ACCELERATED FIRST-ORDER OPTIMIZATION METHODS USING INERTIA AND ERROR BOUNDS

BY

PATRICK ROYCE JOHNSTONE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Pierre Moulin, Chair
Professor Yoram Bresler
Assistant Professor Niao He
Associate Professor Angelia Nedich
Professor Rayadurgam Srikant

ABSTRACT

Optimization is an important discipline of applied mathematics with far-reaching applications. Optimization algorithms often form the backbone of practical systems in machine learning, image processing, signal processing, computer vision, data analysis, and statistics. In an age of massive data sets and huge numbers of variables, a deep understanding of optimization is a necessary condition for developing scalable, computationally inexpensive, and reliable algorithms.

In this thesis we design and analyze efficient algorithms for solving the large-scale nonsmooth optimization problems arising in modern signal processing and machine learning applications. The focus is on first-order methods which have low per-iteration complexity and can exploit problem structure to a high degree. First-order methods have the capacity to address large-scale problems for which all alternative methods fail. However, first-order methods can take many iterations to reach the desired accuracy. This has led optimization researchers to ask the following question: is it possible to improve the convergence rate of first-order methods without jeopardizing their low per-iteration complexity?

In this thesis, we address this question in three areas. Firstly we investigate the use of inertia to accelerate the convergence of proximal gradient methods for convex composite optimization problems. We pay special attention to the famous lasso problem for which we develop an improved version of the well-known Fast Iterative Soft-Thresholding Algorithm. Secondly we investigate the use of inertia for nonconvex composite problems, making use of the Kurdukya-Łojaziewicz inequality in our analysis. Finally, when the objective function satisfies an error bound which is fairly common in practice, we develop stepsize selections for the subgradient method which significantly outperform the classical approach.

The overarching message of this thesis is the following: with careful analysis and design, the convergence rate of first-order methods can be significantly improved.

For Mum and Dad.

ACKNOWLEDGMENTS

I thank my advisor, Prof. Pierre Moulin, for his time and patience. I have learned much from him, but perhaps most importantly I hope to have inherited his rigorous and precise approach to thinking and problem-solving. I am very grateful that he allowed me the space and time to explore my own research interests and to find a topic I find exciting. I thank Prof. Niao He for many interesting and enlightening conversations. Her guidance and advice were most important to me. I thank my office mates throughout the years: Scott, Honghai, Ben, Ethan, Amish, and Furen. Many great discussions were had both in Beckman and later in CSL. Throughout the last few years we had a great group of regular lunch attendees at Beckman cafe with Daphne, Brian, Tom, Jon, Matt, Pooya, and Rob, as well as many others. Those lunches will be missed. Outside of engineering I had a great group of friends who kept me going, especially in the first few years in Champaign: Juan, Neha, Derek, Isaac, Dushyant, and Katherine Curran. And thank you to Katherine Londoño for putting up with me.

Most importantly, this thesis is dedicated to my Mum, Dad, and sister, and no words about them are necessary.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

The purpose of this thesis is to develop and understand algorithms that solve *mathematical optimization problems*. While optimization arises everywhere in engineering and science, we will focus on problems emerging in signal processing and machine learning. In modern times in these areas there has been a trend towards larger problem sizes, which come with unprecedented challenges. Hence the focus of this thesis will be *large-scale* optimization. Modern optimization problems in signal processing and machine learning are so large that only specialized algorithms which utilize the problem's unique structure are feasible. In contrast, *black-box* approaches often fail. The past two decades have seen a considerable amount of research devoted to developing algorithms which exploit problem structure.

During the 1980s and 90s *interior point methods* became popular optimization solvers. However over the last two decades as problem sizes have increased dramatically, these methods have failed to keep up. An important group of alternative methods is actually older than the interior point methods but has gone through a renaissance over the past one or two decades. This is the family of *first-order methods* which earn their name by only extracting (sub)gradients rather than Hessian information from the objective function. These methods succeed by using cheap and scalable computations at each iteration. In contrast with interior point methods, these computations do not include solving large systems of linear equations.

The main drawback of first-order methods is slow convergence rate, meaning that a large number of iterations is required for a moderate to high accuracy solution. This has led to a significant thrust of research in the optimization, machine learning, and signal processing communities aimed at *accelerating* first-order methods without jeopardizing their attractive features. While these acceleration techniques come in all different shapes and sizes, a common thread is the need to take into account detailed problem structure when designing the algorithm. In this thesis, we utilize two types of structure: *composite optimization* and *error bounds*.

Composite optimization in this thesis refers to problems with an additive decomposition into a smooth part and a simple nonsmooth part. This form of objective is ubiquitous in machine learning and signal processing. In signal processing it occurs in compressed sensing and inverse problems such as image deblurring. The smooth term encapsulates the measurement process and the nonsmooth term encapsulates prior information on the object one wishes to reconstruct, such as sparsity in a known basis. In machine learning it occurs in regularized empirical risk minimization where the regularizer is typically a simple nonsmooth function and the empirical risk is typically a smooth function.

In optimization an error bound is an upper bound on the distance of a point to the optimal set by some computable residual function. When an objective function satisfies an error bound, it usually allows for a more precise understanding of the convergence rate of first-order methods. While the study of error bounds goes back to the origin of first-order methods in the 1960s, there has been much renewed interest in the topic recently, with applications to problems arising in machine learning and signal processing. A related concept is the Kurdukya-Łojaziewicz (KL) inequality, which measures the "sharpness" of a function around local minimizers.

1.2 Contributions of the Thesis

The thesis focuses on three major areas which are broken up into Chapters 2, 3, and 4. Section 1.3 provides information on notation and some mathematical background relevant to the entire thesis. Each chapter also discusses notation and the mathematical background specific to that chapter.

In Chapter 2 we consider convex composite optimization problems. The proximal gradient algorithm is an important first-order approach to solving this type of problem. Our first contribution is to show global convergence of an *inertial* variant of the proximal gradient method. Inertia is an acceleration technique for solving quadratic optimization problems and monotone inclusions. Our second contribution in this chapter is to do with the *lasso problem*, a hugely important instance of convex composite optimization. We conduct a local convergence analysis for the inertial proximal gradient method applied to lasso. This result allows us to develop an improved version of the well-known Fast Iterative Soft Thresholding Algorithm (FISTA).

In particular we fix an undesirable local convergence property of FISTA which arises on the lasso problem.

In Chapter 3 we again consider composite optimization problems, but this time we abandon the assumption of convexity. Instead we assume the function satisfies the KL inequality which is common in practice. In fact when the function is semialgebraic it satisfies the KL inequality. The main contribution of this chapter is to determine for the first time the convergence rate of a broad family of inertial proximal gradient methods for solving nonconvex composite problems. The family of methods we study includes several algorithms proposed in the literature for which convergence rates are unknown.

In Chapter 4 we again consider convex optimization but this time under an error bound condition. We study the *subgradient method*, which is a classical approach to nonsmooth optimization going back to the 1970s. Conventional wisdom in optimization says that the subgradient method is slow, simple, intuitive, easy to implement, and scalable. In this chapter, we utilize the error bound condition to address the first element of conventional wisdom. We devise stepsizes which outperform the classical choice and can even obtain a *linear* convergence rate. Linearly convergent subgradient methods under an error bound are not new and were first devised in the 1970s. However our method has the advantage of being able to estimate on-the-fly an unknown error bound parameter.

1.3 Mathematical Background

1.3.1 Notation

For the most part the notation and conventions follow [1]. Thus $\mathcal{H}$ is always a Hilbert space over the reals, $\langle\cdot,\cdot\rangle$ is the inner product and $\|\cdot\|$ is the induced norm. The notation $\mathbb{R}^n$ means the n-dimensional Euclidean Hilbert space. For $\mathbb{R}^n$ we assume the standard Euclidean norm and inner product and use $\|\cdot\|_1$ to denote the ℓ_1-norm. The notation $\mathbb{R}_+$ denotes the set of all nonegative real numbers.

A function is closed if it has a closed epigraph and proper if it has a nonempty domain. Let $\Gamma_0(\mathcal{H})$ be the set of all closed, convex and proper functions from $\mathcal{H}$ to $(-\infty,\infty]$. We will also refer to these functions by saying they are CCP (convex, closed, and proper). For any $g:\mathcal{H}\to(-\infty,\infty]$ and point $x\in\mathcal{H}$, we denote by $\partial g(x)$ the *subdifferential* at x [1, Def. 16.1] defined as the set

$$\partial g(x) \triangleq \{v \in \mathcal{H} : g(y) \geq g(x) + \langle v, y - x\rangle, \forall y \in \mathcal{H}\}.$$

The notation dom $\partial g \subset \mathcal{H}$ represents the set of x such that $\partial g(x)$ is nonempty. If g is CCP, dom(∂g) is a dense subset of dom(f) [1, Cor. 16.29]. When $\partial g(x)$ is a singleton we will call it the (Gâteaux) *gradient* at x, denoted by $\nabla g(x)$.

For $a : \mathbb{R} \to \mathbb{R}$, $b : \mathbb{R} \to \mathbb{R}$, and $c \in [-\infty, +\infty]$, the notation $a(l) = O(b(l))$ (resp. $a(l) = \Omega(b(l))$) means there exists a constant $C \geq 0$ such that $\limsup_{l\to c} |a(l)/b(l)| \leq C$ (resp. $\liminf_{l\to c} |a(l)/b(l)| \geq C$). We will say a sequence $\{x^k\}_{k\in\mathbb{N}} \subset \mathcal{H}$ converges *R-linearly* to $x^* \in \mathcal{H}$ with rate of convergence $q \in (0, 1)$, if $\|x^k - x^*\| = O(q^k)$. We say x^k converges to x^* *Q-linearly* with rate $q \in (0, 1)$ if $\lim_{k\to\infty} \left\{\|x^k - x^*\|/\|x^{k-1} - x^*\|\right\} = q$. Collectively we refer to both Q-linear and R-linear convergence simply as linear convergence. We use $x^k \to x^*$ to denote strong convergence and $x^k \rightharpoonup x^*$ to denote weak convergence.

Given a closed set C and point x, define $d(x, C) \triangleq \min\{\|x - c\| : c \in C\}$. If C is also convex, then there is a unique point, which we denote by $P_C(x)$, such that $\|x - P_C(x)\| = d(x, C)$. If C is a linear subspace of the Hilbert space $\mathcal{H}$, then P_C is a linear operator. The projection satisfies the following nonexpansiveness property: for all $x, y \in \mathcal{H}$, $P_C(x) - P_C(y)\| \leq \|x - y\|$ [1].

For a vector $v \in \mathbb{R}^n$, v^i is the ith element of v for $i = 1, 2, \ldots, n$. Subscripts are used for iteration number, as in x_k.

Some variable names are reused across chapters. For example, we study several different algorithms which produce a sequence of iterates. We will always use $\{x_k\}$ to denote the output of an algorithm and it will always be clear from context and the chapter to which algorithm the iterates belong.

1.3.2 Properties of Convex and Smooth Functions

Now we list some properties of the subdifferential, as well as convex and smooth functions. For the Fréchet and Gâteaux definitions of differentiability we refer to [1, Definition 2.45 and 2.43]. Note that Fréchet differentiability on a neighborhood of a point implies Gâteaux differentiability at that point, and the two derivatives agree [1, Lemma 2.49(i)]. For a Hilbert space $\mathcal{H}$, consider a function $f : \mathcal{H} \to (-\infty, +\infty]$. Then

$$\langle t, u - v\rangle \quad \geq \quad f(u) - f(v), \quad \forall v \in \mathcal{H}, u \in \operatorname{dom} \partial f, \text{ and } t \in \partial f(u), \quad (1.1)$$

and

$$\langle t-p, u-v\rangle \geq 0, \quad \forall u,v \in \operatorname{dom}\partial f,\ t\in\partial f(u) \text{ and } p \in \partial f(v). \tag{1.2}$$

For a proper and convex function which is Gâteaux differentiable everywhere on $\mathcal{H}$, (1.1)–(1.2) hold for all $u,v\in\mathcal{H}$ [1, Prop. 17.10] and $\partial f(x) = \{\nabla f(x)\}$ (i.e. a singleton) everywhere.

We say that a Fréchet differentiable function f has L-Lipschitz continuous gradient if $\|\nabla f(y) - \nabla f(x)\| \leq L\|y-x\|,\ \forall x,y\in\mathcal{H}$. For such a function [1, Thm. 18.15 (iii)]:

$$f(u) - f(v) \quad \leq \quad \langle \nabla f(v), u-v\rangle + \frac{L}{2}\|u-v\|^2, \quad \forall u,v\in\mathcal{H}. \tag{1.3}$$

The gradient ∇f of a convex and Fréchet differentiable function is L-Lipschitz continuous if and only if [1, Cor. 18.16]

$$\langle \nabla f(u) - \nabla f(v), u-v\rangle \geq \frac{1}{L}\|\nabla f(u) - \nabla f(v)\|^2, \quad \forall u,v\in\mathcal{H}. \tag{1.4}$$

This is the celebrated Baillon-Haddad Theorem.

1.3.3 Proximal Operators

The *proximal operator* $\operatorname{prox}_g : \mathcal{H}\to\mathcal{H}$ with respect to a function $g\in\Gamma_0(\mathcal{H})$ is defined implicitly as:

$$y - \operatorname{prox}_g(y) \in \partial g(\operatorname{prox}_g(y)), \tag{1.5}$$

and explicitly as

$$\operatorname{prox}_g(y) = \arg\min_{x\in\mathcal{H}}\left\{\frac{1}{2}\|x-y\|^2 + g(x)\right\}, \quad \forall y\in\mathcal{H}.$$

The proximal operator is a well-defined mapping from $\mathcal{H}$ to $\operatorname{dom}\partial g$ [1, Prop. 23.2, Example 23.3].

CHAPTER 2

AN INERTIAL METHOD FOR CONVEX COMPOSITE PROBLEMS

2.1 Chapter Introduction

The primary problem considered in this chapter is to

$$\underset{x\in\mathcal{H}}{\text{minimize}}\ F(x) = f(x) + g(x) \tag{2.1}$$

where $\mathcal{H}$ is a Hilbert space over the real numbers, the functions $f, g : \mathcal{H} \to (-\infty, +\infty]$ are proper, convex and closed, and in addition f is differentiable everywhere and has a Lipschitz continuous gradient. This problem has come under considerable attention in recent years due to its many applications in areas such as machine learning, compressed sensing and image processing [2, 3, 4, 5, 6, 7, 8, 9, 10]. Of particular interest in this chapter will be the special case where the nonsmooth term is the ℓ_1-norm, i.e.

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \{f(x) + \rho\|x\|_1\} \tag{2.2}$$

where $\rho > 0$, and $\|x\|_1 = \sum_{i=1}^n |x_i|$. As has been widely recognized the ℓ_1-norm encourages "sparse" solutions, i.e. solutions with few nonzero elements, which is its primary attraction [2, 7]. A special case of Prob. (2.2) is

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \left\{\frac{1}{2}\|b - Ax\|_2^2 + \rho\|x\|_1\right\} \tag{2.3}$$

with $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, which is often referred to as sparse least-squares, sparse regression, basis pursuit, or lasso and is of vital importance in many areas [11, 3, 7, 4]. Other important instances of Prob. (2.1) include least-squares with a total-variation [12], nuclear norm [13], or group-sparse [2] regularizer, and minimization of a convex function constrained to a closed and convex set.

2.1.1 Background

The increasing size of Problems (2.1)–(2.3) in modern applications is driving the need for computationally inexpensive and scalable algorithms to find their solutions. In modern applications the number of variables and the number of data can be in the millions [11, 2]. Proximal gradient methods for solving optimization problems including (2.1) are simple and computationally inexpensive, and address the problem by splitting it into simpler subproblems. Hence proximal gradient methods are an example of a *splitting method*. While the overall objective F in Prob. (2.1) may not have desirable properties, each component of the sum can be handled. The function f is smooth which means it can be processed via its gradient, and many popular nonsmooth regularizers can be processed via a computationally tractable proximal operator [5]. Importantly, first-order methods do not rely on or approximate second-order information, which may be prohibitively expensive in high dimensions. The concept of splitting has also been applied to more complicated objectives [14, 15, 16]. These techniques can also be viewed in the broader context of montone inclusion problems and variational inequalities which includes convex optimization as a special case [14, 15, 17, 18, 1, 19].

The celebrated first-order splitting method for Prob. (2.1) is the *proximal forward-backward splitting algorithm* (FBS) [17, 20]. This is also known simply as the proximal gradient method. For this method the convergence rate of the objective function to the optimal value is as good as if the nonsmooth component were not present. Weak convergence of the iterates is also guaranteed and linear convergence occurs on strongly convex problems [1, Cor. 27.9, Ex. 27.12]. Line search techniques allow for when the gradient is not Lipschitz continuous or the Lipschitz constant is unknown [4, 21]. For the special case of Problems (2.2)–(2.3) it is often referred to as the iterative shrinkage and soft-thresholding algorithm (ISTA) due to the form of the proximal operator with respect to the ℓ_1-norm. Other state-of-the-art approaches to solving Prob. (2.1) and Problems (2.2)–(2.3) in particular include coordinate descent [22], ADMM [8], and stochastic methods [23].

2.1.2 Inertial Methods

A class of methods of particular interest in this chapter (and Chapter 3) are *inertial methods* (a.k.a. momentum methods). These are iterative schemes for solving monotone inclusion and optimization problems, as well as computing fixed points, which often have connections to systems of differential equations (e.g. [24, 25, 26, 27, 28]). Their defining property is that the next iterate depends on more than one previous iterate (i.e. they are multistep). A very early example is due to Polyak [28], who introduced the heavy ball with friction method for minimizing a strongly convex quadratic function which can greatly improve upon the convergence speed of the simple gradient method (see also [27, p. 65]). The conjugate gradient method is inertial, as are Nesterov's celebrated accelerated methods, and their variants and extensions [29, 4, 30, 31]. Inertial methods typically have the same per-iteration complexity as their noninertial counterparts. However in certain contexts they can be significantly faster [27, 28, 32, 4].

2.1.3 Chapter Contributions

In this chapter we consider the following *Inertial Forward-Backward Splitting Algorithm* (I-FBS):

$$\begin{aligned} y_{k+1} &= x_k + \zeta_k(x_k - x_{k-1}), && (2.4)\\ x_{k+1} &= \operatorname{prox}_{\lambda_k g}\left(y_{k+1} - \lambda_k \nabla f(y_{k+1})\right) && (2.5) \end{aligned}$$

with $x_0, x_1 \in \mathcal{H}$. The sequences $\{\zeta_k, \lambda_k\}_{k\in\mathbb{N}}$ are in $\mathbb{R}_+$. FBS is recovered when $\zeta_k = 0$. I-FBS is related to FISTA introduced in [4], which is itself related to earlier accelerated methods [29, 32, 30]. FISTA corresponds to a particular choice for the inertia sequence $\{\zeta_k\}_{k\in\mathbb{N}}$ in I-FBS. The goal of our global convergence analysis is different from that of the literature on FISTA in that we are concerned with deriving general conditions on ζ_k which imply convergence of the iterates. For example the choice $\zeta_k = 0.5$ for all $k \in \mathbb{N}$ is not explicitly covered by the FISTA literature but is covered by our analysis. To clarify notation, we will use "I-FBS" to refer to all parameter choices satisfying our convergence criteria given in Corollary 4, and "FISTA" to refer to the parameter choices which guarantee an $O(1/k^2)$ objective function rate (for example [4, 32, 12]). We note that our local analysis for ℓ_1-regularized problems applies to both I-FBS and certain variants of FISTA. For these problems we characterize the local performance of FISTA and provide ways to improve it.

The well-known property of FISTA is the "fast" $O(1/k^2)$ objective function convergence rate for Prob. (2.1). It is important to note that we do not expect this global objective function behavior to hold for I-FBS. Nevertheless the goal of this chapter is not to study objective function convergence rates, but convergence of the iterates $\{x_k\}_{k\in\mathbb{N}}$, which is also important in practice [5, p. 5]. When we do compute convergence rates in the local analysis, they are asymptotically linear rates applicable to the iterates, i.e. $\|x_k - x_*\| \leq Cq^k$ for sufficiently large k, where x_* is an optimal solution, and $q \in (0,1)$. One of the main findings of our local analysis for Prob. (2.2) is that despite the optimal global sublinear convergence rate of FISTA, its local convergence performance can be greatly improved. This is important for applications where a high accuracy solution is needed, such as medical imaging [9, 10].

