Accelerated Inexact First-Order Methods for Solving Nonconvex Composite Optimization Problems

A Dissertation
Presented to
The Academic Faculty

By

Weiwei Kong

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Operations Research

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
May 2021

Accelerated Inexact First-Order Methods for Solving Nonconvex Composite Optimization Problems

Approved by:

Dr. Renato D.C. Monteiro (Advisor)
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Arkadi Nemirovski
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Guanghui Lan
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Santanu S. Dey
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology

Date Approved: April 2, 2021

To my parents, my sister, and

the many friends and mentors that I have met along the way.

ACKNOWLEDGMENTS

First and foremost, I would like to extend my deepest gratitude towards my advisor, Renato D.C. Monteiro, who has provided me immense support throughout my Ph.D. journey. Without his invaluable insight, unrelenting guidance, and steadfast patience, this thesis would not be where it is today. I am also grateful to my collaborator Jefferson G. Melo for his countless conversations about research, life, and career advancement. Special thanks should go to Arkadi Nemirovski, whose numerous suggestions and discussions have immensely influenced this thesis's direction. I would also like to thank all my committee members, including my advisor, Renato D.C. Monteiro, Arkadi Nemirovski, Guanghui "George" Lan, Edmond Chow, and Santanu S. Dey.

Next, I would like to thank faculty members Craig Tovey and the late Shabbir Ahmed for their extensive support during my semesters as a graduate teaching assistant. Professor Tovey provided invaluable feedback on my problem formulations, and I have had many meaningful conversations with Professor Ahmed about career advice and exciting problems in polyhedral theory. I am also grateful for the support that I have received from the ISyE department as a whole. Particular thanks go out to Alan Erera, who has provided valuable guidance in my early years, and Amanda Ford, who has been especially helpful with my studies' administrative side.

The material in this thesis would not have been possible without the generous financial support of the ISyE department, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Institute for Data Engineering and Science (IDEaS), the Transdisciplinary Research Institute for Advancing Data Science (TRIAD), and the Johnson family. Specific grants include the Alexander Graham Bell Postgraduate Scholarship (#PGSD3-516700-2018), the IDEaS-TRIAD Research Scholarship, and the Thomas Johnson Fellowship.

I am incredibly grateful to Bill Cook, whose advanced course in optimization at the University of Waterloo was a driving factor for my decision to begin graduate studies at Georgia

Tech. I am also highly appreciative of his helpful career and research advice. From my summer internships at Google Research, I would like to thank my mentors Aranyak Mehta, D. Sivakumar, Nicolas Mayoraz, Walid Krishne, and my fellow collaborators Christopher Liaw, Steffen Rendle, and Li Zhang. Their insightful discussions have helped shape several of my research and career decisions.

I would now like to extend my gratitude to the many friends and colleagues I have made at Tech. A few of the fellow graduate studies that I would like to thank are Tyler Perini, Cyrus Rich, Andrew ElHabr, Adrian Rivera, Mohamed El Tonbari, Ramon Auad, Ian Herszterg, Edward Yuhang He, Reem Khir, Georgios Kotsalis, Digvijay Boob, Sarah Wiegreffe, and Zhehui Chen. Additionally, I would like to thank my fellow lab member, Jiaming Liang, and my office roommate, Alexander Stroh, for the many memorable conversations. A few of the members of the Georgia Tech Hapkido Club that I would like to thank are Graham Saunders, Jason Ngor Shing Yi, Andrew Schulz, Erik Anderson, Christian Girala, and my instructors Olivia Lodise, Mike Mackenzie, Hung Le, Joel Dunham, Christi Nakajima, Melissa Johnson, Matthieu Bloch, and Grandmaster Nils Onsager.

Additionally, I would like to thank several of my friends from Canada and abroad, including Jeffrey Negrea, Snow Murdock, Jamie Murdoch, Lawson Fulton, Robert Zimmerman, Johnew Zhang, Jess Zhang, Dmytro Korol, Ishan Patel, and Shashanth Shetty. Thank you all for the valuable friendships over the years.

Finally, to my parents Xiaofen "Cindy" Chen and Deying Kong, as well as my sister Willa Kong, I am incredibly grateful for your endless love and patience, continual encouragement, and unconditional trust. This journey would not have been possible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF SYMBOLS AND ABBREVIATIONS

NCO	nonconvex composite optimization
CNCO	constrained nonconvex composite optimization
MCO	min-max composite optimization
SNCO	spectral nonconvex composite optimization
$\mathcal{C}(Z)$	continuously differentiable functions on Z
$\mathcal{C}_L(Z)$	functions in $\mathcal{C}(Z)$ that are L-smooth
$\overline{\mathrm{Conv}}\,(Z)$	proper, closed, convex functions with domain Z
$\mathcal{F}_\mu(Z)$	functions in $\overline{\mathrm{Conv}}\,(Z)$ that are μ-strongly convex
$\mathcal{F}_{\mu,L}(Z)$	functions in $\mathcal{F}_\mu(Z)$ that are L-smooth
$\mathcal{C}_{m,M}(Z)$	functions in $\mathcal{C}(Z)$ that have curvature pair (m, M)
CG	composite gradient
ACG	accelerated composite gradient
GIPP	general inexact proximal point
GD	general descent
CR / SR / PR	composite / specialized / proximal refinement
AIPP	accelerated inexact proximal point
AIP.QP	accelerated inexact proximal quadratic penalty
AIP.AL	accelerated inexact proximal augmented Lagrangian
AICG	accelerated inexact composite gradient
D.AICG	doubly-accelerated composite gradient

SUMMARY

This thesis focuses on developing and analyzing accelerated and inexact first-order methods for solving or finding stationary points of various nonconvex composite optimization (NCO) problems. Our main tools mainly come from variational and convex analysis, and our key results are in the form of iteration complexity bounds and how these bounds compare to other ones in the literature.

Our first study problem is the classic unconstrained NCO problem studied by Mine and Fukushima (1981), and we develop an accelerated inexact proximal point method for finding approximate stationary points of it. By analyzing the method's variational properties, we establish an iteration complexity bound that is optimal in the number of first-order oracle evaluations. As an additional result, we show that our accelerated method and the classic composite/proximal gradient method are instances of a general inexact proximal point framework under different stepsizes and levels of inexactness.

Following our developments for the unconstrained setting, we move to study two instances of a function-constrained NCO problem. The first instance comprises a set of linear set constraints, and we develop a quadratic penalty method for finding approximate stationary points of it. We then establish an iteration complexity bound that is several orders of magnitude better than the previous state-of-the-art bound. As part of the analysis, we show that one can start the method from any point where the objective function is finite (and not necessarily from a near feasible point) and that no regularity conditions are needed to obtain convergence. The second instance consists of a set of nonlinear cone constraints, and we develop a proximal inexact augmented Lagrangian method for finding approximate stationary points of it. We then establish a competitive iteration complexity bound under an easily verifiable Slater-like condition. As part of the analysis, we show that the Lagrange multipliers generated by the method are bounded, without needing to dampen the (dual) multiplier update, and, like in the penalty method, the initial point can be any point where

the objective function is finite.

Before moving on to other problems, we discuss some efficient implementation strategies of the above methods. In particular, we present some efficient line search subroutines, an adaptive stepsize selection scheme, an efficient warm-start strategy, and a discussion about how to relax some algorithms' convexity assumptions. We also present a large number of real-world applications and numerical experiments that highlight our methods' performance against other modern solvers.

Our second-to-last study problem is a class of nonconvex-concave min-max NCO problems, and we develop an accelerated smoothing method for finding two kinds of approximate stationary points of it. Using prior results from our study of the unconstrained NCO problem, we establish iteration bounds that substantially improve on similar ones in the literature. Additionally, we give a brief discussion about how to generalize our smoothing method to solve linearly constrained min-max NCO problems. We then end with some numerical experiments in the unconstrained setting to validate the efficacy of our approach.

Our final study problems are a popular class of spectral NCO problems in which the inputs are general m-by-n real-valued matrices. As part of the study, we develop two inexact composite gradient methods — one based on the classic composite/proximal gradient method and another based on an accelerated variant of it — to find approximate stationary points. Extending some techniques for analyzing accelerated methods, we show that the accelerated variant obtains a competitive convergence rate in the nonconvex setting and an accelerated convergence rate in the convex setting. A vital conclusion of the study is that we show the methods perform nearly all of their iterations over the vector space $\mathbb{R}^{\min\{m,n\}}$ rather than the matrix space $\mathbb{R}^{m\times n}$. We then end with some numerical experiments to show the effectiveness of the previous conclusion.

CHAPTER 1
INTRODUCTION

If everything seems under control, you're just not going fast enough.

-Mario Andretti

Efficient optimization algorithms play a ubiquitous role in both the theory and application of machine learning and scientific computing. From web search engines to facial recognition software, their presence is found in many indispensable systems of modern society.

In this thesis, we contribute to a class of popular continuous optimization algorithms called *first-order methods*, consisting of iterative optimization algorithms that exploit information about the function value and subgradient(s) of the objective function. Since Cauchy's study on the gradient descent method [21] in 1847, these methods have found extensive use in smooth convex minimization (fast gradient methods [81, 83, 84]), nonsmooth convex minimization (subgradient descent [94, 102], mirror descent [8, 79, 83], and bundle methods [10, 44, 54]), and convex-concave saddle-point problems (smoothing methods [85] and mirror prox [80, 82, 83]). Recently, first-order methods have gained a renewed interest due to their ability to obtain cheap (nearly) dimension-free[1] guarantees for large-scale problems in a broad spectrum of disparate fields.

Our focus problems are variants of the following classic smooth nonconvex (additive) composite optimization (NCO) problem, first studied in [71] by Mine and Fukushima:

$$\min_{x\in\mathbb{R}^n} \{\phi(x) = f(x) + h(x)\}, \tag{$\mathcal{NCO}$}$$

where $h : \mathbb{R}^n \mapsto (-\infty, \infty]$ is a closed, proper, convex, but not necessarily differentiable, function and $f : \mathbb{R}^n \mapsto (-\infty, \infty]$ is a function that is continuously differentiable on an open set containing the domain of h, but not necessarily convex. Problems such as $\mathcal{NCO}$

[1] In contrast, interior point methods are known to grow nonlinearly with respect to the dimension, or equivalently, the number of decision variables.

frequently appear in areas such as recommender systems [28, 43], signal processing [17, 24], sparse regularization [34, 108, 112], and compressed sensing [4, 5].

In the forty years following Mine and Fukushima's work, there has been an immense amount of literature devoted to creating efficient methods for finding approximate stationary points[2] of $\mathcal{NCO}$ and its variants. Recent developments, in particular, have focused on generalizing Nesterov's seminal work on *accelerated* gradient methods for smooth convex optimization [84] to the nonconvex setting of $\mathcal{NCO}$ under a structural weak convexity assumption [18, 30, 31, 56, 91], i.e. where we assume that $f + m\|\cdot\|^2/2$ is convex for sufficiently large enough $m > 0$.

Our goal in this thesis is to continue this work and present several accelerated *nonconvex* first-order methods that explicitly take advantage of structural weak convexity in a meaningful way. The main theme that pervades most of our studies is that of variational inclusions, e.g. $0 \in \partial^* \phi(x) = \nabla f(x) + \partial h(x)$ where $\partial^* \phi$ (resp. ∂h) is the Clarke[3] (resp. regular[4]) subdifferential of ϕ (resp. h). By studying the inexact and exact variational properties of several accelerated methods in the convex setting, we construct accelerated methods with similar properties in the nonconvex setting. The efficacy of this approach is validated through competitive iteration complexity bounds, promising numerical experiments, and its utility in established optimization frameworks, e.g. penalty and augmented Lagrangian frameworks.

1.1 Contributions of the Thesis

This section carefully describes the organization and key contributions of this thesis. It it divided into three subsections. The first one is dedicated to optimization algorithms for smooth NCO problems, the second one to efficient implementation strategies, and the last one to optimization algorithms for NCO problems with additional structure.

Throughout this section, we let $\partial^* \phi(x)$ denote the Clarke subdifferential of ϕ at x and

[2] In general, finding even *approximate* minimizers of $\mathcal{NCO}$ is intractable [77, 83, 86].
[3] See Definition 2.1.38.
[4] See Definition 2.1.37.

$\operatorname{dist}(x, C)$ denote the distance between a point x and a set C.

1.1.1 Smooth NCO Problems

In the next three chapters of this thesis, we propose a substantial number of iterative first-order optimization methods for finding approximate stationary points of $\mathcal{NCO}$ in the unconstrained and function-constrained setting. Under the assumption that f is weakly convex and its gradient ∇f is Lipschitz continuous, each method comes with an iteration complexity bound and a comparison with similar methods in the literature. Below, we briefly summarize the contributions of these methods.

Complexity Optimal Proximal Point Method for Unconstrained $\mathcal{NCO}$ Problems. In Chapter 3, we develop a general inexact proximal point framework for finding approximate stationary points of $\mathcal{NCO}$. More specifically, this framework is designed to find a ρ-approximate stationary point $\bar{x} \in \mathbb{R}^n$ satisfying

$$\operatorname{dist}(0, \partial^* \phi(\bar{x})) \leq \rho. \tag{1.1}$$

Using a special inexactness criterion and several variational properties of an accelerated gradient method, we present a specific instance of the framework that is (iteration) complexity optimal in terms of the smoothness parameters of f and the tolerance ρ. It is worth mentioning that this instance does not require the domain of h to be bounded and only requires ϕ_* in $\mathcal{NCO}$ to be finite. Furthermore, the inexactness criterion does not depend on the tolerance ρ but rather on a special proximal residual.

Quadratic Penalty Method for Linearly-Constrained $\mathcal{NCO}$ Problems. In the first section of Chapter 4, we develop a quadratic penalty method for finding approximate stationary

points of linearly set-constrained[5] instances of $\mathcal{NCO}$. More specifically, this method is designed to find a (ρ, η)-approximate stationary point $(\bar{x}, \bar{p})$ satisfying

$$\operatorname{dist}(0, \partial^* \phi(\bar{x}) + A^* \bar{p}) \leq \rho, \quad \operatorname{dist}(A\bar{x}, S) \leq \eta. \tag{1.2}$$

Using our developments in Chapter 3 and some additional properties about penalty functions, we show that the method obtains an $\mathcal{O}(\rho^{-2}\eta^{-1})$ iteration complexity bound, which substantially improves upon the previously known bound of $\mathcal{O}(\rho^{-6})$ that was obtained by a multiblock ADMM-type method [42] for the case of $\rho = \eta$. The main novelty of the proposed method is that the initial starting point z_0 only needs to be in the domain of h, i.e. $h(z_0) < \infty$, and not necessarily feasible with respect to the linear set constraint, i.e. $Az_0 \in S$. It is also worth mentioning that the method does not require any regularity condition on its linear constraints and that the inexactness criterion does not depend on the tolerance pair (ρ, η).

Proximal Augmented Lagrangian Method for Nonlinearly-Constrained $\mathcal{NCO}$ Problems. In the second section of Chapter 4, we develop an inexact proximal augmented Lagrangian method for finding approximate stationary points of nonlinearly cone-constrained instances of $\mathcal{NCO}$ in which: (i) the function h is Lipschitz continuous and its domain is bounded; and (ii) the function g forming the cone constraint $g(x) \preceq_{\mathcal{K}} 0$ is $\mathcal{K}$-convex. More specifically, this method is designed to find a (ρ, η)-approximate stationary point $(\bar{x}, \bar{p})$ satisfying

$$\operatorname{dist}(0, \partial^* \phi(\bar{x}) + \nabla g(x)\bar{p}) \leq \rho, \quad \operatorname{dist}(g(\bar{x}), \mathcal{F}(\bar{p})) \leq \eta, \quad \bar{p} \succeq_{\mathcal{K}^+} 0,$$

where $\mathcal{K}^+$ is the dual cone of $\mathcal{K}$ and the set $\mathcal{F}(\bar{p})$ is given by

$$\mathcal{F}(\bar{p}) := \{g(x) : \langle g(x), \bar{p} \rangle \leq 0, g(x) \preceq_{\mathcal{K}} 0, h(x) < \infty\}.$$

[5] The constraint is of the form $Az \in S$ for some linear operator A and closed convex set S.

Using a special inexactness criterion and several recent developments from convex analysis, we show that the method obtains an $\mathcal{O}(\lceil \eta^{-1/2}\rho^{-2} + \rho^{-3} \rceil \log\lceil \rho^{-1} + \eta^{-1} \rceil))$ iteration complexity bound under a weak Slater-like condition. The contribution of the method is twofold. First, the method proposes a novel way of generating the penalty parameters c_k based on the change in the augmented Lagrangian between consecutive iterations rather than based on the feasibility of a particular iterate[6]. Second, it is shown that the multipliers $\{p_k\}_{k\geq 1}$ generated by the classic (dual) multiplier update are bounded without requiring any normalization[7].

1.1.2 Efficient Implementation Strategies

Following the above developments, we dedicate Chapter 5 to efficient implementation strategies. Additionally, we present iteration complexity bounds for variants of the methods in Chapter 3 and Section 4.1 that use some of these strategies and give several numerical experiments. Below, we highlight some of the most effective strategies.

Adaptive Stepsize Selection. We propose several different approaches of choosing several key "stepsize" parameters based on a finite set of key inequalities. These approaches are designed to adapt to the local geometry of the objective function and improve the convergence rate of the convex *and* nonconvex methods that use them.

Relaxation of Convexity. Several of the methods for the smooth NCO problems rely on the "stepsize" parameters to be within a particular range of values in order to ensure some, not necessarily verifiable, convexity conditions hold. We propose a way to relax some of these conditions to a verifiable set of finite inequalities to allow the "stepsize" parameters to be arbitrarily large or small.

[6] Other methods in the literature [12, 32, 110] usually consider increasing c_k whenever $\|\max\{0, g(x_k)\}\|$ has not sufficiently decreased between iterations

[7] Other methods in the literature [12, 110] usually add a step that projects the multipliers $\{p_k\}_{k\geq 1}$ into a bounded Euclidean box after the classic multiplier update is computed .

Warm-Start Strategy. For methods that operate by finding approximate stationary points of a sequence of optimization subproblems, we propose a warm-start strategy for initializing the starting point of each subproblem. More specifically, we propose a strategy where the current subproblem uses a point obtained from the last iterate of the previous subproblem. We then show that a (convexity) relaxed quadratic penalty method obtains an $\mathcal{O}(\eta^{-2})$ factor improvement in its iteration complexity bound (for finding (ρ, η)-stationary points as in (6.4.1)) when a warm-start strategy is used in place of a cold-start strategy.

1.1.3 NCO Problems with Additional Structure

Following the developments in prior chapters, the last two chapters of this thesis consider variants of $\mathcal{NCO}$ with additional structure and give several numerical experiments. Below, we summarize the contributions of these methods.

Smoothing Methods. In Chapter 6, we first develop a smoothing method for finding approximate stationary points of nonconvex-concave min-max instances of $\mathcal{NCO}$. More specifically, when f is a max function of the form $f(x) = \max_y \Phi(x, y)$, the method is designed to obtain stationary points of two kinds: (i) a δ-approximate directional stationary point x satisfying

$$\exists \bar{x} \text{ s.t. } \inf_{\|d\| \leq 1} \phi'(\bar{x}; d) \leq \delta, \quad \|x - \bar{x}\| \leq \delta,$$

where $\phi'(x; d)$ is the directional derivative of ϕ at x for the direction d, and (ii) a (ρ_x, ρ_y)-approximate primal-dual stationary point $(\bar{x}, \bar{y})$ satisfying

$$\operatorname{dist}(0, \partial^* \psi_{\bar{y}}(\bar{x})) \leq \rho_x, \quad \operatorname{dist}(0, \partial^* \psi_{\bar{x}}(\bar{y})) \leq \rho_y$$

where $\psi_{\bar{x}}(\cdot) := -\Phi(\bar{x}, \cdot)$ and $\psi_{\bar{y}}(\cdot) := \Phi(\cdot, \bar{y}) + h(\cdot)$. Using several results from convex analysis and the efficient method in Chapter 3, we show that the smoothing method obtains

$\mathcal{O}(\delta^{-3})$ and $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ iteration complexity bounds for obtaining δ-approximate directional stationary points and (ρ_x, ρ_y)-approximate primal-dual stationary points, respectively. Following these developments, we propose a quadratic penalty smoothing method for solving linearly-constrained instances of the min-max problem and establish an iteration complexity bound for finding an approximate primal-dual stationary point of the constrained problem. The main contributions are significantly improved complexity bounds (see Tables 6.1 and 6.2) and a new complexity bound for the constrained case. It is worth mentioning that the methods do not assume that the domain of h is bounded.

Spectral Optimization Methods. In Chapter 7, we develop two inexact spectral composite optimization methods, one accelerated and one unaccelerated, for finding ρ-approximate stationary points of $\mathcal{NCO}$ as in (1.1) in which ϕ admits an additional spectral decomposition. More specifically, for a given input point $X \in \mathbb{R}^{m\times n}$, we consider the instances where the composite term h is a function of the singular values of X and the smooth term f can be decomposed as $f = f_1 + f_2$ where f_2 is also a function of the singular values of X. Using a special inexactness criterion and several variational properties of an accelerated gradient method, we show that both methods obtain an $\mathcal{O}(\rho^{-2})$ iteration complexity bound and that the accelerated method obtains an $\mathcal{O}(\rho^{-2/3})$ complexity bound when ϕ is convex. A key contribution is that the methods mainly iterate over a space of singular values rather than the larger space of input matrices.

CHAPTER 2
BACKGROUND

This chapter presents the basic concepts, well-known results, and notational conventions that are used throughout the thesis. **Aside from the notation in Section 2.1.5**, the materials in this chapter are well-established, and hence, may be skipped upon first reading.

Organization

This chapter contains two sections. The first one presents theoretical background material while the second one presents algorithmic background material.

2.1 Theoretical Background

This section presents material that is relevant to the theoretical developments of the thesis.

2.1.1 Basics

This subsection states basic definitions, conventions, and notation.

Sets. We denote $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$, and $\mathbb{C}$ to be the set of real numbers, integers, natural numbers, and complex numbers, respectively. The sets $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the nonnegative a positive numbers, respectively. For sets A, B, we denote their Cartesian product as $A \times B = \{(a, b) : a \in A, b \in B\}$ and their Minkowski sum as $A + B = \{a + b : a \in A, b \in B\}$. For ease of notation, we denote $\{a\} + B \equiv a + B$ and $\lambda A = \{\lambda a : a \in A\}$ for any $a \in A$ and $\lambda \in \mathbb{C}$. For $n \in \mathbb{N}$, we define $A^n = \overbrace{A \times ... \times A}^{n \text{ times}}$. The empty set is denoted by $\varnothing$. For $a, b \in \mathbb{R}^n$ we denote the line interval between a and b as $[a, b] = \{ta + (1 - t)b : 0 \leq t \leq 1\}$. We also denote $[a, b) = [a, b]\backslash\{b\}$, $(a, b] = [a, b]\backslash\{a\}$, and $(a, b) = [a, b]\backslash\{a, b\}$. The set $\{x_i\}_{i=1}^k$ consists of the elements $x_1, ..., x_k$. The set $\{x_i\}_{i \geq 1}$ consists of the elements x_i for every $i \in \mathbb{N}$.

Functions. Let X, Y, and Z be arbitrary sets. We denote $f : X \mapsto Y$ and $F : X \rightrightarrows Y$

to be single-valued and set-valued functions from X to Y, respectively. For any set S, we denote $f(S) = \{f(s) : s \in S\}$. For functions $f : X \mapsto Y$ and $g : Y \mapsto Z$, we denote $g \circ f(x) = g(f(x))$ for every $x \in X$.

Basic Operators. Let $x \in \mathbb{R}$, $f : X \mapsto \mathbb{R}$ be an arbitrary function, and S be an arbitrary set. We denote $\lceil x \rceil$ (resp. $\lfloor x \rfloor$) to be the smallest (resp. largest) element in $\mathbb{Z}$ that is greater (resp. less) than or equal to x. We denote $\sup_{x \in S} f(x)$ (resp. $\inf_{x \in S} f(x)$) as the smallest (resp. largest) element B in $\mathbb{R}$ that satisfies $f(s) \leq B$ (resp. $f(s) \geq B$) for every $s \in S$. The function $\operatorname{sgn}(x)$ takes value +1 if $x \geq 0$ and -1 otherwise. As a convention, we take $a/0 = +\infty$ and $-a/0 = -\infty$ for every $a > 0$.

Computational Complexity. For functions $f, g : \mathbb{R}_{++} \mapsto \mathbb{N}$, we use the following asymptotic notation:

- $f(x) = \mathcal{O}(g(x))$ if there exists $(C, \underline{x}) \in \mathbb{R}^2_{++}$ such that for every $x \geq \underline{x}$ it holds that $f(x) \leq Cg(x)$.
- $f(x) = \Omega(g(x))$ if there exists $(C, \underline{x}) \in \mathbb{R}^2_{++}$ such that for every $x \geq \underline{x}$ it holds that $f(x) \geq Cg(x)$.
- $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $f(x) = \Omega(g(x))$.
- $f(x) = o(g(x))$ if for every $C > 0$ there exists $\underline{x} > 0$ such for every $x \geq \underline{x}$ it holds that $f(x) \leq Cg(x)$.

2.1.2 Analysis

This subsection reviews relevant materials from analysis.

We first start with some basic definitions and notation.

Definition 2.1.1. For a vector space $\mathcal{X}$, an **inner product** $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a mapping that satisfies, for every $x, y, z \in \mathcal{X}$ and $\alpha, \beta \in \mathbb{R}$, the relations:

(i) $\langle x, y\rangle = \langle y, x\rangle$ (symmetry);

(ii) $\langle \alpha x + \beta y, z\rangle = \alpha \langle x, z\rangle + \beta \langle y, z\rangle$ (linearity);

(iii) $\langle x, x\rangle > 0$ if $x \neq 0$ (non-degeneracy).

A vector space equipped with an inner product is said to be a **inner product space**.

Definition 2.1.2. The **induced norm** of an inner product space $\mathcal{X}$, denoted by $\|\cdot\|$, is given by $\|x\| = \langle x, x\rangle$ for every $x \in \mathcal{X}$. It is well-known that every inner product satisfies the **Cauchy-Schwarz** inequality $\langle x, y\rangle \leq \|x\| \cdot \|y\|$ and the triangle inequality $\|x+y\| \leq \|x\| + \|y\|$ for every $x, y \in \mathcal{X}$.

For the rest of this subsection, we let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be inner product spaces with a common inner product $\langle \cdot, \cdot \rangle$. Moreover, we denote $\|\cdot\|$ to be their induced norm.

Definition 2.1.3. For a point $z \in \mathcal{Z}$ and parameter $r > 0$, the **open ball** $\mathcal{B}_r(z)$ and **closed ball** $\overline{\mathcal{B}}_r(z)$ of radius r at z is defined by

$$\mathcal{B}_r(z) := \{z' \in \mathcal{Z} : \|z' - z\| < r\},$$
$$\overline{\mathcal{B}}_r(z) := \{z' \in \mathcal{Z} : \|z' - z\| \leq r\}.$$

A set $Z \subseteq \mathcal{Z}$ is said to be **open** if for every $z \in Z$ there exists $\varepsilon > 0$ such that $\mathcal{B}_\varepsilon(z) \subseteq Z$. A set $\tilde{Z} \subseteq \mathcal{Z}$ is said to be **closed** if the set $\mathcal{Z} \backslash \tilde{Z}$ is open. Finally, a set $Z \subseteq \mathcal{Z}$ is said to be bounded if there exists $r \in \mathbb{R}_{++}$ such that $Z \subseteq \mathcal{B}_r(0)$.

Definition 2.1.4. A set $C \subseteq \mathcal{Z}$ is said to be **compact** if for any collection of open sets $\mathcal{D} = \{D_i\}_{i \in \mathcal{I}}$, for some index set $\mathcal{I}$, satisfying $C \subseteq \bigcup_{i \in \mathcal{I}} D_i$ there exists a finite subcollection $\tilde{\mathcal{D}} = \{\tilde{D}_i\}_{i=1}^k \subseteq \mathcal{D}$ such that $C \subseteq \bigcup_{i=1}^k D_i$. If $\mathcal{Z} = \mathbb{R}^n$, it is well-known that a set $C \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.

Definition 2.1.5. For a sequence $\{z_n\}_{n\geq 1} \subseteq \mathcal{Z}$, we say that z_n converges to z, or equivalently $\lim_{i\to\infty} z_n = z \in \mathcal{Z}$, if for every $\varepsilon > 0$, there exists $\underline{n} \in \mathbb{N}$ such that for every $k \geq \underline{n}$ we have $\|z - z_k\| \leq \varepsilon$.

The next result is a well-known result about bounded sequences.

Theorem 2.1.6. *(Bolzano-Weierstrass) Every bounded sequence in a finite dimensional inner product space has a convergent subsequence.*

We now present definitions and results about some special classes functions.

Definition 2.1.7. A function $\phi : \mathcal{X} \mapsto \mathcal{Y}$ is said to be **continuous** on a set $X \subseteq \mathcal{X}$ if for every $x \in X$ and $\varepsilon > 0$ there exists $\delta > 0$ such that for every $x' \in X$ satisfying $\|x - x'\| \leq \delta$ we have that $\|\phi(x) - \phi(x')\| \leq \varepsilon$. It is well-known that if $\{x_i\}_{i\geq 1} \subseteq X$ is such that $\lim_{i\to\infty} x_i = x \in X$ and ϕ is continuous on X, then $\lim_{i\to\infty} \phi(x_i) = \phi(\lim_{i\to\infty} x_i) = \phi(x)$.

Definition 2.1.8. A function $\phi : \mathcal{X} \mapsto \mathcal{Y}$ is said to be L-**Lipschitz continuous** on a set $X \subseteq \mathcal{X}$ if

$$\|\phi(x) - \phi(x')\| \leq L\|x - x'\| \quad \forall x, x' \in X.$$

Definition 2.1.9. For a closed convex set $Z \subseteq \mathcal{Z}$, the (single-valued) **projection mapping** Π_Z at a point z is defined by

$$\Pi_Z(z) = \operatorname*{argmin}_{u\in Z} \frac{1}{2}\|u - z\|^2.$$

The distance function $\operatorname{dist}(\cdot, Z)$ at a point z is defined by

$$\operatorname{dist}(z, Z) = \|z - \Pi_Z(z)\|.$$

Definition 2.1.10. Let $f : \mathcal{X} \mapsto \mathcal{Y}$ be a function that is well-defined in an open ball around a point $x \in \mathcal{X}$. The function f is said to be **(Fréchet) differentiable** at x if there exists a linear function $Df_x : \mathcal{X} \mapsto \mathcal{Y}$, called the **derivative** of f at x, that approximates the change

$f(x + \Delta x) - f(x)$ up to a residual, called the **first-order Taylor residual**, that is $o(\Delta x)$. More specifically, the function f is differentiable at x if and only if

$$\|f(x + \Delta x) - f(x) - Df_x(\Delta x)\| = o(\Delta x)$$

for every Δx such that $f(x + \Delta x)$ is well-defined.

Definition 2.1.11. A differentiable function $f : \mathcal{X} \mapsto \mathcal{Y}$ is said to be **continuously differentiable** at x if the function $x \mapsto Df_x(\Delta x)$ is continuous for every $\Delta x \in \mathcal{X}$.

Definition 2.1.12. Let $f : \mathcal{X} \mapsto \mathcal{Y}$ be differentiable at a point $x \in \mathcal{Z}$. The **gradient** of f at x is the unique matrix $\nabla f(x)$ that satisfies

$$\nabla f(x)^T u = Df_x(u)$$

for every $u \in \mathcal{X}$ in a neighborhood of x. The **derivative matrix** of f at x is the transpose of $\nabla f(x)$ and is denoted by $f'(x) = \nabla f(x)^T$.

Definition 2.1.13. The **linear approximation** of a differentiable function $f : \mathcal{X} \mapsto \mathcal{Y}$ at a point $x_0 \in \mathcal{X}$ is defined as

$$\ell_f(x; x_0) := f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \quad \forall x \in \mathcal{X}.$$

The next three results present some fundamental properties involving derivatives and gradients and can be found, for example, in [23, 109].

Theorem 2.1.14. *(Chain rule) Let $f : \mathcal{X} \mapsto \mathcal{Y}$ be differentiable at $x \in \mathcal{X}$ and let $g : \mathcal{Y} \mapsto \mathcal{Z}$ be differentiable at $y = f(x) \in \mathcal{Y}$. Then, $g \circ f$ is differentiable at x and*

$$D(g \circ f)_x = Dg_y \circ Df_x.$$

Theorem 2.1.15. *(Mean Value Theorem) For any differentiable function $f : \mathcal{X} \mapsto \mathcal{Y}$ and $x_0, x_1 \in \mathcal{X}$, there exists $t \in [0,1]$ such that*

$$f(x_1) = f(x_0) + \nabla f(x_t)^T (x_1 - x_0),$$

where $x_t = tx_0 + (1-t)x_1$.

Theorem 2.1.16. *(Gradient Theorem) Let $x_0, x_1 \in \mathcal{X}$ and $r : [0,1] \mapsto \mathcal{X}$ be such that $r(0) = x_0$ and $r(1) = x_1$. For any continuously differentiable function $\phi : \mathcal{X} \mapsto \mathbb{R}$, we have*

$$\phi(x_1) - \phi(x_0) = \int_0^1 \nabla \phi(r(t)) \cdot r'(t)\, dt.$$

The below material deals with the convolution of two functions.

Definition 2.1.17. The **convolution** of functions $f, g : \mathcal{X} \mapsto \mathbb{R}$ is

$$(f * g)(x) := \int_{-\infty}^{\infty} f(u) g(x-u)\, du \quad \forall x \in \mathcal{X}$$

The following result can be found, for example, in [14, Chapter 6].

Proposition 2.1.18. *Let $f, g : \mathcal{X} \mapsto \mathcal{Y}$ be continuously differentiable functions. Then, it holds that*

$$D(f * g)_x = Df_x * g = f * Dg_x \quad \forall x \in \mathcal{X}.$$

2.1.3 Linear Algebra

This subsection reviews notation and relevant materials from linear algebra.

We first start with some basic notation and definitions.

For every $(n, m) \in \mathbb{N}^2$, we denote $\mathbb{F}^{n \times m}$ to be the set of matrices with n rows and m columns with entries from $\mathbb{F} \in \{\mathbb{R}, \mathbb{Z}, \mathbb{N}, \mathbb{C}\}$. The entry in the i^{th} row and j^{th} column of A is denoted by $[A]_{ij}$ or A_{ij}.

Definition 2.1.19. For matrices $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{p\times m}$, the **matrix product** $AB \in \mathbb{R}^{n\times m}$ is given by the relation $[AB]_{ij} = \sum_{k=1}^{p} [A]_{ik}[B]_{kj}$.

Definition 2.1.20. The **conjugate transpose (or adjoint)** of a matrix $A \in \mathbb{C}^{m\times n}$, denoted by A^*, is given by the relation $A^*_{ij} = \overline{A_{ij}}$. The **transpose** of a matrix, denoted by A^T, is given by the relation $A^T_{ij} = A_{ij}$. It is well-known that

$$\langle Ax, y\rangle = \langle x, A^*y\rangle \quad \forall (x, y) \in \mathbb{C}^m \times \mathbb{C}^n.$$

If $a_i \in \mathbb{R}^n$ for $i \in \{1, ..., k\}$, then we denote $(a_1, ..., a_k)$ to be the matrix whose i^{th} column is a_i. If A is a linear operator, then we denote $Az \equiv A(z)$.

Definition 2.1.21. A matrix $A \in \mathbb{R}^{m\times n}$ is **symmetric** if $A^* = A$.

Definition 2.1.22. A matrix $A \in \mathbb{R}^{n\times n}$ is **positive (semi-)definite**, or equivalently $A > (\geq)0$, if A is symmetric and for every $x \in \mathbb{R}^n\backslash\{0\}$ we have $x^TAx > (\geq)0$. The **set of positive (semi-)definite matrices** in $\mathbb{R}^{n\times n}$ is denoted by $\mathbb{S}^n_{++}$ ($\mathbb{S}^n_+$).

Definition 2.1.23. The **trace** of a matrix $A \in \mathbb{R}^{n\times n}$ is given by $\mathrm{tr}(A) = \sum_{i=1}^{n} A_{ij}$. It is well-known that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for any matrices $A, B \in \mathbb{R}^{n\times n}$.

Definition 2.1.24. The **identity matrix** of size n, denoted by I_n, is given by $(I_n)_{ij} = 1$ if $i = j$ and 0 if $i \neq j$.

Definition 2.1.25. A matrix $A \in \mathbb{R}^{n\times n}$ is said to be **invertible** (or **non-singular**) if there exists a matrix A^{-1}, called the **inverse** of A, that satisfies $A^{-1}A = AA^{-1} = I_n$.

Definition 2.1.26. A matrix $A \in \mathbb{R}^{n\times n}$ is said to be **orthogonal** if $A^T = A^{-1}$.

Definition 2.1.27. The **determinant** of a matrix $A \in \mathbb{R}^{n\times n}$, denoted by $\det(A)$, is $[A]_{11}$ if $n = 1$ and is given recursively by

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} \det(M_{ij}) = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} \det(M_{ij})$$

for $n \geq 2$, where $M_{ij} \in \mathbb{R}^{(n-1)\times(n-1)}$ is the **minor** that results from removing the i^{th} row and j^{th} column from A. It is well-known that $\det(AB) = \det(A)\det(B)$ and $\det(A) = \det(A^T)$ for any matrices $A, B \in \mathbb{R}^{n\times n}$.

Definition 2.1.28. The **eigenvalues** of a matrix $A \in \mathbb{R}^{n\times n}$ are the roots of the characteristic polynomial $\det(A - \lambda I_n)$ as a univariate function in λ. An **eigenvector** $v \in \mathbb{R}^{n\times n}$ corresponding to some eigenvalue λ is any vector satisfying $Av = \lambda v$. We denote $\lambda_k(A)$ to be the k^{th} **largest eigenvalue** of $A \in \mathbb{R}^{n\times n}$. Moreover, we use the shorthand $\lambda_{\min}(A) = \lambda_n(A)$ and $\lambda_{\max}(A) = \lambda_1(A)$.

Definition 2.1.29. The **singular value decomposition** (SVD) of a matrix $A \in \mathbb{R}^{m\times n}$ is a factorization of the form $A = P\Sigma Q^*$ where $P \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is a rectangular diagonal matrix with nonnegative entries on the diagonal. The diagonal entries $\{\Sigma_{ii}\}_{i\geq 1}$ are known as the **singular values** of A.

The following is a well-known (see, for example, [41, Corollary 4.3.15]) result about eigenvalues of matrix sums.

Theorem 2.1.30. *(Weyl's Inequality) Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices and let $\lambda_k(M)$ denote the k^{th} largest eigenvalue of a matrix M. Then, it holds that*

$$\lambda_k(A) + \lambda_n(B) \leq \lambda_k(A + B) \leq \lambda_k(A) + \lambda_1(B)$$

for every $1 \leq i \leq n$.

2.1.4 Convex and Variational Analysis

This subsection presents relevant material from convex and variational analysis.

We first state some key definitions.

Definition 2.1.31. The interior of a set $Z \subseteq \mathcal{Z}$ is defined as

$$\operatorname{int} Z := \{z \in \mathcal{Z} : \exists \delta > 0 \text{ such that } \mathcal{B}_\delta(z) \subseteq Z\}.$$

Definition 2.1.32. For a convex set $Z \subseteq \mathcal{Z}$, the **affine hull** $\operatorname{aff} Z$ and **relative interior** $\operatorname{ri} Z$ of Z are defined by

$$\operatorname{aff} Z := \left\{ \gamma \in \mathcal{Z} : \gamma = \sum_{i=1}^{k} \alpha_i z_i,\ z_i \in Z,\ \sum_{i=1}^{k} \alpha_i = 1 \text{ for } i \leq k,\ k = 1, 2, \ldots \right\},$$
$$\operatorname{ri} Z := \{ \gamma \in \operatorname{aff} Z : \exists \delta > 0 \text{ such that } \operatorname{aff} Z \cap \mathcal{B}_\delta(\gamma) \subseteq Z \}.$$

Another interpretation of $\operatorname{aff} Z$ is that it is the smallest affine manifold containing Z. Under this interpretation, a point z is in $\operatorname{ri} Z$ if it is in the interior of Z relative to the topology given by $\operatorname{aff} Z$.

Definition 2.1.33. The (effective) **domain** of a function $f : \mathcal{Z} \mapsto (-\infty, \infty]$ is the set

$$\operatorname{dom} f := \{ z \in \mathcal{Z} : f(z) \in \mathbb{R} \}$$

and f is said to be **proper** if $\operatorname{dom} f \neq \emptyset$.

Definition 2.1.34. A proper function $f : \mathcal{Z} \mapsto (-\infty, \infty]$ is said to be **convex** if

$$f(\alpha z + [1-\alpha] z') \leq \alpha f(z) + (1-\alpha) f(z') \quad \forall z, z' \in \mathcal{Z}, \quad \forall \alpha \in (0,1).$$

It is well-known that if f is convex and differentiable, then $f(\cdot) - \ell_f(\cdot; z_0) \geq 0$ for any $z_0 \in \operatorname{dom} f$.

Definition 2.1.35. A proper function $f : \mathcal{Z} \mapsto (-\infty, \infty]$ is said to be μ-**strongly convex** if the function $f - \mu \|\cdot\|^2$ is convex and m-**weakly convex** if the function $f + m\|\cdot\|^2$ is convex.

It is well-known that if f is μ-strongly convex and differentiable, then $f(\cdot) - \ell_f(\cdot; z_0) \geq \mu \|\cdot - z_0\|^2/2$ for every $z_0 \in \operatorname{dom} f$. It is also well-known that if f is m-weakly convex and differentiable, then $f(\cdot) - \ell_f(\cdot; z_0) \geq -m\|\cdot - z_0\|^2/2$ for every $z_0 \in \operatorname{dom} f$.

Definition 2.1.36. A proper convex function $f : \mathcal{Z} \mapsto [-\infty, \infty)$ is said to be **closed** or **lower**

semicontinuous if

$$\liminf_{z \to z_0} f(z) \geq f(z_0) \quad \forall z_0 \in \mathcal{Z}.$$

Definition 2.1.37. For a proper convex function $f : \mathcal{Z} \mapsto [-\infty, \infty)$ and a point $z \in \operatorname{dom} f$, the ε-**subdifferential** of f at z is defined by

$$\partial_\varepsilon f(z) = \left\{ v \in \mathcal{Z} : f(z') \geq f(z) + \langle v, z' - z \rangle \ \ \forall z' \in \mathcal{Z} \right\},$$

and the (regular) **subdifferential** of f at z is $\partial_0 f(z)$ and is commonly denoted by $\partial f(z)$. It is well-known that $z_* \in \operatorname{argmin}_{z' \in \mathcal{Z}} f(z')$ if and only if $0 \in \partial f(z_*)$.

Definition 2.1.38. For a proper function $f : \mathcal{Z} \mapsto [-\infty, \infty)$, the Clarke subdifferential of f at a point $z \in \operatorname{dom} f$ is the set

$$\partial^* \phi(x) := \{ v : \langle v, \cdot \rangle \leq d\phi(x; \cdot) \}$$

where $d\phi(x; u) := \limsup_{t \downarrow 0, y \to x} [\phi(y + tu) - \phi(y)]/t$.

Definition 2.1.39. For a closed convex set $Z \subseteq \mathcal{Z}$ and a point $z \in \mathcal{Z}$, the **indicator function** δ_Z and the normal cone N_Z at a point $z \in \mathcal{Z}$ are given by

$$\delta_Z(z) := \begin{cases} 0, & z \in Z, \\ \infty, & \text{otherwise}, \end{cases}$$

$$N_Z(z) := \left\{ v \in \mathcal{Z} : \langle v, z' - z \rangle \leq 0 \ \forall z' \in Z \right\}.$$

Definition 2.1.40. For a proper, lower semicontinuous function $f : \mathcal{Z} \mapsto [-\infty, \infty)$, a parameter $\lambda > 0$, and a point $z \in \mathcal{Z}$, the **Moreau envelope** $e_\lambda f$ and the **proximal mapping**

$\operatorname{prox}_\lambda f$ of f at z are defined by

$$e_\lambda f(z) := \inf_{z' \in \mathcal{Z}} \left\{ f(z) + \frac{1}{2\lambda}\|z' - z\|^2 \right\} \leq f(z)$$
$$\operatorname{prox}_\lambda f(z) := \operatorname*{Argmin}_{z' \in \mathcal{Z}} \left\{ f(z) + \frac{1}{2\lambda}\|z' - z\|^2 \right\}.$$

The function f is said to be **prox-bounded** if there exists a threshold $\lambda > 0$ such that $e_\lambda f(z_0) > -\infty$ for some $z_0 \in \mathcal{Z}$.

Definition 2.1.41. For an extended real-valued function $f : \mathcal{Z} \mapsto [-\infty, \infty]$, the function $f^* : \mathcal{Z}^* \mapsto [-\infty, \infty]$ given by

$$f^*(u) := \max_{z \in \mathcal{Z}} \{\langle u, z\rangle - f(z)\} \quad \forall u \in \mathcal{Z}^*$$

is called the **conjugate function** of f.

Definition 2.1.42. For $K \subseteq \mathcal{Z}$, the **dual cone** K^+ and **polar cone** K^- are given by

$$K^+ := \{z \in \mathcal{Z} : \langle z, z'\rangle \geq 0 \ \forall z' \in K\},$$
$$K^- := \{z \in \mathcal{Z} : \langle z, z'\rangle \leq 0 \ \forall z' \in K\} = -K^+.$$

We now state some basic properties about the above objects.

The first result, whose proof can be found in [99, Theorem 2.26], describes the continuity of the prox related objects.

Proposition 2.1.43. *For a proper, lower semicontinuous, convex function $f : \mathcal{Z} \mapsto [-\infty, \infty)$ and parameter $\lambda > 0$, the following properties hold:*

(a) the proximal mapping $\operatorname{prox}_\lambda f$ is single-valued and continuous;

(b) the (λ-Moreau) envelope function $e_\lambda f$ is convex, continuously differentiable, and its gra-

dient is given by

$$\nabla e_\lambda f(z) = \frac{1}{\lambda}\left[z - \operatorname{prox}_{\lambda f}(z)\right] \quad \forall z \in \mathcal{Z}.$$

The following proposition, whose proof can be found in [7, Example 3.5] and [7, Theorem 6.24], presents some properties about indicator functions.

Proposition 2.1.44. *For any closed convex set $Z \subseteq \mathcal{Z}$ and point $z \in \mathcal{Z}$, the following properties hold:*

(a) $\partial \delta_Z(z) = N_Z(z)$;

(b) for any $\lambda > 0$, we have $\operatorname{prox}_\lambda \delta_Z(z) = \Pi_Z(z)$.

The next result, whose proof can be found in [39, Proposition XI.1.3.1], presents some basic calculus rules for the approximate subdifferential.

Proposition 2.1.45. *For a proper convex function $f : \mathcal{Z} \mapsto [-\infty, \infty)$, $\varepsilon > 0$, and point $z \in \mathcal{Z}$, the following properties hold:*

(a) for any $\alpha > 0$ and $r \in \mathcal{Z}$, we have $\partial_\varepsilon(\alpha f + r)(z) = \alpha \partial_{\varepsilon/\alpha} f(z)$;

(b) for any $\alpha \neq 0$, we have $\partial_\varepsilon f(\alpha z) = \alpha \partial_\varepsilon f(z)$;

(c) for any $s \in \mathcal{Z}$, we have $\partial_\varepsilon(f + \langle s, \cdot \rangle)(z) = \partial_\varepsilon f(z) + \{s\}$.

The below result, whose proof can be found in [39, Theorem XI.3.1.1], presents a characterization of the approximate subdifferential on sums of functions.

Proposition 2.1.46. *For proper convex functions $f_1, f_2 : \mathcal{Z} \mapsto (-\infty, \infty]$, parameter $\varepsilon > 0$, and $z \in \mathcal{Z}$, it holds that*

$$\partial_\varepsilon(f_1 + f_2)(z) \supseteq \bigcup_{\substack{\varepsilon_1 + \varepsilon_2 \leq \varepsilon, \\ \varepsilon_1, \varepsilon_2 \geq 0}} \left\{\partial_{\varepsilon_1} f_1(z) + \partial_{\varepsilon_2} f_2(z)\right\}.$$

Moreover, if $\operatorname{ri}\operatorname{dom} f_1 \cap \operatorname{ri}\operatorname{dom} f_2 \neq \varnothing$, then the above relation holds at equality.

The following transportation formula can be found in [39, Proposition XI.4.2.2].

Proposition 2.1.47. *(Transportation Formula) For a function $\psi \in \overline{\mathrm{Conv}}\,(\mathcal{Z})$, points $z, \bar{z} \in \mathrm{dom}\,\psi$, and subgradient $s \in \partial\psi(z)$, it holds that $s \in \partial_\varepsilon\psi(\bar{z})$ where $\varepsilon = f(\bar{z}) - f(z) - \langle s, \bar{z} - z\rangle \geq 0$.*

The next result, whose proof can be found in [7, Theorem 6.45], presents a well-known decomposition .

Proposition 2.1.48. *(Extended Moreau Decomposition) Let $f : \mathcal{Z} \mapsto (-\infty, \infty]$ be proper, closed, and convex. Then, for any $z \in \mathcal{Z}$ and $\lambda > 0$, it holds that*

$$\mathrm{prox}_\lambda f(z) + \lambda\,\mathrm{prox}_{\lambda^{-1}} f^*(z/\lambda) = z.$$

2.1.5 Function Classes

This sub-subsection defines some important function classes and their properties.

We first define the key function classes considered in this thesis.

Definition 2.1.49. Let $\mathcal{C}(Z)$ denote the set of continuously differentiable functions from $Z \subseteq \mathcal{Z}$ to $\mathbb{R}$.

Important Note: To be concise, we adopt the convention that if Z is a closed set and $f \in \mathcal{C}(Z)$, then it is implicitly assumed that f is finite on some open set Ω containing Z.

Definition 2.1.50. Let $\mathcal{C}_L(Z)$ denote the set of functions in $\mathcal{C}(Z)$ whose gradient is L-Lipschitz continuous on Z. Such functions are typically called L-**smooth**.

Definition 2.1.51. Let $\mathcal{C}_{m,M}(Z)$ denote the set of functions in $\mathcal{C}(Z)$ that satisfy

$$-\frac{m}{2}\|z - z'\|^2 \leq f(z) - \ell_f(z; z') \leq \frac{M}{2}\|z - z'\|^2 \quad \forall z, z' \in Z. \tag{2.1}$$

A function $f \in \mathcal{C}(Z)$ is said to have a **curvature pair** (m, M) if it is in $\mathcal{C}_{m,M}(Z)$.

Definition 2.1.52. Let $\overline{\mathrm{Conv}}\,(\mathcal{Z})$ be the set of proper, lower semicontinuous, convex functions from $\mathcal{Z}$ to $(-\infty, \infty]$. For a convex set $Z \subseteq \mathcal{Z}$, let $\overline{\mathrm{Conv}}\,(Z)$ be the set of functions in that $\overline{\mathrm{Conv}}\,(\mathcal{Z})$ are real-valued on Z and take value $+\infty$ outside of Z.

Definition 2.1.53. Let $\mathcal{F}_\mu(Z)$ denote the set of functions in $\overline{\mathrm{Conv}}\,(Z)$ that are μ-strongly convex. Let $\mathcal{F}_{\mu,L}(Z)$ denote the set of functions in $\mathcal{F}_\mu(Z)$ that are also L-smooth.

The next set of results present different characterizations of the above classes. The first results is a straightforward consequence of [9, Proposition 6.1.3].

Proposition 2.1.54. *If $f : Z \mapsto \mathbb{R}$ is twice differentiable with $\lambda_{\min}(\nabla^2 f(z)) = -m$ and $\lambda_{\max}(\nabla^2 f(z)) = M$ for every $z \in Z$, then $f \in \mathcal{C}_{m,M}(Z)$.*

The below result[1] relates $\mathcal{C}_{m,M}(Z)$ with $\mathcal{C}_L(Z)$.

Proposition 2.1.55. *Let $f : Z \mapsto \mathbb{R}$ be a continuously differentiable function for some $Z \subseteq \mathcal{Z}$. Then $f \in \mathcal{C}_L(Z)$ if and only if $f \in \mathcal{C}_{L,L}(Z)$.*

Proof. Let $x, y \in Z$ be arbitrary. Suppose $f \in \mathcal{C}_L(Z)$ and define $\boldsymbol{r}(t) = x + t(y - x)$ for every $t \in [0, 1]$. Using the Gradient Theorem, it holds that

$$\begin{aligned} f(y) - f(x) &= \int_{t=0}^{t=1} \nabla f(\boldsymbol{r}(t)) \cdot d\boldsymbol{r}(t) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle \, dt \\ &= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle \, dt. \end{aligned}$$

Using the Cauchy-Schwarz inequality, the above relation, and Lipschitz continuity of ∇f,

[1] Special thanks to Arkadi Nemirovski for helping with this proof.

we now conclude that

$$\begin{aligned}|f(y)-\ell_f(y;x)| &\le \int_0^1 |\langle \nabla f(x+t(y-x))-\nabla f(x), y-x\rangle|\,dt \\ &\le \int_0^1 \|\nabla f(x+t(y-x))-\nabla f(x)\|\cdot\|y-x\|dt \\ &\le \int_0^1 tL\|y-x\|^2 dt = \frac{L}{2}\|y-x\|^2\end{aligned}$$

and hence $f \in \mathcal{C}_{L,L}(Z)$.

Conversely, suppose $f \in \mathcal{C}_{L,L}(Z)$ and let $\{\delta_n\}_{n\ge 1}$ be a sequence of smooth, real-valued, (mollifier) functions over $\mathcal{Z}$ where, for every $n \ge 1$, we have: (i) $\delta_n \ge 0$; (ii) $\int_Z \delta_n(t)\,dt = 1$; and (iii) $\delta_n(t) = 0$ for t satisfying $\|t\| \ge 1/n$. Moreover, for every $n \ge 1$, define $g_n = \delta_n * f$ and denote $d = x - y$. It now follows that

$$\begin{aligned}|g_n(y)-\ell_{g_n}(y;x)| &= |\delta_n * [f(y)-f(x)] + \langle \delta_n * \nabla f(x), d\rangle| \\ &= \left|\int_{\mathcal{Z}} \delta_n(\tau)\left[f(y-\tau)-f(x-\tau)\right]\,d\tau + \left\langle \int_{\mathcal{Z}} \delta_n(\tau)\nabla f(x-\tau)\,d\tau, d\right\rangle\right| \\ &= \left|\int_{\mathcal{Z}} \delta_n(\tau)\left[f(y-\tau)-f(x-\tau)+\langle \nabla f(x-\tau), d\rangle\right]\,d\tau\right| \\ &\le \int_{\mathcal{Z}} \delta_n(\tau)\,|f(y-\tau)-f(x-\tau)+\langle \nabla f(x-\tau), d\rangle|\,d\tau \\ &\le \frac{L}{2}\int_{\mathcal{Z}} \delta_n(\tau)\|d\|^2\,d\tau = \frac{L}{2}\|d\|^2,\end{aligned}$$

and hence that $g_n \in \mathcal{C}_{L,L}(Z)$ as well. Using the smoothness of δ_n (and hence g_n), Taylor's Theorem, and the previous result, it holds that there exists $\xi \in [x, y]$ such that

$$\frac{L}{2} \ge \left|\frac{g_n(y)-\ell_{g_n}(y;x)}{\|d\|^2}\right| = \left|\frac{\langle d, \nabla^2 g_n(\xi) d\rangle}{2\|d\|^2} + \frac{o(\|d\|^2)}{\|d\|^2}\right|.$$

Taking $y \to x$ in the above inequality, we thus conclude that $\|\nabla^2 g_n(z)\| \le L$ for every $z \in Z$, and hence, it holds that $g_n \in \mathcal{C}_L(Z)$. Since $\nabla g_n \to \nabla f$ uniformly, it follows that $f \in \mathcal{C}_L(Z)$ as well. □

2.2 Algorithmic Background

This section presents some fundamental algorithms that will be relevant in the algorithmic developments of the thesis.

Throughout this section, we let $Z \subseteq \mathcal{Z}$ be a nonempty convex set. Moreover, for all the algorithms in the thesis, we use the notation "$\leftarrow$" for scalar or vector variable assignment and "$\Leftarrow$" for function assignment.

2.2.1 Composite Gradient (CG) Method

The composite gradient (CG) method (also known as the proximal gradient method) is a popular optimization algorithm [92] for solving and/or finding stationary points of the problem

$$\min_{z \in \mathcal{Z}} \{\psi(z) := \psi_s(z) + \psi_n(z)\} \tag{$\mathcal{CO}$}$$

where $\psi_n \in \overline{\mathrm{Conv}}\,(Z)$ and $\psi_s \in \mathcal{C}(Z)$. More specifically, it is an iterative method that, at its k^{th} iteration, performs the following update: given $z_{k-1} \in Z$ and $\lambda_k > 0$, compute

$$z_k = \mathrm{prox}_{\lambda_k \psi_n}(z_{k-1} - \lambda_k \nabla \psi_s(z_{k-1})).$$

When $\psi_n = \delta_C$ for some closed convex set C, it is straightforward to see that the CG method (CGM) reduces to the classical projected gradient method for the problem $\min_{z \in C} \psi_s(z)$. For ease of future reference and discussion, we give a description in Algorithm 2.2.1 which includes an important set of auxiliary iterates $\{v_k\}_{k \geq 1}$.

Algorithm 2.2.1: CG Method

Require: $\psi_n \in \overline{\mathrm{Conv}}\,(Z)$, $\psi_s \in \mathcal{C}(Z)$, $z_0 \in Z$, $\{\lambda_k\}_{k \geq 1} \subseteq \mathbb{R}_{++}$;

1: **procedure** CG($\psi_s, \psi_n, z_0, \{\lambda_k\}$)
2: **for** $k = 1, \ldots$ **do**

3: $z_k \leftarrow \underset{u \in \mathcal{Z}}{\operatorname{argmin}} \left\{ \lambda_k \left[\ell_{\psi_s}(u; z_{k-1}) + \psi_n(u) \right] + \frac{1}{2} \|u - z_{k-1}\|^2 \right\}$

4: $v_k \leftarrow \frac{1}{\lambda_k}(z_{k-1} - z_k) + \nabla \psi_s(z_k) - \nabla \psi_s(z_{k-1})$

The proposition below, whose proof can be found in Appendix A, presents some basic properties about the CGM.

Proposition 2.2.1. *Let $\{(z_k, v_k)\}_{k \geq 1}$ be generated by the CGM for some $\{\lambda_k\}_{k \geq 1}$. Then, the following statements hold for every $k \geq 1$:*

(a) $v_k \in \nabla \psi_s(z_k) + \partial \psi_n(z_k)$;

(b) if there exists $L_k \in (0, 2/\lambda_k)$ such that

$$\psi_s(z_k) - \ell_{\psi_s}(z_k; z_{k-1}) \leq \frac{L_k}{2} \|z_k - z_{k-1}\|^2, \tag{2.2}$$

then it holds that

$$\psi(z_k) < \psi(z_k) + \left(\frac{1}{\lambda_k} - \frac{L_k}{2} \right) \|z_{k-1} - z_k\|^2 \leq \psi(z_{k-1}); \tag{2.3}$$

(c) if there exists scalars $\{L_i\}_{i=1}^k \subseteq \mathbb{R}_{++}$ such that

$$\|\nabla \psi_s(z_{i-1}) - \nabla \psi_s(z_i)\| \leq L_i \|z_{i-1} - z_i\|, \quad L_i < \frac{2}{\lambda_i}, \tag{2.4}$$

for every $i \leq k$, then it holds that

$$\min_{i \leq k} \|v_i\|^2 \leq \frac{4 \left[\psi(z_0) - \psi(z_k) \right]}{\sum_{i=1}^k \xi_i \lambda_i}, \tag{2.5}$$

where $\xi_i := (2 - \lambda_i L_i)/(1 + [\lambda_i L_i]^2) > 0$ for every $i \leq k$.

The next proposition, whose proof can also be found in Appendix A, presents additional variational properties about a general iteration in the CGM.

Proposition 2.2.2. *Given* $(\lambda, z^-) \in \mathbb{R}_+ \times \mathcal{Z}$*, define*

$$z := \operatorname*{argmin}_{u \in \mathcal{Z}} \left\{ \lambda \left[\ell_{\psi_s}(u; z^-) + \psi_n(u) \right] + \frac{1}{2} \|u - z^-\|^2 \right\},$$
$$q := \frac{1}{\lambda}(z^- - z), \quad v := q + \nabla \psi_s(z) - \nabla \psi_s(z^-),$$
$$\varepsilon := \psi_n(z^-) - \psi_n(z) + \langle q - \nabla \psi_s(z^-), z - z^- \rangle.$$

Then, the following statements hold:

(a) $q \in \nabla \psi_s(z^-) + \partial_\varepsilon \psi_n(z^-)$ *and* $\varepsilon \geq 0$*;*

(b) it holds that

$$(q, \varepsilon) = \operatorname*{argmin}_{(r,\delta) \in \mathcal{Z} \times \mathbb{R}_+} \left\{ \lambda \|r\|^2 + 2\delta : r \in \nabla \psi_s(z^-) + \partial_\delta \psi_n(z^-) \right\}; \tag{2.6}$$

(c) if there exists $L > 0$ *satisfying*

$$\psi_s(z) - \ell_{\psi_s}(z; z^-) \leq \frac{L}{2} \|z - z^-\|^2,$$

then it holds that

$$\lambda \|q\|^2 + 2\varepsilon \leq 2 \left[\psi(z^-) - \psi(z) \right] + \left(L - \frac{1}{\lambda} \right) \|z^- - z\|^2.$$

2.2.2 Accelerated Composite Gradient (ACG) Method

Accelerated composite gradient (ACG) methods are extensions to the CGM in Section 2.2.1 in which additional computations are performed to improve the rate at which a near optimal solution (or stationary point) is obtained.

The ACG variant that we consider in this thesis is based on the accelerated method in [73]. More specifically, this ACG method (ACGM) assumes that $\psi_s \in \mathcal{F}_\mu(Z)$ for some $\mu \geq 0$

and, at its k^{th} iteration, performs the following update: given $(y_{k-1}, x_{k-1}) \in Z^2$, $A_{k-1} \geq 0$, and $\lambda_k > 0$, compute

$$\begin{aligned}
\tau_{k-1} &= \lambda_k(1 + \mu A_{k-1}), \\
a_{k-1} &= \frac{\tau_{k-1} + \sqrt{\tau_{k-1}^2 + 4\tau_{k-1}A_{k-1}}}{2}, \quad A_k = A_{k-1} + a_{k-1}, \\
\tilde{x}_{k-1} &= \frac{A_{k-1}}{A_k} y_{k-1} + \frac{a_{k-1}}{A_k} x_{k-1}, \\
q_k &\equiv \ell_{\psi_s}(\cdot; \tilde{x}_{k-1}) + \psi_n(\cdot) + \frac{\mu}{2}\|\cdot - \tilde{x}_{k-1}\|^2, \\
y_k &= \operatorname*{argmin}_{y \in \mathcal{Z}} \left\{ \lambda_k q_k(y) + \frac{1}{2}\|y - \tilde{x}_{k-1}\|^2 \right\}, \\
x_k &= x_{k-1} + \frac{a_{k-1}}{1 + \mu A_k} \left[\frac{1}{\lambda_k}(y_k - \tilde{x}_{k-1}) + \mu(y_k - x_{k-1}) \right].
\end{aligned}$$

Using the definition of the proximal operator $\operatorname{prox}_f(\cdot)$, it is straightforward to see that the updates for y_k and x_k can be written as

$$\begin{aligned}
\alpha_k &= \frac{\lambda_k}{1 + \lambda_k \mu}, \quad \beta_k = \frac{a_{k-1}}{1 + \mu A_k}, \\
y_k &= \operatorname{prox}_{\alpha_k \psi_n} \left(\tilde{x}_{k-1} - \alpha_k \nabla \psi_s(\tilde{x}_{k-1}) \right), \\
\gamma_{k-1} &\equiv q_k(y_k) + \frac{1}{\lambda_k} \langle \tilde{x}_{k-1} - y_k, \cdot - y_k \rangle + \frac{\mu}{2}\|\cdot - y_k\|^2, \\
x_k &= x_{k-1} - \beta_k \nabla \gamma_{k-1}(x_{k-1}).
\end{aligned}$$

where it can be shown (see Appendix B) that $\gamma_{k-1} \leq q_k \leq \psi$ for every $k \geq 1$. For ease of future reference and discussion, we give a precise description of this ACG method in Algorithm 2.2.2, which includes an important set of auxiliary iterates $\{(r_k, \tilde{r}_k, \eta_k, \tilde{\eta}_k, L_k)\}_{k \geq 1}$.

Algorithm 2.2.2: ACG Method

Require: $\mu \geq 0$, $\psi_n \in \overline{\operatorname{Conv}}\,(Z)$, $\psi_s \in \mathcal{F}_\mu(Z)$, $y_0 \in Z$, $\{\lambda_k\}_{k \geq 1} \subseteq \mathbb{R}_{++}$;
Initialize: $\psi \leftarrow \psi_s + \psi_n$, $A_0 \leftarrow 0$, $\Gamma_0 \Leftarrow 0$, $x_0 \leftarrow y_0$;

1: **procedure** ACG($\psi_s, \psi_n, \mu, y_0, \mu, \{\lambda_k\}$)

for $k = 1, \ldots$ **do**

PART 1 **Compute** the supporting quantities:

$\tau_{k-1} \leftarrow \lambda_k(1 + \mu A_{k-1})$

$a_{k-1} \leftarrow \dfrac{\tau_{k-1} + \sqrt{\tau_{k-1}^2 + 4\tau_{k-1}A_{k-1}}}{2}$

$A_k \leftarrow A_{k-1} + a_{k-1}$

$\tilde{x}_{k-1} \leftarrow \dfrac{A_{k-1}}{A_k} y_{k-1} + \dfrac{a_{k-1}}{A_k} x_{k-1}$

$q_k \Longleftarrow \ell_{\psi_s}(\cdot; \tilde{x}_{k-1}) + \psi_n(\cdot) + \dfrac{\mu}{2}\|\cdot - \tilde{x}_{k-1}\|^2$

PART 2 **Perform** the accelerated prox steps:

$y_k \leftarrow \operatorname*{argmin}_{y \in \mathcal{Z}} \left\{ \lambda_k q_k(y) + \dfrac{1}{2}\|y - \tilde{x}_{k-1}\|^2 \right\}$

$x_k \leftarrow x_{k-1} + \dfrac{a_{k-1}}{1 + A_k\mu} \left[\dfrac{1}{\lambda_k}(y_k - \tilde{x}_{k-1}) + \mu(y_k - x_{k-1}) \right]$

PART 3 **Compute** the auxiliary quantities:

$\gamma_{k-1} \Longleftarrow q_k(y_k) + \dfrac{1}{\lambda_k} \langle \tilde{x}_{k-1} - y_k, \cdot - y_k \rangle + \dfrac{\mu}{2}\|\cdot - y_k\|^2$

$\Gamma_k \Longleftarrow \dfrac{A_{k-1}}{A_k}\Gamma_{k-1} + \dfrac{a_{k-1}}{A_k}\gamma_{k-1}$

$r_k \leftarrow \dfrac{x_0 - x_k}{A_k}$

$\eta_k \leftarrow \psi(y_k) - \Gamma_k(x_k) - \langle r_k, y_k - x_k \rangle$

$\tilde{r}_k \leftarrow r_k + \mu(y_k - x_k)$

$\tilde{\eta}_k \leftarrow r_k + \dfrac{\mu}{2}\|y_k - x_k\|^2$

The next results, whose proofs are given in Appendix B, present some key properties about the ACGM and its generated iterates.

Proposition 2.2.3. *Let $\{(y_k, r_k, \eta_k)\}_{k\geq 1}$ be generated by the ACGM for some $\{\lambda_k\}_{k\geq 1}$. Then the following statements hold for every $k \geq 1$:*

(a) it holds that $\eta_k \geq 0$ and

$$r_k \in \partial_{\eta_k}\psi(y_k); \tag{2.7}$$

(b) if there exists $L_k > 0$ such that

$$\psi_s(y_k) - \ell_{\psi_s}(y_k; \tilde{x}_{k-1}) \leq \frac{L_k}{2}\|y_k - \tilde{x}_{k-1}\|^2, \quad L_k - \mu \leq \frac{1}{\lambda_k}, \tag{2.8}$$

then it holds that

$$\|A_k r_k + y_k - y_0\|^2 + 2A_k\eta_k \le \|y_k - y_0\|^2; \tag{2.9}$$

(c) it holds that

$$A_k \ge \max\left\{\frac{1}{4}\left(\sum_{i=1}^{k-1}\sqrt{\lambda_{k-1}}\right)^2, \lambda_1\prod_{i=2}^{k}\left(1+\sqrt{\frac{\lambda_{i-1}\mu}{2}}\right)^2\right\};$$

(d) for every minimizer y^* of the problem $\min_{y\in\operatorname{dom}\psi}\psi(y)$, it holds that

$$\psi(y_k) - \psi(y^*) \le \frac{1}{2A_k}\|y^* - y_0\|^2 \quad \forall k \ge 1.$$

Proposition 2.2.4. *Let $\{(y_k, \tilde{r}_k, \tilde{\eta}_k)\}_{k\ge1}$ be generated by the ACGM for some $\{\lambda_k\}_{k\ge1}$. Then the following statements hold for every $k \ge 1$:*

(a) it holds that $\eta_k \ge 0$ and

$$\tilde{r}_k \in \partial_{\tilde{\eta}_k}\left(\psi - \frac{\mu}{2}\|\cdot - y_k\|^2\right)(y_k); \tag{2.10}$$

(b) if there exists $L_k > 0$ such that

$$\psi_s(y_k) - \ell_{\psi_s}(y_k; \tilde{x}_{k-1}) \le \frac{L_k}{2}\|y_k - \tilde{x}_{k-1}\|^2, \quad L_k - \mu \le \frac{1}{\lambda_k}, \tag{2.11}$$

then it holds that

$$\left(\frac{1}{1+\mu A_k}\right)\|A_k\tilde{r}_k + y_k - y_0\|^2 + 2A_k\tilde{\eta}_k \le \|y_k - y_0\|^2; \tag{2.12}$$

2.2.3 Proximal Point Method

The proximal point (PP) method is a classic optimization algorithm [97] for minimizing a function $\psi \in \overline{\mathrm{Conv}}\,(\mathcal{Z})$. More specifically, it is an iterative method that, at its k^{th} iteration, performs the following update: given $z_{k-1} \in \mathrm{dom}\,\psi$ and $\lambda_k > 0$, perform

$$z_k = \mathrm{prox}_{\lambda_k \psi}(z_{k-1}). \tag{2.13}$$

It is well-known (see, for example, [6, Theorem 27.1]) that if $\sum_{k=1}^{\infty} \lambda_k = \infty$, then $\psi(z_k)$ converges to $\inf_{z \in \mathcal{Z}} \phi(z)$. Moreover, if there exists z^* satisfying $\phi(z^*) = \inf_{z \in \mathcal{Z}} \phi(z)$, then z_k converges to the set of minimizers of ϕ.

The following proposition, whose proof can be found in Appendix A, presents some basic properties (cf. Proposition 2.2.1) about the PP method (PPM).

Proposition 2.2.5. *Let $\{z_k\}_{k \geq 1}$ be generated by the PPM for some $\{\lambda_k\}_{k \geq 1}$ and define $v_k := (z_{k-1} - z_k)/\lambda_k$ for every $k \geq 1$. Then, the following statements hold for every $k \geq 1$:*

(a) $v_k \in \partial \psi(z_k)$;

(b) it holds that

$$\psi(z_k) < \psi(z_k) + \frac{1}{\lambda_k} \|z_k - z_{k-1}\|^2 \leq \psi(z_{k-1});$$

(c) it holds that

$$\min_{i \leq k} \|v_i\|^2 \leq \frac{\psi(z_0) - \psi(z_k)}{\sum_{i=1}^{k} \lambda_i}. \tag{2.14}$$

Throughout this thesis, we make reference to the **inexact** proximal point method which is a variant of the PPM in which the update (2.13) is computed inexactly, i.e. z_k approximates the solution of the problem in (2.13) according to some inexactness criterion.

One interesting instance of the proximal point method is when $\psi(x) = (1/2)\langle x, Ax\rangle - \langle b, x\rangle$ where $A \in \mathbb{S}^n_{++}$. Clearly, the optimal solution of $\min_{x \in \mathbb{R}^n} \psi(x)$ is the unique solution

of the linear system of equations $Ax = b$. The proximal update in the case of $\lambda_k = \lambda \in \mathbb{R}_{++}$ for every $k \geq 1$ is

$$x_{k+1} = x_k + (A + \lambda^{-1} I_n)^{-1}(b - Ax_k),$$

which is a well-known algorithm called iterative refinement [92]. The above update is particularly useful when A is ill-conditioned and/or the computation of $A^{-1}b$ is not stable.

CHAPTER 3
UNCONSTRAINED COMPOSITE OPTIMIZATION

Our main goal in this chapter is to describe and establish the iteration complexity of an accelerated **inexact** proximal point (AIPP) method for finding approximate stationary points of the classic NCO problem

$$\phi_* = \min_{z \in \mathcal{Z}} [\phi(z) := f(z) + h(z)], \tag{$\mathcal{NCO}$}$$

where $\mathcal{Z}$ is a finite dimensional inner product space, $h \in \overline{\mathrm{Conv}}\,(Z)$ for some nonempty convex set $Z \subseteq \mathcal{Z}$, and $f \in \mathcal{C}_{m,M}(Z)$ for some $(m, M) \in \mathbb{R}^2_{++}$.

The AIPP method (AIPPM) of this chapter uses an ACGM, specifically Algorithm 2.2.2, to perform the following proximal point-type update to generate its k^{th} iterate: given z_{k-1} and λ, compute

$$z_k \approx \min_{z \in \mathcal{Z}} \left\{ f(z) + h(z) + \frac{1}{2\lambda} \|z - z_{k-1}\|^2 \right\}$$

according to some **relative** inexactness criterion. Throughout our presentation, it is assumed that efficient oracles for evaluating the quantities $f(z)$, $\nabla f(z)$, and $h(z)$ and for obtaining exact solutions of the subproblem

$$\min_{z \in \mathcal{Z}} \left\{ \lambda h(z) + \frac{1}{2} \|z - z_0\|^2 \right\},$$

for any $z_0 \in \mathcal{Z}$ and $\lambda > 0$, are available. Moreover, we define an **oracle call** to be a collection of the above oracles of size $\mathcal{O}(1)$ where each of them appears at least once.

For a given tolerance $\hat{\rho} > 0$ and a suitable choice of λ, the main result of this chapter shows that the AIPPM, started from any point $z_0 \in Z$ obtains a pair $(\hat{z}, \hat{v})$ satisfying the

approximate stationarity condition

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}), \quad \|\hat{v}\| \leq \hat{\rho}, \tag{3.1}$$

in

$$\mathcal{O}\left(\sqrt{\frac{M}{m}+1}\left[\frac{m\cdot\min\{\phi(z_0)-\phi_*, md_0^2\}}{\hat{\rho}^2} + \log_1^+\left(\frac{M}{m}\right)\right]\right) \tag{3.2}$$

oracle calls, where $d_0 = \min_{z\in\mathcal{Z}}\{\|z_0 - z_*\| : \phi(z_*) = \phi_*\}$ and $\log_1^+(\cdot) = \max\{1, \log(\cdot)\}$. It is worth mentioning that this result is obtained under the mild assumption that ϕ_* is finite and neither assumes neither that Z is bounded nor that $\mathcal{NCO}$ has an optimal solution. Near the end of the chapter, we compare the above complexity against ones obtained by other NCO methods.

It is also shown in Section 3.3.3 that the complexity bound in (3.2) is optimal in the sense that it is within the same order of magnitude of a recent established complexity lower bound for finding pairs $(\hat{z}, \hat{v})$ satisfying (3.1) using linear-span first-order methods.

The content of this chapter is based on paper [46] (joint work with Jefferson G. Melo and Renato D.C. Monteiro) and several passages may be taken verbatim from it.

Related Works

The developments in [46] appear to be the first ones to consider an accelerated proximal method for obtaining approximate stationary points as in (3.1) for general h and nonconvex f. Previous developments, which we list below, have only considered the special case of $h = 0$.

Under the assumption that $\operatorname{dom}\phi$ is bounded, paper [30] presents an ACG method applied directly to $\mathcal{NCO}$ which obtains a pair $(\hat{z}, \hat{v})$ satisfying (3.1) in

$$\mathcal{O}\left(\frac{MmD_z^2}{\hat{\rho}^2} + \left[\frac{Md_0}{\hat{\rho}}\right]^{2/3}\right) \tag{3.3}$$

where D_z denotes the diameter of $\operatorname{dom}\phi$. Motivated by the developments in [30], other papers, such as [18, 26, 31, 56, 59, 60, 61, 91], have proposed ACG-like methods under different assumptions on the functions f and h. For example, paper [18] establishes a complexity which is $\mathcal{O}(\sqrt{M}\log M)$ in terms of its dependence on M, but is $\mathcal{O}(\hat{\rho}^{-2}\log\hat{\rho}^{-1})$ in terms of its dependence on $\hat{\rho}$. It should be noted that the second complexity bound in (3.2) in terms of d_0 is new in the context of problem $\mathcal{NCO}$ and follows as a special case of a more general bound, namely (3.3.6), which actually unifies both bounds in (3.2). Moreover, in contrast to the analysis of [30], the analysis in this chapter does not assume that D_z in (3.3) is finite.

Inexact proximal point methods and HPE variants of the ones studied in [74, 104] for solving convex-concave saddle point problems and monotone variational inequalities — which inexactly solve a sequence of proximal subproblems by means of an ACG variant — were previously proposed by [36, 37, 45, 75, 90]. The behavior of an accelerated gradient method near saddle points is studied in [88].

Complexity lower bounds in terms of $\max\{m, M\}$ for finding stationary points as in (3.1) using first-order methods were recently established in [19, 20]. A follow-up work [116] establishes tighter bounds in terms of m and M for the smaller class of linear-span first-order methods.

Organization

This chapter contains three sections. The first one gives some preliminary references and discusses our notion of a stationary point given in (3.1). The second one presents a general inexact proximal point framework which will be important in our analysis of the AIPPM. The third one presents the AIPPM and its iteration complexity. The last one gives a conclusion and some closing comments.

3.1 Preliminaries

This section enumerates the assumptions on problem $\mathcal{NCO}$, states the main problem of interest, and discusses the notion of an approximate stationary point given in (3.1).

It is assumed that (f, h, ϕ) in $\mathcal{NCO}$ satisfy:

(A1) $h \in \overline{\text{Conv}}\,(Z)$ for some nonempty convex set $Z \subseteq \mathcal{Z}$;

(A2) $f \in \mathcal{C}_{m,M}(Z)$ for some $(m, M) \in \mathbb{R}^2_{++}$;

(A3) $\phi_* > -\infty$.

We now make a few remarks about the above assumptions. First, assumption (A1) implies that the effective domain of h is Z. Second, if ∇f is M-Lipschitz continuous, then assumption (A2) holds with $m = M$. Third, it is well-known that a necessary condition for $z^* \in Z$ to be a local minimum of $\mathcal{NCO}$ is that z^* be a stationary point of $f+h$, i.e. $0 \in \nabla f(z^*) + \partial h(z^*)$.

In view of the above assumptions and remarks, we are interested in solving the problem given in Problem 3.1.1.

Problem 3.1.1: Find an approximate stationary point of $\mathcal{NCO}$

Given $\hat{\rho} > 0$, find a pair $(\hat{z}, \hat{v}) \in Z \times \mathcal{Z}$ satisfying condition (3.1).

The next proposition, which follows from Lemma F.1.2, gives another well-known (see, for example, [86]) interpretation of our notion of an approximate stationary point.

Proposition 3.1.1. *Given $\hat{z} \in Z$, there exists $\hat{v} \in \mathcal{Z}$ such that $(\hat{z}, \hat{v})$ satisfies* (3.1) *if and only if* $\inf_{\|d\| \le 1} \phi'(\hat{z}; d) \ge -\hat{\rho}$.

3.2 General Inexact Proximal Point (GIPP) Framework

This section presents and discusses general inexact proximal point framework which will be important in our analysis of the AIPPM. It contains three subsections. The first one presents some important properties of the framework. The second one presents a procedure to turn iterates generated by the framework into iterates that *nearly* solve Problem 3.1.1. The last one gives some instances of the framework.

We begin by first stating the framework in Algorithm 3.2.1.

