_{g_1}(x_1), \ldots, \mathrm{prox}_{g_m}(x_m)\right). \end{aligned} \tag{16.14}$$

The following two scenarios suggest the remarkable advantage achieved by algorithmic solutions to (16.10).

Scenario 1 (Unification of Conditional Optimization Models) Let $f_{\langle\mathcal{D}\rangle} \in \Gamma_0(\mathcal{X})$ and $g_{i\langle\mathcal{D}\rangle} \in \Gamma_0(\mathcal{K}_i)$ $(i = 1, 2, \ldots, m)$ be nonnegative valued functions which are defined with observed data $\mathcal{D}$. Suppose that there exists a well-established data analytic strategy which utilizes with $\Psi \in \Gamma_0(\mathcal{X})$ as

[1] There are many practical conditions for (f, g, A) to guarantee $\mathcal{S}_p \neq \varnothing$, see, e.g., [9, 153] and Fact 16.2 in Section 16.2.1.

$$\text{find } x^\star \in \operatorname*{argmin}_{x\in\mathcal{S}_0} \Psi(x),$$
$$\text{where } \mathcal{S}_0 := \left\{x \in \mathcal{X} \mid f_{\langle\mathcal{D}\rangle}(x)=g_{i\langle\mathcal{D}\rangle}(Ax)=0\ (i=1,2,\ldots,m)\right\}, \quad (16.15)$$

provided that the data $\mathcal{D}$ is consistent, i.e., it satisfies $\mathcal{S}_0 \neq \varnothing$.

However, to deal with more general data $\mathcal{D}$, it is important to establish a mathematically sound extension of the above data analytic strategy to be applicable even to inconsistent data $\mathcal{D}$ s.t. $\mathcal{S}_0 = \varnothing$. One of the most natural extensions of (16.15) would be the following hierarchical formulation:

$$\text{find } x^{\star\star} \in \operatorname*{argmin}_{x^\star\in\mathcal{S}_{\langle\mathcal{D}\rangle}} \Psi(x^\star),$$
$$\text{where } \mathcal{S}_{\langle\mathcal{D}\rangle} := \operatorname*{argmin}_{x\in\mathcal{X}} \left[f_{\langle\mathcal{D}\rangle}(x) + \sum_{i=1}^{m} g_{i\langle\mathcal{D}\rangle}(A_i x) \right], \quad (16.16)$$

because $\mathcal{S}_{\langle\mathcal{D}\rangle} \neq \varnothing$ holds under weaker assumption than $\mathcal{S}_0 \neq \varnothing$, and $\mathcal{S}_{\langle\mathcal{D}\rangle} = \mathcal{S}_0$ holds true if $\mathcal{S}_0 \neq \varnothing$. However, the well-established data analytic strategies only for consistent data $\mathcal{D}$ in the form of (16.15) have often been modified, with the so-called *tuning parameter* $\mathfrak{C} > 0$, to

$$\text{find } \tilde{x}^\star \in \operatorname*{argmin}_{x\in\mathcal{X}} \left[\frac{1}{\mathfrak{C}}\Psi(x) + f_{\langle\mathcal{D}\rangle}(x) + \sum_{i=1}^{m} g_{i\langle\mathcal{D}\rangle}(A_i x) \right], \quad (16.17)$$

which is not really an extension of (16.15) because the model (16.17) unfortunately has no guarantee to produce $x^\star$ in (16.15) even if $\mathcal{D}$ satisfies $\mathcal{S}_0 \neq \varnothing$.

Scenario 2 Suppose that we are in a desired vector in $\mathcal{X}$ at which the functions $f, g_i \circ A_i \in \Gamma_0(\mathcal{X})$ $(i = 1, \ldots, m)$ in (16.2) are known to achieve small values and therefore the model (16.2) has been employed as an estimation strategy. Suppose also that we newly found another effective criterion $\Psi \in \Gamma_0(\mathcal{X})$ which likely to achieve small values around the desired vector to be estimated. In such a case, our common utilization of Ψ, for improvement of the previous strategy, has often been modeled as a new optimization problem:

$$\text{find } \tilde{x}^\star \in \widetilde{\mathcal{S}} := \operatorname*{argmin}_{x\in\mathcal{X}} \left[f(x) + \sum_{i=1}^{m} g_i(A_i x) + \Psi(x) \right] \neq \varnothing. \quad (16.18)$$

However, it is essentially hard to tell which is better between the estimation strategies (16.2) and (16.18) because the criteria in these optimizations are different. Indeed, $\tilde{x}^\star$ does not necessarily achieve best in the sense of the model (16.2) while $x^\star$ certainly achieves best in the sense of the model (16.2). On the other hand, if we formulate a new optimization problem, from a hierarchical optimization point of view, e.g., as

$$\text{find } x^{\star\star} \in \operatorname*{argmin}_{x^\star \in \mathcal{S}} \Psi(x^\star), \text{ where } \mathcal{S} := \operatorname*{argmin}_{x \in \mathcal{X}} \left[f(x) + \sum_{i=1}^{m} g_i(A_i x) \right] \neq \varnothing, \tag{16.19}$$

its solution $x^{\star\star}$ certainly meets more faithfully all the requirements than $x^\star \in \mathcal{S}$ because both $x^{\star\star}, x^\star \in \mathcal{S}$ and $\Psi(x^{\star\star}) \leq \Psi(x^\star)$ are achieved.

The following examples suggest that the hierarchical optimization has been offering well-grounded direction for advancement of computational strategies in inverse problems and data sciences.

Example 16.1 (Hierarchical Convex Optimizations in Real-World Applications[2]*)*

(a) (Generalized inverse/Moore-Penrose inverse [9, 12, 100, 111, 118]) Let $\mathcal{X}$ and $\mathcal{K}$ be real Hilbert spaces, let $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ be such that $\mathrm{ran}(A) := \{A(x) \in \mathcal{K} \mid x \in \mathcal{X}\}$ is closed. Then for every $y \in \mathcal{K}$, $C_y := \{x \in \mathcal{X} \mid \|Ax - y\|_{\mathcal{K}} = \min_{z \in \mathcal{X}} \|Az - y\|_{\mathcal{K}}\} = \{x \in \mathcal{X} \mid A^*A(x) = A^*(y)\} \neq \varnothing$. The generalized inverse (in the sense of Moore-Penrose) $A^\dagger \in \mathcal{B}(\mathcal{K}, \mathcal{X})$ is defined as $A^\dagger : \mathcal{K} \to \mathcal{X} : y \mapsto P_{C_y}(0)$,where P_{C_y} is the orthogonal projection onto C_y. $A^\dagger(y)$ can be seen as the unique solution to the hierarchical convex optimization problem (16.19) for $f(z) := \|A(z) - y\|_{\mathcal{K}}$, $g_i(z) := 0$ $(i = 1, 2, \ldots, m)$ and $\Psi(z) = \frac{1}{2}\|z\|^2_{\mathcal{X}}$. The *Moore-Penrose inverse* $A^\dagger \in \mathcal{B}(\mathcal{K}, \mathcal{X})$ of $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ has been serving as one of the most natural generalizations of the inverse of A, typically in Scenario 1, under the situations where the existence of $A^{-1} \in \mathcal{B}(\mathcal{K}, \mathcal{X})$ is not guaranteed. In particular, for finite dimensional settings, there are many ways to express $A^\dagger$. These include the singular value decomposition of $A^\dagger$ in terms of the singular value decomposition of A.

(b) (Tikhonov approximation [3, 9, 55, 133]) Let Ψ, $f \in \Gamma_0(\mathcal{X})$ and $\mathrm{argmin}(f) \cap \mathrm{dom}(\Psi) \neq \varnothing$ where Ψ is coercive and strictly convex. Then Ψ admits a unique minimizer x_0 over $\mathrm{argmin}(f)$. This x_0 can be seen as the solution of the hierarchical convex optimization in (16.19) for $g_i(z) := 0$ $(i = 1, 2, \ldots, m)$. Moreover, if we define $x_\varepsilon \in \mathcal{X}$ as the unique minimizer of the regularized problem

$$\text{miminize } f(x) + \varepsilon\Psi(x) \text{ subject to } x \in \mathcal{X} \tag{16.20}$$

[2]To the best of the authors' knowledge, little has been reported on the hierarchical *nonconvex* optimization. We remark that the MV-PURE (minimum-variance pseudo-unbiased reduced-rank estimator) (see, e.g., [112, 113, 144]), for the unknown vector possibly subjected to linear constraints, is defined by a closed form solution of a certain hierarchical nonconvex optimization problem which characterizes a natural reduced rank extension of the Gauss-Markov (BLUE) estimator [85, 93] to the case of reduced-rank estimator. It was shown in [113] that specializations of the MV-PURE include Marquardt's reduced rank estimator [97], Chipman-Rao estimator [29], and Chipman's reduced rank estimator [28]. In Section 16.5.2 of this paper, we newly present a special instance of a hierarchical *nonconvex* optimization problem which can be solved through multiple hierarchical *convex* optimization subproblems.

for every $\varepsilon > 0$, the desired x_0 can be approximated as (i) $x_\varepsilon \rightharpoonup x_0$ (as $\varepsilon \downarrow 0$) and (ii) $\Psi(x_\varepsilon) \to \Psi(x_0)$ (as $\varepsilon \downarrow 0$). This fact suggests a strategy for approximating x_0 if we have a practical way of computing x_{ε_n} for positive sequence $(\varepsilon_n)_{n=1}^{\infty}$ satisfying $\varepsilon_n \downarrow 0$ (as $n \to \infty$). Many computational approaches to the hierarchical convex optimization seem to have been designed along this strategy.[3] We remark that many formulations of type (16.17) in Scenario 1 can be seen as instances of (16.20) with $\epsilon = \frac{1}{\mathfrak{C}}$. However, in general, the hierarchical optimality can never be guaranteed by the solution of (16.20) for a fixed constant $\epsilon > 0$.

(c) Assuming differentiability, the iteration of (16.8) for $\mathcal{M} = 1$ can also be interpreted as an implicit discretization of the continuous dynamical system:

$$\dot{x}(t) + \nabla\Phi_0(x(t)) + \varepsilon(t)\nabla\Phi_1(x(t)) = 0, \quad t \geq 0, \tag{16.21}$$

where $\varepsilon : \mathbb{R}_+ \to \mathbb{R}$ is a control parameter tending to 0 when $t \to \infty$. This observation has been motivating explicit discretization of (16.21) for iterative approximation of point in S_1, e.g. by

$$x_{n+1} \in x_n + \lambda_n \partial(\Phi_0 + \varepsilon_n \Phi_0)(x_n), \tag{16.22}$$

and its variations (see, e.g., [78, 79, 127, 128]), where λ_n is a nonnegative stepsize. However this class of algorithms cannot exploit recent advanced proximal splitting techniques for dealing with the constrained set S_0.

(d) Under the assumption that (i) Ψ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi : \mathcal{X} \to \mathcal{X}$, and (ii) $\mathrm{prox}_f : \mathcal{X} \to \mathcal{X}$ is available as a computable tool, the *inertial forward-backward algorithm with vanishing Tikhonov regularization* was proposed [4], along in the frame of accelerated forward-backward methods,[4] for an iterative approximation of the solution of a hierarchical convex optimization in (16.19) for $g_i = 0$ $(i = 1, 2, \ldots, m)$.

(e) In general, the convex optimization problems, especially in the convex feasibility problems [7, 22, 31], have infinitely many solutions that could be considerably different in terms of other criteria. However most iterative algorithms for convex optimization can approximate an anonymous solution of the problem. For pursuing a better solution in some other aspects, *superiorization* [21, 80, 104, 110] introduces proactively designed perturbations into the original algorithms with preserving preferable convergence properties. Essentially, by adopting another criterion Ψ, these methods aim to lower the value of Ψ with incorporating a perturbation involving the descent direction of Ψ. Apparently,

[3] The behavior of $(x_\varepsilon)_{\varepsilon \in (0,1)} \subset \mathcal{X}$ can be analyzed in the context of *approximating curve* for monotone inclusion problem. For recent results combined with Yosida regularization, see [37].

[4] See [4] on the stream of research, to name but a few, [11, 24], originated from Nesterov's seminal paper [103].

as reported in [150], the hierarchical convex optimization can serve as one of the ideal formulations for the superiorization.

(f) Let $\Psi \in \Gamma_0(\mathcal{X})$ be Gâteaux differentiable and its gradient $\nabla\Psi : \mathcal{X} \to \mathcal{X}$ is Lipschitzian. Suppose that $f \in \Gamma_0(\mathcal{X})$ is also Gâteaux differentiable with Lipschitzian gradient $\nabla f : \mathcal{X} \to \mathcal{X}$ and admits $\operatorname{argmin}(f + \iota_K) \neq \varnothing$ for a nonempty closed convex set $K \subset \mathcal{X}$, where ι_K is the indicator function, i.e.,

$$\iota_K(x) := \begin{cases} 0 & \text{if } x \in K, \\ \infty & \text{otherwise.} \end{cases} \tag{16.23}$$

Then

$$\text{minimize } \Psi(x) \ \text{ subject to } x \in \operatorname{argmin}(f + \iota_K) \tag{16.24}$$

can be seen as an instance of the hierarchical convex optimization in (16.19) for $g_1 := \iota_K$ and $g_i = 0$ $(i = 2, 3, \ldots, m)$. By applying the hybrid steepest descent method [52, 141, 142, 146–148] to several expressions of the set $\operatorname{argmin}(f+\iota_K)$ as the fixed point set of certain computable nonexpansive operators $T : \mathcal{X} \to \mathcal{X}$ (see, e.g., [146, Proposition 2.5], [149, Example 17.6(b)]), practical algorithms have been established to produce a sequence $x_n \in \mathcal{X}$ $(n = 0, 1, 2, \ldots)$ which is guaranteed to converge to a solution to Problem (16.24). These cover a version of projected Landweber method [63, 115, 123] for $\Psi(x) := \frac{1}{2}\|x\|_{\mathcal{X}}^2$ and $f(x) := \|A(x) - b\|_{\mathcal{K}}$, where $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ and the metric projection $P_K : \mathcal{X} \to K$ is assumed available as a computational tool. As will be discussed below, the main idea of the present paper specialized for Problem (16.13) (or equivalently Problem (16.10)) is along this simple hierarchical optimization strategy [86, 107, 146, 149, 150] of applying the hybrid steepest descent method (HSDM: see Section 16.2.4) to the precise expressions of the solution sets of the convex optimization problems in terms of fixed point sets of computable nonexpansive operators defined on a certain real Hilbert space $\mathcal{H}$ which is not necessarily the same as the original Hilbert space $\mathcal{X}$.

Apparently, to tackle Problem (16.13) (or equivalently Problem (16.10)), we need to exploit full information on $\mathcal{S}_p$ which is an infinite set in general. Moreover, even by using the recently developed powerful proximal splitting algorithms, specially designed for (16.1), we can produce only some vector sequence that converges to just an anonymous point in $\mathcal{S}_p$, which implies that we need to add further a new twist to the well-known strategies applicable to Problem (16.1).

Fortunately, the unified perspective from the viewpoint of convex analysis and monotone operator theory (see, e.g., [9]) often enables us to enjoy notable characterizations of the solution set $\mathcal{S}_p$ in terms of the set of all fixed points of a computable nonexpansive operator defined on certain real Hilbert spaces. Indeed, almost all existing proximal splitting algorithms for Problem (16.1) more or less rely on the following type of characterizations of $\mathcal{S}_p$:

$$\mathcal{S}_p = \operatorname*{argmin}_{x \in \mathcal{X}} f(x) + g(Ax) = \Xi\left(\mathrm{Fix}(T)\right) := \bigcup_{z \in \mathrm{Fix}(T)} \Xi(z) \subset \mathcal{X}, \tag{16.25}$$

$$\mathrm{Fix}(T) := \{z \in \mathcal{H} \mid T(z) = z\} \qquad \text{(Fixed point set of } T\text{)}, \tag{16.26}$$

where $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}}, \| \cdot \|_{\mathcal{H}})$ is a certain real Hilbert space (not necessarily $\mathcal{H} = \mathcal{X}$), $T : \mathcal{H} \to \mathcal{H}$ is a computable nonexpansive operator, i.e., an operator satisfying

$$(\forall z_1, z_2 \in \mathcal{H}) \quad \|T(z_1) - T(z_2)\|_{\mathcal{H}} \leq \|z_1 - z_2\|_{\mathcal{H}}, \tag{16.27}$$

and $\Xi : \mathcal{H} \to 2^{\mathcal{X}}$ is a certain set valued operator. Examples of such characterizations are found in [62, 150] for the augmented Lagrangian method [81, 116], in [44, 45] for the forward-backward splitting approach [66, 109, 134], in [40, Proposition 18(iii)] for the Douglas-Rachford splitting approach (see Section 16.2.3) [91], in [61, 150] for the alternating direction method of multipliers (ADMM) [66, 76, 91], in [47, 139] for the primal-dual splitting method, and in [150] for a generalized version (see Section 16.2.3) of the linearized augmented Lagrangian method [151].

If we find a computable nonexpansive operator T satisfying (16.25) as well as a computationally tractable way to extract a point in $\Xi(z)(\subset \mathcal{X})$ for a given $z \in \mathrm{Fix}(T)$, we can realize an algorithmic solution to Problem (16.1) by applying the so-called *Krasnosel'skiĭ-Mann Iterative Process* (see Fact 16.6 in Section 16.2.2) to T, and can produce a weak convergent sequence to a fixed point $z \in \mathrm{Fix}T$, followed by a point extraction from $\Xi(z)$. Indeed, the powerful proximal splitting methodologies for Problem (16.2) seem to have been built more or less along this strategy through innovative designs of computable nonexpansive operators by using $\mathrm{prox}_f : \mathcal{X} \to \mathcal{X}$ and $\mathrm{prox}_{g_i} : \mathcal{K}_i \to \mathcal{K}_i$ $(i = 1, \ldots, m)$ as computational tools.

On the other hand, every nonexpansive operator $T : \mathcal{H} \to \mathcal{H}$ can also be plugged into the hybrid steepest descent method for minimizing $\Theta \in \Gamma_0(\mathcal{H})$, whose gradient $\nabla\Theta : \mathcal{H} \to \mathcal{H}$ is Lipschitz continuous, over the fixed point set $\mathrm{Fix}(T) \neq \varnothing$ (see Section 16.2.4).[5] Moreover, for Problem (16.1), if such a computable nonexpansive operator T can be used to express $\mathcal{S}_p$ as in (16.25) but more nicely with some computable bounded linear operator $\Xi \in \mathcal{B}(\mathcal{H}, \mathcal{X})$, we can apply the hybrid steepest descent method to Problem (16.13) after translating it into

$$\text{find } z^\star \in \operatorname*{argmin}_{z \in \mathrm{Fix}(T)} \Theta(z), \tag{16.28}$$

[5] By extending the idea in [75], another algorithm, which we refer to as the *generalized Haugazeau's algorithm*, was developed for minimizing a *strictly convex* function in $\Gamma_0(\mathcal{H})$ over the fixed point set of a certain quasi-nonexpansive operator [33]. In particular, this algorithm was specialized in a clear way for finding the nearest fixed point of a certain quasi-nonexpansive operator [8] and applied successfully to an image recovery problem [39]. If we focus on the case of a nonstrictly convex function, the generalized Haugazeau's algorithm is not applicable, while some convergence theorems of the hybrid steepest descent method suggest its sound applicability *provided that the gradient of the function is Lipschitzian*.

where $\Theta := \Psi \circ \Xi$, because $\Theta \in \Gamma_0(\mathcal{H})$ is certainly Gâteaux differentiable with Lipschitzian gradient $\nabla\Theta : z \mapsto \Xi^*\nabla\Psi(\Xi z)$ and $\Xi(z^\star) \in \mathcal{X}$ is a solution of (16.13).

The goal of this paper is to demonstrate that plugging the modern proximal splitting operators into the hybrid steepest descent method is a powerful computational strategy for solving highly valuable hierarchical convex optimization problems (16.10) in Scenario 1 and Scenario 2. The remainder of the paper is organized as follows. In the next section, as preliminaries, we introduce elements of convex analysis and fixed point theoretic view of the modern proximal splitting algorithms. These include key ideas behind fixed point characterizations of $\mathcal{S}_p$ in Problem (16.13) as well as the hybrid steepest descent method for nonexpansive operators. Section 16.3 contains the main idea of the hierarchical convex optimization based on the hybrid steepest descent method applied to modern proximal splitting operators. In Section 16.4, as a typical example of Scenario 1, we present an application of the proposed strategies to a hierarchical enhancement of *the support vector machine* [48, 135, 136] where we demonstrate how we can compute *the best linear classifier which achieves the maximal margin among all linear classifiers having least empirical hinge loss*. The proposed best linear classifier can be applied to general training data whether it is linearly separable or not. In particular, for linearly separable data, the proposed best linear classifier, *which does not require any parameter tuning*, is guaranteed to reproduce successfully the original support vector machine specially defined in [136]. To the best of the authors' knowledge, such a unified generalization of original support vector machine for linearly separable data has not been achieved by previously reported SVMs (see, e.g., [14, 25, 48, 73, 125, 126, 131] and Section 16.4.2). In Section 16.5, as a typical example along Scenario 2, we present an application of the proposed strategy to a hierarchical enhancement of Lasso [73, 132]. This enhancement is achieved by utilizing maximally the Douglas-Rachford splitting applied to a recently established proximity operator [35, 38] of a perspective function for the TREX problem [89] which is certainly *the state-of-the-art nonconvex formulation* for automatic sparsity control of Lasso. The proposed application can optimize further an additional convex criterion over the all solutions of the TREX problem. Finally, in Section 16.6, we conclude this paper with some remarks on other possible advanced applications of the hybrid steepest descent method.

16.2 Preliminary

Let $\mathcal{X}$ be a real Hilbert space equipped with[6] an inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$, which is denoted by $(\mathcal{X}, \langle\cdot,\cdot\rangle, \|\cdot\|)$. Let $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}}, \|\cdot\|_{\mathcal{K}})$ be another real Hilbert space. Let $A : \mathcal{X} \to \mathcal{K}$ be a bounded linear operator of which the

[6] Often $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ denotes $\langle\cdot,\cdot\rangle$ to explicitly describe its domain.

norm is defined by $\|A\|_{\mathrm{op}} := \sup_{x\in\mathcal{X}:\ \|x\|\le 1} \|Ax\|_{\mathcal{K}}$. For a bounded linear operator $A: \mathcal{X} \to \mathcal{K}$, $A^*: \mathcal{K} \to \mathcal{X}$ denotes its adjoint or conjugate, i.e.,

$$(\forall (x, u) \in \mathcal{X} \times \mathcal{K}) \quad \langle x, A^* u\rangle = \langle Ax, u\rangle_{\mathcal{K}}.$$

16.2.1 Selected Elements of Convex Analysis and Optimization

For readers' convenience, we list minimum elements, in convex analysis, which will be used in the later sections (for their detailed accounts, see, e.g., [7, 9, 35, 38, 44, 64, 82, 122, 143, 152]).

(Convex Set) A set $C \subset \mathcal{X}$ is said to be convex if $\lambda x + (1-\lambda)y \in C$ for all $\lambda \in (0, 1)$ and for all $x, y \in C$.

(Proper Lower Semicontinuous Convex Function; See, e.g., [9, Chapter 9]) A function $f: \mathcal{X} \to (-\infty, \infty]$ is said to be proper if its effective domain $\mathrm{dom}(f) := \{x \in \mathcal{X} \mid f(x) < \infty\}$ is nonempty. A function $f: \mathcal{X} \to (-\infty, \infty]$ is said to be lower semicontinuous if its lower level set $\mathrm{lev}_{\le\alpha} f := \{x \in \mathcal{X} \mid f(x) \le \alpha\} (\subset \mathcal{X})$ is closed for every $\alpha \in \mathbb{R}$. A function $f: \mathcal{X} \to (-\infty, \infty]$ is said to be convex if $f(\lambda x+(1-\lambda)y) \le \lambda f(x)+(1-\lambda) f(y)$ for all $\lambda \in (0, 1)$ and for all $x, y \in \mathrm{dom}(f)$. In particular, f is said to be strictly convex if $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$ for all $\lambda \in (0, 1)$ and for all $x, y \in \mathrm{dom}(f)$ such that $x \ne y$. The set of all proper lower-semicontinuous convex functions defined over the real Hilbert space $\mathcal{X}$ is denoted by $\Gamma_0(\mathcal{X})$.

(Coercivity and Supercoercivity; See, e.g., [9, Chapter 11]) A function $f: \mathcal{X} \to (-\infty, \infty]$ is said to be coercive if

$$\|x\| \to \infty \ \Rightarrow\ f(x) \to \infty$$

and supercoercive if

$$\|x\| \to \infty \ \Rightarrow\ \frac{f(x)}{\|x\|} \to \infty.$$

Obviously, supercoercivity of f implies coercivity of f. Coercivity of $f \in \Gamma_0(\mathcal{X})$ implies that $\mathrm{lev}_{\le\alpha} f = \{x \in \mathcal{X} \mid f(x) \le \alpha\}$ is bounded for every $\alpha \in \mathbb{R}$ as well as $\mathrm{argmin}_{x\in\mathcal{X}} f(x) \ne \varnothing$. Strict convexity of $f \in \Gamma_0(\mathcal{X})$ implies that the set of minimizers is at most singleton.

Fact 16.2 (See, e.g., [9, Section 11.4]) *Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$ and $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ such that $\mathrm{dom}(f) \cap \mathrm{dom}(g \circ A) \ne \varnothing$. Then the following conditions*

(*a*) $\mathrm{argmin}(f + g \circ A)(\mathcal{X})$ *is nonempty, closed, and bounded;*

(*b*) $f + g \circ A$ *is coercive;*

(*c*) *f is coercive, and g is bounded below;*

(*d*) *f is super-coercive;*

satisfy that ((*d*) *or* (*c*)) $\Rightarrow$ (*b*) $\Rightarrow$ (*a*).

(Gâteaux Differential; See, e.g., [9, Section 2.6]) Let U be an open subset of $\mathcal{X}$. Then a function $f : U \to \mathbb{R}$ is said to be Gâteaux differentiable at $x \in U$ if there exists $a(x) \in \mathcal{X}$ such that

$$\lim_{\delta \to 0} \frac{f(x + \delta h) - f(x)}{\delta} = \langle a(x), h \rangle \quad (\forall h \in \mathcal{X}). \tag{16.29}$$

In this case, $\nabla f(x) := a(x)$ is called Gâteaux gradient (or gradient) of f at x. Let $f \in \Gamma_0(\mathcal{X})$ be Gâteaux differentiable at $x_\star \in \mathcal{X}$. Then $x_\star$ is a minimizer of f if and only if $\nabla f(x_\star) = 0$.

(Subdifferential; See, e.g., [9, Chapter 16]) For a function $f \in \Gamma_0(\mathcal{X})$, the subdifferential of f is defined as the set valued operator

$$\partial f : \mathcal{X} \to 2^{\mathcal{X}} : x \mapsto \{u \in \mathcal{X} \mid \langle y - x, u \rangle + f(x) \leq f(y), \forall y \in \mathcal{X}\}.$$

Every element $u \in \partial f(x)$ is called a subgradient of f at x. For a given function $f \in \Gamma_0(\mathcal{X})$, $x_\star \in \mathcal{X}$ is a minimizer of f if and only if $0 \in \partial f(x_\star)$. Note that if $f \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable at $x \in \mathcal{X}$, then $\partial f(x) := \{\nabla f(x)\}$.

(Conjugate Function; See, e.g., [9, Chapter 13 and Chapter 16]) For a function $f \in \Gamma_0(\mathcal{X})$, the conjugate of f is defined by

$$f^* : \mathcal{X} \to [-\infty, \infty] : u \mapsto \sup_{x \in \mathcal{X}} (\langle x, u \rangle - f(x)) = \sup_{x \in \mathrm{dom}(f)} (\langle x, u \rangle - f(x)).$$

Let $f \in \Gamma_0(\mathcal{X})$. Then $f^* \in \Gamma_0(\mathcal{X})$ and $f^{**} = f$ are guaranteed. Moreover, we have

$$(\forall (x, u) \in \mathcal{X} \times \mathcal{X}) \quad u \in \partial f(x) \Leftrightarrow f(x) + f^*(u) = \langle x, u \rangle \Leftrightarrow x \in \partial f^*(u),$$

which implies that $(\partial f)^{-1}(u) := \{x \in \mathcal{X} \mid u \in \partial f(x)\} = \partial f^*(u)$ and $(\partial f^*)^{-1}(x) := \{u \in \mathcal{X} \mid x \in \partial f^*(u)\} = \partial f(x)$. Often

$$(\forall x \in \mathrm{dom}(\partial f))(\forall u \in \partial f(x)) \quad f(x) + f^*(u) = \langle x, u \rangle \tag{16.30}$$

is referred to as Fenchel-Young identity.

For hierarchical enhancement of Lasso in Section 16.5, we exploit the following nontrivial example.

Example 16.3 (Subdifferential of Perspective; See [38, Lemma 2.3]) For given supercoercive $\varphi \in \Gamma_0(\mathbb{R}^N)$, the function

$$\widetilde{\varphi} : \mathbb{R} \times \mathbb{R}^N \to (-\infty, \infty] : (\eta, \mathbf{y}) \mapsto \begin{cases} \eta\varphi(\mathbf{y}/\eta), & \text{if } \eta > 0; \\ \sup\limits_{\mathbf{x} \in \mathrm{dom}(\varphi)} [\varphi(\mathbf{x}+\mathbf{y}) - \varphi(\mathbf{x})], & \text{if } \eta = 0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{16.31}$$

satisfies $\widetilde{\varphi} \in \Gamma_0\left(\mathbb{R} \times \mathbb{R}^N\right)$ and is called the perspective of φ.

The subdifferential of $\widetilde{\varphi}$ is given by

$$\partial\widetilde{\varphi}(\eta, \mathbf{y}) = \begin{cases} \left\{(\varphi(\mathbf{y}/\eta) - \langle \mathbf{y}/\eta, \mathbf{u}\rangle, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^N \mid \mathbf{u} \in \partial\varphi(\mathbf{y}/\eta)\right\}, & \text{if } \eta > 0; \\ \{(\mu, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^N \mid \mu + \varphi^*(\mathbf{u}) \leq 0\}, & \text{if } \eta = 0 \text{ and } \mathbf{y} = \mathbf{0}; \\ \varnothing, & \text{otherwise.} \end{cases} \tag{16.32}$$

(Conical Hull, Span, Convex Sets; See, e.g., [9, Chapter 6]) For a given nonempty set $C \subset \mathcal{X}$, $\mathrm{cone}(C) := \{\lambda x \mid \lambda > 0, x \in C\}$ is called the conical hull of C, and $\mathrm{span}(C)$ denotes the intersection of all the linear subspaces of $\mathcal{X}$ containing C. The closure of $\mathrm{span}(C)$ is denoted by $\overline{\mathrm{span}}(C)$.

The strong relative interior of a convex set $C \subset \mathcal{X}$ is defined by

$$\mathrm{sri}(C) := \{x \in C \mid \mathrm{cone}(C - x) = \overline{\mathrm{span}}(C - x)\},$$

where $C - x := \{y - x \in \mathcal{X} \mid y \in C\}$.

Similarly, the relative interior of a convex set $C \subset \mathcal{X}$ is defined by

$$\mathrm{ri}(C) := \{x \in C \mid \mathrm{cone}(C - x) = \mathrm{span}(C - x)\}.$$

By $\mathrm{cone}(C - x) \subset \mathrm{span}(C - x) \subset \overline{\mathrm{span}}(C - x)$ for every $x \in C$, we have $\mathrm{sri}(C) \subset \mathrm{ri}(C)$. Moreover, $\mathrm{sri}(C) = \mathrm{ri}(C)$ if $\mathrm{span}(C - x) = \overline{\mathrm{span}}(C - x)$ for every $x \in C$, which implies

$$\dim(\mathcal{X}) < \infty \;\Rightarrow\; \mathrm{sri}(C) = \mathrm{ri}(C). \tag{16.33}$$

(Indicator Function) For a nonempty closed convex set $C \subset \mathcal{X}$, the indicator function of C is defined by

$$\iota_C : \mathcal{X} \to (-\infty, \infty] : x \mapsto \begin{cases} 0, & \text{if } x \in C; \\ +\infty, & \text{otherwise,} \end{cases}$$

which belongs to $\Gamma_0(\mathcal{X})$. In particular, for a closed subspace $V \subset \mathcal{X}$,

$$u \in \partial\iota_V(x) \;\Leftrightarrow\; x \in V \text{ and } u \in V^\perp := \{y \in \mathcal{X} \mid (\forall v \in V)\ \langle v, y\rangle = 0\}. \tag{16.34}$$

Furthermore, the indicator function $\iota_{\{0\}} \in \Gamma_0(\mathcal{X})$ of $\{0\} \subset \mathcal{X}$ has the following properties: for all $x, u \in \mathcal{X}$

$$\partial \iota_{\{0\}}(x) = \begin{cases} \mathcal{X}, \text{ if } x = 0; \\ \varnothing, \text{ otherwise,} \end{cases} \tag{16.35}$$

$$\iota^*_{\{0\}}(u) = \sup_{y \in \mathcal{X}} (\langle y, u \rangle - \iota_{\{0\}}(y)) = 0, \tag{16.36}$$

$$\partial \iota^*_{\{0\}}(u) = \{0\}. \tag{16.37}$$

(Fenchel-Rockafellar Duality for Convex Optimization Problem Involving Linear Operator; See, e.g., [9, Definition 15.19]) Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$, and $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$. The primal problem associated with the composite function $f + g \circ A$ is

$$\text{minimize}_{x \in \mathcal{X}} \ f(x) + g(Ax), \tag{16.38}$$

its dual problem is

$$\text{minimize}_{u \in \mathcal{K}} \ f^*(A^*u) + g^*(-u), \tag{16.39}$$

$\mu := \inf_{x \in \mathcal{X}}(f(x) + g(Ax))$ is called the primal optimal value, and $\mu^* := \inf_{u \in \mathcal{K}}(f^*(A^*u) + g^*(-u))$ the dual optimal value.

