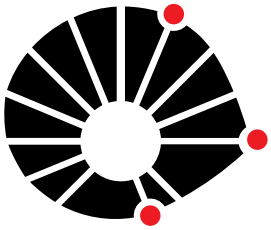


UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Matemática, Estatística e Computação Científica

FELIPE EDUARDO ATENAS MALDONADO

Proximal decomposition methods for optimization problems with structure

Métodos de decomposição proximal para problemas de otimização com estrutura

Campinas

2023

Felipe Eduardo Atenas Maldonado

Proximal decomposition methods for optimization problems with structure

Métodos de decomposição proximal para problemas de otimização com estrutura

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Matemática Aplicada.

Thesis presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Doctor in Applied Mathematics.

Supervisor: Paulo José da Silva e Silva

Co-supervisor: Claudia Alejandra Sagastizabal

Este trabalho corresponde à versão final da Tese defendida pelo aluno Felipe Eduardo Atenas Maldonado e orientada pelo Prof. Dr. Paulo José da Silva e Silva.

Campinas

2023

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

At27p

Atenas Maldonado, Felipe Eduardo, 1994-
Proximal decomposition methods for optimization problems with structure / Felipe Eduardo Atenas Maldonado. – Campinas, SP : [s.n.], 2023.

Orientador: Paulo José da Silva e Silva.
Coorientador: Claudia Alejandra Sagastizabal.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Otimização estocástica. 2. Método de decomposição. 3. Método de descida (Otimização). 4. Método de feixes. 5. Otimização matemática. 6. Programação estocástica. I. Silva, Paulo José da Silva e, 1973-. II. Sagastizabal, Claudia Alejandra, 1961-. III. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. IV. Título.

Informações Complementares

Título em outro idioma: Métodos de decomposição proximal para problemas de otimização com estrutura
Palavras-chave em inglês:
Stochastic optimization
Decomposition method
Method of descent (Optimization)
Bundle methods
Mathematical optimization
Stochastic programming
Área de concentração: Matemática Aplicada
Titulação: Doutor em Matemática Aplicada
Banca examinadora:
Claudia Alejandra Sagastizabal [Coorientador]
Andrzej Piotr Ruszczynski
Radu Ioan Bot
José Mario Martínez Pérez
Welington Luis de Oliveira
Data de defesa: 04-09-2023
Programa de Pós-Graduação: Matemática Aplicada

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-8577-697X
- Currículo Lattes do autor: http://lattes.cnpq.br/2463585840418981

Tese de Doutorado defendida em 04 de setembro de 2023 e aprovada

pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). CLAUDIA ALEJANDRA SAGASTIZABAL

Prof(a). Dr(a). ANDRZEJ PIOTR RUSZCZYNSKI

Prof(a). Dr(a). RADU IOAN BOT

Prof(a). Dr(a). JOSÉ MARIO MARTÍNEZ PÉREZ

Prof(a). Dr(a). WELINGTON LUIS DE OLIVEIRA

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Acknowledgements

A mis padres, Rosa y Hugo, por siempre apoyarme incondicionalmente y acompañarme en mis ñoñerías, desde que era pequeño. Estoy eternamente agradecido por haberme brindado un hogar lleno de amor, en donde pude crecer y desarrollarme hasta llegar a ser la persona que soy ahora.

A mi profesora guía, Claudia Sagastizábal, por ser una orientadora académica y personal sin igual, por siempre tener un buen consejo y por haberme acogido desde la maestría, después acceder a seguir trabajando en conjunto durante el doctorado y por introducirme al fascinante mundo de la investigación.

A mi profesor co-guía, Paulo Silva, por haberme recibido en la Unicamp y siempre brindar su apoyo. También al profesor Jonathan Eckstein, por aceptar guiarme durante la pasantía de investigación y por su comprensión e incentivo en el desarrollo del proyecto. Agradezco también a los profesores Nino –Roberto Andreani– y Sandra Santos, por contribuir a mi formación durante el doctorado.

A los miembros de la banca, profesores Andrzej Ruszczyski, Radu Bo, José Mario Martínez y Welington de Oliveira, por sus comentarios y recomendaciones. También al profesor Luiz-Rafael Santos, miembro suplente de la banca, por sus correcciones y sugerencias. También agradezco al profesor Mikhail Solodov, miembro informal de la banca, por su dedicación cuando hemos colaborado.

O presente trabalho foi desenvolvido com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processos 2019/20023-1 e 2022/02208-7.

Resumo

Problemas complexos de otimização de grande porte geralmente apresentam uma estrutura separável por blocos que permite o uso de técnicas de decomposição. Os métodos de decomposição lidam adequadamente com acoplamentos nas restrições ou nas variáveis, para aproveitar essas estruturas separáveis. Essa tese explora algoritmos de decomposição para problemas de otimização convexos e não convexos que, em cada iteração, primeiro resolvem subproblemas descentralizados e depois coordenam a informação distribuída. Nosso objetivo é duplo. Por um lado, propomos uma análise unificadora de convergência de métodos de descida para otimização não convexa, incluindo taxas de convergência sob hipóteses fracas de regularidade. Os algoritmos de decomposição fornecem frequentemente descida para alguma medida de progresso ao longo das iterações, como a redução de alguma função de mérito ou a distância ao conjunto de soluções. As abordagens incluídas em nossa análise são os métodos de feixe proximal e o método de Douglas-Rachford para otimização fracamente convexa. Por outro lado, estendemos a análise unificadora para problemas de otimização restritos a um subespaço linear. Isso nos permite desenvolver técnicas de decomposição que usufruem da separabilidade por cenários de problemas de otimização estocástica multiestágio, uma vez que as restrições acoplantes desses problemas são representadas por um subespaço linear relacionado à não antecipatividade no processo de decisões. Para problemas convexos, propomos dois novos métodos que aproximam o método de Lagrangiano aumentado, ao induzir separabilidade no problema dual. As restrições acopladoras são modeladas com uma função linear construída usando passos *forward-backward*. Esses dois métodos são variantes do algoritmo *Progressive Hedging* de Rockafellar e Wets, com a importante diferença de serem convergentes também se o comprimento de passo varia com as iterações. Os dois métodos propostos diferem na forma como o progresso ao longo das iterações é avaliado. O primeiro método corresponde a uma variante do tipo feixe proximal do algoritmo *Progressive Hedging*, que mede descida suficiente da função de custo. O segundo método é uma técnica de decomposição que emprega um teste de aceitação de erro relativo que avalia a inviabilidade e a precisão do modelo. Ambos os métodos geram sequências primais-duais que convergem a soluções dos problemas primal e dual, respectivamente, com taxa de convergência linear.

Palavras-chave: métodos de decomposição, métodos de descida, Progressive Hedging, métodos de feixe, convexidade fraca, error bounds, convergência linear.

Abstract

Complex large-scale optimization problems usually display a block-separable structure that allows the use of decomposition techniques. Decomposition methods appropriately handle couplings in the constraints or in the variables in order to exploit the separable structure. This thesis explores decomposition methods for convex and nonconvex optimization problems, that first solve decentralized subproblems, and then coordinate the distributed information. Our goal is two-fold. On one hand, we propose a general unifying framework for convergence analysis of descent methods in nonconvex optimization, including rates of convergence under mild regularity assumptions. Decomposition algorithms frequently provide descent for some improvement measure throughout iterations, such as reduction of some merit function or the distance to the set of solutions. Approaches included in this framework are proximal bundle methods, and the Douglas-Rachford splitting method for weakly convex optimization. On the other hand, we extend the unifying analysis to constrained optimization problems over a linear subspace. This allows us to develop decomposition techniques that capitalize on the scenario separability of multistage stochastic optimization problems, since the linking constraints for these problems are represented by certain linear subspace, related to nonanticipativity in the decision process. For convex problems, we propose two novel methods that replace certain Augmented Lagrangian by separable approximations inducing separability in the dual problem. The coupling constraints are modeled with a linear function constructed using forward-backward steps. These methods are variants of the Progressive Hedging algorithm by Rockafellar and Wets, with the key difference of being convergent also with varying stepsizes along iterations. The two proposed methods differ in the way that improvement along iterations is evaluated. The first method corresponds to a proximal bundle-like adaptation of the Progressive Hedging algorithm, that measures sufficient descent of the cost function. The second one is a splitting technique that employs a relative-error acceptance test, assessing infeasibility and model accuracy. Both methods are shown to generate primal-dual sequences that convergence to solutions to the primal and dual problems, respectively, with linear speed of convergence.

Keywords: Decomposition methods, descent methods, Progressive Hedging, bundle methods, weak convexity, error bounds, linear convergence.

List of Figures

List of Tables

List of abbreviations and acronyms

PPA	Proximal point algorithm
DRS	Douglas-Rachford splitting
PHA	Progressive Hedging algorithm
BPHA	Bundle Progressive Hedging algorithm
DEFB	Dual-embedded forward-backward
DEFBAL	Dual-embedded forward-backward Augmented Lagrangian

List of symbols

$\|\cdot\|$, $\langle\cdot,\cdot\rangle$	Euclidean norm and associated inner product
$\|\cdot\|_S$, $\langle\cdot,\cdot\rangle_S$	Weighted norm and associated inner product
$X \times Y$, $\Pi_{i\in I} X_i$	Cartesian product of X and Y, and of X_i for indices $i \in I$
$\mathrm{int}(C)$, $\mathrm{ri}(C)$	Interior and relative interior of C
$\overline{\mathrm{conv}}(\mathbb{R}^n)$	Set of proper lower semicontinuous convex functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$
$w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$	Set of proper lower semicontinuous $\rho-$ weakly convex functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$
lsc	Lower semicontinuous
$\mathrm{dom}(f)$	Domain of f
f^*	Fenchel conjugate of f
i_C	Indicator function of the set C
P_C	Projection onto the set C
∇f	Gradient of f
$\widehat{\partial} f$, ∂f, $\overline{\partial} f$	Fréchet, limiting, and Clarke subdifferentials of f
$\widehat{N}_C$, N_C	Fréchet and limiting normal cones of C

List of Algorithms

Contents

Introduction

Decomposition-coordination methods, or simply decomposition methods, are used to solve large-scale problems that are difficult or impossible to tackle directly, with classical techniques. Decomposition methods advantageously exploit certain type of structural features present in the problem that make it possible to solve instead a sequence of smaller and simpler subproblems. These subproblems, solved in a decentralized distributed manner, are subsequently coordinated in order to find a solution. Problems well-suited for decomposition appear in multiple applications, including signal and image processing, statistics, energy systems, machine learning, among others. See [1, 2, 3, 4] and references therein.

Modern real-world problems usually involve numerous constraints and variables. Decomposition methods are designed to handle complicating constraints and/or complicating variables, usually presented in a structured way. For instance, in power systems, mathematical optimization models in generation planning include linkage constraints related to different technologies (hydro, wind, thermal). For stochastic formulations, nonanticipativity constraints are incorporated when intermittent sources of energy play a relevant role in the system. We refer to [5, 6, 7, 8] for more details.

Traditional decomposition approaches require mathematical formulations to have a block-separable structure, such as Dantzig-Wolfe decomposition [9], Benders decomposition [10], Douglas-Rachford splitting [11], and the alternating direction method of multipliers (ADMM) [12]. Other methods, not originally designed for decomposition, such as bundle methods [13, 14], can be employed to exploit separable structures through duality.

Many well-known decomposition methods are based on two strategies. These are the dual decomposition technique [15] and the proximal point algorithm [16, 17]. In this thesis, we focus on decomposition techniques of proximal type. These methods usually provide sequential improvement along iterations, in the form of descent of the objective or a merit function, or reducing the distance to the solution set. The descent condition not also serves to evaluate the progress of the method, but also helps to define optimality certificates for the generated sequences.

The classical methods mentioned above were originally devised for linear or convex programming. The rapid growth of challenging problems urges the development of theory and algorithms that match modern necessities, usually involving nonconvex problems. Similarly to the previously mentioned complicating constraints and variables,

nonconvexity frequently appears in a structured, manageable manner. This feature allows the extension of existing methods to new settings.

In this work, we analyze techniques to solve convex and nonconvex optimization problems, including methods of descent and decomposition. We also design new algorithms for contemporary applications, such as inexact proximal methods for multistage stochastic optimization problems. We focus on leveraging structures to enhance the performance of algorithms and finding hidden properties of classical methods to understand them in a more abstract manner.

Our work on descent methods is primarily based on [18] for nonconvex problems, and [19] of methods of $\varepsilon-$subgradient descent. Our first contribution is a comprehensive theory for descent methods for weakly convex optimization, unifying the analysis of explicit and implicit methods, including for the first time proximal bundle methods in a framework akin to [18]. Furthermore, under typical regularity assumptions, we prove linear rates of convergence for methods falling under our unifying approach. Similar general frameworks for descent sequences with subsequential convergence, and global convergence with linear rates under extra assumptions can be found in [18, 20, 21]. The associated published article of our contribution is [22]:

> Atenas, F., Sagastizábal, C., Silva, P. J., & Solodov, M. (2023). A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33(1), 89–115.

We also directly extend this analysis to weakly convex optimization problems with constraints in a linear subspace, as shown in [23] and explained below.

Concerning decomposition methods, we work on two frontiers: first, we develop two scenario-based decomposition methods for stochastic programming, resembling the progressive hedging algorithm [24], but with varying stepsizes. One of the methods is motivated by proximal bundle methods, and the other by an inexact relative-error proximal point method [25]. Both algorithms correspond to extensions of the Progressive Hedging algorithm [24], incorporating an improvement measure that checks the quality of the approximation. The two approaches differ in the acceptance test condition: the proximal bundle sufficient descent condition is at most as strict as the one based on [25]. For both methods, we prove convergence to solutions in the convex case and provide sufficient conditions to obtain linear convergence rates. The first approach has the following associated publication [23]:

Atenas, F., & Sagastizábal, C. (2023). A bundle-like progressive hedging algorithm. *Journal of Convex Analysis*, special issue in honor of R. J-B Wets, 30(2) 453–479.

Similar techniques can be found in [26, 27], for an Augmented Lagrangian scenario decomposition method using separable approximation of the Augmented Lagrangian by fixing variables.

The second frontier involves splitting methods for nonconvex functions, along the lines of [28], but in the weakly convex setting. We extend the previously mentioned ideas of [22] for the Douglas-Rachford method by using a suitable merit function, and also provide convergence guarantees to critical points and rates of convergence for weakly convex problems. Numerical experiments suggest that in this setting, it is still possible to convergence to global minimizers, opening the question of what are the conditions for which the Douglas-Rachford splitting method avoids saddle points.

This thesis is organized as follows. In Chapter 1, we describe fundamental tools of variational analysis, and propose an enlargement for the subdifferential of a weakly convex function that can be employed for implementation purposes. In Chapter 2 we present the central algorithms to solve optimization problems that act as the basis of the ones proposed in this work. We proceed in Chapter 3 to introduce the unifying analysis of descent methods for weakly convex problems of [22], and an extension for constrained optimization problems. We continue with another extension for splitting methods in Chapter 4, in particular, the Douglas-Rachford splitting method for weakly convex optimization problems. Regarding stochastic optimization, in Chapter 5 we propose a variation of the progressive hedging algorithm of proximal bundle-type, by adding an extra sufficient descent test [23]. We develop another variation of the progressive hedging algorithm in Chapter 6, by introducing a relative-error acceptance test that measures feasibility and approximation accuracy. We conclude with Chapter 7 with a discussion of some ongoing work and future research directions.

1 Variational Analysis tools

In this chapter, we introduce the essential notions of variational analysis and optimization we use throughout this dissertation. We closely follow the notation of the book [29].

First, in Section 1.1 we start with elementary properties of functions and operators in the context of optimization. We continue in Section 1.2 stating different subdifferential notions for convex and nonconvex functions, and some algebraic and topological properties. The concept of weak convexity is examined in Section 1.3, highlighting its importance in modern optimization. In particular, in Section 1.3.2, we introduce a new concept of subdifferential, the approximate subdifferential for weakly convex functions, an extension of the $\varepsilon-$subdifferential of Convex Analysis, and we also study a variational principle for it. Section 1.3.2 is ongoing work to be submitted. We end this chapter with a brief survey of error bounds, the Kurdyka-Łojasiewicz inequality and some other regularity properties in the literature in Section 1.4 for convex and nonconvex optimization problems.

1.1 Basic concepts and notation

In this work, we consider functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, possibly nonconvex and nonsmooth. The domain of f is $\operatorname{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$. When $\operatorname{dom}(f)$ is nonempty, f is said to be proper. The graph of f consists of all the points of the form $(x, f(x))$ for $x \in \operatorname{dom}(f)$, while the epigraph of f is the set of points $(x, \alpha) \in \mathbb{R}^n \times \mathbb{R}$ such that $f(x) \leqslant \alpha$, and is denoted by $\operatorname{epi}(f)$.

The study of continuity properties of a function f classically involves the behavior of f on convergent (sub)sequences. For the purpose of still capturing the local properties of a not necessarily continuous function f at a point $\bar{x}$, we use the notion of $f-$attentive convergence, as defined in [29, Chapter 8B, 8(2)]. A sequence $\{x^k\} \subseteq \mathbb{R}^n$ is said to converge to $\bar{x}$ in the $f-$attentive sense, denoted by $x^k \underset{f}{\to} \bar{x}$, if $x^k \to \bar{x}$, and $f(x^k) \to f(\bar{x})$.

A lower semicontinuous (lsc) function f at a point $\bar{x} \in \mathbb{R}^n$ is characterized by the following estimate:

$$f(\bar{x}) \leqslant \liminf_{x \to \bar{x}} f(x).$$

A function f is said to be convex if for all $\lambda \in [0,1]$, $x, y \in \mathbb{R}^n$,

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y). \tag{1.1}$$

We denote by $\overline{\text{conv}}(\mathbb{R}^n)$ the set of all proper lsc convex functions from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$. Functions in $\overline{\text{conv}}(\mathbb{R}^n)$ can be characterized via their subdifferential. For that, we first need to introduce the notion of set-valued operator. A set-valued operator is a mapping T, such that for each $x \in \mathbb{R}^n$, $T(x) \subseteq \mathbb{R}^m$. We denote set-valued operators as $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$.

For a set-valued mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$, $\text{dom}(T)$ denotes the domain of T, and is defined by $\text{dom}(T) = \{x \in \mathbb{R}^n : T(x) \neq \varnothing\}$. The graph of the operator T is defined as the set $\text{gph}(T) = \{(x, g) \in \mathbb{R}^n \times \mathbb{R}^n : g \in T(x)\}$. The inverse operator of T, denoted by T^{-1}, is the set-valued mapping satisfying

$$x \in T^{-1}(y) \iff y \in T(x).$$

A prime example of a maximal monotone operator is the subdifferential of a function $f \in \overline{\text{conv}}(\mathbb{R}^n)$, see (1.2).

Continuity properties of operators are common in Variational Analysis, and they are employed in the convergence analysis of algorithms. In particular, subdifferentials are prominent examples of operators that exhibit some notable continuity properties defined in [29, Definition 5.4].

We recall two concepts of continuity of set-valued operators. A set-valued operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is

1. outer semicontinuous at a point $\bar{x} \in \mathbb{R}^n$, if for some sequences $\{x^k\} \subseteq \mathbb{R}^n$ and $\{v^k\} \subseteq \mathbb{R}^n$, such that $x^k \to \bar{x}$ and $T(x^k) \ni v^k \to v \in \mathbb{R}^n$, it holds $v \in T(\bar{x})$.
2. inner semicontinuous at a point $\bar{x} \in \mathbb{R}^n$, if for any $v \in T(\bar{x})$, and any sequence $x^k \to \bar{x}$, there exists a subsequence $\{v^k\}$ indexed by $K \subseteq \mathbb{N}$, such that $T(x^k) \ni v^k \to v$.

Subdifferentials provide an alternative manner to define convex functions, as mentioned above, through properties of tangents to the graph of f: f is convex if and only if all tangent lines to the graph of f are lower estimates of the function. For $x \in \mathbb{R}^n$, the slopes of the tangent in $(x, f(x))$ define the subdifferential $\partial f(x)$ of the convex function f at x:

$$\partial f(x) = \{v \in \mathbb{R}^n : f(y) \geqslant f(x) + \langle v, y - x\rangle \text{ for all } y \in \mathbb{R}^n\}. \tag{1.2}$$

If $x \notin \text{dom}(f)$, $\partial f(x)$ is empty. For subdifferentials defined in more general settings, see Section 1.2.

An important example of a convex function is the indicator i_C of a set $C \subseteq \mathbb{R}^n$, namely, the function that takes the value 0 on C and $+\infty$ otherwise. When C is a nonempty closed convex set, then $i_C \in \overline{\text{conv}}(\mathbb{R}^n)$. Furthermore, the subdifferential if i_C at a point $x \in C$ is the (convex) normal cone of C at x, denoted $N_C(x)$.

For a nonempty closed set C, we denote the associated (possibly set-valued) projection mapping by

$$P_C(x) = \arg\min_{y \in C} \|y - x\|,$$

while the optimal value of this problem is the distance from x to C:

$$\text{dist}(x, C) = \|x - p\|, \quad \text{for any } p \in P_C(x).$$

For a nonempty closed set C, P_C is a nonempty-valued operator. If, in addition, C is convex, P_C is a single-valued operator.

The Fenchel conjugate $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ of a function f, is defined as

$$f^*(w) = \sup_{x \in \mathbb{R}^n} \{\langle w, x\rangle - f(x)\}. \tag{1.3}$$

We also use the biconjugate $f^{**} : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ of a function f, corresponding to the conjugate of the conjugate, namely

$$f^{**}(x) = \sup_{w \in \mathbb{R}^n} \{\langle x, w\rangle - f^*(w)\}. \tag{1.4}$$

When f is a proper function, both f^*, $f^{**} \in \overline{\text{conv}}(\mathbb{R}^n)$ (see [29, Theorem 11.1]). Furthermore, when $f \in \overline{\text{conv}}(\mathbb{R}^n)$, $f^{**} = f$. From the definition of the Fenchel conjugate, we directly obtain the so-called Fenchel-Young inequality: for all $x, w \in \mathbb{R}^n$, it holds

$$f(x) + f^*(w) \geqslant \langle w, x\rangle. \tag{1.5}$$

The equality in (1.5) characterizes subgradients of f and f^* (see [29, Proposition 11.3]):

$$f(x) + f^*(w) = \langle w, x\rangle \iff w \in \partial f(x) \iff x \in \partial f^*(w), \tag{1.6}$$

and

$$\partial f(x) = \arg\max_{w \in \mathbb{R}^n} \{\langle x, w\rangle - f^*(w)\}, \quad \partial f^*(w) = \arg\max_{x \in \mathbb{R}^n} \{\langle w, x\rangle - f(x)\}.$$

In this way, we can think of ∂f^* as the inverse operator of ∂f.

One particular case is of interest, since it is used in Chapter 6. The Fenchel conjugate of the indicator i_C of a set C takes the following form

$$i_C^*(w) = \sup_{x \in \mathbb{R}^n} \langle w, x\rangle.$$

In particular, for a linear subspace $\mathcal{M}$, the indicators of $\mathcal{M}$ and its orthogonal complement $\mathcal{M}^\perp$ are closely related: $i^*_{\mathcal{M}} = i_{\mathcal{M}^\perp}$. This is due to the fact that a point $w \in \mathbb{R}^n$ satisfies $i^*_{\mathcal{M}}(w) = 0$ if and only if $\langle w, x\rangle = 0$ for all $x \in \mathbb{R}^n$, that is, $w \in \mathcal{M}^\perp$ (see also [29, Example 11.4]).

Another example of interest is the Fenchel conjugate of a strictly convex function. We say a function f is strictly convex if inequality (1.1) holds strictly for $x \neq y$. In this case, $\operatorname{int}(\operatorname{dom}(f)) \neq \varnothing$, and f^* is continuously differentiable on $\operatorname{int}(\operatorname{dom}(f))$. Particularly, $\nabla f^*(w)$ is given by the unique minimizer of $f(\cdot) - \langle w, \cdot\rangle$. Examples of strictly convex functions are strongly convex functions. For a constant $\mu > 0$, we say a function f_μ is $\mu-$strongly convex (cf. Definition 1.3) if $f_\mu(\cdot) - \dfrac{\mu}{2}\|\cdot\|^2$ is convex. For this type of functions, $\operatorname{dom}(f^*) = \mathbb{R}^n$, and ∇f^* is globally $\frac{1}{\mu}-$Lipschitz continuous (see [30, Chapter X, Theorem 4.2.1]), that is, for all $u, v \in \mathbb{R}^n$,

$$\|\nabla f^*(u) - \nabla f^*(v)\| \leqslant \frac{1}{\mu}\|u - v\|.$$

For elementary calculus rules involving conjugate functions, we refer to [30, Chapter X, Proposition 1.3.1]. We summarize, without proof, some of these properties in the next proposition, for future reference.

Proposition 1.1 (Fenchel conjugate calculus rules)**.** *Let $f \in \overline{\operatorname{conv}}(\mathbb{R}^n)$, x_0, $w_0 \in \mathbb{R}^n$, and $\alpha > 0$. Then the following properties hold.*

(i) $(\alpha f)^*(w) = \alpha f^*(w/\alpha)$.

(ii) $(f(\alpha\cdot)^*(w) = f^*(w/\alpha)$.

(iii) $(f(\cdot - x_0))^*(w) = f^*(w) + \langle w, x_0\rangle$.

(iv) $(f(\cdot) + \langle w_0, \cdot\rangle)^*(w) = f^*(w - w_0)$.

(v) $f = f^*$ *if and only if* $f(\cdot) = \dfrac{1}{2}\|\cdot\|^2$.

Regarding optimization problems, we are interested in problems of the form

$$\min_{x\in\mathbb{R}^n} f(x). \tag{1.7}$$

For a nonempty closed set C, we also consider constrained optimization problems with the form

$$\min_{x\in C}\ f(x)\,. \tag{1.8}$$

Constraints can be modeled using indicator functions. More specifically, problem (1.8) is equivalent to

$$\min_{x\in\mathbb{R}^n} f(x) + i_C(x).$$

1.2 Subdifferential concepts

This section presents a brief review of the literature on subdifferentials for nonconvex functions, including calculus and topological properties, as well as the relation of subdifferential with optimization problems.

Consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and $\bar{x} \in \text{dom}(f)$. A point $v \in \mathbb{R}^n$ is called a

1. Fréchet/regular/basic subgradient of f at $\bar{x}$ if

$$\liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geqslant 0.$$

   We denote by $\widehat{\partial} f(\bar{x})$, the Fréchet/regular/basic subdifferential, the set that contains all such subgradients. Equivalently, $v \in \widehat{\partial} f(\bar{x})$ if and only if

$$f(x) \geqslant f(\bar{x}) + \langle v, x - \bar{x} \rangle + o(\|x - \bar{x}\|),$$

   where $\lim_{x \to \bar{x}} \dfrac{o(\|x - \bar{x}\|)}{\|x - \bar{x}\|} = 0.$

2. limiting/general/Mordukhovich subgradient of f at $\bar{x}$ if there exist sequences $\{x^k\} \subseteq \mathbb{R}^n$ and $\{v^k\} \subseteq \mathbb{R}^n$, such that $v^k \in \widehat{\partial} f(x^k)$, $x^k \underset{f}{\to} \bar{x}$, and $v^k \to v$. We denote by $\partial f(\bar{x})$, the limiting/general/ Mordukhovich subdifferential, the set that contains all such subgradients.

3. Clarke subgradient of f at $\bar{x}$ if v can be expressed as the convex combination of points of the form $\lim_{k \to +\infty} \nabla f(x^k)$, where $\mathcal{D} \ni x^k \to \bar{x}$, and $\mathcal{D}$ is any set of Lebesgue measure 0. The set of Clarke subgradients of f at a point $\bar{x}$ is called the Clarke subdifferential, and denoted $\overline{\partial} f(\bar{x})$.

In general, $\widehat{\partial} f(x) \subseteq \partial f(x)$ for $x \in \mathbb{R}^n$. For convex functions, the Fréchet, the limiting, and the Clarke subdifferentials coincide with the subdifferential of Convex Analysis, that is,

$$\widehat{\partial} f(\bar{x}) = \partial f(\bar{x}) = \overline{\partial} f(\bar{x}) = \{v \in \mathbb{R}^n | f(x) \geqslant f(\bar{x}) + \langle v, x - \bar{x} \rangle \text{ for all } x \in \mathbb{R}^n\}.$$

The subdifferential of Convex Analysis is an outer semicontinuous convex-valued mapping. Another important example is when f is (locally) smooth, $\partial f(x)$ reduces to the gradient $\nabla f(x)$.

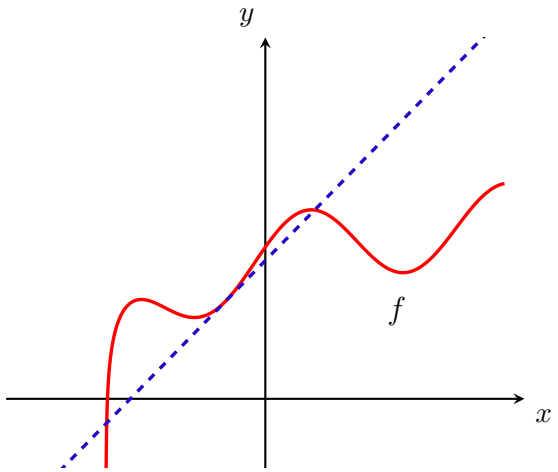
For algorithmic purposes, having a convex-valued subdifferential operator represents an advantage, since convex combinations help us aggregate information of past iterates. In this way, the Clarke subdifferential is preferred for numerical reasons.

Furthermore, the Clarke subdifferential is a well-defined object for a family of functions common in optimization applications. First, recall we say a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is locally Lipschitz continuous around $\bar{x}$, if there exists an open neighborhood U of $\bar{x}$ on which f is Lipschitz continuous, that is, there exists a constant $L_U > 0$, such that for all $x, y \in U$, $|f(x)-f(y)| \leqslant L_U\|x-y\|$. In this case, we say f is L_U-Lipschitz continuous on U. When the estimate holds globally, we simply say f is $L-$Lipschitz continuous. Due to Rademacher's theorem [29, Theorem 9.60], any locally Lipschitz function is almost everywhere differentiable. This property guarantees the Clarke subdifferential to be well-defined [31, Theorem 2.5.1].

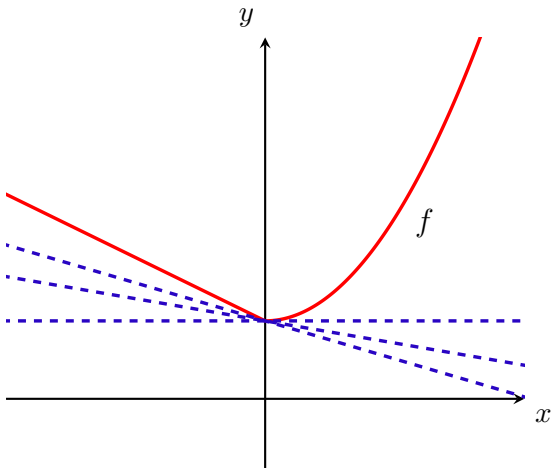
For $x \in \operatorname{dom}(f)$, the Clarke subdifferential $\overline{\partial} f(x)$ of f at x is a nonempty compact convex set [31, Proposition 2.1.2]. Moreover, $\overline{\partial} f$ is an upper semicontinuous mapping; see [31]. Note also that from [32, Proposition 3.1], [32, Theorem 3.6] and Proposition 1.6(iv), the limiting subdifferential ∂f and the Clarke subdifferential $\overline{\partial} f$ coincide for weakly convex functions, the class of interest in Chapter 3.

Figure 1 shows linearizations corresponding to subgradients of different types of functions. Figure 1a presents a nonconvex differentiable function, so that at any point $x \in \mathbb{R}^n$ the subdifferential is the singleton $\nabla f(x)$, although the associated tangent linearization may or may not intersect the graph of f in more than one point. In this particular case, the shown affine function is not a lower linearization, because the function is not convex. Figure 1b shows a convex function nondifferentiable at the origin, and the subgradients are computed at this point. In this case, the subdifferential of the function at $\bar{x} = 0$ is non-singleton set, because the function has a *kink* at $\bar{x} = 0$: $\partial f(\bar{x}) = [-1, 0]$. In view of the convexity of the function, any tangent is a lower linearization of the function that locally approximates it. Figure 1c presents a nonconvex function with (Fréchet) subgradients computed at a point of nondifferentiability, $\bar{x} = 1$. Once again, due to the existence of a kink at this point, the subdifferential of the funcion at $\bar{x} = 1$ is not a singleton: $\partial f(\bar{x}) = [-2, 2]$. Note that some subgradients may define lower tangent linearizations, and others may not. More on subdifferentials of this type of functions can be found in Section 1.3. Finally, Figure 1d exhibits a subgradient of a convex function everywhere differentiable. In this case, the subdifferential is a singleton at every point, the gradient of the function at such point, and all the tangent linearizations lie below the graph of the function.

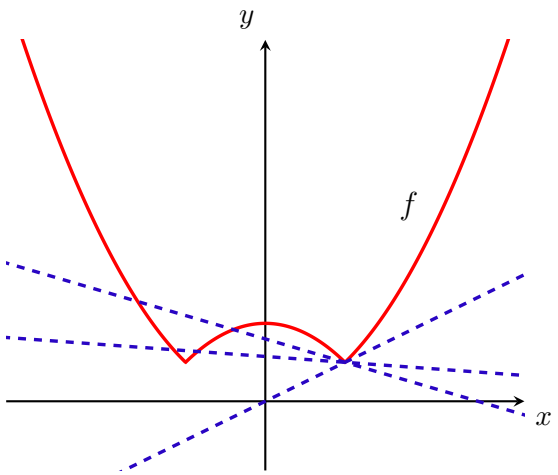
For convex functions, it is well known that global minimizers can be characterized using the subdifferential. More specifically, $\bar{x}$ is a (global) minimizer of $f \in \overline{\operatorname{conv}}(\mathbb{R}^n)$ if and only if $0 \in \partial f(\bar{x})$. For nonconvex functions, this inclusion may fail to characterize global, and even local, minimizers. Nevertheless, zeros of subdifferentials play a central role in nonconvex optimization, since they extend the notion of optimality. For a function



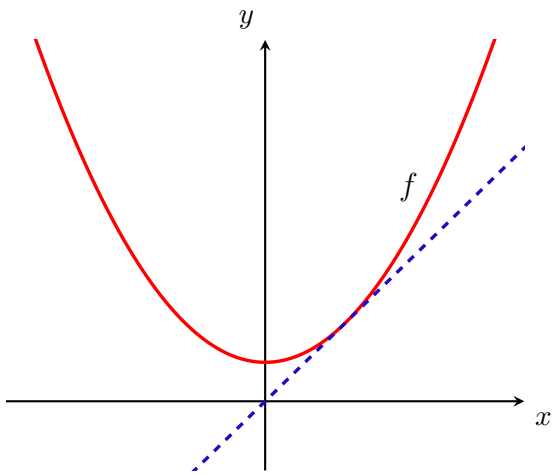
(a) Nonconvex differentiable



(b) Convex nondifferentiable



(c) Nonconvex nondifferentiable



(d) Convex differentiable

Figure 1 – Subgradients of functions. The continuous red line represents the graph of the function, and the dashed blue lines are the linearizations associated with a subgradient at a given point.

$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $\bar{x} \in \mathbb{R}^n$ is said to be a limiting (resp. Clarke) critical point, or simply critical point, of f if $0 \in \partial f(\bar{x})$ (resp. $0 \in \overline{\partial} f(\bar{x})$). Denote by $S := (\partial f)^{-1}(0)$ the set of all critical points of f.

In general, $0 \in \partial f(\bar{x})$ is a necessary condition for $\bar{x}$ to be a local minimizer. Indeed, if there exists $\delta > 0$, such that for all $x \in B(\bar{x}, \delta) \backslash \{\bar{x}\}$, $f(\bar{x}) \leqslant f(x)$, then

$$\frac{f(x) - f(\bar{x}) - \langle 0, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geqslant 0,$$

thus 0 is a Fréchet subgradient of f at $\bar{x}$, and thus, a limiting subgradient of f at $\bar{x}$.

For convex functions, all critical points/local minimizers are global minimizers. Therefore, all critical points have the same critical value, the (global) optimal value. The following property is a local generalization of this idea for the nonconvex case. This prop-

erty is very natural; we refer the readers to [33] for a discussion and sufficient conditions for it to hold.

Definition 1.1 (Proper separation of isocost surfaces)**.** *A lsc function $f : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ has properly separated isocost surfaces if there exists $\varepsilon > 0$ such that*

$$\bar{x}, \bar{y} \in S,\ f(\bar{x}) \neq f(\bar{y}) \implies \|\bar{x} - \bar{y}\| \geqslant \varepsilon\,,$$

where $S = (\partial f)^{-1}(0)$ is the set of critical points of f.

In other words, f satisfies the proper separation of isocost surfaces property if for sufficiently close critical points $\bar{x}$ and $\bar{y}$, the corresponding critical values are the same, namely, $f(\bar{x}) = f(\bar{y})$.

The Fréchet and the limiting subdifferentials satisfy some essential calculus rules. These rules are useful in Section 1.3.2, and are excerpted from [29, Exercise 8.8].

Proposition 1.2 (Subdifferential calculus rules)**.** *Consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, a point $\bar{x} \in \mathrm{dom}(f)$, and a function $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ smooth around $\bar{x}$. Then*

$$\partial(f + g)(\bar{x}) = \partial f(\bar{x}) + \nabla g(\bar{x}),$$

and the same holds for the Fréchet subdifferential $\widehat{\partial}$. In particular, $\widehat{\partial} g(\bar{x}) = \partial g(\bar{x}) = \{\nabla g(\bar{x})\}$.

Differently from Convex Analysis, the subdifferential of a nonconvex function can present a degeneracy in the following sense: for $x^k \underset{f}{\to} \bar{x}$, and $v^k \in \widehat{\partial} f(x^k)$, the sequence $\{v^k\}$ might be unbounded. For instance, consider the non-locally Lipschitz continuous function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^\alpha$, for $\alpha \in (0, 1)$. For $x \neq 0$, $\nabla f(x) = \alpha x^{\alpha-1}$. Given $x^k = k^{-1}$ for $k \in \mathbb{N}$, $f(x^k) = k^{-\alpha}$, and thus $x^k \underset{f}{\to} 0$. On the other hand, $v^k = \nabla f(x^k) = \alpha k^{1-\alpha} \to +\infty$ (cf. Proposition 1.3 3.)

This feature makes it necessary to exclude the *recession* directions, that is, those possible unbounded directions in the subdifferential. For that reason, we introduce the horizon subdifferential.

Definition 1.2 (Horizon subdifferential)**.** *For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and $\bar{x} \in \mathrm{dom}(f)$, a point $v \in \mathbb{R}^n$ is called a horizon/singular subgradient of f at $\bar{x}$ if there exists a sequence $\{x^k\} \subseteq \mathbb{R}^n$, such that $x^k \underset{f}{\to} \bar{x}$, a sequence $\{v^k\} \subseteq \mathbb{R}^n$, such that for all $k \in \mathbb{N}$, $v^k \in \widehat{\partial} f(x^k)$, and a real sequence $\{\lambda^k\} \subseteq \mathbb{R}_+$, such that $\lambda^k \downarrow 0$, satisfying $\lambda^k v^k \to v$. The set $\partial^\infty f(\bar{x})$, called the horizon/singular subdifferential, contains all the horizon/singular subgradients of f at $\bar{x}$.*

The following proposition summarizes some of the topological properties of the horizon subdifferential, extracted from [29, Theorem 8.6, Theorem 8.7, Exercise 8.8, Theorem 9.3].

Proposition 1.3 (Properties of the horizon subdifferential)**.** *Let a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and a point $\bar{x} \in \operatorname{dom}(f)$. Then, the following hold.*

1. *$\partial^\infty f(\bar{x})$ is a closed cone in $\mathbb{R}^n$.*
2. *$\partial^\infty f$ is an outer semicontinuous operator at $\bar{x}$ with respect to the $f-$ attentive convergence.*
3. *If f is locally Lipschitz around $\bar{x}$, $\partial^\infty f(\bar{x}) = \{0\}$.*

Additionally, if f_0 is a smooth function in a neighborhood of $\bar{x}$, and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ a function that is finite at $\bar{x}$, such that $f = f_0 + g$, then

4. *$\partial^\infty f(\bar{x}) = \partial^\infty g(\bar{x})$. In particular, $\partial^\infty f_0(\bar{x}) = \{0\}$.*

The horizon subdifferential possesses a geometric interpretation in connection with normal directions to the epigraph of the function, as proven in [29, Theorem 8.9]. First, we introduce the notion of normal cones in the nonconvex case, and their relations with the previously defined subdifferentials.

Proposition 1.4 (Normal cones through subdifferentials)**.** *Let a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and a point $\bar{x} \in \operatorname{dom}(f)$. Then,*

1. *$\widehat{\partial} f(\bar{x}) = \{v \in \mathbb{R}^n | (v, -1) \in \widehat{N}_{\operatorname{epif}(f)}(\bar{x}, f(\bar{x}))\}$, where for any set $C \subseteq \mathbb{R}^m$, the Fréchet/regular normal cone to C at $\bar{y} \in C$ is defined as*

$$\widehat{N}_C(\bar{y}) = \left\{ u \in \mathbb{R}^m \middle| \limsup_{\substack{y \to \bar{y} \\ y \in C \setminus \{\bar{y}\}}} \frac{\langle u, y - \bar{y} \rangle}{\|y - \bar{y}\|} \leqslant 0 \right\}.$$

2. *$\partial f(\bar{x}) = \{v \in \mathbb{R}^n | (v, -1) \in N_{\operatorname{epif}(f)}(\bar{x}, f(\bar{x}))\}$, where for any set $C \subseteq \mathbb{R}^m$, the limiting/general normal cone to C at $\bar{y} \in C$ is defined as*

$$N_C(\bar{y}) = \{u \in \mathbb{R}^m | \ \exists C \ni y^k \to \bar{y},\ \widehat{N}_C(x^k) \ni u^k \to u\}.$$

If $\bar{x} \notin C$, then $\widehat{N}_C(\bar{x}) = N_C(\bar{x}) = \varnothing$.

For a nonempty closed convex set $C \subseteq \mathbb{R}^n$, the normal cone to C at $\bar{x} \in C$ of Convex Analysis coincides with the Fréchet and limiting normal cones, namely, $\partial i_C(x) = \{v \in \mathbb{R}^n : \langle v, x - \bar{x}\rangle \leqslant 0 \text{ for all } x \in C\}$, coincides with the Fréchet and normal cones of i_C at x.

As proven in [29, Theorem 8.9], horizon subgradients correspond to horizontal normal vectors to the epigraph of the function. For convex functions, the horizon subdifferential takes a special form [29, Proposition 8.12], due to the characterization of normal cones of convex sets.

Proposition 1.5 (Geometry of horizon subdifferentials)**.** *Let a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and a point $\bar{x} \in \operatorname{dom}(f)$. Suppose f is locally lsc at $\bar{x}$, then*

$$\partial^\infty f(\bar{x}) = \{v \in \mathbb{R}^n | (v, 0) \in N_{\operatorname{epif}(f)}\big(\bar{x}, f(\bar{x})\big)\}.$$

If, in addition, f is convex, then $\partial^\infty f(\bar{x}) = N_{\operatorname{dom}(f)}(\bar{x})$.

Note that if $\bar{x} \in \operatorname{int}(\operatorname{dom}(f))$, then $\partial^\infty f(\bar{x}) = \{0\}$. This result is consistent with the fact that for such $\bar{x}$, any function $f \in \overline{\operatorname{conv}}(\mathbb{R}^n)$ is locally Lipschitz continuous around $\bar{x}$ (cf. Proposition 1.3 3.).

One of the applications of the horizon subdifferential is it allows generalizing constraint qualifications. It also provides suitable conditions for subdifferential calculus rules to hold. The following result illustrates one of these applications, corresponding to [29, Theorem 8.15].

Theorem 1.1 (Optimality conditions)**.** *For a proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, and a nonempty closed set $C \subseteq \mathbb{R}^n$, consider problem (1.8). Suppose there exists $\bar{x} \in C$ satisfying the following linear regularity condition:*

$$\partial^\infty f(\bar{x}) \cap -N_C(\bar{x}) = \{0\}.$$

Then, the necessary local optimality condition for $\bar{x}$ is

$$0 \in \partial f(\bar{x}) + N_C(\bar{x}).$$

When f and C are convex, the condition is sufficient for global optimality, without the need for the linear regularity condition to hold.

1.3 The class of weakly convex functions

In this section, we examine the family of weakly convex functions, and provide some examples. Then, we proceed to define a new subdifferential for weakly convex

functions, the approximate subdifferential, using the *convexification* of a weakly convex function. We end this section proving a variational principle for weakly convex functions, and a continuity property of the approximate subdifferential as an application of this principle.

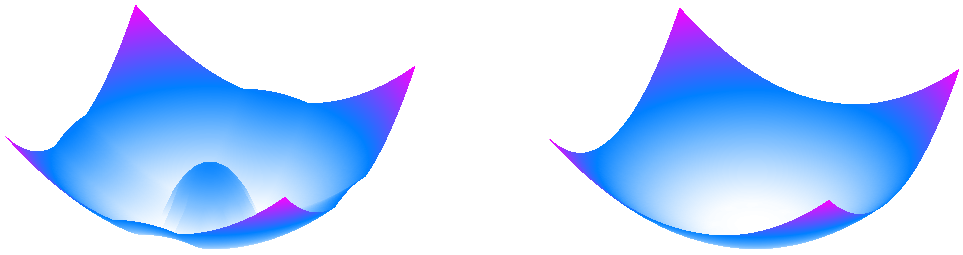
1.3.1 The concept of weak convexity

Weakly convex functions appear naturally in applications. For example, in phase retrieval problems, where the loss function can be chosen to be the ℓ^1-norm [34], or the ℓ^2-norm [35]; and in compressive sensing problems with bilinear/biconvex objectives [36] with a ℓ^1-penalty to induce sparsity. See [37] for more examples.

This subsection on weak convexity is extracted from [22], dealing with the concept of weak convexity and some basic properties.

Definition 1.3 (Weakly convex functions)**.** *We say that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is $\rho-$weakly convex, for $\rho > 0$, if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is a convex function.*

We denote by $w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$ the set of proper lsc $\rho-$weakly convex functions from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$. Figure 2 shows the graph of a function $f \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$ and its "convexification" after adding a quadratic term. In this case, $f : \mathbb{R}^2 \to \mathbb{R}$ is given by $f(x, y) = |x^2 - 1| + |y^2 - 1|$. Note that the convexification no longer presents a "hill" around $(x, y) = (0, 0)$, because the nonconvexity was "repaired".



(a) Graph of f. (b) Graph of $f(\cdot) + (\rho/2)\|\cdot\|^2$.

Figure 2 – Graph of the weakly convex function $f(x, y) = |x^2 - 1| + |y^2 - 1|$ and its "convexification".

The class of weakly convex functions is contained in some larger classes of nonsmooth functions, such as the generalized differentiable functions in the sense of Norkin [38], or the semismooth functions [39]. The following are some equivalent characterizations of weak convexity; see [40, Theorem 2.1], [41, Theorem 3.1].

Proposition 1.6 (Alternative characterizations of weak convexity)**.** *For a proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $\rho > 0$, the following statements are equivalent:*

(i) For any $z \in \mathbb{R}^n$, $f(\cdot) + \dfrac{\rho}{2}\|\cdot - z\|^2$ *is a convex function.*

(ii) For any $x, y \in \mathbb{R}^n$, *such that* $\widehat{\partial} f(y) \neq \varnothing$, *any* $g(y) \in \partial f(y)$ *satisfies*

$$f(y) + \langle g(y), x - y\rangle \leqslant f(x) + \frac{\rho}{2}\|x - y\|^2$$

or, equivalently,

$$\ell_{y,g(y)}(x) \leqslant f(x) + \frac{\rho}{2}\|x - y\|^2 ,$$

where $\ell_{y,g(y)}(\cdot) := f(y) + \langle g(y), \cdot - y\rangle$ *is the linearization of* f *at the point* y.

(iii) For all $x, y \in \mathbb{R}^n$, *and* $\lambda > 0$,

$$f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda) f(y) + \frac{\rho\lambda(1-\lambda)}{2}\|x - y\|^2 .$$

Note that in Proposition 1.6(i), by taking $z = 0$, we retrieve Definition 1.3, which means that f is convex up to a quadratic perturbation. Proposition 1.6(i) is completely equivalent to this way of defining weakly convex functions, since it states that f is convex up to a quadratic perturbation with a linear term. Regarding some other notions of nonconvexity in the literature, it is important to note that for a function to be weakly convex, Proposition 1.6(ii) must hold for all subgradients at all points. By contrast, for prox-regular functions [29, Definition 13.27], also known as lower-$\mathcal{C}^2$ functions, the inequality holds only locally for subgradients, points and functional values. As a result, weak convexity is equivalent to the function being prox-regular everywhere, and the parameter of prox-regularity being the same for all points, or simply uniformly prox-regular.

As already commented, the class of weakly convex functions is quite broad and includes many settings of interest, whose nonconvexity is *benign*, in the parlance of [42]. One example is the class of decomposable functions in [43], that contains max-functions, maximal eigenvalue functions, and norm-1 regularized functions; see also [44] and [45]. The following definition corresponds to the global version of a decomposable function of [43].

Definition 1.4 ($h \circ c$ decomposable functions)**.** *Given a continuously differentiable mapping* $c : \mathbb{R}^n \to \mathbb{R}^m$ *such that* $c(\bar{x}) = 0$, *and a finite-valued sublinear function* $h : \mathbb{R}^m \to \mathbb{R}$, *the real-valued function* f *is* $h \circ c$ *decomposable at* $\bar{x} \in \mathbb{R}^n$, *if for all* $x \in \mathbb{R}^n$,

$$f(x) = f(\bar{x}) + h(c(x)) .$$

If c is a $\mathcal{C}^1$ function with Lipschitz-continuous Jacobian, then such f is weakly convex. To see this, apply [46, Lemma 4.2]. Since h is finite-valued and sublinear, it is

then convex and Lipschitz-continuous (see [30, V(1.2.6)]), while c is $\mathcal{C}^1$ with Lipschitz-continuous Jacobian from the assumptions. Therefore the composition $h \circ c$ and, hence, the function $f(\cdot) = f(\bar{x}) + h \circ c(\cdot)$, are weakly convex.

In association with other notions related to weak convexity, we further remark that all real-valued prox-regular functions (or, in our terminology, real-valued locally weakly convex functions) can also be locally decomposed as a sum of a convex continuous function and a concave quadratic function (in line with Definition 1.3), and can also be expressed as a composition of a convex continuous function with a differentiable function with locally Lipschitz gradient, see [41, Proposition 3.5, Remark 3.6].

We next give an example of weak convexity for extended real-valued functions, that will play a role in Section 3.4 to include the class of feasible descent methods of [33] (for constrained optimization) into the convergence theory developed in Chapter 3.

Proposition 1.7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable, such that the gradient ∇f is $L-$Lipschitz continuous on the nonempty closed convex set $X \subseteq \mathbb{R}^n$. Then, $f + i_X$ is a $L-$weakly convex function.*

Proof. Since f has Lipschitz-continuous gradient with constant L on X, then (e.g., from [47, Lemma A.11]), for all $x, y \in X$ it holds that

$$f(y) + \langle \nabla f(y), x - y \rangle - \frac{L}{2}\|x - y\|^2 \leqslant f(x)\,.$$

Furthermore, for $x \in X$, and $y \in \mathbb{R}^n$ such that $\partial(f + i_X)(y) \neq \varnothing$, that is, for $y \in X$, and for all $w \in N_X(y)$, we have that $\nabla f(y) + w \in \partial(f + i_X)(y)$, and

$$(f + i_X)(y) + \langle \nabla f(y) + w, x - y \rangle - \frac{L}{2}\|x - y\|^2 \leqslant (f + i_X)(x)\,.$$

If $x \notin X$, the above inequality holds trivially, because y needs to be an element of X to ensure that the subdifferential $\partial(f + i_X)(y)$ is nonempty (see Proposition 1.6(ii)). Therefore, $f + i_X$ is $L-$weakly convex. □

1.3.2 Approximate subdifferentials of weakly convex functions

This subsection introduces a novel concept, an approximate subdifferential for weakly convex functions. First, we present the definition of the classical $\varepsilon-$subdifferential of Convex Analysis, the basis of our new concept.

Definition 1.5 ($\varepsilon-$subdifferential – convex case)**.** *For $f \in \overline{\text{conv}}(\mathbb{R}^n)$, a point $\bar{x} \in \text{dom}(f)$, and $\varepsilon \geqslant 0$, a vector $v \in \mathbb{R}^n$ is called an $\varepsilon-$subgradient of f at $\bar{x}$ if for all $x \in \mathbb{R}^n$,*

$$f(x) \geqslant f(\bar{x}) + \langle v, x - \bar{x} \rangle - \varepsilon.$$

The set comprising all such vectors is called the $\varepsilon-$subdifferential of f at $\bar{x}$, and denoted $\partial_\varepsilon f(\bar{x})$.

We introduce the novel concept of approximate subdifferencial for weakly convex functions. For this purpose, for a $\rho-$weakly convex function f, denote by F the "convexification" of f centered at $\bar{x} \in \mathrm{dom}(f)$, namely, $F_\rho^{\bar{x}} = f(\cdot) + \frac{\rho}{2}\|\cdot - \bar{x}\|^2$.

Definition 1.6 (Approximate subdifferential – weakly convex case)**.** *For a function $f \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$, a point $\bar{x} \in \mathrm{dom}(f)$, and $\varepsilon \geqslant 0$, the $\varepsilon-$approximate subdifferential of f centered $\bar{x}$ is the multivalued operator given by*

$$\partial^{\bar{x}}_{\rho,\varepsilon} f(x) = \partial_\varepsilon F^{\bar{x}}_\rho(x) - \rho(x - \bar{x}).$$

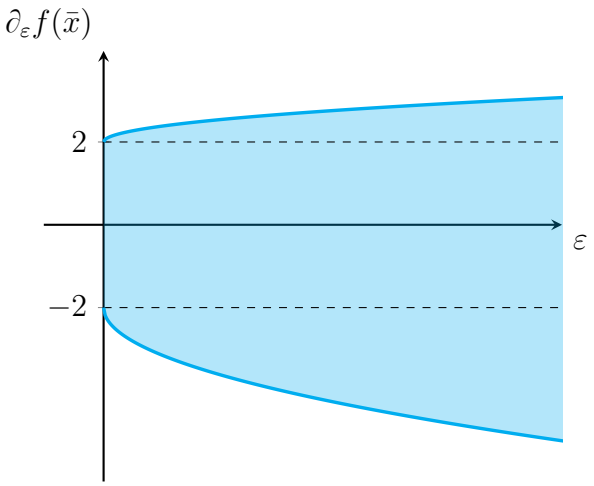


Figure 3 – Approximate subdifferential $\partial^{\bar{x}}_\varepsilon f(\bar{x})$ for $f(x) = |x^2 - 1|$, $\bar{x} = -1$, and varying $\varepsilon \in [0, 4]$.

Figure 3 illustrates the approximate subdifferential of the weakly convex function $f(x) = |x^2 - 1|$, centered at $\bar{x} = -1$ and $\rho = \frac{5}{2}$. Note that $\partial f(\bar{x}) = [-2, 2]$ ($\varepsilon = 0$). In this case,

$$\partial^{\bar{x}}_{\rho,\varepsilon} f(\bar{x}) = [-2 - \sqrt{2\varepsilon(\rho + 2)}, 2 + \sqrt{2\varepsilon(\rho - 2)}].$$

Observe that when $\rho > 2$, then $\partial^{\bar{x}}_{\rho',\varepsilon} f(\bar{x})$ is an enlargement of the subdifferential for any $\rho' > \rho$.

Remark 1.1. *This definition is inspired by Proposition 1.2, because $\partial F^{\bar{x}}(x) = \partial f(x) + \rho(x - \bar{x})$. Additionally, taking $\varepsilon = 0$ in Definition 1.6 yields $\partial^{\bar{x}}_{\rho,0} f(x) = \partial F^{\bar{x}}_\rho(x) - \rho(x - \bar{x})$.*

We now drop the subindex ρ to denote the approximate subdifferential of a $\rho-$weakly convex function, when there is no confusion from the context.

Note that the approximate subdifferential is just an affine translation of the ε-subdifferential in Convex Analysis. Therefore, all the topological and calculus rules

developed for the latter in [30, Ch. XI] are available for the former. Some basic properties are listed in the following result.

Proposition 1.8 (Properties of the approximate subdifferential)**.** *For a function $f \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$, a center $\bar{x} \in \text{dom}(f)$, and $\varepsilon \geqslant 0$, the following holds.*

1. *If $x \in \text{dom}(f)$ and $\varepsilon \geqslant 0$, $\partial_\varepsilon^{\bar{x}} f(x)$ is nonempty closed and convex.*
2. *If $0 \leqslant \varepsilon \leqslant \varepsilon'$, then $\partial_\varepsilon^{\bar{x}} f(x) \subseteq \partial_{\varepsilon'}^{\bar{x}} f(x)$.*
3. *If $x \in \text{int}(\text{dom}(f))$ and $\varepsilon \geqslant 0$, $\partial_\varepsilon^{\bar{x}} f(x)$ is bounded.*
4. *If $0 \in \partial_\varepsilon^{\bar{x}} f(\bar{x})$, then for any $\varepsilon' > \varepsilon$, $\bar{x}$ is an $\varepsilon'-$local minimizer of f.*

Proof. Let us proof statement 4, since $1-3$ are direct from the definition. If $0 \in \partial_\varepsilon^{\bar{x}} f(\bar{x})$, then by definition, $0 \in \partial_\varepsilon F^{\bar{x}}(\bar{x})$. Therefore, $\bar{x}$ is a global minimizer of the convex function $F^{\bar{x}}$: for all $x \in \mathbb{R}^n$,

$$f(\bar{x}) \leqslant f(x) + \varepsilon + \frac{\rho}{2}\|x - \bar{x}\|^2.$$

In particular, for any $\delta > 0$ and $x \in B(\bar{x}, \delta)$,

$$f(\bar{x}) \leqslant f(x) + \varepsilon + \frac{\rho}{2}\delta^2,$$

and the result follows by taking $\varepsilon' = \varepsilon + \frac{\rho}{2}\delta^2$.

□

Variational principles seek to characterize properties of a point that can be extended to a neighborhood. Applications of these principles ca be found in optimization, partial differential equations, equilibrium problems, among others. For example, one of the most acclaimed variational principles, Ekeland's variational principle [29, Proposition 1.43], states that approximate minimizers of functions are close to minimizers of a perturbed problem. Along the same lines, we find the Brøndsted-Rockafellar theorem [48], used in [19] and Chapter 3 to study the convergence of $\varepsilon-$subgradient descent methods.

The next result follows this idea using the concept of approximate subdifferential for weakly convex functions, not only characterizing proximity of points, and subgradients, but also of function values.

Theorem 1.2 (Variational principle)**.** *Given $f \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$, $\alpha > 0$, $\varepsilon \geqslant 0$, consider $\bar{x} \in \text{dom}(f)$, and $v \in \partial_\varepsilon^{\bar{x}} f(x)$. Then, there exist $x_\varepsilon \in \mathbb{R}^n$ and $v_\varepsilon \in \mathbb{R}^n$ such that*

$$v_\varepsilon \in \partial f(x_\varepsilon) + \rho(\sqrt{\varepsilon} + \|x - \bar{x}\|)B(0,1). \tag{1.9}$$

Furthermore, there exists $\gamma \in [-1,1]$*, such that*

$$\begin{aligned} \|x_\varepsilon - x\| + \alpha|\langle v, x_\varepsilon - x\rangle| &\leqslant \sqrt{\varepsilon} \\ \|v_\varepsilon - (1+\alpha\gamma\sqrt{\varepsilon})v\| &\leqslant \sqrt{\varepsilon} \\ |\langle v_\varepsilon - v, x_\varepsilon - x\rangle| &\leqslant \varepsilon \\ |\langle v_\varepsilon, x_\varepsilon - x\rangle| &\leqslant \varepsilon + \alpha^{-1}\sqrt{\varepsilon} \\ |f(x_\varepsilon) - f(x)| &\leqslant (1+\rho)\varepsilon + \alpha^{-1}\sqrt{\varepsilon} + \frac{3\rho}{2}\|x - \bar{x}\|^2. \end{aligned} \tag{1.10}$$

Proof. Apply [49, Proposition 1.1] to $F^{\bar{x}} \in \overline{\text{conv}}(\mathbb{R}^n)$ and $v \in \partial_\varepsilon F^{\bar{x}}(x)$, obtaining the existence of a pair $(x_\varepsilon, v_\varepsilon)$ such that $v_\varepsilon \in \partial F^{\bar{x}}(x_\varepsilon)$, for which the first four estimates in (1.10) hold, and

$$|F^{\bar{x}}(x_\varepsilon) - F^{\bar{x}}(x)| \leqslant \varepsilon + \alpha^{-1}\sqrt{\varepsilon}. \tag{1.11}$$

Furthermore, from the definition $F^{\bar{x}}$

$$v_\varepsilon \in \partial f(x_\varepsilon) + \rho(x_\varepsilon - \bar{x}), \tag{1.12}$$

with

$$\begin{aligned} \|x_\varepsilon - \bar{x}\| &\leqslant \|x_\varepsilon - x\| + \|x - \bar{x}\| \\ &\leqslant \sqrt{\varepsilon} + \|x - \bar{x}\|, \end{aligned}$$

where we use the triangle inequality in the first line, and the first estimate of (1.10) in the second inequality. Thus, (1.9) follows.