Algorithm 3.2.1: GIPP Framework

Require: $h \in \overline{\mathrm{Conv}}\,(Z)$, $f \in \mathcal{C}(Z)$, $z_0 \in Z$, $\sigma \in (0,1)$, $\{\lambda_k\}_{k\geq 1} \subseteq \mathbb{R}_{++}$;

1: **procedure** GIPP($f, h, z_0, \sigma, \{\lambda_k\}_{k\geq 1}$)
2: **for** $k = 1, \ldots$ **do**
3: Find $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k) \in \mathrm{dom}\, h \times \mathcal{Z} \times \mathbb{R}_+$ satisfying:

$$\tilde{v}_k \in \partial_{\tilde{\varepsilon}_k}\left(\lambda_k \phi + \frac{1}{2}\|\cdot - z_{k-1}\|^2\right)(z_k), \tag{3.4}$$

$$\|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k \leq \sigma \|z_{k-1} - z_k + \tilde{v}_k\|^2; \tag{3.5}$$

Observe that the GIPP framework (GIPPF) is not a well-specified algorithm but rather a conceptual framework consisting of (possibly many) specific instances. In particular, it does not specify how the quadruple $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ is computed or even if it exists. Later in this chapter, we will discuss two specific instances of the above GIPPF for solving $\mathcal{NCO}$, namely, the CGM (see Algorithm 2.2.1) and an accelerated proximal point method presented in Section 3.3. In both of these instances, the sequences $\{\tilde{z}_k\}_{k\geq 1}$ and $\{\tilde{\varepsilon}_k\}_{k\geq 1}$ are non-trivial (see Proposition 3.2.6 and Lemma 3.3.4(c)).

3.2.1 Key Properties of the Framework

This subsection presents some key properties of the GIPPF.

Let $\{(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}_{k\geq 1}$ be the sequence generated by an instance of the GIPPF for some $\{\lambda_k\}_{k\geq 1}$, and consider the sequence $\{(v_k, \varepsilon_k)\}_{k\geq 1}$ defined as

$$(v_k, \varepsilon_k) := \frac{1}{\lambda_k}(\tilde{v}_k, \tilde{\varepsilon}_k) \quad \forall k \geq 1. \tag{3.6}$$

Without necessarily assuming that the error condition (3.5) holds, the following technical but straightforward result derives bounds on $\tilde{\varepsilon}_k$ and $\|\tilde{v}_k + z_{k-1} - z_k\|/\lambda_k$ in terms of the quantities

$$\delta_k = \delta_k(\sigma) := \frac{1}{\lambda_k} \max\left\{0, \|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k - \sigma\|z_{k-1} - z_k + \tilde{v}_k\|^2\right\}, \quad \Lambda_k := \sum_{i=1}^{k} \lambda_i \tag{3.7}$$

where $\sigma \in [0, 1)$ is a given parameter. Note that if (3.5) is assumed, then $\delta_k = 0$.

Lemma 3.2.1. *Assume that the sequence* $\{(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$ *satisfies* (3.4) *and let* $\sigma \in (0, 1)$ *be given. Then, for every* $k \geq 1$*, there holds*

$$\frac{1}{\sigma\lambda_k}\left(\|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k - \lambda_k\delta_k\right) \leq \frac{1}{\lambda_k}\|z_{k-1} - z_k + \tilde{v}_k\|^2 \leq \frac{2[\phi(z_{k-1}) - \phi(z_k)] + \delta_k}{1 - \sigma} \tag{3.8}$$

where δ_k *is as in* (3.7).

Proof. First note that the inclusion in (3.4) is equivalent to

$$\lambda_i\phi(z) + \frac{1}{2}\|z - z_{i-1}\|^2 \geq \lambda_i\phi(z_i) + \frac{1}{2}\|z_i - z_{i-1}\|^2 + \langle\tilde{v}_i, z - z_i\rangle - \tilde{\varepsilon}_i \qquad \forall z \in \mathfrak{R}^n.$$

Setting $z = z_{i-1}$ in the above inequality and using the definition of δ_i given in (3.7), we obtain

$$\begin{aligned}\lambda_i(\phi(z_{i-1}) - \phi(z_i)) &\geq \frac{1}{2}\left(\|z_{i-1} - z_i\|^2 + 2\langle\tilde{v}_i, z_{i-1} - z_i\rangle - 2\tilde{\varepsilon}_i\right)\\ &= \frac{1}{2}\left[\|z_{i-1} - z_i + \tilde{v}_i\|^2 - \|\tilde{v}_i\|^2 - 2\tilde{\varepsilon}_i\right] \geq \frac{1}{2}\left[(1-\sigma)\|z_{i-1} - z_i + \tilde{v}_i\|^2 - \lambda_i\delta_i\right]\end{aligned}$$

and hence the proof of the second inequality in (3.8) follows after simple rearrangements.

The first inequality in (3.8) follows immediately from (3.7). □

The next result shows characterizes the approximate optimality of z_k in terms of λ_k, z_{k-1}, and σ.

Lemma 3.2.2. *Let $\{(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$ be generated by an instance of the GIPPF for some $\{\lambda_k\}_{k\geq 1}$. Then, for every $u \in \mathcal{Z}$, it holds that*

$$\phi(z_k) \leq \phi(u) + \frac{1}{2(1-\sigma)\lambda_k}\|z_{k-1} - u\|^2 \quad \forall k \geq 1.$$

Proof. Using some simple algebraic manipulation, it is easy to see that (3.5) yields

$$\langle \tilde{v}_k, z_k - z_{k-1}\rangle + \frac{1}{\sigma}\tilde{\varepsilon}_k - \frac{1}{2}\|z_{k-1} - z_k\|^2 \leq -\frac{1-\sigma}{2\sigma}\|\tilde{v}_k\|^2. \tag{3.9}$$

Now, letting $\theta := (1-\sigma)/\sigma > 0$, recalling the definition of the approximate subdifferential, using (3.4) and (3.9), and the fact that $\langle v, v'\rangle \leq (\theta/2)\|v\|^2 + (1/2\theta)\|v'\|^2$ for all $v, v' \in \mathcal{Z}$, we conclude that

$$\begin{aligned}
\lambda_k[\phi(z_k) - \phi(u)] &\leq \frac{1}{2}\|z_{k-1} - u\|^2 + \langle \tilde{v}_k, z_k - u\rangle + \tilde{\varepsilon}_k - \frac{1}{2}\|z_k - z_{k-1}\|^2 \\
&\leq \frac{1}{2}\|z_{k-1} - u\|^2 + \langle \tilde{v}_k, z_{k-1} - u\rangle - \frac{1-\sigma}{2\sigma}\|\tilde{v}_k\|^2 \\
&\leq \frac{1}{2}\|z_{k-1} - u\|^2 + \left(\frac{\theta}{2}\|\tilde{v}_k\|^2 + \frac{1}{2\theta}\|z_{k-1} - u\|^2\right) - \frac{1-\sigma}{2\sigma}\|\tilde{v}_k\|^2,
\end{aligned}$$

and hence that the conclusion of the lemma holds due to the definition of θ.

Before proceeding, we define the following useful quantity

$$R_\lambda\psi(z_0) := \inf_{u\in\mathcal{Z}}\left[R_\lambda\psi(u; z_0) := \frac{1}{2}\|z_0 - u\|^2 + \lambda\left[\psi(u) - \inf_{\tilde{z}\in\mathcal{Z}}\psi(\tilde{z})\right]\right] \tag{3.10}$$

for any function $\psi : \mathcal{Z} \mapsto (-\infty, \infty]$, scalar $\lambda \geq 0$, and point $z_0 \in \mathcal{Z}$. Clearly, $R_\lambda\psi(u; z_0) \in \mathbb{R}_+$

for all $u \in Z$ and hence $R_\lambda\psi(z_0) \in \mathbb{R}_+$ as well. Moreover, it is easy to see that

$$R_\lambda\psi(z_0) = \lambda\left[e_\lambda\psi(z_0) - \inf_{u\in\mathcal{Z}}\psi(u)\right] \le \lambda\left[\psi(z_0) - \inf_{u\in\mathcal{Z}}\psi(u)\right], \tag{3.11}$$

where $e_\lambda\psi(z_0)$ denotes the λ-Moreau envelope of ψ at z_0.

We now show that the sequence $\{\|z_{k-1} - z_k + \tilde{v}_k\|/\lambda_k\}_{k\ge1}$ contains a subsequence that tends to zero. □

Proposition 3.2.3. *Let $\{(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}_{k\ge1}$ be generated by an instance of the GIPPF for some $\{\lambda_k\}_{k\ge1}$. Then, the following statements hold:*

(a) for every $k \ge 1$,

$$\frac{1-\sigma}{2\lambda_k}\|z_{k-1} - z_k + \tilde{v}_k\|^2 \le \phi(z_{k-1}) - \phi(z_k); \tag{3.12}$$

(b) for every $k \ge 2$, there exists $i \le k$ such that

$$\frac{1}{\lambda_i^2}\|z_{i-1} - z_i + \tilde{v}_i\|^2 \le \frac{2R_{\lambda_1}\phi(z_0)}{(1-\sigma)^2\lambda_1(\Lambda_k - \lambda_1)} = \frac{2\left[e_{\lambda_1}\phi(z_0) - \phi_*\right]}{(1-\sigma)^2(\Lambda_k - \lambda_1)} \tag{3.13}$$

where Λ_k and $R_{\lambda_1}\phi(z_0)$ are as in (3.7) *and* (3.10)*, respectively.*

Proof. (a) This follows immediately from (3.8) and the fact that (3.5) is equivalent to $\delta_k = 0$.

(b) It follows from definitions of ϕ_* and $R_{\lambda_1}\phi(\cdot; z_0)$ in (A3) and (3.10), respectively, part (a) and Lemma 3.2.2 with $k = 1$ that for all $u \in \mathcal{Z}$,

$$\begin{aligned}
\frac{R_{\lambda_1}\phi(u; z_0)}{(1-\sigma)\lambda_1} &= \left(\frac{1}{1-\sigma}\right)\left[\frac{1}{2\lambda_1}\|z_0 - u\|^2 + \phi(u) - \phi_*\right] \\
&\ge \frac{1}{2\lambda_1(1-\sigma)}\|z_0 - u\|^2 + \phi(u) - \phi_* \\
&\ge \phi(z_1) - \phi_* \ge \sum_{i=2}^{k}[\phi(z_{i-1}) - \phi(z_i)] \\
&\ge (1-\sigma)\sum_{i=2}^{k}\frac{\|z_{i-1} - z_i + \tilde{v}_i\|^2}{2\lambda_i} \\
&\ge \frac{(1-\sigma)(\Lambda_k - \lambda_1)}{2}\min_{i\le k}\frac{1}{\lambda_i^2}\|z_{i-1} - z_i + \tilde{v}_i\|^2
\end{aligned}$$

and hence the first inequality of (3.13) holds in view of the definition of $R_{\lambda_1}\phi(z_0)$ in (3.10). The second inequality follows from (3.11).

Note that the above proposition shows the GIPPF enjoys the descent property in Proposition 3.2.3, which many frameworks and/or algorithms for finding approximate stationary points of $\mathcal{NCO}$ also share, e.g. Algorithm 2.2.1. It is worth noting that, under the assumption that ϕ is a KL-function, frameworks and/or algorithms sharing this property have also been developed for example in [2, 3, 22, 29] where it is shown that the generated sequence $\{z_k\}_{k\geq 1}$ converges to some stationary point of $\mathcal{NCO}$ with a well-characterized asymptotic (but not global) convergence rate, as long as $\{z_k\}_{k\geq 1}$ has an accumulation point.

The following result, which follows immediately from Proposition 3.2.3, considers the instances of the GIPPF where λ_k is constant for every $k \geq 1$. For the purpose of stating it, define

$$d_0 := \inf_{z^* \in \mathcal{Z}} \{\|z_0 - z^*\| : \phi(z^*) = \phi_*\}. \tag{3.14}$$

Note that $d_0 < \infty$ if and only if $\mathcal{NCO}$ has an optimal solution, in which case the above infimum can be replaced by a minimum in view of the first assumption following $\mathcal{NCO}$. □

Corollary 3.2.4. *Let $\{(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}$ be generated by an instance the GIPPF with $\lambda_k = \lambda$ for every $k \geq 1$, and define $\{(v_k, \varepsilon_k, r_k)\}$ as in* (3.6)*. Then, the following statements hold:*

(a) for every $k \geq 2$, there exists $i \leq k$ such that

$$\frac{1}{\lambda^2}\|z_{i-1} - z_i + \tilde{v}_i\|^2 \leq \frac{2R_\lambda\phi(z_0)}{\lambda^2(1-\sigma)^2(k-1)} \leq \frac{\min\left\{2[\phi(z_0) - \phi_*], \frac{d_0^2}{\lambda(1-\sigma)}\right\}}{\lambda(1-\sigma)(k-1)} \tag{3.15}$$

where $R_\lambda\phi(z_0)$ and d_0 are as in (3.10) *and* (3.14)*, respectively;*

(b) for any $\tau > 0$, the GIPPF generates a quadruple $(z^-, z, \tilde{v}, \tilde{\varepsilon})$ such that

$$\tilde{v} \in \partial_{\tilde{\varepsilon}}\left(\lambda\phi + \frac{1}{2}\|\cdot - z^-\|^2\right)(z), \quad \frac{1}{\lambda}\|z^- - z + \tilde{v}\| \leq \tau, \quad \frac{1}{\lambda}\tilde{\varepsilon} \leq \left(\frac{\sigma\lambda}{2}\right)\tau^2, \tag{3.16}$$

in a number of iterations bounded by

$$\left\lceil \frac{2R_\lambda\phi(z_0)}{\lambda^2(1-\sigma)^2\tau^2} + 1 \right\rceil. \tag{3.17}$$

Proof. (a) The proof of the first inequality follows immediately from Proposition 3.2.3(b) and the fact that $\lambda_k = \lambda$ for every $k \geq 1$. Now, note that due to (3.10), we have $R_\lambda\phi(z_0) \leq R_\lambda\phi(z_0; z_0) = \lambda[\phi(z_0) - \phi_*]$ and $R_\lambda\phi(z_0) \leq R_\lambda\phi(z^*; z_0) = \|z^* - z_0\|^2/2$ for any z^* satisfying $\phi(z^*) = \phi_*$. The second inequality now follows from the previous observation and the definition of d_0 in (3.14).

(b) This statement follows immediately from the first inequality in (a) and (3.5). □

In the above analysis, we have assumed that ϕ is quite general. For the remainder of this chapter, we derive results that use the composite structure underlying ϕ, i.e. $\phi = f + h$ where f and h satisfy conditions (A1)–(A3).

3.2.2 Generating Stationary Points

In the previous subsection, we established that the GIPPF is able to generate a quadruple $(z^-, z, \tilde{v}, \tilde{\varepsilon})$ which satisfies (3.16) for any $\tau > 0$. In this subsection, we present a refinement procedure that uses the above quadruple to generate a pair $(\hat{z}, \hat{v}) \in Z \times \mathcal{Z}$ which, for sufficiently small enough $\tau > 0$, satisfies (3.1).

We begin by presenting the aforementioned procedure in Algorithm 3.2.2.

Algorithm 3.2.2: CR Procedure

Require: $h \in \overline{\mathrm{Conv}}\,(\mathcal{Z})$, $f \in \mathcal{C}(Z)$, $z \in \mathcal{Z}$, $L > 0$, $\lambda > 0$;
Initialize: $L_\lambda \leftarrow L + \lambda^{-1}$;

1: **procedure** CREF(f, h, z, L, λ)
2: $\quad z_r \leftarrow \operatorname{argmin}_{u \in \mathcal{Z}} \left\{ \ell_f(u; z) + h(u) + \frac{L_\lambda}{2}\|u - z\|^2 \right\}$
3: $\quad q_r \leftarrow L_\lambda(z - z_r)$

4: $v_r \leftarrow q_r + \nabla f(z_r) - \nabla f(z)$
5: $\varepsilon_r \leftarrow h(z) - h(z_r) - \langle q_r - \nabla f(z), z - z_r \rangle$
6: **return** $(z_r, q_r, v_r, \varepsilon_r)$

The result below, whose proof can be found in Appendix D, presents some important properties about the CR procedure (CRP).

Proposition 3.2.5. *Let $(z_r, q_r, v_r, \varepsilon_r)$ and L_λ be generated by the CRP where (f, h) satisfy assumptions (A1)–(A2). Then, the following statements hold:*

(a) $q_r \in \nabla f(z) + \partial_{\varepsilon_r} h(z)$ *and* $\varepsilon_r \geq 0$*;*

(b) $v_r \in \nabla f(z_r) + \partial h(z_r)$ *and*

$$(f + h)(z) - (f + h)(z_r) \geq \frac{L_\lambda}{2}\|z - z_r\|^2;$$

(c) if the inputs f, h, λ, and z satisfy

$$\begin{aligned} \tilde{v} \in \partial_{\tilde{\varepsilon}}\left(\lambda\left[f + h\right] + \frac{1}{2}\|\cdot - z^-\|^2\right)(z), \\ \frac{1}{\lambda}\|z^- - z + \tilde{v}\| \leq \bar{\rho}, \quad \frac{1}{\lambda}\tilde{\varepsilon} \leq \bar{\varepsilon}, \end{aligned} \tag{3.18}$$

for some $(\bar{\rho}, \bar{\varepsilon}) \in \mathbb{R}^2_{++}$ and $(z^-, \tilde{v}, \tilde{\varepsilon}) \in \mathcal{Z} \times \mathcal{Z} \times \mathbb{R}_+$, then

$$\|v_r\| \leq \left(1 + \frac{\max\{m, M\}}{L_\lambda}\right)\|q_r\|, \quad \|q_r\| \leq \bar{\rho} + \sqrt{2\bar{\varepsilon}L_\lambda}. \tag{3.19}$$

The above proposition shows that if $(\tilde{v}, \tilde{\varepsilon}, z, z^-)$ satisfies the inclusion in (3.18) and the residuals $\tilde{\varepsilon}/\lambda$ and $\|z^- - z + \tilde{v}\|/\lambda$ are sufficiently small enough relative to some tolerance $\hat{\rho}$, then the CRP generates a pair $(\hat{z}, \hat{v})$ that solves Problem 3.1.1. Since, Corollary 3.2.4 shows that instances of the GIPPF are able to send the aforementioned residuals to zero along some subsequence, one approach is to iterate an instance of the GIPPF and check if the output

of a call to the CRP, as above, satisfies (3.1). The AIPPM is essentially one method that implements this approach.

3.2.3 Instances of the GIPPF

In this subsection, we briefly discuss some specific instances of the GIPPF.

Recall that, for given stepsize $\lambda > 0$ and initial point $z_0 \in Z$, the CGM in Algorithm 2.2.1 for solving $\mathcal{NCO}$ recursively computes a sequence $\{z_k\}_{k\geq 1}$ given by

$$z_k = \operatorname*{argmin}_{u\in\mathcal{Z}} \left\{ \lambda \left[\ell_g(u; z_{k-1}) + h(u) \right] + \frac{1}{2} \|z - z_{k-1}\|^2 \right\}. \tag{3.20}$$

Note that if h is the indicator function of a closed convex set then the above scheme reduces to the classical projected gradient method.

The following result, whose proof can be found in Appendix A, shows that the CGM with λ sufficiently small is a special case of the GIPPF in which $\lambda_k = \lambda$ for all $k \geq 1$.

Proposition 3.2.6. *Let $\{z_k\}_{k\geq 1}$ be generated by the CGM with $\lambda_k = \lambda \leq 1/m$ and $\lambda < 2/M$ for every $k \geq 1$, and define $\tilde{v}_k := z_{k-1} - z_k$ and*

$$\tilde{\varepsilon}_k := \lambda \left[g(z_k) - \ell_g(z_k; z_{k-1}) + \frac{1}{2\lambda} \|z_k - z_{k-1}\|^2 \right]. \tag{3.21}$$

Then, for every $k \geq 1$, the quadruple $(\lambda_k, z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ satisfies the inclusion (3.4) *with $\phi = g + h$, and the relative error condition* (3.5) *with $\sigma := (\lambda M + 2)/4$. Thus, the CGM can be seen as an instance of the GIPPF.*

Under the assumption that $\lambda < 2/M$ and $g \in \mathcal{C}_M(\mathcal{Z})$, it is well-known that the CGM solves Problem 3.1.1 in $\mathcal{O}([\phi(z_0) - \phi_*]/[\lambda\hat{\rho}^2])$ iterations. On the other hand, under the assumption that $\lambda \leq 1/M$ and $g \in \mathcal{C}_M(\mathcal{Z})$, we can easily see that the above result together with Corollary 3.2.4(b) imply that the CGM solves Problem 3.1.1 in $\mathcal{O}(R_\lambda \phi(z_0)/[\lambda^2\hat{\rho}^2])$ iterations.

We now make a few general remarks about our discussion in this subsection so far. First, the condition on the stepsize λ of Proposition 3.2.6 forces it to be $\mathcal{O}(1/M)$ and hence quite small whenever $M \gg m$. Second, Corollary 3.2.4(b) implies that the larger λ is, the smaller the complexity bound (3.17) becomes. Third, letting $\lambda_k = \lambda$ in the GIPPF for some $\lambda \leq 1/m$ guarantees that the function $\lambda_k\phi + \|\cdot - z_{k-1}\|^2/2$ that appears in (3.4) is convex.

In the remaining part of this subsection, we briefly outline the ideas behind an accelerated instance of the GIPPF which chooses $\lambda = \mathcal{O}(1/m)$. First, note that when $\sigma = 0$, (3.4) and (3.5) imply that $(\tilde{v}_k, \tilde{\varepsilon}_k) = (0, 0)$ and

$$0 \in \partial\left(\lambda_k\phi + \frac{1}{2}\|\cdot - z_{k-1}\|^2\right)(z_k). \tag{3.22}$$

and hence that z_k is an optimal solution of the prox-subproblem

$$z_k = \operatorname*{argmin}_{z\in\mathcal{Z}}\left\{\lambda_k\phi(z) + \frac{1}{2}\|z - z_{k-1}\|^2\right\}. \tag{3.23}$$

More generally, assuming that (3.5) holds for some $\sigma > 0$ gives us an interpretation of z_k, together with $(\tilde{v}_k, \tilde{\varepsilon}_k)$, as being an approximate solution of (3.23) where its (relative) accuracy is measured by the σ-criterion (3.5). Obtaining such an approximate solution is generally difficult unless the objective function of the prox-subproblem (3.23) is convex. This suggests choosing $\lambda_k = \lambda$ for some $\lambda \leq 1/m$ which, according to a remark in the previous paragraph, ensures that $\lambda_k\phi + (1/2)\|\cdot\|^2$ is convex for every k, and then applying an ACGM, e.g. Algorithm 2.2.2, to the (convex) prox-subproblem (3.23) to obtain z_k and a certificate pair $(\tilde{v}_k, \tilde{\varepsilon}_k)$ satisfying (3.5). An accelerated prox-instance of the GIPPF obtained in this manner will be the subject of Section 3.3.

3.3 Accelerated Inexact Proximal Point (AIPP) Method

The main goal of this section is to present another instance of the GIPPF where the triples $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ are obtained by applying an ACGM, e.g. Algorithm 2.2.2, to the subproblem (3.23). It contains two subsections. The first one discusses some new results of the ACGM which will be useful in the analysis of the accelerated GIPP instance. The second one presents the accelerated GIPP instance for solving $\mathcal{NCO}$ and derives its corresponding iteration complexity bound.

3.3.1 Key Properties of the ACGM

The main role of the ACGM is to find an approximate solution z_k of subproblem (3.4) together with a certificate pair $(\tilde{v}_k, \tilde{\varepsilon}_k)$ satisfying (3.4) and (3.5). Indeed, since (3.23) is a special case of $\mathcal{CO}$, we can apply the ACGM (see Algorithm 2.2.2) with $x_0 = z_{k-1}$ to obtain the triple $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ satisfying (3.4) and (3.5).

The following result analyzes the iteration complexity of computing the aforementioned triple.

Lemma 3.3.1. *Let $\{(A_j, y_j, r_j, \eta_j)\}_{j\geq 1}$ be the sequence generated by the ACGM applied to $\mathcal{CO}$, where:*

(i) $\psi_n \in \overline{\mathrm{Conv}}\,(\mathcal{Z})$ and $\psi_s \in \mathcal{F}_{\mu,L}(\mathrm{dom}\,\psi_n)$ for some $L > 0$ and $\mu \geq 0$;

(ii) $\lambda_k = 1/L$ for every $k \geq 1$.

Then, for any $\sigma > 0$ and index j such that $A_j \geq 2(1+\sqrt{\sigma})^2/\sigma$, we have

$$\|r_j\|^2 + 2\eta_j \leq \sigma\|y_0 - y_j + r_j\|^2. \tag{3.24}$$

As a consequence, the ACGM obtains a triple $(y, r, \eta) = (y_j, r_j, \eta_j)$ satisfying

$$r \in \partial_\eta(\psi_s + \psi_n)(y) \quad \|r\|^2 + 2\eta \leq \sigma\|y_0 - y + r\|^2$$

in at most

$$\min\left\{\left\lceil \frac{2\sqrt{2L}(1+\sqrt{\sigma})}{\sqrt{\sigma}} \right\rceil, \left\lceil 1+\sqrt{\frac{2L}{\mu}}\log_1^+\left(\frac{2L\left[1+\sqrt{\sigma}\right]^2}{\sigma}\right)\right\rceil\right\}$$

iterations, where $\log_1^+(\cdot) \coloneqq \max\{\log(\cdot), 1\}$.

Proof. See Appendix B. □

Note that the above lemma holds for any $\mu \geq 0$. On the other hand, the next two results hold only for $\mu > 0$ and derive some important relations satisfied by two distinct iterates of the ACGM.

Lemma 3.3.2. *Let* $\{(A_j, y_j, r_j, \eta_j)\}_{j\geq 1}$ *and* (ψ_s, ψ_n) *be as in Lemma 3.3.1 with* $\mu > 0$. *Then,*

$$\left(1-\left[A_i\mu\right]^{-1/2}\right)\|y^*-y_0\| \leq \|y_j-y_0\| \leq \left(1+\left[A_j\mu\right]^{-1/2}\right)\|y^*-y_0\| \qquad \forall\, j\geq 1, \tag{3.25}$$

where y^* *is the unique solution of* $\mathcal{CO}$. *As a consequence, for all indices* $i, j \geq 1$ *such that* $A_i\mu > 1$, *we have*

$$\|y_j - x_0\| \leq \left(\frac{1+\left[A_j\mu\right]^{-1/2}}{1-\left[A_i\mu\right]^{-1/2}}\right)\|x_i - x_0\|. \tag{3.26}$$

Proof. First note our assumption on ψ_s combined with $\mathcal{CO}$ imply that $\psi \in \mathcal{F}_\mu(Z)$. Hence, it follows from Proposition 2.2.3(d) that

$$\frac{\mu}{2}\|y_j - y^*\|^2 \leq \psi(y_j) - \psi(y^*) \leq \frac{1}{2A_j}\|y^* - y_0\|^2$$

and hence that

$$\|y_j - y^*\| \leq \frac{1}{\sqrt{A_j\mu}}\|y^* - y_0\|. \tag{3.27}$$

The inequalities

$$\|y^* - x_0\| - \|y_j - y^*\| \leq \|y_j - y_0\| \leq \|y_j - y^*\| + \|y^* - y_0\|,$$

which are due to the triangle inequality, together with (3.27) clearly imply (3.25). The last statement of the lemma follows immediately from (3.25). □

As a consequence of Lemma 3.3.2, the following result obtains several important relations on certain quantities corresponding to two arbitrary iterates of the ACGM.

Lemma 3.3.3. *Let $\{(A_j, y_j, r_j, \eta_j)\}_{j\geq 1}$ and (ψ_s, ψ_n) be as in Lemma 3.3.1 with $\mu > 0$. Let i be an index such that $A_i \geq \max\{8, 9/\mu\}$. Then, for every $j \geq i$, we have*

$$\|y_j - y_0\| \leq 2\|y_i - y_0\|, \quad \|r_j\| \leq \frac{4}{A_j}\|y_i - y_0\|, \quad \eta_j \leq \frac{2}{A_j}\|y_i - y_0\|^2, \tag{3.28}$$

$$\|y_0 - y_j + r_j\| \leq \left(4 + \frac{8}{A_j}\right)\|y_0 - y_i + r_i\|, \qquad \eta_j \leq \frac{1}{A_j}8\|y_0 - y_i + r_i\|^2. \tag{3.29}$$

Proof. The first inequality in (3.28) follows from (3.26) and the assumption that $A_i\mu \geq 9$. Now, using Proposition 2.2.3(b) and the triangle inequality for norms, we easily see that

$$\|r_j\| \leq \frac{2}{A_j}\|y_j - y_0\|, \quad \eta_j \leq \frac{1}{2A_j}\|y_j - y_0\|^2$$

which, combined with the first inequality in (3.28), prove the second and the third inequalities in (3.28). Noting that $A_i \geq 8$ by assumption, Lemma 3.3.1 implies that (3.24) holds with $\sigma = 1$ and $j = i$, and hence that

$$\|r_i\| \leq \|y_0 - y_i + r_i\|. \tag{3.30}$$

Using the triangle inequality, the first two inequalities in (3.28) and relation (3.30), we conclude that

$$\begin{aligned}\|y_0 - y_j + r_j\| \leq \|y_0 - y_j\| + \|r_j\| &\leq \left(2 + \frac{4}{A_j}\right)\|y_0 - y_i\| \\ &\leq \left(2 + \frac{4}{A_j}\right)(\|y_0 - y_i + r_i\| + \|r_i\|) \leq \left(4 + \frac{8}{A_j}\right)\|y_0 - y_i + r_i\|,\end{aligned}$$

and that the first inequality in (3.29) holds. Now, the last inequality in (3.28), combined with the triangle inequality for norms and the relation $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ for every $a, b \in \mathcal{Z}$, imply that

$$\eta_j \leq \frac{2}{A_j}\|y_0 - y_i\|^2 \leq \frac{4}{A_j}\left(\|y_0 - y_i + r_i\|^2 + \|r_i\|^2\right).$$

Hence, in view of (3.30), the last inequality in (3.29) follows. □

3.3.2 Statement and Properties of the AIPPM

This subsection presents and analyzes the AIPPM for solving Problem 3.1.1. The main results of this subsection are Theorem 3.3.5 and Corollary 3.3.6 which give the iteration complexity of the AIPPM.

In order to state the method, we first state two ACG instances in Algorithm 3.3.1 that use terminations which are related to (3.5).

Algorithm 3.3.1: ACG Instances for the AIPPM

Require: $\sigma \geq 0$, $(\mu, L) \in \mathbb{R}^2_{++}$, $\psi_n \in \overline{\text{Conv}}\,(\mathcal{Z})$, $\psi_n \in \mathcal{F}_{\mu,L}(Z)$, $y_0 \in Z$;

procedure ACG1($\psi_s, \psi_n, y_0, \sigma, \mu, L$)
  for $k = 1, \ldots$ **do**
    $\lambda_k \leftarrow 1/L$
    Generate (A_k, y_k, r_k, η_k) according to Algorithm 2.2.2.
    if $\|r_k\|^2 + 2\eta_k \leq \sigma\|y_0 - y_k + r_k\|^2$ **and** $A_k \geq \max\{8, 9/\mu\}$ **then**
      return (y_k, r_k)

Require: $(\bar{\eta}, \sigma) \in \mathbb{R}^2_+$, $(\mu, L) \in \mathbb{R}^2_{++}$, $\psi_n \in \overline{\text{Conv}}\,(\mathcal{Z})$, $\psi_n \in \mathcal{F}_{\mu,L}(\text{dom}\,\psi_n)$, $y_0 \in Z$;

procedure ACG2($\psi_s, \psi_n, y_0, \sigma, \bar{\eta}, \mu, L$)
  for $k = 1, \ldots$ **do**
    $\lambda_k \leftarrow 1/L$
    Generate (y_k, r_k, η_k) according to Algorithm 2.2.2.
    if $\|r_k\|^2 + 2\eta_k \leq \sigma\|y_0 - y_k + r_k\|^2$ **and** $\eta_k \leq \bar{\eta}$ **then**
      return (y_k, r_k, η_k)

We now state the AIPPM in Algorithm 3.3.2, which uses the ACGM instances in Al-

gorithm 3.3.1 and the CRP in Algorithm 3.2.2. Given a starting point $z_0 \in Z$ and stepsize $\lambda \in (0, 1/m)$, its main idea is to repeatedly apply the ACGM at its k^{th} iteration to approximately solve the subproblem

$$\min_{z \in Z} \left\{ \lambda(f+h)(z) + \frac{1}{2}\|z - z_{k-1}\|^2 \right\}.$$

This process is iterated until the residuals $\|z_{k-1} - z_k + \tilde{v}_k\|/\lambda$ and $\tilde{\varepsilon}_k$, generated by the ACG call, are sufficiently small relative to $\hat{\rho}$. A call to the CRP is then made to generate a pair $(\hat{z}, \hat{v})$ that solves Problem 3.1.1.

Algorithm 3.3.2: AIPP Method

Require: $\hat{\rho} > 0$, $\sigma \in (0,1)$, $(m, M) \in \mathbb{R}^2_+$, $h \in \overline{\text{Conv}}\,(Z)$, $f \in \mathcal{C}_{m,M}(Z)$, $\lambda \in (0, 1/m)$, $z_0 \in Z$;
Initialize: $\mu \leftarrow 1 - \lambda m$, $L \leftarrow 1 + \lambda M$, $\bar{\rho} \leftarrow \hat{\rho}/4$, $\bar{\varepsilon} \leftarrow \hat{\rho}^2/(32[\max\{m, M\} + \lambda^{-1}])$;

1: **procedure** AIPP($f, h, z_0, \lambda, m, M, \sigma, \hat{\rho}$)
2: **for** $k = 1, \ldots$ **do**
3: PART 1 **Attack** the k^{th} prox subproblem.
4: $\psi_s^k \Leftarrow \lambda f + \|\cdot - z_{k-1}\|^2/2$
5: $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k) \leftarrow$ ACG1($\psi_s^k, \lambda h, z_{k-1}, \sigma, \mu, L$)
6: **if** $\|z_{k-1} - z_k + \tilde{v}_k\| \le \lambda\bar{\rho}/5$ **then**
7: PART 2 **Attack** the last prox subproblem.
8: $(z, \tilde{v}, \tilde{\varepsilon}) \leftarrow$ ACG2($\psi_s^k, \psi_n^k, z_{k-1}, \sigma, \lambda\bar{\varepsilon}, \mu, L$)
9: $(\hat{z}, \hat{q}, \hat{v}, \hat{\varepsilon}) \leftarrow$ CREF($f, h, z, \max\{m, M\}, \lambda$)
10: **return** $(\hat{z}, \hat{v})$

Some comments about the AIPPM are in order. To ease the discussion, let us refer to the ACG iterations performed in Line 5 and Line 8 of the method as **inner iterations** and the iterations over the indices k as **outer iterations**. First, in view of the last statement of Lemma 3.3.1 and the termination conditions given in Algorithm 3.3.1, each ACGM call al-

ways stops and outputs a triple $(z, \tilde{v}, \tilde{\varepsilon})$ satisfying

$$\tilde{v} \in \partial_{\tilde{\varepsilon}} \left(\lambda [f + h] + \frac{1}{2} \| \cdot - z_{k-1} \|^2 \right) (z), \quad \|\tilde{v}\|^2 + 2\tilde{\varepsilon} \leq \sigma \|z_{k-1} - z + \tilde{v}\|^2 \tag{3.31}$$

at the k^{th} outer iteration. Second, in view of the first comment, the outer iterations can be viewed as iterations of the GIPPF applied to $\mathcal{NCO}$. Finally, the goal of the ACGM call in Line 8 is to obtain a triple $(z, \tilde{v}, \tilde{\varepsilon})$ with a possibly smaller $\tilde{\varepsilon}$ while preserving the quality of the quantity $\|z_{k-1} - \tilde{z} + \tilde{v}\|/\lambda$, which at its start is bounded by $(\lambda\bar{\rho})/5$ and, throughout its inner iterations, can be shown to be bounded by $\lambda\bar{\rho}$ (see (3.36)).

The next proposition summarizes some basic facts about the AIPPM.

Lemma 3.3.4. *Let $(\bar{\rho}, \bar{\varepsilon})$ be as in the initialization phase of the AIPPM. Then, the following statements about the AIPPM hold:*

(a) at each outer iteration, its call to the ACGM in Line 5 stops and finds a triple $(z, \tilde{v}, \tilde{\varepsilon})$ satisfying (3.31) in at most

$$k_I := \left\lceil \max \left\{ \frac{2\sqrt{2}(1 + \sqrt{\sigma})}{\sqrt{\sigma}}, \frac{6}{\sqrt{1 - \lambda m}} \right\} \sqrt{1 + \lambda M} \right\rceil \tag{3.32}$$

inner iterations;

(b) its last call to the ACGM in Line 8 stops with an output triple $(z, \tilde{v}, \tilde{\varepsilon})$ satisfying

$$\tilde{v} \in \partial_{\tilde{\varepsilon}} \left(\lambda\phi + \frac{1}{2} \| \cdot - z_{k-1} \|^2 \right) (z), \quad \frac{1}{\lambda} \|z_{k-1} - z + \tilde{v}\| \leq \bar{\rho}, \quad \tilde{\varepsilon} \leq \lambda\bar{\varepsilon} \tag{3.33}$$

in at most

$$k_L := \left\lceil 2\sqrt{2\left(\frac{1 + \lambda M}{1 - \lambda m}\right)} \log_1^+ \left(\frac{2\bar{\rho}\sqrt{2(\lambda M + 1)\lambda}}{5\sqrt{\bar{\varepsilon}}} \right) + 1 \right\rceil \tag{3.34}$$

inner iterations, where $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$;

(c) it is a special implementation of the GIPPF in which $\lambda_k = \lambda$ for every $k \geq 1$;

(d) it stops with an output pair $(\hat{z}, \hat{v})$ that solves Problem 3.1.1 in at most

$$k_O := \left\lceil \frac{25 R_\lambda \phi(z_0)}{(1-\sigma)^2 \lambda^2 \bar{\rho}^2} + 1 \right\rceil \tag{3.35}$$

outer iterations, where $R_\lambda \phi(\cdot)$ is as defined in (3.10);

(e) for every $k \geq 1$, its sequence of iterates $\{z_k\}_{k \geq 1}$ and output point $\hat{z}$ satisfy $\phi(z_1) \geq \phi(z_k) \geq \phi(\hat{z})$.

Proof. All line numbers referenced in this proof are with respect to the AIPPM in Algorithm 3.3.2. Moreover, let (μ, L) be as in the initialization phase of the AIPPM.

(a) In view of assumptions (A1)–(A2), it holds that for every $k \geq 1$ we have $\psi_s^k \in \mathcal{F}_{\mu,L}(Z)$ and $\psi_n^k \in \overline{\mathrm{Conv}}\,(Z)$. Hence, it follows from the last statement of Lemma 3.3.1 and the definition of L that the ACGM obtains a triple $(z, \tilde{v}, \tilde{\varepsilon})$ satisfying (3.31) in at most

$$\left\lceil \left(\frac{2\sqrt{2}[1+\sqrt{\sigma}]}{\sqrt{\sigma}} \right) \sqrt{1+\lambda M} \right\rceil$$

inner iterations. On the other hand, in view of Proposition 2.2.3(c) with $\lambda_i = 1/L$ for every $i \geq 1$ and the definitions of μ and L, the condition $A_k \geq \max\{8, 9/\mu\}$ requires at most

$$\left\lceil \left(\frac{6}{\sqrt{1-\lambda m}} \right) \sqrt{1+\lambda M} \right\rceil$$

inner iterations. Combining the previous two inner iteration bounds yields the desired conclusion.

(b) Consider the triple $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ obtained in the last call to Line 5. In view of the termination criteria in this call, there exists an index $k \geq 1$ such that $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ is the j^{th} iterate of the ACGM started from $y_0 = z_{k-1}$ with $A_j \geq \max\{8, 9/\mu\}$, and hence, the index j satisfies the assumption of Lemma 3.3.3. It then follows from (3.29), Line 6, the first remark

following the AIPPM, and Proposition 2.2.3(c) with $\lambda_i = 1/L$ for every $i \geq 1$, that the call to the ACGM in Line 8 stops and outputs a triple $(z, \tilde{v}, \tilde{\varepsilon})$ satisfying the inclusion in (3.33), the bound

$$\frac{1}{\lambda}\|z_{k-1} - z + \tilde{v}\| \leq \left(4 + \frac{8}{A_j}\right)\frac{\bar{\rho}}{5} \leq \bar{\rho}, \tag{3.36}$$

and the bound

$$\tilde{\varepsilon} \leq \frac{8\lambda^2\bar{\rho}^2}{25A_j} \leq \frac{8L\lambda^2\bar{\rho}^2}{25}\left(1 + \sqrt{\frac{\mu}{2L}}\right)^{-2(j-1)}. \tag{3.37}$$

Using the stopping criterion for the ACGM instance in Line 8, the inequality for $\tilde{\varepsilon}$ above, the definitions of μ and L, and the relation that $\log(1+t) \geq t/2$ for all $t \in [0,1]$, we can easily see that $\tilde{\varepsilon} \leq \lambda\bar{\varepsilon}$ and (b) holds.

(c) This statement is obvious.

(d) The bound on the number of outer iterations follows by combining (c), the stopping criterion in Line 6, and Corollary 3.2.4(b) with $\bar{\rho}$ replaced by $\bar{\rho}/5$.

To show that the output pair $(\hat{z}, \hat{v})$ solves Problem 3.1.1, we first note that part (b) implies that the output $(z, \tilde{v}, \tilde{\varepsilon})$ of Line 8 satisfies (3.18) with $z^- = z_{k-1}$. It now follows from the call to the refinement procedure in Line 9, Proposition 3.2.5(b)–(c) with $(z_r, v_r, z^-) = (\hat{z}, \hat{v}, z_{k-1})$, and the definitions of $\bar{\rho}$ and $\bar{\varepsilon}$, that $\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z})$ and

$$\|\hat{v}\| \leq 2\left[\bar{\rho} + \sqrt{2\bar{\varepsilon}(\max\{m, M\} + \lambda^{-1})}\right] \leq 2\left[\frac{\hat{\rho}}{4} + \frac{\hat{\rho}}{4}\right] \leq \hat{\rho},$$

which is exactly (3.1).

(e) This follows from Line 9, Lemma 3.2.2, and Proposition 3.2.5(b) with $z_r = \hat{z}$. □

We now state one of our main results of this chapter, which is the iteration complexity of the AIPPM for solving Problem 3.1.1. Recall that the AIPPM assumes that $\lambda < 1/m$.

Theorem 3.3.5. *The AIPPM outputs a pair $(\hat{z}, \hat{v})$ that solves Problem 3.1.1 in*

$$\mathcal{O}\left(\sqrt{\frac{\lambda M + 1}{\min\{\sigma, 1 - \lambda m\}}}\left[\frac{R_\lambda \phi(z_0)}{(1-\sigma)^2\lambda^2\hat{\rho}^2} + \log_1^+(\lambda M)\right]\right) \tag{3.38}$$

inner iterations, where $R_\lambda\phi(\cdot)$ *is as in* (3.10) *and* $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

Proof. First, note that the total number of inner iterations in a call of the AIPPM is $k_T := k_I k_O + k_L$ — where k_I, k_O, and k_L are as in Lemma 3.3.4(a), (d), and (b), respectively. Using the fact that $\lambda < 1/m$, and hence $\log_1^+(\lambda \max\{m, M\}) = \mathcal{O}(\log_1^+(\lambda M))$, it is straightforward to verify that k_T is on the same order of magnitude as in (3.38). The fact that $(\hat{z}, \hat{v})$ solves Problem 3.1.1 follows from Lemma 3.3.4(d). □

Note that the AIPP version in which $\lambda = 1/(2m)$ and $\sigma = 1/2$ yields the best complexity bound under the reasonable assumption that, inside the squared bracket in (3.38), the first term is larger than the second one.

The following result describes the number of oracle calls performed by the AIPPM with $\lambda = 1/(2m)$ and $\sigma = 1/2$.

Corollary 3.3.6. *The AIPPM with inputs* $\lambda = 1/(2m)$ *and* $\sigma = 1/2$ *outputs a pair* $(\hat{z}, \hat{v})$ *that solves Problem 3.1.1 in*

$$\mathcal{O}\left(\sqrt{\frac{M}{m}+1}\left[\frac{m^2 R_{1/(2m)}\phi(z_0)}{\hat{\rho}^2} + \log_1^+\left(\frac{M}{m}\right)\right]\right) \tag{3.39}$$

oracle calls, where $R_\lambda\phi(\cdot)$ *is as in* (3.10) *and* $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

Proof. This is immediate from Theorem 3.3.5, the definition of $\log_1^+(\cdot)$, and the fact that the ACGM uses $\mathcal{O}(1)$ oracle calls per iteration. □

We now make a few remarks about the iteration complexity bound (3.39) and its relationship to two other ones obtained in the literature under assumption that: (i) $m \le M$; and (ii) the term $\mathcal{O}(1/\hat{\rho}^2)$ in (3.39) dominates the other one. First, using the definition of $R_\lambda\phi(z_0)$ and the first assumption, it is easy to see that the complexity bound (3.39) is majorized by

$$\mathcal{O}\left(\frac{\sqrt{mM}}{\hat{\rho}^2}\min\left\{\phi(z_0) - \phi_*, md_0^2\right\}\right) \tag{3.40}$$

where d_0 is as in (3.14). Second, since the iteration complexity bound for the CGM with $\lambda = 1/M$ is $\mathcal{O}(M[\phi(z_0) - \phi_*]/\hat{\rho}^2)$ (see the discussion following Proposition 3.2.6), we conclude that (3.40), and hence (3.39), is better than the CGM bound by a factor of $\sqrt{M/m}$. Third, bound (3.40), and hence (3.39), is also better than the one established in [30, Corollary 2] for an ACGM applied directly to $\mathcal{NCO}$ by at least a factor of $\sqrt{M/m}$. Note that the accelerated method of [30] assumes that the diameter of Z is bounded while the AIPPM does not.

3.3.3 Lower Complexity Bounds

Lower complexity bounds have recently been established in [116] for the complexity of finding solutions of Problem 3.1.1. The result below gives its precise statement.

Theorem 3.3.7. *Consider any algorithm $\mathcal{A}$ that solves Problem 3.1.1 under assumptions (A1)–(A3) and the assumption that $h \equiv 0$. For an initial point $z_0 \in Z$, if the iterates $\{z_k\}_{k\geq 1}$ generated by $\mathcal{A}$ satisfy*

$$z_k \in \operatorname{Lin}\{z_0, ..., z_{k-1}, \nabla f(z_0), ..., \nabla f(z_k)\} \quad \forall k \geq 1 \tag{3.41}$$

where $\operatorname{Lin} S$ *denotes the linear span of a set of elements S, then $\mathcal{A}$ requires*

$$\Omega\left(\frac{\sqrt{mM}\,[\phi(z_0) - \phi_*]}{\hat{\rho}^2}\right) \tag{3.42}$$

iterations to generate a solution of Problem 3.1.1.

We now make two remarks about the above result. First, since (3.42) is a lower complexity bound for the case of $h \equiv 0$ it is also a lower complexity bound for the case of $h \in \overline{\operatorname{Conv}}\,(Z)$. Second the linear-span requirement in (3.41) is more restrictive than the one considered in this chapter. Finally, in view of the remarks following Corollary 3.3.6, the AIPPM of this chapter achieves the lower complexity bound (3.42) up to a multiplicative constant.

3.4 Conclusion and Additional Comments

In this chapter, we presented an accelerated inexact proximal point method for obtaining approximate stationary points of an unconstrained NCO problem whose objective function is the sum of two functions $h \in \overline{\text{Conv}}\,(\mathcal{Z})$ and $f \in \mathcal{C}_{m,M}(\text{dom}\, h)$ for some $(m, M) \in \mathbb{R}^2_{++}$. The method consists of inexactly solving a sequence of proximal subproblems using an accelerated composite gradient method. We then established an $\mathcal{O}(\hat{\rho}^{-2})$ iteration complexity bound for finding $\hat{\rho}$-approximate stationary points which was observed to be complexity optimal in terms of m, M, and $\hat{\rho}$ for a large class of linear-span first-order methods.

The next chapter uses the developments in this one to develop methods for solving a class of set-constrained NCO problems.

Additional Comments

We now give a few additional comments about the results in this chapter.

First, the AIPPM improves on the complexity in [18] by a factor of $\log(M/\rho)$. Second, the AIPPM is a variant of the AIPP method in [46]. More specifically, the AIPPM of this chapter checks conditions (3.4) and (3.5) at every inner iteration while the AIPP method in [46] merely prescribes a fixed number of inner iterations per outer iteration.

Future Work

It would be worth investigating if the AIPPM also achieves the lower complexity bound for general first-order methods which do not necessary require condition (3.41). Currently, a lower bound [19, 20] is only known for case where $f \in \mathcal{C}_L(Z)$ for some $L > 0$. Additionally, it would be interesting to see if the behavior of the AIPPM, or a variant of its, under a stochastic oracle (as opposed to a deterministic one). Finally, it would be worth investigating the properties of a non-Euclidean AIPPM which is based on Bregman distances.

CHAPTER 4
FUNCTION CONSTRAINED COMPOSITE OPTIMIZATION

Our main goal in this chapter is to describe and establish the iteration complexity of two methods for finding approximate stationary points of the function constrained NCO (CNCO) problem

$$\min_{z\in\mathcal{Z}} \{\phi(z) := f(z) + h(z) : g(z) \in S\} \tag{$\mathcal{CNCO}$}$$

where $\mathcal{Z}$ is a finite dimensional inner product space, $h \in \overline{\text{Conv}}\ (Z)$ for some $Z \subseteq \mathcal{Z}$, $f \in \mathcal{C}_{m,M}(Z)$ for some $(m, M) \in \mathbb{R}^2_{++}$, $g \in \mathcal{C}(\mathcal{Z})$, and $S \subseteq \mathcal{R}$ is a closed convex set over some finite dimensional inner product space $\mathcal{R}$.

The first method is a **quadratic penalty** method for solved linearly set-constrained instances of $\mathcal{CNCO}$., i.e. g is linear, whereas the second method is an **inexact proximal augmented Lagrangian** method for solving nonlinearly cone-constrained instances of $\mathcal{CNCO}$, i.e. g is (possibly) nonlinear and S is a closed convex cone. Throughout our presentation, it is assumed that efficient oracles for evaluating the quantities $f(z)$, $\nabla f(z)$, $g(z)$, $\nabla g(z)$, and $h(z)$ and for obtaining exact solutions of the subproblems

$$\min_{z\in\mathcal{Z}} \left\{\lambda h(z) + \frac{1}{2}\|z - z_0\|^2\right\}, \quad \min_{r\in S} \|r - r_0\|$$

for any $z_0 \in \mathcal{Z}$, $r \in \mathcal{R}$, and $\lambda > 0$, are available. Moreover, we define an **oracle call** to be a collection of the above oracles of size $\mathcal{O}(1)$ where each of them appears at least once.

Given tolerance pair $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, it is shown that both methods obtain a solution pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfying

$$\begin{gathered} \hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + \nabla g(\hat{z})\hat{p}, \quad g(\hat{z}) + \hat{q} \in S \\ \|\hat{v}\| \le \hat{\rho}, \quad \|\hat{q}\| \le \hat{\eta}, \end{gathered} \tag{4.1}$$

in a number of oracle calls that depends on the tolerance pair $(\hat{\rho}, \hat{\eta})$. More specifically, the

quadratic penalty method obtains the above conditions in $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1})$ oracle calls, while the augmented Lagrangian method does this in $\mathcal{O}([\hat{\eta}^{-1/2}\hat{\rho}^{-2} + \hat{\rho}^{-3}]\log_1^+[\hat{\rho}^{-1} + \hat{\eta}^{-1}])$ oracle calls. It is worth mentioning that the no regularity conditions are needed for the quadratic penalty method and only a Slater-like condition is needed for the augmented Lagrangian method.

The content of this chapter is based on papers [46, 48] (joint work with Jefferson G. Melo and Renato D.C. Monteiro) and several passages may be taken verbatim from it.

Related Works

We first review methods that consider the case where g is linear. The complexity analysis of a first-order quadratic penalty method for the case where f is convex, h is an indicator function, was first given in [51] and further analyzed in [4, 72, 78]. Aside from [46], papers [47, 49, 62] are other works that establish the iteration complexity of quadratic penalty-based methods. For the case where $S = \{b\}$, paper [42] proposes a penalty ADMM approach which introduces an artificial variable y in $\mathcal{CNCO}$ and then penalizes y to obtain the penalized problem

$$\min\left\{f(z) + h(z) + \frac{c}{2}\|y\|^2 : Ax + y = b\right\}, \tag{4.2}$$

which is then solved by a two-block ADMM. It is then shown in [42, Remark 4.3] that the overall number of composite gradient steps performed by the aforementioned two-block ADMM penalty scheme for obtaining an approximate stationary point as in (4.1) is $\mathcal{O}(\hat{\rho}^{-6})$ when: $\hat{\eta} = \hat{\rho}$, the level sets of $f + h$ are bounded, and the initial triple (z_0, y_0, p_0) satisfies $(y_0, p_0) = (0, 0)$, $Az_0 = b$, and $z_0 \in \operatorname{dom} h$.

We now turn our attention to augmented Lagrangian (AL) methods that consider general (possibly nonlinear) functions g. Since AL-based methods for the convex case have been extensively studied in the literature (see, for example, [4, 5, 51, 52, 66, 78, 93, 111]), we focus on papers that deal with nonconvex problems. Moreover, we concentrate on those dealing with proximal augmented Lagrangian (PAL) based methods, i.e. the ones for which

the "inner" subproblems are of (or close to) the form in (4.26), and only those that establish iteration complexities. Paper [40] studies the iteration complexity of a linearized PAL method under the restrictive assumption that $h = 0$. Paper [35] introduces a perturbed θ-AL function, which agrees with the classical one (see (4.25)) when $\theta = 0$, and studies a corresponding unaccelerated PAL method whose iteration complexity is $\mathcal{O}(\hat{\eta}^{-4} + \hat{\rho}^{-4})$ under the strong condition that the initial starting point is feasible with respect to the constraint $g(z) \in S$. Paper [70] analyzes the iteration complexity of an inexact proximal accelerated PAL method based on the aforementioned perturbed AL function and shows, regardless of whether the initial point is feasible, that an approximate stationary point as in (4.1) is obtained in $\mathcal{O}(\hat{\eta}^{-1}\hat{\rho}^{-2}\log\hat{\eta}^{-1})$ ACG iterations and that the latter bound can be improved to $\mathcal{O}(\hat{\eta}^{-1/2}\hat{\rho}^{-2}\log\hat{\eta}^{-1})$ under an additional Slater-like assumption. Both papers [35, 70] assume that $\theta \in (0, 1]$, and hence, their analyses do not apply to the classical PAL method. In fact, as θ approaches zero, the universal constants that appear in the complexity bounds obtained in [35, 70] diverge to infinity. Using a different approach, i.e. one that does not rely on a merit function, paper [69] establishes the iteration complexity of an accelerated PAL method based on the classical augmented Lagrangian (see (4.25)) and Lagrange multiplier update (see (4.27)).

For the case where S is a closed convex cone $-\mathcal{K}$, each component of g is $\mathcal{K}$-convex, and $\mathcal{K} = \{0\} \times \mathbb{R}^k_+$, i.e. the constraint is of the form $g(x) = 0$ and/or $g(x) \leq 0$, papers [58, 101] present PAL methods that perform Lagrange multiplier updates only when the penalty parameter is updated. Hence, if the penalty parameter is never updated (which usually happens when the initial penalty parameter is chosen to be sufficiently large), then these methods never perform Lagrange multiplier updates, and thus they behave more like penalty methods. Paper [57] studies a hybrid penalty/augmented Lagrangian (AL) based method whose penalty iterations are the ones which guarantee its convergence and whose AL iterations are included with the purpose of improving its computational efficiency. For the case where g is not necessarily $\mathcal{K}$-convex and $\mathcal{K} = \{0\}$, i.e. the constraint is of the form $g(x) = 0$, paper [110]

analyzes the complexity of a PAL method under the strong assumption that: (i) $h = 0$; (ii) the smallest singular value of $\nabla g(x)$ is uniformly bounded away from zero everywhere; and, optionally, (iii) the initial starting point is feasible with respect to the constraint $g(z) \in S$.

Finally, we discuss other papers that have motivated the developments in [48] or are tangentially related to it. Paper [13] considers a primal-dual proximal point scheme and analyzes its iteration-complexity under strong conditions on the initial point. Papers [114, 115] present a primal-dual first-order algorithm for solving $\mathcal{CNCO}$ when $h = \delta_P$ and P is a box (in [115]) or more generally a polyhedron (in [114]). They also show that the primal-dual algorithm obtains an approximate stationary point as in (4.1) in $\mathcal{O}(\hat{\rho}^{-2})$ iterations when $\hat{\rho} = \hat{\eta}$.

Organization

This chapter contains two sections. The first one presents an accelerated quadratic penalty method for solving linear set-constrained instances of $\mathcal{CNCO}$. The second one presents an accelerated augmented Lagrangian method for solving nonlinearly cone-constrained instances of $\mathcal{CNCO}$. The last one gives a conclusion and some closing comments.

4.1 Composite Optimization with Linear Set Constraints

The quadratic penalty method is a popular optimization method for solving convex composite optimization problems with functional constraints $g(x) \leq 0$ where $g : \mathbb{R}^n \mapsto \mathbb{R}^m$ is convex in each of its entries. Denoting the function

$$\mathcal{L}_c(x; p) = \phi(x) + \frac{1}{2c}\left[\|\max\{0, p + cg(z)\}\|^2 - \|p\|^2\right] \tag{4.3}$$

as the augmented Lagrangian of the constrained problem $\min_{x\in\mathbb{R}^n}\{\phi(x) : g(x) \le 0\}$, the penalty method generates iterates $\{x_k\}_{k\ge1}$ according to the update

$$x_k = \operatorname*{argmin}_{x\in\mathbb{R}^n} \mathcal{L}_{c_k}(x; p_k), \tag{4.4}$$

for some sequence of penalty parameters $\{c_k\}_{k\ge1}$ and multipliers $\{p_k\}_{k\ge1}$. For the case where h is the indicator of a closed convex set, it is known (see, for example, [11, Proposition 4.2.1]) that if $0 < c_k < c_{k+1}$ for every $k \ge 1$ and $c_k \to \infty$ then every limit point of the sequence $\{x_k\}$ is a global minimum of the constrained problem. Moreover, under some additional regularity conditions, it can be shown (see, for example, [11, Section 4.2.1]) that the sequence $\{\max\{0, p_k + c_k g(x_k)\}\}_{k\ge1}$ converges to a Lagrange multiplier of the constrained problem.

Our main goal in this chapter is to describe and establish the iteration complexity of an accelerated **inexact** proximal quadratic penalty (AIP.QP) method for finding approximate stationary points of the linearly set-constrained NCO problem

$$\hat{\varphi}_* := \min_{z\in\mathcal{Z}} \{\phi(z) := f(z) + h(z) : \mathcal{A}z \in S\}, \tag{$\mathcal{CNCO}[a]$}$$

where $\mathcal{A} : \mathcal{Z} \mapsto \mathcal{R}$ is linear, the feasible set is nonempty, and the functions f and h are as described at the beginning of the chapter.

The AIP.QP method (AIP.QPM) is based on the smooth quadratic penalty function

$$f_c(z) := f(z) + \frac{c}{2}\operatorname{dist}^2(\mathcal{A}z, S) \quad \forall z \in Z, \quad \forall c > 0. \tag{4.5}$$

and it uses the AIPPM of Chapter 3 to generate its ℓ^{th} iterate: given c_ℓ, find an approximate stationary point $\hat{z}$ of the NCO problem

$$\hat{\varphi}_{c_\ell} := \min_{z\in\mathcal{Z}} \{\varphi_{c_\ell}(z) := f_{c_\ell}(z) + h(z)\}, \tag{4.6}$$

and check if it is approximately feasible, i.e. $\operatorname{dist}(\mathcal{A}\hat{z}, S) \approx 0$; if it is not, then multiplicatively increase c_ℓ by some factor and go the next iteration.

For a given tolerance pair $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$ and a suitable choice of λ, the main result of this chapter shows that the AIP.QPM, started from any point $z_0 \in Z$ obtains a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfying the approximate stationarity conditions

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + \mathcal{A}^* \hat{p} \quad \|\hat{v}\| \leq \hat{\rho} \tag{4.7}$$

$$\mathcal{A}\hat{z} + \hat{q} \in S \quad \|\hat{q}\| \leq \hat{\eta} \tag{4.8}$$

in at most

$$\mathcal{O}\left(\sqrt{\frac{\Theta_{\hat{\eta}}}{m} + 1}\left[\frac{m \cdot \min\{\hat{\varphi}_* - \hat{\varphi}_{\hat{c}}, m d_0^2\}}{\hat{\rho}^2} + \log_1^+\left(\frac{\Theta_{\hat{\eta}}}{m}\right)\right]\right)$$

oracle calls, where $d_0 = \min_{z \in \mathcal{Z}}\{\|z_0 - z_*\| : \phi(z_*) = \phi_*\}$, $\log_1^+(\cdot) = \max\{1, \log(\cdot)\}$, $\Theta_{\hat{\eta}} = \mathcal{O}(M + \|\mathcal{A}\|^2/\hat{\eta}^2)$, and $\hat{c}$ is a positive scalar for which $\hat{\varphi}_{\hat{c}}$ as in (4.6) with $c_\ell = \hat{c}$ is finite.

It is worth mentioning that this result neither assumes that Z is bounded nor that $\mathcal{CNCO}[a]$ has an optimal solution.

Organization

This section contains three subsections. The first one gives some preliminary references and discusses our notion of a stationary point given in (4.7) and (4.8). The second one presents some key properties of the penalty approach. The last one presents the AIP.QPM and its iteration complexity.

4.1.1 Preliminaries

This section enumerates the assumptions on problem $\mathcal{CNCO}[a]$, states the main problem of interest, and discusses the notion of an approximate stationary point given in (4.7) and (4.8).

It is assumed that $\phi = f + g$ satisfies assumptions (A1)–(A2) as well as the following assumptions:

(B1) $\mathcal{A} : \mathcal{Z} \mapsto \mathcal{R}$ is a nonzero linear operator, $S \subseteq \mathcal{R}$ is a closed convex set, and the feasible region $\mathcal{F} := \{z \in \mathcal{Z} : \mathcal{A}z \in S\}$ is nonempty;

(B2) there exists $\hat{c} \geq 0$ such that $\hat{\varphi}_{\hat{c}} > -\infty$, where

$$\hat{\varphi}_c := \inf_{z \in \mathcal{Z}} \{\varphi_c(z) := f_c(z) + h(z)\}, \quad \forall c \geq 0, \tag{4.9}$$

where $f_c(\cdot)$ is as in (4.5).

We now make three remarks about the above assumptions. First, the above assumptions imply that the optimal value of $\mathcal{CNCO}[a]$ is finite but not necessarily achieved. Second, assumption (B2) is quite natural in the sense that the penalty approach underlying the AIP.QPM would not make sense without it. Third, it is well-known that a necessary condition for $z^* \in Z$ to be a local minimum of $\mathcal{CNCO}[a]$ is that z^* be a stationary point of $f + h$, i.e. there exists $p^* \in \mathcal{R}$ such that $0 \in \nabla f(z^*) + \partial h(z^*) + \mathcal{A}^* p^*$ and $\mathcal{A}^* z \in S$.

In view of the above assumptions and remarks, we are interested in solving the problem given in Problem 4.1.1.

Problem 4.1.1: Find an approximate stationary point of $\mathcal{CNCO}$[a]

Given $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, find a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}]) \in [Z \times \mathcal{R}] \times [\mathcal{Z} \times \mathcal{R}]$ satisfying conditions (4.7) and (4.8).

4.1.2 Key Properties of the Quadratic Penalty Approach

We begin with some basic properties about the penalty function φ_c and some of its related quantities.

Lemma 4.1.1. *Let (f, h) be a pair of functions satisfying assumptions (A1)–(A2) and (B1)–(B2), $\mathcal{F}$ be as in assumption (B1), $(\hat{c}, \hat{\varphi}_c, \varphi_c)$ be as in assumption (B2), and the functions $R_\lambda\psi(\cdot)$ and $R_\lambda\psi(\cdot,\cdot)$ be as in* (3.10)*. Moreover, define*

$$R^{\mathcal{F}}_\lambda\psi(z_0) := \inf_{u\in\mathcal{F}} R_\lambda\psi(u; z_0), \tag{4.10}$$

for any function $\psi : \mathcal{Z} \mapsto (-\infty, \infty]$, scalar $\lambda \geq 0$, and point $z_0 \in \mathcal{Z}$. Then, the following statements hold for every scalar $c \geq \hat{c}$, scalars $\lambda, \hat{\lambda} \in \mathbb{R}_+$ satisfying $\lambda \geq \hat{\lambda}$, and point $z_0 \in \mathcal{Z}$:

(a) $\hat{\varphi}_c \geq \hat{\varphi}_{\hat{c}} > -\infty$ and $\varphi_c(u) = \varphi_{\hat{c}}(u)$ for every $u \in \mathcal{F}$;

(b) $R_\lambda\varphi_c(u; z_0) \leq R_{\hat{\lambda}}\varphi_{\hat{c}}(u; z_0)$ for every $u \in \mathcal{F}$, and hence, $R^{\mathcal{F}}_\lambda\varphi_c(z_0) \leq R^{\mathcal{F}}_{\hat{\lambda}}\varphi_{\hat{c}}(z_0)$;

(c) if z^* is an optimal solution of $\mathcal{CNCO}[a]$, then

$$R^{\mathcal{F}}_\lambda\varphi_c(z_0) \leq \frac{1}{2}\|z_0 - z^*\|^2 + \lambda\,[\hat{\varphi}_* - \hat{\varphi}_c]$$

where $\hat{\varphi}_*$ is as in $\mathcal{CNCO}[a]$.

Proof. (a) The fact that $\hat{\varphi}_{\hat{c}} > -\infty$ is from assumption (B2). The fact that $\varphi_c(u) = \varphi_{\hat{c}}(u)$ for every $u \in \mathcal{F}$ immediate from the definitions of φ_c and $\mathcal{F}$. The remaining inequality follows from the definition of φ_c and the assumption that $c \geq \hat{c}$.

(b) The first set of inequalities is immediate from part (a) and our assumption on $(\lambda, \hat{\lambda})$. The second one follows from the definition of $R^{\mathcal{F}}_\lambda\varphi_c(\cdot)$ in (4.10).

(c) This is immediate from the definition of $R^{\mathcal{F}}_\lambda\varphi_c(\cdot)$ in (4.10). □

Note that, similar to (3.11), it is straightforward to show that the function $R_{\lambda,\mathcal{F}}\psi(\cdot)$ in (4.10) satisfies

$$R^{\mathcal{F}}_\lambda\psi(z_0) = \lambda\left[e_\lambda(\psi + \delta_{\mathcal{F}})(z_0) - \inf_{u\in\mathcal{Z}}(\psi + \delta_{\mathcal{F}})(u)\right].$$

The next result shows how a solution of Problem 3.1.1 with $f = f_c$ is related to the conditions in Problem 4.1.1.

Lemma 4.1.2. *Given $\hat{\rho} > 0$ and $c > 0$, let $(\hat{z}, \hat{v})$ be a solution of Problem 3.1.1 with $f = f_c$ as in* (4.5)*. Moreover, define the quantities*

$$\hat{p} = c\left[\mathcal{A}\hat{z} - \Pi_S(\mathcal{A}\hat{z})\right], \quad \hat{q} = \Pi_S(\mathcal{A}\hat{z}) - \mathcal{A}\hat{z}.$$

Then the following statements hold:

(a) the pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfies (4.7) *and the inclusion in* (4.8)*;*

(b) it holds that

$$\|\hat{q}\|^2 \leq \frac{2\left[\varphi_c(\hat{z}) - \hat{\varphi}_{\hat{c}}\right]}{c - \hat{c}}.$$

Proof. (a) Using Lemma E.2.1(b) with $\mathcal{K} = S$ and the Chain Rule, it follows that

$$\nabla f_c(\hat{z}) = \nabla f(\hat{z}) + c\mathcal{A}^*\left[\mathcal{A}\hat{z} - \Pi_S(\mathcal{A}\hat{z})\right] = \nabla f(\hat{z}) + \mathcal{A}^*\hat{p},$$

and hence, by the definition of Problem 3.1.1 with f_c, it holds that $(\hat{z}, \hat{p}, \hat{v})$ satisfies (4.7). On the other hand, the inclusion (4.8) follows immediately from the definition of $\hat{q}$.

(b) Using the definition of $\hat{\varphi}_{\hat{c}}$, it holds that

$$\hat{\varphi}_{\hat{c}} + \left(\frac{c - \hat{c}}{2}\right) \cdot \operatorname{dist}^2(\mathcal{A}\hat{z}, S) \leq \varphi_{\hat{c}}(\hat{z}) + \left(\frac{c - \hat{c}}{2}\right) \cdot \operatorname{dist}^2(\mathcal{A}\hat{z}, S) = \varphi_c(\hat{z}).$$

Rearranging the above inequality and using the fact that $\|\hat{q}\| = \operatorname{dist}(A\hat{z}, S)$, it holds that

$$\|\hat{q}\|^2 = \operatorname{dist}^2(A\hat{z}, S) \leq \frac{2\left[\varphi_c(\hat{z}) - \hat{\varphi}_{\hat{c}}\right]}{c - \hat{c}}.$$

□

We now describe the behavior of a GIPP instance (see Chapter 3) applied to (4.9).

Lemma 4.1.3. *Let $\hat{q}$, $\hat{c}$, φ_c, and $R_\lambda^{\mathcal{F}}\varphi_c(\cdot)$ be as in Lemma 4.1.2, and suppose $\{(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)\}_{k\geq 1}$ is a sequence generated by an instance of the GIPPF (see Algorithm 3.2.1) for some $\{\lambda_k\}_{k\geq 1}$ and $z_0 \in Z$ with $\phi = \varphi_c$ for some $c > \hat{c}$. Moreover, let $\hat{\eta} \in \mathbb{R}_{++}$ be given and define*

$$T_{\hat{\eta}}(\lambda) := \hat{c} + \left[\frac{2 \cdot R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)}{\lambda(1-\sigma)}\right]\hat{\eta}^{-2} \quad \forall \lambda \in \mathbb{R}_{++}, \tag{4.11}$$

where $R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(\cdot)$ is as in (4.10)*. Then, for every $\hat{z} \in \mathcal{Z}$ such that $\varphi_c(\hat{z}) \leq \varphi_c(z_1)$, it holds that*

$$\|\hat{q}\|^2 \leq \frac{[T_{\hat{\eta}}(\lambda_1) - \hat{c}]\,\hat{\eta}^2}{c - \hat{c}}. \tag{4.12}$$

As a consequence, if $c \geq T_{\hat{\eta}}(\lambda_1)$ then $\|\hat{q}\| \leq \hat{\eta}$.

Proof. Let $\hat{z} \in \mathcal{Z}$ be such that $\varphi_c(\hat{z}) \leq \varphi_c(z_1)$. Using Lemma 3.2.2 with $k = 1$, the previous bound, and Lemma 4.1.1(a), it holds that

$$\begin{aligned}
\varphi_c(\hat{z}) - \hat{\varphi}_{\hat{c}} &\leq \varphi_c(z_1) - \hat{\varphi}_{\hat{c}} \\
&\leq \varphi_c(u) - \hat{\varphi}_{\hat{c}} + \frac{1}{2(1-\sigma)\lambda_1}\|u - z_0\|^2. \\
&= \varphi_{\hat{c}}(u) - \hat{\varphi}_{\hat{c}} + \frac{1}{2(1-\sigma)\lambda_1}\|u - z_0\|^2 \\
&\leq \frac{1}{\lambda_1(1-\sigma)}\left(\lambda_1\left[\varphi_{\hat{c}}(u) - \hat{\varphi}_{\hat{c}}\right] + \frac{1}{2}\|u - z_0\|^2\right) \quad \forall u \in \mathcal{F}.
\end{aligned}$$

Taking the infimum of the above bound over $u \in \mathcal{F}$ and using the definition of $R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)$, we conclude that

$$\varphi_c(\hat{z}) - \hat{\varphi}_{\hat{c}} \leq \frac{R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)}{\lambda_1(1-\sigma)}. \tag{4.13}$$

Using (4.13), Lemma 4.1.2(b), and the definition in (4.11) yields (4.12). The last conclusion follows immediately from (4.12) and the assumption that $c \geq T_{\hat{\eta}}(\lambda_1)$. □

We now make some remarks about the above result. First, it does not assume that $\mathcal{F}$, and hence Z, is bounded. Also, it does not even assume that $\mathcal{CNCO}[a]$ has an optimal solu-

tion. Second, it implies that all iterates (excluding the starting one) generated by an instance of the GIPPF applied to (4.6) satisfy the feasibility requirement, i.e. the last inequality in (4.8), as long as c_ℓ is sufficiently large, i.e. $c_\ell \geq T_{\hat{\eta}}(\lambda_1)$. Third, since the quantity $R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)$, which appears in the definition of $T_{\hat{\eta}}(\lambda_1)$ is difficult to estimate, a simple way of choosing a penalty parameter c_ℓ such that $c_\ell \geq T_{\hat{\eta}}(\lambda_1)$ is not apparent. This is why the AIP.QPM solves instead a sequence of penalized subproblems (4.6) for a strictly increasing sequence of penalty parameters $\{c_\ell\}_{\ell\geq 1}$. Moreover, despite solving a sequence of penalized subproblems, it is shown that its total number of oracle calls is the same as the one for the ideal method corresponding to solving (4.6) with $c_1 = T_{\hat{\eta}}(\lambda_1)$.

Recall from Lemma 3.3.4 and Theorem 3.3.5 in Chapter 3 that the AIPPM: (i) generates its iterates as an instance of the GIPPF; and (ii) outputs a pair $(\hat{z}, \hat{v})$ that solves Problem 3.1.1 with $\phi(\hat{z}) \leq \phi(z_1)$. In view of these facts, Lemma 4.1.2 and Lemma 4.1.3 show that the AIPPM is a suitable candidate for solving 4.1.1 when it is given $f = f_c$ for a sufficiently large enough $c > 0$. It only remains to show that the AIPPM can be applied to (4.9). Since assumption (A1) is that $h \in \overline{\mathrm{Conv}}\ Z$, we show that f_c satisfies the necessary smoothness requirements in the result below.

Lemma 4.1.4. *Suppose f satisfies assumption (A2) and let f_c be as in* (4.5)*. For any $c \geq 0$, it holds that $f_c \in \mathcal{C}_{m,M_c}(Z)$ where $M_c := M + c\|\mathcal{A}\|^2$.*

Proof. Let $Q(z) := \mathrm{dist}^2(\mathcal{A}z, S)/2$. Using Lemma E.2.1(a)–(b) with $\mathcal{K} = S$ and the Chain Rule, it holds that

$$\begin{aligned}\|\nabla Q(z) - \nabla Q(u)\| &= \|\mathcal{A}^*\left([\mathcal{A}z - \Pi_S(\mathcal{A}z)] - [\mathcal{A}u - \Pi_S(\mathcal{A}z)]\right)\| \\ &\leq \|\mathcal{A}\| \cdot \|[\mathcal{A}z - \Pi_S(\mathcal{A}z)] - [\mathcal{A}u - \Pi_S(\mathcal{A}u)]\| \\ &\leq \|\mathcal{A}\| \cdot \|\mathcal{A}z - \mathcal{A}u\| \leq \|\mathcal{A}\|^2\|z - u\|,\end{aligned}$$

and hence, $Q \in \mathcal{F}_{0,\|\mathcal{A}\|^2}(Z)$. The conclusion now follows from assumption (A2) and the fact that $f_c = f + cQ$. □

4.1.3 Statement and Properties of the AIP.QPM

This subsection describes and establishes the iteration complexity of the AIP.QPM.

We first state the AIP.QPM in Algorithm 4.1.1, which uses the AIPPM in Algorithm 3.3.2. Given $(\sigma, \lambda) \in (0,1) \times (0, 1/m)$ and $z_0 \in Z$, its main idea is to invoke the AIPPM to obtain approximate stationary points of sequence of penalty subproblems of the form

$$\min_{z \in \mathcal{Z}} \{f_{c_\ell}(z) + h(z)\}$$

where $\{c_\ell\}_{\ell \geq 1}$ is a strictly increasing sequence of penalty parameters that tend to infinity. At the end of each AIPPM call, a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ is generated that satisfies (5.32) and the inclusion in (4.7), and the method terminates when the inequality in (4.8) holds.

Algorithm 4.1.1: AIP.QP Method

Require: $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, $\sigma \in (0,1)$, $(m, M) \in \mathbb{R}^2_+$, $h \in \overline{\text{Conv}}\,(Z)$, $f \in \mathcal{C}_{m,M}(Z)$, $\lambda \in (0, 1/m)$, $z_0 \in Z$, $\mathcal{A} \neq 0$, $S \subseteq \mathcal{R}$, $\hat{c} > 0$ satisfying $(B2)$;
Initialize: $c_1 \leftarrow \hat{c} + (M + \lambda^{-1})/\|\mathcal{A}\|^2$;

procedure AIP.QP($f, h, \mathcal{A}, S, z_0, \hat{c}, \lambda, m, M, \sigma, \hat{\rho}, \hat{\eta}$)
  for $\ell = 1, \ldots$ **do**
    PART 1 **Attack** the ℓ^{th} prox penalty subproblem.
    $f_{c_\ell} \Longleftarrow f + \dfrac{c_\ell}{2} \cdot \text{dist}^2(\mathcal{A}(\cdot), S)$
    $M_{c_\ell} \leftarrow M + c_\ell \|\mathcal{A}\|^2$
    $(\hat{z}_\ell, \hat{v}_\ell) \leftarrow \text{AIPP}(f_{c_\ell}, h, z_0, \lambda, m, M_{c_\ell}, \sigma, \hat{\rho})$
    $\hat{p}_\ell \leftarrow c_\ell [\mathcal{A}\hat{z}_\ell - \Pi_S(\mathcal{A}\hat{z}_\ell)]$
    $\hat{q}_\ell \leftarrow \Pi_S(\mathcal{A}\hat{z}_\ell) - \mathcal{A}\hat{z}_\ell$
    PART 2 Either **stop** with a nearly feasible point or **increase** c_ℓ.
    if $\|\hat{q}_\ell\| \leq \hat{\eta}$ **then**
      return $([\hat{z}_\ell, \hat{p}_\ell], [\hat{v}_\ell, \hat{q}_\ell])$
    $c_{\ell+1} \leftarrow 2c_\ell$

Some comments about the AIP.QPM are in order. To ease the discussion, let us refer to the AIPP iterations in each AIPP call as **outer iterations**, the ACG iterations performed

inside each AIPP call as **inner iterations**, and the iterations over the indices ℓ as **cycles**. First, it follows from Lemma 3.3.4(d) that the pair $(\hat{z}, \hat{v}) = (\hat{z}_\ell, \hat{v}_\ell)$ solves Problem 3.1.1 with $f = f_{c_\ell}$. As a consequence, Lemma 4.1.2(a) implies that the output $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfies the (4.7) and the first inequality in (4.8). Second, since every loop of the AIP.QPM doubles c_ℓ, the condition $c_\ell > T_{\hat{\eta}}(\lambda_1)$ will be eventually satisfied. Hence, in view of the previous remark, the $\hat{q}_\ell$ corresponding to this c_ℓ will satisfy the feasibility condition $\|\hat{q}_\ell\| \leq \hat{\eta}$ and the AIP.QPM will stop in view of its stopping criterion in Line 10. Finally, in view of the previous remarks, we conclude that the AIP.QPM terminates with a triple $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfying (4.7) and (4.8).

The next result presents some basic properties of the AIP.QPM in consideration of the above remarks.

Lemma 4.1.5. *Let $T_{\hat{\eta}}(\cdot)$ be as in* (4.11)*. The following statements hold about the AIP.QPM:*

(a) at the ℓ^{th} cycle, its call to the AIPPM in Line 6 stops in

$$\mathcal{O}\left(\sqrt{\frac{\lambda\tilde{M}_\ell + 1}{\min\{\sigma, 1-\lambda m\}}}\left[\frac{R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)}{(1-\sigma)^2\lambda^2\hat{\rho}^2} + \log_1^+\left(\lambda\tilde{M}_\ell\right)\right]\right) \tag{4.14}$$

inner iterations, where $R_\lambda^{\mathcal{F}}\psi(\cdot)$ is as in (4.10)*, $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$, and*

$$\tilde{M}_i = M + 2^{i-1}c_1\|\mathcal{A}\|^2 \quad \forall i \geq 1. \tag{4.15}$$

(b) if ℓ_C is the first cycle where $c_\ell \geq T_{\hat{\eta}}(\lambda)$, then the AIP.QPM stops and outputs with a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ that solves Problem 4.1.1 in at most ℓ_C cycles.

Proof. All line numbers referenced in this proof are with respect to the AIP.QPM in Algorithm 4.1.1.

(a) Let $\ell \geq 1$ and M_{c_ℓ} be as in Line 5. Using the initialization of c_1 in the AIP.QPM, we first remark that

$$M_{c_\ell} = M + c_\ell\|\mathcal{A}\|^2 = M + 2^{\ell-1}c_1\|\mathcal{A}\|^2 = \mathcal{O}(\tilde{M}_\ell). \tag{4.16}$$

Moreover, by the definition of $R_\lambda^{\mathcal{F}}\psi(\cdot)$ and Lemma 4.1.1(b), it follows that $R_\lambda\varphi_{c_\ell}(z_0) \le R_\lambda^{\mathcal{F}}\varphi_{c_\ell}(z_0) \le R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)$. The conclusion result now follows from Lemma 4.1.4, (4.16), the previous bound, and Theorem 3.3.5 with $M = M_{c_\ell}$.