For the sake of generality our global analysis applies to the following scheme which we call the *Generalized Inertial Proximal Splitting Algorithm* (GIPSA). For all $k \in \mathbb{N}$ compute:

$$y_{k+1} = x_k + \beta_k(x_k - x_{k-1}), \tag{2.6}$$

$$z_{k+1} = x_k + \zeta_k(x_k - x_{k-1}), \tag{2.7}$$

$$x_{k+1} = \operatorname{prox}_{\lambda_k g}\left(y_{k+1} - \lambda_k \nabla f(z_{k+1})\right). \tag{2.8}$$

Throughout the chapter we will refer to $\{\zeta_k, \beta_k\}_{k\in\mathbb{N}}$ as the “inertia parameters” and $\{\lambda_k\}_{k\in\mathbb{N}}$ as the “stepsize”. Note that I-FBS is recovered when $\zeta_k = \beta_k$. The main motivation for studying the more general (2.6)–(2.8) is that it unifies several existing schemes which correspond to particular parameter choices [24, 3, 18, 33, 34, 12, 35]. Thus our global convergence analysis of GIPSA unifies and extends the prior art. Certain special cases of GIPSA (e.g. [18, 33]) solve the more general maximal monotone inclusion problem:

$$\text{Find } x \quad \text{s.t. } 0 \in \mathcal{A}(x) + \mathcal{B}(x) \tag{2.9}$$

where $\mathcal{A}$ and $\mathcal{B}$ are maximal monotone and $\mathcal{B}$ is cocoercive.[1] Other special cases were introduced as inertial versions of the Krasnosel’skiĭ-Mann (KM) iterations for finding fixed points [34, 36]. In this chapter we focus on convex optimization, which allows us to obtain less stringent convergence criteria than in those previous studies because we can use properties unique to convex functions. We note that GIPSA was originally suggested in [37], however our convergence conditions are more general. GIPSA is also related (via discretization) to the continuous ODEs studied in [38, 26].

We apply our global analysis to GIPSA rather than the simpler I-FBS in order to unify several previous results under one analysis, and to “fill the

[1] Setting $\mathcal{A} = \partial g$ and $\mathcal{B} = \nabla f$ recovers Prob. (2.1).

gaps" between them. For example [18, 33, 34, 24] correspond to special parameter choices of GIPSA. However note that our primary practical concern is I-FBS, for which our proposed adaptive restart method for Prob. (2.2) outperforms the existing FISTA-type methods.

Our main contributions in this chapter can be summarized as:

1. A global convergence analysis of GIPSA.
2. A local convergence analysis of I-FBS and FISTA for ℓ_1-regularized problems.
3. An adaptive restart modification of FISTA with improved local convergence properties for ℓ_1-regularized problems.

We now explain each contribution in more detail.

Global Analysis

In our global analysis we establish conditions on $\{\zeta_k, \beta_k, \lambda_k\}_{k\in\mathbb{N}}$ that imply the global weak convergence of the iterates $\{x_k, y_k, z_k\}_{k\in\mathbb{N}}$ of GIPSA to a solution of Prob. (2.1). No theoretical convergence study of (2.6)–(2.8) specialized to convex optimization exists. Special cases of GIPSA corresponding to different parameter choices have been studied previously in [18, 33, 34]. However these analyses were not specialized to Prob. (2.1) and therefore impose stricter conditions on the stepsize and inertia parameter than developed here.

Our global analysis builds on the investigation of the inertial proximal algorithm of [24]. This algorithm corresponds to GIPSA when the smooth function f is not present. Essentially our global analysis extends [24, Theorem 3.1] to the composite case. We show that a multistep Lyapunov energy function is nonincreasing and this allows us to establish finiteness of the sum of the squared increments, i.e. $\sum_{k\in\mathbb{N}} \|x_k - x_{k-1}\|^2 < \infty$. This condition is also needed for the local analysis. Weak convergence then follows via Opial techniques adapted from [33].

Local Analysis

The forward-backward nature of I-FBS makes it amenable to a local analysis for Problems (2.2)–(2.3). It has been observed that FBS obtains *local linear convergence* for Prob. (2.2) and others [3, 39, 13, 40, 41]. This means that after finitely many iterations, the iterates are permanently confined to a manifold containing the solution with respect to which the objective function is smooth. Thus after a finite time period, convergence to a solution is linear, so long as the local part of the function is also strongly convex, or a strict complementarity condition holds [13, 3]. For Prob. (2.2) the objective function is smooth with respect to vectors of fixed sign and support.

We extend these results to I-FBS and FISTA. We show that I-FBS achieves local linear convergence and we determine the convergence rate in terms of the local curvature, the stepsize, and the inertia parameter. Importantly our analysis shows that adding the correct inertia term allows for a far better asymptotic convergence rate than is achievable with FBS (or FISTA). The local analysis borrows from the framework developed in [3], however extensive differences emerge in order to incorporate the inertia term.

We note that our local analysis results and techniques differ from what was presented in [42], which used a spectral analysis to study the local behavior of FBS and FISTA applied to Prob. (2.3). In contrast our analysis is based around exploiting the contractive properties of the *soft-thresholding operator*, which is the proximal operator with respect to the ℓ_1-norm. The authors of [42] claim that both algorithms obtain local linear convergence when the minimizer is unique and a strict complementarity condition holds. Some of our results require neither of these conditions (Thms. 5 and 6) while others depend on either strict complementarity (Thm. 9) or solution uniqueness (Cor. 7). Unlike [42], we can compute Q-linear and R-linear convergence rates and this allows us to determine the optimal value for the inertia parameter. Many of our results also hold for the more general Prob. (2.2). Our local analysis is also related to [43] and we discuss this relationship in more detail in Sec. 2.3.4.

We note that it is possible to derive upper bounds on the number of iterations not confined to the optimal smooth manifold within our analysis framework. To the best of our knowledge this is not possible in the competing frameworks [42, 43]. In some situations these upper bounds might be useful, however in general they appear to be overly pessimistic compared to what is observed in practice.

Adaptive Restart

Recently Chambolle and Dossal studied a variant parameter choice of FISTA which we will call FISTA-CD [12]. FISTA-CD was also studied in [38, 44]. This variant has some stronger properties than the original version of FISTA due to Beck and Teboulle [4]. In this chapter we use these strong properties to establish the local convergence behavior of FISTA-CD for Problems (2.2) and (2.3). We prove that FISTA-CD, exactly like I-FBS, obtains finite manifold identification for these problems. Furthermore, we show that after finitely many iterations FISTA-CD reduces to the form of a linear iterative system that has been studied previously in [45], allowing us to determine the asymptotic linear convergence rate. This rate is worse than that of the best choice for the inertia parameter in I-FBS and is comparable with the rate of (non-inertial) FBS. We then propose an adaptive restart for FISTA-CD which obtains the optimal[2] asymptotic convergence rate. Furthermore the restart scheme does not require knowledge of the local curvature parameter. Also important is that our proposed restart scheme preserves the optimal global convergence rate of FISTA-CD while also obtaining the optimal local convergence rate.

We note that restart techniques have been proposed before for accelerated methods, as well as conjugate gradient schemes, but only in the context of smooth and strongly convex problems [45, 26, 46], [47, p. 140]. It has been conjectured that restarting could improve the performance of FISTA even in the presence of nonsmooth regularizers [45, §5.2],[6, p. 36]. Our contribution is to show that this is indeed true for the case of the ℓ_1-norm and to derive explicit convergence rates.

2.1.4 Chapter Organization

The rest of the chapter is organized as follows. In Section 2.2, notation, definitions, assumptions and some preparatory results are presented. In Sections 2.3.1 through 2.3.3 we detail the conclusions of our global analysis of GIPSA for Prob. (2.1). In Sections 2.3.4 through 2.3.8 we give the results of our local convergence analysis of I-FBS and FISTA-CD for Problems (2.2)–(2.3). In Sec. 2.4, a small synthetic numerical experiment on Prob. (2.3) is presented in order to corroborate some of our theoretical findings. Finally the proofs of all our results are given in Sections 2.6 through 2.9.

[2] among first-order methods

2.2 Preliminaries

2.2.1 Chapter Specific Notation

For Prob. (2.1) define the *optimal value* as $F_* \triangleq \inf_{x\in\mathcal{H}} F(x)$ and the *solution set* as $\mathcal{X}_F \triangleq \{x \in \mathcal{H} : F(x) = F_*\}$ which may be empty. For the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by (2.6)–(2.8), let Δ_k denote $x_k - x_{k-1}$ for all $k \in \mathbb{N}$. Given a function $a : \mathbb{R} \to \mathbb{R}$, we say that the *iteration complexity* of a method for minimizing F is $\Omega\left(a(\epsilon)\right)$ if $k = \Omega\left(a\left(\epsilon\right)\right)$ implies $F(x_k) - F_* = O(\epsilon)$ as $\epsilon \to 0$.

For a sorted set $S \subseteq \{1, 2, \ldots, n\}$ with no repeated elements, let $S(i), i = 1, \ldots, |S|$ be the ith element of S, where $|S|$ is the number of elements in S. For a matrix $A \in \mathbb{R}^{m\times n}$, A_S will denote the matrix in $\mathbb{R}^{m\times|S|}$ formed by taking the columns corresponding to the elements of S. That is $A_S(i,j) = A(i, S(j))$. For a vector $v \in \mathbb{R}^n$, v^S will denote the $|S|\times 1$ vector with entries given by $v_S(i) = v(S(i))$. The notation $(v^S, 0)$ will denote the vector in $\mathbb{R}^n$ whose jth entry is $v(j)$ if $j \in S$ and 0 otherwise. The range space and null space of a matrix A are denoted by $\mathcal{R}(A)$ and $\mathcal{N}(A)$ respectively. Given $c \in \mathbb{R}$ and $x \in \mathbb{R}^n$, $\text{sgn}(c)$ is defined as $+1$ if $c \geq 0$ and -1 if $c < 0$, $\text{sgn}(x)$ is simply applying $\text{sgn}(\cdot)$ elementwise. Finally $[c]_+ \triangleq \max(c, 0)$.

The following identity appears in many convergence analyses and we will use it many times in this chapter. For all $x, y, z \in \mathcal{H}$,

$$\langle x - y, x - z\rangle = \frac{1}{2}\|x - y\|^2 + \frac{1}{2}\|x - z\|^2 - \frac{1}{2}\|y - z\|^2. \tag{2.10}$$

2.2.2 Proximal Operators

In light of the implicit definition of the proximal operator given in (1.5) we point out that the update equation for GIPSA given in (2.8) can be written implicitly as

$$0 \in x_{k+1} - y_{k+1} + \lambda_k \partial g(x_{k+1}) + \lambda_k \nabla f(z_{k+1}). \tag{2.11}$$

Now $\rho\|\cdot\|_1 \in \Gamma_0(\mathbb{R}^n)$ and the proximal operator associated with it is the shrinkage and soft-thresholding operator $S_\rho(v) : \mathbb{R} \to \mathbb{R}$, applied elementwise. It is defined as

$$S_\rho(v) \triangleq \left[|v| - \rho\right]_+ \text{sgn}(v) \tag{2.12}$$

$$\{\text{prox}_{\rho\|\cdot\|_1}(z)\}_i = S_\rho(z_i),\ i = 1, 2, \ldots, n. \tag{2.13}$$

2.2.3 Assumptions and Optimality Conditions

Now we are ready to precisely state the assumptions used throughout the chapter.

Assumption 1. (Problems (2.1)–(2.2)). The functions f and g are in $\Gamma_0(\mathcal{H})$, $\operatorname{dom}\partial g$ is nonempty, f is Fréchet differentiable everywhere and has an L_f-Lipschitz continuous gradient with $L_f > 0$, and $F_* > -\infty$.

The optimality conditions for Prob. (2.1) under Assumption 1 are as follows. A vector $x_* \in \mathcal{X}_F$ if and only if [1, Corollary 26.3 (vi)]

$$0 \in \partial F(x_*) = (\partial g + \nabla f)(x_*) = \partial g(x_*) + \{\nabla f(x_*)\}. \tag{2.14}$$

Note that this is equivalent to x_* satisfying

$$x_* = \operatorname{prox}_{\lambda g}(x_* - \lambda \nabla f(x_*)) \tag{2.15}$$

for all $\lambda > 0$ [1, Corollary 26.3 (viii)]. Thus x_* is a solution to Prob. (2.1) if and only if it is a fixed point of the *forward-backward operator*: $T_\lambda(x) \triangleq \operatorname{prox}_{\lambda g}(x - \lambda \nabla f(x))$. Note that T_λ is nonexpansive so long as $0 \leq \lambda < 2/L_f$ [1, Thm. 25.8]

The function $\frac{1}{2}\|Ax-b\|^2$ is differentiable and has gradient equal to $A^\top(Ax-b)$ which is Lipschitz continuous with Lipschitz constant equal to the largest eigenvalue of $A^\top A$. The objective function in Prob. (2.3) is bounded below by 0. As previously stated, $\rho\|\cdot\|_1 \in \Gamma_0(\mathcal{H})$ and $\operatorname{dom}\,\partial\|\cdot\|_1 = \mathbb{R}^n$. Therefore Prob. (2.3) satisfies Assumption 1. Thus results proved for Prob. (2.1) hold for all problems, while results proved for Prob. (2.2) also hold for Prob. (2.3). Note that the solution set $\mathcal{X}_F$ of Prob. (2.3) is always nonempty.

2.2.4 Properties of the Solution Set of Prob. (2.2)

Lemma 1 *Suppose Assumption 1 holds for Prob. (2.2) and $\mathcal{X}_F$ is nonempty, then there exists a vector $h_* \in \mathbb{R}^n$ such that for all $x_* \in \mathcal{X}_F$, $\nabla f(x_*) = h_*$. Furthermore, for all $i \in \{1, 2, \ldots, n\}$ and $x_* \in \mathcal{X}_F$: $-h_*^i x_*^i \geq 0$. Finally*

$$\frac{h_*^i}{\rho} \begin{cases} = -1 : & \textit{if } \exists\ x_* \in \mathcal{X}_F : x_*^i > 0 \\ = +1 : & \textit{if } \exists\ x_* \in \mathcal{X}_F : x_*^i < 0 \\ \in [-1, 1] : & \textit{else.} \end{cases}$$

Proof See Sec. 2.6.

Let $E \triangleq \{i : |h_*^i| = \rho\}$ and note that $E^c = \{i : |h_*^i| < \rho\}$. Throughout the chapter we will assume the elements of E are in increasing order. By Lemma 1, we infer that $\mathrm{supp}(x_*) \subseteq E$ for all $x_* \in \mathcal{X}_F$. The set E will be crucial to our local analysis.

2.2.5 Properties of FISTA-CD

Chambolle and Dossal [12] analyzed a variant parameter choice of FISTA which has the $O(1/k^2)$ global objective function convergence rate and also convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$ to a minimizer (see also [38, 44]). They considered the following parameter choice for GIPSA (more specifically I-FBS), which we refer to as FISTA-CD:

$$x_1 = x_0, \quad \lambda_k = \lambda \in (0, 1/L_f], \quad \zeta_k = \beta_k = \frac{k-1}{k+a}, \quad a > 2, \quad \forall k \in \mathbb{N}. \tag{2.16}$$

For a discussion on how to choose a see [12, §4]. We now detail the important properties of FISTA-CD derived in [12] which we need for our analysis.

Lemma 2 ([12]) *Suppose Assumption 1 holds for Prob. (2.1), $\mathcal{X}_F$ is nonempty, and $\{\lambda_k\}_{k\in\mathbb{N}}$ and $\{\zeta_k\}_{k\in\mathbb{N}}$ are chosen as in (2.16). Then for the iterates $\{x_k\}_{k\in\mathbb{N}}$ of (2.6)–(2.8):*

1. *[12, Theorem 4.1: Eq. (25)]*

$$\sum_{k=1}^{\infty} \sum_{j=1}^{k} \left(\prod_{l=j}^{k} \zeta_l \right) \|x_j - x_{j-1}\|^2 < \infty. \tag{2.17}$$

2. *[12, Theorem 4.1] There exists $\hat{x} \in \mathcal{X}_F$ such that $x_k \rightharpoonup \hat{x}$.*

2.3 Main Results

2.3.1 Global Convergence Analysis of GIPSA

In this section we state conditions on $\{\zeta_k, \beta_k, \lambda_k\}_{k\in\mathbb{N}}$ which imply weak convergence of the iterates $\{x_k, y_k, z_k\}_{k\in\mathbb{N}}$ of (2.6)–(2.8) to a minimizer of Prob. (2.1) under Assumption 1. These conditions also imply finite summability of the squared increments of the sequence, which will be useful in the local analysis. The finite summability result also makes it trivial to prove criticality of the limit points which we include for completeness.

Theorem 3 *For Prob. (2.1), suppose Assumption 1 holds. Assume $\{\lambda_k\}_{k\in\mathbb{N}}$ is positive and nondecreasing, and there exists $\varepsilon > 0$, $0 < \gamma < 2$ and $0 \leq \overline{\beta} < 1$ such that sequences $\{\lambda_k, \zeta_k, \beta_k\}_{k\in\mathbb{N}}$ satisfy:*

$$\begin{aligned} 0 \leq \zeta_k \leq 1, \quad 0 \leq \beta_k \leq \overline{\beta}, \quad \lambda_k\zeta_k \leq \frac{\beta_k}{L_f}, \quad \lambda_k \leq \frac{2-\gamma}{L_f} \\ \text{and} \quad 2 - \lambda_k L_f(1-\zeta_k) - \beta_k - \beta_{k+1} \geq \varepsilon \end{aligned} \tag{2.18}$$

for all $k \in \mathbb{N}$. Then for the iterates $\{x_k, y_k, z_k\}_{k\in\mathbb{N}}$ of (2.6)–(2.8):

(i) $\sum_{k\in\mathbb{N}} \|x_k - x_{k-1}\|^2 < \infty$,

(ii) $d(0, \partial F(x_k)) \to 0$ *as* $k \to \infty$.

(iii) If $\mathcal{X}_F$ is nonempty then there exists $\hat{x} \in \mathcal{X}_F$ such that $x_k \rightharpoonup \hat{x}$, $y_k \rightharpoonup \hat{x}$ and $z_k \rightharpoonup \hat{x}$.

Proof See Sec. 2.5.

With some effort Theorem 3 can be extended to inexact proximal operators through the use of the enlarged subdifferential under a summability condition on the errors [48]. It can also be extended to versions which incorporate a relaxation parameter. To simplify the presentation, proof, and notation, we do not detail these elaborations.

For the special case where $\zeta_k = 0$, Theorem 3 provides more general parameter constraints than existing guarantees derived in [33]. Suppose $\lambda_k = \lambda \in [0, 2/L_f)$, then [33] requires β_k to be nondecreasing and to satisfy $0 \leq \beta_k \leq \overline{\beta}$ where $\overline{\beta} < (2 - \lambda L_f)/6$. On the other hand, Theorem 3 requires: $\beta_k + \beta_{k+1} \leq 2 - \lambda L_f - \varepsilon$, which is satisfied if $\overline{\beta} < (2 - \lambda L_f)/2$. Note that [33] and Theorem 3 have the same requirement on the stepsize.

2.3.2 Specialized Conditions for I-FBS

We now simplify the conditions for the case of I-FBS, i.e. $\zeta_k = \beta_k$. For consistency, let $\overline{\zeta} = \overline{\beta}$. In this case:

$$2 - \zeta_k - \zeta_{k+1} + \lambda_k L_f(\zeta_k - 1) \geq 1 - \zeta_{k+1} \geq 1 - \overline{\zeta} > 0.$$

Therefore $\varepsilon = 1 - \overline{\zeta}$ satisfies (2.18). Next note that if we choose any $\gamma < 1$, the condition on the stepsize simplifies to $\lambda_k \leq 1/L_f$ for all k. In the case of FBS: i.e. $\zeta_k = \beta_k = 0$, Thm. 3 allows for larger stepsizes: $\lambda_k L_f \leq 2 - \gamma < 2$, which agrees with the standard criteria for FBS (e.g. [1, Thm. 25.8]). We formalize this in the following corollary.