Furthermore, from the triangle inequality, we obtain

$$\begin{aligned} |f(x_\varepsilon) - f(x)| - \frac{\rho}{2}\|x_\varepsilon - \bar{x}\|^2 - \frac{\rho}{2}\|x - \bar{x}\|^2 &\leqslant \left| f(x_\varepsilon) - f(x) + \frac{\rho}{2}\|x_\varepsilon - \bar{x}\|^2 - \frac{\rho}{2}\|x - \bar{x}\|^2 \right| \\ &= |F^{\bar{x}}(x_\varepsilon) - F^{\bar{x}}(x)| \\ &\leqslant \varepsilon + \alpha^{-1}\sqrt{\varepsilon}, \end{aligned}$$

where we use the definition of $F^{\bar{x}}$ in the second line, and (1.11) in the third line. This yields

$$|f(x_\varepsilon) - f(x)| \leqslant \varepsilon + \alpha^{-1}\sqrt{\varepsilon} + \frac{\rho}{2}\|x_\varepsilon - \bar{x}\|^2 + \frac{\rho}{2}\|x - \bar{x}\|^2.$$

Moreover, from the triangle inequality, we also have

$$\begin{aligned} \|x_\varepsilon - \bar{x}\|^2 &\leqslant \left(\|x_\varepsilon - x\| + \|x - \bar{x}\|\right)^2 \\ &\leqslant \left(\sqrt{\varepsilon} + \|x - \bar{x}\|\right)^2 \\ &\leqslant 2\varepsilon + 2\|x - \bar{x}\|^2, \end{aligned}$$

where the second inequality follows from the first estimate in (1.10). Hence,

$$|f(x_\varepsilon) - f(x)| \leqslant \varepsilon + \alpha^{-1}\sqrt{\varepsilon} + \rho\varepsilon + \left(\rho + \frac{\rho}{2}\right)\|x - \bar{x}\|^2.$$

This last inequality corresponds to the fifth estimate in (1.10). □

The previous variational principle can be employed to prove an outer semicontinuity property for the subdifferential of weakly convex functions.

Corollary 1.1. *Given $f \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$, consider $\bar{x} \in \mathrm{dom}(f)$. Then, there exist $\{x^k\} \subseteq \mathbb{R}^n$ and $\{v^k\} \subseteq \mathbb{R}^n$, such that $v^k \in \partial f(x^k) + \frac{\rho}{k}B(0,1)$, $x^k \underset{f}{\to} \bar{x}$, and $\langle v^k, x^k - \bar{x}\rangle \to 0$. Moreover, any cluster point of the sequence $\{v^k\}$ satisfying these properties, whenever they exist, belongs to $\partial f(\bar{x})$.*

Proof. Similarly as in the proof of [49, Corollary 1.2], for each $k \in \mathbb{N}\backslash\{0\}$, apply Theorem 1.2 with $\alpha = 1$, $\varepsilon_k = \frac{1}{k^2}$, and $\bar{x} = x$, to some $u^k \in \partial^{\bar{x}}_{\varepsilon_n} f(\bar{x})$, obtaining sequences $\{x^k\}$ and $\{v^k\}$, such that $v^k \in \partial f(x^k) + \frac{\rho}{k}B(0,1)$, and

$$\begin{aligned} \|x^k - \bar{x}\| + |\langle u^k, x^k - \bar{x}\rangle| &\leqslant \frac{1}{k} \\ |f(x^k) - f(\bar{x})| &\leqslant (1+\rho)\frac{1}{k^2} + \frac{1}{k}, \\ |\langle v^k, x^k - \bar{x}\rangle| &\leqslant \frac{1}{k^2} + \frac{1}{k}. \end{aligned}$$

Taking the limit as $k \to +\infty$ yields the convergence properties. Let $v^\star$ be a cluster point of $\{v^k\}$. From (1.12), $v^k \in \partial F^{\bar{x}}(x^k)$, any $x \in \mathrm{dom}(f)$,

$$\begin{aligned} F^{\bar{x}}(x) &\geqslant \limsup_{k\to+\infty} \left\{F^{\bar{x}}(x^k) + \langle v^k, x - x^k\rangle\right\} \\ &= \limsup_{k\to+\infty} \left\{F^{\bar{x}}(x^k) + \langle v^k, x - \bar{x}\rangle + \langle v^k, \bar{x} - x^k\rangle\right\} \\ &\geqslant \liminf_{k\to+\infty} \left\{F^{\bar{x}}(x^k) + \langle v^k, x - \bar{x}\rangle\right\} + \limsup_{k\to+\infty} \langle v^k, \bar{x} - x^k\rangle \\ &= \liminf \left\{F^{\bar{x}}(x^k) + \langle v^k, x - \bar{x}\rangle\right\} \\ &\geqslant F^{\bar{x}}(x) + \langle v^\star, x - \bar{x}\rangle, \end{aligned}$$

where in the second inequality we use algebra of lim inf and lim sup, the fact that $\langle v^k, \bar{x} - x^k\rangle \to 0$ in the second equality, and the identity $F^{\bar{x}}(\bar{x}) = f(\bar{x})$, $x^k \underset{f}{\to} \bar{x}$, and $v^k \to v^\star$ (up to a subsequence, if necessary), in the last line. Therefore, $v^\star \in \partial F^{\bar{x}}(\bar{x}) = \partial f(\bar{x})$. □

1.4 Error bounds and Kurdyka-Łojasiewicz inequality

Error bounds are upper estimates of the distance to solutions (or critical points) of a given optimization problem. Their role is paramount for various reasons, among which is convergence rate analyses; see, e.g., [50, 51, 47, 52], also Proposition 2.3, and Chapters 3–6. The Kurdyka-Łojasiewicz inequality establishes that for a certain family of functions, up to a reparametrization, such functions are *sharp* around critical points. In optimization, the Kurdyka-Łojasiewicz inequality plays a similar role as error bounds to shoe rates of convergence.

1.4.1 Subdifferential-based error bound

This subsection on error bounds is extracted from the introduction of [22]. Some of the notation is modified in order to comply with the rest of the text.

In this work, we shall mostly employ the following subdifferential-based error bound. See, however, Definition 1.8 for the so-called *natural residual* error bound [51] for constrained problems, and its relation with the subdifferential-based bound.

Definition 1.7 (Subdifferential error bound)**.** *We say the subdifferential error bound holds for problem* (1.7) *where* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is bounded below, if for every* $\bar{f} \geqslant \inf_{x \in \mathbb{R}^n} f(x)$*, there exist* $\epsilon, \ell > 0$ *such that whenever* $x \in \mathbb{R}^n$*,* $f(x) \leqslant \bar{f}$*, and* $w \in \partial f(x) \cap B(0, \epsilon)$*, the following is true:*

$$\operatorname{dist}(x, S) \leqslant \ell \|w\| \,,$$

where $S = (\partial f)^{-1}(0)$ *is the set of critical points of* f*.*

The error bound above is related to various other notions that appear in the literature, such as the Kurdyka-Łojasiewicz inequality [53, 54], and quadratic growth of f around the set of its critical points [55, 56], or the set of minimizers when the function is convex [57, 58] (see Section 1.4.3 below). These conditions assure some regularity of the function near a critical point. Furthermore, the subdifferential error bound is related to metric subregularity of ∂f (see Definition 1.13).

We next turn our attention to constrained smooth optimization problems, the framework of [33], dealt with in Section 3.4. Consider the problem

$$\min_{x \in C} \; f(x) \,, \tag{1.13}$$

where C is a closed convex set, and $f : \mathbb{R}^n \to \mathbb{R}$ is finite-valued and smooth. An equivalent problem is to handle constraints by adding to f the indicator function of the feasible set. It turns out that these two equivalent formulations are in fact different when it comes to error bounds, and some subtle issues arise.

Specifically, as is well known, criticality of a point x in the sense of

$$0 \in \partial(f + i_C)(x) = \nabla f(x) + N_C(x)$$

is equivalent to the condition

$$x - P_C(x - \nabla f(x)) = 0 \,.$$

Hence, one can attempt to measure the distance to the set of critical points S by the violation of the projection equality above, or by the violation of the subdifferential inclusion

above. It so happens that, at least in general, these are not the same. We next review the relations between the corresponding error bounds.

The subdifferential error bound would just read exactly the same as in Definition 1.7, using $f + i_C$ instead of f therein (then $w \in \nabla f(x) + N_C(x)$). The projection-based error bound states the following.

Definition 1.8 (Projection error bound)**.** *We say the projection error bound holds for problem (1.13) where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and bounded below, if for every $\bar{f} \geqslant \inf_{x \in C} f(x)$, there exist $\epsilon, \ell > 0$ such that whenever $x \in C$, $\|x - P_C(x - \nabla f(x))\| \leqslant \epsilon$, and $f(x) \leqslant \bar{f}$, the following is true:*

$$\operatorname{dist}(x, S) \leqslant \ell \|x - P_C(x - \nabla f(x))\| .$$

The projection error bound is a natural way to measure violation of stationarity in convexly-constrained problems, used in many developments; see, e.g., [33, 59, 60].

Clearly, for problem (1.13) with smooth f, Definition 1.7 and Definition 1.8 amount to the same if $C = \mathbb{R}^n$ (or if S is in the interior of C). For constrained problems, there are two cases when these error bounds are equivalent. The first one is when the critical point is isolated, see [51, Proposition 6.2.4], [47, Proposition 1.31]. In that case, the projection error bound means the semistability property [47, Definition 1.29]. The second one when the two bounds are equivalent is when C is a generalized box in $\mathbb{R}^n$, i.e., C is defined by bound constraints on the variables (some bounds can be infinite), see [61, Theorem 2]. To the best of our knowledge, in other settings the relations between the subdifferential and projection error bounds are not known. However, the following simple argument shows that when the gradient of f is Lipschitz-continuous, the projection residual is bounded above by a multiple of $\operatorname{dist}(x, S)$, always. Then, if the subdifferential error bound holds, the right-hand side of (1.14) is of order no less than the projection residual. Hence, in principle, the subdifferential error bound can hold when the projection variant does not. Note that this is meant as merely a side observation, to add to the discussion of the comparison between the two error bounds.

Lemma 1.1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with $L-$Lipschitz continuous gradient, and $C \subseteq \mathbb{R}^n$ be a nonempty closed convex set. Therefore, for any $x \in \mathbb{R}^n$, the following inequality holds*

$$\|x - P_C(x - \nabla f(x))\| \leqslant (2 + L) \operatorname{dist}(x, S). \tag{1.14}$$

where $S = (\partial f)^{-1}(0)$ is the set of critical points of f.

Proof. Indeed, for each x let $p(x) \in P_S(x)$. Then,

$$\begin{aligned}\|x - P_C(x - \nabla f(x))\| &= \|x - P_C(x - \nabla f(x)) - [p(x) - P_C(p(x) - \nabla f(p(x))]\| \\ &\leqslant \|x - p(x)\| + \|P_C(x - \nabla f(x)) - P_C(p(x) - \nabla f(p(x)))\| \\ &\leqslant \operatorname{dist}(x, S) + \|x - \nabla f(x) - [p(x) - \nabla f(p(x))]\| \\ &\leqslant (2 + L)\operatorname{dist}(x, S)\,,\end{aligned}$$

where in the first equality we use the fact that $p(x) = P_C(p(x) - \nabla f(p(x)))$, the second inequality follows from the nonexpansiveness of the projection operator P_C, and the last inequality is by the Lipschitz continuity of the gradient of f. □

1.4.2 Kurdyka-Łojasiewicz inequality in optimization

As mentioned in [62], tame functions, and more specifically, definable functions provide a suitable setting on which the variational analysis theory properly works. The importance of such class of functions lies in the fact that the Kurdyka-Łojasiewicz inequality, a generalized form of the Łojasiewicz inequality, is satisfied by definable functions.

The Łojasiewicz inequality was used by the author in [63] to study the convergence of bounded trajectories of gradient dynamical systems to critical points of certain type of C^1 functions. It was first proven valid for real-analytic functions in [64]. For an open set $U \subseteq \mathbb{R}^n$, a function $f : U \to \mathbb{R}$ is called real-analytic on U if for every $\bar{x} \in U$, there exists a neighborhood $V \subseteq U$ of $\bar{x}$, such that f can be represented as a convergent power series in V.

Proposition 1.9. *Let $f : U \to \mathbb{R}$ be a real-analytic function defined on a open domain $U \subseteq \mathbb{R}^n$. For all $\bar{x} \in U$ there exists $\theta \in \left[\frac{1}{2}, 1\right)$ and a neighborhood $V \subseteq U$ such that for all $x \in V$,*

$$\frac{|f(x) - f(\bar{x})|^\theta}{\|\nabla f(x)\|} \quad \textit{remains bounded.}$$

Kurdyka introduced a generalization of the Łojasiewicz inequality in [53], currently known as the Kurdyka-Łojasiewicz inequality (KŁ inequality for short).

Before presenting the formal definition of the KŁ inequality, we need to introduce the concept of desingularizing function (see [65]). For that, we use the following notation: for $\overline{f} \in \mathbb{R}$, the level set of f at level $\overline{f}$ is defined as $[f \leqslant \overline{f}] := \{x \in \mathbb{R}^n : f(x) \leqslant \overline{f}\}$. Additionally, for $\underline{f} \in \mathbb{R}$, the slice of f at levels $\underline{f}$ and $\overline{f}$ is similarly defined as $[\underline{f} \leqslant f \leqslant \overline{f}] := \{x \in \mathbb{R}^n : \underline{f} \leqslant f(x) \leqslant \overline{f}\}$. The sets $[f < \overline{f}]$ and $[\underline{f} < f < \overline{f}]$ can be analogously defined.

Definition 1.9. *Given $r_0 \in (0, +\infty]$, a function $\varphi : [0, r_0) \to \mathbb{R}_+$ is said to be a desingularizing function if*

- $\varphi(0) = 0$,
- *φ is continuous on $[0, r_0)$, and continuously differentiable on $(0, r_0)$, and*
- *for all $r \in (0, r_0)$, $\varphi'(r) > 0$.*

Moreover, for $\bar{x} \in \mathbb{R}^n$ and a neighborhood U of $\bar{x}$, the set

$$U \cap [f(\bar{x}) < f < f(\bar{x}) + \eta]$$

is called a Kurdyka-Łojasiewicz neighborhood of $\bar{x}$.

Originally in [53], the KŁ inequality was formulated for differentiable functions, and it was then extended in [54] for nonsmooth functions.

Definition 1.10 (KŁ inequality)**.** *A proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with open domain is said to satisfy the KŁ inequality around $\bar{x} \in \operatorname{dom}(\partial f)$, if there exist $r_0 \in (0, +\infty]$, a neighborhood U of $\bar{x}$ and a desingularizing function φ defined on $[0, r_0)$, such that for all $x \in U \cap [f(\bar{x}) < f < f(\bar{x}) + \eta]$*

$$\varphi'(f(x) - f(\bar{x})) \operatorname{dist}(0, \partial f(x)) \geqslant 1.$$

The parlance *desingularizing* is related to the fact that the KŁ inequality holds for any noncritical point [66, Remark 4(b)]. The function φ in the above definition is called desingularizing because it reparametrizes the function f in such a way that around any critical point $\bar{x}$, $\partial(\varphi \circ f)(\bar{x})$ does not contain 0, meaning that $\bar{x}$ is an isolated critical point surrounded by nonsingular points.

A noticeable particular case corresponds to taking the desingularizing function to be $\varphi(r) = cr^{1-\theta}$, for some constants $c > 0$ and $\theta \in [0, 1)$ (cf. Proposition 1.9).

Definition 1.11 ($\theta-$KŁ inequality)**.** *A proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with open domain is said to satisfy the KŁ inequality with exponent θ, or simply the $\theta-$KŁ inequality, around $\bar{x} \in \operatorname{dom}(\partial f)$, if there exist $c, \eta > 0$, $\theta \in [0, 1)$, and a neighborhood U of $\bar{x}$, such that for all $x \in U \cap [f(\bar{x}) < f(x) < f(\bar{x}) + \eta]$,*

$$\operatorname{dist}(0, \partial f(x)) \geqslant c(f(x) - f(\bar{x}))^{\theta}.$$

A distinguishable class of functions that satisfy the KŁ inequality is the semi-algebraic family, corresponding to those functions whose graph can be described as the

solutions of finitely many polinomyal equations and inequalities. This result was originally proven for a broader class of functions that generalize the semialgebraic family, the class of functions defined in a o-mininal structure, in the differentiable case in [53], and later extended to the nonsmooth case in [54].

1.4.3 Relationships between regularity conditions

Historically, different regularity conditions have been used to study properties of convergent methods, see [67, 33, 19, 68, 18, 69] and references therein. The following notion, sometimes referred to as zero-order error bound [67], measures the distance to the set of critical points with the corresponding function values.

Definition 1.12 (Quadratic growth)**.** *A proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with a nonempty set of critical points S, is said to have quadratic growth around S, if for any $\bar{x} \in S$, there exist a neighborhood U of $\bar{x}$, and a constant $c > 0$, such that for all $x \in U$,*

$$f(x) \geqslant f(\bar{x}) + \frac{1}{c^2}\operatorname{dist}(x, S)^2$$

Other commonly used notion of regularity involves the subgradients around a critical point, instead of the function values. The following definition resembles the subdifferential error bound, although it only involves close points to a given critical point in the usual sense, and not in the $f-$attentive sense.

Definition 1.13 (Metric subregularity)**.** *A proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with a nonempty set of critical points S, is said to be metrically subregular at $\bar{x} \in (\partial f)^{-1}(0)$, if there exists a neighborhood U of $\bar{x}$ and a neighborhood V of 0, such that for all $x \in U$ and $w \in \partial f(x) \cap V$*

$$\operatorname{dist}(x, S) \leqslant c^2\|w\|$$

In the literature, Definition 1.13 is called metric subregularity of the subdifferential at $\bar{x}$ for $0 \in \partial f(\bar{x})$. Here, by extension, we say f possesses the property, since we are mainly interested in optimization problems. The same observation can be made for Definition 1.14 below.

Regularity assumptions associated with the inverse of the subdifferential can also be studied. For $f \in \overline{\text{conv}}(\mathbb{R}^n)$, we know $(\partial f)^{-1} = \partial f^*$ (cf. (1.6)). The following notion establishes a bound for a localization of the inverse of the subdifferential in terms of subgradients. It is called in [70] as *local upper Lipschitz* property, used in [71] to deduce rate of convergence of the proximal point algorithm, and in [19] to obtain rate of convergence of $\varepsilon-$subgradient methods (referred as *inverse growth condition* therein).

Definition 1.14 (Calmness)**.** *For a proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with nonempty set of critical points S, we say $(\partial f)^{-1}$ is calm at 0 for $\bar{x}$ if there exist $\mu > 0$, a neighborhood U of $\bar{x}$, and a neighborhood V of 0, such that for all $w \in V$*

$$(\partial f)^{-1}(w) \cap U \subseteq S + \mu\|w\|B(0,1) \tag{1.15}$$

We first study the relationship between the aformetioned concepts of regularity in the presence of convexity.

Simplifications in the convex setting

For any function $f \in \overline{\text{conv}}(\mathbb{R}^n)$, employing neighborhoods U of a minimizer $\bar{x}$ is equivalent to using KŁ neighborhoods of the form $U \cap [f(\bar{x}) < f < f(\bar{x}) + \eta]$, for some $\eta > 0$. Indeed, if U is a nontrivial neighborhood of $\bar{x}$ for nonconstant f, there exists $x_0 \in U$ such that $f(\bar{x}) < f(x_0)$. Setting $\eta = f(x_0) - f(\bar{x}) + 1$, then $U \cap [f(\bar{x}) < f < f(\bar{x}) + \eta]$ is a nontrivial neighborhood of $\bar{x}$. The KŁ slice $[f(\bar{x}) < f < f(\bar{x}) + \eta]$ emerges as a necessity for nonconvex functions, to capture the local behavior of f around $\bar{x}$, the same reason for which the use of the $f-$attentive convergence is needed.

The following result summarizes the relationships between different regularity conditions in the convex case. Note that for convex f, the set of critical points corresponds to the set of global minimizers.

Proposition 1.10. *For any function $f \in \overline{\text{conv}}(\mathbb{R}^n)$ with a nonempty set of minimizers S, the following properties are equivalent for a point $\bar{x} \in S$.*

(a) f satisfies the KŁ inequality with exponent $\theta = 1/2$ around $\bar{x}$.

(b) f has quadratic growth around S for $\bar{x}$.

(c) f is metrically subregular at $\bar{x} \in S$.

(d) ∂f^* is calm at 0 for any $\bar{x} \in S$.

Proof. First, if f satisfies the KŁ inequality with exponent $\theta = 1/2$ around $\bar{x}$, the associated desingularizing function $\varphi(r) = cr^{\frac{1}{2}}$ has the *moderate behavior near the origin* property of [67, Lemma 4]. Hence, in view of [67, Theorem 5], (a) and (b) are equivalent. As mentioned in [68] and proved in [57, Theorem 3.3], (b) and (c) are equivalent. Finally, the equivalence between (c) and (d) follows from [72, Theorem 3.2].

□

Nonconvex case

For the more general case, regularity conditions can be expressed in terms of the following notion that extends the idea of the norm of the derivative of a differentiable function, while in the convex case, it corresponds to the minimal subgradient norm (see [55]).

Definition 1.15 (Slope)**.** *For a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and a point $\bar{x} \in \operatorname{dom}(f)$, the slope of f at $\bar{x}$ is defined as*

$$|\nabla f|(\bar{x}) = \limsup_{x \to \bar{x}} \frac{\max\big(f(\bar{x}) - f(x), 0\big)}{\|\bar{x} - x\|}$$

For a function $f \in w - \overline{\operatorname{conv}}_\rho(\mathbb{R}^n)$, the slope of the function at $\bar{x} \in \operatorname{dom}(f)$ coincides with the minimal subgradient norm [65, Lemma 43], namely

$$|\nabla f|(\bar{x}) = \operatorname{dist}(0, \partial f(\bar{x})). \tag{1.16}$$

In view of this identity, the following relations hold for the class of weakly convex functions. Some other relationships between regularity conditions that are commonly used in the literature can be found in [73, Proposition 2] and [74].

Proposition 1.11. *For any function $f \in w - \overline{\operatorname{conv}}_\rho(\mathbb{R}^n)$, and a point $\bar{x} \in \operatorname{dom}(f)$:*

1. *The subdifferential error bound of Definition 1.7 is sufficient for the $1/2-$KŁ inequality to hold around $\bar{x}$.*
2. *The $\theta-$KŁ inequality for some $\theta \in (0, 1)$ implies the following level-error bound: for some neighborhood U of $\bar{x}$, constants $c' > 0$ and η', it holds for all $x \in U \cap [f(\bar{x}) < f < f(\bar{x}) + \eta']$,*
$$\operatorname{dist}(x, [f \leqslant f(\bar{x})]) \leqslant c' \operatorname{dist}(0, \partial f(x))^{\frac{1-\theta}{\theta}}.$$
3. *Additionally, if $\bar{x}$ is a local minimizer, then the subdifferential error bound of Definition 1.7 implies the quadratic growth condition of Definition 1.12.*

Proof. The first relation follows from [55, Proposition 3.8]. The $\theta-$KŁ inequality is a sufficient condition for the level-error bound in view of [55, Theorem 3.7]. The final statement stems from [56, Corollary 3.2]. □

For structured optimization problems, [75, Theorem 4.1] offers a result showing a relationship between the projection error bound and the KŁ inequality. The following proposition is a special case for constrained optimization problems.

Proposition 1.12. *For a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ with locally Lipschitz continuous gradient, and a nonempty closed convex set $C \subseteq \mathbb{R}^n$, consider problem (1.13). Suppose the set of critical points $S = (\partial[f + i_C])^{-1}(0)$ is nonempty, and the projection error bound of Definition 1.8, and the proper separation of isocost surfaces property of Definition 1.1 hold. Then, the $1/2-$KŁ inequality holds true around any critical point $\bar{x}$.*

As pointed out by a referee of [22], possible extensions when using other Łojasiewicz exponents different from $\theta = \frac{1}{2}$ might lead to sublinear or superlinear rates of convergence guarantees for methods complying with the descent methods described in Chapter 3, depending on the value of θ, properties of the model function used to construct the specific algorithm, and possibly other assumptions.

As mentioned in Section 1.4.1 and shown in Chapter 3 below, subdifferential-based error bounds, or first-order error bounds, are utilized to analyze convergence rates of algorithms. Zero-order error bounds, that is, error bounds based on function values, can be used for the same purpose [67, 76]. The following concept can be regarded as a zero-order error bound that measures the distance to a level set, as introduced in [77].

Definition 1.16 (Epi-metric subregularity)**.** *A proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be epi-metrically subregular at $\bar{x} \in \mathrm{dom}(f)$, if there exist $K > 0$ and a neighborhood U of $\bar{x}$, such that for all $x \in U$,*

$$\mathrm{dist}(x, [f \leqslant f(\bar{x})]) \leqslant K \max(f(x) - f(\bar{x}), 0). \tag{1.17}$$

In the following proposition, $\partial^{>} f(x)$ denotes the outer limiting subdifferential at $\bar{x}$, given by

$$\partial^{>} f(\bar{x}) = \{v \in \mathbb{R}^n : \exists x^k \underset{f}{\to} \bar{x}, f(x^k) > f(\bar{x}), \widehat{\partial} f(x^k) \ni v^k \to v\}.$$

Note that $\partial^{>} f(\bar{x}) \subseteq \partial f(\bar{x})$.

Proposition 1.13. *For any $f \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$, the conditions below are equivalent:*

(a) f is epi-metrically subregular at $(\bar{x}, f(\bar{x}))$ with constant K.

(b) $K\,\mathrm{dist}(0, \partial^{>} f(\bar{x})) \geqslant 1$.

(c) $K \liminf \{|\nabla f|(x) : \|x - \bar{x}\| \leqslant \varepsilon\,, f(\bar{x}) < f(x) < f(\bar{x}) + K\varepsilon\} \geqslant 1$.

Proof. The equivalence between items (b) and (c) and that item (c) implies item (a) is [78, Theorem 6.6]. To show the converse, for contradiction purposes, suppose that for $K > 0$

and all x in a neighborhood U of $\bar{x}$, (1.17) holds, but $K\,\mathrm{dist}(0,\partial^{>}f(\bar{x}))<1$. Then, there exist $v\in\partial^{>}f(\bar{x})$ and $\varepsilon>0$ such that $\|v\|\leqslant K^{-1}-\varepsilon$. By definition of $\partial^{>}f(\bar{x})$, for any $k\in\mathbb{N}$, take x^k, $v^k\in\mathbb{R}^n$ such that $v^k\in\widehat{\partial}f(x^k)$, $x^k\underset{f}{\to}\bar{x}$, $f(x^k)>f(\bar{x})$, and $k_0\geqslant 1$ so that for all $k\geqslant k_0$,

$$\|v^k\|\leqslant K^{-1}-\varepsilon. \tag{1.18}$$

Let $\bar{x}^k$ denote the projection of x^k onto the level set $[f\leqslant f(\bar{x})]$. We claim $\bar{x}^k\underset{f}{\to}\bar{x}$. In fact, since there exists $k_1\geqslant k_0$ such that for all $k\geqslant k_1$, $x^k\in U$, then from (1.17) and $f(x^k)>f(\bar{x})$, it follows for all $k\geqslant k_1$, $\mathrm{dist}(x^k,[f\leqslant f(\bar{x})])\leqslant K(f(x^k)-f(\bar{x}))$. As $x^k\underset{f}{\to}\bar{x}$, then $\|x^k-\bar{x}^k\|=\mathrm{dist}(x^k,[f\leqslant f(\bar{x})])\to 0$, and thus $\bar{x}^k-\bar{x}=(\bar{x}^k-x^k)+(x^k-\bar{x})\to 0$. Furthermore, from Proposition 1.6, for any $u\in\widehat{\partial}f(\bar{x})$,

$$f(\bar{x})-f(\bar{x}_k)\leqslant\left\langle u,\bar{x}-\bar{x}^k\right\rangle+\frac{\rho}{2}\|\bar{x}-\bar{x}^k\|^2,$$

and using the Cauchy-Schwarz inequality and the fact that $\bar{x}^k\in[f\leqslant f(\bar{x})]$, it follows

$$|f(\bar{x}_k)-f(\bar{x})|\leqslant\|u\|\|\bar{x}-\bar{x}^k\|+\frac{\rho}{2}\|\bar{x}-\bar{x}^k\|^2.$$

Hence, $\bar{x}^k\to\bar{x}$ yields $|f(\bar{x}_k)-f(\bar{x})|\to 0$, and thus $\bar{x}^k\underset{f}{\to}\bar{x}$.

Moreover, by virtue of Proposition 1.6 and $v^k\in\widehat{\partial}f(x^k)\subseteq\overline{\partial}f(x^k)$,

$$f(\bar{x}^k)+\frac{\rho}{2}\|\bar{x}_k-x_k\|^2\geqslant f(x^k)+\left\langle v^k,\bar{x}^k-x^k\right\rangle.$$

As $\bar{x}_k-x_k\to 0$, for sufficiently large k it holds that $\frac{\rho}{2}\|\bar{x}_k-x_k\|\leqslant\frac{\varepsilon}{2}$, and thus

$$\begin{aligned}
f(x^k) &\leqslant f(\bar{x}^k)-\left\langle v^k,\bar{x}^k-x^k\right\rangle+\frac{\rho}{2}\|\bar{x}^k-x^k\|^2\\
&\leqslant f(\bar{x}^k)+\left(\|v^k\|+\frac{\rho}{2}\|\bar{x}^k-x^k\|\right)\|\bar{x}^k-x^k\|\\
&\leqslant f(\bar{x}^k)+\left(\|v^k\|+\frac{\varepsilon}{2}\right)\|\bar{x}^k-x^k\|\\
&\leqslant f(\bar{x}^k)+\left(K^{-1}-\frac{\varepsilon}{2}\right)\|\bar{x}^k-x^k\|,
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality, and (1.18) yields the fourth inequality. Hence, rearranging terms in the above estimate, for all sufficiently large k,

$$\mathrm{dist}(x_k,[f\leqslant f(\bar{x})])=\|\bar{x}_k-x_k\|\geqslant\left(K^{-1}-\frac{\varepsilon}{2}\right)^{-1}\left(f(x_k)-f(\bar{x}_k)\right)>K\left(f(x_k)-f(\bar{x}_k)\right).$$

contradicting (1.17).

□

2 Computational optimization tools

In this chapter, we review iterative methods of crucial importance for the remainder of this work. In particular, we summarize convergence results of algorithms of proximal type.

In practice, not only convergence guarantees of a method are of interest, but also how fast the method converges to a solution of the problem to be solved. Of particular interest is the linear rate of convergence, defined next. Given a sequence $\{x^k\}$ generated by some algorithm, and a point $\bar{x}$ known to be the limit of $\{x^k\}$, we say $\{x^k\}$ converges to $\bar{x}$

- $Q-$linearly, or just linearly, if $\limsup_{k\to+\infty} \dfrac{\|x^{k+1}-\bar{x}\|}{\|x^k-\bar{x}\|} = q \in (0,1)$.
- $R-$linearly if $\limsup_{k\to+\infty} \dfrac{\|x^k-\bar{x}\|}{q^k} = C$, for some $q \in (0,1)$ and $C > 0$. Equivalently, for all sufficiently large $k \in \mathbb{N}$, $\|x^k - \bar{x}\| \leqslant Cq^k$.
- Superlinearly if $\limsup_{k\to+\infty} \dfrac{\|x^{k+1}-\bar{x}\|}{\|x^k-\bar{x}\|} = 0$.

In order to establish linear rates of convergence, common assumptions in the literature are the properties reviewed in Section 1.4. Examples of how these properties are used can be found in this section and throughout the subsequent chapters. Usually, superlinear convergence requires extra assumptions, related to the precision of the computation of iterates as they approach the limit.

We begin this chapter by examining the proximal point algorithm (PPA) [17] and some inexact variants in Section 2.1, including proximal bundle methods. The PPA serves as the basis of a plethora of methods, including the ones in this chapter. We proceed in Section 2.2 with the study of splitting methods for optimization. In particular, we focus on the Douglas-Rachford splitting (DRS) method, and the Progressive Hedging (PH) algorithm for stochastic optimization problems.

2.1 Proximal-type optimization methods

For a function $f \in \overline{\text{conv}}(\mathbb{R}^n)$, the proximal point algorithm finds solutions to problems of the form (1.7) by iteratively solving a regularized version of the original problem. Problem (1.8) can be addressed similarly, by modeling the constraint $x \in C$ in the

objective function through the indicator function i_C. For the iterative PPA mechanism to be an implementable alternative to solving directly the problem, the regularized subproblems should be simple. Proximal bundle methods rise as an alternative by approximating the objective function using polyhedral models, and then solving a regularization of such models.

2.1.1 The proximal point algorithm

The PPA works in a more general framework than optimization, for maximal monotone operators. A set valued-mapping $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is monotone if $\langle g-g', x-x' \rangle \geqslant 0$, for $g \in T(x)$, $g' \in T(x')$. The operator T is maximal monotone whenever its graph is maximal for the inclusion, among the graphs of the class of monotone operators. The problem of interest in the general setting is to find a point $x \in \mathbb{R}^n$ such that

$$0 \in T(x). \tag{2.1}$$

Equivalently, the problem can be formulated using the inverse operator, namely, $x \in T^{-1}(0)$. The PPA generates a sequence of points $\{x^k\}$ that converges to a zero of T, under appropriate assumptions. When T is the subdifferential of a function $f \in \overline{\text{conv}}(\mathbb{R}^n)$, such a limit of $\{x^k\}$ is a global minimizer of f. All the methods in this dissertation are developed for the operator $T = \partial f$, where $f \in \overline{\text{conv}}(\mathbb{R}^n)$ or $f \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$.

The cornerstone of the PPA is a property of maximal monotone operators, proved in [79]: given a maximal monotone operator T and a scalar $t > 0$, for any $x \in \mathbb{R}^n$, there exists a unique $x^+ \in \mathbb{R}^n$, such that $x \in (I + tT)(x^+)$. This defines the resolvent of T, the map $J_{tT} = (I + tT)^{-1}$, given by $x^+ = J_{tT}(x)$. In other words, for any $x \in \mathbb{R}^n$, x^+ is uniquely determined by the inclusion $\dfrac{x - x^+}{t} \in T(x^+)$. Equivalently, given $x \in \mathbb{R}^n$, there exists a unique pair (x^+, g^+) for which $g^+ \in T(x^+)$, and $x = x^+ + tg^+$.

For $f \in \overline{\text{conv}}(\mathbb{R}^n)$, the unique $x^+ \in \mathbb{R}^n$ such that $x \in (I + t\partial f)(x^+)$ satisfies

$$0 \in \partial f(x^+) + \frac{1}{t}(x^+ - x).$$

This inclusion is the optimality condition of the optimization problem that regularizes the original problem of minimizing f. In the optimization setting, the resolvent is usually called the proximal point operator.

Definition 2.1 (Proximal point operator and Moreau envelope)**.** *For a proper lsc function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $t > 0$, the proximal point operator $\text{prox}_{tf} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is defined as, for all $x \in \mathbb{R}^n$,*

$$\text{prox}_{tf}(x) = \underset{y \in \mathbb{R}^n}{\arg\min} \left\{ f(y) + \frac{1}{2t}\|y - x\|^2 \right\}. \tag{2.2}$$

The optimal value of the minimization problem in (2.2) is called the Moreau envelope of f with stepsize $t > 0$, and denoted by $e_t f$. More precisely,

$$e_t f(x) = \inf_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2t}\|y - x\|^2 \right\}.$$

In the general case, the proximal point operator could be empty-valued and $e_t f(x)$ might take the value $-\infty$. If f is proper, lsc, and prox-bounded, that is, when there exists $t > 0$ and $x \in \mathbb{R}^n$, such that $e_f(x) > -\infty$, then $\operatorname{prox}_{tf}(x)$ is nonempty and compact [29, Theorem 1.25], and e_f is locally Lipschitz continuous at x [29, Example 10.32]. When $f \in \overline{\operatorname{conv}}(\mathbb{R}^n)$, prox_{tf} is a single-valued mapping [29, Theorem 12.12, Theorem 12.17], and $e_t f$ is finite-valued and differentiable, in such a way that

$$\nabla(e_{tf})(x) = \frac{1}{t}\big(x - \operatorname{prox}_{tf}(x)\big). \tag{2.3}$$

In the convex case, $e_t f$ is also known as the Moreau-Yosida regularization of f. Observe the same properties above hold for $f \in w - \overline{\operatorname{conv}}_\rho(\mathbb{R}^n)$ and $0 < t\rho < 1$, see [80].

The proximal point operator takes a simple form when $f = i_C$ for some nonempty closed convex set C. In this case,

$$\begin{aligned} \operatorname{prox}_{tf}(x) &= \arg\min_{y \in \mathbb{R}^n} \left\{ i_C(x) + \frac{1}{2}\|y - x\|^2 \right\} \\ &= \arg\min_{y \in C} \left\{ \frac{1}{2t}\|y - x\|^2 \right\} \\ &= P_C(x), \end{aligned}$$

This identity plays a crucial role in Chapter 6. Observe also in this case

$$e_{tf}(x) = \frac{1}{2t}\operatorname{dist}(x, C)^2,$$

and thus

$$\nabla\big(\operatorname{dist}(x, C)^2\big) = 2\big(x - P_C(x)\big).$$

Exact proximal point algorithm

In the context of maximal monotone operators, for a sequence $\{t_k\} \subseteq (0, +\infty)$, and a starting point $x^0 \in \mathbb{R}^n$, the sequence of proximal points $\{x^k\}$ generated by the PPA is defined as, for all $k \geqslant 0$,

$$x^{k+1} = J_{t_k T}(x^k).$$

In the special case of optimization, the PPA sequence obeys

$$x^{k+1} = \operatorname{prox}_{t_k f}(x^k).$$

The convergence of the PPA has been extensively studied. The first result was provided in [17] tailored for maximal monotone operators.

Proposition 2.1 (Convergence of PPA)**.** *Let $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximal monotone operator, and a sequence $\{t_k\} \subseteq (0, +\infty)$ bounded away from 0, meaning there exists $t_{\min} > 0$, such that for all $k \in \mathbb{N}$, $t_k \geqslant t_{\min}$. Then, any bounded sequence $\{x^k\}$ generated by the PPA converges to a solution of problem (2.1).*

Remark 2.1. *Under the assumptions of the previous proposition, $\{x^k\}$ is bounded if and only if the set of zeros of T is nonempty.*

Inexact proximal point algorithm: first naive approach

Inexact versions of the PPA have also been comprehensively analyzed, including [17]. In inexact variants, the proximal point is approximately computed following specific accuracy rules. A direct extension of the approximate criteria of [17] is examined in [71]. Given $r > 0$, $\{t_k\} \subseteq (0, +\infty)$, $\{\varepsilon_k\} \subseteq [0, +\infty)$ such that $\sum_{k=0}^{+\infty} \varepsilon_k < +\infty$, and a starting point $x^0 \in \mathbb{R}^n$, $\{x^k\}$ is generated satisfying the following approximation criteria

$$\|x^{k+1} - \operatorname{prox}_{t_k f}(x^k)\| \leqslant \varepsilon_k \min\left(1, \|z^{k+1} - z^k\|^r\right). \tag{2.4}$$

The following result corresponds to the convergence of the inexact PPA given by (2.4), and it also provides a condition (cf. Definition 1.13) under which linear/superlinear convergence is obtained.

Proposition 2.2 (Convergence of an inexact PPA)**.** *Let $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a maximal monotone operator, such that the set of zeros is nonempty. Moreover, take a sequence $\{t_k\} \subseteq (0, +\infty)$ bounded away from 0. Then, any bounded sequence $\{x^k\}$ generated by the inexact PPA of (2.4) converges to a solution of problem (2.1).*

In addition, suppose there exist $\ell, \delta > 0$, such that whenever $x \in \mathbb{R}^n$, $w \in T(x) \cap B(0, \delta)$, $\operatorname{dist}(x, T^{-1}(0)) \leqslant \ell\|w\|$. Then, if $\{t_k\}$ is a nondecreasing sequence, $\operatorname{dist}(x^k, T^{-1}(0))$ converges to 0 linearly. Furthermore, if $t_k \to +\infty$, then $\operatorname{dist}(x^k, T^{-1}(0)) \to 0$ superlinearly.

Remark 2.2. *The superlinear convergence result states that, to speed up convergence, it is necessary to drive t_k to $+\infty$, meaning we basically require $e_{tf}(x) \approx \min f$, making the penalization in the proximal point subproblems increasingly neglectable. Additionally, in (2.4) the right-hand side must be null in the limit (the series $\sum_k \varepsilon_k$ is finite), therefore forcing the inexact proximal point calculations to become asymptotically exact.*

Inexact proximal point algorithm: a hybrid extragradient-proximal point version

Among the jungle of inexact variants of the PPA, the hybrid approximate extragradient–proximal point algorithm [25] stands out by using a relative-error criteria

to approximately compute proximal points. This inexact version of PPA employs certain enlargement of operatorst, somewhat akin to the $\varepsilon-$subdifferential in Convex Analysis. Given a maximal monotone operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, the $\varepsilon-$enlargement of T, denoted T^ε, is defined for $x \in \mathbb{R}^n$ as

$$T^\varepsilon(x) = \{g \in \mathbb{R}^n | \langle g' - g, x' - x\rangle \geqslant -\varepsilon,\ x' \in \mathbb{R}^n, g' \in T(x')\}.$$

The enlargement of an operator defines a perturbation of the problem $0 \in T(x)$, while staying close to it. In particular, when $T = \partial f$ for $f \in \overline{\text{conv}}(\mathbb{R}^n)$, then $T^\varepsilon(x) \supseteq \partial_\varepsilon f(x)$ for all $x \in \mathbb{R}^n$.

Algorithm 1 presents the hybrid approximate extragradient–proximal point algorithm. The result that follows presents the convergence of this algorithm. This proposition helps to build the basis of the convergence analysis of the method presented in Chapter 6. Actually, the analysis in Chapter 6 can be deemed as an extension for the optimization case of [25] with an extra projection step.

Algorithm 1 A Hybrid Approximate Extragradient–Proximal Point Algorithm

1: **Initialization**: choose $t_0 > 0$, $\sigma_0 \in [0, 1)$, $\varepsilon_0 \geqslant 0$, and $x^0 \in \mathbb{R}^n$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: **Acceptance test**: find a pair (y^k, v^k) such that $v^k \in T^{\varepsilon_k}(y^k)$, and

$$\|t_k v^k + y^k - x^k\| + 2t_k\varepsilon_k \leqslant \sigma_k^2\|y^k - x^k\|^2.$$

4: **Approximal step**: set $x^{k+1} = x^k - t_k v^k$. Define $t_{k+1} > 0$, $\sigma_{k+1} \in [0, 1)$, and $\varepsilon_{k+1} \geqslant 0$.
5: **end for**

Linear convergence of Algorithm 1 is obtained in [25] by assuming that T^{-1} is locally upper Lipschitzian at 0, a generalization of Definition 1.14 for operators [70]: there exists some constant $L > 0$, and a neighborhood U of 0, such that whenever $g \in T(x) \cap U$, $\|y - x^\star\| \leqslant L\|g\|$, where $x^\star$ is the (unique) solution to (2.1). We refer to Chapter 3, 4, and 6 for similar applications of this type of condition.

Proposition 2.3 (Convergence of Algorithm 1, [25])**.** *For any maximal monotone operator $T : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, suppose problem (2.1) has a nonempty set of solutions. If the sequence of stepsizes $\{t_k\}$ is bounded away from 0, and the error tolerance $\sigma_k = \sigma \in [0, 1)$ is kept fixed, then the sequence $\{x^k\}$ generated by Algorithm 1 converges to a solution to (2.1). Additionally, if T^{-1} is locally upper Lipschitzian at 0, then $\{x^k\}$ converges linearly to the unique solution of (2.1)*

Remark 2.3. *As the authors of [25] mention in a remark after the proof of Theorem 3.2 therein, by taking t_k to infinity and allowing σ_k to tend to 0, the method converges*

superlinearly. Note in passing that driving σ_k to 0, forces ε_k to be asymptotically null too, at a fast rate (the product $t_k\varepsilon_k$ tends to 0, with t_k going to $+\infty$). A more detailed explanation is given in Chapter 6.

2.1.2 Proximal bundle methods: an implementable form of the proximal point algorithm

An acceptance test alternative to implement an inexact proximal step in optimization can be traced back to [81, 82], see also [83]. The idea is to replace f in (2.2) with a simpler model function using the information generated along iterations. Proximal bundle methods construct (polyhedral) models $\check{f}_k$, and define the next iterate to be the proximal point of the model at the *best* candidate generated so far.

Proximal bundle methods correspond to a stabilization of a simpler method called cutting-planes method. We assume there exists a black-box that, given a point $x \in \mathbb{R}^n$, returns the value $f(x)$ and a subgradient $g(x) \in \partial f(x)$. In this way, given the bundle of information

$$\{(x^i, f^i, g^i) : i \in \mathcal{B}_k\},$$

where $f^i = f(x^i)$, $g^i = g(x^i)$, and $\mathcal{B}_k = \{1, \ldots, k\}$, the model defined by the end of iteration $k-1$ and used in iteration k is

$$\check{f}_k(x) = \max_{i \in \mathcal{B}_k} \{f^i + \langle g^i, x - x^i \rangle\}.$$

Due to convexity, $\check{f}_k$ is a lower estimation of f that is improved in each iteration to better represent the graph of f, that is, $\check{f}_k \leqslant \check{f}_{k+1} \leqslant f$ for all $k \in \mathbb{N}$. The advantage of this approach is the use of past accumulated information to construct an approximation of the objective function, different from the memoryless gradient method.

The cutting-planes method defines the next iterate x^{k+1} to be the minimizer of $\check{f}_k(x)$ over some fixed compact set X sufficiently large to contain at least one solution of the original problem. Then, the subgradient linearization of the objective function at x^{k+1} is attached to the model, and the procedure is repeated. Figure 4 illustrates the cutting-planes model for $f(x) = \dfrac{1}{2}x^2 + \dfrac{5}{4}$.

Although convergent [84, 30], the cutting-planes method may present a poor numerical performance, since minimizers of $\check{f}_k$ are difficult to control (see [30, Example 1.1.2]), yielding a non-descent optimization method, possibly unstable, unless the graph of the objective function has a special structure [85, Chapter 9]. For example, for polyhedral functions, the cutting-planes model coincides with the function after a finite number of iterations, and thus the method has finite termination.

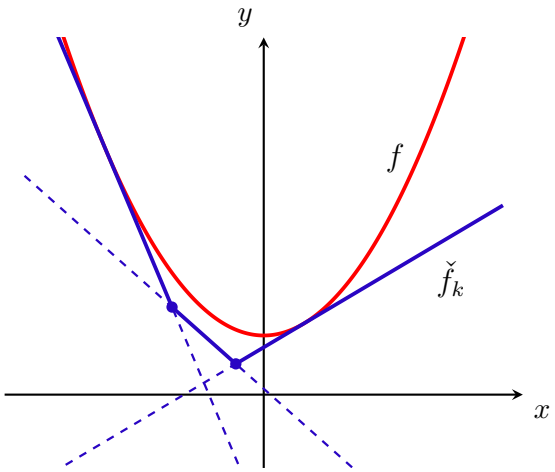


Figure 4 – Cutting-planes model $\check{f}_k$ constructed after 3 iterations of the cutting-planes method applied to $f(x) = \frac{1}{2}x^2 + \frac{5}{4}$, starting from $x^0 = \frac{1}{2}$, and using $X = [-2, 2]$ as the compact set over which minimization of the models is performed.

One option to stabilize the cutting-planes method is to add a quadratic term to the model subproblem objective function. More specifically, given $t_k > 0$, instead of minimizing $\check{f}_k$ over X, proximal bundle methods solve

$$\min_{x \in \mathbb{R}^n} \left\{ \check{f}_k(x) + \frac{1}{2t_k} \|x - \hat{x}^k\|^2 \right\}, \tag{2.5}$$

where $\{\hat{x}^k\}$ is a sequence of centers representing candidate points of *good quality*. This quality is assessed by means of a descent test: given an Armijo-like parameter $m > 0$ and the current center $\hat{x}^k$, once x^{k+1} is obtained as the solution to problem (2.5), check if

$$f(x^{k+1}) - f(\hat{x}^k) \leqslant m\big(\check{f}_k(x^{k+1}) - f(\hat{x}^k)\big). \tag{2.6}$$

If (2.6) holds, declare a *serious step* and set $\hat{x}^{k+1} = x^{k+1}$. Otherwise, declare a *null step* and set $\hat{x}^{k+1} = \hat{x}^k$.

Figure 5 shows the cutting-planes model for $f(x) = \frac{1}{2}x^2 + \frac{5}{4}$ constructed following the proximal bundle idea. Observe that the cutting-planes model constructed by the proximal bundle method represents better the function near the minimizer than the cutting-plane model in Figure 4, due to the fact that proximal bundle methods regularize and stabilize the iterations with a quadratic perturbation.

The sequence $\{\hat{x}^k\}$ of centers, or serious-step iterates, is a subsequence of $\{x^k\}$ of those solutions to (2.5) that provide sufficient descent for the original objective function f compared to its value at the current center, at least a fraction of the decrease the model predicts at the candidate point x^{k+1}. Hence, proximal bundle methods are of descent, since $\{f(\hat{x}^k)\}$ is nonincreasing.

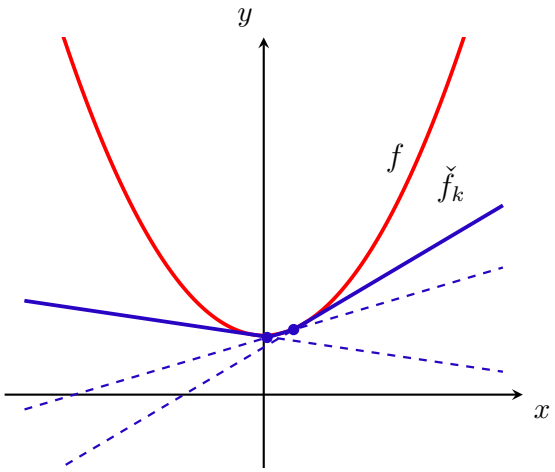


Figure 5 – Cutting-planes model $\check{f}_k$ constructed after 3 iterations of the proximal bundle method applied to $f(x) = \frac{1}{2}x^2 + \frac{5}{4}$, starting from $x^0 = \frac{1}{2}$, using $m = \frac{1}{8}$ and $t_k = \frac{1}{2}$.

Proximal bundle methods can be deemed as an approximation of the PPA. They differ from the inexact versions of the PPA in Section 2.1.1, inasmuch an exact proximal step is performed for an approximation of f, whereas the method of (2.4) or Algorithm 1 compute a point close to the true proximal point satisfying some absolute-error or relative-error condition. See Proposition 6.5 for a relationship between the proximal bundle descent inequality and the relative-error condition of Algorithm 1 in optimization mode. Another difference is, since proximal bundle methods are constructive, convergence of the null steps is provable. In this sense, bundle methods are genuinely implementable, while the method of (2.4) and Algorithm 1 correspond to abstract patterns.

The parlance *implementable form* for proximal bundle methods comes from [83], and refers to the fact that finding the exact proximal point (2.2), and other abstract frameworks, could be, in principle, as costly as directly solving the original problem, while problem (2.5) is much simpler.

Proximal bundle methods enjoy an additional feature, namely, implementable stopping tests. Simply put, whenever the aggregate subgradient and the aggregate error are smaller than a certain tolerance, the center $\hat{x}^k$ is an $\varepsilon-$minimizer of f. More specifically, at iteration k, we define the aggregate gradient

$$\hat{g}^k = \sum_{i \in \mathcal{B}_k} \alpha_i^k g^i,$$

where $\alpha_i^k \in [0,1]$, satisfying $\sum_{i \in \mathcal{B}_k} \alpha_i^k = 1$, are the simplicial multipliers associated with the problem in (2.5). We also define the linearization error at $\hat{x}^k$, for each $i \in \mathcal{B}_k$, as

$$e_i^k = f(\hat{x}^k) - \big(f^i - \langle g^i, \hat{x}^k - x^i \rangle\big),$$

a nonnegative quantity due to the subgradient inequality for $g^i \in \partial f(x^i)$. In turn, we define the aggregate error as

$$\hat{e}^k = \sum_{i \in \mathcal{B}_k} \alpha_i^k e_i^k.$$

Using a transportation argument (see, for instance, [86, Lemma 10.8] or Proposition 3.3 below), $\hat{g}^k \in \partial_{\hat{e}^k} f(\hat{x}^k)$, and thus whenever $\|\hat{g}^k\|$ and $\hat{e}^k$ are sufficiently small, then $\hat{x}^k$ is an approximate solution to $\min f$. We also refer to the discussion pertaining (5.15) for a similar argument.

The following result presents the convergence results of the basic proximal bundle method. We refer the readers to [30, Chapter XV] for a detailed proof.

Proposition 2.4 (Convergence of proximal bundle methods)**.** *Let $f \in \overline{\operatorname{conv}}(\mathbb{R}^n)$ and consider problem (1.7). Suppose the set of solutions of this problem is nonempty.*

(a) Suppose the sequence $\{\hat{x}^k\}$ generated by the proximal bundle method satisfies (2.6) infinitely many times for indices in a set $\mathcal{K}$. If $\sum_{k \in \mathcal{K}} t_k = +\infty$, then $\{\hat{x}^k\}$ is a minimizing sequence. Additionally, if $\{t_k\}_{k \in \mathcal{K}}$ is bounded from above, then $\{\hat{x}^k\}$ converges to a minimizer of problem (1.7).

(b) Suppose the sequence $\{\hat{x}^k\}$ generated by the proximal bundle method satisfies (2.6) finitely many times, namely, there exists a tail of null steps: for some $\hat{k} \in \mathbb{N}$, for all $k \geqslant \hat{k}$, x^k does not satisfy (2.6). If $\{t_k\}_{k > \hat{k}}$ is nonincreasing and $\sum_{k > \hat{k}} \frac{t_k^2}{t_{k-1}} = +\infty$, then $\hat{x}^{\hat{k}}$ is a solution to problem (1.7).

Some adjustments can be made in the proximal bundle method described above. For instance, the bundle $\mathcal{B}_k$ in iteration k can be taken to be a (proper) subset of $\{1, \ldots, k\}$, just keeping the active indices at the solution x^{k+1}, that is, those $i \in \{1, \ldots, k\}$ such that $\check{f}_k(x^{k+1}) = f^i + \langle g^i, x^{k+1} - x^i \rangle$. This would lead to a simpler model, and thus simpler problems to be solved at each iteration from a computational point of view, and still preserve convergence (a feature impossible in cutting-planes). Another typical modification is to aggregate past linearizations, that is, take convex combinations of subgradients g^i and function values f^i for $i \in \mathcal{B}_k$, in order to define the aggregate linearization $\bar{f}_k$. Indeed, any convex model sandwiched between $\bar{f}_k$ and $\check{f}_k$ preserves the convergence properties of Proposition 2.4.

2.2 Operator splitting methods

We saw in Section 2.1.1 that the PPA can be applied to obtain a solution of the problem $0 \in T(x)$, by iteratively solving a perturbation of the problem, namely, finding

a $x^+ \in \mathbb{R}^n$ such that $0 \in T(x^+) + \frac{1}{t}(x^+ - x)$. When T has a block-separable structure, a further specialized method can exploit this particular features in order to simplify the iterations.

Suppose $T = A + B$, where A and B are maximal monotone operators. Computing the resolvent of T can be costly, while separately computing the resolvents J_{tA} and J_{tB} can be more efficient. The Douglas-Rachford (DR) splitting method capitalizes on this assumption, and approximates one proximal step by performing operations for each term separately.

2.2.1 Douglas-Rachford splitting

One iteration of DR is rooted in the following expresion [87, Eq. (DR)]:

$$u^{k+1} \in J_{t_k A}\big(J_{t_k B}(I - t_k A) + t_k A\big)(u^k).$$

Such sequence $\{u^k\}$ is sometimes called shadow DR sequence, since $u^k = J_{t_k A}(s^k)$, where $\{s^k\}$ conforms to

$$s^{k+1} = \big(J_{t_k B}(2J_{t_k A} - I) + I - J_{t_k A}\big)(s^k).$$

The DR method amounts to compute separate proximal steps for each term, once per iteration, and then combine them to generate the next iterate. Algorithm 2 unravels these operations by defining two copies of the same variable in each iteration, u^k and v^k, and then performs a correction/coordination step using the difference $v^k - u^k$ as direction with unit stepsize.

Algorithm 2 Douglas-Rachford splitting method for monotone operators

1: **Initialization**: choose $t_0 > 0$ and $s^0 \in \mathbb{R}^n$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad A-$**proximal step**: $u^k = J_{t_k A}(s^k)$.
4: $\quad B-$**proximal step**: $v^k = J_{t_k B}(2u^k - s^k)$.
5: $\quad$ **Coordination step**: $s^{k+1} = s^k + v^k - u^k$. Define $t_{k+1} > 0$.
6: **end for**

The convergence of Algorithm 2 in shown in [87, Theorem 3.15], and essentially corresponds to proving that $\{s^k\}$ converges to a fixed point of the operator $J_{t_k B}(2J_{t_k A} - I) + I - J_{t_k A}$.

Proposition 2.5 (Convergence of DRS)**.** *Let A, $B : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be two maximal monotone operators, such that $A + B$ has a nonempty set of zeros. In the context of Algorithm 2, if $t_k = t$ for all $k \in \mathbb{N}$, then the sequences $\{u^k\}$ and $\{v^k\}$ converge to a solution $u^\star$ to the problem $0 \in (A + B)(x)$, while the sequence $\{s^k\}$ converges to a point $s^\star = u^\star + tw^\star$ such that $w^\star \in A(u^\star)$, and $-w^\star \in B(u^\star)$.*

In the context of optimization, the DR decomposition approach can be used to solve problems of the form

$$\begin{cases} \min\limits_{x\in\mathbb{R}^n} & \varphi_1(x)+\varphi_2(z) \\ \text{s.t.} & Mx=z\,, \end{cases} \tag{2.7}$$

where $\varphi_1 \in \overline{\text{conv}}(\mathbb{R}^n)$, $\varphi_2 \in \overline{\text{conv}}(\mathbb{R}^m)$, and $M \in \mathbb{R}^{m\times n}$ [87]. For this formulation, φ_2 usually represents the original objective function, and φ_1 is used as a regularization or penalization function.

In order to apply Algorithm 2, we take $A = \partial\varphi_1$ and $B = \partial\big(\varphi_2 \circ M\big)$. Using constant $t_k = t$: starting from $s^0 \in \mathbb{R}^n$, define $u^0 = \text{prox}_{t\varphi_1}(s^0)$. Then, for every $k = 0, 1, 2, \ldots$, define z^k as step 4 of Algorithm 2

$$\begin{aligned} z^k &= \arg\min_{z\in\mathbb{R}^n}\left\{\varphi_2(Mz)+\frac{1}{2t}\|z-(2u^k-s^k)\|^2\right\} \\ &= \arg\min_{x\in\mathbb{R}^n}\left\{\varphi_2(Mz)+\frac{1}{2t}\|z-u^k\|^2+\frac{1}{t}\langle z-u^k, s^k-u^k\rangle+\frac{1}{2t}\|s^k-u^k\|^2\right\} \\ &= \arg\min_{z\in\mathbb{R}^n}\left\{\varphi_2(Mz)+\left\langle\frac{s^k-u^k}{t}, z\right\rangle+\frac{1}{2t}\|z-u^k\|^2\right\}, \end{aligned}$$

where the second equality is obtained after expanding squares, and the third equality yields from discarding constant terms. In the last line, define $w^k = \dfrac{s^k-u^k}{t}$, then line 5 of Algorithm 2 implies

$$\begin{aligned} w^{k+1} &= \frac{s^{k+1}-u^{k+1}}{t} \\ &= \frac{s^k+v^k-u^k-u^{k+1}}{t} \\ &= w^k+\frac{v^k-u^{k+1}}{t}. \end{aligned}$$

Then, we define x^k as step 3 of Algorithm 2 in the following iteration, that is,

$$\begin{aligned} x^k &= u^{k+1} \\ &= \text{prox}_{t\varphi_1}(s^{k+1}) \\ &= \text{prox}_{t\varphi_1}(s^k+z^k-u^k) \\ &= \text{prox}_{t\varphi_1}(z^k+tw^k), \end{aligned}$$

where the third line follows from line 5 of Algorithm 2, and in the last equality we use the definition of w^k.

Algorithm 3 shows DRS applied to the optimization problem (2.7), and summarizes the above calculations. Here, we invert the order of proximal steps in order to first perform the minimization problem associated with the objective function φ_2. and then the minimization problem corresponding to the regularization/penalization φ_1. More on this change of the order of the proximal steps will be discussed in Section 2.2.2.

Algorithm 3 Douglas-Rachford splitting method: optimization mode

Initialization: choose $t > 0$ and $z^{-1} \in \mathbb{R}^n$, $w^0 \in \mathbb{R}^n$.
for $k = 0, 1, 2, \ldots$ **do**
  φ_2**−proximal step**: $z^k = \operatorname{arg\,min}_{z \in \mathbb{R}^n} \left\{ \varphi_2(Mz) + \langle w^k, z \rangle + \frac{1}{2t}\|z - x^{k-1}\|^2 \right\}$.
  φ_1**−proximal step**: $x^k = \operatorname{arg\,min}_{x \in \mathbb{R}^n} \left\{ \varphi_1(x) + \frac{1}{2t}\|x - (z^k + tw^k)\|^2 \right\}$.
  Coordination step: $w^{k+1} = w^k + \frac{z^k - x^k}{t}$.
end for

Convergence of Algorithm 3 follows from Proposition 2.5, under mild regularity assumptions, as shown in [87, Proposition 3.40]. More specifically, we require the following two conditions:

1. $\partial(\varphi_1 + \varphi_2 \circ M) = \partial\varphi_1 + \partial\varphi_2 \circ M$: separability of the subdifferential of the sum of the two involved functions is essential to guarantee that solving $0 \in \partial\varphi_1(x) + \partial(\varphi_2 \circ M)(x)$ is equivalent to solving problem (2.7), so that the splitting is meaningful. The subdifferential of the sum can be separated whenever, for instance,

$$\operatorname{ri}(\operatorname{dom}(\varphi_1)) \cap \operatorname{ri}(M^{-1}(\operatorname{dom}(\varphi_2))) \neq \varnothing. \tag{2.8}$$

   This condition holds, for example, when both φ_1 and φ_2 are polyhedral functions (cf. [87, Proposition 3.23]). This assumption generalizes to the transversality condition $\partial^\infty\varphi_1(x) \cap -\partial^\infty(\varphi_2 \circ M)(x) = \{0\}$ of [29, Corollary 10.9].