Fact 16.4 (See, e.g., [9, Theorem 15.23, Theorem 16.47, Corollary 16.53]) *The condition*

$$\left.\begin{array}{l} 0 \in sri(\mathrm{dom}(g) - A\,\mathrm{dom}(f)) \\ (\text{sri } \textit{can be replaced by} \text{ ri } \textit{in the case of} \dim(\mathcal{K}) < \infty, \textit{ see } (16.33)) \end{array}\right\} \tag{16.40}$$

is the so-called qualification condition *for problem* (16.38).

(a) *The condition* (16.40) *guarantees that the dual problem* (16.39) *has a minimizer and satisfies*

$$\mu = \inf_{x \in \mathcal{X}}(f(x) + g(Ax)) = -\min_{u \in \mathcal{K}}(f^*(A^*u) + g^*(-u)) = -\mu^*; \tag{16.41}$$

(b) *The condition* (16.40) *guarantees that the subdifferential of* $f + g \circ A$ *can be decomposed as*

$$\partial(f + g \circ A) = \partial f + A^* \circ (\partial g) \circ A; \tag{16.42}$$

(c) *The qualification condition* (16.40) *with* $f \equiv 0$ *becomes* $0 \in \mathrm{sri}(\mathrm{dom}(g) - \mathrm{ran}(A))$, *where* $\mathrm{ran}(A) := A(\mathcal{X}) := \{Ax \in \mathcal{K} \mid x \in \mathcal{X}\}$. *Under this condition,* (a), (b), *and* (16.36) *guarantee*

$$\begin{cases} \mu = \inf_{x\in\mathcal{X}} g(Ax) = \inf_{x\in\mathcal{X}} \iota^*_{\{0\}}(x) + g(Ax) = -\min_{u\in\mathcal{K}}(\iota_{\{0\}}(A^*u) + g^*(-u)) \\ \qquad\qquad = -\min_{u\in\mathcal{N}(A^*)} g^*(-u) = -\mu^* \\ \partial(g\circ A) = A^*\circ\partial g\circ A. \end{cases}$$

Fact 16.5 ([9, Theorem 19.1]) *Suppose that* $0 \in \mathrm{dom}(g) - A\,\mathrm{dom}(f)$ *(Note: This condition is not sufficient for* (16.40)*). Let* $(x, u) \in \mathcal{X}\times\mathcal{K}$*. Then the following are equivalent:*

(i) x *is a solution of the primal problem* (16.38)*,* u *is a solution of the dual problem* (16.39)*, and* $\mu = -\mu^*$.
(ii) $A^*u \in \partial f(x)$ *and* $-u \in \partial g(Ax)$.
(iii) $x \in \partial f^*(A^*u) \cap A^{-1}(\partial g^*(-u))$.

16.2.2 Selected Elements of Fixed Point Theory of Nonexpansive Operators for Application to Hierarchical Convex Optimization

For readers' convenience, we list minimum elements in fixed point theory of nonexpansive mapping specially for application to hierarchical convex optimization in this paper (for their detailed accounts, see, e.g., [7, 9, 38, 42, 44, 51, 82, 122, 130, 143]).

(Monotone Operator; See, e.g., [9, Section 20.1]) A set-valued operator $T : \mathcal{X} \to 2^{\mathcal{X}}$ is said to be monotone over $S(\subset \mathcal{X})$ if

$$(\forall x, y \in S)(\forall u \in Tx)(\forall v \in Ty) \quad \langle u - v, x - y\rangle \geq 0.$$

In particular, it is said to be η-strongly monotone over S if

$$(\exists \eta > 0)(\forall x, y \in S)(\forall u \in Tx)(\forall v \in Ty) \quad \langle u - v, x - y\rangle \geq \eta\|x - y\|^2.$$

(Nonexpansive Operator; See, e.g., [7] and [9, Chapter 4]) An operator $T : \mathcal{X} \to \mathcal{X}$ is said to be Lipschitz continuous with Lipschitz constant $\kappa > 0$ (or κ-Lipschitzian) if

$$(\forall x, y \in \mathcal{X}) \quad \|Tx - Ty\| \leq \kappa\|x - y\|.$$

In particular, an operator $T : \mathcal{X} \to \mathcal{X}$ is said to be nonexpansive if it is 1-Lipschitzian, i.e.,

$$(\forall x, y \in \mathcal{X}) \quad \|Tx - Ty\| \leq \|x - y\|.$$

A nonexpansive operator T is said to be α-averaged (or averaged with constant α) [5, 9] if there exist $\alpha \in (0, 1)$ and a nonexpansive operator $\widehat{T} : \mathcal{X} \to \mathcal{X}$ such that

$$T = (1 - \alpha)\mathrm{I} + \alpha\widehat{T}, \tag{16.43}$$

i.e., T is an *average* of the identity operator I and some nonexpansive operator $\widehat{T}$. If (16.43) holds for $\alpha = 1/2$, T is said to be firmly nonexpansive. A nonexpansive operator T is α-averaged if and only if

$$(\forall x, y \in \mathcal{X}) \quad \|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{1-\alpha}{\alpha}\|(x - Tx) - (y - Ty)\|^2. \tag{16.44}$$

Suppose that a nonexpansive operator T has the fixed point set $\mathrm{Fix}(T) := \{x \in \mathcal{X} \mid Tx = x\} \neq \varnothing$. Then $\mathrm{Fix}(T)$ can be expressed as the intersection of closed halfspaces:

$$\mathrm{Fix}(T) = \bigcap_{y \in \mathcal{X}} \left\{ x \in \mathcal{X} \mid \langle y - T(y), x \rangle \leq \frac{\|y\|^2 - \|T(y)\|^2}{2} \right\}$$

and therefore $\mathrm{Fix}(T)$ is closed and convex (see, e.g., [70, Proposition 5.3], [142, Fact 2.1(a)], and [9, Corollary 4.24]). In addition, a nonexpansive operator T with $\mathrm{Fix}(T) \neq \varnothing$ is said to be attracting [7] if

$$(\forall x \notin \mathrm{Fix}(T))(\forall z \in \mathrm{Fix}(T)) \quad \|Tx - z\| < \|x - z\|.$$

The condition (16.44) implies that α-averaged nonexpansive operator T is attracting if $\mathrm{Fix}(T) \neq \varnothing$. Note that other useful properties on α-averaged nonexpansive operators are found, e.g., in [20, 45, 105].

Fact 16.6 (Krasnosel'skiĭ–Mann (KM) Iteration [71] (See Also [9, Section 5.2], [20, 56, 88, 96, 119])) *For a nonexpansive operator $T : \mathcal{X} \to \mathcal{X}$ with $\mathrm{Fix}(T) \neq \varnothing$ and any initial point $x_0 \in \mathcal{X}$, the sequence $(x_n)_{n\in\mathbb{N}}$ generated by*

$$x_{n+1} = (1 - \alpha_n)x_n + \alpha_n T x_n$$

converges weakly[7] *to a point in $\mathrm{Fix}(T)$ if $(\alpha_n)_{n\in\mathbb{N}} \subset [0, 1]$ satisfies $\sum_{n\in\mathbb{N}} \alpha_n(1 - \alpha_n) = \infty$ (Note: The weak limit of $(x_n)_{n\in\mathbb{N}}$ depends on the choices of x_0 and $(\alpha_n)_{n\in\mathbb{N}}$).*[8] *In particular, if T is α-averaged for some $\alpha \in (0, 1)$ (see (16.43)), a simple iteration*

[7](Strong and weak convergences) A sequence $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ is said to converge strongly to a point $x \in \mathcal{X}$ if the real number sequence $(\|x_n - x\|)_{n\in\mathbb{N}}$ converges to 0, and to converge weakly to $x \in \mathcal{X}$ if for every $y \in \mathcal{X}$ the real number sequence $(\langle x_n - x, y\rangle)_{n\in\mathbb{N}}$ converges to 0. If $(x_n)_{n\in\mathbb{N}}$ converges strongly to x, then $(x_n)_{n\in\mathbb{N}}$ converges weakly to x. The converse is true if $\mathcal{X}$ is finite dimensional, hence in finite dimensional case we do not need to distinguish these convergences.

(Sequential cluster point) If a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ possesses a subsequence that strongly (weakly) converges to a point $x \in \mathcal{X}$, then x is a strong (weak) sequential cluster point of $(x_n)_{n\in\mathbb{N}}$. For weak topology of real Hilbert space in the context of Hausdorff space, see [9, Lemma 2.30].

[8]Some extensions to uniformly convex Banach spaces are found in [71, 119].

$$x_{n+1} = Tx_n = (1-\alpha)x_n + \alpha\widehat{T}x_n \tag{16.45}$$

converges weakly to a point in $\mathrm{Fix}(T) = \mathrm{Fix}(\widehat{T})$.

(Proximity Operator [101, 102] (See Also [9, Chapter 24])) The proximity operator of $f \in \Gamma_0(\mathcal{X})$ is defined by

$$\mathrm{prox}_f : \mathcal{X} \to \mathcal{X} : x \mapsto \underset{y\in\mathcal{X}}{\mathrm{argmin}}\ f(y) + \frac{1}{2}\|y-x\|^2.$$

Note that $\mathrm{prox}_f(x) \in \mathcal{X}$ is well defined for all $x \in \mathcal{X}$ due to the coercivity and the strict convexity of $f(\cdot) + \frac{1}{2}\|\cdot - x\|^2 \in \Gamma_0(\mathcal{X})$. It is also well known that prox_f is nothing but the resolvent of ∂f, i.e., $\mathrm{prox}_f = (\mathrm{I}+\partial f)^{-1} =: J_{\partial f}$, which implies that

$$\begin{aligned} x \in \mathrm{Fix}(\mathrm{prox}_f) \ &\Leftrightarrow\ \mathrm{prox}_f(x) = x \ \Leftrightarrow\ (\mathrm{I}+\partial f)^{-1}(x) = x \\ &\Leftrightarrow\ x \in (\mathrm{I}+\partial f)(x) \ \Leftrightarrow\ 0 \in \partial f(x) \ \Leftrightarrow\ x \in \underset{y\in\mathcal{X}}{\mathrm{argmin}}\ f(y). \end{aligned} \tag{16.46}$$

Thanks to this fact, the set of all minimizers of $f \in \Gamma_0(\mathcal{X})$ can be characterized in terms of a single-valued map, i.e., prox_f. Moreover, since the proximity operator is $1/2$-averaged nonexpansive, i.e., $\mathrm{rprox}_f := 2\mathrm{prox}_f - \mathrm{I}$ is nonexpansive, the iteration

$$x_{n+1} = \mathrm{prox}_f(x_n) \tag{16.47}$$

converges weakly to a point in $\mathrm{argmin}_{x\in\mathcal{X}} f(x) = \mathrm{Fix}(\mathrm{prox}_f)$ by (16.45) in Fact 16.6. The iterative algorithm (16.47) is known as *proximal point algorithm* [121] (see (16.7)).

In this paper, $f \in \Gamma_0(\mathcal{X})$ is said to be *proximable* if prox_f is available as a computable operator. Note that if $f \in \Gamma_0(\mathcal{X})$ is proximable, so is $f^* \in \Gamma_0(\mathcal{X})$. This is verified by

$$\mathrm{prox}_{f^*} = J_{\partial f^*} = J_{(\partial f)^{-1}} = \mathrm{I} - J_{\partial f} = \mathrm{I} - \mathrm{prox}_f,$$

which is a special example of *the inverse resolvent identity* [9, Proposition 23.20]. Note that the sum of two proximable convex functions is not necessarily proximable. Moreover, for $A \in \mathcal{B}(\mathcal{X},\mathcal{K})$, the composition $g \circ A \in \Gamma_0(\mathcal{X})$ for a proximable function $g \in \Gamma_0(\mathcal{K})$ is not necessarily proximable. There are many useful formula to compute the proximity operator (see, e.g., [9, Chapter 24], [42]).

Example 16.7

(a) **(Indicator function; see, e.g., [9, Example 12.25])** For a nonempty closed convex set $C \subset \mathcal{X}$,

$$(\forall x \in \mathcal{X}) \quad \mathrm{prox}_{\iota_C}(x) = \underset{y\in\mathcal{X}}{\mathrm{argmin}}\left(\iota_C(y) + \frac{1}{2}\|y-x\|^2\right) = \underset{y\in C}{\mathrm{argmin}}\ \frac{1}{2}\|y-x\|^2 =: P_C(x)$$

holds, which implies that prox_{ι_C} is identical to the *metric projection* onto C. In particular, if ι_C is proximable, C is said to be *simple*.

(b) **(Semi-orthogonal linear transform of proximable function; see, e.g., [9, Proposition 24.14] and [42, Table 10.1])**
For $g \in \Gamma_0(\mathcal{K})$ and $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ such that $AA^* = \nu \mathrm{I}$ with some $\nu > 0$,

$$(\forall x \in \mathcal{X})\ \mathrm{prox}_{g\circ A}(x) = x + \nu^{-1}A^*(\mathrm{prox}_{\nu g}(Ax) - Ax). \tag{16.48}$$

(c) **(Hinge loss function; see, e.g., [1] and [9, Example 24.36])** For $\gamma > 0$ and

$$h : \mathbb{R} \to [0, \infty) : t \mapsto \max\{0, 1 - t\}, \tag{16.49}$$

$$(\forall t \in \mathbb{R}) \quad \mathrm{prox}_{\gamma h}(t) = \min\{t + \gamma, \max\{t, 1\}\}. \tag{16.50}$$

(d) **(ℓ_1 norm; see, e.g., [9, 44])** For $\gamma \geq 0$ and the ℓ_1 norm $\|\cdot\|_1 \in \Gamma_0(\mathbb{R}^N)$

$$(\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N) \quad \|\boldsymbol{x}\|_1 := \sum_{j=1}^{N} |x_i|,$$

the i-th component of the proximity operator of $\gamma\|\cdot\|_1$ is given as

$$(\forall \boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^N) \quad [\mathrm{prox}_{\gamma\|\cdot\|_1}(\boldsymbol{x})]_i = \begin{cases} x_i - \mathrm{sgn}(x_i)\gamma, & \text{if } |x_i| > \gamma; \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathrm{sgn}\colon \mathbb{R} \to \mathbb{R}$ is the signum function, i.e., $\mathrm{sgn}(x) = 0$ if $x = 0$ and $\mathrm{sgn}(x) = x/|x|$ otherwise. $\mathrm{prox}_{\gamma\|\cdot\|_1}$ is also known as soft-thresholding [53, 54].

(e) **(Proximity operator of *perspective* of $\|\cdot\|^q$; see e.g. [38])** Let $\beta > 0$ and $q > 1$. The perspective $\widetilde{\varphi}_q$ of $\varphi_q(\cdot) := \|\cdot\|^q/\beta$ (see also (16.31) in Example 16.3) is given by

$$\widetilde{\varphi}_q : \mathbb{R} \times \mathbb{R}^N \to (-\infty, \infty] : (\eta, \mathbf{y}) \mapsto \begin{cases} \frac{\|\mathbf{y}\|^q}{\beta\eta^{q-1}}, & \text{if } \eta > 0; \\ 0, & \text{if } \eta = 0 \text{ and } \mathbf{y} = \mathbf{0}; \\ +\infty, & \text{otherwise,} \end{cases} \tag{16.51}$$

and its proximity operator can be expressed as

$$\begin{aligned} \mathrm{prox}_{\widetilde{\varphi}_q} &: \mathbb{R} \times \mathbb{R}^N \to \mathbb{R} \times \mathbb{R}^N \\ &: (\eta, \mathbf{y}) \mapsto \begin{cases} \left(\eta + \frac{\varrho}{q^*}\|\mathbf{p}\|^{q^*}, \mathbf{y} - \mathbf{p}\right), & \text{if } q^*\eta + \varrho\|\mathbf{y}\|^{q^*} > 0; \\ (0, \mathbf{0}), & \text{if } q^*\eta + \varrho\|\mathbf{y}\|^{q^*} \leq 0, \end{cases} \end{aligned} \tag{16.52}$$

where $q^* := \dfrac{q}{q-1}$, $\varrho := \left(\beta(1 - 1/q^*)\right)^{q^*-1}$, $\mathbf{p} := \begin{cases} \tau\frac{\mathbf{y}}{\|\mathbf{y}\|}, & \text{if } \mathbf{y} \neq \mathbf{0}; \\ \mathbf{0}, & \text{if } \mathbf{y} = \mathbf{0}, \end{cases}$

and $\tau \in (0, \infty)$ is uniquely determined as the solution to the equation:

$$\tau^{2q^*-1} + \frac{q^*\eta}{\varrho}\tau^{q^*-1} + \frac{q^*\|\mathbf{y}\|}{\varrho^2} = 0.$$

The proximity operator of the translation, by $(a, \mathbf{b}) \in \mathbb{R} \times \mathbb{R}^N$, of $\widetilde{\varphi}_q$

$$_{\tau_{(a,\mathbf{b})}}\widetilde{\varphi}_q : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R} \times \mathbb{R}^N : (\eta, \mathbf{y}) \mapsto \widetilde{\varphi}_q\,(\eta - a, \mathbf{y} - \mathbf{b})\,,$$

which can be expressed as

$$\mathrm{prox}_{_{\tau_{(a,\mathbf{b})}}\widetilde{\varphi}_q} : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R} \times \mathbb{R}^N : (\eta, \mathbf{y}) \mapsto (a, \mathbf{b}) + \mathrm{prox}_{\widetilde{\varphi}_q}\,(\eta - a, \mathbf{y} - \mathbf{b})\,,$$

will play an important role in Section 16.5.

Fact 16.8 (Moreau Envelope (See, e.g., [9, Section 12.4], [101, 102])) *For $f \in \Gamma_0(\mathcal{X})$,*

$$^{\gamma}f : \mathcal{X} \to \mathbb{R} : x \mapsto \min_{y\in\mathcal{X}} \left(f(y) + \frac{1}{2\gamma}\|x - y\|^2 \right)$$

is called the Moreau envelope *(or* Moreau-Yosida regularization*) [101, 102] of f of the index $\gamma > 0$. The function $^{\gamma}f$ is Gâteaux differentiable convex with Lipschitzian gradient*

$$\nabla^{\gamma} f : \mathcal{X} \to \mathcal{X} : x \mapsto \frac{1}{\gamma}(\mathrm{I} - \mathrm{prox}_{\gamma f}(x)).$$

The Moreau envelope of f converges pointwise to f on $\mathrm{dom}(f)$ *as $\gamma \downarrow 0$ (see, e.g., [9, Proposition 12.33(ii)]), i.e.,* $\lim_{\gamma\downarrow 0}{}^{\gamma}f(x) = f(x)$ $(\forall x \in \mathrm{dom}(f))$.

16.2.3 Proximal Splitting Algorithms and Their Fixed Point Characterizations

In this section, we introduce the Douglas-Rachford splitting method[9] (see, e.g., [9, 10, 34, 40, 61, 91]) and the linearized augmented Lagrangian method (see, e.g., [150, 151]) as examples of *the proximal splitting algorithms* built on computable nonexpansive operators with a great deal of potential in their applications to the

[9] See [10, 42] for the history of the Douglas-Rachford splitting method, originated from Douglas-Rachford's seminal paper [57] for solving matrix equations of the form $u = Ax + Bx$, where A and B are positive-definite matrices (see also [137]). For recent applications, of the Douglas-Rachford splitting method, to image recovery, see, e.g., [26, 40, 58, 60], and to data sciences, see, e.g., [38, 67, 68]. Lastly, we remark that it was shown in [61] that *the alternating direction method of multipliers (ADMM)* [17, 62, 66, 91, 150] can be seen as a dual variant of the Douglas-Rachford splitting method.

hierarchical convex optimization problem. As explained briefly just after (16.25–16.27), these proximal splitting algorithms are essentially realized by applying Fact 16.6 (see Section 16.2.2) to certain computable nonexpansive operators.

Proposition 16.9 (DRS Operator and Douglas-Rachford Splitting Method[10]) *Let* $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}})$ *be a real Hilbert space and* $f, g \in \Gamma_0(\mathcal{X})$. *Suppose that*

$$\operatorname{argmin}(f+g)(\mathcal{X}) \neq \varnothing, \tag{16.53}$$

$$\operatorname{argmin}(f^* + g^* \circ (-I))(\mathcal{X}) \neq \varnothing, \tag{16.54}$$

$$\min(f+g)(\mathcal{X}) = -\min(f^* + g^* \circ (-I))(\mathcal{X}). \tag{16.55}$$

Then the DRS operator

$$T_{DRS} := (2\operatorname{prox}_f - I) \circ (2\operatorname{prox}_g - I) \tag{16.56}$$

satisfies:

(a) $\operatorname{prox}_g(\operatorname{Fix}(T_{DRS})) = \operatorname{argmin}(f+g)(\mathcal{X})$;
(b) T_{DRS} *is nonexpansive;*
(c) *By using* $(\alpha_n)_{n\in\mathbb{N}} \subset [0,1]$ *satisfying* $\sum_{n\in\mathbb{N}} \alpha_n(1-\alpha_n) = \infty$ *in Fact 16.6 (see Section 16.2.2), the sequence* $(y_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ *generated by*

$$y_{n+1} = (1-\alpha_n)y_n + \alpha_n T_{DRS}(y_n) \tag{16.57}$$

converges weakly to a point in $\operatorname{Fix}(T_{DRS})$. *Moreover,* $(\operatorname{prox}_g(y_n))_{n\in\mathbb{N}}$ *converges weakly to a point in* $\operatorname{argmin}(f+g)(\mathcal{X})$.

The iterative algorithm to produce $(\operatorname{prox}_g(y_n))_{n\in\mathbb{N}}$ with (16.57) can be seen as a simplest example of the so-called Douglas-Rachford splitting method.

The proof of Proposition 16.9(a) is given in Appendix A because the conditions (16.53–16.55) are newly imposed for applications of T_{DRS} (in (16.56)) to hierarchical convex optimizations in Theorem 16.15 and in Theorem 16.17 (see Remark 16.16(b) and Remark 16.18(b) in Section 16.3.1) and different from [40, Condition (6)] which is also in the context of convex optimization. Proposition 16.9(b) is obvious from the properties of the proximity operator just after (16.46). For weak convergence of $(\operatorname{prox}_g(y_n))_{n\in\mathbb{N}}$ in Proposition 16.9(c), see, e.g., [9, Corollary 28.3(iii)] while the weak convergence of $(y_n)_{n\in\mathbb{N}}$ is obvious from Fact 16.6.

The linearized augmented Lagrangian method (LALM) seems to have been proposed originally as an algorithmic solution to the minimization of the nuclear norm of a matrix subject to a linear constraint [151]. Inspired by the operator defined

[10] We should remark that Proposition 16.9 can also be reproduced from [9, Proposition 26.1(iii) and Theorem 26.11(i)(iii)] in the context of the monotone inclusion problems. For completeness, we present Proposition 16.9 and its proof in the scenario of convex optimization.

as the iterative update [151, (3.7)] in the method for this special convex optimization problem, we extended in [150] the operator to T_{LAL} in (16.61) to be applicable to the general convex optimization problem (16.1) and showed the nonexpansiveness of T_{LAL} for solving efficiently the hierarchical convex optimization (16.13) by plugging the extended operator T_{LAL} into the HSDM.

Proposition 16.10 (LAL Operator and Linearized Augmented Lagrangian Method) *Let* $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}}, \|\cdot\|_{\mathcal{X}})$ *and* $(\mathcal{K}, \langle\cdot,\cdot\rangle_{\mathcal{K}}, \|\cdot\|_{\mathcal{K}})$ *be real Hilbert spaces. Suppose that* $f \in \Gamma_0(\mathcal{X})$, $g = \iota_{\{0\}} \in \Gamma_0(\mathcal{K})$ *and* $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ *satisfy*

$$\mathcal{S}_{pLAL} := \operatorname{argmin}(f + \iota_{\{0\}} \circ A)(\mathcal{X}) \neq \varnothing, \tag{16.58}$$

$$\mathcal{S}_{dLAL} := \operatorname{argmin}(f^* \circ A^*)(\mathcal{K}) \neq \varnothing, \tag{16.59}$$

$$\min(f + \iota_{\{0\}} \circ A)(\mathcal{X}) = -\min(f^* \circ A^*)(\mathcal{K}), \tag{16.60}$$

where $\mathcal{S}_{pLAL}$ *is the solution set of the primal problem and* $\mathcal{S}_{dLAL}$ *is the solution set of the dual problem. Define the LAL operator* $T_{LAL}\colon \mathcal{X} \times \mathcal{K} \to \mathcal{X} \times \mathcal{K}\colon (x, v) \mapsto (x_T, v_T)$ *by*

$$\begin{cases} x_T := \operatorname{prox}_f(x - A^*Ax + A^*v) \\ v_T := v - Ax_T. \end{cases} \tag{16.61}$$

Then

(a) $\operatorname{Fix}(T_{LAL}) = \mathcal{S}_{pLAL} \times \mathcal{S}_{dLAL}$;
(b) T_{LAL} *is nonexpansive if* $\|A\|_{op} \leq 1$;
(c) *By using* $(\alpha_n)_{n\in\mathbb{N}} \subset [0, 1]$ *satisfying* $\sum_{n\in\mathbb{N}} \alpha_n(1 - \alpha_n) = \infty$ *in Fact 16.6 (see Section 16.2.2), the sequence* $(x_n, v_n)_{n\in\mathbb{N}} \subset \mathcal{X} \times \mathcal{K}$ *generated by*

$$(x_{n+1}, v_{n+1}) = (1 - \alpha_n)(x_n, v_n) + \alpha_n T_{LAL}(x_n, v_n) \tag{16.62}$$

converges weakly to a point in $\mathcal{S}_{pLAL} \times \mathcal{S}_{dLAL}$ *if* $\|A\|_{op} \leq 1$;
(d) *If* $\|A\|_{op} < 1$, *the sequence* $(x_n, v_n)_{n\in\mathbb{N}} \subset \mathcal{X} \times \mathcal{K}$ *generated by* (16.62) *with* $\alpha_n = 1$ $(n \in \mathbb{N})$ *converges weakly to a point in* $\mathcal{S}_{pLAL} \times \mathcal{S}_{dLAL}$.

The iterative algorithms, in Proposition 16.10 (c) and (d), to produce $(x_n)_{n\in\mathbb{N}}$ with (16.62) can be seen as simplest examples of the so-called linearized augmented Lagrangian method.

The proof of Proposition 16.10(a) is given in Appendix B for completeness because the conditions (16.58–16.60) are newly imposed for applications of T_{LAL} to hierarchical convex optimizations in Theorem 16.19 and in Theorem 16.21 (see Remark 16.20(b) and Remark 16.22(a) in Section 16.3.2) and different from [150, (32)]. For the proof of Proposition 16.10(b), see [150]. Proposition 16.10(c) is a straightforward application of Fact 16.6 to Proposition 16.10(b). The proof of Proposition 16.10(d) is given in Appendix B.

Remark 16.11 A primitive idea behind the update of the LAL operator T_{LAL} is found in minimization of the augmented Lagrangian function [81, 116]:

$$\mathcal{L}\colon \mathcal{X}\times\mathcal{K}\to(-\infty,\infty]\colon(x,v)\mapsto f(x)-\langle v,Ax\rangle_{\mathcal{K}}+\frac{1}{2}\|Ax\|_{\mathcal{K}}^2. \tag{16.63}$$

Indeed, by introducing

$$(\forall\hat{x}\in\mathcal{X})(\forall\hat{v}\in\mathcal{K})\begin{bmatrix}\mathcal{L}_1^{(\hat{v})}\colon \mathcal{X}\to(-\infty,\infty]\colon x\mapsto\mathcal{L}(x,\hat{v});\\ \mathcal{L}_2^{(\hat{x})}\colon \mathcal{K}\to(-\infty,\infty]\colon v\mapsto\mathcal{L}(\hat{x},v),\end{bmatrix}$$

the zero $(x_\star,v_\star)\in\mathcal{X}\times\mathcal{K}$ of the partial subdifferentials of (16.63) is characterized as

$$\begin{aligned}\begin{bmatrix}0\in\partial\mathcal{L}_1^{(v_\star)}(x_\star)\\ 0\in\partial\mathcal{L}_2^{(x_\star)}(v_\star)\end{bmatrix}&\Leftrightarrow\begin{bmatrix}0\in\partial f(x_\star)-A^*v_\star+A^*(Ax_\star)\\ 0=-Ax_\star\end{bmatrix}\\&\Leftrightarrow\begin{bmatrix}x_\star=\mathrm{prox}_f(x_\star-A^*Ax_\star+A^*v_\star)\\ v_\star=v_\star-Ax_\star\end{bmatrix}\\&\Leftrightarrow(x_\star,v_\star)\in\mathrm{Fix}(T_{\mathrm{LAL}}).\end{aligned}$$

16.2.4 Hybrid Steepest Descent Method

Consider the problem

$$\text{find } x_\star\in\operatorname*{argmin}_{x\in\mathrm{Fix}(T)}\Theta(x)=:\Omega\neq\varnothing, \tag{16.64}$$

where $\Theta\in\Gamma_0(\mathcal{H})$ is Gâteaux differentiable over $T(\mathcal{H})$ and $T\colon\mathcal{H}\to\mathcal{H}$ is a nonexpansive operator with $\mathrm{Fix}(T)\neq\varnothing$. The hybrid steepest descent method (HSDM)

$$x_{n+1}=T(x_n)-\lambda_{n+1}\nabla\Theta(T(x_n)) \tag{16.65}$$

generates a sequence $(x_n)_{n\in\mathbb{N}}$ to approximate successively a solution of Problem (16.64).

Fact 16.12 (Hybrid Steepest Descent Method for Nonexpansive Operators)

I. [142, special case of Theorems 3.2 and 3.3 for more general variational inequality problems] Let $T\colon\mathcal{H}\to\mathcal{H}$ be a nonexpansive mapping with $\mathrm{Fix}(T)\neq\varnothing$. Suppose that the gradient $\nabla\Theta$ is κ-Lipschitzian and η-strongly monotone over $T(\mathcal{H}):=\{T(x)\in\mathcal{H}\mid x\in\mathcal{H}\}$, which guarantees

$|\Omega| = 1$. *Then, by using any sequence* $(\lambda_{n+1})_{n\in\mathbb{N}} \subset [0,\infty)$ *satisfying (W1)* $\lim_{n\to+\infty}\lambda_n = 0$, *(W2)* $\sum_{n\in\mathbb{N}}\lambda_{n+1} = +\infty$, *(W3)* $\sum_{n\in\mathbb{N}}|\lambda_{n+1}-\lambda_{n+2}| < \infty$ *[or* $(\lambda_{n+1})_{n\in\mathbb{N}} \subset (0,\infty)$ *satisfying (L1)* $\lim_{n\to+\infty}\lambda_n = 0$, *(L2)* $\sum_{n\in\mathbb{N}}\lambda_{n+1} = +\infty$, *(L3)* $\lim_{n\to+\infty}(\lambda_n - \lambda_{n+1})\lambda_{n+1}^{-2} = 0$*], the sequence* $(x_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ *generated, for arbitrary* $x_0 \in \mathcal{H}$, *by (16.65) converges strongly to the uniquely existing solution of Problem (16.64).*

II. *(Nonstrictly convex case [105, 106, 149]) Assume that* $\dim(\mathcal{H}) < \infty$. *Suppose that (i)* $T\colon \mathcal{H} \to \mathcal{H}$ *is an attracting nonexpansive operator with bounded* $\mathrm{Fix}(T) \neq \varnothing$, *(ii)* $\nabla\Theta$ *is* κ*-Lipschitzian over* $T(\mathcal{H})$, *which guarantees* $\Omega \neq \varnothing$. *Then, by using*[11] $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell^2_+ \setminus \ell^1_+$, *the sequence* $(x_n)_{n\in\mathbb{N}}$ *generated by (16.65), for arbitrary* $x_0 \in \mathcal{H}$, *satisfies* $\lim_{n\to\infty} d_\Omega(x_n) = 0$, *where* $d_\Omega(x_n) := \min_{y\in\Omega}\|x_n - y\|$.

Remark 16.13

(a) **(Comparison between Fact 16.12(I) and Fact 16.6)** Fact 16.6 in Section 16.2.2 is available for generation of a weak convergent sequence to a point in $\mathrm{Fix}(T)$, where the weak limit depends on the choices of x_0 and $(\alpha_n)_{n\in\mathbb{N}}$. Fact 16.12(I) guarantees *the strong convergence* of $(x_n)_{n\in\mathbb{N}}$ to a point in $\mathrm{Fix}(T)$, where the strong limit is optimal in $\mathrm{Fix}(T)$ because it minimizes Θ able to be designed strategically for many applications. Note that, thanks to Fact 16.12(I), we present that the LAL operator plugged into the HSDM yields an iterative approximation, of a solution of Problem (16.64), whose strong convergence is guaranteed if Θ has the strongly monotone Lipschitzian gradient over $\mathcal{H}$ (see Theorem 16.19 below).