Corollary 4 *Assume $\{\lambda_k\}_{k\in\mathbb{N}}$ is nondecreasing, $\lambda_k \in (0, 1/L_f]$, and $0 \leq \zeta_k \leq \overline{\zeta} < 1$ for all k. Then for the iterates of I-FBS (2.4)–(2.5), for Prob. (2.1), Assumption 1 implies (i) and (ii) of Theorem 3. Assumption 1 and nonemptiness of $\mathcal{X}_F$ imply (iii) of Theorem 3.*

Note that the condition on ζ_k is more general than the requirement on the inertia parameter given in [18] which is

$$1 - 3\zeta_k - \lambda L_f(1-\zeta_k)^2/2 \geq \eta$$

where $\lambda_k = \lambda \in (0, 2/L_f]$ for all k and $\eta > 0$ is some constant. Note that [18] does allow larger values of the stepsize λ, up to $2/L_f$ so long as ζ_k is sufficiently small.

We emphasize that Corollary 4 does not apply to any of the FISTA variants because in all such algorithms $\zeta_k \to 1$. See [12] for a proof of weak convergence of the iterates of FISTA-CD.

It is interesting to note that for FBS the convergence criteria are the same for Prob. (2.9) (monotone inclusion problem) and Prob. (2.1) [1, Thm. 25.8 and Cor. 27.9]. However for GIPSA and I-FBS, this does not appear to be the case.

2.3.3 Discussion of the General Case

We have discussed the special cases $\zeta_k = \beta_k$ and $\zeta_k = 0$. We now discuss the general case. To simplify the discussion, consider fixed choices, i.e. $\{\zeta_k, \beta_k, \lambda_k\} = \{\zeta, \beta, \lambda\}$ for all k. Then (2.18) becomes

$$\zeta \in [0,1], \quad \beta \in [0,1), \quad 0 < \lambda L_f \quad \leq \quad \min\left\{\frac{\beta}{\zeta}, \frac{2(1-\beta)-\varepsilon}{1-\zeta}\right\} \tag{2.19}$$

for some $\varepsilon > 0$ with the convention: $0/0 = \infty$. Now if we set ε to 0, the two arguments to min in (2.19) are equal if $\zeta = \zeta^*(\beta) = \frac{\beta}{2-\beta}$. Substituting this into the expression yields $\lambda L_f < 2 - \beta$. If $\zeta < \zeta^*(\beta)$ then the right-hand expression in the argument of min is the smallest, else it is the left-hand expression. Thus the condition on λ is

$$\lambda L_f < \begin{cases} \frac{2(1-\beta)}{1-\zeta}: & \text{if } 0 \leq \zeta \leq \zeta^*(\beta) \\ \frac{\beta}{\zeta}: & \text{if } \zeta^*(\beta) \leq \zeta \leq 1. \end{cases} \tag{2.20}$$

While $\zeta = \zeta^*(\beta)$ provides the largest range of feasible stepsize parameters according to our theoretical convergence analysis, we do not claim that it is the "best" choice for a given instance of Prob. (2.1). For I-FBS for Prob. (2.2) our local convergence analysis derives some good parameter choices (See Sections 2.3.4–2.3.8). However determining good parameter choices more generally for GIPSA is a topic of future work. Nevertheless it is important to establish general conditions for convergence before attempting to determine appropriate choices via an empirical study or further theoretical analysis.

2.3.4 Finite Convergence Results for I-FBS

We now turn our attention to Problems (2.2)–(2.3) and establish the local convergence behavior of I-FBS and FISTA-CD. The upcoming theorem proves convergence in a finite number of iterations for the components in E^c to 0, and for the components in E to the optimal sign (recall $E \triangleq \{i : |h_*^i| = \rho\}$ where h_* is defined in Lemma 1). Following the terminology of [43, 13] we will refer to this as the "finite active manifold identification" property. The manifold in the ℓ_1-norm setting is the halfspace of vectors with support a subset of E and nonzero components with sign equal to $-h_*^i/\rho$.

Theorem 5 *For Prob. (2.2) suppose that Assumption 1 holds and $\mathcal{X}_F$ is nonempty, thus there exists $h_* \in \mathbb{R}^n$ satisfying the conditions of Lemma 1. Assume that either:*

1. *$\{\lambda_k\}_{k\in\mathbb{N}}$ is nondecreasing, $0 < \lambda_k \leq 1/L_f$, and $0 \leq \zeta_k \leq \overline{\zeta} < 1$ for all $k \in \mathbb{N}$, or*

2. *$\{\zeta_k, \lambda_k\}_{k\in\mathbb{N}}$ are chosen according to (2.16) (i.e. FISTA-CD),*

then for all but finitely many k the iterates $\{x_k, y_k\}_{k\in\mathbb{N}}$ of I-FBS, (2.4)–(2.5), satisfy

$$\operatorname{sgn}\left(y_k^i - \lambda_{k-1}\nabla f(y_k)^i\right) \quad = \quad -\frac{h_*^i}{\rho},\ \forall i : |h_*^i| = \rho, \tag{2.21}$$

and

$$x_k^i = y_k^i = 0,\ \forall i : |h_*^i| < \rho. \tag{2.22}$$

Proof See Sec. 2.7.

Note that if $x_k^i \neq 0$, then (2.12)–(2.13) implies that $\operatorname{sgn}(x_k^i) = \operatorname{sgn}(y_k^i - \lambda_k\nabla f(y_k)^i)$. Note that [3, Theorem 4.5] is recovered when $\zeta_k = 0$ for all k.

The authors of [43] studied finite convergence results for prox-regular and partially smooth functions, which includes Prob. (2.2). Specialized to this problem, the analysis of [43, Theorem 5.3] establishes finite convergence in support and sign for any algorithm which produces a convergent iterate sequence, under the following additional condition: $0 \in \operatorname{rint}(\partial F(x_*))$ for the limit $x_* = \lim_{k\to\infty} x_k$. In the context of Prob. (2.2) this condition is equivalent to the "strict complementarity condition" discussed in Sec. 2.3.7, i.e. $E = \operatorname{supp}(x_*)$. In contrast, Theorem 5 is more general in that it proves finite convergence to 0 on $E^c \subseteq \operatorname{supp}(x_*)^c$ and sign on E. It does not require $E = \operatorname{supp}(x_*)$. However when this is true, Theorem 5 coincides with [43, Theorem 5.3].

Given that I-FBS converges in a finite number of iterations to the optimal manifold, it could be desirable to switch to a local procedure which searches in the space of lower dimension. Indeed for Prob. (2.3), if the solution is unique and the support and sign of the solution are known, then the values of the nonzero entries can be computed by solving a linear system with dimension equal to the number of nonzero entries [2, p. 20]. Theorem 5 also motivates combining two-stage "active-set" strategies such as the one described in [49] with I-FBS or FISTA-CD. Active-set strategies alternate between iterated shrinkage-thresholding updates to identify the active manifold, and local optimization procedures to estimate the nonzero entries. Using I-FBS/FISTA-CD to identify the active manifold within such a framework is an interesting topic for future work.

2.3.5 Reduction to Smooth Minimization

Theorem 5 allows us to characterize the behavior of I-FBS after a manifold identification period of finite duration. In the following theorem, we show that after a finite number of iterations, I-FBS (including FISTA-CD) reduces indefinitely to minimizing a smooth function over E subject to an orthant constraint.

Theorem 6 *For Prob. (2.2) suppose that Assumption 1 holds and $\mathcal{X}_F$ is nonempty, thus there exists $h_* \in \mathbb{R}^n$ satisfying the properties of Lemma 1. Recall $E \triangleq \{i : |h_*^i| = \rho\}$ and let $\phi : \mathbb{R}^{|E|} \to \mathbb{R}$ be defined as*

$$\phi(x^E) \;\triangleq\; -(h_*^E)^\top x^E + f\left((x^E, 0)\right), \tag{2.23}$$

where $x \in \mathbb{R}^n$. Let the set $O_E \subset \mathbb{R}^{|E|}$ be defined as

$$O_E \triangleq \{v \in \mathbb{R}^{|E|} : -\text{sgn}(h_*^{E(j)})\, v_j \geq 0,\ \forall j \in \{1, 2, \ldots, |E|\}\}. \quad (2.24)$$

Assume that either

1. *$\{\lambda_k\}_{k\in\mathbb{N}}$ is nondecreasing, $0 < \lambda_k \leq 1/L_f$, and $0 \leq \zeta_k \leq \overline{\zeta} < 1$, for all k, or*

2. *$\{\zeta_k, \lambda_k\}_{k\in\mathbb{N}}$ are chosen according to FISTA-CD in (2.16),*

then, for all but finitely many k, the iterates $\{x_k, y_k\}_{k\in\mathbb{N}}$ of I-FBS, (2.4)–(2.5), satisfy

$$x_{k+1}^E = P_{O_E}\left(y_{k+1}^E - \lambda_k \nabla\phi(y_{k+1}^E)\right), \quad (2.25)$$

and $F(x_k) = \phi(x_k^E)$, where $F(x) = f(x) + \rho\|x\|_1$ and P_{O_E} is the orthogonal projector onto O_E.

Proof See Sec. 2.8. The result of [3, Corollary 4.6] is recovered when $\zeta_k = 0$ for all k.

2.3.6 Local Linear Convergence Under Local Strong Convexity

The analysis of the previous two sections shows that, after a finite number of iterations, I-FBS reduces to minimizing the function ϕ subject to an orthant constraint. This function can be strongly convex even if f does not have this property. If ϕ is strongly convex, then local linear convergence can be achieved, as we prove in the following corollary. Note that strong (in fact strict) convexity of ϕ implies solution uniqueness for Prob. (2.2).

Corollary 7 *For Prob. (2.2) suppose that Assumption 1 holds and ϕ defined in (2.23) is strongly convex. Let l_E be the strong convexity parameter of ϕ. If $\lambda \in (0, 1/L_f]$, $0 < \mu \leq l_E$,*

$$\lambda_k = \lambda \text{ and } \zeta_k = \frac{1 - \sqrt{\mu\lambda}}{1 + \sqrt{\mu\lambda}} \quad \forall k \in \mathbb{N}, \quad (2.26)$$

then the iterates $\{x_k\}_{k\in\mathbb{N}}$ of I-FBS, (2.4)–(2.5), converge to the unique solution x_* of Prob. (2.2) R-linearly and $F(x_k)$ converges to F_* R-linearly where $F(x) = f(x) + \rho\|x\|_1$. Specifically

$$\|x_k - x_*\|^2 = O\left(\left(1 - \sqrt{\mu\lambda}\right)^k\right) \text{ and } F(x_k) - F_* = O\left(\left(1 - \sqrt{\mu\lambda}\right)^k\right).$$

Proof Recall the definition of $E \triangleq \{i : |h_*^i| = \rho\}$. We consider k large enough that I-FBS has reduced to minimizing the l_E-strongly convex function ϕ, i.e. (2.25) holds, $x_k^{E^c} = 0$, and $F(x_k) = \phi(x_k^E)$. The result can now be seen by considering Nesterov's constant momentum scheme of [32, p. 76], however the variable μ now represents a lower bound for the true strong convexity parameter of ϕ. It can be verified that this does not change the result given in [32, Thm 2.2.3]. Furthermore we allow stepsizes other than $1/L_f$, which is discussed on [32, p. 72]. Finally the minimization is with respect to the orthant O_E defined in (2.24). This simple modification of Nesterov's scheme is discussed in [32, Algorithm (2.2.17)].

Note this local linear convergence result does not depend on strict complementarity (i.e. $E = \text{supp}(x_*)$) unlike the local analysis of FBS in [13, 39]. Suppose $\mu = l_E$ and $\lambda = 1/L_f$, then the convergence rate and iteration complexity are respectively

$$F(x_k) - F_* = O\left(\left(1 - \sqrt{\frac{l_E}{L_f}}\right)^k\right), \text{ iter. comp.} = \Omega\left(\sqrt{\frac{L_f}{l_E}}\log\frac{1}{\epsilon}\right). \quad (2.27)$$

Given the nature of ϕ this iteration complexity is optimal [32]. Indeed it is better than the iteration complexity of FBS [3] (which corresponds to I-FBS with ζ_k equal to 0) which is $\Omega\left((L_f/l_E)\log 1/\epsilon\right)$.

Other parameter choices, such as Constant Scheme III of [32, p. 84], will also achieve local linear convergence with the same rate. However these choices along with (2.26) are difficult to use in practice as they depend on l_E, which is hard to estimate. In Sec. 2.3.8 we will show how the rate and corresponding iteration complexity in (2.27) can be achieved without knowledge of l_E by combining a restart scheme with FISTA-CD.

2.3.7 Local Linear Convergence Under Strict Complementarity

Local linear convergence can also be proved for Prob. (2.3) without requiring solution uniqueness. We require $\lim_{k\to\infty} x_k \triangleq x_* \in \mathcal{X}_F$ to obey the so-called "strict complementarity" condition: $E = \text{supp}(x_*)$, where $E \triangleq \{i : |h_*^i| = \rho\}$. This is a common assumption also used in [3, 13, 42, 43, 39]. Note that this condition is not necessary for x_* to be the unique minimizer for Prob. (2.2) [50, Example (4)]. It is also not sufficient, which can be seen by considering the following instance of Prob. (2.3) taken from [50, Example

(4)]:

$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 2 & -2 \end{bmatrix}, \ b = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}, \ \rho = 1.$$

This example has $E = \{1, 2, 3\}$ and has infinitely many solutions which satisfy strict complementarity, such as $(1/4, 3/8, 1/8)^\top$. The name "strict complementarity" comes from considering the dual problem to Prob. (2.2) [51, §6].

First we state the following proposition which shows that the proximal step (2.8) of I-FBS reduces to a gradient descent step after finitely many iterations, thus the proximal operator may be ignored. The proof follows [3, Lemma 5.3] closely.

Proposition 8 *For Prob. (2.2), suppose Assumption 1 holds and $\mathcal{X}_F$ is nonempty, thus there exists $h_* \in \mathbb{R}^n$ satisfying the conditions of Lemma 1. Let $E \triangleq \{i : |h_*^i| = \rho\}$. Let $\{x_k, y_k\}_{k\in\mathbb{N}}$ be the iterates of I-FBS, (2.4)–(2.5). Assume either:*

1. *$\{\lambda_k\}_{k\in\mathbb{N}}$ is nondecreasing, $0 < \lambda_k \leq 1/L_f$ and $0 \leq \zeta_k \leq \overline{\zeta} < 1$ for all $k \in \mathbb{N}$, or*

2. *$\{\lambda_k, \zeta_k\}_{k\in\mathbb{N}}$ satisfy (2.16).*

Let $x_* = \lim_{k\to\infty} x_k$ which exists by Corollary 3 . Then for all but finitely many k,

$$x_k^i = y_k^i - \lambda_{k-1}(\nabla f(y_k)^i - h_*^i), \quad \forall i \in supp(x_*). \tag{2.28}$$

Proof See Sec. 2.9.

Under strict complementarity ($E = \text{supp}(x_*)$) we will refer to the regime where (2.28) is satisfied and $x_k^{E^c} = 0$ as "the large-k regime" throughout the remainder of the chapter. We refer to the regime where these conditions are not satisfied as "the small-k regime".

Now we consider a simple fixed parameter choice for Prob. (2.3). Under the strict complementarity condition, we can prove local linear convergence for any fixed choice of the inertia parameter in $[0, 1)$ and the stepsize in $(0, 1/L_f]$. The analysis turns out to be fairly elementary in this case since for this problem once in the large-k regime the iterations form a simple second-order linear homogeneous recursion which has been studied before, for example in [45]. Note that we do not require the function ϕ defined in (2.23) to be strongly convex nor the minimizer to be unique.

Theorem 9 *For Prob. (2.3) there exists $h_* \in \mathbb{R}^n$ satisfying the conditions of Lemma 1. Let $E \triangleq \{i : |h_*^i| = \rho\}$. Let $\zeta_k = \zeta \in [0, 1)$ and $\lambda_k = 1/L_f$ for all $k \in \mathbb{N}$. Let the iterates of I-FBS, (2.4)–(2.5), be $\{x_k\}_{k\in\mathbb{N}}$ and $\lim_{k\to\infty} x_k = x_*$, which exists by Corollary 3. Suppose $E = supp(x_*)$ (i.e. strict complementarity holds). Then x_k achieves local Q-linear convergence. In particular there exists $K > 0$, $C > 0$, and $q \in (0, 1)$ such that $\|x_k - x_*\| = Cq^k$ for all $k > K$. Let $\hat{l}_E$ be the smallest nonzero eigenvalue of $A_E^\top A_E$. If $\hat{l}_E > 0$, $0 < \mu \leq \hat{l}_E$ and $\zeta = (1 - \sqrt{\mu/L_f})/(1 + \sqrt{\mu/L_f})$, then $q = \left(1 - \sqrt{\mu/L_f}\right)^{1/2}$. If $\hat{l}_E = 0$ then $q \leq \zeta$. Finally $F(x_k)$ converges to F_* with rate q^2.*

Proof See Sec. 2.9.

This theorem extends [3, Theorem 4.11] to include a momemtum term and shows that if the momentum term is chosen correctly it can accelerate the local Q-linear convergence rate. For simplicity we prove the result only for $\lambda_k = 1/L_f$ but the case $\lambda_k = \lambda \in (0, 1/L_f]$ can also be shown. We stress that in practice the quantities $\hat{l}_E$ and L_f are typically not known. In the next section we show that a simple adaptive restart scheme can be incorporated into FISTA-CD to create a scheme which obtains the optimal iteration complexity without needing knowledge of $\hat{l}_E$.

2.3.8 Asymptotic Behavior of FISTA-CD

We now ask, what is the convergence behavior of FISTA-CD in the large-k regime? For Prob. (2.3) we see that once (2.28) holds, the iterates are in the form of an inhomogeneous second-order linear recurrence which has been studied previously in [45] and [42, §5–6]. It is difficult to analyze this recursion because ζ_k changes at each iteration and to do so rigorously requires a subtle argument following the one presented in [42, §5–6]. A simpler route to understanding the behavior is to use the homogeneous approximation of [45, §4] which sets ζ_k fixed and "close" to 1. This approximation implies that under strict complementarity, once in the large-k regime and for ζ_k sufficiently close to 1 (recall $\zeta_k \to 1$ for this parameter choice), FISTA-CD will exhibit nonmonotone oscillatory behavior in the objective function values with suboptimal Q-linear rate:

$$\exists K, C > 0 : F(x_k) - F_* = C\left(\left(1 - \lambda\hat{l}_E\right)^k\right), \quad \forall k > K, \tag{2.29}$$

where $\hat{l}_E$ is defined in Theorem 9. This is the same as the convergence rate achieved by FBS (I-FBS with $\zeta_k = 0$ and $\lambda_k = 1/L_f$ for all $k \in \mathbb{N}$, although a slightly better rate can be achieved with $\lambda_k = 2/(\hat{l}_E + L_f)$ which nevertheless has the same iteration complexity [3]).

For strongly convex quadratic minimization problems, [45] suggested restarting the inertia sequence of Nesterov's method whenever a certain restart condition is observed. By applying the homogeneous approximation of [45] to FISTA-CD once in the large-k regime, we argue that we can improve the asymptotic convergence rate by incorporating such a restart technique. Thus even though the overall problem is nonsmooth and in general not strongly convex, restarting can improve the convergence properties of FISTA-CD. Restart schemes such as the "speed restart" scheme [26], the "gradient restart" scheme [45], the "objective function" scheme [45], or the more conservative restart scheme of [46] could be incorporated into FISTA-CD. For simplicity we elaborate only the objective function restart scheme of [45] and we call the new method FISTA-CD-RE ("FISTA-CD with restart"). The idea is as follows. Whenever we observe $F(x_{k+1}) > F(x_k)$, set the iteration counter k in (2.16) to 1, and set $x_0 = x_k$ and $x_1 = x_k$. In other words restart FISTA-CD at the current point. We refer the reader to [45] for full details and analysis which can be applied to our situation in the large-k regime (under strict complementarity). The homogeneous approximation of [45] suggests FISTA-CD-RE will have the optimal iteration complexity

$$\text{iter. comp.} = \Omega\left(\sqrt{\frac{L_f}{\hat{l}_E}} \log 1/\epsilon\right)$$

and rate

$$F(x_k) - F_* = C'\left(1 - \sqrt{\hat{l}_E/L_f}\right)^k. \qquad (2.30)$$

Remarkably it achieves this iteration complexity without knowledge of $\hat{l}_E$, the local strong convexity parameter. Thus we do not need to know $\hat{l}_E$ in order to achieve the optimal convergence rate given in Theorem 9 with $\mu = \hat{l}_E$. The method will also have $O(1/k^2)$ convergence rate while no restarts occur. It is also straightforward to incorporate a backtracking line search into FISTA-CD-RE, such as the one described in [4, p. 194], so that the method does not require L_f.