2. Problem (2.7) needs to have at least one solution, so that the set of zeros of $\partial(\varphi_1 + \varphi_2 \circ M)$ is nonempty.

Proposition 2.6 (Convergence of DR in optimization mode: convex case)**.** *Consider problem (2.7) for $\varphi_1 \in \overline{\operatorname{conv}}(\mathbb{R}^n)$, $\varphi_2 \in \overline{\operatorname{conv}}(\mathbb{R}^m)$, such that problem (2.7) has a nonempty set of minimizers, and condition (2.8) holds. Then, the sequences $\{x^k\}$ and $\{z^k\}$ converge to a solution $x^\star$ to problem (2.7), and the sequence $\{w^k\}$ converges to a point $w^\star$ such that $-w^\star \in \partial\varphi_1(x^\star)$, and $w^\star \in \partial(\varphi_2 \circ M)(x^\star)$.*

The DRS method can separately exploit the properties of the functions of problem (2.7) to compute the proximal steps in Algorithm 3. In particular, for stochastic optimization problems, the splitting amounts to decouple the problem, since the objective function has a (further) separable structure, and the constraints couple the constraints. The next subsection describes DRS for stochastic optimization, also known as Progressive Hedging.

2.2.2 Progressive Hedging for stochastic optimization

A general multistage stochastic programming problem can be written as

$$\begin{cases} \min\limits_{x} & F(x) \\ \text{s.t.} & x \in \mathcal{N}\,, \end{cases} \tag{2.9}$$

where $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper lsc function, and $\mathcal{N}$ is the linear subspace of nonanticipative policies. Decisions are made throughout a time horizon divided in T stages, such that in each stage the underlying uncertainty is partially known and progressively revealed.

Figure 6 illustrates nonanticipativity in a scenario tree: in the beginning of stage $t = 1$ (first level/row), no uncertain information is known, and after deciding the policy associated with $t = 1$, part of the information is disclosed, that is, the random variable ξ_1 is realized. This means that in the beginning of the time horizon, all scenarios look alike, since there is no information available to distinguish them. This is represented with the horizontal dashed line in the first row of Figure 6b: all the connected nodes are the same root node of Figure 6a. Then, in state $t = 2$ (second level/row), it is possible to differentiate part of the scenarios using the revealed information, and thus a decision is made taking that into consideration. In Figure 6b we connect with a dashed line those nodes in the second row that in Figure 6a represent just one. After deciding the policy of stage $t = 2$, the random variable ξ_2 is realized, and no uncertain information is left to revealed. In the third and final stage $t = 3$, nonanticipativity does not enforce any constraint on the decisions variables.

The objective function has the special feature of being decomposable for different scenarios, namely,

$$F(x) = \sum_{s=1}^{S} p_s F_s(x_s),$$

where S is the (finite) number of possible scenarios of the underlying random variable of the problem, $p_s > 0$ is the probability of occurrence of scenario s, such that $\sum\limits_{s=1}^{S} p_s = 1$, and $F_s : \mathbb{R}^{n_s} \to \mathbb{R} \cup \{+\infty\}$ are proper lsc functions. The relationship between dimensions is $n = \sum\limits_{s=1}^{S} n_s$.

An extended formulation would be, for instance (5.1), where the scenario-separable constraint sets C_s are explicitly expressed. In this case, for each scenario $s = 1, \ldots, S$, if C_s is nonempty and closed, and $f_s : \mathbb{R}^{n_s} \to \mathbb{R} \cup \{+\infty\}$ is proper and lsc, then each $F_s = f_s + i_{C_s}$ is proper and lsc, and thus so is F. Therefore, (2.9) is as general as (5.1).

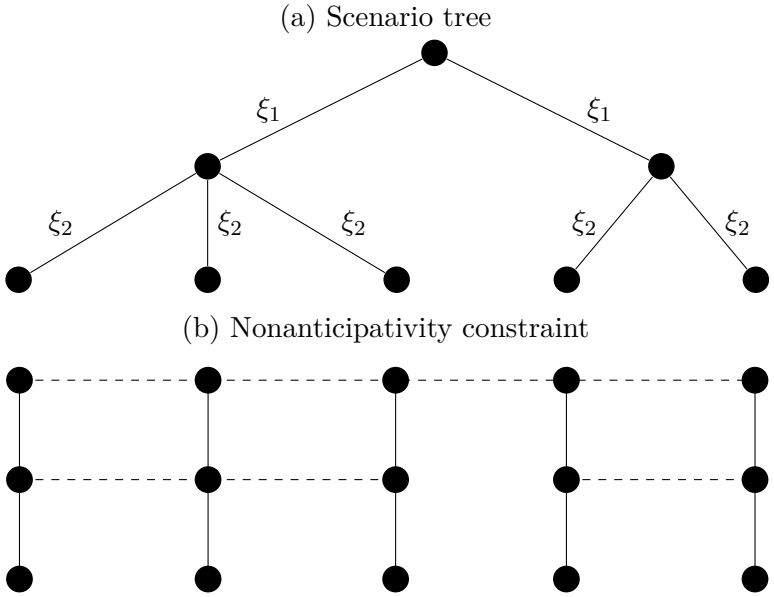


Figure 6 – Scenario tree for 3 stages, depicting with dashed lines the nonanticipativity constraint of stochastic programming.

For stochastic optimization problems, $\mathbb{R}^n$ is endowed with a weighted inner product defined as follows:

$$\langle x, z\rangle_S = \sum_{s=1}^{S} p_s \langle x_s, z_s\rangle,$$

where $x = (x_s)_{s=1}^S$, $z = (z_s)_{s=1}^S$, and $\langle\cdot,\cdot\rangle$ denotes the usual inner product in $\mathbb{R}^{n_s}$ (possibly defined in spaces of different dimension). We equip $\mathbb{R}^n$ with the induced weighted norm $\|\cdot\|_S = \sqrt{\langle\cdot,\cdot,\rangle_S}$, while $\|\cdot\|$ denotes the usual Euclidean norm of $\mathbb{R}^{n_s}$ induced by the inner product $\langle\cdot,\cdot\rangle$.

The Progressive Hedging algorithm (PHA) of Rockafellar and Wets [24] is a scenario-based decomposition method that generates a solution to (2.9). The PHA exploits the decomposable structure of problem (2.9), by appropriately handling the coupling constraint $x \in \mathcal{N}$. In fact, the PHA executes two levels of decomposition: first, it splits (in the sense of Section 2.2.1) the separable objective function F and the non-separable nonanticipativity constraint, and secondly, it further decomposes the objective function for different scenarios.

In Section 5.2.2 below, it is explained that the PHA iteratively solves a scenario-separable approximation of the Augmented Lagranian of problem (2.9), and then proceeds to improve the approximation and guarantee feasibility by projecting onto $\mathcal{N}$. A similar approach is taken in [27], overcoming the nonseparability of the Augmented Lagrangian by means of what the authors call a *nonlinear Jacobi* approach. More precisely, for each scenario s in parallel, in each iteration the Augmented Lagrangian is minimized over the

variables associated with the scenario s only, keeping the rest fixed, and then perform a correction step to improve the approximation. Another approach, called Progressive augmented Lagrangian method [88], uses an Augmented Lagrangian approach to solve an inner approximation of an optimization problem with probabilistic constraints. This approximation is progressively improved throughout iterations, a principle that the method of Chapter 5 also applies.

Originally, the PHA was shown by Rockafellar and Wets [24] to be a particular instance of the Spingarn's splitting method [89]. As shown in [90, Chapter 3], the PHA is an application of Douglas-Rachford splitting of Section 2.2.1. More specifically, the splitting occurs between the decomposable function F, and the simple indicator function $i_{\mathcal{N}}$, in such a way that particular properties of both functions are separately exploited.

Originally, the PHA was derived in [24] by applying the proximal point algorithm to certain maximal monotone operator that captures problem (2.9) in a primal–dual fashion. The PHA can be deduced directly from Algorithm 2 (cf. [90]) or Algorithm 3, as follows. First, we reformulate (2.9) in a ready-to-split manner:

$$\min_{x\in\mathbb{R}^n} F(x) + i_{\mathcal{N}}(x).$$

Then, taking $\varphi_2 = F$, $\varphi_1 = i_{\mathcal{N}}$, and $M = I$ in (2.7), yields Algorithm 5 below. In fact, the φ_1-step from Algorithm 3 corresponds to a projection step:

$$\begin{aligned} x^k &= \arg\min_{x\in\mathbb{R}^n}\left\{ i_{\mathcal{N}}(x) + \frac{1}{2t}\|x-(z^k+tw^k)\|^2 \right\} \\ &= \arg\min_{x\in\mathcal{N}}\left\{ \frac{1}{2t}\|x-(z^k+tw^k)\|^2 \right\} \\ &= P_{\mathcal{N}}[z^k+tw^k] \\ &= P_{\mathcal{N}}[z^k] + tP_{\mathcal{N}}[w^k]. \end{aligned}$$

If $w^0 \in \mathcal{N}^\perp$, by an induction argument, $w^k \in \mathcal{N}^\perp$ for all $k \in \mathbb{N}$. Indeed, suppose $w^k \in \mathcal{N}^\perp$, then $P_{\mathcal{N}}[w^k] = 0$, and $x^k = P_{\mathcal{N}}[z^k]$. Thus $z^k - x^k = P_{\mathcal{N}^\perp}[z^k] \in \mathcal{N}^\perp$. Therefore, $w^{k+1} = w^k + t^{-1}(z^k - x^k) \in \mathcal{N}^\perp$. Bearing in mind this observation, Algorithm 4 below states the Progressive Hedging algorithm of [24] derived from Algorithm 3. Note that in [24] the proximal parameter is $r = \dfrac{1}{t}$.

Note that by swapping the order of the proximal steps in Algorithm 3, in Algorithm 4 we first solve the problem associated with $\varphi_2 = F$, and then we perform the projection step, the step corresponding to φ_1.

As explained in [23] (see Chapter 5 below), the projection step 4 of Algorithm 4 is a simple calculation, since it corresponds to a conditional expectation taking into account the history of the random variable. Furthermore, note that step 3 of Algorithm 4 is

Algorithm 4 Progressive Hedging Algorithm

1: **Initialization**: Choose a primal-dual starting point $(x^0, w^1) \in \mathcal{N} \times \mathcal{N}^\perp$.
2: **for** $k = 1, 2, \ldots$ **do**
3: **Primal subproblems:** for each $s = 1, \ldots, S$, solve

$$z_s^k = \arg\min_{z_s \in \mathbb{R}^{n_s}} \left\{ F_s(z_s) + \langle w_s^k, z_s \rangle + \frac{1}{2t} \|z_s - x_s^{k-1}\|^2 \right\}. \tag{2.10}$$

4: **Primal projection:** $x^k = P_{\mathcal{N}}[z^k]$.
5: **Dual update:** $w_s^{k+1} = w_s^k + \frac{1}{t}(z_s^k - x_s^k)$ for $s = 1, \ldots, S$.
6: **end for**

step 3 of Algorithm 3 after capitalizing on the separable structure of the objective function, the weighted inner product, and the weighted norm. In principle, the S subproblems can be solved in parallel, since they are completely decoupled from each other.

The convergence of Algorithm 4 follows from Proposition 2.6, since the usual Euclidean norm on $\mathbb{R}^n$ and $\|\cdot\|_S$ define equivalent topologies.

Proposition 2.7. *Consider problem (2.9) for $F \in \overline{\text{conv}}(\mathbb{R}^n)$, such that (2.9) has a nonempty set of minimizers, and $\text{ri}(\text{dom}(F)) \cap \mathcal{N} \neq \varnothing$. Then, the sequences $\{x^k\}$ and $\{z^k\}$ converge to a solution $x^\star$ to problem (2.7), and the sequence $\{w^k\}$ converges to a point $w^\star \in -\partial F(x^\star) \cap \mathcal{N}^\perp$, such that $-w^\star$ is a solution to the dual problem of (2.7).*

Proof. It only remains to prove the last statement. Since $x^\star$ is a solution to (2.7), and $w^\star$ is such that $w^\star \in \mathcal{N}^\perp = \partial i_{\mathcal{N}}(x^\star)$, and $-w^\star \in \partial F(x^\star)$, then in view of [87, Proposition 3.26], strong duality holds and thus $w^\star$ is a solution to the dual problem

$$\min_w F^*(-w) + i_{\mathcal{N}^\perp}(w). \tag{2.11}$$

□

Remark 2.4. *Alternatively, we can choose M as the matrix that takes a vector x and returns Mx with the last-stage decision variables erased. This would lead to an analogous convergence analysis, bearing in mind that $\{Mz^k\}$ and $\{x^k\}$ would have the same limit $x^\star$, instead of $\{z^k\}$ and $\{x^k\}$.*

An alternative approach that leads to an equivalent formulation of the PHA is to apply the DRS method to the dual problem (2.11). In Algorithm 3, we choose $\varphi_2 = F^*(-\cdot)$, and $\varphi_1 = i_{\mathcal{N}^\perp}$. Starting from $\nu^0 \in \mathcal{N}$ and $q^{-1} \in \mathcal{N}^\perp$, this approach yields the following problem for F^*:

$$\min_{p \in \mathbb{R}^n} \left\{ F^*(-p) + \langle \nu^k, p \rangle + \frac{1}{2t} \|p - q^{k-1}\|^2 \right\},$$

with unique solution p^k. The problem for $i_{\mathcal{N}^\perp}$ is:

$$\min_{q\in\mathbb{R}^n}\left\{i_{\mathcal{N}^\perp}(q)+\frac{1}{2t}\|q-\left(p^k+t\nu^k\right)\|^2\right\}, \tag{2.12}$$

with unique solution q^k. Finally, the coordination step corresponds to

$$\nu^{k+1}=\nu^k+\frac{1}{t}(p^k-q^k). \tag{2.13}$$

We now proceed to deduce the primal form of this algorithm. The optimality condition of this problem reads

$$\begin{aligned} & & 0 &\in -\partial F^*(-p^k)+\nu^k+\frac{1}{t}(p^k-q^{k-1}).\\ \iff & \nu^k+\frac{1}{t}(p^k-q^{k-1}) &\in& \partial F^*(-p^k)\\ \iff & -p^k &\in& \partial F\left(\nu^k+\frac{1}{t}(p^k-q^{k-1})\right), \end{aligned}$$

where the last line follows from $(\partial F)^{-1}=\partial F^*$. Define $z^k:=\nu^k+\frac{1}{t}(p^k-q^{k-1})$. It thus holds that

$$\begin{aligned} & -\left(\nu^k+\frac{1}{t}(p^k-q^{k-1})\right) &\in& \partial F(z^k).\\ \iff & 0 &\in& \partial F(z^k)+q^{k-1}+t(z^k-\nu^k), \end{aligned}$$

meaning that z^k is the unique solution to

$$\min_{z\in\mathbb{R}^n}\left\{F(z)+\langle q^{k-1},z\rangle+\frac{t}{2}\|z-\nu^k\|^2\right\}.$$

With the identifications $w^k\leftarrow q^{k-1}$, $x^{k-1}\leftarrow\nu^k$, and $\frac{1}{t}\leftarrow t$, z^k coincides with the primal subproblem step 3 of Algorithm 4.

In order to formalize the identifications above, we need to prove that

$$\nu^{k+1}=P_{\mathcal{N}}[z^k], \text{ and } q^k=q^{k-1}+t(z^k-\nu^{k+1}).$$

We proceed by induction. Suppose $\nu^k\in\mathcal{N}$. From the definition of z^k, it holds $\nu^k=z^k+\frac{1}{t}(q^{k-1}-p^k)$. Substituting this relationship in (2.13) yields $P_{\mathcal{N}}[\nu^{k+1}]=P_{\mathcal{N}}[z^k]$. In order to prove the first claim, it only suffices $\nu^{k+1}\in\mathcal{N}$. From the assumption $\nu^k\in\mathcal{N}$ and (2.13), we would only need to prove that $p^k-q^k\in\mathcal{N}$. Since q^k solves (2.12), then

$$\begin{aligned} q^k &= P_{\mathcal{N}^\perp}[p^k+t\nu^k]\\ &= p^k+t\nu^k-P_{\mathcal{N}}[p^k+t\nu^k]. \end{aligned}$$

Therefore,

$$\begin{aligned} p^k - q^k &= P_{\mathcal{N}}[p^k + t\nu^k] - t\nu^k \\ &= P_{\mathcal{N}}[p^k] + t\big(P_{\mathcal{N}}[\nu^k] - \nu^k\big) \\ &= P_{\mathcal{N}}[p^k] - tP_{\mathcal{N}^\perp}[\nu^k] \\ &= P_{\mathcal{N}}[p^k], \end{aligned}$$

where the last line holds in view of the assumption $\nu^k \in \mathcal{N}$. In this way, $p^k - q^k \in \mathcal{N}$. Therefore, (2.13) implies $\nu^{k+1} \in \mathcal{N}$, and $\nu^{k+1} = P_{\mathcal{N}}[z^k]$. For the second claim, from the definition of p^k as the projection of $p^k + t\nu^k$ onto $\mathcal{N}^\perp$, and the definition of z^k that is equivalent to $p^k + t\nu^k = q^{k-1} + tz^k$, it holds that

$$\begin{aligned} q^k &= P_{\mathcal{N}^\perp}[q^{k-1} + tz^k] \\ &= q^{k-1} + tP_{\mathcal{N}^\perp}[z^k] \\ &= q^{k-1} + t\big(z^k - P_{\mathcal{N}}[z^k]\big) \\ &= q^{k-1} + t\big(z^k - \nu^{k+1}\big), \end{aligned} \tag{2.14}$$

where the second line follows from linearity of the projection operator and the fact $q^{k-1} \in \mathcal{N}^\perp$, the third line corresponds to the Moreau identity for the projection, and the last line is the first proved claim.

Hence, using the identifications $w^k \leftarrow q^{k-1}$, $x^{k-1} \leftarrow \nu^k$, and $\frac{1}{t} \leftarrow t$, applying the DRS method to the dual problem yields the PHA. An approach working on the dual problem, in particular, using a proximal bundle-like method to induce scenario-separability, is developed in Chapter 5, produces a bundle-like PHA.

3 A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods

This chapter is an extract of [22]:

Atenas, F., Sagastizábal, C., Silva, P. J., & Solodov, M. (2023). A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, 33(1), 89–115.

The concepts of variational analysis of the introduction of this article were discussed throughout Chapter 1, and Section 2 and Section 3 of said article are included in Section 1.3 and Section 1.4, respectively. Furthermore, Section 3.5 is taken from the appendix of [23]:

Atenas, F., & Sagastizábal, C. (2023). A bundle-like progressive hedging algorithm. *Journal of Convex Analysis*, special issue in honor of R. J-B Wets, 30(2) 453–479.

Some parts have been modified in order to follow the notation and structure of the present work.

Abstract. We present a framework for analyzing convergence and local rates of convergence of a class of descent algorithms, assuming the objective function is weakly convex. The framework is general, in the sense that it combines the possibility of explicit iterations (based on the gradient or a subgradient at the current iterate), implicit iterations (using a subgradient at the next iteration, like in the proximal schemes), as well as iterations when the associated subgradient is specially constructed and does not correspond neither to the current nor the next point (this is the case of descent steps in bundle methods). Under the subdifferential-based error bound on the distance to critical points, linear rates of convergence are established. Our analysis applies, among other techniques, to prox-descent for decomposable functions, the proximal-gradient method for a sum of functions, redistributed bundle methods, and a class of algorithms that can be cast in the feasible descent framework for constrained optimization.

Keywords. weak convexity, descent methods, bundle methods, model-based methods, proximal descent, proximal gradient method, error bound, linear convergence.

MSC codes. 90C30, 90C33, 90C55, 65K05.

3.1 Introduction

We consider algorithmically generated descent sequences that aim at solving problems of the form

$$\min \ f(x), \quad x \in \mathbb{R}^n, \tag{3.1}$$

where

$$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \text{ is weakly convex.}$$

The class of weakly convex functions is fairly broad and covers many problems of interest. It includes convex functions, differentiable functions with Lipschitzian gradient, certain compositions of convex functions with smooth functions, among others. We refer the readers to the discussion in [91], and Section 1.3. The case of constrained optimization will be handled by including into the objective function the indicator function of the feasible set.

Nonsmooth optimization problems like (3.1) arise frequently in applications involving big data and large-scale decision making. Many popular decomposition schemes exploit separable structures by resorting to Lagrangian or Fenchel duals. Typically, iterates are defined by means of certain *model functions*, resulting from some simplification of the objective function. To ensure convergence, primal and dual objects generated by nonsmooth methods must be interwined in very special and sound manner. For this reason, model functions build local approximations not only of the objective function, but also of its subdifferential. The theory presented below establishes a short set of conditions, on the family of model functions, on the primal and dual objects, and on problem (3.1) itself, that provides convergence guarantees for a large family of nonsmooth optimization methods.

More specifically, we are interested in stating conditions that ensure global convergence and local linear convergence rates for algorithms whose sequence of iterates $\{x^k\}$ involves the Clarke's subgradient information about f, possibly collected along iterations. Together with the algorithmically generated sequence $\{x^k\} \subseteq \mathbb{R}^n$, we shall also consider a certain theoretical sequence $\{z^k\} \subseteq \mathbb{R}^n$, with associated perturbation parameters $\{\varepsilon_k\} \subseteq [0, +\infty)$. These objects are introduced to account for the fact that, to compute the iterate x^k, one often minimizes a model/approximation of f. This operation yields a subgradient of the model, which for some methods in general is not a subgradient of f it-

self at any point in the sequence $\{x^k\}$. We show that model subgradients can, however, be "transported" to a nearby point, where they are subgradients of f. For convex functions, this is the well-known transportation formula in [30, Ch.XI, § 4.2]. For weakly convex functions, a similar result requires a delicate construction, given in Section 3.3 below. In particular, we think of $\{z^k\}$ as a (potential) perturbation, not necessarily computed by the algorithm, of the actual sequence $\{x^k\}$ which is computed indeed.

Formally, we shall consider frameworks with the following relations (3.2) valid, for a fixed along iterations k index $\mathtt{i} \in \{0, 1\}$. The index $\mathtt{i} \in \{0, 1\}$ is used to unify the analysis for explicit and implicit options in (3.2). Specifically, $\mathtt{i} = 1$ refers to explicit methods ($z^{k-\mathtt{i}} = z^{k-1}$, so that $g^k \in \partial f(z^{k-1})$), while $\mathtt{i} = 0$ refers to implicit methods ($z^{k-\mathtt{i}} = z^k$, so that $g^k \in \partial f(z^k)$). This feature would also be made more clear in comments and examples that follow (3.2). Again, recall that $\{x^k\}$ is the generated sequence, while $\{z^k\}$ is a theoretical one.

$$f(x^k) + a(\|x^k - x^{k-1}\|^2 + \varepsilon_{k-1}) \leqslant f(x^{k-1}), \text{ for } a > 0\,; \tag{3.2a}$$

$$\exists g^k \in \partial f(z^{k-\mathtt{i}}),\ \|g^k\| \leqslant b\Big(\|x^k - x^{k-1}\| + \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|\Big), \text{ for } b > 0\,; \tag{3.2b}$$

$$\text{both } \|x^k - z^k\| \text{ and } \{\varepsilon_k\} \text{ tend to 0 as } k \to \infty\,. \tag{3.2c}$$

Some remarks are in order. To start with, notice that condition (3.2a) ensures that the sequence of functional values $\{f(x^k)\}$ is non-increasing. By contrast, the theoretical sequence $\{f(z^k)\}$ is not necessarily non-increasing.

To continue, consider first the simplest instance, with $z^k = x^k$ and $\varepsilon_k = 0$. Then the conditions in (3.2c) are automatic, while (3.2b) becomes

$$\|g^k\| \leqslant b\|x^k - x^{k-1}\|\,,$$

for some subgradient g^k of f at either x^{k-1} or x^k. In the first case, it is natural to think of the scheme as being explicit (one obvious example is the gradient descent iteration, if f is differentiable: $x^k = x^{k-1} - t_k \nabla f(x^{k-1})$, with a suitable stepsize $t_k > 0$). In the second case, the scheme is in general implicit, and becomes essentially that of [18, § 2.3] if further $g^k \in \partial f(x^k)$ is taken. A prototypical instance is given by the proximal point iteration:

$$x^k \in \arg\min f(x) + \frac{1}{2t_k}\|x - x^{k-1}\|^2, \text{ for } t_k > 0\,, \tag{3.3}$$

which means that $x^k = x^{k-1} - t_k g^k$, for some $g^k \in \partial f(x^k)$.

Next, note that in the nonsmooth case, even the convex one, an explicit scheme with $g^k \in \partial f(z^{k-1})$ and $z^{k-1} = x^{k-1}$ in (3.2b) does not guarantee the descent condition

(3.2a). Indeed, this would be just the basic subgradient method, which is not of descent. General-purpose algorithms for nonsmooth optimization that build descent sequences are bundle methods [92, 30, 85]. Other nonsmooth methods can also be of descent, if they use more specific problem structure. Some examples are the prox-descent method for composite functions [44] and proximal-gradient methods for sums [93], considered together with the bundle method in Section 3.3 below. It is precisely for treating those type of methods that the theoretical iterate z^k and associated perturbation ε_k were introduced in our framework (3.2). Essentially, such schemes compute the proximal point of a convex *model* of the function f. Thanks to our transportation formula for weakly convex functions, this amounts to performing an explicit step, using a subgradient of f at a perturbed point, that plays the role of z^k in (3.2). This relation holds as long as the model-functions satisfy general conditions stated in Section 3.3. Therein, the process is developed in full details for model-based proximal methods, including serious steps of bundle algorithms for weakly convex functions.

Our convergence analysis recovers, from a unified perspective, various (but not necessarily all) results in sources like [33, 18, 91]. We also give new results, related to bundle methods for weakly convex functions. As stated in the concluding section of [94], developing a convergence theory along the lines of [18] for bundle methods based on practical oracles was an open question. We close this gap in Section 3.3, most notably by stating the linear convergence of descent steps of bundle methods with downshifted models that are typical in the nonconvex setting; we refer to Section 3.3.1 for details. When the objective in (3.1) is convex, linear rates for bundle-like methods can be traced back to [95] and [19]; see also the efficiency estimates in [96]. The topic was revisited more recently in [97] and [98], respectively considering strongly convex functions and multi-cut models, and the classical proximal bundle method for convex optimization. We should make it clear that our linear rate of convergence result for bundle methods concerns the descent iterations only, which themselves are constructed by a subsequence of so-called null steps. The number of null steps needed to produce descent is not part of our development. For strongly convex functions and a fixed prox-parameter, [97] shows that the precision of the solution at null steps is approximately inverse to the number of iterations. Taking into account null steps in the more general setting considered in this work is a challenging matter and should be a subject for future research.

The rest of the chapter is organized as follows. In Section 3.2 we discuss some general global convergence and local linear rate of convergence properties of the framework given by (3.2). In Section 3.3 these results are applied to model-based algorithms, including prox-descent for composite functions, proximal-gradient methods for sums, Taylor-based models, and finally the (serious steps of) bundle methods. In Section 3.4 we show how

our analysis applies to the class of feasible descent methods for constrained optimization considered in [33]. Finally, Section 3.5 extends the theory for constrained optimization problems by adding a projection step, so that constraints are dealt directly.

3.2 General asymptotic relations in the algorithmic pattern

In the sequel, we shall need the following technical result.

Lemma 3.1. *Let $\{a_k\} \subseteq \mathbb{R}^n$ and $\{b_k\} \subseteq [0, +\infty)$ be two sequences such that for all k it holds:*

$$\|a_k - a_{k-1}\| \leqslant \alpha_1 b_{k-1}$$

and

$$b_k \leqslant \alpha_2 b_{k-1}\,,$$

where $\alpha_1 > 0$ and $\alpha_2 \in (0, 1)$.

Then, there exists $a^* \in \mathbb{R}^n$ such that, for any $\bar{k}$, there exist $r \in (0, 1)$ and $c > 0$, such that for all $k \geqslant \bar{k}$,

$$\|a_k - a^*\| \leqslant c\alpha_2^k$$

with $c = \dfrac{\alpha_1 b_0}{1 - \alpha_2}$. In particular, $\{a_k\}$ converges to a^* R-linearly.

Proof. First, by direct induction, for all k it holds that $b_k \leqslant b_0(\alpha_2)^k$. By making a telescopic sum, for all $j \geqslant 1$,

$$\|a_{k+j} - a_k\| \leqslant \sum_{n=k+1}^{k+j} \|a_n - a_{n-1}\| \leqslant \frac{\alpha_1 b_0}{\alpha_2} \sum_{n=k+1}^{k+j} {\alpha_2}^n \leqslant \left(\frac{\alpha_1 b_0}{1 - \alpha_2}\right) {\alpha_2}^k\,, \tag{3.4}$$

where to obtain the last inequality we use that

$$\sum_{n=k+1}^{k+j} {\alpha_2}^n = {\alpha_2}^k \sum_{n=1}^{j} {\alpha_2}^n \leqslant {\alpha_2}^k \frac{\alpha_2}{1 - \alpha_2}\,,$$

since $\alpha_2 \in (0, 1)$. Therefore, $\{a_k\} \subseteq \mathbb{R}^n$ is a Cauchy sequence, and thus $\{a_k\}$ converges to some a^*. By taking the limit in (3.4) when $j \to \infty$, we obtain that $\|a_k - a^*\| \leqslant c\alpha_2^k$, as claimed. □

Regarding our problem of interest, if f in (3.1) is bounded below, the monotonically non-increasing sequence $\{f(x^k)\}$ from (3.2) converges, without any further assumptions (to some value, not necessarily a critical one). We next show that, for weakly convex functions satisfying the subdifferential error bound of Definition 1.7 and the isocost surfaces condition of Definition 1.1, the sequence of functional values of the projections of

the theoretical sequence $\{z^k\}$ onto S stabilizes at a critical value (value of f at a critical point).

In the statements (iv) and (v) below, the index $\mathtt{i} \in \{0,1\}$ is used to unify the analysis for explicit and implicit options in (3.2). Recall that $\mathtt{i} = 1$ refers to explicit methods ($z^{k-\mathtt{i}} = z^{k-1}$, so that $g^k \in \partial f(z^{k-1})$), while $\mathtt{i} = 0$ refers to implicit methods ($z^{k-\mathtt{i}} = z^k$, so that $g^k \in \partial f(z^k)$).

Lemma 3.2 (Convergence to critical points and technical relations)**.** *Let a function $f \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$, such that $\inf f > -\infty$. Then for any algorithmic scheme satisfying (3.2), the following hold:*

(i) $\{f(x^k)\}$ *monotonically converges to some value* $\tilde{f} \in \mathbb{R}$.

(ii) $x^k - x^{k-1} \to 0$, $z^k - z^{k-1} \to 0$ *and* $g^k \to 0$, *as* $k \to +\infty$.

Suppose, in addition, that f satisfies the proper separation of isocost surfaces condition (Definition 1.1) and the subdifferential error bound (Definition 1.7). Then, for any bounded sequence $\{x^k\}$ satisfying (3.2),

(iii) $\{f(z^k)\}$ *converges to* f^*, *where* $f^* \in \mathbb{R}$ *is a critical value (i.e.,* $f^* = f(x)$ *for some* $x \in S$*).*

(iv) For $\boldsymbol{\imath} \in \{0,1\}$, *defining* $\tilde{p}^{k-\imath} \in P_S(z^{k-\imath})$, *for all k sufficiently large the distance from $z^{k-\imath}$ to S can be estimated as*

$$\|z^{k-\imath} - \tilde{p}^{k-\imath}\|^2 \leqslant \frac{2\ell^2 b^2}{a}(f(x^{k-1}) - f(x^k)) + 2\ell^2 b^2 \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 .$$

(v) For the functional value errors $v^k := f(x^k) - f^*$, *it holds that*

$$v^k \leqslant \frac{2\ell b^2}{a}(v^{k-1} - v^k) + 2\ell b^2 \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 + \Theta^{k-\imath} ,$$

where

$$\Theta^{k-\imath} := f(x^{k-\imath}) - f(z^{k-\imath}) + \frac{\rho}{2}\|\tilde{p}^{k-\imath} - z^{k-\imath}\|^2 .$$

Proof. In view of (3.2a) and $\varepsilon_k \geqslant 0$, $\{f(x^k)\}$ is non-increasing. Since f is bounded below, item (i) follows immediately. Then also $f(x^{k-1}) - f(x^k) \to 0$.

As, by (3.2a),

$$\|x^k - x^{k-1}\|^2 \leqslant \frac{1}{a}(f(x^{k-1}) - f(x^k)) - \varepsilon_{k-1} , \tag{3.5}$$

it follows that $x^k - x^{k-1} \to 0$ in item (ii). Then (3.2b) and (3.2c) yield that $g^k \to 0$ and, $z^k - z^{k-1} = (z^k - x^k) + (x^k - x^{k-1}) + (x^{k-1} - z^{k-1}) \to 0$. Item (ii) is proven.

For the remaining items, we apply the subdifferential error bound at the tail of the auxiliary sequence $\{z^k\}$. The starting point is (3.2b), for which we use that f is a $\rho-$weakly convex function, considering the two possibilities $\mathtt{i} = 0$ and $\mathtt{i} = 1$ at the same time. For the rest of the proof, we fix the index $\mathtt{i} \in \{\mathtt{0}, \mathtt{1}\}$.

Since for all $k \geqslant 1$, $g^k \in \partial f(z^{k-i})$, it holds that

$$f(z^{k-\mathtt{i}}) + \langle g^k, x^k - z^{k-\mathtt{i}} \rangle \leqslant f(x^k) + \frac{\rho}{2}\|z^{k-\mathtt{i}} - x^k\|^2 .$$

In view of the fact that $f(x^k)$ decreases to $\tilde{f}$, $g^k \to 0$, $x^k - z^k \to 0$, $z^k - z^{k-1} \to 0$, and thus $z^{k-\mathtt{i}} - x^k \to 0$, we have that for all $\epsilon > 0$, and all sufficiently large k, $f(z^{k-\mathtt{i}}) \leqslant \tilde{f} + \epsilon$ and $g^k \in \partial f(z^{k-\mathtt{i}}) \cap B(0, \epsilon)$. Thus, by the error bound,

$$\|z^{k-i} - \tilde{p}^{k-i}\| = d(z^{k-i}, S) \leqslant \ell \|g^k\| . \tag{3.6}$$

Since $g^k \to 0$, it follows from (3.6) that $z^{k-\mathtt{i}} - \tilde{p}^{k-\mathtt{i}} \to 0$, and then $z^k - \tilde{p}_k \to 0$ as $k \to +\infty$. Combining this with the fact that $z^k - z^{k-1} \to 0$, yields that $\tilde{p}^k - \tilde{p}^{k-1} \to 0$. Moreover, the property of separation of the isocost surfaces implies that $f(\tilde{p}^k) = f^*$ eventually, for a critical value f^* of f. To complete the proof of item (iii), we apply weak convexity of f for $0 \in \partial f(\tilde{p}^k)$, obtaining that for all sufficiently large k it holds that

$$f^* = f(\tilde{p}^k) + \langle 0, z^k - \tilde{p}^k \rangle \leqslant f(z^k) + \frac{\rho}{2}\|z^k - \tilde{p}^k\|^2 .$$

Hence,

$$-\frac{\rho}{2}\|z^k - \tilde{p}^k\|^2 \quad \leqslant \quad f(z^k) - f^* . \tag{3.7}$$

Notice that, in addition, $g^k \in \partial f(z^{k-\mathtt{i}})$ implies that

$$f(z^{k-\mathtt{i}}) + \langle g^k, \tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}} \rangle \leqslant f(\tilde{p}^{k-\mathtt{i}}) + \frac{\rho}{2}\|\tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 = f^* + \frac{\rho}{2}\|\tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 , \tag{3.8}$$

where the last equality holds for all k sufficiently large.

Next, combining (3.7) and (3.8), we obtain that

$$-\frac{\rho}{2}\|z^k - \tilde{p}^k\|^2 \leqslant f(z^k) - f^* \leqslant \langle g^{k+\mathtt{i}}, z^k - \tilde{p}^k \rangle + \frac{\rho}{2}\|\tilde{p}^k - z^k\|^2 .$$

Then, taking the limit as $k \to \infty$ yields that $f(z^k) \to f^*$.

Next, weak convexity implies that for any $d^k \in \partial f(x^{k-\mathtt{i}})$,

$$f(x^{k-\mathtt{i}}) + \langle d^k, z^{k-\mathtt{i}} - x^{k-\mathtt{i}} \rangle \leqslant f(z^{k-\mathtt{i}}) + \frac{\rho}{2}\|z^{k-\mathtt{i}} - x^{k-\mathtt{i}}\|^2 .$$

Also, as $g^k \in \partial f(z^{k-\mathtt{i}})$,

$$f(z^{k-\mathtt{i}}) + \langle g^k, x^{k-\mathtt{i}} - z^{k-\mathtt{i}} \rangle \leqslant f(x^{k-\mathtt{i}}) + \frac{\rho}{2}\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 .$$

Combining the two relations above, we obtain that

$$\begin{aligned}\langle g^k, x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\rangle - \frac{\rho}{2}\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 &\leqslant f(x^{k-\mathtt{i}}) - f(z^{k-\mathtt{i}}) \\ &\leqslant \langle d^k, x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\rangle + \frac{\rho}{2}\|z^{k-\mathtt{i}} - x^{k-\mathtt{i}}\|^2 .\end{aligned}$$

Taking the limit in the last relation as $k \to +\infty$, Lemma 3.2(ii) and (3.2c) imply that $f(x^{k-\mathtt{i}}) - f(z^{k-\mathtt{i}}) \to 0$. Since $\{f(x^k)\}$ is a convergent sequence, and $f(z^k) \to f^*$, the sequences $\{f(x^k)\}$ and $\{f(z^k)\}$ both have the same limit. Thus, $\{f(x^k)\}$ is a non-increasing sequence converging to the critical value f^*, and $\{v^k\}$ is a nonnegative sequence.

To show statements (iv) and (v), recall that $(a+b)^2 \leqslant 2a^2 + 2b^2$, for all real numbers a, b. Then from (3.2b) we obtain that

$$\begin{aligned}\|g^k\|^2 &\leqslant b^2(\|x^k - x^{k-1}\| + \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|)^2 \\ &\leqslant 2b^2\|x^k - x^{k-1}\|^2 + 2b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 \\ &\leqslant \frac{2b^2}{a}(f(x^{k-1}) - f(x^k)) + 2b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 ,\end{aligned} \tag{3.9}$$

where the last inequality follows from (3.5). In this manner, since $g^k \in \partial f(z^{k-\mathtt{i}})$, from (3.6) and (3.9) it follows that

$$\|z^{k-\mathtt{i}} - \tilde{p}^{k-\mathtt{i}}\|^2 \leqslant \frac{2\ell^2 b^2}{a}(f(x^{k-1}) - f(x^k)) + 2\ell^2 b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 ,$$

which is statement (iv).

On the other hand, from (3.8), (3.6), and the fact that for all sufficiently large k it holds that $f(\tilde{p}^{k-\mathtt{i}}) = f^*$, we obtain that

$$\begin{aligned}f(z^{k-\mathtt{i}}) - f^* &\leqslant \|g^k\|\|z^{k-\mathtt{i}} - \tilde{p}^{k-\mathtt{i}}\| + \frac{\rho}{2}\|\tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 \\ &\leqslant \ell\|g^k\|^2 + \frac{\rho}{2}\|\tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 .\end{aligned}$$

Therefore, combining this inequality with (3.9), yields

$$f(z^{k-\mathtt{i}}) - f^* \leqslant \frac{2\ell b^2}{a}(f(x^{k-1}) - f(x^k)) + 2\ell b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 + \frac{\rho}{2}\|\tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 .$$

Hence, as $\{v^k\}$ is non-increasing,

$$v^k \leqslant \frac{2\ell b^2}{a}(v^{k-1} - v^k) + f(x^{k-\mathtt{i}}) - f(z^{k-\mathtt{i}}) + 2\ell b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 + \frac{\rho}{2}\|\tilde{p}^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 .$$

This concludes the proof. □

The relations in Lemma 3.2 lead to the following result, on the convergence speed of both the sequence of functional values and of iterates. The respective rates are linear in the quotient (Q) and root (R) senses, as defined in Chapter 2.

Recall that the index $\mathtt{i} \in \{0, 1\}$ unifies the explicit and implicit options in (3.2).

Theorem 3.1 (Asymptotic results for weakly convex functions)**.** *Let $f \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$ such that $\inf f > -\infty$. Suppose, in addition, that f satisfies the proper separation of isocost surfaces condition (Definition 1.1) and the subdifferential error bound (Definition 1.7).*

For any bounded sequence $\{x^k\}$ and $\{z^k\}$ satisfying (3.2), consider the sequence of functional errors $\{v^k\}$, defined in Lemma 3.2(v). If there exist $C_1, C_2 > 0$, such that, for all sufficiently large k it holds that

$$f(x^{k-\mathtt{i}}) - f(z^{k-\mathtt{i}}) \leqslant C_1(v^{k-1} - v^k) \tag{3.10}$$

and

$$\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 \leqslant C_2(v^{k-1} - v^k)\,, \tag{3.11}$$

then there exist $r \in (0,1)$ and $c > 0$ such that

(i) For all k sufficiently large,

$$v^k \leqslant r v^{k-1}\,,$$

where $r = M/(1+M) \in (0,1)$, and $M = C_1 + \ell b^2(2+\rho\ell)(1/a + C_2)$.

(ii) The sequence of functional errors $\{v^k\}$ monotonically converges to 0 with $Q-$linear rate.

(iii) The sequence $\{x^k\}$ converges $R-$linearly to a critical point x^* of f, such that $f(x^*) = f^* = \lim_{k\to\infty} f(x^k)$. More specifically, for all sufficiently large k,

$$\|x^k - x^*\| \leqslant c\sqrt{r}^k\,,$$

where $c = \dfrac{\sqrt{v_0}}{\sqrt{a}\,(1-\sqrt{r})}$.

Proof. First, convergence of $\{f(x^k)\}$ follows from Lemma 3.2(iii). The rate of convergence of $\{f(x^k)\}$ is derived from the technical estimates of Lemma 3.2. Indeed, combining the definition of $\Theta^{k-\mathtt{i}}$ with Lemma 3.2(iv) and (3.10), for all sufficiently large k it holds that

$$\begin{aligned}\Theta^{k-\mathtt{i}} &\leqslant C_1(v^{k-1} - v^k) + \frac{\rho}{2}\left(\frac{2\ell^2 b^2}{a}(v^{k-1} - v^k) + 2\ell^2 b^2 \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2\right)\\ &= \left(C_1 + \frac{\rho\ell^2 b^2}{a}\right)(v^{k-1} - v^k) + \rho\ell^2 b^2 \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2\,.\end{aligned}$$

Therefore, from Lemma 3.2(v), it further follows that

$$\begin{aligned}v^k &\leqslant \left(C_1 + \frac{\ell b^2}{a}(2+\rho\ell)\right)(v^{k-1} - v^k) + \ell b^2(2+\rho\ell)\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2\\ &\leqslant \left(C_1 + \frac{\ell b^2}{a}(2+\rho\ell)\right)(v^{k-1} - v^k) + \ell b^2(2+\rho\ell)C_2(v^{k-1} - v^k)\,,\end{aligned}$$

where (3.11) is used to obtain the last inequality. Hence, $v^k \leqslant M(v^{k-1} - v^k)$, which gives item (i) with M specified therein.

Using inductively the inequality of item (i), we conclude that there exists $c > 0$ such that for $r = M/(1+M)$ and all sufficiently large k,

$$v^k \leqslant v^0 r^k.$$

To see item (iii), the estimate therein follows from Lemma 3.1. More specifically, there exists a point x^* such that $\{x^k\}$ converges to x^* R-linearly. In particular, from (3.2c), $\{z^{k-\mathtt{i}}\}$ also converges to x^*, for $\mathtt{i} \in \{0,1\}$. Note that, since ∂f is an upper semicontinuous multifunction, Lemma 3.2(ii) and (3.2b) imply that $\partial f(z^{k-\mathtt{i}}) \ni g^k \to 0$, therefore $0 \in \partial f(x^*)$, that is, x^* is a critical point.

Finally, $z^k - \tilde{p}^k \to 0$ implies that $\tilde{p}^k \to x^*$, that is, x^* and $\tilde{p}^k$ are sufficiently close critical points. Therefore, in view of the proper separation of isocost surfaces property, $f(x^*) = f^*$. Hence, the limit of $\{x^k\}$ is a critical point $x^* \in f^{-1}(f^*)$. □

In the final two sections, Theorem 3.1 is applied to show the linear convergence rate of two different families of algorithms, proximal model-based ones akin to (the serious steps of) bundle methods, and the feasible descent framework of [33].

3.3 Bundle and proximal model-based methods

In nonsmooth optimization, satisfaction of (3.2a) is not straightforward. In addition to its role in Theorem 3.1, in this section weak convexity is an important ingredient in showing that iteratively minimizing appropriate approximating models of f indeed generates sequences that are of descent.

Suppose, for the moment, that f is a convex nonsmooth function. In this case, neither subgradient nor cutting-plane methods [85, Part II] fit the algorithmic pattern (3.2), because they do not guarantee the descent condition (3.2a). By contrast, as we shall show, serious steps within a bundle method do satisfy all the requirements. Bundle methods provide an implementable alternative for functions whose proximal point computation in (3.3) is difficult (or impossible). Before briefly reviewing the basic bundling mechanism, we mention that even for smooth functions, computing proximal points of some approximations of f has proven to be a useful technique to exploit decomposable structures. This is the basis of a plethora of approaches, including ADMM, as well as the prox-linear and prox-gradient methods considered below.

Having at hand a family of convex *model functions* for which computing proximal points is computationally implementable, in a bundle method [85, Part II] a candidate

iterate is defined as the proximal point of the model function at x^{k-1}. If the candidate satisfies a condition of sufficient descent for f, it is labeled a *serious step* x^k, and (3.2a) holds; otherwise the candidate is declared a *null step*. At a new iteration, the bundling process improves the model function and/or adjusts the proximal parameter. By this token, at serious steps the approximation of the proximal point is sufficiently good to ensure that errors incurred when replacing f by its model satisfy (3.2c).

For a convex f, a key ingredient in the convergence analysis of bundle methods is to relate the model subgradient associated with the prox-computation to certain ε-subgradient of f. The nonconvex setting precludes the use of approximate subdifferentials in this part of the analysis. For this reason, different ad-hoc approaches have been proposed in the literature. Rather than singling out some specific approach, below we develop a general convergence theory that is applicable to weakly convex functions. The key is to complement the algorithmic pattern of (3.2) with a suitable condition on the model functions used to approximate the proximal point of f. Our proposal unifies the global convergence analysis of a wide variety of methods in the literature, and also provides their linear rate of convergence.

3.3.1 Model function assumptions

Approximating the proximal point scheme (3.3) involves defining a family of simpler (than f) model functions whose proximal point is computed at each iteration. Often, a trade-off must be found between simplicity (fast prox-computation) and accuracy (increased chances of accepting the candidate as a serious step, i.e., satisfying (3.2a)).

Given $x \in \mathbb{R}^n$, consider modelling the function $f - f(x)$ by a convex function $\varphi_x : \mathbb{R}^n \to \mathbb{R}$. Note that f might be extended real-valued, while its model is finite everywhere. The most synthetic model uses the linearization introduced in Proposition 1.6,

$$\varphi_x^{\mathtt{sg}}(\cdot) = \ell_{x,g(x)}(\cdot) - f(x)\,.$$

Incidentally, computing the proximal point of this model amounts to one subgradient iteration, with stepsize given by the inverse of the prox-parameter.

A cutting-plane model is richer, as it takes the maximum over several linearizations, generated with past iterates x_i for $i \in \mathbb{B}$, the *bundle*:

$$\begin{aligned}\varphi_x^{\mathtt{cp}}(\cdot) \;&:=\; \max_{i\in\mathbb{B}} \left\{\ell_{x_i,g(x_i)}(\cdot) - f(x)\right\} = \max_{i\in\mathbb{B}} \left\{-e_i(x) + \langle g(x_i), \cdot - x\rangle\right\}\,,\\ &\qquad \text{where we define } e_i(x) := f(x) - \ell_{x_i,g(x_i)}(x)\,.\end{aligned}$$

The term $e_i(x)$, called *linearization error* in the bundle terminology, measures the quality of the linearization with respect to the reference point x. For convex f, the error is

nonnegative and the cutting-plane model satisfies $\varphi_x^{\mathtt{cp}} \leqslant f - f(x)$. But for nonconvex f this inequality cannot be ensured. To address this problem, a common approach is to downshift negative linearization errors, making them nonnegative. This can be done in different ways; typically, the error term

$$e_i(x) \text{ is replaced by } e_i^q(x) := \max\{e_i(x), \frac{q}{2}\|x_i - x\|^2\} \text{ for } q > 0 \text{ sufficiently large;}$$

see [99, 92] and, more recently, [100, 101, 102, 103]. The approach in [104, 105, 106] differs from those works, as it handles nonconvexity using *redistributed* models that, in addition to downshifting, tilt the slopes, as in Proposition 3.1 below.

In order to account for many alternative models in the literature, we shall assume that the family of model functions satisfies the following property. In the sequel, we shall show that it holds for many methods of interest.

Definition 3.1 (Models 1QA)**.** *A convex proper function $\varphi_x : \mathbb{R}^n \to \mathbb{R}$ is said to model f at x with one-sided quadratical accuracy, if*

$$\exists q > 0 : \forall y \in \mathbb{R}^n \quad \varphi_x(y) \leqslant f(y) - f(x) + \frac{q}{2}\|y - x\|^2 . \tag{3.12}$$

The property 1QA is a weakened form of the two-sided models considered in [68] and [55]. Making the condition unilateral is crucial for including bundle methods in the analysis (even when f is convex; see Figure 7 for an illustration).

The key role of convex 1QA models φ_x in convergence analyses is that they allow to *transport* subgradients, a mechanism that is not available for the nonconvex function f directly. Also, 1QA models are quite general, as the condition (3.12) can be satisfied both by cutting-plane-like models, where linearizations are oblivious to possible further information about f, and also by models that use structure. When a function has known structure, it is appealing to make the model inherit some of this feature. We next provide some examples.

3.3.1.1 Models defined using linearizations

For weakly convex functions, the simplest model $\varphi_x^{\mathtt{sg}}$ is clearly 1QA , taking $q = \rho$, the weak convexity parameter, but as already commented, the descent condition (3.2a) is not guaranteed for such a model, as it gives just a subgradient iteration. By contrast, the cutting-plane model with downshifted errors satisfies (3.12), as long as the iterates remain in a bounded set. The case of the more sophisticated model from [104, 105, 106] is analyzed below. Note that the model associated with the following result is equivalent to constructing a cutting-plane model for the "convexified" function $f(\cdot) + \frac{\rho}{2}\|\cdot - x\|^2$.

Proposition 3.1 (Redistributed models are 1QA). *Let $f \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$ and $x \in \mathbb{R}^n$. Given bundle elements $x_i, f(x_i), g(x_i) \in \partial f(x_i)$ for $i \in \mathbb{B}$, consider the downshifted linearization errors and tilted subgradients, respectively defined by*

$$e_i^\rho(x) := f(x) - \ell_{x_i,g(x_i)}(x) + \frac{\rho}{2}\|x - x_i\|^2 \quad \text{and} \quad g_i^\rho(x) := g(x_i) - \rho(x - x_i)\,.$$

Then the associated model $\varphi_x^\rho(\cdot) := \max\limits_{i\in\mathbb{B}} \{-e_i^\rho(x) + \langle g_i^\rho(x), \cdot - x\rangle\}$ is 1QA.

Proof. The model is convex, as the maximum of affine functions.

For any bundle element, weak convexity implies that, for all y,

$$f(y) + \frac{\rho}{2}\|y - x_i\|^2 \geqslant \ell_{x_i,g(x_i)}(y) = f(x_i) + \langle g(x_i), y - x_i\rangle\ .$$

Since $e_i(x) = f(x) - \ell_{x_i,g(x_i)}(x)$, rearranging terms, we obtain that

$$f(y) - f(x) \geqslant -e_i(x) + \langle g(x_i), y - x\rangle - \frac{\rho}{2}\|y - x_i\|^2\,.$$

Adding $\frac{\rho}{2}\|y - x\|^2$ to both sides yields

$$f(y) - f(x) + \frac{\rho}{2}\|y - x\|^2 \geqslant -e_i(x) + \langle g(x_i), y - x\rangle + \frac{\rho}{2}\Big(\|y - x\|^2 - \|y - x_i\|^2\Big)\,.$$

As

$$\frac{\rho}{2}\Big(\|y - x\|^2 - \|y - x_i\|^2\Big) = -\frac{\rho}{2}\|x - x_i\|^2 - \langle \rho(x - x_i), y - x\rangle\ ,$$

it follows that

$$\begin{aligned} & f(y) - f(x) + \frac{\rho}{2}\|y - x\|^2 \\ \geqslant\ & -\Big(e_i(x) + \frac{\rho}{2}\|x - x_i\|^2\Big) + \Big\langle \Big(g(x_i) - \rho(x - x_i)\Big), y - x\Big\rangle \\ =\ & -e_i^\rho(x) + \langle g_i^\rho(x), y - x\rangle\ . \end{aligned}$$

Since each of the terms defining the model φ_x^ρ satisfies (3.12), so does the model. □

In the redistributed proximal bundle method [105] iterates are generated with a model $\varphi_x^{\rho_k}$ whose augmentation parameter ρ_k is updated along the process, without knowing ρ beforehand. It is shown in [104] that unless x^{k-1} is critical, the procedure generates a serious step after a finite number of null iterations for weakly convex functions (f is uniformly prox-bounded in the language of that work). In [105] the serious step sequence is shown to be globally convergent under the same assumptions. Thanks to the theory developed in Section 3.3.2, based in Theorem 3.1, in addition to global convergence, we can now prove that serious steps converge at the linear rate. To the best of our knowledge, this is the first result on linear convergence rates for nonconvex bundle methods.

3.3.1.2 Decomposable functions, prox-descent and composite bundle methods

Recalling Definition 1.4, for decomposable functions $f = h \circ c$ the *ProxDescent* iterates [44, Algorithm 1] are defined by computing the proximal point of the model that is created by replacing the smooth mapping c with its Taylor expansion:

$$\varphi_x^{\mathtt{lw}}(\cdot) := h(c(x) + \nabla c(x)^\top(\cdot - x)) - f(x)\,.$$

In [55], the associated method is called *prox-linear*. We next show that the model $\varphi_x^{\mathtt{lw}}$ is 1QA under our assumptions (it should be noted that in [44] the outer function h can be more general, specifically extended-valued prox-regular).

Proposition 3.2 (Models for decomposable functions are 1QA)**.** *Let $h : \mathbb{R}^m \to \mathbb{R}$ be convex, finite-valued and positively homogeneous, and let $c : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable with its Jacobian being Lipschitz-continuous.*

Then the model $\varphi_x^{\mathtt{lw}}$ is 1QA.

Proof. Under the stated assumptions, $\varphi_x^{\mathtt{lw}}$ is convex.

As h is convex positive homogeneous and finite, it is the support function of a compact convex set D (that coincides with its subdifferential at 0), see [30, Chapter V] or [29, Theorem 8.24]. That is

$$h(d) = \max_{s\in D}\langle s, d\rangle.$$

Moreover, let L be the Lipschitz constant of the Jacobian of c. It follows that, for all $y, x \in \mathbb{R}^n$,

$$\|c(y) - c(x) - \nabla c(x)^\top(y - x)\| \leqslant \frac{L}{2}\|y - x\|^2.$$

Hence,

$$\begin{aligned}
h(c(x) + \nabla c(x)^\top(y - x)) &= h(c(x) + \nabla c(x)^\top(y - x) - c(y) + c(y)) \\
&= \max_{s\in D}\langle s, c(x) + \nabla c(x)^\top(y - x) - c(y) + c(y)\rangle \\
&\leqslant \max_{s\in D}\langle s, c(y)\rangle + \max_{s\in D}\langle s, c(x) + \nabla c(x)^\top(y - x) - c(y)\rangle \\
&\leqslant \max_{s\in D}\langle s, c(y)\rangle + \max_{s\in D}\|s\|\|c(x) + \nabla c(x)^\top(y - x) - c(y)\| \\
&\leqslant h(c(y)) + \frac{\max_{s\in D}\|s\|L}{2}\|y - x\|^2.
\end{aligned}$$

After adding $-f(x)$ on both sides, this is (3.12) with $q = \max_{s\in D}\|s\|L$. □

When computing the proximal point of $\varphi_x^{\mathtt{lw}}$ is computationally expensive, an alternative is to employ the composite proximal bundle method of [45]. The proposal

therein is to replace the outer function h by its cutting-plane model $h^{\mathtt{cp}}$, thereby computing the proximal point of the model

$$\varphi_x^{\mathtt{cs}}(\cdot) := h^{\mathtt{cp}}(c(x) + \nabla c(x)^\top(\cdot - x)) - f(x)\,.$$

By convexity of h, $\varphi_x^{\mathtt{cs}} \leqslant \varphi_x^{\mathtt{lw}}$. This model is also 1QA , by Proposition 3.2.

3.3.1.3 Sum of functions and prox-gradient method

Given a C^2-function f_1 with Lipschitz-continuous gradient and a convex function f_2, the proximal gradient method [93] minimizes $f := f_1 + f_2$ computing the proximal point of f_2 at $x^k - t_k \nabla f_1(x^k)$, $t_k > 0$. This is equivalent to computing the proximal point of the model that makes a Taylor linearization of f_1 and keeps f_2:

$$\varphi_x^{\mathtt{pg}}(\cdot) := f_1(x) + \langle \nabla f_1(x), \cdot - x\rangle + f_2(\cdot) - f(x)\,.$$

If f_2 is convex, then so is $\varphi_x^{\mathtt{pg}}$. Also, the 1QA property for the model follows directly from the Lipschitz-continuity of the gradient of f_1.

3.3.1.4 Taylor-like models

The theory in [55] uses powerful tools in Variational Analysis, including Ekeland's variational principle, to prove convergence of a variety of algorithmic schemes. Like in this work, the iterates are generated by computing a proximal point of some model. An important difference, however, is [55, relation (1.4)], which requires the model to approximate f not only uniformly but also *bilaterally* (from above and from below). Specifically, with our notation, the theory presented in [55] requires that

$$\exists q > 0 : \forall y \in \mathbb{R}^n \quad f(y) - f(x) - \frac{q}{2}\|y - x\|^2 \leqslant \varphi_x(y) \leqslant f(y) - f(x) + \frac{q}{2}\|y - x\|^2\,.$$

While this condition holds in several situations described in [55] (related to Taylor-like models), the two-sided quadratic requirement excludes cutting-plane models from the analysis. The reason is that, even for a convex f, linearizations in the cutting-plane model $\varphi_x^{\mathtt{cp}}$, the key ingredient in a bundle algorithm, may deviate from below from f in a non-polynomial manner. Figure 7 illustrates this phenomenon. Note that according to (3.13b), a lower bound condition for the model actually does hold, but it is related to the points x^k and x^{k-1} only. In particular, this requirement is weaker than asking for global quadratic accuracy from below, such as Taylor-like models.

3.3.2 Convergence theory for model-based methods

Using 1QA models φ_{x^k} approximating f, we shall consider the following algorithmic scheme, that will be shown to fit the framework of (3.2).

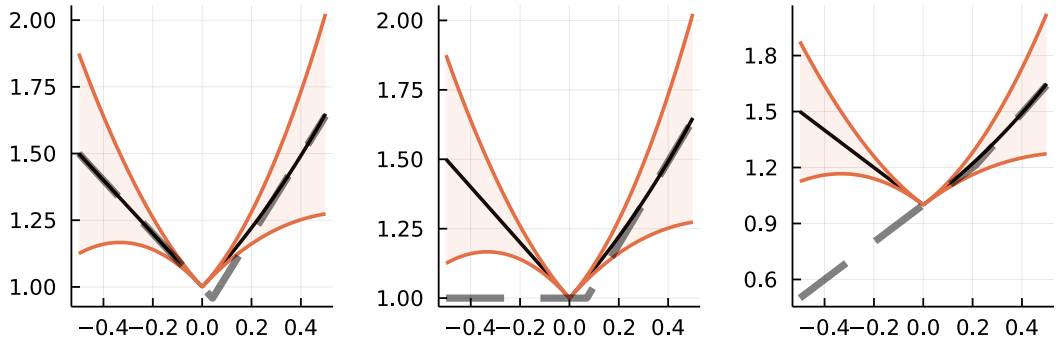


Figure 7 – For the function $f(x) = \begin{cases} 1-x & x \leqslant 0 \\ e^x & x > 0 \end{cases}$ plotted with a continuous dark line, three cutting-plane models are shown in dashed lines. These are all 1QA models, because they remain under the thick curved line in the top. By contrast, bilateral models with quadratic accuracy considered in [55] must lie in the shaded region. Even for this simple convex function, none of the cutting-plane models satisfies the two-sided condition in [55].