(b) **(Boundedness assumption of $\mathrm{Fix}(T)$ in Fact 16.12(II))** For readers who get worried about the boundedness assumption in Fact 16.12(II), we present some sufficient conditions, in Section 16.3.3, to guarantee the boundedness for $\mathrm{Fix}(T)$ in the context of DRS operators and LAL operators. These conditions hold automatically in the application to the hierarchical enhancement of the Lasso, in Section 16.5.2. However, the boundedness assumption in Fact 16.12(II) may not be restrictive for most practitioners by just modifying our original target (16.64) into

$$\text{minimize } \Theta(x) \text{ subject to } x \in \overline{B}(0,r) \cap \mathrm{Fix}(T) \neq \varnothing \tag{16.66}$$

with a sufficiently large closed ball $\overline{B}(0,r)$. Note that Fact 16.12(II) is applicable to (16.66) because $P_{\overline{B}(0,r)} \circ T$ is nonexpansive and satisfies $\mathrm{Fix}(P_{\overline{B}(0,r)} \circ T) = \overline{B}(0,r) \cap \mathrm{Fix}(T)$ (see [145, Proposition 1(d)]). Similar strategy will be utilized in the application to the hierarchical enhancement of the SVM in Section 16.4.2.

[11] ℓ^1_+ denotes the set of all summable nonnegative sequences. ℓ^2_+ denotes the set of all square-summable nonnegative sequences.

(c) **(Conditions for Θ)** The condition for $\Theta \in \Gamma_0(\mathcal{H})$ in (16.64), where it is required to have the Lipschitzian gradient $\nabla\Theta$, may not be restrictive as well for practitioners just by passing through the smooth regularizations, e.g. Moreau-Yosida regularization (see (16.12) and Fact 16.8 in Section 16.2.2).

Remark 16.14 (On the Hybrid Steepest Descent Method)

(a) The HSDM was established originally as a generalization of the so-called Halpern-type iteration (or anchor method) [6, 72, 90] for iteratively computing $P_{\mathrm{Fix}(T)}(x)$ for a nonexpansive operator $T\colon \mathcal{H} \to \mathcal{H}$ and $x \in \mathcal{H}$. Indeed, by choosing $\Psi(\cdot) := \frac{1}{2}\|\cdot - x\|^2$, the iteration (16.65) is reduced to the Halpern-type iteration.

(b) One can relax (L3) to $\lim_{n\to\infty} \frac{\lambda_n}{\lambda_{n+1}} = 1$ in [140]. Moreover, if T is an averaged nonexpansive operator it was shown in [83] that only (W1) and (W2) are required to guarantee the strong convergence.

(c) The HSDM can be robustified against the numerical errors produced possibly in the computation of T [146].

(d) Parallel versions of the HSDM were developed in [129]. Specifically, convex optimization over the Cartesian product of the intersections of the fixed point sets of nonexpansive operators is considered, where strong convergence theorems are established under a certain contraction assumption with respect to the weighted maximum norm.

(e) The HSDM has been extended for the variational inequality problems over the fixed point set of certain class of quasi-nonexpansive operators including subgradient projection operators [145, 149] and has been applied to signal processing problems (see, e.g., [108, 149]).

(f) The mathematical properties of the HSDM, e.g., in [142, 145] have been studied extensively in various directions by many mathematicians (see, e.g., [27, 94] for extensions in Banach spaces).

16.3 Hierarchical Convex Optimization with Proximal Splitting Operators

In this section, we present our central strategy for iterative approximation of the solution of the hierarchical convex optimization (16.13) by plugging proximal splitting operators into the HSDM. For simplicity, we focus on the DRS and the LAL operators as such proximal splitting operators.[12] Assume that Problem (16.13) has a solution, i.e., there exists at least one minimizer of Ψ over

[12] In [149, Sec. 17.5], the authors introduced briefly the central strategy of plugging the Douglas-Rachford splitting operator into the HSDM for hierarchical convex optimization. For applications of the HSDM to other proximal splitting operators, e.g., the forward-backward splitting operator [44], the primal-dual splitting operator [47, 139] for the hierarchical convex optimization of different types from (16.13), see [107, 149].

$\mathcal{S}_p$, and that (f, g, A) satisfies its qualification condition (16.40) (Note: The condition (16.40) holds automatically for many instances of (16.1), see, e.g., Section 16.4.2 [(16.148)] and Section 16.5.2 [Lemma 16.27 and (16.205)]). As explained briefly just around (16.28) in Section 16.1, for applications of the HSDM (16.65) to Problem (16.13), we need characterization of the constraint set as $\mathcal{S}_p = \Xi(\mathrm{Fix}(T))$ with a computable nonexpansive operator $T : \mathcal{H} \to \mathcal{H}$ and with a bounded linear operator $\Xi \in \mathcal{B}(\mathcal{H}, \mathcal{X})$ which ensures the Gâteaux differentiability of $\Theta := \Psi \circ \Xi \in \Gamma_0(\mathcal{H})$ with Lipschitzian gradient $\nabla\Theta$. In the following, we introduce three examples of such pair of computable nonexpansive operator $T : \mathcal{H} \to \mathcal{H}$ and $\Xi \in \mathcal{B}(\mathcal{H}, \mathcal{X})$.

16.3.1 Plugging DRS Operators into Hybrid Steepest Descent Method

We introduce a nonexpansive operator called $\mathbf{T}_{\mathrm{DRS_I}}$ of Type-I, as an instance of the DRS operator, that can characterize $\mathcal{S}_p$ (see (16.79)) and demonstrate how this nonexpansive operator can be plugged into the HSDM for (16.13).

Theorem 16.15 (HSDM with the DRS Operator in Product Space of Type-I) *Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$, and $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ in Problem (16.13) satisfy $\mathcal{S}_p \neq \varnothing$ and the qualification condition (16.40). Suppose that $\Psi \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi$ over $\mathcal{X}$ and that $\Omega := \operatorname{argmin}_{x^\star \in \mathcal{S}_p} \Psi(x^\star) \neq \varnothing$. Then the operator*

$$\mathbf{T}_{DRS_I} \colon \mathcal{X} \times \mathcal{K} \to \mathcal{X} \times \mathcal{K} \colon (x, y) \mapsto (x_T, y_T), \tag{16.67}$$

where

$$\begin{cases} p = x - A^*(I + AA^*)^{-1}(Ax - y) \\ (x_{1/2}, y_{1/2}) = (2p - x, 2Ap - y) \\ (x_T, y_T) = (2\,\mathrm{prox}_f(x_{1/2}) - x_{1/2}, 2\,\mathrm{prox}_g(y_{1/2}) - y_{1/2}), \end{cases} \tag{16.68}$$

can be plugged into the HSDM (16.65), with any $\alpha \in (0, 1)$ and any $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell^2_+ \setminus \ell^1_+$, as

$$\begin{cases} (x_{n+1/2}, y_{n+1/2}) = (1 - \alpha)(x_n, y_n) + \alpha \mathbf{T}_{DRS_I}(x_n, y_n) \\ x^\star_{n+1} = x_{n+1/2} - A^*(I + AA^*)^{-1}(Ax_{n+1/2} - y_{n+1/2}) \\ x_{n+1} = x_{n+1/2} - \lambda_{n+1}(I - A^*(I + AA^*)^{-1}A) \circ \nabla\Psi(x^\star_{n+1}) \\ y_{n+1} = y_{n+1/2} - \lambda_{n+1}((I + AA^*)^{-1}A) \circ \nabla\Psi(x^\star_{n+1}). \end{cases} \tag{16.69}$$

The algorithm (16.69) *generates, for any* $(x_0, y_0) \in \mathcal{X} \times \mathcal{K}$, *a sequence* $(x^\star_{n+1})_{n\in\mathbb{N}} \subset \mathcal{X}$ *which satisfies*

$$\lim_{n\to\infty} d_{\Omega}(x^\star_n) = 0 \tag{16.70}$$

if $\dim(\mathcal{X} \times \mathcal{K}) < \infty$ *and* $\mathrm{Fix}(\mathbf{T}_{DRS_l})$ *is bounded.*

Remark 16.16 (Idea Behind the Derivation of Theorem 16.15)

(a) The operator $\mathbf{T}_{\mathrm{DRS_I}}$ in (16.67) can be expressed as[13]

$$\mathbf{T}_{\mathrm{DRS_I}} = (2\,\mathrm{prox}_F - \mathrm{I}) \circ (2\,\mathrm{prox}_{\iota_{\mathcal{N}(\check{A})}} - \mathrm{I}) = (2\,\mathrm{prox}_F - \mathrm{I}) \circ (2P_{\mathcal{N}(\check{A})} - \mathrm{I}) \tag{16.71}$$

which is nothing but the DRS operator in the sense of Proposition 16.9 (see Section 16.2.3) specialized for

$$\text{minimize } (F + \iota_{\mathcal{N}(\check{A})})(\mathcal{X} \times \mathcal{K}), \tag{16.72}$$

where

$$F\colon\ \mathcal{X} \times \mathcal{K} \to (-\infty, \infty]\colon (x, y) \mapsto f(x) + g(y), \tag{16.73}$$

$$\check{A}\colon\ \mathcal{X} \times \mathcal{K} \to \mathcal{K}\colon (x, y) \mapsto Ax - y, \tag{16.74}$$

and $\mathcal{N}(\check{A})$ stands for the null space of $\check{A} \in \mathcal{B}(\mathcal{X} \times \mathcal{K}, \mathcal{K})$. Note that exactly in the same way as in (16.14), $\mathrm{prox}_F\colon \mathcal{X} \times \mathcal{K} \to \mathcal{X} \times \mathcal{K}\colon (x, y) \mapsto (\mathrm{prox}_f(x), \mathrm{prox}_g(y))$ can be used as a computational tool if prox_f and prox_g are available. Moreover, $\mathrm{prox}_{\iota_{\mathcal{N}(\check{A})}} = P_{\mathcal{N}(\check{A})}\colon \mathcal{X} \times \mathcal{K} \to \mathcal{N}(\check{A})\colon (x, y) \mapsto (p, Ap)$ is also available if p in (16.68) is computable, hence Problem (16.72) is minimization of the sum of two proximable functions. Obviously, Problem (16.72) is a reformulation of Problem (16.13) in a higher dimensional space in the sense of

$$\mathcal{S}_p[\text{in (16.13)}] = \mathcal{Q}_{\mathcal{X}}\left[\operatorname*{argmin}_{(x,y)\in\mathcal{X}\times\mathcal{K}} (F(x, y) + \iota_{\mathcal{N}(\check{A})}(x, y))\right], \tag{16.75}$$

where

$$\mathcal{Q}_{\mathcal{X}}\colon\ \mathcal{X} \times \mathcal{K} \to \mathcal{X}\colon (x, y) \mapsto x, \tag{16.76}$$

[13] The use of the DRS operator in a product space as in (16.71) is found explicitly or implicitly in various applications, mainly for solving (16.2) (see, e.g., [23, 41, 43, 59, 67, 68, 117]).

which is verified by

$$\begin{aligned}&\operatorname{argmin}_{x\in\mathcal{X}} f(x)+g(Ax)\\ &=\mathcal{Q}_{\mathcal{X}}\left[\operatorname{argmin}_{(x,y)\in\mathcal{X}\times\mathcal{K}} f(x)+g(y)+\iota_{\{0\}}(Ax-y)\right]\\ &=\mathcal{Q}_{\mathcal{X}}\left[\operatorname{argmin}_{(x,y)\in\mathcal{X}\times\mathcal{K}} F(x,y)+\iota_{\mathcal{N}(\check{A})}(x,y)\right].\end{aligned}$$

(b) For application of the HSDM (based on Fact 16.12(II) in Section 16.2.4), Theorem 16.15 uses the convenient expression:

$$\mathcal{S}_p[\text{in (16.13)}] \overset{\text{see below}}{=} \mathcal{Q}_{\mathcal{X}}(\operatorname{prox}_{\iota_{\mathcal{N}(\check{A})}}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})) \tag{16.77}$$

$$=\mathcal{Q}_{\mathcal{X}}(P_{\mathcal{N}(\check{A})}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}})) \tag{16.78}$$

$$=\mathcal{E}_{\mathrm{DRS_I}}(\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS_I}}))=\mathcal{E}_{\mathrm{DRS_I}}(\operatorname{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}})) \tag{16.79}$$

in terms of attracting operator $(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}}$ with $\alpha\in(0,1)$ (see (16.44)), where

$$\mathcal{E}_{\mathrm{DRS_I}}:=\mathcal{Q}_{\mathcal{X}}\circ P_{\mathcal{N}(\check{A})}\in\mathcal{B}(\mathcal{X}\times\mathcal{K},\mathcal{X}). \tag{16.80}$$

Note that the characterization (16.79) is illustrated in Figure 16.3 (see Section 16.5.1) and is utilized, in Section 16.5.2, in the context of the hierarchical enhancement of Lasso. To prove (16.77) based on Proposition 16.9(a) in Section 16.2.3, we need:

Claim 16.15 If $f\in\Gamma_0(\mathcal{X})$, $g\in\Gamma_0(\mathcal{K})$, and $A\in\mathcal{B}(\mathcal{X},\mathcal{K})$ in Problem (16.13) satisfy $\mathcal{S}_p\neq\varnothing$ and the qualification condition (16.40), we have

$$\operatorname{argmin}(F+\iota_{\mathcal{N}(\check{A})})(\mathcal{X}\times\mathcal{K})\neq\varnothing, \tag{16.81}$$

$$\operatorname{argmin}(F^*+\iota^*_{\mathcal{N}(\check{A})}\circ(-\mathrm{I}))(\mathcal{X}\times\mathcal{K})\neq\varnothing, \tag{16.82}$$

$$\min(F+\iota_{\mathcal{N}(\check{A})})(\mathcal{X}\times\mathcal{K})=-\min(F^*+\iota^*_{\mathcal{N}(\check{A})}\circ(-\mathrm{I}))(\mathcal{X}\times\mathcal{K}). \tag{16.83}$$

Note that (16.81–16.83) correspond to (16.53–16.55) in Proposition 16.9 for minimization of $F+\iota_{\mathcal{N}(\check{A})}$ and therefore Claim 16.15 is the main step in the proof of Theorem 16.15.

(c) To plug the operator $\mathbf{T}_{\mathrm{DRS_I}}:\mathcal{H}\to\mathcal{H}$, with $\mathcal{H}:=\mathcal{X}\times\mathcal{K}$, into the HSDM based on Fact 16.12(II) in Section 16.2.4, the characterization $\mathcal{S}_p=\mathcal{E}_{\mathrm{DRS_I}}(\operatorname{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}}))$ in (16.79) is utilized in the translation [exactly in the same way as in (16.28)]:

$$\Omega[\text{in Theorem 16.15}] = \Xi_{\mathrm{DRS_I}}(\Omega_{\mathrm{DRS_I}}), \tag{16.84}$$

$$\text{where } \Omega_{\mathrm{DRS_I}} := \operatorname*{argmin}_{\mathbf{z}\in \mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_I}})} \Theta_{\mathrm{DRS_I}}(\mathbf{z}) = \operatorname*{argmin}_{\mathbf{z}\in \mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}})} \Theta_{\mathrm{DRS_I}}(\mathbf{z}), \tag{16.85}$$

and $\Theta_{\mathrm{DRS_I}} = \Psi \circ \Xi_{\mathrm{DRS_I}} \in \Gamma_0(\mathcal{X}\times\mathcal{K})$.

(d) Application of the HSDM to (16.85) yields

$$\left[\begin{array}{l} \mathbf{z}_{n+1/2} = [(1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{DRS_I}}](\mathbf{z}_n), \\ \mathbf{z}_{n+1} = \mathbf{z}_{n+1/2} - \lambda_{n+1}\nabla\Theta_{\mathrm{DRS_I}}(\mathbf{z}_{n+1/2}) \\ \qquad = \mathbf{z}_{n+1/2} - \lambda_{n+1}\Xi^*_{\mathrm{DRS_I}}\nabla\Psi(\Xi_{\mathrm{DRS_I}}\mathbf{z}_{n+1/2}), \end{array}\right. \tag{16.86}$$

where $\Xi^*_{\mathrm{DRS_I}}$ is the conjugate of $\Xi_{\mathrm{DRS_I}}$ in (16.80). By letting $\mathbf{z}_n =: (x_n, y_n) \in \mathcal{X}\times\mathcal{K}$, $\mathbf{z}_{n+1/2} =: (x_{n+1/2}, y_{n+1/2}) \in \mathcal{X}\times\mathcal{K}$, and $x^\star_{n+1} := \Xi_{\mathrm{DRS_I}}\mathbf{z}_{n+1/2} \in \mathcal{X}$, as well as, by noting

$$\Xi^*_{\mathrm{DRS_I}} = P_{\mathcal{N}(\check{A})}\circ Q^*_{\mathcal{X}} \colon \mathcal{X} \to \mathcal{X}\times\mathcal{K} \colon x \mapsto ((\mathrm{I}-A^*(\mathrm{I}+AA^*)^{-1}A)x, (\mathrm{I}+AA^*)^{-1}Ax),$$

we can verify the equivalence between (16.86) and (16.69).

(e) Fact 16.12(II) in Section 16.2.4 guarantees $\lim_{n\to\infty} d_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_n) = 0$. Moreover, by noting that $\Xi_{\mathrm{DRS_I}}P_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_{n+1/2}) \in \Omega$ (see (16.84)) and $\Omega_{\mathrm{DRS_I}} \subset \mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_I}})$ (see (16.85)), (16.70) is verified as

$$\begin{aligned} d_\Omega(x^\star_{n+1}) &= d_\Omega(\Xi_{\mathrm{DRS_I}}\mathbf{z}_{n+1/2}) \\ &\le \|\Xi_{\mathrm{DRS_I}}\mathbf{z}_{n+1/2} - \Xi_{\mathrm{DRS_I}}P_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_{n+1/2})\|_{\mathcal{X}} \\ &\le \|\Xi_{\mathrm{DRS_I}}\|_{\mathrm{op}}\|\mathbf{z}_{n+1/2} - P_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_{n+1/2})\|_{\mathcal{H}} \\ &\le \|\Xi_{\mathrm{DRS_I}}\|_{\mathrm{op}} d_{\Omega_{\mathrm{DRS_I}}}(\mathbf{z}_n) \to 0 \ (n\to\infty). \end{aligned}$$

(The proof of Theorem 16.15 is given in Appendix C).

Next, we introduce another nonexpansive operator called $\mathbf{T}_{\mathrm{DRS_{II}}}$ of Type-II, as an instance of the DRS operator, that can characterize $\mathcal{S}_p$ (see (16.99)) and demonstrate how this nonexpansive operator can be plugged into the HSDM for (16.13). The operator $\mathbf{T}_{\mathrm{DRS_{II}}}$ is designed based on Example 16.7(b) in Section 16.2.2.

Theorem 16.17 (HSDM with the DRS Operator in Product Space of Type-II) *Let $\mathcal{K} = \mathbb{R}^m$. Let $f \in \Gamma_0(\mathcal{X})$, $g = \bigoplus_{i=1}^m g_i \in \Gamma_0(\mathcal{K})$, $A\colon \mathcal{X}\to\mathcal{K}\colon x \mapsto Ax = (A_1x, A_2x, \ldots, A_mx)$ with $A_i \in \mathcal{B}(\mathcal{X},\mathbb{R})\setminus\{0\}$ $(i = 1, 2, \ldots, m)$ in Problem (16.13) satisfy $\mathcal{S}_p \neq \varnothing$ and the qualification condition (16.40). Suppose that $\Psi \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi$ over $\mathcal{X}$ and that $\Omega := \operatorname*{argmin}_{x^\star\in\mathcal{S}_p} \Psi(x^\star) \neq \varnothing$. Then the operator*

$$\mathbf{T}_{DRS_{II}} \colon \mathcal{X}^{m+1} \to \mathcal{X}^{m+1} \colon (x^{(1)}, x^{(2)}, \ldots, x^{(m+1)}) \mapsto (x_T^{(1)}, x_T^{(2)}, \ldots, x_T^{(m+1)}), \tag{16.87}$$

where

$$\begin{bmatrix} \bar{x} = \frac{1}{m+1}\sum_{j=1}^{m+1} x^{(j)} \\ x_T^{(i)} = (2\bar{x} - x^{(i)}) + 2(A_i A_i^*)^{-1} A_i^* (\mathrm{prox}_{(A_i A_i^*)g_i}[A_i(2\bar{x} - x^{(i)})] - A_i(2\bar{x} - x^{(i)})) \\ \qquad (i = 1, 2, \ldots, m) \\ x_T^{(m+1)} = 2\,\mathrm{prox}_f(2\bar{x} - x^{(m+1)}) - (2\bar{x} - x^{(m+1)}), \end{bmatrix}$$

can be plugged into the HSDM (16.65)*, with any* $\alpha \in (0, 1)$ *and any* $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell_+^2 \setminus \ell_+^1$*, as*

$$\begin{bmatrix} \left(x_{n+1/2}^{(1)}, \ldots, x_{n+1/2}^{(m+1)}\right) = (1-\alpha)\left(x_n^{(1)}, \ldots, x_n^{(m+1)}\right) \\ \qquad +\alpha \mathbf{T}_{DRS_{II}}\left(x_n^{(1)}, \ldots, x_n^{(m+1)}\right) \\ x_{n+1}^{\star} = \frac{1}{m+1}\sum_{j=1}^{m+1} x_{n+1/2}^{(j)} \\ x_{n+1}^{(i)} = x_{n+1/2}^{(i)} - \frac{\lambda_{n+1}}{m+1}\nabla\Psi(x_{n+1}^{\star}) \quad (i = 1, 2, \ldots, m+1). \end{bmatrix} \tag{16.88}$$

The algorithm (16.88) *generates, for any* $\left(x_0^{(1)}, \ldots, x_0^{(m+1)}\right) \in \mathcal{X}^{m+1}$*, a sequence* $(x_{n+1}^{\star})_{n\in\mathbb{N}} \subset \mathcal{X}$ *which satisfies*

$$\lim_{n\to\infty} d_{\Omega}(x_n^{\star}) = 0 \tag{16.89}$$

if $\dim(\mathcal{X}) < \infty$ *and* $\mathrm{Fix}(\mathbf{T}_{DRS_{II}})$ *is bounded.*

Remark 16.18 (Idea Behind the Derivation of Theorem 16.17)

(a) The operator $\mathbf{T}_{\mathrm{DRS_{II}}}$ in (16.87) can be expressed as

$$\mathbf{T}_{\mathrm{DRS_{II}}} = (2\,\mathrm{prox}_H - \mathrm{I}) \circ (2\,\mathrm{prox}_{\iota_D} - \mathrm{I}) = (2\,\mathrm{prox}_H - \mathrm{I}) \circ (2P_D - \mathrm{I}) \tag{16.90}$$

which is the DRS operator in the sense of Proposition 16.9 (see Section 16.2.3) specialized for

$$\text{minimize } (H + \iota_D)(\mathcal{X}^{m+1}), \tag{16.91}$$

where

$$H\colon \mathcal{X}^{m+1} \to (-\infty, \infty]\colon (x^{(1)}, \ldots, x^{(m+1)}) \mapsto \sum_{i=1}^{m} g_i(A_i x^{(i)}) + f(x^{(m+1)}), \tag{16.92}$$

$$D := \{(x^{(1)}, \ldots, x^{(m+1)}) \in \mathcal{X}^{m+1} \mid x^{(i)} = x^{(j)}\ (i, j = 1, 2, \ldots, m+1)\}. \tag{16.93}$$

Note that exactly in the same way as in (16.14),

$$\begin{aligned}&\mathrm{prox}_H(x^{(1)}, x^{(2)}, \ldots, x^{(m+1)})\\ &= (\mathrm{prox}_{g_1\circ A_1}(x^{(1)}), \ldots, \mathrm{prox}_{g_m\circ A_m}(x^{(m)}), \mathrm{prox}_f(x^{(m+1)}))\end{aligned}$$

can be used with (16.48), in Example 16.7(b) (see Section 16.2.2), as a computational tool if prox_f and $\mathrm{prox}_{A_i A_i^* g}$ $(i = 1, 2, \ldots, m)$ are available. Moreover, $\mathrm{prox}_{\iota_D} = P_D\colon \mathcal{X}^{m+1} \to \mathcal{X}^{m+1}\colon (x^{(1)}, x^{(2)}, \ldots, x^{(m+1)}) \mapsto (\bar{x}, \ldots, \bar{x})$ with $\bar{x} := \frac{1}{m+1}\sum_{i=1}^{m+1} x^{(i)}$ is also available. Hence Problem (16.91) is minimization of the sum of two proximable functions (Note: Thanks to $A_i A_i^* \in \mathbb{R}_{++} := \{r \in \mathbb{R} \mid r > 0\}$, the computation of $\mathbf{T}_{\mathrm{DRS_{II}}}$ in (16.90) does not require any matrix inversion). Obviously, Problem (16.91) is a reformulation of Problem (16.13) in a higher dimensional space in the sense of

$$\mathcal{S}_p[\text{in (16.13)}] = \mathcal{Q}_{\mathcal{X}^{(1)}}\left[\operatorname*{argmin}_{(x^{(1)},\ldots,x^{(m+1)})\in\mathcal{X}^{m+1}} (H + \iota_D)(x^{(1)}, \ldots, x^{(m+1)}))\right], \tag{16.94}$$

where

$$\mathcal{Q}_{\mathcal{X}^{(1)}}\colon \mathcal{X}^{m+1} \to \mathcal{X}\colon (x^{(1)}, \ldots, x^{(m+1)}) \mapsto x^{(1)}, \tag{16.95}$$

which is verified by

$$\begin{aligned}&\operatorname{argmin}_{x\in\mathcal{X}}\, g(Ax) + f(x)\\ &= \operatorname{argmin}_{x\in\mathcal{X}} \sum_{i=1}^{m} g_i(A_i x) + f(x)\\ &= \mathcal{Q}_{\mathcal{X}^{(1)}}\left[\operatorname{argmin}_{(x^{(1)},\ldots,x^{(m+1)})\in\mathcal{X}^{m+1}} (H + \iota_D)(x^{(1)}, \ldots, x^{(m+1)})\right].\end{aligned} \tag{16.96}$$

(b) For application of the HSDM (based on Fact 16.12(II) in Section 16.2.4), Theorem 16.17 uses the convenient expression:

$$\begin{aligned}&\mathcal{S}_p[\text{in (16.13)}]\\ &\overset{\text{see below}}{=} \mathcal{Q}_{\mathcal{X}^{(1)}}(\mathrm{prox}_{\iota_D}(\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})) &(16.97)\\ &= \mathcal{Q}_{\mathcal{X}^{(1)}}(P_D(\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})) &(16.98)\\ &= \Xi_{\mathrm{DRS_{II}}}(\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})) = \Xi_{\mathrm{DRS_{II}}}(\mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{DRS_{II}}})) &(16.99)\end{aligned}$$

in terms of attracting operator $(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_{II}}}$ with $\alpha \in (0, 1)$ (see (16.44)), where

$$\Xi_{\mathrm{DRS_{II}}} := \Omega_{\mathcal{X}^{(1)}} \circ P_D \in \mathcal{B}(\mathcal{X}^{m+1}, \mathcal{X}). \tag{16.100}$$

To prove (16.97) based on Proposition 16.9(a) in Section 16.2.3, we need:

Claim 16.17 If $\dim(\mathcal{K}) < \infty$, $f \in \Gamma_0(\mathcal{X})$, $g = \bigoplus_{i=1}^m g_i \in \Gamma_0(\mathcal{K})$, $A\colon \mathcal{X} \to \mathcal{K}\colon x \mapsto Ax = (A_1x, A_2x, \ldots, A_mx)$ with $A_i \in \mathcal{B}(\mathcal{X}, \mathbb{R}) \setminus \{0\}$ $(i = 1, 2, \ldots, m)$ in Problem (16.13) satisfy $\mathcal{S}_p \neq \varnothing$ and the qualification condition (16.40), we have

$$\operatorname{argmin}(H + \iota_D)(\mathcal{X}^{m+1}) \neq \varnothing, \tag{16.101}$$

$$\operatorname{argmin}(H^* + \iota_D^* \circ (-\mathrm{I}))(\mathcal{X}^{m+1}) \neq \varnothing, \tag{16.102}$$

$$\min(H + \iota_D)(\mathcal{X}^{m+1}) = -\min(H^* + \iota_D^* \circ (-\mathrm{I}))(\mathcal{X}^{m+1}). \tag{16.103}$$

Note that (16.101–16.103) correspond to (16.53–16.55) in Proposition 16.9 for minimization of $H + \iota_D$ and therefore Claim 16.17 is the main step in the proof of Theorem 16.17.

(c) To plug the operator $\mathbf{T}_{\mathrm{DRS_{II}}}\colon \mathcal{H} \to \mathcal{H}$, with $\mathcal{H} := \mathcal{X}^{m+1}$, into the HSDM based on Fact 16.12(II) in Section 16.2.4, the characterization $\mathcal{S}_p = \Xi_{\mathrm{DRS_{II}}}(\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_{II}}}))$ in (16.99) is utilized in the translation [exactly in the same way as in (16.28)]:

$$\begin{aligned} &\Omega[\text{in Theorem 16.17}] = \Xi_{\mathrm{DRS_{II}}}(\Omega_{\mathrm{DRS_{II}}}),\\ &\text{where } \Omega_{\mathrm{DRS_{II}}} := \operatorname*{argmin}_{\mathbf{X}\in\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})} \Theta_{\mathrm{DRS_{II}}}(\mathbf{X}) = \operatorname*{argmin}_{\mathbf{X}\in\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{DRS_{II}}})} \Theta_{\mathrm{DRS_{II}}}(\mathbf{X}), \end{aligned} \tag{16.104}$$

and $\Theta_{\mathrm{DRS_{II}}} = \Psi \circ \Xi_{\mathrm{DRS_{II}}} \in \Gamma_0(\mathcal{X}^{m+1})$.

(d) Application of the HSDM to (16.104) yields

$$\left[\begin{aligned} \mathbf{X}_{n+1/2} &= [(1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{DRS_{II}}}](\mathbf{X}_n),\\ \mathbf{X}_{n+1} &= \mathbf{X}_{n+1/2} - \lambda_{n+1}\nabla\Theta_{\mathrm{DRS_{II}}}(\mathbf{X}_{n+1/2})\\ &= \mathbf{X}_{n+1/2} - \lambda_{n+1}\Xi^*_{\mathrm{DRS_{II}}}\nabla\Psi(\Xi_{\mathrm{DRS_{II}}}\mathbf{X}_{n+1/2}), \end{aligned}\right. \tag{16.105}$$

where $\Xi^*_{\mathrm{DRS_{II}}}$ is the conjugate of $\Xi_{\mathrm{DRS_{II}}}$ in (16.100). By letting $\mathbf{X}_n =: (x_n^{(1)}, \ldots, x_n^{(m+1)}) \in \mathcal{X}^{m+1}$, $\mathbf{X}_{n+1/2} =: (x_{n+1/2}^{(1)}, \ldots, x_{n+1/2}^{(m+1)}) \in \mathcal{X}^{m+1}$, and $x_{n+1}^\star := \Xi_{\mathrm{DRS_{II}}}\mathbf{X}_{n+1/2} \in \mathcal{X}$, as well as, by noting

$$\Xi^*_{\mathrm{DRS_{II}}} = P_D \circ \Omega^*_{\mathcal{X}^{(1)}}\colon \mathcal{X} \to \mathcal{X}^{m+1}\colon x \mapsto \frac{1}{m+1}(x, x, \ldots, x),$$

we can verify the equivalence between (16.105) and (16.88).

(e) In the same way as in Remark 16.16(e), Fact 16.12(II) in Section 16.2.4 guarantees $\lim_{n\to\infty} d_{\Omega_{\mathrm{DRS}_{\mathrm{II}}}}(\mathbf{X}_n) = 0$ and (16.89).

(The proof of Theorem 16.17 is given in Appendix D).

16.3.2 Plugging LAL Operator into Hybrid Steepest Descent Method

We introduce a nonexpansive operator called $\mathbf{T}_{\mathrm{LAL}}$, as an instance of the LAL operator, that can characterize $\mathcal{S}_p$ (see (16.114)) and demonstrate how this nonexpansive operator can be plugged into the HSDM for (16.13). In particular, if $\nabla\Psi$ is strongly monotone over $\mathcal{X}$, $\mathbf{T}_{\mathrm{LAL}}$ can be plugged into the HSDM based on Fact 16.12(I) in Section 16.2.4, which results in a strongly convergent iterative algorithm for (16.13) (see Theorem 16.19). Of course, $\mathbf{T}_{\mathrm{LAL}}$ can also be plugged into the HSDM based on Fact 16.12(II) (see Theorem 16.21).