We stress that the convergence rates given in (2.29) and (2.30) can be proved rigorously using arguments in the spirit of [42, §5–6]. For simplicity we omit the details. Our main contribution is to show that, for all but finitely many iterations, FISTA-CD reduces to a form that has been studied previously in [42, 45], from which convergence behavior can be extracted.

2.4 Numerical Results

We now provide a synthetic experiment to corroborate the theoretical findings of this chapter.

2.4.1 Experiment Details

We consider a randomly generated instance of Prob. (2.3). The parameters of the experiment are $n = 2000$, $m = 1000$ and $\rho = 0.1$. The entries of A are drawn i.i.d. from the normal distribution with mean 0 and variance 0.01. The vector b is given by Ax_0, where x_0 has 260 nonzero entries generated i.i.d. from the 0-mean unit variance normal distribution, and support set chosen uniformly at random. Recall that l_E denotes the smallest nonzero eigenvalue of $A_E^\top A_E$ where E is defined in Sec. 2.2.4. Note that for such a randomly generated problem where the entries of A are drawn from a continuous probability distribution, $l_E > 0$, and thus the solution is unique, with probability 1 [7]. We run (2.6)–(2.8) with several choices for the parameters. For the most general form GIPSA we consider four parameter choices and choose the stepsize $\lambda_k = \lambda$ satisfying (2.20) with equality minus a small constant 0.01. For I-FBS, we considered three parameter choices and chose the stepsize as $1/L_f$. The Lipschitz constant L_f is the largest eigenvalue of $A^\top A$ and is computed via the SVD. These parameter choices and their identifiers are given in Table 2.1 where ζ^* is the locally optimal choice from Thm. 9: $(1-\sqrt{l_E/L_f})/(1+\sqrt{l_E/L_f})$. We estimate E via the interior point solver of [11] which we use to find an approximate solution x_* such that the relative objective function error is no greater than 10^{-6}. We then compute $h_* = \nabla f(x_*)$, and estimate E as the set of all i such that $\rho - |h_*^i|$ is smaller than 10^{-4}. We then use the SVD of A_E to estimate l_E. Using this approach, ζ^* is estimated as 0.77 for this experiment. Note that this is obviously not a practical method for estimating the optimal inertia parameter. The purpose of this experiment is simply to test the theoretical findings of Sections 2.3. In fact this experiment demonstrates that our proposal, FISTA-CD-RE, has the same asymptotic convergence rate as I-FBS with the optimal inertia parameter yet does not need to estimate l_E. We run FISTA, which is

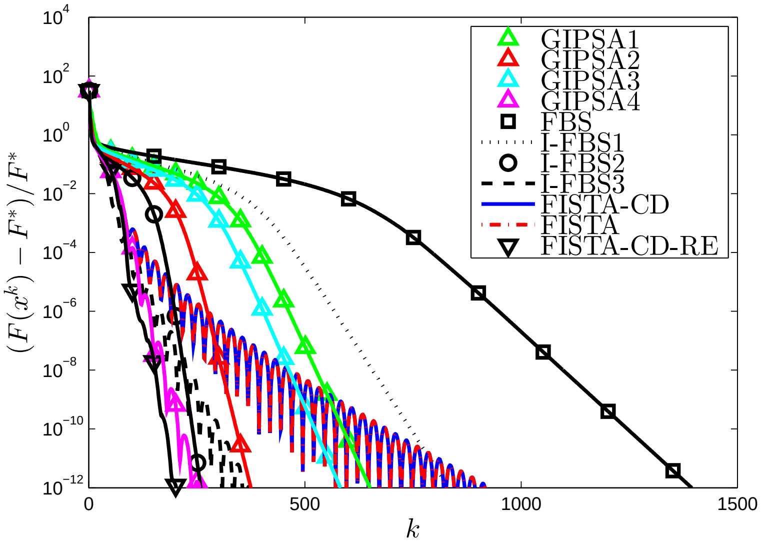

Figure 2.1: Simulation results: showing relative error $(F(x_k) - F_*)/F_*$ versus iteration k for Experiment 1.

parameter choice [4, Eq. (4.2)–(4.3)] with $\zeta_k = \beta_k$. We also run FISTA-CD which is (2.6)–(2.8) with the parameter choice given in (2.16) with $\lambda = 1/L_f$ and $a = 2.1$. We run FISTA-CD-RE with the same values for λ and a as FISTA-CD. All algorithms are initialized to $x_1 = x_0 = 0$. The results are shown in Fig. 2.1 where we plot the relative error $(F(x_k) - F_*)/F_*$ versus k. Note the y-axis is logarithmic.

Table 2.1: The fixed parameter choices

Algorithm Identifier	$\zeta_k = \zeta$	$\beta_k = \beta$
FBS	0	0
GIPSA1	0	0.8
GIPSA2	0.4	ζ^*
GIPSA3	1	0.9
GIPSA4	1	0.7
I-FBS1	0.4	0.4
I-FBS2	ζ^*	ζ^*
I-FBS3	0.95	0.95

2.4.2 Repeated Trials

We repeat this experiment 1000 times with different randomly drawn A and x_0 from the distributions described above. For each trial we record the number of iterations until after which the relative error remains below tol,

Table 2.2: Results for repeated trials (Sec. 2.4.2). The algorithm with the lowest average # of iterations is boxed.

Algorithm	Average # iterations to rel. err. 10^{-2} (1000 trials)	Average # iterations to rel. err. 10^{-6} (1000 trials)
FBS	564	916
GIPSA1	293	457
GIPSA2	175	268
GIPSA3	248	395
GIPSA4	73	137
I-FBS1	339	549
I-FBS2	133	201
I-FBS3	$\boxed{\mathbf{51}}$	160
FISTA	65	255
FISTA-CD	65	255
FISTA-CD-RE	65	$\boxed{\mathbf{111}}$

i.e. $k : (F(x_j) - F_*)/F_* \leq \text{tol}, \forall j \geq k$.[3] The average of this number across the 1000 trials is given in Table 2.2 for tol $\in \{10^{-2}, 10^{-6}\}$ and all algorithms.

2.4.3 Observations

First let's look at Fig. 2.1. Although the figure shows objective function values, since the minimizer is unique, convergence of the iterates is implied, which corroborates Theorem 3. All tested parameter choices for I-FBS transition from a manifold identification period to a local linear convergence period, corroborating Theorem 5. Interestingly the GIPSA parameter choices also exhibit local linear convergence suggesting it is possible to extend some of the results of Theorem 5 to these choices. Furthermore, adding inertia does improve the asymptotic rate and using ζ^* achieves the best asymptotic rate. However, our proposed FISTA-CD-RE essentially achieves the same asymptotic rate despite not needing to know l_E. The upper bound for the asymptotic convergence rate of I-FBS with inertial parameter ζ^* is computed using (2.27) to be 0.89, which compares with an empirically determined rate of 0.83. However the fixed choice $\zeta_k = \zeta^*$ is outperformed by the larger choice $\zeta = 0.95$, along with GIPSA4, FISTA, FISTA-CD and FISTA-CD-RE in the small-k regime (i.e. before linear convergence commences). In the large-k regime FISTA-CD exhibits nonmonotone oscillatory

[3] F_* is approximated by the smallest objective function value among all tested algorithms after 1500 iterations.

behavior and suboptimal asymptotic convergence as predicted in Sec. 2.3.8.

Now we look at Table 2.2. For a "low accuracy" solution (defined here as rel. err. less than 10^{-2}), I-FBS with $\zeta = 0.95$ performs best and there is no difference between FISTA, FISTA-CD and FISTA-CD-RE. GIPSA4 is also competitive. FISTA-CD and FISTA-CD-RE are identical because a restart had not yet occurred in any of the 1000 trials. The strong performance of I-FBS with $\zeta = 0.95$ in the low accuracy regime is interesting and we cannot explain it with the existing theory. However for such large values of the inertia parameter we expect the performance to be approximately similar to FISTA and its variants. For a "high accuracy" solution (defined here as rel. err. less than 10^{-6}), our proposed FISTA-CD-RE outperforms all other algorithms. It requires on average fewer than half as many iterations as FISTA or FISTA-CD at essentially the same per-iteration cost.[4] Furthermore our proposal does not require one to tune the momentum parameters based on the local curvature constant. In contrast we see that fixed choices for I-FBS and GIPSA are highly sensitive to the curvature.

2.5 Proof of Theorem 3

Before proving the theorem, we give three lemmas, beginning with the celebrated lemma due to Opial.

Lemma 10 ([52], Opial's lemma) *Suppose $\{x_k\}$ is a sequence in $\mathcal{H}$ and $S \subset \mathcal{H}$ is a nonempty set such that:*

1. $\lim_{k\to\infty} \|x_k - x_*\|$ *exists for every $x_* \in S$,*
2. *Every weakly convergent subsequence of $\{x_k\}_{k\in\mathbb{N}}$ weakly converges to some $x_* \in S$.*

Then there exists $\hat{x} \in S$ such that $x_k \rightharpoonup \hat{x}$.

It is trivial to verify that the second condition of Opial's lemma holds for GIPSA, so long as $x_k - x_{k-1} \to 0$. We do this in the following lemma.

Lemma 11 *For Prob. (2.1) suppose Assumption 1 holds and $\mathcal{X}_F$ is nonempty. Let $\{x_k\}_{k\in\mathbb{N}}$ be the sequence generated by (2.6)–(2.8). Suppose $x_{v_k} \rightharpoonup x$ for some subsequence $\{v_k\}_{k\in\mathbb{N}} \subseteq \mathbb{N}$, and $x_k - x_{k-1} \to 0$. Then $x \in \mathcal{X}_F$.*

[4] Despite having an additional function evaluation per iteration, FISTA-CD-RE only requires one matrix multiply per iteration, which is the same as FISTA-CD and FISTA since the matrix multiply is the dominant cost.

Proof The proof follows the techniques of [33, Theorem 2.1]. Thanks to (2.6) and the assumption that $x_k - x_{k-1} \to 0$, we know that $y_{k+1} - x_k \to 0$ and thus $x_k - y_k \to 0$. Similarly by (2.7) we see that $z_{k+1} - x_k \to 0$ and therefore $z_k - x_k \to 0$. Now by (2.11)

$$\begin{aligned} &-\frac{1}{\lambda_{v_k-1}}(x_{v_k} - y_{v_k}) + \nabla f(x_{v_k}) - \nabla f(z_{v_k}) \\ &\in \{\nabla f(x_{v_k})\} + \partial g(x_{v_k}). \end{aligned} \tag{2.31}$$

Now passing to the limit $v_k \to \infty$, using the fact that ∇f is Lipschitz continuous, and [48, Proposition 3.4(b)], we infer that $0 \in \partial g(x) + \{\nabla f(x)\}$, therefore $x \in \mathcal{X}_F$ by optimality condition (2.14).

The final Lemma is standard in the analysis of inertial methods.

Lemma 12 *Let $\{\varphi_k, \delta_k, \sigma_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$. If $\varphi_{k+1} - \varphi_k \le \sigma_k(\varphi_k - \varphi_{k-1}) + \delta_k$ for all k where $\sigma_k \le \overline{\sigma} < 1$ and $\sum_{k\in\mathbb{N}} \delta_k < \infty$, then $\lim_{k\to\infty} \varphi_k$ exists.*

Proof We refer to [24, Thm 3.1].

We now turn our attention to Theorem 3. We prove statement (i) by using the multistep Lyapunov function from [24] which is shown to be nonincreasing. The proof of (ii) is trivial. Finally to prove (iii) we use Lemma 12 to prove the first condition of Opial's lemma holds (the second condition of Opial's lemma holds by Lemma 11).

Proof of Theorem 3 statement (i)

Recall the notation: $\Delta_k \triangleq x_k - x_{k-1}$. Define the Lyapunov energy function: $V_k \triangleq F(x_k) + \frac{\beta_k}{2\lambda_k}\|\Delta_k\|^2$. We will show that V_k is nonincreasing. Using (1.1) and (1.3), first note that

$$\begin{aligned} F(x_{k+1}) - F(x_k) &= f(x_{k+1}) - f(x_k) + g(x_{k+1}) - g(x_k) \\ &= f(x_{k+1}) - f(z_{k+1}) + f(z_{k+1}) - f(x_k) \\ &\quad +g(x_{k+1}) - g(x_k) \\ &\le \langle \nabla f(z_{k+1}), x_{k+1} - z_{k+1}\rangle + \frac{L_f}{2}\|x_{k+1} - z_{k+1}\|^2 \\ &\quad + \langle \nabla f(z_{k+1}), z_{k+1} - x_k\rangle + \langle v, x_{k+1} - x_k\rangle \\ &\quad \forall v \in \partial g(x_{k+1}) \\ &= \langle \nabla f(z_{k+1}) + v, \Delta_{k+1}\rangle \\ &\quad +\frac{L_f}{2}\|x_{k+1} - z_{k+1}\|^2. \end{aligned} \tag{2.32}$$

Using the fact that λ_k is nondecreasing and (2.32), we write

$$\begin{aligned} V_{k+1} - V_k &\leq F(x_{k+1}) - F(x_k) + \frac{1}{2\lambda_k}\left(\beta_{k+1}\|\Delta_{k+1}\|^2 - \beta_k\|\Delta_k\|^2\right) \\ &\leq \langle \nabla f(z_{k+1}) + v, \Delta_{k+1}\rangle + \frac{L_f}{2}\|z_{k+1} - x_{k+1}\|^2 \\ &\quad + \frac{1}{2\lambda_k}\left(\beta_{k+1}\|\Delta_{k+1}\|^2 - \beta_k\|\Delta_k\|^2\right), \\ &\quad \forall v \in \partial g(x_{k+1}). \end{aligned} \tag{2.33}$$

Note that by the definition of the prox operator, $x_{k+1} \in \operatorname{dom}\partial g$ which is nonempty by Assumption 1. Using (2.11) in (2.33) implies

$$\begin{aligned} V_{k+1} - V_k &\leq \frac{1}{\lambda_k}\langle y_{k+1} - x_{k+1}, \Delta_{k+1}\rangle + \frac{L_f}{2}\|z_{k+1} - x_{k+1}\|^2 \\ &\quad + \frac{1}{2\lambda_k}\left(\beta_{k+1}\|\Delta_{k+1}\|^2 - \beta_k\|\Delta_k\|^2\right). \end{aligned} \tag{2.34}$$

Now using (2.6) and (2.7) we derive:

$$y_{k+1} - x_{k+1} = \beta_k\Delta_k - \Delta_{k+1} \quad \text{and} \quad z_{k+1} - x_{k+1} = \zeta_k\Delta_k - \Delta_{k+1}. \tag{2.35}$$

Substituting (2.35) into (2.34) yields

$$\begin{aligned} V_{k+1} - V_k &\leq \left(\frac{\beta_k - \zeta_k\lambda_k L_f}{\lambda_k}\right)\langle \Delta_{k+1}, \Delta_k\rangle + \frac{\zeta_k^2\lambda_k L_f - \beta_k}{2\lambda_k}\|\Delta_k\|^2 \\ &\quad + \left(\frac{\beta_{k+1}}{2\lambda_k} + \frac{L_f}{2} - \frac{1}{\lambda_k}\right)\|\Delta_{k+1}\|^2 \\ &= -\left(\frac{\beta_k - \zeta_k\lambda_k L_f}{2\lambda_k}\right)\|\Delta_{k+1} - \Delta_k\|^2 - \frac{\zeta_k(1-\zeta_k)L_f}{2}\|\Delta_k\|^2 \\ &\quad - \frac{2 - \lambda_k L_f(1-\zeta_k) - \beta_k - \beta_{k+1}}{2\lambda_k}\|\Delta_{k+1}\|^2. \end{aligned} \tag{2.36}$$

Now (2.18) implies that the coefficients of $\|\Delta_{k+1} - \Delta_k\|^2$, $\|\Delta_k\|^2$ and $\|\Delta_{k+1}\|^2$ are nonpositive. Furthermore, from condition (2.18) we see that

$$\frac{2 - \lambda_k L_f(1-\zeta_k) - \beta_k - \beta_{k+1}}{2\lambda_k} \geq \frac{\varepsilon}{2\lambda_k} > \frac{\varepsilon L_f}{4} > 0.$$

Therefore telescoping (2.36) implies

$$\frac{\varepsilon L_f}{4}\sum_{k=1}^{M}\|\Delta_{k+1}\|^2 < V_1 - V_{M+1} < \infty, \quad \forall M \in \mathbb{N}.$$

Thus statement (i) is proven.

Proof of Theorem 3 statement (ii)

Consider (2.31) with the subsequence chosen as $v_k = k$. Clearly this implies statement (ii) since the left-hand side of (2.31) goes to 0 as k goes to ∞.

Proof of Theorem 3 statement (iii)

Assume $\mathcal{X}_F$ is nonempty and x_{v_k} is a subsequence which weakly converges to x'. We note that statement (i) implies $\Delta_k \to 0$, thus from Lemma 11, $x' \in \mathcal{X}_F$. Therefore the second condition of Opial's lemma is satisfied.