Starting from some $x_0 \in \mathbb{R}^n$, for all $k \geqslant 1$,

$$x^k = x^{k-1} - t_{k-1}G^{k-1}, \qquad \text{for } G^{k-1} \in \partial\varphi_{x^{k-1}}(x^k)\,, \tag{3.13a}$$

$$f(x^k) - f(x^{k-1}) \leqslant m\varphi_{x^{k-1}}(x^k), \qquad \text{for } m \in (0,1)\,. \tag{3.13b}$$

In particular, the new iterate is obtained computing the proximal point of the model, and the descent is measured using the value of the model at the new point. This is one of the characteristics of bundle methods. Other methods can also be recast in this manner. Below, we show that the sequences associated to the models described in Section 3.3.1.2 and Section 3.3.1.3 are of descent, both in the original sense of (3.2) and in the model-based sense of (3.13). Regarding the Taylor-like models in Section 3.3.1.4, the proposal in [55] does not consider a specific type of problem to be tackled by a particular method. So, as long as we are able to generate a descent sequence in the sense of (3.13), the results in Proposition 3.3 below would hold, since Taylor-like models are bilateral, while 1QA models are one-sided (in this sense, more general).

3.3.2.1 Decomposable functions and prox-descent method

Let f be a decomposable function as in Section 3.3.1.2, and consider the model $\varphi_x^{\mathtt{lw}}$ defined therein. Let $\{x^k\}$ be a prox-descent sequence as in [44, Algorithm 1].

First, (3.13a) is a direct consequence of the definition of the next iterate in [44, Algorithm 1] with stepsize $t_k := 1/\mu$. In order to see this, it suffices to recall that the step

$d := x^k - x^{k-1}$ is characterized by the relations

$$\nabla c(x^{k-1})^\top v + \frac{1}{t_{k-1}} d = 0, \quad v \in \partial h(c(x^{k-1}) + \nabla c(x^{k-1})^\top d).$$

Setting $G^{k-1} := \nabla c(x^{k-1})^\top v$, it holds that

$$x^k - x^{k-1} = d = -t_{k-1} G^{k-1}, \quad G^{k-1} \in \partial \varphi_{x^{k-1}}(x^k),$$

which is (3.13a). As for (3.13b), it is the same as the acceptance criterion for the step in [44, Algorithm 1] with $m = \sigma$.

Note also that it is proven in [44, Theorem 5.4] that [44, Algorithm 1] generates stepsizes t_k that are bounded away from zero. Thus, the algorithm satisfies the assumptions in Proposition 3.3 below.

3.3.2.2 Sum of functions and prox-gradient method

Let $f = f_1 + f_2$ be as in Section 3.3.1.3. The proximal gradient method conforms to the algorithmic pattern of (3.2) if $t_{\min} \leqslant t_k \leqslant 1/L_{f_1}$, where L_{f_1} is the Lipschitz constant of ∇f_1. Indeed, (3.2a) with $\varepsilon_{k-1} = 0$ and $a = L_{f_1}/2$ is a direct consequence of the decent properties of this algorithm; see, e.g., [107, Proposition 6.3.2]. As for (3.2b), we know that x^k minimizes

$$\varphi_x^{\mathrm{pg}}(\cdot) + \frac{1}{2t_k} \| \cdot - x^{k-1} \|^2.$$

Hence, there is $g_2^k \in \partial f_2(x^k)$ such that

$$\begin{aligned} 0 &= \nabla f_1(x^{k-1}) + g_2^k + \frac{1}{t_k}(x^k - x^{k-1}) \\ &= \nabla f_1(x^k) + g_2^k + \nabla f_1(x^{k-1}) - \nabla f_1(x^k) + \frac{1}{t_k}(x^k - x^{k-1}). \end{aligned}$$

Defining $z^k := x^k$, we have $g^k := \nabla f_1(x^k) + g_2^k \in \partial f(x^k)$ and

$$\begin{aligned} \|g^k\| &= \left\| \nabla f_1(x^{k-1}) - \nabla f_1(x^k) + \frac{1}{t_k}(x^k - x^{k-1}) \right\| \\ &\leqslant (L_{f_1} + 1/t_{\min}) \|x^k - x^{k-1}\|. \end{aligned}$$

This is (3.2b) with $b = L_{f_1} + 1/t_{\min}$. Finally, (3.2c) holds trivially.

3.3.2.3 Convergence of sequences generated by model-based methods

To continue with our analysis, we need to exhibit the errors ε_k and the theoretical sequence $\{z^k\}$ from (3.2) that are associated with the bundle-like scheme (3.13). We start by transporting subgradients of convex models of nonconvex functions to the convex function obtained from f, by weak convexity. This relation and Theorem 3.2 below yield z^k as a perturbation of the iterate x^k, as desired.

Proposition 3.3 (Transportation of subgradients and the validity of (3.2a))**.** *Consider the minimization of a function $f \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n)$ applying the model-based proximal scheme in (3.13) with models φ_x that are 1QA with parameter $q \leqslant \rho$ in Definition 3.1, and let $G^k \in \partial\varphi_{x^k}(x^{k+1})$ as in (3.13a). The following holds for all k.*

(i) The model aggregate error at x^k,

$$E^k := -t_k\|G^k\|^2 - \varphi_{x^k}(x^{k+1})\,,$$

satisfies $E^k \geqslant 0$.

(ii) If for all $x \in \mathbb{R}^n$, $F_x(\cdot)$ denotes the (convex) function $f(\cdot) + \frac{\rho}{2}\|\cdot - x\|^2$, then a subgradient G^k in (3.13a) can be transported to be the convex E^k-subgradient of F_{x^k} at x^k:

$$G^k \in \partial_{E^k} F_{x^k}(x^k).$$

Suppose, in addition, that $\inf f > -\infty$, and the proximal step sizes are bounded:: $t_{\max} \geqslant t_k \geqslant t_{\min} > 0$. Then,

(iii) both $\{G^k\}\,, \{E^k\}$ converge to 0 as $k \to \infty$, and

(iv) condition (3.13b) is equivalent to (3.2a) written with $a = m/t_{\max}$ and $\varepsilon_k = t_{k-1}E^{k-1}$.

Proof. Since the models are 1QA , taking $x = y = x^k$ in (3.12) gives that $\varphi_{x^k}(x^k) \leqslant 0$. By the convexity of the model and the iterate definition in (3.13a), it holds that

$$\begin{aligned} 0 &\geqslant \varphi_{x^{k-1}}(x^{k-1}) \\ &\geqslant \varphi_{x^{k-1}}(x^k) + \left\langle G^{k-1}, x^{k-1} - x^k \right\rangle \\ &= \varphi_{x^{k-1}}(x^k) + t_{k-1}\|G^{k-1}\|^2 = -E^{k-1}\,, \end{aligned}$$

and $E^k \geqslant 0$ for all k, as stated in item (i).

To show item (ii), because the model is 1QA , we have that

$$\begin{aligned} \varphi_{x^{k-1}}(x) &\leqslant f(x) - f(x^{k-1}) + \frac{\rho}{2}\|x - x^{k-1}\|^2 \\ &= F_{x^{k-1}}(x) - f(x^{k-1})\,. \end{aligned}$$

Combining now the model convexity with (i) yields

$$\begin{aligned} F_{x^{k-1}}(x) \geqslant f(x^{k-1}) + \varphi_{x^{k-1}}(x) &\geqslant f(x^{k-1}) + \varphi_{x^{k-1}}(x^k) + \left\langle G^{k-1}, x - x^k \right\rangle \\ &= f(x^{k-1}) + \left\langle G^{k-1}, x - x^{k-1} \right\rangle \\ &\quad + \varphi_{x^{k-1}}(x^k) + \left\langle G^{k-1}, x^{k-1} - x^k \right\rangle \\ &= F_{x^{k-1}}(x^{k-1}) + \left\langle G^{k-1}, x - x^{k-1} \right\rangle - E^{k-1}\,. \end{aligned}$$

As the last relation is (ii) written with k replaced by $k-1$, the desired result follows.

To show item (iii), note that the descent condition (3.13b), written using the aggregate gradient and error definitions, gives

$$m(E^{k-1} + t_{k-1}\|G^{k-1}\|^2) \leqslant f(x^{k-1}) - f(x^k)\,. \tag{3.14}$$

As $\{f(x^k)\}$ is non-increasing and f is bounded below, this sequence is convergent. Hence, $f(x^{k-1}) - f(x^k) \to 0$ as $k \to \infty$. Then from (3.14) and $t_k \geqslant t_{\min} > 0$, it follows that $E^k \to 0$ and $G^k \to 0$.

Finally, rewriting the descent condition (3.13b) using the aggregate gradient and error definitions yields (iv), as

$$f(x^k) + \frac{m}{t_{k-1}}(\|x^k - x^{k-1}\|^2 + t_{k-1}E^{k-1}) \leqslant f(x^{k-1})\,.$$

□

Proposition 1.6 shows that for weakly convex functions Clarke's and proximal subgradients are equivalent concepts. As noted by one reviewer, the transportation result in item (ii) of Proposition 3.3 yields an ε-proximal subdifferential for f at x, satisfying satisfying for all y the inequality

$$f(y) \geqslant f(x) + \langle g, y - x\rangle + \frac{\rho}{2}\|y - x\|^2 - \varepsilon\,.$$

To complete formulating (3.13) in the format of the algorithmic pattern in (3.2), we show the validity of (3.2b) and (3.2c). This is achieved applying the error bound inequality in Definition 1.7, noting that it involves the exact (Clarke) subgradients of f. We have just shown that the transported model subgradient is an E^k-subgradient of the auxiliary convex function F_{x^k} at x^k. The connection with the original function f is done by means of the following result, reproduced from [19, Theorem 2].

Theorem 3.2 (Brøndsted-Rockafellar's like relation)**.** *Let $F \in \overline{\text{conv}}(\mathbb{R}^n)$. Suppose that $E \geqslant 0$ and that $G \in \partial_E F(x)$. Then, for each $\gamma > 0$, there is a unique $y = y(\gamma)$ such that*

$$G - \frac{1}{\gamma}y \in \partial F(x + \gamma y)\,, \quad \|y\| \leqslant \sqrt{E}\,.$$

By the above result, any $\varepsilon-$subgradient of a convex function can be perturbed to obtain an exact subgradient of the same function, at a perturbed point. Since weak convexity gives an explicit relation between f and the convex function F_x, we shall be able to relate the respective subgradients, and apply the subdifferential error bound for f using the perturbed points.

Lemma 3.3 (Casting (3.13) in the format of (3.2))**.** *Under the assumptions of Proposition 3.3, suppose f satisfies the subdifferential error bound of Definition 1.7 and the sequence of stepsizes $\{t_k\}$ in (3.13a) is bounded: $t_{\max} \geqslant t_k \geqslant t_{\min} > 0$. Then there exists a theoretical sequence $\{z^k\}$ such that all conditions in (3.2) hold, with $\|z^k - x^k\| \leqslant \sqrt{\ell E^k}$.*

Proof. The validity of (3.2a) was already shown in Proposition 3.3(iv).

To derive the expression for z^k, apply Theorem 3.2 written with $G := G^k \in \partial_{E^k} F_{x^k}(x^k)$ for the convex function $F := F_{x^k}$, $E := E^k$, taking $\gamma := \sqrt{\ell} > 0$, where $\ell > 0$ is the constant of the subdifferential error bound in Definition 1.7. It follows that there exists a unique y^k such that

$$\|y^k\| \leqslant \sqrt{E^k} \quad \text{and} \quad G^k - \frac{1}{\sqrt{\ell}} y^k \in \partial F_{x^k}\left(x^k + \sqrt{\ell} y^k\right) = \partial f\left(x^k + \sqrt{\ell} y^k\right) + \rho\sqrt{\ell} y^k \,,$$

by the definition of F_{x^k}. Therefore, letting

$$z^{k-1} := x^{k-1} + \sqrt{\ell} y^{k-1} \text{ it holds that } g^{k-1} := G^{k-1} - \left(\frac{1+\rho\ell}{\sqrt{\ell}}\right) y^{k-1} \in \partial f(z^{k-1}) \,.$$

To show that condition (3.2c) holds, first notice that

$$\frac{1}{\sqrt{\ell}} \|z^{k-1} - x^{k-1}\| = \|y^{k-1}\| \leqslant \sqrt{E^{k-1}} \,.$$

Since $E^k \to 0$ by Proposition 3.3(iii), this means that $z^{k-1} - x^{k-1} \to 0$. The remaining condition $\varepsilon_k \to 0$ follows from the expression $\varepsilon_{k-1} = t_{k-1} E^{k-1}$ in Proposition 3.3(iv), combined with the boundedness assumption on t_k, using once more that $E^k \to 0$.

To show that the sequence $\{g^k \in \partial f(z^{k-1})\}$ satisfies condition (3.2b), notice that

$$\begin{aligned} \|G^{k-1}\| &\leqslant \|G^{k-1}\| + \left(\frac{1+\rho\ell}{\sqrt{\ell}}\right) \|y^{k-1}\| \\ &= \frac{1}{t_k}\|x^k - x^{k-1}\| + \left(\frac{1+\rho\ell}{\sqrt{\ell}}\right) \frac{1}{\sqrt{\ell}} \|z^{k-1} - x^{k-1}\| \\ &\leqslant \frac{1}{t_{\min}}\|x^k - x^{k-1}\| + \left(\frac{1+\rho\ell}{\ell}\right) \|z^{k-1} - x^{k-1}\| \,. \end{aligned}$$

Hence, (3.2b) holds with $b := \max\{1/t_{\min}, (1+\rho\ell)/\ell\}$. □

Thanks to Lemma 3.3, we are now in position of applying Theorem 3.1 to show that the general scheme based on models considered in this section converges, with a rate that is R-linear for the iterates and Q-linear for the functional values.

Theorem 3.3 (Global convergence of (3.13) and local linear rate)**.** *Let $f \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$ such that $\inf f > -\infty$. Suppose, in addition, that f satisfies the proper separation of isocost surfaces (Definition 1.1) and the subdifferential error bound (Definition 1.7), and that the*

sequence of stepsizes $\{t_k\}$ *in* (3.13a) *is bounded:* $t_{\max} \geqslant t_k \geqslant t_{\min} > 0$. *The following holds for the model-based proximal scheme in* (3.13)*, as long as the models* φ_x *therein are 1QA (Definition 3.1 with parameter* $q \leqslant \rho$*), and the generated sequence* $\{x^k\}$ *is bounded.*

(i) $\{f(x^k)\}$ *monotonically converges to some critical value* f^**, such that the sequence of functional errors* $\{v^k = f(x^k) - f^*\}$ *converges to 0 with Q-linear rate:*

$$\exists r \in (0,1) : v^k \leqslant q v^{k-1} \text{ for all sufficiently large } k.$$

(ii) The sequence of iterates $\{x^k\}$ *converges to a critical point* x_* *of* f *with R-linear rate:*

$$\exists r \in (0,1) \text{ and } c > 0 : \|x^k - x^*\| \leqslant c\sqrt{q}^k \text{ for all sufficiently large } k.$$

Proof. To see item (i), we apply Theorem 3.1. First, from the definition of the aggregate error E^k and (3.14), it follows that

$$E^{k-1} \leqslant \frac{1}{m}(v^{k-1} - v^k)\,,$$

$$\|G^{k-1}\|^2 \leqslant \frac{1}{m t_{k-1}}(v^{k-1} - v^k)\,.$$

The first inequality combined with the definition of z^k imply that

$$\|x^{k-1} - z^{k-1}\|^2 \leqslant \ell E^{k-1} \leqslant \frac{\ell}{m}(v^{k-1} - v^k)\,.$$

Moreover, combining the last inequalities with $G^{k-1} \in \partial_{E^{k-1}} F_{x^{k-1}}(x^{k-1})$, the definition of z^k, and the fact that t_k is bounded away from 0, we obtain that

$$\begin{aligned} f(x^{k-1}) - f(z^{k-1}) &\leqslant \frac{\rho}{2}\|z^{k-1} - x^{k-1}\|^2 - \langle G^{k-1}, z^{k-1} - x^{k-1}\rangle + E^{k-1} \\ &\leqslant \frac{\rho\ell}{2m}(v^{k-1} - v^k) + \|G^{k-1}\|\sqrt{\frac{\ell}{m}(v^{k-1} - v^k)} + \frac{1}{m}(v^{k-1} - v^k) \\ &\leqslant \frac{1}{m}\left(\frac{\rho\ell}{2} + \sqrt{\frac{\ell}{t_{\min}}} + 1\right)(v^{k-1} - v^k)\,. \end{aligned}$$

Since (3.10) and (3.11) in Theorem 3.1 hold for

$$C_1 = \frac{1}{m}\left(\frac{\rho\ell}{2} + \sqrt{\frac{\ell}{t_{\min}}} + 1\right) \quad \text{and} \quad C_2 = \frac{\ell}{m}\,,$$

items (i) and (ii) follow. □

3.4 The theory applied to constrained smooth optimization

Another application of our unified analysis is the feasible descent framework of [33] (see also [108]). Consider the constrained optimization problem (1.13), where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable with Lipschitz-continuous gradient on the nonempty closed convex set $C \subseteq \mathbb{R}^n$.

The work [33] considers iterative sequences $\{x^k\}$ satisfying

$$x^k = P_C(x^{k-1} - t_{k-1}\nabla f(x^{k-1}) + E^{k-1}), \quad t_{k-1} \geqslant t_{\min} > 0, \tag{3.15a}$$

$$\|E^{k-1}\| \leqslant \alpha\|x^k - x^{k-1}\|, \quad \alpha \in (0,1). \tag{3.15b}$$

This setting is quite broad. It includes, of course, the basic gradient projection method, taking $E^k = 0$ for all k. But, depending on the form of the mapping e that gives E^{k-1} in (3.2), it includes many other algorithms for solving problem (1.13). Some examples are the extragradient method, the proximal point method, coordinate descent, and several splitting techniques; see [33] and references therein.

We next show that our general analysis of (3.2) is applicable to methods given by (3.15) as well. We consider (3.2) for the function $f + i_C$ and take, for all $k \geqslant 1$, $\varepsilon_k = 0$ and $x^k = z^k$ (note that (3.2c) is then automatic). Under the stated assumptions, $f + i_C$ is weakly convex; see Proposition 1.7. We next need to show that (3.15) implies (3.2a) and (3.2b) for $f + i_C$. Once this is done, we apply Theorem 3.1 for the weakly convex function $f + i_C$.

The proof below that the sequence $\{x^k\}$ from (3.15) satisfies the descent condition (3.2a) for $f + i_C$ is essentially a similar argument as in [33] for f, because by (3.15a) it holds that $x^k \in C$ for all k (and so $(f + i_C)(x^k) = f(x^k)$). We include this part of the proof here mostly for completeness. Note, however, that the subgradients of f and of $(f + i_C)$ are *not* the same. Also, our rate of convergence analysis is different, as our results are based on the subdifferential error bound (Definition 1.7), while [33] uses the projection error bound (Definition 1.8). Therefore, our results are new when the error bounds are different (see the comments in Section 1.4.1 regarding the comparisons of the error bounds in question).

Proposition 3.4 (The feasible descent framework (3.15) fits (3.2))**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with $L-$Lipschitz continuous gradient on the nonempty closed convex set $C \subseteq \mathbb{R}^n$. Then any sequence $\{x^k\}$ satisfying (3.15) is a sequence of descent for the function $f + i_C$ in the sense of (3.2). More specifically,*

(i) For all k,

$$f(x^k) + \left(\frac{1-\alpha}{t^*} - \frac{L}{2}\right)\|x^k - x^{k-1}\|^2 \leqslant f(x^{k-1}),$$

whenever $t_k \leqslant t^* \leqslant 2(1-\alpha)/L$. I.e., (3.2a) holds for $f + i_C$ (recall that $x^k \in C$).

(ii) For all k, there exists $u^k \in N_C(x^k)$ such that

$$\|\nabla f(x^k) + u^k\| \leqslant \left(\frac{1+\alpha}{t_{\min}} + L\right)\|x^k - x^{k-1}\|\,,$$

i.e., (3.2b) holds for $f + i_C$.

Proof. From (3.15a) and the characterization of the projection operator, for all $y \in C$ it holds that

$$\langle x^{k-1} - t_{k-1}\nabla f(x^{k-1}) + E^{k-1} - x^k, y - x^k\rangle \leqslant 0\,.$$

Taking $y = x^{k-1}$ in this inequality and rearranging terms, we obtain that

$$\|x^{k-1} - x^k\|^2 - t_{k-1}\langle \nabla f(x^{k-1}), x^{k-1} - x^k\rangle \leqslant \langle E^{k-1}, x^k - x^{k-1}\rangle\,.$$

Using the Cauchy-Schwarz inequality and (3.15b) on the right-hand side, it holds that

$$\|x^{k-1} - x^k\|^2 - t_{k-1}\langle \nabla f(x^{k-1}), x^{k-1} - x^k\rangle \leqslant \alpha\|x^{k-1} - x^k\|^2\,.$$

It follows that

$$\langle \nabla f(x^{k-1}), x^k - x^{k-1}\rangle \leqslant \frac{\alpha - 1}{t_{k-1}}\|x^{k-1} - x^k\|^2\,.$$

Since the function is differentiable with Lipschitz-continuous gradient with constant L, by [47, Lemma A.11] we have that

$$f(x^k) - f(x^{k-1}) \leqslant \langle \nabla f(x^{k-1}), x^k - x^{k-1}\rangle + \frac{L}{2}\|x^k - x^{k-1}\|^2\,.$$

Combining the last two inequalities above gives

$$f(x^k) - f(x^{k-1}) \leqslant \left(\frac{\alpha - 1}{t_{k-1}} + \frac{L}{2}\right)\|x^{k-1} - x^k\|^2\,,$$

from which item (i) follows.

We next prove item (ii), i.e., condition (3.2b) for $f + i_C$. Again, from (3.15a) and the characterization of the projection operator, there exists $\nu^k \in N_C(x^k)$ such that

$$x^{k-1} - t_{k-1}\nabla f(x^{k-1}) + E^{k-1} - x^k = \nu^k\,.$$

Defining $u^k = \nu^k/t_{k-1} \in N_C(x^k)$, we have that

$$t_{k-1}u^k = x^{k-1} - x^k + E^{k-1} - t_{k-1}\nabla f(x^{k-1})\,,$$

and

$$t_{k-1}(\nabla f(x^k) + u^k) = x^{k-1} - x^k + E^{k-1} + t_{k-1}(\nabla f(x^k) - \nabla f(x^{k-1}))\,.$$

Define $\bar{u}^k = \nabla f(x^k) + u^k \in \partial(f + i_C)(x^k)$. We then obtain that

$$\begin{aligned} \|\bar{u}^k\| &\leqslant \frac{1}{t_{k-1}}\|x^{k-1} - x^k + E^{k-1}\| + \|\nabla f(x^k) - \nabla f(x^{k-1})\| \\ &\leqslant \left(\frac{1+\alpha}{t_{\min}} + L\right)\|x^k - x^{k-1}\|, \end{aligned}$$

where the triangle inequality, (3.15b), and the Lipschitz-continuity of the gradient of f were used. The proof is complete. □

Due to Proposition 3.4 and Proposition 1.7, we are now in position to apply our unified analysis for weakly convex functions to obtain estimates for the rate of convergence in (3.15).

Theorem 3.4 (Linear rate of convergence of (3.15))**.** *Under the assumptions of Proposition 3.4, if f is bounded from below, the subdifferential error bound (Definition 1.7) and the proper separation of isocost surfaces condition (Definition 1.1) hold, then for any bounded iterates $\{x^k\}$ satisfying (3.15) it holds that:*

(i) There exists some critical value $f^* \in \mathbb{R}$ of f such that $f(x^k) \to f^*$. For $v^k := f(x^k) - f^*$, there exists $r \in (0,1)$ such that for all sufficiently large k,

$$v^k \leqslant r v^{k-1}.$$

(ii) $\{x^k\}$ converges R-linearly to a critical point x^* of f with $f(x^*) = f^*$. More specifically, there exists $c > 0$ such that for all k sufficiently large,

$$\|x^k - x^*\| \leqslant c\sqrt{r}^k$$

Proof. By Proposition 1.7, $f + i_C$ is a weakly convex function. By Proposition 3.4, any sequence $\{x^k\}$ satisfying (3.15) conforms to (3.2) and all the conditions of Theorem 3.1, with $x^k = z^k$, $\varepsilon_k = 0$ for all k, and $g^k \in \partial(f + i_C)(x^k)$. Then the assertions follow from Theorem 3.1, with

$$r = \frac{M}{1+M}, \quad M = \frac{2\ell\left(\frac{1+\alpha}{t_{\min}+L}\right)^2}{\frac{1-\alpha}{t^*} - \frac{L}{2}(1+L\ell)}, \quad c = \frac{\sqrt{v_0}}{\sqrt{\frac{1-\alpha}{t^*} - \frac{L}{2}}\left(1-\sqrt{q}\right)}.$$

□

Note that while the scheme (3.15) is explicit in our terminology, as it uses the gradient of f at x^{k-1}, it is cast in our framework (3.2) as being implicit, as the subgradient of $f + i_C$ is taken therein at x^k.

3.5 Projective variant for constrained optimization

As mentioned in the beginning of the chapter, the following section is extracted from [23]:

> Atenas, F., & Sagastizábal, C. (2023). A bundle-like progressive hedging algorithm. *Journal of Convex Analysis*, special issue in honor of R. J-B Wets, 30(2) 453–479.

Given a function $h \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$ and $\mathcal{M}$ a nonempty closed convex set, onto which computing a projection is not costly, we are interested in solving the following constrained minimization problem

$$\min_{w \in \mathcal{M}} h(w) \tag{3.16}$$

by means of a bundle-like method. For the multistage program (5.1), the constraints are given by $\mathcal{M} = \mathcal{N}^\perp$ in problem (5.7), but any easy-to-project feasible set could be considered (for instance a nonnegative orthant, $\mathcal{M} = \{w \geqslant 0\}$).

In the setting of problem (3.16), a family of *model functions* $\{\varphi^k_{w^k}\}$, built along iterations, is available. Denoting by $i_\mathcal{M}$ the indicator function of the feasible set, given $w^k \in \mathcal{M}$,

$$\text{the model is a convex function } \varphi^k_{w^k}(\cdot) \text{ approximating } h(\cdot) + i_\mathcal{M}(\cdot) - h(w^k)\,. \tag{3.17}$$

If the objective function h is convex, a typical construct is to create a cutting-plane model for the sum $h + i_\mathcal{M}$. For nonconvex h, cutting-plane models need to be tilted and/or downshifted to be adequate. If the parameter of weak convexity ρ is known, this is an easy task. Otherwise, estimates of such parameter must be "guessed" and suitably updated along the iterative scheme, as in the redistributed proximal bundle method from [104, 105].

In our development, to generate iterates with a special subsequence converging with linear rate, the following property should be satisfied by the models. Recall from Section 3.3.1, that a family of finite-valued convex functions $\varphi^k_{w^k} : \mathbb{R}^n \to \mathbb{R}$ satisfying (3.17) is one-sided with quadratic accuracy for $h + i_\mathcal{M}$ if there exists $q > 0$ such that

$$\forall w^k \in \mathcal{M}\,, \forall w \in \mathbb{R}^n\,, \quad \varphi^k_{w^k}(w) \leqslant h(w) + i_\mathcal{M}(w) - h(w^k) + \frac{q}{2}\|w - w^k\|^2\,. \tag{3.18}$$

It is shown in Proposition 3.1 that the redistributed approach from [104, 105], which defines piecewise linear models for $h(w) + \frac{q}{2}\|w - w^k\|^2$, satisfies the 1QA property.

Model functions are useful to define iterates with low computational burden. Finding a new iterate in our proposal amounts to solving

$$\min_{w} \left\{ \varphi_{w^k}^k(w) + \frac{1}{2t_k}\|w - w^k\|^2 \right\}, \tag{3.19}$$

a problem that is a simple quadratic program if the model is piecewise linear. The first-order optimality condition for problem (3.19),

$$0 \in \partial\varphi_{w^k}^k(w) + \frac{1}{t_k}(w - w^k),$$

characterizes the implicit updating rule (3.20a) given below.

Given parameters $m \in (0,1)$, $\eta > 0$, and some initial $w^0 \in \mathcal{M}$, for all $k \geqslant 0$, our algorithmic scheme defines

$$w^{k+\frac{1}{2}} = w^k - t_k G^k, \qquad \text{for some } G^k \in \partial\varphi_{w^k}^k(w^{k+\frac{1}{2}}), \tag{3.20a}$$

$$u^{k+1} = P_{\mathcal{M}}(w^{k+\frac{1}{2}}), \tag{3.20b}$$

$$\text{if } h(u^{k+1}) - h(w^k) \leqslant m\varphi_{w^k}^k(w^{k+\frac{1}{2}}), \quad \text{declare a serious step.} \quad \text{Set } w^{k+1} = u^{k+1} \text{ and } t_{k+1} \geqslant t_{\min}. \tag{3.20c}$$

$$\text{Otherwise,} \quad \text{declare a null step.} \quad \text{Set } w^{k+1} = w^k \text{ and choose } t_{k+1}. \tag{3.20d}$$

In both cases select a new model $\varphi_{w^{k+1}}^{k+1}$.

In this algorithmic pattern, iterates satisfying (3.20a)-(3.20c) form the so-called *serious-step subsequence* in bundle methods. As indicated by (3.20c), at serious steps the objective functional values decrease. Iterates satisfying (3.20a) and (3.20b) but not (3.20c) form the subsequence of *null steps*.

Since iterates in bundle methods can be of two types, the convergence theory splits the asymptotic analysis into two parts, depending on whether the serious step sequence is finite or infinite. We state below the corresponding result for the latter. Note that the result extends the theory of descent methods in Section 3.3, to take into consideration the extra projective step in (3.20b).

For the next result, by construction (cf. (3.20c)), recall that the stepsizes t_k are bounded from below by $t_{\min} > 0$ whenever a serious step is performed.

Theorem 3.5 (Global convergence and local linear rate of serious subsequence)**.** *Consider problem* (3.16) *with the following assumptions:*

$$h \in w - \overline{\mathrm{conv}}_\rho(\mathbb{R}^n) \tag{A1}$$

$$\mathcal{M} \text{ is a nonempty closed convex set;} \tag{A2}$$

$$h \text{ is bounded below on } \mathcal{M}, \text{ that is } \inf_{\mathcal{M}} h > -\infty\,; \tag{A3}$$

$$h + i_{\mathcal{M}} \text{ satisfies the property of Definition 1.1;} \tag{A4}$$

$$h + i_{\mathcal{M}} \text{ satisfies the error bound of Definition 1.7.} \tag{A5}$$

Assume that in the pattern (3.20) *the functions* $\varphi^k_{w^k}$ *are 1QA models having parameter* $q \leqslant \rho$ *as in* (3.12)*, and that the stepsizes* t_k *are bounded above by* $t_{\max} > 0$ *in step* (3.20c)*.*

If there is an infinite subsequence of serious steps, that is $\{w^k\}$ *satisfying* (3.20a)-(3.20c)*, the following holds.*

(i) The subsequence of functional values $\{h(w^k)\}$ *monotonically converges to some critical value* h^* *of* $h + i_{\mathcal{M}}$*, such that the sequence of functional errors* $\{v^k = h(w^k) - h^*\}$ *converges to 0 with Q-linear rate: there exists* $r \in (0,1)$ *such that for all sufficiently large* k*,*

$$v^{k+1} \leqslant r v^k\,.$$

(ii) The subsequences of iterates $\{w^k\}$ *and intermediate points* $\{w^{k+\frac{1}{2}}\}$ *converge to a critical point* w^* *of* $h + i_{\mathcal{M}}$ *with R-linear rate: there exists* $r \in (0,1)$*, and* $c > 0$ *such that for all sufficiently large* k

$$\|w^k - w^*\| \leqslant c\sqrt{r}^k, \quad \|w^{k+\frac{1}{2}} - w^*\| \leqslant c(2-\sqrt{r})\sqrt{r}^k$$

□

Similarly to Section 3.3 the first step is to exploit the model convexity to transport the model subgradient to an approximate subgradient of the "convexified" function associated to weak convexity. Then, a Brøndsted-Rockafellar's like relation makes the connection with the objective function in (3.16). The corresponding results are gathered in the following statement, listing several technical relations. None of those relations involves the projection step, they can be shown following the arguments in Lemma 3.4 and Lemma 3.3, applied to the function $H = h + i_{\mathcal{M}}$.

Lemma 3.4 (Transportation of subgradients, descent, theoretical sequence)**.** *Let* $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ *and* $\mathcal{M}$ *satisfy* (A1)−(A2)*. Consider solving problem* (3.16) *applying the model-based proximal scheme in* (3.20)*. If the models* $\varphi^k_{w^k}$ *are of type 1QA with parameter* $q \leqslant \rho$*, as in Definition 3.1, the following hold for all* k*.*

(i) The aggregate error $E^k := -t_k\|G^k\|^2 - \varphi^k_{w^k}(w^{k+\frac{1}{2}})$ *satisfies* $E^k \geqslant 0$*.*

(ii) $G^k \in \partial_{E^k} H_{w^k}(w^k)$ *where* $H_{w^k}(\cdot)$ *denotes the "convexified" function* $h(\cdot) + i_{\mathcal{M}}(\cdot) + \frac{\rho}{2}\|\cdot - x\|^2$

If (A3) *holds, and the proximal stepsizes* t_k *are bounded,* $t_k \in [t_{\min}, t_{\max}]$*, then*

(iii) the sequences of model subgradient and errors $\{G^k\}$ *and* $\{E^k\}$ *converge to* 0 *as* $k \to \infty$*, and*

(iv) condition (3.20c) *is equivalent to*

$$h(w^{k+1}) + \frac{m}{t_{\max}}\left(\|w^{k+\frac{1}{2}} - w^k\|^2 + t_k E^k\right) \leqslant h(w^k)\,. \tag{3.22}$$

If, in addition, (A5) *holds, then*

(v) there exists a theoretical auxiliary sequence $\{z^k\}$ *such that* $\|z^k - w^k\| \leqslant \sqrt{\ell E^k}$ *and*

$$\exists g^k \in \partial h(z^k) \text{ and } \nu^k \in N_{\mathcal{M}}(z^k),\ \|g^k + \nu^k\| \leqslant b\Big(\|w^{k+\frac{1}{2}} - w^k\| + \|w^k - z^k\|\Big), \text{ for } b > 0\,. \tag{3.23}$$

Proof. Items (i)−(iv) follow directly from Proposition 3.3(i)-(iv) applied to $h + i_{\mathcal{M}}$. Item (v) follows from Lemma 3.3 applied to $h + i_{\mathcal{M}}$. Indeed, apply [19, Theorem 2] to the function H_{w^k} and the pair of points (G^k, w^k) such that $G^k \in \partial_{E^k} H_{w^k}(w^k)$, it follows that there exists d^k such that $\|d^k\| \leqslant \sqrt{E^k}$ and

$$G^k - \frac{1}{\sqrt{\ell}} d^k \in \partial H_{w^k}(w^k + \sqrt{\ell} d^k),$$

where $\ell > 0$ is the constant of condition (A5). Defining $z^k := w^k + \sqrt{\ell} d^k$, we have that $\|z^k - w^k\| \leqslant \sqrt{\ell E^k}$, and $\partial H_{w^k}(z^k) = \partial(h + i_{\mathcal{M}})(z^k) + \rho\sqrt{\ell} d^k$. Therefore, for $g^k_{\mathcal{M}} := G^k + (1 + \rho\ell)\sqrt{\ell^{-1}} d^k$, it holds that

$$g^k_{\mathcal{M}} \in \partial(h + i_{\mathcal{M}})(z^k) = \partial h(z^k) + N_{\mathcal{M}}(z^k).$$

Thus, there exists $g^k \in \partial h(z^k)$ and $\nu^k \in N_{\mathcal{M}}(z^k)$ such that $g^k_{\mathcal{M}} = g^k + \nu^k$. Moreover, from the definition of $g^k_{\mathcal{M}}$, d^k, and (3.20a), we also have

$$\begin{aligned}\|g^k + \nu^k\| &\leqslant \|G^k\| + (1+\rho\ell)\sqrt{\ell^{-1}}\|d^k\| \\ &= \frac{1}{t_k}\|w^k - w^{k+\frac{1}{2}}\| + (1+\rho\ell)\ell^{-1}\|z^k - w^k\| \\ &\leqslant \frac{1}{t_{\min}}\|w^k - w^{k+\frac{1}{2}}\| + (1+\rho\ell)\ell^{-1}\|z^k - w^k\|\,,\end{aligned}$$

then condition (3.23) holds for $b := \max\{t_{\min}^{-1}, (1+\rho\ell)\ell^{-1}\}$.

□

Lemma 3.5. *Let $h : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ and $\mathcal{M}$ satisfy (A1)–(A3), and that the stepsizes t_k remain in $[t_{\min}, t_{\max}]$. Then the following holds:*

(i) $\{h(w^k)\}$ monotonically converges to some value $\tilde{h} \in \mathbb{R}$.

(ii) $w^{k+1} - w^k \to 0$, $z^{k+1} - z^k \to 0$ and $g^k + \nu^k \to 0$, as $k \to +\infty$.

Suppose, in addition, (A4) and (A5) hold. Then

(iii) $\{h(z^k)\}$ and $\{h(w^k)\}$ both converge to f^*, where $f^* \in \mathbb{R}$ is a critical value (i.e., $f^* = h(w)$ for some $w \in [\partial(h + i_{\mathcal{M}})]^{-1}(0)$).

(iv) Defining $\tilde{p}^k \in P_{\mathcal{S}}(z^k)$, for all k sufficiently large, the distance from z^k to S can be estimated as

$$\|z^k - \tilde{p}^k\|^2 \leqslant \frac{2\ell^2 b^2}{a}(h(w^k) - h(w^{k+1})) + 2\ell^2 b^2 \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2\,.$$

(v) For the functional value errors $v^k := h(w^k) - f^*$, it holds that

$$v^{k+1} \leqslant \frac{2\ell b^2}{a}(v^k - v^{k+1}) + 2\ell b^2 \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 + \Theta^k\,,$$

where

$$\Theta^k := h(w^k) - h(z^k) + \frac{\rho}{2}\|\tilde{p}^k - z^k\|^2\,.$$

Proof. The following proof is an adaptation of the proof of Lemma 3.2, taking into consideration the extra projection step in the present setting. To see item (i), notice that, since $\mathcal{M}$ is convex, then $P_{\mathcal{M}}$ is a non-expansive operator, and together with $w^k \in \mathcal{M}$, imply

$$\|w^{k+1} - w^k\| = \|P_{\mathcal{M}}(w^{k+\frac{1}{2}}) - P_{\mathcal{M}}(w^k)\| \leqslant \|w^{k+\frac{1}{2}} - w^k\|$$

Combined with (3.22), it yields for some constant $a > 0$

$$h(w^{k+1}) + a(\|w^{k+1} - w^k\|^2 + \varepsilon_k) \leqslant h(w^k) \tag{3.24}$$

which, in particular, implies that the sequence of function values of $\{w^k\}$ is non-increasing. Since h is bounded below on $\mathcal{M}$, monotonicity of $\{h(w^k)\}$ implies that $h(w^k) \to \tilde{h}$, for some $\tilde{h} \in \mathbb{R}$. Note that inequality (3.24) corresponds to (3.2a) for $\varepsilon_k = t_{\min} E^k$.

For item (ii), observe that (3.24) also implies, combined with the fact that $h(w^{k+1}) - h(w^k) \to 0$, that $w^{k+1} - w^k \to 0$. Note that $w^{k+\frac{1}{2}} - w^k \to 0$ follows similarly from (3.22). Together with Lemma 3.4(v), imply $g^k + \nu^k \to 0$.

Item (iii) follows directly from Lemma 3.2(iii), because this original result does not depend directly on (3.23). In particular, since $g^k + \nu^k \to 0$ and $h(z^k) \to f^*$, we can apply the error bound to obtain, for all sufficiently large k,

$$\|z^k - \tilde{p}^k\| \leqslant \ell \|g^k + \nu^k\| \,. \tag{3.25}$$

Regarding item (iv), from (3.23) and (3.24), there holds

$$\begin{aligned} \|g^k + \nu^k\|^2 &\leqslant b^2(\|w^{k+\frac{1}{2}} - w^k\| + \|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|)^2 \\ &\leqslant 2b^2\|w^{k+\frac{1}{2}} - w^k\|^2 + 2b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 \\ &\leqslant \frac{2b^2}{a}(h(w^k) - h(w^{k+1})) + 2b^2\|x^{k-\mathtt{i}} - z^{k-\mathtt{i}}\|^2 \,, \end{aligned} \tag{3.26}$$

and the results follows from (3.25). The final item (v) is just Lemma 3.2(v). □

Finally, with the tools constructed above, we prove Theorem 3.5.

Proof. The results of global convergence and local rate of convergence are similar to Theorem 3.3, now taking into account the projection step (3.20b).

First, let $\{z^k\}$ be defined as in Lemma 3.4(v), and consider the sequence of functional errors $\{v^k\}$, defined in Lemma 3.5(v). Then, there exist constants $C_1\,, C_2 > 0$, such that for all sufficiently large k, there holds

$$h(w^k) - h(z^k) \leqslant C_1(v^k - v^{k+1}) \tag{3.27}$$

and

$$\|w^k - z^k\|^2 \leqslant C_2(v^k - v^{k+1}) \,. \tag{3.28}$$

Indeed, these two estimates directly follow from the proof of Theorem 3.3. Estimate (3.28) comes from the definition of z^k, and the descent condition (3.22), while estimate (3.27) follows from the E^k-subgradient inequality of H_{w^k} for G^k and (3.28).

The proof of item (i) in Theorem 3.5 is exactly the same as for Theorem 3.1(i), due to the fact that it depends on Lemma 3.5. Indeed, use (3.27), (3.28), and Lemma 3.5(iv)

in the estimate of Lemma 3.5(v), and then rearrange terms to deduce (i) for $r = M/(1+M) \in (0,1)$, and $M = C_1 + \ell b^2(2+\rho\ell)(1/a + C_2)$.

To deduce Theorem 3.5(ii), we can apply Lemma 3.1, using (3.24) (which is exactly the same as (3.2a)) and item (i). Therefore, $\{w^k\}$ converges to some w^* with a $R-$linear rate: for all sufficiently large k,

$$\|w^k - w^*\| \leqslant c\sqrt{r}^k, \text{ where } c = \frac{\sqrt{v^0}}{\sqrt{a}\,(1-\sqrt{r})}.$$

The fact that w^* is critical can be similarly proven as in the original analysis, using Lemma 3.4(v) and the fact that the subdifferential ∂h is an upper semicontinuous multi-function.

Regarding the sequence of intermediate points, for all k,

$$\|w^{k+\frac{1}{2}} - w^*\| \leqslant \|w^{k+\frac{1}{2}} - w^k\| + \|w^k - w^*\|$$

In the right-hand side, the second term is bounded by $c\sqrt{r}^k$, therefore it only remains to bound the first one. From (3.24) and the fact that $\varepsilon_k, v^k \geqslant 0$, it follows that

$$\|w^{k+\frac{1}{2}} - w^k\|^2 \leqslant \frac{1}{a}(h(w^k) - h(w^{k+1})) = \frac{1}{a}(v^k - v^{k+1}) \leqslant \frac{1}{a}v^k$$

From (ii), it holds that

$$\|w^{k+\frac{1}{2}} - w^k\|^2 \leqslant \frac{v^0}{a}r^k$$

Therefore,

$$\|w^{k+\frac{1}{2}} - w^*\| \leqslant c(2-\sqrt{r})\sqrt{r}^k,$$

which concludes the proof of Theorem 3.5. □

It remains to analyze the asymptotic behavior of the "tail of null steps". To do so, the specific definition of the model functions plays a crucial role. Since this feature is problem dependent, we consider the particular instance of interest, that is problem (5.1)-(5.7), and give a complete convergence analysis for the bundle PH algorithm presented in Chapter 5.

3.6 Final remarks

In this chapter, we presented a framework that merges the analysis of explicit and implicit methods of descent for weakly convex optimization. First, we provided convergence results for a general abstract scheme as introduced in [22], and then a particular case of model-based methods, resembling the serious steps of proximal bundle methods.

The convergence results comprise global convergence to a critical point, and local rates of convergence by assuming a subdifferential-based error bound. This way of reasoning bears a resemblance with the seminal work [18] for functions satisfying the KŁ inequality. However, and differently from our case, the framework considered by the authors prevents the inclusion of bundle-like methods.

We continued with an extension of [22] for constrained optimization problems, where the constraints are directly treated by means of an extra projection step onto the feasible set. This approach corresponds to the appendix of [23], and resembles the basic idea of the splitting methods: we separately handle the objective function and the constraints modeled with an indicator function, since they have different structural properties. In this case, we retrieved analogous convergence results as the original analysis.

When the projective variant of [22] is applied to the dual formulation of a stochastic programming problem, it yields a scenario-based decomposition method, which will be discussed in Chapter 5. Additionally, similar arguments can be carried out on a merit function capturing the essential properties of the problem, and obtain similar convergence results. That is the case of the Douglas-Rachford splitting method, examined in Chapter 4.

It is important to observe that in this work, we use a notion of weak convexity that is global. In the literature [109, 41, 110], sometimes a weaker and seemingly more general condition is assumed, namely, that around each point x, there exists a neighborhood U, and a constant ρ_x, such that $f(\cdot) + \frac{\rho_x}{2}\|\cdot\|^2$ is convex on U. When the analysis is restricted over a compact set C, then this weaker condition is actually equivalent to Definition 1.3 due to compacity of C, with the same weak convexity parameter ρ over all C.

4 Nonconvex Douglas-Rachford splitting via descent of merit functions

Abstract. We analyze Douglas-Rachford splitting techniques applied to solving weakly convex optimization problems. Under mild regularity assumptions, and by the token of a suitable merit function called the Douglas-Rachford envelope, we show convergence to critical points and local linear rates of convergence. The resulting iterates can be interpreted to be generated by a descent method applied to this merit function. The Douglas-Rachford envelope plays here an analogous role as the Moreau envelope for the proximal point algorithm. This feature allows us to extend to the nonconvex nondifferentiable setting arguments employed in the analysis of the gradient descent method in convex differentiable optimization.

4.1 Introduction and motivation

Decomposition techniques are fundamental to deal with complex systems represented by large-scale sophisticated model formulations. Decomposition can be achieved by separating problems in simpler and smaller subproblems, depending on the involved variables and constraints, and also on the number of possible outcomes. Separability can also stem from structural properties, for instance, splitting smooth and nonsmooth, or convex and nonconvex parts, in the objective function of the problem of interest. Splitting methods have been successfully applied in signal processing, image processing, and machine learning, see [111, 112, 113, 12, 114] for a few illustrations of applications.

Operator splitting methods decompose complex structured problems into simpler individual pieces. A solution of the original problem is obtained by iteratively solving separate subproblems for each involved function, or more generally, operator. Prominent instances suitable for composite optimization are the Douglas-Rachford and the Peaceman-Rachford (PRS) splittings, the Alternating Direction Method of Multipliers (ADMM), the Spingarn's partial inverse and the Forward-Backward methods; we refer to [11, 115, 87, 12, 116] and references therein for details. When applied to optimization problems, these methods were originally studied for linear, and more generally, convex programming.

The cornerstone of most operator splitting methods is the proximal point algorithm, introduced by [16] and thoroughly studied by [17] to find a zero of a maximal monotone operator, see also Section 2.1.1. In the context of convex optimization, for

$\varphi_1, \varphi_2 \in \overline{\text{conv}}(\mathbb{R}^n)$, the problem boils down to minimizing the composite function $\varphi_1 + \varphi_2$. As explained in [87] and Section 2.2.1, recall one DRS iteration, given in the scheme (4.2) below and Algorithm 2, amounts to applying the PPA with constant stepsize equal to 1, to the auxiliary maximal monotone operator

$$[\text{prox}_{c\varphi_1} \circ (2\,\text{prox}_{c\varphi_2} - I) + (I - \text{prox}_{c\varphi_2})]^{-1} - I, \text{ with } c > 0,$$

where I is the identity map, and the notation $\text{prox}_{c\varphi}$ stands for the proximal point operator of Definition 2.1. Thanks to this reformulation, DRS convergence rates can be derived from those available for the PPA.

Classical approaches study the sum of two functions or the sum of two monotone operators, corresponding to convex functions in the optimization setting. More recently, extensions to the sum of $d \geqslant 2$ functions/monotone operators have been proposed, see [117, 118] for some examples. Our approach is applicable to the sum of two terms in $w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$. Accordingly, we consider the following minimization problem (cf. (2.7))

$$\min_{x \in \mathbb{R}^n} \varphi(x) = \varphi_1(x) + \varphi_2(x), \tag{4.1}$$

where $\varphi_1, \varphi_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are proper lsc functions, not necessarily convex.

Following [28], we examine relaxed DRS variants. Given a relaxation parameter $\lambda > 0$, a stepsize $\gamma > 0$, and an initial $s^0 \in \mathbb{R}^n$, define one iteration of relaxed DRS as below:

$$\begin{cases} u^k = \text{prox}_{\gamma\varphi_1}(s^k) \\ v^k \in \text{prox}_{\gamma\varphi_2}(2u^k - s^k) \\ s^{k+1} = s^k + \lambda(v^k - u^k). \end{cases} \tag{4.2}$$

Note that for $\lambda = 1$, scheme (4.2) reduces to DRS (cf. Algorithm 2), while for $\lambda = 2$ it corresponds to PRS. As stated, one iteration amounts to performing successively proximal steps, computed separately for each term in the sum, followed by a gradient step.

The iterative approach (4.2) has a long history. Lions and Mercier in [11] studied convergence properties and speed of convergence for splitting methods to find a zero of the sum of two maximally monotone operators defined on a Hilbert space. When applied to the optimization problem (5.1) for convex proper lower semicontinuous functions φ_1 and φ_2, the corresponding operators are the subdifferentials of convex analysis, $\partial\varphi_1$ and $\partial\varphi_2$. Under mild regularity assumptions, the DRS sequence $\{s^k\}$ converges to some $s^\star$, for which $u^\star = \text{prox}_{\gamma\varphi_1}(s^\star)$ solves (5.1), and both $\{u^k\}$ and $\{v^k\}$ converge to $u^\star$ [87, Theorem 3.15,Proposition 3.40]. Additionally, if φ_1 is differentiable and strongly convex, with Lipschitz continuous gradient, then $\{s^k\}$ $Q-$linearly converges to $s^\star$, and $\{u^k\}$ $Q-$linearly converges to the unique solution to (5.1). For varying stepsizes and inexact proximal evaluations, see [119, Theorem 7]. A similar analysis was carried out for the PRS method in

[87, 11]. More recently, the authors in [120] studied the convergence speed of a relaxed PRS for convex problems, under the assumption of strong convexity of one of the functions and Lipschitz continuity of its gradient. Rates of convergence are provided, including the standard DRS and PRS as special cases. Furthermore, the method applied to the dual formulation yields convergence rates for a relaxed ADMM. In [121] the authors derive a global linear rate of convergence for ADMM variants for the convex case, assuming one of the two functions is strongly convex with Lipschitz continuous gradient.

The aforementioned works are typically based on monotonicity of the sequence of iterate distances to the solution set. In a DRS, however, functional values are not monotonically increasing, and for this reason [122] introduced a special merit function, called *Douglas-Rachford envelope* (DRE). For convex composite problems with one convex quadratic function, the DRE is real-valued and continuously differentiable. Furthermore, one DRS iteration corresponds to one gradient step applied to minimizing the DRE. In a manner similar to how the Moreau envelope sheds a light on the PPA, the DRE gives an insight on DRS. In particular, because DRS provides (sufficient) descent for the merit function DRE, a variable metric gradient method for the DRE yields complexity estimates and rates of convergence for DRS iterates. A point crucial for this type of analysis is that DRE critical points are related to minimizers of the original convex problem. For convex composite objective functions with one $L-$smooth and strongly convex term, a similar approach is adopted in [123] to analyze Forward-Backward methods by means of a suitably defined envelope.

The literature is much more scarce for nonconvex problems, the setting considered in this work. We can mention the DR splitting proposed in [124, 125] for the sum of a differentiable function with Lipschitz continuous gradient, and a proper lower semicontinuous function with an easily computable proximal point. By defining a merit function related to the DRE, global subsequential convergence to a critical point is obtained, as well as eventual convergence rate under some extra assumptions, namely, that the functions satisfy the KL inequality [53, 54], a concept related to error bounds [75, Theorem 4.1]. These two notions are often used in the literature to establish local rates of convergence [22, 33, 19, 21, 126, 67, 75].

Our contribution refers to deriving local convergence rates for weakly convex Douglas-Rachford splitting mechanisms. This is achieved by combining the machinery developed in [28] for the DRE with the unifying framework for descent methods from Chapter 3. In some sense, we generalize the latter work, since when applying arguments of the corresponding theory to the DRE merit function, we succeed in showing convergence properties for another sequence of iterates, the DRS method applied to the original function. Our results resemble the ones briefly referred without proof in [28, page 15] for

semialgebraic functions.

The remainder of this chapter is organized as follows. We introduce the definition of the Douglas-Rachford envelope and some properties for convex and nonconvex functions in Section 4.2. Next, in Section 4.3 we show the necessary components to follow the ideas of Chapter 3 to obtain convergence and local rate of convergence of nonconvex Douglas-Rachford. We end this chapter with some preliminary numerical experiments of the DRS applied to a nonconvex consensus optimization problem.

4.2 Douglas-Rachford envelope

In this section, we examine the Douglas-Rachford envelope, a merit function that plays the role of the Moreau envelope (or Moreau-Yosida regularization) in the proximal point algorithm. We first discuss the properties of the Douglas-Rachford envelope in the nonconvex case, and then review the precursor setting, the convex case.

4.2.1 Douglas-Rachford envelope for nonconvex functions

We adopt the assumptions and notation in [28].

Assumption 4.1. *For problem (4.1), consider the following conditions:*

- $\varphi_1 : \mathbb{R}^n \to \mathbb{R}$ *is a* $L-$*smooth function, for (known)* $L > 0$*, that is,* φ_1 *is differentiable and its gradient* $\nabla\varphi_1$ *is a* $L-$*Lipschitz continuous function.*
- $\varphi_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is a proper lower semicontinuous function.*
- *the set of solutions of problem (5.1) is nonempty.*

As mentioned in [28, Remark 3.1], due to Assumption 4.1, the scheme in (4.2) is well-defined for any $0 < \gamma < \dfrac{1}{L}$, since for such choice of γ, both

$$\varphi_1(\cdot) + \frac{1}{2\gamma}\|\cdot\|^2 \text{ and } \varphi_2(\cdot) + \frac{1}{2\gamma}\|\cdot\|^2$$

are bounded from below. Furthermore, since φ_1 is $L-$smooth, $\operatorname{prox}_{\gamma\varphi_1}$ is a single-valued operator [28, Proposition 2.3(i)], and from the update rule for u^k in (4.2),

$$u^k = \operatorname{prox}_{\gamma\varphi_1}(s^k) \iff 0 = \gamma\nabla\varphi_1(u^k) + u^k - s^k. \tag{4.3}$$

Combining this last identity with the update rule for v^k in (4.2), v^k corresponds to a solution to the following problem

$$\min_v \left\{ \varphi_2(v) + \frac{1}{2\gamma}\|v - \left(u^k - \gamma\nabla\varphi_1(u^k)\right)\|^2 \right\}. \tag{4.4}$$

After expanding squares in the last expression, we end up with the original form of the *Douglas-Rachford envelope*, used in the analysis of [28] to show DRS convergence results.

Definition 4.1 (Douglas-Rachford envelope)**.** *For any $s \in \mathbb{R}^n$ the DRE function $\varphi_\gamma^{DR} : \mathbb{R}^n \to \mathbb{R}$ is defined as*

$$\varphi_\gamma^{DR}(s) = \min_v \left\{ \varphi_1(u) + \langle \nabla\varphi_1(u), v - u\rangle + \varphi_2(v) + \frac{1}{2\gamma}\|v - u\|^2 \right\}, \tag{4.5}$$

where $u = \operatorname{prox}_{\gamma\varphi_1}(s)$.

Problem (4.5) can be interpreted as yielding an approximate value of $\operatorname{prox}_{\gamma\varphi}(u)$, and thus yielding an approximation of the Moreau envelope $e_{\gamma\varphi}(u)$. Namely, in the sum $\varphi = \varphi_1 + \varphi_2$, the first term is replaced by a first-order Taylor model of φ_1 at u.

Thanks to the DRE, properties of the splitting (4.2) can be analyzed by resorting to techniques of descent methods. To this aim, we first recall some relations between the DRE and the scheme (4.2) stated in [28, Propositions 3.2, Theorem 3.4].

Remark 4.1. *Observe that the definition of the Douglas-Rachford envelope requires, in principle, the knowledge of the Lipschitz constant L. We refer to [28] for the analysis of an adaptive variant of the method examined in this chapter, that is, when the constant L is not assumed to be known.*

Proposition 4.1 (General properties of the envelope)**.** *For a function $\varphi = \varphi_1 + \varphi_2$ that satisfies Assumption 4.1, and φ_γ^{DR} defined in (4.5), for any $0 < \gamma < \frac{1}{L}$ the following holds.*

(i) The DRE satisfies the relation for all $s \in \mathbb{R}^n$

$$\varphi_\gamma^{DR}(s) = \big(\varphi_2^\gamma \circ (Id - \gamma\nabla\varphi_1) \circ \operatorname{prox}_{\gamma\varphi_1}\big)(s),$$

and is a real-valued and locally Lipschitz function.

(ii) $\inf_{x\in\mathbb{R}^n} \varphi(x) = \inf_{s\in\mathbb{R}^n} \varphi_\gamma^{DR}(s)$.

(iii) $\arg\min \varphi = \operatorname{prox}_{\gamma\varphi_1}(\arg\min \varphi_\gamma^{DR})$.

Proof. To prove item (i), in [28, Proposition 2.3] is shown that $\operatorname{prox}_{\gamma\varphi_1}$ is a Lipschitz continuous operator. Additionally, φ_2^γ is locally Lipschitz [29, Example 10.32], and so is $Id - \gamma\nabla\varphi_1$ due to Assumpption 4.1(i), and the result follows. Items (ii) and (iii) are [28, Theorem 3.4]. □

The Lipschitz continuity of the envelope is useful not only to be able to compute subgradients (by applying the chain rule to the expression in item (i) above), but also in the convergence analysis, since the DRE is thus a continuous function. Moreover, items (ii) and (iii) in Proposition 4.1 explicitly relate $\varphi_\gamma^{\mathrm{DR}}$ with the original objective function φ, through the proximal mapping of the first term φ_1. In particular, whenever φ is bounded below, so is the envelope $\varphi_\gamma^{\mathrm{DR}}$.

The next result, corresponding to [28, Proposition 3.3], relates the objective function φ with its regularization $\varphi_\gamma^{\mathrm{DR}}$, by using the definition of the DRE, and $L-$smoothness of φ_1. These relations are crucial to prove convergence of function values in Theorem 4.3. We include a proof for the sake of clarity.

Proposition 4.2 (Relations between DRS and DRE iterates)**.** *For a function $\varphi = \varphi_1 + \varphi_2$ that satisfies Assumption 4.1, φ_γ^{DR} defined in (4.5), and $\{(u^k, v^k, s^k)\}$ generated by (4.2), for any $0 < \gamma < \dfrac{1}{L}$ it holds*

(i) $\varphi_\gamma^{DR}(s^k) \leqslant \varphi(u^k)$.

(ii) $\varphi(v^k) \leqslant \varphi_\gamma^{DR}(s^k) - \dfrac{1-\gamma L}{2\gamma}\|u^k - v^k\|^2$.

Furthermore, any limit point $(s^\star, u^\star, v^\star)$ of the sequence $\{(u^k, v^k, s^k)\}$, whenever they exist, satisfy

(iii) $\varphi_\gamma^{DR}(s^\star) \leqslant \varphi(u^\star)$, *and* $\varphi(v^\star) \leqslant \varphi_\gamma^{DR}(s^\star) - \dfrac{1-\gamma L}{2\gamma}\|u^\star - v^\star\|^2$.

Proof. Item (i) directly follows from (4.5), by evaluating $\varphi_\gamma^{\mathrm{DR}}$ at $s = s^k$, and the minimand at $v = u^k$. Item (ii) follows similarly as [127, Proposition 4.3(ii)]. Indeed, since v^k minimizes the problem in (4.5) for $s = s^k$,

$$\varphi_\gamma^{\mathrm{DR}}(s^k) = \varphi_1(u^k) + \langle \nabla\varphi_1(u^k), v^k - u^k\rangle + \varphi_2(v^k) + \frac{1}{2\gamma}\|v^k - u^k\|^2.$$

The right-hand side can be bounded using the descent lemma [128, Proposition A.24], namely,

$$|\varphi_1(v^k) - \varphi_1(u^k) - \langle \nabla\varphi_1(u^k), v^k - u^k\rangle| \leqslant \frac{L}{2}\|u^k - v^k\|^2. \tag{4.6}$$

This yields $\varphi_\gamma^{\mathrm{DR}}(s^k) \geqslant \varphi_1(v^k) - \dfrac{L}{2}\|v^k - u^k\|^2 + \varphi_2(v^k) + \dfrac{1}{2\gamma}\|v^k - u^k\|^2$, giving the desired result.