Theorem 16.19 (Strong Convergence Achieved by HSDM with LAL Operator) *Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$ and $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ in Problem (16.13) satisfy not only $\mathcal{S}_p \neq \varnothing$ and the qualification condition (16.40) but also $\|\check{A}\|_{op} \leq \frac{1}{\mathfrak{u}}$ $(\exists \mathfrak{u} > 0)$ with $\check{A}$ in (16.74). Suppose also that $\Psi \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable with Lipschitzian as well as strongly monotone gradient $\nabla\Psi$ over $\mathcal{X}$. Then the operator*

$$\mathbf{T}_{LAL}\colon \mathcal{X}\times\mathcal{K}\times\mathcal{K} \to \mathcal{X}\times\mathcal{K}\times\mathcal{K} \tag{16.106}$$

$$\colon \begin{pmatrix} x \\ y \\ v \end{pmatrix} \mapsto \begin{pmatrix} x_T \\ y_T \\ v_T \end{pmatrix} = \begin{pmatrix} \mathrm{prox}_f(x - \mathfrak{u}^2(A^*Ax - A^*y) + \mathfrak{u}A^*v) \\ \mathrm{prox}_g(y - \mathfrak{u}^2(-Ax + y) - \mathfrak{u}v) \\ v - \mathfrak{u}(Ax_T - y_T) \end{pmatrix}$$

can be plugged into the HSDM (16.65), with any $\alpha \in (0, 1]$ and any $\eta_{xy}, \eta_v > 0$, as

$$\begin{cases} (x_{n+1/2}, y_{n+1/2}, v_{n+1/2}) = (1-\alpha)(x_n, y_n, v_n) + \alpha\mathbf{T}_{LAL}(x_n, y_n, v_n) \\ x_{n+1} = x_{n+1/2} - \lambda_{n+1}(\nabla\Psi(x_{n+1/2}) + \eta_{xy}A^*(Ax_{n+1/2} - y_{n+1/2})) \\ y_{n+1} = y_{n+1/2} + \lambda_{n+1}\eta_{xy}(Ax_{n+1/2} - y_{n+1/2}) \\ v_{n+1} = v_{n+1/2} - \lambda_{n+1}\eta_v v_{n+1/2}. \end{cases} \tag{16.107}$$

The algorithm (16.107) generates, for any $(x_0, y_0, v_0) \in \mathcal{X}\times\mathcal{K}\times\mathcal{K}$, a sequence $(x_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ which converges strongly to the uniquely existing solution of Problem (16.13) if $(\lambda_{n+1})_{n\in\mathbb{N}} \subset [0, \infty)$ satisfies conditions (W1–W3) [or $(\lambda_{n+1})_{n\in\mathbb{N}} \subset (0, \infty)$ satisfies (L1–L3)] in Fact 16.12(I) in Section 16.2.4.

Remark 16.20 (Idea Behind the Derivation of Theorem 16.19)

(a) The operator $\mathbf{T}_{\mathrm{LAL}}$ in (16.106) can be expressed as

$$(\mathbf{z}, \nu) \mapsto (\mathbf{z}_T, \nu_T) \text{ with } \begin{cases} \mathbf{z}_T = \mathrm{prox}_F(\mathbf{z} - (\mathrm{u}\check{A})^*(\mathrm{u}\check{A})\mathbf{z} + (\mathrm{u}\check{A})^*\nu) \\ \nu_T = \nu - \mathrm{u}\check{A}\mathbf{z}_T \end{cases} \tag{16.108}$$

by introducing $\mathbf{z} := (x, y)$ and $\mathbf{z}_T := (x_T, y_T)$, which is the LAL operator of Proposition 16.10 (see Section 16.2.3) specialized for

$$\text{minimize } (F + \iota_{\{0\}} \circ (\mathrm{u}\check{A}))(\mathcal{X} \times \mathcal{K}), \tag{16.109}$$

where F and $\check{A}$ are defined, respectively, in (16.73) and in (16.74). Note that exactly in the same way as in (16.14), $\mathrm{prox}_F \colon \mathcal{X} \times \mathcal{K} \to \mathcal{X} \times \mathcal{K} \colon (x, y) \mapsto (\mathrm{prox}_f(x), \mathrm{prox}_g(y))$ can be used as a computational tool if prox_f and prox_g are available. Obviously, Problem (16.109) is a reformulation of Problem (16.13) in a higher dimensional space in the sense of

$$\mathcal{S}_p[\text{in (16.13)}] = \mathcal{Q}_{\mathcal{X}} \left[\operatorname*{argmin}_{(x,y)\in\mathcal{X}\times\mathcal{K}} F(x, y) + \iota_{\{0\}}(\mathrm{u}\check{A}(x, y)) \right], \tag{16.110}$$

where $\mathcal{Q}_{\mathcal{X}}$ is defined as in (16.76), which is verified by

$$\begin{aligned} \mathcal{S}_p &= \operatorname{argmin}_{x\in\mathcal{X}} f(x) + g(Ax) \\ &= \mathcal{Q}_{\mathcal{X}} \left[\operatorname{argmin}_{(x,y)\in\mathcal{X}\times\mathcal{K}} f(x) + g(y) + \iota_{\{0\}}(Ax - y) \right] \\ &= \mathcal{Q}_{\mathcal{X}} \left[\operatorname{argmin}_{(x,y)\in\mathcal{X}\times\mathcal{K}} F(x, y) + \iota_{\{0\}}(\mathrm{u}\check{A}(x, y)) \right]. \end{aligned} \tag{16.111}$$

(b) For application of the HSDM (based on Fact 16.12(I) in Section 16.2.4), Theorem 16.19 uses the convenient expression:

$$\mathcal{S}_p[\text{in (16.13)}]$$

$$= \mathcal{Q}_{\mathcal{X}} \circ \mathcal{Q}_{\mathcal{X}\times\mathcal{K}} \left[\operatorname{argmin}(F + \iota_{\{0\}} \circ (\mathrm{u}\check{A}))(\mathcal{X} \times \mathcal{K}) \times \operatorname{argmin}(F^* \circ (\mathrm{u}\check{A})^*)(\mathcal{K}) \right] \tag{16.112}$$

$$\overset{\text{see below}}{=} \mathcal{Q}_{\mathcal{X}} \circ \mathcal{Q}_{\mathcal{X}\times\mathcal{K}}(\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})) \tag{16.113}$$

$$= \Xi_{\mathrm{LAL}}(\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})) = \Xi_{\mathrm{LAL}}(\mathrm{Fix}((1 - \alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}})) \tag{16.114}$$

in terms of nonexpansive operator $(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}}$ with $\alpha \in (0, 1]$ (Note: The nonexpansiveness of $\mathbf{T}_{\mathrm{LAL}}$ is ensured by Proposition 16.10(b) in Section 16.2.3 with $\|\mathrm{u}\check{A}\|_{\mathrm{op}} \leq 1$)

$$\mathfrak{Q}_{\mathcal{X}\times\mathcal{K}}\colon \mathcal{X}\times\mathcal{K}\times\mathcal{K}\to\mathcal{X}\times\mathcal{K}\colon (x,y,v)\mapsto(x,y) \tag{16.115}$$

$$\Xi_{\mathrm{LAL}} := \mathfrak{Q}_{\mathcal{X}}\circ\mathfrak{Q}_{\mathcal{X}\times\mathcal{K}} \in \mathcal{B}(\mathcal{X}\times\mathcal{K}\times\mathcal{K},\mathcal{X}), \tag{16.116}$$

where $\mathfrak{Q}_{\mathcal{X}}$ is defined as in (16.76). To prove (16.113) based on Proposition 16.10(a) in Section 16.2.3, we need:

Claim 16.19 If $f\in\Gamma_0(\mathcal{X})$, $g\in\Gamma_0(\mathcal{K})$ and $A\in\mathcal{B}(\mathcal{X},\mathcal{K})$ in Problem (16.13) satisfy not only $\mathcal{S}_p\neq\varnothing$ and the qualification condition (16.40) but also $\|\check{A}\|_{\mathrm{op}}\leq\frac{1}{\mathfrak{u}}$ $(\exists\mathfrak{u}>0)$ with $\check{A}$ in (16.74), we have

$$\operatorname{argmin}(F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A}))(\mathcal{X}\times\mathcal{K})\neq\varnothing, \tag{16.117}$$

$$\operatorname{argmin}(F^*\circ(\mathfrak{u}\check{A})^*)(\mathcal{K})\neq\varnothing, \tag{16.118}$$

$$\min(F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A}))(\mathcal{X}\times\mathcal{K}) = -\min(F^*\circ(\mathfrak{u}\check{A})^*)(\mathcal{K}). \tag{16.119}$$

Note that (16.117–16.119) correspond to (16.58–16.60) in Proposition 16.10 for minimization of $F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A})$ and therefore Claim 16.19 is the main step in the proof of Theorem 16.19. In Claim 16.19, we also remark that $(\mathfrak{u}\check{A})^*$ in (16.118) is the conjugate of $\mathfrak{u}\check{A}$ and given by

$$(\mathfrak{u}\check{A})^*\colon\mathcal{K}\to\mathcal{X}\times\mathcal{K}\colon v\mapsto(\mathfrak{u}A^*v,-\mathfrak{u}v). \tag{16.120}$$

(c) To plug the operator $\mathbf{T}_{\mathrm{LAL}}\colon\mathcal{H}\to\mathcal{H}$, with $\mathcal{H}:=\mathcal{X}\times\mathcal{K}\times\mathcal{K}$, into the HSDM based on Fact 16.12(I) in Section 16.2.4, the characterization $\mathcal{S}_p=\Xi_{\mathrm{LAL}}(\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}}))$ in (16.114) is utilized in the translation:

$$\Omega[\text{in Theorem 16.19}] = \Xi_{\mathrm{LAL}}(\Omega^{\mathrm{reg}}_{\mathrm{LAL}}), \tag{16.121}$$

$$\text{where } \Omega^{\mathrm{reg}}_{\mathrm{LAL}} := \operatorname*{argmin}_{\mathbf{w}\in\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})}\Theta^{\mathrm{reg}}_{\mathrm{LAL}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in\mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}})}\Theta^{\mathrm{reg}}_{\mathrm{LAL}}(\mathbf{w}), \tag{16.122}$$

$$\Theta^{\mathrm{reg}}_{\mathrm{LAL}}\colon\mathcal{X}\times\mathcal{K}\times\mathcal{K}\to\mathbb{R}$$
$$\colon\mathbf{w}_\star\mapsto\Psi(\Xi_{\mathrm{LAL}}\mathbf{w}_\star)+\frac{\eta_{xy}}{2}\|\check{A}\circ\mathfrak{Q}_{\mathcal{X}\times\mathcal{K}}\mathbf{w}_\star\|^2_{\mathcal{K}}+\frac{\eta_v}{2}\|\mathfrak{Q}_{\mathcal{K}}\mathbf{w}_\star\|^2_{\mathcal{K}},$$

for $\eta_{xy},\eta_v>0$ with $\mathfrak{Q}_{\mathcal{K}}\colon\mathcal{X}\times\mathcal{K}\times\mathcal{K}\colon(x,y,v)\mapsto v$. Note that, since $\nabla\Psi$ is strongly monotone over $\mathcal{X}$, the gradient $\nabla\Theta^{\mathrm{reg}}_{\mathrm{LAL}}$ is strongly monotone over $\mathcal{X}\times\mathcal{K}\times\mathcal{K}$ (for the proof, see [150, Theorem 2(d)]).

(d) Application of the HSDM to (16.122) yields

$$\begin{bmatrix}\mathbf{w}_{n+1/2}=[(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}}](\mathbf{w}_n)\\ \mathbf{w}_{n+1}=\mathbf{w}_{n+1/2}-\lambda_{n+1}\nabla\Theta^{\mathrm{reg}}_{\mathrm{LAL}}(\mathbf{w}_{n+1/2}).\end{bmatrix} \tag{16.123}$$

By letting $\mathbf{w}_n := (x_n, y_n, v_n) \in \mathcal{X} \times \mathcal{K} \times \mathcal{K}$ and $\mathbf{w}_{n+1/2} := (x_{n+1/2}, y_{n+1/2}, v_{n+1/2}) \in \mathcal{X} \times \mathcal{K} \times \mathcal{K}$, as well as, by noting

$$\begin{aligned} \Xi^*_{\mathrm{LAL}} = \mathfrak{Q}^*_{\mathcal{X}\times\mathcal{K}} \circ \mathfrak{Q}^*_{\mathcal{X}} \colon \mathcal{X} &\to \mathcal{X} \times \mathcal{K} \times \mathcal{K} \colon x \mapsto (x, 0, 0), \\ (\check{A} \circ \mathfrak{Q}_{\mathcal{X}\times\mathcal{K}})^* \colon \mathcal{K} &\to \mathcal{X} \times \mathcal{K} \times \mathcal{K} \colon y \mapsto (A^* y, -y, 0), \\ \mathfrak{Q}^*_{\mathcal{K}} \colon \mathcal{K} &\to \mathcal{X} \times \mathcal{K} \times \mathcal{K} \colon v \mapsto (0, 0, v), \end{aligned} \tag{16.124}$$

we can verify the equivalence between (16.123) and (16.107).

(e) Fact 16.12(I) in Section 16.2.4 guarantees that $(\mathbf{w}_n)_{n\in\mathbb{N}}$ converges strongly to a point in $\Omega^{\mathrm{reg}}_{\mathrm{LAL}}$. Hence, $(\Xi_{\mathrm{LAL}}\mathbf{w}_n (= x_n))_{n\in\mathbb{N}}$ also converges strongly to a point in Ω (see (16.121)).

(The proof of Theorem 16.19 is given in Appendix E).

Theorem 16.21 (HSDM with the LAL Operator Based on Fact 16.12(II)) *Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$ and $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ in Problem (16.13) satisfy not only $\mathcal{S}_p \neq \varnothing$ and the qualification condition (16.40) but also $\|\check{A}\|_{op} \leq \frac{1}{\mathfrak{u}}$ $(\exists \mathfrak{u} > 0)$ with $\check{A}$ in (16.74). Suppose also that $\Psi \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable with Lipschitzian gradient $\nabla\Psi$ over $\mathcal{X}$ and that $\Omega := \operatorname{argmin}_{x^\star \in \mathcal{S}_p} \Psi(x^\star) \neq \varnothing$. Then the operator $\mathbf{T}_{LAL}$ in (16.106) can be plugged into HSDM (16.65), with any $\alpha \in (0, 1)$ and any $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell^2_+ \setminus \ell^1_+$, as*

$$\begin{cases} (x_{n+1/2}, y_{n+1}, v_{n+1}) = (1-\alpha)(x^\star_n, y_n, v_n) + \alpha \mathbf{T}_{LAL}(x^\star_n, y_n, v_n) \\ x^\star_{n+1} = x_{n+1/2} - \lambda_{n+1} \nabla\Psi(x_{n+1/2}). \end{cases} \tag{16.125}$$

The algorithm (16.125) generates, for any $(x^\star_0, y_0, v_0) \in \mathcal{X} \times \mathcal{K} \times \mathcal{K}$, a sequence $(x^\star_n)_{n\in\mathbb{N}} \subset \mathcal{X}$ which satisfies

$$\lim_{n\to\infty} d_\Omega(x^\star_n) = 0 \tag{16.126}$$

if $\dim(\mathcal{X} \times \mathcal{K} \times \mathcal{K}) < \infty$ and $\mathrm{Fix}(\mathbf{T}_{LAL})$ is bounded.

Remark 16.22 (Idea Behind the Derivation of Theorem 16.21)

(a) Following Remark 16.20(a)(b), we obtain the characterization

$$\mathcal{S}_p[\text{in (16.13)}] = \Xi_{\mathrm{LAL}}(\mathrm{Fix}((1-\alpha)\mathrm{I} + \alpha \mathbf{T}_{\mathrm{LAL}})),$$

in (16.114) (see also Ξ_{LAL} in (16.116)), with the attracting operator $(1-\alpha)\mathrm{I} + \alpha\mathbf{T}_{\mathrm{LAL}}$ for $\alpha \in (0, 1)$ (see (16.44)). This characterization is utilized, to plug $\mathbf{T}_{\mathrm{LAL}} \colon \mathcal{H} \to \mathcal{H}$ $(\mathcal{H} := \mathcal{X} \times \mathcal{K} \times \mathcal{K})$ into the HSDM based on Fact 16.12(II) in Section 16.2.4, in the translation [see also (16.28)]:

$$\Omega[\text{in Theorem 16.21}] = \Xi_{\mathrm{LAL}}(\Omega_{\mathrm{LAL}}),$$

$$\text{where } \Omega_{\mathrm{LAL}} := \operatorname*{argmin}_{\mathbf{w}\in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})} \Theta_{\mathrm{LAL}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}\in \mathrm{Fix}((1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}})} \Theta_{\mathrm{LAL}}(\mathbf{w}) \tag{16.127}$$

and $\Theta_{\mathrm{LAL}} := \Psi \circ \Xi_{\mathrm{LAL}} \in \Gamma_0(\mathcal{X}\times\mathcal{K}\times\mathcal{K})$.

(b) Application of the HSDM to (16.127) yields

$$\begin{cases} \mathbf{w}_{n+1/2} = [(1-\alpha)\mathrm{I}+\alpha\mathbf{T}_{\mathrm{LAL}}](\mathbf{w}_n), \\ \mathbf{w}_{n+1} = \mathbf{w}_{n+1/2} - \lambda_{n+1}\nabla\Theta_{\mathrm{LAL}}(\mathbf{w}_{n+1/2}) \\ \qquad = \mathbf{w}_{n+1/2} - \lambda_{n+1}\Xi^*_{\mathrm{LAL}}\nabla\Psi(\Xi_{\mathrm{LAL}}\mathbf{w}_{n+1/2}), \end{cases} \tag{16.128}$$

where Ξ^*_{LAL} is the conjugate of Ξ_{LAL}. By letting $\mathbf{w}_n =: (x^\star_n, y_n, v_n) \in \mathcal{X}\times\mathcal{K}\times\mathcal{K}$ and $\mathbf{w}_{n+1/2} =: (x_{n+1/2}, y_{n+1/2}, v_{n+1/2}) \in \mathcal{X}\times\mathcal{K}\times\mathcal{K}$, as well as, by noting (16.124), we can verify the equivalence between (16.128) and (16.125).

(c) In the same way as in Remark 16.16(e), Fact 16.12(II) in Section 16.2.4 guarantees $\lim_{n\to\infty} d_{\Omega_{\mathrm{LAL}}}(\mathbf{w}_n) = 0$ and (16.126).

(The proof of Theorem 16.21 is omitted, see Remark 16.22).

16.3.3 *Conditions for Boundedness of Fixed Point Sets of DRS and LAL Operators*

In Theorems 16.15, 16.17, and 16.21, the boundednesses of $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$, $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_{II}}})$, and $\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})$ are required for the algorithms (16.69), (16.88), and (16.125) to produce $(x^\star_{n+1})_{n\in\mathbb{N}}$ satisfying $\lim_{n\to\infty} d_\Omega(x^\star_n) = 0$. Theorem 16.23 below presents sufficient conditions for the boundednesses of these fixed point sets. Corollary 16.24 below presents a stronger condition which will be used in Section 16.5.2 to guarantee the boundedness of $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS_I}})$ in the context of the hierarchical enhancement of Lasso.

Theorem 16.23 *Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$ and $A \in \mathcal{B}(\mathcal{X},\mathcal{K})$ in Problem (16.13) satisfy $\mathcal{S}_p \neq \varnothing$ and the qualification condition (16.40). Let $(A^*)^{-1}\colon \mathcal{X} \to 2^{\mathcal{K}} : x \mapsto \{y \in \mathcal{K} \mid x = A^*y\}$. Then we have*

(a) *$\mathcal{S}_p$, $\partial f(\mathcal{S}_p)$, and $\bigcup_{x\in\mathcal{S}_p}\left([-(A^*)^{-1}(\partial f(x))]\cap\partial g(Ax)\right) \subset \mathcal{K}$ are bounded*

$$\Rightarrow \mathrm{Fix}\left(\mathbf{T}_{DRS_I}\right) \subset \mathcal{X}\times\mathcal{K} \text{ in Theorem 16.15 is bounded.}$$

(b) *If $\check{A}$ in (16.74) satisfies $\|\check{A}\|_{op} \leq \frac{1}{\mathfrak{u}}$ $(\exists \mathfrak{u} > 0)$, then $\mathcal{S}_p$ and $\bigcup\limits_{x\in\mathcal{S}_p}\left([-(A^*)^{-1}(\partial f(x))]\cap\partial g(Ax)\right) \subset \mathcal{K}$ are bounded*

$$\Rightarrow \mathrm{Fix}\left(\mathbf{T}_{LAL}\right) \subset \mathcal{X}\times\mathcal{K}\times\mathcal{K} \text{ in Theorem 16.21 is bounded.}$$

(c) *If, in particular,* $\mathcal{K} = \mathbb{R}^m$, $g = \bigoplus_{i=1}^m g_i \in \Gamma_0(\mathbb{R}^m)$, $A : \mathcal{X} \to \mathbb{R}^m : x \mapsto Ax = (A_1x, A_2x, \ldots, A_mx)$ *with* $A_i \in \mathcal{B}(\mathcal{X}, \mathbb{R}) \setminus \{0\}$ $(i = 1, 2, \ldots, m)$ *in Problem* (16.13)*, then*
$\mathcal{S}_p$ *and* $\bigcup_{x\in\mathcal{S}_p} \left(\left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_jx)\right] \times \left[\partial f(x) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_ix)\right)\right]\right)$ *are bounded*

$$\Rightarrow \operatorname{Fix}\left(\mathbf{T}_{DRS_{II}}\right) \subset \mathcal{X}^{m+1} \text{ in Theorem 16.17 is bounded.}$$

(The proof of Theorem 16.23 is given in Appendix F.)
The following simple relations

$$\begin{aligned}
&\bigcup_{x\in\mathcal{S}_p} \left([-(A^*)^{-1}(\partial f(x))] \cap \partial g(Ax)\right) \text{ [in Theorem 16.23(a)(b)]} \\
&\quad \subset \left[(-(A^*)^{-1}(\partial f(\mathcal{S}_p))) \cap \partial g(\mathcal{K})\right] \text{ and} \\
&\bigcup_{x\in\mathcal{S}_p} \left(\left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_jx)\right] \times \left[\partial f(x) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_ix)\right)\right]\right) \\
&\qquad \text{[in Theorem 16.23(c)]} \\
&\quad \subset \left[\bigtimes_{j=1}^m A_j^* \partial g_j(\mathbb{R})\right] \times \left[-\sum_{i=1}^m A_i^* \partial g_i(\mathbb{R})\right]
\end{aligned}$$

lead to the corollary below.

Corollary 16.24 *Let* $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$, $A \in \mathcal{B}(\mathcal{X}, \mathcal{K})$ *in Problem* (16.13)*, and* $\check{A}$ *in* (16.74) *satisfy* $\mathcal{S}_p \neq \varnothing$*, the qualification condition* (16.40)*, and* $\|\check{A}\|_{op} \leq \frac{1}{\mathfrak{u}}$ $(\exists \mathfrak{u} > 0)$*. Then we have*
(a) $\mathcal{S}_p$, $\partial f(\mathcal{S}_p)$ *and* $(-(A^*)^{-1}(\partial f(\mathcal{S}_p))) \cap \partial g(\mathcal{K})$ *are bounded* (16.129)

$$\Rightarrow \begin{bmatrix} \operatorname{Fix}\left(\mathbf{T}_{DRS_I}\right) \subset \mathcal{X} \times \mathcal{K} \text{ in Theorem 16.15 is bounded;} \\ \operatorname{Fix}\left(\mathbf{T}_{LAL}\right) \subset \mathcal{X} \times \mathcal{K} \times \mathcal{K} \text{ in Theorem 16.21 is bounded.} \end{bmatrix}$$

(b) If, in particular, $\mathcal{K} = \mathbb{R}^m$, $g = \bigoplus_{i=1}^m g_i \in \Gamma_0(\mathbb{R}^m)$, $A : \mathcal{X} \to \mathbb{R}^m : x \mapsto Ax = (A_1x, A_2x, \ldots, A_mx)$ *with* $A_i \in \mathcal{B}(\mathcal{X}, \mathbb{R}) \setminus \{0\}$ $(i = 1, 2, \ldots, m)$ *in Problem* (16.13)*, then*

$$\mathcal{S}_p \text{ and } \partial g_i(\mathbb{R}) \ (i = 1, 2, \ldots, m) \text{ are bounded}$$
$$\Rightarrow \operatorname{Fix}\left(\mathbf{T}_{DRS_{II}}\right) \subset \mathcal{X}^{m+1} \text{ in Theorem 16.17 is bounded.}$$

16.4 Application to Hierarchical Enhancement of Support Vector Machine

16.4.1 Support Vector Machine

Consider a supervised learning problem for estimating a binary function

$$\mathfrak{L} : \mathbb{R}^p \to \{-1, 1\} \tag{16.130}$$

with a given training dataset $\mathcal{D} := \{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, 1\} \mid i = 1, 2, \ldots, N\}$, where y_i is a possibly corrupted version of the label $\mathfrak{L}(\mathbf{x}_i)$ of the point $\mathbf{x}_i$. The *support vector machine* (SVM) has been recognized as one of the most successful supervised machine learning algorithms for such a learning problem. For simplicity, we focus on the linear SVM because the nonlinear SVM exploiting the so-called *Kernel trick* can be viewed as an instance of the linear classifiers in the *Reproducing Kernel Hilbert Spaces* (RKHS).

The dataset $\mathcal{D}$ is said to be *linearly separable* if there exists $(\mathbf{w}, b) \in (\mathbb{R}^p \setminus \{\mathbf{0}\}) \times \mathbb{R}$ defining a $(p-1)$-dimensional hyperplane

$$\Pi_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b = 0\} = \Pi_{t(\mathbf{w},b)} \quad (\forall t > 0) \tag{16.131}$$

which satisfies

$$\left.\begin{aligned} \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, 1) \in \mathcal{D}\} \subset \Pi^+_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b > 0\} \\ \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, -1) \in \mathcal{D}\} \subset \Pi^-_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b < 0\} \end{aligned}\right\}. \tag{16.132}$$

In such a case, the so-called linear classifier is defined as a mapping

$$\mathcal{L}_{(\mathbf{w},b)} : \mathbb{R}^p \to \{-1, 1\} : \mathbf{x} \mapsto \begin{cases} +1 & \text{if } \mathbf{x} \in \Pi^+_{(\mathbf{w},b)}, \\ -1 & \text{if } \mathbf{x} \in \Pi^-_{(\mathbf{w},b)}, \end{cases} \tag{16.133}$$

which is hopefully a good approximation of the function $\mathfrak{L}$ observed partially through the training dataset $\mathcal{D}$. If $\mathcal{D}$ is linearly separable, there also exists infinitely many $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ satisfying

$$\left.\begin{aligned} \mathcal{D}_+ := \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, 1) \in \mathcal{D}\} \subset \Pi^{\geq 1}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b \geq 1\} \\ \mathcal{D}_- := \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, -1) \in \mathcal{D}\} \subset \Pi^{\leq -1}_{(\mathbf{w},b)} := \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top \mathbf{x} - b \leq -1\} \end{aligned}\right\}, \tag{16.134}$$

which is confirmed by rescaling $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ in (16.131) with a constant $t \geq 1/\min\{|\mathbf{w}^\top \mathbf{x}_i - b|\}_{i=1}^N > 0$.

The half-spaces $\Pi^{\geq 1}_{(\mathbf{w},b)}$ and $\Pi^{\leq -1}_{(\mathbf{w},b)}$ defined in (16.134) are main players in the following consideration on the linear classifier $\mathcal{L}_{(\mathbf{w},b)}$ even for linearly non-separable data $\mathcal{D}$. In this paper, the *margin* of the linear classifier $\mathcal{L}_{(\mathbf{w},b)}$ in (16.133) is defined by

$$\frac{1}{2}\mathrm{dist}\left(\Pi^{\geq 1}_{(\mathbf{w},b)}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right) = \frac{1}{2} \min_{\mathbf{x}_+ \in \Pi^{\geq 1}_{(\mathbf{w},b)}, \mathbf{x}_- \in \Pi^{\leq -1}_{(\mathbf{w},b)}} \|\mathbf{x}_+ - \mathbf{x}_-\| = \frac{1}{\|\mathbf{w}\|}. \tag{16.135}$$

By using the function h in (16.49) and

$$(\forall \mathbf{z} \in \mathbb{R}^p) \quad \begin{bmatrix} \mathrm{d}\left(\mathbf{z}, \Pi^{\geq 1}_{(\mathbf{w},b)}\right) = \begin{cases} \frac{|\mathbf{w}^\top \mathbf{z} - b - 1|}{\|\mathbf{w}\|} = \frac{1-(\mathbf{w}^\top \mathbf{z} - b)}{\|\mathbf{w}\|} & \text{if } \mathbf{z} \notin \Pi^{\geq 1}_{(\mathbf{w},b)}, \\ 0 & \text{otherwise,} \end{cases} \\ \mathrm{d}\left(\mathbf{z}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right) = \begin{cases} \frac{|\mathbf{w}^\top \mathbf{z} - b + 1|}{\|\mathbf{w}\|} = \frac{1+(\mathbf{w}^\top \mathbf{z} - b)}{\|\mathbf{w}\|} & \text{if } \mathbf{z} \notin \Pi^{\leq -1}_{(\mathbf{w},b)}, \\ 0 & \text{otherwise,} \end{cases} \end{bmatrix} \tag{16.136}$$

we deduce

$$\|\mathbf{w}\| \left[\sum_{\mathbf{z} \in \mathcal{D}_+} \mathrm{d}\left(\mathbf{z}, \Pi^{\geq 1}_{(\mathbf{w},b)}\right) + \sum_{\mathbf{z} \in \mathcal{D}_-} \mathrm{d}\left(\mathbf{z}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right) \right] = \sum_{i=1}^{N} h\left(y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right)\right) \tag{16.137}$$

which clarifies the geometric interpretation of "the *empirical hinge loss* of $\mathcal{L}_{(\mathbf{w},b)}$" defined in the right-hand side of (16.137) and ensures

$$\text{Condition (16.134)} \Leftrightarrow \sum_{i=1}^{N} h\left(y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right)\right) = 0. \tag{16.138}$$

For linearly separable data $\mathcal{D}$, among all $\mathcal{L}_{(\mathbf{w},b)}$ satisfying (16.134), the *Support Vector Machine (SVM)* $\mathcal{L}_{(\mathbf{w}^\star,b^\star)}$ was proposed in 1960s by Vapnik (see, e.g.,[135, 136]) as a special linear classifier which achieves maximal margin, i.e.,

$$\frac{1}{2}\mathrm{dist}\left(\Pi^{\geq 1}_{(\mathbf{w}^\star,b^\star)}, \Pi^{\leq -1}_{(\mathbf{w}^\star,b^\star)}\right) = \max_{(\mathbf{w},b) \text{ satisfies (16.138)}} \frac{1}{2}\mathrm{dist}\left(\Pi^{\geq 1}_{(\mathbf{w},b)}, \Pi^{\leq -1}_{(\mathbf{w},b)}\right). \tag{16.139}$$

Therefore the SVM $\mathcal{L}_{(\mathbf{w}^\star,b^\star)}$ for linearly separable $\mathcal{D}$ is given as the solution of the following convex optimization problem:

$$\text{minimize } \|\mathbf{w}\|^2 \text{subject to } \sum_{i=1}^{N} h\left(y_i \left(\mathbf{w}^\top \mathbf{x}_i - b\right)\right) = 0 \tag{16.140}$$

$$\Updownarrow$$

$$\text{minimize } \|\mathbf{w}\|^2 \text{subject to } (\mathbf{w}, b) \in \operatorname*{argmin}_{(\hat{\mathbf{w}},\hat{b}) \in \mathbb{R}^p \times \mathbb{R}} \sum_{i=1}^{N} h\left(y_i \left(\hat{\mathbf{w}}^\top \mathbf{x}_i - \hat{b}\right)\right), \tag{16.141}$$

where the last equivalence holds true under the linear separability of $\mathcal{D}$ because of the nonnegativity of h in (16.49).

The SVM defined equivalently in (16.139) or (16.140) or (16.141) for linearly separable training data has been extended for applications to even possibly linearly nonseparable training data $\mathcal{D}$ where the existence of $(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}$ satisfying (16.134) is no longer guaranteed. One of the most widely accepted extensions of (16.141) is known as *the soft margin hyperplane* [14, 25, 48, 73] which is characterized as a solution to the optimization problem:

$$\text{minimize, w.r.t. } (\mathbf{w}, b), \quad \frac{1}{2}\|\mathbf{w}\|^2 + \mathfrak{C}\sum_{i=1}^{N} h\left(y_i(\mathbf{w}^\top \mathbf{x}_i - b)\right) \tag{16.142}$$

or equivalently

$$\begin{aligned} &\text{minimize, w.r.t. } (\mathbf{w}, b, \xi), \quad \frac{1}{2}\|\mathbf{w}\|^2 + \mathfrak{C}\sum_{i=1}^{N} \xi_i \\ &\text{subject to } y_i\left(\mathbf{w}^\top \mathbf{x} - b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \ (i = 1, 2, \ldots, N), \end{aligned} \tag{16.143}$$

where $\mathfrak{C} > 0$ is a tuning parameter, and ξ_i $(i = 1, 2, \ldots, N)$ are slack variables.

Along the Cover's theorem (on the capacity of a space in linear dichotomies) [49], saying that the probability of any grouping of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \in \mathbb{R}^l$, in general position, into two classes to be linearly separable tends to unity as $l \to \infty$, another extension of the strategy $\mathcal{L}_{(\mathbf{w}^\star, b^\star)}$ of (16.139) into higher dimensional spaces was made in [16, 49], for application to possibly linearly nonseparable training data $\mathcal{D}$, by passing through a certain nonlinear transform $\mathfrak{N}\colon \mathbb{R}^p \to \mathbb{R}^l$ $(l \gg p)$ of the original training data $\mathcal{D} := \{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{-1, 1\} \mid i = 1, 2, \ldots, N\}$ to $\mathfrak{D} := \{(\mathfrak{N}(\mathbf{x}_i), y_i) \in \mathbb{R}^l \times \{-1, 1\} \mid i = 1, 2, \ldots, N\}$, where the nonlinear transform $\mathfrak{N}$ is defined usually in terms of kernel built in the theory of *the Reproducing Kernel Hilbert Space (RKHS)* [2, 124, 126] for exploiting the so-called kernel trick.