We now proceed to show the first condition of Opial's lemma, i.e. for any $x_* \in \mathcal{X}_F$, the limit of $\{\|x_k - x_*\|\}_{k\in\mathbb{N}}$ exists. The key will be to derive a recursion in the form of Lemma 12. This part of the proof has been adapted from [33] which studies the special case where $\zeta_k = 0$ for all k. Fix $x_* \in \mathcal{X}_F$ (which is nonempty by assumption) and let $\varphi_k \triangleq \frac{1}{2}\|x_k - x_*\|^2$. Now using (2.10) we see that

$$\langle x_{k+1} - x_k, x_* - x_{k+1}\rangle = \varphi_k - \varphi_{k+1} - \frac{1}{2}\|\Delta_{k+1}\|^2.$$

Combining this with (2.6) yields

$$\begin{aligned}\varphi_k - \varphi_{k+1} \quad = \quad & \frac{1}{2}\|\Delta_{k+1}\|^2 + \langle x_{k+1} - y_{k+1}, x_* - x_{k+1}\rangle \\ & + \beta_k\langle x_k - x_{k-1}, x_* - x_{k+1}\rangle. \end{aligned} \tag{2.37}$$

Now (2.11) implies

$$-\left(x_{k+1} - y_{k+1} + \lambda_k \nabla f(z_{k+1})\right) \in \lambda_k \partial g(x_{k+1}).$$

On the other hand optimality condition (2.14) implies

$$-\lambda_k \nabla f(x_*) \in \lambda_k \partial g(x_*).$$

Using these facts and (1.2) gives

$$\langle x_{k+1} - y_{k+1} + \lambda_k(\nabla f(z_{k+1}) - \nabla f(x_*)), x_* - x_{k+1}\rangle \geq 0$$

which implies

$$\langle x_{k+1} - y_{k+1}, x_* - x_{k+1}\rangle \geq \lambda_k\langle \nabla f(z_{k+1}) - \nabla f(x_*), x_{k+1} - x_*\rangle. \tag{2.38}$$

Substituting (2.38) into (2.37) yields

$$\begin{aligned}\varphi_{k+1}-\varphi_k &\leq -\frac{1}{2}\|\Delta_{k+1}\|^2-\lambda_k\langle\nabla f(z_{k+1})-\nabla f(x_*),x_{k+1}-x_*\rangle \\ &\quad +\beta_k\langle x_k-x_{k-1},x_{k+1}-x_*\rangle. \qquad (2.39)\end{aligned}$$

Now using (2.10) again

$$\begin{aligned}\langle x_k-x_{k-1},x_{k+1}-x_*\rangle = \quad & \varphi_k-\varphi_{k-1}+\frac{1}{2}\|\Delta_k\|^2 \\ & +\langle x_k-x_{k-1},x_{k+1}-x_k\rangle. \qquad (2.40)\end{aligned}$$

On the other hand using (1.4)

$$\begin{aligned}& \langle\nabla f(z_{k+1})-\nabla f(x_*),x_{k+1}-x_*\rangle \\ \geq\ & \frac{1}{L_f}\left(\|\nabla f(z_{k+1})-\nabla f(x_*)\|^2+L_f\left\langle\nabla f(z_{k+1}-\nabla f(x_*),x_{k+1}-z_{k+1}\right\rangle\right) \\ =\ & \frac{1}{L_f}\|\nabla f(z_{k+1})-\nabla f(x_*)+\frac{L_f}{2}(x_{k+1}-z_{k+1})\|^2-\frac{L_f}{4}\|x_{k+1}-z_{k+1}\|^2 \\ \geq\ & -\frac{L_f}{4}\|x_{k+1}-z_{k+1}\|^2. \qquad (2.41)\end{aligned}$$

Therefore by substituting (2.40) and (2.41) into (2.39) and using (2.35), we get

$$\begin{aligned}\varphi_{k+1}-\varphi_k-\beta_k(\varphi_k-\varphi_{k-1})\leq \quad & -\frac{\zeta_k}{2}\|\Delta_{k+1}\|^2+o_k\|\Delta_k\|^2 \\ & +d_k\langle\Delta_k,\Delta_{k+1}\rangle, \qquad (2.42)\end{aligned}$$

where $\zeta_k=(1-\lambda_k L_f/2)$, $o_k=(\beta_k/2+\zeta_k^2\lambda_k L_f/4)$, and $d_k=\beta_k-\zeta_k\lambda_k L_f/2$. Note that (2.18) implies $\zeta_k\geq\gamma/2>0$, $o_k\in[0,1)$ and $|d_k|<1$. Now if we let $\Theta_k\triangleq\varphi_k-\varphi_{k-1}$ then (4.71) implies

$$\Theta_{k+1}-\beta_k\Theta_k \quad\leq\quad -\frac{\zeta_k}{2}\left\|\Delta_{k+1}-\frac{d_k}{\zeta_k}\Delta_k\right\|^2+\left(o_k+\frac{(d_k)^2}{2\zeta_k}\right)\|\Delta_k\|^2\leq\delta_k,$$

with $\delta_k\triangleq(1+1/\gamma)\|\Delta_k\|^2$. Note that (i) of this Theorem implies $\sum_{k\in\mathbb{N}}\delta_k<\infty$. Now since $\beta_k\leq\overline{\beta}<1$, we can apply Lemma 12, which implies $\lim_{k\to\infty}\|x_k-x_*\|$ exists for any $x_*\in\mathcal{X}_F$. Therefore both conditions of Opial's lemma hold and $\{x_k\}_{k\in\mathbb{N}}$ converges weakly to some minimizer $\hat{x}$. Now repeating (2.6): $y_{k+1}=x_k+\beta_k(x_k-x_{k-1})$ for all $k\in\mathbb{N}$. Therefore for any $h\in\mathcal{H}$, $\langle h,y_{k+1}\rangle=\langle h,x_k\rangle+\langle h,\beta_k(x_k-x_{k-1})\rangle\to\langle h,\hat{x}\rangle$, which proves $y_k\rightharpoonup\hat{x}$. In exactly the same way we can show $z_k\rightharpoonup\hat{x}$ using (2.7).

2.6 Proof of Lemma 1

We commence by proving that the gradient with respect to f is constant at all optimal points. The proof follows by considering [1, Corollary 26.3(vii)]. Note that condition (a) of this Corollary holds trivially because $\mathrm{dom} f = \mathcal{H}$ and $\mathrm{dom}\,\partial g \subseteq \mathrm{dom}\, g$ is nonempty. Now statement (vii) of Corollary 26.3 states the following. Given $x \in \mathcal{X}_F$

$$\langle x - y, \nabla f(x)\rangle + g(x) \leq g(y) \quad \forall y \in \mathcal{H}. \tag{2.43}$$

Consider $x^1, x^2 \in \mathcal{X}_F$, then (2.43) implies $\langle x^1 - x^2, \nabla f(x^1)\rangle + g(x^1) \leq g(x^2)$ and $\langle x^2 - x^1, \nabla f(x^2)\rangle + g(x^2) \leq g(x^1)$. Adding these two together yields

$$\langle \nabla f(x^1) - \nabla f(x^2), x^1 - x^2\rangle \leq 0.$$

From this point on the proof is identical to [1, Prop. 26.10], which implies $\nabla f(x^1) = \nabla f(x^2) \triangleq h_*$. The rest of the Lemma follows by examining the structure of the optimality condition (2.14) for the special case of Prob. (2.2). We refer the reader to [3, Thm 2.1 (ii) and (iii)].

2.7 Proof of Theorem 5

Before proving the theorem, we need several lemmas. The first lemma details the contractive properties of the soft-thresholding operator.

Lemma 13 ([3], Lemma 3.2) *Fix any a and b in $\mathbb{R}$, and $\nu \geq 0$:*

(i) [3, Lemma 3.2 (3.7)] The function S_ν defined in (2.12)–(2.13) is non-expansive. That is, $|S_\nu(a) - S_\nu(b)| \leq |a - b|$.

(ii) [3, Lemma 3.2 statement (5)] If $|b| \geq \nu$ and $\mathrm{sgn}(a) \neq \mathrm{sgn}(b)$ then $|S_\nu(a) - S_\nu(b)| \leq |a - b| - \nu$.

(iii) [3, Lemma 3.2 statement (6)] If $S_\nu(a) \neq 0 = S_\nu(b)$ then $|S_\nu(a) - S_\nu(b)| \leq |a - b| - (\nu - |b|)$.

Next we derive some technical properties of the solution set for Prob. (2.2).

Lemma 14 *For Prob. (2.2) suppose Assumption 1 holds and $\mathcal{X}_F$ is nonempty, $x_* \in \mathcal{X}_F$ and $\lambda > 0$. Then there exists a vector $h_* \in \mathbb{R}^n$ satisfying the conditions of Lemma 1. Furthermore*

$$|x_*^i - \lambda h_*^i| \geq \rho\lambda, \text{ and } \mathrm{sgn}(x_*^i - \lambda h_*^i) = -h_*^i/\rho, \quad \forall i : |h_*^i| = \rho. \tag{2.44}$$

Proof. Recall that $E \triangleq \{i : |h_*^i| = \rho\}$. For $i \in \text{supp}(x_*)$, (2.12)–(2.13) and (2.15) imply

$$0 \neq x_*^i = \text{sgn}\left(x_*^i - \lambda h_*^i\right)\left[|x_*^i - \lambda h_*^i| - \rho\lambda\right]_+. \tag{2.45}$$

Therefore $|x_*^i - \lambda h_*^i| > \rho\lambda$ for all $i \in \text{supp}(x_*)$. On the other hand, if $i \in E \setminus \text{supp}(x_*)$, then $|x_*^i - \lambda h_*^i| = \lambda|h_*^i| = \rho\lambda$. Recall that $\text{supp}(x_*) \subseteq E$. Therefore the first part of (2.44) is proven.

Looking at (2.45) it can be seen that

$$\text{sgn}(x_*^i) = \text{sgn}(x_*^i - \lambda h_*^i), \quad \forall i \in \text{supp}(x_*). \tag{2.46}$$

Note by Lemma 1, if $i \in \text{supp}(x_*)$, then $\text{sgn}(x_*^i) = -h_*^i/\rho$. Else if $i \in E \setminus \text{supp}(x_*)$ then

$$\text{sgn}(x_*^i - \lambda h_*^i) = \text{sgn}(-\lambda h_*^i) = -\text{sgn}(h_*^i) = -\frac{h_*^i}{\rho}. \tag{2.47}$$

since $|h_*^i| = \rho$. Combining (2.46) and (2.47) yields the second part of (2.44).

The final lemma before we proceed with the proof of Theorem 5 is a crucial finite summability result.

Lemma 15 *For Prob. (2.2) suppose Assumption 1 holds. Assume either*

1. $\{\lambda_k\}_{k\in\mathbb{N}}$ *is nondecreasing,* $0 < \lambda_k \leq 1/L_f$*, and* $0 \leq \zeta_k \leq \overline{\zeta} < 1$ *for all* $k \in \mathbb{N}$*, or*
2. $\mathcal{X}_F$ *is nonempty and* $\{\zeta_k, \lambda_k\}_{k\in\mathbb{N}}$ *are chosen according to FISTA-CD in (2.16).*

Furthermore assume the iterates $\{x_k, y_k\}_{k\in\mathbb{N}}$ *of (2.4)–(2.5) satisfy, for all* $k \in \mathbb{N}$*:*

$$\|x_k - x\|^2 \leq \|y_k - x\|^2 - N_k \tag{2.48}$$

for some $x \in \mathbb{R}^n$ *and* $\{N_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$*. Then:* $\sum_{k=1}^{\infty} N_k < \infty$*.*

Proof. Substituting (2.4) into (2.48) yields

$$\begin{aligned} \|x_{k+1} - x\|^2 &\leq \|x_k - x + \zeta_k\Delta_k\|^2 - N_{k+1} \\ &= \|x_k - x\|^2 + \zeta_k^2\|\Delta_k\|^2 + 2\zeta_k\langle x_k - x, \Delta_k\rangle \\ &\quad -N_{k+1}. \end{aligned} \tag{2.49}$$

Let $\varphi_k = \frac{1}{2}\|x_k - x\|^2$ and $\Theta_k = \varphi_k - \varphi_{k-1}$. Using (2.10) we write $\langle x_k - x, \Delta_k\rangle = \varphi_k - \varphi_{k-1} + \frac{1}{2}\|\Delta_k\|^2$. Using this in (2.49) yields

$$\Theta_{k+1} \leq \zeta_k\Theta_k + \delta_k - \frac{1}{2}N_{k+1}, \tag{2.50}$$

where $\delta_k = \frac{1}{2}\zeta_k(1+\zeta_k)\|\Delta_k\|^2$. Note that $0 \leq \delta_k \leq \zeta_k\|\Delta_k\|^2 \leq \|\Delta_k\|^2$.

We first prove the lemma for parameter choice 1. For this parameter choice by Theorem 3(i), $\sum_{k\in\mathbb{N}}\delta_k < \infty$. Let $\underline{\zeta} = \inf_k \zeta_k$ and note that $\underline{\zeta} \in [0, \overline{\zeta}]$. Thus using (2.50):

$$\Theta_{k+1} \quad \leq \quad \overline{\zeta}^k|\Theta_1| + \sum_{j=1}^{k}\overline{\zeta}^{k-j}\delta_j - \frac{1}{2}\sum_{j=1}^{k}\underline{\zeta}^{k-j}N_{j+1}.$$

Therefore, for all $M \in \mathbb{N}$,

$$\begin{aligned}
\varphi_M \quad &= \quad \varphi_0 + \sum_{k=1}^{M}\Theta_k \leq \varphi_0 + \frac{1}{1-\overline{\zeta}}\left(|\Theta_1| + \sum_{k=1}^{M-1}\delta_k\right) - \frac{1}{2}\sum_{k=1}^{M-1}N_{k+1} \\
\implies \quad &\sum_{k=1}^{M-1}N_{k+1} \leq 2\varphi_0 + \frac{2}{1-\overline{\zeta}}\left(|\Theta_1| + \sum_{k=1}^{\infty}\delta_k\right) < \infty, \quad \forall M \in \mathbb{N}.
\end{aligned}$$

Now for parameter choice 2, we proceed as follows. Note that since $x_1 = x_0$, $\Theta_1 = 0$ for this parameter choice. From (2.50), and $\delta_k \leq \zeta_k\|\Delta_k\|^2$, we infer (using the convention: $\prod_{j=a}^{b}\zeta_j = 1$ if $a > b$):

$$\begin{aligned}
\Theta_{k+1} \quad &\leq \quad \left(\prod_{i=1}^{k}\zeta_i\right)\Theta_1 + \sum_{j=1}^{k}\left(\prod_{l=j}^{k}\zeta_l\right)\|\Delta_j\|^2 - \frac{1}{2}\sum_{j=1}^{k}\left(\prod_{l=j+1}^{k}\zeta_l\right)N_{j+1} \\
&\leq \quad \sum_{j=1}^{k}\left(\prod_{l=j}^{k}\zeta_l\right)\|\Delta_j\|^2 - \frac{1}{2}\sum_{j=1}^{k}\zeta_2^{k-j}N_{j+1},
\end{aligned}$$

where we have used the fact that $\zeta_2 < \zeta_k$ for all $k > 2$. Thus for all $M \in \mathbb{N}$

$$\begin{aligned}
\varphi_M \quad &= \quad \varphi_0 + \sum_{k=1}^{M}\Theta_k \\
&\leq \quad \varphi_0 + \sum_{k=1}^{M-1}\sum_{j=1}^{k}\left(\prod_{l=j}^{k}\zeta_l\right)\|\Delta_j\|^2 - \frac{1}{2}\sum_{k=1}^{M-1}N_{k+1}\left(\sum_{j=0}^{M-1-k}\zeta_2^j\right) \\
&\leq \quad \varphi_0 + \sum_{k=1}^{\infty}\sum_{j=1}^{k}\left(\prod_{l=j}^{k}\zeta_l\right)\|\Delta_j\|^2 - \frac{1}{2}\sum_{k=1}^{M-1}N_{k+1}.
\end{aligned}$$

Now by applying (2.17) of Lemma 2 and noting that $\varphi_M \geq 0$, we infer $\sum_{k=1}^{\infty} N_k < \infty$.

We are now ready to prove Theorem 5. Note that parameter choice 1 satisfies the requirements of Corollary 4. Furthermore, by assumption, $\mathcal{X}_F$ is nonempty, thus all conclusions of Corollary 4 hold. For parameter choice 2 (FISTA-CD) we note that both conclusions of Lemma 2 hold.

Throughout the proof, fix an arbitrary $x_* \in \mathcal{X}_F$. We will use the contractive properties of S_ν given in Lemma 13 to construct a recursion in the form of (2.48) of Lemma 15. That lemma allows us to argue that the number of iterations such that (2.21)–(2.22) do not hold is finite.

Proof of (2.21) of Theorem 5

Recall from Lemma 1 there exists a vector h_* such that $\nabla f(x_*) = h_*$ for all $x_* \in \mathcal{X}_F$, and $\text{supp}(x_*) \subseteq E$, where $E \triangleq \{i : |h_*^i| = \rho\}$. Fix $k \in \mathbb{N}$. Now (2.5) and optimality condition (2.15) imply

$$\begin{aligned} & |x_{k+1}^i - x_*^i|^2 \\ = \;& \left|S_{\rho\lambda_k}(y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i) - S_{\rho\lambda_k}(x_*^i - \lambda_k h_*^i)\right|^2 \end{aligned} \qquad (2.51)$$

for all $i \in [n]$, using the notation $[n] \triangleq \{1, 2, \ldots, n\}$. Consider the following condition:

$$\text{sgn}\left(y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i\right) \neq \text{sgn}(x_*^i - \lambda_k h_*^i) \quad \text{for some } i \in E. \qquad (2.52)$$

(Note that $\text{sgn}(x_*^i - \lambda_k h_*^i) = -h_*^i/\rho$ from Lemma 14.) Now recall Lemma 14 implies $|x_*^i - \lambda_k h_*^i| \geq \lambda_k \rho$ for all $i \in E$. Therefore we can apply Lemma 13 (ii) to (2.51) to say the following. If (2.52) holds, then

$$\begin{aligned} |x_{k+1}^i - x_*^i|^2 &\leq \left(|y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i - (x_*^i - \lambda_k h_*^i)| - \rho\lambda_k\right)^2 \\ &\leq \left|y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i - (x_*^i - \lambda_k h_*^i)\right|^2 - \rho^2 \lambda_k^2. \end{aligned} \qquad (2.53)$$

Inequality (2.53) follows because of the following fact:

$$a \geq b \geq 0 \implies (a-b)^2 \leq a^2 - b^2 \qquad (2.54)$$

which applies because

$$|(y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i) - (x_*^i - \lambda_k h_*^i)| \quad \geq |(x_*^i - \lambda_k h_*^i)| \geq \rho\lambda_k > 0 (2.55)$$

where we have used (2.52) and Lemma 14 to prove (4.70).

Now define for $k \in \mathbb{N}$,

$$\mathcal{P}_k \triangleq \{i \in E : \text{sgn}(y_k^i - \lambda_{k-1}\nabla f(y_k)^i) \neq -h_*^i/\rho\}$$

and recall the standard notation $|\mathcal{P}_k|$ for the number of elements in $\mathcal{P}_k$. For all $k \in \mathbb{N}$:

$$\begin{aligned}
\|x_{k+1} - x_*\|^2 &= \sum_{j \in \mathcal{P}_{k+1}} |x_{k+1}^j - x_*^j|^2 + \sum_{j \in [n] \backslash \mathcal{P}_{k+1}} |x_{k+1}^j - x_*^j|^2 \\
&\leq \sum_{j \in \mathcal{P}_{k+1}} \left\{ \left| y_{k+1}^j - \lambda_k \nabla f(y_{k+1})^j - (x_*^j - \lambda_k h_*^j) \right|^2 - \rho^2 \lambda_k^2 \right\} \\
&\quad + \sum_{j \in [n] \backslash \mathcal{P}_{k+1}} \left| y_{k+1}^j - \lambda_k \nabla f(y_{k+1})^j - (x_*^j - \lambda_k h_*^j) \right|^2 \qquad (2.56) \\
&= \|y_{k+1} - \lambda_k \nabla f(y_{k+1}) - (x_* - \lambda_k h_*)\|^2 - \rho^2 \lambda_k^2 |\mathcal{P}_{k+1}| \\
&\leq \|y_{k+1} - x_*\|^2 - \rho^2 \lambda_1^2 |\mathcal{P}_{k+1}|. \qquad (2.57)
\end{aligned}$$

Inequality (2.56) follows from (2.53) and the elementwise nonexpansiveness of $S_{\rho\lambda_k}$ (i.e. Lemma 13(i)). To deduce (2.57), we used the fact that $I - \lambda\nabla f$ is nonexpansive for $0 < \lambda < 2/L_f$ [1, Pro. 4.33], and $\{\lambda_k\}_{k \in \mathbb{N}}$ is nondecreasing. Now (2.57) is in the form of (2.48) of Lemma 15 with $x = x_*$ and $N_k = \rho^2 \lambda_1^2 |\mathcal{P}_k|$. Since we assumed $\rho > 0$ and $\lambda_1 > 0$ it follows that $\sum_{k \in \mathbb{N}} |\mathcal{P}_k| < \infty$ for either parameter choice 1 or 2. This implies $|\mathcal{P}_k|$ is nonzero for only finitely many iterations, thus (2.21) is proved.