(b) This follows immediately from Lemma 4.1.2(a) and Lemma 4.1.3. □

We now state one of our main results of this section, which is the iteration complexity of the AIP.QPM for solving Problem 4.1.1. Recall that the AIP.QPM assumes that $\lambda < 1/m$.

Theorem 4.1.6. *Let $T_{\hat{\eta}}(\cdot)$ be as in* (4.11) *and define*

$$\Theta_{\hat{\eta}} := M + T_{\hat{\eta}}(\lambda)\|\mathcal{A}\|^2 \quad \forall (\hat{\eta}, \lambda) \in \mathbb{R}^2_{++}. \tag{4.17}$$

The AIP.QPM outputs a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ that solves Problem 4.1.1 in

$$\mathcal{O}\left(\sqrt{\frac{\lambda\Theta_{\hat{\eta}} + 1}{\min\{\sigma, 1 - \lambda m\}}}\left[\frac{R_\lambda^{\mathcal{F}}\varphi_{\hat{c}}(z_0)}{(1-\sigma)^2\lambda^2\hat{\rho}^2} + \log_1^+(\lambda\Theta_{\hat{\eta}})\right]\right) \tag{4.18}$$

inner iterations, where $R_\lambda^{\mathcal{F}}\psi(\cdot)$ is as in (4.10) *and* $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

Proof. The fact that the output of the AIP.QPM solves Problem 4.1.1 is an immediate consequence of Lemma 4.1.5(b).

Let us now prove the desired complexity bound. Let $\tilde{M}_i$ and ℓ_C be as in (4.15) and Lemma 4.1.5(b), respectively. In view of the AIPP call in Line 6 and Lemma 4.1.5(b), it follows that the number of inner iterations performed by the AIP.QPM is on the order given by the sum of the bound in (4.14) from $\ell = 1$ to ℓ_C. To show that this sum is exactly (4.18), we prove that

$$\sum_{i=1}^{\ell_C}(\lambda\tilde{M}_i + 1)^{1/2} = \mathcal{O}\left(\left[\lambda\Theta_{\hat{\eta}} + 1\right]^{1/2}\right), \quad \log_1^+\left(\lambda\tilde{M}_\ell\right) = \mathcal{O}\left(\log_1^+\left[\lambda\Theta_{\hat{\eta}}\right]\right) \quad \forall \ell \ge 1. \tag{4.19}$$

To begin, observe that the definition of c_1 implies that

$$M + \lambda^{-1} \leq c_1 \|\mathcal{A}\|^2 \leq 2^{i-1} c_1 \|\mathcal{A}\|^2 \quad \forall i \geq 1, \tag{4.20}$$

and the definitions of $\Theta_{\hat{\eta}}$, $T_{\hat{\eta}}(\cdot)$, and c_1 yield

$$\begin{aligned} \lambda \tilde{M}_1 + 1 =& \lambda \left(M + \lambda^{-1} + c_1 \|\mathcal{A}\|^2 \right) \leq 2\lambda c_1 \|\mathcal{A}\|^2 = 2\lambda \left(M + \lambda^{-1} + \hat{c} \|\mathcal{A}\|^2 \right). \\ &= \lambda \left[M + T_{\hat{\eta}}(\lambda) \|\mathcal{A}\|^2 \right] + 1 = \lambda \Theta_{\hat{\eta}} + 1. \end{aligned} \tag{4.21}$$

Using (4.21), it follows that the bounds in (4.19) hold for $\ell_C = 1$ or $\ell = 1$. Suppose now that $\ell_C > 1$. The definition of ℓ_C implies that $c_1 \cdot 2^{\ell_C - 1} \leq 2T_{\hat{\eta}}$, or equivalently, $2^{\ell_C/2} \leq 2 \left[T_{\hat{\eta}}(\lambda)/c_1 \right]^{1/2}$. Using the previous bound, (4.20), and the definition of $\Theta_{\hat{\eta}}$, it follows that

$$\begin{aligned} \sum_{i=1}^{\ell_C} (\lambda \tilde{M}_i + 1)^{1/2} &= \sum_{i=1}^{\ell_C} \left(\lambda \left[M + \lambda^{-1} + 2^{i-1} c_1 \|\mathcal{A}\|^2 \right] + 1 \right)^{1/2} \\ &\leq \sum_{i=1}^{\ell_C} \left(2\lambda \left[M + \lambda^{-1} + 2^{i-1} c_1 \|\mathcal{A}\|^2 \right] \right)^{1/2} \\ &= 2 \left(\lambda c_1 \|\mathcal{A}\|^2 \right)^{1/2} \sum_{i=1}^{\ell_C} 2^{(i-1)/2} = \mathcal{O}\left(\left[\lambda c_1 \|\mathcal{A}\|^2 \right]^{1/2} 2^{\ell_C/2} \right) \\ &= \mathcal{O}\left(\left[\lambda T_{\hat{\eta}}(\lambda) \|\mathcal{A}\|^2 \right]^{1/2} \right) = \mathcal{O}\left(\left[\lambda \Theta_{\hat{\eta}} + 1 \right]^{1/2} \right). \end{aligned} \tag{4.22}$$

Similarly, using the fact that $\{c_i\}_{i \geq 1}$ is monotone increasing, the previous bound on $2^{\ell_C/2}$, (4.20), and the definition of $\Theta_{\hat{\eta}}$, it holds that

$$\log\left(\lambda \tilde{M}_i \right) \leq \log\left(\lambda \tilde{M}_{\ell_C} \right) = \log\left(\lambda 2^{\ell_C} c_1 \|\mathcal{A}\|^2 \right) = \log\left(\lambda T_{\hat{\eta}}(\lambda) \|\mathcal{A}\|^2 \right) = \log\left(\lambda \Theta_{\hat{\eta}} \right). \tag{4.23}$$

Using (4.22) and (4.23), it follows that the bounds in (4.19) hold for $\ell_C \geq 2$ or $\ell \geq 2$. □

The following result describes the number of oracle calls performed by the AIP.QPM with $\lambda = 1/(2m)$ and $\sigma = 1/2$.

Corollary 4.1.7. *The AIP.QPM with inputs* $\lambda = 1/(2m)$ *and* $\sigma = 1/2$ *outputs a* $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ *that solves Problem 4.1.1 in*

$$\mathcal{O}\left(\sqrt{\frac{\Theta_{\hat{\eta}}}{m} + 1}\left[\frac{m^2 R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(z_0)}{\hat{\rho}^2} + \log_1^+\left(\frac{\Theta_{\hat{\eta}}}{m}\right)\right]\right)$$

oracle calls, where $R^{\mathcal{F}}_{\lambda}\psi(\cdot)$ *is as in* (4.10) *and* $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

Proof. This follows immediately from Theorem 4.1.6, the definition of $\log_1^+(\cdot)$, and the fact that every iteration of the ACGM performs $\mathcal{O}(1)$ oracle calls. □

4.2 Composite Optimization with Nonlinear Cone Constraints

The augmented Lagrangian method [38, 95] is an well-known extension of the quadratic penalty method (see Section 4.1) applied to the problem $\min_{x\in\mathbb{R}^n}\{\phi(x) : g(x) \leq 0\}$ in which a multiplier update is added to every iteration of the method. More specifically, recalling the Lagrangian $\mathcal{L}(\cdot;\cdot)$ in (4.3) and denoting

$$\ell_k(p; p_{k-1}) = \mathcal{L}_{c_k}(x_k; p_{k-1}) + \left\langle \nabla_p \mathcal{L}_{c_k}(x_k; p_{k-1}), p - p_{k-1}\right\rangle,$$

to be the linear approximation of the function $p \mapsto \mathcal{L}_{c_k}(x_k; p)$ at $p = p_{k-1}$, the multiplier update is given by

$$\begin{aligned} p_k &= \operatorname*{argmax}_{p\geq 0}\left\{c_k \ell_k(p; p_{k-1}) + \frac{1}{2}\|p - p_{k-1}\|^2\right\}, \\ &= \max\left\{0, p_{k-1} + c_k g(x_k)\right\}. \end{aligned} \tag{4.24}$$

For the case where $h \equiv 0$, it is known [11, Proposition 4.2.3] that if the generated sequence $\{p_k\}_{k\geq 1}$ is bounded, the penalty parameter c_k is sufficiently large enough after a certain index k, and some additional regularity conditions hold, then x_k and p_k converge to a global minimum and Lagrange multiplier of the constrained problem, respectively.

Our main goal in this section is to describe and establish the iteration complexity of an accelerated **inexact** proximal augmented Lagrangian (AIP.AL) method for finding approximate stationary points of the nonlinearly cone-constrained NCO problem

$$\varphi_* = \min_{z \in \mathcal{Z}} \{\phi(z) = f(z) + h(z) : g(z) \preceq_{\mathcal{K}} 0\} \tag{$\mathcal{CNCO}[b]$}$$

where $\mathcal{K}$ is a closed convex cone, the feasible set is nonempty, and the functions f, h, and g are as described in the beginning of the chapter. We will also assume that g is $\mathcal{K}$-convex function, i.e.

$$g(tu + [1-t]z) \preceq_{\mathcal{K}} tg(u) + [1-t]g(z) \quad \forall (t, u, z) \in [0,1] \times \mathcal{Z} \times \mathcal{Z},$$

with a Lipschitz continuous gradient, h is Lipschitz continuous on its domain $Z \subseteq \mathcal{Z}$, the set Z is convex compact, and that we have an oracle for computing the projection onto the dual cone of $\mathcal{K}$, which is denoted by $\mathcal{K}^+$ and included in the oracles that make up the oracle call mentioned at the beginning of this chapter. Here, the relation $g(z) \preceq_{\mathcal{K}} 0$ means that $g(z) \in -\mathcal{K}$.

The AIP.AL method (AIP.ALM) is based on the generalized (cf. [66] and [99, Section 11.K]) augmented Lagrangian function

$$\mathcal{L}_c(z; p) := f(z) + h(z) + \frac{1}{2c}\left[\operatorname{dist}^2(p + cg(z), -\mathcal{K}) - \|p\|^2\right], \tag{4.25}$$

and it uses an ACGM, e.g. Algorithm 2.2.2, to perform the following proximal point-type update to generate its k^{th} iterate: given (z_{k-1}, p_{k-1}) and (λ, c_k), compute

$$z_k \approx \operatorname*{argmin}_u \left\{\lambda \mathcal{L}_{c_k}(u; p_{k-1}) + \frac{1}{2}\|u - z_{k-1}\|^2\right\}, \tag{4.26}$$

$$p_k = \Pi_{\mathcal{K}^+}(p_{k-1} + c_k g(z_k)), \tag{4.27}$$

where $\mathcal{K}^+$ denotes the dual cone of $\mathcal{K}$ and the inexactness in the z_k update is according to some **relative** inexactness criterion. At the end of the k^{th} iteration above, it also performs a novel test to decide whether c_k is left unchanged or doubled.

Under a generalized Slater assumption[1] and a suitable choice of the inputs (λ, c), the main result of this section shows that for any $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, the AIP.ALM obtains a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + \nabla g(\hat{z})\hat{p}, \quad \langle g(\hat{z}) + \hat{q}, \hat{p} \rangle = 0, \quad g(\hat{z}) + \hat{q} \preceq_{\mathcal{K}} 0, \quad \hat{p} \succeq_{\mathcal{K}^+} 0 \tag{4.28}$$

$$\|\hat{v}\| \leq \hat{\rho}, \quad \|\hat{q}\| \leq \hat{\eta}, \tag{4.29}$$

in $\mathcal{O}([\hat{\eta}^{-1/2}\hat{\rho}^{-2} + \hat{\rho}^{-3}] \log_1^+[\hat{\rho}^{-1} + \hat{\eta}^{-1}])$ oracle calls, where $\log_1^+(\cdot) = \max\{1, \log(\cdot)\}$. Moreover, this complexity result is shown without requiring that the initial point z_0 be feasible with respect to the nonlinear constraint, i.e. $g(z_0) \preceq_{\mathcal{K}} 0$. A key fact about AIP.AL is that its generated sequence of Lagrange multipliers is always bounded, and this conclusion strongly uses the fact that its constraint function g is $\mathcal{K}$-convex.

Organization

This section contains four subsections. The first one gives some preliminary references and discusses our notion of a stationary point given in (4.28) and (4.29). The second one presents some key properties of the augmented Lagrangian approach. The third one presents the AIP.ALM and its iteration complexity. The last one gives the proof of the main result in this section.

4.2.1 Preliminaries

It is assumed that $\phi = f + h$ satisfies assumptions (A1)–(A2) with $m \leq M$, as well as the following assumptions:

[1] See Proposition 4.2.1.

(C1) h is also K_h-Lipschitz continuous for some $K_h > 0$, and Z is also compact with diameter $D_z := \sup_{u,z\in Z} \|u - z\|$;

(C2) $g : \mathcal{Z} \mapsto \mathbb{R}^\ell$ is continuously differentiable, $\mathcal{K}$-convex, and there exists $L_g > 0$ such that

$$\|\nabla g(u) - \nabla g(z)\| \le L_g \|u - z\| \quad \forall u, z \in \mathcal{Z};$$

(C3) there exists $\bar{z} \in \operatorname{int} Z$ and $\tau \in (0, 1]$ such that $g(\bar{z}) \preceq_{\mathcal{K}} 0$ and

$$\max\{\|\nabla g(z)p\|, |\langle g(\bar{z}), p\rangle|\} \ge \tau \|p\| \quad \forall z \in Z, \quad \forall p \succeq_{\mathcal{K}^+} 0; \tag{4.30}$$

We now give three remarks about the above assumptions. First, since Z is compact by (C1), the image of any continuous $\mathbb{R}^\ell$-valued function on Z is bounded. In view of this observation, we introduce the useful notation for any continuously differentiable function $\Psi : Z \mapsto \mathbb{R}^\ell$:

$$B_\Psi^{(0)} := \sup_{z\in\mathcal{H}} \|\Psi(z)\| < \infty, \quad B_\Psi^{(1)} := \sup_{z\in\mathcal{H}} \|\nabla\Psi(z)\| < \infty. \tag{4.31}$$

Second, it is well-known that if g is differentiable and $\mathcal{K}$-convex, then for every $z, u \in \mathcal{Z}$ it holds that

$$g'(z)(u - z) \preceq_{\mathcal{K}} g(u) - g(z).$$

Third, it is also well-known that a necessary condition for a point z^* to be a local minimum of $\mathcal{CNCO}[b]$ is that there exists a multiplier $p^* \in \mathbb{R}^\ell$ that satisfies the stationarity conditions

$$0 \in \nabla f(z^*) + \partial h(z^*) + \nabla g(z^*)p^*, \quad \langle g(z^*), p^*\rangle = 0, \quad g(z^*) \preceq_{\mathcal{K}} 0, \quad p^* \succeq_{\mathcal{K}^+} 0. \tag{4.32}$$

Moreover, the last three conditions in (4.32) (resp. (4.28)) are equivalent[2] to the inclusion $g(z^*) \in N_{\mathcal{K}^+}(p^*)$ (resp. the inequality $\operatorname{dist}(g(\hat{z}), N_{\mathcal{K}^+}(\hat{p})) \le \hat{\eta}$). In view of the above, (4.28) and (4.28) are clearly relaxations of (4.32). For the ease of future reference, let us formally

[2] See, for example, [99, Example 11.4] with $\bar{x} = g(z^*)$ and $\bar{v} = p^*$.

state the problem of finding a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfying (4.28) and (4.28) in Problem 4.2.1.

Problem 4.2.1: Find an approximate stationary point of $\mathcal{CNCO}[b]$

Given $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, find a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}]) \in [Z \times \mathbb{R}^\ell] \times [\mathcal{Z} \times \mathbb{R}^\ell]$ satisfying conditions (4.28) and (4.29).

It is also worth mentioning that the conditions in (C3) can be viewed as a generalization of a Slater-like assumption with respect to g, as shown in Proposition 4.2.1 below.

Proposition 4.2.1. *(Slater-like Assumption) Assume that the constraint $g(z) \preceq_{\mathcal{K}} 0$ is of the form*

$$g_\iota(z) \preceq_{\mathcal{J}} 0 \quad g_e(z) = 0 \tag{4.33}$$

where $\mathcal{J} \subseteq \mathbb{R}^s$ is a closed convex cone, $g_\iota : \mathbb{R}^n \to \mathbb{R}^s$ is continuously differentiable, and $g_e : \mathbb{R}^n \to \mathbb{R}^t$ is an onto affine map (and hence $g = (g_\iota, g_e)$ and $\mathcal{K} = \mathcal{J} \times \{0\}$). Assume also that there exists $\bar{z} \in \mathcal{H}$ such that $g_\iota(\bar{z}) \prec_{\mathcal{J}} 0$ and $g_e(\bar{z}) = 0$. Then, there exists $\tau > 0$ such that $(\bar{z}, \tau)$ satisfies (4.30)*. If, in addition, $\bar{z} \in \operatorname{int} Z$, then $(\bar{z}, \tau)$ satisfies (C3).*

Proof. Since g_e is affine and onto, its gradient matrix $G_e := \nabla g_e$ is independent of z and has full column rank.Hence, there exists $\tau_e > 0$ such that

$$\|G_e p_e\| \geq \tau_e \|p_e\|_1 \quad \forall p_e \in \mathbb{R}^s. \tag{4.34}$$

On the other hand, the assumption that $g_\iota(\bar{z}) \prec_{\mathcal{J}} 0$, and Lemma E.2.2 with $\mathcal{K} = \mathcal{J}$ and $x = -g_\iota(\bar{z}) \in \mathcal{J}$, imply that there exists $\tau_\iota > 0$ such that

$$-\langle p_\iota, g_\iota(\bar{z}) \rangle \geq \tau_\iota \|p_\iota\| \quad \forall p_\iota \in \mathcal{J}^+.$$

Using the previous inequality and the fact that $\|\nabla g_\iota(z)\|$ is bounded on $\mathcal{H}$, we conclude that

there exists $\gamma > 0$ such that

$$-\|\nabla g_\iota(z)p_\iota\| - 2\gamma\langle p_\iota, g_\iota(\bar{z})\rangle \geq [2\gamma\tau_\iota - \|\nabla g_\iota(z)\|] \cdot \|p_\iota\| \geq \tau_\iota\|p_\iota\|_1 \quad \forall z \in Z \tag{4.35}$$

Relations (4.34), (4.35), and the reverse triangle inequality, then imply that for every $z \in Z$,

$$\begin{aligned}
&\|\nabla g(z)p\| - 2\gamma\,\langle p, g(\bar{z})\rangle = \|\nabla g_\iota(z)p_\iota + G_e p_e\| - 2\gamma\,\langle p_\iota, g_\iota(\bar{z})\rangle \\
&\geq \|G_e p_e\| - \|\nabla g_\iota(z)p_\iota\| - 2\gamma\,\langle p_\iota, g_\iota(\bar{z})\rangle \geq \tau_e\|p_e\|_1 + \tau_\iota\|p_\iota\|_1 \\
&\geq \tau_c\|p\|_1 \geq \tau_c\|p\|,
\end{aligned}$$

where $\tau_c := \min\{\tau_e, \tau_\iota, 1\}$. It is now straightforward to see that the above inequality yields inequality (4.30) with $\tau = \tau_c/(1+2\gamma) \in (0, 1]$. The last part of the proposition now follows from the statement of assumption (C3) and the previous conclusion. □

Some additional comments about Proposition 4.2.1 are in order. First, the assumption that g_ι is $\mathcal{J}$-convex and g_e is affine implies that g is $\mathcal{K}$-convex. Second, the Slater condition is with regards to a single point $\bar{z} \in Z$, as opposed to condition (4.30) which involves inequality (4.30) at all pairs $(z, p) \in Z \times \mathcal{K}^+$. Third, (C3) can be replaced by the Slater-like assumption of Proposition 4.2.1 since the former is implied by the latter. Actually, a slightly more involved analysis can be done to show that the assumption that g_e is onto (which is part of the assumption of Proposition 4.2.1) can be removed at the expense of obtaining a weaker version of (C3), namely: inequality (4.30) holds for every pair $(z, p) \in Z \times (\mathcal{J}^+ \times \operatorname{Im} \nabla g_e)$, instead of $(z, p) \in Z \times (\mathcal{J}^+ \times \mathbb{R}^t) = Z \times \mathcal{K}^+$. Finally, since the analysis of this chapter can be easily adapted to this slightly weaker version of (C3), the Slater-like condition of Proposition 4.2.1 without g_e assumed to be onto (or equivalently, ∇g_e to have full column rank) can be used in place of (C3) in order to guarantee that all of the results derived in this chapter for the AIP.ALM hold.

4.2.2 Key Properties of the Augmented Lagrangian Approach

This subsection presents some technical results about the augmented Lagrangian approach.

The first result describes some properties about the smooth part of the Lagrangian in (4.25).

Lemma 4.2.2. *Define the function*

$$\widetilde{\mathcal{L}}_c(z;p) := f(z) + \frac{1}{2c}\left[\operatorname{dist}^2(p + cg(z), -\mathcal{K}) - \|p\|^2\right] \quad \forall (z,p,c) \in Z \times \mathbb{R}^\ell \times \mathbb{R}_{++}. \quad (4.36)$$

Then, for every $c > 0$ and $p \in \mathbb{R}^\ell$, the following properties hold:

(a) $\widetilde{\mathcal{L}}_c(\cdot;p)$ is convex, differentiable, and its gradient is given by

$$\nabla_z \widetilde{\mathcal{L}}_c(z;p) = \nabla f(z) + \nabla g(z)\Pi_{\mathcal{K}^+}(p + cg(z)) \quad \forall z \in \mathbb{R}^n;$$

(b) $\widetilde{\mathcal{L}}_c(\cdot;p) \in \mathcal{C}_{m,\widetilde{L}}(Z)$ where

$$\widetilde{L} = \widetilde{L}(c,p) := M + L_g\|p\| + c\left(B_g^{(0)}L_g + [B_g^{(1)}]^2\right), \quad (4.37)$$

and the quantities L_g and $(B_g^{(0)}, B_g^{(1)})$ are as in (C2) and (4.31)*, respectively.*

Proof. We first state that the case of $f \equiv 0$ and $M = 0$ has been previously shown in [66, Proposition 5] under the condition that $B_g^{(1)}$ is a Lipschitz constant of g. Hence, in view of assumption (C2) and the definition of $\mathcal{L}_c$, it suffices to verify the aforementioned condition. Indeed, using the Mean Value Inequality and the definition of $B_g^{(1)}$ in (4.31) we have that

$$\|g(z') - g(z)\| \le \sup_{t\in[0,1]} \|\nabla g(tz' + [1-t]z)\| \cdot \|z' - z\| \le B_g^{(1)}\|z' - z\| \quad \forall z', z \in \mathcal{H},$$

and hence that g is $B_g^{(1)}$-Lipschitz continuous. □

The next result, whose proof can be found in Appendix D, describes how the refinement procedure in Algorithm 3.2.2 yields a pair $([\hat{z},\hat{p}],[\hat{v},\hat{q}])$ that *nearly* solves Problem 4.2.1 when given inputs that satisfy conditions similar to (3.5) and (3.18).

Proposition 4.2.3. *Given $(c,\sigma) \in \mathbb{R}^2_{++}$, $(\lambda, z, p^-) \in \mathbb{R}_{++} \times Z \times \mathbb{R}^\ell$, and (f,h) satisfying assumptions (A1)–(A2), suppose there exists $\bar{\rho} \geq 0$ and $(z^-,\tilde{v},\tilde{\varepsilon}) \in Z \times \mathcal{Z} \times \mathbb{R}_+$ such that*

$$\begin{aligned} \tilde{v} &\in \partial_{\tilde{\varepsilon}} \left(\lambda \left[\widetilde{\mathcal{L}}_c(\cdot, p) + h \right] + \frac{1}{2} \| \cdot - z^- \|^2 \right)(z) \\ \|\tilde{v}\|^2 + 2\tilde{\varepsilon} &\leq \sigma \|z^- - z + \tilde{v}\|^2, \quad \frac{1}{\lambda} \|z^- - z + \tilde{v}\|^2 \leq \bar{\rho}, \end{aligned} \tag{4.38}$$

where $\widetilde{\mathcal{L}}_c(\cdot,\cdot)$ is as in (4.36)*. Moreover, using $\widetilde{L}(\cdot,\cdot)$ in* (4.37)*, define*

$$\begin{aligned} L^\psi &:= \lambda \widetilde{L}(c, p^-) + 1, \quad p := \Pi_{\mathcal{K}^+} \left(p^- + c g(z) \right), \\ \hat{p} &:= \Pi_{\mathcal{K}^+} \left(p^- + c g(\hat{z}) \right), \quad \hat{q} := \frac{1}{c}(p - \hat{p}), \end{aligned} \tag{4.39}$$

and using Algorithm 3.2.2, consider the assigned triple

$$(\hat{z}, \hat{w}, \hat{v}, \varepsilon) \leftarrow \mathit{CREF}(\widetilde{\mathcal{L}}_c(\cdot, p), h, z, L^\psi, \lambda).$$

Then, the following properties hold:

(a) the tuple $(w, \varepsilon, p, L^\psi)$ satisfies

$$\begin{aligned} \hat{w} &\in \nabla f(z) + \partial_\varepsilon h(z) + \nabla g(z) p, \\ \|\hat{w}\| &\leq \left(1 + \sqrt{\sigma L^\psi} \right) \bar{\rho}, \quad \varepsilon \leq \frac{\sigma}{2} \bar{\rho}^2; \end{aligned} \tag{4.40}$$

(b) the tuples $(\hat{z},\hat{p},\hat{v},\hat{q})$ and (p, L^ψ) satisfy (4.28) *and*

$$\|\hat{v}\| \leq 2 \left(1 + \sqrt{\sigma L^\psi} \right) \bar{\rho}, \quad \|\hat{q}\| \leq \frac{B_g^{(1)}}{L^\psi} \left(1 + \sqrt{\sigma L^\psi} \right) \bar{\rho} + \frac{1}{c} \|p - p^-\|, \tag{4.41}$$

where $B_g^{(1)}$ is given by (4.31)*.*

Proof. (a) The inclusion follows from Proposition 3.2.5(a) with $(z_r, q_r) = (\hat{z}, \hat{w})$ and $f = \widetilde{\mathcal{L}}_c(\cdot; p^-)$, Lemma 4.2.2(a) and the definition of p in (4.39). To show the bound on ε, observe that its definition and the inequalities in (4.38) yield

$$\varepsilon = \frac{1}{\lambda}\tilde{\varepsilon} \le \frac{\sigma}{2\lambda}\|z^- - z + \tilde{v}\|^2 \le \frac{\sigma}{2}\bar{\rho}^2.$$

To show that Proposition 3.2.5(c) with $(L_\lambda, q_r) = (L^\psi, \hat{w})$ and $\bar{\varepsilon} = \sigma\bar{\rho}^2/(2\lambda)$ imply that

$$\|\hat{w}\| \le \bar{\rho} + \sqrt{L^\psi \sigma \bar{\rho}^2} = \left(1 + \sqrt{L^\psi \sigma}\right)\bar{\rho}.$$

(b) The inclusion in (4.28) follows from Proposition 3.2.5(b) with $(z_r, v_r) = (\hat{z}, \hat{v})$ and $f = \widetilde{\mathcal{L}}_c(\cdot; p^-)$, Lemma 4.2.2(a), and the definition of $\hat{p}$ in (4.39). To show the remaining relations in (4.28), observe that Lemma E.2.1(b) with $u = p^- + cg(\hat{z})$ and the definitions of $\hat{q}$ and $\hat{p}$ in (4.39) imply that

$$g(\hat{z}) + \hat{q} = \frac{1}{c}\left[p^- + cg(\hat{z}) - \hat{p}\right] \in N_{\mathcal{K}^+}(\hat{p}).$$

Combining the above relations and Lemma E.2.1(c) with $u = g(\hat{z}) + \hat{q}$ and $p = \hat{p}$, we conclude that the remaining relations in (4.41) hold.

We now show the bounds in (4.41). The bound on $\|\hat{v}\|$ follows immediately from part (a) and Proposition 3.2.5(a) with $(z_r, v_r, q_r, f) = (\hat{z}, \hat{v}, \hat{w}, \widetilde{\mathcal{L}}_c(\cdot; p^-))$. To show the bound on $\hat{q}$, we first use the definitions of $\hat{p}$ and p in (4.39), the definition of $B_g^{(1)}$ given by (4.31), the Mean Value Inequality, and Lemma E.2.1(a) to obtain

$$\begin{aligned}\frac{1}{c}\|\hat{p} - p\| &= \frac{1}{c}\left\|\Pi_{\mathcal{K}^+}\left(p^- + cg(\hat{z})\right) - \Pi_{\mathcal{K}^+}\left(p^- + cg(z)\right)\right\| \le \frac{1}{c}\|cg(\hat{z}) - cg(z)\| \\ &\le \sup_{t\in[0,1]} \|\nabla g(t\hat{z} + [1-t]z)\| \cdot \|\hat{z} - z\| \le B_g^{(1)}\|\hat{z} - z\|. \end{aligned} \tag{4.42}$$

Now, since $w = L^\psi(\hat{z} - z)$ (see the definition of q_r in Algorithm 3.2.2), it follows from the

triangle inequality, the definition of $\hat{q}$ given in (4.39), part (a), and (4.42), that

$$\begin{aligned}\|\hat{q}\| = \frac{1}{c}\|\hat{p} - p^-\| &\leq \frac{1}{c}\|\hat{p} - p\| + \frac{1}{c}\|p - p^-\| \\ &\leq B_g^{(1)}\|\hat{z} - z\| + \frac{1}{c}\|p - p^-\| \leq \frac{B_g^{(1)}}{L^\psi}\|w\| + \frac{1}{c}\|p - p^-\| \\ &\leq \frac{B_g^{(1)}}{L^\psi}\left(1 + \sqrt{\sigma L^\psi}\right)\bar{\rho} + \frac{1}{c}\|p - p^-\|.\end{aligned}$$

□

Two comments about Proposition 4.2.3 are in order. First, the relations in (a) will be used to establish the boundedness of the sequence $\{p_k\}_{k\geq 1}$. Second, in view of (b), the quadruple $(\hat{z}, \hat{p}, \hat{w}, \hat{q})$ always satisfies (4.28). Hence, in order to solve Problem 4.2.1, it remains only to guarantee that condition (4.29) will eventually be satisfied. The inequalities in (4.41) will be essential to show the latter fact.

4.2.3 Statement and Properties of the AIP.ALM

This subsection describes and establishes the iteration complexity of the AIP.ALM.

Before presenting the method, we present an ACG subroutine in Line 7 that is used to approximate solve its key subproblems.

Algorithm 4.2.1: ACGM Instance for the AIP.ALM

Require: $\sigma \geq 0,\ \ (\mu, L) \in \mathbb{R}^2_{++},\ \ \psi_n \in \overline{\mathrm{Conv}}\,(\mathcal{Z}),\ \ \psi_n \in \mathcal{F}_{\mu,L}(Z),\ \ y_0 \in Z;$

1: **procedure** ACG3($\psi_s, \psi_n, y_0, \sigma, \mu, L$)
2: **for** $k = 1, \ldots$ **do**
3: $\lambda_k \leftarrow 1/L$
4: Generate (A_k, y_k, r_k, η_k) according to Algorithm 2.2.2.
5: **if** $\|r_k\|^2 + 2\eta_k \leq \sigma\|y_0 - y_k + r_k\|^2$ **then**
6: **return** (y_k, r_k, η_k)

We now state the AIP.ALM in Algorithm 4.2.2, which uses the ACG subroutine in Al-

gorithm 4.2.1, the refinement procedure in Algorithm 3.2.2, and the Lagrangian $\mathcal{L}_c(\cdot;\cdot)$ in (4.25). Given $(\lambda,\theta) \in \mathbb{R}_{++}\times(0,1/\sqrt{2}]$ and $z_0 \in Z$, its main idea is to invoke at its k^{th} iteration an ACG variant to obtain the inexact update

$$z_k \approx \operatorname*{argmin}_u \left\{ \lambda \mathcal{L}_{c_k}(u; p_{k-1}) + \frac{1}{2}\|u - z_{k-1}\|^2 \right\}.$$

More specifically, this ACG call obtains a triple $(z_k, v_k, \varepsilon_k)$ satisfying

$$\begin{gathered} v_k \in \partial_{\varepsilon_k}\left(\lambda\left[\widetilde{\mathcal{L}}_{c_k}(\cdot; p_{k-1}) + h\right] + \frac{1}{2}\|\cdot - z_{k-1}\|^2\right)(z_k), \\ \|v_k\|^2 + 2\varepsilon_k \le \frac{\theta^2}{L^{\psi}_{k-1}}\|v_k + z_{k-1} - z_k\|^2 \end{gathered} \tag{4.43}$$

where

$$L^{\psi}_{k-1} = \lambda \widetilde{L}(c_k, p_{k-1}) + 1, \tag{4.44}$$

and $\widetilde{L}(\cdot,\cdot)$ and $\widetilde{\mathcal{L}}_{c_k}(\cdot;\cdot)$ are as in (4.37) and (4.36), respectively. Using this triple $(z_k, v_k, \varepsilon_k)$, and the available data $(\lambda, z_{k-1}, p_{k-1}, \theta, L^{\psi}_{k-1})$, it then generates a refined point $(\hat{z}, \hat{p}, \hat{v}, \hat{q}) = (\hat{z}_k, \hat{p}_k, \hat{v}_k, \hat{q}_k)$ satisfying all the conditions in (4.28). If this quadruple also satisfies the bounds in (4.29), then the method stops and outputs $(\hat{z}, \hat{p}, \hat{v}, \hat{q})$. Otherwise, p_k is updated according to

$$p_k = \Pi_{\mathcal{K}^*}(p_{k-1} + c_k g(z_k)),$$

a novel test is invoked to check if c_k needs to be doubled, and the method continues to the $(k+1)^{\text{th}}$ iteration.

Algorithm 4.2.2: AIP.AL Method

Require: $(\hat{\rho},\hat{\eta}) \in \mathbb{R}^2_{++}$, $(m, M) \in \mathbb{R}^3_+$, $L_g > 0$, $h \in \overline{\text{Conv}}\,(Z)$, $f \in \mathcal{C}_{m,M}(Z)$, g satisfying (C2), $\lambda \in (0, 1/m)$, $\theta \in (0, 1/\sqrt{2}]$, $c_1 > 0$, $(z_0, p_0) \in Z \times \mathbb{R}^{\ell}$, $\mathcal{K} \subseteq \mathbb{R}^{\ell}$;

Initialize: $\mu \leftarrow 1 - \lambda m$, $\hat{k} \leftarrow 0$;

1: **procedure** AIP.AL($f, h, g, z_0, p_0, \lambda, m, M, L_g, \theta, \hat{\rho}, \hat{\eta}$)

2: **for** $k = 1, \ldots$ **do**

3: PART 1 **Attack** the k^{th} prox subproblem.

4: $\psi_s^k \Leftarrow \lambda \widetilde{\mathcal{L}}_{c_k}(\cdot; p_{k-1}) + \|\cdot - z_{k-1}\|^2/2$ ▷ See (4.36).

5: $L_{k-1}^{\psi} \leftarrow \lambda \widetilde{L}(c_k, p_{k-1}) + 1$ ▷ See (4.37).

6: $\sigma_{k-1} \leftarrow \theta^2 / L_{k-1}^{\psi}$

7: $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k) \leftarrow \text{ACG3}(\psi_s^k, \lambda h, z_{k-1}, \sigma_{k-1}, \mu, L_{k-1}^{\psi})$

8: PART 2 **Compute** and **check** the candidate output pair.

9: $(\hat{z}_k, \hat{w}_k, \hat{v}_k, \hat{\varepsilon}_k) \leftarrow \text{CREF}(\mathcal{L}^k, h, z_k, \max\{m, L_{k-1}^{\psi}\}, \lambda)$

10: $p_k \leftarrow \Pi_{\mathcal{K}^+}(p_{k-1} + c_k g(z_k))$

11: $\hat{p}_k \leftarrow \Pi_{\mathcal{K}^+}(p_k + c_k g(z_k))$

12: $\hat{q}_k \leftarrow (\hat{p}_k - p_{k-1})/c_k$

13: **if** $\|\hat{v}_k\| \le \hat{\rho}$ and $\|\hat{q}_k\| \le \hat{\eta}$ **then**

14: **return** $([\hat{z}_k, \hat{p}_k], [\hat{v}_k, \hat{q}_k])$

15: PART 3 **Check** if we need to increase c_k.

16: $\Delta_k \leftarrow [\mathcal{L}_{c_k}(z_{\hat{k}+1}; p_{\hat{k}+1}) - \mathcal{L}_{c_k}(z_k; p_k)] / (k - \hat{k} + 1)$

17: **if** $k > \hat{k} + 1$ and $\Delta_k \le \lambda(1-\theta)\hat{\rho}^2/36$ **then**

18: $c_{k+1} \leftarrow 2c_k$

19: $\hat{k} \leftarrow k$

20: **else**

21: $c_{k+1} \leftarrow c_k$

Some remarks about the AIP.ALM are in order. To ease the discussion, let us refer to the ACG iterations performed in Line 7 as **inner iterations** and the iterations over the indices k as **outer iterations**. First, its input z_0 can be any element in Z and does not necessarily need to be a point satisfying the constraint $g(z_0) \preceq_{\mathcal{K}} 0$. Second, its ACG call in Line 7 generates an output $(z_k, v_k, \varepsilon_k)$ that satisfies (4.43), which corresponds to the approximate update in (4.26). Finally, in view of Proposition 4.2.3(b) and the comments following it, AIP.AL stops if and only if the quadruple $(\hat{z}, \hat{p}, \hat{w}, \hat{q})$ solves Problem 4.2.1.

We now discuss the notion of a cycle. Define the l^{th} cycle $\mathcal{K}_l$ as the l^{th} set of consecutive indices k for which c_k remains constant, i.e.

$$\mathcal{C}_l := \left\{ k \ge 1 : c_k = \tilde{c}_l := 2^{l-1} c_1 \right\}. \tag{4.45}$$

For every $l \geq 1$, we let k_l denote the largest index in $\mathcal{C}_l$. Hence,

$$\mathcal{C}_l = \{k_{l-1} + 1, \ldots, k_l\} \quad \forall l \geq 1$$

where $k_0 := 0$. Clearly, the different values of $\hat{k}$ that arise in Line 19 are exactly the indices in the index set $\{k_l\}_{l\geq 1}$. Moreover, in view of the test performed in Line 17, we have that $k_l - k_{l-1} \geq 2$ for every $l \geq 1$, or equivalently, every cycle contains at least two indices. While generating the indices in the l^{th} cycle, if an index $k \geq k_{l-1} + 2$ satisfying the bound on Δ_k in Line 17 is found, k becomes the last index k_l in the l-th cycle and the $(l+1)^{\text{th}}$ cycle is started at iteration $k_l + 1$ with the penalty parameter set to $\tilde{c}_{l+1} = 2\tilde{c}_l$, where $\tilde{c}_l$ is as in (4.45).

The following result, whose proof is deferred to Section 4.2.4, describes the inner iteration complexity of AIP.AL. Its iteration complexity bound is expressed in terms of its inputs and the following auxiliary constants:

$$\bar{d} := \operatorname{dist}(\bar{z}, \partial Z), \quad \phi_* := \inf_{z\in\mathcal{Z}} \phi(z), \quad R_\phi := \hat{\varphi}_* - \phi_* + \frac{1}{\lambda} D_z^2, \tag{4.46}$$

$$\kappa_0 := 2\left[K_h + B_f^{(1)}\right] D_z + \left[\frac{\theta^2}{2(1-\theta)^2} + 2\left(\frac{1+\theta}{1-\theta}\right)\right] \frac{D_z^2}{\lambda}, \tag{4.47}$$

$$\kappa_1 := M + \frac{\kappa_0 L_g}{\min\{1, \bar{d}\}\tau}, \quad \kappa_2 := B_g^{(0)} L_g + [B_g^{(1)}]^2, \tag{4.48}$$

$$\kappa_3 := \frac{16(1+2\theta)^2 \max\{\|p_0\|, \kappa_0\}^2}{\lambda(1-\theta^2)\min\{1, \bar{d}^2\}\tau^2}, \quad \kappa_4 := \frac{\theta D_z}{\lambda(1-\sigma)B_g^{(1)}} + \frac{2\max\{\|p_0\|, \kappa_0\}^2}{\min\{1, \bar{d}\}\tau}, \tag{4.49}$$

where $\hat{\varphi}_*$, $(B_g^{(0)}, B_g^{(1)}, B_f^{(1)})$, (K_h, D_z), M, L_g, and $(\tau, \bar{z})$ are as in $\mathcal{CNCO}[b]$, (4.31), (C1), (A2), (C2), and (C3), respectively, and ∂Z denotes the boundary of the set Z.

Theorem 4.2.4. *Let the scalars $\{\kappa_i\}_{i=1}^4$ and R_ϕ be as in* (4.46), (4.48), *and* (4.49). *Moreover, define*

$$\bar{c}(\hat{\rho}, \hat{\eta}) := \frac{\kappa_3}{\hat{\rho}^2} + \frac{\kappa_4}{\hat{\eta}}, \quad \mathcal{T}_{\hat{\eta},\hat{\rho}} := [\lambda(\kappa_1 + c_1\kappa_2) + 1] \frac{\max\{c_1, 2\bar{c}(\hat{\rho}, \hat{\eta})\}}{c_1}. \tag{4.50}$$

Then, the AIP.ALM outputs a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ *that solves Problem 4.2.1 in*

$$\mathcal{O}\left(\frac{R_\phi}{\lambda\hat{\rho}^2}\sqrt{\frac{\mathcal{T}_{\hat{\eta},\hat{\rho}}}{1-\lambda m}}\log_1^+\left[\frac{\mathcal{T}_{\hat{\eta},\hat{\rho}}}{\theta}\right]\right) \tag{4.51}$$

inner iterations, where $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

The result below presents the iteration complexity of the AIP.ALM with inputs $\theta = 1/\sqrt{2}$ and $\lambda = 1/(2m)$.

Corollary 4.2.5. *Let* $T_{\hat{\eta},\hat{\rho}}$, $\hat{\varphi}_*$, ϕ_*, *and* D_z *be as in Theorem 4.2.4,* $\mathcal{CNCO}[b]$, (4.46), *and assumption (C1), respectively. Then, the AIP.ALM with inputs* $\lambda = 1/(2m)$ *and* $\theta = 1/\sqrt{2}$ *outputs a pair* $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ *that solves Problem 4.2.1 in*

$$\mathcal{O}\left(\left[\frac{m\left(\hat{\varphi}_* - \phi_*\right) + m^2 D_z^2}{\hat{\rho}^2}\right]\sqrt{\mathcal{T}_{\hat{\eta},\hat{\rho}}}\log_1^+\mathcal{T}_{\hat{\eta},\hat{\rho}}\right) \tag{4.52}$$

inner iterations, where $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$.

Proof. This follows immediately from Theorem 4.2.4 with $\lambda = 1/(2m)$ and $\theta = 1/\sqrt{2}$ and the definition of R_ϕ in (4.46). □

Note that the iteration complexity bound in (4.51) solely in terms of the tolerance pair $(\hat{\rho}, \hat{\eta})$ is

$$\mathcal{O}\left(\left[\frac{1}{\sqrt{\hat{\eta}}\hat{\rho}^2} + \frac{1}{\hat{\rho}^3}\right]\log_1^+\left[\frac{1}{\hat{\eta}} + \frac{1}{\hat{\rho}^2}\right]\right).$$

4.2.4 Proof of Theorem 4.2.4

This subsection presents several technical results that are needed to establish Theorem 4.2.4. It is divided into two subsections. The first one presents a bound on the sequence of multipliers $\{p_k\}_{k\geq 1}$ generated by the AIP.AL, while the second one is devoted to proving Theorem 4.2.4.

Bounding on the Sequence of Lagrangian Multipliers

The goal of this subsection is to show that the sequence of multipliers $\{p_k\}_{k\geq 1}$ generated by the AIP.ALM is bounded.

The first result presents a key inclusion and some basic bounds on several residuals generated by AIP.ALM.

Lemma 4.2.6. *Let $\{(\hat{w}_k, v_k, z_k, \varepsilon_k)\}_{k\geq 1}$ and $\{\sigma_{k-1}\}_{k\geq 1}$ be generated by the AIP.ALM, and consider D_z and τ as in assumptions (C1) and (C3), respectively. Moreover, define for every $k \geq 1$, the quantities*

$$\xi_k = \hat{w}_k - \nabla f(z_k) - \nabla g(z_k)p_k, \quad r_k = v_k + z_{k-1} - z_k. \tag{4.53}$$

Then, for every $k \geq 1$, it holds that

$$\xi_k \in \partial_{\delta_k} h(z_k), \quad \|r_k\| \leq \frac{D_z}{1-\theta}, \quad \varepsilon_k \leq \frac{1}{2}\left(\frac{\theta D_z}{1-\theta}\right)^2, \quad \|\hat{w}_k\| \leq \frac{(1+\theta)D_z}{\lambda(1-\theta)}. \tag{4.54}$$

Proof. Let $k \geq 1$ be fixed. Using Proposition 4.2.3(a) and the definition of ξ_k yields the required inclusion. On the other hand, the definitions of r_k and σ_{k-1}, the inequality in (4.43), and the fact that $z_k, z_{k-1} \in Z$ imply that

$$\|r_k\| = \|v_k + z_{k-1} - z_k\| \leq \|v_k\| + D_z \leq \sigma_{k-1}^{1/2}\|r_k\| + D_z \leq \theta\|r_k\| + D_z,$$

which, after a simple re-arrangement, yields the desired bound on $\|r_k\|$. Consequently, the definition of ε_k, the aforementioned bound on $\|r_k\|$, the fact that $L_{k-1}^{\psi} \geq 1$, and the inequality in (4.43) gives the bound on ε_k. Finally, the definitions of w_k and σ_{k-1}, the fact that $\theta \leq 1$, and Proposition 4.2.3(a) with $\bar{\rho} = \|r_k\|/\lambda$ and $(\sigma, L^{\psi}) = (\sigma_{k-1}, L_{k-1}^{\psi})$ yield

$$\|\hat{w}_k\| \leq \frac{1}{\lambda}\left(1 + \sqrt{\sigma_{k-1}L_{k-1}^{\psi}}\right)\|r_k\| = \frac{1+\theta}{\lambda}\|r_k\|,$$

which, combined with the previous bound on $\|r_k\|$, gives the desired bound on $\|\hat{w}_k\|$. □

The next result presents some important properties about the iterates generated by the AIP.ALM.

Lemma 4.2.7. *Let $\{(z_k, p_k, c_k)\}_{k\geq 1}$ be generated by the AIP.ALM and define, for every $k \geq 1$,*

$$s_k := \Pi_{-\mathcal{K}}(p_{k-1} + c_k g(z_k)). \tag{4.55}$$

Then, the following relations hold for every $k \geq 1$:

$$p_{k-1} + c_k g(z_k) = p_k + s_k, \quad \langle p_k, s_k \rangle = 0, \quad (p_k, s_k) \in \mathcal{K}^+ \times (-\mathcal{K}), \tag{4.56}$$

$$\mathcal{L}_{c_k}(z_k, p_{k-1}) = \phi(z_k) + \frac{1}{2c_k}\left(\|p_k\|^2 - \|p_{k-1}\|^2\right). \tag{4.57}$$

Proof. Let $\mathcal{K}^-$ denote the polar of $\mathcal{K}$. The two identities in (4.56) follow from the definitions of p_k and s_k in (4.27) and (4.55), respectively, the fact that $(\mathcal{K}^+)^- = -\mathcal{K}$, and [100, Exercise 2.8] with $\mathcal{K} = \mathcal{K}^-$ and $x = p_{k-1} + c_k g(z_k)$. On the other hand, using the definitions of $\mathcal{L}_c(\cdot;\cdot)$ and s_k in (4.25) and (4.55), respectively, it holds that

$$\mathcal{L}_{c_k}(z_k, p_{k-1}) = \phi(z_k) + \frac{1}{2c_k}\left[\|p_{k-1} + c_k g(z_k) - s_k\|^2 - \|p_{k-1}\|^2\right]$$

which, in view of the first identity in (4.56), immediately implies (4.57). □

The following technical result, whose proof can be found in [69, Lemma 4.7], plays an important role in the proof of Proposition 4.2.9 below.

Lemma 4.2.8. *Let h be a function as in (C1). Then, for every $u, z \in Z$, $\delta > 0$, and $\xi \in \partial_\delta h(z)$, we have*

$$\|\xi\|\mathrm{dist}(u, \partial Z) \leq [\mathrm{dist}(u, \partial Z) + \|z - u\|]\, K_h + \langle \xi, z - u \rangle + \delta,$$

where ∂Z denotes the boundary of the set Z.

We are now ready to prove the main result of this subsection, namely, that the sequence $\{p_k\}_{k\geq 1}$ is bounded.

Proposition 4.2.9. *Consider the sequence $\{(p_k, c_k)\}_{k\geq 1}$ generated by the AIP.ALM and let κ_0, τ, and $\bar{d}$ be as in* (4.47)*, (C3), and* (4.46)*, respectively. Then, the following statements hold:*

(a) for every $k \geq 1$, we have

$$\min\{1, \bar{d}\}\tau\|p_k\| + \frac{\|p_k\|^2}{c_k} \leq \kappa_0 + \frac{1}{c_k}\langle p_k, p_{k-1}\rangle;$$

(b) for every $k \geq 0$, we have

$$\|p_k\| \leq C_0 := \frac{\max\{\|p_0\|, \kappa_0\}}{\min\{1, \bar{d}\}\tau}. \tag{4.58}$$

Proof. (a) Let $k \geq 1$ be fixed and $\bar{z}$ be as in (C3). Moreover, let $(\xi_k, \varepsilon_k, z_k)$ be as in Lemma 4.2.6. It follows from Lemma 4.2.6 that $\xi_k \in \partial_{\varepsilon_k} h(z_k)$ for every $k \geq 1$. Hence, assumption (C1), the fact that $\bar{d} \leq D_z$ and $z_k \in Z$, Lemma 4.2.8 with $\xi = \xi_k$, $z = z_k$, $u = \bar{z}$ and $\delta = \varepsilon_k$, and the bound on $\|\varepsilon_k\|$ in Lemma 4.2.6, imply that

$$\bar{d}\|\xi_k\| \leq 2D_z K_h + \frac{\theta^2 D_z^2}{2\lambda(1-\theta)^2} + \langle \xi_k, z_k - \bar{z}\rangle. \tag{4.59}$$

On the other hand, using the assumption that g is $\mathcal{K}$-convex (see (C2)), the fact that $p_k \in \mathcal{K}^+$, the definition of ξ_k in (4.53), the bound on $\|\hat{w}_k\|$ in Lemma 4.2.6, and the Cauchy-Schwarz inequality, we conclude that

$$\begin{aligned}
\langle \xi_k, z_k - \bar{z}\rangle &= \langle \hat{w}_k - \nabla f(z_k) - \nabla g(z_k)p_k, z_k - \bar{z}\rangle \\
&= \langle \hat{w}_k - \nabla f(z_k), z_k - \bar{z}\rangle + \langle p_k, g'(z_k)(\bar{z} - z_k)\rangle \\
&\leq \langle \hat{w}_k - \nabla f(z_k), z_k - \bar{z}\rangle + \langle p_k, g(\bar{z}) - g(z_k)\rangle \\
&\leq B_f^{(1)} D_h + \frac{(1+\theta)D_h^2}{\lambda(1-\theta)} + \langle p_k, g(\bar{z}) - g(z_k)\rangle
\end{aligned} \tag{4.60}$$

where $B_f^{(1)}$ is as in (4.31). Now, defining

$$\kappa := \left[2K_h + B_f^{(1)}\right] D_h + \left[\frac{\theta^2}{2(1-\theta)^2} + \frac{1+\theta}{1-\theta}\right] \frac{D_z^2}{\lambda}, \tag{4.61}$$

and using (4.59), (4.60), together with the relations in (4.56), we conclude that

$$\begin{aligned}\bar{d}\|\xi_k\| - \langle p_k, g(\bar{z})\rangle \leq \kappa - \langle p_k, g(z_k)\rangle &= \kappa - \frac{1}{c_k}\langle p_k, s_k + p_k - p_{k-1}\rangle \\ &= \kappa - \frac{\|p_k\|^2}{c_k} + \frac{1}{c_k}\langle p_k, p_{k-1}\rangle\end{aligned}$$

where s_k is as in (4.55). Noting that the definition of ξ_k and the reverse triangle inequality yield

$$\|\xi_k\| = \|\nabla f(z_k) - \hat{w}_k + \nabla g(z_k)p_k\| \geq -\|\nabla f(z_k) - \hat{w}_k\| + \|\nabla g(z_k)p_k\|,$$

it follows that

$$\bar{d}\|\nabla g(z_k)p_k\| - \langle p_k, g(\bar{z})\rangle \leq \kappa - \frac{\|p_k\|^2}{c_k} + \frac{1}{c_k}\langle p_k, p_{k-1}\rangle + \bar{d}\|\nabla f(z_k) - \hat{w}_k\|. \tag{4.62}$$

Using now the triangle inequality, assumption (C3), (4.61), (4.62), the fact that $\bar{d} \leq D_z$, and the definition of κ_0 in (4.47), we finally conclude that

$$\min\{1, \bar{d}\}\tau\|p_k\| + \frac{\|p_k\|^2}{c_k} \leq \kappa + B_f^{(1)} D_h + \frac{(1+\sigma)D_h^2}{\lambda(1-\sigma)} + \frac{1}{c_k}\langle p_k, p_{k-1}\rangle = \kappa_0 + \frac{1}{c_k}\langle p_k, p_{k-1}\rangle.$$

(b) This statement is proved by induction. Since $\tau \leq 1$, inequality (4.58) trivially holds for $k = 0$. Assume that (4.58) holds with $k = i - 1$ for some $i \geq 1$. This assumption, together with the bound obtained in the latter result and the Cauchy-Schwarz inequality, then imply

that

$$\left(\min\{1,\bar{d}\}\tau+\frac{\|p_i\|}{c_i}\right)\|p_i\|\leq\kappa_0+\frac{\|p_i\|\cdot\|p_{i-1}\|}{c_i}\leq\kappa_0+\frac{\|p_i\|C_0}{c_i}$$
$$\leq\left(\min\{1,\bar{d}\}\tau+\frac{\|p_i\|}{c_i}\right)C_0,$$

which implies that $\|p_i\|\leq C_0$. Then, (4.58) also holds with $k=i$ and hence, by induction, we conclude that (4.58) holds for the whole sequence $\{p_k\}_{k\geq1}$. □

Proving Theorem 4.2.4

The main goal of this sub-subsection is to present the proof of Theorem 4.2.4.

The proof of Theorem 4.2.4 requires several technical results. The first one characterizes the change in the augmented Lagrangian between consecutive iterations of the AIP.ALM.

Lemma 4.2.10. *The sequence $\{(z_k,p_k)\}_{k\geq1}$ generated by AIP.AL satisfies the relations*

$$\mathcal{L}_{c_k}(z_k;p_k)\leq\mathcal{L}_{c_k}(z_k;p_{k-1})+\frac{1}{c_k}\|p_k-p_{k-1}\|^2, \tag{4.63}$$
$$\mathcal{L}_{c_k}(z_k;p_k)\leq\mathcal{L}_{c_k}(z_{k-1};p_{k-1})-\left(\frac{1-\theta^2}{2\lambda}\right)\|r_k\|^2+\frac{1}{c_k}\|p_k-p_{k-1}\|^2, \tag{4.64}$$

for every $k\geq1$, where r_k is as in (D.1).

Proof. Let s_k be as in (4.55). Using (4.57), the definition of $\mathcal{L}_c(\cdot;\cdot)$ in (4.25), the fact that

$s_k \in -\mathcal{K}$ and $p_{k-1} + c_k g(z_k) = p_k + s_k$ in view of (4.56), we have that

$$\begin{aligned}
&\mathcal{L}_{c_k}(z_k, p_k) - \mathcal{L}_{c_k}(z_k, p_{k-1}) \\
&= \mathcal{L}_{c_k}(z_k, p_k) - \phi(z_k) - \frac{1}{2c_k}\left(\|p_k\|^2 - \|p_{k-1}\|^2\right) \\
&= \frac{1}{2c_k}\left(\operatorname{dist}^2(p_k + c_k g(z_k), -\mathcal{K}) - \|p_k\|^2\right) - \frac{1}{2c_k}\left(\|p_k\|^2 - \|p_{k-1}\|^2\right) \\
&\le \frac{1}{2c_k}\left(\|p_k + c_k g(z_k) - s_k\|^2 - \|p_k\|^2\right) - \frac{1}{2c_k}\left(\|p_k\|^2 - \|p_{k-1}\|^2\right) \\
&= \frac{1}{2c_k}\left(\|2p_k - p_{k-1}\|^2 - 2\|p_k\|^2 + \|p_{k-1}\|^2\right),
\end{aligned}$$

which immediately implies (4.63). Now, in view of the definition of the approximate subdifferential and the fact that $(z_k, v_k, \varepsilon_k)$ satisfies both the inclusion and the inequality in (4.43), we conclude that

$$\begin{aligned}
&\lambda\mathcal{L}_{c_k}(z_k, p_{k-1}) - \lambda\mathcal{L}_{c_k}(z_{k-1}, p_{k-1}) \le -\frac{1}{2}\|z_k - z_{k-1}\|^2 + \langle v_k, z_k - z_{k-1}\rangle + \varepsilon_k \\
&= -\frac{1}{2}\|v_k + z_k - z_{k-1}\|^2 + \frac{1}{2}\|v_k\|^2 + \varepsilon_k \le -\left(\frac{1-\sigma_{k-1}}{2}\right)\|r_k\|^2 \le -\left(\frac{1-\theta^2}{2}\right)\|r_k\|^2, \qquad (4.65)
\end{aligned}$$

where the last inequality follows from the fact that $\sigma_{k-1} \le \theta$. Inequality (4.64) now follows by combining (4.63) with (4.65). □

Recall that the l^{th} cycle $\mathcal{C}_l$ and the penalty constants $\{\tilde{c}_l\}_{l\ge1}$ are defined in (4.45). The next results present some properties of the iterates generated during an AIP.AL cycle. The first one below establishes an upper bound on the augmented Lagrangian function along the iterates within an AIP.AL cycle.

Lemma 4.2.11. *Consider the sequences $\{(z_k, p_k)\}_{k\in\mathcal{C}_l}$ and $\{\tilde{c}_l\}_{l\ge1}$ generated during the l^{th} cycle of the AIP.ALM. Then, for every $k \in \mathcal{C}_l$, we have*

$$\mathcal{L}_{\tilde{c}_l}(z_k; p_k) \le R_\phi + \phi_* + \frac{4C_0^2}{\tilde{c}_l}, \qquad (4.66)$$

where (ϕ_*, R_ϕ)*,* $\tilde{c}_l$*, and* C_0 *are as in* (4.46)*,* (4.45)*, and* (4.58)*, respectively.*

Proof. First note that for any $k \in \mathcal{C}_l$, we have $c_k = \tilde{c}_l = 2^{l-1}c_1$. Moreover, $(\lambda, z_k, v_k, \varepsilon_k, \theta)$ satisfies the inclusion and the inequality in (4.43). Hence, it follows from Lemma E.1.1 with $s = 1$, $\tilde{\sigma} = \sigma_{k-1}$ and $\tilde{\phi} = \lambda\mathcal{L}_{\tilde{c}_l}(\cdot, p_{k-1})$, and assumption (C1) that for every $z \in Z$, we have

$$\begin{aligned}\lambda\mathcal{L}_{\tilde{c}_l}(z_k, p_{k-1}) + \frac{1 - 2\sigma_{k-1}^2}{2}\|r_k\|^2 &\le \lambda\mathcal{L}_{\tilde{c}_l}(z, p_{k-1}) + \|z - z_{k-1}\|^2 \\ &\le \lambda\mathcal{L}_{\tilde{c}_l}(z, p_{k-1}) + D_z^2 \end{aligned} \tag{4.67}$$

where r_{k_0} is as in (4.53) with $k = k_0$. Now, observe that the definitions of σ_{k-1} and L_{k-1}^{ψ} imply that $\sigma_{k-1} \le \theta \in (0, 1/\sqrt{2}]$ and that the definition of $\mathcal{L}_c$ in (4.25) implies that $\mathcal{L}_{\tilde{c}_l}(z, p_{k-1}) \le \phi(z)$ for every $z \in \mathcal{F} := \{z \in Z : g(z) \preceq_{\mathcal{K}} 0\}$. Using then the definition of $\hat{\varphi}_*$ given in $\mathcal{CNCO}[b]$, the aforementioned observations, and the minimization of the right-hand-side of (4.67) with respect to $z \in \mathcal{F}$, we get

$$\mathcal{L}_{\tilde{c}_l}(z_k, p_{k-1}) \le \hat{\varphi}_* + \frac{D_h^2}{\lambda} = R_\phi + \phi_*$$

where the last equality is due to the definition of R_ϕ in (4.46). Combining the above inequality, (4.63) and the bound $(a + b)^2 \le 2a^2 + 2b^2$ for every $a, b \in \mathbb{R}$, we have

$$\begin{aligned}\mathcal{L}_{\tilde{c}_l}(z_k, p_k) &\le \mathcal{L}_{\tilde{c}_l}(z_k, p_{k-1}) + \frac{1}{\tilde{c}_l}\|p_k - p_{k-1}\|^2 \\ &\le \mathcal{L}_{\tilde{c}_l}(z_k, p_{k-1}) + \frac{2}{\tilde{c}_l}\left(\|p_k\|^2 + \|p_{k-1}\|^2\right) \\ &\le R_\phi + \phi_* + \frac{4C_0^2}{\tilde{c}_l},\end{aligned}$$

and hence the conclusion of the lemma follows. □

The next result presents some bounds on the sequences $\{\|r_k\|\}_{k\in\mathcal{C}_l}$ and $\{\Delta_k\}_{k\in\mathcal{C}_l}$.

Lemma 4.2.12. *Let* $\{(z_k, v_k, \varepsilon_k, \Delta_k)\}_{k\in\mathcal{C}_l}$ *and* $\{\tilde{c}_l\}_{l\ge1}$ *be generated during the* l^{th} *cycle of the AIP.ALM and consider* $\{r_k\}_{k\in\mathcal{C}_l}$ *as in* (4.53)*. Then, for every* $k \in \mathcal{C}_l$ *such that* $k \ge k_{l-1} + 2$*, we*

have

$$\min_{k_{l-1}+2\le j\le k} \|r_j\|^2 \le \frac{2\lambda}{1-\theta^2}\left(\Delta_k + \frac{4C_0^2}{\tilde{c}_l}\right), \tag{4.68}$$

$$\Delta_k \le \frac{1}{k-k_{l-1}-1}\left(R_\phi + \frac{9C_0^2}{2\tilde{c}_l}\right), \tag{4.69}$$

where C_0 *is as in* (4.58).

Proof. Relations (4.58), (4.64), the fact that $c_k = \tilde{c}_l$ for every $k \in \mathcal{C}_l$, and the inequality $\|p_k - p_{k-1}\|^2 \le 2\|p_k\|^2 + 2\|p_{k-1}\|^2$, imply that for any $k \in \mathcal{C}_l$ such that $k \ge k_{l-1} + 2$ the following inequalities hold:

$$\begin{aligned}
&\frac{(1-\theta^2)(k-k_{l-1}-1)}{2\lambda} \min_{k_{l-1}+2\le j\le k} \|r_j\|^2 \le \frac{(1-\theta^2)}{2\lambda} \sum_{j=k_{l-1}+2}^{k} \|r_j\|^2 \\
&\le \mathcal{L}_{\tilde{c}_l}(z_{k_{l-1}+1}; p_{k_{l-1}+1}) - \mathcal{L}_{\tilde{c}_l}(z_k; p_k) + \frac{1}{\tilde{c}_l} \sum_{j=k_{l-1}+2}^{k} \|p_j - p_{j-1}\|^2 \\
&\le \mathcal{L}_{\tilde{c}_l}(z_{k_{l-1}+1}; p_{k_{l-1}+1}) - \mathcal{L}_{\tilde{c}_l}(z_k; p_k) + \frac{4(k-k_{l-1}-1)C_0^2}{\tilde{c}_l},
\end{aligned}$$

and hence that (4.68) holds, in view of the definition of Δ_k. Now, in view of the definitions of $\mathcal{L}_c$ and ϕ_* given in (4.25) and (4.46), respectively, we have

$$\mathcal{L}_{\tilde{c}_l}(z_k; p_k) = \phi(z_k) + \frac{1}{2\tilde{c}_l}\left[\operatorname{dist}^2(p_k + \tilde{c}_l g(z_k), -\mathcal{K}) - \|p_k\|^2\right] \ge \phi_* - \frac{\|p_k\|^2}{2\tilde{c}_l}.$$

It follows from the above inequality, (4.66) with $k = k_{l-1} + 1$, and the definition of Δ_k that

$$\Delta_k \le \frac{1}{k-k_{l-1}-1}\left(R_\phi + \phi_* + \frac{4C_0^2}{\tilde{c}_l} + \frac{\|p_k\|^2}{2\tilde{c}_l} - \phi_*\right),$$

which proves (4.69) in view of (4.58). □

The next technical lemma presents some additional properties of the refined iterates generated by the AIP.ALM.

Lemma 4.2.13. *Consider the sequences* $\{(c_k, z_k, p_k, v_k, \varepsilon_k)\}_{k\in\mathcal{C}_l}$, $\{(\sigma_{k-1}, L_{k-1}^{\psi})\}_{k\in C_l}$, $\{\tilde{c}_l\}_{l\ge 1}$,

and $\{(\hat{z}_k, \hat{p}_k, \hat{v}_k, \hat{q}_k)\}_{k\in\mathcal{C}_l}$ *generated during the* l^{th} *cycle of the AIP.ALM. Then, the following statements hold:*

(a) for every $k \in \mathcal{C}_l$*, the quadruple* $(\hat{z}, \hat{p}, \hat{v}, \hat{q}) = (\hat{z}_k, \hat{p}_k, \hat{v}_k, \hat{q}_k)$ *satisfies* (4.28) *and* (4.41) *with*

$$(c, p^-, \sigma, L^\psi), = (c_k, p_{k-1}, \sigma_{k-1}, L^\psi_{k-1}), \quad \bar{\rho} = \frac{1}{\lambda}\|z_{k-1} - z_k + v_k\|;$$

(b) for every $k \in \mathcal{C}_l$ *and* $k \geq k_{l-1} + 2$*, there exists an index* $i \in \{k_{l-1} + 2, \ldots, k\}$ *such that*

$$\|\hat{v}_i\|^2 \leq \frac{2(1+2\sigma)^2 R_\phi}{\lambda(1-\sigma^2)(k - k_{l-1} - 1)} + \frac{\kappa_3}{2\tilde{c}_l}, \quad \|\hat{q}_i\| \leq \frac{\kappa_4}{\tilde{c}_l}, \tag{4.70}$$

where R_ϕ *and* (κ_3, κ_4) *are as in* (4.46) *and* (4.49)*, respectively.*

Proof. (a) In view of the ACG call in Line 7 of the method, we have that (λ, θ), L^ψ_{k-1}, (z_{k-1}, p_{k-1}), and $(z_k, v_k, \varepsilon_k)$ satisfy (4.43). The conclusion now follows from Proposition 4.2.3(b)–(c).

(b) Let $k \in \mathcal{C}_l$ such that $k \geq k_{l-1} + 2$. In view of Lemma 4.2.12, there exists an index $i \in \{k_{l-1} + 2, \ldots, k\}$ such that

$$\|r_i\|^2 \leq \frac{2\lambda}{1-\sigma^2}\left[\frac{R_\phi}{k - k_{l-1} - 1} + \frac{4C_0^2}{\tilde{c}_l}\right], \tag{4.71}$$

where C_0 is as in (4.58). The bound on $\|\hat{w}_i\|^2$ now follows from combining (4.71), the first inequality in (4.41), the definitions of κ_3 and C_0 in (4.49) and (4.58), and the fact that $\sigma_{k-1}L^\psi_{k-1} = \theta^2$.

Now, recall that for any $k \in \mathcal{C}$ it holds that $c_k = \tilde{c}_l$. Hence, in view of the second inequality in (4.41), (4.58), the triangle inequality for norms, and the facts that $\sigma_{k-1}L^\psi_{k-1} = \theta$ and $L^\psi_{k-1} \geq$

$\lambda\tilde{c}_l[B_g^{(1)}]^2$ (see their definitions in the AIP.ALM), we have

$$\begin{aligned}\|\hat{q}_i\| &\leq \frac{B_g^{(1)}\sigma_{k-1}}{\sqrt{L_{k-1}^{\psi}}}\|r_i\| + \frac{1}{\tilde{c}_l}\left(\|p_i\| + \|p_{i-1}\|\right) \leq \frac{B_g^{(1)}\theta}{L_{k-1}^{\psi}}\|r_i\| + \frac{2C_0}{\tilde{c}_l} \\ &\leq \frac{\theta D_h}{\lambda(1-\theta)B_g^{(1)}\tilde{c}_l} + \frac{2C_0}{\tilde{c}_l} = \frac{\kappa_4}{\tilde{c}_l}, \end{aligned} \tag{4.72}$$

where the last relation is due to the definitions of κ_4 and C_0 in (4.49) and (4.58), respectively.

□

The next result establishes some bounds on the number of inner and outer iterations performed during an AIP.AL cycle. It also shows that if the penalty parameter is sufficiently large, then AIP.AL generates a solution of Problem 4.2.1.

Lemma 4.2.14. *Let R_ϕ, (κ_1, κ_2), and $\bar{c}(\hat{\rho}, \hat{\eta})$ be as in* (4.46), (4.48), *and* (4.50), *respectively. Then, the following statements hold about the AIP.ALM:*

(a) at every outer iteration k within the l^{th} cycle, it performs at most

$$\left\lceil 1 + 4\sqrt{\frac{2\left[1+\lambda(\kappa_1+\tilde{c}_l\kappa_2)\right]}{1-\lambda m}}\ \log_1^+\left(\frac{4\left[1+\lambda(\kappa_1+\tilde{c}_l\kappa_2)\right]}{\theta}\right)\right\rceil$$

inner iterations, where $\log_1^+(\cdot) := \max\{\log(\cdot), 1\}$;

(b) every cycle performs $\mathcal{O}(R_\phi/[\lambda\hat{\rho}^2])$ outer iterations;

(c) if $\tilde{c}_l \geq \bar{c}(\hat{\rho}, \hat{\eta})$ then the AIP.ALM must stop in the l^{th} cycle with a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ that solves Problem 4.2.1.

Proof. (a) Note that within the l^{th} cycle, $c_k = \tilde{c}_l$. Hence, in view of (4.58) and the definitions

of L_{k-1}^{ψ}, (κ_1, κ_2), and $\bar{L}_c^{\psi}$, we have

$$\begin{aligned} L_{k-1}^{\psi} &= \lambda\left[L_f + L_g\|p_{k-1}\| + c_k\left(B_g^{(0)}L_g + [B_g^{(1)}]^2\right)\right] + 1 \\ &\le \lambda\left[L_f + L_g C_0 + \tilde{c}_l\left(B_g^{(0)}L_g + [B_g^{(1)}]^2\right)\right] + 1 \\ &= \lambda(\kappa_1 + \kappa_2\tilde{c}_l) + 1. \end{aligned} \tag{4.73}$$

Using the fact that the AIP.ALM invokes Algorithm 4.2.1 in Line 7 with $(L, \mu) = (L_{k-1}^{\psi}, 1 - \lambda m)$, (4.73), the fact that $\sigma_{k-1} = \theta^2/L_{k-1}^{\psi} \le 1$, and Lemma 3.3.1, it holds that the number of inner iterations performed within this cycle is at most

$$\begin{aligned} &\left\lceil 1 + \sqrt{\frac{2L_{k-1}^{\psi}}{1-\lambda m}}\log_1^+\left(\frac{2L_{k-1}^{\psi}\left[1+\sqrt{\sigma_{k-1}}\right]^2}{\sigma_{k-1}}\right)\right\rceil \\ &\le \left\lceil 1 + \sqrt{\frac{2L_{k-1}^{\psi}}{1-\lambda m}}\ \log^+\left(16\left[\frac{L_{k-1}^{\psi}}{\theta}\right]^2\right)\right\rceil \le \left\lceil 1 + 2\sqrt{\frac{2L_{k-1}^{\psi}}{1-\lambda m}}\log^+\left(\frac{4L_{k-1}^{\psi}}{\theta}\right)\right\rceil \\ &\le \left\lceil 1 + 4\sqrt{\frac{2\left[1+\lambda(\kappa_1+\tilde{c}_l\kappa_2)\right]}{1-\lambda m}}\ \log_1^+\left(\frac{4\left[1+\lambda(\kappa_1+\tilde{c}_l\kappa_2)\right]}{\theta}\right)\right\rceil. \end{aligned}$$

(b) Fix a cycle $l \ge 1$ and let C_0 be as in (4.58). It follows from (4.69) that, for every $k \in \mathcal{C}_l$, we have $k \ge k_{l-1} + 2$, and

$$\Delta_k \le \frac{1}{k - k_{l-1} - 1}\left(R_\phi + \frac{9C_0^2}{2\tilde{c}_l}\right).$$

Hence, since $\tilde{c}_l \ge c_1$, it is easy to see that if k satisfies

$$k > k_{l-1} + 1 + \frac{4(1+2\theta)^2}{\lambda(1-\theta^2)\hat{\rho}^2}\left(R_\phi + \frac{9C_0^2}{2c_1}\right)$$

then the condition on Δ_k in Line 17 of the method holds, ending the l^{th} cycle. Since the cycle starts at $k_{l-1} + 1$, statement (b) follows immediately from the above bound.

(c) From the definition of $\bar{c}(\cdot,\cdot)$ and the fact that $\tilde{c}_l \geq \bar{c}(\cdot,\cdot)$, we have

$$\tilde{c}_l \geq \frac{\kappa_3}{\hat{\rho}^2}, \quad \tilde{c}_l \geq \frac{\kappa_4}{\hat{\eta}}, \tag{4.74}$$

where κ_3 and κ_4 are as in (4.49). Now, let $\bar{k} \geq k_{l-1} + 2$ be the smallest index such that

$$\frac{2(1+2\sigma)^2 R_\phi}{\lambda(1-\theta^2)(\bar{k} - k_{l-1} - 1)} \leq \frac{\hat{\rho}^2}{2}. \tag{4.75}$$

Hence, in view of (4.74), (4.75), and Lemma 4.2.13(b), there exists an index $i \in \{k_{l-1} + 2, \ldots, \bar{k}\}$ such that

$$\|\hat{v}_i\| \leq \hat{\rho}, \quad \|\hat{q}_i\| \leq \hat{\eta}$$

which implies that the AIP.ALM must stop at iteration i, in view its Line 13. Hence, the proof of the statement in (c) follows. □

We are now ready give the proof of Theorem 4.2.4.

Proof of Theorem 4.2.4. For a fixed $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, first define

$$\bar{c} = \bar{c}(\hat{\rho}, \hat{\eta}), \quad L^{\psi}_{\tilde{c}_l} = 1 + \lambda(\kappa_1 + \tilde{c}_l \kappa_2), \quad \forall l \geq 1,$$

where $\bar{c}(\cdot,\cdot)$ and $\tilde{c}_l$ are as in (4.50) and (4.45), respectively. Moreover, let $\bar{l}$ be the first index l such that $\tilde{c}_l \geq \bar{c}$, and recall from (4.45) that in the l^{th} cycle of the AIP.ALM, we have $c_k = \tilde{c}_l = 2^{l-1} c_1$, for every $l \geq 1$. In view of Lemma 4.2.14(c), we see that the AIP.AL obtains a solution of Problem 4.2.1 within the $\bar{l}^{\text{th}}$ cycle. Moreover, it follows by Lemma 4.2.14(a)–(b) that the total number of inner iterations performed by AIP.ALM is $\mathcal{O}(T_I)$ where

$$T_I := \frac{R_\phi}{\lambda \hat{\rho}^2} \sum_{l=1}^{\bar{l}} \sqrt{\frac{\bar{L}^{\psi}_{\tilde{c}_l}}{1 - \lambda m}} \log_1^+ \left[\frac{\bar{L}^{\psi}_{\tilde{c}_l}}{\theta} \right]. \tag{4.76}$$

Since c_k is doubled every time the cycle is changed, we have in view of the definitions of $\tilde{c}_l$

and $\bar{l}$ that

$$\tilde{c}_l \leq \max\{c_1, 2\bar{c}\}, \quad \forall l = 1, \ldots, \bar{l}. \tag{4.77}$$

Hence, it holds that

$$\begin{aligned} \bar{L}^{\psi}_{\tilde{c}_l} &= 1 + \lambda(\kappa_1 + \tilde{c}_l\kappa_2) \\ &\leq [\lambda(\kappa_1 + c_1\kappa_2) + 1]\frac{\max\{c_1, 2\bar{c}\}}{c_1}. \end{aligned} \tag{4.78}$$

Moreover, using (4.77) and the fact that $\tilde{c}_l = 2^{l-1}c_1$, it holds that

$$\begin{aligned} \sum_{l=1}^{\bar{l}} \sqrt{\bar{L}^{\psi}_{\tilde{c}_l}} = \sum_{l=1}^{\bar{l}} \sqrt{\lambda(\kappa_1 + \tilde{c}_l\kappa_2) + 1} &\leq \sqrt{\lambda(\kappa_1 + c_1\kappa_2) + 1} \sum_{l=1}^{\bar{l}} \sqrt{2}^{l-1} \\ &\leq 8\sqrt{\lambda(\kappa_1 + c_1\kappa_2) + 1}\left(\frac{\bar{c}_{\bar{l}}}{c_1}\right)^{1/2} \\ &\leq 8\sqrt{\lambda(\kappa_1 + c_1\kappa_2) + 1}\left(\frac{\max\{c_1, 2\bar{c}\}}{c_1}\right)^{1/2}. \end{aligned}$$

Hence, (4.51) then follows by combining (4.50), (4.76), (4.78), and the above inequalities.

□

4.3 Conclusion and Additional Comments

In this chapter, we presented two optimization methods for finding approximate stationary points for two classes of set-constrained optimization problems with constraints of the form $g(z) \in S \subseteq \mathcal{R}$. More specifically, a quadratic penalty method was proposed for a class of linear set-constrained NCO problems and a proximal augmented Lagrangian method was proposed for a class of nonlinearly cone-constrained NCO problems. We then established $\mathcal{O}(\hat{\eta}^{-1}\hat{\rho}^{-2})$ and $\mathcal{O}([\hat{\eta}^{-1/2}\hat{\rho}^{-2} + \hat{\rho}^{-3}]\log_1^+[\hat{\rho}^{-1} + \hat{\eta}^{-1}])$ iteration complexity bounds, in each of the respective methods, for finding $\hat{\rho}$-approximate stationary points that are $\hat{\eta}$ feasible, i.e. points $\bar{z}$ satisfying $\operatorname{dist}(g(\bar{z}), S) \leq \hat{\eta}$.

The next chapter continues the developments in Chapter 3 to develop a smoothing

method for solving min-max NCO problems.

Additional Comments

We now give some additional comments about the results and assumptions in this chapter.

First, it is worth stressing that the regularity condition in assumption (C3), which is a generalization of the weak Slater condition (see Proposition 4.2.1), is generally easier to verify compared to other conditions in the literature. For example, paper [58] requires a regularity condition to hold at every point generated by their proposed algorithm and paper [13] requires either the Mangasarian-Fromovitz constraint qualification or strong feasibility to hold. It is worth mentioning that we do **not** assume any regularity conditions on the linear set constraints in Section 4.1.

Second, we comment on the contributions of the AIP.QPM to the literature. The AIP.QPM and the QP-AIPP method from [46] appear to be the first methods to consider an infeasible starting point with a guaranteed complexity bound under the general assumptions in this chapter. Moreover, these methods have substantially improved on the previous state-of-art complexity bound of $\mathcal{O}(\hat{\rho}^{-6})$ which was obtained in [42] under the assumption that Z is bounded and $\hat{\rho} = \hat{\eta}$.

Third, we comment on how the AIP.ALM compares with the works [35, 40, 58, 69, 101, 110]. The IAPIAL method of [69] is designed to solve the special instance of $\mathcal{CNCO}[b]$ in which $\mathcal{K} = \{0\}$. In contrast to the AIP.ALM, the IAPIAL method sets p_k to p_0 every time the penalty parameter c_k is increased, and hence it is not a full warm-start proximal augmented Lagrangian method. Compared to [58, 101], the multiplier update in (4.27) is performed at every prox iteration, regardless of whether the penalty parameter is updated. Unlike the methods in [35, 40, 110], which require the initial point z_0 to be feasible, i.e. $g(z_0) \preceq_{\mathcal{K}} 0$, the AIP.ALM only requires z_0 to be in Z.

Future Work

Several recent works present improved complexity bounds (compared to the ones in this chapter) for obtaining approximate stationary points of linearly-constrained [114, 115] and nonlinearly-constrained [58, 62] NCO problems under different conditions and multiplier updates. For example, papers [114, 115] assume that h is the indicator of a polyhedron and [58] requires the Lagrange multiplier and penalty updates be performed simultaneously. It would be worth investigating whether the methods in this chapter, or some variant of them, can obtain these improved rates. Comparing the AIP.ALM to the AIP.QPM, the former assumes that the composite function h has bounded domain and is Lipschitz continuous, whereas the latter does not. It would be interesting to see if the AIP.ALM, or some variant of it, can still obtain approximate stationary points when the above conditions are removed.