16.4.2 Optimal Margin Classifier with Least Empirical Hinge Loss

As suggested in [48, Sec.3], the original goal behind the soft margin hyperplane in (16.142) or (16.143) seems to determine $(\mathbf{w}^{\star\star}, b^{\star\star}) \in (\mathbb{R}^p \setminus \{\mathbf{0}\}) \times \mathbb{R}$ as the solution of the following nonconvex hierarchical optimization:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}^\star\|^2 \tag{16.144}$$

$$\text{subject to } (\mathbf{w}^\star, b^\star) \in \operatorname*{argmin}_{(\mathbf{w}, b)} |\mathcal{E}(\mathbf{w}, b)|, \tag{16.145}$$

where $|\cdot|$ stands for the cardinality of a set and $\mathcal{E}(\mathbf{w}, b) \subset \mathcal{D}_+ \cup \mathcal{D}_-$ is the training error set defined as

$$\mathcal{E}(\mathbf{w}, b) := \left\{\mathbf{z} \in \mathcal{D}_+ \mid \mathrm{d}\left(\mathbf{z}, \Pi^{\geq 1}_{(\mathbf{w}, b)}\right) > 0\right\} \cup \left\{\mathbf{z} \in \mathcal{D}_- \mid \mathrm{d}\left(\mathbf{z}, \Pi^{\leq -1}_{(\mathbf{w}, b)}\right) > 0\right\}, \tag{16.146}$$

i.e., determining a special hyperplane $(\mathbf{w}^{\star\star}, b^{\star\star})$, which achieves maximal margin in the set $\operatorname{argmin}_{(\mathbf{w}, b)} |\mathcal{E}(\mathbf{w}, b)|$, is desired.

Unfortunately, since the problem to determine $(\mathbf{w}^\star, b^\star)$ in (16.145) is in general NP-hard [15, 48, 84] and since (16.137) implies that

$$(16.145) \Leftrightarrow (\mathbf{w}^\star, b^\star) \in \operatorname*{argmin}_{(\mathbf{w},b)} \sum_{i=1}^{N} \left[\lim_{\sigma \downarrow 0} h^\sigma \left(y_i (\mathbf{w}^\top \mathbf{x}_i - b) \right) \right],$$

the original goal set in (16.144–16.146) was replaced in [48, Sec.3] by a realistic goal (16.142) [or (16.143)] for a sufficiently large constant $\mathfrak{C} > 0$. However, unlike the desired solution of (16.144–16.146), the soft margin hyperplane in (16.142) applied to linearly separable data has no guarantee to reproduce the original SVM in (16.139).

The above observations induce a natural question:
Is the solution of (16.142) for general training data really a mathematically sound extension of the original SVM defined equivalently in (16.139) or (16.140) or (16.141) specialized for linearly separable training data?[14]

Clearly, this question comes from essentially common concern as seen in Scenario 1, therefore, an alternative natural extension of the original SVM in (16.141) would be the solution of the optimization problem:

$$\text{minimize } \frac{1}{2}\|\mathbf{w}^\star\|^2 \text{ subject to } (\mathbf{w}^\star, b^\star) \in \Gamma := \operatorname*{argmin}_{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}} \sum_{i=1}^{N} h\left(y_i (\mathbf{w}^\top \mathbf{x}_i - b) \right) \tag{16.147}$$

which does not seem different from (16.141) at a glance but is defined even possibly for linearly nonseparable training data $\mathcal{D}$. Remark that the hierarchical convex optimization problem (16.147) is a more faithful convex relaxation of (16.144–16.146) than the convex optimization (16.142) [or its equivalent formulation (16.143) with slack variables.[15]] for the soft margin hyperplane. This is because the solution of (16.147) for linearly separable data certainly reproduces the original SVM in (16.139). As remarked in Example 16.1(b) in Section 16.1, in general, the soft

[14] This question is common even for the soft margin SVM applied to the transformed data $\mathfrak{D}$ employed in [16] because the linear separability of $\mathfrak{D}$ is not always guaranteed.

[15] In terms of slack variables, Problem (16.147) can also be restated as

$$\begin{aligned}
&\text{minimize } \frac{1}{2}\|\mathbf{w}^\star\|^2 \\
&\text{subject to } (\mathbf{w}^\star, b^\star, \xi^\star) \in \operatorname*{argmin}_{(\mathbf{w},b,\xi)\in\mathbb{R}^p\times\mathbb{R}\times\mathbb{R}^N} \sum_{i=1}^{N} \left[\xi_i + \iota_{S_i}(\mathbf{w}, b, \xi) + \iota_{\Xi_i}(\mathbf{w}, b, \xi) \right], \\
&\text{where } S_i := \left\{ (\mathbf{w}, b, \xi) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^N \mid y_i \left(\mathbf{w}^\top \mathbf{x}_i - b \right) \geq 1 - \xi_i \right\} \\
&\text{and } \Xi_i := \left\{ (\mathbf{w}, b, \xi) \in \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^N \mid \xi_i \geq 0 \right\} \quad (i = 1, 2, \ldots, N).
\end{aligned}$$

margin hyperplane via (16.142) for a fixed constant $\mathfrak{C} > 0$ does not achieve the hierarchical optimality in the sense of (16.147). Fortunately, Problem (16.147) falls in the class of the hierarchical convex optimization problems of type (16.13).

In the following, we demonstrate how Problem (16.147) can be solved by a proposed strategy in Section 16.3. Let $\mathcal{X} := \mathbb{R}^p \times \mathbb{R}$, $A_i \colon \mathcal{X} \to \mathbb{R} \colon (\mathbf{w}, b) \mapsto y_i(\mathbf{x}_i^\top \mathbf{w} - b)$ $(i = 1, 2, \ldots, N)$, $f \colon \mathcal{X} \to \mathbb{R} \colon (\mathbf{w}, b) \mapsto h(A_N(\mathbf{w}, b))$, $g := \bigoplus_{i=1}^{N-1} h$, and $A \colon \mathbb{R}^{p+1} \to \mathbb{R}^{N-1} \colon (\mathbf{w}, b) \mapsto (A_i(\mathbf{w}, b))_{i=1}^{N-1}$. By using these translations, we can express the hinge loss function, in the form of the first stage cost function of (16.13), as

$$\sum_{i=1}^{N} h\left(y_i(\mathbf{w}^\top \mathbf{x}_i - b)\right) = g \circ A(\mathbf{w}, b) + f(\mathbf{w}, b)$$

and its associated qualification condition (see (16.40)) is verified by

$$\operatorname{ri}\left(\operatorname{dom}(g) - A\operatorname{dom}(f)\right) = \operatorname{ri}\left(\mathbb{R}^{N-1} - A\operatorname{dom}(f)\right) = \operatorname{ri}\left(\mathbb{R}^{N-1}\right) = \mathbb{R}^{N-1} \ni 0. \tag{16.148}$$

Note that, for any $\gamma \in \mathbb{R}_{++}$, the proximity operator $\operatorname{prox}_{\gamma h}$ can be computed as (16.50) in Example 16.7(c) (see Section 16.2.2) and therefore $\operatorname{prox}_f = \operatorname{prox}_{h \circ A_N}$ can also be computed by applying (16.48) and (16.50) in Example 16.7(b)(c). Moreover, by introducing $\Psi \colon \mathbb{R}^p \times \mathbb{R} \to \mathbb{R} \colon (\mathbf{w}, b) \mapsto \frac{1}{2}\|\mathbf{w}\|^2$, we can regard Problem (16.147) as an instance of Problem (16.13) under the assumption of $\mathcal{S}_p := \Gamma \neq \varnothing$. In fact, we can apply Theorem 16.15, Theorem 16.17, and Theorem 16.21 to (16.147) because Ψ is not strictly convex. In the following numerical experiment, we applied Theorem 16.17 to (16.147) with slight modification.[16]

[16] If we need to guarantee $\mathcal{S}_p$[in (16.13)] $\neq \varnothing$, we recommend the following slight modification of (16.147):

$$\underset{\mathbf{w}^\star \in \widetilde{\Gamma}}{\text{minimize}} \ \frac{1}{2}\|\mathbf{w}^\star\|^2 \text{ subject to } \widetilde{\Gamma} := \underset{(\mathbf{w}, b) \in \mathbb{R}^p \times \mathbb{R}}{\operatorname{argmin}} \left[\Phi(\mathbf{w}, b) := \iota_{\overline{B}(0,r)}(\mathbf{w}, b) + \sum_{i=1}^{N} h\left(y_i(\mathbf{w}^\top \mathbf{x}_i - b)\right) \right]$$

with a sufficiently large closed ball $\overline{B}(0, r)$, where $\mathcal{S}_p := \widetilde{\Gamma} \neq \varnothing$ is guaranteed due to the coercivity of Φ. Fortunately, our strategies in Section 16.3 are still applicable to this modified problem because it is also an instance of (16.10) which can be translated into (16.13) as explained in Section 16.1. In the application of Theorem 16.17 in Section 16.3.1 to this modification, the boundedness of $\operatorname{Fix}(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})$ is automatically guaranteed because of Corollary 16.24(b) (see Section 16.3.3) and the boundedness of both $\widetilde{\Gamma} \subset \overline{B}(0, r)$ and $\partial h(\mathbb{R}) = \partial h(\mathbb{R} \setminus \{1\}) \cup \partial h(\{1\}) = [-1, 0]$.

16.4.3 *Numerical Experiment: Margin Maximization with Least Empirical Hinge Loss*

We demonstrate that, as an extension of the original SVM in (16.139), the hierarchical enhancement of the SVM in (16.147) is more faithful to the original SVM than the soft margin SVM (16.142). In our experiment, we applied the original SVM in (16.139), the soft margin SVM in (16.142), and the proposed hierarchical enhancement of the SVM in (16.147) to the Iris dataset which is a famous dataset used firstly in Fisher's paper [65]. This data set has 150 sample points, which are divided into three classes (I(setosa), II(versicolor), III(virginica)), and each sample point has four features (sepal length, sepal width, petal length, and petal width). From Iris dataset, we construct two datasets: separable $\mathcal{D}_{\mathrm{sep}} \subset \mathbb{R}^2 \times \{-1, 1\}$ with $|\mathcal{D}_{\mathrm{sep}}| = 100$ comprising all the samples of Class I and Class II having only sepal length and sepal width; and non-separable $\mathcal{D}_{\mathrm{nsep}} \subset \mathbb{R}^2 \times \{-1, 1\}$ with $|\mathcal{D}_{\mathrm{nsep}}| = 100$ comprising all the samples of Class II and Class III having only petal length and petal width. For each linear classifier $\mathcal{L}_{(\mathbf{w},b)}$ of our interest, the three hyperplanes $\Pi_{(\mathbf{w},b)}$, $\Pi_{(\mathbf{w},b+1)}$, $\Pi_{(\mathbf{w},b-1)}$ (see (16.131) and (16.134)) are drawn in Figs. 16.1 and 16.2, in cyan for "Original SVM," in green for "Soft Margin SVM," and in magenta for the proposed "M^{2}LEHL" (which stands for *the Margin Maximization with Least Empirical Hinge Loss*), respectively, where $(\mathbf{w}, b)_{\mathrm{org}}$ is obtained by applying a quadratic programming solver `quadprog` in Matlab to (16.139), $(\mathbf{w}, b)_{\mathrm{soft}}$ is obtained by applying a soft margin SVM solver `fitcsvm` (with the default setting, i.e., $\mathfrak{C} = 1$) in Matlab to (16.142), and $(\mathbf{w}, b)_{\mathrm{M^2LEHL}}$ is obtained by

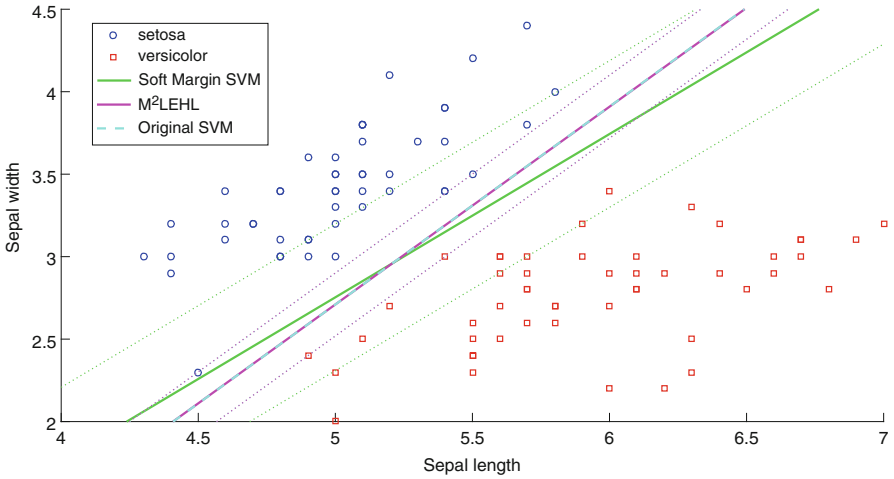


Fig. 16.1 Comparison between M^{2}LEHL, Original SVM, and Soft Margin SVM (Case of a separable training dataset $\mathcal{D}_{\mathrm{sep}}$): M^{2}LEHL reproduces Original SVM

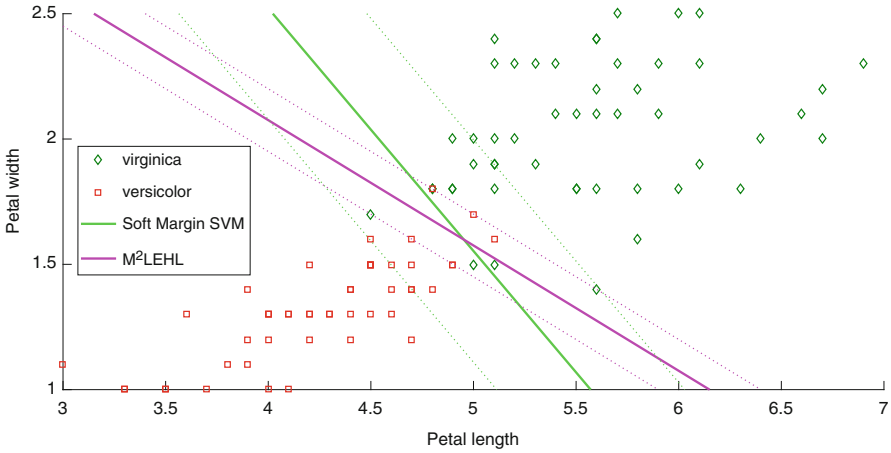


Fig. 16.2 Comparison between M^{2}LEHL and Soft Margin SVM (Case of a nonseparable training dataset $\mathcal{D}_{\text{nsep}}$)

applying the proposed algorithmic solution in Section 16.4.2, to (16.147), designed based on Theorem 16.17 with slight modification.[17]

Figure 16.1 illustrates the resulting separating hyperplanes for the separable dataset $\mathcal{D}_{\text{sep}}$. Since the magenta lines are completely overlapped with cyan lines, "M^{2}LEHL" reproduces "Original SVM," as explained in just after (16.147). "Soft Margin SVM" does not succeed in maximizing the margin, i.e., (16.147) is a more faithful extension of the original SVM in (16.139) than the soft margin SVM (16.142).

Figure 16.2 illustrates the resulting separating hyperplanes for the nonseparable dataset $\mathcal{D}_{\text{nsep}}$. Since the original SVM (16.139) has no solution, "Original SVM" is not depicted. As the performance measure, we employ the number of errors $|\mathcal{E}(\cdot)|$ defined in (16.146) along the original goal (16.144) (as suggested in [48, Sec. 3]). Though "Soft Margin SVM" has 21 errors, "M^{2}LEHL" achieves only 6 errors, which demonstrates that (16.147) is more effective formulation for approaching to the original goal (16.144) than the soft margin SVM (16.142).

16.5 Application to Hierarchical Enhancement of Lasso

16.5.1 TREX: A Nonconvex Automatic Sparsity Control of Lasso

Consider the estimation of a sparse vector $\mathbf{b}^{\text{tru}} \in \mathbb{R}^p$ in the standard linear regression model:

[17] See footnote 16.

$$\mathbf{z} = \mathbf{X}\mathbf{b}^{\mathrm{tru}} + \sigma\mathbf{e}, \tag{16.149}$$

where $\mathbf{z} = (z_1, \ldots, z_N)^\top \in \mathbb{R}^N$ is a response vector, $\mathbf{X} \in \mathbb{R}^{N\times p}$ a design matrix, $\sigma > 0$ a constant, $\mathbf{e} = (\varepsilon_1, \ldots, \varepsilon_N)^\top$ the noise vector, each ε_i is the realization of a random variable with mean zero and variance 1.

The Lasso (Least Absolute Shrinkage and Selection Operator) [132] has been used widely as one of the most well-known sparsity aware statistical estimation methods [73, 74]. The Lasso for (16.149) is defined as a minimizer of the least squares criterion with ℓ_1 penalty, i.e.,

$$\mathbf{b}_{\mathrm{Lasso}}(\lambda) \in \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2N}\|\mathbf{z} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1, \tag{16.150}$$

where the tuning parameter $\lambda > 0$ aims at controlling the sparsity of $\mathbf{b}_{\mathrm{Lasso}}(\lambda)$. However selection of $\lambda > 0$ is highly influential to $\mathbf{b}_{\mathrm{Lasso}}(\lambda)$ and therefore its reliable way of selection has been strongly desired. Among many efforts toward automatic sparsity control of Lasso, the following prediction bound offers a firm basis and has been applied widely in recent strategies including [50, 69, 77, 89].

Fact 16.25 (A Prediction Bound of Lasso [87, 120]) *For* $\lambda \geq \frac{2\|\mathbf{X}^\top(\mathbf{z}-\mathbf{X}\mathbf{b}^{tru})\|_\infty}{N}$, *it holds* $\frac{\|\mathbf{X}\mathbf{b}_{Lasso}(\lambda)-\mathbf{X}\mathbf{b}^{tru}\|^2}{N} \leq 2\lambda\|\mathbf{b}^{tru}\|_1$.

The TREX (Tuning-free Regression that adapts to the Entire noise $\sigma\mathbf{e}$ and the design matrix $\mathbf{X}$) [89] is one of the state-of-the-art strategies based on Fact 16.25. The TREX is defined as a solution of a nonconvex optimization problem:

$$\text{find } \mathbf{b}_{\mathrm{TREX}} \in \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p} \frac{\|\mathbf{X}\mathbf{b} - \mathbf{z}\|^2}{\left\|\mathbf{X}^\top(\mathbf{X}\mathbf{b} - \mathbf{z})\right\|_\infty} + \beta\|\mathbf{b}\|_1, \tag{16.151}$$

where $\left\|\mathbf{X}^\top(\mathbf{X}\mathbf{b} - \mathbf{z})\right\|_\infty = \max_{1\leq j\leq p}\left|\mathbf{X}_{:j}^\top(\mathbf{X}\mathbf{b} - \mathbf{z})\right|$, $\mathbf{X}_{:j}$ denotes the jth column of $\mathbf{X}$, and the parameter β can be set to a constant value ($\beta = 1/2$ being the default choice).

The authors in [13] cleverly decomposed the nonconvex optimization (16.151) into $2p$ subproblems:

$$\text{find } \mathbf{b}_{\mathrm{TREX}}^{(j)} \in \operatorname*{argmin}_{\substack{\mathbf{b}\in\mathbb{R}^p \\ \mathbf{x}_j^\top(\mathbf{X}\mathbf{b}-\mathbf{z})>0}} \left[\frac{\|\mathbf{X}\mathbf{b} - \mathbf{z}\|^2}{\beta\mathbf{x}_j^\top(\mathbf{X}\mathbf{b} - \mathbf{z})} + \|\mathbf{b}\|_1\right], \tag{16.152}$$

where

$$\mathbf{x}_j = \begin{cases} \mathbf{X}_{:j} & (j = 1, 2, \ldots, p); \\ -\mathbf{X}_{:j-p} (j = p+1, p+2, \ldots, 2p). \end{cases} \tag{16.153}$$

More precisely, $\mathbf{b}_{\mathrm{TREX}}$ in (16.151) is characterized as

$$\mathbf{b}_{\mathrm{TREX}} \in \widehat{\Omega}_{\mathbb{R}^p} \left[\operatorname*{argmin}_{\substack{(\mathbf{b}, j) \in \mathbb{R}^p \times \{1,2,\ldots,2p\} \\ \mathbf{x}_j^\top (\mathbf{X}\mathbf{b} - \mathbf{z}) > 0}} \left(\frac{\|\mathbf{X}\mathbf{b} - \mathbf{z}\|^2}{\beta \mathbf{x}_j^\top (\mathbf{X}\mathbf{b} - \mathbf{z})} + \|\mathbf{b}\|_1 \right) \right], \tag{16.154}$$

where

$$\widehat{\Omega}_{\mathbb{R}^p} : \mathbb{R}^p \times \{1, 2, \ldots, 2p\} \to \mathbb{R}^p : (\mathbf{b}, j) \mapsto \mathbf{b}. \tag{16.155}$$

Remarkably, each subproblem (16.152) was shown to be a convex optimization and solved in [13] with a second-order cone program (SOCP) [92].

Recently, for sound extensions of the subproblem (16.152) as well as for sound applications of proximal splitting, a successful reformulation of (16.154) was made for general $q > 1$ in [38] as

$$\mathbf{b}_{\mathrm{TREX_q}} \in \mathcal{S}_{\mathrm{TREX_q}} := \widehat{\Omega}_{\mathbb{R}^p} \left[\operatorname*{argmin}_{(\mathbf{b}, j) \in \mathbb{R}^p \times \{1,2,\ldots,2p\}} g_{(j,q)}(\mathbf{M}_j \mathbf{b}) + \|\mathbf{b}\|_1 \right] \tag{16.156}$$

whose solution $\mathbf{b}_{\mathrm{TREX_q}}$ is given, by passing through $2p$ convex subproblems, as $\mathbf{b}^{(j^\star)}_{\mathrm{TREX_q}}$, where

$$\begin{bmatrix} \mathbf{b}^{(j)}_{\mathrm{TREX_q}} \in \mathcal{S}_{(j,q)} := \operatorname*{argmin}_{\mathbf{b} \in \mathbb{R}^p} \left[g_{(j,q)}(\mathbf{M}_j \mathbf{b}) + \|\mathbf{b}\|_1 \right] \ (j = 1, 2, \ldots, 2p); \\ j^\star \in \operatorname*{argmin}_{j \in \{1,2,\ldots,2p\}} \left[g_{(j,q)}(\mathbf{M}_j \mathbf{b}^{(j)}_{\mathrm{TREX_q}}) + \|\mathbf{b}^{(j)}_{\mathrm{TREX_q}}\|_1 \right], \end{bmatrix} \tag{16.157}$$

$$g_{(j,q)} : \mathbb{R} \times \mathbb{R}^N \to (-\infty, \infty] : (\eta, \mathbf{y}) \mapsto \begin{cases} \frac{\|\mathbf{y} - \mathbf{z}\|^q}{\beta (\eta - \mathbf{x}_j^\top \mathbf{z})^{q-1}}, & \text{if } \eta > \mathbf{x}_j^\top \mathbf{z}; \\ 0, & \text{if } \mathbf{y} = \mathbf{z} \text{ and } \eta = \mathbf{x}_j^\top \mathbf{z}; \\ +\infty, & \text{otherwise} \end{cases} \tag{16.158}$$

is a proper lower semicontinuous convex function, and

$$\mathbf{M}_j : \mathbb{R}^p \to \mathbb{R} \times \mathbb{R}^N : \mathbf{b} \mapsto \left(\mathbf{x}_j^\top \mathbf{X}\mathbf{b}, \mathbf{X}\mathbf{b} \right) \tag{16.159}$$

is a bounded linear operator. The estimator $\mathbf{b}_{\mathrm{TREX}_{\mathrm{q}}}$ in (16.156) is called the generalized TREX in [38] where, as its specialization, $\mathbf{b}_{\mathrm{TREX}_2}$ is also called TREX. Note that, in view of Example 16.7(d)(e) in Section 16.2.2 and a relation between $g_{(j,q)}$ and $\widetilde{\varphi}_q$ in (16.51) [38, in Sec. 4.3.2]:

$$(\forall(\eta,\mathbf{y})\in\mathbb{R}\times\mathbb{R}^N)\quad g_{(j,q)}(\eta,\mathbf{y})=\tau_{\left(\mathbf{x}_j^\top\mathbf{z},\mathbf{z}\right)}\widetilde{\varphi}_q\,(\eta,\mathbf{y})=\widetilde{\varphi}_q\left(\eta-\mathbf{x}_j^\top\mathbf{z},\mathbf{y}-\mathbf{z}\right), \tag{16.160}$$

each convex subproblem in (16.157) is an instance of Problem (16.1).

For the subproblem (16.157), the Douglas-Rachford splitting method (see Proposition 16.9 in Section 16.2.3) was successfully applied in [38]. For completeness, we reproduce this result in the style of (16.25–16.27) followed by application of Fact 16.6 (in Section 16.2.2) to the characterization (16.78). Suppose that for (16.157) the qualification condition (see (16.40))

$$0\in\mathrm{ri}(\mathrm{dom}(g_{(j,q)})-\mathbf{M}_j\,\mathrm{dom}(\|\cdot\|_1)) \tag{16.161}$$

holds.[18] Then, by using

$$\begin{cases}\check{\mathbf{M}}_j\colon\mathbb{R}^p\times\mathbb{R}^{N+1}\to\mathbb{R}^{N+1}\colon(\mathbf{b},\mathbf{c})\mapsto\mathbf{M}_j\mathbf{b}-\mathbf{c},\\ \mathcal{Q}_{\mathbb{R}^p}\colon\mathbb{R}^p\times\mathbb{R}^{N+1}\to\mathbb{R}^p\colon(\mathbf{b},\mathbf{c})\mapsto\mathbf{b},\end{cases} \tag{16.162}$$

we obtain

$$\mathcal{S}_{(j,q)}=\mathcal{Q}_{\mathbb{R}^p}\circ P_{\mathcal{N}(\check{\mathbf{M}}_j)}\left(\mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)}\right)\right) \tag{16.163}$$

(which is a specialization of (16.78) for (16.157), see Figure 16.3), where

$$\mathbf{T}_{\mathrm{DRS_I}}^{(j,q)}\colon\mathbb{R}^p\times\mathbb{R}^{N+1}\to\mathbb{R}^p\times\mathbb{R}^{N+1}\colon(\mathbf{b},\mathbf{c})\mapsto(\mathbf{b}_T,\mathbf{c}_T) \tag{16.164}$$

is the DRS operator of Type-I (c.f., (16.68) and (16.71)) specialized for (16.157) and is defined by

$$\begin{cases}\mathbf{p}=\mathbf{b}-\mathbf{M}_j^*(\mathbf{I}+\mathbf{M}_j\mathbf{M}_j^*)^{-1}(\mathbf{M}_j\mathbf{b}-\mathbf{c})\\ (\mathbf{b}_{1/2},\mathbf{c}_{1/2})=(2\mathbf{p}-\mathbf{b},2\mathbf{M}_j\mathbf{p}-\mathbf{c}))\\ (\mathbf{b}_T,\mathbf{c}_T)=(2\,\mathrm{prox}_{\|\cdot\|_1}(\mathbf{b}_{1/2})-\mathbf{b}_{1/2},2\,\mathrm{prox}_{g_{(j,q)}}(\mathbf{c}_{1/2})-\mathbf{c}_{1/2}),\end{cases} \tag{16.165}$$

[18] In [38], the qualification condition (16.161) seems to be assumed implicitly. If we assume additionally that $\mathbf{X}\in\mathbb{R}^{N\times p}$ has no zero column, it is automatically guaranteed as will be shown in Lemma 16.27 in Section 16.5.2.

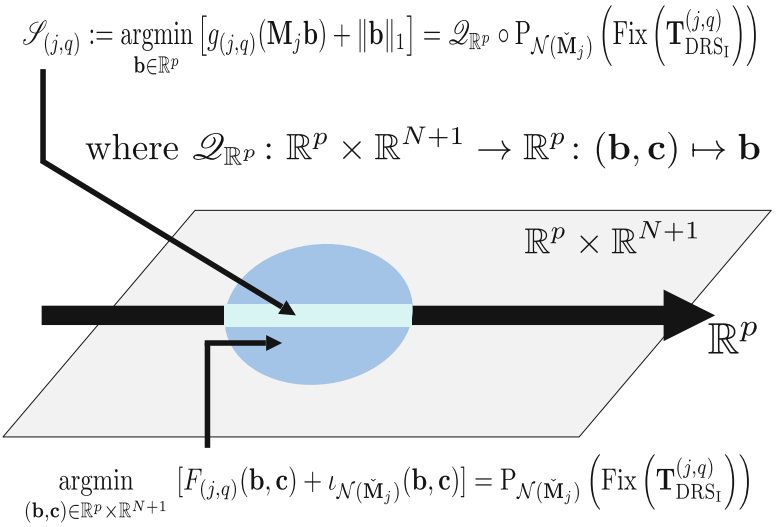


Fig. 16.3 Illustration of the fixed point characterization of $\mathcal{S}_{(j,q)}$ in (16.157) via the Douglas-Rachford splitting operator $\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}$ in (16.164)

or equivalently by

$$\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}} := (2\,\mathrm{prox}_{F_{(j,q)}} - \mathbf{I}) \circ (2P_{\mathcal{N}(\check{\mathbf{M}}_j)} - \mathbf{I}) \tag{16.166}$$

with $F_{(j,q)} : \mathbb{R}^p \times \mathbb{R}^{N+1} \to (-\infty, \infty] : (\mathbf{b}, \mathbf{c}) \mapsto g_{(j,q)}(\mathbf{c}) + \|\mathbf{b}\|_1$. Note that $\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}$ can be computed efficiently if $(\mathbf{I} + \mathbf{M}_j\mathbf{M}_j^*)^{-1}$ is available as a computational tool.

The above characterization (16.163) and Fact 16.6 (see Section 16.2.2) lead to the following algorithmic solution of (16.157).

Fact 16.26 (Douglas–Rachford Splitting Method for Subproblems of Generalized TREX) *Under the qualification condition* (16.161) *for* $\mathcal{S}_{(j,q)}$ *in* (16.157)*, the sequence* $(\mathbf{b}_n, \mathbf{c}_n)_{n\in\mathbb{N}} \subset \mathbb{R}^p \times \mathbb{R}^{N+1}$ *generated, with* $(\alpha_n)_{n\in\mathbb{N}} \subset [0, 1]$ *satisfying* $\sum_{n\in\mathbb{N}} \alpha_n(1-\alpha_n) = \infty$ *in Fact 16.6 (see Section 16.2.2) and* $(\mathbf{b}_0, \mathbf{c}_0) \in \mathbb{R}^p \times \mathbb{R}^{N+1}$*, by*

$$(\mathbf{b}_{n+1}, \mathbf{c}_{n+1}) = (1 - \alpha_n)(\mathbf{b}_n, \mathbf{c}_n) + \alpha_n \mathbf{T}^{(j,q)}_{DRS_I}(\mathbf{b}_n, \mathbf{c}_n) \tag{16.167}$$

converges to a point $(\mathbf{b}_\star, \mathbf{c}_\star)$ *in* $\mathrm{Fix}\left(\mathbf{T}^{(j,q)}_{DRS_I}\right)$ *as well as the sequence* $(\mathcal{Q}_{\mathbb{R}^p} \circ P_{\mathcal{N}(\check{\mathbf{M}}_j)}(\mathbf{b}_n, \mathbf{c}_n))_{n\in\mathbb{N}}$ *converges to* $\mathcal{Q}_{\mathbb{R}^p} \circ P_{\mathcal{N}(\check{\mathbf{M}}_j)}(\mathbf{b}_\star, \mathbf{c}_\star) \in \mathcal{S}_{(j,q)}$*, where*

$$\mathcal{Q}_{\mathbb{R}^p} \circ P_{\mathcal{N}(\check{\mathbf{M}}_j)} : \mathbb{R}^p \times \mathbb{R}^{N+1} \to \mathbb{R}^p : (\mathbf{b}, \mathbf{c}) \mapsto \mathbf{b} - \mathbf{M}_j^*(\mathbf{I} + \mathbf{M}_j\mathbf{M}_j^*)^{-1}(\mathbf{M}_j\mathbf{b} - \mathbf{c}).$$

Note that the sequence $(\mathfrak{Q}_{\mathbb{R}^p} \circ P_{\mathcal{N}(\check{\mathbf{M}}_j)}(\mathbf{b}_n, \mathbf{c}_n))_{n\in\mathbb{N}}$ can be generated efficiently by (16.167) if $(\mathbf{I} + \mathbf{M}_j\mathbf{M}_j^*)^{-1}$ is available as a computational tool.