Proof of (2.22) of Theorem 5

For E^c nonempty, define $\omega \triangleq \min\{\rho - |h_*^i| : i \in E^c\} \in (0, \rho]$. If E^c is empty, (2.22) is trivially true, therefore assume E^c is nonempty and note that

$$\omega \lambda_k = \min\{\rho\lambda_k - \lambda_k |h_*^i| : i \in E^c\} > 0. \qquad (2.58)$$

Consider $i \in E^c$ (which implies $i \notin \text{supp}(x_*)$). If $x_{k+1}^i \neq 0$, then Lemma 13 (iii), (2.5) and optimality condition (2.15) imply

$$\begin{aligned}
|x_{k+1}^i|^2 &= |S_{\rho\lambda_k}\left(y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i\right) - S_{\rho\lambda_k}\left(-\lambda_k h_*^i\right)|^2 \\
&\leq \left[|y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i + \lambda_k h_*^i| - \left(\rho\lambda_k - \lambda_k |h_*^i|\right) \right]^2 \\
&\leq |y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i + \lambda_k h_*^i|^2 - \left(\rho\lambda_k - \lambda_k |h_*^i|\right)^2 \qquad (2.59) \\
&\leq |y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i + \lambda_k h_*^i|^2 - \omega^2 \lambda_k^2. \qquad (2.60)
\end{aligned}$$

To derive (2.60) we used (2.58). To derive (2.59) we used (2.54) which applies because

$$\begin{aligned} |y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i + \lambda_k h_*^i| &\geq |y_{k+1}^i - \lambda_k \nabla f(y_{k+1})^i| \\ &\quad -\lambda_k |h_*^i| \qquad (2.61) \\ &> \rho\lambda_k - \lambda_k |h_*^i|, \qquad (2.62) \end{aligned}$$

which is greater than 0 by (2.58). Note that (2.61) follows from the identity:

$$|a+b| \geq |a| - |b|, \quad \forall\, a, b \in \mathbb{R}$$

and (2.62) follows from the fact that $0 \neq x_{k+1}^i = S_{\rho\lambda_k}(y_{k+1}^i - \lambda_k \nabla f(y_{k+1}))^i$. Analogously to the definition of $\mathcal{P}_k$, define for all $k \in \mathbb{N}$,

$$\mathcal{Q}_k \triangleq \{i \in E^c : x_k^i \neq 0\}.$$

Thus for all $k \in \mathbb{N}$,

$$\begin{aligned} \|x_{k+1} - x_*\|^2 &= \sum_{j \in [n] \setminus \mathcal{Q}_{k+1}} |x_{k+1}^j - x_*^j|^2 + \sum_{j \in \mathcal{Q}_{k+1}} |x_{k+1}^j|^2 \\ &\leq \sum_{j \in [n] \setminus \mathcal{Q}_{k+1}} |y_{k+1}^j - \lambda_k \nabla f(y_{k+1})^j - (x_*^j - \lambda_k h_*^j)|^2 \\ &\quad + \sum_{j \in \mathcal{Q}_{k+1}} \left\{ |y_{k+1}^j - \lambda_k \nabla f(y_{k+1})^j + \lambda_k h_*^j|^2 - \omega^2 \lambda_k^2 \right\} \\ &\leq \|y_{k+1} - x_*\|^2 - \omega^2 \lambda_1^2 |\mathcal{Q}_{k+1}|. \end{aligned}$$

This recursion is in the form of (2.48) in Lemma 15 with $x = x_*$ and $N_k = \omega^2 \lambda_1^2 |\mathcal{Q}_k|$. Since ω and λ_1 are both greater than 0 we have $\sum_{k \in \mathbb{N}} |\mathcal{Q}_k| < \infty$. Thus $\mathcal{Q}_k$ is nonempty for only finitely many iterations. Note that by (2.4), if x_k^i and x_{k-1}^i are equal to 0, then $y_{k+1}^i = 0$. Thus (2.22) is proved.

2.8 Proof of Theorem 6

We first prove (2.25). From Theorem 5, there exists $K > 0$ such that for all $k > K$, (2.21) and (2.22) hold for either parameter choice 1 or 2. For $i \in E$,

$k > K$, we calculate the quantity

$$\begin{aligned} u^i_{k+1} &\triangleq y^i_{k+1} - \lambda_k \nabla\phi(y^E_{k+1})^i \\ &= y^i_{k+1} - \lambda_k(-h^i_* + \nabla f(y_{k+1})^i) \qquad (2.63) \\ &= y^i_{k+1} - \lambda_k \nabla f(y_{k+1})^i + \rho\lambda_k \left(\frac{h^i_*}{\rho}\right) \\ &= \operatorname{sgn}\left(y^i_{k+1} - \lambda_k \nabla f(y_{k+1})^i\right) \\ &\quad \times(|y^i_{k+1} - \lambda_k \nabla f(y_{k+1})^i| - \rho\lambda_k). \qquad (2.64) \end{aligned}$$

Equation (2.63) follows from $\operatorname{supp}(y_{k+1}) \subseteq E$. Equation (2.64) follows from (2.21). Therefore, for $i \in E$, $k > K$,

$$x^i_{k+1} = S_{\rho\lambda_k}\left(y^i_{k+1} - \lambda_k \nabla f(y_{k+1})^i\right) = \begin{cases} u^i_{k+1} : -h^i_* u^i_{k+1} \geq 0 \\ 0 : \text{else} \end{cases},$$

which proves (2.25). Now (2.21) implies $\operatorname{sgn}(x^i_k) = -h^i_*/\rho$ for all $i \in E$, $k > K$, and $x^i_k \neq 0$. Therefore $-h^i_* x^i_k = \rho|x^i_k|$, for all $i \in E$, $k > K$. Therefore since $x^{E_c}_k = 0$ for $k > K$, $-(h^E_*)^\top x^E_k = \rho\|x_k\|_1$, which implies $F(x_k) = \phi(x^E_k)$.

2.9 Proofs of Sec. 2.3.7

Proof of Proposition 8

Corollary 4 implies that $\lim_{k\to\infty} x_k \triangleq x_*$ exists and $x_* \in \mathcal{X}_F$ for parameter choice 1. On the other hand Lemma 2 implies this is true for parameter choice 2. Theorem 5 states that there exists a finite K such that for $k > K$ (2.21) holds for all $i \in E$, and recall that $\operatorname{supp}(x_*) \subseteq E$. Now since $x^i_k \to x^i_* \neq 0$ for all $i \in \operatorname{supp}(x_*)$, there exists some $K' > 0$ such that for all $k > K'$, $x^i_k \neq 0$. Combining this with (2.5), (2.13) and (2.21) implies that for all $k > \max(K, K')$, and $i \in \operatorname{supp}(x_*)$,

$$\begin{aligned} x^i_k &= \operatorname{sgn}(y^i_k - \lambda_{k-1}\nabla f(y_k)^i)(|y^i_k - \lambda_{k-1}\nabla f(y_k)^i| - \lambda_{k-1}\rho) \\ &= y^i_k - \lambda_{k-1}(\nabla f(y_k)^i - h^i_*). \end{aligned}$$

Proof of Theorem 9

Recall that Prob (2.3) satisfies Assumption 1, therefore all conclusions of Lemma 1 hold. Further recall that $\mathcal{X}_F$ is nonempty for Prob. (2.3). Therefore $\lim_{k\to\infty} x_k \triangleq x_*$ exists and $x_* \in \mathcal{X}_F$ by Corollary 4. Recall that $E = \{i : |h_*^i| = \rho\}$ and also by the strict complementarity assumption: $E = \text{supp}(x_*)$. Proposition 8 proves that there exists $K > 0$ such that for all $k > K$

$$x_k^E = y_k^E - \frac{1}{L_f}(\nabla f(y_k)^E - h_*^E) = y_k^E - \frac{1}{L_f}((A^\top A y_k)^E - (A^\top A x_*)^E).(2.65)$$

On the other hand Theorem 5 proved that there exists $K' > 0$ such that for all $k > K'$, $x_k^{E^c} = y_k^{E^c} = 0$. Therefore for all $k > \max(K, K') \triangleq K''$ both conditions hold. Let $Q = (A_E^\top A_E)$ and $P_{\mathcal{R}(Q)}$ be the orthogonal projector for the range space of Q.

We first consider the part of the error in the nullspace of $P_{\mathcal{R}(Q)}$. Equation (2.65) implies

$$(I - P_{\mathcal{R}(Q)})(x_k^E - x_*^E) = (I - P_{\mathcal{R}(Q)})(y_k^E - x_*^E), \quad \forall k > K''.$$

Combining this with (2.4) implies: $t_{k+1} = (1+\zeta)t_k - \zeta t_{k-1}$ where $t_k = (I - P_{\mathcal{R}(Q)})(y_k^E - x_*^E)$ for all $k > K''$. This is a linear homogeneous recursion with solution:

$$\tilde{t}_M^i = \tilde{t}_0^i + \frac{(\tilde{t}_1^i - \tilde{t}_0^i)(1 - \zeta_M)}{1 - \zeta}, \quad \forall M \in \mathbb{N},$$

where $\tilde{t}_k = t_{k+\lceil K'' \rceil}$. Now $\lim_{M\to\infty} \tilde{t}_M^i = (\tilde{t}_1^i - \zeta \tilde{t}_0^i)/(1-\zeta)$. On the other hand, Thm. 3 (iii) implies $\tilde{t}_M^i \to 0$ as $M \to \infty$. Therefore either $\tilde{t}_M = 0$ for all $M \in \mathbb{N}$ or $\tilde{t}_M^i = \zeta \tilde{t}_{M-1}^i$ for all M. Therefore $(I - P_{\mathcal{R}(Q)})(y_k^E - x_*^E) = (I - P_{\mathcal{R}(Q)})(x_k^E - x_*^E)$ converges to 0 R-linearly with rate ζ.

Next we consider $P_{\mathcal{R}(Q)}(x_k^E - x_*^E)$. Note that Q is symmetric thus $\mathcal{R}(Q) = \mathcal{N}(Q)^\perp$. Thus, for all $k > K''$

$$\begin{aligned} P_{\mathcal{R}(Q)}(x_k^E - x_*^E) &= P_{\mathcal{R}(Q)}(y_k^E - x_*^E) - \frac{1}{L_f} P_{\mathcal{R}(Q)} Q (y_k^E - x_*^E) \\ &= P_{\mathcal{R}(Q)}(y_k^E - x_*^E) - \frac{1}{L_f} Q P_{\mathcal{R}(Q)}(y_k^E - x_*^E). \quad (2.66) \end{aligned}$$

Let $\hat{l}_E$ be the smallest nonzero eigenvalue of Q, which is also the smallest eigenvalue of Q restricted $\mathcal{R}(P_{\mathcal{R}(Q)})$. If $\hat{l}_E = 0$, then $P_{\mathcal{R}(Q)}$ is the all-zero matrix and x_k converges to x_* R-linearly with rate ζ. Assume $\hat{l}_E > 0$. Restating (2.4) we have for all $k \geq K''$: (4.42) holds and

$$P_{\mathcal{R}(Q)} y_{k+1}^E = P_{\mathcal{R}(Q)} x_k^E + \zeta (P_{\mathcal{R}(Q)} x_k^E - P_{\mathcal{R}(Q)} x_{k-1}^E).$$

This is exactly the same recursion as studied in [45, §4.2–4.3] with respect to the sequences $\{P_{\mathcal{R}(Q)}(x_k^E - x_*^E)\}$ and $\{P_{\mathcal{R}(Q)}(y_k^E - x_*^E)\}$. Note that ϕ restricted to $\mathcal{R}(P_{\mathcal{R}(Q)})$ is a strongly-convex quadratic function. By looking at the eigenvalues and eigenvectors of Q restricted to $\mathcal{R}(P_{\mathcal{R}(Q)})$, one can see that Q-linear convergence of $P_{\mathcal{R}(Q)}x_k^E$ is obtained and the rate $(1 - \sqrt{\mu/L_f})^{1/2}$ is achieved by the choice: $\zeta = (1 - \sqrt{\mu/L_f})/(1 + \sqrt{\mu/L_f})$. We refer to [45] for all the details. Note that the rate of x_k is the same as $P_{\mathcal{R}(Q)}x_k^E$ since x_k is zero outside E for $k > K''$ and $(I - P_{\mathcal{R}(Q)})(x_k^E - x_*^E)$ has R-linear convergence to 0 with rate ζ, which is faster than the rate $(1 - \sqrt{\mu/L_f})^{1/2}$. Finally the fact that ϕ is quadratic for this problem gives the objective function rates.

CHAPTER 3

AN INERTIAL METHOD FOR NONCONVEX COMPOSITE PROBLEMS

3.1 Chapter Introduction

In this chapter we are interested in solving the following optimization problem:

$$\min_{x\in\mathbb{R}^n} \Phi(x) = \Phi_1(x) + \Phi_2(x) \tag{3.1}$$

where $\Phi_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is closed and $\Phi_1 : \mathbb{R}^n \to \mathbb{R}$ is differentiable with Lipschitz continuous gradient. This is a composite optimization problem like Problem (2.1) studied in Chapter 2. The difference is that in Problem (3.1) we make no assumption of convexity. We do assume that Φ is *semialgebraic* [53], meaning there exist integers $p, q \geq 0$ and polynomial functions $U_{ij}, W_{ij} : \mathbb{R}^{n+1} \to \mathbb{R}$ such that

$$\{(x,y) : y \geq \Phi(x)\} = \bigcup_{j=1}^{p} \bigcap_{i=1}^{q} \{z \in \mathbb{R}^{n+1} : U_{ij}(z) = 0, W_{ij}(z) < 0\}.$$

Semialgebraic objective functions in the form of (3.1) are widespread in machine learning, image processing, compressed sensing, matrix completion, and computer vision [54, 55, 56, 57, 58, 59, 60]. We will list a few examples below.

In this chapter we focus on the application of Prob. (3.1) to *sparse least-squares* and regression. This problem arises when looking for a sparse solution to a set of underdetermined linear equations. Such problems occur in compressed sensing, computer vision, machine learning and many other related fields. Suppose we observe $y = Ax + b$ where b is noise and wish to recover x which is known to be sparse, however the matrix A is "fat" or poorly conditioned. One approach is to solve (3.1) with Φ_1 a loss function modeling the noise b and Φ_2 a regularizer modeling prior knowledge of x, in this case sparsity. The correct choice for Φ_1 will depend on the noise model and may be nonconvex. Examples of appropriate nonconvex semialgebraic choices for r are the ℓ_0 pseudo-norm, and the smoothly clipped absolute deviation (SCAD) [61]. The prevailing convex choice is the ℓ_1 norm which is also semialgebraic. This results in the lasso problem considered in Chapter 2. SCAD has the advantage over the ℓ_1-norm that it leads to nearly unbiased estimates of large coefficients [61]. Furthermore unlike the ℓ_0 norm SCAD leads to a solution which is continuous in the data matrix A.

In this chapter, much like Chapter 2, we are interested in inertial first-order methods. For nonconvex problems it has been observed that using inertia can help the algorithm escape local minima and saddle points that would capture other first-order algorithms [62, Sec 4.1]. A prominent example of the use of inertia in nonconvex optimization is training neural networks, which goes by the name of *back propagation with momentum* [63].

Over the past decade the Kurdyka–Łojaziewicz (KL) inequality has come to prominence in the optimization community as a powerful tool for studying both convex and nonconvex problems. It is very general, applicable to almost all problems encountered in real applications, and powerful because it allows researchers to precisely understand the local convergence properties of first-order methods. The inequality goes back to [64, 65]. In [66, 67, 68] the KL inequality was used to derive convergence rates of descent-type first-order methods. The KL inequality was used to study convex optimization problems in [69, 70].

Nonconvex optimization has traditionally been challenging for researchers to study since generally they cannot distinguish a local minimum from a global minimum. Nevertheless, for some applications such as empirical risk minimization in machine learning, finding a good local minimum is all that is required of the optimization solver [71, Sec. 3]. In other problems local minima have been shown to be global minima [72].

3.1.1 Chapter Contributions

The main contribution of this chapter is to determine for the first time the local convergence rate of a broad family of inertial proximal gradient splitting methods for solving Prob. (3.1). The family of methods we study includes several algorithms proposed in the literature for which convergence rates are unknown. The family was proposed in [73], where it was proved that the iterates converge to a critical point. However the *convergence rate*, e.g. how fast the iterates converge, was not determined. In fact in [73], local linear convergence was shown under a partial smoothness assumption. In contrast we do not assume partial smoothness and our results are far more general.

We use the KL inequality and show finite, linear, or sublinear convergence, depending on the KL exponent (see Sec. 2). The main inspiration for our work is [68] which studied convergence rates of several *noninertial* schemes using the KL property. However, the analysis of [68] cannot be applied to inertial methods. Our approach is to extend the framework of [68] to the inertial setting. This is done by proving convergence rates of a multistep Lyapunov potential function which upper bounds the objective function. We also prove convergence rates of the iterates and extend a result of [70, Thm. 3.7] to show that our multistep Lyapunov potential has the same KL exponent as the objective function. Finally we include experiments to illustrate the derived convergence rates.

3.2 Preliminaries

In this section we give an overview of the mathematical concepts relevant to this chapter. The *Fréchet subdifferential* of a closed function $\Phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ at a point $x \in \text{dom}(\Phi)$ is defined as

$$\partial^F \Phi(x) \triangleq \left\{ v : \liminf_{z \to x} \left(\Phi(z) - \Phi(x) - \langle v, z - x \rangle \geq 0 \right) \right\}.$$

The (limiting) subdifferential is defined as

$$\partial^L \Phi(x) \triangleq \{ v : \exists x_k \to x, \Phi(x_k) \to \Phi(x), v_k \in \partial^F \Phi(x_k) \to v \}.$$

Note that $\partial^F \Phi(x) \subset \partial^L \Phi(x)$ and $\partial^L \Phi(x)$ is closed. For more details and properties we refer to [53, Sec 2.1]. A necessary (but not sufficient) condition for x to be a minimizer of Φ is $0 \in \partial^L \Phi(x)$. The set of critical points of Φ is $\text{crit}(\Phi) \triangleq \{x : 0 \in \partial^L \Phi(x)\}$. In the case where Φ is convex, $\partial^L \Phi$ coincides with the normal subdifferential $\partial \Phi$ as defined in Section 1.3.

We use the same definition of *proximal operator* as defined in Section 1.3, except we do not require the function to be convex. To repeat, the proximal operator w.r.t. a closed proper function Φ_2 is defined as

$$\text{prox}_{\Phi_2}(x) = \arg\min_{x' \in \mathbb{R}^n} \; \Phi_2(x') + \frac{1}{2} \|x - x'\|^2.$$

Note that, unlike the convex case, this operator is not necessarily single-valued. However it is always a nonempty set.

Definition A function $\Phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is said to have the Kurdyka–Łojaziewicz (KL) property at $x^* \in \text{dom}\, \partial^L \Phi$ if there exists $\eta \in (0, +\infty]$, a neighborhood

U of x^*, and a continuous and concave function $\varphi : [0, \eta) \to \mathbb{R}_+$ such that

(i) $\varphi(0) = 0$,

(ii) φ is C^1 on $(0, \eta)$ and for all $s \in (0, \eta)$, $\varphi'(s) > 0$,

(iii) for all $x \in U \cap \{x : \Phi(x^*) < \Phi(x) < \Phi(x^*) + \eta\}$ the KL inequality holds:

$$\varphi'(\Phi(x) - \Phi(x^*))d(0, \partial^L \Phi(x)) \geq 1. \tag{3.2}$$

If Φ is semialgebraic, then it has the KL property at all points in $\operatorname{dom} \partial^L \Phi$, and $\varphi(t) = \frac{c_\theta}{\theta} t^\theta$ for $\theta \in (0, 1]$.

In the semialgebraic case we will refer to θ as the *KL exponent* (note that some other papers use $1 - \theta$ [70]). For the special case where Φ is smooth, (3.2) can be rewritten as $\|\nabla(\varphi \circ (\Phi(x) - \Phi(x^*))\| \geq 1$, which shows why φ is called a "desingularizing function". The slope of φ near the origin encodes information about the "flatness" of the function about a point, thus the KL exponent provides a way to quantify convergence rates of iterative first-order methods.

For example the 1D function $\Phi(x) = |x|^p$ for $p \geq 2$ has desingluarizing function $\varphi(t) = t^{\frac{1}{p}}$. The larger p, the flatter Φ is around the origin, and the slower gradient-based methods will converge. In general, functions with smaller exponent θ have slower convergence near a critical point [68]. Thus, determining the KL exponent of an objective function holds the key to assessing convergence rates near critical points. Note that for most prominent optimization problems, determining the KL exponent is an open problem. Nevertheless many important examples have been determined recently, such as least-squares and logistic regression with an ℓ_1, ℓ_0, or SCAD penalty [70]. A very interesting recent work showed that for convex functions the KL property is equivalent to an error bound condition which is often easier to check in practice [69].

We now precisely state our assumptions on Problem (3.1), which will be in effect throughout the rest of the chapter.

Assumption 2. (Problem (3.1)). The function $\Phi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is semialgebraic, bounded from below, and has desingularizing function $\varphi(t) = \frac{c_\varphi}{\theta} t^\theta$ where $c_\varphi > 0$ and $\theta \in (0, 1]$. The function $\Phi_2 : \mathbb{R}^n \to \overline{\mathbb{R}}$ is closed, and $\Phi_1 : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous gradient with constant L_{Φ_1}.