CHAPTER 5
EFFICIENT IMPLEMENTATION STRATEGIES

The main goal of this chapter is to present efficient implementation strategies of some procedures and methods presented in prior chapters for smooth NCO problems. For the iterative methods, in particular, the variants in this chapter consider two key improvements. First, they apply efficient line search subroutines to adaptively choose parameters that directly affect convergence rates, such as stepsize parameters. Second, the convex subproblems that are solved in each of the iterative method are relaxed to (possibly) nonconvex subproblems. The degree of relaxation in these subproblems is determined by checking a finite set of novel descent inequalities which are guaranteed to hold when the subproblems are convex. We then demonstrate the effectiveness of these strategies on many of optimization problems in the literature.

The content of this chapter is based on paper [47] (joint work with Jefferson G. Melo and Renato D.C. Monteiro) and several passages may be taken verbatim from it.

Organization

This chapter contains six sections. The first one presents an efficient refinement procedure. The second one presents a relaxed ACG variant. The third one presents a relaxed AIPP variant and its iteration complexity. The fourth one presents a relaxed AIP.QPM variant and its iteration complexity. The fifth one presents a large collection of numerical experiments The last one gives a conclusion and some closing comments.

5.1 Proximal Refinement Procedure

This section presents a refinement procedure that is generally more effective in practice than the refinement procedure (the CRP) in Algorithm 3.2.2.

We first state the procedure in Algorithm 5.1.1, which follows a similar approach as in the CRP.

Algorithm 5.1.1: PR Procedure

Require: $h \in \overline{\text{Conv}}\,(\mathcal{Z})$, $f \in \mathcal{C}(Z)$, $(z, z^-, v) \in \mathcal{Z}^3$, $L > 0$, $\lambda > 0$;

Initialize: $L_\lambda \leftarrow \lambda L + 1$, $f_\lambda \leftarrow \lambda f + \frac{1}{2}\|\cdot - z^-\|^2 - \langle v, \cdot\rangle$, $h_\lambda \leftarrow \lambda h$;

1: **procedure** PREF($f, h, z, z^-, v, L, \lambda$)

2: $\quad z_r \leftarrow \underset{u\in\mathcal{Z}}{\text{argmin}} \left\{ \ell_{f_\lambda}(u; z) + h_\lambda(u) + \frac{L_\lambda}{2}\|u - z\|^2 \right\}$

3: $\quad v_r \leftarrow \frac{1}{\lambda}\left[(v + z^- - z) + L_\lambda(z - z_r)\right] + \nabla f(z_r) - \nabla f(z)$

4: $\quad \varepsilon_r \leftarrow (f_\lambda + h_\lambda)(z) - (f_\lambda + h_\lambda)(z_r)$

5: $\quad$ **return** $(z_r, v_r, \varepsilon_r)$

The result below, whose proof can be found in Appendix D, presents the some important properties of the PR procedure (PRP).

Proposition 5.1.1. *Let $(z_r, v_r, \varepsilon_r)$ and L_λ be generated by the PRP where (f, h) satisfy assumptions (A1)–(A2). Then, the following properties hold:*

(a) $\varepsilon_r \geq L_\lambda\|z - z_r\|^2/2$;

(b) $v_r \in \nabla f(z_r) + \partial h(z_r)$ *and*

$$\|v_r\| \leq \frac{1}{\lambda}\|v + z^- - z\| + \left(\frac{1}{\lambda} + \frac{\max\{m, M\}}{L_\lambda}\right)\sqrt{2\varepsilon_r L_\lambda};$$

(c) if the inputs f, h, λ, and (z, z^-, v) satisfy

$$\begin{aligned} v &\in \partial_{\varepsilon_r}\left(\lambda\left[f + h\right] + \frac{1}{2}\|\cdot - z^-\|^2\right)(z), \\ \frac{1}{\lambda}\|z^- - z + v\| &\leq \bar{\rho}, \quad \frac{1}{\lambda}\varepsilon \leq \bar{\varepsilon}, \end{aligned} \tag{5.1}$$

for some $(\bar{\rho}, \bar{\varepsilon}) \in \mathbb{R}^2_{++}$ *and* $\varepsilon > 0$*, it holds that*

$$\|v_r\| \leq \bar{\rho} + \left(\frac{1}{\lambda} + \frac{\max\{m, M\}}{L_\lambda}\right)\sqrt{2\lambda\bar{\varepsilon}L_\lambda}. \tag{5.2}$$

The result above is analogous to Proposition 3.2.5, which describes properties of the CRP. In view of this link, we now make a comparison between the PRP and the aforementioned CRP. First, the PRP requires two extra points, z^- and v, as part of its input compared to the CRP. Second, Proposition 3.2.5(b) shows that the CRP obtains a point v_r satisfying the inclusion in Proposition 5.1.1(b). Finally, under the same conditions in (5.1), Proposition 3.2.5(c) shows that the point v_r obtained by the CRP satisfies

$$\|v_r\| \leq \left(1 + \frac{\max\{m, M\}}{L_\lambda}\right)\left(\bar{\rho} + \sqrt{2\bar{\varepsilon}L_\lambda}\right), \tag{5.3}$$

which is analogous to the bound in (5.2). Note that, compared to (5.2), the above bound has a larger constant in front of $\bar{\rho}$ and a possibly larger constant in front of $\bar{\varepsilon}$ depending on the relationships between λ, M, m, and L.

5.2 Relaxed ACG (R.ACG) Method

This section presents a relaxed ACG (R.ACG) variant that is generally more efficient in practice than the ACGM in Algorithm 2.2.2.

We first state the R.ACG variant in Algorithm 5.2.1. It main idea is to start with a possibly large stepsize λ_1 and adaptively update this stepsize by checking a particular descent inequality at every iteration.

Algorithm 5.2.1: R.ACG Method

Require: $\psi_n \in \overline{\text{Conv}}\,(\mathcal{Z})$, $\psi_n \in \mathcal{C}(\text{dom}\,\psi_n)$, $y_0 \in \text{dom}\,\psi_n$, $(\mu, L_{\text{est}}) \in \mathbb{R}^2_{++}$, $L_{\min} \in (0, L_{\text{est}}]$;

Initialize: $L_1 \leftarrow L_{\text{est}}$

procedure R.ACG($\psi_s, \psi_n, y_0, \mu, L_{\min}, L_{\text{est}}$)
  for $k = 1, \ldots$ **do**
    $L \leftarrow L_k$
    do
      $\lambda_k \leftarrow 1/L$
      Generate $(A_k, y_k, \tilde{x}_{k-1}, r_k, \eta_k)$ according to Algorithm 2.2.2.
      $L \leftarrow 2(L - L_{\min}) + L_{\min}$
    while $\psi_s(y_k) - \ell_{\psi_s}(y_k; \tilde{x}_{k-1}) > \dfrac{L}{2}\|y_k - \tilde{x}_{k-1}\|^2$
    $L_{k+1} \leftarrow L$

We now make two remarks about the above R.ACG method (R.ACGM). First, if $\psi_s \in \mathcal{C}_{m,\overline{L}}(\operatorname{dom}\psi_n)$ for some $(m, \overline{L}) \in \mathbb{R}^2_{++}$ and $L_{\text{est}} \geq \overline{L}$, then $L_k = L_{\text{est}}$ for every $k \geq 1$. On the other hand, if $L_{\text{est}} < \overline{L}$ then L_k is doubled at most

$$\left\lceil 1 + \log_2\left(\frac{L - L_{\text{est}}}{L_{\text{est}} - L_{\min}}\right)\right\rceil$$

times and $L_1 \leq L_k \leq 2\overline{L}$ for every $k \geq 1$. Second, if $(L - L_{\text{est}})/(L_{\text{est}} - L_{\min}) = \mathcal{O}(1)$, then the iteration complexities of the R.ACGM and ACGM in Algorithm 2.2.2 are on the same order of magnitude when given a common termination condition.

It is worth mentioning that the above line search idea has been explored in many other works in the literature. For example, [86] considers applying a similar line search subroutine in which the stepsize parameter λ_k is increased whenever a key descent inequality holds and decreased otherwise.

5.3 Relaxed AIPP (R.AIPP) Method

This section establishes an iteration complexity bound for a relaxed AIPPM (R.AIPPM) that is generally more efficient in practice than the AIPPM in Algorithm 3.3.2.

Before proceeding, we first state the main problem of the R.AIPPM and its key assumptions. Consider the NCO problem

$$\phi_* = \min_{z \in \mathcal{Z}} \left[\phi(z) := f(z) + h(z) \right], \tag{$\mathcal{NCO}$}$$

where $\mathcal{Z}$ is a finite dimensional inner product space, and it is assumed that

(D1) $h \in \overline{\text{Conv}}\,(Z)$ for some nonempty convex set $Z \subseteq \mathcal{Z}$;

(D2) $f \in \mathcal{C}_M(Z)$ for some $M > 0$;

(D3) $\phi_* > -\infty$.

Moreover, like in Chapter 3, assume that efficient oracles for evaluating the quantities $f(z)$, $\nabla f(z)$, and $h(z)$ and for obtaining exact solutions of the subproblem

$$\min_{z \in \mathcal{Z}} \left\{ \lambda h(z) + \frac{1}{2}\|z - z_0\|^2 \right\},$$

for any $z_0 \in \mathcal{Z}$ and $\lambda > 0$, are available.

The AIPPM considers finding approximate stationary points of $\mathcal{NCO}$ as in Problem 3.1.1, i.e. given $\hat{\rho} > 0$, find $(\hat{z}, \hat{v}) \in Z \times \mathcal{Z}$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}), \quad \|\hat{v}\| \leq \hat{\rho}. \tag{5.4}$$

For the sake of future referencing, let us state the problem of finding $(\hat{z}, \hat{v})$ satisfying (5.4) in Problem 5.3.1.

Problem 5.3.1: Find an approximate stationary point of $\mathcal{NCO}$

Given $\hat{\rho} > 0$, find a pair $(\hat{z}, \hat{v}) \in Z \times \mathcal{Z}$ satisfying condition (5.4).

To ease the notation in later sections, let us conclude by defining the useful quantity

$$\underline{m} := \inf_{m > 0} \left\{ f(u) - \ell_f(u; z) \geq -\frac{m}{2}\|u - z\|^2 \quad \forall u, z \in Z \right\}. \tag{5.5}$$

5.3.1 General Descent (GD) Framework

This subsection presents a general descent (GD) framework that relaxes the GIPP framework from Chapter 3. We later show that the R.AIPPM is a special instance of GD framework (GDF) in which each prox subproblem is approximate solved by invoking the R.ACGM in Algorithm 2.2.2.

Recall that for an IPP framework with stepsizes $\{\lambda_k\}_{k\geq 1}$, the larger λ_k is the faster the IPP framework converges to a desirable approximate solution. While λ_k is required to be at most $1/\underline{m}$ in the GIPPF of Chapter 3, the GDF of this subsection considers choosing λ_k significantly larger than $1/\underline{m}$ despite a possible loss of convexity. More specifically, it adaptively chooses its stepsizes based on two key inequalities that are checked at the end of its iterations.

We first start by stating the GDF in Algorithm 5.3.1.

Algorithm 5.3.1: GD Framework

Require: $h \in \overline{\text{Conv}}\,(Z)$, $f \in \mathcal{C}(Z)$, $z_0 \in Z$, $(\theta,\tau) \in \mathbb{R}^2_{++}$, $L > 0$, $\{\lambda_k\}_{k\geq 1} \subseteq \mathbb{R}_{++}$;
Initialize: $L_\lambda \leftarrow \lambda M + 1$, $\phi \Leftarrow f + h$;

1: **procedure** GD$(f, h, z_0, \theta, \tau, M)$
2: **for** $k = 1, \ldots$ **do**
3: **Find** $(z_k, v_k, \lambda_k) \in Z \times \mathcal{Z} \times \mathbb{R}_{++}$ such that its corresponding refined triple

$$(\hat{z}_k, \hat{v}_k, \hat{\varepsilon}_k) \leftarrow \text{PREF}(f, h, z_k, z_{k-1}, v_k, M, \lambda_k) \tag{5.6}$$

4: satisfies the bounds

$$\|v_k + z_{k-1} - z_k\|^2 \leq \theta\lambda_k\left[\phi(z_{k-1}) - \phi(z_k)\right], \tag{5.7}$$

$$2L_\lambda\hat{\varepsilon}_k \leq \tau\|v_k + z_{k-1} - z_k\|^2. \tag{5.8}$$

We now give two remarks about the above framework. First, no termination criterion is added so as to be able to discuss convergence rate results about its generated sequence. A discussion of how to terminate it is given after Proposition 5.3.1 below. Second, its Line 3 should be viewed as an oracle in that it does not specify how to compute the triple (λ_k, z_k, v_k). Third,

Corollary 5.3.4 below shows that if the stepsize λ_k is chosen so that the prox subproblem

$$\min_{z\in\mathcal{Z}}\left\{\lambda_k(f+h)(z)+\frac{1}{2}\|z-z_{k-1}\|^2\right\} \tag{5.9}$$

is a strongly convex composite problem, i.e. $\lambda_k \in (0, 1/\underline{m})$, the point z_k is chosen as its unique optimal solution, and v_k is set to zero, then the triple (λ_k, z_k, v_k) satisfies (5.7) and (5.8) with $\theta = 2$ and $\tau = 0$. Thus, when $(\theta,\tau) \in [2,\infty)\times[0,\infty)$, we conclude that: (i) there always exists a triple satisfying (5.7) and (5.8); and (ii) the GD framework can be viewed as an IPP method.

In Section 5.3.3, we show that the R.AIPPM is a special instance of the GD framework, and hence, can be viewed as a relaxed IPP method which chooses (θ,τ) in the open rectangle $(2,\infty)\times(0,\infty)$. In particular, it applies an instance of the R.ACGM in Algorithm 5.2.1 to problem (5.9) in order to obtain a triple (λ_k, z_k, v_k) satisfying (5.7) and (5.8).

We now present an important property about the sequence of iterates $\{(\lambda_k,\hat{z}_k,\hat{v}_k)\}_{k\geq1}$.

Proposition 5.3.1. *The sequences of stepsizes $\{\lambda_k\}_{k\geq1}$ and iterate pairs $\{(\hat{z}_k,\hat{v}_k)\}_{k\geq1}$ satisfy*

$$\hat{v}_k \in \nabla g(\hat{z}_k)+\partial h(\hat{z}_k), \quad \min_{i\leq k}\|\hat{v}_i\|^2 \leq \theta\left(1+2\sqrt{\tau}\right)^2\frac{[\phi(z_0)-\phi_*]}{\Lambda_k}, \tag{5.10}$$

for every $k\geq1$, where $\Lambda_k := \sum_{i=1}^k\lambda_i$.

Proof. Let $k\geq1$ be fixed. The inclusion in (5.10) follows from Proposition 5.1.1 with $(\hat{z},\hat{v}) = (\hat{z}_k,\hat{v}_k)$ and the definitions of $\hat{z}_k$ and $\hat{v}_k$ in (5.6). To show the inequality in (5.10), first observe that (5.7) and the definition of ϕ_* in $\mathcal{NCO}$ implies that

$$\begin{aligned}\phi(z_0)-\phi_* &\geq \sum_{i=1}^k[\phi(z_{i-1})-\phi(z_i)] \geq \sum_{i=1}^k\frac{\|v_i+z_{i-1}-z_i\|^2}{\theta\lambda_i}\\ &\geq \frac{\Lambda_k}{\theta}\min_{i\leq k}\frac{1}{\lambda_i^2}\|v_i+z_{i-1}-z_i\|^2.\end{aligned} \tag{5.11}$$

Now, let $i\geq1$ be arbitrary. Using (5.6), (5.8) with $k=i$, and Proposition 5.1.1 with $\lambda=\lambda_i$,

$(z^-, z) = (z_{i-1}, z_i)$, and $(v, v_r) = (v_i, \hat{v}_i)$, it holds that

$$\begin{aligned}\|\hat{v}_i\| &\leq \frac{1}{\lambda_i}\|v_i + z_{i-1} - z_i\| + \left(\frac{1}{\lambda_i} + \frac{M}{\lambda_i M + 1}\right)\sqrt{2(\lambda_i M + 1)\hat{\varepsilon}_i} \\ &\leq \frac{1}{\lambda_i}\|v_i + z_{i-1} - z_i\| + \frac{2}{\lambda_i}\sqrt{2(\lambda_i M + 1)\hat{\varepsilon}_i} \qquad (5.12) \\ &\leq \left(\frac{1 + 2\sqrt{\tau}}{\lambda_i}\right)\|v_i + z_{i-1} - z_i\|. \qquad (5.13)\end{aligned}$$

The inequality in (5.10) now follows by combining (5.11) and (5.13). □

We now make three additional remarks about the GDF in light of Proposition 5.3.1. First, if the GDF stops when a pair $(\hat{z}_k, \hat{v}_k)$ such that $\|\hat{v}_k\| \leq \hat{\rho}$ is found, then it follows from the inclusion in (5.10) that $(\hat{z}_k, \hat{v}_k)$ solves Problem 5.3.1. Second, if the sequence of stepsizes $\{\lambda_i\}$ satisfies $\lim_{k\to\infty} \Lambda_k = \infty$, then it follows from the inequality in (5.10) and assumption (D3) that the GDF indeed stops according to the above termination criterion. Third, (5.10) indicates that the larger the stepsizes λ_k are, the faster the quantity $\min_{i\leq k} \|\hat{v}_i\|$ approaches zero.

For the remainder of this section, our goal is to show that the GDF can be seen as a relaxation of the GIPPF from Section 3.2. The proof of this fact is not essential in establishing any results pertaining to the R.AIPPM in this section and may skipped without any loss of continuity.

Recall that, for a given $z_0 \in Z$ and $\sigma \in [0, 1)$, the GIPPF in Section 3.2 considers a sequences $\{\lambda_k\}_{k\geq 1}$ and $\{(z_k, v_k, \varepsilon_k)\}_{k\geq 1}$ satisfying

$$v_k \in \partial_{\varepsilon_k}\left(\lambda_k\phi + \frac{1}{2}\|\cdot - z_{k-1}\|^2\right)(z_k), \quad \|v_k\|^2 + 2\varepsilon_k \leq \sigma\|v_k + z_{k-1} - z_k\|^2, \qquad (5.14)$$

for every $k \geq 1$. We begin by presenting a simple technical result that will be used both here and in the analysis of the R.AIPPM.

Lemma 5.3.2. *Assume that* $\varepsilon \geq 0$ *and* $(\lambda, z^-, z, v) \in \mathbb{R}_{++} \times \mathcal{Z} \times Z \times \mathcal{Z}$ *satisfy*

$$v \in \partial_\varepsilon \left(\lambda\phi + \frac{1}{2}\|\cdot - z^-\|^2 \right)(z). \tag{5.15}$$

Then, the quantity ε_r *computed in Algorithm 5.1.1 satisfies* $\varepsilon_r \leq \varepsilon$.

Proof. Let (z_r, ε_r) be computed as in Algorithm 5.1.1. It follows from the definition of the approximate subdifferential and (5.15) that

$$\lambda\phi(u) + \frac{1}{2}\|u - z^-\|^2 \geq \lambda\phi(z) + \frac{1}{2}\|z - z^-\|^2 + \langle v, u - z\rangle - \varepsilon \quad \forall u \in \mathcal{Z}.$$

Considering the above inequality at the point $u = z_r$, along with some algebraic manipulation, we have

$$\varepsilon \geq \left[\lambda\phi(z) + \frac{1}{2}\|z - z^-\|^2 - \langle v, z\rangle \right] - \left[\lambda\phi(z_r) + \frac{1}{2}\|z_r - z^-\|^2 - \langle v, z_r\rangle \right] = \varepsilon_r,$$

where the last equality is due to the definitions f_λ, h_λ, and ε_r given in Algorithm 5.1.1. □

The following result shows the relationship between the GIPPF of Section 3.2 and the GDF of this section.

Proposition 5.3.3. *If, for some* $z_{k-1} \in Z$, *constant* $\sigma \in [0, 1)$, *and index* $k \geq 1$, *the quadruple* $(\lambda_k, z_k, v_k, \varepsilon_k)$ *satisfies* (5.14), *then* (λ_k, z_k, v_k) *satisfies* (5.7) *and* (5.8) *for any* $\theta \geq 2/(1-\sigma)$ *and* $\tau \geq \sigma(\lambda_k M + 1)$. *As a consequence, if* $\sup_{k\geq 1} \lambda_k < \infty$, *then every instance of the GIPPF is an instance of the GDF for any* (θ, τ) *satisfying*

$$\theta \geq \frac{2}{1-\sigma}, \quad \tau \geq \sup_{k\geq 1} \left[\sigma(\lambda_k M + 1) \right]. \tag{5.16}$$

Proof. The fact that (λ_k, z_k, v_k) satisfies (5.7) with $\theta = 2/(1-\sigma)$ follows from Proposition 3.2.3(a). Now, let $k \geq 1$ and observe that from Lemma 5.3.2 with $(\lambda, z^-, z, v) = (\lambda_k, z_{k-1}, z_k, v_k)$ and $\varepsilon = \varepsilon_k$ we have $\hat{\varepsilon}_k \leq \varepsilon_k$. It follows from the last inequality and the

inequality in (5.14) that $2\hat{\varepsilon}_k \leq \sigma\|v_k + z_{k-1} - z_k\|^2$. Combining the previous inequality with the assumption on τ now shows that (λ_k, z_k, v_k) satisfies (5.8). The second part of the proposition follows immediately from the first part and condition (5.16). □

[add remarks]

Corollary 5.3.4. *Let $z_{k-1} \in Z$ and $\lambda_k \in (0, 1/\underline{m})$ be given, where $\underline{m}$ is as in* (5.5)*. Then, $\mathcal{NCO}$ has a unique global minimum z_k and the triple (λ_k, z_k, v_k) satisfies* (5.7) *and* (5.8) *with $(\theta, \tau, v_k) = (2, 0, 0)$.*

Proof. The existence and unique uniqueness of z_k follows from the fact that $\phi + \|\cdot - z_{k-1}\|^2/\lambda_k$ is strongly convex. Moreover, the fact that z_k is the unique global minimum of $\mathcal{NCO}$ implies that the quadruple $(\lambda_k, z_k, v_k, \varepsilon_k)$, where $(v_k, \varepsilon_k) = (0, 0)$, satisfies (5.14) with $\sigma = 0$. The conclusion of the corollary now follows immediately from the first part of Proposition 5.3.3 with $\sigma = 0$. □

5.3.2 Key Properties of the R.ACGM

This subsection describes how the R.ACGM in Algorithm 5.2.1 can be used to implement a single iteration of the GDF in Section 5.3.1.

Consider the R.ACGM inputs

$$\begin{gathered} \psi_s = \lambda f + \frac{1}{2}\|\cdot - z_{k-1}\|^2, \quad \psi_n = \lambda h, \quad y_0 = z_{k-1}, \\ \mu = 1, \quad L_{\min} = 1, \quad L_{\text{est}} = L_\lambda := \lambda M + 1, \end{gathered} \tag{5.17}$$

and the termination criteria

$$2\max\{0, L_\lambda \eta_j\} \leq \tau\|y_0 - y_j + r_j\|^2, \tag{5.18}$$

$$\|y_0 - y_j + r_j\|^2 \leq \lambda\theta\left[\phi(y_0) - \phi(y_j)\right], \tag{5.19}$$

for some $(\theta, \tau) \in \mathbb{R}^2_{++}$. In the following lemma, we show that if the conditions (5.18) and

(5.19)

$$\|A_i r_i + y_i - y_0\|^2 + 2\max\{0, A_i\eta_i\} \le \|y_i - y_0\|^2, \tag{5.20}$$

$$\psi(y_0) \ge \psi(y_i) + \langle r_i, y_0 - y_i\rangle - \max\{0, \eta_i\}, \tag{5.21}$$

hold at every iteration of R.ACGM, then the conditions (5.18) and (5.19) will be obtained in a finite number of R.ACG iterations.

Lemma 5.3.5. *Let $\phi = f + h$ be a function satisfying assumptions (D1)–(D2), L_λ be as in* (5.17)*, and $(z_{k-1}, \lambda) \in Z \times \mathbb{R}_{++}$. Moreover, suppose the R.ACGM is called with $(\psi_s, \psi_n, y_0, L_{\min}, L_{\text{est}})$ as in* (5.17) *and generates the sequence of iterates $\{(A_i, y_i, r_i, \eta_i)\}_{i\ge1}$. Then, the following statements hold:*

(a) if the inequalities (5.20) *and* (5.21) *hold for every $i \ge 1$, then for any $\theta > 2$ and $\tau > 0$ the R.ACGM generates an iterate (y_j, r_j, η_j) satisfying* (5.18) *and* (5.19) *in*

$$\left\lceil 1 + \sqrt{2L_\lambda}\,\log_1^+(2C_{\theta,\tau}L_\lambda)\right\rceil \tag{5.22}$$

iterations, where

$$C_{\theta,\tau} := \max\left\{\left[1 + \sqrt{\frac{L_\lambda}{\tau}}\right]^2, \left[1 + \sqrt{\frac{\theta}{\theta-2}}\right]^2\right\}. \tag{5.23}$$

(b) if $\lambda \le 1/\underline{m}$, then (5.20)*,* (5.21)*, and the inclusion $r_j \in \partial_{\max\{\eta_j,0\}}\psi(y_j)$ hold for every $j \ge 1$.*

Proof. (a) See Appendix C.

(b) If $\lambda \le 1/\underline{m}$, it follows that $\psi_s \in \mathcal{F}_{0,L_\lambda}(Z)$, and hence, (ψ_s, ψ_n) satisfy the requirements of the ACGM (see Algorithm 2.2.2) with $L = L_\lambda$. The conclusion now follows from Proposition 2.2.3 and the definition of the approximate subdifferential. □

The next result shows that conditions (5.18) and (5.19) are sufficient to implement a single iteration of the GDF in Section 5.3.1.

Lemma 5.3.6. *Let $\phi = f + h$ and (z_{k-1}, λ) be as in Lemma 5.3.5 and (ψ_s, ψ_n, y_0) be as in* (5.17)*. If (y_j, r_j, η_j) satisfy* (5.18)*,* (5.19)*, and $r_j \in \partial_{\max\{\eta_j, 0\}}(y_j)$, then the assigned triples*

$$(\lambda_k, z_k, v_k) \leftarrow (\lambda, y_j, r_j), \quad (\hat{z}_k, \hat{v}_k, \hat{\varepsilon}_k) \leftarrow \textit{PREF}(f, h, y_j, z_{k-1}, r_j, M, \lambda) \tag{5.24}$$

satisfy (5.7) *and* (5.8)*.*

Proof. The fact that $(\lambda_k, z_{k-1}, z_k, v_k)$ satisfies (5.7) follows immediately from (5.19) and (5.24). On the other hand, using Lemma 5.3.2 with $(z, v, \varepsilon) = (y_j, r_j, \max\{\eta_j, 0\})$ and the definition of (ψ_s, ψ_n), we have that $\hat{\varepsilon}_k \leq r_j$. Using the previous bound and (5.18) yields (5.8). □

We now conclude by discussing alternative choices for the R.ACG input $(L_{\text{est}}, L_{\min})$ in (5.17). First, note that if $L_{\text{est}} = \lambda\alpha M + 1$ for some $\alpha \in (0, 1)$ and $L_{\min} = 1$, then

$$\frac{L_\lambda - L_{\min}}{L_{\text{est}} - L_{\min}} = \frac{\lambda M + 1 - 1}{\lambda\alpha M + 1 - 1} = \frac{1}{\alpha}.$$

Hence, in view of the above identity and the discussion following the R.ACGM in Algorithm 2.2.2, choosing $L_{\text{est}} = \lambda\alpha M + 1$ with $\alpha^{-1} = \mathcal{O}(1)$ in an R.ACG call yields an complexity that is on the same order of magnitude as an R.ACG call with $L_{\text{est}} = \lambda M + 1$.

5.3.3 Statement and Properties of the R.AIPPM

This subsection describes and gives the iteration complexity of the R.AIPPM.

We first state the R.ACG instance in Algorithm 5.3.2 that implements the approach described in the preceding subsection. More specifically, this instance chooses $L_{\min} = 1$ and $L_{\text{est}} = \lambda M/100 + 1$, uses the termination conditions (5.18) and (5.19), and uses (5.20) and (5.21) to check for failure of the method. The variable π_S is used to store the termination

status of the method where $\pi_S =$ `true` if the method outputs a solution satisfying (5.18) and (5.19) and $\pi_S =$ `false` otherwise.

Algorithm 5.3.2: R.ACG Instance for the R.AIPPM

Require: $\psi_n \in \overline{\mathrm{Conv}}\,(\mathcal{Z})$, $\psi_n \in \mathcal{C}(Z)$, $y_0 \in Z$, $(\theta, \tau) \in \mathbb{R}^2_{++}$, $L_1 > 0$, $L_{\min} \in (0, L_1)$;
Initialize: $L_1 \leftarrow L_{\mathrm{est}}$, $\pi_S \leftarrow$ `true`, $\psi \Leftarrow \psi_s + \psi_n$,

1: **procedure** R.ACG1($\psi_s, \psi_n, \phi, y_0, \theta, \tau, L_{\min}, L_{\max}, L_{\mathrm{est}}$)
2: **for** $k = 1, \ldots$ **do**
3: Generate (A_k, y_k, r_k, η_k) according to Algorithm 5.2.1.
4: $\eta_k^+ \leftarrow \max\{\eta_k, 0\}$
5: **if** (5.18) **and** (5.19) hold with $j = k$ **then**
6: **return** $(y_k, r_k, \eta_k^+, \pi_S)$
7: **if** (5.20) **or** (5.21) do not hold with $i = k$ **then**
8: $\pi_S \leftarrow$ `false`
9: **return** $(y_0, \infty, \infty, \pi_S)$

Using the R.ACGM instance in Algorithm 5.3.2 and the refinement procedure in Algorithm 5.1.1, we now state the R.AIPPM in Algorithm 5.3.3. Given $\lambda_0 > 0$ and $z_0 \in Z$, its main idea is to apply the R.ACGM to obtain the approximate update for the k^{th} iteration

$$z_k \approx \min_{z \in \mathcal{Z}} \left\{ \lambda_k (f + h)(z) + \frac{1}{2} \|z - z_{k-1}\|^2 \right\}$$

for a suitable stepsize λ_k, and implement one iteration of the GDF in Section 5.3.1. The iterate z_k is then refined using the PRP in Algorithm 5.1.1 and termination of the method occurs when a refined iterate solving Problem 5.3.1 is found.

Algorithm 5.3.3: R.AIPP Method

Require: $\hat{\rho} > 0$, $M > 0$, $h \in \overline{\mathrm{Conv}}\,(Z)$, $f \in \mathcal{C}_M(Z)$, $\lambda_0 > 0$, $(\theta, \tau) \in (2, \infty) \times \mathbb{R}_{++}$, $z_0 \in Z$;
Initialize: $\phi \Leftarrow f + h$;

1: **procedure** R.AIPP($f, h, z_0, \lambda_0, \theta, \tau, M, \hat{\rho}$)

2: **for** $k = 1, \ldots$ **do**

3: PART 1 **Find** the right λ_k and **attack** the k^{th} prox subproblem.

4: $\lambda \leftarrow \lambda_{k-1}$

5: $\psi_s^k \Leftarrow \lambda f + \frac{1}{2}\|\cdot - z_{k-1}\|^2$

6: **repeat**

7: $(L_{\min}, L_{\max}, L_{\text{est}}) \leftarrow (1, \lambda M + 1, \lambda\lfloor M/100 \rfloor + 1)$

8: $(z_k, v_k, \varepsilon_k, \pi_k^{\text{acg}}) \leftarrow \text{R.ACG1}(\psi_s^k, \lambda h, \phi, y_0, \theta, \tau, L_{\min}, L_{\max}, L_{\text{est}})$

9: $(\hat{z}_k, \hat{v}_k, \hat{\varepsilon}_k) \leftarrow \text{PREF}(f, h, z_k, z_{k-1}, v_k, M, \lambda)$

10: **if** $\neg(\pi_k^{\text{acg}})$ **or** $2L_{\max}\hat{\varepsilon}_k > \tau\|v_k + z_{k-1} - z_k\|^2$ **then**

11: $\lambda \leftarrow \lambda/2$

12: **until** π_k^{acg} and $2L_{\max}\hat{\varepsilon}_k \le \tau\|v_k + z_{k-1} - z_k\|^2$

13: $\lambda_k \leftarrow \lambda$

14: PART 2 **Check** the termination condition.

15: **if** $\|\hat{v}_k\| \le \hat{\rho}$ **then**

16: **return** $(\hat{z}_k, \hat{v}_k)$

Some comments about the R.AIPPM are in order. To ease the discussion, let us refer to the ACG iterations performed in Line 8 of the method as **inner iterations** and the iterations over the indices k as **outer iterations**. First, the failure checks in the R.ACG instances and Line 10 of the method immediately imply that a single iteration of the R.AIPPM implements a single iterations of the GDF. Second, the termination condition in Line 15 and Proposition 5.1.1(b), with $(\lambda, z^-, z, v) = (\lambda_k, z_{k-1}, z_k, v_k)$, imply that the required solution, i.e. a pair $(\hat{z}, \hat{v})$ that solves Problem 5.3.1, is obtained when the R.AIPPM terminates. Third, since the R.AIPP iterates correspond to iterates of the GDF, and the sequence $\{\lambda_k\}$ is bounded below (see Lemma 5.3.7(c) below), Proposition 5.3.1 implies that the sequence $\{\hat{v}_k\}$ generated by the R.AIPPM has a subsequence approaching zero, and thus the method must terminate in Line 15. Fifth, although the R.AIPPM does not necessarily generate proximal subproblems with convex objective functions, it is shown in Theorem 5.3.8 below that it has an iteration complexity similar to that of the AIPPM of Section 3.3. Finally, in contrast to the aforementioned AIPPM, the R.AIPPM neither requires an upper bound on the quantity $\underline{m}$ in (5.5) as part of its input nor does it place any restriction on the initial stepsize λ_0.

Each iteration of the R.AIPPM may call the R.ACGM multiple times (possibly just one

time). Invocations of the R.ACGM algorithm that stop with $\pi_k^{\text{acg}} = \texttt{true}$ are said to be of type S while the other invocations are said to be of type F. Let k_S (resp., k_F) denote the total number of R.ACG calls of type S (resp., type F). The following technical result provides some basic facts about k_S, k_F, and the sequence of stepsizes $\{\lambda_k\}_{k\geq 1}$.

Lemma 5.3.7. *The following statements hold for the R.AIPPM:*

(a) if the stepsize $\lambda_{\bar{k}} \leq 1/\underline{m}$ for some $\bar{k} \geq 1$, then every iteration $k \geq \bar{k}$ is of type S and, as a consequence, $\lambda_k = \lambda_{\bar{k}}$ for every $k > \bar{k}$;

(b) k_F can be bounded as $2^{k_F} \leq \max\{1, 2\lambda_0\underline{m}\}$;

(c) $\{\lambda_k\}_{k\geq 1}$ is non-increasing and satisfies

$$\xi := \max\left\{\frac{1}{\lambda_0}, 2\underline{m}\right\} \geq \frac{1}{\lambda_k} \quad \forall k \geq 1. \tag{5.25}$$

Proof. (a) Since $\lambda_{\bar{k}} \leq 1/\underline{m}$, the definition of $\underline{m}$ in (5.5) implies that $\lambda_{\bar{k}} f + \|\cdot - z_{k-1}\|^2/2$ is convex. Hence, it follows from Lemma 5.3.5(b) that $\pi_k^{\text{acg}} = \texttt{true}$, which is to say that this iteration is of type S. Since $\{\lambda_k\}_{k\geq 1}$ is clearly nonincreasing, the same conclusion holds true for every iteration $k \geq \bar{k}$. Moreover, as λ is not halved for subsequent iterations following $\bar{k}$, it follows that $\lambda_k = \lambda_{\bar{k}}$ for every $k > \bar{k}$.

(b) Using the fact that immediately before each iteration of type F, the stepsize λ is halved, we see that the condition $\lambda_{\bar{k}} \leq 1/\underline{m}$ in part (a) would eventually be satisfied for some iteration $\bar{k} \geq 1$, and hence k_F is finite. Now, note that if $k_F = 0$ then the inequality in part (b) follows immediately. Assume then that $k_F \geq 1$. It now follows from part (a) and the definition of k_F that $\lambda_0/2^{k_F-1} > 1/\underline{m}$, which clearly implies the inequality in part (b).

(c) The first statement follows trivially from the update rule of λ_k in the R.AIPPM. Now, note that the definition of k_F together with the update rule for λ_k imply, for every $k \geq 1$, that $\lambda_0/2^{k_F} \leq \lambda_k$. The inequality in part (c) then follows from the inequality in part (b). □

In view of Lemma 5.3.7(a), choosing an initial stepsize λ_0 satisfying $\lambda_0 \leq 1/(2\underline{m})$ results in an R.AIPP variant with constant stepsize, which resembles the AIPPM described in Section 3.3.

The following theorem presents a worst-case iteration complexity bound on the number of inner iterations of the R.AIPPM with respect to the inputs M, λ_0, z_0, the quantity $\underline{m}$ in (5.5), and the tolerance $\hat{\rho}$.

Theorem 5.3.8. *The R.AIPPM outputs a pair $(\hat{z}, \hat{v})$ that solves Problem 5.3.1 in*

$$\mathcal{O}\left(\sqrt{M+\xi}\left(\frac{\sqrt{\xi}\theta\,[1+\tau]\,[\phi(z_0)-\phi_*]}{\hat{\rho}^2}+\sqrt{\lambda_0}\right)\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right) \tag{5.26}$$

inner iterations, where $C_{\theta,\tau}$ and ξ are as (5.23) *and* (5.25)*, respectively.*

Proof. The fact that its output solves Problem 5.3.1 follows from the termination condition in Line 15, Line 9, and Proposition 5.1.1.

To show the desired complexity, we let TI_S (resp. TI_F) denote the total number of inner iterations performed during all calls of type S (resp. type F) (see the paragraph preceding Lemma 5.3.7). Clearly, the total number of inner iterations is $\mathrm{TI} := \mathrm{TI}_S + \mathrm{TI}_F$. We now bound each one of the quantities TI_S and TI_F separately by using the fact that the inputs given to every R.ACG call and Lemma 5.3.5(a) imply that the number of inner iterations performed during each R.ACG call is

$$\mathcal{O}\left(\sqrt{\lambda M+1}\log_1^+\left[C_{\theta,\tau}(\lambda M+1)\right]\right),$$

where λ is the value of λ just before the call and C is as in (5.23) with $L_\lambda = \lambda M + 1$.

We first consider TI_F. Note that Lemma 5.3.7(b) implies that k_F is finite. Since $\mathrm{TI}_F = 0$ when $k_F = 0$, we may assume without loss of generality that $k_F \geq 1$. Note that the values of λ just before the k_F calls of type F are exactly $\lambda_0, \lambda_0/2, \ldots, \lambda_0/2^{k_F-1}$. Hence, we conclude

that

$$\begin{aligned}\mathrm{TI}_F &= \mathcal{O}\left(\sum_{i=1}^{k_F}\sqrt{\frac{\lambda_0 M}{2^{i-1}}+1}\,\log_1^+\left[C_{\theta,\tau}\left(\frac{\lambda_0 M}{2^{i-1}}\right)\right]\right)\\ &= \mathcal{O}\left(\sum_{i=1}^{k_F}\sqrt{\frac{\lambda_0\,(M+\xi)}{2^{i-1}}}\,\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right)\\ &= \mathcal{O}\left(\sqrt{\lambda_0\,(M+\xi)}\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right) \end{aligned} \tag{5.27}$$

where the second identity is due the fact that Lemma 5.3.7(b) implies $2^{i-1} \le 2^{k_F-1} \le 2\lambda_0\xi$ for every $i \le k_F$.

We now bound TI_S. Suppose that $k_S > 1$ and observe that the termination criterion $\|\hat{v}_k\| \le \hat{\rho}$ is not satisfied in the first $k_S - 1$ iterations. Since the R.AIPPM is an instance of the GDF, it follows from Proposition 5.3.1 that

$$\hat{\rho}^2 < \min_{j\le k_S-1}\|\hat{v}_j\|^2 \le \theta\left(1+2\sqrt{\tau}\right)^2\frac{[\phi(z_0)-\phi_*]}{\sum_{j=1}^{k_S-1}\lambda_j}. \tag{5.28}$$

Using the fact that Lemma 5.3.7(c) implies $1/\lambda_j \le \xi$ and $\lambda_j \le \lambda_0$ for every $j \ge 1$, we obtain

$$\begin{aligned}\mathrm{TI}_S &= \mathcal{O}\left(\sum_{j=1}^{k_S}\sqrt{\lambda_j M+1}\,\log_1^+\left[C_{\theta,\tau}\lambda_j M\right]\right)\\ &= \mathcal{O}\left(\sum_{j=1}^{k_S}\sqrt{\lambda_j(M+\xi)}\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right)\\ &= \mathcal{O}\left(\sqrt{M+\xi}\left[\sum_{j=1}^{k_S}\sqrt{\lambda_j}\right]\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right)\\ &= \mathcal{O}\left(\sqrt{M+\xi}\left[\sum_{j=1}^{k_S}\frac{\lambda_j}{\sqrt{\lambda_j}}+\sqrt{\lambda_0}\right]\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right). \end{aligned} \tag{5.29}$$

Now, using 5.28, the bound $(a+b)^2 \le 2a^2+2b^2$ for every $a, b \in \mathbb{R}$, and the previous bound $1/\lambda_j \le \xi$ for every $j \ge 1$, it holds that

$$\sum_{j=1}^{k_S-1}\frac{\lambda_j}{\sqrt{\lambda_j}} \le \sqrt{\xi}\sum_{j=1}^{k_S-1}\lambda_j = \mathcal{O}\left(\frac{\sqrt{\xi}\theta\,(1+\tau)\,[\phi(z_0)-\phi_*]}{\hat{\rho}^2}\right). \tag{5.30}$$

Hence, combining (5.29) and (5.30), we conclude that

$$\mathrm{TI}_S = \mathcal{O}\left(\sqrt{M+\xi}\left(\frac{\sqrt{\xi}\theta\left[1+\tau\right]\left[\phi(z_0)-\phi_*\right]}{\hat{\rho}^2}+\sqrt{\lambda_0}\right)\log_1^+\left[C_{\theta,\tau}\lambda_0 M\right]\right). \tag{5.31}$$

It can be easily seen that the bound in (5.31) trivially holds when $k_S \leq 1$ in view of the last term in it. Indeed, to prove this, just assume that $\sum_{j=1}^{k_S-1}\lambda_j = 0$ in the above argument bounding TI_S. Now, since $\mathrm{TI} = \mathrm{TI}_F + \mathrm{TI}_S$, the bound in (5.26) follows by adding (5.27) and (5.31). □

The result below presents the iteration complexity of the R.AIPPM with inputs $(\theta,\tau) = (4,2)$ and $\lambda = 1/\underline{m}$.

Corollary 5.3.9. *The R.AIPPM with inputs $(\theta,\tau) = (4,2)$ and $\lambda = 1/\underline{m}$ outputs a pair $(\hat{z},\hat{v})$ that solves in Problem 5.3.1 in*

$$\mathcal{O}\left(\sqrt{\frac{M}{\underline{m}}+1}\left[\frac{\underline{m}\left[\phi(z_0)-\phi_*\right]}{\hat{\rho}^2}+1\right]\log_1^+\left[\frac{M}{\underline{m}}\right]\right)$$

oracle calls, where ξ is as in (5.25).

Proof. This follows immediately from Theorem 5.3.8 with $(\theta,\tau) = (4,2)$ and $\lambda = 1/\underline{m}$ together with the fact that the R.ACGM uses $\mathcal{O}(1)$ oracle calls at the end of every one of its iterations. □

We now briefly discuss alternative update rules for the stepsize λ_k. To begin, one could consider an update in which the intermediate variable λ in Line 4 of the R.AIPPM is initialized with $\lambda \leftarrow \beta\lambda_{k-1}$ for some $\beta > 1$. For larger values of β, this might result in larger number of inner iterations per outer iteration due to a (possibly) large number of R.ACG calls that result in $\pi_k^{\mathrm{acg}} = \texttt{false}$. A modification of this approach is to fix this multiplier β to be 1 for all iterations following one in which $\pi_k^{\mathrm{acg}} = \texttt{false}$. This modification results in a bitonic stepsize sequence (as opposed to a monotonic one) and is only slightly more con-

servative than the first approach. The second approach will be used in our computational experiments in Section 5.5.

5.4 Relaxed AIP.QP (R.AIP.QP) Method

This section establishes an iteration complexity bound for a relaxed AIP.QPM (R.AIP.QPM) that is generally more efficient in practice than the AIP.QPM in Section 4.1.

Before proceeding, we first recall the main problem of the R.AIP.QPM and its key assumptions. Consider the CNCO problem

$$\hat{\varphi}_* := \min_{z \in \mathcal{Z}} \{\phi(z) := f(z) + h(z) : \mathcal{A}z \in S\} \qquad (\mathcal{CNCO}[a])$$

where $\mathcal{Z}$ is a finite dimensional inner product space and it is assumed that $\phi = f + h$ satisfies assumptions (D1)–(D3) and:

(E1) $\mathcal{A} : \mathcal{Z} \mapsto \mathcal{R}$ is a nonzero linear operator for some finite dimensional inner product space $\mathcal{R}$, the quantity $S \subseteq \mathcal{R}$ is a closed convex set, and $\mathcal{F} := \{z \in \mathcal{Z} : \mathcal{A}z \in S\} \neq \varnothing$;

(E2) Z is compact.

Moreover, like in Chapter 4, it is assumed that efficient oracles for evaluating the quantities $f(z)$, $\nabla f(z)$, $\mathcal{A}z$, and $h(z)$ and for obtaining exact solutions of the subproblems

$$\min_{z \in \mathcal{Z}} \left\{ \lambda h(z) + \frac{1}{2}\|z - z_0\|^2 \right\}, \quad \min_{r \in S} \|r - r_0\|$$

for any $z_0 \in \mathcal{Z}$, $r \in \mathcal{R}$, and $\lambda > 0$, are available.

The R.AIP.QPM considers finding approximate stationary points of 5.4.1 as in Prob-

lem 4.1.1, i.e. given $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, find $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}]) \in [Z \times \mathcal{R}] \times [\mathcal{Z} \times \mathcal{R}]$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + \mathcal{A}^* \hat{p} \quad \|\hat{v}\| \leq \hat{\rho}, \tag{5.32}$$

$$\mathcal{A}\hat{z} + \hat{q} \in S \quad \|\hat{q}\| \leq \hat{\eta}. \tag{5.33}$$

For the sake of future referencing, let us state the problem of finding $(\hat{z}, \hat{v})$ satisfying (5.32) in Problem 5.3.1.

Problem 5.4.1: Find an approximate stationary point of $\mathcal{CNCO}$[a]

Given $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$, find a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}]) \in [Z \times \mathcal{R}] \times [\mathcal{Z} \times \mathcal{R}]$ satisfying conditions (5.32) and (5.33).

5.4.1 Key Properties of the Quadratic Penalty Approach

This subsection presents some key properties of a quadratic penalty function that is used in the R.AIP.QPM. Its properties mirror those in Section 4.1.2.

We first introduce some useful quantities. First, the diameter of Z is denoted by

$$D_z := \sup_{u,z \in Z} \|u - z\|. \tag{5.34}$$

We define the following important quantity for future reference:

$$\hat{\varphi}_c := \inf_{z \in Z} \{\varphi_c(z) := f_c(z) + h(z)\}. \tag{5.35}$$

where $f_c(\cdot)$ is a quadratic penalty function given by

$$f_c(z) := f(z) + \frac{c}{2} \operatorname{dist}^2(\mathcal{A}z, S) \quad \forall z \in Z. \tag{5.36}$$

Note that using Lemma 4.1.1(a) and the definition of $\hat{\varphi}_*$ is as in $\mathcal{CNCO}[a]$, it is easily seen

that

$$\hat{\varphi}_* \geq \hat{\varphi}_{\bar{c}} \geq \hat{\varphi}_c \quad \forall \bar{c} > c \geq 0. \tag{5.37}$$

The following result shows how a solution of Problem 5.3.1 with $f = f_c$ yields a solution of Problem 5.4.1 when the penalty parameter c is sufficiently large.

Proposition 5.4.1. *Given $\hat{\rho} > 0$ and $c > 0$, let $(\hat{z}, \hat{v})$ be a solution of Problem 3.1.1 with $f = f_c$ as in* (5.36)*. Moreover, let $\underline{m}$ be as in* (5.5) *and define the quantities*

$$\begin{aligned} \hat{p} &:= c\left[\mathcal{A}\hat{z} - \Pi_S(\mathcal{A}\hat{z})\right], \quad \hat{q} := \Pi_S(\mathcal{A}\hat{z}) - \mathcal{A}\hat{z}, \\ T_{\hat{\eta}} &:= \left[2(\hat{\varphi}_* - \hat{\varphi}_0 + \hat{\rho}D_z) + \underline{m}D_z^2\right]\hat{\eta}^{-2}, \quad M_c := M + c\|\mathcal{A}\|^2 \end{aligned} \tag{5.38}$$

where $\hat{\varphi}_*$, $\hat{\varphi}_0$, and D_z are as in $\mathcal{CNCO}[a]$, (5.35)*, and* (5.34)*, respectively. Then the following statements hold:*

(a) it holds that $f_c \in \mathcal{C}_{m,M_c}(Z)$;

(b) the pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfies (4.7)*, the inclusion in* (4.8)*, and*

$$\|\hat{q}\|^2 \leq \frac{1}{c}\left(2\left[\hat{\varphi}_* - \varphi(\hat{z}) + \hat{\rho}D_z\right] + \underline{m}D_z^2\right);$$

(c) if $c \geq T_{\hat{\eta}}$, then $\|\hat{q}\| \leq \hat{\eta}$.

Proof. (a) See Lemma 4.1.4.

(b) See Lemma 4.1.2(a) for the proof that the pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ satisfies (4.7), the inclusion in (4.8). To show the desired inequality on $\|\hat{q}\|$, let $p_S(z) = (c/2)\operatorname{dist}^2(\mathcal{A}z, S)$ for every $z \in Z$. Using the inclusion in (4.8), the convexity of p_S, the definition of $\hat{p}$, and Lemma E.2.1(b), it follows that $\hat{v} \in \nabla f(\hat{z}) + \partial\left[h + p_S\right](\hat{z})$, or equivalently,

$$h(u) + p_S(u) \geq h(\hat{z}) + p_S(\hat{z}) + \langle \hat{v} - \nabla f(\hat{z}), u - \hat{z}\rangle \quad \forall u \in Z. \tag{5.39}$$

Considering (5.39) at any $u \in \mathcal{F}$ and using the fact that $p_S(u) = 0$ for any $u \in \mathcal{F}$, the definition of $\underline{m}$ in (5.5), and the definitions of p_S and $\hat{q}$, we conclude that

$$\begin{aligned}\frac{c}{2}\|\hat{q}\|^2 =& \frac{c}{2}\|\Pi_S(\mathcal{A}\hat{z}) - \mathcal{A}\hat{z}\|^2 = p_S(\mathcal{A}\hat{z}) \\ &\leq h(u) - h(\hat{z}) + \langle \nabla f(\hat{z}), u - \hat{z}\rangle - \langle \hat{v}, u - \hat{z}\rangle \\ &\leq (f+h)(u) - (f+h)(\hat{z}) + \|\hat{v}\|\|u - \hat{z}\| + \frac{1}{2}\left(\underline{m}\|u - \hat{z}\|^2\right) \\ &\leq \varphi(u) - \varphi(\hat{z}) + \hat{\rho} D_h + \frac{1}{2}\left(\underline{m} D_h^2\right).\end{aligned}$$

Taking the infimum over $u \in \mathcal{F}$ immediately yields the desired bound.

(c) Suppose $c \geq T_{\hat{\eta}}$. Using the previous bound on c, the fact that $\varphi(\hat{z}) \geq \hat{\varphi}_0$, and the definition of $T_{\hat{\eta}}$, it follows from part (b) that

$$\|\hat{q}\|^2 \leq \frac{1}{c}\left(2\left[\hat{\varphi}_* - \hat{\varphi}_0 + \hat{\rho} D_h\right] + \underline{m} D_h^2\right) = \frac{1}{c}\left[\hat{\eta}^2 T_{\hat{\eta}}\right] \leq \hat{\eta}^2.$$

□

In view of the above proposition, we now outline a static penalty method for solving Problem 5.4.1. First, let $z_0 \in Z$ be given and select a penalty parameter $c = \mathcal{O}(\hat{\eta}^{-2})$ satisfying $c \geq T_{\hat{\eta}}$. Second, obtain a point $(\hat{z}, \hat{v})$ solving Problem 5.3.1 with $f = f_c$ (see (5.36)) using the R.AIPPM of Section 5.3 with z_0, $(m, M) = (m, M_c)$, $(\theta, \tau) = (4, 2)$, and $\lambda = 1/\underline{m}$, where M_c is as in Proposition 5.4.1(b). Finally, compute the pair $(\hat{p}, \hat{q})$ according to (5.38) and output the pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$, which solves Problem 5.4.1 in view of Proposition 5.4.1(a) and (d). Using the fact that $c = \mathcal{O}(\hat{\eta}^{-2})$, and Corollary 5.3.9 with $(f, M) = (f_c, M_c)$, it is easy to see that the inner iteration complexity of the outlined method is

$$\mathcal{O}\left(\sqrt{\frac{M_c}{\underline{m}} + 1}\left[\frac{\underline{m}\left[\varphi_c(z_0) - \hat{\varphi}_c\right]}{\hat{\rho}^2} + 1\right]\log_1^+\left[\frac{M_c}{\underline{m}}\right]\right) = \mathcal{O}\left(\hat{\rho}^{-2}\hat{\eta}^{-3}\log_1^+\hat{\eta}^{-1}\right), \tag{5.40}$$

where the last quantity ignores any constants aside from the tolerances. A drawback of this

static penalty method is that it requires in its first step the selection of a single parameter c, which is generally difficult to obtain. This issue can be circumvented by considering a dynamic cold-started penalty method in which the static penalty method is repeated for a sequence of increasing values of c and common starting point z_0. It can be shown that the resulting cold-started dynamic penalty method has an ACG iteration complexity that is still on the same order as (5.40). Note that the bound (5.40) is actually $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1}\log_1^+\hat{\eta}^{-1})$ when $z_0 \in \mathcal{F}$, but our interest lies in the case where $z_0 \notin \mathcal{F}$ since an initial point $z_0 \in \mathcal{F}$ is generally not known.

The AIP.QPM of Section 4.1 is a modified cold-started dynamic penalty method like the one just outlined, but which replaces each R.AIPP call of the static penalty method with the AIPPM of Section 5.3. It has been shown in Theorem 4.1.6 that its inner iteration complexity bound for solving is $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1})$. This bound is established without assuming that Z is bounded and is clearly better than the one in (5.40).

The next subsection considers a warm-started dynamic penalty method, similar to the one described immediately after Proposition 5.4.1, in which the input z_0 to the R.AIPP call for solving the next penalty subproblem is chosen to be the output $\hat{z}$ from the R.AIPP call for solving the current one. It is shown in Theorem 5.4.3 that its inner iteration complexity is $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1}\log_1^+\hat{\eta}^{-1})$, which is the same as the one for the AIP.QPM up to a logarithmic factor. As a side remark, we note that although a warm-started version of the AIP.QPM in Section 4.1 can be also considered, the aforementioned $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1})$ inner iteration complexity bound was derived for its cold-started version.

5.4.2 Statement and Properties of the R.AIP.QPM

This subsection describes and establishes the iteration complexity of the R.AIP.QPM.

We first state the R.AIP.QPM in Algorithm 5.4.1. Given $(\theta, \tau) \in (2, \infty) \times \mathbb{R}_{++}$ and $z_0 \in Z$, its main idea is to call the R.AIPPM of Section 5.3 to obtain approximate stationary points

for a sequence of penalty subproblems of the form

$$\min_{z\in\mathcal{Z}} \{f_{c_\ell}(z) + h(z)\}$$

where $\{c_\ell\}_{\ell\geq 1}$ is a strictly increasing sequence that tends to infinity. At the end of each R.AIPPM call, a pair $([\hat{z},\hat{p}],[\hat{v},\hat{q}])$ is generated that satisfies (5.32) and the inclusion in (5.33), and the method terminates when the inequality in (5.33) holds.

Algorithm 5.4.1: R.AIP.QP Method

Require: $(\hat{\rho},\hat{\eta}) \in \mathbb{R}^2_{++}$, $M > 0$, $h \in \overline{\text{Conv}}\,(Z)$, $f \in \mathcal{C}_M(Z)$, $\lambda > 0$, $(\theta,\tau) \in (2,\infty)\times\mathbb{R}_{++}$, $z_0 \in Z$, $\mathcal{A} \neq 0$, $S \subseteq \mathcal{R}$;
Initialize: $\mathbb{C}_1 \leftarrow M/\|\mathcal{A}\|^2$, $\hat{z}_0 \leftarrow z_0$;

procedure R.AIP.QP($f, h, \mathcal{A}, S, z_0, c_1, \lambda, m, M, \theta, \tau, \hat{\rho}, \hat{\eta}$)
  for $\ell = 1, \ldots$ **do**
    PART 1 **Attack** the ℓ^{th} prox penalty subproblem.
    $f_{c_\ell} \Leftarrow f + \frac{c_\ell}{2}\cdot \text{dist}^2(\mathcal{A}(\cdot), S)$
    $M_{c_\ell} \leftarrow M + c_\ell\|\mathcal{A}\|^2$
    $(\hat{z}_\ell, \hat{v}_\ell) \leftarrow$ R.AIPP($f_{c_\ell}, h, \hat{z}_{\ell-1}, \lambda, \theta, \tau, M_{c_\ell}, \hat{\rho}$)
    $\hat{p}_\ell \leftarrow c_\ell\,[\mathcal{A}\hat{z}_\ell - \Pi_S(\mathcal{A}\hat{z}_\ell)]$
    $\hat{q}_\ell \leftarrow \Pi_S(\mathcal{A}\hat{z}_\ell) - \mathcal{A}\hat{z}_\ell$
    PART 2 Either **stop** with a nearly feasible point or **increase** c_ℓ.
    if $\|\hat{q}_\ell\| \leq \hat{\eta}$ **then**
      return $([\hat{z}_\ell,\hat{p}_\ell],[\hat{v}_\ell,\hat{q}_\ell])$
    $c_{\ell+1} \leftarrow 2c_\ell$

We now make three comments about the R.AIP.QPM. To ease the discussion, let us refer to the R.AIPP iterations in each R.AIPP call as **outer iterations**, the R.ACG iterations performed inside each R.AIPP call as **inner iterations**, and the iterations over the indices ℓ as **cycles**. First, it follows from Proposition 5.4.1(b) that, for every $\ell \geq 1$, the pair $([\hat{z},\hat{p}],[\hat{v},\hat{q}]) = ([\hat{z}_\ell,\hat{p}_\ell],[\hat{v}_\ell,\hat{q}_\ell])$ satisfies (5.32) and the first inclusion of (5.33). Second, since every cycle of the R.AIP.QPM doubles c, the condition on c in Proposition 5.4.1(c) will

be eventually satisfied. Hence, the residual $\hat{q}$ corresponding to this c will satisfy the condition $\|\hat{q}\| \leq \hat{\eta}$ and the R.AIP.QPM will stop in view of its stopping criterion in Line 10. Finally, in view of the first and second comments, we conclude that the R.AIP.QPM always outputs a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ that solves Problem 5.4.1.

Before deriving the inner iteration complexity of the R.AIP.QPM, we note that the number of inner iterations needed in the $(\ell + 1)^{\text{th}}$ execution of the R.AIPPM depends on the quantity $\varphi_{c_\ell}(\hat{z}_l) - \hat{\varphi}_{c_\ell}$ (see the left-hand-side of (5.40) with $(c, z_0) = (c_\ell, \hat{z}_\ell)$). The result below shows that the warm-start strategy in Line 6 of the method together with the boundedness of Z imply that the aforementioned quantity has an upper bound that is independent of the size of the parameter c_ℓ.

Lemma 5.4.2. *Let c_1 be as in the initialization of the R.AIP.QPM and define*

$$\begin{aligned} S(z_0) &:= \varphi_{c_1}(z_0) - \hat{\varphi}_{c_1}, \\ Q(z_0) &:= S(z_0) + 2\left[\hat{\varphi}_* - \hat{\varphi}_0 + \hat{\rho} D_z + \frac{1}{2}\underline{m} D_z^2\right], \end{aligned} \tag{5.41}$$

where $\hat{\varphi}_*$ and $\hat{\varphi}_0$are as in $\mathcal{CNCO}[a]$ and (5.35)*,, respectively. Then, for every $\ell \geq 1$, we have*

$$\varphi_{c_\ell}(\hat{z}_\ell) - \hat{\varphi}_{c_\ell} \leq Q(\hat{z}_0). \tag{5.42}$$

Proof. All line numbers referenced in this proof are with respect to Algorithm 5.4.1. The case in which $\ell = 1$ follows trivially from the definition of $S(z_0)$. Consider now the case in which $\ell \geq 2$. Remark that Line 12 and Proposition 5.4.1 respectively imply that $c_\ell = 2c_{\ell-1}$ and $(\hat{z}_\ell, \hat{v}_\ell)$ solves Problem 5.3.1 with $f = f_{c_{\ell-1}}$. It now follows from the aforementioned remarks, the last inequality in (5.37) with $c = c_\ell$, the definition of $\hat{q}_\ell$, and Proposition 5.4.1(b)

with $(\hat{z}, c) = (\hat{z}_\ell, c_\ell)$, that

$$\begin{aligned}\varphi_{c_\ell}(\hat{z}_\ell) - \hat{\varphi}_{c_\ell} &\leq \varphi_{c_\ell}(\hat{z}_\ell) - \hat{\varphi}_0 = \varphi(\hat{z}_\ell) + 2\left[\frac{c_{\ell-1}}{2}\|\hat{q}_\ell\|^2\right] - \hat{\varphi}_0 \\ &\leq \varphi(\hat{z}_\ell) + 2\left[\hat{\varphi}_* - \varphi(\hat{z}_\ell) + \hat{\rho}D_z + \frac{1}{2}\underline{m}D_z^2\right] - \hat{\varphi}_0 \\ &= 2\hat{\varphi}_* - \varphi(\hat{z}_\ell) - \hat{\varphi}_0 + 2\left[\hat{\rho}D_z + \frac{1}{2}\underline{m}D_z^2\right] \\ &\leq 2\left[\hat{\varphi}_* - \hat{\varphi}_0 + \hat{\rho}D_z + \frac{1}{2}\underline{m}D_z^2\right] \leq Q(z_0). \end{aligned} \tag{5.43}$$

□

We now establish the iteration complexity of the R.AIP.QPM in the following result.

Theorem 5.4.3. *Let $T_{\hat{\eta}}$ be as in* (5.38) *and define*

$$\Xi_{\hat{\eta}} := M + T_{\hat{\eta}}\|A\|^2 \quad \forall \hat{\eta} > 0. \tag{5.44}$$

Then, the R.AIP.QPM outputs a pair $([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ that solves Problem 5.4.1 in

$$\mathcal{O}\left(\sqrt{\Xi_{\hat{\eta}} + \xi}\left(\frac{\sqrt{\xi}\theta\,[1+\tau]\,Q(z_0)}{\hat{\rho}^2} + \sqrt{\lambda_0}\right)\log_1^+\left(C_{\theta,\tau}\lambda_0\Xi_{\hat{\eta}}\right)\right), \tag{5.45}$$

inner iterations, where $C_{\theta,\tau}$, ξ, and $Q(z_0)$ are as in (5.23), (5.25), *and* (5.41), *respectively.*

Proof. The fact that the output of the R.AIP.QPM solves Problem 5.4.1 is an immediate consequence of Proposition 5.4.1 and the termination condition in Line 10 of the method.

Let us now prove the desired complexity bound. Let $\bar{\ell} \geq 1$ be the smallest index such that $c_{\bar{\ell}} \geq T_{\hat{\eta}}$. Since the R.AIP.QPM calls the R.AIPPM with $(M, f) = (M_{c_\ell}, f_{c_\ell})$ at every cycle, it follows from Lemma 5.4.2 and Theorem 5.3.8, with $M = M_{c_\ell}$, that the total number of inner iterations at the ℓ^{th} cycle of the R.AIP.QPM is on the order of

$$\mathcal{O}\left(\sqrt{\left[1 + \frac{\xi}{M}\right]\Xi_{\hat{\eta}}}\left[\frac{\sqrt{\xi}\theta\,[1+\tau]\,Q(z_0)}{\hat{\rho}^2} + \sqrt{\lambda_0}\right]\log_1^+\left[C_{\theta,\tau}\lambda_0 M_{c_\ell}\right]\right). \tag{5.46}$$

Hence, the R.AIP.QPM method stops in a total number of inner iterations bounded above by the sum of the quantity in (5.46) over $\ell = 1, \ldots, \bar{\ell}$.

We now focus on simplifying some quantities in the aforementioned sum. Using the fact that $M = c_1\|\mathcal{A}\|^2$, we obtain the bound

$$\begin{aligned} M_{c_\ell} &= M + c_\ell\|\mathcal{A}\|^2 = M + 2^{\ell-1}c_1\|\mathcal{A}\|^2 \\ &\leq 2^{\ell-1}\left(M + c_1\|\mathcal{A}\|^2\right) = 2^\ell c_1\|\mathcal{A}\|^2. \end{aligned} \tag{5.47}$$

Now, if $\bar{\ell} = 1$, then the above inequality implies that $M_{c_{\bar{\ell}}} \leq 2c_1\|\mathcal{A}\|^2 = 2M = \mathcal{O}(\Xi_{\hat{\eta}})$. Assume then that $\bar{\ell} \geq 2$. Observe that the definition of $\bar{\ell}$ implies that $2^{\bar{\ell}-1}c_1 = c_{\bar{\ell}} \leq T_{\hat{\eta}}$ or, equivalently, $\sqrt{c_1}\sqrt{2}^{\bar{\ell}} \leq \sqrt{2T_{\hat{\eta}}}$. Combining the previous inequality with (5.47), we conclude that

$$\begin{aligned} &\sum_{k=1}^{\bar{\ell}}\sqrt{M_{c_k} + \xi} \leq \sum_{k=1}^{\bar{\ell}}\sqrt{2^k c_1\|\mathcal{A}\|^2 + \xi} \leq \sqrt{2}^{\bar{\ell}}\left(1 + \sqrt{2}\right)\sqrt{2c_1\|\mathcal{A}\|^2 + \xi} \\ &\leq 8\sqrt{T_{\hat{\eta}}\left(\|\mathcal{A}\|^2 + \frac{\xi}{c_1}\right)} = \sqrt{\|\mathcal{A}\|^2 T_{\hat{\eta}}\left(1 + \frac{\xi}{M}\right)} \\ &= \mathcal{O}\left(\sqrt{\left[1 + \frac{\xi}{M}\right]\Xi_{\hat{\eta}}}\right), \end{aligned} \tag{5.48}$$

and also that

$$\log_1^+\left(M_{c_\ell}\right) \leq \log_1^+\left(2^{\bar{\ell}}c_0\|A\|^2\right) \leq \log_1^+\left(T_{\hat{\eta}}\|A\|^2\right) = \mathcal{O}\left(\log_1^+\Xi_{\hat{\eta}}\right). \tag{5.49}$$

It now follows from (5.46), (5.48), and (5.49) that the R.AIP.QPM stops in a total number of inner bounded by the quantity in (5.45). □

The result below presents the iteration complexity of the R.AIP.QPM with inputs $(\theta, \tau) = (4, 2)$ and $\lambda = 1/\underline{m}$.

Corollary 5.4.4. *The R.AIP.QPM with inputs $(\theta, \tau) = (4, 2)$ and $\lambda = 1/\underline{m}$ outputs a pair*

$([\hat{z}, \hat{p}], [\hat{v}, \hat{q}])$ *that solves in Problem 5.4.1 in*

$$\mathcal{O}\left(\sqrt{\left[1 + \frac{m}{M}\right]\Xi_{\hat{\eta}}}\left[\frac{\sqrt{m}Q(z_0)}{\hat{\rho}^2} + \frac{1}{\sqrt{m}}\right]\log_1^+\left[\frac{\Xi_{\hat{\eta}}}{\underline{m}}\right]\right) \tag{5.50}$$

inner iterations, where $\Xi_{\hat{\eta}}$ *and* $Q(z_0)$ *are as in* (5.44) *and* (5.41)*, respectively.*

Note that in terms of the tolerance pair $(\hat{\rho}, \hat{\eta})$, it is $\mathcal{O}(\hat{\rho}^{-2}\hat{\eta}^{-1}\log_1^+\hat{\eta}^{-1})$, which improves upon the complexity in (5.40) by a $\Theta(\hat{\eta}^{-2})$ multiplicative factor.

We now end this section by discussing how the above R.AIP.QP instance in Corollary 5.4.4 compares to the AIP.QP instance in Corollary 4.1.7. First, recall that the AIP.QPM requires the knowledge of an upper bound m on $\underline{m}$. Under the same assumptions of this section, it can be shown, using the bound $m \leq M$ and Theorem 4.1.6 with $\hat{c} = 0$, that AIP.QPM instance iteration complexity is

$$\begin{aligned}&\mathcal{O}\left(\sqrt{\frac{\Xi_{\hat{\eta}}}{m}}\left[\frac{mQ(z_0)}{\hat{\rho}^2} + \log_1^+\left(\frac{\Xi_{\hat{\eta}}}{m}\right)\right]\right)\\ &= \mathcal{O}\left(\frac{\sqrt{m\Xi_{\hat{\eta}}}Q(z_0)}{\hat{\rho}^2} + \sqrt{\frac{\Xi_{\hat{\eta}}}{m}}\log_1^+\left[\frac{\Xi_{\hat{\eta}}}{m}\right]\right)\end{aligned} \tag{5.51}$$

On the other hand, using the bound $m \leq M$ it can be shown that (5.50) reduces to

$$\begin{aligned}&\mathcal{O}\left(\sqrt{\Xi_{\hat{\eta}}}\left[\frac{\sqrt{\underline{m}}Q(z_0)}{\hat{\rho}^2} + \sqrt{\frac{1}{\underline{m}}}\right]\log_1^+\left[\frac{\Xi_{\hat{\eta}}}{\underline{m}}\right]\right)\\ &= \mathcal{O}\left(\frac{\sqrt{\underline{m}\Xi_{\hat{\eta}}}Q(z_0)}{\hat{\rho}^2}\log_1^+\left[\frac{\Xi_{\hat{\eta}}}{\underline{m}}\right] + \sqrt{\frac{\Xi_{\hat{\eta}}}{\underline{m}}}\log_1^+\left[\frac{\Xi_{\hat{\eta}}}{\underline{m}}\right]\right).\end{aligned} \tag{5.52}$$

Note that (5.52) is as good as (5.51) when $\Xi_{\hat{\eta}}/\underline{m} = \mathcal{O}(1)$ and is only worse by a factor of $\log\hat{\eta}^{-1}$ when $m = \bar{m}$.

5.5 Numerical Experiments

This section presents several numerical experiments that use the various algorithms developed in this chapter. All experiments are run on Linux 64-bit machines each containing Xeon E5520 processors and at least 8 GB of memory using MATLAB 2020a. Supporting code for some of the benchmarked solvers was generously denoted by the original authors Jiaming Liang, Saeed Ghadimi, and Guanghui "George" Lan. It is worth mentioning that the complete code for reproducing the experiments is freely available online[1].

5.5.1 Unconstrained Optimization Problems

This subsection examines the performance of several solvers for finding approximate stationary points of $\mathcal{NCO}$ where (f, h) satisfy assumptions (A1)–(A3) of Chapter 3.

The algorithms benchmarked in this section are as follows.

- **AIPP**: a variant of Algorithm 5.3.3 with starting inputs $\lambda_0 = 1/m$, $\theta = 2$, and $\tau = 10(\lambda_0 M + 1)$. More specifically, this variant adaptively changes the value of τ based on the update rule in [49], uses the bitonic stepsize update rule described at the end of Section 5.3.3, and initializes L_0 for each R.ACGM call as follows: at the k^{th} outer iteration, if L_{-1} denotes either $\lambda_0 M + 1$ for $k = 1$ or the last obtained estimate of L_k from a previous R.ACG call for $k > 1$, then L_0 of the current R.ACG call is set to $L_0 = \lambda_k(L_1 - 1)/[100\lambda_{\max\{k-1,1\}}] + 1$. Moreover, at the k^{th} iterate, it uses the z_{k-1} as the initial starting point for its k^{th} R.ACG call.

- **AG**: an instance of [30, Algorithm 2] in which $L_\Psi = \max\{m, M\}$ and the sequences $\{\alpha_k\}_{k\geq 1}$, $\{\beta_k\}_{k\geq 1}$, and $\{\lambda_k\}_{k\geq 1}$ are as in [30, Corollary 1].

- **NC-FISTA**: an instance of the algorithm in [61, Section] in which, defining $\xi = 1.05m$, we have $A_0 = 2\xi(\xi + m)/(\xi - m)^2$.

[1] See the code in `./tests/thesis/` from the GitHub repository https://github.com/wwkong/nc_opt/

- **UPFAG**: an instance of [31, Algorithm 1] in which $H = \max\{m, M\}$, $\nu = 1$, $(\gamma_1, \gamma_2, \gamma_3) = (0.4, 0.4, 1)$, $\delta = 10^{-3}$, $\hat{\lambda}_0 = H$, $\hat{\beta}_0 = 1/H$, and the line search method the Barzilai-Borwein method given in [31, Equation 2.12] with $\sigma = 10^{-10}$.

Given a tolerance $\hat{\rho} > 0$ and an initial point $z_0 \in Z$, each algorithm above seeks a pair $(\hat{z}, \hat{v}) \in Z \times \mathcal{Z}$ satisfying

$$\hat{v} \in \nabla g(\hat{z}) + \partial h(\hat{z}), \quad \frac{\|\hat{v}\|}{\|\nabla g(z_0)\| + 1} \leq \hat{\rho}. \tag{5.53}$$

Moreover, each algorithm is given a time limit of 4000 seconds. Iteration counts are not reported for instances which were unable to obtain $(\hat{z}, \hat{v})$ as above. The bold numbers in each of the tables in this section highlight the algorithm that performed the most efficiently in terms of iteration count or total runtime.

5.5.1.1 Quadratic Matrix Problem

This sub-subsection presents computational results for the unconstrained quadratic matrix (QM) problem considered in [46]. More specifically, given a pair of dimensions $(l, n) \in \mathbb{N}^2$, scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$, linear operators $\mathcal{B} : \mathbb{S}^n_+ \mapsto \mathbb{R}^n$ and $\mathcal{C} : \mathbb{S}^n_+ \mapsto \mathbb{R}^l$ defined by

$$[\mathcal{B}(z)]_j = \langle B_j, z \rangle_F, \quad [\mathcal{C}(z)]_i = \langle C_i, z \rangle_F,$$

for matrices $\{B_j\}_{j=1}^n, \{C_i\}_{i=1}^l \subseteq \mathbb{R}^{n \times n}$, positive diagonal matrix $D \in \mathbb{R}^{n \times n}$, and vector $d \in \mathbb{R}^l$, we consider the QM problem

$$\begin{aligned} \min_{Z \in \mathbb{R}^{n \times n}} \ & \frac{\alpha_1}{2} \|\mathcal{C}Z - d\|^2 - \frac{\alpha_2}{2} \|D\mathcal{B}Z\|^2 \\ \text{subject to } & Z \in P_n \end{aligned}$$

where $P_n = \{Z \in \mathbb{S}^n_+ : \operatorname{tr} z = 1\}$ denotes the n-dimensional spectraplex.

We now describe the experiment parameters for the instances considered. First, the

dimensions were set to be $(l, n) = (50, 200)$ and only 2.5% of the entries of the submatrices B_j and C_i being nonzero. Second, the entries of B_j, C_i, and d (resp., D) are generated by sampling from the uniform distribution $\mathcal{U}[0, 1]$ (resp., $\mathcal{U}\{1, ..., 1000\}$). Third, the initial starting point is $z_0 = I_n/n$. Fourth, with respect to the termination criterion (5.53), the key problem parameters, for every $Z \in \mathbb{S}^n_+$, are

$$f(Z) = \frac{\alpha_1}{2}\|\mathcal{C}Z - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}Z\|^2, \quad h(z) = \delta_{P_n}(z), \quad \hat{\rho} = 10^{-7}.$$

Finally, each problem instance considered is based on a specific curvature pair $(m, M) \in \mathbb{R}^2_{++}$ for which the scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$ is selected so that $M = \lambda_{\max}(\nabla^2 g)$ and $-m = \lambda_{\min}(\nabla^2 g)$. In Appendix H, we describe how to generate the pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$ under the reasonable assumption that $\mathcal{B}, \mathcal{C}$, and D are nonzero.

The table of iteration counts and total runtimes are given in Table 5.1 and Table 5.2, respectively.

Table 5.1: Iteration counts for QM problems.

(m, M)		**Iteration Count**			
m	M	UPFAG	NC-FISTA	AG	AIPP
10^1	10^3	4766	**1463**	4139	2420
10^1	10^4	7768	**1820**	3439	1851
10^1	10^5	10452	3873	3326	**898**
10^1	10^6	11422	4432	3316	**801**

Table 5.2: Runtimes for QM problems.

(m, M)		**Runtime**			
m	M	UPFAG	NC-FISTA	AG	AIPP
10^1	10^3	242.67	**32.83**	123.54	71.42
10^1	10^4	377.05	**40.57**	102.11	54.86
10^1	10^5	485.79	89.18	102.01	**26.24**
10^1	10^6	499.48	107.1	106.56	**26.37**

5.5.1.2 Support Vector Machine Problem

This sub-subsection presents computational results for the support vector machine (SVM) considered in [31]. More specifically, given a pair of dimensions $(n, k) \in \mathbb{N}^2$, matrix $U \in \mathbb{R}^{n\times k}$, and vector $v \in \{-1, +1\}^n$, this subsection considers the (sigmoidal) SVM problem

$$\min_{z\in\mathbb{R}^n} \frac{1}{k}\sum_{i=1}^{k}[1 - \tanh(v_i \langle u_i, z\rangle)] + \frac{1}{2k}\|z\|^2,$$

where u_i denotes the i^{th} column of U.

We now describe the experiment parameters for the instances considered. First, the entries of U are generated by sampling from the uniform distribution $\mathcal{U}[0, 1]$, with only 5% of the entries being nonzero, and $v = \text{sgn}(U^T x)$ where the entries of x are sampled from the uniform distribution over the k–dimensional ball centered at 0 with radius 50. Second, the initial starting point is $z_0 = 0$. Third, the curvature parameters for each problem instance are $m = M = (4\sqrt{3}\|U\|_F^2)/(9k) + 1/k$. Fourth, with respect to the termination criterion (5.53), the key problem parameters, for every $z \in \mathbb{R}^n$, are

$$f(z) = \frac{1}{k}\sum_{i=1}^{k}[1 - \tanh(v_i \langle u_i, z\rangle)] + \frac{1}{2k}\|z\|^2, \quad h(z) = 0, \quad \hat{\rho} = 10^{-3}.$$

Finally, each problem instance considered is based on a specific dimension pair $(n, k) \in \mathbb{N}^2$.

The table of iteration counts and total runtimes are given in Table 5.1 and Table 5.2, respectively.

Table 5.3: Iteration counts for SVM problems.

(n, k)		**Iteration Count**			
n	k	UPFAG	NC-FISTA	AG	AIPP
1000	500	**80**	3024	782	145
2000	1000	**194**	8360	1191	234
4000	2000	1112	22485	1346	**392**
8000	4000	**327**	-	1646	782

Table 5.4: Runtimes for SVM problems.

(n,k)		**Iteration Count**			
n	k	UPFAG	NC-FISTA	AG	AIPP
1000	500	5.46	71.64	19.11	**5.03**
2000	1000	35.88	570.19	84.85	**21.14**
4000	2000	775.77	3447.60	179.31	**66.26**
8000	4000	**659.85**	4000.00	1286.05	780.07

5.5.2 Function Constrained Optimization Problems

This section examines the performance of several solvers for finding approximate stationary points of $\mathcal{CNCO}$ where (f,h,g,S) satisfy (A1)–(A2) and either (B1)–(B2) or (C1)–(C3) of Chapter 4.

The algorithms benchmarked in this section are as follows.

- **AIP.QP**: a variant of Algorithm 5.4.1 in which the R.AIPPM is replaced with the R.AIPP variant described in Section 5.5.1 and $c_0 = \max\left\{1, \hat{c} + L_f/[B_g^{(1)}]^2\right\}$.
- **AIP.AL**: an variant of Algorithm 4.2.2 in which the parameter inputs for the S.ACGM and the variant are given by

$$c_1 = \max\left\{1, \frac{L_f}{\left[B_g^{(1)}\right]^2}\right\}, \quad \theta = \frac{1}{\sqrt{2}}, \quad \sigma = \min\left\{\frac{\nu}{\sqrt{L_{k-1}^{\psi}}}, \theta\right\},$$
$$\nu = \sqrt{\theta\,(\lambda M + 1)}, \quad \lambda = \frac{1}{2m}, \quad p_0 = 0,$$

and the condition on Δ_k in Line 16 of Algorithm 4.2.2 is replaced by

$$\Delta_k \leq \frac{\lambda(1-\theta^2)\hat{\rho}^2}{4(1+2\nu)^2}.$$

- **AG.QP**: a variant of Algorithm 5.4.1 in which the R.AIPPM is replaced with the AG method described in Section 5.5.1 and $c_0 = \max\left\{1, \hat{c} + L_f/[B_g^{(1)}]^2\right\}$.