16.5.2 Enhancement of Generalized TREX Solutions with Hierarchical Optimization

Along Scenario 2 in Section 16.1, suppose that we found newly an effective criterion $\Psi \in \Gamma_0(\mathbb{R}^p)$ whose gradient is Lipschitzian over $\mathbb{R}^p$ and we hope to select a most desirable vector, in the sense of Ψ, from the solution set $\mathcal{S}_{\mathrm{TREX_q}}$ in (16.156). This task is formulated as a *hierarchical nonconvex optimization problem* (see (16.155), (16.158), and (16.159) for $\widehat{\mathfrak{Q}}_{\mathbb{R}^p}$, $g_{(j,q)}$, and $\mathbf{M}_j$):

$$\begin{aligned} &\text{minimize} \quad \Psi(\mathbf{b}^\star) \\ &\text{subject to} \quad \mathbf{b}^\star \in \mathcal{S}_{\mathrm{TREX_q}} = \widehat{\mathfrak{Q}}_{\mathbb{R}^p}\left[\operatorname*{argmin}_{(\mathbf{b},j)\in\mathbb{R}^p\times\{1,2,\ldots,2p\}} g_{(j,q)}(\mathbf{M}_j\mathbf{b}) + \|\mathbf{b}\|_1\right] \end{aligned} \tag{16.168}$$

whose solution $\mathbf{b}_{\mathrm{HTREX_q}}$ is given, by passing through $2p$ (*hierarchical convex* optimization) subproblems, as $\mathbf{b}^{(j^{\star\star})}_{\mathrm{HTREX_q}}$, where

$$\left[\begin{array}{l} \mathbf{b}^{(j)}_{\mathrm{HTREX_q}} \in \mathit{\Omega}^{(j,q)}_{\mathrm{DRS_I}} := \operatorname*{argmin}_{\mathbf{b}^\star\in\mathcal{S}_{(j,q)}} \Psi(\mathbf{b}^\star), \\ \mathcal{S}_{(j,q)} = \operatorname*{argmin}_{\mathbf{b}\in\mathbb{R}^p}\left[g_{(j,q)}(\mathbf{M}_j\mathbf{b}) + \|\mathbf{b}\|_1\right] \quad (j = 1, 2, \ldots, 2p) \text{ [in (16.157)]}; \\ \mathfrak{J}^\star := \operatorname*{argmin}_{j\in\{1,2,\ldots,2p\}}\left[g_{(j,q)}(\mathbf{M}_j\mathbf{b}^{(j)}_{\mathrm{HTREX_q}}) + \|\mathbf{b}^{(j)}_{\mathrm{HTREX_q}}\|_1\right]; \\ j^{\star\star} \in \operatorname*{argmin}_{j^\star\in\mathfrak{J}^\star} \Psi(\mathbf{b}^{(j^\star)}_{\mathrm{HTREX_q}}). \end{array}\right. \tag{16.169}$$

Note that the coercivity of $\|\cdot\|_1$ and the nonnegativity of $g_{(j,q)}$ ensure that $\mathcal{S}_{(j,q)}$ is nonempty and bounded (see Fact 16.2(c) in Section 16.2.1), which also guarantees $\mathit{\Omega}^{(j,q)}_{\mathrm{DRS_I}} = \operatorname{argmin}(\iota_{\mathcal{S}_{(j,q)}} + \Psi)(\mathbb{R}^p) \neq \varnothing$ $(j = 1, 2, \ldots, 2p)$ by the classical Weierstrass theorem.

In the following, we focus on how to compute the solution $\mathbf{b}^{(j)}_{\mathrm{HTREX_q}}$ $(j = 1, 2, \ldots, 2p)$ in (16.169) by a proposed strategy in Section 16.3. We assume that the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ in (16.149) has no zero column, to guarantee the qualification condition (16.161) for $\mathcal{S}_{(j,q)}$ in (16.157) for each $j = 1, 2, \ldots, 2p$.

Lemma 16.27 *Suppose that the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ has no zero column. Then the qualification condition (16.161) for $\mathcal{S}_{(j,q)}$ in (16.157) is guaranteed automatically for each $j = 1, 2, \ldots, 2p$.*

(The proof of Lemma 16.27 is given in Appendix G).

Theorem 16.28 (Algorithmic Solution to Hierarchical TREX$_q$) *Suppose that* $\mathbf{X}$ *has no zero column and* $\Psi \in \Gamma_0(\mathbb{R}^p)$ *is Gâteaux differentiable with Lipschitzian gradient* $\nabla\Psi$ *over* $\mathbb{R}^p$. *Then, for* $\mathbf{T}^{(j,q)}_{DRS_I}$ *in* (16.164) *(*$j = 1, 2, \ldots, 2p$*),*

(a) $\mathrm{Fix}(\mathbf{T}^{(j,q)}_{DRS_I})$ *is bounded;*

(b) $\mathbf{T}^{(j,q)}_{DRS_I}$ *can be plugged into the HSDM* (16.65)*, with any* $\alpha \in (0,1)$ *and* $(\lambda_{n+1})_{n\in\mathbb{N}} \in \ell^2_+ \setminus \ell^1_+$, *as*

$$\begin{cases} (\mathbf{b}_{n+1/2}, \mathbf{c}_{n+1/2}) = (1-\alpha)(\mathbf{b}_n, \mathbf{c}_n) + \alpha \mathbf{T}^{(j,q)}_{DRS_I}(\mathbf{b}_n, \mathbf{c}_n) \\ \mathbf{b}^{\star}_{n+1} = \mathbf{b}_{n+1/2} - \mathbf{M}^*_j(\mathbf{I} + \mathbf{M}_j\mathbf{M}^*_j)^{-1}(\mathbf{M}_j\mathbf{b}_{n+1/2} - \mathbf{c}_{n+1/2}) \\ \mathbf{b}_{n+1} = \mathbf{b}_{n+1/2} - \lambda_{n+1}(\mathbf{I} - \mathbf{M}^*_j(\mathbf{I} + \mathbf{M}_j\mathbf{M}^*_j)^{-1}\mathbf{M}_j) \circ \nabla\Psi(\mathbf{b}^{\star}_{n+1}) \\ \mathbf{c}_{n+1} = \mathbf{c}_{n+1/2} - \lambda_{n+1}((\mathbf{I} + \mathbf{M}_j\mathbf{M}^*_j)^{-1}\mathbf{M}_j) \circ \nabla\Psi(\mathbf{b}^{\star}_{n+1}). \end{cases} \tag{16.170}$$

The algorithm (16.170) *generates, for any* $(\mathbf{b}_0, \mathbf{c}_0) \in \mathbb{R}^p \times \mathbb{R}^{N+1}$, *a sequence* $(\mathbf{b}^{\star}_{n+1})_{n\in\mathbb{N}} \subset \mathbb{R}^p$ *which satisfies*

$$\lim_{n\to\infty} d_{\Omega^{(j,q)}_{DRS_I}}(\mathbf{b}^{\star}_n) = 0,$$

where $\Omega^{(j,q)}_{DRS_I} \neq \varnothing$ *is defined in* (16.169).

Remark 16.29 (Idea Behind Derivation of Theorem 16.28)

(a) Recall that $\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}$ is a DRS operator of Type-I (see (16.164) and Theorem 16.15 in Section 16.3.1). By applying Corollary 16.24(a) in Section 16.3.3 to Lemma 16.27, the boundedness of $\mathcal{S}_{(j,q)} \neq \varnothing$, and the boundedness of the image of $\partial\|\cdot\|_1 \colon \mathbb{R}^p \to [-1,1]^p \colon \mathbf{b} = (b_1, b_2, \ldots, b_p) \mapsto \bigtimes_{i=1}^p \partial|\cdot|(b_i)$, we deduce the relation:

$$\left[-(\mathbf{M}^\top_j)^{-1}(\partial\|\cdot\|_1(\mathcal{S}_{(j,q)}))\right] \cap \partial g_{(j,q)}(\mathbb{R}^{N+1}) \text{ is bounded} \tag{16.171}$$
$$\Rightarrow \mathrm{Fix}(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}}) \text{ is bounded},$$

where $(\mathbf{M}^\top_j)^{-1} \colon \mathbb{R}^p \to 2^{\mathbb{R}^{N+1}} \colon \mathbf{b} \mapsto \{\mathbf{c} \in \mathbb{R}^{N+1} \mid \mathbf{b} = \mathbf{M}^\top_j\mathbf{c}\}$ (see (16.159) for $\mathbf{M}_j$). Now, by $\partial g_{(j,q)}(\mathbb{R}^{N+1}) = \partial\widetilde{\varphi}_q(\mathbb{R}^{N+1})$ (due to (16.160)) and the supercoercivity of φ_q and φ^*_q (due to [9, Example 13.2 and Example 13.8]), for proving the boundedness of $\mathrm{Fix}(\mathbf{T}^{(j,q)}_{\mathrm{DRS_I}})$ from (16.171), it is sufficient to show the following claim:

Claim 16.28 Suppose that $\mathbf{X}$ has no zero column. Let $S \subset \mathbb{R}^p$ be bounded, and $\varphi \in \Gamma_0(\mathbb{R}^N)$ a supercoercive function having supercoercive $\varphi^* \in \Gamma_0(\mathbb{R}^N)$. Then $(\mathbf{M}^\top_j)^{-1}(S) \cap \partial\widetilde{\varphi}(\mathbb{R}^{N+1})$ is bounded.
Note that Claim 16.28 is the main step in the proof of Theorem 16.28.

(b) We have already confirmed the qualification condition (16.161) in Lemma 16.27, $\mathcal{S}_{(j,q)} \neq \varnothing$, and $\Omega^{(j,q)}_{\mathrm{DRS_I}} \neq \varnothing$ $(j = 1, 2, \ldots, 2p)$ (see the short remark just after (16.169)). Therefore, application of Theorem 16.15 (in Section 16.3.1) to the subproblems to compute $\mathbf{b}^{(j)}_{\mathrm{HTREX_q}}$ $(j = 1, 2, \ldots, 2p)$ in (16.169) guarantees the statement of Theorem 16.28(b).

(The proof of Theorem 16.28 is given in Appendix H).

16.5.3 Numerical Experiment: Hierarchical TREX$_2$

We demonstrate that the proposed estimator $\mathbf{b}_{\mathrm{HTREX}_2}$, i.e., Hierarchical TREX$_2$ in (16.168) (see Section 16.5.2) can enhance further the estimation accuracy achieved by $\mathbf{b}_{\mathrm{TREX}_2}$ in (16.156) if we can exploit another new criterion $\Psi : \mathbb{R}^p \to \mathbb{R}$ for promoting characteristics, of $\mathbf{b}^{\mathrm{tru}}$, which is not utilized in TREX$_2$. Consider the situation where we like to estimate unknown vector

$$\mathbf{b}^{\mathrm{tru}} = \frac{1}{\sqrt{p}}(0, 0, 0, 1, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p \tag{16.172}$$

from the noisy observation $\mathbf{z} \in \mathbb{R}^N$ in (16.149). We suppose to know that $\mathbf{b}^{\mathrm{tru}}$ is not only sparse but also *fairly flat*. Here, the fairly flatness of $\mathbf{b}^{\mathrm{tru}}$ means that the energy of oscillations (i.e., the sum of the squared gaps between the adjacent components) of $\mathbf{b}^{\mathrm{tru}}$ is small, which is supposed to be our additional knowledge not utilized in the TREX$_2$ and Lasso estimators. If we have such prior knowledge, suppression of

$$\Psi : \mathbb{R}^p \to \mathbb{R} : \mathbf{b} \mapsto \frac{1}{2}\|\mathbf{D}\mathbf{b}\|^2, \quad \text{with } \mathbf{D} := \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ & & \ddots & \ddots & & \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}, \tag{16.173}$$

is expected to be effective for estimation of $\mathbf{b}^{\mathrm{tru}}$ because Ψ can distinguish $\mathbf{b}^{\mathrm{tru}}$ from $\widetilde{\mathbf{b}} := \frac{1}{\sqrt{p}}(0, 1, 0, 0, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^p$ of the same sparsity as $\mathbf{b}^{\mathrm{tru}}$ (i.e., $\|\mathbf{b}^{\mathrm{tru}}\|_0 = \|\widetilde{\mathbf{b}}\|_0$ and $\|\mathbf{b}^{\mathrm{tru}}\|_1 = \|\widetilde{\mathbf{b}}\|_1$) by $\Psi(\mathbf{b}^{\mathrm{tru}}) < \Psi(\widetilde{\mathbf{b}})$. Now, our new goal for enhancement of TREX$_2$ is to minimize Ψ while keeping the optimality of the TREX$_2$ in the sense of (16.156) for $q = 2$ (see Scenario 2 in Section 16.1). This goal is achieved by solving the hierarchical nonconvex optimization (16.168) for $q = 2$.

In our experiments, the design matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$ in (16.149) is given to satisfy $\mathbf{X}_{:2} = \mathbf{X}_{:3} = \mathbf{X}_{:4}$ with a sample of zero-mean Gaussian random variable followed by normalization to satisfy $\|\mathbf{X}_{:j}\| = \sqrt{N}$ $(j = 1, \ldots, p)$. The additive noise

$\mathbf{e} \in \mathbb{R}^N$ in (16.149) is drawn from the unit white Gaussian distribution. We tested the performances of the estimators under $(\mathrm{SNR}) = 10\log\left(\frac{\|\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{\|\sigma\mathbf{e}\|^2}\right) \in [10, 1000] \cup \{+\infty\}$, where $\sigma \in \mathbb{R}$ is adjusted to obtain a specific SNR. Note that, in this setting, $\left\{\mathbf{b} \in \mathbb{R}^p \mid \mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{b}^{\mathrm{tru}} \text{ and } \|\mathbf{b}\|_1 = \|\mathbf{b}^{\mathrm{tru}}\|_1\right\}$ is apparently an infinite set containing both $\mathbf{b}^{\mathrm{tru}}$ and $\widetilde{\mathbf{b}}$.

We compared the performances of $\mathbf{b}_{\mathrm{TREX}_2}$ in (16.156) ($\beta = 1/2$) and $\mathbf{b}_{\mathrm{HTREX}_2}$ in (16.168) employing Ψ in (16.173). To approximate iteratively $\mathbf{b}^{(j)}_{\mathrm{TREX}_2}$ ($j = 1, 2, \ldots, 2p$) for (16.157) and $\mathbf{b}^{(j)}_{\mathrm{HTREX}_2}$ ($j = 1, 2, \ldots, 2p$) for (16.169), we used respectively TREX_2 (16.167) (Fact 16.26 with $\alpha_n = 1.95$ ($n \in \mathbb{N}$)) and the proposed algorithm (16.170) (HTREX_2 with $\alpha = 1.95$ and $\lambda_n = \frac{1}{n}$ for $n \in \mathbb{N}$). As performance measures, we used, in Figures 16.4 and 16.5,

$$\begin{bmatrix} \textbf{Function Value} \text{ (see (16.168))} & \min_{j=1,\ldots,2p}(g_{(j,2)}(\mathbf{M}_j\mathbf{b}_n) + \|\mathbf{b}_n\|_1), \\ \textbf{Distance} & \|\mathbf{b}_n - \mathbf{b}^{\mathrm{tru}}\|. \end{bmatrix} \tag{16.174}$$

The experiments were performed both in an over-determined case ($N = 30$ and $p = 20$) in Figure 16.4 and an under-determined case ($N = 20$ and $p = 30$) in Figure 16.5.

Figures 16.4(a) and 16.5(a) illustrate the process of convergences of TREX_2 and HTREX_2 in the absence of noise, i.e., $\mathbf{e} = \mathbf{0} \in \mathbb{R}^N$. From these figures, we observe that (i) Function Values of TREX_2 and HTREX_2 converge to the same level, and that (ii) Distance (to $\mathbf{b}^{\mathrm{tru}}$) of HTREX_2 converges to a lower level than that of TREX_2. Figures 16.4(b) and 16.5(b) summarize the behavior of Distance (to $\mathbf{b}^{\mathrm{tru}}$), against various SNR, by TREX_2 and HTREX_2 after 10000 iterations. For all the SNR, HTREX_2 seems to succeed in improving the performance of TREX_2.

16.6 Concluding Remarks

In this paper, we have demonstrated how the modern proximal splitting operators can be plugged nicely into the hybrid steepest descent method (HSDM) for their applications to the hierarchical convex optimization problems which require further strategic selection of a most desirable vector from the set of all solutions of the standard convex optimization. For simplicity as well as for broad applicability, we have chosen to cast our target in the iterative approximation of a viscosity solution of the standard convex optimization problem, where the 1st stage cost function is given as a superposition of multiple nonsmooth convex functions, involving linear operators, while its viscosity solution is a minimizer of the 2nd stage cost function which is Gâteaux differentiable convex function with Lipschitzian gradient. The key ideas for the successful collaboration between the proximal splitting operators and the HSDM are not only in (i) the previously known expressions of the solution set of the standard convex optimization problem as the fixed point set of computable

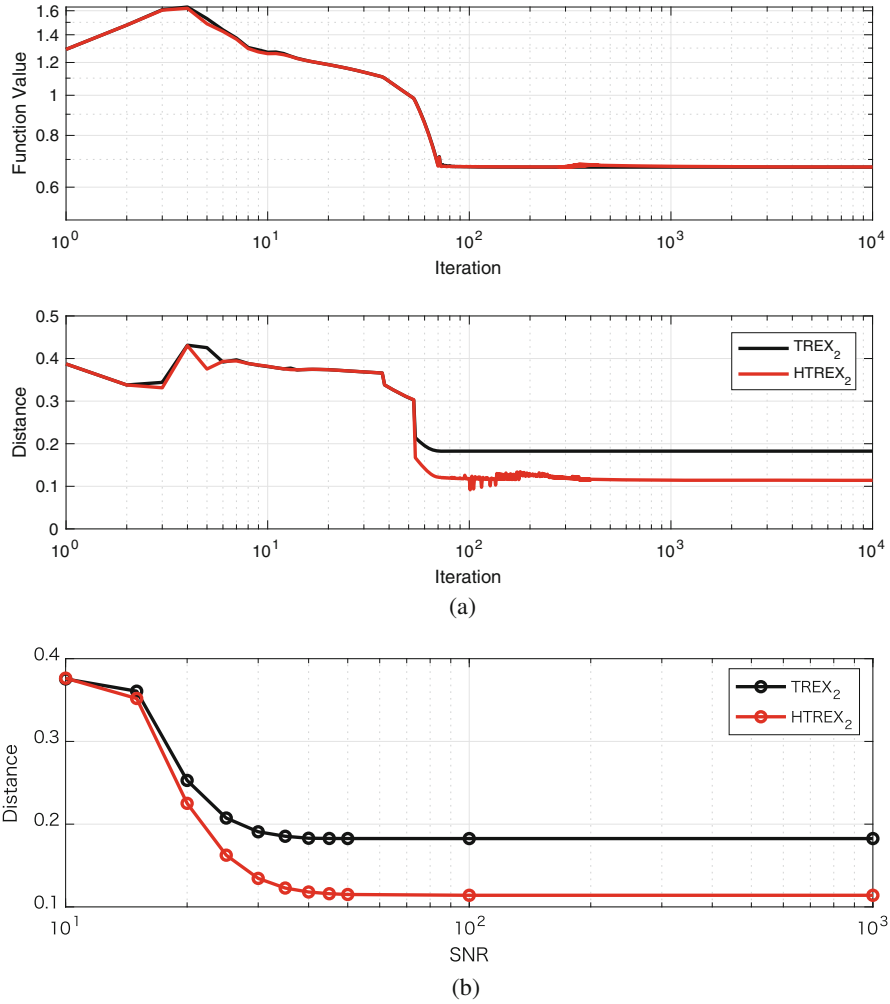


Fig. 16.4 Transient performances in an over-determined case; Criteria "Function Value," "Distance" are given in (16.174) and $(\mathrm{SNR}) = 10\log\left(\frac{\|\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{\|\sigma\mathbf{e}\|^2}\right)$ [dB]. (**a**) Comparison of TREX_2 and HTREX_2 in the process of convergences under the noise $\mathbf{e} = 0$. (**b**) Estimation accuracy achieved by TREX_2 and HTREX_2 for various SNR

nonexpansive operators but also in (ii) linear relations built strategically between the solution set and the fixed point set. Fortunately, we have shown that such key ideas can be achieved by extending carefully the strategies behind the Douglas-Rachford splitting operators as well as the LAL operators defined in certain product Hilbert spaces. We have also presented applications of the proposed algorithmic strategies to certain unexplored hierarchical enhancements of the support vector machine and the Lasso estimator.

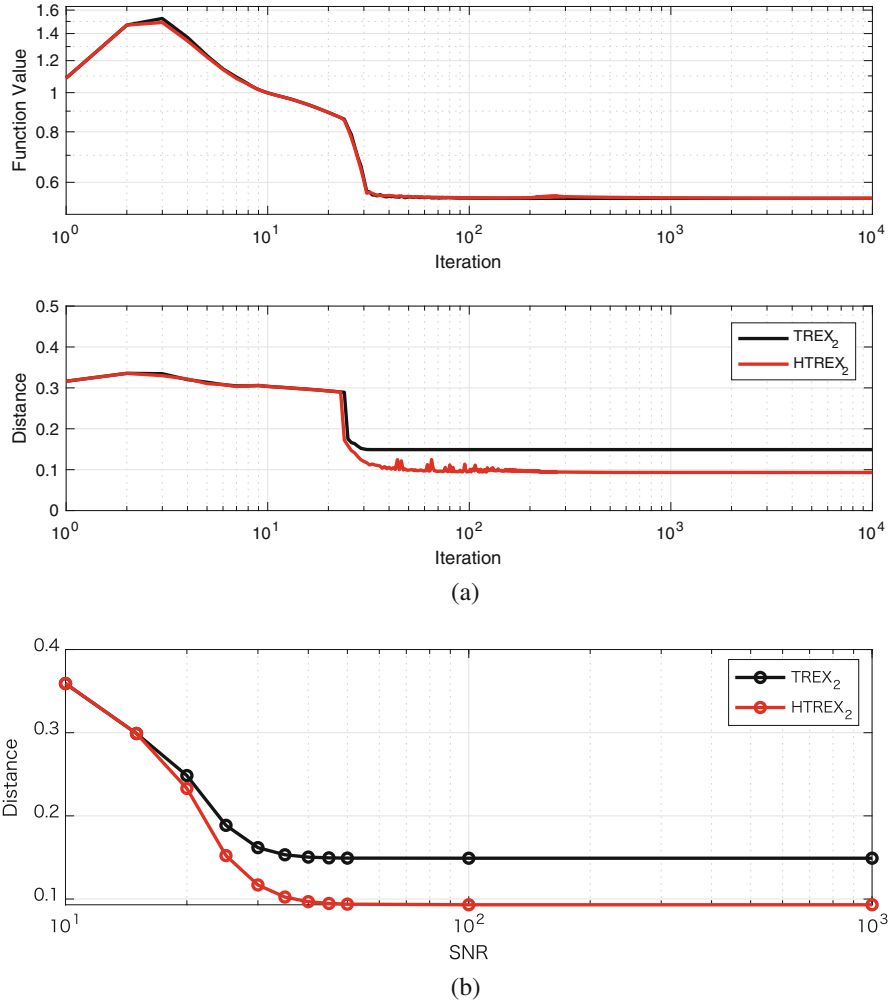


Fig. 16.5 Transient performances in an under-determined case; Criteria "Function Value," "Distance" are given in (16.174) and $(\mathrm{SNR}) = 10\log\left(\frac{\|\mathbf{X}\mathbf{b}^{\mathrm{tru}}\|^2}{\|\sigma\mathbf{e}\|^2}\right)$ [dB]. (**a**) Comparison of TREX_2 and HTREX_2 in the process of convergences under the noise $\mathbf{e} = 0$. (**b**) Estimation accuracy achieved by TREX_2 and HTREX_2 for various SNR

Acknowledgements Isao Yamada would like to thank Heinz H. Bauschke, D. Russell Luke, and Regina S. Burachik for their kind encouragement and invitation of the first author to the dream meeting: *Splitting Algorithms, Modern Operator Theory, and Applications* (September 17–22, 2017) in Oaxaca, Mexico where he had a great opportunity to receive insightful deep comments by Hédy Attouch. He would also like to thank Patrick Louis Combettes and Christian L. Müller for

their invitation of the first author to a special mini-symposium *Proximal Techniques for High-Dimensional Statistics* in the SIAM conference on Optimization 2017 (May 22–25, 2017) in Vancouver. Their kind invitations and their excellent approach to the TREX problem motivated very much the authors to study the application of the proposed strategies to the hierarchical enhancement of Lasso in this paper. Isao Yamada would also like to thank Raymond Honfu Chan for his kind encouragement and invitation to *the Workshop on Optimization in Image Processing* (June 27–30, 2016) at the Harvard University. Lastly, the authors thank to Yunosuke Nakayama for his help in the numerical experiment related to the proposed hierarchical enhancement of the SVM.

Appendices

A: Proof of Proposition 16.9(a)

Fact 16.5(i)⇔(ii) in Section 16.2.1 yields

$$\begin{aligned}(16.53\text{–}16.55) &\Leftrightarrow (\exists \nu_\star \in \mathcal{K})\ \nu_\star \in \partial f(x_\star) \text{ and } -\nu_\star \in \partial g(x_\star)\\ &\Leftrightarrow 0 \in \partial f(x_\star) + \partial g(x_\star).\end{aligned}$$

The remaining follows from the proof in [40, Proposition 18]. □

B: Proof of Proposition 16.10(a)(d)

(a) From (16.58) and (16.59), there exists $(x_\star, \nu_\star) \in \mathcal{S}_{\mathrm{pLAL}} \times \mathcal{S}_{\mathrm{dLAL}}$. Fact 16.5(i)⇔(ii) in Section 16.2.1 yields the equivalence

$$\begin{aligned}&(x_\star, \nu_\star) \in \mathcal{S}_{\mathrm{pLAL}} \times \mathcal{S}_{\mathrm{dLAL}} \text{ and } (16.60)\\ \Leftrightarrow\ & A^*\nu_\star \in \partial f(x_\star) \text{ and } -\nu_\star \in \partial \iota_{\{0\}}(Ax_\star) && (16.175)\\ \Leftrightarrow\ & A^*\nu_\star \in \partial f(x_\star) \text{ and } Ax_\star = 0 && (16.176)\\ \Leftrightarrow\ & x_\star = \mathrm{prox}_f(x_\star - A^*Ax_\star + A^*\nu_\star) \text{ and } \nu_\star = \nu_\star - Ax_\star\\ \Leftrightarrow\ & (x_\star, \nu_\star) \in \mathrm{Fix}(T_{\mathrm{LAL}}). && (16.177)\end{aligned}$$

□

(d) Choose arbitrarily $(\bar{x}, \bar{\nu}) \in \mathrm{Fix}(T_{\mathrm{LAL}})$, i.e.,

$$(\bar{x}, \bar{\nu}) = T_{\mathrm{LAL}}(\bar{x}, \bar{\nu}) = \Big(\mathrm{prox}_f(\bar{x} - A^*A\bar{x} + A^*\bar{\nu}),\ \bar{\nu} - A\bar{x}\Big).$$

Let $(x_n, v_n)_{n\in\mathbb{N}} \subset \mathcal{X}\times\mathcal{K}$ be generated, with any $(x_0, v_0) \in \mathcal{X}\times\mathcal{K}$, by

$$(x_{n+1}, v_{n+1}) = T_{\mathrm{LAL}}(x_n, v_n) = \Big(\mathrm{prox}_f(x_n - A^*Ax_n + A^*v_n),\, v_n - Ax_{n+1}\Big). \tag{16.178}$$

Then [150, (B.3)] yields

$$\begin{aligned}
0 \le{}& \|x_n - \bar{x}\|_{\mathcal{X}}^2 - \|x_{n+1} - \bar{x}\|_{\mathcal{X}}^2 - \|(x_{n+1} - x_n) - (\bar{x} - \bar{x})\|_{\mathcal{X}}^2 \\
&+ \|Ax_{n+1} - A\bar{x} - (Ax_n - A\bar{x})\|_{\mathcal{K}}^2 - \|Ax_n - A\bar{x}\|_{\mathcal{K}}^2 \\
&- \|Ax_{n+1} - A\bar{x} + \bar{v} - v_n\|_{\mathcal{K}}^2 + \|\bar{v} - v_n\|_{\mathcal{K}}^2 \\
={}& \|x_n - \bar{x}\|_{\mathcal{X}}^2 - \|x_{n+1} - \bar{x}\|_{\mathcal{X}}^2 - \|x_{n+1} - x_n\|_{\mathcal{X}}^2 \\
&+ \|Ax_{n+1} - Ax_n\|_{\mathcal{K}}^2 - \|Ax_n\|_{\mathcal{K}}^2 - \|\bar{v} - v_{n+1}\|_{\mathcal{K}}^2 + \|\bar{v} - v_n\|_{\mathcal{K}}^2 \\
\le{}& (\|x_n - \bar{x}\|_{\mathcal{X}}^2 + \|v_n - \bar{v}\|_{\mathcal{K}}^2) - (\|x_{n+1} - \bar{x}\|_{\mathcal{X}}^2 + \|v_{n+1} - \bar{v}\|_{\mathcal{K}}^2) \\
&+ (\|A\|_{\mathrm{op}}^2 - 1)\|x_{n+1} - x_n\|_{\mathcal{X}}^2 - \|Ax_n\|_{\mathcal{K}}^2.
\end{aligned} \tag{16.179}$$

Equation (16.179) and $\|A\|_{\mathrm{op}} < 1$ imply that $(\|x_n - \bar{x}\|_{\mathcal{X}}^2 + \|v_n - \bar{v}\|_{\mathcal{K}}^2)_{n\in\mathbb{N}}$ decreases monotonically, i.e., $(x_n, v_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T_{\mathrm{LAL}})$, and $(\|x_n - \bar{x}\|_{\mathcal{X}}^2 + \|v_n - \bar{v}\|_{\mathcal{K}}^2)_{n\in\mathbb{N}}$ converges to some $c \ge 0$. From this observation, we have

$$\begin{aligned}
&\sum_{n=0}^{N}\Big[(1 - \|A\|_{\mathrm{op}}^2)\|x_{n+1} - x_n\|_{\mathcal{X}}^2 + \|Ax_n\|_{\mathcal{K}}^2\Big] \\
&\le \sum_{n=0}^{N}\Big[(\|x_n - \bar{x}\|_{\mathcal{X}}^2 + \|v_n - \bar{v}\|_{\mathcal{K}}^2) - (\|x_{n+1} - \bar{x}\|_{\mathcal{X}}^2 + \|v_{n+1} - \bar{v}\|_{\mathcal{K}}^2)\Big] \\
&= (\|x_0 - \bar{x}\|_{\mathcal{X}}^2 + \|v_0 - \bar{v}\|_{\mathcal{K}}^2) - (\|x_{N+1} - \bar{x}\|_{\mathcal{X}}^2 + \|v_{N+1} - \bar{v}\|_{\mathcal{K}}^2) \\
&\to (\|x_0 - \bar{x}\|_{\mathcal{X}}^2 + \|v_0 - \bar{v}\|_{\mathcal{K}}^2) - c < \infty \quad (N \to \infty)
\end{aligned}$$

and thus

$$\lim_{n\to\infty}\|x_{n+1} - x_n\|_{\mathcal{X}} = 0 \text{ and } \lim_{n\to\infty}\|Ax_n\|_{\mathcal{K}} = 0. \tag{16.180}$$

By [51, Theorem 9.12], the bounded sequence of $(x_n, v_n)_{n\in\mathbb{N}}$ has some subsequence $(x_{n_j}, v_{n_j})_{j\in\mathbb{N}}$ which converges weakly to some point, say $(x_\star, v_\star)$, in the Hilbert

space $\mathcal{X} \times \mathcal{K}$. Therefore, by applying [9, Theorem 9.1(iii)$\Leftrightarrow$(i)] to $f \in \Gamma_0(\mathcal{X})$, we have

$$f(x_\star) \leq \liminf_{j\to\infty} f(x_{n_j}) \tag{16.181}$$

and, by the Cauchy-Schwarz inequality and (16.180),

$$\begin{aligned}\|Ax_\star\|_{\mathcal{K}}^2 &= \langle Ax_\star - Ax_{n_j}, Ax_\star\rangle_{\mathcal{K}} + \langle Ax_{n_j}, Ax_\star\rangle_{\mathcal{K}}\\ &\leq \langle x_\star - x_{n_j}, A^*Ax_\star\rangle_{\mathcal{X}} + \|Ax_{n_j}\|_{\mathcal{K}}\|Ax_\star\|_{\mathcal{K}} \to 0 \quad (j\to\infty),\end{aligned}$$

which implies $Ax_\star = 0$.