3.3 A Family of Inertial Algorithms

We study the family of inertial algorithms proposed in [73]. In what follows $s \geq 1$ is an integer, and $I = \{0, 1, \ldots, s-1\}$.

```
Algorithm 1: Multi-step Inertial Forward-Backward splitting (MiFB)
Require: $x_0 \in \mathbb{R}^n$, $0 < \underline{\gamma} \leq \overline{\gamma} < 1/L_{\Phi_1}$.
  Set $x_{-s} = \ldots = x_{-1} = x_0$, $k = 1$
  repeat
    Choose $0 < \underline{\gamma} \leq \gamma_k \leq \overline{\gamma} < 1/L_{\Phi_1}$, $\{a_{k,0}, a_{k,1}, \ldots\} \in (-1, 1]^s$,
    $\{b_{k,0}, b_{k,1}, \ldots\} \in (-1, 1]^s$.
    $y_{a,k} = x_k + \sum_{i \in I} a_{k,i}(x_{k-i} - x_{k-i-1})$
    $y_{b,k} = x_k + \sum_{i \in I} b_{k,i}(x_{k-i} - x_{k-i-1})$
    $x_{k+1} \in \mathrm{prox}_{\gamma_k \Phi_1}(y_{a,k} - \gamma_k \nabla \Phi_1(y_{b,k}))$
    $k = k + 1$
  until convergence
```

Note the algorithm as stated leaves open the choice of the parameters $a_{k,i}$, $b_{k,i}$, and γ_k. For convergence conditions on the parameters we refer to Section 3.4 and [74, Thm. 1].

The algorithm is very general and covers several inertial algorithms proposed in the literature as special cases. For instance the inertial forward-backward method proposed in [62] corresponds to MiFB with $s = 1$, and $b_{k,0} = 0$. The well-known iPiano algorithm also corresponds to this same parameter choice, however the original analysis of this algorithm assumed r was convex [75]. The *heavy-ball method* is an early and prominent inertial first-order method which also corresponds to this parameter choice when $\Phi_2(x) = 0$. The heavy-ball method was originally proposed for strongly convex quadratic problems but was considered in the context of nonconvex problems in [76]. The analysis of [77] applies to MiFB for the special case when $s = 1$ and $a_{k,0} = b_{k,0}$. However [77] only derived convergence rates of the iterates and not the function values, which are our main interest.[1] Furthermore [77] used a different proof technique to the one used here. This same parameter choice has been considered for convex optimization in [74, 18], albeit without the sharp convergence rates derived here. In both the convex and nonconvex settings, employing inertia has been found to improve either the convergence rate or the quality of the obtained local minimum in several studies [62, 75, 73, 74].

General convergence rates have not been derived for MiFB under noncon-

[1] Note that the objective function is not assumed to be Lipschitz continuous so rates derived for the iterates do not immediately imply rates for the objective.

vexity and semialgebraicity assumptions. The convergence rate of iPiano has been examined in a limited situation where the KL exponent $\theta = 1/2$ in [70, Thm 5.2]. Note that the primary motivation for studying this framework is its generality - allowing our analysis to cover many special cases from the literature. However the case $s = 1$ is the most interesting in practice and corresponds to the most prominent inertial algorithms.

3.4 Convergence Rate Analysis

Throughout the analysis, Assumption 2 is in effect. Before providing our convergence rate analysis, we need a few results from [73].

Theorem 16 *Fix $s \geq 1$ and recall $I = \{0, 1, \ldots, s-1\}$. Fix $\{\gamma_k\}$, $\{a_{k,i}\}$ and $\{b_{k,i}\}$ for $k \in \mathbb{N}$ and $i \in I$. Fix $\xi_1, \xi_2 > 0$ and define*

$$\begin{aligned} \Lambda_k &\triangleq \frac{1 - \gamma_k L_{\Phi_1} - \xi_1 - \xi_2\gamma_k}{2\gamma_k}, \quad \underline{\Lambda} \triangleq \liminf_{k\in\mathbb{N}} \Lambda_k, \\ \Pi_{k,i} &\triangleq \frac{s a_{k,i}^2}{2\gamma_k \xi_1} + \frac{s b_{k,i}^2 L_{\Phi_1}^2}{2\xi_2}, \quad \overline{\Pi}_i \triangleq \limsup_{k\in\mathbb{N}} \Pi_{k,i}, \end{aligned}$$

and $z_k \triangleq (x_k^\top, x_{k-1}^\top, \ldots, x_{k-s}^\top)^\top$ where $\{x_k\}$ is the sequence of iterates generated by MiFB. Define the multi-step Lyapunov function as

$$\Psi(z_k) \triangleq \Phi(x_k) + \sum_{i\in I}\left(\sum_{j=i}^{s-1}\overline{\Pi}_j\right)\|x_{k-i} - x_{k-i-1}\|^2. \tag{3.3}$$

and

$$\xi_3 \triangleq \underline{\Lambda} - \sum_{i\in I}\overline{\Pi}_i > 0. \tag{3.4}$$

If the parameters of MiFB are chosen so that $\xi_3 > 0$ then

(i) for all k, $\Psi(z_{k+1}) \leq \Psi(z_k) - \xi_3\|x_{k+1} - x_k\|^2$,

(ii) for all k, there is a $\sigma > 0$ such that $d(0, \partial^L\Psi(z_k)) \leq \sigma \sum_{j=k+1-s}^{k} \|x_j - x_{j-i-1}\|$,

(iii) If $\{x_k\}$ is bounded there exists $x^* \in crit(\Phi)$ such that $x_k \to x^*$ and $\Phi(x_k) \to \Phi(x^*)$.

Proof Statements (i) and (ii) are shown in [73, Lemma A.5] and [73, Fact (R.2)] respectively. The fact that $\Phi(x_k) \to \Phi(x^*)$ is shown in [73, Lemma A.6]. The fact that $x_k \to x^*$ is the main result of [73, Thm 2.2].

The assumption that $\{x_k\}$ is bounded is standard in the analysis of algorithms for nonconvex optimization and is guaranteed under ordinary conditions such as coercivity. Since the set of semialgebraic functions is closed under addition, Ψ is semialgebraic [78]. We now give our convergence result.

Theorem 17 *Assume the parameters of MiFB are chosen such that $\xi_3 > 0$ where ξ_3 is defined in (3.4), thus there exists a critical point x^* such that $x_k \to x^*$, where $\{x_k\}$ are the iterates of MiFB. Let θ be the KL exponent of Ψ defined in (3.3).*

(a) If $\theta = 1$, then x_k converges to x^* in a finite number of iterations.

(b) If $\frac{1}{2} \leq \theta < 1$, then $\Phi(x_k) \to \Phi(x^*)$ linearly.

(c) If $0 < \theta < 1/2$, then $\Phi(x_k) - \Phi(x^*) = O\left(k^{\frac{1}{2\theta-1}}\right)$.

Proof The starting point is the KL inequality applied to the multi-step Lyapunov function defined in (3.3). Let $z^* \triangleq ((x^*)^\top, \ldots, (x^*)^\top)^\top$. Suppose $\Psi(z_K) = \Psi(z^*)$ for some $K > 0$. Then the descent property of Thm. 1(i), along with the fact that $\Psi(z_k) \to \Psi(z^*)$, implies that $\|x_{K+1} - x_K\| = 0$ and therefore $\Psi(z_k) = \Psi(z^*)$ holds for all $k > K$. Therefore assume $\Psi(z_k) > \Psi(z^*)$. Now since $z_k \to z^*$ and $\Psi(z_k) \to \Psi(z^*)$, there exists $k_0 > 0$ such that for $k > k_0$ (3.2) holds with $f = \Psi$. Assume $k > k_0$. Squaring both sides of (3.2) yields

$$\varphi'^2(\Psi(z_k) - \Psi(z^*))d(0, \partial^L \Psi(z_k))^2 \geq 1, \tag{3.5}$$

Now substituting Thm.1 (ii) into (3.5) yields

$$\sigma^2 \varphi'^2(\Psi(z_k) - \Psi(z^*)) \left(\sum_{j=k+1-s}^{k} \|x_j - x_{j-1}\| \right)^2 \geq 1. \tag{3.6}$$

Now

$$\begin{aligned}
\left(\sum_{j=k+1-s}^{k} \|x_j - x_{j-1}\| \right)^2 &\leq s \sum_{j=k+1-s}^{k} \|x_j - x_{j-1}\|^2 \\
&\leq \frac{s}{\xi_3} \sum_{j=k+1-s}^{k} \left(\Psi(z_{j-1}) - \Psi(z_j)\right) \\
&= \frac{s}{\xi_3} \left(\Psi(z_{k-s}) - \Psi(z_k)\right),
\end{aligned}$$

where in the first inequality we have used the fact that $(\sum_{i=1}^s a_i)^2 \leq s \sum_{i=1}^n a_i^2$, and in the second inequality we have used Thm. 1(i). Substituting this into

(3.6) yields

$$\frac{\sigma^2 s}{\xi_3}\varphi'^2(\Psi(z_k)-\Psi(z^*))\left(\Psi(z_{k-s})-\Psi(z_k)\right)\geq 1,$$

from which convergence rates can be derived by extending the arguments in [68, Thm 4].

Proceeding, let $r_k \triangleq \Psi(z_k)-\Psi(z^*)$, and $C_1 = \frac{\xi_3}{\sigma^2 c_\varphi^2 s}$, then using $\varphi'(t) = c_\varphi t^{\theta-1}$, we get

$$r_{k-s}-r_k \geq C_1 r_k^{2(1-\theta)}. \tag{3.7}$$

If $\theta = 1$, then the recursion becomes $r_{k-s}-r_k \geq C_1, \quad \forall k > k_0$. Since by Theorem 16 (iii), r_k converges, this would require $C_1 = 0$, which is a contradiction. Therefore there exists k_1 such that $r_k = 0$ for all $k > k_1$.

Suppose $\theta \geq 1/2$, then since $r_k \to 0$, there exists k_2 such that for all $k > k_2$, $r_k \leq 1$, and $r_k^{2(1-\theta)} \geq r_k$. Therefore for all $k > k_2$,

$$\begin{aligned} r_{k-s}-r_k \geq C_1 r_k \implies r_k &\leq (1+C_1)^{-1} r_{k-s} \\ &\leq (1+C_1)^{-p_1} r_{k_2}, \end{aligned} \tag{3.8}$$

where $p_1 \triangleq \lfloor \frac{k-k_2}{s} \rfloor$. Note that $p_1 > \frac{k-k_2-s}{s}$. Therefore $r_k \to 0$ linearly. Note that if $\theta = \frac{1}{2}$, $2(1-\theta) = 1$ and (3.8) holds for all $k \geq k_0$.

Finally suppose $\theta < 1/2$. Define $\phi(t) \triangleq \frac{D}{1-2\theta} t^{2\theta-1}$ where $D > 0$, so $\phi'(t) = -Dt^{2\theta-2}$. Now

$$\phi(r_k)-\phi(r_{k-s}) = \int_{r_{k-s}}^{r_k} \phi'(t)dt \quad = \quad D\int_{r_k}^{r_{k-s}} t^{2\theta-2}dt.$$

Therefore since $r_{k-s} \geq r_k$ and $t^{2\theta-2}$ is nonincreasing,

$$\phi(r_k)-\phi(r_{k-s}) \geq D(r_{k-s}-r_k) r_{k-s}^{2\theta-2}.$$

Now we consider two cases.

Case 1: suppose $2r_{k-s}^{2\theta-2} \geq r_k^{2\theta-2}$, then

$$\phi(r_k)-\phi(r_{k-s}) \geq \frac{D}{2}(r_{k-s}-r_k) r_k^{2\theta-2} \geq \frac{C_1 D}{2}, \tag{3.9}$$

where in the second inequality we have used (3.7).

Case 2: suppose that $2r_{k-s}^{2\theta-2} < r_k^{2\theta-2}$. Now $2\theta-2 < 2\theta-1 < 0$, therefore

$(2\theta-1)/(2\theta-2)>0$, thus $r_k^{2\theta-1} > q r_{k-s}^{2\theta-1}$ where $q = 2^{\frac{2\theta-1}{2\theta-2}} > 1$. Thus

$$\begin{aligned}
\phi(r_k)-\phi(r_{k-s}) &= \frac{D}{1-2\theta}\left(r_k^{2\theta-1}-r_{k-s}^{2\theta-1}\right) \\
&> \frac{D}{1-2\theta}(q-1)r_{k-s}^{2\theta-1} \\
&\geq \frac{D}{1-2\theta}(q-1)r_{k_0}^{2\theta-1} \triangleq C_2. \qquad (3.10)
\end{aligned}$$

Thus putting together (3.9) and (3.10) yields $\phi(r_k) \geq \phi(r_{k-s}) + C_3$ where $C_3 = \max(C_2, \frac{C_1 D}{2})$. Therefore

$$\phi(r_k) \quad \geq \quad \phi(r_k) - \phi(r_{k-p_2 s}) \geq p_2 C_3,$$

where $p_2 \triangleq \lfloor \frac{k-k_0}{s} \rfloor$. Therefore

$$r_k \leq \left(\frac{1-2\theta}{D}\right)^{\frac{1}{2\theta-1}} (p_2 C_3)^{\frac{1}{2\theta-1}} \leq C_4 \left(\frac{k-s-k_0}{s}\right)^{\frac{1}{2\theta-1}},$$

where $C_4 = \left(\frac{C_3(1-2\theta)}{D}\right)^{\frac{1}{2\theta-1}}$. To end the proof, note that $\Phi(x_k) \leq \Psi(z_k)$.

In the case where Φ_1 and Φ_2 are also convex, we can use parameter choices specified in [74, Thm. 1].

3.5 Convergence Rates of the Iterates

The convergence rates of $\|x_k - x^*\|$ can also be quantified. To do so we need another result from [73].

Lemma 18 *Recall the notation of Sec. 3.4 which defines $r_k \triangleq \Psi(z_k) - \Psi(z^*)$, where Ψ and z_k are defined in (3.3), and $\{x_k\}$ are the iterates of MiFB. Let $v_k \triangleq \frac{\sigma}{\xi_3}(\varphi(r_k) - \varphi(r_{k+1}))$ where σ is defined in Theorem 16 (ii) and ξ_3 in (3.4). Assume the parameters of MiFB are chosen to so that $\xi_3 > 0$ and $\{x_k\}$ is bounded. Fix $\xi_4 > 0$ so that $\xi_4 < 2/s$. Then there exists a $k_0 > 0$ such that for all $k > k_0$*

$$r_k > 0 \implies \|x_k - x_{k-1}\| \leq \frac{\xi_4}{2}\sum_{j=k-s}^{k-1}\|x_j - x_{j-1}\| + \frac{1}{2\xi_4}v_{k-1}. \qquad (3.11)$$

Proof This inequality is proved on page 14 of [73] as part of the proof of [73, Thm 2.2].

We now state our result.

Theorem 19 *Assume the iterates $\{x_k\}$ of MiFB are bounded and the parameters of MiFB are chosen so that $\xi_3 > 0$ where ξ_3 is defined in (3.4). Let θ be the KL exponent of Ψ defined in (3.3). Then*

(a) If $\theta = 1$, then $x_k = x^*$ after finitely many iterations.

(b) If $\frac{1}{2} \leq \theta < 1$, $x_k \to x^*$ linearly.

(c) If $0 < \theta < \frac{1}{2}$, $\|x_k - x^*\| = O\left(k^{\frac{\theta}{2\theta-1}}\right)$.

Proof Statement (a) follows trivially from the fact that $r_k = 0$ after finitely many iterations, and therefore $\|x_k - x_{k-1}\| = 0$. We proceed to prove statements (b) and (c). As with Theorem 17 the basic idea is to extend the techniques of [68] to allow for the inertial nature of the algorithm. The starting point is (3.11). Fix $K > k_0$. Then

$$\begin{aligned}
\sum_{k\geq K}\|x_k - x_{k-1}\| &\leq \frac{\xi_4}{2}\sum_{k\geq K}\sum_{j=k-s}^{k-1}\|x_j - x_{j-1}\| + \frac{1}{2\xi_4}\sum_{k\geq K} v_{k-1} \\
&\leq \frac{\xi_4 s}{2}\sum_{k\geq K-s}\|x_k - x_{k-1}\| + \frac{1}{2\xi_4}\sum_{k\geq K} v_{k-1}.
\end{aligned}$$

Let $C = \frac{\xi_4 s}{2}$ and note that $0 < C < 1$. Therefore subtracting $C\sum_{k\geq K}\|x_k - x_{k-1}\|$ from both sides yields

$$\sum_{k\geq K}\|x_k - x_{k-1}\| \leq \frac{1}{1-C}\left(C\sum_{k=K-s}^{K-1}\|x_k - x_{k-1}\| + \frac{1}{2\xi_4}\sum_{k\geq K} v_{k-1}\right).$$

Next note that

$$\begin{aligned}
\sum_{k=K-s}^{K-1}\|x_k - x_{k-1}\| &\leq \sqrt{\frac{s}{\xi_3}}\left(\Psi(z_{K-s-1}) - \Psi(z_{K-1})\right)^{1/2} \\
&\leq \sqrt{\frac{s}{\xi_3}}\sqrt{r_{K-s-1}}.
\end{aligned}$$

Let $C' \triangleq C\sqrt{\frac{s}{\xi_3}}$ then using $\sum_{k\geq K} v_{k-1} = \frac{\sigma}{\xi_3}\varphi(r_{K-1})$,

$$\begin{aligned}
\sum_{k\geq K}\|x_k - x_{k-1}\| &\leq \frac{1}{1-C}\left(C'\sqrt{r_{K-s-1}} + \frac{\sigma}{\xi_3}\varphi(r_{K-1})\right) \\
&\leq \frac{1}{1-C}\left(C'\sqrt{r_{K-s-1}} + \frac{\sigma}{\xi_3}\varphi(r_{K-s-1})\right),
\end{aligned}$$

where in the second inequality we used the fact that r_k is nonincreasing and φ is a monotonic increasing function. Thus using the triangle inequality and the fact that $\lim_k \|x_k - x^*\| = 0$,

$$\|x_K - x^*\| \leq \sum_{k \geq K} \|x_k - x_{k-1}\| \leq \frac{1}{1-C}\left(C'\sqrt{r_{K-s-1}} + \frac{\sigma}{\xi_3}\varphi(r_{K-s-1})\right).$$

Hence if $r_k \to 0$ linearly, then so does $\|x_k - x^*\|$, which proves (b). On the other hand if $0 < \theta < 1/2$, for k sufficiently large we see that $\|x_k - x^*\| = O(\varphi(r_{k-s-1}))$, which proves statement (c).

3.6 KL Exponent of the Lyapunov Function

We now extend the result of [70, Thm 3.7] so that it covers the Lyapunov function defined in (3.3).

Theorem 20 *Let $s \geq 1$, and consider*

$$\Psi^{(s)}(x_1, x_2, \ldots, x_s) \triangleq \Phi(x_1) + \sum_{i=1}^{s-1} \pi_i \|x_{i+1} - x_i\|^2, \tag{3.12}$$

where $\pi_i \geq 0$. If Φ has KL exponent $\theta \in (0, 1/2]$ at $\bar{x}$ then $\Psi^{(s)}$ has KL exponent θ at $[\bar{x}, \bar{x}, \ldots, \bar{x}]^\top$.

Proof Before commencing, note that if Φ has desingularizing function $\varphi(t) = \frac{c_\theta}{\theta} t^\theta$, the KL inequality (3.2) can be written in the equivalent form:

$$d(0, \partial^L \Phi(x))^{\frac{1}{1-\theta}} \geq c_\theta^{-1}(\Phi(x) - \Phi(x^*)).$$

We now show that this bound holds for the Lyapunov function in (3.12).

The key is to notice the recursive nature of the Lyapunov function. In particular for all $s \geq 2$

$$\begin{aligned} \Psi^{(s)}(x_s^1) &= \Psi^{(s-1)}(x_1^{s-1}) \\ &+ \pi_{s-1}\|x_{s-1} - x_s\|^2, \end{aligned}$$

with $\Psi^{(1)}(x_1^1) \triangleq \Phi(x_1)$, and $x_1^s \triangleq [x_1^\top, \ldots, x_s^\top]^\top$. Since Φ has KL exponent θ at $\bar{x}$, $\Psi^{(1)}$ has KL exponent θ at $\bar{x}$. We will prove the following inductive step for $s \geq 2$: If $\Psi^{(s-1)}$ has KL exponent θ (with constant c_θ^{-1}) at $\bar{x}_1^{s-1}$, then $\Psi^{(s)}$ has KL exponent θ at $\bar{x}_1^s$ where $\bar{x}_1^s \triangleq [\bar{x}, \bar{x}, \ldots, \bar{x}]^\top$ where $\bar{x}$ is repeated s times.