Let $\{(s^{k_j}, u^{k_j}, v^{k_j})\}_j$ be a subsequence such that $(s^{k_j}, u^{k_j}, v^{k_j}) \to (s^\star, u^\star, v^\star)$ as $j \to +\infty$. Since $\nabla\varphi_1$ is continuous, then taking the limit in (4.3), it holds that $0 =$

$\gamma\nabla\varphi_1(u^\star) + u^\star - s^\star$, which is equivalent to $u^\star = \operatorname{prox}_{\gamma\varphi_1}(s^\star)$. Therefore, it follows from (4.5) that

$$\varphi_\gamma^{\mathrm{DR}}(s^\star) \leqslant \varphi_1(u^\star) + \langle \nabla\varphi(u^\star), u^\star - u^\star\rangle + \varphi_2(u^\star) + \frac{1}{2\gamma}\|u^\star - u^\star\|^2 = \varphi(u^\star).$$

Furthermore, since φ is lower semicontinuous, and $\varphi_\gamma^{\mathrm{DR}}$ is continuous (Proposition 4.1(i)), it follows from Proposition 4.2(ii)

$$\begin{aligned}
\varphi(v^\star) &\leqslant \liminf_{j\to+\infty} \varphi(v^{k_j}) \\
&\leqslant \liminf_{j\to+\infty} \left\{ \varphi_\gamma^{\mathrm{DR}}(s^{k_j}) - \frac{1-\gamma L}{2\gamma}\|u^{k_j} - v^{k_j}\|^2 \right\} \\
&= \varphi_\gamma^{\mathrm{DR}}(s^\star) - \frac{1-\gamma L}{2\gamma}\|u^\star - v^\star\|^2.
\end{aligned}$$

□

4.2.2 Gradient method applied to the Douglas-Rachford envelope: convex case

For the smooth convex case, the authors in [122] provide an alternative analysis of the Douglas-Rachford splitting method as a variable-metric gradient method applied to the DRE. More specifically, we momentarily assume φ_1 is twice continuously differentiable strongly convex with $L-$Lipschitz continuous gradient, and φ_2 is convex. Under this hypothesis,

$$\|\nabla^2\varphi_1(u)\|_2 \leqslant L,$$

where ∇^2 denotes the Hessian matrix, and $\|\cdot\|_2$ is the operator norm induced by the ℓ^2-norm in $\mathbb{R}^n$. Furthermore, for $\gamma \in (0, 1/L)$, the Hessian of $e_\gamma\varphi_1$ exists everywhere and

$$\nabla^2(e_\gamma\varphi_1)(s) = \gamma^{-1}\Big(I - \big(I + \gamma\nabla^2\varphi_1(\operatorname{prox}_{\gamma\varphi_1}(s))\big)^{-1}\Big).$$

In this manner, the DRE is differentiable and

$$\nabla\varphi_\gamma^{\mathrm{DR}}(s) = \big(I - 2\gamma\nabla^2(e_\gamma\varphi_1)(s)\big)\big(\nabla(e_\gamma\varphi_1)(s) + (e_\gamma\varphi_2)(s) - 2\gamma\nabla(e_\gamma\varphi_1)(s)\big),$$

where, as stated in (2.3),

$$\nabla(e_\gamma\varphi_1)(s) = \gamma^{-1}\big(s - \operatorname{prox}_{\gamma\varphi_1}(s)\big).$$

Starting from s^0, one iteration of DRS (4.2) for $\lambda = 1$ is equivalent to one iteration of the following variable-metric gradient step on $\varphi_\gamma^{\mathrm{DR}}$

$$s^{k+1} = s^k - D^k\nabla\varphi_\gamma^{\mathrm{DR}}(s^k), \tag{4.7}$$

where

$$D^k = \gamma\big(2[I + \gamma\nabla^2\varphi_1\big(\operatorname{prox}_{\gamma\varphi_1}(s^k)\big)]^{-1} - I\big)^{-1}.$$

The key observation in the current convex setting is that $u^\star$ is a solution to (4.1) if and only if $u^\star = \operatorname{prox}_{\gamma\varphi_1}(s^\star)$ for some $s^\star$ such that

$$\operatorname{prox}_{\gamma\varphi_2}\big(2\operatorname{prox}_{\gamma\varphi_1}(s^\star) - s^\star\big) - \operatorname{prox}_{\gamma\varphi_1}(s^\star) = 0. \tag{4.8}$$

When φ_1 is additionally convex quadratic, then for any $\gamma \in (0, 1/L)$, $\varphi_\gamma^{\mathrm{DR}}$ is strongly convex with $L-$Lipschitz continuous gradient. Hence, from the classical convergence results of the gradient descent method, the sequence $\{u^k\}$ defined by the $u-$step of (4.2) with $\lambda = 1$, and $\{s^k\}$ generated by (4.7), converges to some minimizer $u^\star$ of φ, and $\{\varphi(v^k)\}$ converges to min φ. The convergence properties are deduced from the following estimate:

$$\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s^\star) \leqslant \frac{1}{(2\gamma\lambda)k}\|s^0 - s^\star\|^2.$$

In view of Proposition 4.2(ii)–(iii), it thus holds

$$\varphi(v^k) - \min\ \varphi \leqslant \frac{1}{(2\gamma\lambda)k}\|s^0 - s^\star\|^2$$

Under the more general nonconvex Assumption 4.1, it is still possible to obtain descent for $\varphi_\gamma^{\mathrm{DR}}$ along the DRS iterations, resembling gradient descent, but in a nondifferentiable setting.

4.3 Convergence of nonconvex Douglas-Rachford splitting

In order to obtain convergence properties of DRS using arguments for descent methods under Assumption 4.1, we make use of the DRE. The work in [28] constructs the tools to employ the analysis in Chapter 3.

4.3.1 Convergence analysis as a descent method for the Douglas-Rachford envelope

For the purpose of using the properties of φ_1 and φ_2 of Assumption 4.1, note that $\varphi_\gamma^{\mathrm{DR}}$ can be computed directly using the iterates of (4.2), by first reformulating problem (5.1) as follows

$$\min_{u,v\in\mathbb{R}^n} \varphi_1(u) + \varphi_2(v) \quad \text{s.t.} \quad u - v = 0.$$

The augmented Lagrangian of this reformulation is, for $\beta > 0$:

$$\mathcal{L}_\beta(u, v, y) = \varphi_1(u) + \varphi_2(v) + \langle y, u - v\rangle + \frac{\beta}{2}\|u - v\|^2,$$

where $y \in \mathbb{R}^n$ is a Lagrange multiplier associated with the constraint $u - v = 0$. Therefore, due to (4.5), it holds

$$\varphi_\gamma^{\mathrm{DR}}(s^k) = \mathcal{L}_{\gamma^{-1}}\big(u^k, v^k, \gamma^{-1}(u^k - s^k)\big). \tag{4.9}$$

Following Chapter 3, the first main ingredient is to prove that DRS is a descent method for the Lagrangian evaluated at the primal-dual point

$$\sigma^k := \big(u^k, v^k, \gamma^{-1}(u^k - s^k)\big).$$

The following result states that $\{\varphi_\gamma^{\mathrm{DR}}(s^k)\}$ satisfies a condition of sufficient decrease with respect to both $\|s^k - s^{k+1}\|^2$ and $\|u^k - u^{k+1}\|^2$, proven in [28, Theorem 4.1] for $\{\varphi_\gamma^{\mathrm{DR}}(s^k)\}$.

Theorem 4.2 (Descent properties of DRS)**.** *Suppose that $\varphi = \varphi_1 + \varphi_2$ satisfies Assumption 4.1. For $\lambda \in (0, 2)$, and $0 < \gamma < \dfrac{2-\lambda}{2L}$, the iterates $\{(u^k, v^k, s^k)\}$ generated by (4.2) satisfy,*

$$\varphi_\gamma^{DR}(s^k) - \varphi_\gamma^{DR}(s^{k+1}) \geqslant c \max\left\{\frac{1}{(1+\gamma L)^2}\|s^k - s^{k+1}\|^2, \|u^k - u^{k+1}\|^2\right\}, \tag{4.10}$$

where

$$c = \frac{2-\lambda}{2\lambda\gamma} - \frac{L}{\lambda} > 0.$$

Furthermore, for $\sigma^k = \big(u^k, v^k, \gamma^{-1}(u^k - s^k)\big)$, it holds

$$g^k := \big(\gamma^{-1}(u^k - v^k), 0, u^k - v^k\big) \in \partial\mathcal{L}_{\gamma^{-1}}(\sigma^k), \tag{4.11}$$

and

$$\|g^k\| = \lambda^{-1}\sqrt{\gamma^{-2} + 1}\|s^{k+1} - s^k\|. \tag{4.12}$$

Proof. First, the estimate (4.10) for $\|s^k - s^{k+1}\|^2$ corresponds to [28, (4.2)] after using the identity (4.9), for $\sigma_{\varphi_1} = -L$. The estimate for $\|u^k - u^{k+1}\|^2$ appears in the proof of [28, Theorem 4.1].

Furthermore, the subdifferential of $\mathcal{L}_{\gamma^{-1}}$ at $\sigma = \sigma^k$ can be computed taking partial derivatives with respect to the different components of the primal-dual vector σ, as follows:

- Since φ_1 is differentiable, $\dfrac{\partial\mathcal{L}_{\gamma^{-1}}}{\partial u}(u, v, y) = \nabla\varphi_1(u) + y + \gamma^{-1}(u - v)$. Then, evaluating this last identity at $(u, v, y) = \sigma^k$, and using (4.3), it follows that

$$\begin{aligned}\frac{\partial\mathcal{L}_{\gamma^{-1}}}{\partial u}(\sigma^k) &= \nabla\varphi_1(u^k) + \gamma^{-1}(u^k - s^k) + \gamma^{-1}(u^k - v^k)\\ &= \gamma^{-1}(u^k - v^k).\end{aligned}$$

- From the optimality condition of v^k for problem (4.5),

$$0 \in \partial\varphi_2(v^k) + \nabla\varphi_1(u^k) + \gamma^{-1}(v^k - u^k),$$

it holds that

$$\frac{\partial \mathcal{L}_{\gamma^{-1}}}{\partial v}(\sigma^k) = \partial\varphi_2(v^k) - \gamma^{-1}(u^k - s^k) + \gamma^{-1}(v^k - u^k) \ni 0.$$

- Since $\mathcal{L}_{\gamma^{-1}}$ only depends on y linearly, then $\dfrac{\partial \mathcal{L}_{\gamma^{-1}}}{\partial y}(\sigma^k) = \gamma^{-1}(u^k - s^k)$.

Therefore, from $\partial\mathcal{L}_{\gamma^{-1}}(\sigma^k) = \dfrac{\partial \mathcal{L}_{\gamma^{-1}}}{\partial u}(\sigma^k) \times \dfrac{\partial \mathcal{L}_{\gamma^{-1}}}{\partial v}(\sigma^k) \times \dfrac{\partial \mathcal{L}_{\gamma^{-1}}}{\partial y}(\sigma^k)$, identity (4.11) follows. To prove (4.12), note that due to the update rule for $\{s^k\}$ in (4.2), it follows

$$\begin{aligned} \|g^k\|^2 &= \gamma^{-2}\|u^k - v^k\|^2 + \|u^k - v^k\|^2 \\ &= (\gamma^{-2} + 1)\|u^k - v^k\|^2 \\ &= (\gamma^{-2} + 1)\lambda^{-2}\|s^k - s^{k+1}\|^2. \end{aligned}$$

□

Our result is not a straightforward application of the general scheme in Chapter 3. As made clear in the proof, in our setting, the DRE functional decrease is measured only in terms of some components of the primal-dual term $\|\sigma^k - \sigma^{k+1}\|^2$. This feature prevents us to directly apply the unifying convergence theory of Chapter 3.

We now give an alternative proof of [28, Theorem 4.3], based on the developments in Chapter 3. The result states subsequential convergence of the iterates to critical points of φ, and convergence of $\varphi_\gamma^{\mathrm{DR}}(s^k)$ to a critical value of φ.

Theorem 4.3 (Subsequential convergence of DRS)**.** *Suppose that $\varphi = \varphi_1 + \varphi_2$ satisfies Assumption 4.1. For $\lambda \in (0, 2)$, and $0 < \gamma < \dfrac{2-\lambda}{2L}$, then for any bounded sequence $\{(u^k, v^k, s^k)\}$ generated by (4.2),*

(i) The sequence $\{\varphi_\gamma^{DR}(s^k)\}$ monotonically converges to a critical value $\varphi^\star$ of φ, and the sequence $\{\varphi_1(u^k) + \varphi_2(v^k)\}$ converges to the same value $\varphi^\star$.

(ii) $u^k - v^k \to 0$, $u^k - u^{k+1} \to 0$, $v^k - v^{k+1} \to 0$, and $s^k - s^{k+1} \to 0$, as $k \to +\infty$.

(iii) All cluster points of $\{u^k\}$ and $\{v^k\}$ coincide, and are also critical points of φ, with same critical value $\varphi^\star = \lim\limits_{k\to\infty} \varphi_\gamma^{DR}(s^k) = \lim\limits_{k\to\infty} \varphi_1(u^k) + \varphi_2(v^k)$.

Proof. First, Assumption 4.1(iii) implies that φ is bounded from below. Then, from Proposition 4.1(ii), $\varphi_\gamma^{\mathrm{DR}}$ is also bounded from below, and so is the sequence $\{\varphi_\gamma^{\mathrm{DR}}(s^k)\}$. Furthermore, the descent condition (4.10) implies that $\{\varphi_\gamma^{\mathrm{DR}}(s^k)\}$ is a non-increasing real sequence. Thus, there exists $\varphi^\star \in \mathbb{R}$ such that $\varphi_\gamma^{\mathrm{DR}}(s^k) \to \varphi^\star$. In turn, (4.10) then yields $s^k - s^{k+1} \to 0$, and $u^k - u^{k+1} \to 0$. These results have a couple of consequences:

- $u^k - v^k \to 0$, due to the update rule for $\{s^k\}$ in (4.2), and thus $\{u^k\}$ and $\{v^k\}$ have the same limit points.
- From (4.12), it follows that $g^k \to 0$.
- Furthermore, $v^k - v^{k+1} = v^k - u^k + u^k - u^{k+1} + u^{k+1} - v^{k+1} \to 0$.

As for item (ii), let $u^\star, v^\star$, and $s^\star$ be limit points of the sequences $\{u^k\}, \{v^k\}$, and $\{s^k\}$, respectively. Note that $v^\star = u^\star$, because $\{u^k\}$ and $\{v^k\}$ have the same limit points, and thus $\varphi(v^\star) = \varphi(u^\star)$. Then, up to a subsequence, $v^k \to v^\star$, and following the arguments in [28]:

$$\begin{aligned}
\varphi(v^\star) &\leqslant \liminf_{k\in K} \varphi(v^k) && \text{Assumption 4.1(i)-(ii)}\\
&\leqslant \limsup_{k\in K} \varphi(v^k)\\
&\leqslant \limsup_{k\in K} \varphi_\gamma^{\mathrm{DR}}(s^k) && \text{Proposition 4.2(ii)}\\
&= \varphi_\gamma^{\mathrm{DR}}(s^\star) && \text{Proposition 4.1(i)}\\
&\leqslant \varphi(u^\star) && \text{Proposition 4.2(iii)}\\
&= \varphi(v^\star)
\end{aligned}$$

Therefore, $\varphi(v^k) \to \varphi(v^\star)$, and $\mathcal{L}_{\gamma^{-1}}(\sigma^k) = \varphi_\gamma^{\mathrm{DR}}(s^k) \to \varphi(v^\star)$ through the same subsequence, with $\varphi(u^\star) = \varphi(v^\star) = \varphi_\gamma^{\mathrm{DR}}(s^\star) = \varphi^\star$. Note that since $\{u^k - s^k\}$ is bounded, and $u^k - v^k \to 0$ as $k \to +\infty$, then $\langle \gamma^{-1}(u^k - s^k), u^k - v^k\rangle + \dfrac{1}{2\gamma}\|u^k - v^k\|^2 \to 0$, and thus

$$\varphi_1(u^k) + \varphi_2(v^k) = \mathcal{L}_{\gamma^{-1}}(\sigma^k) - \left(\langle \gamma^{-1}(u^k - s^k), u^k - v^k\rangle + \frac{1}{2\gamma}\|u^k - v^k\|^2\right) \to \varphi^\star.$$

Furthermore, from the definition of the augmented Lagrangian,

$$\begin{aligned}
\mathcal{L}_{\gamma^{-1}}(u^\star, u^\star, \gamma^{-1}(u^\star - s^\star)) &= \varphi_1(u^\star) + \varphi_2(u^\star) + \langle \gamma^{-1}(u^\star - s^\star), u^\star - u^\star\rangle + \frac{1}{2\gamma}\|u^\star - u^\star\|^2\\
&= \varphi(u^\star),
\end{aligned}$$

that is, $\{\mathcal{L}_{\gamma^{-1}}(\sigma^k)\}$ (subsequentially) converges to $\mathcal{L}_{\gamma^{-1}}(u^\star, u^\star, \gamma^{-1}(u^\star - s^\star))$, as $\sigma^k \to (u^\star, u^\star, \gamma^{-1}(u^\star - s^\star)$. Therefore, taking the limit in (4.11) (passing through a subsequence if necessary) gives

$$0 \in \partial\mathcal{L}_{\gamma^{-1}}(u^\star, v^\star, \gamma^{-1}(u^\star - s^\star)),$$

which is equivalent to the following criticality conditions

$$\begin{cases} 0 = \nabla\varphi_1(u^\star) + \gamma^{-1}(u^\star - s^\star) + \gamma^{-1}(u^\star - v^\star) \\ 0 \in \partial\varphi_2(v^\star) - \gamma^{-1}(u^\star - s^\star) + \gamma^{-1}(u^\star - v^\star) \\ 0 = v^\star - u^\star \end{cases}$$

Adding the first two relations, and using that $v^\star = u^\star$, the final result follows, as we obtain $0 \in \nabla\varphi_1(u^\star) + \partial\varphi_2(u^\star)$.

□

Remark 4.2. *Some comments about Theorem 4.3 are in order.*

- *By items (i) and (ii), the sequence of DRE functional values $\varphi_\gamma^{DR}(s^k)$ converges monotonically to a critical value of φ. By contrast, the sequence of functional values $\varphi_1(u^k) + \varphi_2(v^k)$ converges to the same critical value, but not necessarily in a monotone manner.*
- *Boundedness of the iterates $\{(u^k, v^k, s^k)\}$ generated by (4.2) is usually ensured by assuming φ has bounded level sets [28, Theorem 4.3(iii)], which is equivalent to φ_γ^{DR} having the same property [28, Theorem 3.4(iii)].*

4.3.2 Rate of convergence for nonconvex Douglas-Rachford splitting

The analysis of rate of convergence requires additional assumptions. This section is an extension of [28], by applying the machinery of Chapter 3 to (4.2) through the envelope $\varphi_\gamma^{\mathrm{DR}}$ and the augmented Lagrangian $\mathcal{L}_{\gamma^{-1}}$.

We recall two concepts related to criticality and the study of rates of convergence in the following, namely Definition 1.7 and Definition 1.1. We say a function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ satisfy a local error bound, if for any $\bar{\varphi} \geqslant \inf\varphi > -\infty$, there exists $\varepsilon > 0, \ell > 0$, such that whenever $\varphi(v) \leqslant \bar{\varphi}$,

$$\operatorname{dist}\big(x, (\partial\varphi)^{-1}(0)\big) \leqslant \ell \operatorname{dist}\big(0, \partial\varphi(v) \cap B(0,\varepsilon)\big). \tag{4.13}$$

Furthermore, we say a function $\varphi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ satisfy the proper separation of isocost surfaces property if there exists $\delta > 0$, such that if

$$\forall u, v \in (\partial\varphi)^{-1}(0),\ \|u - v\| \leqslant \delta \implies \varphi(u) = \varphi(v). \tag{4.14}$$

The next result relates the sequences generated by (4.2) with the error bound condition (4.13), resulting in an estimate crucial to obtain a local rate of convergence.

Proposition 4.3. *Suppose that $\varphi = \varphi_1 + \varphi_2$ satisfies Assumption 4.1, as well as a local error bound (4.13). For $\lambda \in (0,2)$, and $0 < \gamma < \dfrac{2-\lambda}{2L}$, then for any bounded sequence $\{(u^k, v^k, s^k)\}$ generated by (4.2), and any $\varepsilon, \ell > 0$, there exists $\bar{\varphi} > 0$, and a sequence $\{\tilde{g}^k\}$, such that $\tilde{g}^k \in \partial\varphi(v^k) \cap B(0,\varepsilon)$, $\varphi(v^k) \leqslant \bar{\varphi}$, and*

$$\operatorname{dist}\big(v^k, (\partial\varphi)^{-1}(0)\big) \leqslant \ell\|\tilde{g}^k\|. \tag{4.15}$$

Proof. First, since $\{\varphi_\gamma^{\mathrm{DR}}(s^k)\}$ monotonically converges to $\varphi^\star$, then for any $\epsilon > 0$, and for any sufficiently large k, $\mathcal{L}_{\gamma^{-1}}(\sigma^k) = \varphi_\gamma^{\mathrm{DR}}(s^k) \leqslant \varphi^\star + \epsilon$. From the definition of the augmented Lagrangian, we then have

$$\varphi_1(u^k) + \varphi_2(v^k) + \gamma^{-1}\langle u^k - s^k, u^k - v^k\rangle + \frac{1}{2\gamma}\|u^k - v^k\|^2 \leqslant \varphi^\star + \epsilon. \tag{4.16}$$

Furthermore, from the bounded assumption of the generated sequences, $\{\nabla\varphi_1(u_k)\}$ is bounded, that is, there exists $M_1 > 0$, such that for all k, $\|\nabla\varphi_1(u^k)\| \leqslant M_1$. From (4.6) and Theorem 4.3(ii), for any $\eta > 0$, and for all sufficiently large k,

$$|\varphi_1(v^k) - \varphi_1(u^k)| \leqslant M_1\eta + \frac{L}{2}\eta^2.$$

Therefore, $\varphi(v^k) \leqslant \varphi_1(u^k) + \varphi_2(v^k) + M_1\eta + \dfrac{L}{2}\eta^2$. Combining this last inequality with (4.16), it holds

$$\begin{aligned}\varphi(v^k) &\leqslant \varphi^\star + \epsilon - \gamma^{-1}\langle u^k - s^k, u^k - v^k\rangle - \frac{1}{2\gamma}\|u^k - v^k\|^2 + M_1\eta + \frac{L}{2}\eta^2\\ &\leqslant \varphi^\star + \epsilon + \gamma^{-1}\|u^k - s^k\|\|u^k - v^k\| + M_1\eta + \frac{L}{2}\eta^2\end{aligned}$$

Since the generated sequences are bounded, then there exists $M_2 > 0$ such that $\|u^k - s^k\| \leqslant M_2$. Therefore, for all sufficiently large k,

$$\varphi(v^k) \leqslant \varphi^\star + \epsilon + \gamma^{-1}M_2\eta + M_1\eta + \frac{L}{2}\eta^2 =: \bar{\varphi}.$$

Furthermore, from the optimality conditions of (4.4) (as shown in the proof of [28, Theorem 4.3]), it follows that $\tilde{g}^k := \gamma^{-1}(u^k - v^k) - (\nabla\varphi_1(u^k) - \nabla\varphi_1(v^k)) \in \partial\varphi(v^k)$. Since $\nabla\varphi_1$ is $L-$Lipschitz continuous, then

$$\|\tilde{g}^k\| \leqslant \gamma^{-1}(\|u^k - v^k\| + \gamma L\|u^k - v^k\|) = \gamma^{-1}(1 + \gamma L)\|u^k - v^k\|. \tag{4.17}$$

In this manner, from Theorem 4.3(ii), $\tilde{g}^k \to 0$. Then, for any sufficiently large k, $\tilde{g}^k \in \partial\varphi(v^k) \cap B(0,\varepsilon)$, and $\varphi(v^k) \leqslant \bar{\varphi}$. This allow us to apply (4.13) to obtain (4.15). □

Before giving the most important result, first we need some technical estimates deduced from Proposition 4.3.

Lemma 4.1. *Suppose the conditions of Proposition 4.3 hold. For any* $p_v^k \in proj_{(\partial\varphi)^{-1}(0)}(v^k)$, *define*

$$p^k = (p_v^k, p_v^k, \gamma^{-1}(p_v^k - s^k)).$$

Then, there exists $\overline{C} > 0$, *such that for all* k,

$$\|p^k - \sigma^k\|^2 \leqslant \overline{C}(\varphi_\gamma^{DR}(s^k) - \varphi_\gamma^{DR}(s^{k+1})). \tag{4.18}$$

Proof. From the definition of p^k and σ^k, we have

$$\begin{aligned}\|p^k - \sigma^k\|^2 &= \|p_v^k - u^k\|^2 + \|p_v^k - v^k\|^2 + \|\gamma^{-1}(p_v^k - s^k) - \gamma^{-1}(u^k - s^k)\|^2 \\ &= (1+\gamma^{-2})\|p_v^k - u^k\|^2 + \|p_v^k - v^k\|^2.\end{aligned}$$

From (4.15), it follows that $\|p_v^k - v^k\|^2 \leqslant \ell\|\tilde{g}^k\|$. As for $\|p_v^k - u^k\|$, it holds that

$$\begin{aligned}\|p_v^k - u^k\|^2 &\leqslant (\|p_v^k - v^k\| + \|v^k - u^k\|)^2 \\ &\leqslant 2\|p_v^k - v^k\|^2 + 2\|v^k - u^k\|^2 \\ &\leqslant 2\ell^2\|\tilde{g}^k\|^2 + 2\|v^k - u^k\|^2 \\ &= 2\ell^2\|\tilde{g}^k\|^2 + 2\lambda^{-2}\|s^k - s^{k+1}\|^2\end{aligned}$$

where for the first inequality we apply the triangle inequality, for the second inequality we use the estimate $(a+b)^2 \leqslant 2(a^2+b^2)$, for the third inequality $\|p_v^k - v^k\|^2 \leqslant \ell\|\tilde{g}^k\|$ is used, and for the last equality we use the update rule for $\{s^k\}$ in (4.2). Therefore,

$$\|p^k - \sigma^k\|^2 \leqslant 2(1+\gamma^{-2})(\ell^2\|\tilde{g}^k\|^2 + \lambda^{-2}\|s^k - s^{k+1}\|^2) + \ell^2\|\tilde{g}^k\|^2.$$

Now, we bound the terms in the right-hand side of the above estimate.

- First, note that the descent condition (4.10) implies

$$\|s^k - s^{k+1}\|^2 \leqslant \frac{(1+\gamma L)^2}{c}(\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s^{k+1})). \tag{4.19}$$

- Furthermore, (4.17) together with (4.2) and (4.19) imply

$$\begin{aligned}\|\tilde{g}^k\|^2 &= \gamma^{-2}(1+\gamma L)^2\|u^k - v^k\|^2 \\ &= (\lambda\gamma)^{-2}(1+\gamma L)^2\|s^k - s^{k+1}\|^2 \\ &\leqslant (\lambda\gamma)^{-2}(1+\gamma L)^2\frac{(1+\gamma L)^2}{c}(\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s^{k+1}))\end{aligned}$$

Hence, follows (4.18) for

$$\overline{C} = \frac{(1+\gamma L)^2}{c\lambda^2}\left(\frac{\big(2(1+\gamma^{-2})+1\big)(1+\gamma L)^2\ell^2}{\gamma^2} + 2(1+\gamma^{-2})\right).$$

□

In order to relate an error bound for φ with the descent properties of Theorem 4.2, we need the following assumption for φ_2.

Assumption 4.4. *Assume that* $\varphi_2 \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$.

The use of weakly convex regularizers can present an advantage by reducing the bias introduced by, for example, the ℓ^1-norm, meaning that nonzero entries of the computed solution are not underestimated, as pointed out in [129]. On the other hand, allowing weakly convex objective functions broadens the applicability of splitting methods with convergence guarantees.

Under Assumption 4.4, both $\varphi_1,\ \varphi_2 \in w - \overline{\text{conv}}_\rho(\mathbb{R}^n)$ for $\rho \geqslant L$, due to Proposition 1.7. Hence, φ is locally Lipschitz on its domain and weakly convex. Therefore, as mentioned before, limiting (or Clarke) subgradients of φ can be characterized as proximal subgradients in the whole space, and $\hat{\partial}\varphi = \partial\varphi = \overline{\partial}\varphi$. This particular form of subgradients allows us to take full advantage of Lemma 4.1.

Theorem 4.5 (Rate of convergence of nonconvex DRS)**.** *Suppose that* $\varphi = \varphi_1 + \varphi_2$ *satisfies Assumptions 4.1 and 4.4, a local error bound* (4.13)*, and property* (4.14)*. Then, for* $\lambda \in (0,2)$, $0 < \gamma < \dfrac{2-\lambda}{2L}$, *and any bounded sequence* $\{(u^k, v^k, s^k)\}$ *generated by* (4.2)*,*

(i) The sequence $\{\varphi_\gamma^{DR}(s^k)\}$ *Q−linearly converges to a critical value* $\varphi^\star$ *of* φ*, and the sequence* $\{\varphi_1(u^k) + \varphi_2(v^k)\}$ *R−linearly converges to the same value* $\varphi^\star$*.*

(ii) The sequences $\{u^k\}$ *and* $\{v^k\}$ *R−linearly converge to a critical point* $u^\star$ *of* φ*, and* $\{s^k\}$ *R−linearly converges to a point* $s^\star$*, such that* $u^\star = \text{prox}_{\gamma\varphi_1}(s^\star)$*.*

Proof. First, from Proposition 4.3, $\tilde{g}^k \to 0$ implies $v^k - p_v^k \to 0$, which in turn implies $p_v^k - p_v^{k+1} \to 0$, in view of Theorem 4.3(ii). Then, applying the proper separation of isocost surfaces property (4.14), for all sufficiently large k, $\varphi(p_v^k) = \varphi(p_v^{k+1})$, and thus $\varphi(p_v^k) = \varphi^\star$, for some critical value $\varphi^\star$ of φ. From Theorem 4.3(iii), up to a subsequence, $v^k \to u^\star$, for a critical point $u^\star$ of φ. Therefore, $p_v^k \to u^\star$ and $\varphi(p_v^k) \to \varphi(u^\star)$, for the same subsequence. Thus $\varphi^\star = \varphi(u^\star)$.

Furthermore, from the definition of the augmented Lagrangian, $\mathcal{L}_{\gamma^{-1}}(p^k) = \varphi(p_v^k)$. Moreover, in view of (4.11)

$$\begin{aligned} \varphi_\gamma^{\text{DR}}(s^k) - \varphi_\gamma^{\text{DR}}(p_v^k) &= \mathcal{L}_{\gamma^{-1}}(\sigma^k) - \mathcal{L}_{\gamma^{-1}}(p^k) \\ &\leqslant -\langle g^k, p^k - \sigma^k\rangle + \frac{\rho}{2}\|p^k - \sigma^k\|^2. \end{aligned}$$

In particular, for all sufficiently large k,

$$\varphi_\gamma^{\text{DR}}(s^k) - \varphi^\star \leqslant \|g^k\|\|p^k - \sigma^k\| + \frac{\rho}{2}\|p^k - \sigma^k\|^2. \tag{4.20}$$

Note that from (4.12) and (4.19),

$$\begin{aligned}\|g^k\|^2 &= \lambda^{-2}(1+\gamma^{-2})\|s^k - s^{k+1}\|^2 \\ &\leqslant \lambda^{-2}(1+\gamma^{-2})\frac{(1+\gamma L)^2}{c}(\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s^{k+1}))\end{aligned}$$

Combining the last estimate with (4.20) and (4.18), yields

$$\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi^\star \leqslant \left(\tilde{C} + \overline{C}\frac{\rho}{2}\right)(\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s^{k+1})) \tag{4.21}$$

for $\tilde{C} = \lambda^{-1}\sqrt{1+\gamma^{-2}}(1+\gamma L)\sqrt{\dfrac{\overline{C}}{c}}$. Set $\hat{C} := \tilde{C} + \overline{C}\dfrac{\rho}{2}$, $r = \dfrac{\hat{C}}{1+\hat{C}} \in (0,1)$, and $V^k := \varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi^\star$. Note that monotonicity of $\{\varphi_\gamma^{\mathrm{DR}}(s^k)\}$ implies $V^{k+1} \leqslant V^k$. Thus, from (4.21), for all sufficiently large k,

$$V^{k+1} \leqslant \hat{C}(V^k - V^{k+1}) \iff V^{k+1} \leqslant rV^k,$$

from which the first part of item (i) follows.

For item (ii), suppose that above estimate holds for all $k \geqslant k_0$. Then,

$$V^{k+1} \leqslant (V^{k_0} r^{1-k_0}) r^k,$$

or equivalently, for $q = V^{k_0} r^{-k_0}$ and all $k \geqslant k_0 + 1$

$$V^k \leqslant q r^k. \tag{4.22}$$

From the descent condition of Equation (4.10), it follows

$$\|s^k - s^{k+1}\| \leqslant \frac{1+\gamma L}{\sqrt{c}}\sqrt{V^k}, \quad \text{and} \quad \|u^k - u^{k+1}\| \leqslant \frac{1}{\sqrt{c}}\sqrt{V^k}. \tag{4.23}$$

Therefore, from Lemma 3.1, there exists $m > 0$, $\alpha \in (0,1)$, $s^\star \in \mathbb{R}^n$ such that for all sufficiently large k

$$\|s^k - s^\star\| \leqslant m\alpha^k, \quad \|u^k - u^\star\| \leqslant m\alpha^k.$$

Note that $\{u^k\}$ converges to the critical point $u^\star$, since $\{u^k\}$ and $\{v^k\}$ have the same limit points. Observe that since φ_1 is $L-$smooth, then $\operatorname{prox}_{\gamma\varphi_1}$ is Lipschitz continuous [28, Proposition 2.3(ii)], therefore $u^k = \operatorname{prox}_{\gamma\varphi_1}(s^k) \to \operatorname{prox}_{\gamma\varphi_1}(s^\star)$, and thus $u^\star = \operatorname{prox}_{\gamma\varphi_1}(s^\star)$.

In addition, the rate of convergence of $\{v^k\}$ can be deduced as follows. Using the triangle inequality, the update rule for $\{s^k\}$ in (4.2), and (4.23), it holds

$$\begin{aligned}\|v^k - v^{k+1}\| &\leqslant \|v^k - u^k\| + \|u^k - u^{k+1}\| + \|u^{k+1} - v^{k+1}\| \\ &= \lambda^{-1}\|s^k - s^{k+1}\| + \|u^k - u^{k+1}\| + \lambda^{-1}\|s^{k+1} - s^{k+2}\| \\ &\leqslant \lambda^{-1}\frac{1+\gamma L}{\sqrt{c}}\sqrt{V^k} + \frac{1}{\sqrt{c}}\sqrt{V^k} + \lambda^{-1}\frac{1+\gamma L}{\sqrt{c}}\sqrt{V^{k+1}}.\end{aligned}$$

Since $\{V^k\}$ is nonincreasing, then

$$\|v^k - v^{k+1}\| \leqslant \left(\frac{2\lambda^{-1}(1+\gamma L)+1}{\sqrt{c}}\right)\sqrt{V^k}.$$

Then, from Lemma 3.1 it follows that for all sufficiently large k,

$$\|v^k - v^\star\| \leqslant \bar{m}\bar{\alpha}^k.$$

for some $\bar{m} > 0$ and $\bar{\alpha} \in (0,1)$.

Finally, to obtain the rate of convergence of $\{\varphi_1(u^k) + \varphi_2(v^k)\}$, first note that since $\varphi_\gamma^{\mathrm{DR}}(s^{k+1}) \geqslant \varphi^\star$, then from (4.10), (4.22), and (4.2), it follows

$$\lambda^2\|u^k - v^k\|^2 = \|s^k - s^{k+1}\|^2 \leqslant \left[\frac{(1+\gamma L)^2 q}{c}\right] r^k. \tag{4.24}$$

Furthermore, in view of (4.9),

$$|\varphi_1(u^k) + \varphi_2(v^k) - \varphi^\star| \leqslant \varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi^\star + \frac{1}{\gamma}\|u^k - s^k\|\|u^k - v^k\| + \frac{1}{2\gamma}\|u^k - v^k\|^2.$$

By assumption, $\{u^k\}$ and $\{s^k\}$ are bounded sequences, therefore there exists $M_2 > 0$, such that for all k, $\|u^k - s^k\| \leqslant M_2$. Substituting this estimate, (4.22) and (4.24) in the above inequality, yields

$$|\varphi_1(u^k) + \varphi_2(v^k) - \varphi^\star| \leqslant qr^k + \left(\frac{M_2(1+\gamma L)}{\gamma\lambda}\sqrt{\frac{q}{c}}\right)\sqrt{r}^k + \left(\frac{(1+\gamma L)^2 q}{2\gamma c\lambda^2}\right) r^k.$$

Since $r \in (0,1)$, then $r \leqslant \sqrt{r}$, and thus

$$|\varphi_1(u^k) + \varphi_2(v^k) - \varphi^\star| \leqslant \tilde{K}\sqrt{r}^k,$$

where

$$\tilde{K} = q + \frac{M_2(1+\gamma L)}{\gamma\lambda}\sqrt{\frac{q}{c}} + \frac{(1+\gamma L)^2 q}{2\gamma c\lambda^2}.$$

This proves the second part of item (i).

□

4.4 Numerical results

The numerical experiments are performed for the phase retrieval problem described in [37, §5.1]. The goal of the phase retrieval problem is to find a point $x^\star$ that (approximately) simultaneously solves the equations $\langle a_i, x\rangle^2 = b_i$, for $i = 1, \ldots, N$, $a_i \in \mathbb{R}^n$ and $b_i \geqslant 0$. As an optimization problem, we seek to find a point $x^\star$ that solves

$$\min_{x\in\mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} |\langle a_i, x\rangle^2 - b_i|. \tag{4.25}$$

Note that each $x \mapsto |\langle a_i, x\rangle^2 - b_i|$ is a weakly convex function, as the composition of the Lipschitz continuous convex function $|\cdot|$ with the smooth function $x \mapsto \langle a_i, x\rangle^2 - b_i$ with Lipschitz continuous derivative (see [46, Lemma 4.2]).

For the numerical examples, we draw the slopes a_i from a standard Gaussian distribution in $\mathbb{R}^n$, for $i = 1, \ldots, N$. We choose a target $\bar{x} \in \mathbb{R}^n$, and define $b_i = \langle a_i, \bar{x}\rangle^2$. Since problem (4.25) has a decomposable structure, we create N copies x_i of the variable x, and define $\mathcal{N} = \{(x_i)_{i=1}^N : x_1 = \cdots = x_N\}$, corresponding to a 1−stage nonanticipative subsapce from the perspective of a stochastic optimization problem. Therefore, problem (4.25) can be reformulated as the following consensus optimization problem:

$$\begin{cases} \min & \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} |\langle a_i, x_i\rangle^2 - b_i| \\ \text{s.t.} & (x_i)_{i=1}^N \in \mathcal{N}. \end{cases} \tag{4.26}$$

Since DRS applies to problems with objective functions expressed as a sum of two functions, we rewrite problem (4.26) accordingly. One possibility, explored in [23], is to add the indicator function $i_{\mathcal{C}}$ to the objective. Being non-differentiable, having a term φ_1 given by the indicator function escapes the framework of the present chapter. Instead, inspired by the approach [124] for nonconvex feasibility problems, we consider the squared distance to the feasible set as a penalty function. Namely, given a scaling/penalty factor $\mu > 0$, we work with the following formulation

$$\min_{x \in \mathbb{R}^n} \frac{1}{N}\sum_{i=1}^{N} |\langle a_i, x_i\rangle^2 - b_i| + \frac{\mu}{2} d_{\mathcal{N}}^2(x). \tag{4.27}$$

Hence, we take

$$\varphi_1(x) = \frac{\mu}{2} d_{\mathcal{N}}^2(x) \text{ and } \varphi_2(x) = \sum_{i=1}^{N} |\langle a_i, x_i\rangle^2 - b_i|$$

in (5.1). Since

$$P_{\mathcal{N}}[(x_i)_{i=1}^N] = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

in view of [130, Corollary 12.30], the derivative of the penalty function can be explicitly computed as

$$\nabla\left(\frac{\mu}{2} d_{\mathcal{N}}^2\right) = \mu(I - P_{\mathcal{N}}), \tag{4.28}$$

and the first two conditions in our blanket Assumption 4.1 hold, with $L = \mu$.

The battery of randomly generated problems parses the three settings described in [37, §5.1]:

$$(N, n) \in \{(30, 10), (150, 50), (300, 100)\},$$

for 15 uniformly randomly generated initial points in the unit sphere. We also define 5000 as the maximum number of iterations, with a target accuracy of 10^{-6} for the objective function value to stop iterations. We set the scaling parameter $\mu = \dfrac{\sqrt{N}}{2}$, since it provides the best performance out of the tests we run. We use 20 equally spaced values of λ between 0.05 and 1.95, and 5 equidistant values of γ, from 0.01 to $0.99\left(\dfrac{2-\lambda}{2\mu}\right)$.

Furthermore, the objective function φ_2 is separable, and in view of (4.4)–(4.5), at DRS iteration k the proximal subproblem to be solved for component $i = 1, \ldots, N$ is

$$\min_{x_i \in \mathbb{R}^n} |\langle a_i, x_i\rangle^2 - b_i| + \langle w_i^k, x_i\rangle + \frac{1}{2\gamma}|x_i - u_i^k|^2, \tag{4.29}$$

where $w^k = \mu\big(u^k - P_{\mathcal{N}}[u^k]\big)$, and

$$u^k = \left(\frac{1}{1+\gamma\mu}\right) s^k + \left(\frac{\gamma\mu}{1+\gamma\mu}\right) P_{\mathcal{N}}[s^k].$$

The optimality conditions of subproblems (4.29) yield four explicit critical points for each $i = 1, \ldots, N$, that are candidates to subproblem solutions, namely,

$$u_i^k - \gamma\left(w_i^k - 2\left[\frac{\gamma\langle w_i^k, a_i\rangle - \langle u_i^k, a_i\rangle}{2\gamma\|a_i\|^2 \pm 1}\right] a_i\right),$$

and

$$u_i^k - \gamma\left(w_i^k - \left[\frac{\gamma\langle w_i^k, a_i\rangle - \langle u_i^k, a_i\rangle \pm \sqrt{b_i}}{\|a_i\|^2}\right] a_i\right).$$

Note that when the dual iterate $w_i^k = 0$, we retrieve the solutions obtained by applying the stochastic proximal point method of [37, Section 5.1].

Table 1 shows the accuracy achieved for the objective function of problem (4.27) by the DRS iterates. It is clear that all problems were solved at best with a low accuracy of 10^{-1}, while nearly 38% of them reached the target accuracy of 10^{-6} or better. These results include all the values of λ and γ we tested, that is, various possible methods (by varying λ) and several stepsizes. Observe that despite being a weakly convex problem, the method does approach a global minimizer. Similar behavior has been reported for a proximal-type method in [44] for weakly convex problems.

accuracy	$< 10^{-1}$	$< 10^{-2}$	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-5}$	$< 10^{-6}$
%	100	99.365	68.254	62.857	53.016	37.778

Table 1 – Percentage of problems solved by DRS for different levels of accuracy, and different values of $\lambda \in (0, 2)$, and $\gamma \in \left(0, \dfrac{2-\lambda}{2\mu}\right)$.

As a rule of thumb, better accuracies are obtained whenever λ and γ increase within their bounds. For instance, Table 2 shows the results for $\lambda = 1.95$ and

$\gamma = 0.99\left(\frac{2-\lambda}{2\mu}\right)$. Observe that in this case, almost 96% of the problems attained the target accuracy of 10^{-6} or beyond. Choosing the best pair of parameters (λ, γ) could be a problem-dependent issue, and further investigation is required.

accuracy	$< 10^{-1}$	$< 10^{-2}$	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-5}$	$< 10^{-6}$
%	100	100	100	100	100	95.556

Table 2 – Percentage of problems solved by the DRS method for different levels of accuracy, $\lambda = 1.95$, and $\gamma = 0.99\left(\frac{2-\lambda}{2\mu}\right)$.

The benchmark in Figure 8 shows the performance of DRS for problem (4.26), the Elicited Progressive Decoupling method [131, 132] for problem (4.26), and the stochastic prox-linear method described in [37], applied to problem (4.25). The two latter methods only require setting a stepsize γ. For our numerical tests, we use 100 equally distributed stepsizes in the interval $(10^{-4}, 1)$, as in [37]. For the Elicited Progressive Decoupling, we set the elicitation parameter $e = 0$, following the numerical results and suggestions of the authors in [132], in order to accelerate convergence. The starting points, maximum number of iterations, and target accuracy are the same as described before.

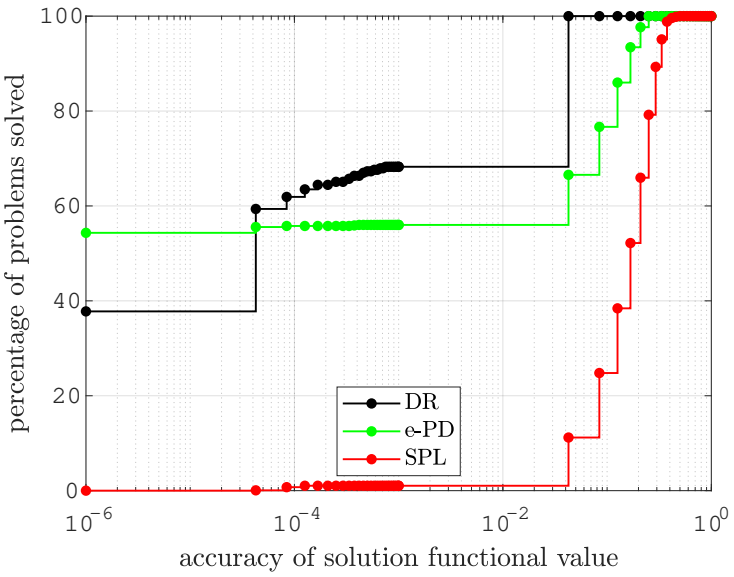


Figure 8 – Performance profile of the accuracy of the best objective function value found along iterations, for DRS using different values of λ and γ, corresponding to Table 1, Elicited Progressive Decoupling (e-PD), and stochastic prox-linear (SPL) method.

We observe that, on the considered battery of functions and accuracies, the Douglas-Rachford splitting method and the Elicited Progressive Decoupling method al-

ways outperform the prox-linear method, the one that provided the best results in [37]. This can be explained by the fact that the latter method does not inherently exploit the structure of the problem. The nonconvex version of DRS, for different values of λ, can solve approximately 40% of the problems with an accuracy of order 10^{-6}, while the Elicited Progressive Decoupling solve around 55% of the problems with this accuracy. Contrast these results with the best performance found for our proposal inTable 2.

For accuracy of order between 10^{-4} and 10^{-1}, our proposed method shows a better behavior than the Elicited Progressive Hedging. This is because the latter presents a relatively polarized performance, since it solves approximately half of the problems with an accuracy of 10^{-6} or better, and the other half just achieved an accuracy worse than 10^{-2}. In this sense, the nonconvex DRS method possesses a more distributed performance, achieving different levels of accuracy.

4.5 Final remarks

In this chapter, we presented the convergence analysis of the Douglas-Rachford splitting method in the context of optimization of weakly convex objective functions. Originally, the authors of [28] showed subsequential convergence of the DRS method in a nonconvex setting, where one of the functions of the sum is continuously differentiable with Lipschitz continuous gradient, and the other just lsc.

The fundamental feature of the analysis is that one DRS iteration can be interpreted as applying *some* descent method to the Douglas-Rachford envelope, which in the smooth convex setting coincides with the gradient descent method. Two properties conform the pillars of the analysis: a sufficient descent estimate for the Dogulas-Rachford envelope, and an estimate of a subgradient. This configuration makes it possible to apply the reasoning of [22], and thus we obtained global convergence of the iterates to critical points of the original objective function. Furthermore, when the problem displays more structure, namely, when it is weakly convex, then a subdifferential error bound allowed us to prove local linear rates of convergence to critical points.

One of the properties of the DRE is that there is a direct relation between global minimizers of the envelope and the original objective function, and their respective minimal values. However, for nonconvex problems, it is most common and less restrictive to characterize critical points as either local minimizers or saddle points. As shown in [133], smoothness conditions of the objective function, namely, the behavior of the Hessian of the objective function in a neighborhood of critical points, guarantee that DRS iterates converge to local minimizers almost surely. It would be more desirable to follow the pathway of more recent works of saddle point avoidance for simpler methods, as the

proximal point algorithm [134] or the subgradient method [135], where it is proven that the methods themselves *naturally* lead the iterates to a manifold where the objective function behaves nicely, that is, the objective function restricted to the manifold has fitting properties alongside the iterates that allow the avoidance of saddle points. It is hypothesized that the DRS method enjoy analogous properties, since it is described as a sequence of proximal steps.

Regarding the numerical experiments, the penalization term μ in problem (4.27) could be dynamically updated in order to control the weight of the penalization throughout the iterations. However, this would update the objective function in every iteration, a feature not supported by our analysis. Two lines of research could be explored in this case.

The first option would be to consider the epi-smoothing approach of [136], where a family of models approximates the objective function of our problem in the epi-convergence sense. In particular, these models can handle variable penalization parameters μ_k when the objective function is smooth, and defines a sequence $\{\varphi_k^{\mathrm{DRE}}\}$ of Douglas-Rachford envelopes parametrized by the sequence of penalization parameters $\{\mu_k\}$. This approach generates a sequence that subsequentially converges to critical points. We still need to investigate if these results can be generalized under Assumption 4.1.

Another alternative would be to use the notion of exact penalty representation of [29, Definition 11.60]. As long as the parameter $\mu > 0$ is sufficiently large, under regularity conditions akin to strong duality, solving problem (4.27) is equivalent to solving the original problem (4.26). In this case, the DRS method generates a sequence that converges to a global minimizer, which initially could be considered a bit restrictive. Nonetheless, as we saw in practice with the phase retrieval problem, weak convexity is seemingly a manageable form of nonconvexity that could allow optimization methods to obtain global minimizers.

5 A bundle-like progressive hedging algorithm

This section corresponds to [23]:

Atenas, F., & Sagastizábal, C. (2023). A bundle-like progressive hedging algorithm. *Journal of Convex Analysis*, special issue in honor of R. J-B Wets, 30(2) 453–479.

Some parts have been modified in order to follow the notation and structure of this thesis.

Abstract. For convex multistage programming problems, we propose a variant for the Progressive Hedging algorithm inspired from bundle methods. Like in the original algorithm, iterates are generated by first solving separate problems for each scenario, and then performing a projective step to ensure non-anticipativity. An additional test checks the quality of the approximation, splitting iterates into two subsequences, akin to the dichotomy between bundle serious and null steps. The method is shown to converge in both cases, and the convergence rate is linear for the serious subsequence. Our bundle-like approach endows the Progressive Hedging algorithm with an implementable stopping test. Moreover, it is possible to vary the augmentation parameter along iterations without impairing convergence. Such enhancements with respect to the original Progressive Hedging algorithm are obtained at the expense of the solution of additional subproblems at each iteration, one per scenario.

5.1 Introduction and motivation

Multistage stochastic programs represent an important source of large-scale optimization problems. This is because the number of variables and constraints needed to formulate the problem grows with the number of scenarios. As illustrated by energy applications in [6], decomposition methods are essential for solving effectively this type of problems; see also [137].

A very popular approach that deals with scenario decomposition is the Progressive Hedging algorithm (PHA), introduced in [24] and later extended to handle risk measures in [138]. The setting is such that the problem uncertainty, represented by a set of S scenarios, reveals progressively, in T stages. For each scenario s, the decision variable

is $x_s \in \mathbb{R}^n$, the convex objective function is $f_s : \mathbb{R}^n \to \mathbb{R}$ and a nonempty convex compact feasible set $C_s \subseteq \mathbb{R}^n$ is given. Consider the following stochastic multistage optimization problem,

$$\begin{cases} \min\limits_{x} & \mathbb{E}[f(x)] := \sum\limits_{s=1}^{S} p_s f_s(x_s) \\ \text{s.t.} & x_s \in C_s, \quad \text{for all } s = 1, \ldots, S, \\ & x \in \mathcal{N}\,, \end{cases} \tag{5.1}$$

where, in the last inclusion, the linear subspace $\mathcal{N}$ gathers the so-called non-anticipativity constraints. Such constraints ensure that, at each stage t, the decision making process depends only on information of the uncertainty that is available at time t.

Non-anticipativity constraints couple decisions along scenarios in a structured manner. The PHA decouples those constraints so that, at each iteration, individual subproblems can be solved separately for each scenario. This makes the approach very suitable for parallel implementations. The parallel phase of separate scenario subproblems is followed by a synchronization step that yields a non-anticipative vector by projecting onto the linear subspace $\mathcal{N}$.

The PHA is in fact a Douglas-Rachford splitting, in a space endowed with a weighted scalar product; see for example [139] and Section 2.2.2. Early work on splitting methods, dealing with the classical problem of finding a zero of a sum of maximal monotone operators, can be traced back to [11, 87, 140]. The more recent family of projective splitting methods [141, 142, 143, 117] expanded significantly the reach of Douglas-Rachford approaches. The projective hedging algorithm [139], in particular, can operate in block-iterative and partially asynchronous manner. For a randomized asynchronous variant of the PHA, we refer to [144].

Another important advance of projective splitting methods is the ability to dynamically update certain proximal parameter involved in the calculations. Along iterations, the PHA also generates a dual sequence, that lies in $\mathcal{N}^{\perp}$, the orthogonal complement of the non-anticipative space. The update of the dual variable amounts to maximizing certain dual function, by applying a gradient method with fixed stepsize (related to the aforementioned proximal parameter). Empirically, splitting methods are generally observed to converge linearly, but at a rate that becomes asymptotically slow and is heavily dependent on the parameter choice. In order to speed up the process, practitioners resort to various heuristic techniques that can be computationally expensive and may not always prove successful [145, 146, 147]. This difficulty is not a surprise as, for the specific PHA context, the parameter in question must strike a good balance between optimality and feasibility. Clearly, such a goal is not easy to attain with a value that is kept fixed along iterations.

Our proposal is to employ, instead, a proximal bundle method [85, Part II], tailored to maximize the dual function over $\mathcal{N}^\perp$. To this aim, we consider a more general setting, the constrained minimization of weakly convex functions. For tackling such problems, assuming that projecting onto the feasible set is an easy operation, we propose to apply a bundle algorithm of projective type.

This chapter is organized as follows. After introducing some notation and definitions, we recall the PHA and formulate a problem dual to (5.1). This dual problem is a particular instance of a constrained weakly convex problem considered in Section 3.5. Therein, the subspace $\mathcal{N}^\perp$ is replaced by a general constraint set $\mathcal{M}$, onto which projecting points is assumed to be easy. The general methodology developed in Section 3.5 is particularized to (5.1) in Section 5.2. The resulting Bundle Progressive Hedging algorithm (BPHA) is formulated in both primal and dual forms. Convergence to solutions of (5.1) and (5.7) is shown in Section 5.3, using the dual form. An equivalent primal formulation is useful to compare our new approach with the original PHA. The final Section 5.4 discusses similarities and differences between the original Progressive Hedging algorithm and our proposal.

Notation and some definitions

Our notation is standard, following mainly [29], [30] and Section 2.2.2. Given S scenarios, for a certain scenario realization $s \in \{1, \dots, S\}$, the probability of occurrence is denoted by $p_s > 0$. Without loss of generality, we can assume that the dimension of the decision variable of scenario $s = 1, \dots, S$ is $n_s = n$, and thus $\sum_{s=1}^{S} n_s = nS$. For a vector $v \in \mathbb{R}^{nS}$ with components $v_s \in \mathbb{R}^n$ for all $s = 1, \dots, S$, the expected value and the conditional expectation at stage t are respectively denoted by

$$\mathbb{E}[v] = \sum_{s=1}^{S} p_s v_s, \text{ and } \mathbb{E}_{[t-1]}[v]\,,$$

where the uncertainty realization at stages $1, \dots, t-1$ is known.

Considering that all vectors are column vectors, the inner product employed in the space of decision variables is

$$\forall u, v \in \mathbb{R}^{nS} : \langle u, v \rangle_S = \sum_{s=1}^{S} p_s \langle u_s, v_s \rangle\,.$$

Note that this inner product uses the (nonnegative) probability of each scenario as a weight, a crucial feature in the analysis. Recall we denote by $\|\cdot\|_S$ its corresponding induced norm in $\mathbb{R}^{nS}$, while $\|\cdot\|$ stands for the usual Euclidean norm in $\mathbb{R}^n$. Throughout the text, the symbol $\langle \cdot, \cdot \rangle$ refers to any either inner product (with or without weights), and similarly for the corresponding norms, when it is clear from the context.

On non-anticipativity.

In the stochastic setting, decisions have to be taken progressively, as uncertainty is revealed: at stage t, decisions should only depend on the information that became available in stages $1, \dots, t-1$. Accordingly, if $x_{t,s}$ denotes the decision made in stage t for scenario s, the following relations, called of *non-anticipativity*, need to hold for all t and s:

$$x_{t,s} = \mathbb{E}_{[t-1]}[x_t]\,.$$

In particular, for the first stage, this constraint states that $x_{1,s}$ are equal for all scenario s. Non-anticipativity constraints define a linear subspace $\mathcal{N}$ of decision variables characterized by conditional expected values. In a manner similar, the projection operator $P_{\mathcal{N}}$ onto $\mathcal{N}$ is characterized by the following simple algebraic relations:

$$\text{for each stage } t,\ P_{\mathcal{N}}[x]_t = \mathbb{E}_{[t-1]}[x_t].$$

Being a self-adjoint operator, the following relation holds for the projection: $P_{\mathcal{N}^\perp} = I - P_{\mathcal{N}}$.

Primal and dual formulations of the multistage program

To induce separability, the PHA relaxes the non-anticipativity constraint in (5.1), namely

$$x \in \mathcal{N} \iff x = P_{\mathcal{N}}[x]\,,$$

by means of the following Lagrangian:

$$\mathcal{L}(x, w) = \mathbb{E}[f(x)] + \langle w, x - P_{\mathcal{N}}[x]\rangle_S\,.$$

Because of the identity $P_{\mathcal{N}^\perp} = I - P_{\mathcal{N}}$, the Lagrangian multiplier $w = (w_s)_{s=1}^S \in \mathbb{R}^{nS}$ can be assumed to satisfy

$$w \in \mathcal{N}^\perp \iff P_{\mathcal{N}}[w] = 0\,. \tag{5.2}$$

Furthermore, by perpendicularity, the linear term in the relaxation can be simplified:

$$\langle w, x - P_{\mathcal{N}}[x]\rangle_S = \langle w, x\rangle_S - \langle w, P_{\mathcal{N}}[x]\rangle_S = \langle w, x\rangle_S.$$

Therefore, the Lagrangian

$$\mathcal{L}(x, w) = \mathbb{E}[f(x)] + \langle w, x\rangle_S$$

is decomposable along scenarios. More specifically,

$$\mathcal{L}(x, w) = \sum_{s=1}^{S} p_s \mathcal{L}_s(x_s, w_s)\,, \tag{5.3}$$

where for each scenario $s = 1, \ldots, S$, we defined the $s-$Lagrangian

$$\mathcal{L}_s(x_s, w_s) = f_s(x_s) + \langle w_s, x_s \rangle . \tag{5.4}$$

Throughout our development, subindices s refer to scenario components (of functions, sets, or vectors). To ease the understanding, Table 3 below summarizes the main elements in our notation and clarifies their dimensionality.

In Algorithm 5, with the PHA, the s-Lagrangians (5.4) are used to construct separate subproblems in the primal space, one per scenario. Algorithm 5 is Algorithm 4 written with the notation of this chapter.

Algorithm 5 Progressive Hedging Algorithm

1: **Initialization**: Choose a primal-dual starting point $(x^0, w^0) \in \mathcal{N} \times \mathcal{N}^\perp$.
2: **for** $k = 0, 1, \ldots$ **do**
3: **Primal subproblems:** solve for $s = 1, \ldots, S$,

$$x_s^{k+\frac{1}{2}} = \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2. \tag{5.5}$$

4: **Primal projection:** $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$.
5: **Dual update:** $w_s^{k+1} = w_s^k + t_k(x_s^{k+\frac{1}{2}} - x_s^{k+1})$ for $s = 1, \ldots, S$.
6: **end for**

Notice that, by (5.4) and (5.5), the dual updating rule can be interpreted as an approximate gradient step for maximizing an augmented s-dual function.

A notable feature of the PHA is that feasibility is achieved both in the primal and dual iterates by performing simple calculations. On the primal space, the vector formed by collecting the subproblem solutions is projected onto $\mathcal{N}$. On the dual space, the difference $x^{k+\frac{1}{2}} - x^{k+1}$, which measures primal feasibility, lies in $\mathcal{N}^\perp$. As a result, dual feasibility is guaranteed throughout the iterative process, as long as the starting dual point belongs to $\mathcal{N}^\perp$. The Bundle Progressive Hedging algorithm proposed in this work preserves this characteristic, introducing some modifications in the PHA scheme by resorting to duality. More precisely, consider the convex dual function derived from the Lagrangian, that is

$$h(w) = \sum_{s=1}^{S} p_s h_s(w_s)\,, \text{ where } h_s(w_s) = \max_{x_s \in C_s}(-\mathcal{L}_s)(x_s, w_s)\,. \tag{5.6}$$

With this notation, the problem dual to (5.1) has the expression

$$\begin{cases} \min\limits_{w} & h(w) \\ \text{s.t.} & w \in \mathcal{N}^\perp \end{cases} \tag{5.7}$$

Since the multistage stochastic program (5.1) is convex with a linear constraint, there is no duality gap. As a result, a convergent dual method applied to (5.7) yields solutions to the original problem. Note that in view of weak duality and the fact that the feasible set of the primal problem is compact and the primal objective function is continuous, then $\inf_{\mathcal{N}^\perp} h > -\infty$.

	vector in $\mathbb{R}^{nS}$	scenario s subvector in $\mathbb{R}^n$
primal variable	x	x_s
dual variable	w	w_s
	full function	scenario s subfunction
primal objective	$f : \mathbb{R}^{nS} \to \mathbb{R}$	$f_s : \mathbb{R}^n \to \mathbb{R}$
Lagrangian	$\mathcal{L} : \mathbb{R}^{nS} \times \mathbb{R}^{nS} \to \mathbb{R}$	$\mathcal{L}_s : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$
dual objective	$h : \mathbb{R}^{nS} \to \mathbb{R}$	$h_s : \mathbb{R}^n \to \mathbb{R}$

Table 3 – Notation and dimensionality of variables and functions.