Meanwhile, by (16.178), we have

$$\begin{aligned}x_{n_j} &= \operatorname{prox}_f(x_{n_j-1} - A^*Ax_{n_j-1} + A^*v_n)\\ &= (\mathrm{I} + \partial f)^{-1}(x_{n_j-1} - A^*Ax_{n_j-1} + A^*v_{n_j-1})\\ &\Leftrightarrow x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*v_{n_j-1} \in \partial f(x_{n_j})\\ &\Leftrightarrow (\forall x \in \mathcal{X})\ f(x_{n_j}) + \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*v_{n_j-1}, x - x_{n_j}\rangle_{\mathcal{X}} \leq f(x),\end{aligned} \tag{16.182}$$

where the inner product therein satisfies

$$\lim_{j\to\infty} \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*v_{n_j-1}, x - x_{n_j}\rangle_{\mathcal{X}} = \langle A^*v_\star, x - x_\star\rangle_{\mathcal{X}}, \tag{16.183}$$

which is verified by $Ax_\star = 0$, the triangle inequality, the Cauchy-Schwarz inequality, and (16.180), as follows:

$$\begin{aligned}&(\forall x \in \mathcal{X})\\ &|\langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*v_{n_j-1}, x - x_{n_j}\rangle_{\mathcal{X}} - \langle A^*v_\star, x - x_\star\rangle_{\mathcal{X}}|\\ &= |\langle x_{n_j-1} - x_{n_j}, x - x_{n_j}\rangle_{\mathcal{X}} - \langle Ax_{n_j-1}, A(x - x_{n_j})\rangle_{\mathcal{K}}\\ &\qquad + \langle v_{n_j-1}, A(x - x_{n_j})\rangle_{\mathcal{K}} - \langle v_\star, Ax\rangle_{\mathcal{K}}|\\ &= |\langle x_{n_j-1} - x_{n_j}, x - x_{n_j}\rangle_{\mathcal{X}} - \langle Ax_{n_j-1}, A(x - x_{n_j})\rangle_{\mathcal{K}}\\ &\qquad + \langle v_{n_j-1}, -Ax_{n_j}\rangle_{\mathcal{K}} - \langle v_{n_j} - v_{n_j-1}, Ax\rangle_{\mathcal{K}} - \langle v_\star - v_{n_j}, Ax\rangle_{\mathcal{K}}|\\ &\leq (\|x_{n_j-1} - x_{n_j}\|_{\mathcal{X}}\|x - x_{n_j}\|_{\mathcal{X}} + \|Ax_{n_j-1}\|_{\mathcal{K}}\|A(x - x_{n_j})\|_{\mathcal{K}}\\ &\qquad + \|v_{n_j-1}\|_{\mathcal{K}}\| - Ax_{n_j}\|_{\mathcal{K}} + \|Ax_{n_j}\|_{\mathcal{K}}\|Ax\|_{\mathcal{K}} + |\langle v_\star - v_{n_j}, Ax\rangle_{\mathcal{K}}|)\\ &\to 0 \quad (j\to\infty).\end{aligned}$$

Now, by (16.182), (16.181), and (16.183), we have for any $x \in \mathcal{X}$

$$\begin{aligned} f(x) &\geq f(x_\star) + \liminf_{j\to\infty} \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*v_{n_j-1}, x - x_{n_j}\rangle_{\mathcal{X}} \\ &= f(x_\star) + \lim_{j\to\infty} \langle x_{n_j-1} - x_{n_j} - A^*Ax_{n_j-1} + A^*v_{n_j-1}, x - x_{n_j}\rangle_{\mathcal{X}} \\ &= f(x_\star) + \langle A^*v_\star, x - x_\star\rangle_{\mathcal{X}}, \end{aligned}$$

which implies

$$A^*v_\star \in \partial f(x_\star). \tag{16.184}$$

By recalling (16.176)⇔(16.177), (16.184) and $Ax_\star = 0$ prove $(x_\star, v_\star) \in \mathrm{Fix}(T_{\mathrm{LAL}})$. The above discussion implies that every weak sequential cluster point (see Footnote 7 in Section 16.2.2) of $(x_n, v_n)_{n\in\mathbb{N}}$, which is Fejér monotone with respect to $\mathrm{Fix}(T_{\mathrm{LAL}})$, belongs to $\mathrm{Fix}(T_{\mathrm{LAL}})$. Therefore, [9, Theorem 5.5] guarantees that $(x_n, v_n)_{n\in\mathbb{N}}$ converges weakly to a point in $\mathrm{Fix}(T_{\mathrm{LAL}})$. □

C: Proof of Theorem 16.15

Now by recalling Proposition 16.9 in Section 16.2.3 and Remark 16.16 in Section 16.3.1, it is sufficient to prove Claim 16.15. Let $x_\star \in \mathcal{S}_p \neq \varnothing$. Then the Fermat's rule, Fact 16.4(b) (applicable due to the qualification condition (16.40)) in Section 16.2.1, $\check{A}^*\colon \mathcal{K} \to \mathcal{X}\times\mathcal{K}\colon v \mapsto (A^*v, -v)$ for $\check{A}$ in (16.74), the property of $\iota_{\{0\}}$ in (16.35), the straightforward calculations, and Fact 16.5(ii)⇔(i) (in Section 16.2.1) yield

$$\begin{aligned} x_\star \in \mathcal{S}_p &\Leftrightarrow 0 \in \partial(f + g\circ A)(x_\star) = \partial f(x_\star) + A^*\partial g(Ax_\star) \\ &\Leftrightarrow\ y_\star = Ax_\star \text{ and } 0 \in \partial f(x_\star) + A^*\partial g(y_\star) \\ &\Leftrightarrow (\exists v_\star \in \mathcal{K})\ y_\star = Ax_\star \text{ and } \begin{cases} A^*v_\star \in \partial f(x_\star) \\ -v_\star \in \partial g(y_\star) \end{cases} \\ &\Leftrightarrow (\exists v_\star \in \mathcal{K})\ \check{A}(x_\star, y_\star) = 0 \text{ and } \check{A}^*v_\star \in \partial F(x_\star, y_\star) \\ &\Leftrightarrow (\exists v_\star \in \mathcal{K})\ -v_\star \in \partial\iota_{\{0\}}(\check{A}(x_\star, y_\star)) \text{ and } \check{A}^*v_\star \in \partial F(x_\star, y_\star) \\ &\Rightarrow (\exists v_\star \in \mathcal{K})\ -\check{A}^*v_\star \in \check{A}^*\partial\iota_{\{0\}}(\check{A}(x_\star, y_\star)) \text{ and } \check{A}^*v_\star \in \partial F(x_\star, y_\star) \\ &\Rightarrow (\exists v_\star \in \mathcal{K})\ -\check{A}^*v_\star \in \partial(\iota_{\{0\}}\circ\check{A})(x_\star, y_\star) \text{ and } \check{A}^*v_\star \in \partial F(x_\star, y_\star) \\ &\Leftrightarrow (\exists v_\star \in \mathcal{K})\ -\check{A}^*v_\star \in \partial\iota_{\mathcal{N}(\check{A})}(x_\star, y_\star) \text{ and } \check{A}^*v_\star \in \partial F(x_\star, y_\star) \end{aligned}$$

$$\Leftrightarrow (\exists \nu_\star \in \mathcal{K}) \begin{cases} (x_\star, y_\star) \in \operatorname{argmin}(F + \iota_{\mathcal{N}(\check{A})})(\mathcal{X} \times \mathcal{K}) \\ \check{A}^* \nu_\star \in \operatorname{argmin}(F^* + \iota^*_{\mathcal{N}(\check{A})} \circ (-\mathrm{I}))(\mathcal{X} \times \mathcal{K}) \\ \min(F + \iota_{\mathcal{N}(\check{A})})(\mathcal{X} \times \mathcal{K}) = -\min(F^* + \iota^*_{\mathcal{N}(\check{A})} \circ (-\mathrm{I}))(\mathcal{X} \times \mathcal{K}), \end{cases}$$

which confirms Claim 16.15. □

D: Proof of Theorem 16.17

Now by recalling Proposition 16.9 in Section 16.2.3 and Remark 16.18 in Section 16.3.1, it is sufficient to prove (16.97) by verifying Claim 16.17. We will use

$$A^* \circ \partial g \circ A = \sum_{i=1}^m A_i^* \circ \partial g_i \circ A_i = \sum_{i=1}^m \partial(g_i \circ A_i) \tag{16.185}$$

which is verified by $g = \bigoplus_{i=1}^m g_i$, Fact 16.4(c) (see Section 16.2.1), and $\operatorname{ri}(\operatorname{dom}(g_j) - \operatorname{ran}(A_j)) = \operatorname{ri}(\operatorname{dom}(g_j) - \mathbb{R}) = \mathbb{R} \ni 0$ $(j = 1, 2, \ldots, m)$. Let $x_\star^{(m+1)} \in \mathcal{S}_p \neq \varnothing$. Then by using the Fermat's rule, Fact 16.4(b) (applicable due to (16.40)), (16.185), D in (16.93), and H in (16.92), we deduce the equivalence

$$\begin{aligned} & x_\star^{(m+1)} \in \mathcal{S}_p \\ \Leftrightarrow\ & 0 \in \partial(f + g \circ A)(x_\star^{(m+1)}) = \partial f(x_\star^{(m+1)}) + A^* \partial g(A x_\star^{(m+1)}) \\ & \qquad = \partial f(x_\star^{(m+1)}) + \sum_{i=1}^m \partial(g_i \circ A_i)(x_\star^{(m+1)}) \\ \Leftrightarrow\ & (j = 1, \ldots, m)\ x_\star^{(j)} = x_\star^{(m+1)} \text{ and } 0 \in \partial f(x_\star^{(m+1)}) + \sum_{i=1}^m \partial(g_i \circ A_i)(x_\star^{(i)}) \\ \Leftrightarrow\ & (\exists \nu^{(1)}, \ldots, \nu^{(m)} \in \mathcal{X})(j = 1, \ldots, m) \begin{cases} x_\star^{(j)} = x_\star^{(m+1)} \\ \nu^{(j)} \in \partial(g_j \circ A_j)(x_\star^{(j)}) \\ -\sum_{i=1}^m \nu^{(i)} \in \partial f(x_\star^{(m+1)}) \end{cases} \\ \Leftrightarrow\ & (\exists \nu^{(1)}, \ldots, \nu^{(m)} \in \mathcal{X}) \\ & \begin{cases} (x_\star^{(1)}, \ldots, x_\star^{(m+1)}) \in D \\ \left(\nu^{(1)}, \ldots, \nu^{(m)}, -\sum_{i=1}^m \nu^{(i)}\right) \in \left[\bigtimes_{j=1}^m \partial(g_j \circ A_j)(x_\star^{(j)})\right] \times \partial f(x_\star^{(m+1)}) \\ \qquad\qquad = \partial H(x_\star^{(1)}, \ldots, x_\star^{(m+1)}). \end{cases} \end{aligned} \tag{16.186}$$

Then by $-\big(\nu^{(1)},\ldots,\nu^{(m)},-\sum_{i=1}^m \nu^{(i)}\big) \in D^\perp = \partial\iota_D(x_\star^{(1)},\ldots,x_\star^{(m+1)})$ (see (16.34)) and by Fact 16.5(ii)⇔(i) in Section 16.2.1, we have

$$
\begin{aligned}
x_\star^{(m+1)} \in \mathcal{S}_p \;&\Leftrightarrow (\exists \nu^{(1)},\ldots,\nu^{(m)} \in \mathcal{X})\\
&\qquad \begin{cases} -\big(\nu^{(1)},\ldots,\nu^{(m)},-\sum_{i=1}^m \nu^{(i)}\big) \in \partial\iota_D(x_\star^{(1)},\ldots,x_\star^{(m+1)})\\ \big(\nu^{(1)},\ldots,\nu^{(m)},-\sum_{i=1}^m \nu^{(i)}\big) \in \partial H(x_\star^{(1)},\ldots,x_\star^{(m+1)}) \end{cases}\\
&\Leftrightarrow (\exists \nu^{(1)},\ldots,\nu^{(m)} \in \mathcal{X})\\
&\qquad \begin{cases} (x_\star^{(1)},\ldots,x_\star^{(m+1)}) \in \operatorname{argmin}(H+\iota_D)(\mathcal{X}^{m+1})\\ \big(\nu^{(1)},\ldots,\nu^{(m)},-\sum_{i=1}^m \nu^{(i)}\big) \in \operatorname{argmin}(H^*+\iota_D^*\circ(-\mathrm{I}))(\mathcal{X}^{m+1})\\ \min(H+\iota_D)(\mathcal{X}^{m+1}) = -\min(H^*+\iota_D^*\circ(-\mathrm{I}))(\mathcal{X}^{m+1}), \end{cases}
\end{aligned}
$$

which confirms Claim 16.17. □

E: Proof of Theorem 16.19

Now by recalling Proposition 16.10 in Section 16.2.3 and Remark 16.20 in Section 16.3.2, it is sufficient to prove Claim 16.19. Let $x_\star \in \mathcal{S}_p \neq \varnothing$. Then the Fermat's rule, Fact 16.4(b) (applicable due to (16.40)) in Section 16.2.1, $\check{A}^*\colon \mathcal{K} \to \mathcal{X}\times\mathcal{K}\colon \nu \mapsto (A^*\nu,-\nu)$ for $\check{A}$ in (16.74), the property of $\iota_{\{0\}}$ in (16.35), the straightforward calculations, and Fact 16.5(ii)⇔(i) (in Section 16.2.1) yield

$$
\begin{aligned}
x_\star \in \mathcal{S}_p &\Leftrightarrow 0 \in \partial(f+g\circ A)(x_\star) = \partial f(x_\star) + A^*\partial g(Ax_\star)\\
&\Leftrightarrow\ y_\star = Ax_\star \text{ and } 0 \in \partial f(x_\star) + A^*\partial g(y_\star)\\
&\Leftrightarrow (\exists \nu_\star \in \mathcal{K})\ y_\star = Ax_\star \text{ and } \begin{cases} \mathfrak{u}A^*\nu_\star \in \partial f(x_\star)\\ -\mathfrak{u}\nu_\star \in \partial g(y_\star) \end{cases}\\
&\Leftrightarrow (\exists \nu_\star \in \mathcal{K})\ (\mathfrak{u}\check{A})(x_\star,y_\star) = 0 \text{ and } (\mathfrak{u}\check{A})^*\nu_\star \in \partial F(x_\star,y_\star)\\
&\Leftrightarrow (\exists \nu_\star \in \mathcal{K})\ -\nu_\star \in \partial\iota_{\{0\}}((\mathfrak{u}\check{A})(x_\star,y_\star)) \text{ and } (\mathfrak{u}\check{A})^*\nu_\star \in \partial F(x_\star,y_\star)\\
&\Leftrightarrow (\exists \nu_\star \in \mathcal{K})\\
&\qquad \begin{cases} (x_\star,y_\star) \in \operatorname{argmin}(F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A}))(\mathcal{X}\times\mathcal{K})\\ \nu_\star \in \operatorname{argmin}(F^*\circ(\mathfrak{u}\check{A})^*)(\mathcal{K})\\ \min(F+\iota_{\{0\}}\circ(\mathfrak{u}\check{A}))(\mathcal{X}\times\mathcal{K}) = -\min(F^*\circ(\mathfrak{u}\check{A})^*)(\mathcal{K}), \end{cases}
\end{aligned}
$$

which confirms Claim 16.19. □

F: Proof of Theorem 16.23

(a) We have seen in (16.78) that, under the assumptions of Theorem 16.23(a), for any vector $x_\star \in \mathcal{X}$,

$$x_\star \in \mathcal{S}_p[\text{in (16.13)}] \text{ if and only if } (x_\star, y_\star) = P_{\mathcal{N}(\check{A})}(\zeta_\star) \tag{16.187}$$

for some $y_\star \in \mathcal{X}$ and some $\zeta_\star \in \operatorname{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right)$, where $\check{A}\colon \mathcal{X}\times\mathcal{K} \to \mathcal{K}\colon (x, y) \mapsto Ax - y$ (see (16.74)), $\mathcal{N}(\check{A}) = \{(x, Ax) \in \mathcal{X} \times \mathcal{K} \mid x \in \mathcal{X}\}$, and $\mathbf{T}_{\mathrm{DRS_I}} = (2\operatorname{prox}_F - \mathrm{I}) \circ (2P_{\mathcal{N}(\check{A})} - \mathrm{I})$ for $F\colon \mathcal{X}\times\mathcal{K} \to (-\infty, \infty]\colon (x, y) \mapsto f(x)+g(y)$ (see (16.71) and (16.73)).
Choose $\zeta_\star := (\zeta_\star^x, \zeta_\star^y) \in \operatorname{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right)$ arbitrarily and let $\mathbf{z}_\star := (x_\star, y_\star) := P_{\mathcal{N}(\check{A})}(\zeta_\star)$. Then we have

$$\begin{aligned}
&\zeta_\star \in \operatorname{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right) \text{ and } P_{\mathcal{N}(\check{A})}(\zeta_\star) = \mathbf{z}_\star \\
&\Leftrightarrow (2\operatorname{prox}_F - \mathrm{I}) \circ (2P_{\mathcal{N}(\check{A})} - \mathrm{I})(\zeta_\star) = \zeta_\star \text{ and } P_{\mathcal{N}(\check{A})}(\zeta_\star) = \mathbf{z}_\star && (16.188)\\
&\Rightarrow (2\operatorname{prox}_F - \mathrm{I})(2\mathbf{z}_\star - \zeta_\star) = \zeta_\star \ \Leftrightarrow\ \operatorname{prox}_F(2\mathbf{z}_\star - \zeta_\star) = \mathbf{z}_\star \\
&\Leftrightarrow (\mathrm{I} + \partial F)^{-1}(2\mathbf{z}_\star - \zeta_\star) = \mathbf{z}_\star \ \Leftrightarrow\ 2\mathbf{z}_\star - \zeta_\star \in \mathbf{z}_\star + \partial F(\mathbf{z}_\star) \\
&\Leftrightarrow \mathbf{z}_\star - \zeta_\star \in \partial F(\mathbf{z}_\star) = \partial f(x_\star) \times \partial g(y_\star) && (16.189)\\
&\Leftrightarrow x_\star - \zeta_\star^x \in \partial f(x_\star) \text{ and } y_\star - \zeta_\star^y \in \partial g(y_\star). && (16.190)
\end{aligned}$$

Meanwhile, we have

$$\begin{aligned}
\mathbf{z}_\star = P_{\mathcal{N}(\check{A})}(\zeta_\star) &\Leftrightarrow (\forall \mathbf{z} = (x, Ax) \in \mathcal{N}(\check{A}))\ \langle \zeta_\star - \mathbf{z}_\star, \mathbf{z}\rangle_{\mathcal{X}\times\mathcal{K}} = 0 \\
&\Leftrightarrow (\forall x \in \mathcal{X})\ \left\langle \zeta_\star^x - x_\star, x\right\rangle_{\mathcal{X}} + \left\langle \zeta_\star^y - y_\star, Ax\right\rangle_{\mathcal{K}} = 0 \\
&\Leftrightarrow (\forall x \in \mathcal{X})\ \left\langle (\zeta_\star^x - x_\star) + A^*(\zeta_\star^y - y_\star), x\right\rangle_{\mathcal{X}} = 0 \\
&\Leftrightarrow A^*(\zeta_\star^y - y_\star) = -(\zeta_\star^x - x_\star). && (16.191)
\end{aligned}$$

Equations (16.191) and (16.190) imply

$$\begin{aligned}
&\zeta_\star \in \operatorname{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right) \text{ and } P_{\mathcal{N}(\check{A})}(\zeta_\star) = \mathbf{z}_\star \\
&\Rightarrow x_\star - \zeta_\star^x \in \partial f(x_\star) \text{ and } y_\star - \zeta_\star^y \in (-(A^*)^{-1}(\partial f(x_\star))) \cap \partial g(y_\star) \\
&\Rightarrow \zeta_\star = (\zeta_\star^x, \zeta_\star^y) \in (x_\star, y_\star) - \left(\partial f(x_\star) \times [(-(A^*)^{-1}(\partial f(x_\star))) \cap \partial g(y_\star)]\right). && (16.192)
\end{aligned}$$

Moreover, by noting that (16.187) ensures $x_\star \in \mathcal{S}_p$ and $y_\star = Ax_\star$, we have from (16.192)

$$\zeta_\star \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right) \text{ and } (x_\star, Ax_\star) = P_{\mathcal{N}(\check{A})}(\zeta_\star)$$

$$\Rightarrow \zeta_\star \in (x_\star, Ax_\star) - \Big(\partial f(x_\star) \times [(-(A^*)^{-1}(\partial f(x_\star))) \cap \partial g(Ax_\star)]\Big)$$

$$\Rightarrow \zeta_\star \in \bigcup_{x' \in \mathcal{S}_p} (x', Ax') - \bigcup_{x'' \in \mathcal{S}_p} \Big(\partial f(x'') \times [(-(A^*)^{-1}(\partial f(x''))) \cap \partial g(Ax'')]\Big)$$

Since $\zeta_\star$ is chosen arbitrarily from $\mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right)$, we have

$$\mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_I}}\right) \subset \bigcup_{x' \in \mathcal{S}_p} (x', Ax') - \bigcup_{x'' \in \mathcal{S}_p} \big(\partial f(x'') \times [(-(A^*)^{-1}(\partial f(x''))) \cap \partial g(Ax'')]\Big), \tag{16.193}$$

from which Theorem 16.23(a) is confirmed.

(b) We have seen in (16.113) that, under the assumptions of Theorem 16.23(b), for any vector $x_\star \in \mathcal{X}$,

$$x_\star \in \mathcal{S}_p[\text{in (16.13)}] \text{ if and only if } (x_\star, y_\star, v_\star) \in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}}) \tag{16.194}$$

for some $(y_\star, v_\star) \in \mathcal{K} \times \mathcal{K}$, where

$$\mathbf{T}_{\mathrm{LAL}} \colon \mathcal{X} \times \mathcal{K} \times \mathcal{K} \to \mathcal{X} \times \mathcal{K} \times \mathcal{K}$$

$$: \begin{pmatrix} x \\ y \\ v \end{pmatrix} = \begin{pmatrix} \mathbf{z} \\ v \end{pmatrix} \mapsto \begin{pmatrix} x_T \\ y_T \\ v_T \end{pmatrix} = \begin{pmatrix} \mathbf{z}_T \\ v_T \end{pmatrix} = \begin{pmatrix} \mathrm{prox}_F(\mathbf{z} - (\mathrm{u}\check{A})^*(\mathrm{u}\check{A})\mathbf{z} + (\mathrm{u}\check{A})^* v) \\ v - \mathrm{u}\check{A}\mathbf{z}_T \end{pmatrix}$$

and $(\mathrm{u}\check{A})^* \colon \mathcal{K} \to \mathcal{X} \times \mathcal{K} \colon v \mapsto (\mathrm{u}A^* v, -\mathrm{u}v)$ (see (16.108) and (16.120)). Choose $(\mathbf{z}_\star, v_\star) \in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})$ arbitrarily and denote $\mathbf{z}_\star = (x_\star, y_\star) \in \mathcal{X} \times \mathcal{K}$. By passing similar steps in (16.177)⇔(16.176), we deduce

$$(\mathbf{z}_\star, v_\star) \in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})$$

$$\Leftrightarrow (\mathrm{u}\check{A})^* v_\star \in \partial F(\mathbf{z}_\star) = \partial f(x_\star) \times \partial g(y_\star) \text{ and } \mathrm{u}\check{A}(\mathbf{z}_\star) = 0, \tag{16.195}$$

and then, from (16.195), straightforward calculations yield

$$(x_\star, y_\star, v_\star) \in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}}) \Leftrightarrow \begin{bmatrix} \mathrm{u}A^* v_\star = A^*(\mathrm{u}v_\star) \in \partial f(x_\star) \\ -\mathrm{u}v_\star \in \partial g(y_\star) \\ Ax_\star = y_\star \end{bmatrix}$$

$$\Rightarrow -\mathrm{u}v_\star \in \Big[-(A^*)^{-1}(\partial f(x_\star))\Big] \cap \partial g(Ax_\star) \text{ and } Ax_\star = y_\star$$

$$\Leftrightarrow -\mathrm{u}(x_\star, y_\star, v_\star) \in \{-\mathrm{u}(x_\star, Ax_\star)\} \times [-(A^*)^{-1}(\partial f(x_\star)) \cap \partial g(Ax_\star)]. \tag{16.196}$$

Moreover, by noting that (16.194), we have from (16.196)

$$\begin{aligned}&(x_\star, y_\star, v_\star) \in \mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})\\ \Rightarrow\ &-\mathfrak{u}(x_\star, y_\star, v_\star) \in \bigcup_{x\in\mathcal{S}_p}\{-\mathfrak{u}(x, Ax)\}\times[-(A^*)^{-1}(\partial f(x))\cap\partial g(Ax)].\end{aligned}$$

Since $(x_\star, y_\star, v_\star)$ is chosen arbitrarily from $\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}})$, we have

$$-\mathfrak{u}\,\mathrm{Fix}(\mathbf{T}_{\mathrm{LAL}}) \subset \bigcup_{x\in\mathcal{S}_p}\{-\mathfrak{u}(x, Ax)\}\times\Big[-(A^*)^{-1}(\partial f(x))\cap\partial g(Ax)\Big]$$

from which Theorem 16.23(b) is confirmed.

(c) We have seen in (16.98) that, under the assumptions of Theorem 16.23(c), for any vector $x_\star \in \mathcal{X}$,

$$x_\star \in \mathcal{S}_p[\text{in (16.13)}] \text{ if and only if } (x_\star, x_\star, \ldots, x_\star) = P_D(\mathfrak{X}_\star) \tag{16.197}$$

for some $\mathfrak{X}_\star \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_{II}}}\right)$, where $D = \{(x^{(1)}, \ldots, x^{(m+1)}) \in \mathcal{X}^{m+1} \mid x^{(i)} = x^{(j)}\ (i, j = 1, 2, \ldots, m+1)\}$ (see (16.93)), $H\colon \mathcal{X}^{m+1} \to (-\infty, \infty]\colon (x^{(1)}, \ldots, x^{(m+1)}) \mapsto \sum_{i=1}^m g_i(A_i x^{(i)}) + f(x^{(m+1)})$ (see (16.92)), and $\mathbf{T}_{\mathrm{DRS_{II}}} = (2\,\mathrm{prox}_H - \mathrm{I}) \circ (2P_D - \mathrm{I})$ (see (16.90)) [For the availability of prox_H and P_D as computational tools, see Remark 16.18(a)].

Choose $\mathfrak{X}_\star := (\zeta_\star^{(1)}, \ldots, \zeta_\star^{(m+1)}) \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_{II}}}\right)$ arbitrarily, and let $\mathbf{X}_\star := (x_\star, \ldots, x_\star) = P_D(\mathfrak{X}_\star)$. Then we have

$$\begin{aligned}&\mathfrak{X}_\star \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_{II}}}\right) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star\\ \Leftrightarrow\ &(2\,\mathrm{prox}_H - \mathrm{I}) \circ (2P_D - \mathrm{I})(\mathfrak{X}_\star) = \mathfrak{X}_\star \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star.\end{aligned} \tag{16.198}$$

Now, by passing similar steps for (16.188)⇒(16.189), we deduce that

$$\begin{aligned}&\mathfrak{X}_\star \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS_{II}}}\right) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star\\ \Rightarrow\ &\mathbf{X}_\star - \mathfrak{X}_\star \in \partial H(\mathbf{X}_\star) = \left[\bigtimes_{j=1}^m \partial(g_j \circ A_j)(x_\star)\right] \times \partial f(x_\star)\\ \Leftrightarrow\ &(j = 1, 2, \ldots, m)\ x_\star - \zeta_\star^{(j)} \in \partial(g_j \circ A_j)(x_\star) \text{ and } x_\star - \zeta_\star^{(m+1)} \in \partial f(x_\star)\\ \Leftrightarrow\ &(j = 1, 2, \ldots, m)\ x_\star - \zeta_\star^{(i)} \in A_j^* \partial g_j(A_j x_\star) \text{ and } x_\star - \zeta_\star^{(m+1)} \in \partial f(x_\star),\end{aligned} \tag{16.199}$$

where the last equivalence follows from Fact 16.4(c) (applicable due to $\mathrm{ri}(\mathrm{dom}(g_j) - \mathrm{ran}(A_j)) = \mathrm{ri}(\mathrm{dom}(g_j) - \mathbb{R}) = \mathbb{R} \ni 0$). Meanwhile, we have

$$\mathbf{X}_\star = P_D(\mathfrak{X}_\star) \Leftrightarrow x_\star = \frac{1}{m+1}\sum_{i=1}^{m+1}\zeta_\star^{(i)} \Leftrightarrow x_\star - \zeta_\star^{(m+1)} = -\sum_{i=1}^{m}(x_\star - \zeta_\star^{(i)}). \tag{16.200}$$

Equations (16.200) and (16.199) imply

$$\begin{aligned}
&\mathfrak{X}_\star \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}\right) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star \\
\Rightarrow &\begin{cases} (j = 1, 2, \ldots, m)\ x_\star - \zeta_\star^{(j)} \in A_j^* \partial g_j(A_j x_\star) \\ x_\star - \zeta_\star^{(m+1)} \in \partial f(x_\star) \cap \left[-\sum_{i=1}^m A_i^* \partial g_i(A_i x_\star)\right] \end{cases} \\
\Rightarrow &\mathbf{X}_\star - \mathfrak{X}_\star \in \left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_j x_\star)\right] \times \left[\partial f(x_\star) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_i x_\star)\right)\right].
\end{aligned} \tag{16.201}$$

Moreover, by noting that (16.197) ensures $x_\star \in \mathcal{S}_p$, we have from (16.201)

$$\begin{aligned}
&\mathfrak{X}_\star \in \mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}\right) \text{ and } P_D(\mathfrak{X}_\star) = \mathbf{X}_\star = (x_\star, \ldots, x_\star) \\
\Rightarrow &\mathfrak{X}_\star \in \mathcal{S}_p^{m+1} - \bigcup_{x \in \mathcal{S}_p} \left(\left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_j x)\right] \times \left[\partial f(x) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_i x)\right)\right]\right).
\end{aligned}$$

Since $\mathfrak{X}_\star$ is chosen arbitrarily from $\mathrm{Fix}(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}})$, we have

$$\begin{aligned}
\mathrm{Fix}\left(\mathbf{T}_{\mathrm{DRS}_{\mathrm{II}}}\right) \subset\ &\mathcal{S}_p^{m+1} \\
&- \bigcup_{x \in \mathcal{S}_p} \left(\left[\bigtimes_{j=1}^m A_j^* \partial g_j(A_j x)\right] \times \left[\partial f(x) \cap \left(-\sum_{i=1}^m A_i^* \partial g_i(A_i x)\right)\right]\right),
\end{aligned} \tag{16.202}$$

from which Theorem 16.23(c) is confirmed. □

G: Proof of Lemma 16.27

Obviously, we have from (16.158)

$$(j = 1, 2, \ldots, 2p) \quad \mathrm{dom}(g_{(j,q)}) \supset \{\eta \in \mathbb{R} \mid \eta > \mathbf{x}_j^\top \mathbf{z}\} \times \mathbb{R}^N. \tag{16.203}$$

By recalling $0 \neq \mathbf{x}_j \in \mathbb{R}^N$ in (16.153) and $\mathbf{M}_j \in \mathbb{R}^{(N+1)\times p}$ in (16.159), we have

$$(j = 1, 2, \ldots, 2p) \begin{bmatrix} \|\mathbf{X}^\top \mathbf{x}_j\| \geq \|\mathbf{x}_j\|^2 > 0 \\ t(\|\mathbf{X}^\top \mathbf{x}_j\|)^{-2} \mathbf{M}_j \mathbf{X}^\top \mathbf{x}_j = \begin{pmatrix} t \\ t \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} \quad (\forall t \in \mathbb{R}), \end{bmatrix}$$

and therefore

$$(j = 1, 2, \ldots, 2p) \quad \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1) = \mathbf{M}_j(\mathbb{R}^p) \supset \operatorname{span} \begin{pmatrix} 1 \\ \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix}. \tag{16.204}$$

To prove $\operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1) = \mathbb{R} \times \mathbb{R}^N$, choose arbitrarily $(\eta, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^N$. Then (16.203) and (16.204) guarantee

$$\begin{aligned} \begin{pmatrix} \eta \\ \mathbf{y} \end{pmatrix} &= \begin{pmatrix} \mathbf{x}_j^\top \mathbf{z} + 1 \\ \mathbf{y} + (\mathbf{x}_j^\top \mathbf{z} + 1 - \eta) \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} - \begin{pmatrix} \mathbf{x}_j^\top \mathbf{z} + 1 - \eta \\ (\mathbf{x}_j^\top \mathbf{z} + 1 - \eta) \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} \\ &\in \{\tilde{\eta} \in \mathbb{R} \mid \tilde{\eta} > \mathbf{x}_j^\top \mathbf{z}\} \times \mathbb{R}^N - \operatorname{span} \begin{pmatrix} 1 \\ \frac{\mathbf{X}\mathbf{X}^\top \mathbf{x}_j}{\|\mathbf{X}^\top \mathbf{x}_j\|^2} \end{pmatrix} \\ &\subset \operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1), \end{aligned}$$

implying thus

$$\operatorname{ri}(\operatorname{dom}(g_{(j,q)}) - \mathbf{M}_j \operatorname{dom}(\|\cdot\|_1)) = \operatorname{ri}(\mathbb{R} \times \mathbb{R}^N) = \mathbb{R} \times \mathbb{R}^N \ni 0. \tag{16.205}$$

□

H: Proof of Theorem 16.28

By recalling Remark 16.29 in Section 16.5.2, it is sufficient to prove Claim 16.28, for which we use the following inequality: for each $j = 1, 2, \ldots, 2p$,

$$(\forall (\eta, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^N) \quad \left\| \mathbf{M}_j^\top \begin{pmatrix} \eta \\ \mathbf{y} \end{pmatrix} \right\| \geq \left| \eta \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{y} \rangle \right|, \tag{16.206}$$

where $\mathbf{x}_j \in \mathbb{R}^N$ in (16.153) and $\mathbf{M}_j \in \mathbb{R}^{(N+1)\times p}$ in (16.159). Equation (16.206) is confirmed by

$$(j = 1, 2, \ldots, 2p)(\forall (\eta, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^N) \quad \mathbf{M}_j^\top \begin{pmatrix} \eta \\ \mathbf{y} \end{pmatrix} = \eta \mathbf{X}^\top \mathbf{x}_j + \mathbf{X}^\top \mathbf{y} \tag{16.207}$$

and

$$\begin{cases} \left[\eta \mathbf{X}^\top \mathbf{x}_j + \mathbf{X}^\top \mathbf{y}\right]_j = \eta \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{y} \rangle & \text{if } j \in \{1, 2, \ldots, p\} \\ \left[\eta \mathbf{X}^\top \mathbf{x}_j + \mathbf{X}^\top \mathbf{y}\right]_{j-p} = -\eta \|\mathbf{x}_j\|^2 - \langle \mathbf{x}_j, \mathbf{y} \rangle & \text{if } j \in \{p+1, p+2, \ldots, 2p\}. \end{cases}$$

Let $U_S := \sup\{\|\mathbf{b}\| \mid \mathbf{b} \in S\} (< \infty)$. By supercoercivity of φ and Example 16.3, the subdifferential of its perspective $\widetilde{\varphi}$ at each $(\eta, \mathbf{y}) \in \mathbb{R} \times \mathbb{R}^N$ can be expressed as (16.32), and thus, to prove Claim 16.28, it is sufficient to show

(i) $(\mathbf{M}_j^\top)^{-1}(S) \cap \partial \widetilde{\varphi}(\mathbb{R}_{++} \times \mathbb{R}^N)$ is bounded;

(ii) $(\mathbf{M}_j^\top)^{-1}(S) \cap \partial \widetilde{\varphi}(0, 0)$ is bounded.