- **iALM**: an instance of [58, Algorithm 3] in which

$$\sigma = 5, \quad \beta_0 = \max\left\{1, \frac{L_f}{\left[B_g^{(1)}\right]^2}\right\}, \quad w_0 = 1, \quad \boldsymbol{y}^0 = 0, \quad \gamma_k = \frac{(\log 2)\,\|c(x^1)\|}{(k+1)\,[\log(k+2)]^2},$$

  for every $k \geq 1$, and the starting point given to the k^{th} APG call is set to be $\boldsymbol{x}^{k-1}$, which is the prox center for the k^{th} prox subproblem.

Given a tolerance pair $(\hat{\rho}, \hat{\eta}) \in \mathbb{R}^2_{++}$ and an initial point $z_0 \in Z$, each algorithm in this section seeks a pair $([\hat{z}, \hat{p}], [\hat{p}, \hat{q}]) \in [Z \times \mathcal{R}] \times [\mathcal{Z} \times \mathcal{R}]$ satisfying

$$\begin{gathered} \hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + \nabla g(\hat{z})\hat{p}, \quad g(\hat{z}) + \hat{q} \in S \\ \|\hat{v}\| \leq \hat{\rho}, \quad \|\hat{q}\| \leq \hat{\eta}. \end{gathered} \tag{5.54}$$

For cone-constrained problems, i.e. where S is a closed convex cone $-\mathcal{K}$, the following additional conditions are also required:

$$\langle g(\hat{z}) + \hat{q}, \hat{p} \rangle = 0, \quad \hat{p} \succeq_{\mathcal{K}^+} 0,$$

where $\mathcal{K}^+$ denotes the dual cone of $\mathcal{K}$. Moreover, each algorithm is given a time limit of 4000 seconds. Iteration counts are not reported for instances which were unable to obtain $([\hat{z}, \hat{p}], [\hat{p}, \hat{q}])$ as above. The bold numbers in each of the tables in this section highlight the algorithm that performed the most efficiently in terms of iteration count or total runtime.

It is worth mentioning that for problems where S is a pointed convex cone $-\mathcal{K}$, the iALM method attempts to solve the equivalent problem with equality constraints under an additional slack variable. More specifically, it introduces an additional slack variable s, and

considers the equivalent problem

$$\min_{(z,s)\in\mathcal{Z}\times\mathcal{R}} \{f(z) + h(z) : c(z) + s = 0, s \succeq_{\mathcal{K}} 0\}.$$

5.5.2.1 Linearly-Constrained Quadratic Matrix Problem

This sub-subsection presents computational results for the linearly-constrained quadratic matrix (LC-QM) problem considered in [46]. More specifically, given a pair of dimensions $(l, n) \in \mathbb{N}^2$, scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$, linear operators $\mathcal{A} : \mathbb{S}^n_+ \mapsto \mathbb{R}^l$, $\mathcal{B} : \mathbb{S}^n_+ \mapsto \mathbb{R}^n$, and $\mathcal{C} : \mathbb{S}^n_+ \mapsto \mathbb{R}^l$ defined by

$$[\mathcal{A}Z]_i = \langle A_i, Z\rangle_F, \quad [\mathcal{B}Z]_j = \langle B_j, Z\rangle_F, \quad [\mathcal{C}Z]_i = \langle C_i, Z\rangle_F,$$

for matrices $\{A_i\}_{i=1}^l, \{B_j\}_{j=1}^n, \{C_i\}_{i=1}^l \subseteq \mathbb{R}^{n\times n}$, positive diagonal matrix $D \in \mathbb{R}^{n\times n}$, and vector pair $(b, d) \in \mathbb{R}^l \times \mathbb{R}^l$, we consider the LC-QM problem

$$\begin{aligned} \min_{Z\in\mathbb{R}^{n\times n}} \ & \frac{\alpha_1}{2}\|\mathcal{C}Z - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}Z\|^2 \\ \text{subject to } & \mathcal{A}Z \in \{b\}, \\ & Z \in P_n, \end{aligned}$$

where $P_n = \{Z \in \mathbb{S}^n_+ : \operatorname{tr} Z = 1\}$ denotes the n-dimensional spectraplex.

We now describe the experiment parameters for the instances considered. First, the dimensions were set to be $(\ell, n) = (10, 50)$ and only 1.0% of the entries of the submatrices A_i, B_j, and C_i being nonzero. Second, the entries of A_i, B_j, C_i, b, and d (resp., D) were generated by sampling from the uniform distribution $\mathcal{U}[0, 1]$ (resp., $\mathcal{U}\{1, ..., 1000\}$). Third, the initial starting point z_0 was chosen to be a random point in $\mathbb{S}^n_+$. More specifically, three unit vectors $\nu_1, \nu_2, \nu_3 \in \mathbb{R}^n$ and three scalars $e_1, e_2, e_2 \in \mathbb{R}_+$ are first generated by sampling vectors $\widetilde{\nu}_i \sim \mathcal{U}^n[0, 1]$ and scalars $\widetilde{d}_i \sim \mathcal{U}[0, 1]$ and setting $\nu_i = \widetilde{\nu}_i/\|\widetilde{\nu}_i\|$ and $e_i = \widetilde{e}_i/(\sum_{j=1}^3 \widetilde{e}_i)$

for $i = 1, 2, 3$. The initial iterate for the first subproblem is then set to $z_0 = \sum_{i=1}^{3} e_i \nu_i \nu_i^T$. Fourth, key problem parameters, for every $z \in S_+^n$, are

$$f(Z) = \frac{\alpha_1}{2}\|\mathcal{C}Z - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}Z\|^2, \quad h(Z) = \delta_{P_n}(Z),$$
$$g(Z) = \mathcal{A}(z), \quad S = \{b\}, \quad \hat{\rho} = 10^{-4}, \quad \hat{\eta} = 10^{-4}.$$

Sixth, using the fact that $\|Z\|_F \leq 1$ for every $Z \in P_n$, the constant hyperparameters for the AIP.ALM and iALM are

$$L_g = 0, \quad B_g^{(1)} = \|\mathcal{A}\|, \quad L_j = 0, \quad \rho_j = 0, \quad B_j = \|A_j\|_F.$$

Finally, each problem instance considered is based on a specific curvature pair $(m, M) \in \mathbb{R}_{++}^2$ for which the scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}_{++}^2$ is selected so that $M = \lambda_{\max}(\nabla^2 f)$ and $-m = \lambda_{\min}(\nabla^2 f)$. More specifically, the pair $(\alpha_1, \alpha_2) \in \mathbb{R}_{++}^2$ is generated using the approach in Section 5.5.1.1.

The table of iteration counts and total runtimes are given in Table 5.5 and Table 5.6, respectively.

Table 5.5: Iteration Counts for LC-QM problems.

(m, M)		**Iteration Count**			
m	M	iALM	AIP.QP	AG.QP	AIP.AL
10^1	10^2	65780	2211	6891	**366**
10^1	10^3	34629	1839	6672	**217**
10^1	10^4	54469	1906	6667	**644**
10^1	10^5	136349	**1966**	6667	2175
10^1	10^6	371276	**2222**	6666	13831

Table 5.6: Runtimes for LC-QM problems.

(m, M)		**Runtime**			
m	M	iALM	AIP.QP	AG.QP	AIP.AL
10^1	10^2	407.46	23.71	76.17	**5.02**
10^1	10^3	214.04	19.81	73.39	**2.88**
10^1	10^4	337.36	20.58	72.81	**7.59**
10^1	10^5	971.32	**21.35**	73.82	25.00
10^1	10^6	2493.30	**25.35**	77.11	162.56

5.5.2.2 Sparse Principal Component Analysis

This subsection presents computational results for the sparse principal component analysis (SPCA) problem considered in [34]. More specifically, given an integer k, positive scalar pair $(\nu, b) \in \mathbb{R}^2_{++}$, and matrix $\Sigma \in S^n_+$, we consider the SPCA problem

$$\begin{aligned} \min_{\Pi,\Phi} \ & \langle \Sigma, \Pi \rangle_F + \sum_{i,j=1}^{n} q_\nu(\Phi_{ij}) + \nu \sum_{i,j=1}^{n} |\Phi_{ij}| \\ \text{subject to } & \Pi - \Phi = 0, \\ & (\Pi, \Phi) \in \mathcal{F}^k \times \mathbb{R}^{n \times n}, \end{aligned}$$

where $\mathcal{F}^k = \{z \in S^n_+ : 0 \preceq z \preceq I, \operatorname{tr} M = k\}$ denotes the k–Fantope and q_ν is the min-max concave penalty function given by

$$q_\nu(t) := \begin{cases} -t^2/(2b), & \text{if } |t| \le b\nu, \\ b\nu^2/2 - \nu|t|, & \text{if } |t| > b\nu, \end{cases} \qquad \forall t \in \mathbb{R}.$$

We now describe the experiment parameters for the instances considered. First, the scalar parameters are chosen to be $(\nu, n, k, b) = (100, 100, 1, 0.1)$. Second, the matrix Σ is generated according to an eigenvalue decomposition $\Sigma = P\Lambda P^T$, based on a parameter pair (s, k), where k is as in the problem description and s is a positive integer. In particular, we choose $\Lambda = (100, 1, ..., 1)$, the first column of P to be a sparse vector whose first s entries

are $1/\sqrt{s}$, and the other entries of P to be sampled randomly from the standard Gaussian distribution. Third, the initial starting point is $(\Pi_0, \Phi_0) = (D_k, 0)$ where D_k is a diagonal matrix whose first k entries are 1 and whose remaining entries are 0. Fourth, the curvature parameters for each problem instance are $m = M = 1/b$. Fifth, the key problem parameters, the inputs, for every $(\Pi, \Phi) \in S_+^n \times \mathbb{R}^{n\times n}$, are

$$f(\Pi, \Phi) = \langle \Sigma, \Pi \rangle_F + \sum_{i,j=1}^{n} q_\nu(\Phi_{ij}), \quad h(\Pi, \Phi) = \delta_{\mathcal{F}^k}(\Pi) + \nu \sum_{i,j=1}^{n} |\Phi_{ij}|,$$
$$g(\Pi, \Phi) := \Pi - \Phi, \quad S = \{0\}, \quad \hat{\eta} = 10^{-3}, \quad \hat{\rho} = 10^{-6}.$$

Finally, each problem instance considered is based on a specific choice of s (see the description above).

The table of iteration counts and total runtimes are given in Table 5.7 and Table 5.8, respectively.

Table 5.7: Iteration counts for SPCA problems.

	Iteration Count	
s	AIP.QP	AG.QP
5	**5254**	25871
10	**5328**	27074
15	**5492**	26664

Table 5.8: Runtimes for SPCA problems.

	Runtime	
s	AIP.QP	AG.QP
5	**76.81**	295.78
10	**72.88**	310.87
15	**86.89**	361.03

5.5.2.3 Box-Constrained Matrix Completion

This subsection presents computational results for the box-constrained matrix completion (BC-MC) problem considered in [112]. More specifically, given a dimension pair $(p, q) \in \mathbb{N}^2$, positive scalar triple $(\beta, \mu, \theta) \in \mathbb{R}^3_{++}$, scalar pair $(u, l) \in \mathbb{R}^2$, matrix $A \in \mathbb{R}^{p\times q}$, and indices Ω, we consider the BC-MC problem:

$$\begin{aligned}\min_X \ & \frac{1}{2}\|P_\Omega(X - A)\|^2 + \mu \sum_{i=1}^{\min\{p,q\}} [\kappa(\sigma_i(X)) - \kappa_0\sigma_i(X)] + \mu\kappa_0\|X\|_* \\ \text{s.t. } & l \le X_{ij} \le u \quad \forall (i,j) \in \{1,...,p\} \times \{1,...,q\},\end{aligned}$$

where $\|\cdot\|_*$ denotes the nuclear norm, the function P_Ω is the linear operator that zeros out any entry not in Ω, the function $\sigma_i(X)$ denotes the i^{th} largest singular value of X, and

$$\kappa_0 := \frac{\beta}{\theta}, \quad \kappa(t) := \beta \log\left(1 + \frac{|t|}{\theta}\right) \quad \forall t \in \mathbb{R}.$$

We now describe the experiment parameters for the instances considered. First, the matrix A is the user-movie ratings data matrix of the Jester dataset[2], the index set Ω is the set of nonzero entries in A, the dimension pair is set to be $(p, q) = (24938, 100)$, and the fixed scalar parameters are $(\mu, \theta) = (2, \sqrt{2})$. Second, the initial starting point was chosen to be $X_0 = 0$. Third, the curvature parameters for each problem instance are $m = 2\beta\mu/\theta^2$ and $M = \max\{1, m\}$ and the bounds are set to $(l, u) = (0, 5)$. Fourth, the key problem parameters, for every $X \in \mathbb{R}^{n\times n}$, are

$$f(X) = \frac{1}{2}\|P_\Omega(X - A)\|^2 + \mu \sum_{i=1}^{\min\{p,q\}} [\kappa(\sigma_i(X)) - \kappa_0\sigma_i(X)], \quad h(X) = \mu\kappa_0\|X\|_*,$$

$$g(X) = X, \quad S = \{Z \in \mathbb{R}^{p\times q} : l \le Z_{ij} \le u,\ (i,j) \in \{1,...,p\} \times \{1,...,q\}\},$$

$$\hat{\eta} = 10^{-2}, \quad \hat{\rho} = 10^{-2}.$$

[2]The ratings in the file "jester_dataset_1_1.zip" from `http://eigentaste.berkeley.edu/dataset/`..

Finally, each problem instance considered is based on a specific scalar parameter $\beta > 0$.

The table of iteration counts and total runtimes are given in Table 5.9 and Table 5.10, respectively.

Table 5.9: Iteration counts for BC-MC problems.

	Iteration Count	
β	AIP.QP	AG.QP
1/2	470	**100**
1	447	**100**
2	420	**100**

Table 5.10: Iteration counts for BC-MC problems.

	Runtime	
β	AIP.QP	AG.QP
1/2	509.79	**98.563**
1	466.05	**124.45**
2	486.5	**117.26**

5.5.2.4 Quadratically-Constrained Quadratic Matrix Problem

This subsection presents computational results for the nonconvex quadratically constrained quadratic matrix (QC-QM) problem considered in [48]. More specifically, given a dimension pair $(\ell, n) \in \mathbb{N}^2$, matrices $P, Q, R \in \mathbb{R}^{n\times n}$, and the quantities (α, β), $\mathcal{B}$, $\mathcal{C}$, $\{B_j\}_{j=1}^n$, $\{C_i\}_{i=1}^\ell$, D, d as in Section 5.5.2.1, we consider the QC-QM problem

$$\begin{aligned}
\min_Z \; & -\frac{\alpha}{2}\|D\mathcal{B}(Z)\|^2 + \frac{\beta}{2}\|\mathcal{C}(Z) - d\|^2 \\
\text{s.t. } & \frac{1}{2}(PZ)^*PZ + \frac{1}{2}Q^*QZ + \frac{1}{2}ZQ^*Q \preceq R^*R, \\
& 0 \le \lambda_i(Z) \le \frac{1}{\sqrt{n}}, \qquad\qquad i \in \{1, ..., n\}, \\
& Z \in \mathbb{S}^n_+,
\end{aligned}$$

where $\lambda_i(Z)$ denotes the i^{th} largest eigenvalue of Z and the constraint $M \preceq 0$ is equivalent to $-M \in \mathbb{S}^n_+$.

We now describe the experiment parameters for the instances considered. First, the dimensions are set to $(\ell, n) = (10, 50)$. Second, the quantities $\mathcal{B}$, $\mathcal{C}$, D, and d were generated in the same way as Section 5.5.2.1. On the other hand, the matrix R is set to I/n and the entries of matrices P and Q are sampled from the uniform distributions $\mathcal{U}[0, 1/\sqrt{n}]$ and $\mathcal{U}[0, 1/n]$, respectively. Third, the initial starting point z_0 is set to the zero matrix. Fourth, the key problem parameters, for every $Z \in \mathbb{R}^{n\times n}$ are

$$f(Z) = -\frac{\alpha_1}{2}\|D\mathcal{B}(Z)\|^2 + \frac{\alpha_2}{2}\|\mathcal{C}(Z) - d\|^2, \quad h(Z) = \delta_S(z),$$
$$g(Z) = \frac{1}{2}(PZ)^*PZ + \frac{1}{2}Q^*QZ + \frac{1}{2}ZQ^*Q \preceq R^*R,$$
$$\mathcal{K} = \mathbb{S}^n_+, \quad \hat{\rho} = 10^{-2}, \quad \hat{\eta} = 10^{-2},$$

where $S = \{Z \in \mathbb{S}^n_+ : 0 \le \lambda_i(Z) \le 1/\sqrt{n}, i = 1, ..., n\}$. Fifth, using the fact that $\|Z\|_F \le 1$ for every $Z \in S$, the constant hyperparameters for the iALM and AIP.AL are

$$L_g = \|P\|_F^2, \quad B_g^{(1)} = \frac{1}{2}\|P\|_F^2 + \|Q\|_F^2,$$
$$L_{ij} = \left|[P^*P]_{ij}\right|, \quad \rho_{ij} = 0, \quad B_j = \frac{1}{2}\left|[P^*P]_{ij}\right| + \left|[Q^*Q]_{ij}\right|,$$

for $1 \le i, j \le n$. Finally, each problem instance considered was based on a specific curvature pair (m, M) for which the scalar pair (α_1, α_2) is selected so that $M = \lambda_{\max}(\nabla^2 f)$ and $-m = \lambda_{\min}(\nabla^2 f)$. More specifically, the pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$ is generated using the approach in Section 5.5.1.1.

The table of iteration counts and total runtimes are given in Table 5.11 and Table 5.12, respectively.

Table 5.11: Iteration counts for QC-QM problems.

(m, M)		**Iteration Count**	
m	M	iALM	AIP.ALM
10^1	10^2	**2127**	2373
10^1	10^3	4196	**283**
10^1	10^4	10075	**1130**
10^1	10^5	21428	**5657**

Table 5.12: Runtimes for QC-QM problems.

(m, M)		**Runtime**	
m	M	iALM	AIP.ALM
10^1	10^2	**21.46**	42.24
10^1	10^3	41.60	**4.53**
10^1	10^4	97.28	**18.61**
10^1	10^5	216.33	**88.40**

5.5.3 Discussion of the Results

We see that the methods in this chapter are competitive against other modern solvers and that they especially perform well when the curvature ratio M/m is large. Additionally, we see that each method scales well across problem dimensions and parameters. Comparing the AIP.QPM with the AIP.ALM, in particular, we see that the former is scales better across different curvature ratios whereas the latter performs substantially better on some problem instances than others.

We conjecture that the efficiency of these efficient methods is attributed to three facts: (i) the use of efficient ACGM subroutines which adaptively choose the sequence of stepsizes; (ii) the implementation of several termination criteria that allow certain methods to stop early; and (iii) the relaxation of certain convex proximal subproblems to nonconvex ones (which is generally known to improve convergence).

5.6 Conclusion and Additional Comments

In this chapter, we presented several implementation strategies of the methods in previous chapters. More specifically, we presented practical variants of the CRP in Algorithm 3.2.2, the ACGM in Algorithm 2.2.2, the AIPPM in Algorithm 3.3.2, and the AIP.QPM in Algorithm 4.1.1. For the iterative methods in particular, we devised new schemes in which the "stepsize" parameter is chosen in a relaxed and adaptive manner. Additionally, for the AIP.QPM variant, we showed how using a warm-start strategy between penalty prox subproblems made substantial improvements to the derived complexity (compared to using a simple cold-start strategy). Finally, numerical experiments were given to validate the efficacy of our implementation strategies

Additional Comments

We now make several comments about the results in this chapter.

Similar to how the R.AIPPM (resp. R.AIP.QPM) of Section 5.3 (resp. Section 5.4) is a relaxation of the AIPPM of Section 3.3 (resp. AIP.QPM of Section 4.1) that uses an efficient R.ACGM (resp. R.AIPPM) to solve its key subproblems, one could also consider similarly relaxed versions of methods in prior chapters. We briefly describe some of these relaxations. First, recall that the AIPP.SM in Section 6.3 uses a single AIPP call to obtain an approximate stationary point as in Problem 5.3.1. Hence, one could consider a relaxation of the AIPP.SM in which the single AIPP call is replaced by an R.AIPP call. Second, recall that the AIP.QP.SM in Section 6.4 uses a single AIP.QP call to obtain an approximate stationary point as in (5.4.1). Hence, similar to the first relaxation, on could consider a relaxation of the AIPP.QPM in which the single AIP.QP call is replaced by an R.AIP.QP call.

Observing the arguments used in the proofs of Proposition 5.4.1, Lemma 5.4.2, and Theorem 5.4.3, it is straightforward to see that the assumption of Z being bounded can be relaxed to assuming that the iterates $\{\hat{z}_\ell\}_{\ell\geq 1}$ generated by R.AIP.QPM be bounded. Explicitly

assuming that the iterates satisfy $\|\hat{z}_\ell\| \leq B$, for every $\ell \geq 1$ and some $B > 0$, the resulting oracle complexity of R.AIP.QPM method is (5.45) with $Q(z_0)$ replaced by the quantity

$$\varphi_{c_1}(z_0) - \hat{\varphi}_{c_1} + \hat{\varphi}_* - \hat{\varphi}_0 + \hat{\rho}\left[d_0 + B\right] + \underline{m}\left[d_0^2 + B^2\right],$$

where $d_0 := \inf\{\|u - \hat{z}_0\| : z \in \mathcal{F}\}$ and the quantity $\underline{m}$ is as in (5.5).

Note that the description of the R.AIPPM (resp. R.AIP.QPM) does not actually require knowledge of an upper bound m on the parameter $\underline{m}$ in (5.5). This is in contrast to the AIPPM (resp. AIP.QPM) method of Chapter 3 (resp. Chapter 4), which requires m in order to establish its validity and iteration complexity. In addition, one could consider an R.AIPPM and AIP.QPM variant in which the quantity M is adaptively inferred from its iterates rather than requiring knowledge of its value beforehand. While for the sake of brevity we omit the formal description and analysis of such a variant in this thesis, we conjecture that the iteration complexity of the R.AIPPM (resp. R.AIP.QPM) variant is as in (5.26) (resp. (5.45)) with M replaced with a quantity that lower bounds it, e.g. the maximum of the lower estimates of M which are inferred by the generated iterates.

Future Work

A future avenue of research is to investigate whether the iteration complexity of R.AIP.QPM can still be established when Z is unbounded.

CHAPTER 6
NONCONVEX-CONCAVE MIN-MAX COMPOSITE OPTIMIZATION

Smoothing methods are a broad class of optimization algorithms that consider applying an smooth optimization method to a smooth approximation of a nonsmooth optimization problem. An important class of optimization problems that have particularly benefited from the use of smoothing methods is the class of convex-concave min-max problems of the form $\min_{x\in\mathbb{R}^n} \max_{y\in\mathbb{R}^m} \Phi(x,y)$. In particular, several works [85, 107, 113] consider smoothing the nonsmooth primal function $p(x) = \max_{y\in\mathbb{R}^m} \Phi(x,y)$ and applying an efficient solver to the resulting smooth problem under a careful choice of the smoothing parameter.

Our main goal in this chapter is to describe and establish the iteration complexity of an accelerated **inexact** proximal point **smoothing** (AIPP.S) method for finding approximate stationary points of the **nonconvex**-concave min-max composite optimization (MCO) problem

$$\hat{p}_* := \min_{x\in X} \{\hat{p}(x) := p(x) + h(x)\} \qquad (\mathcal{MCO})$$

where X is a nonempty convex set, $h \in \overline{\text{Conv}}\,(X)$, and p is a max function given by

$$p(x) := \max_{y\in Y} \Phi(x,y) \quad \forall x \in X, \tag{6.1}$$

for some nonempty compact convex set Y and function Φ which, for some scalar $m > 0$ and open set $\Omega \supseteq X$, is such that: (i) Φ is continuous on $\Omega \times Y$; (ii) the function $-\Phi(x,\cdot) \in \overline{\text{Conv}}\,(Y)$ for every $x \in X$; and (iii) for every $y \in Y$, the function $\Phi(\cdot,y) + m\|\cdot\|^2$ is convex, differentiable, and its gradient is Lipschitz continuous on $X\times Y$. Here, the objective function is the sum of a convex function h and the pointwise supremum of differentiable functions which is generally a nonsmooth function.

When Y is a singleton, the max term in $\mathcal{MCO}$ becomes smooth and $\mathcal{MCO}$ reduces to the smooth NCO problem in Chapter 3 which may be solved by the AIPPM in Section 3.3.

When Y is not a singleton, $\mathcal{MCO}$ can no longer be directly solved by the AIPPM due to the nonsmoothness of the max term. The AIPP.S method (AIPP.SM) developed in this chapter is instead based on a perturbed version of $\mathcal{MCO}$ in which the max term in $\mathcal{MCO}$ is replaced by a smooth approximation and the resulting smooth NCO problem is solved by the aforementioned AIPPM.

Throughout our presentation, it is assumed that efficient oracles for evaluating the quantities $\Phi(x,y)$, $\nabla_x\Phi(x,y)$, and $h(x)$ and for obtaining exact solutions of the problems

$$\min_{x\in X}\left\{\lambda h(x)+\frac{1}{2}\|x-x_0\|^2\right\},\quad \max_{y\in Y}\left\{\lambda\Phi(x_0,y)-\frac{1}{2}\|y-y_0\|^2\right\} \tag{6.2}$$

for any (x_0,y_0) and $\lambda>0$, are available. Throughout this chapter, the terminology **oracle call** is used to refer to a collection of the above oracles of size $\mathcal{O}(1)$ where each of them appears at least once.

We first develop an instance of the AIPP.SM that obtains a stationary point based on a primal-dual formulation of $\mathcal{MCO}$. More specifically, given a tolerance pair $(\rho_x,\rho_y)\in\mathbb{R}^2_{++}$, it is shown that an instance of this scheme obtains a pair $([\bar{x},\bar{y}],[\bar{u},\bar{v}])$ such that

$$\begin{pmatrix}\bar{u}\\ \bar{v}\end{pmatrix}\in\begin{pmatrix}\nabla_x\Phi(\bar{x},\bar{y})\\ 0\end{pmatrix}+\begin{pmatrix}\partial h(\bar{x})\\ \partial\left[-\Phi(\bar{x},\cdot)\right](\bar{y})\end{pmatrix},\quad \|\bar{u}\|\le\rho_x,\quad \|\bar{v}\|\le\rho_y \tag{6.3}$$

in $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ oracle calls. We then show that another instance of the AIPP.SM can obtain an approximate stationary point based on the directional derivative of $\hat{p}$. More specifically, given a tolerance pair $\delta>0$, it is shown that this instance computes a point $x\in X$ such that

$$\exists\hat{x}\in X \text{ s.t. } \inf_{\|d\|\le 1}\hat{p}'(\hat{x};d)\ge-\delta,\quad \|\hat{x}-x\|\le\delta, \tag{6.4}$$

in $\mathcal{O}(\delta^{-3})$ oracle calls.

A secondary goal of this chapter is to develop an accelerated inexact proximal quadratic

penalty smoothing (AIP.QP.S) method to obtain approximate stationary points of a linearly constrained version of $\mathcal{MCO}$, namely the min-max constrained composite optimization (MCCO) problem

$$\min_{x\in X}\{p(x)+h(x):\mathcal{A}x=b\} \tag{$\mathcal{MCCO}$}$$

where p is as in (6.1), $\mathcal{A}$ is a linear operator, and b is in the range of $\mathcal{A}$. Similar to the approach used for the AIPP.SM, the AIP.QP.S method (AIP.QP.SM) considers a perturbed variant of $\mathcal{MCCO}$ in which the objective function is replaced by a smooth approximation and the resulting CNCO problem is solved by the AIP.QPM in Section 4.1. For a given tolerance triple $(\rho_x,\rho_y,\eta)\in\mathbb{R}^3_{++}$, it is shown that the method computes a pair $([\bar{x},\bar{y},\bar{r}],[\bar{u},\bar{v}])$ satisfying

$$\begin{gathered}\begin{pmatrix}\bar{u}\\ \bar{v}\end{pmatrix}\in\begin{pmatrix}\nabla_x\Phi(\bar{x},\bar{y})+\mathcal{A}^*\bar{r}\\ 0\end{pmatrix}+\begin{pmatrix}\partial h(\bar{x})\\ \partial\left[-\Phi(\bar{x},\cdot)\right](\bar{y})\end{pmatrix},\\ \|\bar{u}\|\le\rho_x,\quad \|\bar{v}\|\le\rho_y,\quad \|\mathcal{A}\bar{x}-b\|\le\eta.\end{gathered} \tag{6.5}$$

in $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2}+\rho_x^{-2}\eta^{-1})$ oracle calls.

It is worth mentioning that all the above complexities are obtained under the mild assumption that the optimal values of the optimization problems $\mathcal{MCO}$ and $\mathcal{MCCO}$ are bounded below. Moreover, it is neither assumed that X be bounded nor that $\mathcal{MCO}$ or $\mathcal{MCCO}$ have an optimal solution.

The content of this chapter is based on paper [49] (joint work with Renato D.C. Monteiro) and several passages may be taken verbatim from it.

Related Works

Since the case when $\Phi(\cdot,\cdot)$ is convex-concave has been well-studied in the literature (see, for example, [1, 37, 45, 82, 85, 98]), we will make no more mention of it here. Instead, we will focus on papers that consider the case where $\Phi(\cdot,y)$ is differentiable and nonconvex for

every $y \in Y$ and there are mild conditions on $\Phi(x, \cdot)$ for every $x \in X$.

Denoting $\rho = \min\{\rho_x, \rho_y\}$, D_x to be the diameter of x, and C to be a general closed convex set, we present Tables 6.1 and 6.2, which compare our contributions to past [87, 96] and subsequent [63, 89, 106] works. It is worth mentioning that the above works consider termination conditions that are slightly different from the ones in this chapter. In Section 6.1, we show that they are equivalent to the ones in this chapter up to a multiplicative constant that is independent of the tolerances, i.e. ρ_x, ρ_y, δ.

Table 6.1: Comparison of iteration complexities and possible use cases under notions equivalent to (6.3) with $\rho := \min\{\rho_x, \rho_y\}$.

Algorithm	Oracle Complexity	Use Cases			
		$D_x = \infty$	$h \equiv 0$	$h \equiv \delta_C$	$h \in \overline{\text{Conv}}\ X$
PGSF [87]	$\mathcal{O}(\rho^{-3})$	✗	✓	✓	✗
Minimax-PPA [63]	$\mathcal{O}(\rho^{-2.5}\log^2(\rho^{-1}))$	✗	✓	✓	✗
FNE Search [89]	$\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2}\log(\rho^{-1}))$	✓	✓	✓	✗
AIPP.S	$\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$	✓	✓	✓	✓

Table 6.2: Comparison of iteration complexities and possible use cases under notions equivalent to (6.4).

Algorithm	Oracle Complexity	Use Cases			
		$D_x = \infty$	$h \equiv 0$	$h \equiv \delta_C$	$h \in \overline{\text{Conv}}\ X$
PG-SVRG [96]	$\mathcal{O}(\delta^{-6}\log\delta^{-1})$	✗	✓	✓	✓
Minimax-PPA [63]	$\mathcal{O}(\delta^{-3}\log^2(\delta^{-1}))$	✗	✓	✓	✗
Prox-DIAG [106]	$\mathcal{O}(\delta^{-3}\log^2(\delta^{-1}))$	✓	✓	✗	✗
AIPP.S	$\mathcal{O}(\delta^{-3})$	✓	✓	✓	✓

To the best of our knowledge, this chapter and [49] are the first works to analyze the complexity of a smoothing scheme for finding approximate stationary points as in (6.5).

Organization

This chapter contains six sections. The first one gives some preliminary references and discusses our notion of an approximate stationary point given in (6.3) and (6.4). The second one presents properties of a smooth approximation to the primal function p in (6.1). The

third one presents the AIPP.SM and its iteration complexity. The fourth one presents the AIP.QP.SM and its iteration complexity. The fifth one presents some numerical experiments. The last one gives a conclusion and some closing comments.

6.1 Preliminaries

This section describes the assumptions and four notions of stationary points for problem $\mathcal{MCO}$. It is worth mentioning that the complexities of the smoothing method of this chapter are presented with respect to two of these notions. In order to understand how these results can be translated to the other two alternative notions, which have been used in a few papers dealing with problem $\mathcal{MCO}$, we also present a few results discussing some useful relations between all these notions.

Throughout our presentation, we let $\mathcal{X}$ and $\mathcal{Y}$ be finite dimensional inner product spaces. We also make the following assumptions on problem $\mathcal{MCO}$:

(F1) $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$ are nonempty convex sets, and Y is also compact;

(F2) there exists an open set $\Omega \supseteq X$ such that $\Phi(\cdot,\cdot)$ is finite and continuous on $\Omega \times Y$; moreover, $\nabla_x \Phi(x,y)$ exists and is continuous at every $(x,y) \in \Omega \times Y$;

(F3) $h \in \overline{\mathrm{Conv}}(X)$ and $-\Phi(x,\cdot) \in \overline{\mathrm{Conv}}(Y)$ for every $x \in \Omega$;

(F4) there exist scalars $(L_x, L_y) \in \mathbb{R}^2_{++}$, and $m \in (0, L_x]$ such that

$$\Phi(x,y) - [\Phi(x',y) + \langle \nabla_x \Phi(x',y), x - x' \rangle] \geq -\frac{m}{2}\|x - x'\|^2, \tag{6.6}$$

$$\|\nabla_x \Phi(x,y) - \nabla_x \Phi(x',y')\| \leq L_x \|x - x'\| + L_y \|y - y'\|, \tag{6.7}$$

for every $x, x' \in X$ and $y, y' \in Y$;

(F5) $\hat{p}_* > -\infty$;

We make three remarks about the above assumptions. First, it is well-known that condition (6.7) implies that

$$\Phi(x', y) - [\Phi(x, y) + \langle \nabla_x \Phi(x, y), x' - x \rangle] \leq \frac{L_x}{2} \|x' - x\|^2, \tag{6.8}$$

for every $(x', x, y) \in X \times X \times Y$. Third, the weak convexity condition in (F4) implies that, for any $y \in Y$, the function $\Phi(\cdot, y) + m\|\cdot\|^2/2$ is convex, and hence $p + m\|\cdot\|^2/2$ is as well. Note that while $\hat{p}$ is generally nonconvex and nonsmooth, it has the nice property that $\hat{p} + m\|\cdot\|^2/2$ is convex.

We now discuss two stationarity conditions of $\mathcal{MCO}$ under assumptions (F1)–(F3). First, denoting

$$\hat{\Phi}(x, y) := \Phi(x, y) + h(x) \quad \forall (x, y) \in X \times Y, \tag{6.9}$$

it is well-known that problem $\mathcal{MCO}$ is related to the saddle-point problem which consists of finding a pair $(x^*, y^*) \in X \times Y$ such that

$$\hat{\Phi}(x^*, y) \leq \hat{\Phi}(x^*, y^*) \leq \hat{\Phi}(x, y^*), \tag{6.10}$$

for every $(x, y) \in X \times Y$. More specifically, (x^*, y^*) satisfies (6.10) if and only if x^* is an optimal solution of $\mathcal{MCO}$, y^* is an optimal solution of the dual of $\mathcal{MCO}$, and there is no duality gap between the two problems. Using the composite structure described above for $\hat{\Phi}$, it can be shown that a necessary condition for (6.10) to hold is that (x^*, y^*) satisfy the stationarity condition

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \nabla_x \Phi(x^*, y^*) \\ 0 \end{pmatrix} + \begin{pmatrix} \partial h(x^*) \\ \partial \left[-\Phi(x^*, \cdot) \right](y^*) \end{pmatrix}. \tag{6.11}$$

When $m = 0$, the above condition also becomes sufficient for (6.10) to hold. Second, it can be shown that $p'(x^*; d)$ is well-defined for every $d \in \mathcal{X}$ and that a necessary condition for

$x^* \in X$ to be a local minimum of $\mathcal{MCO}$ is that it satisfies the stationarity condition

$$\inf_{\|d\| \le 1} \hat{p}'(x^*; d) \ge 0. \tag{6.12}$$

When $m = 0$, the above condition also becomes sufficient for x^* to be a global minimum of $\mathcal{MCO}$. Moreover, in view of Lemma F.2.1 in Appendix F.2 with $(\bar{u}, \bar{v}, \bar{x}, \bar{y}) = (0, 0, x^*, y^*)$, it follows that x^* satisfies (6.12) if and only if there exists $y^* \in Y$ such that (x^*, y^*) satisfies (6.11).

Note that finding points that satisfy (6.11) or (6.12) exactly is generally a difficult task. Hence, in this section and the next one, we only consider approximate versions of (6.11) or (6.12), which are (6.3) and (6.4), respectively. For ease of future reference, we say that:

(i) a pair $([\bar{x}, \bar{y}], [\bar{u}, \bar{v}])$ is a (ρ_x, ρ_y)-**primal-dual stationary point** of $\mathcal{MCO}$ if it satisfies (6.3);

(ii) a point $\hat{x}$ is a δ-**directional stationary point** of $\mathcal{MCO}$ if it satisfies the first inequality in (6.4).

It is worth mentioning that (6.4) is generally hard to verify for a given point $x \in X$. This is primarily because the definition requires us to check an infinite number of directional derivatives for a (potentially) nonsmooth function at points $\hat{x}$ near $\bar{x}$. In contrast, the definition of an approximate primal-dual stationary point is generally easier to verify because the quantities $\|\bar{u}\|$ and $\|\bar{v}\|$ can be measured directly, and the inclusions in (6.3) are easy to verify when the prox oracles for h and $\Phi(x, \cdot)$, for every $x \in X$, are readily available.

The next result, whose proof is given in Appendix F.2, shows that a (ρ_x, ρ_y)-primal-dual stationary point, for small enough ρ_x and ρ_y, yields a point x satisfying (6.4). Its statement makes use of the diameter of Y defined as

$$D_y := \inf_{y, y' \in Y} \|y - y'\|. \tag{6.13}$$

Proposition 6.1.1. *If the pair* $([\bar{x}, \bar{y}], [\bar{u}, \bar{v}])$ *is a* (ρ_x, ρ_y)*-primal-dual stationary point of* $\mathcal{MCO}$*, then there exists a point* $\hat{x} \in X$ *such that*

$$\inf_{\|d\| \le 1} \hat{p}'(\hat{x}; d) \ge -\rho_x - 2\sqrt{2mD_y\rho_y}, \quad \|\bar{x} - \hat{x}\| \le \sqrt{\frac{2D_y\rho_y}{m}}.$$

The iteration complexities in this chapter (see Section 6.3) are stated with respect to the two notions of stationary points (6.3) and (6.4). However, it is worth discussing below two other notions of stationary points that are common in the literature as well as some results that relate all four notions.

Given $(\lambda, \varepsilon) \in \mathbb{R}^2_{++}$, a point x is said to be a (λ, ε)-prox stationary point of $\mathcal{MCO}$ if the function $\hat{p} + \|\cdot\|^2/(2\lambda)$ is strongly convex and

$$\frac{1}{\lambda}\|x - x_\lambda\| \le \varepsilon, \quad x_\lambda = \operatorname*{argmin}_{u \in \mathcal{X}} \left\{ \hat{P}_\lambda(u) := \hat{p}(u) + \frac{1}{2\lambda}\|u - x\|^2 \right\}. \tag{6.14}$$

The above notion is considered, for example, in [63, 96, 106]. The result below, whose proof is given in Appendix F.2, shows how it is related to (6.4).

Proposition 6.1.2. *For any given* $\lambda \in (0, 1/m)$*, the following statements hold:*

(a) for any $\varepsilon > 0$*, if* $x \in X$ *satisfies* (6.4) *and*

$$0 < \delta \le \frac{\lambda^3 \varepsilon}{\lambda^2 + 2(1 - \lambda m)(1 + \lambda)}, \tag{6.15}$$

then x *is a* (λ, ε)*-prox stationary point;*

(b) for any $\delta > 0$*, if* $x \in X$ *is a* (λ, ε)*-prox stationary point for some* $\varepsilon \le \delta \cdot \min\{1, 1/\lambda\}$*, then* x *satisfies* (6.4) *with* $\hat{x} = x_\lambda$*, where* x_λ *is as in* (6.14)*.*

Note that for a fixed $\lambda \in (0, 1/m)$ such that $\max\{\lambda^{-1}, (1 - \lambda m)^{-1}\} = \mathcal{O}(1)$, the largest δ in part (a) is $\mathcal{O}(\varepsilon)$. Similarly, for part (b), if $\lambda^{-1} = \mathcal{O}(1)$ then largest ε in part (b) is $\mathcal{O}(\delta)$.

Combining these two observations, it follows that (6.14) and (6.4) are equivalent (up to a multiplicative factor) under the assumption that $\delta = \Theta(\varepsilon)$.

Given $(\rho_x, \rho_y) \in \mathbb{R}^2_{++}$, a pair $(\bar{x}, \bar{y})$ is said to be a (ρ_x, ρ_y)-first-order Nash equilibrium point of $\mathcal{MCO}$ if

$$\inf_{\|d_x\| \le 1} \mathcal{S}'_{\bar{y}}(\bar{x}; d_x) \ge -\rho_x, \qquad \sup_{\|d_y\| \le 1} \mathcal{S}'_{\bar{x}}(\bar{y}; d_y) \le \rho_y, \tag{6.16}$$

where $\mathcal{S}_{\bar{y}} := \Phi(\cdot, \bar{y}) + h(\cdot)$ and $\mathcal{S}_{\bar{x}} := \Phi(\bar{x}, \cdot)$. The above notion is considered, for example, in [63, 87, 89]. The next result, whose proof is given in Appendix F.2, shows that (6.16) is equivalent to (6.3).

Proposition 6.1.3. *A pair $(\bar{x}, \bar{y})$ is a (ρ_x, ρ_y)-first-order Nash equilibrium point if and only if there exists $(\bar{u}, \bar{v}) \in \mathcal{X} \times \mathcal{Y}$ such that $([\bar{x}, \bar{y}], [\bar{u}, \bar{v}])$ is a (ρ_x, ρ_y)-primal-dual stationary point.*

We now briefly discuss some approaches for finding approximate stationary points of $\mathcal{MCO}$. One approach is to apply a proximal descent type method directly to problem $\mathcal{MCO}$, but this would lead to subproblems with nonsmooth convex composite functions. A second approach is based on first applying a smoothing method to $\mathcal{MCO}$ and then using a prox-convexifying descent method such as the AIPPM in Section 3.3 to solve the perturbed unconstrained smooth problem. An advantage of the second approach, which is the one pursued in this chapter, is that it generates subproblems with smooth convex composite objective functions. The next subsection describes one possible way to smooth the (generally) nonsmooth function p in $\mathcal{MCO}$.

Before ending this section, we formally state the problem of finding primal-dual and directional stationary points in Problem 6.1.1 and Problem 6.1.2, respectively .

Problem 6.1.1: Find an approximate primal-dual stationary point of $\mathcal{MCO}$

Given $(\rho_x, \rho_y) \in \mathbb{R}^2_{++}$, find a pair $([\bar{x}, \bar{y}], [\bar{u}, \bar{v}]) \in [X \times Y] \times [\mathcal{X} \times \mathcal{Y}]$ satisfying condition (6.3).

Problem 6.1.2: Find an approximate directional stationary point of $\mathcal{MCO}$

Given $\delta > 0$, find a point $x \in X$ satisfying condition (6.4).

6.2 Smooth Approximation

This subsection presents a smooth approximation of the function p in $\mathcal{MCO}$.

For every $\xi > 0$, consider the smoothed function p_ξ defined by

$$p_\xi(x) := \max_{y \in Y} \left\{ \Phi_\xi(x, y) := \Phi(x, y) - \frac{1}{2\xi} \|y - y_0\|^2 \right\} \quad \forall x \in X, \tag{6.17}$$

for some $y_0 \in Y$. The following proposition presents the key properties of p_ξ and its related quantities.

Proposition 6.2.1. *Let $\xi > 0$ be given and assume that the function Φ satisfies conditions (F1)–(F4). Let $p_\xi(\cdot)$ and $\Phi_\xi(\cdot, \cdot)$ be as defined in* (6.17) *and define*

$$y_\xi(x) := \operatorname*{argmax}_{y' \in Y} \Phi_\xi(x, y') \quad \forall x \in X. \tag{6.18}$$

Then, the following properties hold:

(a) $y_\xi(\cdot)$ is Q_ξ-Lipschitz continuous on X where

$$Q_\xi := \xi L_y + \sqrt{\xi(L_x + m)}; \tag{6.19}$$

(b) $p_\xi(\cdot)$ is continuously differentiable on X and $\nabla p_\xi(x) = \nabla_x \Phi(x, y_\xi(x))$ for every $x \in X$;

(c) $\nabla p_\xi(\cdot)$ is L_ξ-Lipschitz continuous on X where

$$L_\xi := L_y Q_\xi + L_x \leq \left(L_y \sqrt{\xi} + \sqrt{L_x} \right)^2; \tag{6.20}$$

(d) for every $x, x' \in X$, we have

$$p_\xi(x) - [p_\xi(x') + \langle \nabla p_\xi(x'), x - x' \rangle] \geq -\frac{m}{2}\|x - x'\|^2; \tag{6.21}$$

Proof. The inequality in (6.20) follows from (a), the fact that $m \leq L_x$, and the bound

$$L_\xi = L_y \left[\xi L_y + \sqrt{\xi(L_x + m)} \right] + L_x \leq \xi L_y^2 + 2\sqrt{\xi L_x} + L_x = \left(L_y \sqrt{\xi} + \sqrt{L_x} \right)^2.$$

The other conclusions of (a)–(c) follow from Lemma E.3.1 and Proposition E.3.2 in Appendix E.3 with $(\Psi, q, y) = (\Phi_\xi, p_\xi, y_\xi)$. We now show that the conclusion of (d) is true. Indeed, if we consider (6.6) at $(y, x') = (y_\xi(x'), x')$, the definition of Φ_ξ, and use the definition of ∇p_ξ in (b), then

$$\begin{aligned}
&-\frac{m}{2}\|x - x'\|^2 \leq \Phi(x', y_\xi(x)) - [\Phi(x, y_\xi(x)) + \langle \nabla_x \Phi(x, y_\xi(x)), x' - x \rangle] \\
&= \Phi_\xi(x', y_\xi(x)) - [p_\xi(x) + \langle \nabla p_\xi(x), x' - x \rangle] \\
&\leq p_\xi(x') - [p_\xi(x) + \langle \nabla p_\xi(x), x' - x \rangle],
\end{aligned}$$

where the last inequality follows from the optimality of y. □

We now make two remarks about the above properties. First, the Lipschitz constants of y_ξ and ∇p_ξ depend on the value of ξ while the weak convexity constant m in (6.21) does not. Second, as $\xi \to \infty$, it holds that $p_\xi \to p$ pointwise and $Q_\xi, L_\xi \to \infty$. These remarks are made more precise in the next result.

Lemma 6.2.2. *For every $\xi > 0$, it holds that*

$$-\infty < p(x) - \frac{D_y^2}{2\xi} \leq p_\xi(x) \leq p(x) \quad \forall x \in X,$$

where D_y is as in (6.13).

Proof. The fact that $p(x) > -\infty$ follows immediately from assumption (F5). To show the other bounds, observe that for every $y_0 \in Y$, we have

$$\Phi(x,y) + h(x) \geq \Phi(x,y) - \frac{1}{2\xi}\|y - y_0\|^2 + h(x) \geq \Phi(x,y) - \frac{D_y^2}{2\xi} + h(x)$$

for every $(x,y) \in X \times Y$. Taking the supremum of the bounds over $y \in Y$ and using the definitions of p and p_ξ yields the remaining bounds. □

6.3 Accelerated Inexact Proximal Point Smoothing (AIPP.S) Method

This section presents the AIPP.SM for finding stationary points of $\mathcal{MCO}$ as in (6.3) and (6.4).

We first state the AIPP.SM in Algorithm 6.3.1. Given $(x_0, y_0) \in X \times Y$, its main idea is to apply an instance of the AIPPM in Section 3.3 to the NCO problem

$$\min_{x \in X} \{\hat{p}_\xi(x) := p_\xi(x) + h(x)\}, \tag{6.22}$$

where p_ξ is as in (6.17) and ξ is a positive scalar that will depend on the tolerances in (6.3) and (6.4). It is stated in an incomplete manner in the sense that it does not specify how the parameter ξ and the tolerance ρ used in its AIPPM call are chosen. Two invocations of this method, with different choices of ξ and ρ, are considered in Propositions 6.3.2 and 6.3.3, which describe the iteration complexities for finding approximate stationary points as in (6.3) and (6.4), respectively.

Algorithm 6.3.1: AIPP.S Method

Require: $\rho > 0$, $\xi > 0$, $(m, L_x, L_y) \in \mathbb{R}^3_+$, $h \in \overline{\text{Conv}}\,(Z)$, Φ satisfying $(F2)$ – $-(F4)$, $(x_0, y_0) \in X \times Y$;
Initialize: $\lambda \leftarrow 1/(2m)$, $\sigma \leftarrow 1/2$

1: **procedure** AIPP.S($\Phi, h, x_0, y_0, m, L_x, L_y, \rho$)
2: $p_\xi \Leftarrow \max_{y \in Y} \Phi_\xi(\cdot, y)$ ▷ See (6.17).

3: $L_\xi \leftarrow L_y\left[\xi L_y + \sqrt{\xi(L_x+m)}\right] + L_x$

4: $(x,u) \leftarrow \text{AIPP}(p_\xi, h, x_0, \lambda, m, L_\xi, \sigma, \rho)$

5: **return** (x,u)

We now give four remarks about the above method. First, it follows from Corollary 3.3.6 that the AIPPM invoked in Line 4 stops and outputs a pair (x,u) satisfying

$$u \in \nabla p_\xi(x) + \partial h(x), \quad \|u\| \le \rho. \tag{6.23}$$

Second, since the AIPP.SM is a one-pass method (as opposed to an iterative method), the complexity of the AIPP.SM is essentially that of the AIPPM. Third, similar to the smoothing scheme of [85] which assumes $m = 0$, the AIPP.SM is also a smoothing scheme for the case in which $m > 0$. On the other hand, in contrast to the algorithm of [85] which uses an ACG variant, the AIPP.SM invokes the AIPPM to solve (6.22) due to its nonconvexity. Finally, while the AIPPM in Line 4 is called with $(\sigma, \lambda) = (1/2, 1/(2m))$, it can also be called with any $\sigma \in (0,1)$ and $\lambda \in (0, 1/m)$ to establish the desired termination.

For the remainder of this subsection, our goal will be to show that a careful selection of the parameter ξ and the tolerance ρ will allow the AIPP.SM to generate approximate stationary points as in (6.4) and (6.3).

We first recall the quantity $R_\lambda\psi(z_0)$ in (3.10) of Chapter 3. The result below presents a bound on $R_\lambda\hat{p}_\xi(x_0)$ in terms of the data in problem $\mathcal{MCO}$.

Lemma 6.3.1. *For every $\xi > 0$ and $\lambda \ge 0$, it holds that*

$$R_\lambda\hat{p}_\xi(x_0) \le R_\lambda\hat{p}(x_0) + \frac{\lambda D_y^2}{2\xi}, \tag{6.24}$$

where $R_\lambda\psi(\cdot)$ and D_y are as in (3.10) *and* (6.13)*, respectively.*

Proof. Using Lemma 6.2.2 and the definitions of $\hat{p}$ and $\hat{p}_\xi$, it holds that

$$\hat{p}_\xi(x) - \inf_{x'} \hat{p}_\xi(x') \le \hat{p}(x) - \inf_{x'} \hat{p}(x') + \frac{D_y^2}{2\xi}, \quad \forall x \in X. \tag{6.25}$$

Multiplying the above expression by $(1-\sigma)\lambda$ and adding the quantity $\|x_0 - x\|^2/2$ yields the inequality

$$\begin{aligned} &\frac{1}{2}\|x_0 - x\|^2 + (1-\sigma)\lambda\left[\hat{p}_\xi(x) - \inf_{x'} \hat{p}_\xi(x')\right] \\ &\le \frac{1}{2}\|x_0 - x\|^2 + (1-\sigma)\lambda\left[\hat{p}(x) - \inf_{\tilde{x}} \hat{p}(x')\right] + (1-\sigma)\frac{\lambda D_y^2}{2\xi} \quad \forall x \in X, \end{aligned} \tag{6.26}$$

Taking the infimum of the above expression, and using the definition of $R_\lambda\psi(\cdot)$ in (3.10) yields the desired conclusion. □

We now show how the AIPP.SM generates a (ρ_x, ρ_y)-primal-dual stationary point, i.e. a pair that solves Problem 6.1.1.

Proposition 6.3.2. *For a given tolerance pair $(\rho_x, \rho_y) \in \mathbb{R}^2_{++}$, let (x, u) be the pair output by the AIPP.SM with input parameter ξ and tolerance ρ satisfying*

$$\xi \ge \frac{D_y}{\rho_y}, \quad \rho = \rho_x. \tag{6.27}$$

Moreover, define

$$(\bar{u}, \bar{v}) := \left(u, \frac{y_0 - y_\xi(x)}{\xi}\right), \quad (\bar{x}, \bar{y}) := (x, y_\xi(x)), \tag{6.28}$$

where y_ξ is as in (6.18)*. Then, the following statements hold:*

(a) the AIPP.SM performs

$$\mathcal{O}\left(\Omega_\xi\left[\frac{m^2 R_{1/(2m)}\hat{p}(x_0)}{\rho_x^2} + \frac{mD_y^2}{\xi\rho_x^2} + \log_1^+(\Omega_\xi)\right]\right) \tag{6.29}$$

oracle calls, where $R_\lambda\psi(\cdot)$ *and* D_y *are as in* (3.10) *and* (6.13)*, respectively,* $\log_1^+(\cdot) := \max\{1, \log(\cdot)\}$*, and*

$$\Omega_\xi := 1 + \frac{\sqrt{\xi}L_y + \sqrt{L_x}}{\sqrt{m}}; \tag{6.30}$$

(b) the pair $([\bar{x}, \bar{y}], [\bar{u}, \bar{v}])$ *is a* (ρ_x, ρ_y)*-primal-dual stationary point of* $\mathcal{MCO}$*, and hence, solves Problem 6.1.1.*

Proof. (a) Using the inequality in (6.20), it holds that

$$\sqrt{\frac{L_\xi}{4m} + 1} \le 1 + \sqrt{\frac{L_\xi}{4m}} \le 1 + \frac{\sqrt{\xi}L_y + \sqrt{L_x}}{2\sqrt{m}} = \Theta(\Omega_\xi). \tag{6.31}$$

Moreover, using Corollary 3.3.6 with $(\phi, M) = (\hat{p}_\xi, L_\xi)$, Lemma 6.3.1, and bound (6.31), it follows that the number of oracle calls performed by the AIPP.SM is on the order given by (6.29).

(b) It follows from the definitions of p_ξ, tolerance ρ, and $(\bar{y}, \bar{u})$ in (6.17), (6.27), and (6.28), respectively, Proposition 6.2.1(b), and the inclusion in (6.23) that $\|\bar{u}\| \le \rho_x$ and

$$\bar{u} \in \nabla p_\xi(\bar{x}) + \partial h(\bar{x}) = \nabla_x \Phi(\bar{x}, y_\xi(\bar{x})) + \partial h(\bar{x}) = \nabla_x \Phi(\bar{x}, \bar{y}) + \partial h(\bar{x}).$$

Hence, we conclude that the top inclusion and the upper bound on $\|\bar{u}\|$ in (6.3) hold. Next, the optimality condition of $\bar{y} = y_\xi(\bar{x})$ as a solution to (6.17) and the definition of $\bar{v}$ in (6.17) give

$$0 \in \partial\left[-\Phi(\bar{x}, \cdot)\right](\bar{y}) + \frac{\bar{y} - y_0}{\xi} = \partial\left[-\Phi(\bar{x}, \cdot)\right](\bar{y}) - \bar{v} \tag{6.32}$$

Moreover, the definition of ξ implies that

$$\|\bar{v}\| = \frac{\|\bar{y} - y_0\|}{\xi} \le \frac{D_y}{D_y/\rho_y} = \rho_y. \tag{6.33}$$

Hence, combining (6.32) and (6.33), we conclude that the bottom inclusion and the upper bound on $\|\bar{v}\|$ in (6.3) hold. □

We now make two remarks about Proposition 6.3.2. First, under the assumption that (6.27) is satisfied as equality, the complexity of AIPP.SM reduces to

$$\mathcal{O}\left(m^{3/2}R_{1/(2m)}\hat{p}(x_0)\left[\frac{L_x^{1/2}}{\rho_x^2}+\frac{L_yD_y^{1/2}}{\rho_x^2\rho_y^{1/2}}\right]\right) \tag{6.34}$$

under the reasonable assumption that the $\mathcal{O}(\rho_x^{-2}+\rho_x^{-2}\rho_y^{-1/2})$ term in (6.29) dominates the other terms. Second, recall from the last remark following the previous proposition that when Y is a singleton, $\mathcal{MCO}$ becomes a special instance of $\mathcal{NCO}$ and the AIPP.SM becomes equivalent to the AIPPM of Section 3.3. It then follows that the complexity in (6.34) reduces to

$$\mathcal{O}\left(\frac{m^{3/2}L_x^{1/2}R_{1/(2m)}\hat{p}(x_0)}{\rho_x^2}\right) \tag{6.35}$$

and, in view of this remark, the $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ term in (6.34) is attributed to the (possible) nonsmoothness in $\mathcal{MCO}$.

We next show how the AIPP.SM generates a point that is *near* a δ-directional stationary point, i.e. a point that solves Problem 6.1.2.

Proposition 6.3.3. *Let a tolerance pair $\delta > 0$ be given and consider the AIPP.SM with input parameter ξ and tolerance ρ satisfying*

$$\xi \geq \frac{D_y}{\tau}, \quad \rho = \frac{\delta}{2}, \quad \tau \leq \min\left\{\frac{m\delta^2}{2D_y}, \frac{\delta^2}{32mD_y}\right\}. \tag{6.36}$$

Then, the following statements hold:

(a) the AIPP.SM performs

$$\mathcal{O}\left(\Omega_\xi\left[\frac{R_{1/(2m)}\hat{p}(x_0)}{\lambda^2\delta^2}+\frac{D_y^2}{\lambda\xi\delta^2}+\log_1^+(\Omega_\xi)\right]\right) \tag{6.37}$$

oracle calls where Ω_ξ, $R_\lambda\psi(\cdot)$, and D_y are as in (6.30), (3.10), *and* (6.13), *respectively, and* $\log_1^+(\cdot) := \max\{1, \log(\cdot)\}$,;

(b) the first argument x in the pair output by the AIPP.SM satisfies (6.4)*, and hence, solves Problem 6.1.2.*

Proof. (a) Using Proposition 6.3.2 with $(\rho_x, \rho_y) = (\delta/2, \tau)$ and the bound on τ in (6.36) it follows that the AIPP.SM stops in a number of oracle calls bounded above by (6.37).

(b) Let (x, u) be the pair generated by the AIPPM with ξ and $\bar{\rho}$ satisfying (6.36). Defining $(\bar{v}, \bar{y})$ as in (6.28), it follows from Proposition 6.3.2 with $(\rho_x, \rho_y) = (\delta/2, \tau)$ that $(u, \bar{v}, x, \bar{y})$ is a $(\delta/2, \tau)$-primal-dual stationary point of $\mathcal{MCO}$. As a consequence, it follows from Proposition 6.1.1 with $(\rho_x, \rho_y) = (\delta/2, \tau)$ that there exists a point $\hat{x}$ satisfying

$$\|\hat{x} - x\| \leq \sqrt{\frac{2D_y\tau}{m}}, \quad \inf_{\|d\|\leq 1} \hat{p}'(\hat{x}; d) \geq -\frac{\delta}{2} - 2\sqrt{2mD_y\tau}. \tag{6.38}$$

Combining the above bounds with the bound on τ in (6.36) yields the desired conclusion in view of (6.4). □

We now give three remarks about the above result. Second, Proposition 6.3.3(b) states that, while x not a stationary point itself, it is near a δ-directional stationary point $\hat{x}$. Second, under the assumption that (6.36) is satisfied as equality, the complexity of the AIPP.SM reduces to

$$\mathcal{O}\left(m^{3/2} R_{1/(2m)} \hat{p}(x_0) \left[\frac{L_x^{1/2}}{\delta^2} + \frac{L_y D_y}{\delta^3}\right]\right) \tag{6.39}$$

under the reasonable assumption that the $\mathcal{O}(\delta^{-2} + \delta^{-3})$ term in (6.37) dominates the other $\mathcal{O}(\delta^{-1})$ terms. Fourth, when Y is a singleton, it is easy to see that $\mathcal{MCO}$ becomes a special instance of $\mathcal{NCO}$, the AIPP.SM becomes equivalent to the AIPPM of Section 3.3, and the complexity in (6.39) reduces to

$$\mathcal{O}\left(\frac{m^{3/2} L_x^{1/2} R_{1/(2m)} \hat{p}(x_0)}{\delta^2}\right). \tag{6.40}$$

In view of the last remark, the $\mathcal{O}(\delta^{-3})$ term in (6.39) is attributed to the (possible) nonsmoothness in $\mathcal{MCO}$.

6.4 Accelerated Inexact Proximal Quadratic Penalty Smoothing (AIP.QP.S) Method

This section presents the AIP.QP.SM for finding stationary points of $\mathcal{MCCO}$ as in (6.5).

Since the AIP.QP.SM applies the AIP.QPM of Section 4.1 to a relaxation of $\mathcal{MCCO}$, we assume that (Φ, h, X, Y) satisfies assumptions (F1)–(F4) of Section 6.3 as well as the following ones:

(G1) $\mathcal{A} : \mathcal{X} \mapsto \mathcal{R}$ is a nonzero linear operator, b is in the range of $\mathcal{A}$, and the feasible region $\mathcal{F} := \{x \in \mathcal{X} : \mathcal{A}x = b\}$ is nonempty;

(G2) there exists $\hat{c} \geq 0$ such that $\hat{\varphi}_{\hat{c}} > -\infty$, where

$$\hat{\varphi}_{c,\xi} := \inf_{z \in \mathcal{Z}} \left\{ \varphi_{c,\xi}(z) := p_\xi(x) + \frac{c}{2}\|\mathcal{A}x - b\|^2 + h(z) \right\}, \quad \forall c \geq 0, \tag{6.41}$$

where $p_\xi(\cdot)$ is as in (6.17).

For ease of referencing, we also state the problem of finding a pair satisfying (6.5) in Problem 6.4.1.

Problem 6.4.1: Find an approximate primal-dual stationary point of $\mathcal{MCCO}$

Given $(\rho_x, \rho_y, \eta) \in \mathbb{R}^2_{++}$, find a pair $([\bar{x}, \bar{y}, \bar{r}], [\bar{u}, \bar{v}]) \in [X \times Y \times \mathcal{R}] \times [\mathcal{X} \times \mathcal{Y}]$ satisfying condition (6.5).

We now state the AIP.QP.SM in Algorithm 6.4.1. Given $(x_0, y_0) \in X \times Y$ and $\hat{c} > 0$, its main idea is to its main idea is to apply an instance of the AIP.QPM in Section 4.1 to the CNCO problem

$$\min_{x \in X} \{\hat{p}_\xi(x) := p_\xi(x) + h(x) : \mathcal{A}x = b\}, \tag{6.42}$$

where p_ξ is as in (6.17) and ξ is a positive scalar that will depend on the tolerances in (6.5). The resulting output of this AIP.QP call is then similarly transformed like the AIPP.SM of Section 6.3 to obtain a pair that solves Problem 6.4.1.

Algorithm 6.4.1: AIP.QP.S Method

Require: $(\rho_x, \rho_y, \eta) \in \mathbb{R}^3_{++}$, $\xi > D_y/\rho_y$, $(m, L_x, L_y) \in \mathbb{R}^3_+$, $h \in \overline{\text{Conv}}\,(Z)$, Φ as in (F2)–(F4), $\hat{c} > 0$, $(x_0, y_0) \in X \times Y$;
Initialize: $\lambda \leftarrow 1/(2m)$, $\sigma \leftarrow 1/2$

1: **procedure** AIP.QP.S($\Phi, h, x_0, y_0, m, L_x, L_y, \rho$)
2: $y_\xi \Leftarrow \operatorname{argmax}_{y \in Y} \Phi_\xi(\cdot, y)$ ▷ See (6.17).
3: $p_\xi \Leftarrow \max_{y \in Y} \Phi_\xi(\cdot, y)$ ▷ See (6.17).
4: $L_\xi \leftarrow L_y \left[\xi L_y + \sqrt{\xi(L_x + m)}\right] + L_x$
5: $([\bar{x}, \bar{r}], [\bar{u}, \bar{q}]) \leftarrow$ AIP.QP$(p_\xi, h, \mathcal{A}, \{b\}, x_0, \hat{c}, \lambda, m, L_\xi, \sigma, \rho_y, \eta)$
6: $\bar{y} \leftarrow y_\xi(\bar{x})$
7: $\bar{v} \leftarrow \dfrac{y_0 - y_\xi(x)}{\xi}$
8: **return** $([\bar{x}, \bar{y}, \bar{r}], [\bar{u}, \bar{v}])$.

We give two remarks about the AIP.QP.SM. First, it follows from Corollary 4.1.7 that the AIP.QPM invoked in Line 5 stops and outputs a pair $([\bar{x}, \bar{r}], [\bar{u}, \bar{q}])$ satisfying

$$\bar{u} \in \nabla p_\xi(\bar{x}) + \partial h(\bar{x}) + A^* \bar{r}, \quad \|\bar{u}\| \le \rho_x, \quad \|\mathcal{A}\bar{x} - b\| \le \eta.$$

Second, since it is a one-pass algorithm (as opposed to an iterative algorithm), the complexity of the AIP.QP.SM is essentially that of the AIP.QPM.

We now show how the AIP.QP.SM generates a point $([\bar{x}, \bar{y}, \bar{r}], [\bar{u}, \bar{v}])$ satisfying (6.5).

Proposition 6.4.1. *Let a tolerance triple $(\rho_x, \rho_x, \eta) \in \mathbb{R}^3_{++}$ be given and let $([\bar{x}, \bar{y}, \bar{r}], [\bar{u}, \bar{v}])$ be the output obtained by the QP-AIPP.SM. Then, the following properties hold:*

(a) the AIP.QP.SM performs

$$\mathcal{O}\left(\Omega_{\xi,\eta}\left[\frac{m^2 R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)}{\rho_x^2}+\frac{mD_y^2}{\xi\rho_x^2}+\log_1^+\left(\Omega_{\xi,\eta}\right)\right]\right) \tag{6.43}$$

oracle calls, where

$$\begin{aligned}\varphi_{\hat{c}} &:= \hat{p}(x)+\frac{c}{2}\|\mathcal{A}x-b\|^2,\\ \Omega_{\xi,\eta} &:= \Omega_\xi+\left(R_{\hat{c}}(\hat{p};1/(4m))+\frac{D_y^2}{m\xi}\right)^{1/2}\frac{\|\mathcal{A}\|}{\eta},\end{aligned} \tag{6.44}$$

and Ω_ξ, $R^{\mathcal{F}}_\lambda\psi(\cdot)$, and D_y are as in (6.30), (4.10), and (6.13), respectively;

(b) the pair $([\bar{x},\bar{y},\bar{r}],[\bar{u},\bar{v}])$ solves Problem 6.4.1.

Proof. (a) Let Θ_η be as in (4.17) with $M = L_\xi$. Using the same arguments as in Lemma 6.3.1, it is easy to see that

$$R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c},\xi}(x_0) \le R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)+\frac{D_y^2}{8m\xi}. \tag{6.45}$$

where $\varphi_{\hat{c},\xi}$ is as in (6.41). Hence, using (6.31) and (6.45), we have

$$\begin{aligned}\sqrt{\frac{\Theta_\eta}{4m}+1} &\le 1+\sqrt{\frac{L_\xi}{4m}}+\sqrt{\frac{4R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c},\xi}(x_0)\|\mathcal{A}\|^2}{\eta^2}}\\ &\le 1+\frac{\sqrt{\xi}L_y+\sqrt{L_x}}{2\sqrt{m}}+2\left(R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)+\frac{D_y^2}{8m\xi}\right)^{1/2}\frac{\|\mathcal{A}\|}{\eta}\\ &= \Theta(\Omega_{\xi,\eta}).\end{aligned} \tag{6.46}$$

The complexity in (6.43) now follows from Corollary 4.1.7 with $(\phi,f,M) = (p,p_\xi,L_\xi)$, (6.46), and (6.45).

(b) The top inclusion and bounds involving $\|\bar{u}\|$ and $\|\mathcal{A}\bar{x}-b\|$ in (6.5) follow from Proposition 6.2.1(b), the definition of $\bar{y}$ in Line 6 of the method, and Corollary 4.1.7 with $f = p_\xi$. The bottom inclusion and bound involving $\|\bar{v}\|$ follow from similar arguments given for

Proposition 6.3.2(b). □

We now make two remarks about the above complexity bound. First, under the assumption that $\xi = D_y/\rho_y$, the complexity of the AIP.QP.SM reduces to

$$\mathcal{O}\left(m^{3/2}R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)\left[\frac{L_x^{1/2}}{\rho_x^2}+\frac{L_yD_y^{1/2}}{\rho_y^{1/2}\rho_x^2}+\frac{m^{1/2}\|\mathcal{A}\|}{\eta\rho_x^2}\sqrt{R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)}\right]\right), \tag{6.47}$$

under the reasonable assumption that the $\mathcal{O}(\rho_x^{-2}+\eta^{-1}\rho_x^{-2}+\rho_y^{-1/2}\rho_x^{-2})$ term in (6.43) dominates the other terms. Third, when Y is a singleton, it is easy to see that $\mathcal{MCCO}$ becomes a special instance of the CNCO problem $\mathcal{CNCO}$, the AIP.QP.SM of this subsection becomes equivalent to the AIP.QPM of Section 4.1, and the complexity in (6.47) reduces to

$$\mathcal{O}\left(m^{3/2}R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)\left[\frac{L_x^{1/2}}{\rho_x^2}+\frac{m^{1/2}\|\mathcal{A}\|}{\eta\rho_x^2}\sqrt{R^{\mathcal{F}}_{1/(2m)}\varphi_{\hat{c}}(x_0)}\right]\right). \tag{6.48}$$

In view of the last remark, the $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ term in (6.47) is attributed to the (possible) nonsmoothness in $\mathcal{MCCO}$.

Let us now conclude this section with a remark about the penalty subproblem

$$\min_{x\in X}\left\{p_\xi(x)+h(x)+\frac{c}{2}\|\mathcal{A}x-b\|^2\right\}, \tag{6.49}$$

which is what the AIPPM considers every time it is called in the AIP.QPM (see Line 5 of the AIP.QP.SM). First, observe that $\mathcal{MCCO}$ can be equivalently reformulated as

$$\min_{x\in X}\max_{y\in Y, r\in\mathcal{U}}\left[\Psi(x,y,r):=\Phi(x,y)+h(x)+\langle r,\mathcal{A}x-b\rangle\right]. \tag{6.50}$$

Second, it is straightforward to verify that problem (6.49) is equivalent to

$$\min_{x\in X}\left\{\hat{p}_{c,\xi}(x):=p_{c,\xi}(x)+h(x)\right\}, \tag{6.51}$$

where the function $p_{c,\xi} : X \mapsto \mathbb{R}$ is given by

$$p_{c,\xi}(x) := \max_{y \in Y, r \in \mathcal{U}} \left\{ \Psi(x, y, r) - \frac{1}{2c}\|r\|^2 - \frac{1}{2\xi}\|y - y_0\|^2 \right\} \quad \forall x \in X \tag{6.52}$$

with Ψ as in (6.50). As a consequence, problem (6.51) is similar to (6.22) in that a smooth approximate is used in place of the nonsmooth component of the underlying saddle function Ψ. On the other hand, observe that we cannot directly apply the smoothing scheme developed in Section 6.3 to (6.51) as the set $\mathcal{U}$ is generally unbounded. One approach that avoids this problem is to invoke the AIPPM of Section 3.3 to solve a sequence subproblems of the form in (6.51) for increasing values of c. However, in view of the equivalence of (6.49) and (6.51), this is exactly the approach taken by the AIP.QP.SM of this section.

6.5 Numerical Experiments

This section examines the performance of several solvers for finding approximate stationary points of $\mathcal{MCO}$ where (X, Y, Φ, h) satisfy assumptions (F1)–(F5) of Chapter 6. Each problem is chosen so that the computation of the function y_ξ in (6.18) is easy, and the justification for the curvature constants in this section, e.g. m, L_x, and L_y, can be found in Appendix I. All experiments are run on Linux 64-bit machines each containing Xeon E5520 processors and at least 8 GB of memory using MATLAB 2020a. It is worth mentioning that the complete code for reproducing the experiments is freely available online[1].

The algorithms benchmarked in this section are as follows.

- **PGSF**: a variant of [87, Algorithm 2] in which the input parameters are as in [87, Theorem 4.2] and which explicitly evaluates the argmax function $\alpha^*(\cdot)$ in [87, Section 4] instead of applying an ACG variant to estimate its evaluation.

- **AG.S**: an instance of Algorithm 6.3.1 in which the AIPPM is replaced by the AG

[1] See the code in `./tests/thesis/` from the GitHub repository https://github.com/wwkong/nc_opt/

method in Section 5.5.1.

- **AIPP.S**: an instance of Algorithm 6.3.1 in which the AIPPM is replaced by the R.AIPP variant in Section 5.5.1.

Given a tolerance pair $(\rho_x, \rho_y) \in \mathbb{R}^2_{++}$ and an initial point $(x_0, y_0) \in X \times Y$, each algorithm in this section seeks a pair $([\bar{x}, \bar{y}], [\bar{u}, \bar{v}]) \in [X \times Y] \times [\mathcal{X} \times \mathcal{Y}]$ satisfying

$$\begin{gathered} \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix} \in \begin{pmatrix} \nabla_x \Phi(\bar{x}, \bar{y}) \\ 0 \end{pmatrix} + \begin{pmatrix} \partial h(\bar{x}) \\ \partial \left[-\Phi(\bar{x}, \cdot) \right](\bar{y}) \end{pmatrix}, \\ \frac{\|\bar{u}\|}{\|\nabla p_\xi(z_0)\| + 1} \le \rho_x, \quad \|\bar{v}\| \le \rho_y, \end{gathered} \tag{6.53}$$

is obtained, where $\xi = D_y/\rho_y$ and p_ξ is as in (6.17). Moreover, each algorithm is given a time limit of 4000 seconds. Iteration counts are not reported for instances which were unable to obtain $([\hat{x}, \hat{y}], [\hat{u}, \hat{v}])$ as above. The bold numbers in each of the tables in this section highlight the algorithm that performed the most efficiently in terms of iteration count or total runtime.

6.5.1 Maximum of Nonconvex Quadratic Forms

This subsection presents computational results for the min-max quadratic vector problem (MQV) problem considered in [49]. More specifically, given a dimension triple $(n, l, k) \in \mathbb{N}^3$, a set of parameters $\{(\alpha_i, \beta_i)\}_{i=1}^k \subseteq \mathbb{R}^2_{++}$, a set of vectors $\{d_i\}_{i=1}^k \subseteq \mathbb{R}^l$, a set of diagonal matrices $\{D_i\}_{i=1}^k \subseteq \mathbb{R}^{n \times n}$, and matrices $\{C_i\}_{i=1}^k \subseteq \mathbb{R}^{l \times n}$ and $\{B_i\}_{i=1}^k \subseteq \mathbb{R}^{n \times n}$, we consider the MQV problem

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^k} \left\{ \delta_{\Delta^n}(x) + \sum_{i=1}^k y_i g_i(x) : y \in \Delta^k \right\},$$

where, for every index $1 \le i \le k$, integer $p \in \mathbb{N}$, and $x \in \mathbb{R}^n$,

$$f_i(x) := \frac{\alpha_i}{2} \|C_i x - d_i\|^2 - \frac{\beta_i}{2} \|D_i B_i x\|^2, \quad \Delta^p := \left\{ z \in \mathbb{R}^p_+ : \sum_{i=1}^p z_i = 1, z \ge 0 \right\}.$$

We now describe the experiment parameters for the instances considered. First, the dimensions are set to be $(n, l, k) = (200, 10, 5)$ and only 5.0% of the entries of the submatrices B_i and C_i are nonzero. Second, the entries of B_i, C_i, and d_i (resp., D_i) are generated by sampling from the uniform distribution $\mathcal{U}[0, 1]$ (resp., $\mathcal{U}\{1, ..., 1000\}$). Third, the initial starting point is $z_0 = I_n/n$. Fourth, the key problem parameters, for every $(x, y) \in \mathbb{R}^n \times \mathbb{R}^k$, are

$$\Phi(x, y) = \sum_{i=1}^{k} y_i f_i(x), \quad h(x) = \delta_{\Delta^n}(x),$$
$$\rho_x = 10^{-2}, \quad \rho_y = 10^{-1}, \quad Y = \Delta^k.$$

Fifth, each problem instance considered is based on a specific curvature pair $(m, M) \in \mathbb{R}^2_{++}$ satisfying $m \leq M$, for which each scalar pair $(\alpha_i, \beta_i) \in \mathbb{R}^2_{++}$ is selected so that $M = \lambda_{\max}(\nabla^2 f_i)$ and $-m = \lambda_{\min}(\nabla^2 f_i)$ for $1 \leq i \leq k$. Moreover, the method for obtaining each pair (α_i, β_i) is the same as in Section 5.5.1.1. Finally, the Lipschitz and curvature constants selected are

$$m = m, \quad L_x = M, \quad L_y = M\sqrt{k} + \|P\|, \tag{6.54}$$

where P is an n-by-k matrix whose i^{th} column is equal to $\alpha_i C_i^T d_i$.

The table of iteration counts and total runtimes are given in Table 6.3 and Table 6.4, respectively.

Table 6.3: Iteration Counts for MQV problems.

(m, M)		**Iteration Count**		
m	M	PGSF	AG.S	AIPP.S
10^1	10^2	21462	1824	**81**
10^1	10^3	159682	6280	**267**
10^1	10^4	-	28966	**793**
10^1	10^5	-	28966	**793**

Table 6.4: Runtimes for MQV problems.

(m, M)		**Runtime**		
m	M	PGSF	AG.S	AIPP.S
10^1	10^2	358.24	40.17	**1.86**
10^1	10^3	2896.70	179.27	**6.36**
10^1	10^4	4000.00	698.52	**15.21**
10^1	10^5	4000.00	835.17	**18.76**

6.5.2 Truncated Robust Regression

This subsection presents computational results for the truncated robust regression (TRR) problem in [96]. More specifically, given a dimension pair $(n, k) \in \mathbb{N}^2$, a set of n data points $\{(a_j, b_j)\}_{i=1}^n \subseteq \mathbb{R}^k \times \{1, -1\}$ and a parameter $\alpha > 0$, we consider the TRR problem

$$\min_{x \in \mathbb{R}^k} \max_{y \in \mathbb{R}^n} \left\{ \sum_{j=1}^n y_j (\phi_\alpha \circ \ell_j)(x) : y \in \Delta^n \right\}$$

where Δ^n is as in (7.10) with $p = n$ and, for every $(\alpha, t, x) \in \mathbb{R}_{++} \times \mathbb{R}_{++} \times \mathbb{R}^k$,

$$\phi_\alpha(t) := \alpha \log\left(1 + \frac{t}{\alpha}\right), \quad \ell_j(x) := \log\left(1 + e^{-b_j \langle a_j, x \rangle}\right).$$

We now describe the experiment parameters for the instances considered. First, α is set to 10 and the data points $\{(a_i, b_i)\}$ are taken from different datasets in the LIBSVM library[2] for which each problem instance is based off of (see the "data name" column in the table below, which corresponds to a particular LIBSVM dataset). Second, the initial starting point is $z_0 = 0$. Third, the key problem parameters, for every $(x, y) \in \mathbb{R}^k \times \mathbb{R}^n$, are

$$\Phi(x, y) = \sum_{j=1}^n y_j (\phi_\alpha \circ \ell_j)(x), \quad h(x) = 0, \quad \rho_x = 10^{-5}, \quad \rho_y = 10^{-3}, \quad Y = \Delta^n.$$

[2] See https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html.

Finally, the Lipschitz and curvature constants selected are

$$m = L_x = \frac{1}{\alpha} \max_{1 \le j \le n} \|a_j\|^2, \quad L_y = \sqrt{\sum_{j=1}^{n} \|a_j\|^2}. \tag{6.55}$$

The table of iteration counts and total runtimes are given in Table 6.5 and Table 6.6, respectively.

Table 6.5: Iteration Counts for TRR problems.