5.2 Bundle Progressive Hedging

Our motivation to consider algorithmic schemes of the form (3.20a)-(3.20c) is to exploit decomposable structures in the objective function in (3.16). The challenge is to define model functions that inherit h's structure and, at the same time, incorporate information on the feasible set $\mathcal{M}$ without destroying separability. We now explain how to build suitable model functions for (5.7) and present the bundle variant of the PHA.

5.2.1 Building separable models

As mentioned, usually bundle methods define linearizations for the sum $h+i_{\mathcal{N}^\perp}$, but such model functions are not separable. To illustrate this difficulty, consider the following simple instance of (3.16), where $w_1, w_2 \in \mathbb{R}$ are decision variables:

$$\begin{cases} \min\limits_{w_1,w_2} & h_1(w_1) + h_2(w_2) \\ \text{s.t.} & w_1 + w_2 = 0\,. \end{cases}$$

The equality constraint represents $(w_1, w_2) \in \mathcal{N}^\perp$ for a uniformly distributed random variable with two scenarios. A classical model for this problem takes cutting-plane approximations for each term, say $\check{h}_s(w_s)$ for $s = 1, 2$, and adds the indicator function:

$$\varphi^k_{w^k}(w_1, w_2) := \check{h}_1(w_1) + \check{h}_2(w_2) - h_1(w^k_1) - h_2(w^k_2) + i_{\mathcal{N}^\perp}(w_1, w_2)\,.$$

With such a model, problem (3.19), whose solution yields (3.20a), is not separable:

$$\min_{w_1+w_2=0} \left\{ \varphi^k_{w^k}(w_1, w_2) + \frac{1}{2t_k} \left\| \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} - \begin{pmatrix} w^k_1 \\ w^k_2 \end{pmatrix} \right\|^2_S \right\}.$$

By contrast, if a separable model function was available for the indicator function $i_{\mathcal{N}^\perp}$ evaluated at (w_1, w_2), the calculations required in (3.20a) could be performed separately, by solving, for $s = 1, 2$,

$$\min_{w_s} \left\{ \varphi^k_{w^k_s}(w_s) + \frac{1}{2t_k} \|w_s - w^k_s\|^2 \right\} . \tag{5.8}$$

When the objective function h involves the sum of more than two terms, as in the case of multistage programs with many scenarios, using separable models improves significantly the computational performance (the solution to (5.8) can be computed in parallel for the different scenarios).

Following the espirit of Table 3, and because the ability of defining separable models depends on the constraints $\mathcal{M}$ under consideration, we recall in Table 4 the feasible sets involved in the primal and dual problems.

space	nature	notation
general framework of Section 3.5	-	$\mathcal{M}$
nonanticipative	primal	$\mathcal{N}$
nonanticipative orthogonal	dual	$\mathcal{N}^\perp$

Table 4 – Spaces involved in the general framework of Section 3.5, and in the BPHA analysis.

Regarding the PHA setting, that is when (3.16) has the format (5.6)−(5.7), the feasible set therein is given by the non-anticipativity constraints written in dual form. Hence, as in (5.2),

$$\mathcal{M} = \mathcal{N}^\perp = \{w : P_{\mathcal{N}}[w] = 0\} .$$

This subspace has a favorable structure, in particular, the projection step (3.20b) is straightforward:

$$P_{\mathcal{M}}[w] = w - P_{\mathcal{N}}[w] .$$

Recall that, by (5.6), $h(w) = \sum_{s=1}^{S} p_s h_s(w_s)$. Our proposal is to work with the weighted sum of special model functions, defined for each scenario and derived from (5.5).

Proposition 5.1 (Separable 1QA models)**.** *Given $(x^k, w^k) \in \mathcal{N} \times \mathcal{N}^\perp$, for $s = 1, \ldots, S$ consider the approximate s-Lagrangian*

$$\mathcal{L}^k_s(x_s, w_s) = f(x_s) + \langle w_s, x_s - x^k_s \rangle = \mathcal{L}_s(x_s, w_s) - \langle w_s, x^k_s \rangle , \tag{5.9}$$

and the individual functions

$$\varphi^k_{w^k_s}(w_s) := \max_{x_s \in C_s} (-\mathcal{L}^k_s)(x_s, w_s) - h_s(w^k_s) . \tag{5.10}$$

The corresponding weighted sum

$$\varphi^k_{w^k}(w) = \sum_{s=1}^{S} p_s \varphi^k_{w^k_s}(w_s) = h(w) - h(w^k) + \left\langle w, x^k \right\rangle_S$$

defines a model for $(h + i_{\mathcal{N}^\perp})(w)$ *at* w^k *that is separable, convex and of type 1QA with* $q = 0$.

Proof. First, the model $\varphi^k_{w^k}$ is separable by construction, and convex, since each $\varphi^k_{w^k_s}$ is the maximum of affine functions $-\mathcal{L}_s(x_s, w_s) - h_s(w^k_s)$ in w_s. From (5.10) it follows that

$$\varphi^k_{w^k_s}(w_s) \leqslant h_s(w_s) + \langle w_s, x^k_s \rangle - h_s(w^k_s).$$

After multiplying by p_s and adding all scenarios, this means that

$$\varphi^k_{w^k}(w) \leqslant h(w) + \langle w, x^k \rangle_S - h(w^k).$$

Note that since $x^k \in \mathcal{N}$, then $\langle w, x^k \rangle_S \leqslant i_{\mathcal{N}^\perp}(w)$ for all w. Indeed, if $w \in \mathcal{N}^\perp$, $\langle w, x^k \rangle_S = 0 \leqslant 0 = i_{\mathcal{N}^\perp}(w)$. On the other hand, if $w \in \mathcal{N} \backslash \{0\}$, $\langle w, x^k \rangle$ is a finite value, thus $\langle w, x^k \rangle_S < +\infty = i_{\mathcal{N}^\perp}(w)$. Therefore,

$$\varphi^k_{w^k}(w) \leqslant h(w) + i_{\mathcal{N}^\perp}(w) - h(w^k).$$

Hence, $\varphi^k_{w^k}$ is a 1QA model for $h + i_{\mathcal{N}^\perp}$ at w^k with $q = 0$. □

The approximate s-Lagrangians put together build a 1QA model for $h + i_{\mathcal{N}^\perp}$. In fact, as

$$\mathcal{L}^k_s(x_s, w_w) = \mathcal{L}_s(x_s, w_s) - \langle x^k_s, w_s \rangle,$$

the term $\langle x^k, w \rangle_S$ corresponds to a lower linearization of the indicator function $i_{\mathcal{N}^\perp}$. Using the wording from [148], the models $\varphi^k_{w^k}$ are of lower type; see also [149].

5.2.2 Comparison with subproblems in the Progressive Hedging algorithm

To understand the given definition for the individual models, first recall we are interested in dealing with separate subproblems (5.8). Plugging (5.10) therein yields

$$\min_{w_s} \left\{ \max_{x_s \in C_s} (-\mathcal{L}^k_s)(x_s, w_s) - h_s(w^k_s) + \frac{1}{2t_k} \|w_s - w^k_s\|^2 \right\}$$

$$= \min_{w_s} \left\{ \max_{x_s \in C_s} (-\mathcal{L}_s)(x_s, w_s) + \langle x^k_s, w_s \rangle - h_s(w^k_s) + \frac{1}{2t_k} \|w_s - w^k_s\|^2 \right\}.$$

For this (convex-concave) saddle-point problem it is equivalent to solve

$$\max_{x_s \in C_s} \min_{w_s} \left\{ (-\mathcal{L}_s)(x_s, w_s) + \langle x_s^k, w_s \rangle - h_s(w_s^k) + \frac{1}{2t_k} \|w_s - w_s^k\|^2 \right\}$$

$$= \quad -h_s(w_s^k) + \max_{x_s \in C_s} \left\{ \min_{w_s} \left[-f_s(x_s) - \langle w_s, x_s - x_s^k \rangle + \frac{1}{2t_k} \|w_s - w_s^k\|^2 \right] \right\} .$$

The expression between brackets is minimized at $w_s = w_s^k + t_k(x_s - x_s^k)$. Therefore,

$$\begin{aligned}
\text{(5.8) is equivalent to} \quad & \max_{x_s \in C_s} \left\{ -f_s(x_s) - \langle w_s^k + t_k(x_s - x_s^k), x_s - x_s^k \rangle + \frac{t_k}{2} \|x_s - x_s^k\|^2 \right\} \\
= \quad & \max_{x_s \in C_s} \left\{ -f_s(x_s) - \langle w_s^k, x_s - x_s^k \rangle - t_k \|x_s - x_s^k\|^2 + \frac{t_k}{2} \|x_s - x_s^k\|^2 \right\} \\
= \quad & - \min_{x_s \in C_s} \left\{ f_s(x_s) + \langle w_s^k, x_s - x_s^k \rangle + \frac{t_k}{2} \|x_s - x_s^k\|^2 \right\} \\
= \quad & \langle w_s^k, x_s^k \rangle - \min_{x_s \in C_s} \mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2} \|x_s - x_s^k\|^2 .
\end{aligned} \tag{5.11}$$

This last minimization problem is practically identical to the PHA original subproblem (5.5). The difference is that the stepsize t_k, in this case, can be dynamically updated, as in a bundle method.

Additionally, note that the Lagrangian $\mathcal{L}_s^k$ is an approximation of the original Lagrangian obtained when relaxing the non-anticipativity constraint. More specifically, the Lagrangian

$$\mathcal{L}(x, w) = \sum_{s=1}^{S} p_s \left(f_s(x_s) + \langle w_s, x_s - P_{\mathcal{N}}[x]_s \rangle \right)$$

is approximated with

$$\mathcal{L}^k(x, w) = \sum_{s=1}^{S} p_s \left(f_s(x_s) + \langle w_s, x_s - x_s^k \rangle \right), \quad x_s^k \in \mathcal{N}.$$

Consequently, the non-separable augmented Lagrangian subproblems

$$\sum_{s=1}^{S} p_s \min_{x_s \in C_s} \left\{ f_s(x_s) + \langle w_s, x_s - P_{\mathcal{N}}[x]_s \rangle + \frac{t_k}{2} \|x_s - P_{\mathcal{N}}[x]_s\|^2 \right\}$$

are approximated (in both PHA and BPH) by the separable subproblems

$$\sum_{s=1}^{S} p_s \min_{x_s \in C_s} \left\{ f_s(x_s) + \langle w_s, x_s - x_s^k \rangle + \frac{t_k}{2} \|x_s - x_s^k\|^2 \right\}$$

The difference is that, with our approach, the quality of this approximation is measured by means of the serious/null step test (3.20c). Performing the descent test is not free, as it requires an extra dual function evaluation, that is, solving another set of subproblems in parallel. This is made clear in the descent test in Algorithm 6 given below.

5.2.3 Statement of the algorithm in dual form

Recall that the PHA keeps separate subproblems for each scenario, performing afterwards a coordination step, projecting the primal candidates onto the linear subspace of non-anticipative constraints. The Bundle Progressive Hedging Algorithm 6 maintains those features, adding a descent test to measure the quality of the model approximation. The main steps of the algorithm are as follows.

- Iteration k starts by solving, for each scenario $s = 1, \ldots, S$, a dual subproblem, yielding dual intermediate iterates.
- Intermediate primal points are derived from the dual intermediate solutions. As shown in Lemma 5.1(i), such primal points actually correspond to minimizing an augmented s-Lagrangian in the primal space.
- The projection of the dual intermediate points onto the orthogonal subspace of the non-anticipativity ensures dual feasibility.
- The projected dual iterates are evaluated, to determine if there was some decrease in the dual function h:
  - if there is sufficient descent, a serious step is made. This point then becomes the best candidate point generated so far, or
  - if there is no sufficient descent, a null step is made.

The Bundle Progressive Hedging Algorithm 6 results from applying the scheme (3.20) with the 1QA model in Proposition 5.1, to the dual problem (5.7). Being a particular instance of the general algorithmic pattern given in Section 3.5, convergence of the serious dual subsequence generated by BPHA in therefore ensured by Theorem 3.5.

Table 5 compares the notation employed in Algorithms 6 and 7 to generate, respectively, the primal and dual sequences of the bundle approach.

	vector in $\mathbb{R}^{nS}$	scenario s subvector in $\mathbb{R}^n$
intermediate primal	$x^{k+\frac{1}{2}}$	$x_s^{k+\frac{1}{2}}$
projected primal	x^k	x_s^k
intermediate dual	$w^{k+\frac{1}{2}}$	$w_s^{k+\frac{1}{2}}$
projected dual	u^k	u_s^k
dual serious step	w^k	w_s^k
minimizer of $\mathcal{L}_s(\cdot, u_s^k)$ over C_s	y^k	y_s^k

Table 5 – Notation for primal and dual sequences generated by Algorithms 6 and 7.

Algorithm 6 Bundle Progressive Hedging Algorithm in dual form

1: **Initialization**: Given a stopping tolerance $\texttt{TOL} \geqslant 0$ and parameters $m \in (0,1)$ and $t_{\min} > 0$, choose a primal-dual starting point $(x^0, w^0) \in \mathcal{N} \times \mathcal{N}^{\perp}$ and an initial stepsize $t_0 > 0$.
Compute the dual value $h_s(w_s^0) = -\mathcal{L}_s(\widehat{y}_s^0, w_s^0)$, by finding

$$\widehat{y}_s^0 \in \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, w_s^0) \text{ for } s = 1, \ldots, S\,.$$

2: **for** $k = 0, 1, \ldots$ **do**
3: **Dual subproblems:** given $\varphi_{w_s^k}^k(w_s) = \max_{x_s \in C_s}(-\mathcal{L}_s^k)(x_s, w_s) - h_s(w_s^k)$, solve

$$w_s^{k+\frac{1}{2}} = \arg\min_{w_s} \left\{ \varphi_{w_s^k}^k(w_s) + \frac{1}{2t_k}\|w_s - w_s^k\|^2 \right\} \text{ for } s = 1, \ldots, S\,. \tag{5.12}$$

The nominal decrease $\varphi_{w_s^k}^k(w_s^{k+\frac{1}{2}})$ is available.
4: **Primal projection:** $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$, where

$$x_s^{k+\frac{1}{2}} = x_s^k + \frac{1}{t_k}(w_s^{k+\frac{1}{2}} - w_s^k) \text{ for } s = 1, \ldots, S\,. \tag{5.13}$$

5: **Stopping test**: if $-\varphi_{w^k}^k(w^{k+\frac{1}{2}}) \leqslant \texttt{TOL}$, stop and return (x^k, w^k).
6: **Dual projection:** $u^{k+1} = P_{\mathcal{N}^{\perp}}[w^{k+\frac{1}{2}}] = w^{k+\frac{1}{2}} - P_{\mathcal{N}}[w^{k+\frac{1}{2}}]$.
7: **Descent test:** Compute the dual value $h_s(u_s^{k+1}) = -\mathcal{L}_s(y_s^{k+1}, u_s^{k+1})$, by finding

$$y_s^{k+1} \in \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, u_s^{k+1}) \text{ for } s = 1, \ldots, S\,.$$

If $-\mathcal{L}(y^{k+1}, u^{k+1}) \leqslant -\mathcal{L}(\widehat{y}^k, w^k) + m\varphi_{w^k}^k(w^{k+\frac{1}{2}})$ declare a serious step:

$$\begin{cases} w^{k+1} = u^{k+1} \\ \widehat{y}^{k+1} = y^{k+1} \\ t_{k+1} \geqslant t_{\min} \end{cases}$$

Otherwise, declare a null step: set $w^{k+1} = w^k$, $\widehat{y}^{k+1} = \widehat{y}^k$ and choose $t_{k+1} \geqslant t_{\min}$.
8: **end for**

Before passing to Section 5.3, with BPHA's convergence analysis, some remarks are in order.

- Notice that the descent test is performed with the full Lagrangian $\mathcal{L}$, and not the individual s-Lagrangians.
- In (3.20), consider h to be the dual function defined in (5.6), the linear subspace $\mathcal{M} = \mathcal{N}^{\perp}$, and $N = nS$. First, the optimality conditions of the problem in (5.12)

correspond to (3.20a), for

$$G^k = -(x^k - x^{k+\frac{1}{2}}), \tag{5.14}$$

according to (5.13). Furthermore, step 6 of Algorithm 6 is exactly (3.20c), while the descent test in step 7 of Algorithm 6 is equivalent to (3.20c), by using the definition of the dual function h in (5.6), the construction of y^{k+1}, and the definition of $\hat{y}^k$. Barring the stopping test, Algorithm 6 is a particular instance of the pattern in (3.20).

– The stopping test of step 5 in Algorithm 7 yields indeed approximate solutions. To see this, recall that the BPHA stops when the aggregate error and the aggregate subgradient defined in Lemma 3.4(i) are sufficiently small. Since, see Lemma 3.4(ii), $G^k \in \partial_{E^k} \varphi^k_{w^k_s}(w^k_s)$ for $E^k \geqslant 0$, and $\varphi^k_{w^k_s}(w^k_s) = 0$, this means that for all w

$$h(w) - h(w^k) + \left\langle w, x^k \right\rangle_S \geqslant \left\langle G^k, w - w^k \right\rangle_S - E^k \,, \text{ where } t_k \|G^k\|_S^2 \,, E^k \leqslant \texttt{TOL}.$$

In particular, for all $w \in \mathcal{N}^\perp$ the linear term in the left-hand side vanishes, and w^k is an approximate minimizer of the dual problem (5.7), for $\eta = \max\left\{\texttt{TOL}, \sqrt{\frac{\texttt{TOL}}{t_{\min}}}\right\}$:

$$h(w) \geqslant h(w^k) - \eta \|w - w^k\|_S - \eta \,. \tag{5.15}$$

5.2.4 Relation with Progressive Hedging and primal formulation

The BPHA in Algorithm 6 differs from the PHA in Algorithm 5 in the implementation of the descent test. Notwithstanding, both methods also share some features. As shown in (5.11), for each scenario, the subproblem (5.12) solved by BPHA is dual to the subproblem of the PHA in Algorithm 5. In both PHA and BPHA, primal points are projected onto the set of non-anticipativity constraints; and the nature of the dual update is the same, projecting onto $\mathcal{N}^\perp$ to guarantee dual feasibility.

The resemblance between the two methods becomes more apparent after formulating Algorithm 6 in primal terms, exploiting in addition to (5.11), primal-dual relation stated in Lemma 5.2.2. The primal formulation of the Bundle Progressive Hedging given in Algorithm 7 allows a straightforward comparison with the PHA in Algorithm 5. However, the dual form Algorithm 6 is more handy for the convergence analysis.

Lemma 5.1 (From dual to primal BPHA formulations)**.** *Consider the approximate s-Lagrangian $\mathcal{L}^k_s$ defined in (5.9), and the model $\varphi^k_{w^k}$ of (5.10) that defines subproblems (5.12) in Algorithm 6. The following holds.*

(i) For all scenario $s = 1, \dots, S$, the intermediate step $x_s^{k+\frac{1}{2}}$ defined in (5.13) can be equivalently computed as follows

$$x_s^{k+\frac{1}{2}} \in \arg\min_{x_s \in C_s} \left\{ \mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2 \right\}.$$

In particular, $x_s^{k+\frac{1}{2}} \in C_s$.

(ii) The dual projection rule $u^{k+1} = w^{k+\frac{1}{2}} - P_{\mathcal{N}}[w^{k+\frac{1}{2}}]$ can be equivalently performed by doing

$$u^{k+1} = w^k + t_k(x^{k+\frac{1}{2}} - x^{k+1}).$$

(iii) The dual model $\varphi_{w^k}^k$ evaluated at the dual intermediate point $w^{k+\frac{1}{2}}$ can be written in primal-dual terms as follows

$$\varphi_{w^k}^k(w^{k+\frac{1}{2}}) = \mathcal{L}(\widehat{y}^k, w^k) - \mathcal{L}^k(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}).$$

Proof. To show item (i), recall the relations (5.11). Since $w_s^{k+\frac{1}{2}}$ minimizes $\varphi_{w_s^k}^k(w_s) + \frac{1}{2t_k}\|w_s - w_s^k\|^2$, and it also has the form $w_s^{k+\frac{1}{2}} = w_s^k + t_k(x_s^{k+\frac{1}{2}} - x_s^k)$, then $x_s^{k+\frac{1}{2}}$ minimizes over C_s

$$f_s(x_s) + \langle w_s^k, x_s - x_s^k\rangle + \frac{t_k}{2}\|x_s - x_s^k\|^2 = \mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2.$$

Regarding item (ii), since $w^{k+\frac{1}{2}} = w^k + t_k(x^{k+\frac{1}{2}} - x^k)$, and $w^k \in \mathcal{N}^\perp$, then

$$\begin{aligned} w^{k+\frac{1}{2}} - P_{\mathcal{N}}[w^{k+\frac{1}{2}}] &= w^k + t_k(x^{k+\frac{1}{2}} - x^k) - P_{\mathcal{N}}[w^k + t_k(x^{k+\frac{1}{2}} - x^k)] \\ &= w^k + t_k(x^{k+\frac{1}{2}} - x^k) - t_k(x^{k+1} - x^k) \\ &= w^k + t_k(x^{k+\frac{1}{2}} - x^{k+1}), \end{aligned}$$

where in the second equality we use $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$.

Finally, note that (5.13) and (5.12) imply that $x_s^{k+\frac{1}{2}}$ solves

$$\max_{x_s \in C_s}(-\mathcal{L}_s^k)(x_s, w_s^{k+\frac{1}{2}}).$$

Therefore, from the definition of $\varphi_{w_s^k}^k$, it holds that

$$\varphi_{w_s^k}^k(w_s^{k+\frac{1}{2}}) = (-\mathcal{L}_s^k)(x_s^{k+\frac{1}{2}}, w_s^{k+\frac{1}{2}}) - h_s(w_s^k).$$

Moreover, by construction, $\widehat{y}_s^k$ solves $\max_{x_s \in C_s}(-\mathcal{L}_s)(x_s, w_s^k)$, then

$$\varphi_{w_s^k}^k(w_s^{k+\frac{1}{2}}) = (-\mathcal{L}_s^k)(x_s^{k+\frac{1}{2}}, w_s^{k+\frac{1}{2}}) + \mathcal{L}_s(\widehat{y}_s, w_s^k).$$

Taking the expected value in this last formula gives item (iii). □

By Lemma 5.1, the intermediate primal point $x_s^{k+\frac{1}{2}}$ can be computed as either a minimizer over C_s of the augmented approximate s-Lagrangian

$$\mathcal{L}_s^k(x_s, w_s^k) + \frac{t_k}{2}\|x_s - x_s^k\|^2,$$

evaluated at the serious dual point $w_s = w_s^k$, or as a minimizer over C_s of the Lagrangian

$$\mathcal{L}_s(x_s, w_s^{k+\frac{1}{2}}),$$

evaluated at the intermediate dual point $w_s = w_s^{k+\frac{1}{2}}$. Together with the primal updating rule (5.13), this means that the dual subproblems of Algorithm 6 can be written in primal terms, followed by a dual updating rule for $w_s^{k+\frac{1}{2}}$ deduced from (5.13). Algorithm 7, given next, is the primal version of Algorithm 6.

Algorithm 7 Bundle Progressive Hedging Algorithm in primal form

1: **Initialization**: Given a stopping tolerance $\texttt{TOL} \geqslant 0$ and parameters $m \in (0,1)$ and $t_{\min} > 0$, choose a primal-dual starting point $(x^0, w^0) \in \mathcal{N} \times \mathcal{N}^\perp$. Compute

$$\widehat{y}_s^0 \in \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, w_s^0) \text{ for } s = 1, \ldots, S\,.$$

2: **for** $k = 0, 1, \ldots$ **do**

3: **Primal subproblems:** solve

$$x_s^{k+\frac{1}{2}} = \arg\min_{x_s \in C_s} \left\{ f_s(x_s) + \langle w_s^k, x_s \rangle + \frac{t_k}{2}\|x_s - x_s^k\|^2 \right\}, \text{ for } s = 1, \ldots, S. \tag{5.16}$$

4: **Primal projection:** $x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]$.

5: **Dual update:** Compute $w_s^{k+\frac{1}{2}} = w_s^k + t_k(x_s^{k+\frac{1}{2}} - x_s^k)$ for $s = 1, \ldots, S\,.$

6: **Stopping test:** if $\mathcal{L}^k(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}) - \mathcal{L}(\widehat{y}^k, w^k) \leqslant \texttt{TOL}$, stop and return (x^k, w^k).

7: **Dual projection:** $u^{k+1} = w^k + t_k(x^{k+\frac{1}{2}} - x^{k+1})$.

8: **Descent test:** Compute

$$y_s^{k+1} \in \arg\min_{x_s \in C_s} \mathcal{L}_s(x_s, u_s^{k+1}) \text{ for } s = 1, \ldots, S\,.$$

If $\mathcal{L}(\widehat{y}^k, w^k) - \mathcal{L}(y^{k+1}, u^{k+1}) \leqslant m(\mathcal{L}(\widehat{y}^k, w^k) - \mathcal{L}^k(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}))$, declare a serious step: set $w^{k+1} = u^{k+1}$, $\widehat{y}^{k+1} = y^{k+1}$, and take $t_{k+1} \geqslant t_{\min}$.
Otherwise, declare a null step: set $w^{k+1} = w^k$, $\widehat{y}^{k+1} = \widehat{y}^k$ and choose $t_{k+1} \geqslant t_{\min}$.

9: **end for**

We now show that Bundle Progressive Hedging algorithm finds primal and dual solutions, in the case of finite termination, infinite number of serious steps or when there is a tail of null steps.

5.3 Convergence analysis of the Bundle Progressive Hedging

The primal and dual versions of the Bundle Progressive Hedging algorithm are equivalent because:

- Steps 3 and 4 of Algorithm 6 are equivalent to steps 3 and 4 of Algorithm 7, due to Lemma 5.1(i).
- Step 5 of Algorithm 6, the stopping test in dual terms, is equivalent to step 5 of Algorithm 7, due to Lemma 5.1(iii).
- Step 6 of Algorithm 6 is equivalent to step 6 of Algorithm 7, due to Lemma 5.1(ii).
- Step 7 of Algorithm 6 is equivalent to step 7 of Algorithm 7. Indeed, the descent test is exactly the same by using Lemma 5.1(iii). As for the construction of the model, in the dual version of the algorithm (Algorithm 6) it is explicitly done by using the serious step w^{k+1}, while in the primal case, Algorithm 7, it is implicitly performed by using the projection x^{k+1} in the quadratic term of the primal subproblem, and using w^{k+1} as the cost of the linear term of the objective of the subproblem.

Throughout this section, we assume the stopping tolerance is set to $\texttt{TOL} = 0$. By the stopping test in Algorithm 6, this means that when the test is triggered, it must hold that

$$0 \leqslant -\varphi^k_{w^k}(w^{k+\frac{1}{2}}) \leqslant \texttt{TOL} = 0\,,$$

by Lemma 3.4(i). With this setting, either Algorithm 6 stops after a finite number of iterations with $\varphi^k_{w^k}(w^{k+\frac{1}{2}}) = 0$, or the algorithm runs indefinitely. In this case, two more options arise: either the serious subsequence is infinite (after each serious iterate, only a finite number of null iterations occur), or a last serious iterate is generated at iteration $\hat{k} - 1$, say $\hat{w} = w^{\hat{k}}$, and afterwards all iterates are declared null steps.

5.3.1 Cases of finite termination and infinite number of serious steps

We first consider the case of a finite termination. To this aim, it is useful to characterize the subdifferential of the dual function h in terms of primal information. Specifically, given a scenario $s \in \{1, \dots, S\}$, each function $h_s(w_s)$ defined in (5.6) is the maximum of a family of affine functions of w_s, therefore it is convex [30, Chapter I, Proposition 2.1.2]. Furthermore, letting

$$C_s(w_s) = \{x_s \in C_s : -\mathcal{L}_s(x_s, w_s) = h_s(w_s)\}\,,$$

then according to [30, Chapter VI, Theorem 4.4.2],

$$\partial h_s(\bar{w}_s) = \operatorname{co}\left(\bigcup_{\bar{x}_s \in C_s(\bar{w}_s)} \partial(-\mathcal{L}_s)(\bar{x}_s, \bar{w}_s)\right).$$

Moreover, since $\mathcal{L}_s$ is differentiable with respect to w_s, then $\partial_w(-\mathcal{L}_s)(\bar{x}_s, \bar{w}_s) = \{-\bar{x}_s\}$, for any $\bar{x}_s \in C_s(\bar{w}_s)$, that is, any $\bar{x}_s \in C_s$ that solves the problem in (5.6) for $w_s = \bar{w}_s$. In other words, $C_s(\bar{w}_s)$ is exactly the set of maximizers of the problem in (5.6) for $w_s = \bar{w}_s$. Therefore,

$$\partial h_s(\bar{w}_s) = \operatorname{co}\{-\bar{x}_s : \bar{x}_s \text{ solves the problem in (5.6) for } w_s = \bar{w}_s\}. \tag{5.17}$$

Theorem 5.1 (Finite termination of BPHA)**.** *Let $h : \mathbb{R}^{nS} \to \mathbb{R}$ in (5.6) be the dual function associated with the primal problem (5.1). If Algorithm 6 stops after finitely many iterations, then x^k and $x^{k+\frac{1}{2}}$ are equal and both are a solution of primal problem (5.1), and w^k and $w^{k+\frac{1}{2}}$ are equal, and both are a solution of dual problem (5.7).*

Proof. If the algorithm stops, then $\varphi^k_{w^k}(w^{k+\frac{1}{2}}) = 0$. Thus, the aggregate error and gradient defined in Lemma 3.4 (stated in Section 3.5) are null: $E^k = 0$ and $G^k = 0$. In particular, we also have that $w^k = w^{k+\frac{1}{2}}$ and $x^k = x^{k+\frac{1}{2}}$. Then, from Lemma 3.4(ii), we have that $0 \in \partial(h + i_{\mathcal{N}^\perp})(w^k)$, that is, $w^k = w^{k+\frac{1}{2}} \in \mathcal{N}^\perp$ is a dual solution.

Furthermore, since $G^k \in \partial\varphi^k_{w^k}(w^{k+\frac{1}{2}})$, we have that $G^k - x^k \in \partial h(w^{k+\frac{1}{2}})$. Thus $-x^k \in \partial h(w^{k+\frac{1}{2}})$, because $G^k = 0$. It follows from (5.17) that for all $s = 1, \ldots, S$, $x^k_s \in C_s$ and it also solves $\min_{x_s \in C_s} \mathcal{L}_s(x_s, w^{k+\frac{1}{2}})$. Note that this means that x^k is primal feasible. Hence, since $w^{k+\frac{1}{2}}$ is a dual solution, then $x^k = x^{k+\frac{1}{2}}$ is a primal solution, as stated. □

We continue with the analysis of an infinite serious subsequence, and show the generated primal and dual points asymptotically solve the primal and dual problems (5.1) and (5.7). Recall that the dual form of the BPHA fits the framework (3.20). In this case, the proper separation of isocost surfaces (1.1) is trivially satisfied because the function $H := h + i_{\mathcal{N}^\perp}$ is convex. Regarding the error bound (1.7), the condition is equivalent to requiring that, for every $v \geqslant \inf_{w \in \mathcal{N}^\perp} h(w)$, there exists $\varepsilon, \ell > 0$, such that whenever $w \in \mathcal{N}^\perp$, $h(w) \leqslant v$, and $x \in \mathbb{R}^{nS}$, with $\|x - P_{\mathcal{N}}[x]\|_S < \varepsilon$, and the $s-$component x_s solves the problem in (5.6), there holds that

$$d(w, \mathcal{S}) \leqslant \ell \|x\|_S, \tag{5.18}$$

where $\mathcal{S}$ is the set of minimizers of the dual problem.

Theorem 5.2 (Convergence of serious steps)**.** *Let $h : \mathbb{R}^{nS} \to \mathbb{R}$ in (5.6) be the dual function associated with the primal problem (5.1). Suppose, in addition, that the subdifferential*

error bound (5.18) *is valid. In Algorithm 6, let*

$$K_{\mathtt{ser}} := \{k : w^{k+1} \text{ was declared a serious step}\}$$

and recall that, by construction $\{t_k\}_{K_{\mathtt{ser}}}$ *is bounded below by* $t_{\min} > 0$. *Assume, additionally, that* $\{t_k\}_{K_{\mathtt{ser}}}$ *is bounded above by* $t_{\max} > 0$.

If the set $K_{\mathtt{ser}}$ *is infinite, the following hold.*

(i) $\{h(w^k)\}_{K_{\mathtt{ser}}}$ *monotonically converges to the dual optimal value* h^*, *such that the sequence of functional errors* $\{v^k = h(w^k) - h^*\}_{K_{\mathtt{ser}}}$ *converges to 0 with Q-linear rate: there exists* $r \in (0,1)$, *such that for all sufficiently large* $k \in K_{\mathtt{ser}}$,

$$v^{k+1} \leqslant r v^k.$$

(ii) The sequence of serious-step iterates $\{w^k\}_{K_{\mathtt{ser}}}$, *as well as the intermediate points* $\{w^{k+\frac{1}{2}}\}_{K_{\mathtt{ser}}}$, *converge to a minimizer* w^* *of the dual problem* (5.7) *with R-linear rate: there exists* $r \in (0,1)$, *and* $c > 0$ *such that for all sufficiently large* $k \in K_{\mathtt{ser}}$,

$$\|w^k - w^*\|_S \leqslant c\sqrt{r}^k, \quad \|w^{k+\frac{1}{2}} - w^*\|_S \leqslant c(2-\sqrt{r})\sqrt{r}^k.$$

(iii) The primal sequences $\{x^k\}_{K_{\mathtt{ser}}}$ *and* $\{x^{k+\frac{1}{2}}\}_{K_{\mathtt{ser}}}$ *sub-sequentially converge to a solution of the primal problem* (5.1). *The functional values* $\{f(x^k)\}_{K_{\mathtt{ser}}}$ *and* $\{f(x^{k+\frac{1}{2}})\}_{K_{\mathtt{ser}}}$ *sub-sequentially converge to the optimal value of problem* (5.1).

Proof. Throughout, iterations parse $k \in K_{\mathtt{ser}}$. Items (i) and (ii) follow from Theorem 3.5, applied to the convex function h, which is 0−weakly convex, and $\mathcal{M} = \mathcal{N}^\perp$. For item (iii), note that from Lemma 3.4(iii), $G^k \to 0$. Therefore, the primal update in Algorithm 6 and (3.20a) imply $x^k - x^{k+\frac{1}{2}} \to 0$. Since both sequences are bounded, because $\{x^{k+\frac{1}{2}}\}$ belongs to a compact set and $P_{\mathcal{N}}$ is continuous, then both sequences $\{x^k\}$ and $\{x^{k+\frac{1}{2}}\}$ have the same accumulation points. In particular, any accumulation point x^* of these sequences belongs to $\mathcal{N}$, and for all scenario $s = 1, \ldots, S$, $x_s^* \in C_s$, from Lemma 5.1(i).

Furthermore, (3.20a) implies $G^{k_i} - x^{k_i} \in \partial h(w^{k_i+\frac{1}{2}})$, where $\{x^{k_i}\}$ is a subsequence that converges to x^*. Therefore, since ∂h is outer semicontinuous, $-x^* \in \partial h(w^*)$. In particular,

$$-f(x^{k_i+\frac{1}{2}}) - \langle w^{k_i+\frac{1}{2}}, x^{k_i+\frac{1}{2}}\rangle_S = h(w^{k_i+\frac{1}{2}}) \geqslant h(w^*) + \langle -x^*, w^{k_i+\frac{1}{2}} - w^*\rangle_S.$$

Taking the limit when $i \to +\infty$, it follows that $-f(x^*) \geqslant h(w^*)$, because $(x^*, w^*) \in \mathcal{N} \times \mathcal{N}^\perp$. Therefore, weak duality implies $-f(x^*) = h(w^*)$, with x^* being primal feasible. Hence, x^* is primal optimal.

Finally, sub-sequential convergence of $\{f(x^k)\}$ and $\{f(x^{k+\frac{1}{2}})\}$ to the optimal primal value $f(x^*)$ follows from continuity. □

To conclude, we show that when there are finitely many serious steps, the generated sequences also provide solutions for problems (5.1) and (5.7).

5.3.2 Tail of null steps

It remains to analyze the case when there is a last serious step $\widehat{w}$, performed at iteration $k = \widehat{k} - 1$. Accordingly, we let $K_{\texttt{null}} := \{k > \widehat{k} : w^k \text{ was declared a null step}\}$ denote the corresponding iteration index set. Note that for all $k \in K_{\texttt{null}}$, $w^k = \widehat{w}$. We will also assume that the stepsizes eventually stabilize along the tail of null steps. The next result shows that the last serious step is a solution of the dual problem, and that the accumulation points of the primal sequences are solutions of the primal problem.

Theorem 5.3 (Convergence of null steps)**.** *Let $h : \mathbb{R}^{nS} \to \mathbb{R}$ in (5.6) be the dual function associated with the primal problem (5.1). Assume there is a last serious step $\widehat{w}$, followed by a tail of nulls steps. If the corresponding stepsizes eventually stabilize ($t_k = \bar{t} \geqslant t_{\min}$ for $k \in K_{\texttt{null}}$ sufficiently large), the following holds for Algorithm 6.*

(i) Both $\{w^{k+\frac{1}{2}}\}_{K_{\texttt{null}}}$ and $\{u^k\}_{K_{\texttt{null}}}$ converge to $\widehat{w}$, which is also a solution for the dual problem (5.7). Furthermore, the sequences $\{h(w^{k+\frac{1}{2}})\}_{K_{\texttt{null}}}$ and $\{h(u^k)\}_{K_{\texttt{null}}}$ converge to the optimal value of the dual problem (5.7).

(ii) The primal sequences $\{x^k\}_{K_{\texttt{null}}}$ and $\{x^{k+\frac{1}{2}}\}_{K_{\texttt{null}}}$ subsequentially converge to a solution of the primal problem. The corresponding functional values $\{f(x^k)\}_{K_{\texttt{null}}}$ and $\{f(x^{k+\frac{1}{2}})\}_{K_{\texttt{null}}}$ subsequentially converge to the optimal value of problem (5.1).

Proof. Throughout, consider $k \in K_{\texttt{null}}$ for which $t_k = \bar{t}$. We claim that $\{x^k\}$ is the result of applying a projected gradient method with constant stepsize $\dfrac{1}{\bar{t}}$ to the following problem

$$\begin{cases} \min & F(x) \\ \text{s.t.} & x \in \mathcal{N} \end{cases} \quad \text{for} \quad F(x) = \min_{y \in X} \left\{ f(y) + \langle \widehat{w}, y - x \rangle_S + \frac{\bar{t}}{2} \|y - x\|^2 \right\}. \tag{5.19}$$

By convexity of f, the objective function in (5.19) is strongly convex with modulus $\dfrac{1}{\bar{t}}$, and thus F has a Lipschitz continuous gradient with modulus $\bar{t}$. In particular, when $x = x^k$, by Lemma 5.1(i),

$$x^{k+\frac{1}{2}} = \arg\min_{y \in X} \left\{ f(y) + \langle \widehat{w}, y - x^k \rangle_S + \frac{\bar{t}}{2} \|y - x^k\|_S^2 \right\},$$

and

$$F(x^k) = -\widehat{w} + \bar{t} \left(x^k - x^{k+\frac{1}{2}} \right).$$

Combined with the identity $\nabla F(x^k) = -w^{k+\frac{1}{2}}$ from (5.13), we see that

$$x^k - \frac{1}{\bar{t}}\nabla F(x^k) = x^k + \frac{1}{\bar{t}}w^{k+\frac{1}{2}} = x^{k+\frac{1}{2}} + \frac{1}{\bar{t}}\widehat{w}.$$

Projecting over $\mathcal{N}$ and recalling that $\hat{w} \in \mathcal{N}^\perp$ yields the claim, because

$$P_{\mathcal{N}}\left[x^k - \frac{1}{\bar{t}}\nabla F(x^k)\right] = P_{\mathcal{N}}[x^{k+\frac{1}{2}}] + \frac{1}{\bar{t}}P_{\mathcal{N}}[\widehat{w}] = P_{\mathcal{N}}[x^{k+\frac{1}{2}}] = x^{k+1}\,.$$

By compacteness of X, the sequence $\{x^k\}$ is bounded and the assumptions in [150, Theorem 4.1.3], stating convergence properties of gradient projected methods, are satisfied. Thus, $x^{k+1} - x^k \to 0$, and all accumulation points of $\{x^k\}$ are solutions to (5.19). Passing to limit as $k \to +\infty$ in the identity $P_{\mathcal{N}}[w^{k+\frac{1}{2}}] = \bar{t}(x^{k+1} - x^k)$ ensures that $P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \to 0$. In turn,

$$u^{k+1} - w^{k+\frac{1}{2}} = P_{\mathcal{N}^\perp}[w^{k+\frac{1}{2}}] - w^{k+\frac{1}{2}} = -P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \to 0.$$

The sequence $\{w^{k+\frac{1}{2}} = \widehat{w} + \bar{t}(x^{k+\frac{1}{2}} - x^k)\}$ is bounded and so is $\{u^{k+1}\}$. Being convex and finite, h is Lipschitz continuous in any compact set that contains u^{k+1} and $w^{k+\frac{1}{2}}$, therefore

$$|h(u^{k+1}) - h(w^{k+\frac{1}{2}})| \leqslant L\|u^{k+1} - w^{k+\frac{1}{2}}\|_S \to 0, \text{ as } k \to +\infty.$$

Non-satisfaction of (3.20c) amounts to $h(u^{k+1}) - h(\widehat{w}) > m\big(h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}}\rangle_S - h(\widehat{w})\big)$, which is equivalent to $h(u^{k+1}) - h(w^{k+\frac{1}{2}}) - m\langle x^k, w^{k+\frac{1}{2}}\rangle >_S (m-1)\big(h(w^{k+\frac{1}{2}}) - h(\widehat{w})\big)$. On the left-hand side, the inner product satisfies the relations

$$\langle x^k, w^{k+\frac{1}{2}}\rangle_S = \langle x^k, P_{\mathcal{N}}[w^{k+\frac{1}{2}}]\rangle_S \to 0, \tag{5.20}$$

because $P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \to 0$ and $\{x^k\}$ is bounded. Passing to the limit,

$$0 \geqslant (m-1)\liminf_k \left\{h(w^{k+\frac{1}{2}}) - h(\widehat{w})\right\} \iff h(\widehat{w}) \leqslant \liminf_k h(w^{k+\frac{1}{2}}), \tag{5.21}$$

because $m \in (0,1)$. On the other hand, by definition of $w^{k+\frac{1}{2}}$,

$$h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}}\rangle_S + \frac{1}{2\bar{t}}\|w^{k+\frac{1}{2}} - \widehat{w}\|_S^2 \leqslant h(\widehat{w}).$$

Passing once again to the limit, using (5.21) and (5.20),

$$h(\widehat{w}) \leqslant \liminf_k h(w^{k+\frac{1}{2}}) \leqslant \liminf_k \left\{h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}}\rangle_S + \frac{1}{2\bar{t}}\|w^{k+\frac{1}{2}} - \widehat{w}\|_S^2\right\} \leqslant h(\widehat{w}).$$

Take any accumulation point $w^\star$ of the bounded $\{w^{k+\frac{1}{2}}\}$ and consider a subsequence $w^{k_j+\frac{1}{2}} \to w^\star$, whenever $j \to +\infty$. Then $h(w^\star) + \frac{1}{2\bar{t}}\|w^\star - \widehat{w}\|_S^2 = h(\widehat{w})$. Moreover, from (5.21) it also holds that $h(\widehat{w}) \leqslant h(w^\star)$, and thus $h(w^\star) + \frac{1}{2\bar{t}}\|w^\star - \widehat{w}\|_S^2 \leqslant h(w^\star)$, which necessarily implies $w^\star = \widehat{w}$. This means that any accumulation point of $\{w^{k+\frac{1}{2}}\}$ is equal

to the last serious step $\widehat{w}$, and thus $\{w^{k+\frac{1}{2}}\}$ converges to $\widehat{w}$, and since $u^{k+1} - w^{k+\frac{1}{2}} \to 0$, then $\{u^k\}$ also converges to $\widehat{w}$.

To prove that $\widehat{w}$ solves the dual problem, note that from (5.14) and (5.13),

$$G^k = x^{k+\frac{1}{2}} - x^k = \frac{1}{t}(w^{k+\frac{1}{2}} - \widehat{w}) \to 0, \text{ as } k \to +\infty.$$

By continuity of h and (5.20), $\varphi^k_{\widehat{w}}(w^{k+\frac{1}{2}}) \to 0$. Lemma 3.4(i) ensures that $E^k \to 0$ and Lemma 3.4(ii) yields in the limit that $0 \in \partial(h + i_{\mathcal{N}^\perp})(\widehat{w})$, which means that $\widehat{w}$ minimizes $h + i_{\mathcal{N}^\perp}$.

As for the primal problem, since $x^{k+\frac{1}{2}} - x^k \to 0$, both sequences $\{x^k\}$ and $\{x^{k+\frac{1}{2}}\}$ have the same accumulation points. The remaining assertions follow as in the proof of Theorem 5.2(iii), now taking $k \in K_{\texttt{null}}$.

□

5.4 Final remarks

We have introduced a new projective bundle method that, when applied to multistage programs, exploits parallelism and generates a serious subsequence converging with linear rate. The resulting Bundle Progressive Hedging, both in primal (Algorithm 7) and dual (Algorithm 6) forms, preserves the main features of the Progressive Hedging algorithm by T. Rockafellar and R. Wets, [24]. In particular, both methods solve separate scenario subproblems per iteration. These subproblems are strongly convex, thanks to the addition of a quadratic term related to an approximate augmented Lagrangian of the problem. Furthermore, the PHA and its bundle BPHA variant both use projections onto $\mathcal{N}$ and $\mathcal{N}^\perp$ to respectively ensure feasibility in the primal and dual spaces.

Besides these similarities, our proposal adds features typical from the bundle methodology to measure the quality of the approximation of the augmented Lagrangian of the full problem. By contrast, the original Progressive Hedging method uses the dual information obtained with an approximate Lagrangian without further ado.

Thanks to the bundle techniques, it is possible to dynamically adjust the augmentation parameter in the PHA without impairing its convergence. Figure 9 gives a simple, yet illustrative, instance of a randomly generated linear problem with 50 scenarios. The impact of keeping t_k fixed along the iterative procedure is clear. With $t_k = t_0$, the PHA approach only reaches a good accuracy if the parameter is sufficiently large ($t_0 = 100$, named "very large t" in the figure). The Bundle Progressive Hedging method, with its adaptive adjustment of stepsizes, seems less sensitive to the initial value of t_0.

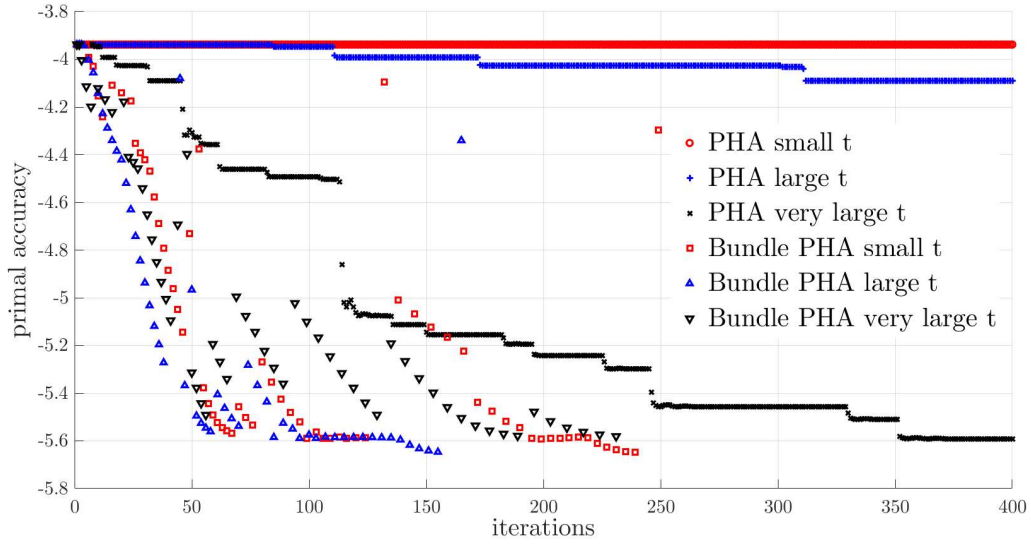


Figure 9 – Accuracy in the primal solution for the PHA and the BPHA (the ordinate reports the negative of the number of digits obtained at a given iteration). Both methods were run for the same random instance with 50 scenarios, taking initial proximal stepsizes $t_0 \in \{1, 10, 100\}$, being small t, large t, and very large t, respectively. The values of t_k are maintained fixed to t_0 with the PHA variant, yielding good accuracy only for the largest value $t_0 = 100$. With BPHA, t_k varies according to the serious/null step rules and the performance is more stable.

With the bundle approach, optimality primal and dual certificates are available, based on the aggregate information constructed along iterations (the PHA, by contrast, measures distance to the primal-dual set of minimizers). However, those enhancements require additional evaluations of the dual functions at each iteration. This involves a new set of computations, that can be done in parallel and are similar to the PHA subproblems, barring the quadratic term in (5.8).

The projective bundle method in Section 3.5 is general and can be used to extend existing approaches to the weakly convex setting. As often in bundle methods, the crucial point is the ability to show that, when the method enters an infinite loop of null steps, the family of model functions drives iterates to the last generated serious point. The stumbling block lies in the suitable definition of the models, as it was done in [45] for Taylor models in composite optimization, or in Section 5.2 for the particular dual function (5.6).

6 A dual-embedded forward-backward scenario decomposition method for convex stochastic programming

Abstract. We investigate convergence properties of new decomposition methods for large-scale optimization problems, combining a generalized approximate proximal point algorithm with forward-backward steps for a dual version of the augmented Lagrangian subproblem that allows decomposition for different scenarios of the random variable of a stochastic optimization problem. The resulting analysis can be compared with convergence rate frameworks for descent methods and operator splitting from previous chapters. In particular, we study convergence guarantees for a variant of the progressive hedging algorithm, and also conditions to obtain linear and superlinear convergence with varying stepsizes.

Introduction

A considerable amount of problems of interest involve the minimization of the sum of compositions of a convex function and a linear operator, including applications in machine learning and industrial applications of stochastic programming. The functions utilized in such problems may not necessarily be differentiable, and the number of variables involved might be exceedingly large. A variety of methods referred to as *splitting methods* have been developed to solve these problems; see [140, 11, 87, 137, 6] for some founding contributions and examples. These methods solve a sequence of simpler subproblems, each one of them related to a single term from the sum comprising the original problem formulation. An important family of such methods is made up of *proximal splitting methods*, based on the proximal point algorithm [17], including Douglas-Rachford splitting [140] of Section 2.2.1 and Chapter 4, and its common special case, the alternating direction method of multipliers (ADMM) [151, 152]. This family of algorithms has spawned numerous variants that exploit specific structural attributes of different types of problems.

However, operator splitting methods exhibit some practical limitations. Specifically, it has been empirically and theoretically observed that they typically demonstrate slow rates of linear convergence. Despite this shortcoming, these methods have gained widespread popularity in recent decades, primarily due to the escalating demand to tackle

large-scale problems for which low-accuracy solutions are deemed sufficient. To enhance their performance and broaden their applicability, improving these methods' tail convergence could potentially yield significant practical advantages. The importance of better convergence rates for such methods would be reflected by higher numerical accuracy, fewer subproblems to solve to achieve a specified accuracy, and more potential applications.

The algorithm proposed in this chapter is strongly related to the classical augmented Lagrangian method, but adds two new features: a relative-error criterion to either accept or reject a candidate subproblem solution, and also a systematic way to construct such candidate solutions using forward-backward steps, even when the original problem would not appear amenable to forward-backward algorithms. As a by-product, we propose a scenario-decomposition method for convex stochastic programming, resembling the PH of Section 2.2.2, and actually corresponding variant of the bundle-like progressive hedging algorithm of Chapter 5. We also prove that our method enjoys local linear rates of convergence, and also superlinear rates when the stepsizes increase with no bound and the error tolerance is driven to 0. The new method we have developed has the potential to be applied in a more general context, namely minimizing an extended-real-valued function constrained to a linear subspace. We first describe the method in this general setting.

6.1 Embedded forward-backward method applied to dual-type problems

We closely follow the notation of Chapter 5. In particular, for a finite-valued function $h \in \overline{\text{conv}}(\mathbb{R}^n)$, and a linear subspace $\mathcal{M} \subseteq \mathbb{R}^n$, consider the following constrained minimization problem

$$\begin{cases} \min\limits_{w} & h(w) \\ \text{s.t.} & w \in \mathcal{M}. \end{cases} \tag{6.1}$$

We assume that problem (6.1) has a nonempty set of solutions, denoted by $W^\star$.

Note that (6.1) corresponds to the dual problem (5.7) in the stochastic optimization setting. More on this observation is discussed in Section 6.2.

The idea of our method is to iteratively solve problems involving surrogate models of $h + i_{\mathcal{M}}$. The original idea can be found in Chapter 5, used to develop the bundle-like progressive hedging algorithm. The "approximations" of $h+i_{\mathcal{M}}$ take the simple form

$$h^k(w) = h(w) + \langle x^k, w \rangle_S, \tag{6.2}$$

where $x^k \in \mathcal{M}^\perp$ is computed in each iteration of the algorithm. Note that $h^k : \mathbb{R}^n \to \mathbb{R}$

shares the separability attributes of h (cf. equation (5.6)):

$$h^k(w) = \sum_{s=1}^{S} p_s h_s^k(w_s), \text{ where } h_s^k(w_s) = h_s(w_s) + \langle x_s^k, w_s \rangle.$$

Moreover, it is important to observe that the linear map $w \mapsto \langle x^k, w \rangle_s$ is a loose approximation of the indicator $i_{\mathcal{M}}$: since $x^k \in \mathcal{M}^\perp$, then for any $w \in \mathbb{R}^n$,

$$\langle x^k, w \rangle_S = \langle x^k, P_{\mathcal{M}}[w] + P_{\mathcal{M}^\perp}[w] \rangle_S = \langle x^k, P_{\mathcal{M}^\perp}[w] \rangle_S.$$

If $w \in \mathcal{M}$, then $\langle x^k, w \rangle_S = 0 = i_{\mathcal{M}}(w)$, but if $w \in \mathcal{M}^\perp$, $\langle x^k, w \rangle_S < +\infty = i_{\mathcal{M}}(w)$. In this way, roughly speaking, the close w is to $\mathcal{M}$, the better $w \mapsto \langle x^k, w \rangle_S$ approximates $i_{\mathcal{M}}$, which is the case of our interest for problem (6.1).

Using this family of models, we propose Algorithm 8, which we call the Dual Embedded Forward-Backward (DEFB) method. Each iteration of this method solves a proximal subproblem for the model h^k, and then performs a projection step, like the Bundle Progressive Hedging method of Chapter 5 does. To assess the quality of the subproblem solution, a relative-error condition evaluates model accuracy and the violation of the subspace constraint, instead of the bundle-like descent condition (3.20c) of the model-based scheme of Section 3.5. At the end of each iteration, a new model is constructed following a projected-gradient rule.

For notational simplicity and to emphasize its resemblance to bundle methods, the algorithm is presented as having a single loop, indexed by k. However, it can also be viewed as having two nested loops, with (6.5) being the termination condition for the inner loop. The inner loops consist of the steps (6.3), $u^{k+1} = P_{\mathcal{M}}[w^{k+1/2}]$, and (6.4)-(6.7), until condition (6.5) holds. At this point, the proximal center w^{k+1} is updated and an iteration of the outer loop can be considered to occur.

The following result summarizes properties that Algorithm 8 share with Algorithm 6 and Algorithm 7, and some other new features.

Proposition 6.1. *The following holds for Algorithm 8.*

(i) The convex model function h^k defined in (6.2) is a lower estimate of $h + i_{\mathcal{M}}$.

(ii) The error defined in (6.4) satisfies $\varepsilon^k \geqslant 0$.

(iii) There exists $\{x^{k+\frac{1}{2}}\}$ such that for all $k \in \mathbb{N}$, $-x^{k+\frac{1}{2}} \in \partial h(w^{k+\frac{1}{2}})$, and

$$w^{k+\frac{1}{2}} = w^k + t_k(x^{k+\frac{1}{2}} - x^k). \tag{6.8}$$

Proof. First, item (i) can be similarly proven as Proposition 5.1, since it is the same type of model, and here we provide an alternative proof. From line 4 of Algorithm 8,

Algorithm 8 Dual Embedded Forward-Backward method

1: **Initialization**: choose $w^0 \in \mathbb{R}^n$, $x^0 \in \mathcal{M}^\perp$ and $\sigma^0 \in [0,1)$, $t_{\min} > 0$, $t_0 \geqslant t_{\min}$, and a stopping tolerance $\texttt{TOL} > 0$.
2: **for** $k = 0, 1, \ldots$ **do**
3: **Subproblem:** for each $s = 1, \ldots, S$, compute $w_s^{k+\frac{1}{2}}$ as the unique solution of

$$\min_{w_s \in \mathbb{R}^{n_s}} \left\{ h_s^k(w_s) + \frac{1}{2t_k}\|w_s - w_s^k\|^2 \right\}, \tag{6.3}$$

and define $w^{k+\frac{1}{2}} = (w_s^{k+\frac{1}{2}})_{s=1}^S$.
4: **Projection:** define $u^{k+1} = P_{\mathcal{M}}[w^{k+\frac{1}{2}}]$.
5: **Acceptance test:** Define

$$\varepsilon_k = h(u^{k+1}) - h^k(w^{k+\frac{1}{2}}) - \frac{1}{t_k}\langle w^k - w^{k+\frac{1}{2}}, u^{k+1} - w^{k+\frac{1}{2}}\rangle_S. \tag{6.4}$$

If

$$\|w^{k+\frac{1}{2}} - u^{k+1}\|_S^2 + 2t_k\varepsilon_k \leqslant \sigma_k^2\|u^{k+1} - w^k\|_S^2, \tag{6.5}$$

execute an **outer step**: that is, let $w^{k+1} = u^{k+1}$ and choose $t_{k+1} \geqslant t_{\min}$ and $\sigma_{k+1} \in [0,1)$. Otherwise, execute an **inner step**: set $w^{k+1} = w^k$, $t_{k+1} = t_k$, and $\sigma_{k+1} = \sigma_k$.
6: **Stopping test**: if $\|w^{k+\frac{1}{2}} - w^k\|_S \leqslant t_k\texttt{TOL}$ and $\varepsilon_k \leqslant \texttt{TOL}$, stop and return u^{k+1}.
7: **Model update:** define

$$x^{k+1} = x^k + \frac{1}{t_k}(w^{k+\frac{1}{2}} - u^{k+1}), \tag{6.6}$$

and define the function

$$h^{k+1} : w \mapsto h(w) + \langle x^{k+1}, w\rangle_S. \tag{6.7}$$

8: **end for**

$w^{k+\frac{1}{2}} - u^{k+1} = P_{\mathcal{M}^\perp}[w^{k+\frac{1}{2}}] \in \mathcal{M}^\perp$, and since $x^0 \in \mathcal{M}^\perp$, then $x^k \in \mathcal{M}^\perp$ in view of (6.6). Thus

$$i_{\mathcal{M}}(\bar{w}) + \langle x^k, w - \bar{w}\rangle_S,$$

is a lower linearization of $i_{\mathcal{M}}$ at any $\bar{w} \in \mathcal{M}$, because $\mathcal{M}^\perp = N_{\mathcal{M}}(\bar{w}) = \partial i_{\mathcal{M}}(\bar{w})$. Furthermore, due to orthogonality, $x^k \in \mathcal{M}^\perp$ and $\bar{w} \in \mathcal{M}$ imply

$$i_{\mathcal{M}}(\bar{w}) + \langle x^k, w - \bar{w}\rangle_S = \langle x^k, w\rangle_S,$$

and thus $w \mapsto \langle x^k, w\rangle_S$ is a lower subgradient estimate of $i_{\mathcal{M}}$ at any $\bar{w} \in \mathcal{M}$. To prove (ii), note that the optimality condition of (6.3) reads

$$0 \in \partial h^k(w^{k+\frac{1}{2}}) + \frac{1}{t_k}(w^{k+\frac{1}{2}} - w^k). \tag{6.9}$$

Therefore, there exists $g^k \in \partial h^k(w^{k+\frac{1}{2}})$, such that $0 = t_k g^k + w^{k+\frac{1}{2}} - w^k$, or equivalently

$$w^{k+\frac{1}{2}} = w^k - t_k g^k. \tag{6.10}$$

The subgradient inequality for g^k evaluated at $w = u^{k+1}$ gives

$$h^k(u^{k+1}) - \left(h^k(w^{k+\frac{1}{2}}) + \langle g^k, u^{k+1} - w^{k+\frac{1}{2}} \rangle_S\right) \geqslant 0.$$

Since $u^{k+1} \in \mathcal{M}$, then $h^k(u^{k+1}) = h(u^{k+1})$, and thus it follows from (6.4) and (6.10) that $\varepsilon^k \geqslant 0$. Finally, since $g^k \in \partial h^k(w^{k+\frac{1}{2}})$, there exists $x^{k+\frac{1}{2}} \in \mathbb{R}^n$ such that $-x^{k+\frac{1}{2}} \in \partial h(w^{k+\frac{1}{2}})$, and $g^k = -x^{k+\frac{1}{2}} + x^k$. Thus, (6.9) implies item (iii) holds.