Proof of (i) Choose $(\eta, \mathbf{y}) \in \mathbb{R}_{++} \times \mathbb{R}^N$ arbitrarily. Then, from (16.32), every $\mathbf{c}_{(\eta,\mathbf{y})} \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial \widetilde{\varphi}(\eta, \mathbf{y}) \subset \mathbb{R} \times \mathbb{R}^N$ can be expressed with some $\mathbf{u} \in \partial \varphi(\mathbf{y}/\eta)$ as

$$\mathbf{c}_{(\eta,\mathbf{y})} = (\varphi(\mathbf{y}/\eta) - \langle \mathbf{y}/\eta, \mathbf{u} \rangle, \mathbf{u}) = (-\varphi^*(\mathbf{u}), \mathbf{u}), \tag{16.208}$$

where the last equality follows from $\varphi(\mathbf{y}/\eta) + \varphi^*(\mathbf{u}) = \langle \mathbf{y}/\eta, \mathbf{u} \rangle$ due to the Fenchel-Young identity (16.30). By $\mathbf{M}_j^\top \mathbf{c}_{(\eta,\mathbf{y})} \in S$ and by applying the inequality (16.206) to (16.208), we have

$$\begin{aligned} U_S \geq \|\mathbf{M}_j^\top \mathbf{c}_{(\eta,\mathbf{y})}\| = \left\| \mathbf{M}_j^\top \begin{pmatrix} -\varphi^*(\mathbf{u}) \\ \mathbf{u} \end{pmatrix} \right\| &\geq \left| (-\varphi^*(\mathbf{u})) \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{u} \rangle \right| \\ &= |\Upsilon(\mathbf{u})| \geq \Upsilon_+(\mathbf{u}), \end{aligned} \tag{16.209}$$

where $\Upsilon \colon \mathbb{R}^N \to \mathbb{R} \colon \mathbf{v} \mapsto \varphi^*(\mathbf{v}) \|\mathbf{x}_j\|^2 - \langle \mathbf{x}_j, \mathbf{v} \rangle$ and $\Upsilon_+ \colon \mathbb{R}^N \to \mathbb{R} \colon \mathbf{v} \mapsto \max\{\Upsilon(\mathbf{v}), 0\}$ are coercive convex functions (see Section 16.2.1) and independent from the choice of $(\eta, \mathbf{y})$. The coercivity of Υ_+ ensures the existence of an open ball $B(0, \hat{U}_{(\mathrm{i})})$ of radius $\hat{U}_{(\mathrm{i})} > 0$ such that $\mathrm{lev}_{\leq U_S} \Upsilon_+ := \{\mathbf{v} \in \mathbb{R}^N \mid \Upsilon_+(\mathbf{v}) \leq U_S\} \subset B(0, \hat{U}_{(\mathrm{i})})$, and thus (16.209) implies

$$\|\mathbf{u}\| \leq \hat{U}_{(\mathrm{i})}. \tag{16.210}$$

Moreover, by $\mathbf{x}_j \neq 0$, the triangle inequality, the Cauchy-Schwarz inequality, (16.209), and (16.210), we have

$$|\varphi^*(\mathbf{u})| = \left| \frac{\Upsilon(\mathbf{u})}{\|\mathbf{x}_j\|^2} + \frac{\langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{x}_j\|^2} \right| \leq \left| \frac{\Upsilon(\mathbf{u})}{\|\mathbf{x}_j\|^2} \right| + \left| \frac{\langle \mathbf{x}_j, \mathbf{u} \rangle}{\|\mathbf{x}_j\|^2} \right|$$
$$\leq \left| \frac{\Upsilon(\mathbf{u})}{\|\mathbf{x}_j\|^2} \right| + \frac{\|\mathbf{u}\|}{\|\mathbf{x}_j\|} \leq \frac{U_S}{\|\mathbf{x}_j\|^2} + \frac{\hat{U}_{(\mathrm{i})}}{\|\mathbf{x}_j\|} =: U_{(\mathrm{i})}, \tag{16.211}$$

which yields $\mathbf{c}_{(\eta,\mathbf{y})} = (-\varphi^*(\mathbf{u}), \mathbf{u}) \in [-U_{(\mathrm{i})}, U_{(\mathrm{i})}] \times B(0, \hat{U}_{(\mathrm{i})})$. Since $(\eta, \mathbf{y}) \in \mathbb{R}_{++} \times \mathbb{R}^N$ is chosen arbitrarily and $\mathbf{c}_{(\eta,\mathbf{y})} \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(\eta, \mathbf{y})$ is also chosen arbitrarily, we have

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(\mathbb{R}_{++} \times \mathbb{R}^N) \subset [-U_{(\mathrm{i})}, U_{(\mathrm{i})}] \times B(0, \hat{U}_{(\mathrm{i})}),$$

which confirms the statement (i).

Proof of (ii) By introducing

$$\mathfrak{B} := \left\{ \mathbf{v} \in \mathbb{R}^N \;\middle|\; \left| \left\langle \frac{2}{\|\mathbf{x}_j\|^2} \mathbf{x}_j, \mathbf{v} \right\rangle \right| > |\varphi^*(\mathbf{v})| \right\}, \tag{16.212}$$

we can decompose the set $(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0)$ into

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}) \text{ and } (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}^c). \tag{16.213}$$

In the following, we show the boundedness of each set in (16.213).

First, we show the boundedness of $\mathfrak{B}$ by contradiction. Suppose that $\mathfrak{B} \not\subset B(0, r)$ for all $r > 0$. Then there exists a sequence $(\mathbf{u}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ such that

$$(\forall k \in \mathbb{N})\ \frac{2}{\|\mathbf{x}_j\|} \geq \left| \left\langle \frac{2}{\|\mathbf{x}_j\|^2} \mathbf{x}_j, \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|} \right\rangle \right| > \frac{|\varphi^*(\mathbf{u}_k)|}{\|\mathbf{u}_k\|} \text{ and } \|\mathbf{u}_k\| \geq k, \tag{16.214}$$

which contradicts the supercoercivity of φ^*, implying thus the existence of $r_* > 0$ such that $\mathfrak{B} \subset B(0, r_*)$.

Next, we show the boundedness of the former set in (16.213). Choose arbitrarily

$$(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}). \tag{16.215}$$

By $\mathbf{x}_j \neq 0$, $\mathbf{M}_j^\top (\mu, \mathbf{u}^\top)^\top \in S \subset B(0, U_S)$, the inequality (16.206), the triangle inequality, the Cauchy-Schwarz inequality, and $\mathbf{u} \in \mathfrak{B} \subset B(0, r_*)$, we have

$$\frac{U_S}{\|\mathbf{x}_j\|^2} \geq \frac{1}{\|\mathbf{x}_j\|^2} \left\| \mathbf{M}_j^\top \begin{pmatrix} \mu \\ \mathbf{u} \end{pmatrix} \right\| \geq \frac{1}{\|\mathbf{x}_j\|^2} |\mu \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{u} \rangle|$$

$$\geq |\mu| - \left|\left\langle \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|^2}, \mathbf{u} \right\rangle\right| \geq |\mu| - \frac{\|\mathbf{u}\|}{\|\mathbf{x}_j\|} \geq |\mu| - \frac{r_*}{\|\mathbf{x}_j\|}$$

which yields

$$\hat{U}_{(\mathrm{iia})} := \frac{U_S}{\|\mathbf{x}_j\|^2} + \frac{r_*}{\|\mathbf{x}_j\|} \geq |\mu|.$$

Therefore, we have $(\mu, \mathbf{u}) \in [-\hat{U}_{(\mathrm{iia})}, \hat{U}_{(\mathrm{iia})}] \times B(0, r_\star)$. Since $(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B})$ is chosen arbitrarily, we have

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}) \subset [-\hat{U}_{(\mathrm{iia})}, \hat{U}_{(\mathrm{iia})}] \times B(0, r_\star). \tag{16.216}$$

Finally, we show the boundedness of the latter set in (16.213). Let

$$(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}^c). \tag{16.217}$$

From (16.32), we have

$$\partial\widetilde{\varphi}(0,0) = \{(\mu', \mathbf{u}') \in \mathbb{R} \times \mathbb{R}^N \mid \mu' + \varphi^*(\mathbf{u}') \leq 0\}. \tag{16.218}$$

Note that coercivity of φ^* ($\Rightarrow \exists \min \varphi^*(\mathbb{R}^N) \in \mathbb{R}$, see Fact 16.2) and (16.218) yield $\varphi^*(\mathbf{u}) \in [\min \varphi^*(\mathbb{R}^N), -\mu]$ and thus

$$|\varphi^*(\mathbf{u})| \leq \max\{|\min \varphi^*(\mathbb{R}^N)|, |\mu|\} \leq |\min \varphi^*(\mathbb{R}^N)| + |\mu|. \tag{16.219}$$

By $\mathbf{x}_j \neq 0$, $\mathbf{M}_j^\top (\mu, \mathbf{u}^\top)^\top \in S \subset B(0, U_S)$ (see (16.217)), the inequality (16.206), the triangle inequality, $\mathbf{u} \in \mathfrak{B}^c$ (see (16.217) and (16.212)), and (16.219), we have

$$\begin{aligned}
\frac{2}{\|\mathbf{x}_j\|^2} U_S &\geq \frac{2}{\|\mathbf{x}_j\|^2} \left\| \mathbf{M}_j^\top \begin{pmatrix} \mu \\ \mathbf{u} \end{pmatrix} \right\| \geq \frac{2}{\|\mathbf{x}_j\|^2} |\mu \|\mathbf{x}_j\|^2 + \langle \mathbf{x}_j, \mathbf{u} \rangle| \\
&\geq 2|\mu| - \left| \left\langle \frac{2}{\|\mathbf{x}_j\|^2} \mathbf{x}_j, \mathbf{u} \right\rangle \right| \geq 2|\mu| - |\varphi^*(\mathbf{u})| \\
&\geq 2|\mu| - |\min \varphi^*(\mathbb{R}^N)| - |\mu| = |\mu| - |\min \varphi^*(\mathbb{R}^N)|
\end{aligned}$$

and thus, with (16.219),

$$\hat{U}_{(\mathrm{iib})} := \frac{2}{\|\mathbf{x}_j\|^2} U_S + 2|\min \varphi^*(\mathbb{R}^N)| \geq |\mu| + |\min \varphi^*(\mathbb{R}^N)| \geq |\varphi^*(\mathbf{u})| \geq \varphi^*(\mathbf{u}). \tag{16.220}$$

Hence, we have

$$(\mu, \mathbf{u}) \in [-\hat{U}_{(\mathrm{iib})}, \hat{U}_{(\mathrm{iib})}] \times \mathrm{lev}_{\leq \hat{U}_{(\mathrm{iib})}}(\varphi^*).$$

Since $(\mu, \mathbf{u}) \in (\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}^c)$ is chosen arbitrarily, we have

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \cap (\mathbb{R} \times \mathfrak{B}^c) \subset [-\hat{U}_{(\mathrm{iib})}, \hat{U}_{(\mathrm{iib})}] \times \mathrm{lev}_{\leq \hat{U}_{(\mathrm{iib})}}(\varphi^*). \tag{16.221}$$

Consequently, by using (16.216) and (16.221) and by letting $U_{(\mathrm{ii})} := \max\{\hat{U}_{(\mathrm{iia})}, \hat{U}_{(\mathrm{iib})}\}$, we have

$$(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0) \subset [-U_{(\mathrm{ii})}, U_{(\mathrm{ii})}] \times [\mathrm{lev}_{\leq U_{(\mathrm{ii})}}(\varphi^*) \cup B(0, r_\star)],$$

which guarantees the boundedness of $(\mathbf{M}_j^\top)^{-1}(S) \cap \partial\widetilde{\varphi}(0,0)$, due to the coercivity of φ^*, implying thus finally the statement (ii).

□

References

1. Argyriou, A., Baldassarre, L., Micchelli, C.A., Pontil, M.: On sparsity inducing regularization methods for machine learning. In: B. Schölkopf, Z. Luo, V. Vovk (eds.) Empirical Inference, pp. 205–216. Springer Berlin, Heidelberg (2013)
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)
3. Attouch, H.: Viscosity solutions of minimization problems. SIAM J. Optim. **6**, 769–806 (1996)
4. Attouch, H., Cabot, A., Chbani, Z., Riahi, H.: Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization. (preprint)
5. Baillon, J.-B., Bruck, R.E., Reich, S.: On the asymptotic behavior of nonexpansive mappings and semigroups in Banach spaces. Houst. J. Math. **4**, 1–9 (1978)
6. Bauschke, H.H.: The approximation of fixed points of compositions of nonexpansive mappings in Hilbert space. J. Math. Anal. Appl. **202**, 150–159 (1996)
7. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Rev. **38**, 367–426 (1996)
8. Bauschke, H.H., Combettes, P.L.: A weak-to-strong convergence principle for Fejér monotone methods in Hilbert space. Math. Oper. Res. **26**, 248–264 (2001)
9. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Space, 2nd edn. Springer (2017)
10. Bauschke, H.H., Moursi, M.: On the Douglas-Rachford algorithm. Math. Program. **164**, 263–284 (2017)
11. Beck, A., Teboulle, M.: Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. IEEE Trans. Image Process. **18**, 2419–2434 (2009)
12. Ben-Israel, A., Greville, T.N.E.: Generalized Inverses: Theory and Applications, 2nd edn. Springer-Verlag (2003)
13. Bien, J., Gaynanova, I., Lederer, J., Müller, C.L.: Non-convex global minimization and false discovery rate control for the TREX. J. Comput. Graph. Stat. **27**, 23–33 (2018)

14. Bishop, C.M.: Machine Learning and Pattern Recognition. Information Science and Statistics. Springer, Heidelberg (2006)
15. Blum, A., Rivest, R.L.: Training a 3-node neural network is NP-complete. Neural Networks **5**, 117–127 (1992)
16. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proc. the 5th Annual ACM Workshop on Computational Learning Theory (COLT), pp. 144–152 (1992)
17. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**, 1–122 (2011)
18. Cabot, A.: Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchical minimization. SIAM J. Optim. **15**, 555–572 (2005)
19. Candler, W., Norton, R.: Multilevel programming. Technical Report 20, World Bank Development Research Center, Washington D.C., USA (1977)
20. Cegielski, A.: Iterative Methods for Fixed Point Problems in Hilbert Spaces. Springer (2012)
21. Censor, Y., Davidi, R., Herman, G.T.: Perturbation resilience and superiorization of iterative algorithms. Inverse Probl. **26**, 065008 (2010)
22. Censor, Y., Zenios, S.A.: Parallel Optimization: Theory, Algorithm, and Optimization. Oxford University Press (1997)
23. Chaari, L., Ciuciu, P., Mériaux, S., Pesquet, J.C.: Spatio-temporal wavelet regularization for parallel MRI reconstruction: Application to functional MRI. Magn. Reson. Mater. Phys. Biol. Med. **27**, 509–529 (2014)
24. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". J. Optim. Theory Appl. **166**, 968–982 (2015)
25. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 1–27 (2011)
26. Chaux, C., Pesquet, J.C., Pustelnik, N.: Nested iterative algorithms for convex constrained image recovery problems. SIAM J. Imaging Sci. **2**, 730–762 (2009)
27. Chidume, C.: Geometric Properties of Banach Spaces and Nonlinear Iterations (Chapter 7: Hybrid steepest descent method for variational inequalities). vol. 1965 of Lecture Notes in Mathematics. Springer (2009)
28. Chipman, J.S.: Linear restrictions, rank reduction, and biased estimation in linear regression. Linear Algebra Appl. **289**, 55–74 (1999)
29. Chipman, J.S., Rao, M.M.: The treatment of linear restrictions in regression analysis. Econometrics **32**, 198–204 (1964)
30. Coloson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. Ann. Oper. Res. **153**, 235–256 (2007)
31. Combettes, P.L.: The foundations of set theoretic estimation. Proc. IEEE **81**, 182–208 (1993)
32. Combettes, P.L.: Inconsistent signal feasibility problems: Least squares solutions in a product space. IEEE Trans. Signal Process. **42**, 2955–2966 (1994)
33. Combettes, P.L.: Strong convergence of block-iterative outer approximation methods for convex optimization. SIAM J. Control Optim. **38**, 538–565 (2000)
34. Combettes, P.L.: Iterative construction of the resolvent of a sum of maximal monotone operators. J. Convex Anal. **16**, 727–748 (2009)
35. Combettes, P.L.: Perspective functions: Properties, constructions, and examples. Set-Valued Var. Anal. **26**, 247–264 (2017)
36. Combettes, P.L., Bondon, P.: Hard-constrained inconsistent signal feasibility problems. IEEE Trans. Signal Process. **47**, 2460–2468 (1999)
37. Combettes, P.L., Hirstoaga, S.A.: Approximating curves for nonexpansive and monotone operators. J. Convex Anal. **13**, 633–646 (2006)
38. Combettes, P.L., Müller, C.L.: Perspective functions: Proximal calculus and applications in high-dimensional statistics. J. Math. Anal. Appl. **457**, 1283–1306 (2018)
39. Combettes, P.L., Pesquet, J.-C.: Image restoration subject to a total variation constraint. IEEE Trans. Image Process. **13**, 1213–1222 (2004)

40. Combettes, P.L., Pesquet, J.-C.: A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. IEEE J. Sel. Top. Signal Process. **1**, 564–574 (2007)
41. Combettes, P.L., Pesquet, J.-C.: A proximal decomposition method for solving convex variational inverse problems. Inverse Probl. **24**, 065014 (2008)
42. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: H.H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, H. Wolkowicz (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer-Verlag (2011)
43. Combettes, P.L., Pesquet, J.-C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim. **25**, 1221–1248 (2015)
44. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. SIAM Multiscale Model. Simul. **4**, 1168–1200 (2005)
45. Combettes, P.L., Yamada, I.: Compositions and convex combinations of averaged nonexpansive operators. J. Math. Anal. Appl. **425**, 55–70 (2015)
46. Cominetti, R., Courdurier, M.: Coupling general penalty schemes for convex programming with the steepest descent and the proximal point algorithm. SIAM J. Optim. **13**, 745–765 (2002)
47. Condat, L.: A primal-dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**, 460–479 (2013)
48. Cortes, C., Vapnik, V.N.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
49. Cover, T.M.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. **14**, 326–334 (1965)
50. Dalalyan, A.S., Hebiri, M., Lederer, J.: On the prediction performance of the Lasso. Bernoulli **23**, 552–581 (2017)
51. Deutsch, F.: Best Approximation in Inner Product Spaces. New York: Springer-Verlag (2001)
52. Deutsch, F., Yamada, I.: Minimizing certain convex functions over the intersection of the fixed point sets of nonexpansive mappings. Numer. Funct. Anal. Optim. **19**, 33–56 (1998)
53. Donoho, D.L.: De-noising by soft-thresholding. IEEE Trans. Inf. Theory **41**, 613–627 (1995)
54. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation via wavelet shrinkage. Biometrika **81**, 425–455 (1994)
55. Dontchev, A.L., Zolezzi, T.: Well-posed optimization problems. vol. 1543 of Lecture Notes in Mathematics. Springer-Verlag (1993)
56. Dotson Jr., W.G.: On the Mann iterative process. Trans. Amer. Math. Soc. **149**, 65–73 (1970)
57. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two or three space variables. Trans. Amer. Math. Soc. **82**, 421–439 (1956)
58. Dupé, F.X., Fadili, M.J., Starck, J.-L.: A proximal iteration for deconvolving Poisson noisy images using sparse representations. IEEE Trans. Image Process. **18**, 310–321 (2009)
59. Dupé, F.X., Fadili, M.J., Starck, J.-L.: Deconvolution under Poisson noise using exact data fidelity and synthesis or analysis sparsity priors. Stat. Methodol. **9**, 4–18 (2012)
60. Durand, S., Fadili, M.J., Nikolova, M.: Multiplicative noise removal using L1 fidelity on frame coefficients. J. Math. Imaging Vision **36**, 201–226 (2010)
61. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
62. Eckstein, J., Yao, W.: Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. Pac. J. Optim. **11**, 619–644 (2015)
63. Eicke, B.: Iteration methods for convexly constrained ill-posed problems in Hilbert space. Numer. Funct. Anal. Optim. **13**, 413–429 (1992)
64. Ekeland, I., Themam, R.: Convex Analysis and Variational Problems. Classics in Applied Mathematics 28. SIAM (1999)
65. Fisher, A.R.: The use of multiple measurements in taxonomic problems. Ann. Hum. Genet. **7**, 179–188 (1936)
66. Gabay, D.: Applications of the method of multipliers to variational inequalities. In: M. Fortin, R. Glowinski (eds.) Augmented Lagrangian Methods: Applications to the solution of boundary value problems. North-Holland, Amsterdam (1983)

67. Gandy, S., Recht, B., Yamada, I.: Tensor completion and low-n-rank tensor recovery via convex optimization. Inverse Probl. **27**, 025010 (2011)
68. Gandy, S., Yamada, I.: Convex optimization techniques for the efficient recovery of a sparsely corrupted low-rank matrix. J. Math-For-Industry **2**, 147–156 (2010)
69. van de Geer, S., Lederer, J.: The Lasso, correlated design, and improved oracle inequalities. IMS Collections **9**, 303–316 (2013)
70. Goebel, K., Reich, S.: Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings. Marcel Dekker, New York (1984)
71. Groetsch, C. W.: A note on segmenting Mann iterates. J. Math. Anal. Appl. **40**, 369–372 (1972)
72. Halpern, B.: Fixed points of nonexpanding maps. Bull. Amer. Math. Soc. **73**, 957–961 (1967)
73. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer Series in Statistics (2009)
74. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC press (2015)
75. Haugazeau, Y.: Sur les inéquations variationnelles et la minimisation de fonctionnelles convexes. Thèse, Universite de Paris (1968)
76. He, B., Yuan, X.: On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. SIAM J. Numer. Anal. **50**, 700–709 (2012)
77. Hebiri, M., Lederer, J.: How correlations influence Lasso prediction. IEEE Trans. Inf. Theory **59**, 1846–1854 (2013)
78. Helou, E.S., De Pierro, A.R.: On perturbed steepest descent methods with inexact line search for bilevel convex optimization. Optimization **60**, 991–1008 (2011)
79. Helou, E.S., Simões, L.E.A.: ϵ-subgradient algorithms for bilevel convex optimization. Inverse Probl. **33**, 055020 (2017)
80. Herman, G.T., Garduño, E., Davidi, R., Censor, Y.: Superiorization: An optimization heuristic for medical physics. Med. Phys. **39**, 5532–5546 (2012)
81. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**, 303–320 (1969)
82. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms. Springer (1993)
83. Iemoto, S., Takahashi, W.: Strong convergence theorems by a hybrid steepest descent method for countable nonexpansive mappings in Hilbert spaces. Sci. Math. Jpn. **69**, 227–240 (2009)
84. Judd, J.S.: Learning in networks is hard. In: Proc. 1st Int. Conf. Neural Networks, pp. 685–692 (1987)
85. Kailath, T., Sayed, A.H., Hassibi, B.: Linear Estimation. Prentice-Hall (2000)
86. Kitahara, D., Yamada, I.: Algebraic phase unwrapping based on two-dimensional spline smoothing over triangles. IEEE Trans. Signal Process. **64**, 2103–2118 (2016)
87. Koltchinskii, V., Lounici, K., Tsybakov, A.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist. **39**, 2302–2329 (2011)
88. Krasnosel'skiĭ, M.A.: Two remarks on the method of successive approximations. Uspekhi Mat. Nauk **10**, 123–127 (1955)
89. Lederer, J., Müller, C.L.: Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. In: Proc. Twenty-Ninth AAAI Conf. Artif. Intell., pp. 2729–2735 (2015)
90. Lions, P.L.: Approximation de points fixes de contractions. C. R. Acad. Sci. Paris Sèrie A-B **284**, 1357–1359 (1977)
91. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
92. Lobo, M.S., Vandenberghe, L., Boyd, S., Lebret, H.: Applications of second-order cone programming. Linear Algebra Appl. **284**, 193–228 (1998)
93. Luenberger, D.G.: Optimization by Vector Space Methods. Wiley (1969)
94. Mainge, P.E.: Extension of the hybrid steepest descent method to a class of variational inequalities and fixed point problems with nonself-mappings. Numer. Funct. Anal. Optim. **29**, 820–834 (2008)

95. Mangasarian, O.L.: Iterative solution of linear programs. SIAM J. Numer. Amal. **18**, 606–614 (1981)
96. Mann, W.: Mean value methods in iteration. Proc. Amer. Math. Soc. **4**, 506–510 (1953)
97. Marquardt, D.W.: Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics **12**, 591–612 (1970)
98. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle **4**, 154–159 (1970)
99. Martinet, B.: Détermination approchée d'un point fixe d'une application pseudo-contractante. C. R. Acad. Sci. Paris Ser. A-B **274**, 163–165 (1972)
100. Moore, E.H.: On the reciprocal of the general algebraic matrix. Bull. Amer. Math. Soc. **26**, 394–395 (1920)
101. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. Acad. Sci. Paris Ser. A Math. **255**, 2897–2899 (1962)
102. Moreau, J.J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France **93**, 273–299 (1965)
103. Nesterov, Y.: A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Math. Dokl. **27**, 372–376 (1983)
104. Nikazad, T., Davidi, R., Herman, G.T.: Accelerated perturbation-resilient block-iterative projection methods with application to image reconstruction. Inverse Probl. **28**, 035005 (2012)
105. Ogura, N., Yamada, I.: Non-strictly convex minimization over the fixed point set of the asymptotically shrinking nonexpansive mapping. Numer. Funct. Anal. Optim. **23**, 113–137 (2002)
106. Ogura, N., Yamada, I.: Non-strictly convex minimization over the bounded fixed point set of nonexpansive mapping. Numer. Funct. Anal. Optim. **24**, 129–135 (2003)
107. Ono, S., Yamada, I.: Hierarchical convex optimization with primal-dual splitting. IEEE Trans. Signal Process. **63**, 373–388 (2014)
108. Ono, S., Yamada, I.: Signal recovery with certain involved convex data-fidelity constraints. IEEE Trans. Signal Process. **63**, 6149–6163 (2015)
109. Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**, 383–390 (1979)
110. Penfold, S.N., Schulte, R.W., Censor, Y., Rosenfeld, A.B.: Total variation superiorization schemes in proton computed tomography image reconstruction. Med. Phys. **37**, 5887–5895 (2010)
111. Penrose, R.: A generalized inverse for matrices. Proc. Cambridge Philos. Soc. **51**, 406–413 (1955)
112. Piotrowski, T., Cavalcante, R., Yamada, I.: Stochastic MV-PURE estimator? Robust reduced-rank estimator for stochastic linear model. IEEE Trans. Signal Process. **57**, 1293–1303 (2009)
113. Piotrowski, T., Yamada, I.: MV-PURE estimator: Minimum-variance pseudo-unbiased reduced-rank estimator for linearly constrained ill-conditioned inverse problems. IEEE Trans. Signal Process. **56**, 3408–3423 (2008)
114. Polyak, B.T.: Sharp minimum. International Workshop on Augmented Lagrangians (1979)
115. Potter, L.C., Arun, K.S.: A dual approach to linear inverse problems with convex constraints. SIAM J. Control Optim. **31**, 1080–1092 (1993)
116. Powell, M.J.D.: A method for nonlinear constraints in minimization problems. In: R. Fretcher (ed.) Optimization, pp. 283–298. Academic Press (1969)
117. Pustelnik, N., Chaux, C., Pesquet, J.-C.: Parallel proximal algorithm for image restoration using hybrid regularization. IEEE Trans. Image Process. **20**, 2450–2462 (2011)
118. Rao, C.R., Mitra, S.K.: Generalized Inverse of Matrices and Its Applications. John Wiley & Sons (1971)
119. Reich, S.: Weak convergence theorems for nonexpansive mappings in Banach spaces. J. Math. Anal. Appl. **67**, 274–276 (1979)
120. Rigollet, P., Tsybakov, A.: Exponential screening and optimal rates of sparse estimation. Ann. Statist. **39**, 731–771 (2011)

121. Rockafellar, R.T.: Monotone operators and proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)
122. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, 1st edn. Springer (1998)
123. Sabharwal, A., Potter, L.C.: Convexly constrained linear inverse problems: Iterative least-squares and regularization. IEEE Trans. Signal Process. **46**, 2345–2352 (1998)
124. Saitoh, S.: Theory of Reproducing Kernels and Its Applications. Longman Scientific & Technical, Harlow (1988)
125. Schölkopf, B., Luo, Z., Vovk, V.: Empirical Inference. Springer-Verlag (2013)
126. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT press (2002)
127. Solodov, M.: An explicit descent method for bilevel convex optimization. J. Convex Anal. **14**, 227–237 (2007)
128. Solodov, M.: A bundle method for a class of bilevel nonsmooth convex minimization problems. SIAM J. Optim. **18**, 242–259 (2008)
129. Takahashi, N., Yamada, I.: Parallel algorithms for variational inequalities over the cartesian product of the intersections of the fixed point sets of nonexpansive mappings. J. Approx. Theory **153**, 139–160 (2008)
130. Takahashi, W.: Nonlinear Functional Analysis—Fixed Point Theory and its Applications. Yokohama Publishers (2000)
131. Theodoridis, S.: Machine Learning: Bayesian and Optimization Perspective. Academic Press (2015)
132. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B **58**, 267–288 (1996)
133. Tikhonov, A.N.: Solution of incorrectly formulated problems and the regularization method. Soviet Math. Dokl. **4**, 1035–1038 (1963)
134. Tseng, P.: Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim. **29**, 119–138 (1991)
135. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons (1998)
136. Vapnik, V.N., Lerner, A.: Pattern recognition using generalized portrait method. Automat. Rem. Contr. **24**, 774–780 (1963)
137. Varga, R.S.: Matrix Iterative Analysis, 2nd edn. Springer, New York (2000)
138. Vicente, L.N., Calamai, P.H.: Bilevel and multilevel programming: A bibliography review. J. Global Optim. **5**, 291–306 (1994)
139. Vu, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**, 667–681 (2013)
140. Xu, H.K., Kim, T.H.: Convergence of hybrid steepest descent methods for variational inequalities. J. Optim. Theory Appl. **119**, 185–201 (2003)
141. Yamada, I.: Approximation of convexly constrained pseudoinverse by hybrid steepest descent method. In: Proc. IEEE ISCAS (1999)
142. Yamada, I.: The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings. In: D. Butnariu, Y. Censor, S. Reich (eds.) Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications, pp. 473–504. Elsevier (2001)
143. Yamada, I.: Kougaku no Tameno Kansu Kaiseki (Functional Analysis for Engineering). Suurikougaku-Sha/Saiensu-Sha, Tokyo (2009)
144. Yamada, I., Elbadraoui, J.: Minimum-variance pseudo-unbiased low-rank estimator for ill-conditioned inverse problems. In: Proc. IEEE ICASSP, III, pp. 325–328 (2006)
145. Yamada, I., Ogura, N.: Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings. Numer. Funct. Anal. Optim. **25**, 619–655 (2004)
146. Yamada, I., Ogura, N., Shirakawa, N.: A numerically robust hybrid steepest descent method for the convexly constrained generalized inverse problems. In: Z. Nashed, O. Scherzer (eds.) Inverse Problems, Image Analysis, and Medical Imaging, *Contemporary Mathematics*, vol. 313, pp. 269–305. AMS (2002)

147. Yamada, I., Ogura, N., Yamashita, Y., Sakaniwa, K.: An extension of optimal fixed point theorem for nonexpansive operator and its application to set theoretic signal estimation. Technical Report of IEICE, DSP96-106, pp. 63–70 (1996)
148. Yamada, I., Ogura, N., Yamashita, Y., Sakaniwa, K.: Quadratic optimization of fixed points of nonexpansive mappings in Hilbert space. Numer. Funct. Anal. Optim. **19**, 165–190 (1998)
149. Yamada, I., Yukawa, M., Yamagishi, M.: Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings. In: H.H. Bauschke, R. Burachik, P. Combettes, V. Elser, D. Luke, H. Wolkowicz (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 345–390. Springer (2011)
150. Yamagishi, M., Yamada, I.: Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization. Inverse Probl. **33**, 044003 (2017)
151. Yang, J., Yuan, X.: Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. Math. Comp. **82**, 301–329 (2013)
152. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific (2002)
153. Zeidler, E.: Nonlinear Functional Analysis and its Applications, III - Variational Methods and Optimization. Springer (1985)