	Iteration Count		
data name	PGSF	AG.S	AIPP.S
heart	6415	1746	**506**
diabetes	3721	1641	**463**
ionosphere	54545	8327	**1262**
sonar	-	96208	**69464**

Table 6.6: Runtimes for TRR problems.

	Runtime		
data name	PGSF	AG.S	AIPP.S
heart	10.24	3.24	**2.08**
diabetes	5.98	3.77	**1.67**
ionosphere	104.75	18.94	**4.58**
sonar	4000.00	**97.56**	107.42

It is worth mentioning that [96] also presents a min-max algorithm for obtaining a stationary point as in (6.53). However, its iteration complexity, which is $\mathcal{O}(\rho^{-6})$ when $\rho = \rho_x = \rho_y$, is significantly worse than the other algorithms considered in this section and, hence, we choose not to include this algorithm in our benchmarks.

6.5.3 Power Control in the Presence of a Jammer

This subsection presents computational results for the power control (PC) problem in [65]. More specifically, given a dimension pair $(N, K) \in \mathbb{N}^2$, a pair of parameters $(\sigma, R) \in \mathbb{R}^2_{++}$, a

3D tensor $\mathcal{A} \in \mathbb{R}_+^{K\times K\times N}$, and a matrix $B \in \mathbb{R}_+^{K\times N}$, we consider the PC problem

$$\min_{X\in\mathbb{R}^{K\times N}} \max_{y\in\mathbb{R}^N} \left\{ \sum_{k=1}^{K}\sum_{n=1}^{N} f_{k,n}(X,y) : 0 \le X \le R, 0 \le y \le \frac{N}{2}, \right\},$$

where, for every $(X,y) \in \mathbb{R}^{K\times N} \times \mathbb{R}^N$,

$$f_{k,n}(X,y) := -\log\left(1 + \frac{\mathcal{A}_{k,k,n}X_{k,n}}{\sigma^2 + B_{k,n}y_n + \sum_{j=1,j\neq k}^{K} \mathcal{A}_{j,k,n}X_{j,n}}\right).$$

We now describe the experiment parameters for the instances considered. First, the scalar parameters are set to be $(\sigma, R) = (1/\sqrt{2}, K^{1/K})$ and the quantities $\mathcal{A}$ and B are set to be the squared moduli of the entries of two Gaussian sampled complex–valued matrices $\mathcal{H} \in \mathbb{C}^{K\times K\times N}$ and $P \in \mathbb{C}^{K\times N}$. More precisely, the entries of $\mathcal{H}$ and P are sampled from the standard complex Gaussian distribution $\mathcal{CN}(0,1)$ and

$$\mathcal{A}_{j,k,n} = |\mathcal{H}_{j,k,n}|^2, \quad B_{k,n} = |P_{k,n}|^2 \quad \forall (j,k,n).$$

Second, the initial starting point is $z_0 = 0$. Third, with respect to the termination criterion (6.53), the inputs, for every $(X,y) \in \mathbb{R}^{K\times N} \times \mathbb{R}^N$, are

$$\Phi(X,y) = \sum_{k=1}^{K}\sum_{n=1}^{N} f_{k,n}(X,y), \quad h(X) = \delta_{Q_R^{K\times N}}(X),$$
$$\rho_x = 10^{-1}, \quad \rho_y = 10^{-1}, \quad Y = Q_{N/2}^{N\times 1}.$$

where $Q_T^{U\times V} := \{z \in \mathbb{R}^{p\times q} : 0 \le z \le T\}$ for every $T > 0$ and $(U,V) \in \mathbb{N}^2$. Fourth, each problem instance considered is based on a specific dimension pair (N,K). Finally, the Lipschitz and curvature constants selected are

$$m = L_x = \frac{2}{\min\{\sigma^4, \sigma^6\}} \max_{\substack{1\le k\le K\\ 1\le n\le N}} \sum_{j=1}^{K} \mathcal{A}_{k,j,n}^2, \quad L_y = \frac{2}{\min\{\sigma^4, \sigma^6\}} \max_{\substack{1\le k\le K\\ 1\le n\le N}} \sum_{j=1}^{K} B_{j,n}\mathcal{A}_{k,j,n}. \quad (6.56)$$

The table of iteration counts and total runtimes are given in Table 6.7 and Table 6.8,

respectively.

Table 6.7: Iteration Counts for PC problems.

(N, K)		**Iteration Count**		
N	K	PGSF	AG.S	AIPP.S
5	5	-	322831	**38**
10	10	-	33398	**62**
25	25	-	161716	**187**
50	50	-	-	**572**

Table 6.8: Runtimes for PC problems.

(N, K)		**Runtime**		
N	K	PGSF	AG.S	AIPP.S
5	5	4000.00	3166.40	**0.65**
10	10	4000.00	509.47	**0.74**
25	25	4000.00	3907.10	**4.89**
50	50	4000.00	4000.00	**30.29**

It is worth mentioning that [65] also presents a min-max algorithm for obtaining stationary points for the aforementioned problem. However, its termination criterion and notion of stationarity are significantly different from what is being considered in this chapter and, hence, we choose not to include the algorithm of [65] in our benchmarks.

6.5.4 Discussion of the Results

We see that the smoothing method in this chapter are competitive against other modern solvers and that they especially perform well when the curvature ratio M/m is large. Additionally, we see that the method scales well across problem dimensions and parameters.

6.6 Conclusion and Additional Comments

In this chapter, we presented a smoothing method for finding approximate stationary points of a class of min-max NCO problems. The method consists of applying the accelerated method of Chapter 3 to a smooth approximation of the original nonsmooth min-max problem. We then established an $\mathcal{O}(\delta^{-3})$ iteration complexity bound for finding δ-directional stationary points and an $\mathcal{O}(\rho_x^{-2}\rho_y^{-1/2})$ iteration bound for finding (ρ_x, ρ_y)-primal-dual stationary points. Additionally, we combined our developments with those in Section 4.1 to present a quadratic penalty smoothing method for finding approximate stationary points of a linearly-constrained variant of the original class of min-max NCO problems. We then established a $\mathcal{O}(\rho_x^{-2}[\rho_y^{-1/2} + \eta^{-1}])$ iteration complexity bound for finding (ρ_x, ρ_y)-primal-dual stationary points that were η feasible, i.e. the points $\bar{x}$ satisfy $\|\mathcal{A}\bar{x} - b\| \leq \eta$ for a particular linear constraint $\mathcal{A}x = b$.

The next chapter uses a framework similar to the one in Chapter 3 to develop methods for finding stationary points of a class of spectral NCO problems.

Additional Comments

We now give a few additional comments about the results in this chapter.

First, recall that the main idea of the AIPP.SM is to call the AIPPM of Chapter 3 to obtain a pair satisfying (6.23), or equivalently[3],

$$\inf_{\|d\| \leq 1} (\hat{p}_\xi)'(x; d) \geq -\rho. \tag{6.57}$$

Moreover, using Proposition 6.3.2 with $(\rho_x, \rho_y) = (\rho, D_y/\xi)$, it straightforward to see that the number of oracle calls, in terms of (ξ, ρ), is $\mathcal{O}(\rho^{-2}\xi^{1/2})$, which reduces to $\mathcal{O}(\rho^{-2.5})$ if ξ is chosen so as to satisfy $\xi = \Theta(\rho^{-1})$. The latter complexity bound improves upon the one obtained for an algorithm in [87] which obtains a point x satisfying (6.57) with $\xi = \Theta(\rho^{-1})$

[3] See Lemma F.1.2 with $f = p_\xi$.

in $\mathcal{O}(\rho^{-3})$ oracle calls.

Second, similar to Chapter 3, we neither assume that the set X in (F1) is bounded nor that the min-max NCO problem $\mathcal{MCO}$ has an optimal solution. Also, both the AIPP.SM and AIP.QP.SM only require that their starting point x_0 be in X and the AIP.QP.SM, in particular, makes no assumption about the feasibility of x_0.

Future Work

It is worth investigating whether complexity results for the AIPP.SM can be derived for the case where Y is unbounded or for the case in which assumption (F2) is relaxed to the condition that there exists $m_y > 0$ such that $-\Phi(x,\cdot)$ is m_y-weakly convex for every $x \in X$. It would also be interesting to see if the notions of stationary points in Section 6.1 are related to first-order stationary points[4] of the related mathematical program with equilibrium constraints:

$$\min_{(x,y)\in X\times Y} \{\Phi(x,y) + h(y) : 0 \in \partial[-\Phi(\cdot,y)](x)\}.$$

Finally, it would be worth investigating if a complexity as in Proposition 6.3.3 and Proposition 6.3.2 can still be obtained if the exact proximal oracle for $\Phi(x,\cdot)$ in Equation (6.2) is replaced with an inexact one.

[4] See, for example, [67, Chapter 3].

CHAPTER 7
SPECTRAL COMPOSITE OPTIMIZATION

Over the past decade, there has been a tremendous interest [17, 33, 53, 64, 68, 108] in developing iterative optimization algorithms for solving large-scale matrix NCO problems. Moreover, a large majority of the NCO problems in these works are such that the composite term h is a function of the singular values of its inputs and the smooth term f can be decomposed as $f_1 + f_2$ where f_2 is also a function of the singular values of its input. In this sense, these problems admit a sort of **spectral** decomposition.

Our main goal in this chapter is to describe and establish the iteration complexity of two efficient inexact composite gradient (ICG) methods for finding approximate stationary points of the spectral NCO (SNCO) problem

$$\min_{U \in \mathbb{R}^{m \times n}} \left\{ \phi(U) := f_1(U) + (f_2^{\mathcal{V}} \circ \sigma)(U) + (h^{\mathcal{V}} \circ \sigma)(U) \right\}, \tag{$\mathcal{SNCO}$}$$

where, denoting $r = \min\{m, m\}$, the function $\sigma : \mathbb{R}^{m \times n} \mapsto \mathbb{R}^r$ maps a matrix to its singular value vector in nonincreasing order of magnitude, $h^{\mathcal{V}} \in \overline{\text{Conv}}\, Z$ for some nonempty convex set $Z \subseteq \mathbb{R}^r$, $f_1 \in \mathcal{C}_{m_1,M_1}(\mathbb{R}^{m \times n})$ for some $(m_1, M_1) \in \mathbb{R}^2_{++}$, and $f_2^{\mathcal{V}} \in \mathcal{C}_{m_2,M_2}(Z)$ for some $(m_2, M_2) \in \mathbb{R}^2_{++}$. Moreover, we also assume that both $f_2^{\mathcal{V}}$ and $h^{\mathcal{V}}$ are absolutely symmetric in their arguments, i.e. they do not depend on the ordering or the sign of their arguments.

A standard approach for finding stationary points of $\mathcal{SNCO}$ is to apply the CGM (see Algorithm 2.2.1), or an accelerated version of it, to problem $\mathcal{SNCO}$ where $f = f_1 + f_2^{\mathcal{V}} \circ \sigma$ and $h = h^{\mathcal{V}} \circ \sigma$. The two ICG methods in this chapter generalize this approach by exploiting the spectral structure underlying the objective function. For example, one of the methods, called the accelerated ICG (AICG) method inexactly solves a sequence of matrix prox subproblems

of the form

$$\min_{U\in\mathbb{R}^{m\times n}} \left\{ \lambda \left[\langle \nabla f_1(Z_{k-1}), U\rangle + (f_2^{\mathcal{V}} \circ \sigma)(U) + (h^{\mathcal{V}} \circ \sigma)(U) \right] + \frac{1}{2}\|U - Z_{k-1}\|^2 \right\} \tag{7.1}$$

where $\lambda > 0$ and the point Z_{k-1} is the previous iterate. It is shown (see Section 7.5.1) that the effort of finding the required inexact solution Z_k of (7.1) consists of computing one SVD and applying an ACG method to the related vector prox subproblem

$$\min_{u\in\mathbb{R}^r} \left\{ \lambda \left[f_2^{\mathcal{V}}(u) - \langle c_{k-1}, u\rangle + h^{\mathcal{V}}(u) \right] + \frac{1}{2}\|u\|^2 \right\} \tag{7.2}$$

where $r = \min\{m, n\}$ and $c_{k-1} = \sigma(Z_{k-1} - \lambda\nabla f_1(Z_{k-1}))$. Note that (7.2) is a problem over the vector space $\mathbb{R}^r$, and hence, has significantly fewer dimensions than (7.1) which is a problem over the matrix space $\mathbb{R}^{m\times n}$. The other ICG method, called the doubly accelerated ICG (D.AICG) method, solves a similar prox subproblem as in (7.1) but with Z_{k-1} selected in an accelerated manner (and hence its qualifier of "doubly accelerated") and some additional mild assumptions.

Throughout our presentation, it is assumed that efficient oracles for evaluating the quantities $f_1(U)$, $f_2^{\mathcal{V}}(u)$, $\nabla f_1(U)$, $\nabla f_2^{\mathcal{V}}(u)$, and $h^{\mathcal{V}}(u)$ and for obtaining exact solutions of the subproblems

$$\min_{u\in\mathbb{R}^r} \left\{ \lambda h^{\mathcal{V}}(u) + \frac{1}{2}\|u - z_0\|^2 \right\},$$

for any $z_0 \in \mathbb{R}^r$ and $\lambda > 0$, are available. Moreover, we define an **oracle call** to be a collection of the above oracles of size $\mathcal{O}(1)$ where each of them appears at least once.

Given $\hat{\rho} > 0$ and a suitable choice of λ, the main result of this chapter shows that both of the ICG methods, started from any point $Z_0 \in Z$, obtain a pair $(\hat{Z}, \hat{V})$ satisfying the approximate stationarity condition

$$\hat{V} \in \nabla f_1(\hat{Z}) + \nabla \left(f_2^{\mathcal{V}} \circ \sigma\right)(\hat{Z}) + \partial \left(h^{\mathcal{V}} \circ \sigma\right)(\hat{Z}), \quad \|\hat{V}\| \le \hat{\rho} \tag{7.3}$$

in $\mathcal{O}(\hat{\rho}^{-2})$ oracle calls. When f_1 and $f_2^{\mathcal{Y}}$ are convex, it is shown that the D.AICGM obtains a pair $(\hat{Z}, \hat{V})$ satisfying in $\mathcal{O}(\hat{\rho}^{-2/3})$ oracle calls.

It is worth mentioning that the AICG method (AICGM) can be viewed an inexact version of the CGM applied to $\mathcal{SNCO}$, which solves a sequence of subproblems

$$\min_{U\in\mathbb{R}^{m\times n}} \left\{\lambda\left[\left\langle\nabla\left[f_1 + f_2^{\mathcal{Y}}\circ\sigma\right](Z_{k-1}), U\right\rangle + (h^{\mathcal{Y}}\circ\sigma)(U)\right] + \frac{1}{2}\|U - Z_{k-1}\|^2\right\}, \tag{7.4}$$

where $\lambda > 0$ and the point Z_{k-1} is the previous iterate. Similarly, the D.AICG method (D.AICGM) can be viewed as an inexact version of a monotone ACGM, which also solves a sequence of subproblems (7.4) but with Z_{k-1} chosen in an accelerated manner.

For high-dimensional instances of $\mathcal{SNCO}$ where $\min\{m, n\}$ is large, and hence, SVDs are expensive to compute, it will be shown that the larger the Lipschitz constant of $\nabla f_2^{\mathcal{Y}}$ is, the better the performance of the ICG methods is compared to that of their exact counterparts. This is due to the following facts: (i) solving (7.4) or (7.1) involves a single SVD computation; (ii) even though (7.4) requires fewer resolvent evaluations to solve than (7.1), the cost of solving these subproblems is comparable due to the fact that the aforementioned SVD is the bottleneck step; and (iii) the larger the Lipschitz constant of $\nabla f_2^{\mathcal{Y}}$, is the smaller the stepsize λ in (7.4) must be, and hence, the more subproblems of form (7.4) need to be solved during the execution of the exact counterparts.

The content of this chapter is based on paper [50] (joint work with Renato D.C. Monteiro) and several passages may be taken verbatim from it. To the best of our knowledge, paper [50] is the first one to present ICG methods that exploit both the spectral and composite structure in $\mathcal{SNCO}$.

Organization

This chapter contains seven sections. The first one gives some preliminary references and discusses our notion of a stationary point given in (7.3). The second one presents some spe-

cialized subroutines that are used in the ICG methods. The third one presents the AICGM and its iteration complexity. The fourth one presents the D.AICGM and its iteration complexity. The fifth one presents an ACG variant that exploits the spectral structure underlying the subproblems, i.e. (7.1), that each of the ICG methods solve. The sixth one presents some numerical experiments. The last one gives a conclusion and some closing comments.

7.1 Preliminaries

This subsection describes the general problem that the ICG methods solve and outlines their general structure.

The ICG methods consider the NCO problem

$$\phi_* = \min_{u \in \mathcal{Z}} [\phi(u) := f_1(u) + f_2(u) + h(u)] \qquad (\mathcal{NCO}_2)$$

where $\mathcal{Z}$ is an finite dimensional inner product space and the functions $f_1, f_2,$ and h are assumed to satisfy the following assumptions:

(H1) $h \in \overline{\mathrm{Conv}}\,(Z)$ for some nonempty convex set $Z \subseteq \mathcal{Z}$;

(H2) $f_1 \in \mathcal{C}_{m_1,M_1}(Z)$ and $f_2 \in \mathcal{C}_{m_2,M_2}(Z)$ for some $(m_1, M_1) \in \mathbb{R}^2$ and $(m_2, M_2) \in \mathbb{R}^2$;

(H3) $\phi_* > -\infty$.

We now make a few remarks about $\mathcal{NCO}_2$ and the above assumptions. First, $\mathcal{SNCO}$ is an instance of $\mathcal{NCO}_2$ in which $f_2 = f_2^{\mathcal{V}}$ and $h = h^{\mathcal{V}}$, and hence, any results developed in this section immediately apply for $\mathcal{SNCO}$. Second, it is well-known that a necessary condition for z^* to be a local minimum of $\mathcal{SNCO}$ is that z^* be a stationary point of ϕ, i.e. $0 \in \nabla f_1(z^*) + \nabla f_2(z^*) + \partial h(z^*)$.

In view of the above remarks, our goal is to find an approximate stationary point $(\hat{z}, \hat{v})$

of $\mathcal{NCO}_2$ in the following sense: given $\hat{\rho} > 0$, find a pair $(\hat{z}, \hat{v})$ that satisfies

$$\hat{v} \in \nabla f_1(\hat{z}) + \nabla f_2(\hat{z}) + \partial h(\hat{z}), \quad \|\hat{v}\| \leq \hat{\rho}. \tag{7.5}$$

For ease of future reference, let us state the problem of finding this pair in Problem 7.1.1.

Problem 7.1.1: Find an approximate stationary point of $\mathcal{NCO}_2$

Given $\hat{\rho} > 0$, find a pair $(\hat{z}, \hat{v}) \in Z \times \mathcal{Z}$ satisfying condition (7.5).

We now outline the ICG methods. Given a starting point $z_0 \in Z$ and a special stepsize $\lambda > 0$, each method continually calls an ACG variant, i.e. based on Algorithm 2.2.2, to find an approximate solution of a prox-linear form of $\mathcal{NCO}_2$. More specifically, each ACG call is used to tentatively find an inexact solution of

$$\min_{u \in \mathcal{Z}} \left\{ \lambda \left[\ell_{f_1}(u; w) + f_2(u) + h(u) \right] + \frac{1}{2} \|u - w\|^2 \right\}, \tag{7.6}$$

for some reference point w. For the AICGM, the point w is z_0 for the first ACG call and is the last obtained point for the other ACG calls. For the D.AICGM, the point w is chosen in an accelerated manner. From the output of the k^{th} ACG call, a refined pair $(\hat{z}, \hat{v}) = (\hat{z}_k, \hat{v}_k)$ is generated which: (i) always satisfies the inclusion of (7.5); and (ii) is such that $\min_{i \leq k} \|\hat{v}_i\| \to 0$ as $k \to \infty$.

The next section details the inexactness criterion considered by the ACG variant as well as how the refined pair $(\hat{z}, \hat{v})$ is generated. Before proceeding, we introduce the function

$$L_\Psi(u, z) := \begin{cases} \dfrac{\|\nabla \Psi(u) - \nabla \Psi(z)\|}{\|u - z\|}, & u \neq z, \\ 0, & u = z, \end{cases} \qquad \forall (u, z) \in Z,$$

for any differentiable function Ψ on Z, and the shorthand notation

$$\begin{gathered} M_i^+ := \max\{0, M_i\}, \quad m_i^+ := \max\{0, m_i\}, \quad L_i := \max\{m_i^+, M_i^+\} \\ L_i(u,z) = L_{f_i}(u,z) \quad \forall u, z \in Z, \end{gathered} \tag{7.7}$$

for $i \in \{1,2\}$, to keep the presentation of future results concise.

7.2 Specialized Refinement and ACG Procedures

Recall from the beginning of this chapter that our interest is in solving $\mathcal{SNCO}$ by repeated solving a sequence of prox subproblems as in (7.1). This subsection presents some background material regarding (7.1).

Consider the NCO problem

$$\min_{u \in \mathcal{Z}} \{\psi(u) = \psi_s(u) + \psi_n(u)\}, \tag{7.8}$$

where $\mathcal{Z}$ is a finite dimensional inner product space, $\psi_n \in \overline{\mathrm{Conv}}\,(Z)$, and $\psi_s \in \mathcal{C}_{m,L}(Z)$ for some $(m, L) \in \mathbb{R} \times \mathbb{R}_{++}$. Clearly, problem (7.6) and (7.1) are special cases of (7.8), and hence any definition or result that is stated in the context of (7.8) applies to (7.6) and/or (7.1).

We now discuss the inexactness criterion under which the subproblems (7.1) are solved. The criterion is described in the context of (7.8) as follows.

Problem $\mathcal{A}$: Given $(\mu, \theta) \in \mathbb{R}_{++}\times$ and $z_0 \in \mathcal{Z}$, find $(z, v, \varepsilon) \in Z \times \mathcal{Z} \times \mathbb{R}_+$ such that

$$v \in \partial_\varepsilon \left(\psi - \frac{\mu}{2}\|\cdot - z\|^2\right)(z), \quad \|v\|^2 + 2\varepsilon \le \theta^2\|z - z_0\|^2. \tag{7.9}$$

Some remarks about the above problem are in order. First, if (z, v, ε) solves Problem $\mathcal{A}$ with $\theta = 0$, then $(v, \varepsilon) = (0, 0)$, and z is an exact solution of (7.8). Hence, the output (z, v, ε) of Problem $\mathcal{A}$ can be viewed as an inexact solution of (7.8) when $\theta \in \mathbb{R}_{++}$. Second, the input

z_0 is arbitrary for the purpose of this section. However, the two methods described in the next two sections for solving $\mathcal{NCO}_2$ repeatedly solve (7.1) according to Problem $\mathcal{A}$ with the input z_0 at the k^{th} iteration determined by the iterates generated at the $(k-1)^{\text{th}}$ iteration. Third, defining the function

$$\Delta_\mu(u; z, v) := \psi(z) - \psi(u) - \langle v, u - z\rangle + \frac{\mu}{2}\|u - z\|^2 \quad \forall u, z \in Z, \tag{7.10}$$

another way to express the inclusion in (7.9) is $\Delta_\mu(u; z, v) \leq \varepsilon$ for every $u \in Z$. Finally, the ACG variant presented later in this section will be shown to solve Problem $\mathcal{A}$ when $\psi_s \in \mathcal{F}_\mu(Z)$. Moreover, it solves a weaker version of Problem $\mathcal{A}$ involving Δ_μ (see Problem $\mathcal{B}$ later on) whenever $\psi_s \notin \mathcal{F}_\mu(Z)$ and as long as some key inequalities are satisfied during its execution.

A technical issue in our analysis in this chapter lies in the ability of refining the output of Problem $\mathcal{A}$ to an point $(\hat{z}, \hat{v})$ satisfying the inclusion in (7.5), in which $\|\hat{v}\|$ is nicely bounded. The follow two results establish a way to obtain such a point.

The first result presents some properties of a composite gradient step made on (7.8).

Lemma 7.2.1. *Let a quadruple $(z_0, z, v, \varepsilon) \in \mathcal{Z} \times Z \times \mathcal{Z} \in \mathbb{R}_+$ and functions $\psi_n \in \overline{\text{Conv}}\,(Z)$ and $\psi_s \in \mathcal{C}_{\mu,L}(Z)$ for some $(mu, L) \in \mathbb{R} \times \mathbb{R}_{++}$ be given. Moreover, let $\psi = \psi_s + \psi_n$, the function $\Delta_\mu(\cdot;\cdot,\cdot)$ be as in* (7.10)*, and consider the pair $(\hat{z}, v_r)$ given by*

$$\begin{aligned} \hat{z} &:= \operatorname*{argmin}_{u \in \mathcal{Z}} \left\{ \ell_{\psi_s}(u; z) - \langle v, u\rangle + \frac{L}{2}\|u - z\|^2 + \psi_n(u) \right\}, \\ v_r &:= v + L(z - \hat{z}) + \nabla\psi_s(\hat{z}) - \nabla\psi_s(z), \end{aligned} \tag{7.11}$$

Then, the following statements hold:

(a) $v_r \in \nabla\psi_s(\hat{z}) + \partial\psi_n(\hat{z})$;

(b) for every $s \in Z$ we have $\Delta_\mu(u; z, v) \geq 0$ and, in particular,

$$\Delta_\mu(\hat{z}; z, v) \geq \frac{L}{2}\|\hat{z} - z\|^2; \tag{7.12}$$

(c) if $\Delta_\mu(\hat{z}; z, v) \leq \varepsilon$ and (z, v, ε) satisfy the inequality in (7.9)*, then*

$$\|v_r\| \leq \theta\left[1 + \frac{L + L_{\psi_s}(z, \hat{z})}{\sqrt{L}}\right]\|z - z_0\|; \tag{7.13}$$

(d) if (z, v, ε) solves Problem $\mathcal{A}$, then $\Delta_\mu(u; z, v) \leq \varepsilon$ for every $u \in Z$, and, as a consequence, bound (7.13) *holds.*

Proof. (a) The optimality condition of $\hat{z}$ is

$$0 \in \nabla\psi_s(z) - v + L(\hat{z} - z) + \partial\psi_n(\hat{z})$$

which, together with the definition of v_r, yields the desired inclusion.

(b) The fact that $\Delta_\mu(u; z, v) \geq 0$ for every $u \in Z$ follows from the optimality of $\hat{z}$ and the fact that $\psi_s \leq \ell_{\psi_s}(\cdot; z) + L\|\cdot - z\|^2/2$. The bound (7.12) follows from Proposition 2.2.2(c) with $\lambda = 1/L$ and $(z, z^-) = (\hat{z}, z)$.

(c) Using the assumption that $\Delta_\mu(\hat{z}; z, v) \leq \varepsilon$, part (b), and the inequality in (7.9), we have that

$$\|z - \hat{z}\| \leq \sqrt{\frac{2\Delta_\mu(\hat{z}; z, v)}{L}} \leq \sqrt{\frac{2\varepsilon}{L}} \leq \frac{\theta}{\sqrt{L}}\|z - z_0\|. \tag{7.14}$$

Using the triangle inequality, the definitions of $L(\cdot, \cdot)$ and v_r, (7.14), and the inequality in

(7.9), we conclude that

$$\begin{aligned}\|v_r\| &= \|v + L_\lambda(z - \hat{z}) + \nabla\psi_s(\hat{z}) - \nabla\psi_s(z)\| \\ &\leq \|v\| + \left[L + L_{\tilde{f}}(z, \hat{z})\right]\|z - \hat{z}\| \\ &\leq \theta\left[1 + \frac{L + L_{\tilde{f}}(z, \hat{z})}{\sqrt{L}}\right]\|z - z_0\|.\end{aligned}$$

(d) The fact that $\Delta_\mu(u; z, v) \leq \varepsilon$ for every $u \in Z$ follows immediately from the inclusion in (7.9) and the definition of Δ_μ in (7.10). The fact that (7.13) holds now follows from part (c). □

The next result specializes the above lemma to the context of $\mathcal{NCO}_2$ and describes the desired pair $(\hat{z}, \hat{v})$.

Proposition 7.2.2. *Let functions f_1, f_2, and h functions satisfying assumptions (H1)–(H2) and a quadruple $(z_0, z, v) \in Z \times Z \times \mathcal{Z} \in \mathbb{R}_+$ be given. Moreover, let $\Delta_\mu(\cdot;\cdot,\cdot)$ and $(\hat{z}, v_r)$ be as in Lemma 7.2.1 with*

$$\psi_s = \lambda\left[\ell_{f_1}(\cdot; z_0) + f_2\right] + \frac{1}{2}\|\cdot - z_0\|^2, \quad \psi_n = \lambda h, \quad L = \lambda M_2^+ + 1,$$

and define

$$\begin{aligned}\hat{v} &:= \frac{1}{\lambda}(v_r + z_0 - \hat{z}) + \nabla f_1(\hat{z}) - \nabla f_1(z_0), \\ C_\lambda(u, z) &:= \frac{2 + \lambda\left[M_2^+ + L_1(u, z) + L_2(u, z)\right]}{\sqrt{1 + \lambda M_2^+}},\end{aligned} \tag{7.15}$$

for every $u, z \in \mathcal{Z}$. Then, the following statements hold:

(a) $\hat{v} \in \nabla f_1(\hat{z}) + \nabla f_2(\hat{z}) + \partial h(\hat{z})$;

(b) if $\Delta_\mu(\hat{z}; z, v) \le \varepsilon$ *and* (z, v, ε) *satisfy the inequality in* (7.9)*, then it holds that*

$$\|\hat{v}\| \le \left[L_1(z_0, z) + \frac{2 + \theta C_\lambda(z, \hat{z})}{\lambda} \right] \|z - z_0\|; \tag{7.16}$$

Proof. (a) It follows from Lemma 7.2.1(a) and the definition of $\hat{v}$ that

$$\begin{aligned}
\hat{v} &= \frac{1}{\lambda}(v_r + z_0 - \hat{z}) + \nabla f_1(\hat{z}) - \nabla f_1(z_0) \\
&\in \frac{1}{\lambda}[\nabla \psi_s(\hat{z}) + \partial \psi_n(\hat{z})] + \frac{1}{\lambda}(v_r + z_0 - \hat{z}) + \nabla f_1(\hat{z}) - \nabla f_1(z_0) \\
&= \nabla f_1(\hat{z}) + \nabla f_2(\hat{z}) + \partial h(\hat{z}).
\end{aligned}$$

(b) It follows from (7.14) with $L := \lambda M_2^+ + 1$, the triangle inequality that

$$\begin{aligned}
&\frac{1}{\lambda}\|z_0 - \hat{z}\| + \|\nabla f_1(z_0) - \nabla f_1(\hat{z})\| \\
&\le \frac{1}{\lambda}[1 + \lambda L_1(z, \hat{z})]\,\|z - \hat{z}\| + [1 + \lambda L_1(z_0, z)]\,\|z - z_0\| \\
&\le \frac{1}{\lambda}\left(1 + \lambda L_1(z_0, z) + \theta\left[\frac{1 + \lambda L_1(z, \hat{z})}{\sqrt{1 + \lambda M_2^+}}\right]\right)\|z - z_0\|.
\end{aligned}$$

Using the above bound, Lemma 7.2.1(c) with $L = \lambda M_2^+ + 1$ and $L_{\psi_s}(\cdot, \cdot) = \lambda L_2(\cdot, \cdot) + 1$, the definition of $C_\lambda(\cdot, \cdot)$, and the fact that $\theta \le 1$, we conclude that

$$\begin{aligned}
\|\hat{v}\| &\le \frac{1}{\lambda}\|v_r\| + \frac{1}{\lambda}\|z_0 - \hat{z}\| + \|\nabla f_1(z_0) - \nabla f_1(\hat{z})\| \\
&\le \frac{1}{\lambda}\left(1 + \theta + \lambda L_1(z_0, z) + \theta\left[\frac{2 + \lambda M_2^+ + \lambda L_1(z, \hat{z}) + \lambda L_2(z, \hat{z})}{\sqrt{1 + \lambda M_2^+}}\right]\right)\|z - z_0\| \\
&\le \left[L_1(z_0, z) + \frac{2 + \theta C_\lambda(z, \hat{z})}{\lambda} \right]\|z - z_0\|.
\end{aligned}$$

□

We make a few remarks about Proposition 7.2.2. First, it follows from (a) that $(\hat{z}, \hat{v})$ satisfies the inclusion in (7.5). Second, it follows from (a) and (c) that if $\theta = 0$, then $(\hat{z}, \hat{v}) =$

$(0,0)$, and hence $\hat{z}$ is an exact stationary point of $\mathcal{NCO}_2$. In general, (7.16) implies that the residual $\|\hat{v}\|$ is directly proportional to $\|z - z_0\|$, and hence, becomes smaller as this quantity approaches *zero*.

For the sake of future referencing, we state the specialized refinement procedure (SRP) for generating $(\hat{z}, \hat{v})$ in Algorithm 7.2.1.

Algorithm 7.2.1: SR Procedure

Require: $(m_1, M_1, m_2, M_2) \in \mathbb{R}^4$, $h \in \overline{\mathrm{Conv}}\,(\mathcal{Z})$, $f_1 \in \mathcal{C}_{m_1,M_1}(Z)$, $f_2 \in \mathcal{C}_{m_2,M_2}(Z)$, $(z, z_0, v) \in Z \times \mathcal{Z} \times \mathcal{Z}$, $\lambda > 0$;

Initialize: $\psi_s \Leftarrow \lambda\,[\ell_{f_1}(\cdot; z_0) + f_2] + \dfrac{1}{2}\|\cdot - z_0\|^2$, $\psi_n \Leftarrow \lambda h$, $L \leftarrow \lambda M_2^+ + 1$ (see (7.7));

1: **procedure** SREF($f_1, f_2, h, z, z_0, v, M_2, \lambda$)

2: $\quad \hat{z} \leftarrow \underset{u \in \mathcal{Z}}{\mathrm{argmin}} \left\{ \ell_{\psi_s}(u; z) - \langle v, u \rangle + \dfrac{L}{2}\|u - z\|^2 + \psi_n(u) \right\}$

3: $\quad v_r \leftarrow v + L(z - \hat{z}) + \nabla\psi_s(\hat{z}) - \nabla\psi_s(z)$

4: $\quad \hat{v} \leftarrow \dfrac{1}{\lambda}(v_r + z_0 - \hat{z}) + \nabla f_1(\hat{z}) - \nabla f_1(z_0)$

5: $\quad$ **return** $(\hat{z}, \hat{v})$

Inequalities (7.13) and (7.16) play an important technical role in the complexity analysis of the two prox-type methods of the next two sections. Sufficient conditions for their validity are provided in Lemma 7.2.1(c)–(d), with (c) being the weaker one, in view of (d). When $\psi_s \in \mathcal{F}_\mu(Z)$, it is shown that every iterate of our proposed ACG variant always satisfies the inclusion in (7.9), and hence, verifying the validity of the sufficient condition in (c) amounts to simply checking whether the inequality in (7.9) holds. When $\psi_s \notin \mathcal{F}_\mu(Z)$, verification of the inclusion in (7.9), and hence the sufficient condition in (d), is generally not possible, while the one in (c) is. This is a major advantage of the sufficient condition in (c), which is exploited in this chapter towards the development of adaptive prox-type methods which attempt to approximately solve (7.8) when $\psi_s \notin \mathcal{F}_\mu(Z)$.

To ease future referencing, we state below the problem for finding a triple (z, v, ε) satisfying the sufficient condition in Lemma 7.2.1(c).

Problem $\mathcal{B}$: Given the same inputs as in Problem $\mathcal{A}$, find $(z, v, \varepsilon) \in Z \times \mathcal{Z} \times \mathbb{R}_+$ satisfying the inequality in (7.9) and

$$\Delta_\mu(\hat{z}; z, v) \leq \varepsilon, \tag{7.17}$$

where $\Delta_\mu(\cdot;\cdot,\cdot)$ is as in (7.10) and the point $\hat{z}$ is given by (7.11).

We now present the specialized ACG (S.ACG) method in Algorithm 7.2.2, which solves Problem $\mathcal{A}$ when $\psi_s \in \mathcal{F}_\mu(Z)$ and solves Problem $\mathcal{B}$ whenever two key inequalities are always satisfied, one at every iteration and one at the end of its execution. The termination status of the method is stored in the variable π_S which is `true` if the method solves Problem $\mathcal{B}$ and `false` otherwise.

Algorithm 7.2.2: S.ACG Method

Require: $(\mu, L) \in \mathbb{R}^2_{++}$, $\psi_n \in \overline{\text{Conv}}\,(\mathcal{Z})$, $\psi_s \in \mathcal{C}_L(Z)$, $y_0 \in Z$, $\theta \in (0,1)$;
Initialize: $\pi_S \leftarrow$ `true`, $\psi \Leftarrow \psi_s + \psi_n$,

procedure S.ACG($\psi_s, \psi_n, y_0, \theta, \mu, L$)
  for $k = 1, \ldots$ **do**
    $\lambda_k \leftarrow 1/L$
    Generate $(A_k, y_k, \tilde{x}_{k-1}, \tilde{r}_k, \tilde{\eta}_k)$ according to Algorithm 2.2.2.
    PART 1 **Check** the first failure point.
    if $\dfrac{1}{1+\mu A_k}\|A_k\tilde{r}_k + y_k - y_0\|^2 + 2A_k\tilde{\eta}_k \leq \|y_k - y_0\|^2$ **then**
      $\pi_S \leftarrow$ `false`
      return $(y_0, \infty, \infty, \pi_S)$
    if $\|\tilde{r}_k\|^2 + 2\tilde{\eta}_k \leq \theta^2\|y_k - y_0\|^2$ **then**
      $\hat{y}_k \leftarrow \operatorname{argmin}_{u \in \mathcal{Z}} \left\{ \ell_{\psi_s}(u; z) - \langle v, u\rangle + \dfrac{M}{2}\|u - z\|^2 + \psi_n(u) \right\}$
      PART 2 **Check** the second failure point.
      if $\Delta_\mu(\hat{y}_k; y_k, \tilde{r}_k) > \tilde{\eta}_k$ **then** ▷ See (7.10)
        $\pi_S \leftarrow$ `false`
        return $(y_0, \infty, \infty, \pi_S)$
      else
        return $(y_k, r_k, \tilde{\eta}_k, \pi_S)$

The next result presents the key properties of the S.ACG method (S.ACGM).

Proposition 7.2.3. *The following properties hold about the S.ACGM:*

(a) it stops in

$$\mathcal{O}\left(\left[1+\sqrt{\frac{L}{\mu}}\right]\log_1^+\left[LK_\theta(1+\mu K_\theta)\right]\right) \tag{7.18}$$

iterations, where $K_\theta = 1+\sqrt{2}/\theta$*;*

(b) if its stops with a quadruple $(z,v,\varepsilon,\pi_S) = (y_k,\tilde{r}_k,\tilde{\eta}_k,\pi_S)$ *where* $\pi_S =$ `true`*, then the triple* (z,v,ε) *solves Problem* $\mathcal{B}$*;*

(c) if $\psi_S \in \mathcal{F}_\mu(Z)$*, then it always stops with a quadruple* $(z,v,\varepsilon,\pi_S) = (y_k,\tilde{r}_k,\tilde{\eta}_k,\pi_S)$ *where* $\pi_S =$ `true`*, and the triple* (z,v,ε) *solves Problem* $\mathcal{A}$*.*

Proof. (a) See Appendix C.

(b) Using the successful checks in Line 9 and Line 15 of the method, it follows that the triple (z,v,ε) solves Problem $\mathcal{B}$.

(c) Using Proposition 2.2.3(a)–(b) and the definition of the approximate subdifferential, it follows that the method always stops with $\pi_S =$ `true` when $\psi_s \in \mathcal{F}_\mu(Z)$. On the other hand, Proposition 2.2.3(a), the definition of the approximate subdifferential, and the successful check in Line 9 of the method imply that the triple (z,v,ε) solves Problem $\mathcal{A}$. □

It is worth recalling that in the applications we consider, the cost of the ACG call is small compared to SVD computation that is performed before solving each subproblem as in (7.6). Hence, in the analysis that follows, we present complexity results related to the number of subproblems solved rather than the total number of ACG iterations. We do note, however, that the number of ACG iterations per subproblem is finite in view of Proposition 7.2.3(a).

7.3 Accelerated Inexact Composite Gradient (AICG) Method

This section presents the static AICGM and its dynamic variant.

We first state the static AICGM in Algorithm 7.3.1, which uses Algorithm 7.2.1 and Algorithm 7.2.2 as subroutines. Given $z_0 \in Z$ and a special choice of $\lambda > 0$, its main idea is to attempt to generate its k^{th} iterate by using the S.ACGM to obtain the inexact update

$$z_k \approx \min_{u \in Z} \left\{ \lambda \left[\ell_{f_1}(u; z_{k-1}) + f_2(u) + h(u) \right] + \frac{1}{2} \|u - z_{k-1}\|^2 \right\}.$$

The iterate is then refined using the SRP in Algorithm 7.2.1 and termination of the method occurs when either: (i) a refined iterate solving Problem 7.1.1 is found; or (ii) a failure condition has been triggered. The termination status of the method is store in a variable π_S which is `true` if the former scenario occurs and `false` the latter scenario occurs.

Algorithm 7.3.1: Static AICG Method

Require: $\hat{\rho} > 0$, $(m_1, M_1, m_2, M_2) \in \mathbb{R}^4$, $h \in \overline{\text{Conv}}\,(Z)$, $f_1 \in \mathcal{C}_{m_1,M_1}(Z)$, $f_2 \in \mathcal{C}_{m_2,M_2}(Z)$, $(\lambda, \theta) \in \mathbb{R}^2_{++}$ s.t. $\lambda M_1 + \theta^2 < 1/2$, $z_0 \in Z$;
Initialize: $\mu \leftarrow 1$, $L \leftarrow \lambda M_2^+ + 1$(see (7.7)), $\pi_S \leftarrow$ `true`

1: **procedure** St.AICG($f_1, f_2, h, z_0, \lambda, \theta, M_2, \hat{\rho}$)
2: **for** $k = 1, \ldots$ **do**
3: Part 1 **Attack** the k^{th} prox-linear subproblem.
4: $\psi_s^k \Longleftarrow \lambda \left[\ell_{f_1}(\cdot; z_{k-1}) + f_2 \right] + \frac{1}{2} \|\cdot - z_{k-1}\|^2$
5: $(z_k, v_k, \varepsilon_k, \pi_k^{\text{acg}}) \leftarrow$ S.ACG($\psi_s^k, \lambda h, z_{k-1}, \theta, \mu, L$)
6: Part 2 **Check** a special convexity condition.
7: **if** $\neg(\pi_k^{\text{acg}})$ **or** $\Delta_\mu(z_{k-1}; z_k, v_k) > \varepsilon_k$ **then** ▷ See (7.10)
8: $\pi_S \leftarrow$ `false`
9: **return** (z_0, ∞, π_S)
10: Part 3 **Check** the termination condition.
11: $(\hat{z}_k, \hat{v}_k) \leftarrow$ SREF($f_1, f_2, h, z_k, z_{k-1}, v_k, M_2, \lambda$)
12: **if** $\|\hat{v}_k\| \leq \hat{\rho}$ **then**
13: **return** $(\hat{z}_k, \hat{v}_k, \pi_S)$

Some remarks about this method are in order. To ease the discussion, let us refer to the ACG iterations performed in Line 5 of the method as **inner iterations** and the iterations

over the indices k as **outer iterations**. First, in view of the requirement on (λ, θ), if $M_1 > 0$ then $0 < \lambda < (1 - 2\theta^2)/(2M_1)$ whereas if $M_1 \leq 0$ then $0 < \lambda < \infty$. Second, it may fail to obtain a pair satisfying (7.5), i.e. when $\pi_S = \texttt{false}$. In Theorem 7.3.1(c) below, we state that a sufficient condition for the method to stop successfully is that f_2 be convex. This property will be important when we present the dynamic AICGM, which: (i) repeatedly calls the static method; and (ii) incrementally transfers convexity from f_1 to f_2 between each call until a termination where $\pi_S = \texttt{true}$ is achieved.

The next result, whose proof is deferred to Section 7.3.1, summarizes some facts about the static AICGM. Before proceeding, we first define some useful quantities. For and $\lambda > 0$ and $u, w \in \mathcal{Z}$, define

$$\widetilde{\ell}_\phi(u; w) := \ell_{f_1}(u; w) + f_2(u) + h(u), \quad \overline{C}_\lambda := \frac{1 + \lambda(M_2^+ + L_1 + L_2)}{\sqrt{1 + \lambda M_2^+}}. \tag{7.19}$$

Theorem 7.3.1. *The following statements hold about the static AICGM:*

(a) it stops in

$$\mathcal{O}\left(\left[\sqrt{\lambda}L_1 + \frac{1 + \theta\overline{C}_\lambda}{\sqrt{\lambda}}\right]^2 \left[\frac{\phi(z_0) - \phi_*}{\hat{\rho}^2}\right]\right) \tag{7.20}$$

outer iterations, where ϕ_* is as in (H3);

(b) if it stops with $\pi_S = \texttt{true}$, then the first two arguments of its output triple $(\hat{z}, \hat{v}, \pi_S)$ solve Problem 7.1.1;

(c) if f_2 is convex, then it always stops with $\pi_S = \texttt{true}$.

We now make three remarks about the above results. First, if $\theta = \mathcal{O}(1/\overline{C}_\lambda)$ then (7.20) reduces to

$$\mathcal{O}\left(\left[\sqrt{\lambda}L_1 + \frac{1}{\sqrt{\lambda}}\right]^2 \left[\frac{\phi(z_0) - \phi_*}{\hat{\rho}^2}\right]\right). \tag{7.21}$$

Moreover, comparing the above complexity to the iteration complexity of the CGM (see

Algorithm 2.2.1), which is known [86] to solve Problem 7.1.1 in

$$\mathcal{O}\left(\left[\sqrt{\lambda}(L_1 + L_2) + \frac{1}{\sqrt{\lambda}}\right]^2 \left[\frac{\phi(z_0) - \phi_*}{\hat{\rho}^2}\right]\right) \tag{7.22}$$

iterations, we see that (7.21) is smaller than (7.22) in magnitude when L_2 is large. Second, Theorem 7.3.1(b) shows that if the method stops with π_S = `true`, regardless of the convexity of f_2, then its output pair $(\hat{z}, \hat{v})$ is always a solution of Problem 7.1.1. Third, it is shown in Proposition 7.3.4, that the quantities L_1 and $\overline{C}_\lambda$ in all the previous complexity results can be replaced by their averaged counterparts in (7.24). As these averaged quantities only depend on $\{(z_i, \hat{z}_i)\}_{i=1}^k$, we can infer that the static AICG method adapts to the local geometry of its input functions.

We now state the (dynamic) AICG variant in Algorithm 7.3.2, which address the possibility of failure by repeatedly calling the static AICGM.

Algorithm 7.3.2: AICG Method

Require: $\hat{\rho} > 0$, $(m_1, M_1, m_2, M_2) \in \mathbb{R}^4$, $h \in \overline{\text{Conv}}\,(Z)$, $f_1 \in \mathcal{C}_{m_1,M_1}(Z)$, $f_2 \in \mathcal{C}_{m_2,M_2}(Z)$, $(\lambda, \theta) \in \mathbb{R}^2_{++}$ s.t. $\lambda M_1 + \theta^2 < 1/2$, $z_0 \in Z$, $\xi_0 > 0$;
Initialize: $\mu \leftarrow 1$, $L \leftarrow \lambda M_2^+ + 1$(see (7.7))

procedure AICG($f_1, f_2, h, z_0, \lambda, \theta, M_2, \xi_1, \hat{\rho}$)
  for $k = 1, \ldots$ **do**
    PART 1 **Call** the static AICGM with perturbed inputs.
    $f_1^k \Leftarrow f_1 - \frac{\xi_k}{2}\|\cdot\|^2$
    $f_2^k \Leftarrow f_2 + \frac{\xi_k}{2}\|\cdot\|^2$
    $(\hat{z}, \hat{v}, \pi_S) \leftarrow$ St.AICG($f_1^k, f_2^k, h, z_0, \lambda, \theta, M_2 + \xi_k, \hat{\rho}$)
    PART 2 Either **stop** with a solution or **increase** ξ_k for the next AICG call.
    if π_S **then**
      return $(\hat{z}, \hat{v})$
    else
      $\xi_{k+1} \leftarrow 2\xi_k$

Some remarks about the above method are in order. First, in view of the requirement

on (λ, θ) and the fact that the upper curvature of f_1^k is monotonically decreasing in k, the parameter λ does not need to be changed for each static AICG call. Second, in view Theorem 7.3.1(c), every static AICG call always terminates with success whenever f_2^k is convex. As a consequence, assumption (H2) implies that the total number of static AICG calls is at most $\lceil \log(2m_2^+/\xi_1) \rceil$. Third, in view of the second remark and Theorem 7.3.1(b), the methods always obtains a solution of Problem 7.1.1 in a finite number of static AICG outer iterations. Finally, in view of second remark again, the total number of static AICG outer iterations is as in Theorem 7.3.1(a) but with: (i) an additional multiplicative factor of $\lceil \log(2m_2^+/\xi_0) \rceil$; and (ii) the constants m_1 and M_2 replaced with $(m_1 + 2m_2^+)$ and $(M_2 + 2m_2^+)$, respectively. It is worth mentioning that a more refined analysis, such as the one in Section 5.3, can be applied in order to remove the factor of $\lceil \log(2m_2^+/\xi_0) \rceil$ from the previously mentioned complexity.

7.3.1 AICG Properties and Iteration Complexity

This subsection establishes the key properties of the static AICGM and gives the proof of Theorem 7.3.1.

We first start with a technical lemma that describes the progress, in terms of function value, between consecutive iterations. Its statement, and the statement of subsequent results, will make use of the key constants in (7.7).

Lemma 7.3.2. *Let $\{(z_i, \hat{z}_i, \hat{v}_i)\}_{i=1}^k$ be the collection of iterates generated by the static AICGM. For every $i \geq 1$, we have*

$$\frac{1}{4\lambda}\|z_{i-1} - z_i\|^2 \leq \phi(z_{i-1}) - \widetilde{\ell}_\phi(z_i; z_{i-1}) - \frac{M_1}{2}\|z_i - z_{i-1}\|^2 \leq \phi(z_{i-1}) - \phi(z_i), \tag{7.23}$$

where $\widetilde{\ell}_\phi$ is as in (7.19).

Proof. Let $i \geq 1$ be fixed, define

$$\mu := 1, \quad \psi_s := \lambda\left[\ell_{f_1}(\cdot; z_{i-1}) + f_2\right] + \frac{1}{2}\|\cdot - z_{i-1}\|^2, \quad \psi_n := \lambda h,$$

and let $(z_i, v_i, \varepsilon_i, \pi_i)$ be the output of the i^{th} call to the S.ACG algorithm. Moreover, let $\Delta_\mu(\cdot;\cdot,\cdot)$ be as in (7.10) with (ψ_s, ψ_n) as above. Using the definition of $\widetilde{\ell}_\phi$ and fact that $(z, v, \varepsilon) = (z_i, v_i, \varepsilon_i)$ solves Problem $\mathcal{B}$ in Section 7.2, we have that

$$\begin{aligned}\varepsilon_i &\geq \Delta_1(z_{i-1}; z_i, v_i)\\ &= \lambda\widetilde{\ell}_\phi(z_i; z_{i-1}) - \lambda\phi(z_{i-1}) - \langle v_i, z_i - z_{i-1}\rangle + \|z_i - z_{i-1}\|^2.\end{aligned}$$

Rearranging the above inequality and using assumption (H2), the requirement on (λ, θ) (in the AICGM), and the fact that $\langle a, b\rangle \geq -\|a\|^2/2 - \|b\|^2/2$ for every $a, b \in \mathcal{Z}$ yields

$$\begin{aligned}&\lambda\phi(z_{i-1}) - \lambda\widetilde{\ell}_\phi(z_i; z_{i-1}) \geq \langle v_i, z_{i-1} - z_i\rangle - \varepsilon_i + \|z_i - z_{i-1}\|^2\\ &= \frac{1}{2}\|z_i - z_{i-1}\|^2 - \frac{1}{2}\left(\|v_i\|^2 + 2\varepsilon_i\right) \geq \left(\frac{1-\sigma^2}{2}\right)\|z_i - z_{i-1}\|^2\\ &= \frac{\lambda M_1}{2}\|z_i - z_{i-1}\|^2 + \left(\frac{1 - \lambda M_1 - \sigma^2}{2}\right)\|z_i - z_{i-1}\|^2\\ &= \frac{\lambda M_1}{2}\|z_i - z_{i-1}\|^2 + \frac{1}{4}\|z_i - z_{i-1}\|^2.\end{aligned}$$

Rearranging terms yields the first inequality of (7.23). The second inequality of (7.23) follows from the first inequality, the fact that $\widetilde{\ell}_\phi(z_i; z_{i-1}) + M_1\|z_i - z_{i-1}\|^2/2 \geq \phi(z_i)$ from assumption (H2), and the definition of $\widetilde{\ell}_\phi$. □

The next results establish the rate at which the residual $\|\hat{v}_i\|$ tends to 0.

Lemma 7.3.3. *Let $p > 1$ be given. Then, for every $a, b \in \mathbb{R}^k$, we have*

$$\min_{1\leq i\leq k}\{|a_i b_i|\} \leq k^{-p}\,\|a\|_1\,\|b\|_{1/(p-1)}\,.$$

Proof. Let $p > 1$ and $a, b \in \mathbb{R}^k$ be fixed and let $q \geq 1$ be such that $p^{-1} + q^{-1} = 1$. Using the fact that $\langle x, y\rangle \leq \|x\|_p\|y\|_q$ for every $x, y \in \mathbb{R}^k$, and denoting $\tilde{a}$ and $\tilde{b}$ to be vectors with entries

$|a_i|^{1/p}$ and $|b_i|^{1/p}$, respectively, we have that

$$k \min_{1\le i\le k} \{|a_i b_i|\}^{1/p} \le \sum_{i=1}^{k} |a_i b_i|^{1/p}$$
$$\le \|\tilde{a}\|_p \|\tilde{b}\|_q = \|a\|_1^{1/p} \left(\sum_{i=1}^{k} |b_i|^{q/p} \right)^{1/q} = \left(\|a\|_1 \|b\|_{q/p} \right)^{1/p}.$$

Dividing by k, taking the p^{th} power on both sides, and using the fact that $p/q = p - 1$, yields

$$\min_{1\le i\le k} \{|a_i b_i|\} \le k^{-p} \|a\|_1 \|b\|_{q/p} = k^{-p} \|a\|_1 \|b\|_{1/(p-1)}.$$

□

Proposition 7.3.4. *Let $\{(z_i, \hat{z}_i, \hat{v}_i)\}_{i=1}^k$ be as in Lemma 7.3.2 and define the quantities*

$$L_{1,k}^{\text{avg}} := \frac{1}{k} \sum_{i=1}^{k} L_1(z_i, z_{i-1}), \quad C_{\lambda,k}^{\text{avg}} := \frac{1}{k} \sum_{i=1}^{k} C_\lambda(\hat{z}_i, z_i), \tag{7.24}$$

where $C_\lambda(\cdot,\cdot)$ and $\overline{C}_\lambda$ are as in (7.15) *and* (7.19)*, respectively. Then, we have*

$$\min_{i\le k} \|\hat{v}_i\| = \mathcal{O}\left(\left[\sqrt{\lambda} L_{1,k}^{\text{avg}} + \frac{1 + \theta C_{\lambda,k}^{\text{avg}}}{\sqrt{\lambda}} \right] \left[\frac{\phi(z_0) - \phi_*}{k} \right]^{1/2} \right) + \frac{\hat{\rho}}{2}.$$

Proof. Using Proposition 7.2.2 with $(z, z_0) = (z_i, z_{i-1})$ and the fact that $C_\lambda(\cdot,\cdot) \le \overline{C}_\lambda$ and $L_1(\cdot,\cdot) \le L_1$, we have $\|\hat{v}_i\| \le \mathcal{E}_i \|z_i - z_{i-1}\|$, for every $i \le k$, where

$$\mathcal{E}_i := \frac{2 + \lambda L_1(z_i, z_{i-1}) + \theta C_\lambda(\hat{z}_i, z_i)}{\lambda} \quad \forall i \ge 1.$$

As a consequence, using the sum of the second bound in Lemma 7.3.2 from $i = 1$ to k, the definitions in (7.24), and Lemma 7.3.3 with $p = 3/2$, $a_i = \mathcal{E}_i$, and $b_i = \|z_i - z_{i-1}\|$ for $i = 1$ to

k, yields

$$\begin{aligned}\min_{i\le k}\|\hat{v}_i\| &\le \min_{i\le k}\mathcal{E}_i\|z_i - z_{i-1}\| \le \frac{1}{k^{3/2}}\left(\sum_{i=1}^{k}\mathcal{E}_i\right)\left(\sum_{i=1}^{k}\|z_i - z_{i-1}\|^2\right)^{1/2} \\ &= \mathcal{O}\left(\left[\sqrt{\lambda}L_{1,k}^{\text{avg}} + \frac{1+\theta C_{\lambda,k}^{\text{avg}}}{\sqrt{\lambda}}\right]\left[\frac{\phi(z_0)-\phi_*}{k}\right]^{1/2}\right).\end{aligned}$$

□

We are now ready to give the proof of Theorem 7.3.1.

Proof of Theorem 7.3.1. (a) This follows from Proposition 7.3.4, the fact that $C_\lambda(\cdot,\cdot) \le \overline{C}_\lambda$ and $L_{f_1}(\cdot,\cdot) \le L_1$, and the termination condition in Line 12 of the AICGM.

(b) The fact that $(\hat{z},\hat{v}) = (\hat{z}_k,\hat{v}_k)$ satisfies the inclusion of (7.5) follows from Proposition 7.2.2 with $(z,v,z_0) = (z_k,v_k,z_{k-1})$. The fact that $\|\hat{v}\| \le \hat{\rho}$ follows from the termination condition in Line 12 of the AICGM.

(c) This follows from Proposition 7.2.3(c) and the fact that method stops in finite number of iterations from part (a). □

7.4 Doubly-Accelerated Inexact Composite Gradient (D.AICG) Method

This subsection presents the static D.AICGM, but omits its dynamic variant for the sake of brevity. We do argue, however, that the dynamic variant can be stated in the same way as the dynamic AICG variant in Section 7.3 but with the call to the static AICGM replaced with a call to the static D.AICGM of this subsection.

We start by stating some additional assumptions. It is assumed that:

(i) the set Z is closed;

(ii) there exists a bounded set $\Omega \supseteq Z$ for which a projection oracle exists.

We first state the static D.AICGM in Algorithm 7.4.1, which uses Algorithm 7.2.1 and Algorithm 7.2.2 as subroutines. Given $z_0 \in Z$ and a special choice of $\lambda > 0$, its main idea is to attempt to generate its k^{th} iterate using the S.ACGM and project oracle of Ω to obtain accelerated updates

$$
\begin{aligned}
a_k &= \frac{1 + \sqrt{1 + 4A_{k-1}}}{2}, \quad A_k = A_{k-1} + a_k, \\
\tilde{y}_k &= \frac{A_{k-1}}{A_k} z_{k-1} + \frac{a_{k-1}}{A_k} y_{k-1}, \\
z_k^a &\approx \min_{u \in \mathcal{Z}} \left\{ \lambda \left[\ell_{f_1}(u; z_{k-1}) + f_2(u) + h(u) \right] + \frac{1}{2} \|u - z_{k-1}\|^2 \right\}, \\
y_k &= \operatorname*{argmin}_{u \in \Omega} \frac{1}{2} \left\| u - \left[y_{k-1} - a_{k-1} \left(v_k + \tilde{y}_{k-1} - z_k^a \right) \right] \right\|^2, \\
z_k &= \operatorname*{argmin}_{u \in \{z_{k-1}, z_k^a\}} \phi(u),
\end{aligned}
$$

where $y_0 = z_0$, $A_0 = 0$, and v_k is a residual that is obtained from computing z_k^a. In particular, the S.ACGM is used in the inexact update of z_k^a. The iterate is then refined using the SRP in Algorithm 7.2.1 and termination of the method occurs when either: (i) a refined iterate solving Problem 7.1.1 is found; or (ii) a failure condition has been triggered. The termination status of the method is store in a variable π_S which is `true` if the former scenario occurs and `false` the latter scenario occurs.

Algorithm 7.4.1: Static D.AICG Method

Require: $\hat{\rho} > 0$, $(m_1, M_1, m_2, M_2) \in \mathbb{R}^4$, $h \in \overline{\text{Conv}}\,(Z)$, $f_1 \in \mathcal{C}_{m_1,M_1}(Z)$, $f_2 \in \mathcal{C}_{m_2,M_2}(Z)$, $(\lambda, \theta) \in \mathbb{R}^2_{++}$ s.t. $\lambda M_1 + \theta^2 < 1/2$, $z_0 \in Z$;
Initialize: $\mu \leftarrow 1$, $L \leftarrow \lambda M_2^+ + 1$(see (7.7)), $\pi_S \leftarrow$ `true`, $y_0 \leftarrow z_0$, $A_0 \leftarrow 0$, $\phi \Leftarrow f_1 + f_2 + h$;

1: **procedure** St.D.AICG($f_1, f_2, h, z_0, \lambda, \theta, M_2, \hat{\rho}$)
2: **for** $k = 1, \ldots$ **do**
3: Part 1 **Attack** the k^{th} prox-linear subproblem.
4: $a_{k-1} \leftarrow \frac{1 + \sqrt{1 + 4A_{k-1}}}{2}$
5: $A_k \leftarrow A_{k-1} + a_{k-1}$

6: $\tilde{y}_{k-1} \leftarrow \dfrac{A_{k-1}z_{k-1} + a_{k-1}y_{k-1}}{A_k}$;

7: $\psi_s^k \Leftarrow \lambda\left[\ell_{f_1}(\cdot; z_{k-1}) + f_2\right] + \dfrac{1}{2}\|\cdot - z_{k-1}\|^2$

8: $(z_k^a, v_k, \varepsilon_k, \pi_k^{\text{acg}}) \leftarrow \text{S.ACG}(\psi_s^k, \lambda h, \tilde{y}_{k-1}, \theta, \mu, L)$

9: PART 2 **Check** a special convexity condition.

10: **if** $\neg(\pi_k^{\text{acg}})$ **or** $\Delta_\mu(z_{k-1}; z_k^a, v_k) > \varepsilon_k$ **then** ▷ See (7.10)

11: $\pi_S \leftarrow$ `false`

12: **return** (z_0, ∞, π_S)

13: PART 3 **Check** the termination condition.

14: $(\hat{z}_k, \hat{v}_k) \leftarrow \text{SREF}(f_1, f_2, h, z_k, \tilde{y}_{k-1}, v_k, M_2, \lambda)$

15: **if** $\|\hat{v}_k\| \le \hat{\rho}$ **then**

16: **return** $(\hat{z}_k, \hat{v}_k, \pi_S)$

17: PART 4 **Compute** an accelerated prox step.

18: $y_k \leftarrow \underset{u\in\Omega}{\text{argmin}} \dfrac{1}{2}\left\|u - \left[y_{k-1} - a_{k-1}\left(v_k + \tilde{y}_{k-1} - z_k^a\right)\right]\right\|^2$

19: $z_k \leftarrow \underset{u\in\{z_{k-1}, z_k^a\}}{\text{argmin}} \phi(u)$

Some remarks about this method are in order. To ease the discussion, let us refer to the ACG iterations performed in Line 8 of the method as **inner iterations** and the iterations over the indices k as **outer iterations**. First, similar to the static AICGM, the static D.AICGM may fail without obtaining a pair that solves Problem 7.1.1. Theorem 7.4.1(c) shows that a sufficient condition for the method to stop successfully is that f_2 be convex. Using arguments similar to the ones employed to derive the dynamic AICG variant, a dynamic D.AICG variant can also be developed that repeatedly invokes the static D.AICGM in place of the static AICGM. Second, in view of the update for z_k in Line 19, the collection of function values $\{\phi(z_i)\}_{i=0}^k$ is non-increasing. Third, in view of the requirement on (λ, θ), if $M_1 > 0$ then $0 < \lambda < (1 - 2\theta^2)/(2M_1)$ whereas if $M_1 \le 0$ then $0 < \lambda < \infty$.

The next result summarizes some facts about the D.AICGM. Before proceeding, we in-

troduce the useful constants

$$D_z \coloneqq \sup_{u,z\in Z} \|u-z\|, \quad D_\Omega \coloneqq \sup_{u,z\in\Omega} \|u-z\|, \quad \Delta_\phi^0 \coloneqq \phi(z_0)-\phi_*,$$
$$d_0 \coloneqq \inf_{u^*\in\mathcal{Z}} \{\|z_0-u^*\| : \phi(u^*)=\phi_*\}, \quad E_{\lambda,\theta} \coloneqq \sqrt{\lambda}L_1 + \frac{1+\theta\overline{C}_\lambda}{\sqrt{\lambda}}. \tag{7.25}$$

Theorem 7.4.1. *The following statements hold about the static D.AICGM:*

(a) it stops in

$$\mathcal{O}\left(\frac{E_{\lambda,\theta}^2[m_1^+ D_z^2+\Delta_\phi^0]}{\hat{\rho}^2} + \frac{E_{\lambda,\theta}[m_1^+ + 1/\lambda]^{1/2} D_\Omega}{\hat{\rho}}\right) \tag{7.26}$$

outer iterations;

(b) if it stops with π_S = `true`, then the first two arguments of its output triple $(\hat{z},\hat{v},\pi_S)$ solve Problem 7.1.1;

(c) if f_2 is convex, then it always stops with π_S = `true` in

$$\mathcal{O}\left(\frac{E_{\lambda,\theta}^2 m_1^+ D_z^2}{\hat{\rho}^2} + \frac{E_{\lambda,\theta}[m_1^+]^{1/2} D_\Omega}{\hat{\rho}} + \frac{E_{\lambda,\theta}^{2/3} d_0^{2/3} \lambda^{-1/3}}{\hat{\rho}^{2/3}}\right) \tag{7.27}$$

outer iterations.

We now make three remarks about the above results. First, in the "best" scenario of $\max\{m_1,m_2\}\le 0$, we have that (7.27) reduces to

$$\mathcal{O}\left(\left[L_1+\frac{1}{\lambda}\right]^{2/3}\left[\frac{d_0^{2/3}}{\hat{\rho}^{2/3}}\right]\right),$$

which has a smaller dependence on $\hat{\rho}$ when compared to (7.21). In the "worst" scenario of $\min\{m_1,m_2\}>0$, if we take $\theta=\mathcal{O}(1/\overline{C}_\lambda)$, then (7.26) reduces to

$$\mathcal{O}\left(\left[\sqrt{\lambda}L_1+\frac{1}{\sqrt{\lambda}}\right]^2\left[\frac{m_1^+ D_z^2+\phi(z_0)-\phi_*}{\hat{\rho}^2}\right]\right),$$

which has the same dependence on $\hat{\rho}$ as in (7.21). Second, part (c) shows that if the method stops with an output pair $(\hat{z}, \hat{v})$, regardless of the convexity of f_2, then that pair is always an approximate solution of $\mathcal{NCO}_2$. Third, Proposition 7.4.9 shows that the quantities L_1 and $\overline{C}_\lambda$ in all the previous complexity results can be replaced by their averaged counterparts in (7.43). As these averaged quantities only depend on $\{(z_i^a, \hat{z}_i, \tilde{y}_{i-1})\}_{i=1}^k$, we can infer that the static D.AICGM, like the static AICGM of the previous subsection, also adapts to the local geometry of its input functions.

7.4.1 D.AICG Properties and Iteration Complexity

This subsection establishes several key properties of static D.AICGM and gives the proof of Theorem 7.4.1.

To avoid repetition, we assume throughout this subsection that $k \geq 1$ denotes an arbitrary successful outer iteration of the D.AICGM and let

$$\{(a_i, A_i, z_i, z_i^a, y_i, \tilde{y}_{i-1}, \hat{z}_i, \hat{v}_i, v_i, \varepsilon_i)\}_{i=1}^k$$

denote the sequence of all iterates generated by it up to and including the k^{th} iteration. Observe that this implies that the i^{th} D.AICG outer iteration for any $1 \leq i \leq k$ has π_S = `true`, i.e. the (only) S.ACG call in this iteration does not stop with π_i^{acg} = `false` and $\Delta_1(z_{i-1}; z_i^a, v_i) \leq \varepsilon_i$. Moreover, throughout this subsection we let

$$\widetilde{\gamma}_i(u) = \ell_{f_1}(u; \tilde{y}_{i-1}) + f_2(u) + h(u), \quad \gamma_i(u) = \widetilde{\gamma}_i(z_i^a) + \frac{1}{\lambda}\langle v_i + \tilde{y}_{i-1} - z_i^a, u - z_i^a \rangle. \quad (7.28)$$

The first set of results present some basic properties about the functions $\widetilde{\gamma}_i$ and γ_i as well as the iterates generated by the method.

Lemma 7.4.2. *The following statements hold for any $s \in Z$ and $1 \leq i \leq k$:*

(a) $\gamma_i(z_i^a) = \widetilde{\gamma}_i(z_i^a)$;

(b) $y_i = \operatorname{argmin}_{u\in\Omega}\{\lambda a_{i-1}\gamma_i(u) + \|u - y_{i-1}\|^2/2\}$;

(c) $z_i^a - v_i = \operatorname{argmin}_{u\in\mathcal{Z}}\{\lambda\gamma_i(u) + \|u - \tilde{y}_{i-1}\|^2/2\}$;

(d) $-M_1\|u - \tilde{y}_{i-1}\|^2/2 \le \widetilde{\gamma}_i(u) - \phi(u) \le m_1\|u - \tilde{y}_{i-1}\|^2/2$;

(e) $\phi(z_{i-1}) \ge \phi(z_i)$ *and* $\phi(z_i^a) \ge \phi(z_i)$.

Proof. To keep the notation simple, denote

$$\begin{gathered}(z_+^a, z_+, z, \tilde{y}) = (z_i^a, z_i, z_{i-1}, \tilde{y}_{i-1}), \quad (y_+, y) = (y_i, y_{i-1}),\\ (A_+, A, a) = (A_i, A_{i-1}, a_{i-1}), \quad (v, \varepsilon) = (v_i, \varepsilon_i).\end{gathered} \tag{7.29}$$

(a) This is immediate from the definitions of γ and $\widetilde{\gamma}$ in (7.28).

(b) Define $\widehat{y}_i := y_{k-1} - a_{k-1}\left(v_k + \tilde{y}_{k-1} - z_k^a\right)$. Using the definition of γ in (7.28), we have that

$$\begin{aligned}\operatorname*{argmin}_{u\in\Omega}\left\{\lambda a\gamma(u) + \frac{1}{2}\|u - y\|^2\right\} &= \operatorname*{argmin}_{u\in\Omega}\left\{a\langle v + \tilde{y} - z_+^a, u - x\rangle + \frac{1}{2}\|u - y\|^2\right\}\\ &= \operatorname*{argmin}_{u\in\Omega}\frac{1}{2}\left\|u - \left(y - a\left[v + \tilde{y} - z_+^a\right]\right)\right\|^2\\ &= \operatorname*{argmin}_{u\in\Omega}\frac{1}{2}\left\|u - \widehat{y}_+\right\|^2 = y_+.\end{aligned}$$

(c) Using the definition of γ in (7.28), we have that

$$\lambda\nabla\gamma\left(z_+^a - v\right) + \left(z_+^a - v\right) - \tilde{y} = \left(v + \tilde{y} - z_+^a\right) + \left(z_+^a - v\right) - \tilde{y} = 0,$$

and hence, the point $z_+^a - v$ is the global minimum of $\lambda\gamma + \|\cdot - \tilde{y}\|^2/2$.

(d) This follows from the fact that $f_1 \in \mathcal{C}_{m_1,M_1}(Z)$ and the definition of $\widetilde{\gamma}$ in (7.28).

(e) This follows immediately from the update rule of z_i in Line 19 of the D.AICGM. □

Lemma 7.4.3. *Let $w = \tilde{y}_{i-1}$ and $\Delta_1(\cdot;\cdot,\cdot)$ be as in* (7.10) *with*

$$\psi_s = \lambda\left[\ell_{f_1}(\cdot; z_{k-1}) + f_2\right] + \frac{1}{2}\|\cdot - z_{k-1}\|^2, \quad \psi_n = \lambda h. \tag{7.30}$$

Then, following statements hold:

(a) the triple $(z_i^a, v_i, \varepsilon_i)$ solves Problem $\mathcal{B}$ and satisfies $\Delta_1(z_{i-1}; z_i^a, v_i) \le \varepsilon$, and hence

$$\|v_i\| + 2\varepsilon_i \le \sigma^2\|z_i^a - \tilde{y}_{i-1}\|^2, \quad \Delta_1(u; z_i^a, v_i) \le \varepsilon_i \quad \forall u \in \{\hat{z}_i, z_{i-1}\}, \tag{7.31}$$

(b) if f_2 is convex, then $(z_i^a, v_i, \varepsilon_i)$ solves Problem $\mathcal{A}$;

(c) $\Delta_1(s; z_i^a, v_i) = \lambda[\gamma_i(s) - \widetilde{\gamma}_i(s)]$;

(d) $\Delta_1(z_i; z_i^a, v_i) \le \varepsilon$.

Proof. (a) This follows from Line 10 of the D.AICGM and Proposition 7.2.3(b).

(b) This follows from the S.ACG call in Line 8 of the D.AICGM, the fact that h is convex, and Proposition 7.2.3(c) with $\psi_s = \widetilde{\gamma}_i + \|\cdot - \tilde{y}_{i-1}\|^2/2$.

(c) Using the definitions of (ψ_s, ψ_n) and $(\gamma, \widetilde{\gamma})$, we have that

$$\begin{aligned}
\Delta_1(s; z_+^a, v) &= (\psi_s + \psi_n)(z_+^a) - (\psi_s + \psi_n)(s) - \langle v, z_+^a - s\rangle + \frac{1}{2}\|s - z_+^a\|^2 \\
&= \left[\lambda\widetilde{\gamma}(z_+^a) + \frac{1}{2}\|z_+^a - \tilde{x}\|^2\right] - \left[\lambda\widetilde{\gamma}(s) + \frac{1}{2}\|s - \tilde{x}\|^2\right] - \langle v, z_+^a - s\rangle + \frac{1}{2}\|s - z_+^a\|^2 \\
&= \left[\lambda\gamma(s) + \frac{1}{2}\|s - \tilde{x}\|^2\right] - \left[\lambda\widetilde{\gamma}(s) + \frac{1}{2}\|s - \tilde{x}\|^2\right] \\
&= \lambda\gamma(s) - \lambda\widetilde{\gamma}(s).
\end{aligned}$$

(d) If $z_i = z_{i-1}$, then this follows from Line 10 of the method. On the other hand, if $z_i = z_i^a$, then this follows from part (c). □

We now state some well-known (see, for example, Lemma B.0.2 with $\lambda_k = \tau_k = 1$) properties of A_i and a_{i-1}.

Lemma 7.4.4. *For every $1 \le i \le k$, we have that:*

(a) $a_{i-1}^2 = A_i$;

(b) $i^2/4 \le A_i \le i^2$.

The next two lemmas are technical results that are needed to establish the key inequality in Proposition 7.4.7.

Lemma 7.4.5. *For every $u \in Z$ and $1 \le i \le k$, we have that*

$$\frac{1}{2}\left(A_{i-1}\|z_{i-1} - \tilde{y}_{i-1}\|^2 + a_{i-1}\|u - \tilde{y}_{i-1}\|^2\right) \le 2D_\Omega^2 + a_{i-1}D_z^2.$$

Proof. Throughout the proof, we use the notation in (7.29). Using the relation $(p+q)^2 \le 2p^2 + 2q^2$ for every $p, q \in \mathbb{R}$, Lemma 7.4.4(a), the fact that $A \le A^+$, $x \in \Omega$, and $y \in Z$, and the definition of $\tilde{y}$, and the definitions of D_Ω and D_z in (7.25), we conclude that

$$\begin{aligned}
A\|z - \tilde{y}\|^2 + a\|u - \tilde{y}\|^2 &= A\left\|\frac{a}{A_+}(z-y)\right\|^2 + a\left\|\frac{A}{A_+}(u-z) + \frac{a}{A_+}(u-y)\right\|^2 \\
&\le \frac{A}{A_+}\left(\|(z-u)+(u-y)\|^2 + 2a\left[\frac{A^2}{A_+^2}\|u-z\|^2 + \frac{a^2}{A_+^2}\|u-y\|^2\right]\right) \\
&\le \frac{2A}{A^+}\left(\|u-z\|^2 + \|u-y\|^2\right) + 2a\|u-z\|^2 + \frac{2a}{A_+}\|u-y\|^2 \\
&\le 2\left[\|u-x\|^2 + (1+a)\|u-y\|^2\right] \\
&\le 2\left[D_\Omega^2 + (1+a)D_z^2\right].
\end{aligned}$$

The conclusion now follows from dividing both sides of the above inequalities by 2 and using the fact that $D_z \le D_\Omega$. □

Lemma 7.4.6. *For every $u \in Z$ and $1 \le i \le k$, we have that*

$$\begin{aligned}
&A_i\left[\phi(z_i) + \left(\frac{1-\lambda M_1}{2\lambda}\right)\|z_i^a - \tilde{y}_{i-1}\|^2 - \frac{\|v_i\|^2}{2\lambda}\right] + \frac{1}{2\lambda}\|u - y_i\|^2 \\
&\le A_{i-1}\gamma_i(z_{i-1}) + a_{i-1}\gamma_i(u) + \frac{1}{2\lambda}\|u - y_{i-1}\|^2. \qquad (7.32)
\end{aligned}$$

Proof. Throughout the proof, we use the notation in (7.29). We first present two key expressions. First, using the definition of γ in (7.28) and Lemma 7.4.2(c), it follows that

$$\begin{aligned}\min_{u\in\mathcal{Z}}\left\{\lambda\gamma\left(u\right)+\frac{1}{2}\|u-\tilde{y}\|^2\right\} &= \lambda\widetilde{\gamma}(z_+^a)-\left\langle v+\tilde{y}-z_+^a,v\right\rangle+\frac{1}{2}\left\|v+\tilde{y}-z_+^a\right\|^2\\ &= \lambda\widetilde{\gamma}(z_+^a)-\|v\|^2-\left\langle v,\tilde{y}-z_+^a\right\rangle+\frac{1}{2}\left\|v+\tilde{y}-z_+^a\right\|^2\\ &= \lambda\widetilde{\gamma}(z_+^a)-\frac{1}{2}\|v\|^2+\frac{1}{2}\|\tilde{y}-z_+^a\|^2. \end{aligned}\tag{7.33}$$

Second, Lemma 7.4.2(b) and the fact that the function $a\gamma+\|\cdot-y\|^2/(2\lambda)$ is $(1/\lambda)$-strongly convex imply that

$$a\gamma\left(y_+\right)+\frac{1}{2\lambda}\|y_+-y\|^2\le a\gamma\left(u\right)+\frac{1}{2\lambda}\|u-y\|^2-\frac{1}{2\lambda}\|u-y_+\|^2.\tag{7.34}$$

Using (7.33), Lemma 7.4.2(d)–(e), Lemma 7.4.4(a), and the fact that γ is affine, we have that

$$\begin{aligned}&A_+\left[\phi(z_+)+\left(\frac{1-\lambda M_1}{2\lambda}\right)\|z_+^a-\tilde{y}\|^2\right]\\ &\le A_+\left[\widetilde{\gamma}\left(z_+^a\right)+\frac{1}{2\lambda}\|z_+^a-\tilde{y}\|^2\right] &(7.35)\\ &= A_+\left[\min_{u\in\mathcal{Z}}\left\{\gamma\left(u\right)+\frac{1}{2\lambda}\|u-\tilde{y}\|^2\right\}+\frac{\|v\|^2}{2\lambda}\right]\\ &\le A_+\left[\gamma\left(\frac{Az+ay_+}{A_+}\right)+\frac{1}{2\lambda}\left\|\frac{Az+ay_+}{A_+}-\frac{Az+ay}{A_+}\right\|^2+\frac{\|v\|^2}{2\lambda}\right]\\ &= A\gamma\left(z\right)+a\gamma\left(y_+\right)+\frac{a^2}{2\lambda A_+}\|y-y_+\|^2+\frac{A_+}{2\lambda}\|v\|^2\\ &= A\gamma\left(z\right)+a\gamma\left(y_+\right)+\frac{1}{2\lambda}\|y-y_+\|^2+\frac{A_+}{2\lambda}\|v\|^2 &(7.36)\end{aligned}$$

The conclusion now follows from combining (7.34) with (7.36). □

We now present an inequality that plays an important role in the analysis of the D.AICGM.