□

Remark 6.1. *Proposition 6.1 gathers common properties of the DEFB method and the BPHA: as mentioned in the proof, Proposition 6.1(i) is Proposition 5.1, Proposition 6.1(ii) is analogous to Lemma 3.4(i), and Proposition 6.1(iii) corresponds to* (3.20a) *and* (5.14).

A key feature of Algorithm 8 is that the model update is equivalent to a forward-backward step, similarly as in the proof of Theorem 5.3 of Section 5.3.2, since the model is updated using the same rule of the BPHA: see (5.13) of Algorithm 6, and line 4 of Algorithm 7. More precisely, a direct application of the proximal point algorithm to problem (6.1) amounts to solve

$$\begin{aligned}
&\min_{w \in \mathbb{R}^n} \left\{ h(w) + i_{\mathcal{M}}(w) + \frac{1}{2t_k}\|w - w^k\|_S^2 \right\} \\
&= \min_{w \in \mathbb{R}^n} \left\{ \left(h(w) + \frac{1}{2t_k}\|w - w^k\|_S^2 \right) + i_{\mathcal{M}}(w) \right\},
\end{aligned}$$

where the second line is obtained by regrouping terms. The Fenchel dual of the problem in the last line can be written as

$$\begin{aligned}
&\min_{x \in \mathbb{R}^n} \left\{ \left(h(\cdot) + \frac{1}{2t_k}\|\cdot - w^k\|_S^2 \right)^* (-x) + i^*_{\mathcal{M}}(x) \right\} \\
&= \min_{x \in \mathbb{R}^n} \left\{ \left(h(\cdot) + \frac{1}{2t_k}\|\cdot - w^k\|_S^2 \right)^* (-x) + i_{\mathcal{M}^\perp}(x) \right\},
\end{aligned}$$

where we use the fact that $i^*_{\mathcal{M}} = i_{\mathcal{M}^\perp}$, since $\mathcal{M}$ is a linear subspace. Therefore, the Fenchel dual problem is equivalent to solving

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} & \left(h(\cdot) + \frac{1}{2t_k}\|\cdot - w^k\|_S^2 \right)^* (-x) \\ \text{s.t.} & x \in \mathcal{M}^\perp. \end{cases} \tag{6.11}$$

The following result states that the update rule of the slopes x^k in Algorithm 8 is actually a step of the forward-backward method (also known as proximal-gradient) applied to problem (6.11).

Proposition 6.2. *The slope x^{k+1} of Algorithm 8 defined in (6.6) is the result of applying one step of the forward-backward method to (6.11), using t_k as stepsize and x^k as starting point.*

Proof. Observe that the objective function of this problem is the convex conjugate of

$$h(\cdot) + \frac{1}{2t_k}\|\cdot - w^k\|_S^2,$$

and thus differentiable with t_k-Lipschitz gradient, since the argument function is strongly convex. Starting from x^k, applying the forward-backward method stepsize t_k to (6.11) yields

$$\begin{aligned} x^+ &= \operatorname{prox}_{i_{\mathcal{M}^\perp}}\left[x^k - \frac{1}{t_k}\left(-\nabla\left(h + \frac{1}{2t_k}\|\cdot - w^k\|_S^2\right)^*(-x^k)\right)\right] \\ &= P_{\mathcal{M}^\perp}\left[x^k - \frac{1}{t_k}\left(-\nabla\left(h + \frac{1}{2t_k}\|\cdot - w^k\|_S^2\right)^*(-x^k)\right)\right], \end{aligned}$$

where

$$\begin{aligned} \nabla\left(h + \frac{1}{2t_k}\|\cdot - w^k\|_S^2\right)^*(-x^k) &= \arg\max_w \left\{\langle -x^k, w\rangle_S - h(w) - \frac{1}{2t_k}\|w - w^k\|_S^2\right\} \\ &= \arg\min_w \left\{h(w) + \langle x^k, w\rangle_S + \frac{1}{2t_k}\|w - w^k\|_S^2\right\} \\ &= \left\{w^{k+\frac{1}{2}}\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} x^+ &= P_{\mathcal{M}^\perp}\left[x^k + \frac{1}{t_k}w^{k+\frac{1}{2}}\right] \\ &= x^k + \frac{1}{t_k}P_{\mathcal{M}^\perp}\left[w^{k+\frac{1}{2}}\right] \\ &= x^k + \frac{1}{t_k}\left(w^{k+\frac{1}{2}} - u^{k+1}\right), \end{aligned}$$

where in the second equality we inductively assume that $x^k \in \mathcal{M}^\perp$, and use the linearity of the projection operator $P_{\mathcal{M}^\perp}$, and in the last equality we use the Moreau identity for projections. This proves that the model update for x^{k+1} in Algorithm 8 is a projected-gradient step applied to the Fenchel dual of the proximal point objective function associated with $h + i_{\mathcal{M}}$. □

Algorithm 8 constructs in each iteration an approximate subgradient of the function $h + i_{\mathcal{M}}$, that helps to define the stopping test of the algorithm. This observation is based in the following result.

Lemma 6.1. *Consider $\{g^k\}$ defined in (6.10). For any k, $g^k = \frac{1}{t_k}(w^k - w^{k+\frac{1}{2}})$, and $g^k \in \partial_{\varepsilon^k}(h + i_{\mathcal{M}})(u^{k+1})$.*

Proof. The first identity follows from (6.10). Moreover, for all $w \in \mathbb{R}^n$

$$\begin{aligned}(h + i_{\mathcal{M}})(w) &\geqslant h^k(w) \\ &\geqslant h^k(w^{k+\frac{1}{2}}) + \langle g^k, w - w^{k+\frac{1}{2}} \rangle \\ &= h(u^{k+1}) + \langle g^k, w - u^{k+1} \rangle - (h(u^{k+1}) - h^k(w^{k+\frac{1}{2}}) - \langle g^k, u^{k+1} - w^{k+\frac{1}{2}} \rangle) \\ &= h(u^{k+1}) + \langle g^k, w - u^{k+1} \rangle - \varepsilon_k,\end{aligned}$$

where in the first inequality we use that h^k is a lower model of $h + i_{\mathcal{M}}$, the second line follows from the subgradient inequality for $g^k \in \partial h^k(w^{k+\frac{1}{2}})$, and the last line is obtained from the definition of ε_k in (6.4). It follows that g^k is an ε_k subgradient of $h + i_{\mathcal{M}}$ at u^{k+1}.

□

The precise form of problem (6.11) at iteration $k+1$ depends on the type of step performed in the previous iteration. This divides the convergence analysis of Algorithm 8 in two cases, determined by the two types of steps:

- An inner step: the solution $w^{k+\frac{1}{2}}$ of subproblem (6.3) does not satisfy the relative-error condition (6.5), which means that the point is far from being dual feasible or that the model is not accurate enough at the projection of the calculated solution. The center w^k is not updated and the model of $i_{\mathcal{M}}$ is improved.
- An outer step: a solution $w^{k+\frac{1}{2}}$ of subproblem (6.3) that satisfies the relative-error condition (6.5) is found, therefore a new center w^{k+1} is defined as the projection of $w^{k+\frac{1}{2}}$ onto $\mathcal{M}$. The linear model of $i_{\mathcal{M}}$ is also updated.

Note the resemblance with the serious/null steps analysis of bundle methods, in particular, of the method in Chapter 5.

6.1.1 Convergence of the dual-embedded forward-backward method

In this section, we prove the convergence of Algorithm 8 in three different cases. We commence by examining the finite termination case, in order to establish that the output of the method is a solution of the problem when the tolerance is null, and that it is an approximate solution whenever the tolerance is positive. We then continue with the case of a tail of inner steps, where the vital idea is that we keep the parameters fixed and thus convergence follows from the theory of forward-backward methods. We finally proceed to investigate the convergence of the outer loop iterates by using a relative-error approximate proximal point algorithm.

6.1.1.1 Finite termination case

We first address the case of finite termination of Algorithm 8, that is, when at some iteration k,

$$\|w^{k+\frac{1}{2}} - w^k\| \leqslant t_k \texttt{TOL}, \quad \text{and} \quad \varepsilon_k \leqslant \texttt{TOL}.$$

This stopping test serves as an approximate optimality certificate, similarly to the case of the bundle-like progressive hedging algorithm of Chapter 5. Using Lemma 6.1, the stopping test is equivalent to

$$\|g^k\| \leqslant \texttt{TOL}, \ \varepsilon_k \leqslant \texttt{TOL}, \ g^k \in \partial_{\varepsilon^k}(h + i_{\mathcal{M}})(u^{k+1}).$$

If $\texttt{TOL} = 0$, then u^{k+1} is an exact global solution to problem (6.1), because in this case $0 \in \partial(h + i_{\mathcal{M}})(u^{k+1})$. Otherwise, if $\texttt{TOL} > 0$, then u^{k+1} is an approximate solution to (6.1), similarly to (5.15) for the bundle-like progressive hedging algorithm of Chapter 5. More specifically, the inequality of the proof of Lemma 6.1 implies for all $w \in \mathcal{M}$

$$h(w) \geqslant h(u^{k+1}) - \texttt{TOL}\|w - u^{k+1}\| - \texttt{TOL}.$$

6.1.1.2 Convergence: infinite loop of inner steps

In this section, we consider the case in which the outer loop is executed finitely many times, meaning that there exists $\widehat{k} \in \mathbb{N}$ such that for all $k > \widehat{k}$, the acceptance test (6.5) is not satisfied, and the last time it is satisfied is in iteration $k = \widehat{k}$. Therefore, the center, the stepsize, and the error tolerance are not updated: for all $k > \widehat{k}$, one has $w^k = \widehat{w} := w^{\widehat{k}}$, $t_k = \widehat{t} := t_{\widehat{k}}$, and $\sigma_k = \widehat{\sigma} := \sigma_{\widehat{k}}$. In this situation, one has for all $k > \widehat{k}$:

$$\left\|w^{k+\frac{1}{2}} - u^{k+1}\right\|^2 + 2\widehat{t}\varepsilon_k > \widehat{\sigma}^2\|u^{k+1} - \widehat{w}\|^2. \tag{6.12}$$

Furthermore, problem (6.11) takes the following fixed form

$$\begin{cases} \min\limits_{x} & \left(h(\cdot) + \dfrac{1}{2\widehat{t}}\|\cdot - \widehat{w}\|^2\right)^*(-x) \\ \text{s.t.} & x \in \mathcal{M}^\perp, \end{cases} \tag{6.13}$$

The following result presents the convergence results of the inner loop of Algorithm 8 (cf. Theorem 5.3).

Theorem 6.1. *Suppose there exists a last outer step in Algorithm 8 at iteration $k = \widehat{k}$. Then*

(i) The sequence of slopes $\{x^k\}_{k>\widehat{k}}$ converges to a solution $x^\star$ of problem (6.13).

(ii) The sequence $\{w^{k+\frac{1}{2}}\}_{k>\widehat{k}}$ is asymptotically feasible, that is, $\{P_{\mathcal{M}^\perp}[w^{k+\frac{1}{2}}]\}_{k>\widehat{k}}$ converges to 0.

Proof. As shown in Proposition 6.2, x^{k+1} can be viewed as one forward-backward step performed for problem (6.13). Under the assumption of a tail of inner steps of Algorithm 8, the function for which the forward-backward step is performed does not vary throughout iterations, and thus [93, Theorem 10.24] implies that item (i) follows. Moreover, from (6.6), there holds

$$\begin{aligned} P_{\mathcal{M}^\perp}[w^{k+\frac{1}{2}}] &= P_{\mathcal{M}^\perp}[u^{k+1} + \widehat{t}(x^{k+1} - x^k)] \\ &= P_{\mathcal{M}^\perp}[u^{k+1}] + \widehat{t}P_{\mathcal{M}^\perp}[x^{k+1} - x^k] \\ &= \widehat{t}P_{\mathcal{M}^\perp}[x^{k+1} - x^k] \\ &= \widehat{t}(x^{k+1} - x^k), \end{aligned} \tag{6.14}$$

where in the second line we use the linearity of the projection operator, in the third line the fact that $u^{k+1} \in \mathcal{M}$, and in the last line $x^{k+1} - x^k \in \mathcal{M}$, since $x^k \in \mathcal{M}$ for all k and $\mathcal{M}$ is a linear subspace. Thus, in view of item (i), $x^{k+1} - x^k \to 0$, and item (ii) follows. □

When problem (6.1) corresponds to the dual of a problem, we will show in the next section that $\{x^k\}$ actually converges to a solution of the primal problem. Regardless, in the general case, $\{w^{k+\frac{1}{2}}\}$ not only is asymptotically feasible, but also converges to a solution to (6.1). The following result closely follows Theorem 5.3 of Chapter 5, and we give a proof for completeness. We need a mild extra assumption, and in the following section we explain why it does not pose a restriction in the analysis.

Theorem 6.2 (Convergence to a solution – inner loop)**.** *Consider problem* (6.1)*, such that the conjugate h^* of the objective function has compact domain. Suppose there exists a last outer step in Algorithm 8 at iteration $k = \widehat{k}$. Then both $\{u^k\}_{k>\widehat{k}}$ and $\{w^{k+\frac{1}{2}}\}_{k>\widehat{k}}$ converge to the last center $\widehat{w}$, which in turn is a solution of* (6.1)*.*

Proof. Note that when the inner loop executes indefinitely, equation (6.8) reads

$$w^{k+\frac{1}{2}} = \widehat{w} + \widehat{t}(x^{k+\frac{1}{2}} - x^k). \tag{6.15}$$

Due to Proposition 6.1(iii), $-x^{k+\frac{1}{2}} \in \operatorname{dom}(h^*)$, and thus the assumption on h^* implies $\{x^{k+\frac{1}{2}}\}$ is bounded. Therefore, in view of (6.15) and Theorem 6.1(i), $\{w^{k+\frac{1}{2}}\}$ is also a bounded sequence, and so is $\{P_{\mathcal{M}}[w^{k+\frac{1}{2}}]\}$. Since h convex and finite, then it is locally Lipschitz continuous. Hence, h is Lipschitz continuous on a compact set that contains $w^{k+\frac{1}{2}}$ and $P_{\mathcal{M}}[w^{k+\frac{1}{2}}]$. Therefore, there exists a constant $L > 0$ such that

$$|h(P_{\mathcal{M}}[w^{k+\frac{1}{2}}]) - h(w^{k+\frac{1}{2}})| \leqslant L\|P_{\mathcal{M}}[w^{k+\frac{1}{2}}] - w^{k+\frac{1}{2}}\|_S,$$

which implies $h(P_{\mathcal{M}}[w^{k+\frac{1}{2}}]) - h(w^{k+\frac{1}{2}}) \to 0$, by using Theorem 6.1(ii), and the identity $P_{\mathcal{M}^\perp}[w^{k+\frac{1}{2}}] = w^{k+\frac{1}{2}} - P_{\mathcal{M}}[w^{k+\frac{1}{2}}]$.

Furthermore, using again Theorem 6.1, $\mathcal{M}^\perp \ni x^k \to x^\star$ and $P_{\mathcal{M}^\perp}\left[w^{k+\frac{1}{2}}\right] \to 0$ imply $\langle x^k, w^{k+\frac{1}{2}}\rangle_S = \langle x^k, P_{\mathcal{M}^\perp}[w^{k+\frac{1}{2}}]\rangle_S \to 0$. Therefore,

$$\varepsilon^k = h(P_{\mathcal{M}}[w^{k+\frac{1}{2}}]) - h(w^{k+\frac{1}{2}}) - \langle x^k, w^{k+\frac{1}{2}}\rangle_S - \frac{1}{\widehat{t}}\langle \widehat{w} - w^{k+\frac{1}{2}}, P_{\mathcal{M}}[w^{k+\frac{1}{2}}] - w^{k+\frac{1}{2}}\rangle_S \to 0,$$

where in the right-most inner product we use that $\{w^{k+\frac{1}{2}}\}$ is bounded. In this way, the left-hand side of (6.12) converges to 0, thus $P_{\mathcal{M}}[w^{k+\frac{1}{2}}] = u^{k+1} \to \widehat{w}$, and

$$w^{k+\frac{1}{2}} = w^{k+\frac{1}{2}} - P_{\mathcal{M}}[w^{k+\frac{1}{2}}] + u^{k+1} = P_{\mathcal{M}^\perp}[w^{k+\frac{1}{2}}] + u^{k+1} \to \widehat{w}.$$

Finally, by definition, $w^{k+\frac{1}{2}}$ minimizes $h(\cdot) + \langle x^k, \cdot\rangle_S + \frac{1}{2\widehat{t}}\|\cdot - \widehat{w}\|_S^2$. Then, for all $w \in \mathcal{M}$, since $x^k \in \mathcal{M}^\perp$

$$h(w) + \frac{1}{2\widehat{t}}\|w - \widehat{w}\|_S^2 \geqslant h(w^{k+\frac{1}{2}}) + \langle x^k, w^{k+\frac{1}{2}}\rangle_S + \frac{1}{2\widehat{t}}\|w^{k+\frac{1}{2}} - \widehat{w}\|_S^2.$$

Taking the limit as $k \to +\infty$, there holds for all $w \in \mathcal{M}$, as h is continuous and $\widehat{w} \in \mathcal{M}$:

$$(h + i_{\mathcal{M}})(w) + \frac{1}{2\widehat{t}}\|w - \widehat{w}\|^2 \geqslant (h + i_{\mathcal{M}})(\widehat{w}),$$

which means that $\widehat{w} = \operatorname{prox}_{\widehat{t}(h+i_{\mathcal{M}})}(\widehat{w})$. Hence, $\widehat{w}$ is a global minimizer of $h + i_{\mathcal{M}}$. □

6.1.1.3 Convergence: infinite loop of outer steps

In this subsection, we assume the outer loop is executed an infinite number of times, meaning that the acceptance criterion (6.5) fails to hold only a finite number of tries after each outer step. In other words, the algorithm sets $w^{k+1} = u^{k+1}$ infinitely many times. In this situation, $\{w^k\}$ is an infinite sequence, and thus we consider the infinite set of indices

$$K_O = \{k \in \mathbb{N} : u^{k+1} \text{ satisfies (6.5)}\}.$$

For all $k \in K_O$, $u^{k+1} = w^{k+1}$.

The convergence analysis of the outer loop is an application of the inexact proximal point method for finding zeros of a maximal monotone set-valued operator presented in [25], corresponding to Algorithm 1 described in Section 2.1.1.

The following result defines the sense in which the approximate proximal steps are performed in each iteration of the outer loop, in terms of the analysis in [25]. Loosely speaking, the outer loops of Algorithm 8 correspond to approximately solve the inclusion

$$0 \in \partial(h + i_{\mathcal{M}})(w) + \frac{1}{t_k}(w - w^k),$$

which corresponds to applying the proximal point algorithm to the maximal monotone operator $T = \partial(h + i_{\mathcal{M}})$ with a relative-error criterion, since we seek to minimize $h + i_{\mathcal{M}}$.

Proposition 6.3. *Consider u^{k+1} defined in line 4 of Algorithm 8, and g^k defined in Lemma 6.1. For $k \in K_O$, the pair (u^{k+1}, g^k) is an approximate solution with error tolerance σ_k of $0 \in t_k \partial h(\cdot) + (\cdot - w^k)$, in the sense that for $\varepsilon^k \geqslant 0$,*

(i) $g^k \in \partial_{\varepsilon^k}(h + i_{\mathcal{M}})(u^{k+1})$,

(ii) there exists $r_k \in \mathbb{R}^n$, *such that* $t_k g^k + u^{k+1} - w^k = r^k$, *and*

(iii) $\|r^k\|^2 + 2t_k\varepsilon^k \leqslant \sigma_k^2 \|u^{k+1} - w^k\|^2$.

Proof. First, item (i) corresponds to Lemma 6.1. Moreover, applying the projection operator $P_{\mathcal{M}}$ to both sides of (6.10), it follows

$$\begin{aligned} u^{k+1} &= P_{\mathcal{M}}[w^k - t_k g^k] \\ &= P_{\mathcal{M}}[w^k] - t_k P_{\mathcal{M}}[g^k] \\ &= w^k - t_k P_{\mathcal{M}}[g^k] \\ &= w^k - t_k g^k + r^k, \end{aligned}$$

where in the second equality we use the linearity of the projection operator, in the third equality we use the fact $w^k \in \mathcal{M}$, and in the last equality we use the identity $g^k = P_{\mathcal{M}}[g^k] + P_{\mathcal{M}^\perp}[g^k]$, and define $r^k = t_k P_{\mathcal{M}^\perp}[g^k]$.

The third of the relations follows directly from [25, Definition 2.1], using the relations in Table 6 and the fact that K_O gathers all the steps at which the acceptance test (6.5) holds.

Notation in [25]	Progressive Hedging notation
T	$\partial(h + i_{\mathcal{M}})$
y^k	u^{k+1}
v^k	g^k
x^k	w^k

Table 6 – Relations between notations in Algorithm 1 and in Algorithm 8.

□

Remark 6.2. *The analysis for Algorithm 1 in [25] is performed for a generalization of the $\varepsilon-$subdifferential, the $\varepsilon-$enlargement of a maximal monotone operator. Since $h \in \overline{\text{conv}}(\mathbb{R}^n)$ and $\mathcal{M}$ a linear subspace, then $\partial(h + i_{\mathcal{M}})$ is a maximal monotone operator. The ε_k-enlargement operator of $\partial(h + i_{\mathcal{M}})$ at u^{k+1} contains $\partial_{\varepsilon^k}(h + i_{\mathcal{M}})(u^{k+1})$, although in this case we avoid working with the enlargement of the subdifferential, since the use of the $\varepsilon-$subdifferential suffices for the optimization case.*

To establish convergence of the generated sequences, we prove that $\{w^k\}_{k\in K_O}$ is Fejér monotone with respect to the set of dual solutions $W^\star$, a classic property in the literature.

Lemma 6.2. *For any $w^\star \in W^\star$, and $k \in K_O$, it holds*

$$\|w^{k+1} - w^\star\|^2 \leqslant \|w^k - w^\star\|^2 - (1-\sigma_k^2)\|u^{k+1} - w^k\|^2.$$

Proof. First, [25, Lemma 4.1] implies for all $k \in \mathbb{N}$,

$$\|w^{k+\frac{1}{2}} - w^\star\|^2 \leqslant \|w^k - w^\star\|^2 - (1-\sigma_k^2)\|u^{k+1} - w^k\|^2. \tag{6.16}$$

Furthermore, for $k \in K_O$, it holds that $w^{k+1} = P_{\mathcal{M}}[w^{k+\frac{1}{2}}]$, and since the projection operator is nonexpansive and $W^\star \subseteq \mathcal{M}$, then

$$\begin{aligned} \|w^{k+1} - w^\star\| &= \|P_{\mathcal{M}}[w^{k+\frac{1}{2}}] - P_{\mathcal{M}}[w^\star]\| \\ &\leqslant \|w^{k+\frac{1}{2}} - w^\star\|. \end{aligned}$$

Substituting this last inequality in the above estimate gives the desired result. □

Based on Féjer monotonicity, the following proposition shows convergence properties for the candidate solutions $\{u^k\}$, the centers $\{w^k\}$, the aggregate subgradients $\{g^k\}$, and the errors $\{\varepsilon^k\}$.

Proposition 6.4. *Assume the outer loop of Algorithm 8 is performed infinitely many times. The following hold.*

(i) $\{w^{k+1}\}_{k\in K_O}$ *is bounded.*

(ii) $\sum_{k\in K_O} (1-\sigma_k^2)\|u^{k+1} - w^k\|^2 < +\infty.$

In addition, if $\{\sigma_k\}_{k\in K_O} \subseteq [0,1)$ stays bounded away from 1, then

(iii) $\{u^{k+1}\}_{k\in K_O}$, *and $u^{k+1} - w^k \to 0$ as $K_O \ni k \to +\infty$.*

(iv) $\sum_{k\in K_O} t_k^2\|g^k\|^2 < +\infty.$

(v) $\sum_{k\in K_O} t_k\varepsilon^k < +\infty.$

Proof. This result corresponds to [25, Corollary 4.2], using Lemma 6.2, a varying error tolerance σ_k and the relations in Table 6. Indeed, since Lemma 6.2 implies that $\{\|w^k -$

$w^\star\|\}_{k\in K_O}$ is a nonincreasing sequence for any $w^\star \in W^\star$, item (i) holds. Moreover, for any $n \in \mathbb{N}$

$$\begin{aligned}\sum_{\substack{k=0\\k\in K_O}}^{n}(1-\sigma_k^2)\|u^{k+1}-w^k\|^2 &\leqslant \sum_{\substack{k=0\\k\in K_O}}^{n}\|w^k-w^\star\|^2-\|w^{k+1}-w^\star\|^2\\ &= \sum_{k=0}^{n}\|w^k-w^\star\|^2-\|w^{k+1}-w^\star\|^2\\ &= \|w^0-w^\star\|^2-\|w^{n+1}-w^\star\|^2\\ &\leqslant \|w^0-w^\star\|^2,\end{aligned}$$

where in the first inequality we use Lemma 6.2, in the first equality we use that for any $k \notin K_O$, $w^{k+1} = w^k$, and in the third line we use the telescopic sum. Therefore, item (ii) follows by taking $K_O \ni n \to \infty$. As for item (iii), note that the assumption implies that $1-\sigma_k^2$ stays bounded away from 0, thus $\sum_{k\in K_O}\|u^{k+1}-w^k\|^2 < +\infty$. In turn, this implies $u^{k+1}-w^k \to 0$ as $K_O \ni k \to \infty$, and in particular, in view of item (i), $\{u^{k+1}\}_{k\in K_O}$ stays bounded. Furthermore, using the triangle inequality, the definition of r^k, Proposition 6.3(iii), and the fact that $\sigma_k < 1$, it holds that for $k \in K_O$

$$\begin{aligned}t_k\|g^k\| &\leqslant \|t_kg^k+u^{k+1}-w^k\|+\|u^{k+1}-w^k\|\\ &= \|r^k\|+\|u^{k+1}-w^k\|\\ &\leqslant (\sigma_k+1)\|u^{k+1}-w^k\|\\ &\leqslant 2\|u^{k+1}-w^k\|.\end{aligned}$$

Hence, $\sum_{k\in K_O} t_k^2\|g^k\|^2 < +\infty$. Finally, item (v) follows similarly from Proposition 6.3(iii). □

The key to prove global convergence of the iterates is noting that by taking the limit $K_O \ni k \to +\infty$ in the inclusion

$$g^k \in \partial_{\varepsilon_k}(h+i_{\mathcal{M}})(u^{k+1}),$$

implies that any accumulation point of $\{u^{k+1}\}_{k\in K_O}$ is a critical point of $h+i_{\mathcal{M}}$, that is, a solution to the convex problem (6.1). Observe that, by construction, Algorithm 8 guarantees $\{t_k\}_{k\in K_O}$ is bounded away from 0: $t_k \geqslant t_{\min}$ for all $k \in K_O$.

Theorem 6.3. *Assume the outer loop of Algorithm 8 is performed infinitely many times, and $\{\sigma_k\}_{k\in K_O} \subseteq [0,1)$ is bounded away from 1. Then, the sequences $\{w^{k+1}\}_{k\in K_O}$ and $\{w^{k+\frac{1}{2}}\}_{k\in K_O}$ converge to a minimizer of $h+i_{\mathcal{M}}$, and $\{h(w^{k+1})\}_{k\in K_O}$ and $\{h(w^{k+\frac{1}{2}})\}_{k\in K_O}$ converge to the optimal value of this minimization problem.*

Proof. Let $\bar{w}$ be any limit point of $\{w^{k+1}\}_{k\in K_O}$. Such a point exists by virtue of Proposition 6.4(i). Let $\{w^{k_j+1}\}$ be a subsequence of $\{w^{k+1}\}_{k\in K_O}$ such that $w^{k_j+1} \to \bar{w}$ as $j \to +\infty$.

Since $k_j \in K_O$, then $u^{k_j+1} \to \bar{w}$ as well. From Proposition 6.4 (iv) and (v), $t_k\|g^k\| \to 0$ and $t_k\varepsilon^k \to 0$ as $K_O \ni k \to +\infty$. Moreover, in view of $t_k \geqslant t_{\min}$ for all $k \in K_O$, then $g^k \to 0$ and $\varepsilon^k \to 0$. Therefore, we have

$$g^{k_j} \in \partial_{\varepsilon^{k_j}}(h + i_{\mathcal{M}})(u^{k_j+1}) \implies 0 \in \partial(h + i_{\mathcal{M}})(\bar{w}).$$

Therefore, any limit point of $\{w^{k+1}\}_{k\in K_O}$ is a solution to problem (6.1). Next, we prove that $\{w^{k+1}\}_{k\in K_O}$ has a unique accumulation point, following the reasoning of [17]. We include the proof here for completeness. Indeed, let $\overline{w}_1$, and $\overline{w}_2$ be two limit points of $\{w^{k+1}\}_{k\in K_O}$. Therefore, $\overline{w}_1$, $\overline{w}_2 \in W^*$, and thus Lemma 6.2 implies for all $k \in K_O$,

$$\|w^{k+1} - \overline{w}_j\|^2 \leqslant \|w^k - \overline{w}_j\|^2, \text{ for } j = 1, 2.$$

Hence, $\{\|w^k - \overline{w}_j\|\}_{k\in K_O}$, for $j = 1, 2$, is a monotone bounded sequence. Thus,

$$\lim_{k\in K_O} \|w^k - \overline{w}_j\| = \mu_j$$

exists for $j = 1, 2$. Fix $j = 1, 2$, and denote by $i = 3 - j$ the other index. Then

$$\begin{aligned} \|w^k - \overline{w}_j\|^2 &= \|w^k - \overline{w}_i + \overline{w}_i - \overline{w}_j\|^2 \\ &= \|w^k - \overline{w}_i\|^2 + 2\langle w^k - \overline{w}_i, \overline{w}_i - \overline{w}_j\rangle + \|\overline{w}_i - \overline{w}_j\|^2. \end{aligned}$$

Therefore, by taking the limit, we obtain

$$\mu_j^2 = \mu_i^2 + 2\lim_{k\in K_O}\langle w^k - \overline{w}_i, \overline{w}_i - \overline{w}_j\rangle + \|\overline{w}_i - \overline{w}_j\|^2.$$

Since $\overline{w}_i$ is an accumulation point of $\{w^k\}_{k\in K_O}$, then the limit above needs to vanish, and thus

$$\mu_j^2 - \mu_i^2 = \|\overline{w}_i - \overline{w}_j\|^2 > 0.$$

Switching indexes, it also holds

$$\mu_i^2 - \mu_j^2 = \|\overline{w}_j - \overline{w}_i\|^2 > 0,$$

a contradiction. Hence, there exists a unique accumulation point, the limit of $\{w^{k+1}\}_{k\in K_O}$, which is also a solution to problem (6.1) as proven above. Note that Proposition 6.4(ii), since $u^{k+1} = w^{k+1}$ for $k \in K_O$, implies $\{w^k\}_{k\in K_O}$ converges to $\overline{w}$ as well.

Finally, with respect to the sequence of intermediate points $\{w^{k+\frac{1}{2}}\}_{k\in K_O}$, note that (6.16) holds for $w^\star = \overline{w}$. Thus, taking the limit as $K_O \ni k \to +\infty$ in that estimate, and using Proposition 6.4(ii), it follows that $w^{k+\frac{1}{2}} \to \overline{w}$ as $K_O \ni k \to +\infty$. The convergence of $\{h(w^{k+1}\}_{k\in K_O}$ and $\{h(w^{k+\frac{1}{2}})\}_{k\in K_O}$ follows from the continuity of h. □

Local rate of convergence: infinite loop of outer steps

In order to establish linear rates of convergence, an extra assumption is needed, in this case, an error bound (cf. Section 1.4). Specifically, we use a generalization of the condition in [25, Theorem 3.2]: assume that for any $\overline{h} \geqslant \inf_{w\in\mathcal{M}} h(w)$, there exists some $L > 0$ and $\delta > 0$ such that whenever $w \in \mathcal{M}$, and $h(w) \leqslant \overline{h}$,

$$g \in \big(\partial h(w) + \mathcal{M}^{\perp}\big) \cap B(0,\delta) \implies \operatorname{dist}(w, W^{\star}) \leqslant L\|g\|. \tag{6.17}$$

This estimate actually corresponds to the subdifferential-based error bound of Definition 1.7 applied to $h + i_{\mathcal{M}}$, and as such, it is used to prove the speed of convergence of the proposed method. Recall that Algorithm defines $\{t_k\}_{k\in K_O}$ bounded away from zero by construction.

Theorem 6.4. *Assume the outer loop of Algorithm 8 is performed infinitely many times, $\{\sigma_k\}_{k\in K_O} \subseteq [0,1)$ is bounded away from 1, and the error bound condition (6.17) holds. Then*

(i) If $\{t_k\}_{k\in K_O}$ is bounded above, then $\{\operatorname{dist}(w^k, W^{\star})\}_{k\in K_O}$ converges linearly to 0.

(ii) If $t_k \to +\infty$ and $\sigma_k \to 0$ as $K_O \ni k \to +\infty$, then $\{\operatorname{dist}(w^k, W^{\star})\}_{k\in K_O}$ converges superlinearly to 0, and for $k \in K_O$,

$$\operatorname{dist}(w^{k+1}, W^{\star})^2 \leqslant \left(1 - \frac{1-\sigma_k}{(1+Lt_k^{-1})^2(1+\sigma_k)}\right) \operatorname{dist}(w^k, W^{\star})^2.$$

Proof. We follow the proof of [25, Theorem 3.2]. In fact, from Proposition 6.4(iii) and (iv), we know that

$$t_k\|g^k\| \to 0, \text{ and } w^{k+1} - w^k = u^{k+1} - w^k \to 0, \text{ as } K_O \ni k \to +\infty.$$

Since $\{t_k\}_{k\in K_O}$ is bounded away from 0, then $g^k \to 0$. Denote $z^k = \operatorname{prox}_{t_k(h+i_{\mathcal{M}})}(w^k)$, then [25, Lemma 2.2] implies for $k \in K_O$,

$$\|z^k - u^{k+1}\| \leqslant \sigma_k\|w^k - u^{k+1}\|. \tag{6.18}$$

Therefore, using the triangle inequality and (6.18), it follows for $k \in K_O$

$$\begin{aligned} \|z^k - w^k\| &\leqslant \|z^k - u^{k+1}\| + \|u^{k+1} - w^k\| \\ &\leqslant (1+\sigma_k)\|u^{k+1} - w^k\|. \end{aligned} \tag{6.19}$$

From the definition of z^k, it also holds $z^k \in \mathcal{M}$ and

$$\frac{1}{t_k}(w^k - z^k) \in \partial(h + i_{\mathcal{M}})(z^k).$$

This subgradient satisfies the following inequality, due to (6.19):

$$\left\|\frac{1}{t_k}(w^k - z^k)\right\| \leqslant \frac{1+\sigma_k}{t_k}\|u^{k+1} - w^k\| \leqslant \frac{2}{t_{\min}}\|u^{k+1} - w^k\|,$$

where the second inequality follows from $\sigma_k \in [0,1)$, and $t_{\min} \leqslant t_k$. Therefore, $u^{k+1} - w^k \to 0$ as $K_O \ni k \to +\infty$ implies

$$\partial(h + i_{\mathcal{M}})(z^k) \ni \frac{1}{t_k}(w^k - z^k) \to 0 \text{ as } K_O \ni k \to +\infty. \tag{6.20}$$

Additionally, from the definition of z^k, it holds

$$h(z^k) + \frac{t_{\min}}{2}\left\|\frac{1}{t_k}(z^k - w^k)\right\|^2 \leqslant h(z^k) + \frac{1}{2t_k}\|z^k - w^k\|^2 \leqslant h(w^k). \tag{6.21}$$

In view of Theorem 6.3, since $w^{k+1} - w^k \to 0$ as $K_O \ni k \to +\infty$, then $w^k \to \overline{w}$ and $h(w^k) \to \inf\limits_{w\in\mathcal{M}} h(w)$. Hence

$$\begin{aligned}
\inf_{w\in\mathcal{M}} h(w) &\leqslant \liminf_{k\in K_O} h(z^k) \\
&= \liminf_{k\in K_O} h(z^k) + \frac{t_{\min}}{2}\left\|\frac{1}{t_k}(z^k - w^k)\right\|^2 \\
&\leqslant \limsup_{k\in K_O} h(z^k) + \frac{t_{\min}}{2}\left\|\frac{1}{t_k}(z^k - w^k)\right\|^2 \\
&\leqslant \lim_{k\in K_O} h(w^k) \\
&= \inf_{w\in\mathcal{M}} h(w),
\end{aligned}$$

where the first inequality follows from the fact $z^k \in \mathcal{M}$, the first equality follows from (6.20), and (6.21) yields the fourth line. Moreover, (6.20) also implies

$$\limsup_{k\in K_O} h(z^k) + \frac{t_{\min}}{2}\left\|\frac{1}{t_k}(z^k - w^k)\right\|^2 = \limsup_{k\in K_O} h(z^k).$$

Then, the chain of estimates above yields

$$\inf_{w\in\mathcal{M}} h(w) \leqslant \liminf_{k\in K_O} h(z^k) \leqslant \limsup_{k\in K_O} h(z^k) \leqslant \inf_{w\in\mathcal{M}} h(w),$$

meaning $h(z^k) \to \inf\limits_{w\in\mathcal{M}} h(w)$.

In this manner, for any $\overline{h} > \inf\limits_{w\in M} h(w)$ and $\delta > 0$ from (6.17), for all sufficiently large $k \in K_O$, $h(z^k) \leqslant \overline{h}$, $z^k \in \mathcal{M}$, and

$$\frac{1}{t_k}(w^k - z^k) \in \partial(h + i_{\mathcal{M}})(z^k) \cap B(0,\delta).$$

The error bound (6.17) gives

$$\operatorname{dist}(z^k, W^\star) \leqslant \frac{L}{t_k}\|w^k - z^k\|. \tag{6.22}$$

Next, we use the triangle inequality, (6.22), and (6.19) to obtain for all sufficiently large $k \in K_O$

$$\begin{aligned} \operatorname{dist}(w^k, W^\star) &\leqslant \operatorname{dist}(z^k, W^\star) + \|w^k - z^k\| \\ &\leqslant \left(1 + \frac{L}{t_k}\right) \|w^k - z^k\| \\ &\leqslant \left(1 + \frac{L}{t_k}\right) (1 + \sigma_k) \|u^{k+1} - w^k\| \end{aligned} \tag{6.23}$$

Moreover, Lemma 6.2 implies for $k \in K_O$

$$\operatorname{dist}(w^{k+1}, W^\star)^2 \leqslant \operatorname{dist}(w^k, W^\star)^2 - (1 - \sigma_k^2)\|u^{k+1} - w^k\|^2.$$

Taking squares in (6.23) and multiplying by -1, then we can bound $\|u^{k+1} - w^k\|^2$ for all sufficiently large $k \in K_O$ in the last inequality to obtain

$$\begin{aligned} \operatorname{dist}(w^{k+1}, W^\star)^2 &\leqslant \left(1 - \frac{1 - \sigma_k^2}{\left(1 + \frac{L}{t_k}\right)^2 (1 + \sigma_k)^2}\right) \operatorname{dist}(w^k, W^\star)^2 \\ &= \left(1 - \frac{1 - \sigma_k}{\left(1 + \frac{L}{t_k}\right)^2 (1 + \sigma_k)}\right) \operatorname{dist}(w^k, W^\star)^2. \end{aligned}$$

This estimate is the cornerstone of the rate of convergence of Algorithm 8. To prove (i), since $\sigma_k \in [0, 1)$, and $t_{\min} \leqslant t_k$, then

$$1 - \frac{1 - \sigma_k}{\left(1 + \frac{L}{t_k}\right)^2 (1 + \sigma_k)} \leqslant 1 - \frac{1}{2\left(1 + \frac{L}{t_{\min}}\right)^2} < 1.$$

In this way, for all sufficiently large $k \in K_O$

$$\operatorname{dist}(w^{k+1}, W^\star)^2 \leqslant \left(1 - \frac{1}{2\left(1 + \frac{L}{t_{\min}}\right)^2}\right) \operatorname{dist}(w^k, W^\star)^2,$$

and thus

$$\limsup_{k \in K_O} \frac{\operatorname{dist}(w^{k+1}, W^\star)^2}{\operatorname{dist}(w^k, W^\star)^2} < 1.$$

To prove (ii), note that when $t_k \to +\infty$, and $\sigma_k \to 0$, then

$$1 - \frac{1 - \sigma_k}{\left(1 + \frac{L}{t_k}\right)^2 (1 + \sigma_k)} \to 0,$$

therefore,

$$\limsup_{k \in K_O} \frac{\operatorname{dist}(w^{k+1}, W^\star)^2}{\operatorname{dist}(w^k, W^\star)^2} = 0.$$

□

6.1.2 Comparison between the dual-embedded forward-backward and Bundle Progressive Hedging methods

Algorithm 8 bears some resemblance to PH and the progressive-hedging-like method proposed in Chapter 5. Algorithm 6 and DEFB present some key differences with the classical PH, the most important one being the capacity of changing the stepsize parameter along iterations. PH keeps the stepsize fixed, making it difficult to obtain (empirical) superlinear convergence results. The method presented in this section, by contrast, adjusts x^k using a forward-backward step, and does not alter t_k when the surrogate model function of $h + i_{\mathcal{M}}$ appears to be too inaccurate. This difference makes it at least theoretically possible to attain a form of superlinear convergence.

The main difference between DEFB and the algorithm proposed in Chapter 5 is the acceptance test. After solving all the proximal subproblem, DEFB checks the accuracy of the iterates in terms of their feasibility (how far they are from being an element of $\mathcal{M}$), and the accuracy of the model, using a relative-error criteria. Therefore, DEFB directly addresses feasibility and the quality of the approximations. On the other hand, the progressive-hedging-like algorithm of Chapter 5 checks descent of the iterates, that is, it verifies sufficient objective improvement compared to the last solution candidate, with a fixed Armijo-like parameter. The following result shows that DEFB actually provides descent for the objective function, satisfying a sufficient descent estimate with a varying Armijo-like parameter.

Proposition 6.5 (Sufficient descent of DEFB)**.** *The sequence $\{w^{k+1}\}_{k\in K_O}$ generated by Algorithm 8 satisfies the following sufficient descent estimate:*

$$h(w^{k+1}) - h(w^k) \leqslant \left(\frac{\sigma_k^2 - 1}{2t_k}\right) \|w^{k+1} - w^k\|^2. \tag{6.24}$$

Proof. By expanding squares, for all k,

$$\|w^{k+\frac{1}{2}} - u^{k+1}\|^2 = \|w^{k+\frac{1}{2}} - w^k\|^2 + \|w^k - u^{k+1}\|^2 + 2\langle w^{k+\frac{1}{2}} - w^k, w^k - u^{k+1}\rangle.$$

In this way, the left-hand side of the acceptance test (6.5) can be written as

$$\begin{aligned}
\|w^{k+\frac{1}{2}} - u^{k+1}\|^2 + 2t_k\varepsilon_k &= \|w^{k+\frac{1}{2}} - w^k\|^2 + \|w^k - u^{k+1}\|^2 + 2\langle w^{k+\frac{1}{2}} - w^k, w^k - u^{k+1}\rangle \\
&\quad +2t_k\varepsilon_k \\
&= \|w^{k+\frac{1}{2}} - w^k\|^2 + \|w^k - u^{k+1}\|^2 + 2\langle w^{k+\frac{1}{2}} - w^k, w^k - u^{k+1}\rangle \\
&\quad +2t_k\big(h(u^{k+1}) - h^k(w^{k+\frac{1}{2}})\big) + 2\langle w^{k+\frac{1}{2}} - w^k, u^{k+1} - w^{k+\frac{1}{2}}\rangle \\
&= \|w^{k+\frac{1}{2}} - w^k\|^2 + \|w^k - u^{k+1}\|^2 + 2\langle w^{k+\frac{1}{2}} - w^k, w^k - w^{k+\frac{1}{2}}\rangle \\
&\quad +2t_k\big(h(u^{k+1}) - h^k(w^{k+\frac{1}{2}})\big) \\
&= -\|w^{k+\frac{1}{2}} - w^k\|^2 + \|w^k - u^{k+1}\|^2 + 2t_k\big(h(u^{k+1}) - h^k(w^{k+\frac{1}{2}})\big).
\end{aligned}$$

Therefore, (6.5) is equivalent to

$$h(u^{k+1}) - \left[h^k(w^{k+\frac{1}{2}}) + \frac{1}{2t_k}\|w^{k+\frac{1}{2}} - w^k\|^2\right] \leqslant \left(\frac{\sigma_k^2 - 1}{2t_k}\right)\|w^k - u^{k+1}\|^2.$$

By construction, since $w^{k+\frac{1}{2}}$ solves (6.3), thus

$$h^k(w^{k+\frac{1}{2}}) + \frac{1}{2t_k}\|w^{k+\frac{1}{2}} - w^k\|^2 \leqslant h^k(w^k),$$

yielding

$$h(u^{k+1}) - h^k(w^k) \leqslant \left(\frac{\sigma_k^2 - 1}{2t_k}\right)\|w^k - u^{k+1}\|^2.$$

The result follows from the fact $u^{k+1} = w^{k+1}$ whenever $k \in K_O$. □

Remark 6.3. *The estimate in (6.24) holds for $k \notin K_O$ as well, since in this case $w^{k+1} = w^k$, and thus the left-hand side vanishes.*

Unlike DEFB, the method in Chapter 5 also requires bounded stepsizes t_k (cf. Theorem 5.2), precluding the use of standard arguments for demonstrating a superlinear rate of convergence (cf. [17, Theorem 2]). On the other hand, DEFB allows the stepsizes to be driven to infinity, which leads to superlinear convergence. In practice, this means that once DEFB is close to a solution, one can increase its stepsizes to accelerate the rate at which it gains accuracy, by paying the cost of having to solve a problem close to the original one. Furthermore, (6.24) allows the term $m_k = \dfrac{\sigma_k^2 - 1}{2t_k}$ to vary along iterations, while the classical proximal bundle acceptance test $h(w^{k+1}) - h(w^k) \leqslant m\big(h^k(w^{k+1}) - h(w^k)\big)$ keeps the parameter $m \in (0, 1)$ fixed.

6.2 Dual embedded forward-backward method for stochastic programming

The method presented in Section 6.1 can be viewed as a splitting method to individually exploit the properties of the objective function h of (6.1), and the feasible set $\mathcal{M}$, a linear subspace. This configuration materializes in the dual formulation of a convex stochastic programming problem.

In the outer-loop convergence case of Algorithm 8, the iterations conform to a form of proximal point algorithm for problem (6.1). When proximal point algorithms are applied to dual problems, they lead to augmented Lagrangian methods, as established in [153]. Thus, when Algorithm 8 is applied to a dual problem, it leads to a type of augmented Lagrangian algorithm. We call such methods "Dual-embedded forward-backward Augmented Lagrangian" (DEFBAL) algorithms.

In this section, we apply the theory of Section 6.1 to obtain a method resembling the progressive hedging algorithm for convex stochastic optimization problems. We follow the notation of Section 2.2.2 and Chapter 5.

We introduce some extra notation in order to directly treating last-stage variables, as in [139]. Recall we consider S possible T-stage scenario realizations for the underlying stochastic process of the problem. For each scenario $s = 1, \ldots, S$, let $(x_s, x_{sT}) \in \mathbb{R}^{n'_s}$ denote the scenario-s decision variable vector, including all stages, in which $x_s \in \mathbb{R}^{n_s}$ covers stages $t = 1, \ldots, T-1$, and x_{sT} denotes the decision variable for the last stage T. Furthermore, $x = (x_s)_{s=1}^S \in \mathbb{R}^n$ represents the vector that gathers decision variables for all scenarios and all stages, except the last stage $t = T$, while $x_T \in \mathbb{R}^{n'-n}$ denotes the vector of decision variables at stage $t = T$ and all scenarios.

For each $s = 1, \ldots, S$, let $p_s > 0$ denote the probability of scenario s, hence $\sum_{s=1}^{S} p_s = 1$. Moreover, for each $s = 1, \ldots, S$, let $f_s \in \overline{\text{conv}}(\mathbb{R}^{n'_s})$, and

$$f(x, x_T) = \sum_{s=1}^{S} p_s f_s(x_s, x_{sT}),$$

be the objective function to be minimized.

Consider the problem

$$\begin{cases} \min & f(x, x_T) \\ \text{s.t.} & (x_s, x_{sT}) \in C_s, \ s = 1, \ldots, S, \\ & x \in \mathcal{N}, \end{cases} \tag{6.25}$$

where $C_s \subseteq \mathbb{R}^{n'_s}$ is a nonempty compact convex set, and $\mathcal{N}$ denotes the nonanticipative subspace. We denote $\mathcal{C} = \Pi_{s=1}^S C_s$.

Define, for each scenario $s = 1, \ldots, S$, the marginal function $F_s : \mathbb{R}^{n_s} \to \mathbb{R} \cup \{+\infty\}$ as

$$F_s(x_s) = \inf_{x_{sT}} \{ f_s(x_s, x_{sT}) : (x_s, x_{sT}) \in C_s \}. \tag{6.26}$$

Additionally, define the function $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ as

$$F(x) = \sum_{s=1}^{S} p_s F_s(x_s), \tag{6.27}$$

allowing problem (6.25) to be equivalently reformulated as (cf. (2.9))

$$\begin{cases} \min\limits_{x} & F(x) \\ \text{s.t.} & x \in \mathcal{N}. \end{cases} \tag{6.28}$$

The solution of problems (6.28) and (6.25) can be easily related, as the following simple result shows.

Proposition 6.6. *Suppose $x^\star$ is a solution to (6.28), and $x_T^\star$ satisfies*

$$F(x^\star) = f(x^\star, x_T^\star), \quad (x^\star, x_T^\star) \in \mathcal{C}.$$

Then, $(x^\star, x_T^\star)$ is a solution to (6.25).

Proof. For all $x \in \mathcal{N}$, and all x_T such that $(x, x_T) \in \mathcal{C}$, there holds

$$f(x^\star, x_T^\star) = F(x^\star) \leqslant F(x) \leqslant f(x, x_T).$$

□

Our main goal is to take advantage of the separable structure of f. For that, we relax the constraint $x \in \mathcal{N}$, obtaining the separable Lagrangian (cf. (5.3)–(5.4))

$$\mathcal{L}(x, w) = \sum_{s=1}^{S} p_s \mathcal{L}_s(x_s, w_s), \text{ where } \mathcal{L}_s(x_s, w_s) = F_s(x_s) + \langle x_s, w_s \rangle,$$

with the corresponding dual function given by (cf. (5.6))

$$h(w) = \sum_{s=1}^{S} p_s h_s(w_s), \text{ where } h_s(w_s) = -\inf_{x_s} \mathcal{L}_s(x_s, w_s). \tag{6.29}$$

The corresponding dual problem is defined as (cf. (6.1))

$$\begin{cases} \min\limits_{w} & h(w) \\ \text{s.t.} & w \in \mathcal{N}^\perp. \end{cases} \tag{6.30}$$

Note that problem (6.30) is problem (6.1) for $\mathcal{M} = \mathcal{N}^\perp$. This relationship goes beyond solely the feasible sets. The following result states some basic properties that the objective functions of the primal and dual problems satisfy.

Proposition 6.7. *Consider function $F : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined in (6.26)–(6.27) and (6.29), respectively. The following hold.*

(i) $F \in \overline{\text{conv}}(\mathbb{R}^n)$ and is finite.

(ii) $h \in \overline{\text{conv}}(\mathbb{R}^n)$ and is finite, such that for each $s = 1, \ldots, S$,

$$h_s(w_s) = -\begin{cases} \min\limits_{x_s, x_{sT}} & f_s(x_s, x_{sT}) + \langle x_s, w_s \rangle \\ s.t. & (x_s, x_{sT}) \in C_s \end{cases}$$

Proof. From (6.26), the $s-$dual function can be equivalently formulated as

$$\begin{aligned} h_s(w_s) &= -\inf_{x_s}\left(\inf_{x_{sT}}\{F_s(x_s,x_{sT}) : (x_s,x_{sT}) \in C_s\} + \langle x_s, w_s\rangle\right) \\ &= -\inf_{x_s,x_{sT}}\{F_s(x_s,x_{sT}) + \langle x_s, w_s\rangle : (x_s,x_{sT}) \in C_s\}. \end{aligned}$$

In this manner, h_s is a convex function, since it is the negative of a function defined as the infimum of convex functions $F_s(x_s,x_{sT}) + \langle x_s, w_s\rangle$. Furthermore, since f_s is lsc and C_s is compact, then from [130, Lemma 1.2], F_s is lsc as a marginal function of f_s, and the problem in (6.26) attains its minimum, which implies that $F_s(x_s)$ is finite for all x_s. From [130, Proposition 8.26], F_s is convex because f_s is convex. Finally, h_s is lsc and finite from [130, Lemma 1.2], since each C_s is compact, and $F_s(x_s,x_{sT}) + \langle x_s, w_s\rangle$ lsc as a function of (x_s,x_{sT}).

□

We now apply the theory of Section 6.1 to the convex dual function h and $\mathcal{M} = \mathcal{N}^\perp$, observing that problem (6.30) has the same structure of problem (6.1). In this setting, Algorithm 8 takes the form shown in Algorithm 9, adding the extra feature of scenario-based decomposition.

The introduction of the marginal function in (6.27) is crucial in (6.31), since in this way the last-stage decision variables x_{sT} are not carried to the linear term associated with the multiplier, nor the quadratic penalization term. In practice, this means that the vacuous nonanticipative constraint on the last state does not impact the formulation of the problem.

The direct relationship between Algorithm 9 and Algorithm 8 is stated in the following result.

Proposition 6.8. *Consider Algorithm 8 and Algorithm 9. The following hold.*

(i) For each $s = 1,\dots,S$, defining $w_s^{k+\frac{1}{2}}$ via solving (6.3) is equivalent to perform (6.31)–(6.32) to obtain $(x_s^{k+\frac{1}{2}}, x_{sT}^{k+1})$, where $x_s^{k+\frac{1}{2}}$ coincides with the primal point of Proposition 6.1(iii).

(ii) The formulas (6.4) and (6.35) are two identical ways to define the error ε_k.

(iii) The acceptance tests (6.5) and (6.36) are equivalent.

(iv) The stopping tests (6.5) and (6.33) are equivalent.

(v) The rules (6.6) and (6.33) to update x^k are equivalent.

Algorithm 9 A progressive-hedging-like algorithm derived from DEFBAL

1: **Initialization**: choose $x^0 \in \mathcal{N}$, $w^0 \in \mathcal{N}^\perp$ and $\sigma^0 \in [0,1)$, $t_{\min} > 0$, and $t_0 \geqslant t_{\min}$. Set $\texttt{TOL} > 0$, and $\varepsilon_0 > \texttt{TOL}$.

2: **for** $k = 0, 1, \ldots$ **do**

3: **Primal subproblems:** for each scenario $s = 1, \ldots, S$, compute $(x_s^{k+\frac{1}{2}}, x_{sT}^{k+1})$ as the unique solution of

$$\min_{(x_s, x_{sT}) \in C_s} \left\{ f_s(x_s, x_{sT}) + \langle w_s^k, x_s \rangle + \frac{t_k}{2} \|x_s - x_s^k\|^2 \right\}. \tag{6.31}$$

4: **Dual update:** Define

$$w_s^{k+\frac{1}{2}} = w_s^k + t_k (x_s^{k+\frac{1}{2}} - x_s^k) \quad \text{for } s = 1, \ldots, S, \tag{6.32}$$

and

$$u^{k+1} = P_{\mathcal{N}^\perp}[w^{k+\frac{1}{2}}].$$

5: **Center update:** define

$$x^{k+1} = P_{\mathcal{N}}[x^{k+\frac{1}{2}}]. \tag{6.33}$$

6: **Stopping test**: $\|x^{k+\frac{1}{2}} - x^k\| \leqslant \texttt{TOL}$ and $\varepsilon_k \leqslant \texttt{TOL}$, stop and return (w^k, x^k).

7: **Acceptance test:** For each $s = 1, \ldots, S$, find

$$y_s^{k+1} \in \arg\min_{z_s} \mathcal{L}_s(z_s, u_s^{k+1}). \tag{6.34}$$

and define

$$\varepsilon_k = -\mathcal{L}(y^{k+1}, u^{k+1}) + \mathcal{L}(x^{k+\frac{1}{2}}, w^{k+\frac{1}{2}}) - \langle x^k, w^{k+\frac{1}{2}} \rangle_S - t_k \langle x^k - x^{k+\frac{1}{2}}, x^k - x^{k+1} \rangle_S. \tag{6.35}$$

If

$$\|x^k - x^{k+1}\|_S^2 + 2t_k^{-1}\varepsilon_k \leqslant \sigma_k^2 \|x^{k+\frac{1}{2}} - x^{k+1}\|_S^2, \tag{6.36}$$

perform an **outer step**: define $w^{k+1} = u^{k+1}$ and choose $t_{k+1} \geqslant t_{\min}$ and $\sigma_{k+1} \in [0,1)$. Otherwise, perform an **inner step**: set $w^{k+1} = w^k$, $t_{k+1} = t_k$, and $\sigma_{k+1} = \sigma_k$.

8: **end for**

Proof. Item (i) follows by using classic primal-dual arguments, as the one provided in Section 5.2.2. Here, we give an alternative proof. First, problem (6.31) is equivalent to

$$\min_{x_s} \left\{ F_s(x_s) + \langle w_s^k, x_s \rangle + \frac{t_k}{2} \|x_s - x_s^k\|^2 \right\}$$

in view of (6.26). The Fenchel dual of the problem above is

$$\min_{w_s} \left\{ F_s^*(-w_s) + \left(\langle w_s^k, \cdot \rangle + \frac{t_k}{2} \| \cdot - x_s^k \|^2 \right)^* (w_s) \right\}.$$

Each one of the terms in the sum can be rewritten as follows:

$$\begin{aligned} F_s^*(-w_s) &= \sup_{x_s}\left\{\langle -w_s^k, x_s\rangle - F_s(x_s)\right\} \\ &= -\inf_{x_s}\left\{F_s(x_s) + \langle w_s^k, x_s\rangle\right\} \\ &= h_s(w_s), \end{aligned} \tag{6.37}$$

and by using the rules in Proposition 1.1:

$$\begin{aligned} \left(\langle w_s^k, \cdot\rangle + \frac{t_k}{2}\|\cdot - x_s^k\|^2\right)^*(w_s) &= \left(\frac{t_k}{2}\|\cdot - x_s^k\|^2\right)^*(w_s - w_s^k) \\ &= \left(\frac{t_k}{2}\|\cdot\|^2\right)^*(w_s - w_s^k) + \langle x_s^k, w_s - w_s^k\rangle \\ &= t_k\left(\frac{t_k}{2}\|\cdot\|^2\right)^*\left(\frac{1}{t_k}(w_s - w_s^k)\right) + \langle x_s^k, w_s - w_s^k\rangle \\ &= \frac{t_k}{2}\left\|\frac{1}{t_k}(w_s - w_s^k)\right\|^2 + \langle x_s^k, w_s - w_s^k\rangle \\ &= \frac{1}{2t_k}\|w_s - w_s^k\|^2 + \langle x_s^k, w_s - w_s^k\rangle. \end{aligned}$$

In this way, after discarding constant terms, the Fenchel dual problem is equivalent to (6.31). Relationship (6.32) is obtained via duality as well: from the optimality condition (6.9) of problem (6.3), there exists $x_s^* \in \mathbb{R}^{n_s}$ such that $-x_s^* \in \partial h_s(w_s^{k+\frac{1}{2}})$, and

$$0 = -x_s^* + x^k + \frac{1}{t_k}(w_s^{k+\frac{1}{2}} - w_s^k).$$

Since $F_s^*(-w_s) = h_s(w_s)$, then $-x_s^* \in \partial h_s(w_s^{k+\frac{1}{2}})$ is equivalent to $-w_s^{k+\frac{1}{2}} \in \partial F_s(x_s^*)$. Therefore, the above identity implies

$$\begin{aligned} 0 &= -w_s^{k+\frac{1}{2}} + w_s^k + t_k(x_s^* - x_s^k) \\ &\in \partial F_s(x_s^*) + w_s^k + t_k(x_s^* - x_s^k), \end{aligned} \tag{6.38}$$

corresponding to the optimality condition of problem (6.31). From the uniqueness of the solution of this problem (due to strong convexity of the objective function), $x_s^* = x_s^{k+\frac{1}{2}}$, and thus (6.32) follows.

Moreover, from (6.29) and (6.34), $h_s(u_s^{k+1}) = -\mathcal{L}_s(y_s^{k+1}, u_s^{k+1})$. In view of (6.38) and (6.32), $x_s^{k+\frac{1}{2}}$ minimizes $\mathcal{L}_s(\cdot, w_s^{k+\frac{1}{2}})$, and thus $h_s(w_s^{k+\frac{1}{2}}) = -\mathcal{L}_s(\cdot, w_s^{k+\frac{1}{2}})$. Using again (6.32), we can see that the following primal-dual relationships hold: $t_k(x^k - x^{k+\frac{1}{2}}) = w^k - w^{k+\frac{1}{2}}$, and $t_k(x^k - x^{k+1}) = u^{k+1} - w^{k+\frac{1}{2}}$ by applying $P_{\mathcal{N}}$ on both sides. Thus, from (6.35)

$$\begin{aligned} \varepsilon_k &= \sum_{s=}^{S} p_s\left(-\mathcal{L}_s(u_s^{k+1}, u_s^{k+1}) + \mathcal{L}_s(x_s^{k+\frac{1}{2}}, w_s^{k+\frac{1}{2}}) - \langle x_s^k, w_s^{k+\frac{1}{2}}\rangle - t_k\langle x_s^k - x_s^{k+\frac{1}{2}}, x_s^k - x_s^{k+1}\rangle\right) \\ &= \sum_{s=1}^{S} p_s\left(h_s(y_s^{k+1}) - h_s(w_s^{k+\frac{1}{2}}) - \langle x_s^k, w_s^{k+\frac{1}{2}}\rangle - \frac{1}{t_k}\langle w_s^k - w_s^{k+\frac{1}{2}}, u_s^{k+1} - w_s^{k+1}\rangle\right) \\ &= \sum_{s=}^{S} p_s\left(h_s(y_s^{k+1}) - h_s^k(w_s^{k+\frac{1}{2}}) - \frac{1}{t_k}\langle w_s^k - w_s^{k+\frac{1}{2}}, u_s^{k+1} - w_s^{k+1}\rangle\right), \end{aligned}$$

from which item (ii) follows.

Regarding the acceptance tests, multiplying (6.36) by t_k^2 yields

$$\begin{aligned} (6.36) \quad &\iff \|t_k(x^k - x^{k+1})\|_S^2 + 2t_k\varepsilon_k \leqslant \sigma_k^2\|t_k(x^{k+\frac{1}{2}} - x^{k+1})\|_S^2 \\ &\iff \|u^{k+1} - w^{k+\frac{1}{2}}\|_S^2 + 2t_k\varepsilon_k \leqslant \sigma_k^2\|w^k - w^{k+\frac{1}{2}}\|_S^2, \end{aligned}$$

and thus item (iii) follows. The same primal-dual relationship gives the validity of item (iv).

Finally, applying $P_{\mathcal{N}}$ to (6.32) and using (6.33), yields

$$\begin{aligned} x^{k+1} &= P_{\mathcal{N}}\left[x^k + \frac{1}{t_k}(w^{k+\frac{1}{2}} - w^k)\right] \\ &= P_{\mathcal{N}}[x^k] + \frac{1}{t_k}(P_{\mathcal{N}}[w^{k+\frac{1}{2}}] - P_{\mathcal{N}}[w^k]) \\ &= x^k + \frac{1}{t_k}P_{\mathcal{N}}[w^{k+\frac{1}{2}}] \\ &= x^k + \frac{1}{t_k}(w^{k+\frac{1}{2}} - P_{\mathcal{N}^\perp}[w^{k+\frac{1}{2}}]) \\ &= x^k + \frac{1}{t_k}(w^{k+\frac{1}{2}} - u^{k+1}), \end{aligned}$$

where in the second line we use linearity of the projection operator, in the third line we use the fact $x^k \in \mathcal{N}$ and $w^k \in \mathcal{N}^\perp$, in the fourth line we use the Moreau identity for the projection, and the last line follows from the definition of u^{k+1}. This proves item (v).

□

In order to deduce convergence of Algorithm 9, we capitalize on on the theory described in Section 6.1. As for primal iterates, we require the following standard result, for which we provide a short proof.

Lemma 6.3. *Let $w^\star$ be a solution to the dual problem (6.30). Then, any $x^\star \in \mathcal{N}$ is a solution to primal problem (6.28) whenever $-x^\star \in \partial h(w^\star)$.*

Proof. First, the Fenchel-Young inequality implies

$$h(w^\star) + h^*(-x^\star) = -\langle x^\star, w^\star\rangle_S, \tag{6.39}$$

where the right-hand side inner product is 0, since $x^\star \in \mathcal{N}$, and $w^\star \in \mathcal{N}^\perp$. Furthermore, note that fue to (6.37), $F^*(-w) = h(w)$, and thus

$$\begin{aligned} h^*(-x^*) &= \sup_w\{\langle w, -x^*\rangle_S - h(w)\} \\ &= \sup_w\{\langle -w, x^*\rangle_S - F^*(-w)\} \\ &= \sup_w\{\langle w, x^*\rangle_S - F^*(w)\} \\ &= F^{**}(x^*) \\ &= F(x^*), \end{aligned} \tag{6.40}$$

where in the third equality we use the change of variables $w \leftarrow -w$, in the fourt line we use the definition of the Fenchel conjugate, and in the last line we use the fact that $F \in \overline{\text{conv}}(\mathbb{R}^n)$ (Proposition 6.7(i)). Substituting this relation in (6.39) gives, for $w^* \in \mathcal{N}^\perp$, $x^* \in \mathcal{N}$,

$$h(w^*) + F(x^*) = 0.$$

Hence, the result follows from strong duality.

□

Now we proceed to show the convergence properties of Algorithm 9, following the reasoning of Section 6.1.1. We first analyze the case of finite termination, and then two cases when the algorithm runs indefinitely: the first case corresponds to the tail of inner steps, and the second one is when an infinite number of outer steps are performed.

Finite termination: primal-dual case

Similarly to the dual finite termination case of Section 6.1.1, when Algorithm 9 stops after finitely many iterations when $\texttt{TOL} = 0$, due to Proposition 6.8(iv), then $0 \in \partial(h + i_{\mathcal{N}^\perp})(u^{k+1})$, $x^{k+\frac{1}{2}} = x^k$ and $\varepsilon_k = 0$. In particular, in view of (6.32), $w^{k+\frac{1}{2}} = w^k \in \mathcal{N}^\perp$, and thus $w^{k+\frac{1}{2}} = u^{k+1}$. Therefore, (6.10) implies $0 = g^k \in \partial h(w^k) + x^k$. To summarize, $w^k = u^{k+1}$ is a dual solution, and $-x^k \in \partial h(w^k) \cap \mathcal{N}$. In this way, Lemma 6.3 implies x^k is a primal solution.

Convergence of infinite loop of inner steps: primal-dual case

In this section we adopt the assumptions and notation of Section 6.1.1.2. In particular, let $\widehat{k} \in \mathbb{N}$ be the last iteration the acceptance test (6.5) is satisfied, and thus for all $k > \widehat{k}$, $w^k = \widehat{w}$, $t_k = \widehat{t}$, $\sigma_k = \widehat{\sigma}$, and (6.12) holds.

Remark 6.4. *The dual convergence of the tail of inner steps of Theorem 6.2 requires* $\text{dom}(h^*)$ *to be compact. In view of* (6.40)*,* $F(-\cdot) = h^*(\cdot)$*, F has compact domain if and only if h^* has compact domain. Furthermore, taking* (6.26) *into account, F has compact domain whenever the sets $C_s \subset \mathbb{R}^{n_s}$, for all $s = 1, \ldots, S$, are compact in problem* (6.25)*. In particular,* $\text{dom}(F_s)$ *is compact.*

The following primal-dual convergence theorem is an extension of Theorem 6.2, using the primal-dual relationships established in Proposition 6.8.

Theorem 6.5 (Primal-dual convergence – inner loop)**.** *Consider the primal problem* (6.28) *with its respective dual* (6.30)*. Moreover, consider problem* (6.25) *such that C_s*

is a nonempty compact convex set for each $s = 1, \dots, S$. Suppose there exists a last outer step in Algorithm 9 at iteration $k = \widehat{k}$. Then the following hold.

(i) The primal sequences $\{x^k\}_{k>\widehat{k}}$ and $\{x^{k+\frac{1}{2}}\}_{k>\widehat{k}}$ converge to a solution of the primal problem (6.28).

(ii) All accumulation points of the sequences $\{(x^k, x_T^{k+1})\}_{k>\widehat{k}}$ and $\{(x^{k+\frac{1}{2}}, x_T^{k+1})\}_{k>\widehat{k}}$ are solutions of primal problem (6.25).

(iii) Both $\{u^k\}_{k>\widehat{k}}$ and $\{w^{k+\frac{1}{2}}\}_{k>\widehat{k}}$ converge to the last center $\widehat{w}$, a solution of (6.30).

Proof. Dual convergence, that is, item (iii), follows from Theorem 6.2, in view of the equivalence between Algorithm 8 and Algorithm 9 of Proposition 6.8.

When the inner loop executes indefinitely, equation (6.32) reads

$$w^{k+\frac{1}{2}} = \widehat{w} + \widehat{t}(x^{k+\frac{1}{2}} - x^k). \tag{6.41}$$

Due to (6.41), there holds $x^{k+\frac{1}{2}} - x^k \to 0$. In view of Theorem 6.1, $x^k \to x^\star \in \mathcal{N}$, and thus $\mathrm{dom}(F) \ni x^{k+\frac{1}{2}} \to x^\star$, and $x^\star \in \mathrm{dom}(F)$. It remains to prove for item (i) that $x^\star$ is a primal solution. From the optimality condition (6.9), $g^k \in \partial h^k(w^{k+\frac{1}{2}})$, then from (6.2), $g^k - x^k \in \partial h(w^{k+\frac{1}{2}})$. In view of (6.10),

$$g^k = \frac{1}{\widehat{t}}(\widehat{w} - w^{k+\frac{1}{2}}) \to 0,$$

therefore $-x^\star \in \partial h(\widehat{w})$ for $x^\star \in \mathcal{N}$. Then, $x^\star$ is a solution to the primal problem (6.28) by applying Lemma 6.3 with $x^\star$ and $w^\star = \widehat{w}$. This proves item (i).

Regarding item (ii), for each $s = 1, \dots, S$, let $x_{sT}^\star \in \mathbb{R}^{(n'_s - n_s)}$ be a point where the minimum in (6.26) is attained for $x_s = x_s^\star$, that is, such that $F_s(x_s^\star) = f_s(x_s^\star, x_{sT}^\star)$, and $(x_s^\star, x_{sT}^\star) \in C_s$. This point exists because C_s is compact. Note that $(x^\star, x_T^\star)$ is thus a solution of problem (6.25) due to Proposition 6.6. Furthermore, in view of compactness of C_s, there exists $\overline{x}_{sT}$, and a subsequence $\{x_{sT}^{k_j+1}\}$ of $\{x_{sT}^{k+1}\}$, such that $x_{sT}^{k_j+1} \to \overline{x}_{sT}$ as $j \to +\infty$. From the definition of $(x^{k+\frac{1}{2}}, x_T^{k+1})$,

$$f_s(x_s^{k_j+\frac{1}{2}}, x_{sT}^{k_j+1}) + \langle \widehat{w}_s, x_s^{k_j+\frac{1}{2}} \rangle + \frac{\widehat{t}}{2}\|x^{k_j+\frac{1}{2}} - x^{k_j}\|^2 \leqslant f_s(x_s^\star, x_{sT}^\star) + \langle \widehat{w}_s, x_s^\star \rangle + \frac{\widehat{t}}{2}\|x_s^\star - x_s^{k_j}\|^2.$$

Taking the limit as $j \to +\infty$, multiplying by p_s, summing over $s = 1, \dots, S$, and using $(\widehat{w}, x^\star) \in \mathcal{N}^\perp \times \mathcal{N}$, $x^k - x^{k+\frac{1}{2}} \to 0$ and $x^k \to x^\star$ as $k \to +\infty$, it follows that

$$f(x^\star, \overline{x}_T) \leqslant f(x^\star, x_T^\star).$$

Thus, $(x^\star, \overline{x}_T)$ is a solution to problem (6.25), because $(x_s^\star, \overline{x}_{sT}) \in C_s$ for all $s = 1, \dots, S$, and $x^\star \in \mathcal{N}$.

□

Convergence of loop of outer steps: primal-dual case

Recall that when the acceptance test is satisfied infinitely many times, we define

$$K_O = \{k \in \mathbb{N} : u^{k+1} \text{ satisfies (6.5)}\},$$

and thus for all $k \in K_O$, $u^{k+1} = w^{k+1}$.

Dual convergence and dual linear rate of convergence follow from Theorem 6.3 and Theorem 6.4, respectively, in view of Proposition 6.8.

We write the error bound (6.17) in primal-dual terms: assume for any $\overline{h} \geqslant \inf_{w \in \mathcal{M}} F^*(w)$, there exists some $L > 0$ and $\delta > 0$ such that whenever $w \in \mathcal{M}$, and $F^*(w) \leqslant \overline{h}$, there exists $x \in \mathbb{R}^n$ such that

$$y \in B(0,\delta),\ x + y \in \mathcal{N},\ w \in \partial F(x) \implies \operatorname{dist}(w, W^\star) \leqslant L\|y\|. \tag{6.42}$$

This estimate is equivalent to (6.17), because for (w, x, y) satisfying (6.42), we have $x \in \partial F^*(w) = -\partial h(-w)$, where the equality follows from (6.40), and there exists $z \in \mathcal{N}$ such that $z = x + y$, and thus $y = -x + z \in \big(\partial h(-w) + \mathcal{N}\big) \cap B(0,\delta)$. In this manner, we can apply the change of variables $w \leftarrow -w$ to retrieve (6.17) for $\mathcal{M} = \mathcal{N}^\perp$.

Theorem 6.6 (Primal-dual convergence – outer loop)**.** *Consider the primal problem (6.28) with its respective dual (6.30). Moreover, consider problem (6.25) such that C_s is a nonempty compact convex set for each $s = 1, \ldots, S$. Suppose Algorithm 9 performs infinitely many outer steps, in such a way $\{\sigma_k\}_{k \in K_O} \subseteq [0, 1)$ stays bounded away from 1. Recall $\{t_k\}_{k \in K_O}$ is bounded away from 0 by construction. Then the following hold.*

(i) All accumulation points of the sequences $\{x^{k+1}\}_{k \in K_O}$ and $\{x^{k+\frac{1}{2}}\}_{kv}$ are solutions of the primal problem (6.28).

(ii) All accumulation points of the sequences $\{(x^{k+1}, x_T^{k+1})\}_{k \in K_O}$ and $\{(x^{k+\frac{1}{2}}, x_T^{k+1})\}_{k \in K_O}$ are solutions of primal problem (6.25).

(iii) Both $\{w^{k+1}\}_{k \in K_O}$ and $\{w^{k+\frac{1}{2}}\}_{k \in K_O}$ converge to a solution of the dual problem (6.30).

In addition, if (6.42) is satisfied, then the dual linear and superlinear convergence rates of Theorem 6.4 hold.

Proof. Due to Proposition 6.8, Theorem 6.3 is valid: $\{w^{k+1}\}_{k \in K_O}$ and $\{w^{k+\frac{1}{2}}\}_{k \in K_O}$ converge to a solution $w^\star$ of (6.30).

With respect to the primal sequences, both $\{x^{k+\frac{1}{2}}\}_{k \in K_O}$ and $\{x^{k+1}\}_{k \in K_O} \subseteq \mathcal{N}$ have nonempty sets of accumulation points. Indeed, in view of Remark 6.4, since

$\mathrm{dom}(F) \supseteq \{x^{k+\frac{1}{2}}\}_{k\in K_O}$ is compact, then there exists $M > 0$, such that for all $k \in K_O$, $\|x^{k+\frac{1}{2}}\| \leqslant M$, and

$$\begin{aligned}\|x^{k+1}\| &\leqslant \|x^{k+1} - x^1\| + \|x^1\| \\ &= \|P_{\mathcal{N}}[x^{k+\frac{1}{2}}] - P_{\mathcal{N}}[x^{\frac{1}{2}}]\| + \|x^1\| \\ &\leqslant \|x^{k+\frac{1}{2}} - x^{\frac{1}{2}}\| + \|x^1\| \\ &\leqslant \|x^{k+\frac{1}{2}}\| + \|x^{\frac{1}{2}}\| + \|x^1\| \\ &\leqslant M + \|x^{\frac{1}{2}}\| + \|x^1\|,\end{aligned}$$

where in the first and fourth line we use the triangle inequality, in the second line we use (6.33), and in the third line we use nonexpansiveness of the projection operator. Hence, both $\{x^{k+\frac{1}{2}}\}_{k\in K_O}$ and $\{x^{k+1}\}_{k\in K_O} \subseteq \mathcal{N}$ are bounded sequences.

Using (6.32), since $t_k \geqslant t_{\min}$ and $w^{k+\frac{1}{2}} - w^k \to w^\star - w^\star = 0$ as $K_O \ni k \to +\infty$, then

$$\|x^k - x^{k+\frac{1}{2}}\| \leqslant \frac{1}{t_{\min}}\|w^{k+\frac{1}{2}} - w^k\| \to 0,$$

and $\{x^{k+\frac{1}{2}}\}_{k\in K_O}$ and $\{x^k\}_{k\in K_O}$ have the same accumulation points. Moreover, this also implies together with (6.8) and (6.10), $g^k \to 0$ for $K_O \ni k \to +\infty$.

Let $\overline{x} \in \mathcal{N}$ and $\{x^{k_j+1}\}_j$ be a subsequence of $\{x^{k+1}\}_{k\in K_O}$, such that $x^{k_j+1} \to \bar{x}$ as $j \to +\infty$. From (6.9), $g^{k+1} \in \partial h^{k+1}(w^{k+\frac{3}{2}})$, then from (6.2) it holds

$$g^{k_j+1} \in \partial h(w^{k_j+\frac{3}{2}}) + x^{k_j+1} \implies 0 \in \partial h(w^\star) + \overline{x}.$$

Thus, the result in item (i) follows from Lemma 6.3, by taking $w = w^\star$, and $x^\star = \overline{x}$.

For item (ii), first note that $w^{k+\frac{1}{2}} - w^k \to 0$ as $K_O \ni k \to +\infty$ implies $t_k(x^{k+\frac{1}{2}} - x^k) \to 0$. Since $\{(x_s^{k+\frac{1}{2}}, x_{sT}^{k+1})\} \subseteq C_s$, take any limit point $(\overline{x}_s, \overline{x}_{sT}) \in C_s$ of the sequence $\{(x_s^{k+\frac{1}{2}}, x_{sT}^{k+1})\}_{k\in K_O}$, and a subsequence such that $(x_s^{k_j+\frac{1}{2}}, x_{sT}^{k_j+1}) \to (\overline{x}_s, \overline{x}_{sT})$ as $j \to +\infty$. Next, take the limit (up to the corresponding subsequence) in the optimality conditions of problem (6.31)

$$0 \in \partial f_s\left(x_s^{k+\frac{1}{2}}, x_{sT}^{k+1}\right) + N_{C_s}\left(x_s^{k+\frac{1}{2}}, x_{sT}^{k+1}\right) + \begin{pmatrix} w_s^k \\ 0 \end{pmatrix} + t_k \begin{pmatrix} x_s^{k+1} - x_s^k \\ 0 \end{pmatrix},$$

to obtain

$$0 \in \partial f_s(\overline{x}_s, \overline{x}_{sT}) + N_{C_s}(\overline{x}_s, \overline{x}_{sT}) + \begin{pmatrix} \overline{w_s} \\ 0 \end{pmatrix},$$

for $s = 1, \ldots, S$. Therefore, $0 \in \partial f(\overline{x}, \overline{x}_T) + N_{\mathcal{C}}(\overline{x}, \overline{x}_T) + \mathcal{N}^\perp \times \{0\}$, and in view of [30, CH. III, Proposition 5.3.1], $0 \in \partial f(\overline{x}, \overline{x}_T) + N_{(\mathcal{N}\times\{0\})\cap\mathcal{C}}(\overline{x}, \overline{x}_T)$. Since $f \in \overline{\mathrm{conv}}(\mathbb{R}^n)$ and $\mathcal{C}$ is closed, it follows from Theorem 1.1 that $(\overline{x}, \overline{x}_T)$ is a solution of primal problem (6.25). $\square$

6.3 Final remarks

In this chapter, we investigated a scenario-based decomposition method for convex multistage stochastic optimization problems. The central idea of the method is to apply an approximate proximal point algorithm to the dual formulation of the problem, and use a relative-error criteria to determine the quality of the candidate point. As in Chapter 5, we model the dual objective function $h + i_{\mathcal{N}^\perp}$ by using a linear surrogate function for the indicator $i_{\mathcal{N}^\perp}$. The advantage of this simple model is it allows decomposing the resulting model h^k for different scenario realizations, just as the Progressive Hedging algorithm. Other common models, such as a linear-by-parts function, prevent decomposition.

The relative-error acceptance test is verified in every iteration. When it is not satisfied, inner iterations are performed with constant parameters in order to improve the linear model of $i_{\mathcal{N}^\perp}$. Such step corresponds to a forward-backward step on the primal space, applied to the Fenchel dual of the dual proximal point subproblem. The method is shown to generate primal-dual sequences that converge to solutions to the primal and dual problem, respectively. When the acceptance test holds infinitely many times, the dual sequence converges with linear speed.

The relationship between DEFBAL and BPHA of Section 6.1.2 suggests that the acceptance test of DEFBAL can be viewed as a *super serious* step of the BPHA. This is confirmed by the preliminary numerical experiments (not shown in this thesis) performed for small linear stochastic optimization problems: DEFBAL executes fewer outer steps than BPH declares serious steps. Further tests will be performed in order to measure any possible advantage of DEFBAL over PHA/BPHA for problems of large size.

One possible direct extension of DEFBAL and BPHA for convex multistage stochastic optimization is to risk-averse problems. The formulations we examined were risk-neutral, meaning the objective function is an expected cost value. For risk-averse problems that use the conditional value-at-risk (CVaR) risk-measure,

$$\mathrm{CVaR}_\alpha[X] = \min_{u\in\mathbb{R}} \left\{ u + \frac{1}{1-\alpha}\mathbb{E}\big[\max\big(0, X-u\big)\big] \right\}, \tag{6.43}$$

for a random variable X and some level $\alpha \in (0,1)$. The CVaR_α of X, in simple words, corresponds to the expectation of X over the $\alpha-$tail of its distribution. For problem (5.1), a risk-averse formulation would have the following objective function for some $\lambda \in [0,1]$:

$$f_{\mathtt{obj}}(x) = \lambda \mathrm{CVaR}_\alpha[f(x)] + (1-\lambda)\mathbb{E}[f(x)].$$

As explained in [138], the objective function can still be decomposed for different scenarios, by creating S copies of the variable u in (6.43), and interpreting it as a first-stage decision

variable, that is, attaching to the nonanticipativity space the constraints $u = u_s$ for $s = 1, \ldots, S$. Therefore,

$$f_{\texttt{obj}}(x) = \sum_{s=1}^{S} p_s \left[\lambda \left(u_s + \frac{1}{1-\alpha} \max \left(0, f_s(x_s) - u_s\right) \right) + (1-\lambda) f_s(x_s) \right]$$

The decomposable feature of the objective function is thus inherited to the proximal subproblems of Algorithm 6 and Algorithm 7.

7 Conclusion and future work

In this work, we investigated decomposition methods for stochastic optimization problems. More specifically, for a convex multistage stochastic programming problem, we applied an approximate proximal point algorithm to its dual formulation, by means of modeling the nonanticipative constraints with a linear term. Using this model allows us to decompose the problem for different scenarios in the proximal subproblems for the model function, since the coupling constraints are dealt with a separable model. These subproblems correspond to minimize a quadratic regularization of the dual model.

We presented two approaches developed for convex problems, which can be deemed as variants of the Progressive Hedging algorithm. The difference between the two is how we judge the quality of the generated iterates: in Chapter 5, we proposed a bundle-like descent condition that measures the descent provided by the iterates when compared to the decrease predicted by the model, while in Chapter 6 we formulated a relative-error condition that measures dual feasibility and accuracy of the model at the current iterate. In both cases, when the acceptance test is satisfied infinitely many times, we proved that the iterates converge with linear rate to a solution to the problem. Otherwise, when the acceptance test holds one last iteration, we showed that the tail of iterates converges to a solution of the problem with linear rate as well. Furthermore, the acceptance condition of DEFB(AL) in Chapter 6 can be interpreted as a *super serious* step condition for the one of BPHA in Chapter 5, meaning that the former is at least as strict as the latter.

The BPHA approach can be extended to a more general setting, based on the Elicited Progressive Decoupling Algorithm of [132], mentioned in Chapter 4. This method allows giving different weights to primal and dual progress, by modifying the proximal parameter using an *elicitation* parameter. Incorporating the proximal bundle descent test would also potentially allow extending the BPHA ideas to nonconvex problems.

Regarding DEFBAL, its essential ideas can be applied to a more general setting. For convex composite optimization problems, ongoing work shows that the DEFBAL approach leads us to a method resembling the prominent Alternating Direction Method of Multipliers with a condition to accept or reject candidate points. More precisely, consider the minimization problem

$$\min_{x\in\mathbb{R}^n} f(x) + g(Mx), \tag{7.1}$$

for $f \in \overline{\text{conv}}(\mathbb{R}^n)$, $g \in \overline{\text{conv}}(\mathbb{R}^m)$, and a matrix $M \in \mathbb{R}^{m\times n}$, with the corresponding Fenchel dual problem given by

$$\min_{p\in\mathbb{R}^n} f^*(-M^\top p) + g^*(p). \tag{7.2}$$

Instead of applying the proximal point algorithm to (7.2), that is, solving

$$\min_{p\in\mathbb{R}^n}\left\{f^*(-M^\top p)+g^*(p)+\frac{1}{2c_k}\|p-p^k\|^2\right\}, \tag{7.3}$$

we apply the forward-backward method to its Fenchel dual, that is, we iteratively solve

$$\min_{x\in\mathbb{R}^n}\left(\left(f^*\circ\left(-M^\top\right)\right)(\cdot)+\frac{1}{2c_k}\|\cdot-p^k\|^2\right)^*(-x)+g(x). \tag{7.4}$$

To evaluate the quality of the solution of (7.4) we can either use an absolute-error or a relative-error acceptance test, creating two possibilities in each iteration: an outer step is performed, meaning that the acceptance test is satisfied, or an inner step is performed, that is, the acceptance test is not true for the current iterate. Note that the method described in Chapter 6 is a special case of (7.1) by taking $g=i_{\mathcal{N}}$, causing the $g-$proximal step to be simpler, it would merely be a projection onto $\mathcal{N}$. This line of research is currently being developed, and can be compared to [154], where extra acceptance tests for Douglas-Rachford (and consequently, for ADMM) are examined.

The convergence analysis in the context of stochastic optimization of the serious steps of the BPHA, and the outer steps of DEFBAL, fit in a more general framework for methods of descent, the one presented in Chapter 3. Actually, this framework works beyond convexity: for weakly convex problems, the basic ingredients to achieve convergence to critical points are: (1) a sufficient descent condition for the objective function throughout the iterates, and (2) an estimate of a subgradient using the step size, namely, the difference between two consecutive iterates. These two conditions are satisfied by proximal bundle methods in weakly convex optimization, due to the fact that weak convexity represent a "harmless" form of nonconvexity, still capturing a wide range of modern applications.

In Chapter 4, we studied the Douglas-Rachford splitting method for weakly convex optimization, obtaining global convergence to critical points, and local linear rates of convergence under common regularity assumptions. The analysis is a byproduct of the unifying framework of Chapter 3, where the sufficient descent condition and the subgradient estimate of the previous paragraph are satisfied by a merit function that represents the primal objective function in the dual space. Current ongoing work suggests that DRS can be applied to weakly convex stochastic optimization problems in the primal space, yielding a variant of the Progressive Hedging algorithm with a relaxed projection step. This proposal consists of a quadratic penalization method for the nonanticipativity constraints, and thus the limit of the generated sequences are critical points of the penalized problem. We are currently developing a convergence analysis to critical points of the original non-penalized problem, and naturally, avoid saddle points to obtain local minimizers, as observed in the final remarks of Chapter 4.

The methods of $\varepsilon-$subgradient descent for nonconvex optimization complying with (3.13a)–(3.13b) of Section 3.3 could be considered to define a discretization of a solution trajectory to a dynamical system in continuous time, such as in [155, 156, 157, 158] and references ibid. The direction that model-based methods Section 3.3 take, given by (3.13a), is a subgradient of the model function at the next iterate, and thus after transporting, it is also an $\varepsilon-$subgradient of the convexification of the objective function at the current iterate, as Proposition 3.3(ii) shows. This fact suggests the study of the continuous-time version of

$$\frac{x^{k+1} - x^k}{t_k} \in -\partial_{\varepsilon_k} f(x^k),$$

namely, the dynamical system for the $\varepsilon-$subdifferential

$$x'(t) \in -\partial_{\varepsilon(t)} f\big(x(t)\big). \tag{7.5}$$

A fundamental step to study continuous time systems consists on characterizing the functions that strictly decrease alongside the trajectories $x(t)$. In turn, a chain rule for the subdifferential valid for such trajectories allows proving a sufficient descent property in continuous time [156]. In this way, we would need a chain rule for the $\varepsilon-$subdifferential valid for trajectories satisfying (7.5), in order to obtain a similar descent properties for methods of $\varepsilon-$descent. Research in this direction would need to use the approximate subdifferential introduced in Section 1.3.2 for the weakly convex case. In particular, we would analyze proximal bundle methods in continuous time.

Bibliography

1 TROPP, J. A. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, IEEE, v. 52, n. 3, p. 1030–1051, 2006. Cited on page 16.

2 DONOHO, D. L. Compressed sensing. *IEEE Transactions on information theory*, IEEE, v. 52, n. 4, p. 1289–1306, 2006. Cited on page 16.

3 CONEJO, A. J.; CASTILLO, E.; MINGUEZ, R.; GARCIA-BERTRAND, R. *Decomposition techniques in mathematical programming: engineering and science applications*. Germany: Springer Science & Business Media, 2006. Cited on page 16.

4 LEMARÉCHAL, C.; SAGASTIZÁBAL, C.; PELLEGRINO, F.; RENAUD, A. Bundle methods applied to the unit-commitment problem. In: SPRINGER. *System Modelling and Optimization: Proceedings of the Seventeenth IFIP TC7 Conference on System Modelling and Optimization, 1995*. Netherlands, 1996. p. 395–402. Cited on page 16.

5 BELLONI, A.; LIMA, A. D. S.; MACEIRA, M. P.; SAGASTIZÁBAL, C. A. Bundle relaxation and primal recovery in unit commitment problems. The brazilian case. *Annals of Operations Research*, Springer, v. 120, p. 21–44, 2003. Cited on page 16.

6 SAGASTIZÁBAL, C. Divide to conquer: Decomposition methods for energy optimization. *Math. Program.*, v. 134, n. 1, p. 187–222, 2012. Cited 3 times on pages 16, 117, and 138.

7 GOLLMER, R.; NOWAK, M. P.; RÖMISCH, W.; SCHULTZ, R. Unit commitment in power generation–a basic model and some extensions. *Annals of Operations Research*, Springer, v. 96, p. 167–189, 2000. Cited on page 16.

8 TAKRITI, S.; BIRGE, J. R.; LONG, E. A stochastic model for the unit commitment problem. *IEEE Transactions on Power Systems*, IEEE, v. 11, n. 3, p. 1497–1508, 1996. Cited on page 16.

9 DANTZIG, G. B.; WOLFE, P. Decomposition principle for linear programs. *Operations research*, INFORMS, v. 8, n. 1, p. 101–111, 1960. Cited on page 16.

10 BENDERS, J. Partitioning procedures for solving mixed-variables programming problems. *Numer. Math*, v. 4, n. 1, p. 238–252, 1962. Cited on page 16.

11 LIONS, P. L.; MERCIER, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, v. 16, n. 6, p. 964–979, dez. 1979. ISSN 0036-1429, 1095-7170. Cited 6 times on pages 16, 95, 96, 97, 118, and 138.

12 BOYD, S.; PARIKH, N.; CHU, E.; PELEATO, B.; ECKSTEIN, J. et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, Now Publishers, Inc., v. 3, n. 1, p. 1–122, 2011. Cited 2 times on pages 16 and 95.

13 LEMARÉCHAL, C. Constructing bundle methods for convex optimization. In: *Mathematics Studies*. North-Holland: Elsevier, 1986. v. 129, p. 201–240. Cited on page 16.

14 KIWIEL, K. C. Approximations in proximal bundle methods and decomposition of convex programs. *Journal of Optimization Theory and applications*, Springer, v. 84, n. 3, p. 529–548, 1995. Cited on page 16.

15 EVERETT III, H. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research*, INFORMS, v. 11, n. 3, p. 399–417, 1963. Cited on page 16.

16 MARTINET, B. Régularisation d'inéquations variationnelles par approximations successives. *Revue Francaise d'informatique et de Recherche operationelle*, v. 4, p. 154–159, 1970. Cited 2 times on pages 16 and 95.

17 ROCKAFELLAR, R. T. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, SIAM, v. 14, n. 5, p. 877–898, 1976. Cited 8 times on pages 16, 45, 47, 48, 95, 138, 151, and 156.

18 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. *Math. Program.*, Springer, v. 137, n. 1-2, p. 91–129, 2013. Cited 5 times on pages 17, 40, 65, 66, and 94.

19 ROBINSON, S. M. Linear convergence of epsilon-subgradient descent methods for a class of convex functions. *Math. Program.*, Springer, v. 86, n. 1, p. 41–50, 1999. Cited 7 times on pages 17, 33, 40, 66, 81, 90, and 97.

20 BOŢ, R. I.; DAO, M. N.; LI, G. Inertial proximal block coordinate method for a class of nonsmooth sum-of-ratios optimization problems. *SIAM Journal on Optimization*, SIAM, v. 33, n. 2, p. 361–393, 2023. Cited on page 17.

21 FRANKEL, P.; GARRIGOS, G.; PEYPOUQUET, J. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, Springer, v. 165, p. 874–900, 2015. Cited 2 times on pages 17 and 97.

22 ATENAS, F.; SAGASTIZÁBAL, C.; SILVA, P. J.; SOLODOV, M. A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods. *SIAM Journal on Optimization*, SIAM, v. 33, n. 1, p. 89–115, 2023. Cited 10 times on pages 17, 18, 29, 36, 43, 63, 93, 94, 97, and 115.

23 ATENAS, F.; SAGASTIZÁBAL, C. A bundle-like progressive hedging algorithm. *J. Convex Anal.*, v. 30, n. 2, p. 453–479, 2023. Cited 8 times on pages 17, 18, 59, 63, 87, 94, 112, and 117.

24 ROCKAFELLAR, R. T.; WETS, R. J.-B. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of operations research*, INFORMS, v. 16, n. 1, p. 119–147, 1991. Cited 5 times on pages 17, 58, 59, 117, and 136.

25 SOLODOV, M. V.; SVAITER, B. F. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, Springer, v. 7, n. 4, p. 323–345, 1999. Cited 7 times on pages 17, 48, 49, 147, 148, 149, and 152.

26 RUSZCZYŃSKI, A. On convergence of an Augmented Lagrangian decomposition method for sparse convex optimization. *Mathematics of Operations Research*, INFORMS, v. 20, n. 3, p. 634–656, 1995. Cited on page 18.

27 ROSA, C. H.; RUSZCZYŃSKI, A. On Augmented Lagrangian decomposition methods for multistage stochastic programs. *Annals of Operations Research*, Springer, v. 64, p. 289–309, 1996. Cited 2 times on pages 18 and 58.

28 THEMELIS, A.; PATRINOS, P. Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization*, SIAM, v. 30, n. 1, p. 149–181, 2020. Cited 14 times on pages 18, 96, 97, 98, 99, 100, 102, 103, 104, 105, 106, 107, 110, and 115.

29 ROCKAFELLAR, R. T.; WETS, R. J.-B. *Variational analysis*. Germany: Springer Science & Business Media, 2009. v. 317. Cited 16 times on pages 19, 20, 21, 22, 24, 26, 27, 28, 30, 33, 47, 56, 76, 99, 116, and 119.

30 HIRIART-URRUTY, J.; LEMARÉCHAL, C. *Convex Analysis and Minimization Algorithms I and II*. Germany: Springer-Verlag, 1996. (Grundlehren der mathematischen Wissenschaften, 305 and 306). Cited 12 times on pages 22, 31, 33, 50, 53, 65, 66, 76, 119, 131, 132, and 166.

31 CLARKE, F. H. *Optimization and nonsmooth analysis*. Wiley New York: SIAM, 1990. Cited on page 24.

32 NGAI, H. V.; LUC, D. T.; THÉRA, M. Approximate convex functions. *J. Nonlinear Convex Anal.*, v. 1, n. 2, p. 155–176, 2000. Cited on page 24.

33 LUO, Z.-Q.; TSENG, P. Error bounds and convergence analysis of feasible descent methods: a general approach. *Ann. Oper. Res.*, Springer, v. 46, n. 1, p. 157–178, 1993. Cited 10 times on pages 26, 31, 36, 37, 40, 66, 67, 72, 84, and 97.

34 DUCHI, J. C.; RUAN, F. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Information and Inference: A Journal of the IMA*, Oxford University Press, v. 8, n. 3, p. 471–529, 2019. Cited on page 29.

35 CANDES, E. J.; LI, X.; SOLTANOLKOTABI, M. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, IEEE, v. 61, n. 4, p. 1985–2007, 2015. Cited on page 29.

36 LING, S.; STROHMER, T. Self-calibration and biconvex compressive sensing. *Inverse Problems*, IOP Publishing, v. 31, n. 11, p. 115002, 2015. Cited on page 29.

37 DAVIS, D.; DRUSVYATSKIY, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, SIAM, v. 29, n. 1, p. 207–239, 2019. Cited 6 times on pages 29, 111, 112, 113, 114, and 115.

38 NORKIN, V. Generalized-differentiable functions. *Cybernetics*, Springer, v. 16, n. 1, p. 10–12, 1980. Cited on page 29.

39 MIFFLIN, R. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control Optim.*, Society for Industrial and Applied Mathematics, v. 15, n. 6, p. 959–972, 1977. Cited on page 29.

40 DAVIS, D.; GRIMMER, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM J. Optim.*, SIAM, v. 29, n. 3, p. 1908–1930, 2019. Cited on page 29.

41 DANIILIDIS, A.; MALICK, J. Filling the gap between lower$-\mathcal{C}^1$ and lower$-\mathcal{C}^2$ functions. *J. Convex Anal.*, v. 12, n. 2, p. 315–329, 2005. Cited 3 times on pages 29, 31, and 94.

42 WRIGHT, S. *Some Perspectives on Nonconvex Optimization*. 2020. Available at: <http://helper.ipam.ucla.edu/publications/lco2020/lco2020_16234.pdf>. IPAM workshop on Intersections between Control, Learning and Optimization. Cited on page 30.

43 SHAPIRO, A. On a class of nonsmooth composite functions. *Math. Oper. Res.*, INFORMS, v. 28, n. 4, p. 677–692, 2003. Cited on page 30.

44 LEWIS, A.; WRIGHT, S. A proximal method for composite minimization. *Math. Program.*, Springer, v. 158, n. 1-2, p. 501–546, 2015. Cited 6 times on pages 30, 66, 76, 78, 79, and 113.

45 SAGASTIZÁBAL, C. Composite proximal bundle method. *Math. Program.*, Springer-Verlag, v. 140, n. 1, p. 189–233, 2013. ISSN 0025-5610. Cited 3 times on pages 30, 76, and 137.

46 DRUSVYATSKIY, D.; PAQUETTE, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.*, Springer, v. 178, n. 1, p. 503–558, 2019. Cited 2 times on pages 30 and 112.

47 IZMAILOV, A.; SOLODOV, M. *Newton-type methods for optimization and variational problems*. Germany: Springer, Cham, 2014. (Springer Series in Operations Research and Financial Engineering). Cited 4 times on pages 31, 35, 37, and 85.

48 BRØNDSTED, A.; ROCKAFELLAR, R. T. On the subdifferentiability of convex functions. *Proceedings of the American Mathematical Society*, v. 16, n. 4, p. 605–611, 1965. Cited on page 33.

49 PENOT, J.-P. Subdifferential calculus without qualification assumption. *Journal of Convex Analysis*, Heldermann Verlag, v. 3, p. 207–220, 1996. Cited 2 times on pages 34 and 35.

50 PANG, J.-S. Error bounds in mathematical programming. *Math. Program.*, v. 79, p. 299–332, 1997. Cited on page 35.

51 FACCHINEI, F.; PANG, J.-S. *Finite-dimensional variational inequalities and complementarity problems*. USA: Springer Science & Business Media, 2007. Cited 3 times on pages 35, 36, and 37.

52 ZHOU, Z.; SO, A. M.-C. A unified approach to error bounds for structured convex optimization problems. *Math. Program.*, v. 165, n. 2, p. 689–728, 2017. Cited on page 35.

53 KURDYKA, K. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, Cellule MathDoc/CEDRAM, v. 48, n. 3, p. 769–783, 1998. Cited 5 times on pages 36, 38, 39, 40, and 97.

54 BOLTE, J.; DANIILIDIS, A.; LEWIS, A.; SHIOTA, M. Clarke subgradients of stratifiable functions. *SIAM J. Optim.*, SIAM, v. 18, n. 2, p. 556–572, 2007. Cited 4 times on pages 36, 39, 40, and 97.

55 DRUSVYATSKIY, D.; IOFFE, A.; LEWIS, A. Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria. *Math. Program.*, Springer Science and Business Media LLC, v. 185, n. 1-2, p. 357–383, 2019. Cited 6 times on pages 36, 42, 74, 76, 77, and 78.

56 DRUSVYATSKIY, D.; MORDUKHOVICH, B. S.; NGHIA, T. T. A. Second-order growth, tilt stability, and metric regularity of the subdifferential. *J. Convex Anal. 21(4), 11651192*, 2014. Cited 2 times on pages 36 and 42.

57 ARTACHO, F. A.; GEOFFROY, M. H. Characterization of metric regularity of subdifferentials. *J. Convex Anal.*, Heldermann Verlag, v. 15, n. 2, p. 365–380, 2008. Cited 2 times on pages 36 and 41.

58 ZHANG, R.; TREIMAN, J. Upper-lipschitz multifunctions and inverse subdifferentials. *Nonlinear Anal., Theory Methods Appl.*, Elsevier, v. 24, n. 2, p. 273–286, 1995. Cited on page 36.

59 SOLODOV, M.; TSENG, P. Modified projection-type methods for monotone variational inequalities. *SIAM J. Control Optim.*, v. 34, p. 1814–1830, 1996. Cited on page 37.

60 SOLODOV, M. Convergence rate analysis of iterative algorithms for solving variational inequality problems. *Math. Program.*, v. 96, p. 513–528, 2003. Cited on page 37.

61 FISCHER, A. Local behavior of an iterative framework for generalized equations with nonisolated solutions. *Math. Program.*, Springer, v. 94, n. 1, p. 91–124, 2002. Cited on page 37.

62 IOFFE, A. An invitation to tame optimization. *SIAM Journal on Optimization*, SIAM, v. 19, n. 4, p. 1894–1917, 2009. Cited on page 38.

63 LOJASIEWICZ, S. Sur les trajectoires du gradient dune fonction analytique. *Seminari di geometria*, v. 1983, p. 115–117, 1982. Cited on page 38.

64 _____. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, v. 117, p. 87–89, 1963. Cited on page 38.

65 BOLTE, J.; DANIILIDIS, A.; LEY, O.; MAZET, L. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, v. 362, n. 6, p. 3319–3363, 2010. Cited 2 times on pages 38 and 42.

66 ATTOUCH, H.; BOLTE, J.; REDONT, P.; SOUBEYRAN, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of operations research*, INFORMS, v. 35, n. 2, p. 438–457, 2010. Cited on page 39.

67 BOLTE, J.; NGUYEN, T. P.; PEYPOUQUET, J.; SUTER, B. W. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, Springer, v. 165, n. 2, p. 471–507, 2017. Cited 4 times on pages 40, 41, 43, and 97.

68 DRUSVYATSKIY, D.; LEWIS, A. S. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, INFORMS, v. 43, n. 3, p. 919–948, 2018. Cited 3 times on pages 40, 41, and 74.

69 KRUGER, A. Y.; LUKE, D. R.; THAO, N. H. Set regularities and feasibility problems. *Mathematical Programming*, Springer, v. 168, p. 279–311, 2018. Cited on page 40.

70 ROBINSON, S. M. *Some continuity properties of polyhedral multifunctions*. USA: Springer Berlin Heidelberg, 1981. 206–214 p. Cited 2 times on pages 40 and 49.

71 LUQUE, F. J. Asymptotic convergence analysis of the proximal point algorithm. *SIAM Journal on Control and Optimization*, SIAM, v. 22, n. 2, p. 277–293, 1984. Cited 2 times on pages 40 and 48.

72 DONTCHEV, A. L.; ROCKAFELLAR, R. T. Regularity and conditioning of solution mappings in variational analysis. *Set-Valued Analysis*, Springer, v. 12, n. 1-2, p. 79–109, 2004. Cited on page 41.

73 YE, J. J.; YUAN, X.; ZENG, S.; ZHANG, J. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. *Set-Valued and Variational Analysis*, Springer, v. 29, n. 4, p. 803–837, 2021. Cited on page 42.

74 BAI, S.; LI, M.; LU, C.; ZHU, D.; DENG, S. The equivalence of three types of error bounds for weakly and approximately convex functions. *Journal of Optimization Theory and Applications*, Springer, v. 194, n. 1, p. 220–245, 2022. Cited on page 42.

75 LI, G.; PONG, T. K. Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, Springer, v. 18, n. 5, p. 1199–1232, 2018. Cited 2 times on pages 42 and 97.

76 LI, G.; MORDUKHOVICH, B. S.; NGHIA, T. T.; PHM, T. Error bounds for parametric polynomial systems with applications to higher-order stability analysis and convergence rates. *Mathematical Programming*, Springer, v. 168, p. 313–346, 2018. Cited on page 43.

77 IOFFE, A. D. Metric regularity survey part II. Applications. *Journal of the Australian Mathematical Society*, Cambridge University Press, v. 101, n. 3, p. 376–417, 2016. Cited on page 43.

78 _____. Metric regularity survey part I. Theory. *Journal of the Australian Mathematical Society*, Cambridge University Press, v. 101, n. 2, p. 188–243, 2016. Cited on page 43.

79 MINTY, G. J. Monotone (nonlinear) operators in Hilbert space. 1962. Cited on page 46.

80 HOHEISEL, T.; LABORDE, M.; OBERMAN, A. On proximal point-type algorithms for weakly convex functions and their connection to the backward Euler method. *Optimization Online*. Cited on page 47.

81 AUSLENDER, A. Numerical methods for nondifferentiable convex optimization. *Nonlinear Analysis and Optimization*, Springer, p. 102–126, 1987. Cited on page 50.

82 FUKUSHIMA, M.; MINE, H. A generalized proximal point algorithm for certain non-convex minimization problems. *International Journal of Systems Science*, Taylor & Francis, v. 12, n. 8, p. 989–1000, 1981. Cited on page 50.

83 CORREA, R.; LEMARÉCHAL, C. Convergence of some algorithms for convex minimization. *Mathematical Programming*, Springer, v. 62, p. 261–275, 1993. Cited 2 times on pages 50 and 52.

84 KELLEY JR, J. E. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics*, SIAM, v. 8, n. 4, p. 703–712, 1960. Cited on page 50.

85 BONNANS, J.; GILBERT, J.; LEMARÉCHAL, C.; SAGASTIZÁBAL, C. *Numerical Optimization: Theoretical and Practical Aspects*. Berlin, Germany: Springer, 2006. Second Edition. Cited 4 times on pages 50, 66, 72, and 119.

86 BONNANS, J.-F.; GILBERT, J. C.; LEMARÉCHAL, C.; SAGASTIZÁBAL, C. A. *Numerical optimization: theoretical and practical aspects*. Germany: Springer Science & Business Media, 2006. Cited on page 53.

87 ECKSTEIN, J. Splitting methods for monotone operators, with applications to parallel optimization – Ph.d. thesis, MIT. 1989. Cited 9 times on pages 54, 55, 56, 60, 95, 96, 97, 118, and 138.

88 DENTCHEVA, D.; MARTINEZ, G. Regularization methods for optimization problems with probabilistic constraints. *Mathematical Programming*, Springer, v. 138, n. 1-2, p. 223–251, 2013. Cited on page 59.

89 SPINGARN, J. E. Applications of the method of partial inverses to convex programming: decomposition. *Mathematical Programming*, Springer, v. 32, n. 2, p. 199–223, 1985. Cited on page 59.

90 RUSZCZYŃSKI, A.; SHAPIRO, A. *Stochastic Programming (Handbooks in Operations Research and Management Science, 10)*. Amsterdam: Elsevier, 2003. Cited on page 59.

91 DRUSVYATSKIY, D.; DAVIS, D. Subgradient methods under weak convexity and tame geometry. *SIAG/OPT Views and News*, v. 28, n. 1, p. 1–10, 2020. Cited 2 times on pages 64 and 66.

92 KIWIEL, K. *Methods of descent for nondifferentiable optimization*. Berlin: Springer-Verlag, 1985. vi+362 p. Cited 2 times on pages 66 and 74.

93 BECK, A. *First-order methods in optimization*. USA: Society for Industrial and Applied Mathematics, 2017. (MOS-SIAM Series on Optimization). Cited 3 times on pages 66, 77, and 146.

94 NOLL, D. Convergence of non-smooth descent methods using the Kurdyka–Łojasiewicz inequality. *J. Optim. Theory Appl.*, Springer Science and Business Media LLC, v. 160, n. 2, p. 553–572, 2013. Cited on page 66.

95 KIWIEL, K. A phase I-phase II method for inequality constrained minimax problems. *Control. Cybern.*, v. 12, n. 1-2, p. 55–75, 1983. Cited on page 66.

96 _____. Efficiency of proximal bundle methods. *J. Optim. Theory Appl.*, Springer Science and Business Media LLC, v. 104, n. 3, p. 589–603, 2000. Cited on page 66.

97 DU, Y.; RUSZCZYŃSKI, A. Rate of convergence of the bundle method. *J. Optim. Theory Appl.*, Springer, v. 173, n. 3, p. 908–922, 2017. Cited on page 66.

98 DÍAZ, M.; GRIMMER, B. Optimal convergence rates for the proximal bundle method. *SIAM Journal on Optimization*, SIAM, v. 33, n. 2, p. 424–454, 2023. Cited on page 66.

99 MIFFLIN, R. A modification and extension of Lemaréchal's algorithm for nonsmooth minimization. *Math. Programming Stud.*, v. 17, p. 77–90, 1982. Cited on page 74.

100 MAKELA, M.; NEITTAANMAKI, P. *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*. Singapore: World Scientific: Singapore, 1992. Cited on page 74.

101 LUKŠAN, L.; VLČEK, J. Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. *J. Optim. Theory Appl.*, v. 2, p. 407–430, 2001. Cited on page 74.

102 FUDULI, A.; GAUDIOSO, M.; GIALLOMBARDO, G. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM J. Optim.*, v. 14, n. 3, p. 743–756, 2003. ISSN 1095-7189. Cited on page 74.

103 NOLL, D.; PROT, O.; RONDEPIERRE, A. A proximity control algorithm to minimize nonsmooth and nonconvex functions. *Pac. J. Optim.*, v. 4, n. 3, p. 569 – 602, 2008. Cited on page 74.

104 HARE, W.; SAGASTIZÁBAL, C. Computing proximal points of nonconvex functions. *Math. Program.*, v. 116, n. 1-2, p. 221–258, 2009. Cited 3 times on pages 74, 75, and 87.

105 _____. A redistributed proximal bundle method for nonconvex optimization. *SIAM J. Optim.*, v. 20, n. 5, p. 2442–2473, 2010. Cited 3 times on pages 74, 75, and 87.

106 HARE, W.; SAGASTIZÁBAL, C.; SOLODOV, M. A proximal bundle method for nonsmooth nonconvex functions with inexact information. *Comput. Optim. Appl.*, v. 63, p. 1–28, 2016. Cited on page 74.

107 BERTSEKAS, D. P. *Convex Optimization Algorithms*. USA: Athena scientific, 2015. ISBN 978-1-886529-28-1. Cited on page 79.

108 SOLODOV, M. Convergence analysis of perturbed feasible descent methods. *J. Optim. Theory Appl.*, v. 93, p. 337–353, 1997. Cited on page 84.

109 ROCKAFELLAR, R. T. Favorable classes of lipschitz continuous functions in subgradient optimization. WP-81-001, 1981. Cited on page 94.

110 POLIQUIN, R.; ROCKAFELLAR, R. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, v. 348, n. 5, p. 1805–1838, 1996. Cited on page 94.

111 COMBETTES, P. L.; PESQUET, J.-C. A Douglas–Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, v. 1, n. 4, p. 564–574, 2007. Cited on page 95.

112 CAI, J.-F.; OSHER, S.; SHEN, Z. Split Bregman methods and frame based image restoration. *Multiscale modeling & simulation*, SIAM, v. 8, n. 2, p. 337–369, 2010. Cited on page 95.

113 COMBETTES, P. L.; PESQUET, J.-C. Proximal splitting methods in signal processing. *Fixed-point algorithms for inverse problems in science and engineering*, Springer, p. 185–212, 2011. Cited on page 95.

114 GLOWINSKI, R.; OSHER, S. J.; YIN, W. *Splitting methods in communication, imaging, science, and engineering*. 1. ed. (2017): Springer. Cited on page 95.

115 SPINGARN, J. E. Partial inverse of a monotone operator. *Applied mathematics and optimization*, Springer, v. 10, n. 1, p. 247–265, 1983. Cited on page 95.

116 DAVIS, D.; YIN, W. Convergence rate analysis of several splitting schemes. *Splitting methods in communication, imaging, science, and engineering*, Springer, p. 115–163, 2016. Cited on page 95.

117 ECKSTEIN, J. A simplified form of block-iterative operator splitting and an asynchronous algorithm resembling the multi-block alternating direction method of multipliers. *Journal of Optimization Theory and Applications*, Springer, v. 173, n. 1, p. 155–182, 2017. Cited 2 times on pages 96 and 118.

118 MALITSKY, Y.; TAM, M. K. Resolvent splitting for sums of monotone operators with minimal lifting. *Mathematical Programming*, Springer, p. 1–32, 2022. Cited on page 96.

119 ECKSTEIN, J.; BERTSEKAS, D. P. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical programming*, Springer, v. 55, p. 293–318, 1992. Cited on page 96.

120 DAVIS, D.; YIN, W. Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions. *Mathematics of Operations Research*, INFORMS, v. 42, n. 3, p. 783–805, 2017. Cited on page 97.

121 DENG, W.; YIN, W. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, Springer, v. 66, p. 889–916, 2016. Cited on page 97.

122 PATRINOS, P.; STELLA, L.; BEMPORAD, A. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In: IEEE. *53rd IEEE Conference on Decision and Control*. USA, 2014. p. 4234–4239. Cited 2 times on pages 97 and 101.

123 _____. Forward-backward truncated newton methods for convex composite optimization. *arXiv preprint arXiv:1402.6655*, 2014. Cited on page 97.

124 LI, G.; PONG, T. K. Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical programming*, Springer, v. 159, p. 371–401, 2016. Cited 2 times on pages 97 and 112.

125 _____. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, SIAM, v. 25, n. 4, p. 2434–2460, 2015. Cited on page 97.

126 DAVIS, D. Convergence rate analysis of the forward-Douglas-Rachford splitting scheme. *SIAM Journal on Optimization*, SIAM, v. 25, n. 3, p. 1760–1786, 2015. Cited on page 97.

127 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*, SIAM, v. 28, n. 3, p. 2274–2303, 2018. Cited on page 100.

128 BERTSEKAS, D. P. Nonlinear programming. *Journal of the Operational Research Society*, Taylor & Francis, v. 48, n. 3, p. 334–334, 1997. Cited on page 100.

129 BÖHM, A.; WRIGHT, S. J. Variable smoothing for weakly convex composite functions. *Journal of optimization theory and applications*, Springer, v. 188, p. 628–649, 2021. Cited on page 109.

130 BAUSCHKE, H. H.; COMBETTES, P. L. et al. *Convex analysis and monotone operator theory in Hilbert spaces*. Germany: Springer, 2011. v. 408. Cited 2 times on pages 112 and 159.

131 ROCKAFELLAR, R. T. Progressive decoupling of linkages in optimization and variational inequalities with elicitable convexity or monotonicity. *Set-Valued and Variational Analysis*, Springer, v. 27, n. 4, p. 863–893, 2019. Cited on page 114.

132 SUN, J.; ZHANG, M. The elicited progressive decoupling algorithm: A note on the rate of convergence and a preliminary numerical experiment on the choice of parameters. *Set-Valued and Variational Analysis*, Springer, v. 29, n. 4, p. 997–1018, 2021. Cited 2 times on pages 114 and 169.

133 LIU, Y.; YIN, W. An envelope for Davis–Yin splitting and strict saddle-point avoidance. *Journal of Optimization Theory and Applications*, Springer, v. 181, p. 567–587, 2019. Cited on page 115.

134 DAVIS, D.; DRUSVYATSKIY, D. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, Springer, v. 22, n. 2, p. 561–606, 2022. Cited on page 116.

135 DAVIS, D.; DRUSVYATSKIY, D.; JIANG, L. Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality. *arXiv preprint arXiv:2108.11832*, 2021. Cited on page 116.

136 BURKE, J. V.; HOHEISEL, T. Epi-convergent smoothing with applications to convex composite functions. *SIAM Journal on Optimization*, SIAM, v. 23, n. 3, p. 1457–1479, 2013. Cited on page 116.

137 TSENG, P. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, v. 29, n. 1, p. 119–138, 1991. Cited 2 times on pages 117 and 138.

138 ROCKAFELLAR, R. T. Solving stochastic programming problems with risk measures by progressive hedging. *Set-Valued and Variational Analysis*, Springer, v. 26, n. 4, p. 759–768, 2018. Cited 2 times on pages 117 and 167.

139 ECKSTEIN, J.; WATSON, J.-P.; WOODRUFF, D. L. Projective hedging algorithms for distributed optimization under uncertainty. 2022. Cited 2 times on pages 118 and 157.

140 GLOWINSKI, R.; TALLEC, P. L. *Augmented Lagrangian Methods for the Solution of Variational Problems*. Philadelphia: Society for Industrial and Applied Mathematics, 1989. (Studies in Applied and Numerical Mathematics). Cited 2 times on pages 118 and 138.

141 ECKSTEIN, J.; SVAITER, B. F. A family of projective splitting methods for the sum of two maximal monotone operators. *Math. Program.*, v. 111, n. 1-2, p. 173–199, 2008. Cited on page 118.

142 _____. General projective splitting methods for sums of maximal monotone operators. *SIAM J. Control Optim.*, v. 48, n. 2, p. 787–811, 2009. ISSN 0363-0129. Cited on page 118.

143 COMBETTES, P. L.; ECKSTEIN, J. Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. *Math. Program.*, v. 126, n. 1–2, p. 645–672, 2018. Cited on page 118.

144 BAREILLES, G.; LAGUEL, Y.; GRISHCHENKO, D.; IUTZELER, F.; MALICK, J. Randomized progressive hedging methods for multi-stage stochastic programming. *Annals of Operations Research*, Springer Science and Business Media LLC, v. 295, n. 2, p. 535–560, sep 2020. Cited on page 118.

145 DOS SANTOS, M. L.; DA SILVA, E. L.; FINARDI, E. C.; GONÇALVES, R. E. Practical aspects in solving the medium-term operation planning problem of hydrothermal power systems by using the Progressive Hedging method. *Int. J. Electr. Power Energy Syst.*, v. 31, n. 9, p. 546–552, 2009. ISSN 0142-0615. Power Systems Computation Conference (PSCC) 2008. Cited on page 118.

146 GONÇALVES, R. E.; FINARDI, E. C.; DA SILVA, E. L. Applying different decomposition schemes using the progressive hedging algorithm to the operation planning problem of a hydrothermal system. *Electr. Power Syst. Res.*, v. 83, n. 1, p. 19–27, 2012. ISSN 0378-7796. Cited on page 118.

147 KNUEVEN, B.; MILDEBRATH, D.; MUIR, C.; SIIROLA, J. D.; WATSON, J.-P.; WOODRUFF, D. L. A parallel hub-and-spoke system for large-scale scenario-based optimization under uncertainty. *Optimization Online*, p. 11–8088, 2020. Cited on page 118.

148 ACKOOIJ, W.; SAGASTIZÁBAL, C. Constrained bundle methods for upper inexact oracles with application to joint chance constrained energy problems. *SIAM J. Optim.*, v. 24, n. 2, p. 733–765, 2014. Cited on page 124.

149 ACKOOIJ, W. van; BERGE, V.; OLIVEIRA, W. de; SAGASTIZÁBAL, C. Probabilistic optimization via approximate p-efficient points and bundle methods. *Computers & Operations Research*, v. 77, p. 177 – 193, 2017. ISSN 0305-0548. Cited on page 124.

150 IZMAILOV, A.; SOLODOV, M. *Otimização, volume 2: métodos computacionais*. Rio de Janeiro: IMPA, 2007. Cited on page 135.

151 FORTIN, M.; GLOWINSKI, R. *On Decomposition-Coordination Methods Using an Augmented Lagrangian*. Amsterdam: North-Holland, 1983. 97–146 p. (Studies in Mathematics and its Applications, v. 15). Cited on page 138.

152 GABAY, D. *Applications of the Method of Multipliers to Variational Inequalities*. Netherlands: Elsevier, 1983. 299–331 p. (Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, v. 15). Cited on page 138.

153 ROCKAFELLAR, R. T. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, INFORMS, v. 1, n. 2, p. 97–116, 1976. Cited on page 156.

154 ECKSTEIN, J.; YAO, W. Relative-error approximate versions of Douglas–Rachford splitting and special cases of the ADMM. *Mathematical Programming*, Springer, v. 170, p. 417–444, 2018. Cited on page 170.

155 RUSZCZYŃSKI, A. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optim. Lett.*, Springer, p. 1–11, 2020. Cited on page 171.

156 DAVIS, D.; DRUSVYATSKIY, D.; KAKADE, S.; LEE, J. D. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, Springer, v. 20, n. 1, p. 119–154, 2020. Cited on page 171.

157 BOLTE, J.; DANIILIDIS, A.; LEWIS, A. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, SIAM, v. 17, n. 4, p. 1205–1223, 2007. Cited on page 171.

158 BOŢ, R. I.; CSETNEK, E. R. A forward-backward dynamical approach to the minimization of the sum of a nonsmooth convex with a smooth nonconvex function. *ESAIM: Control, Optimisation and Calculus of Variations*, EDP Sciences, v. 24, n. 2, p. 463–477, 2018. Cited on page 171.