Proposition 7.4.7. *Let $\Delta_1(\cdot;\cdot,\cdot)$ be as in* (7.10) *with (ψ_s, ψ_n) as in* (7.30)*, and define*

$$\theta_i(u) := A_i\left[\phi(z_i) - \phi(u)\right] + \frac{1}{2\lambda}\|u - y_i\|^2 \quad \forall i \geq 0. \tag{7.37}$$

For every $u \in Z$ satisfying $\Delta_1(u; z_i^a, v_i) \leq \varepsilon$ and $1 \leq i \leq k$, we have that

$$\frac{A_i}{4\lambda}\|z_i^a - \tilde{y}_{i-1}\|^2 \leq m_1^+\left(a_{i-1}D_z^2 + 2D_\Omega^2\right) + \theta_{i-1}(u) - \theta_i(u). \tag{7.38}$$

Proof. Throughout the proof, we use the notation in (7.29) together with the notation $\theta = \theta_{i-1}$ and $\theta_+ = \theta_i$. Let $u \in \operatorname{dom} h$ be such that $\Delta_1(u; z_+^a, v) \leq \varepsilon$. Subtracting $A\phi(u)$ from both sides of the inequality in (7.32) and using the definition of θ_+ we have

$$\begin{aligned}
&\frac{A_+}{2\lambda}\left[(1-\lambda M_1)\|z_+^a - \tilde{y}\|^2 - \|v\|^2\right] + \theta_+(u)\\
&= \frac{A_+}{2\lambda}\left[(1-\lambda M_1)\|z_+^a - \tilde{y}\|^2 - \|v\|^2\right] + A_+\left[\phi(z_+) - \phi(u)\right] + \frac{1}{2\lambda}\|u - z_+^a\|^2\\
&\leq A\gamma(z) + a\gamma(u) - A\phi(u) + \frac{1}{2\lambda}\|u - y\|^2\\
&= a\left[\gamma(u) - \phi(u)\right] + A\left[\gamma(z) - \phi(z)\right] + \theta(u).
\end{aligned} \tag{7.39}$$

Moreover, using Lemma 7.4.3(a) and (c), and with our assumption that $\Delta_1(u; z_+^a, v) \leq \varepsilon$, we have that

$$\gamma(s) - \phi(s) = \widetilde{\gamma}(s) - \phi(s) + \frac{\Delta_1(s; z_+^a, v)}{\lambda} \leq \frac{m_1^+}{2}\|s - \tilde{y}\|^2 + \frac{\varepsilon}{\lambda} \quad \forall s \in \{u, z\}. \tag{7.40}$$

Combining (7.39), (7.40), and Lemma 7.4.5 then yields

$$\begin{aligned}
&\frac{A_+}{2\lambda}\left[(1-\lambda M_1)\|z_+^a - \tilde{y}\|^2 - \|v\|^2\right] + \theta_+(u)\\
&\leq \frac{m_1^+}{2}\left[a\|u - \tilde{y}\|^2 + A\|z - \tilde{y}\|^2\right] + \frac{\varepsilon A_+}{\lambda} + \theta(u)\\
&\leq m_1^+\left(aD_z^2 + 2D_\Omega^2\right) + \frac{\varepsilon A_+}{\lambda} + \theta(u).
\end{aligned}$$

Re-arranging the above terms and using the restriction on (λ, θ) (in the D.AICGM) together with the first inequality in (7.31), we conclude that

$$\begin{aligned}
&m_1^+ \left(aD_h^2 + 2D_\Omega^2\right) + \theta(u) - \theta_+(u) \\
&\geq \frac{A_+}{2\lambda}\left[(1-\lambda M_1)\|z_+^a - \tilde{y}\|^2 - \|v\|^2 - 2\varepsilon\right] \\
&\geq \frac{A_+(1-\lambda M_1 - \sigma^2)}{2\lambda}\|z_+^a - \tilde{y}\|^2 \\
&\geq \frac{A_+}{4\lambda}\|z_+^a - \tilde{y}\|^2.
\end{aligned}$$

□

The following result describes some important technical bounds obtained by summing (7.38) for two different choices of u (possibly changing with i) from $i = 1$ to k.

Proposition 7.4.8. *Let Δ_ϕ^0 and d_0 be as in* (7.25) *and define*

$$S_k := \frac{1}{4\lambda}\sum_{i=1}^{k} A_i\|z_i^a - \tilde{y}_{i-1}\|^2. \tag{7.41}$$

Then, the following statements hold:

(a) $S_k = \mathcal{O}_1(k^2[m_1^+ D_z^2 + \Delta_\phi^0] + k[m_1^+ + 1/\lambda]D_\Omega^2)$;

(b) if f_2 is convex, then $S_k = \mathcal{O}_1(k^2 m_1^+ D_z^2 + k m_1^+ D_\Omega^2 + d_0^2/\lambda)$.

Proof. (a) Let $\Delta_1(\cdot;\cdot,\cdot)$ be defined as in (7.10) with (ψ_s, ψ_n) given by (7.30). Using (7.37), the fact that $y_i, z_i^a \in \Omega$, the fact that A_i is nonnegative and increasing, and the definitions of θ_i and D_Ω in (7.37) and (7.25), respectively, we have that

$$\begin{aligned}
\sum_{i=1}^{k}[\theta_{i-1}(z_i) - \theta_i(z_i)] &\leq \sum_{i=1}^{k} A_{i-1}[\phi(z_{i-1}) - \phi(z_i)] + \frac{1}{2\lambda}\sum_{i=1}^{k}\|z_i - y_{i-1}\|^2 \\
&\leq A_k\sum_{i=1}^{k}[\phi(z_{i-1}) - \phi(z_i)] + \frac{k}{2\lambda}D_\Omega^2 \\
&\leq A_k[\phi(z_0) - \phi_*] + \frac{k}{2\lambda}D_\Omega^2.
\end{aligned} \tag{7.42}$$

Moreover, noting Lemma 7.4.3(d) and using Proposition 7.4.7 with $u = y_i$, we conclude that (7.38) holds with $u = y_i$ for every $1 \le i \le k$. Summing these k inequalities and using (7.42), the definition of S_k in (7.41), and Lemma 7.4.4(b) yields the desired conclusion.

(b) Assume now that f_2 is convex and let z_* be a point such that $\phi(z_*) = \phi_*$ and $\|z_0 - z_*\| = d_0$. It then follows from Lemma 7.4.3(b) and Lemma 7.2.1(d) with $(z, v) = (z_i^a, v_i)$ that $\Delta_1(z_*; z_i^a, v_i) \le \varepsilon$ for every $1 \le i \le k$. The conclusion now follows by using an argument similar to the one in (a) but which instead sums (7.38) with $u = z_*$ from $i = 1$ to k, and uses the fact that

$$\sum_{i=1}^{k} [\theta_{i-1}(z_*) - \theta_i(z_*)] = \theta_0(z_*) - \theta_k(z_*) \le \frac{1}{2\lambda}\|z_0 - z_*\|^2 = \frac{d_0}{2\lambda},$$

where the inequality is due to the fact that $\theta_k(z_*) \ge 0$ (see (7.37)) and $A_0 = 0$. □

We now establish the rate at which the residual $\|\hat{v}_i\|$ tends to 0.

Proposition 7.4.9. *Let S_k be as in* (7.41)*. Moreover, define the quantities*

$$L_{1,k}^{\text{avg}} := \frac{1}{k}\sum_{i=1}^{k} L_1(z_i^a, \tilde{y}_{i-1}), \quad C_{\lambda,k}^{\text{avg}} := \frac{1}{k}\sum_{i=1}^{k} C_\lambda(\hat{z}_i, z_i^a), \tag{7.43}$$

where $C_\lambda(\cdot,\cdot)$ and $\overline{C}_\lambda$ are as in (7.15) *and* (7.19)*, respectively. Then, we have*

$$\min_{i \le k} \|\hat{v}_i\| = \mathcal{O}_1\left(\left[\sqrt{\lambda}L_{1,k}^{\text{avg}} + \frac{1 + \theta C_{\lambda,k}^{\text{avg}}}{\sqrt{\lambda}}\right]\left[\frac{S_k}{k^3}\right]^{1/2}\right) + \frac{\hat{\rho}}{2}.$$

Proof. Let $\ell = \lceil k/2 \rceil$. Using Proposition 7.2.2 with $(z, w) = (z_i^a, \tilde{y}_{i-1})$ and the bounds $C_\lambda(\cdot,\cdot) \le \overline{C}_\lambda$ and $L_1(\cdot,\cdot) \le L_1$ we have that $\|\hat{v}_i\| \le \mathcal{E}_i\|z_i^a - \tilde{y}_{i-1}\|$, for every $\ell \le i \le k$, where

$$\mathcal{E}_i = \frac{2 + \lambda L_1(z_i^a, \tilde{y}_{i-1}) + \theta C_\lambda(\hat{z}_i, z_i^a)}{\lambda} \quad \forall i \ge 1.$$

As a consequence, using the definition of S_k in (7.41), the definitions in (7.43), Lemma 7.3.3 with $p = 3/2$, $a_i = \mathcal{E}_i/\sqrt{A_i}$, and $b_i = \sqrt{A_i}\|z_i^a - \tilde{y}_{i-1}\|$ for $i \in \{\ell, \ldots, k\}$, Lemma 7.4.4(b), and

the fact that $(k - \ell + 1) \geq k/2$, yields

$$\begin{aligned}
\min_{\ell \leq i \leq k} \|\hat{v}_i\| &\leq \min_{\ell \leq i \leq k} \mathcal{E}_i \|z_i^a - \tilde{y}_{i-1}\| \\
&\leq \frac{1}{(k-\ell+1)^{3/2}} \left(\sum_{i=\ell}^{k} \frac{\mathcal{E}_i}{\sqrt{A_i}} \right) \left(\sum_{i=\ell}^{k} A_i \|z_i^a - \tilde{y}_{i-1}\|^2 \right)^{1/2} \\
&\leq \frac{2^{3/2}}{k^{3/2}} \left(\frac{2}{k} \sum_{i=1}^{k} \mathcal{E}_i \right) (4\lambda S_k)^{1/2} \\
&= \mathcal{O}_1 \left(\left[\sqrt{\lambda} L_{1,k}^{\text{avg}} + \frac{1 + \theta C_{\lambda,k}^{\text{avg}}}{\sqrt{\lambda}} \right] \left[\frac{S_k}{k^3} \right]^{1/2} \right).
\end{aligned}$$

□

We are now ready to give the proof of Theorem 7.4.1.

Proof of Theorem 7.4.1. (a) This follows from Proposition 7.4.9, Proposition 7.4.8(a), the fact that $C_\lambda(\cdot,\cdot) \leq \overline{C}_\lambda$ and $L_{f_1}(\cdot,\cdot) \leq L_1$, and the termination condition in Line 15 of the D.AICGM.

(b) The fact that $(\hat{z}, \hat{v}) = (\hat{z}_k, \hat{v}_k)$ satisfies the inclusion of (7.5) follows from Proposition 7.2.2 with $(z, v, z_0) = (z_k^a, v_k, \tilde{y}_{k-1})$. The fact that $\|\hat{v}\| \leq \hat{\rho}$ follows the termination condition in Line 15 of the D.AICGM.

(c) The fact that the method does not stop with $\pi_S = \texttt{false}$ follows from Proposition 7.2.3(c). The bound in (7.27) follows from a similar argument as in part (a) except that Proposition 7.4.8(a) is replaced with Proposition 7.4.8(b). □

7.5 Exploiting the Spectral Decomposition

Recall that at every outer iteration of the ICG methods in the previous sections, a call to the S.ACG algorithm is made to tentatively solve the Problem $\mathcal{B}$ (see Section 7.2) associated with (7.6). Our goal in this section is to present a significantly more efficient ACG variant (based on the idea outlined at the beginning of this chapter) for solving the same problem when the underlying problem of interest is $\mathcal{SNCO}$.

Throughout our presentation, we make use of the functions $\mathrm{dg} : \mathbb{R}^r \mapsto \mathbb{R}^{r\times r}$ and $\mathrm{Dg} : \mathbb{R}^{m\times n} \mapsto \mathbb{R}^r$ given pointwise by

$$[\mathrm{dg}\, z]_{ij} = \begin{cases} z_i, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \qquad [\mathrm{Dg}\, Z]_i = Z_{ii}, \tag{7.44}$$

for every $z \in \mathbb{R}^r$, $Z \in \mathbb{R}^{m\times n}$, and $(i, j) \in \{1, ..., r\}^2$.

The content of this section is divided into two subsections. The first one presents the aforementioned algorithm, whereas the second one proves its key properties.

7.5.1 Spectral ACG Method

This subsection presents an efficient spectral ACG method (σ.ACGM), which utilizes the S.ACGM of Section 7.2, for solving the Problem $\mathcal{B}$ associated with (7.6).

Throughout our presentation, we let Z_0 represent the starting point given to the S.ACGM by the two ICG methods. Moreover, we assume that we have a method `SVD(...)` that returns a triple $(P, \sigma(Z), Q)$ representing the SVD of its input Z. More specifically, if `(P, s, Q) ← SVD(Z)` then it holds that $Z = P[\mathrm{dg}\, s]Q^*$.

We now state the σ.ACGM in Algorithm 7.5.1, which uses the S.ACGM of Section 7.2 and the aforementioned SVD method as subroutines.

Algorithm 7.5.1: σ.ACG Method

Require: $M_2 \in \mathbb{R}_{++}$, $h^{\mathcal{V}} \in \mathcal{C}(\mathbb{R}^r)$, $f_1 \in \mathcal{C}(\mathrm{dom}[h^{\mathcal{V}} \circ \sigma])$, $f_2^{\mathcal{V}} \in \mathcal{C}_{M_2}(\mathrm{dom}\, h)$, $Z_0 \in \mathbb{R}^{m\times n}$, $\theta \in (0, 1)$;
Initialize: $\mu \leftarrow 1$, $L \leftarrow \lambda M_2 + 1$, $\pi_S \leftarrow$ `true`, $\psi_n^{\mathcal{V}} \leftarrow \lambda h^{\mathcal{V}}$

1: **procedure** σ.ACG($f_1, f_2^{\mathcal{V}}, h^{\mathcal{V}}, Z_0, \theta, \mu, L$)
2: PART 1 **Attack** a vectorized prox-linear subproblem using the S.ACGM.
3: $Z_0^{\lambda} \leftarrow Z_0 - \lambda \nabla f_1(Z_0)$
4: $(P, s, Q) \leftarrow \mathrm{SVD}(Z_0^{\lambda})$

5: $\psi_s^{\mathcal{V}} \Leftarrow \lambda f_2^{\mathcal{V}} - \langle s, \cdot \rangle + \frac{1}{2}\|\cdot\|^2$

6: $(z, v, \varepsilon, \pi_S) \leftarrow \text{S.ACG}(\psi_s^{\mathcal{V}}, \psi_n^{\mathcal{V}}, \text{Dg}(P^* Z_0 Q), \theta, \mu, L$

7: PART 1 **Terminate** based on the status of the S.ACGM call

8: **if** π_S **then**

9: $Z \leftarrow P(\text{dg}\, z)Q^*$

10: $V \leftarrow P(\text{dg}\, v)Q^*$

11: **return** $(Z, V, \varepsilon, \pi_S)$

12: **else**

13: **return** $(Z_0, \infty, \infty, \pi_S)$

We now make two remarks about the method. First, since it calls the S.ACGM in Line 6, its iteration complexity is the same as the one given for the S.ACGM, i.e. as in Proposition 7.2.3(a). Second, because the functions $\psi_s^{\mathcal{V}}$ and $\psi_n^{\mathcal{V}}$ used in its S.ACG call have vector inputs over $\mathbb{R}^r$, the steps in the σ.ACGM are significantly less costly than the ones in an analogous S.ACGM call, which use functions with matrix inputs over $\mathbb{R}^{m\times n}$.

The following result, whose proof is deferred to the next subsection, presents the key properties of the σ.ACGM.

Proposition 7.5.1. *Let $(Z, V, \varepsilon, \pi_S)$ be the output of a call to the σ.ACGM. Then, the following properties hold:*

(a) if $\pi_S =$ `true`, then the triple (Z, V, ε) solves the Problem $\mathcal{B}$ associated with (7.6)*;*

(b) if f_2 is convex, then $\pi_S =$ `true` and the triple (Z, V, ε) solves the Problem $\mathcal{A}$ associated with (7.6)*.*

7.5.2 Proof of Proposition 7.5.1

This subsection gives the proof of Proposition 7.5.1.

Let the quantities (P, Q) and $(\psi_s^{\mathcal{V}}, \psi_n^{\mathcal{V}})$ be generated by a call of the σ.ACGM. Moreover,

for every $(u, U) \in \mathbb{R}^r \times \mathbb{R}^{m\times n}$, define the functions

$$\begin{gathered} f_2(U) := f_2^{\mathcal{V}} \circ \sigma(U), \quad h := h^{\mathcal{V}} \circ \sigma, \quad \psi^{\mathcal{V}}(u) := \psi_s^{\mathcal{V}}(u) + \psi_n^{\mathcal{V}}(u) \\ \mathcal{M}(u) := P(\operatorname{dg} u)Q^*, \quad \mathcal{V}(U) := \operatorname{Dg}(P^*UQ). \end{gathered} \tag{7.45}$$

The result below relates the function triple $(\psi_s^{\mathcal{V}}, \psi_n^{\mathcal{V}}, \psi^{\mathcal{V}})$ to the function triple (ψ_s, ψ_n, ψ) given by

$$\psi_s := \lambda\left[\ell_{f_1}(\cdot, Z_0) + f_2 \circ \sigma\right] + \frac{1}{2}\|\cdot - Z_0\|^2, \quad \psi_n := \lambda(h \circ \sigma), \quad \psi = \psi_s + \psi_n.$$

Lemma 7.5.2. *Let $(z, v, \varepsilon, \pi_S)$ and (Z, V) be generated by a call to the σ.ACGM in which $\pi_S =$* `true`*. Then, the following properties hold:*

(a) we have

$$\psi_n^{\mathcal{V}}(z) = \psi_n(Z), \quad \psi_s^{\mathcal{V}}(z) + B_0^\lambda = \psi_s(Z),$$

where $B_0^\lambda := \lambda f_1(Z_0) - \lambda\langle \nabla f_1(Z_0), Z_0\rangle + \|Z_0\|_F^2/2$;

(b) we have

$$V \in \partial_\varepsilon\left(\psi - \frac{1}{2}\|\cdot - Z\|_F^2\right)(Z) \iff v \in \partial_\varepsilon\left(\psi^{\mathcal{V}} - \frac{1}{2}\|\cdot - z\|^2\right)(z). \tag{7.46}$$

Proof. (a) The relationship between $\psi_n^{\mathcal{V}}$,and ψ_n is immediate. On the other hand, using the definitions of Z, f_2, and B_0^λ, we have

$$\begin{aligned} \psi_s^{\mathcal{V}}(z) + B_0^\lambda &= \lambda f_2(Z) - \langle Z_0^\lambda, Z\rangle + \frac{1}{2}\|Z\|_F^2 + B_0^\lambda \\ &= \lambda\left[f_2(Z) + f_1(Z_0) + \langle \nabla f_1(Z_0), Z - Z_0\rangle\right] + \frac{1}{2}\|Z - Z_0\|_F^2 \\ &= \psi_s(Z). \end{aligned}$$

(b) Let $S_0 = V + Z_0^\lambda - Z$ and $s_0 = v + \sigma(Z_0^\lambda) - z$, and note that $S_0 = \mathcal{M}(s_0)$. Moreover, in

view of part (a) and the definition of ψ, observe that the left inclusion in (7.46) is equivalent to $S_0 \in \partial_\varepsilon(\lambda[f_2+h])(Z)$. Using this observation, the fact that S_0 and Z have a simultaneous SVD, and Theorem G.0.3 with $(S,s) = (S_0, s_0)$, $\Psi = \lambda[f_2+h]$, and $\Psi^{\mathcal{V}} = \lambda[f_2^{\mathcal{V}}+h^{\mathcal{V}}]$, we have that the left inclusion in (7.46) is also equivalent to $s_0 \in \partial_\varepsilon(\lambda[f_2^{\mathcal{V}} + h^{\mathcal{V}}])(z)$. The conclusion now follows from the observing that the latter inclusion is equivalent to the right inclusion in (7.46). □

We are now ready to give the proof of Proposition 7.5.1.

Proof of Proposition 7.5.1. (a) Let $(z,v) = (\mathcal{V}(Z), \mathcal{V}(V))$ and remark that the successful termination of the algorithm implies that the inequality in (7.9) and (7.17) hold. Using this remark, the fact that $\|V\|_F^2 = \|v\|^2$, and the bound

$$\begin{aligned}\sigma^2\|z_j - z_0\|^2 &= \sigma^2\left(\|z_j\|^2 - 2\langle z_j, \mathcal{V}(z_0)\rangle + \|Z_0\|_F^2\right) + \sigma^2(\|\mathcal{V}(z_0)\|^2 - \|Z_0\|_F^2)\\ &\le \sigma^2\left(\|Z_j\|^2 - 2\langle Z_j, Z_0\rangle + \|Z_0\|_F^2\right) = \sigma^2\|Z_j - Z_0\|_F^2, \end{aligned} \tag{7.47}$$

we then have that the inequality in (7.9) also holds with $(z,v) = (Z,V)$.

To show the corresponding inequality for (7.17), let $L = \lambda M_2 + 1$ and consider the refined quantities

$$\begin{aligned}\hat{Z} &= \operatorname*{argmin}_{U\in\mathbb{R}^{n\times m}}\left\{\ell_{\psi_s}(U;Z) - \langle V, U\rangle + \frac{L}{2}\|U - Z\|^2 + \psi_n(U)\right\}\\ \hat{z} &= \operatorname*{argmin}_{u\in\mathbb{R}^r}\left\{\ell_{\psi_s^{\mathcal{V}}}(u;z) - \langle v, u\rangle + \frac{L}{2}\|u - z\|^2 + \psi_n^{\mathcal{V}}(u)\right\}\end{aligned}$$

as well as the corresponding residuals

$$\begin{aligned}V_r &= V + L(Z - \hat{Z}) + \nabla\psi_s(\hat{Z}) - \nabla\psi_s(Z),\\ v_r &= v + L(z - \hat{z}) + \nabla\psi_s^{\mathcal{V}}(\hat{z}) - \nabla\psi_s^{\mathcal{V}}(z).\end{aligned}$$

Moreover, let $\Delta_1^{\mathcal{V}}(\cdot;\cdot,\cdot)$ be as in (7.10) with $(\psi_s, \psi_n) = (\psi_s^{\mathcal{V}}, \psi_n^{\mathcal{V}})$ and $\Delta_1(\cdot;\cdot,\cdot)$ as in (7.10).

Using Lemma G.0.2 with $\Psi = \psi_n$ and $S = V + MZ - \nabla\psi_s(Z)$ and Lemma G.0.1(b) we have that Z_r, V_r, Z, and V have a simultaneous SVD. As a consequence, it follows from Lemma 7.5.2(a) that

$$\begin{aligned}\varepsilon \geq \Delta_1^{\mathcal{V}}(\hat{z}; z, v) &= \psi^{\mathcal{V}}(z) - \psi^{\mathcal{V}}(\hat{z}) - \langle v, \hat{z} - z\rangle + \frac{1}{2}\|\hat{z} - z\|^2 \\ &= \psi(Z) - \psi(\hat{Z}) - \langle V, \hat{Z} - Z\rangle + \frac{1}{2}\|\hat{Z} - Z\|^2 \\ &= \Delta_1(\hat{Z}; Z, V).\end{aligned}$$

The conclusion now follows from the above and the definition of the specialized refinement procedure in Section 7.2.

(b) This follows from part (a), Proposition 7.2.3(c), and Lemma 7.5.2(b). □

7.6 Numerical Experiments

This section examines the performance of several solvers for finding approximate stationary points of $\mathcal{SNCO}$ where $(f_1, f_2^{\mathcal{V}}, h^{\mathcal{V}})$ satisfy assumptions (H1)–(H3) of Chapter 7 with $(f_2, h) = (f_2^{\mathcal{V}} \circ \sigma, h^{\mathcal{V}} \circ \sigma)$. All experiments are run on Linux 64-bit machines each containing Xeon E5520 processors and at least 8 GB of memory using MATLAB 2020a. It is worth mentioning that the complete code for reproducing the experiments is freely available online[1].

The algorithms benchmarked in this section are as follows.

- **AICG**: an instance of Algorithm 7.3.2 in which $\xi = M_1$, $\lambda = 5/M_1$, $\sigma = (9/10 - \max\{\lambda(M_1 - \xi, 0\})$, the ACG call is replaced by an R.ACG call with $L_0 = \lambda(M/100) + 1$.

- **CG**: an instance of Algorithm 2.2.1 in which $\lambda_k = 1/(M_1 + M_2)$ for every $k \geq 1$.

[1]See the code in `./tests/thesis/` from the GitHub repository https://github.com/wwkong/nc_opt/

- **D.AICG**: an instance of the dynamic version of Algorithm 7.4.1 in which $\xi = M_1$, $\lambda = 5/M_1$, $\sigma = (1/2 - \max\{\lambda(M_1 - \xi, 0\})$, the ACG call is replaced by an R.ACG call with $L_0 = \lambda(M/100) + 1$.

- **AG**: a variant of the AG method described in Section 5.5.1 in which $\{(\alpha_k, \beta_k, \lambda_k)\}_{k\geq 1}$ are as in [30, Corollary 1] with $L_\Psi = M_1 + M_2$.

Given a tolerance $\hat{\rho} > 0$ and an initial point $Z_0 \in Z$, each algorithm in this section seeks a pair $(\hat{Z}, \hat{V}) \in Z \times \mathbb{R}^{m\times n}$ satisfying

$$\hat{V} \in \nabla f_1(\hat{Z}) + \nabla(f_2^{\mathcal{V}} \circ \sigma)(\hat{Z}) + \partial(h^{\mathcal{V}} \circ \sigma)(\hat{Z}),$$
$$\frac{\|\hat{V}\|}{\|\nabla f_1(Z_0) + (f_2^{\mathcal{V}} \circ \sigma)(Z_0)\| + 1} \leq \hat{\rho}.$$

Moreover, each algorithm is given a time limit of either 10800 or 7200 seconds. The bold numbers in each of the tables in this section highlight the algorithm that performed the most efficiently in terms of function value.

7.6.1 Ball-Constrained Matrix Completion

This subsection presents computational results for the ball-constrained matrix (BC-MC) problem in [50]. More specifically, given a quadruple $(\alpha, \beta, \mu, \theta) \in \mathbb{R}^4_{++}$, a data matrix $A \in \mathbb{R}^{m\times n}$, and indices Ω, this subsection considers the BC-MC problem

$$\min_{U\in\mathbb{R}^{m\times n}} \frac{1}{2}\|P_\Omega(U - A)\|_F^2 + \kappa_\mu \circ \sigma(U) + \tau_\alpha \circ \sigma(U)$$
$$\text{s.t. } \|U\|_F^2 \leq \sqrt{mn} \cdot \max_{i,j} |A_{ij}|,$$

where P_Ω is the linear operator that zeros out any entry that is not in Ω and

$$\kappa_\mu(z) = \frac{\mu\beta}{\theta}\sum_{i=1}^{n} \log\left(1 + \frac{|z_i|}{\theta}\right), \quad \tau_\alpha(z) = \alpha\beta\left[1 - \exp\left(-\frac{\|z\|_2^2}{2\theta}\right)\right]$$

for every $z \in \mathbb{R}^n$. Here, the function $\kappa_\mu + \tau_\alpha$ is a nonconvex generalization of the convex elastic net regularizer [105], and it is well-known [112] that the function $\kappa_\mu - \mu\|\cdot\|_*$ is concave, differentiable, and has a $(2\beta\mu/\theta^2)$-Lipschitz continuous gradient.

We now describe the different data matrices that are considered. Each matrix $A \in \mathbb{R}^{m\times n}$ is obtained from a different collaborative filtering system where each row represents a unique user, each column represents a unique item, and each entry represents a particular rating. Table 7.1 lists the names of each data set, where the data originates from (in the footnotes), and some basic statistics about the matrices.

Table 7.1: Description of the BC-MC data matrices.

Name	m	n	% nonzero	$\min_{i,j} A_{ij}$	$\max_{i,j} A_{ij}$
Jester[2]	24938	100	24.66%	-9.95	10
Anime[3]	506	9437	10.50%	1	10
MovieLens 100K[4]	610	9724	1.70%	0.5	5
FilmTrust[5]	1508	2071	1.14%	0.5	8
MovieLens 1M[6]	6040	3952	4.19%	1	5

We now describe the experiment parameters considered. First the starting point Z_0 is randomly generated from a shifted binomial distribution that closely follows the data matrix A. More specifically, the entries of Z_0 are distributed according to a $\textsc{Binomial}(n, \mu/n) - \underline{A}$ distribution, where μ is the sample average of the nonzero entries in A, the integer n is the ceiling of the range of ratings in A, and $\underline{A}$ is the minimum rating in A. Second, the decomposition of the objective function is as follows

$$f_1 = \frac{1}{2}\|P_\Omega(\cdot - A)\|_F^2, \quad f_2^{\mathcal{V}} = \mu\left[\kappa_\mu(\cdot) - \frac{\beta}{\theta}\|\cdot\|_1\right] + \tau_\alpha(\cdot), \quad h^{\mathcal{V}} = \frac{\mu\beta}{\theta}\|\cdot\|_1 + \delta_{\mathcal{F}}(\cdot),$$

where $\mathcal{F} = \{U \in \mathbb{R}^{m\times n} : \|U\|_F \le \sqrt{mn}\cdot\max_{i,j}|A_{ij}|\}$ is the set of feasible solutions. Third,

[2]The ratings in the file "jester_dataset_1_1.zip" from `http://eigentaste.berkeley.edu/dataset/`.

[3]A subset of the ratings from `https://www.kaggle.com/CooperUnion/anime-recommendations-database` where each user has rated at least 720 items.

[4]The ratings in the file "ml-latest-small.zip" from `https://grouplens.org/datasets/movielens/`.

[5]See the ratings in the file "ratings.txt" under the FilmTrust section in `https://www.librec.net/datasets.html`.

[6]See the ratings in the file "ml-1m.zip" from `https://grouplens.org/datasets/movielens/`.

in view of the previous decomposition, the curvature parameters are set to be

$$m_1 = 0, \quad M_1 = 1, \quad m_2 = \frac{2\beta\mu}{\theta^2} + \frac{2\alpha\beta}{\theta}\exp\left(\frac{-3\theta}{2}\right), \quad M_2 = \frac{\alpha\beta}{\theta},$$

where it can be shown that the smallest and largest eigenvalues of $\nabla^2\tau_\alpha(z)$ are bounded below and above by $-2\alpha\beta\exp(-3\theta/2)/\theta$ and $\alpha\beta/\theta$, respectively, for every $z \in \mathbb{R}^n$. Fourth, each problem instance uses a specific data matrix A from Table 7.1, the hyperparameters $(\alpha, \beta, \mu, \theta) = (10, 20, 2, 1)$ and $\hat{\rho} = 10^{-6}$, and Ω to be the index set of nonzero entries in the chosen matrix A. Finally, a cutoff time of 10800 seconds is used for the MovieLens 1M dataset and a cutoff time of 7200 seconds is used for the other datasets.

Figure Figure 7.1 contains the plots of the log objective function value against the runtime, listed in increasing order of the smallest dimension in the data matrix.

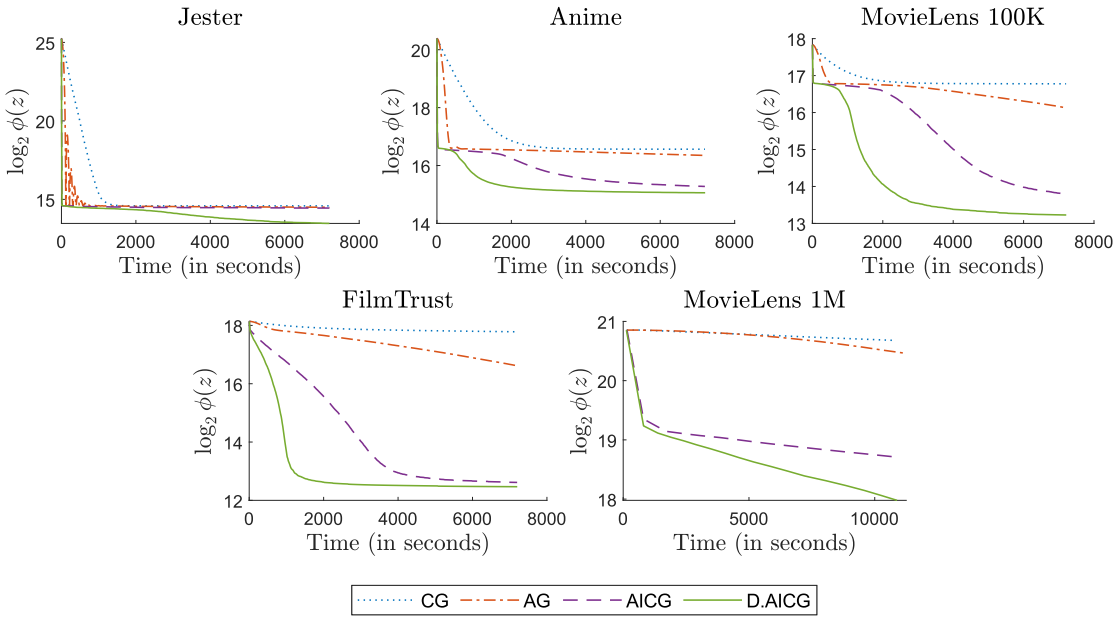

Figure 7.1: Function value vs. runtime for the BC-MC problems.

7.6.2 Multiblock Ball-Constrained Matrix Completion

This subsection presents computational results for the multiblock ball-constrained matrix (MBC-MC) problem in [50]. Given a quadruple $(\alpha, \beta, \mu, \theta) \in \mathbb{R}^4_{++}$, a block decomposable

data matrix $A \in \mathbb{R}^{m \times n}$ with blocks $\{A_i\}_{i=1}^k \subseteq \mathbb{R}^{p \times q}$, and indices Ω, this subsection considers the MBC-MC problem:

$$\begin{aligned} \min_{U \in \mathbb{R}^{m \times n}} \; & \frac{1}{2}\|P_\Omega(U - A)\|_F^2 + \sum_{i=1}^{k} \left[\kappa_\mu \circ \sigma(U_i) + \tau_\alpha \circ \sigma(U_i)\right] \\ \text{s.t.} \; & \|U\|_F^2 \leq \sqrt{mn} \cdot \max_{i,j} |A_{ij}|, \end{aligned}$$

where P_Ω, κ_μ, and τ_α are as in Section 7.6.1 and $U_i \in \mathbb{R}^{p \times q}$ is the i^{th} block of U with the same indices as A_i with respect to A.

We now describe the two classes of data matrices that are considered. Every data matrix is a 5-by-5 block matrix consisting of 50-by-100 sized submatrices. Every submatrix contains only 25% nonzero entries and each data matrix generates its submatrix entries from different probability distributions. More specifically, for a sampled probability $p \sim \textsc{Uniform}[0, 1]$ specific to a fixed submatrix, one class uses a $\textsc{Binomial}(n, p)$ distribution with $n = 10$, while the other uses a $\textsc{TruncatedNormal}(\mu, \sigma)$ distribution with $\mu = 10p$, $\sigma^2 = 10p(1 - p)$, and upper and lower bounds 0 and 10, respectively.

We now describe the experiment parameters considered. First, the decomposition of the objective function and the quantities Z_0, (m_1, M_1), (m_2, M_2), $\hat{\rho}$, and Ω are the same as in Section 7.6.1. Second, we fix $(\beta, \theta) = (20, 1)$ and vary (α, μ, A) across the different problem instances. Finally, a cutoff time of 7200 seconds is used for all problem instances tested.

Figure 7.2 contains the plots of the log objective function value against the runtime for the binomial data set, listed in increasing order of M_2. The corresponding plots for the truncated normal data set are similar to the binomial plots, so we omit them for the sake of brevity. Table 7.2 and Table 7.3 respectively contain the last function values of each algorithm for the binomial and truncated normal data sets, listed in increasing order of M_2. Moreover, each row of these tables corresponds to a different choice of (μ, α) and the bolded numbers highlight which algorithm performed the best in terms of the last function value.

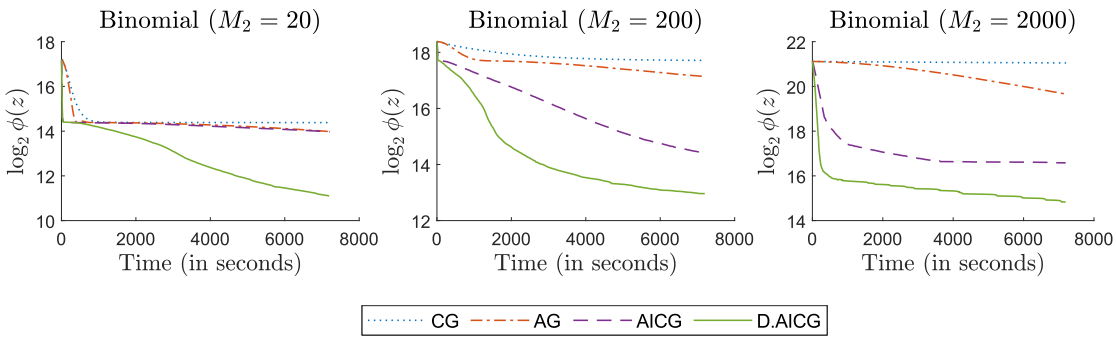

Figure 7.2: Function value vs. runtime for the binomial MBC-MC problems.

Table 7.2: Last function values for the binomial MBC-MC problems.

Parameters		**Last Function Value**			
(μ, α)	M_2	CG	AG	AICG	D.AICG
$(1, 0.2)$	20	2.13E+04	1.62E+04	1.61E+04	**2.20E+03**
$(10, 2)$	200	2.15E+05	1.44E+05	2.19E+04	**7.98E+03**
$(100, 20)$	2000	2.17E+06	8.24E+05	9.82E+04	**2.92E+04**

7.6.3 Discussion of the Results

We see that the D.AICGM and AICGM are generally more efficient than the AG and CG methods, respectively. The D.AICGM method, in particular, appears to escape local minima more quickly than the other methods. Moreover, the larger the constant M_2 is, the more efficient the ICG methods are compared to the benchmark methods. Curiously, the larger the smallest dimension of the matrix space is, the more efficient the inexact methods are compared to the exact ones.

We conjecture that the efficiency of the spectral methods is attributed to the fact that the main iterations of the methods are performed within the space of singular values rather than in the space of matrices.

7.7 Conclusion and Additional Comments

In this chapter, we presented two methods for finding approximate stationary points of a class of spectral NCO problems. More specifically, the methods are inexact variants of the

Table 7.3: Last function values for the truncated normal MBC-MC problems.

Parameters		**Last Function Value**			
(μ, α)	M_2	CG	AG	AICG	D.AICG
$(1, 0.2)$	20	2.14E+04	8.92E+03	1.26E+04	**1.25E+03**
$(10, 2)$	200	2.21E+05	1.75E+05	3.29E+04	**1.16E+04**
$(100, 20)$	2000	2.27E+06	1.71E+06	1.06E+05	**4.50E+04**

CGM (see Algorithm 2.2.1) and an accelerated monotonic CGM. We established an $\mathcal{O}(\hat{\rho}^{-2})$ iteration complexity bound for finding $\hat{\rho}$-approximate stationary points for both methods and an $\mathcal{O}(\hat{\rho}^{-2/3})$ bound for the accelerated method when the objective function is convex. Through several new results about spectral functions, we also developed a variant of the ACGM in Algorithm 2.2.2 which is especially efficient for spectral NCO problems.

The next chapter presents some practical improvements on the methods and procedures developed in this and previous chapters.

Additional Comments

It is worth mentioning that the outer iteration scheme of the D.AICGM is a monotonic and inexact generalization of the AG method in [30]. More specifically, the AG method can be viewed as a version of the D.AICGM where: (i) $\theta = 0$; (ii) the S.ACG call in Line 8 is replaced by an exact solver of (7.6); and (iii) the update of y_k in Line 18 is replaced by an update involving a prox evaluation of the function $a_{k-1}(f_2 + h)$. Hence, the D.AICGM can be significantly more efficient when its S.ACG call is more efficient than an exact solver of (7.6) and/or when the projection onto Ω is more efficient than the proximal evaluation of $a_{k-1}(f_2 + h)$.

Future Work

It would be worth investigating if the developments in Section 7.5 are applicable to other first-order iterative optimization algorithms and/or other classes of NCO problems.

APPENDIX A
PROPERTIES OF THE PPM AND CGM

This appendix presents the proofs of propositions related to the PPM and CGM.

Proof of Proposition 2.2.5. (a) The optimality of z_k and the definition of v_k immediately yield $v_k = (z_{k-1} - z_k)/\lambda_k \in \partial\psi(z_k)$

(b) Using the inclusion in (a) and the fact that $\lambda_k \geq 0$, we immediately have

$$\psi(z_{k-1}) \geq \psi(z_k) + \langle v_k, z_{k-1} - z_k \rangle = \psi(z_k) + \frac{1}{\lambda_k}\|z_{k-1} - z_k\|^2 > \psi(z_k).$$

(c) Summing the inequality in (b) from indices 1 to k yields

$$\begin{aligned}\left(\sum_{i=1}^{k}\lambda_i\right)\cdot\min_{i\leq k}\|v_i\|^2 \leq \sum_{i=1}^{k}\lambda_i\|v_i\|^2 \leq \sum_{i=1}^{k}\frac{1}{\lambda_i}\|z_{i-1} - z_i\|^2 &\leq \sum_{i=1}^{k}[\psi(z_{i-1}) - \psi(z_i)] \\ &= \psi(z_0) - \psi(z_k),\end{aligned}$$

which implies the desired inequality. □

Proof of Proposition 2.2.1. (a) The optimality of z_k and the definitions of v_k and ℓ_{ψ_s} imply that

$$\nabla\psi_s(z_k) + \partial\psi_n(z_k) \ni \nabla\psi_s(z_k) - \nabla\psi_s(z_{k-1}) + \frac{1}{\lambda_k}(z_{k-1} - z_k) = v_k.$$

(b) The above inclusion in part (a) and (2.2) imply that

$$\begin{aligned}\psi_n(z_{k-1}) &\geq \psi_n(z_k) + \langle v_k - \psi_s(z_k), z_{k-1} - z_k\rangle \\ &= \psi_n(z_k) + \langle \nabla\psi_s(z_{k-1}), z_k - z_{k-1}\rangle + \frac{1}{\lambda_i}\|z_{k-1} - z_k\|^2 \\ &\geq \psi_n(z_k) + [\psi_s(z_k) - \psi_s(z_{k-1})] + \left(\frac{1}{\lambda_k} - \frac{L_k}{2}\right)\|z_{k-1} - z_k\|^2,\end{aligned}$$

which implies the rightmost inequality. The leftmost inequality follows from the assumption that $L_k < 2/\lambda_k$.

(c) Using the inequality in (b) from indices 1 to k, the definition of v_k, (2.4), and the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$, it holds that

$$\begin{aligned}
&\left(\frac{1}{4}\sum_{i=1}^{k}\lambda_i\xi_i\right)\min_{i\le k}\|v_i\|^2 \\
&\le \sum_{i=1}^{k}\frac{1}{2}\left(\frac{2-\lambda_i L}{2\lambda_i}\right)\left(\frac{1}{\lambda_i^2}+L_i^2\right)^{-1}\|v_i\|^2 \\
&= \sum_{i=1}^{k}\frac{1}{2}\left(\frac{2-\lambda_i L}{2\lambda_i}\right)\left(\frac{1}{\lambda_i^2}+L_i^2\right)^{-1}\left\|\frac{1}{\lambda_i}(z_{i-1}-z_i)+\nabla\psi_s(z_i)-\nabla\psi_s(z_{i-1})\right\|^2 \\
&\le \sum_{i=1}^{k}\left(\frac{2-\lambda_i L}{2\lambda_i}\right)\left(\frac{1}{\lambda_i^2}+L_i^2\right)^{-1}\left[\frac{1}{\lambda_i^2}\|z_{i-1}-z_i\|^2+\|\nabla\psi_s(z_i)-\nabla\psi_s(z_{i-1})\|^2\right] \\
&\le \sum_{i=1}^{k}\left(\frac{2-\lambda_i L}{2\lambda_i}\right)\|z_{i-1}-z_i\|^2 \\
&\le \sum_{i=1}^{k}[\psi(z_{i-1})-\psi(z_i)] = \psi(z_0)-\psi(z_k),
\end{aligned}$$

which clearly implies the inequality in (2.5). □

Proof of Proposition 2.2.2. (a) It follows from Proposition 2.2.1 and the definitions of q and v that $q \in \nabla\psi_s(z^-) + \partial\psi_n(z)$. The desired inclusion and inequality now follow from Proposition 2.1.47 with $(s,\varepsilon,\bar{z}) = (q-\nabla\psi_s(z^-),\varepsilon,z^-)$ and $\psi = \psi_n$.

(b) Clearly, part (a) shows that (q,ε) is feasible to (2.6). Assume now that (r,δ) satisfies $r \in \nabla\psi_s(z^-) + \partial_\delta\psi_n(z^-)$, or equivalently

$$\psi_n(u) \ge \psi_n(z^-) + \langle r-\nabla\psi_s(z^-), u-z^-\rangle - \delta \quad \forall u \in \mathcal{Z}.$$

Using the above inequality with $u = z$ and the definitions of q and ε, we then conclude that

$$\begin{aligned}\lambda\|q\|^2 + 2\varepsilon &= \frac{1}{\lambda}\|z - z^-\|^2 + 2\left[\psi_n(z^-) - \psi_n(z) + \langle q - \nabla\psi_s(z^-), z - z^-\rangle\right] \\ &= 2\left[\psi_n(z^-) - \psi_n(z) - \langle \nabla\psi_s(z^-), z - z^-\rangle\right] - \frac{1}{\lambda}\|z - z^-\|^2 \\ &\le 2\delta - 2\langle r, z - z^-\rangle - \frac{1}{\lambda}\|z - z^-\|^2 \\ &= 2\delta - 2\lambda\langle r, q\rangle - \lambda\|q\|^2 \\ &\le 2\delta + \lambda\|r\|^2 + \lambda\|q\|^2 - \lambda\|q\|^2 = \lambda\|r\|^2 + 2\delta,\end{aligned}$$

where the last inequality follows from the inequality $2\langle a, b\rangle \le \|a\|^2 + \|b\|^2$ for every $a, b \in \mathcal{Z}$. Since (r, δ) are feasible to (2.6), the result follows.

(c) Using (2.2) and the definitions of q and ε yield

$$\begin{aligned}\lambda\|q\|^2 + 2\varepsilon &= 2\left[\psi_n(z^-) - \psi_n(z) - \langle \nabla\psi_s(z^-), z - z^-\rangle\right] - \frac{1}{\lambda}\|z^- - z\|^2 \\ &= 2\left[\psi(z^-) - \psi(z)\right] + 2\left[\psi_s(z) - \ell_{\psi_s}(z; z^-)\right] - \frac{1}{\lambda}\|z^- - z\|^2 \\ &\le 2\left[\psi(z^-) - \psi(z)\right] + \left(L - \frac{1}{\lambda}\right)\|z^- - z\|^2.\end{aligned}$$

□

Proof of Proposition 3.2.6. Define the quantities

$$\Psi_\lambda = \Psi_{\lambda,k} := g + \frac{1}{2\lambda}\|\cdot - z_{k-1}\|^2, \quad r_k := \frac{z_{k-1} - z_k}{\lambda}, \tag{A.1}$$

and note that $\nabla\Psi_\lambda(z_{k-1}) = \nabla g(z_{k-1})$, and that Ψ_λ is convex due to (A2) and the assumption $\lambda < 1/m$. Hence Proposition 2.1.47 and $\nabla g(z_{k-1}) = \nabla\Psi_\lambda(z_{k-1}) \in \partial_{\varepsilon_k}\Psi_\lambda(z_k)$, where $\varepsilon_k = \Psi_\lambda(z_k) - \Psi_\lambda(z_{k-1}) - \langle \nabla\Psi_\lambda(z_{k-1}), z_k - z_{k-1}\rangle \ge 0$. The previous inclusion combined with the optimality of z_k and definition of r_k imply that $r_k \in \partial h(z_k) + \partial_{\varepsilon_k}\Psi_\lambda(z_k) \subset \partial_{\varepsilon_k}(h + \Psi_\lambda)(z_k)$ where the last inclusion follows immediately from the definition of the operator ∂_{ε_k} and

convexity of h. Hence, since $(\tilde{\varepsilon}_k, \tilde{v}_k) = \lambda(\varepsilon_k, r_k)$ (see (3.21) and (A.1)), it follows from the above inclusion and the definition of Ψ_λ that the triple $(z_k, \tilde{v}_k, \tilde{\varepsilon}_k)$ satisfies the inclusion in (3.4) with $\phi = g + h$ and $\lambda_k = \lambda$.

Now, to prove that the inequality in (3.5) holds, first note that the definitions of ε_k and Ψ_λ together with property (A2), imply that $\varepsilon_k \leq (\lambda M + 1)\|z_{k-1} - z_k\|^2/(2\lambda)$. Combining the previous inequality with the relations $\tilde{v}_k = z_{k-1} - z_k$ and $\tilde{\varepsilon}_k = \lambda\varepsilon_k$, we obtain

$$\begin{aligned}\|\tilde{v}_k\|^2 + 2\tilde{\varepsilon}_k &= \|z_{k-1} - z_k\|^2 + 2\lambda\varepsilon_k \leq \|z_{k-1} - z_k\|^2 + (\lambda M + 1)\|z_{k-1} - z_k\|^2 \\ &= (\lambda M + 2)\|z_{k-1} - z_k\|^2 = \frac{\lambda M + 2}{4}\|z_{k-1} - z_k + \tilde{v}_k\|^2.\end{aligned}$$

Hence, since $\lambda M < 2$, we conclude that $\sigma = (\lambda M + 2)/4 < 1$ and that (3.5) holds. □

APPENDIX B
PROPERTIES OF THE ACGM

This appendix presents important properties and proofs related to the ACGM in Chapter 2. Throughout this appendix, we assume that the iterates

$$\{(x_k, y_k, r_k, \eta_k)\}_{k\geq 1}, \quad \{(\tau_k, a_k, A_k, \gamma_k, q_k, \Gamma_k)\}_{k\geq 1},$$

are generated by the ACGM and the quantities μ, $\{\lambda_k\}_{k\geq 1}$, and ψ are from its input and initialization, respectively.

We first present some basic properties involving the function pairs $\{(\gamma_k, q_k)\}_{k\geq 1}$.

Lemma B.0.1. *The following statements hold for every $k \geq 1$:*

(a) $\gamma_{k-1}(y_k) = q_k(y_k)$ and $\gamma_{k-1} \leq q_k \leq \psi$;

(b) it holds that

$$\min_{u\in\mathcal{Z}} \left\{ q_k(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_{k-1}\|^2 \right\} = \min_{u\in\mathcal{Z}} \left\{ \gamma_{k-1}(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_{k-1}\|^2 \right\}.$$

Proof. (a) The fact that $\gamma_{k-1}(y_k) = q_k(y_k)$ is immediate from the definitions of γ_k and q_k. The fact that $\psi \geq q_k$ follows from the assumption that $\psi_s \in \mathcal{F}_\mu(Z)$. To show that $\gamma_{k-1} \leq q_k$, observe that the optimality of y_k and the fact that $q_k \in \mathcal{F}_\mu(Z)$ imply that

$$\begin{aligned} &\lambda_k q_k(y_k) + \frac{1}{2}\|y_k - \tilde{x}_{k-1}\|^2 + \left(\frac{\lambda_k\mu + 1}{2}\right)\|u - y_k\|^2 \\ &\leq \lambda_k q_k(u) + \frac{1}{2}\|u - \tilde{x}_{k-1}\|^2 \quad \forall u \in \mathcal{Z}. \end{aligned}$$

Rearranging terms and using the definition γ_{k-1}, we conclude that

$$\begin{aligned} q_k(u) &\geq q_k(y_k) + \frac{1}{2\lambda_k}\|y_k - \tilde{x}_{k-1}\|^2 + \left(\frac{\lambda_k\mu + 1}{2\lambda_k}\right)\|u - y_k\|^2 - \frac{1}{2\lambda_k}\|u - \tilde{x}_{k-1}\|^2 \\ &= q_k(y_k) + \frac{1}{2\lambda_k}\left[\|y_k - \tilde{x}_{k-1}\|^2 + \|u - y_k\|^2 - \|u - \tilde{x}_{k-1}\|^2\right] + \frac{\mu}{2}\|u - y_k\|^2 \\ &= q_k(y_k) + \frac{1}{2\lambda_k}\left(2\|y_k - \tilde{x}_{k-1}\|^2 + 2\left\langle \tilde{x}_{k-1} - y_k, u - \tilde{x}_{k-1}\right\rangle\right) + \frac{\mu}{2}\|u - y_k\|^2 \\ &= \gamma_{k-1}(u) + \frac{1}{\lambda_k}\|y_k - \tilde{x}_{k-1}\|^2 \geq \gamma_{k-1}(u) \quad \forall u \in \mathcal{Z}. \end{aligned}$$

(b) Recall that y_k is an optimal solution of the left problem. Suppose that $\bar{y}$ is an optimal solution of the right problem. Since γ_k is a smooth convex function, the optimality of $\bar{y}$ and the definition of γ_k imply that

$$\begin{aligned} 0 &= \nabla\gamma_{k-1}(\bar{y}) + \frac{1}{\lambda_k}(\bar{y} - \tilde{x}_{k-1}) = \frac{1}{\lambda_k}(\tilde{x}_{k-1} - y_k) + \mu(\bar{y} - y_k) + \frac{1}{\lambda_k}(\bar{y} - \tilde{x}_{k-1}) \\ &= \left(\mu + \frac{1}{\lambda_k}\right)(\bar{y} - y_k), \end{aligned}$$

which, since $\mu, \lambda_k > 0$, implies that $\bar{y} = y_k$. □

We next present properties involving the scalars $\{(\lambda_k, a_k, A_k)\}_{k\geq 1}$.

Lemma B.0.2. *The following statements hold for every $k \geq 1$:*

(a) $a_k^2 = \tau_k A_{k+1}$;

(b) it holds that

$$A_k \geq \max\left\{\frac{1}{4}\left(\sum_{i=1}^{k}\sqrt{\lambda_{i-1}}\right)^2, \lambda_1\prod_{i=2}^{k}\left(1 + \sqrt{\frac{\lambda_{i-1}\mu}{2}}\right)^2\right\}.$$

Proof. (a) Let $k \geq 1$ be fixed. It is easy to see that a_k is a root of the quadratic function $x \mapsto x^2 - \tau_k x + \tau_k A_k$ and hence, using the definitions of τ_k and the update rule of A_{k+1}, it

holds that

$$0 = a_k^2 - \tau_k(a_k + A_k) = a_k^2 - \tau_k A_{k+1},$$

which implies the desired identity.

(b) We first make the observation that

$$a_{k-1} = \frac{\tau_{k-1} + \sqrt{\tau_{k-1}^2 + 4\tau_{k-1}A_{k-1}}}{2} \geq \frac{\tau_{k-1}}{2} + \sqrt{\tau_{k-1}A_{k-1}} \tag{B.1}$$

We now show that A_k is bounded below by the first term in the max. Using (B.1) and the update rule for A_k, it holds that

$$\begin{aligned} A_k = A_{k-1} + a_{k-1} &\geq \frac{\tau_{k-1}}{2} + \sqrt{\tau_{k-1}A_{k-1}} + A_{k-1} \\ &\geq \left(\sqrt{A_{k-1}} + \frac{\sqrt{\tau_{k-1}}}{2}\right)^2 \end{aligned}$$

which, by taking square roots on both sides, yields

$$\sqrt{A_k} \geq \sqrt{A_{k-1}} + \frac{\sqrt{\tau_{k-1}}}{2} \geq \sqrt{A_{k-1}} + \frac{\sqrt{\lambda_{k-1}}}{2}.$$

Applying this relationship recursively, and squaring the resulting relation yields the desired bound on A_k.

We next show that A_k is bounded below by the second term in the max on the right-hand-side. If $k = 1$, the inequality follows immediately from the fact that $A_1 = \lambda_1$. Instead,

suppose that $k \geq 2$. Using (B.1) and the update rule for A_k, it holds that

$$\begin{aligned}
A_k =& A_{k-1} + a_{k-1} \geq \frac{\tau_{k-1}}{2} + \sqrt{\tau_{k-1} A_{k-1}} + A_{k-1} \\
=& \left(\sqrt{A_{k-1}} + \sqrt{\frac{\tau_{k-1}}{2}} \right)^2 + \frac{\tau_{k-1}}{4} \\
\geq& \left(\sqrt{A_{k-1}} + \sqrt{\frac{\mu \lambda_{k-1} A_{k-1}}{2}} \right)^2 + \frac{\mu \lambda_{k-1} A_{k-1}}{4} \\
=& A_{k-1} \left[\left(1 + \sqrt{\frac{\mu \lambda_{k-1}}{2}} \right)^2 + \frac{\mu \lambda_{k-1} A_{k-1}}{4} \right] \geq A_{k-1} \left(1 + \sqrt{\frac{\mu \lambda_{k-1}}{2}} \right)^2 .
\end{aligned}$$

Applying the above relationship recursively yields the desired relation

$$A_k \geq A_1 \prod_{i=2}^{k} \left(1 + \sqrt{\frac{\lambda_{i-1} \mu}{2}} \right)^2 = \lambda_1 \prod_{i=2}^{k} \left(1 + \sqrt{\frac{\lambda_{i-1} \mu}{2}} \right)^2 .$$

□

We now present properties involving the iterates $\{(x_k, y_k)\}_{k \geq 1}$ generated by the method.

Lemma B.0.3. *The following statements hold for every $k \geq 1$:*

(a) $\Gamma_k \in \mathcal{F}_\mu(\mathcal{Z})$ is a quadratic function and $\Gamma_k \leq \psi$;

(b) $x_k = \operatorname{argmin}_{u \in \mathcal{Z}} \{A_k \Gamma_k(u) + \|u - x_0\|^2/2\}$;

(c) if there exists $L_k > 0$ satisfying (2.8)*, then it holds that*

$$A_k \psi(y_k) \leq \min_{u \in \mathcal{Z}} \left\{ A_k \Gamma_k(u) + \frac{1}{2} \|u - x_0\|^2 \right\} .$$

Proof. (a) Observe that recursively applying the definition of Γ_k yields the identity $\Gamma_k = \sum_{i=0}^{k-1} a_i \gamma_i / (\sum_{i=0}^{k-1} a_i)$, which shows that Γ_k is a convex combination of the functions $\{\gamma_i\}_{i=0}^{k-1}$. The desired conclusion now follows from the definition of γ_k and Lemma B.0.1(a).

(b) We proceed by induction on k. The case of $k = 0$ is obvious. Suppose instead that $x_{k-1} = \operatorname{argmin}_{u\in\mathcal{Z}} \{A_{k-1}\Gamma_{k-1}(u) + \|u - x_0\|^2/2\}$ for some $k \geq 2$. The optimality of x_{k-1} implies that

$$x_{k-1} - x_0 + A_{k-1}\nabla\Gamma_{k-1}(x_{k-1}) = 0. \tag{B.2}$$

Moreover, since $\Gamma_k \in \mathcal{F}_\mu(\mathcal{Z})$ is a quadratic function (see part (a)), it holds that

$$\nabla\Gamma_k(x_k) = \nabla\Gamma_k(x_{k-1}) + \mu(x_k - x_{k-1}). \tag{B.3}$$

Let us now verify the optimality condition on x_k. Using (B.3), (B.2), the update rule for x_k, the definition of γ_{k-1}, and our hypothesis, we have that

$$\begin{aligned}
&x_k - x_0 + A_k\nabla\Gamma_k(x_k)\\
&= x_k - x_0 + A_k\left[\nabla\Gamma_k(x_{k-1}) + \mu(x_k - x_{k-1})\right]\\
&= x_k - x_0 + A_{k-1}\nabla\Gamma_{k-1}(x_{k-1}) + a_{k-1}\nabla\gamma_{k-1}(x_{k-1}) + \mu A_k(x_k - x_{k-1})\\
&= \left[x_k - x_{k-1}\right] + \left[x_{k-1} - x_0 + A_{k-1}\nabla\Gamma_{k-1}(x_{k-1})\right] +\\
&\qquad \left[a_{k-1}\nabla\gamma_{k-1}(x_{k-1}) + \mu A_k(x_k - x_{k-1})\right]\\
&= (1 + \mu A_k)(x_k - x_{k-1}) + a_{k-1}\nabla\gamma_{k-1}(x_{k-1})\\
&= -a_{k-1}\left[\frac{1}{\lambda_k}(\tilde{x}_{k-1} - y_k) + \mu(x_{k-1} - y_k)\right] + a_{k-1}\nabla\gamma_{k-1}(x_{k-1})\\
&= -a_{k-1}\nabla\gamma_{k-1}(z_{k-1}) + a_{k-1}\nabla\gamma_{k-1}(x_{k-1}) = 0,
\end{aligned}$$

which implies that $x_k = \operatorname{argmin}_{u\in\mathcal{Z}} \{A_k\Gamma_k(u) + \|u - x_0\|^2/2\}$.

(c) We proceed by induction by k. The case of $k = 0$ is obvious. Suppose instead that

$$A_{k-1}\psi(y_{k-1}) \leq \min_{u\in\mathcal{Z}} \left\{A_{k-1}\Gamma_{k-1}(u) + \frac{1}{2}\|u - x_0\|^2\right\}$$

for some $k \geq 2$. Using the fact that Γ is μ-strongly convex, the optimality of x_{k-1} in part (b),

and our hypothesis, we have that

$$\begin{aligned}
&\psi(y_{k-1}) + \frac{A_{k-1}\mu + 1}{2}\|u - x_{k-1}\|^2 \\
&\leq A_{k-1}\Gamma_{k-1}(x_{k-1}) + \frac{1}{2}\|x_{k-1} - x_0\|^2 + \left(\frac{A_{k-1}\mu + 1}{2}\right)\|u - x_{k-1}\|^2 \\
&\leq A_{k-1}\Gamma_{k-1}(u) + \frac{1}{2}\|u - x_0\|^2, \qquad \text{(B.4)}
\end{aligned}$$

for every $u \in \mathcal{Z}$. Combining (B.4), Lemma B.0.2(a), Lemma B.0.1(a), and the fact that γ_{k-1} is convex yields

$$\begin{aligned}
&\min_{u\in\mathcal{Z}}\left\{A_k\Gamma_k(u) + \frac{1}{2}\|u - x_0\|^2\right\} \\
=&\min_{u\in\mathcal{Z}}\left\{A_{k-1}\Gamma_{k-1}(u) + a_{k-1}\gamma_{k-1}(u) + \frac{1}{2}\|u - x_0\|^2\right\} \\
\geq&\min_{u\in\mathcal{Z}}\left\{A_{k-1}\psi(y_{k-1}) + \left(\frac{A_{k-1}\mu + 1}{2}\right)\|u - x_{k-1}\|^2 + a_{k-1}\gamma_{k-1}(x)\right\} \\
\geq&\min_{u\in\mathcal{Z}}\left\{A_{k-1}\gamma_{k-1}(y_{k-1}) + \left(\frac{A_{k-1}\mu + 1}{2}\right)\|u - x_{k-1}\|^2 + a_{k-1}\gamma_{k-1}(x)\right\} \\
=&\min_{u\in\mathcal{Z}}\left\{(A_{k-1} + a_{k-1})\gamma_{k-1}\left(\underbrace{\frac{A_{k-1}y_{k-1} + a_{k-1}u}{A_{k-1} + a_{k-1}}}_{:=\tilde{u}}\right) + \left(\frac{A_{k-1}\mu + 1}{2}\right)\|u - x_{k-1}\|^2\right\} \\
=&\min_{\tilde{u}\in\mathcal{Z}}\left\{A_k\gamma_{k-1}(\tilde{u}) + \left(\frac{A_{k-1}\mu + 1}{2}\right)\left(\frac{A_k^2}{a_{k-1}^2}\right)\|\tilde{u} - \tilde{x}_{k-1}\|^2\right\} \\
=&A_k\min_{\tilde{u}\in\mathcal{Z}}\left\{\gamma_{k-1}(\tilde{u}) + \left(\frac{A_{k-1}\mu + 1}{2}\right)\left(\frac{A_k}{a_{k-1}^2}\right)\|\tilde{u} - \tilde{x}_{k-1}\|^2\right\} \\
=&A_k\min_{\tilde{u}\in\mathcal{Z}}\left\{\gamma_{k-1}(\tilde{u}) + \frac{1}{2\lambda_k}\|\tilde{u} - \tilde{x}_{k-1}\|^2\right\}. \qquad \text{(B.5)}
\end{aligned}$$

On the other hand, using (2.8), the definition of y_k, and Lemma B.0.1(b), it holds that

$$\begin{aligned}
\psi(y_k) &\le l_{\psi_s}(y_k;\tilde{x}_{k-1}) + \psi_n(y_k) + \frac{L_k}{2}\|y_k - \tilde{x}_{k-1}\|^2 \\
&= q_k(y_k) + \frac{1}{2}\|y_k - \tilde{x}_{k-1}\|^2 + \frac{1}{2}\left([L_k - \mu] - \frac{1}{\lambda_k}\right)\|y_k - \tilde{x}_{k-1}\|^2 \\
&\le q_k(y_k) + \frac{1}{2\lambda_k}\|y_k - \tilde{x}_{k-1}\|^2 \\
&= \min_{u\in\mathcal{Z}}\left\{q_k(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_{k-1}\|^2\right\} \\
&= \min_{u\in\mathcal{Z}}\left\{\gamma_{k-1}(u) + \frac{1}{2\lambda_k}\|u - \tilde{x}_{k-1}\|^2\right\}. \qquad \text{(B.6)}
\end{aligned}$$

Combining (B.5) and (B.6), we conclude that

$$A_k\psi(y_k) \le \min_{u\in\mathcal{Z}}\left\{A_k\Gamma_k(u) + \frac{1}{2}\|u - x_0\|^2\right\}.$$

□

We are now ready to give the proofs of Proposition 2.2.3 and Proposition 2.2.4.

Proof of Proposition 2.2.3. (a) The optimality of x_k and the definition of r_k imply that $r_k \in \partial\Gamma_k(x_k)$. Using the previous inclusion, the definition of η_k, and the fact that $\psi \ge \Gamma_k$ yields

$$\psi(u) \ge \Gamma_k(u) \ge \Gamma(x_k) + \langle r_k, u - x_k\rangle = \psi(y_k) + \langle r_k, u - y_k\rangle - \eta_k$$

for every $u \in \mathcal{Z}$, which is exactly the desired inclusion. The fact that $\eta \ge 0$ follows from the above relationship evaluated at $u = y_k$.

(b) Using Lemma B.0.3(b) and (c) and the definitions of η_k and r_k yields

$$\begin{aligned}
0 &\le \Gamma_k(x_k) + \frac{1}{2A_k}\|x_k - x_0\|^2 - \psi(y_k) \\
&= -\eta_k - \langle r_k, y_k - x_k\rangle + \frac{1}{2A_k}\|x_k - x_0\|^2 \\
&= -\eta_k + \left[-\frac{1}{A_k}\langle x_0 - x_k, y_k - x_k\rangle + \frac{1}{2A_k}\|x_k - x_0\|^2\right] \\
&= -\eta_k + \left[\frac{1}{2A_k}\|y_k - x_0\|^2 - \frac{1}{2A_k}\|y_k - x_k\|^2\right] \\
&= -\eta_k + \frac{1}{2A_k}\|y_k - x_0\|^2 - \frac{1}{2A_k}\|A_k r_k + y_k - x_0\|^2
\end{aligned}$$

which, together with the identity $x_0 = y_0$, yields the desired inequality.

(c) Let y^* be an optimal solution of $\mathcal{CO}$. Using Lemma B.0.3(a) and (c) it holds that

$$A_k\psi(y_k) \le A_k\Gamma_k(y^*) + \frac{1}{2}\|y^* - y_0\|^2 \le A_k\psi(y^*) + \frac{1}{2}\|y^* - y_0\|^2,$$

which implies the desired inequality. □

Proof of Proposition 2.2.4. (a) Using Lemma B.0.3(b) we first observe that $(x_k - x_0)/A_k \in \Gamma_k(x_k)$ and hence, by Lemma B.0.3(a) and the definition of r_k, it holds that

$$\tilde{r}_k = \frac{x_k - x_0}{A_k} + \mu(y_k - x_k) \in \partial\left(\Gamma_k - \frac{\mu}{2}\|\cdot - y_k\|^2\right)(x_k)$$

Using the above inclusion, the fact that $\Gamma_k - \mu\|\cdot\|^2/2$ is affine, and the definition of η_k, we conclude that for every $u \in \mathcal{Z}$ it holds that

$$\begin{aligned}
\psi(u) - \frac{\mu}{2}\|u - y_k\|^2 \ge \Gamma_k(u) - \frac{\mu}{2}\|u - y_k\|^2 \\
&= \Gamma_k(y_k) - \frac{\mu}{2}\|y_k - x_k\|^2 + \langle \tilde{r}_k, u - x_k\rangle \\
&= \psi(y_k) + \langle \tilde{r}_k, u - y_k\rangle - \tilde{\eta}_k, \qquad \text{(B.7)}
\end{aligned}$$

which is equivalent to (2.7). The fact that $\tilde{\eta}_k \geq 0$ follows from the above inequality with $u = y_k$.

(b) Using Lemma B.0.3(b) and (c) and the definitions of $\tilde{\eta}_k$ and $\tilde{r}_k$ yields

$$\begin{aligned}
0 &\leq \Gamma_k(x_k) + \frac{1}{2A_k}\|x_k - x_0\|^2 - \psi(y_k) \\
&= -\tilde{\eta}_k - \frac{\mu}{2}\|y_k - x_k\|^2 - \langle \tilde{r}_k, y_k - x_k\rangle + \frac{1}{2A_k}\|x_k - x_0\|^2 \\
&= -\tilde{\eta}_k - \frac{\mu}{2}\|y_k - x_k\|^2 + \left[-\frac{1}{A_k}\langle x_0 - x_k, y_k - x_k\rangle + \frac{1}{2A_k}\|x_k - x_0\|^2\right] \\
&= -\tilde{\eta}_k - \frac{\mu}{2}\|y_k - x_k\|^2 + \left[\frac{1}{2A_k}\|y_k - x_0\|^2 - \frac{1}{2A_k}\|y_k - x_k\|^2\right] \\
&= -\tilde{\eta}_k + \frac{1}{2A_k}\|y_k - x_0\|^2 - \frac{1}{2A_k}(1 + \mu A_k)\|y_k - x_k\|^2 \\
&= -\tilde{\eta}_k + \frac{1}{2A_k}\|y_k - x_0\|^2 - \frac{1}{2A_k(1 + \mu A_k)}\|A_k\tilde{r}_k + y_k - x_0\|^2
\end{aligned}$$

which, together with the identity $x_0 = y_0$, yields the desired inequality. □

We next give the proof of Lemma 3.3.1.

Proof of Lemma 3.3.1. Let j be such that

$$A_j \geq \frac{2(1 + \sqrt{\sigma})^2}{\sigma}.$$

Using the triangle inequality, the previous bound on A_j, the relation $(a+b)^2 \leq 2a^2 + 2b^2$ for every $a, b \in \mathbb{R}$, and Proposition 2.2.3(b), we obtain

$$\begin{aligned}
\|r_j\|^2 + 2\eta_j &\leq \max\left\{1/A_j^2, 1/(2A_j)\right\}\left(\|A_j r_j\|^2 + 4A_j\eta_j\right) \\
&\leq \max\left\{1/A_j^2, 1/(2A_j)\right\}\left(2\|A_j r_j + y_j - y_0\|^2 + 2\|y_j - y_0\|^2 + 4A_j\eta_j\right) \\
&\leq \max\left\{(2/A_\ell)^2, 2/A_\ell\right\}\|y_j - y_0\|^2 \leq \frac{\sigma}{(1 + \sqrt{\sigma})^2}\|y_j - y_0\|^2.
\end{aligned}$$

On the other hand, the triangle inequality and simple calculations yield

$$\|y_j - y_0\|^2 \leq (1+\sqrt{\sigma})\|y_0 - y_j + r_j\|^2 + \left(1 + \frac{1}{\sqrt{\sigma}}\right)\|r_j\|^2.$$

Combining the previous bounds, we obtain

$$\|r_j\|^2 + 2\eta_j \leq \frac{\sigma}{1+\sqrt{\sigma}}\|y_0 - y_j + r_j\|^2 + \frac{\sqrt{\sigma}}{1+\sqrt{\sigma}}\|u_j\|^2$$

which easily implies (3.24).

Let us now show what conditions on j yield $A_j \geq 2(1+\sqrt{\sigma})^2/\sigma$. Using the first bound in Lemma B.0.2 with $\lambda_k = 1/L$, it is straightforward to show that the condition

$$j \geq \left\lceil \frac{2\sqrt{2L}(1+\sqrt{\sigma})}{\sqrt{\sigma}} \right\rceil$$

suffices.

On the other hand, using the second bound in Lemma B.0.2 with $\lambda_k = 1/L$ and the bound $\log(1+t) \geq t/2$ for $t \in [0,1]$, it is straightforward to show that the condition

$$j \geq \left\lceil 1 + \sqrt{\frac{2L}{\mu}}\log_1^+\left(\frac{2L\left[1+\sqrt{\sigma}\right]^2}{\sigma}\right) \right\rceil$$

suffices. The conclusion now follows from combining the previous bounds on j. □

APPENDIX C
PROPERTIES OF THE S.ACGM AND R.ACGM

This appendix contains proofs related to the S.ACGM and R.ACGM in Chapters 5 and 7, respectively.

We first give the proof of Proposition 7.2.3(a).

Proof of Proposition 7.2.3(a). Let ℓ be the first iteration where

$$\min\left\{\frac{A_\ell^2}{4(1+\mu A_\ell)}, \frac{A_\ell}{2}\right\} \geq K_\theta^2 \tag{C.1}$$

and suppose that the method has not stopped with π_S = `false` before iteration ℓ. We show that it must stop with π_S = `true` at the end of the ℓ^{th} iteration. Combining the triangle inequality, the successful checks in Line 6 of the method, (C.1), and the relation $(a+b)^2 \leq 2a^2 + 2b^2$ for all $a, b \in \mathbb{R}$, we first have that

$$\begin{aligned}
&\|r_\ell\|^2 + 2\eta_\ell \\
&\leq \max\left\{\frac{1+\mu A_\ell}{A_\ell^2}, \frac{1}{2A_\ell}\right\}\left(\frac{1}{1+\mu A_\ell}\|A_\ell \tilde{r}_\ell\|^2 + 4A_\ell \eta_\ell\right) \\
&\leq \max\left\{\frac{1+\mu A_\ell}{A_\ell^2}, \frac{1}{2A_\ell}\right\}\left(\frac{2}{1+\mu A_\ell}\|A_\ell \tilde{r}_\ell + z_\ell - z_0\|^2 + 2\|z_\ell - z_0\|^2 + 4A_\ell \tilde{\eta}_\ell\right) \\
&\leq \max\left\{\frac{4(1+\mu A_\ell)}{A_\ell^2}, \frac{2}{A_\ell}\right\}\|z_\ell - z_0\|^2 \leq \frac{1}{K_\theta^2}\|z_\ell - z_0\|^2 \leq \theta^2\|z_\ell - z_0\|^2,
\end{aligned}$$

and hence the method must terminate at the ℓ^{th} iteration. We now bound ℓ based on the requirement in (C.1). Solving for the quadratic in A_ℓ in the first bound of (C.1), it is easy to see that $A_\ell \geq 4\mu K_\theta^2 + 2K_\theta$ implies (C.1). On the other hand, for the second condition in (C.1), it is immediate that $A_\ell \geq 2K_\theta^2$ implies (C.1). In view of Proposition 2.2.3(c) with $\lambda_i = 1/L$ for every $i \geq 1$, and the previous two bounds, it follows that

$$A_\ell \geq \frac{1}{L}\left(1 + \sqrt{\frac{\mu}{2L}}\right)^{2(\ell-1)} \geq 2K_\theta(1 + 2\mu K_\theta^2)$$

implies (C.1). Using the bound $\log(1+t) \geq t/(1+t)$ for $t \geq 0$ and the above bound on ℓ, it is straightforward to see that ℓ is on the same order of magnitude as in (7.18). $\square$

We next give the proof of Lemma 5.3.5.

Proof of Lemma 5.3.5. Using our assumption that $f \in \mathcal{C}_M(Z)$, and hence $\psi_s \in \mathcal{C}_{L_\lambda}(Z)$, together with the check in Line 8 of Algorithm 5.2.1, it holds that the stepsizes $\{\lambda_i\}_{i\geq 1}$ in the R.ACGM are constant with a value of $1/L_\lambda$. Hence, using Proposition 2.2.3 and (5.17), it holds that

$$A_i \geq \frac{1}{L_\lambda}\left(1+\sqrt{\frac{1}{2L_\lambda}}\right)^{2(i-1)} \qquad \forall i \geq 1. \tag{C.2}$$

Now, let ℓ denote the quantity in (5.22), and suppose the R.ACGM has performed ℓ iterations in which (5.20) and (5.21) hold for every $i \leq \ell$. Using (C.2), the definition of $C_{\theta,\tau}$ in (5.23), and the fact that $\log(1+t) \geq t/2$ for all $t \in [0,1]$, it holds that

$$A_\ell \geq \frac{1}{L_\lambda}\left(1+\sqrt{\frac{1}{2L_\lambda}}\right)^{2(\ell-1)} \geq 2C_{\theta,\tau} > 2.$$

Combining the triangle inequality, (5.20), the bounds $2/A_\ell \leq 1/C$ and $(2/A_\ell)^2 < 2/A_\ell < 1$ from above, and the relation $(a+b)^2 \leq 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$, we obtain

$$\begin{aligned}
\|r_\ell\|^2 + 2\eta_\ell &\leq \max\{1/A_\ell^2, 1/(2A_\ell)\}(\|A_\ell r_\ell\|^2 + 4A_\ell\eta_\ell) \\
&\leq \max\{1/A_\ell^2, 1/(2A_\ell)\}(2\|A_\ell r_\ell + y_\ell - y_0\|^2 + 2\|y_\ell - x_0\|^2 + 4A_\ell\eta_\ell) \\
&\leq \max\{(2/A_\ell)^2, 2/A_\ell\}\|y_\ell - y_0\|^2 \leq \frac{1}{C_{\theta,\tau}}\|y_\ell - y_0\|^2.
\end{aligned}$$

On the other hand, using the triangle inequality and the fact that $(a+b)^2 \leq (1+s)a^2 + (1+1/s)b^2$ for every $(a,b,s) \in \mathbb{R}\times\mathbb{R}\times R_{++}$ (under the choice of $s = 1/(\sqrt{C}-1)$), we obtain

$$\|y_\ell - y_0\|^2 \leq \frac{\sqrt{C_{\theta,\tau}}}{\sqrt{C_{\theta,\tau}}-1}\|y_0 - y_\ell + r_\ell\|^2 + \sqrt{C_{\theta,\tau}}\|r_\ell\|^2.$$

Combining the previous estimates, we then conclude that

$$\|u_\ell\|^2 + 2\eta_\ell \leq \frac{1}{C_{\theta,\tau} - \sqrt{C_{\theta,\tau}}} \|x_0 - x_\ell + u_\ell\|^2 + \frac{1}{\sqrt{C_{\theta,\tau}}} \|u_\ell\|^2,$$

which, after a simple algebraic manipulation, easily implies that

$$\begin{aligned} \frac{1}{\sqrt{C_{\theta,\tau}} - 1} \|x_0 - x_\ell + u_\ell\|^2 &\geq 2\sqrt{C_{\theta,\tau}}\eta_\ell + \left(\sqrt{C_{\theta,\tau}} - 1\right) \|u_\ell\|^2 \\ &\geq \left(\sqrt{C_{\theta,\tau}} - 1\right) \left(\|u_\ell\|^2 + 2\eta_\ell\right). \end{aligned} \tag{C.3}$$

Using the first term in the maximum of (5.23) together with the second inequality of (C.3) immediately implies that (5.18) holds with $j = \ell$. To show that (5.19) holds at $j = \ell$, observe that the definition of ψ in (5.17), (5.21) with $j = \ell$, the second inequality of (C.3), and the second term in the maximum of (5.23) imply that

$$\begin{aligned} \lambda \left[\phi(x_0) - \phi(x_\ell)\right] &\geq \langle r_\ell, y_0 - y_\ell \rangle + \eta_\ell + \frac{1}{2}\|y_\ell - y_0\|^2 \\ &= \frac{1}{2}\left[\|y_0 - y_\ell + r_\ell\|^2 - \left(\|r_\ell\|^2 + 2\eta_\ell\right)\right] \\ &\geq \frac{1}{2}\left[1 + \left(\sqrt{C_{\theta,\tau}} - 1\right)^{-2}\right] \|y_0 - y_\ell + r_\ell\|^2 \geq \frac{1}{\theta}\|y_0 - y_\ell + r_\ell\|^2. \end{aligned}$$

□

APPENDIX D
PROPERTIES OF THE CRP

This appendix contains proofs related to the CRP in Chapter 3.

We first give the proof of Proposition 3.2.5.

Proof of Proposition 3.2.5. (a) This follows immediately from Proposition 2.2.2(a) with $(\psi_s, \psi_n, z_{k-1}) = (f, h, z)$ and $(q_k, \varepsilon_k) = (q_r, \varepsilon_r)$.

(b) Using the definition of ε_r, it follows that $q_r \in \nabla f(z) + \partial_{\varepsilon_r} h(z)$ if and only if

$$\begin{aligned} h(u) &\geq h(z) + \langle q_r - \nabla f(z), u - z \rangle - \varepsilon_r \\ &= h(z_r) + \langle q_r - \nabla f(z), u - z_r \rangle \quad \forall u \in \mathcal{Z}, \end{aligned}$$

or equivalently, $q_r \in \nabla f(z) + \partial h(z_r)$. The desired inclusion now follows from the previous inclusion and the definition of v_r. The desired inequality follows from (A2) and Proposition 2.2.2(b)–(c) with

$$(\psi_s, \psi_n, z_{k-1}) = (f, h, z), \quad (q_k, \varepsilon_k) = (q_r, \varepsilon_r), \quad L = L_\lambda, \quad \lambda = \frac{1}{L_\lambda}.$$

(c) Let $(\bar{\rho}, \bar{\varepsilon})$ and $(z^-, \tilde{v}, \tilde{\varepsilon})$ satisfying (3.18) be given, and define the function

$$\psi_s(u) := f(u) + \frac{1}{2\lambda}\|u - z^-\|^2 - \frac{1}{\lambda}\langle \tilde{v}, u \rangle \quad \forall u \in Z. \tag{D.1}$$

Clearly, the inclusion in (3.18) holds if and only if $0 \in \partial_{\tilde{\varepsilon}}(\psi_s + h)(z)$, or equivalently, $(\psi_s + h)(u) \geq (\psi_s + h)(z) - \tilde{\varepsilon}$ for every $u \in \mathcal{Z}$. In particular, for $u = z_r$, we have $(\psi_s + h)(z) - (\psi_s + h)(z_r) \leq \tilde{\varepsilon}$. Using the previous bound, the second inequality in (3.18), and Proposition 2.2.2(c) with

$$(z^-, \psi_n) = (z, h), \quad L = L_\lambda, \quad \lambda = \frac{1}{L_\lambda},$$

it holds that there exists $(q_\psi, \varepsilon_\psi) \in \mathcal{Z} \times \mathbb{R}_+$ satisfying $q_\psi \in \nabla\psi_s(z) + \partial_{\varepsilon_\psi} h(z)$ and

$$\|q_\psi\|^2 + L_\lambda \varepsilon_\psi \le L_\lambda \left[(\psi_s + h)(z) - (\psi_s + h)(z_r)\right] \le L_\lambda \tilde{\varepsilon} \le L_\lambda \bar{\varepsilon}.$$

Since the previous inclusion implies that $q_\psi + (z^- - z + \tilde{v})/\lambda \in \nabla f(z) + \partial_{\varepsilon_\psi} h(z)$, it follows from Proposition 2.2.2(b) with

$$(z^-, q, \varepsilon) = (z, q_r, \varepsilon_r), \quad (\psi_s, \psi_n) = (f, h), \quad L = L_\lambda, \quad \lambda = \frac{1}{L_\lambda},$$

the first inequality in (3.18), and the triangle inequality, that

$$\begin{aligned}
\|q_r\|^2 \le \|q_r\|^2 + 2L_\lambda \varepsilon_r &\le \left\| q_\psi + \frac{1}{\lambda}(z^- - z + \tilde{v}) \right\|^2 + 2L_\lambda \varepsilon_\psi \\
&\le \left(\|q_\psi\| + \frac{1}{\lambda}\|z^- - z + \tilde{v}\| \right)^2 + 2L_\lambda \varepsilon_\psi \\
&\le \left(\|q_\psi\|^2 + 2L_\lambda \varepsilon_\psi \right) + 2\rho\|q_\psi\| + \rho^2 \\
&\le 2L_\lambda \bar{\varepsilon} + 2\rho\sqrt{2L_\lambda} + \rho^2 = \left(\bar{\rho}^2 + L_\lambda \bar{\varepsilon}\right)^2,
\end{aligned} \tag{D.2}$$

which implies the second inequality in (3.19). On the other hand, using assumption (A2), i.e,. ∇f is $\max\{m, M\}$-Lipschitz continuous, and the definitions of v_r and q_r yields

$$\begin{aligned}
\|v_r\| - \|q_r\| \le \|v_r - q_r\| &= \|\nabla f(z_r) - \nabla f(z)\| \le \max\{m, M\}\|z_r - z\| \\
&= \frac{\max\{m, M\}}{L_\lambda}\|q_r\|,
\end{aligned}$$

which, together with (D.2), implies the second inequality in (3.19). □

Proof of Proposition 5.1.1. (a) This follows from assumptions (A1)–(A2), the definition of ε_r, and Proposition 2.2.2(b) with

$$(\psi_s, \psi_n, z_{k-1}) = (f_\lambda, h_\lambda, z), \quad L = L_\lambda, \quad \lambda = \frac{1}{L_\lambda}.$$

(b) The optimality of z_r implies that

$$\begin{aligned}\partial h(z_r) \ni -\frac{1}{\lambda}\nabla f_\lambda(z) + \frac{L_\lambda}{\lambda}(z_r - z) \\ &= -\nabla f(z) + \frac{1}{\lambda}(z^- - z + v) + \frac{L_\lambda}{\lambda}(z - z_r) \\ &= v_r - \nabla f(z_r)\end{aligned}$$

which immediately implies the desired inclusion. To show the desired inequality, we use part (a), the triangle inequality, assumption (A2), and the definition of v_r to conclude that

$$\begin{aligned}\|v_r\| &\le \frac{1}{\lambda}\|z^- - z + v\| + \left\|\frac{L_\lambda}{\lambda}(z - z_r) + \nabla f(z_r) - \nabla f(z)\right\| \\ &\le \frac{1}{\lambda}\|z^- - z + v\| + \left(\frac{L_\lambda}{\lambda} + \max\{m, M\}\right)\|z - z_r\| \\ &\le \frac{1}{\lambda}\|z^- - z + v\| + \left(\frac{L_\lambda}{\lambda} + \max\{m, M\}\right)\sqrt{\frac{2\varepsilon_r}{L_\lambda}} \\ &= \frac{1}{\lambda}\|z^- - z + v\| + \left(\frac{1}{\lambda} + \frac{\max\{m, M\}}{L_\lambda}\right)\sqrt{2\varepsilon_r L_\lambda}.\end{aligned}$$

(c) Using the inclusion in (5.1) and the definition of ε_r, it holds that

$$\begin{aligned}\varepsilon_r &= (f_\lambda + h_\lambda)(z) - (f_\lambda + h_\lambda)(z_r) \\ &= \lambda(f + h)(z) - \left[\lambda(f + h)(z_r) + \frac{1}{2}\|z - z_r\|^2\right] + \langle v, z_r - z\rangle \\ &\le \varepsilon.\end{aligned}$$

Combining the above bound with part (b) and the inequalities in (5.1) yields

$$\begin{aligned}\|v_r\| &\le \frac{1}{\lambda}\|z^- - z + v\| + \left(\frac{1}{\lambda} + \frac{\max\{m, M\}}{L_\lambda}\right)\sqrt{2\varepsilon_r L_\lambda} \\ &\le \bar{\rho} + \left(\frac{1}{\lambda} + \frac{\max\{m, M\}}{L_\lambda}\right)\sqrt{2\lambda\bar{\varepsilon}L_\lambda}.\end{aligned}$$

□

APPENDIX E
CONVEX FUNCTIONS AND CONVEX SETS

This appendix consists of several appendices that contain results related to convex functions and convex sets.

E.1 Properties of Subdifferentials

The below technical result presents a fact about approximate subdifferentials, and its proof can be found, for example, in [69, Lemma A.2].

Lemma E.1.1. *Let proper function* $\tilde{\phi} : \mathfrak{R}^n \to (-\infty, \infty]$, *scalar* $\tilde{\sigma} \in (0,1)$ *and* $(z_0, z_1) \in Z \times \operatorname{dom} \tilde{\phi}$ *be given, and assume that there exists* (v_1, ε_1) *such that*

$$v_1 \in \partial_{\varepsilon_1}\left(\tilde{\phi} + \frac{1}{2}\|\cdot - z_0\|^2\right)(z_1), \quad \|v_1\|^2 + 2\varepsilon_1 \le \tilde{\sigma}^2\|v + z_0 - z_1\|^2. \tag{E.1}$$

Then, for every $z \in Z$ *and* $s > 0$, *we have*

$$\tilde{\phi}(z_1) + \frac{1}{2}\left[1 - \tilde{\sigma}^2(1 + s^{-1})\right]\|v_1 + z_0 - z_1\|^2 \le \tilde{\phi}(z) + \frac{s+1}{2}\|z - z_0\|^2.$$

E.2 Properties of Convex Cones

The first result presents some well-known (see, for example, [7, Chapter 6] and [99, Example 11.4]) properties about the projection and distance functions over a closed convex set.

Lemma E.2.1. *Let* $\mathcal{K} \subseteq \mathcal{Z}$ *be a closed convex set. Then the following properties hold:*

(a) for every $u, z \in \mathcal{Z}$, *we have* $\|\Pi_{\mathcal{K}}(u) - \Pi_{\mathcal{K}}(u)\| \le \|u - z\|$;

(b) the function $d(\cdot) := \operatorname{dist}^2(\cdot, \mathcal{K})/2$ *is differentiable, and its gradient, given by*

$$\nabla d(u) = u - \Pi_{\mathcal{K}}(u) \in N_{\mathcal{K}}(\Pi_{\mathcal{K}}(u)) \quad \forall u \in \mathbb{R}^n, \tag{E.2}$$

is 1-Lipschitz continuous;

(c) if $\mathcal{K}$ is a cone, then holds that $u \in N_{\mathcal{K}^+}(p)$ if and only if $\langle u, p\rangle = 0$, $u \in -\mathcal{K}$, and $p \in \mathcal{K}^+$.

The next result presents a well-known fact (see, for example, [25, Sub-subsection 2.13.2]) about closed convex cones.

Lemma E.2.2. *For any closed convex cone $\mathcal{K}$, we have that $x \in \operatorname{int}\mathcal{K}$ if and only if*

$$\inf_{p\in\mathcal{K}^+} \{\langle p, x\rangle : \|p\| = 1\} > 0.$$

E.3 Properties of Max Functions

This appendix contains results about functions that can be described be as the maximum of a family of differentiable functions.

The technical lemma below, which is a special case of [27, Theorem 10.2.1], presents a key property about max functions.

Lemma E.3.1. *Assume that the triple (Ψ, X, Y) satisfies (F1)–(F2) in Section 6.1 with $\Phi = \Psi$. Moreover, define*

$$q(x) := \sup_{y\in Y} \Psi(x, y), \quad Y(x) := \{y \in Y : \Psi(x, y) = q(x)\}, \quad \forall x \in X. \tag{E.3}$$

Then, for every $(x, d) \in X \times \mathcal{X}$, it holds that

$$q'(x; d) = \max_{y\in Y(x)} \langle \nabla_x \Psi(x; y), d\rangle.$$

Moreover, if $Y(x)$ reduces to a singleton, say $Y(x) = \{y(x)\}$, then q is differentiable at x and $\nabla q(x) = \nabla_x \Psi(x, y(x))$.

Under assumptions (F1)–(F4) in Section 6.1, the next result establishes Lipschitz continuity of the gradient of q. It is worth mentioning that it generalizes related results in [7,

Theorem 5.26] (which covers the case where Ψ is bilinear) and [76, Proposition 4.1] (which makes the stronger assumption that $\Psi(\cdot, y)$ is convex for every $y \in Y$).

Proposition E.3.2. *If the triple (Ψ, X, Y) satisfies (F1)–(F4) in Section 6.1 with $\Phi = \Psi$ and it holds that $-\Psi(x, \cdot) \in \mathcal{F}_\mu(Y)$ for some $\mu > 0$ and every $x \in X$, then the following properties hold:*

(a) the function $y(\cdot)$ given by

$$y(x) := \operatorname*{argmax}_{y \in Y} \Psi(x, y) \quad \forall x \in X$$

is Q_μ-Lipschitz continuous on X, where

$$Q_\mu := \frac{L_y}{\mu} + \sqrt{\frac{L_x + m}{\mu}}; \tag{E.4}$$

(b) $\nabla q(\cdot)$ is L_μ-Lipschitz continuous on X, where q is as in (E.3) *and*

$$L_\mu := L_y Q_\mu + L_x. \tag{E.5}$$

Proof. (a) Let $x, \tilde{x} \in X$ be given and denote $(y, \tilde{y}) = (y(x), y(\tilde{x}))$. Define

$$\alpha(u) := \Psi(u, y) - \Psi(u, \tilde{y}) \quad \forall u \in X. \tag{E.6}$$

and observe that the optimality conditions of y and $\tilde{y}$ imply that

$$\alpha(x) \geq \frac{\mu}{2}\|y - \tilde{y}\|^2, \quad -\alpha(\tilde{x}) \geq \frac{\mu}{2}\|y - \tilde{y}\|^2. \tag{E.7}$$

Using (E.7), (6.6), (6.7), (6.8), and the Cauchy-Schwarz inequality, we conclude that

$$\begin{aligned}\mu\|y-\tilde{y}\|^2 \le \alpha(x)-\alpha(\tilde{x}) &\le \langle \nabla_x\Psi(x,y)-\nabla_x\Psi(x,\tilde{y}),x-\tilde{x}\rangle + \frac{L_x+m}{2}\|x-\tilde{x}\|^2\\ &\le \|\nabla_x\Psi(x,y)-\nabla_x\Psi(x,\tilde{y})\|\cdot\|x-\tilde{x}\| + \frac{L_x+m}{2}\|x-\tilde{x}\|^2\\ &\le L_y\|y-\tilde{y}\|\cdot\|x-\tilde{x}\| + \frac{L_x+m}{2}\|x-\tilde{x}\|^2.\end{aligned}$$

Considering the above as a quadratic inequality in $\|\tilde{y}-y\|$ yields the bound

$$\begin{aligned}\|y-\tilde{y}\| &\le \frac{1}{2\mu}\left[L_y\|x-\tilde{x}\| + \sqrt{L_y^2\|x-\tilde{x}\|^2 + 4\mu(L_x+m)\|x-\tilde{x}\|^2}\right]\\ &\le \left[\frac{L_y}{\mu} + \sqrt{\frac{L_x+m}{\mu}}\right]\|x-\tilde{x}\| = Q_\mu\|x-\tilde{x}\|\end{aligned}$$

which is the conclusion of (a).

(b) Let $x,\tilde{x}\in X$ be given and denote $(y,\tilde{y}) = (y(x),y(\tilde{x}))$. Using part (a), Lemma E.3.1, and (6.7) we have that

$$\begin{aligned}\|\nabla q(x)-\nabla q(\tilde{x})\| &= \|\nabla_x\Psi(x,y)-\nabla_x\Psi(\tilde{x},\tilde{y})\|\\ &\le \|\nabla_x\Psi(x,y)-\nabla_x\Psi(x,\tilde{y})\| + \|\nabla_x\Psi(x,\tilde{y})-\nabla_x\Psi(\tilde{x},\tilde{y})\|\\ &\le L_y\|y-\tilde{y}\| + L_x\|x-\tilde{x}\| \le (L_yQ_\mu+L_x)\|x-\tilde{x}\| = L_\mu\|x-\tilde{x}\|,\end{aligned}$$

which is the conclusion of (b). □

APPENDIX F
NOTIONS OF STATIONARY POINTS

This appendix contains technical results about different notions of stationary points in an optimization problem.

F.1 Directional and Primal-Dual Stationarity

The main goal of this appendix is to prove Propositions F.1.4 and F.1.5, which are used in the proofs of Propositions 6.1.1 to 6.1.3 given in Appendix F.2. Several technical lemmas are stated and proved to accomplish the above goal. Some of these technical results (e.g. Lemma F.1.1(a) and Lemma F.1.3) are stated without proof as they are broadly available in the convex analysis literature. Others (e.g. Lemma F.1.1(b) and Lemma F.1.2) are given proofs because we could not find a suitable reference for them.

The first technical lemma presents some general results about proper convex functions and nonempty closed convex sets.

Lemma F.1.1. *Let ψ be a convex function and let $C \subseteq \mathcal{X}$ be a nonempty closed convex set. Then, the following statements hold:*

(a) $\inf_{\|d\|\leq 1} \sigma_C(d) = [-\min_{u\in C} \|u\|]$;

(b) if $C \cap \mathrm{ri}(\mathrm{dom}\, \psi) \neq \emptyset$, *then* $\inf_{x\in C} \mathrm{cl}\, \psi(x) = \inf_{x\in C} \psi(x) < \infty$.

Proof. (a) See, for example, the proof of [15, Lemma 5.1] with $g = 0$.

(b) Define $\psi_* := \inf_{x\in C} \psi(x), \quad \psi_*^{\mathrm{cl}} := \inf_{x\in C} \mathrm{cl}\, \psi(x)$. Then, note that the assumption of (b) implies that $\psi_* < \infty$. Now, assume for contradiction that the conclusion of (b) does not hold. Since $\mathrm{cl}\, \psi \leq \psi$, and hence $\psi_*^{\mathrm{cl}} \leq \psi_*$, we must have $\psi_*^{\mathrm{cl}} < \psi_*$. Hence, due to a well-known infimum property, there exists $\bar{x} \in C$ such that $\mathrm{cl}\, \psi(\bar{x}) < \psi_* < \infty$. In particular, it follows that $\psi_* \in \mathbb{R}$, and hence that $\psi(x) > -\infty$ for every $x \in C$, in view of the definition of ψ_*. Now, by assumption, there exists $x_0 \in C \cap \mathrm{ri}(\mathrm{dom}\, \psi)$ which, in view of the previous

conclusion, satisfies $\psi(x_0) > -\infty$. As $x_0 \in \mathrm{ri}(\mathrm{dom}\,\psi)$, this implies that ψ is proper due to [98, Theorem 7.2]. Hence, in view of [98, Theorem 7.5] with $f = \psi$, we have

$$\mathrm{cl}\,\psi(\bar{x}) = \lim_{y \in (\bar{x}, x_0],\, y \to \bar{x}} \psi(y)$$

where $(\bar{x}, x_0] := \{tx_0 + (1-t)\bar{x} : t \in (0,1]\}$. On the other hand, as $x_0, \bar{x} \in C$ and C is convex, we have $(\bar{x}, x_0] \subseteq C$. This inclusion and the definition of ψ_* then imply that the above limit, and hence $\mathrm{cl}\,\psi(\bar{x})$, is greater than or equal to ψ_*, which contradicts the previously obtained inequality $\psi_* > \mathrm{cl}\,\psi(\bar{x})$. □

The following technical lemma presents an important property about the directional derivative of a composite function $(f + h)$.

Lemma F.1.2. *Let $h : \mathcal{X} \mapsto (-\infty, \infty]$ be a proper convex function and let f be a differentiable function on* $\mathrm{dom}\,h$*. Then, for any $x \in \mathrm{dom}\,h$, it holds that*

$$\inf_{\|d\| \le 1} (f+h)'(x;d) = \inf_{\|d\| \le 1} \left[\langle \nabla f(x), d \rangle + \sigma_{\partial h(x)}(d) \right] = - \inf_{u \in \nabla f(x) + \partial h(x)} \|u\|. \tag{F.1}$$

Proof. Let $x \in \mathrm{dom}\,h$ be fixed and define $\tilde{h}(\cdot) := \langle \nabla f(x), \cdot \rangle + h(\cdot)$. We first claim that $\inf_{\|d\| \le 1} \tilde{h}'(x;d) = \inf_{\|d\| \le 1} [\mathrm{cl}\,\tilde{h}'(x;\cdot)](d)$. Before showing this claim, let us show how it proves the desired conclusion. Since the definition of $\tilde{h}$ implies that $(f+h)'(x;\cdot) = \tilde{h}'(x;\cdot)$ and $\partial\tilde{h}(x) = \nabla f(x) + \partial h(x)$, it follows from our previous claim and [98, Theorem 23.2] with $f = \tilde{h}$ that

$$\begin{aligned} \inf_{\|d\| \le 1} (f+h)'(x;d) &= \inf_{\|d\| \le 1} \tilde{h}'(x;d) = \inf_{\|d\| \le 1} [\mathrm{cl}\,\tilde{h}'(x;\cdot)](d) \\ &= \inf_{\|d\| \le 1} \sigma_{\partial\tilde{h}(x)}(d) = \inf_{\|d\| \le 1} \left[\langle \nabla f(x), d \rangle + \sigma_{\partial h(x)}(d) \right], \end{aligned} \tag{F.2}$$

which gives the first identity in (F.1). The second identity in (F.1) follows from Lemma F.1.1 with $C = \partial\tilde{h}(x)$ and the last identity in (F.2).

To complete the proof, we now justify the claim made the in the previous paragraph. Define $\mathcal{B} := \{d \in \mathcal{X} : \|d\| \leq 1\}$ and $\psi(\cdot) := \tilde{h}'(x;\cdot)$. In view of Lemma F.1.1 with $C = \mathcal{B}$, it suffices to show that $\mathcal{B} \cap \mathrm{ri}(\mathrm{dom}\,\psi) \neq \varnothing$. To show this, note that the convexity of $\tilde{h}$ and the discussion following [98, Theorem 23.1] imply that $\mathrm{dom}\,\psi = \bigcup_{t>0}(\mathrm{dom}\,h - x)/t$, which is a nonempty convex cone. Hence, it follows from [98, Theorem 6.2] and the discussion in the second paragraph following [98, Corollary 6.8.1] that $\mathrm{ri}(\mathrm{dom}\,\psi)$ is also a nonempty convex cone. This conclusion clearly implies that $\mathcal{B} \cap \mathrm{ri}(\mathrm{dom}\,\psi) \neq \varnothing$. □

It is worth mentioning that the result above is a generalization of the one given in [16, Lemma 5.1], which only considers the case where $(f+h)$ is real-valued and locally Lipschitz.

The next technical lemma, which can be found in [103, Corollary 3.3], presents a well-known min-max identity.

Lemma F.1.3. *Let a convex set $D \subseteq \mathcal{X}$ and compact convex set $Y \subseteq \mathcal{Y}$ be given. Moreover, let $\psi : D \times Y \mapsto \mathbb{R}$ be a function in which $\psi(\cdot, y)$ is convex lower semicontinuous for every $y \in Y$ and $\psi(d,\cdot)$ is concave upper semicontinuous for every $d \in D$. Then,*

$$\inf_{d\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \psi(d,y) = \sup_{y\in\mathcal{Y}} \inf_{d\in\mathcal{X}} \psi(d,y).$$

The next result establishes an identity similar to Lemma F.1.2 but for the case where f is a max function.

Proposition F.1.4. *Assume the quadruple (Ψ, h, X, Y) satisfies assumptions (F1)–(F4) of Section 6.1 with $\Phi = \Psi$. Moreover, suppose that $\Psi(\cdot, y)$ is convex for every $y \in Y$, and let q and $Y(\cdot)$ be as in Lemma E.3.1. Then, for every $\bar{x} \in X$, it holds that*

$$\inf_{\|d\|\leq 1} (q+h)'(\bar{x};d) = -\inf_{u\in Q(\bar{x})} \|u\| \tag{F.3}$$

where

$$Q(\bar{x}) := \partial h(\bar{x}) + \bigcup_{y\in Y(\bar{x})} \{\nabla_x \Psi(\bar{x},y)\}. \tag{F.4}$$

Moreover, if $\partial h(\bar{x})$ is nonempty, then the infimum on the right-hand side of (F.3) *is achieved.*

Proof. Let $\bar{x} \in X$ and define

$$\psi(d, y) := (\Psi_y + h)'(\bar{x}; d), \quad \forall (d, x, y) \in \mathcal{X} \times \Omega \times Y. \tag{F.5}$$

We claim that ψ in (F.5) satisfies the assumptions on ψ in Lemma F.1.3 with $Y = Y(\bar{x})$ and D given by

$$D := \{d \in \mathcal{Z} : \|d\| \leq 1, d \in F_X(\bar{x})\},$$

where $F_X(\bar{x}) := \{t(x - \bar{x}) : x \in X, t \geq 0\}$ is the set of feasible directions at $\bar{x}$. Before showing this claim, we use it to show that (F.3) holds. First observe that (F2) and Lemma E.3.1 imply that $q'(\bar{x}; d) = \sup_{y \in Y} \Psi'_y(\bar{x}; d)$ for every $d \in \mathcal{X}$. Using then Lemma F.1.3 with $Y = Y(\bar{x})$, Lemma F.1.2 with $(f, x) = (\Psi_{\bar{y}}, \bar{x})$ for every $\bar{y} \in Y(\bar{x})$, and the previous observation, we have that

$$\begin{aligned}
&\inf_{\|d\| \leq 1} (q + h)'(\bar{x}; d) = \inf_{d \in D} (q + h)'(\bar{x}; d) = \inf_{d \in D} \sup_{y \in Y(\bar{x})} (\Psi_y + h)'(\bar{x}; d) \\
&= \inf_{d \in D} \sup_{y \in Y(\bar{x})} \psi(d, y) = \sup_{y \in Y(\bar{x})} \inf_{d \in D} \psi(d, y) = \sup_{y \in Y(\bar{x})} \inf_{\|d\| \leq 1} (\Psi_y + h)'(\bar{x}; d) \\
&= \sup_{y \in Y(\bar{x})} \left[- \inf_{u \in \nabla_x \Phi(\bar{x}, y) + \partial h(\bar{x})} \|u\| \right] = \left[- \inf_{u \in Q(\bar{x})} \|u\| \right].
\end{aligned} \tag{F.6}$$

Let us now assume that $\partial h(\bar{x})$ is nonempty, and hence, $Q(\bar{x})$ is nonempty as well. Note that continuity of the function $\nabla_x \Psi(\bar{x}, \cdot)$ from assumption (F2) and the compactness of $Y(\bar{x})$ imply that Q is closed. Moreover, since $\|u\| \geq 0$, it holds that any sequence $\{u_k\}_{k \geq 1}$ where $\lim_{k \to \infty} \|u_k\| = \inf_{u \in Q(\bar{x})} \|u\|$ is bounded. Combining the previous two remarks with the Bolzano-Weierstrass Theorem, we conclude that $\inf_{u \in Q(\bar{x})} \|u\| = \min_{u \in Q(\bar{x})} \|u\|$, and hence (F.3) holds.

To complete the proof, we now justify the above claim on ψ. First, for any given $y \in Y(\bar{x})$,

it follows from [98, Theorem 23.1] with $f(\cdot) = \Psi_y(\cdot)$ and the definitions of q and $Y(\bar{x})$ that

$$\psi(d, \bar{y}) = \Psi'_{\bar{y}}(\bar{x}; d) = \inf_{t>0} \frac{\Psi_y(\bar{x} + td) - q(\bar{x})}{t} \quad \forall d \in \mathcal{X}. \tag{F.7}$$

Since assumption (F3) implies that $\Psi(\bar{x}, \cdot)$ is upper semicontinuous and concave on Y, it follows from (F.7), [98, Theorem 5.5], and [98, Theorem 9.4] that $\psi(d, \cdot)$ is upper semicontinuous and concave on Y for every $d \in \mathcal{X}$. On the other hand, since $\Psi(\cdot, y)$ is assumed to be lower semicontinuous and convex on X for every $y \in Y$, it follows from (F.7), the fact that $\bar{x} \in \operatorname{int} \Omega$, and [98, Theorem 23.4], that $\psi(\cdot, y)$ is lower semicontinuous and convex on $\mathcal{X}$, and hence $D \subseteq \mathcal{X}$, for every $y \in Y(\bar{x})$. □

The last technical result is a specialization of the one given in [39, Theorem 4.2.1].

Proposition F.1.5. *Let a proper closed function $\phi : \mathcal{X} \mapsto (-\infty, \infty]$ and assume that $[\phi + \|\cdot\|^2/2\lambda] \in \mathcal{F}_\mu(\mathcal{X})$ for some scalars $\mu, \lambda > 0$. If a quadruple $(x^-, x, u, \varepsilon) \in \mathcal{X} \times \operatorname{dom} \phi \times \mathcal{X} \times \mathbb{R}_+$ together with λ satisfy*

$$u \in \partial_\varepsilon \left(\phi + \frac{1}{2\lambda} \|\cdot - x^-\|^2 \right)(x), \tag{F.8}$$

then the point $\hat{x} \in \operatorname{dom} \phi$ given by

$$\hat{x} := \operatorname*{argmin}_{x'} \left\{ \phi_\lambda(x') := \phi(x') + \frac{1}{2\lambda} \|x' - x^-\|^2 - \langle u, x' \rangle \right\} \tag{F.9}$$

satisfies

$$\inf_{\|d\| \le 1} \phi'(\hat{x}; d) \ge -\frac{1}{\lambda} \|x^- - x + \lambda u\| - \sqrt{\frac{2\varepsilon}{\lambda^2 \mu}}, \quad \|\hat{x} - x\| \le \sqrt{\frac{2\varepsilon}{\mu}}. \tag{F.10}$$

Proof. We first observe that (F.8) implies that

$$\phi_\lambda(x') \ge \phi_\lambda(x) - \varepsilon \quad \forall x' \in \mathcal{X}. \tag{F.11}$$

Remark that (F.11) at $x' = \hat{x}$, the optimality of $\hat{x}$, and the μ–strong convexity of ϕ_λ imply that

$$\frac{\mu}{2}\|\hat{x} - x\|^2 \leq \phi_\lambda(x) - \phi_\lambda(\hat{x}) \leq \varepsilon$$

from which we conclude that $\|\hat{x} - x\| \leq \sqrt{2\varepsilon/\mu}$, i.e. the second inequality in (F.10). On the other hand, using the definition of ϕ_λ, the triangle inequality, and the previous bound on $\|\hat{x} - x\|$, we obtain

$$\begin{aligned} 0 \leq \inf_{\|d\|\leq 1} \phi'_\lambda(\hat{x}; d) &= \inf_{\|d\|\leq 1} \phi'(\hat{x}; d) - \frac{1}{\lambda}\langle d, \lambda u + x^- - \hat{x}\rangle \\ &\leq \inf_{\|d\|\leq 1} \phi'(\hat{x}; d) + \frac{\|x^- - x + \lambda u\|}{\lambda} + \frac{\|x - \hat{x}\|}{\lambda} \\ &\leq \inf_{\|d\|\leq 1} \phi'(\hat{x}; d) + \frac{\|x^- - x + \lambda u\|}{\lambda} + \sqrt{\frac{2\varepsilon}{\lambda^2\mu}}, \end{aligned} \tag{F.12}$$

which clearly implies the first inequality in (F.10). □

F.2 Equivalent Notions of Stationarity

This appendix presents the proofs of Propositions 6.1.1 to 6.1.3.

The first technical result shows that an approximate primal-dual stationary point is equivalent to an approximate directional stationary point of a perturbed version of problem $\mathcal{MCO}$.

Lemma F.2.1. *Suppose the quadruple (Φ, h, X, Y) satisfies assumptions (F1)–(F4) of Section 6.1 and let $(\bar{x}, \bar{u}, \bar{v}) \in X \times \mathcal{X} \times \mathcal{Y}$ be given. Then, there exists $\bar{y} \in Y$ such that the quadruple $(\bar{u}, \bar{v}, \bar{x}, \bar{y})$ satisfies the inclusion in* (6.3) *if and only if*

$$\inf_{\|d\|\leq 1} (p_{\bar{u},\bar{v}} + h)'(\bar{x}; d) \geq 0, \tag{F.13}$$

where the function $p_{\bar{u},\bar{v}}$ is given by

$$p_{\bar{u},\bar{v}}(x) := \max_{y\in Y}\left[\Phi(x,y) + \langle \bar{v}, y\rangle - \langle \bar{u}, x\rangle\right] \quad \forall x \in \Omega. \tag{F.14}$$

Proof. Let $(\bar{x},\bar{u},\bar{v}) \in X \times \mathcal{X} \times \mathcal{Y}$ be given, define

$$\Psi(x,y) := \Phi(x,y) + \langle \bar{v}, y\rangle - \langle \bar{u}, x\rangle + m\|x-\bar{x}\|^2 \quad \forall (x,y) \in \Omega \times Y, \tag{F.15}$$

and let q and $Y(\cdot)$ be as in Lemma E.3.1. It is easy to see that $q = p_{\bar{u},\bar{v}}$, the function Ψ satisfies the assumptions on Ψ in Proposition F.1.4, and $\bar{x}$ satisfies (F.13) if and only if $\inf_{\|d\|\le 1}(q + h)'(\bar{x};d) \ge 0$. The desired conclusion follows from Proposition F.1.4, the previous observation, and the fact that $\bar{y} \in Y(\bar{x})$ if and only if $\bar{v} \in \partial[-\Phi(\bar{x},\cdot)](\bar{y})$. □

We are now ready to give the proof of Proposition 6.1.1.

Proof of Proposition 6.1.1. Suppose $([\bar{x},\bar{y}],[\bar{u},\bar{v}])$ is a (ρ_x,ρ_y)-primal-dual stationary point of $\mathcal{MCO}$. Moreover, let Ψ, q, and D_y be as in (F.15), (E.3) and (6.13), respectively, and define

$$\hat{q}(x) := q(x) + h(x) \quad \forall x \in X.$$

Using Lemma F.2.1, we first observe that $\inf_{\|d\|\le 1}\hat{q}(\bar{x};d) \ge 0$. Since $\hat{q}$ is convex from assumption (F4), it follows from the previous bound and Lemma F.1.2 with $(f,h) = (0,\hat{q})$, that $\min_{u\in\partial\hat{q}(\bar{x})}\|u\| \le 0$, and hence, $0 \in \partial\hat{q}(\bar{x})$. Moreover, using the Cauchy-Schwarz inequality, the second inequality in (6.3), the previous inclusion, and the definition of q and Ψ, it follows that for every $x \in \mathcal{X}$,

$$\hat{p}(x) + D_y\rho_y - \langle \bar{u}, x\rangle + m\|x-\bar{x}\|^2 \ge \hat{q}(x) \ge \hat{q}(\bar{x}) \ge \hat{p}(\bar{x}) - D_y\rho_y - \langle \bar{u}, \bar{x}\rangle,$$

and hence that $\bar{u} \in \partial_\varepsilon(\hat{p} + m\|\cdot - \bar{x}\|^2)(\bar{x})$ where $\varepsilon = 2D_y\rho_y$. Using now the first inequality in (6.3), Proposition F.1.5 with $(\phi, x, x^-, u) = (\hat{p},\bar{x},\bar{x},\bar{u})$ and also $(\varepsilon,\lambda,\mu) =$

$(D_y\rho_y, 1/(2m), m)$, we conclude that there exists $\hat{x}$ such that $\|\hat{x} - \bar{x}\| \le \sqrt{2D_y\rho_y/m}$ and

$$\inf_{\|d\|\le 1} \hat{p}'(\hat{x}; d) \ge -\|\bar{u}\| - 2\sqrt{2mD_y\rho_y} \ge -\rho_x - 2\sqrt{2mD_y\rho_y}.$$

□

We next give the proof of Proposition 6.1.2.

Proof of Proposition 6.1.2. (a) We first claim that $\hat{P}_\lambda \in \mathcal{F}_\alpha(X)$, where $\alpha = 1/\lambda - m$. To see this, note that $\Phi(\cdot, y) + m\|\cdot\|^2/2$ is convex for every $y \in Y$ from assumption (F4). The claim now follows from assumption (F3), the fact that the supremum of a collection of convex functions is also convex, and the definition of $\hat{p}$ in $\mathcal{MCO}$.

Suppose the pair (x, δ) satisfies (6.4) and (6.15). If $\hat{x} = x_\lambda$ in (6.4), then clearly the second inequality in (6.4), the fact that $\lambda < 1/m$, and (6.15) imply the inequality in (6.14), and hence, that x is a (λ, ε)-prox stationary point. Suppose now that $\hat{x} \ne x_\lambda$. Using the convexity of $\hat{P}_\lambda$, we first have that $\hat{P}'_\lambda(\hat{x}; d) = \inf_{t>0} \left[\hat{P}_\lambda(\hat{x} + td) - \hat{P}_\lambda(\hat{x})\right]/t$ for every $d \in \mathcal{X}$. Hence, using both inequalities in (6.4) and the previous identity, it holds that

$$\begin{aligned}\frac{\hat{P}_\lambda(x_\lambda) - \hat{P}_\lambda(\hat{x})}{\|x_\lambda - \hat{x}\|} \ge \hat{P}'_\lambda\left(\hat{x}; \frac{x_\lambda - \hat{x}}{\|x_\lambda - \hat{x}\|}\right) &= \hat{p}'\left(\hat{x}; \frac{x_\lambda - \hat{x}}{\|x_\lambda - \hat{x}\|}\right) + \frac{1}{\lambda}\left\langle \frac{x_\lambda - \hat{x}}{\|x_\lambda - \hat{x}\|}, \hat{x} - x\right\rangle \\ &\ge -\delta - \frac{1}{\lambda}\|\hat{x} - x\| \ge -\delta\left(\frac{1+\lambda}{\lambda}\right).\end{aligned}$$

Using the optimality of x_λ, the α-strong convexity of $\hat{P}_\lambda$ (see our claim on $\hat{p}$ in the first paragraph), and the above bound, we conclude that

$$\frac{1}{2\alpha}\|\hat{x} - x_\lambda\|^2 \le \hat{P}_\lambda(\hat{x}) - \hat{P}_\lambda(x_\lambda) \le \delta\left(\frac{1+\lambda}{\lambda}\right)\|\hat{x} - x_\lambda\|.$$

Thus, $\|\hat{x} - x_\lambda\| \le 2\alpha\delta(1+\lambda)/\lambda$. Using the previous bound, the second inequality in (6.4),

and (6.15) yields

$$\|x - x_\lambda\| \le \|x - \hat{x}\| + \|\hat{x} - x_\lambda\| \le \left(1 + 2\alpha\left[\frac{1+\lambda}{\lambda}\right]\right)\delta \le \lambda\varepsilon,$$

which implies (6.14), and hence, that x is a (λ, ε)-prox stationary point.

(b) Suppose that the point x is a (λ, ε)-prox stationary point with $\varepsilon \le \delta \cdot \min\{1, 1/\lambda\}$. Then the optimality of x_λ, the fact that $\hat{P}_\lambda$ is convex (see the beginning of part (a)), the inequality in (6.14), and the Cauchy-Schwarz inequality imply that

$$0 \le \inf_{\|d\|\le 1}\left[\hat{p}'(x_\lambda; d) + \frac{1}{\lambda}\langle d, x_\lambda - x\rangle\right] \le \inf_{\|d\|\le 1}\hat{p}'(x_\lambda; d) + \varepsilon \le \inf_{\|d\|\le 1}\hat{p}'(x_\lambda; d) + \delta,$$

which, together with the fact that $\lambda\varepsilon \le \delta$, imply that x satisfies (6.4) with $\hat{x} = x_\lambda$. □

Finally, we give the proof of Proposition 6.1.3.

Proof of Proposition 6.1.3. This follows by using Lemma F.1.2 with $(f, h) = (\Phi(\cdot, \bar{y}), h)$ and $(f, h) = (0, -\Phi(\bar{x}, \cdot))$. □

APPENDIX G
SPECTRAL FUNCTIONS

This section presents some results about spectral functions as well as the proof of Proposition 7.5.1. It is assumed that the reader is familiar with the key quantities given in Section 7.5.1, e.g. (7.45), and the functions in (7.44).

We first state two well-known results [7, 55] about spectral functions.

Lemma G.0.1. *Let $\Psi = \Psi^{\mathcal{V}} \circ \sigma$ for some absolutely symmetric function $\widetilde{\Psi} : \mathbb{R}^r \mapsto \mathbb{R}$. Then, the following properties hold:*

(a) $\Psi^* = \left(\Psi^{\mathcal{V}} \circ \sigma\right)^* = \left(\Psi^{\mathcal{V}}\right)^* \circ \sigma$;

(b) $\nabla\Psi = \left(\nabla\Psi^{\mathcal{V}}\right) \circ \sigma$;

Lemma G.0.2. *Let $(\Psi, \Psi^{\mathcal{V}})$ be as in Lemma G.0.1, the pair $(S, Z) \in \mathcal{Z} \times \operatorname{dom}\Psi$ be fixed, and the decomposition $S = P[\operatorname{dg}\sigma(S)]Q^*$ be an SVD of S, for some $(P, Q) \in \mathcal{U}^m \times \mathcal{U}^n$. If $\Psi \in \overline{\operatorname{Conv}}\,\mathbb{R}^{m\times n}$ and $\Psi^{\mathcal{V}} \in \overline{\operatorname{Conv}}\,\mathbb{R}^r$, then for every $M > 0$, we have*

$$S \in \partial\left(\Psi + \frac{M}{2}\|\cdot\|_F^2\right)(Z) \iff \begin{cases} \sigma(S) \in \partial\left(\Psi^{\mathcal{V}} + \frac{M}{2}\|\cdot\|^2\right)(\sigma(Z)), \\ Z = P[\operatorname{dg}\sigma(Z)]Q^*. \end{cases}$$

We now present a new result about spectral functions.

Theorem G.0.3. *Let $(\Psi, \Psi^{\mathcal{V}})$ be as in Lemma G.0.1 and the point $Z \in \mathbb{R}^{m\times n}$ be such that $\sigma(Z) \in \operatorname{dom}\Psi^{\mathcal{V}}$. Then for every $\varepsilon \geq 0$, we have $S \in \partial_\varepsilon\Psi(Z)$ if and only if $\sigma(S) \in \partial_{\varepsilon(S)}\Psi^{\mathcal{V}}(\sigma(Z))$, where*

$$\varepsilon(S) := \varepsilon - \left[\langle\sigma(Z), \sigma(S)\rangle - \langle Z, S\rangle\right] \geq 0. \tag{G.1}$$

Moreover, if S and Z have a simultaneous SVD, then $\varepsilon(S) = \varepsilon$.

Proof. Using Lemma G.0.1(a), (G.1), and the well-known fact that $S \in \partial_\varepsilon \Psi(Z)$ if and only if $\varepsilon \geq \Psi(Z) + \Psi^*(S) - \langle Z, S \rangle$, we have that $S \in \partial_\varepsilon \Psi(Z)$ if and only if

$$\begin{aligned}
\varepsilon(S) &= \varepsilon - [\langle \sigma(Z), \sigma(S) \rangle - \langle Z, S \rangle] \\
&\geq \Psi(Z) + \Psi^*(S) - \langle Z, S \rangle - [\langle \sigma(Z), \sigma(S) \rangle - \langle Z, S \rangle] \\
&= \Psi^{\mathcal{V}}(\sigma(Z)) + (\Psi^{\mathcal{V}})^*(\sigma(S)) - \langle \sigma(Z), \sigma(S) \rangle ,
\end{aligned}$$

or, equivalently, $\sigma(S) \in \partial_{\varepsilon(S)} \Psi^{\mathcal{V}}(\sigma(Z))$ and $\varepsilon(S) \geq 0$.

To show that the existence of a simultaneous SVD of S and Z implies $\varepsilon(S) = \varepsilon$ it suffices to show that $\langle \sigma(S), \sigma(Z) \rangle = \langle S, Z \rangle$. Indeed, if $S = P[\operatorname{dg} \sigma(S)]Q^*$ and $Z = P[\operatorname{dg} \sigma(Z)]Q^*$, for some $(P, Q) \in \mathcal{U}^m \times \mathcal{U}^n$, then we have

$$\langle S, Z \rangle = \langle \operatorname{dg} \sigma(S), P^* P[\operatorname{dg} \sigma(Z)]Q^* Q \rangle = \langle \operatorname{dg} \sigma(S), \operatorname{dg} \sigma(Z) \rangle = \langle \sigma(S), \sigma(Z) \rangle.$$

□

APPENDIX H
COMPUTATIONAL DETAILS

This appendix presents technical details about the numerical experiments considered in Section 5.5.

Generating Parameters for the Quadratic Matrix Problem

In the unconstrained QM problem of Section 5.5, recall that the objective function is of the form

$$f(Z) := \frac{\alpha_1}{2}\|\mathcal{C}Z - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}Z\|^2 \tag{H.1}$$

where $\mathcal{B}$ and $\mathcal{C}$ are linear operators, D is a diagonal matrix, and d is a vector. This appendix describes how, for a given $(m, M) \in \mathbb{R}^2_{++}$, the parameters α_1, α_2 are chosen so that $M = \lambda_{\max}(\nabla^2 f(x))$ and $-m = \lambda_{\min}(\nabla^2 f(x))$.

Suppose $\mathcal{B}$ and $\mathcal{C}$ are full rank. Define the Hessian matrix

$$H_{\xi,\tau} := \alpha_1 \mathcal{C}^*\mathcal{C} - \alpha_2 \mathcal{B}^* D^2 \mathcal{B} = \nabla^2 f(x)$$

and note that the operators $\mathcal{B}^* D^2 \mathcal{B}$ and $\mathcal{C}^*\mathcal{C}$ are symmetric positive semidefinite. By Weyl's inequality, it holds that for any $\gamma > 0$ we have

$$\lambda_k(H_{\xi,\tau} - \gamma \mathcal{B}^* D^2 \mathcal{B}) \leq \lambda_k(H_{\xi,\tau})$$
$$\lambda_k(H_{\xi,\tau}) \leq \lambda_k(H_{\xi,\tau} + \gamma \mathcal{C}^*\mathcal{C})$$

for $k = 1, \ldots, n$. The above two inequalities imply that $H_{\xi,\tau}$ is monotonically decreasing in ξ

and monotonically increasing in τ. In addition, if $\mathcal{B}$, $\mathcal{C}$, and D are nonzero, then

$$\lim_{\gamma\to\infty} \lambda_1(H_{\xi,\tau} + \gamma\mathcal{C}^*\mathcal{C}) = +\infty$$

$$\lim_{\gamma\to\infty} \lambda_n(H_{\xi,\tau} - \gamma\mathcal{C}^*\mathcal{C}) = -\infty$$

$$\lim_{\gamma\to\infty} \lambda_1(H_{\xi,\tau} + \gamma\mathcal{B}^* D^2\mathcal{B}) = +\infty$$

$$\lim_{\gamma\to\infty} \lambda_n(H_{\xi,\tau} - \gamma\mathcal{B}^* D^2\mathcal{B}) = -\infty$$

Thus, for a fixed $\xi_0 > 0$, we can find a $\tau_0 > 0$ such that $\lambda_{\max}(H_{\xi_0,\tau_0})/\lambda_{\min}(H_{\xi_0,\tau_0}) = -M/m$ by bisection search and set $(\xi, \tau) = (\xi_0, \tau_0) \cdot (M/\tau_0)$ to obtain the desired conditions $M = \lambda_{\max}(H_{\xi,\tau})$ and $-m = \lambda_{\min}(H_{\xi,\tau})$.

In Section 5.5, we implement the above approach, we with $\xi_0 = 10^{-6}$ and $\tau_0 = 1$ as an initial candidate solution.

APPENDIX I
CURVATURE CONSTANTS

This appendix presents the description of y_ξ and justification for the constants m, L_x, and L_y for each of the optimization problems in Section 5.5.

Maximum of a finite number of nonconvex functions

Recall that

$$M = \lambda_{\max}(\nabla^2 f_i), \quad -m = \lambda_{\min}(\nabla^2 f_i) \quad \forall i \in \{1, ..., k\}. \tag{I.1}$$

Since $Y = \Delta^k$, it is easy to verify that

$$y_\xi(x) = \operatorname*{argmax}_{y'} \left\{ \|y' - \xi g_i(x)\| : y' \in \Delta^k \right\} \quad \forall x \in \mathbb{R}^n.$$

For the validity of the constants m, L_x, and L_y, we first define, for every $1 \le i \le k$, the quantities

$$P_i = \alpha_i C_i^T d_i, \quad Q_i^x := \alpha_i C_i^T C_i x - \beta_i B_i^T D_i^T D_i B_i x \quad \forall x \in \mathbb{R}^n,$$

and observe that $\nabla_x \Phi(x, y) = \sum_{i=1}^k (Q_i^x + P_i) y_i$. Now, using the fact that $y \in \Delta^k$, (I.1), and defining $N_i := \alpha_i C_i^T C_i - \beta_i B_i^T D_i^T D_i B_i$, we then have that

$$\lambda_{\max}(\nabla^2_{xx} \Phi) \le \sum_{i=1}^k y_i \lambda_{\max}(N_i) = \sum_{i=1}^k y_i \lambda_{\max}(\nabla^2 g_i) \le M = L_x,$$
$$\lambda_{\min}(\nabla^2_{xx} \Phi) \ge \sum_{i=1}^k y_i \lambda_{\min}(N_i) = \sum_{i=1}^k y_i \lambda_{\min}(\nabla^2 g_i) \ge -m \ge -L_x,$$

and hence we conclude that the choice of m and L_x in (6.54) is valid. On the other hand, using the fact that $\|x\| \le 1$ for every $x \in \Delta^n$ and (I.1), we then have that for every $y, y' \in Y$,

$$\|\nabla_x \Phi(x,y) - \nabla_x \Phi(x,y')\| = \left\| \sum_{i=1}^{k} (Q_i^x + P_i)(y_i - y_i') \right\|$$
$$\le \left(\sqrt{\sum_{i=1}^{k} M^2 \|x\|^2} + \|P\| \right) \|y - y'\| \le L_y \|y - y'\|,$$

where P is a an n–by–k matrix whose i^{th} column is $\alpha_i C_i^T d_i$, and hence we conclude that the choice of L_y in (6.54) is valid.

Truncated robust regression

Since $Y = \Delta^n$, it is easy to verify that

$$y_\xi(x) = \operatorname*{argmax}_{y'} \{ \|y' - \xi g_i(x)\| : y' \in \Delta^n \} \quad \forall x \in \mathbb{R}^k.$$

For the validity of the constants m, L_x, and L_y, we first define for every $1 \le i \le k$ the function

$$\tau_j(x) := \left[e^{-b_j \langle a_j, x \rangle} \right] \left[1 + e^{-b_j \langle a_j, x \rangle} \right]^{-1} \left[\alpha + \ell_j(x) \right]^{-1} \quad \forall x \in \mathbb{R}^k,$$

and observe that $\nabla_x \Phi(x,y) = -\alpha \sum_{j=1}^{n} \left[y_j b_j \tau_j(x) \right] a_j$ and also that

$$\sup_{x \in \mathbb{R}^k} |\tau_j(x)| \le \alpha^{-1}, \tag{I.2}$$

for every $1 \le j \le n$. Now, using the fact that $y \in \Delta^n$, the bound (I.2), and the Mean Value Theorem applied to τ_j, we have that for every $x, x' \in \mathbb{R}^k$,

$$\begin{aligned}
&\|\nabla_x \Phi(x,y) - \nabla_x \Phi(x',y)\| \le \alpha \sum_{j=1}^{n} y_j \|a_j \left[\tau_j(x) - \tau_j(x')\right]\| \\
&\le \alpha \max_j \left(\|a_j \left[\tau_j(x) - \tau_j(x')\right]\|\right) = \alpha \max_{1\le j\le n} \left[\|a_j\| \cdot |\tau_j(x) - \tau_j(x')|\right] \\
&\le \alpha \max_{1\le j\le n} \left[\|a_j\| \sup_{x\in\mathbb{R}^k} \|\nabla \tau_j(x)\| \|x - x'\|\right] \\
&= \alpha \max_{1\le j\le n} \left[\|a_j\|^2 \sup_{x\in\mathbb{R}^k} \left|\frac{\tau_j(z)}{\alpha + \ell_j(z)}\right|\right] \|x - x'\| \\
&\le \frac{1}{\alpha} \max_{1\le j\le n} \|a_j\|^2 \|x - x'\| = L_x \|x - x'\|,
\end{aligned}$$

and hence we conclude that the choice of $m = L_x$ in (6.55) is valid. On the other hand, using the bound (I.2), we have that for every $y, y' \in \mathbb{R}^n$,

$$\|\nabla_x \Phi(x,y) - \nabla_x \Phi(x,y')\| = \alpha \left\| \sum_{j=1}^{n} b_j \tau_j(x) a_j [y_j - y'_j] \right\| \le L_y \|y - y'\|,$$

and hence we conclude that the choice of L_y in (6.55) is valid.

Power control in the presence of a jammer

For every $1 \le k \le K$ and $1 \le n \le N$, we first define the quantities

$$S^-_{k,n}(X,y) := \sigma^2 + B_{k,n} y_n + \sum_{j=1, j\neq k}^{K} \mathcal{A}_{j,k,n} X_{j,n}, \quad S_{k,n}(X,y) := \mathcal{A}_{k,k,n} X_{k,n} + S^-_{k,n},$$

as well as

$$T_{j,n}(X,y) := \left[S^-_{j,n}(X,y) + S_{j,n}(X,y)\right] / \left[S_{j,n}(X,y) S^-_{j,n}(X,y)\right]^2,$$

for every $(X, y) \in \mathbb{R}^{K\times N} \times \mathbb{R}^N$. Observe now that

$$\frac{\partial \Phi}{\partial y_n}(X, y) = \frac{B_{k,n}}{S_{k,n}(X, y)S^-_{k,n}(X, y)} \quad \forall n \in \{1, ..., N\}. \tag{I.3}$$

The form in (I.3) implies that $\nabla_y \Phi(X, y)$ is a separable function in y where each component is a monotonically decreasing function in its argument. Hence, since $Y = Q^{N\times 1}_{N/2}$, the computation of y_ξ reduces to an N–dimensional bisection search on the functions

$$F_n(y; \xi) = \left[\sum_{k=1}^{K} \frac{B_{k,n}}{S_{k,n}(X, y)S^-_{k,n}(X, y)}\right] - \frac{y_n}{\xi} \quad \forall n \in \{1, ..., N\}.$$

For the validity of the constants $m, L_x,$ and L_y, we first observe that, for every $1 \le k \le K$ and $1 \le n \le N$ and also $(X, y) \in \mathbb{R}^{K\times N} \times \mathbb{R}^N$, we have

$$\frac{\partial \Phi}{\partial X_{k,n}}(X, y) = -\frac{A_{k,k,n}}{S_{k,n}(X, y)} + \sum_{j=1, j\neq k}^{K} \frac{A_{k,j,n}}{S_{j,n}(X, y)S^-_{j,n}(X, y)} \quad \forall (X, y) \in \mathbb{R}^{K\times N} \times \mathbb{R}^N.$$

Using the Mean Value Theorem with respect to $X_{k,n}$ on $\partial\Phi/\partial X_{k,n}$, we have that for every $X, X' \in \mathbb{R}^{K\times N}$,

$$\begin{aligned}
&\left|\frac{\partial}{\partial X_{k,n}} f(X, y) - \frac{\partial}{\partial X_{k,n}} f(X', y)\right| \le \sup_{(X,y)\in\mathbb{R}^{K\times N}\times\mathbb{R}^N} \left|\frac{\partial^2}{\partial X^2_{k,n}} f(X, y)\right| |X_{k,n} - X'_{k,n}| \\
&= \sup_{(X,y)\in\mathbb{R}^{K\times N}\times\mathbb{R}^N} \left|\frac{\mathcal{A}^2_{k,k,n}}{S_{k,n}(X, y)} - \sum_{j=1, j\neq k}^{K} \mathcal{A}^2_{k,j,n} T_{j,n}(X, y)\right| |X_{k,n} - X'_{k,n}| \\
&\le \frac{2\sum_{j=1}^{K} \mathcal{A}^2_{k,j,n}}{\min\{\sigma^4, \sigma^6\}} |X_{k,n} - X'_{k,n}| \le L_x |X_{k,n} - X'_{k,n}|,
\end{aligned}$$

and hence we conclude that the choice of L_x in (6.56) is valid. On the other hand, using the Mean Value Theorem with respect to y_n on $\partial\Phi/\partial X_{k,n}$, we have that for every $y, y' \in \mathbb{R}^{K\times N}$,

$$\begin{aligned}
&\left|\frac{\partial}{\partial X_{k,n}} f(X,y) - \frac{\partial}{\partial X_{k,n}} f(X',y)\right| \leq \sup_{(X,y)\in\mathbb{R}^{K\times N}\times\mathbb{R}^N} \left|\frac{\partial^2}{\partial y_n X_{k,n}} f(X,y)\right| |y_n - y_n'| \\
&= \sup_{(X,y)\in\mathbb{R}^{K\times N}\times\mathbb{R}^N} \left|\frac{B_{k,n}\mathcal{A}_{k,k,n}}{S_{k,n}(X,y)} - \sum_{j=1,j\neq k}^{K} B_{k,n}\mathcal{A}_{k,j,n} T_{j,n}(X,y)\right| |y_n - y_n'| \\
&\leq \frac{2\sum_{j=1}^{K} B_{k,n}\mathcal{A}_{k,j,n}}{\min\{\sigma^4,\sigma^6\}} |y_n - y_n'| \leq L_y |y_n - y_n'|,
\end{aligned}$$

and hence we conclude that the choice of L_y in (6.56) is valid.

REFERENCES

[1] K. J. Arrow et al. *Studies in Linear and Non-Linear Programming*. Vol. 67. 2. Stanford University Press, 1958, p. 229.

[2] H. Attouch and J. Bolte. "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features". In: *Math. Program.* 116.1-2 (2009), pp. 5–16. ISSN: 00255610. DOI: 10.1007/s10107-007-0133-5.

[3] H. Attouch, J. Bolte, and B. F. Svaiter. "Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods". In: *Math. Program.* 137.1-2 (2013), pp. 91–129. ISSN: 14364646. DOI: 10.1007/s10107-011-0484-9.

[4] N. S. Aybat and G. Iyengar. "A first-order smoothed penalty method for compressed sensing". In: *SIAM J. Optim.* 21.1 (2011), pp. 287–313. ISSN: 10526234. DOI: 10.1137/090762294.

[5] N. S. Aybat and G. Iyengar. "A first-order augmented lagrangian method for compressed sensing". In: *SIAM J. Optim.* 22.2 (2012), pp. 429–459. ISSN: 10526234. DOI: 10.1137/100786721. arXiv: 1005.5582.

[6] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory*. 2010. ISBN: 9781441975140.

[7] A. Beck. *First-order methods in optimization*. Society for Industrial and Applied Mathematics, 2017.

[8] A. Beck and M. Teboulle. "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Oper. Res. Lett.* 31.3 (2003), pp. 167–175. ISSN: 01676377. DOI: 10.1016/S0167-6377(02)00231-6.

[9] A. Ben-tal and A. Nemirovski. *Optimization III: Convex Analysis, Nonlinear Programming Theory, Standard Nonlinear Programming Algorithms*. 2020.

[10] A. Ben-Tal and A. S. Nemirovski. "Non-euclidean restricted memory level method for large-scale convex optimization". In: *Math. Program.* 102.3 (2005), pp. 407–456. ISSN: 00255610. DOI: 10.1007/s10107-004-0553-4.

[11] D. P. Bertsekas. *Nonlinear Programming*. Vol. 9. 1999.

[12] E. G. Birgin and J. M. Martínez. "Complexity and performance of an Augmented Lagrangian algorithm". In: *Optim. Methods Softw.* 35.5 (2020), pp. 885–920. ISSN: 10294937. DOI: 10.1080/10556788.2020.1746962. arXiv: 1907.02401.

[13] D. Boob, Q. Deng, and G. Lan. "Stochastic first-order methods for convex and nonconvex functional constrained optimization". In: *arXiv* (2019). ISSN: 23318422. arXiv: 1908.02734.

[14] R. Bracewell. *The Fourier transform and its applications*. McGraw Hill, 2000, p. 616.

[15] J. V. Burke. "Exact penalization viewpoint of constrained optimization". In: *SIAM J. Control Optim.* 29.4 (1991), pp. 968–998. ISSN: 03630129. DOI: 10.1137/0329054.

[16] J. V. Burke and J. J. Moré. "On the Identification of Active Constraints". In: *SIAM J. Numer. Anal.* 25.5 (1988), pp. 1197–1211.

[17] E. J. Candès et al. "Phase retrieval via matrix completion". In: *SIAM J. Imaging Sci.* 6.1 (2013), pp. 199–225. ISSN: 19364954. DOI: 10.1137/110848074.

[18] Y. Carmon et al. "Accelerated methods for nonconvex optimization". In: *SIAM J. Optim.* 28.2 (2018), pp. 1751–1772. ISSN: 10526234. DOI: 10.1137/17M1114296. arXiv: 1611.00756.

[19] Y. Carmon et al. "Lower bounds for finding stationary points I". In: *Math. Program.* 184.1-2 (2020), pp. 71–120. ISSN: 14364646. DOI: 10.1007/s10107-019-01406-y. arXiv: 1710.11606.

[20] Y. Carmon et al. "Lower bounds for finding stationary points II: first-order methods". In: *Math. Program.* 185.1-2 (2021), pp. 315–355. ISSN: 14364646. DOI: 10.1007/s10107-019-01431-x. arXiv: 1711.00841.

[21] A.-L. Cauchy. "Méthode générale pour la résolution des systèmes d'équations simultanées". In: *Compte Rendu des Seances L'Académie des Sci.* 25.2 (1847), pp. 536–538.

[22] E. Chouzenoux, J. C. Pesquet, and A. Repetti. "A block coordinate variable metric forward-backward algorithm". In: *J. Glob. Optim.* 66.3 (2016), pp. 457–485. ISSN: 15732916. DOI: 10.1007/s10898-016-0405-9.

[23] R. Coleman. *Calculus on normed vector spaces*. Springer Science & Business Media, 2012. DOI: 10.1007/978-1-4614-3894-6_1.

[24] P. L. Combettes and J. C. Pesquet. "Proximal splitting methods in signal processing". In: *Springer Optim. Its Appl.* 49 (2011), pp. 185–212. ISSN: 19316836. DOI: 10.1007/978-1-4419-9569-8_10. arXiv: 0912.3522.

[25] J. Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing, 2005, p. 570. ISBN: 0976401304.

[26] D. Drusvyatskiy and C. Paquette. "Efficiency of minimizing compositions of convex functions and smooth maps". In: *Math. Program.* 178.1-2 (2019), pp. 503–558. ISSN: 14364646. DOI: 10.1007/s10107-018-1311-3. arXiv: 1605.00125.

[27] F. Facchinei and J.-S. Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer New York, 2008, p. 693.

[28] H. Fang et al. "Improved bounded matrix completion for large-scale recommender systems". In: *Int. Jt. Conf. Artif. Intell.* 0 (2017), pp. 1654–1660. ISSN: 10450823. DOI: 10.24963/ijcai.2017/229.

[29] P. Frankel, G. Garrigos, and J. Peypouquet. "Splitting Methods with Variable Metric for Kurdyka-Łojasiewicz Functions and General Convergence Rates". In: *J. Optim. Theory Appl.* 165.3 (2015), pp. 874–900. ISSN: 15732878. DOI: 10.1007/s10957-014-0642-3.

[30] S. Ghadimi and G. Lan. "Accelerated gradient methods for nonconvex nonlinear and stochastic programming". In: *Math. Program.* 156.1-2 (2016), pp. 59–99. ISSN: 14364646. DOI: 10.1007/s10107-015-0871-8. arXiv: 1310.3787.

[31] S. Ghadimi, G. Lan, and H. Zhang. "Generalized Uniformly Optimal Methods for Nonlinear Programming". In: *J. Sci. Comput.* 79.3 (2019), pp. 1854–1881. ISSN: 08857474. DOI: 10.1007/s10915-019-00915-4. arXiv: 1508.07384.

[32] G. N. Grapiglia and Y. Yuan. "On The Complexity of An Augmented Lagrangian Method for Nonconvex Optimization". In: *arXiv* (2019). ISSN: 23318422. DOI: 10.1093/imanum/draa021. arXiv: 1906.05622.

[33] K. Greenewald, S. Zhou, and A. Hero. "Tensor graphical Lasso (TeraLasso)". In: *J. R. Stat. Soc. Ser. B (Statistical Methodol.* 81.5 (2019), pp. 901–931. ISSN: 23318422.

[34] Q. Gu, Z. Wang, and H. Liu. "Sparse PCA with oracle property". In: *Adv. Neural Inf. Process. Syst.* 2.January (2014), pp. 1529–1537. ISSN: 10495258.

[35] D. Hajinezhad and M. Hong. "Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization". In: *Math. Program.* 176.1-2 (2019), pp. 207–245. ISSN: 14364646. DOI: 10.1007/s10107-019-01365-4.

[36] Y. He and R. D. C. Monteiro. "Accelerating block-decomposition first-order methods for solving composite saddle-point and two-player Nash equilibrium problems". In: *SIAM J. Optim.* 25.4 (2015), pp. 2182–2211. ISSN: 10526234. DOI: 10.1137/130943649.

[37] Y. He and R. D. C. Monteiro. "An accelerated hpe-type algorithm for a class of composite convex-concave saddle-point problems". In: *SIAM J. Optim.* 26.1 (2016), pp. 29–56. ISSN: 10526234. DOI: 10.1137/14096757X.

[38] M. R. Hestenes. "Multiplier and gradient methods". In: *J. Optim. Theory Appl.* 4 (1969), pp. 303–320.

[39] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Springer Berlin Heidelberg, 1993, p. 348.

[40] M. Hong. "Decomposing Linearly Constrained Nonconvex Problems by a Proximal Primal Dual Approach: Algorithms, Convergence, and Applications". In: (2016). arXiv: 1604.00543.

[41] R. A. Horn and C. R. Johnson. *Matrix Analysis*. 2ed. Cambridge university press, 2012. DOI: 10.1017/cbo9780511810817.

[42] B. Jiang et al. "Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis". In: *Comput. Optim. Appl.* 72.1 (2019), pp. 115–157. ISSN: 15732894. DOI: 10.1007/s10589-018-0034-y. arXiv: 1605.02408.

[43] S. Jiang, K. Li, and R. Y. D. Xu. "Magnitude bounded matrix factorisation for recommender systems". In: *IEEE Trans. Knowl. Data Eng.* (2020). ISSN: 23318422. DOI: 10.1109/tkde.2020.2998218. arXiv: 1807.05515.

[44] K. C. Kiwiel. "Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities". In: *Math. Program.* 69.1-3 (1995), pp. 89–109. ISSN: 00255610. DOI: 10.1007/BF01585554.

[45] O. Kolossoski and R. D. C. Monteiro. "An accelerated non-Euclidean hybrid proximal extragradient-type algorithm for convex-concave saddle-point problems". In: *Optim. Methods Softw.* 32.6 (2017), pp. 1244–1272. ISSN: 10294937. DOI: 10.1080/10556788.2016.1266355.

[46] W. Kong, J. G. Melo, and R. D. C. Monteiro. "Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs". In: *SIAM J. Optim.* 29.4 (2019), pp. 2566–2593. ISSN: 10526234. DOI: 10.1137/18M1171011. arXiv: 1802.03504.

[47] W. Kong, J. G. Melo, and R. D. C. Monteiro. "An efficient adaptive accelerated inexact proximal point method for solving linearly constrained nonconvex composite problems". In: *Comput. Optim. Appl.* 76.2 (2020), pp. 305–346. ISSN: 15732894. DOI: 10.1007/s10589-020-00188-w. arXiv: 1812.06352.

[48] W. Kong, J. G. Melo, and R. D. C. Monteiro. "Iteration-complexity of a proximal augmented Lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints". In: *arXiv* (2020). ISSN: 23318422. arXiv: 2008.07080.

[49] W. Kong and R. D. C. Monteiro. "An accelerated inexact proximal point method for solving nonconvex-concave min-max problems". In: *arXiv* (2019). ISSN: 23318422. arXiv: 1905.13433.

[50] W. Kong and R. D. C. Monteiro. "Accelerated inexact composite gradient methods for nonconvex spectral optimization problems". In: *arXiv* (2020). ISSN: 23318422. arXiv: 2007.11772.

[51] G. Lan and R. D. C. Monteiro. "Iteration-complexity of first-order penalty methods for convex programming". In: *Math. Program.* 138.1-2 (2013), pp. 115–139. ISSN: 00255610. DOI: 10.1007/s10107-012-0588-x.

[52] G. Lan and R. D. C. Monteiro. "Iteration-complexity of first-order augmented Lagrangian methods for convex programming". In: *Math. Program.* 155.1-2 (2016), pp. 511–547. ISSN: 14364646. DOI: 10.1007/s10107-015-0861-x.

[53] A. Lanza et al. "Sparsity-inducing nonconvex nonseparable regularization for convex image processing". In: *SIAM J. Imaging Sci.* 12.2 (2019), pp. 1099–1134. ISSN: 19364954. DOI: 10.1137/18M1199149.

[54] C. Lemaréchal, J. Strodiot, and A. Bihain. "On a Bundle Algorithm for Nonsmooth Optimization". In: *Nonlinear Program. 4* (1981), pp. 245–282. DOI: 10.1016/b978-0-12-468662-5.50015-x.

[55] A. S. Lewis. "The Convex Analysis of Unitarily Invariant Matrix Functions". In: *J. Convex Anal.* 2.1 (1995), pp. 173–183. ISSN: 09446532.

[56] H. Li and Z. Lin. "Accelerated proximal gradient methods for nonconvex programming". In: *Adv. Neural Inf. Process. Syst.* 2015-Janua (2015), pp. 379–387. ISSN: 10495258.

[57] Z. Li and Y. Xu. "Augmented Lagrangian based first-order methods for convex and nonconvex programs: Nonergodic convergence and iteration complexity". In: *arXiv* (2020). ISSN: 23318422. arXiv: 2003.08880.

[58] Z. Li et al. "Rate-improved Inexact Augmented Lagrangian Method for Constrained Nonconvex Optimization". In: *arXiv* (2020). ISSN: 23318422. arXiv: 2007.01284.

[59] J. Liang and R. D. C. Monteiro. "A doubly accelerated inexact proximal point method for nonconvex composite optimization problems". In: *arXiv* (2018). ISSN: 23318422. arXiv: 1811.11378.

[60] J. Liang and R. D. C. Monteiro. "An Average Curvature Accelerated Composite Gradient Method for Nonconvex Smooth Composite Optimization Problems". In: *SIAM J. Optim.* 31.1 (2021), pp. 217–243. ISSN: 23318422. DOI: 10.1137/19m1294277.

[61] J. Liang, R. D. C. Monteiro, and C. K. Sim. "A FISTA-type accelerated gradient algorithm for solving smooth nonconvex composite optimization problems". In: *arXiv* (2019). ISSN: 23318422. arXiv: 1905.07010.

[62] Q. Lin, R. Ma, and Y. Xu. "Inexact proximal-point penalty methods for non-convex optimization with non-convex constraints". In: *arXiv* (2019). ISSN: 23318422. arXiv: 1908.11518.

[63] T. Lin, C. Jin, and M. I. Jordan. "Near-optimal algorithms for minimax optimization". In: *arXiv*. 2020, pp. 2738–2779. arXiv: 2002.02417.

[64] P. L. Loh and M. J. Wainwright. "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima". In: *J. Mach. Learn. Res.* 16 (2015), pp. 559–616. ISSN: 15337928. arXiv: 1305.2436.

[65] S. Lu, I. Tsaknakis, and M. Hong. "Block Alternating Optimization for Non-convex Min-max Problems: Algorithms and Applications in Signal Processing and Communications". In: *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 2019-May (2019), pp. 4754–4758. ISSN: 15206149. DOI: 10.1109/ICASSP.2019.8683795.

[66] Z. Lu and Z. Zhou. "Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming". In: *arXiv* (2018). ISSN: 23318422. arXiv: 1803.09941.

[67] Z.-Q. Luo, J.-S. Pang, and D. Ralph. *Mathematical Programs with Equilibrium Constraints*. 1996. DOI: 10.1017/cbo9780511983658.

[68] R. Mazumder, D. F. Saldana, and H. Weng. "Matrix completion with nonconvex regularization: Spectral operators and scalable algorithms". In: *Stat. Comput.* 30 (2020), pp. 1113–1138. ISSN: 23318422.

[69] J. G. Melo and R. D. C. Monteiro. "Iteration-complexity of an inner accelerated inexact proximal augmented lagrangian method based on the classical lagrangian function and a full lagrange multiplier update". In: *arXiv* (2020), arXiv:2008.00562. ISSN: 23318422. arXiv: 2008.00562.

[70] J. G. Melo, R. D. C. Monteiro, and H. Wang. "Iteration-complexity of an inexact proximal accelerated augmented lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems". In: *arXiv* (2020). ISSN: 23318422. arXiv: 2006.08048.

[71] H. Mine and M. Fukushima. "A minimization method for the sum of a convex function and a continuously differentiable function". In: *J. Optim. Theory Appl.* 33.1 (1981), pp. 9–23. ISSN: 00223239. DOI: 10.1007/BF00935173.

[72] C. Molinari, J. Peypouquet, and F. Roldan. "Alternating forward-backward splitting for linearly constrained optimization problems". In: *Optim. Lett.* 14.5 (2020), pp. 1071–1088. ISSN: 18624480. DOI: 10.1007/s11590-019-01388-y.

[73] R. D. C. Monteiro, C. Ortiz, and B. F. Svaiter. "An adaptive accelerated first-order method for convex optimization". In: *Comput. Optim. Appl.* 64.1 (2016), pp. 31–73. ISSN: 15732894. DOI: 10.1007/s10589-015-9802-0.

[74] R. D. C. Monteiro and B. F. Svaiter. "On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean". In: *SIAM J. Optim.* 20.6 (2010), pp. 2755–2787. ISSN: 10526234. DOI: 10.1137/090753127.

[75] R. D. C. Monteiro and B. F. Svaiter. "Iteration-complexity of a newton proximal extragradient method for monotone variational inequalities and inclusion problems". In: *SIAM J. Optim.* 22.3 (2012), pp. 914–935. ISSN: 10526234. DOI: 10.1137/11083085X.

[76] R. D. C. Monteiro and B. F. Svaiter. "Iteration-complexity of block-decomposition algorithms and the alternating direction method of multipliers". In: *SIAM J. Optim.* 23.1 (2013), pp. 475–507. ISSN: 10526234. DOI: 10.1137/110849468.

[77] K. G. Murty and S. N. Kabadi. "Some NP-complete problems in quadratic and nonlinear programming". In: *Math. Program.* 39.2 (1987), pp. 117–129. ISSN: 14364646. DOI: 10.1007/BF02592948.

[78] I. Necoara, A. Patrascu, and F. Glineur. "Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming". In: *Optim. Methods Softw.* 34.2 (2019), pp. 305–335. ISSN: 10294937. DOI: 10.1080/10556788.2017.1380642. arXiv: 1506.05320.

[79] A. S. Nemirovski. "Efficient methods for large-scale convex problems". In: *Ekon. i Mat. Metod.* 15 (1979).

[80] A. S. Nemirovski. "Efficient iterative algorithms for variational inequalities with monotone operators". In: *Ekon. i Mat. Metod.* 17.2 (1981), pp. 344–359.

[81] A. S. Nemirovski. “Orth-method for smooth convex optimization”. In: *Izv. AN SSSR, Ser. Tekhnicheskaya Kibern.* 2 (1982).

[82] A. S. Nemirovski. “Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems”. In: *SIAM J. Optim.* 15.1 (2005), pp. 229–251. ISSN: 10526234. DOI: 10.1137/S1052623403425629.

[83] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983, p. 388.

[84] Y. Nesterov. “A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$”. In: *Dokl. Akad. Nauk SSSR* 269 (1983), pp. 543–547.

[85] Y. Nesterov. “Smooth minimization of non-smooth functions”. In: *Math. Program.* 103.1 (2005), pp. 127–152. ISSN: 00255610. DOI: 10.1007/s10107-004-0552-5.

[86] Y. Nesterov. “Gradient methods for minimizing composite functions”. In: *Math. Program.* 140.1 (2013), pp. 125–161. ISSN: 00255610. DOI: 10.1007/s10107-012-0629-5.

[87] M. Nouiehed et al. “Solving a class of non-convex min-max games using iterative first order methods”. In: *arXiv*. 2019. arXiv: 1902.08297.

[88] M. O’Neill and S. J. Wright. “Behavior of accelerated gradient methods near critical points of nonconvex functions”. In: *Math. Program.* 176.1 (2019), pp. 403–427. ISSN: 23318422.

[89] D. M. Ostrovskii, A. Lowy, and M. Razaviyayn. “Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems”. In: *arXiv* (2020). ISSN: 23318422. arXiv: 2002.07919.

[90] Y. Ouyang et al. “An accelerated linearized alternating direction method of multipliers”. In: *SIAM J. Imaging Sci.* 8.1 (2015), pp. 644–681. ISSN: 19364954. DOI: 10.1137/14095697X.

[91] C. Paquette et al. “Catalyst acceleration for gradient-based non-convex optimization”. In: *arXiv* (2017). ISSN: 23318422. arXiv: 1703.10993.

[92] N. Parikh and S. Boyd. “Proximal Algorithms”. In: *Found. Trends Optim.* 1.3 (2014), pp. 127–239. DOI: 10.1561/9781601987174.

[93] A. Patrascu, I. Necoara, and Q. Tran-Dinh. “Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization”. In: *Optim. Lett.* 11.3 (2017), pp. 609–626. ISSN: 18624480. DOI: 10.1007/s11590-016-1024-6.

[94] B. T. Polyak. "A general method for solving extremal problems". In: *Sov. Math. Dokl.* 174 (1967), pp. 33–36.

[95] M. J. D. Powell. "A method for nonlinear constraints in minimization problems". In: *Optimization* (1969), pp. 283–298.

[96] H. Rafique et al. "Non-convex min-max optimization: Provable algorithms and applications in machine learning". In: *arXiv* (2018). ISSN: 23318422. arXiv: 1810.02060.

[97] R. T. Rockafellar. "Monotone Operators and the Proximal Point Algorithm." In: *SIAM J. Control Optim.* 14.5 (1976), pp. 877–898. ISSN: 03630129. DOI: 10.1137/0314056.

[98] T. R. Rockafellar. *Convex analysis*. Princeton University Press, 1997.

[99] T. R. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Science & Business Media, 2009.

[100] A. P. Ruszczyński. *Nonlinear optimization*. Princeton University Press, 2011.

[101] M. F. Sahin et al. "An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints". In: *arXiv* (2019). ISSN: 23318422. arXiv: 1906.11357.

[102] N. Z. Shor. "Generalized gradient descent with application to block programming". In: *Kibernetika* 3 (1967), pp. 53–55.

[103] M. Sion. "On general minimax theorems". In: *Pacific J. Math.* 8.1 (1958), pp. 171–176. ISSN: 00308730. DOI: 10.2140/pjm.1958.8.171.

[104] M. V. Solodov and B. F. Svaiter. "A unified framework for some inexact proximal point algorithms". In: *Numer. Funct. Anal. Optim.* 22.7-8 (2001), pp. 1013–1035. ISSN: 01630563. DOI: 10.1081/NFA-100108320.

[105] T. Sun and C. H. Zhang. "Calibrated elastic regularization in matrix completion". In: *Adv. Neural Inf. Process. Syst.* 2 (2012), pp. 863–871. ISSN: 10495258. arXiv: 1211.2264.

[106] K. K. Thekumparampil et al. "Efficient algorithms for smooth minimax optimization". In: *arXiv* (2019), arXiv:1907.01543. ISSN: 23318422. arXiv: 1907.01543.

[107] A. Tsoukalas, P. Parpas, and B. Rustem. "A smoothing algorithm for finite min-max-min problems". In: *Optim. Lett.* 3.1 (2009), pp. 49–62. ISSN: 18624472. DOI: 10.1007/s11590-008-0090-9.

[108] F. Wen et al. “Robust PCA Using Generalized Nonconvex Regularization”. In: *IEEE Trans. Circuits Syst. Video Technol.* 30.6 (2020), pp. 1497–1510. ISSN: 15582205. DOI: 10.1109/TCSVT.2019.2908833.

[109] R. Williamson and H. Trotter. *Multivariable Mathematics*. Pearson Education, Inc, 2004.

[110] Y. Xie and S. J. Wright. “Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints”. In: *arXiv* (2019). ISSN: 23318422. DOI: 10.1007/s10915-021-01409-y. arXiv: 1908.00131.

[111] Y. Xu. “Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming”. In: *Math. Program.* 185 (2019), pp. 199–244. ISSN: 23318422.

[112] Q. Yao and J. T. Kwok. “Efficient learning with a family of nonconvex regularizers by redistributing nonconvexity”. In: *J. Mach. Learn. Res.* 18 (2018), pp. 1–52. ISSN: 15337928. arXiv: 1606.03841.

[113] I. Zang. “A smoothing-out technique for min-max optimization”. In: *Math. Program.* 19.1 (1980), pp. 61–77. ISSN: 00255610. DOI: 10.1007/BF01581628.

[114] J. Zhang and Z. Q. Luo. “A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization”. In: *arXiv* (2020). ISSN: 23318422.

[115] J. Zhang and Z. Q. Luo. “A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization”. In: *SIAM J. Optim.* 30.3 (2020), pp. 2272–2302. ISSN: 23318422.

[116] D. Zhou and Q. Gu. “Lower bounds for smooth nonconvex finite-sum optimization”. In: *Int. Conf. Mach. Learn.* 2019, pp. 7574–7583.

VITA

Weiwei "William" Kong is a Canadian national born on July 16, 1992, in Guangzhou, China. His family immigrated to Canada in the Summer of 1999 to the suburban district of Scarborough in Toronto, Ontario. In December 2014, he earned a B. Math. degree from the University of Waterloo with distinction. After obtaining his undergraduate degree, he then spent two years as a statistical modeler in Canada. In August 2016, he began his Ph.D. studies in Operations Research at Georgia Tech, where he received the Thomas Johnson Fellowship. His research has been funded in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Institute for Data Engineering and Science (IDEaS), and Transdisciplinary Research Institute for Advancing Data Science (TRIAD). During his doctoral studies, he obtained an M.Sc. in Computational Science and Engineering at Georgia Tech in May 2019 and two internships at Google Research as a Research and Software Engineering Intern.