Proceeding, for $s \geq 2$ assume $x_1, x_2, \ldots, x_s$ are such that $\|x_s - x_{s-1}\| \leq 1$

and the KL inequality (3.2) applies to $\Psi^{(s-1)}$ at $\bar{x}_1^s$. Then

$$\partial^L \Psi^{(s)}(\bar{x}_s^1) \ni \begin{pmatrix} u_1^{s-2} \\ u_{s-1} \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ \pi_{s-1}(x_{s-1} - x_s) \\ \pi_{s-1}(x_s - x_{s-1}) \end{pmatrix},$$

where $(u_1^{s-2}, u_{s-1}) \in \partial^L \Psi^{(s-1)}(x_1^{s-2}, x_{s-1})$. Therefore

$$\begin{aligned}
& d(0, \partial^L \Psi^{(s)}(x_1^s))^{\frac{1}{1-\theta}} \\
\overset{(a)}{\geq}\ & C_1 \Bigg(\inf_{(u_1^{s-2}, u_{s-1}) \in \partial^L \Psi^{(s-1)}(x_1^{s-2}, x_{s-1})} \|u_{s-1}^c\|^{\frac{1}{1-\theta}} \\
& + \|u_{s-1} + \pi_{s-1}(x_{s-1} - x_s)\|^{\frac{1}{1-\theta}} + \|\pi_{s-1}(x_s - x_{s-1})\|^{\frac{1}{1-\theta}} \Bigg) \\
\overset{(b)}{\geq}\ & C_1 \Bigg(\inf_{(u_{s-1}^c, u_{s-1}) \in \partial^L \Psi^{(s-1)}(x_1^{s-2}, x_{s-1})} \|u_{s-1}^c\|^{\frac{1}{1-\theta}} + \eta_1 \|u_{s-1}\|^{\frac{1}{1-\theta}} \\
& - \eta_2 \|\pi_{s-1}(x_{s-1} - x_s)\|^{\frac{1}{1-\theta}} + \|\pi_{s-1}(x_s - x_{s-1})\|^{\frac{1}{1-\theta}} \Bigg) \\
\overset{(c)}{\geq}\ & C_2 \Bigg(\inf_{(u_{s-1}^c, u_{s-1}) \in \partial^L \Psi^{(s-1)}(x_1^{s-2}, x_{s-1})} \|u_{s-1}^c\|^{\frac{1}{1-\theta}} + \|u_{s-1}\|^{\frac{1}{1-\theta}} \\
& + \frac{\pi_{s-1} c_\theta^{-1}}{2} \|x_s - x_{s-1}\|^{\frac{1}{1-\theta}} \Bigg) \\
\overset{(d)}{\geq}\ & C_3 \Bigg(\inf_{(u_{s-1}^c, u_{s-1}) \in \partial^L \Psi^{(s-1)}(x_1^{s-2}, x_{s-1})} \left\| \begin{matrix} u_{s-1}^c \\ u_{s-1} \end{matrix} \right\|^{\frac{1}{1-\theta}} \\
& + \frac{\pi_{s-1} c_\theta^{-1}}{2} \|x_s - x_{s-1}\|^{\frac{1}{1-\theta}} \Bigg) \\
\overset{(e)}{\geq}\ & C_3 c_\theta^{-1} \Big(\Psi^{(s-1)}(x_1^{s-1}) - \Psi^{(s-1)}(\bar{x}_1^{s-1}) \\
& + \frac{\pi_{s-1}}{2} \|x_s - x_{s-1}\|^{\frac{1}{1-\theta}} \Big) \\
\overset{(f)}{\geq}\ & C_3 c_\theta^{-1} \Big(\Psi^{(s-1)}(x_1^{s-1}) - \Psi^{(s-1)}(\bar{x}_1^{s-1}) \\
& + \frac{\pi_{s-1}}{2} \|x_s - x_{s-1}\|^2 \Big) \\
=\ & C_3 c_\theta^{-1} \Big(\Psi^{(s)}(x_1^s) - \Psi^{(s)}(\bar{x}_1^s) \Big).
\end{aligned}$$

Now (a) and (d) follow from [70, Lemma 2.2], and (b) follows from [70, Lemma 3.1]. Next (c) follows because $\eta_1 > 0$, $0 < \eta_2 < 1$, and we have decreased C_2 to compensate for factoring out these coefficients. Further (e) follows by the KL inequality. Finally (f) follows because $\|x_s - x_{s-1}\| \leq 1$ and $(1-\theta)^{-1} \in (1, 2]$. Since $\Psi^{(1)}$ has KL exponent θ at $\bar{x}$, then so does $\Psi^{(s)}$

at $[\bar{x}, \bar{x}, \dots, \bar{x}]^\top$ (of length s) for all $s \geq 2$, which concludes the proof.

This theorem says that when the KL exponent of the objective function Φ is known, the same exponent applies to the Lyapunov function in (3.3). This allows us to exactly determine the convergence rate of MiFB via Theorems 17 and 19.

3.7 Numerical Results

3.7.1 One-Dimensional Polynomial

This simple experiment verifies the convergence rates derived in Theorem 17 for MiFB. Consider the one-dimensional function $\Phi_1(x) = |x|^p$ for $p > 2$. Use $\Phi_2(x) = +\infty$ if $|x| > 1$ and 0 otherwise. The proximal operator is simple projection and Φ_1 is $p(p-1)$-smooth on this set. The function $\Phi = h + r$ is semialgebraic with $\varphi(t) = pt^{1/p}$, i.e. $\theta = 1/p$. Therefore Theorem 2 predicts $O\left(k^{-\frac{p}{p-2}}\right)$ rates for MiFB, which is verified in Fig. 3.1 for three parameter choices in the cases $p = 4, 18$. For simplicity we ignore constants and focus on the sublinear order. For $p \leq 4$ this convergence rate is better than that of Nesterov's accelerated method [32], for which only $O(1/k^2)$ worst-case rate is known. Faster rates are achievable due to the additional knowledge of the KL exponent.

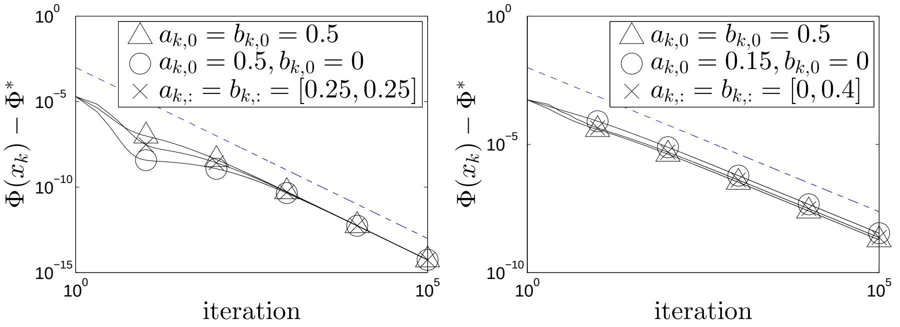

Figure 3.1: (Left) $p = 4$, (Right) $p = 18$, $\Phi^* = 0$. The dotted line is the slope of the predicted $O\left(k^{-\frac{p}{p-2}}\right)$ rate (i.e. ignoring constants). Note $a_{k,:} \triangleq [a_{k,0}, a_{k,1}]$ and these are log-log plots.

3.7.2 SCAD and ℓ_1 regularized Least-Squares

We solve Prob. (3.1) with $\Phi_1(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $\Phi_2(x) = \sum_{i=1}^{n} \Phi_0(x^i)$ where Φ_2 is: 1) the SCAD regularizer defined as

$$\Phi_0(x^i) = \begin{cases} \lambda|x^i| & \text{if } |x^i| \leq \lambda \\ -\frac{|x^i|^2 - 2a\lambda|x^i| + \lambda^2}{2(a-1)} & \text{if } \lambda < |x^i| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |x^i| > a\lambda, \end{cases}$$

and 2) the absolute value $\Phi_0(x^i) = \lambda|x^i|$ leading to the ℓ_1-norm. In both cases the proximal operator w.r.t. Φ_2 is easily computed. It was shown in [70, Sec. 5.2] and [69, Lemma 10] that both of these objective functions are KL functions with exponent $\theta = 1/2$.

We choose $A \in \mathbb{R}^{500\times 1000}$ having i.i.d. $\mathcal{N}(0, 10^{-4})$ entries, and $b = Ax_0$, where $x_0 \in \mathbb{R}^{1000}$ has 50 nonzero $\mathcal{N}(0,1)$-distributed entries. For SCAD we use $a = 5$ and $\lambda = 1$ and for the ℓ_1 norm we use $\lambda = 0.01$.

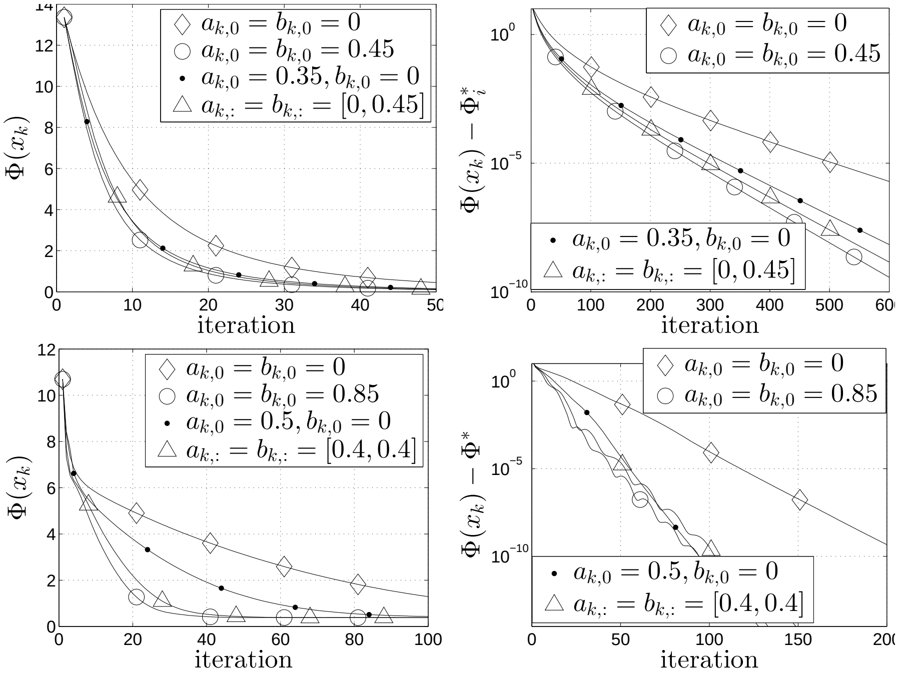

Figure 3.2: (Top Left) Plot of $\Phi(x_k)$ for SCAD least-squares. (Top Right) Plot of $\Phi(x_k) - \Phi_i^*$ with a logarithmic y-axis for SCAD least-squares. As SCAD least-squares is a nonconvex problem, each of the four considered parameter choices may converge to a different objective function value Φ_i^* for $i = 1, 2, 3, 4$. (Bottom Left) Plot of $\Phi(x_k)$ for ℓ_1 least-squares. (Bottom Right) Plot of $\Phi(x_k) - \Phi^*$ with a logarithmic y-axis for ℓ_1 least-squares.

We consider four valid parameter choices. To isolate the effect of inertia, all choices used the same randomly chosen starting point and fixed stepsize,

$\gamma_k = 0.1/L_{\Phi_1}$ for SCAD and $\gamma_k = 1/L_{\Phi_1}$ for ℓ_1. The inertial parameters were chosen so that $u_3 > 0$ (defined in (3.4)) for SCAD and to satisfy [74, Thm. 1] for the ℓ_1 problem. The two figures on the right corroborate Theorem 2 in that all considered parameter choices converge linearly to their limit, which was estimated by using the attained objective function value after 1000 iterations. For the nonconvex SCAD this is a new result. For ℓ_1-regularized least squares, inertial methods have been shown to achieve *local* linear convergence in [74, 37] under additional strict complementarity or restricted strong convexity assumptions. However, our analysis, which is based on the KL inequality, does not explicitly require these additional assumptions, as the objective function always has a KL exponent of 1/2 [69, Lemma 10]. Furthermore our result proves *global* linear convergence, in that the KL inequality (3.2) holds for all k, implying $k_0 = 1$ in (3.5) and (3.8) holds for all k. In addition the two left figures show that the inertial choices appear to provide acceleration relative to the standard non-inertial choice which for SCAD is a new observation. This does not conflict with Theorem 2 which only shows that both non-inertial and inertial methods will converge *linearly*, however the convergence factor may be different. Estimating the factor is beyond the scope of this paper and we leave it for future work. Finally we mention that FISTA [79] and other Nesterov-accelerated methods [32] are not applicable to SCAD as it is nonconvex.

CHAPTER 4

FASTER SUBGRADIENT METHODS UNDER AN ERROR BOUND

4.1 Chapter Introduction

4.1.1 Motivation and Background

In this chapter we consider the problem

$$\min_{x \in \mathcal{C}} h(x), \tag{4.1}$$

where $\mathcal{H}$, as in Chapter 2, is a Hilbert space, $h : \mathcal{H} \to \mathbb{R}$ is a convex and closed function, and $\mathcal{C}$ is a convex, closed, and nonempty subset of $\mathcal{H}$. We do not assume h is smooth or strongly convex. Solving Problem (4.1) arises in many applications such as image processing, machine learning, compressed sensing, statistics, and computer vision [5, 80, 81, 82, 83].

As in the previous chapters, we are interested in first-order methods for solving this problem. Specifically, we focus on the class of *subgradient methods*, which were first studied in the 1970s [84, 85]. Since then, these methods have been used extensively in nonsmooth convex optimization because of their simplicity, and low-complexity [84, 85, 86, 87, 88, 89]. However in general these methods have a slow worst-case convergence rate of $h(\hat{x}_k) - \min_{x \in \mathcal{C}} h(x) \leq O(1/\sqrt{k})$ after k subgradient evaluations for a particular averaged point $\hat{x}_k$. In this chapter we show how a structural assumption for Problem (4.1) that is commonly satisfied in practice yields faster variants of the subgradient method.

The structural assumption we consider is the *Hölder error bound* (throughout referred to as either HEB or HEB(c, θ)). We assume that h satisfies

$$h(x) - h^* \geq c d(x, \mathcal{X}_h)^{\frac{1}{\theta}}, \quad \forall x \in \mathcal{C},$$

for some $\theta \in (0, 1]$ and $c > 0$, where $h^* = \min_{x \in \mathcal{C}} h(x)$, and $\mathcal{X}_h \triangleq \{x \in \mathcal{C} : h(x) = h^*\}$ is the solution set (assumed to be nonempty). In general, an "error bound" is an upper bound on the distance of a point to the optimal set by some residual function. The study of error bounds has a long tradition in optimization, sensitivity analysis, systems of inequalities, projection methods, and convergence rate estimation [90, 91, 92, 93, 69, 94, 95, 96, 97, 98, 99, 100, 101, 102] In recent years there has been much renewed interest in the topic. HEB is often referred to as the *Lojaziewicz error bound* [103]. HEB is also related to the KL inequality utilized in Chapter 3. In fact in [69] it was shown that the KL inequality is equivalent to HEB for CCP functions.

There are three main motivations for studying the behavior of algorithms for problems satisfying HEB. Firstly HEB holds for many problems arising in various applications. In fact for a semialgebraic function HEB is guaranteed to hold on any compact set for some θ and c [69]. Secondly, many algorithms have been shown to achieve significantly faster convergence behavior when HEB is satisfied. Thirdly, under HEB it has been possible to develop even faster methods.

The two most common instances of HEB in practice are $\theta = 1/2$ and $\theta = 1$. The case $\theta = 1/2$ is often referred to as the *quadratic growth condition* (QG) [100]. The case $\theta = 1$ is often referred to by saying the function has *weakly sharp minima* (WS) [99]. If the minimum is unique, then it is simply a sharp minimum. In this chapter we will also refer to this case by saying that the function is weakly sharp. There are also a small number of applications where $\theta \neq 1/2$ or 1, such as L_p regression with $p \neq 1, 2$.

Due to its prevalence in applications, many recent papers have studied QG (the $\theta = 1/2$ case). QG has been used to show a *linear* convergence rate of the objective function values for various algorithms that would otherwise only guarantee sublinear convergence [104, 101, 105, 106, 92, 107, 100]. Many papers have discovered connections between QG and other error bounds and conditions known in the literature. Most importantly it was shown in [100, Appendix A] that for convex functions, QG is equivalent to the Luo-Tseng error bound [97], the *Polyak-Lojaziewicz* condition [100], and the *restricted secant inequality* [95].

Weakly sharp functions (i.e. $\theta = 1$) have been studied in many papers, for example [99, 98, 96, 84, 87, 108, 109, 110, 53]. For such functions [98] showed that the proximal point method converges to a minimum in a *finite* number of iterations. This is interesting because these methods would otherwise only guarantee an $O(1/k)$ rate.

4.1.2 Applications satisfying HEB

Strongly and uniformly convex functions.
A uniformly convex function satisfies [111] for some $\mu_{uc} > 0$ and $d \geq 2$

$$h(y) \geq h(x) + \langle g, y - x \rangle + \frac{\mu_{uc}}{2} \|y - x\|^d \quad \forall x, y \in \mathcal{H}, g \in \partial h(x). \tag{4.2}$$

This corresponds to strong convexity when $d = 2$ which is the most important special case. For a minimizer x^* in the interior of $\mathcal{C}$, $0 \in \partial h(x^*)$. Substituting $g = 0$ into (4.2) yields HEB with $\theta = 1/d$. Applications with $d > 2$ include L_d norm regression (discussed below) and polynomial convex optimization [90].

Least squares and Logistic Regression.
The paper [100] showed that functions of the form $h(x) = h_0(Ax)$ where h_0 is strongly convex and A is a matrix satisfy QG. This includes the ubiquitous least-squares objective. Logistic regression is in the form $h(x) = h_0(Ax)$, however h_0 is only strictly convex. Nevertheless, it is strongly convex on any bounded set.

Lasso (ℓ_1 regularized Least-squares). The ℓ_1-regularized least squares problem considered in Chapter 2 was shown in [69, Lemma 10] to satisfy HEB on the set $\{x : \|x\|_1 \leq R\}$ for sufficiently large R. QG is also shown to be locally satisfied by the group lasso penalized least-squares and logistic regression in [91, Theorem 2].

Composite Optimization The paper [92] considers the problem

$$\min_{x \in \mathcal{H}} h_0(Ax) + P(x)$$

where h is strongly convex on any bounded set and P is polyhedral or the group lasso penalty. Rather surprisingly, they showed in [92] that this function satisfies a local version of QG. The result also applies when P is the nuclear norm so long as a strict complementarity condition is satisfied.

$_d$ **Norm Regression Estimators**
The goal of linear regression is to estimate a vector $\beta_{L_d} \in \mathbb{R}^n$ given a noisy version of its linear measurements $y = X^\top \beta_{L_d} + e$ where e is an unknown noise term. If e conforms to a Gaussian distribution, then the least squares estimate is the maximum likelihood estimator. If the noise is not Gaussian, then the performance of the least squares estimator can be significantly degraded. The L_d estimator with $d \neq 2$ has been considered as an alternative

[112, 113, 114]. It is given by

$$\arg\min_{\beta_{L_d}} \sum_{i=1}^{m} |X(i)^\top \beta_{L_d} - y_i|^d \tag{4.3}$$

for $d \geq 1$, where $X(i)$ is the ith column of X. The case $d = 2$ corresponds to least squares, and $d = 1$ to least absolute deviation. Other choices of d have been considered in [112, 113, 114]. It is not hard to see that (4.3) satisfies the KL inequality given in Chapter 3 with $\theta = 1/d$. Therefore by [69, Thm 5] it satisfies HEB with $\theta = 1/d$.

Polyhedral Convex Optimization.

Suppose that the function h in Problem (4.1) has a polyhedral epigraph (i.e. is piecewise linear), then Problem (4.1) is called a polyhedral convex optimization (PCO) problem. In this case, [109] showed that WS is satisfied globally. Many applications are instances of PCO. For instance both the hinge loss used in SVM classification and the ℓ_1 loss/regularizer used in robust regression are polyhedral. Linear programming is PCO. Another very important application is submodular optimization. The Lovász extension is a convex relaxation for submodular optimization problems which is PCO [115]. Finally note that the sum of polyhedral functions is polyhedral.

4.1.3 Subgradient Methods under HEB

There were a few early works that studied the subgradient method under conditions related to HEB with $\theta = 1$. In [84, Thm 2.7, Sec. 2.3], Shor proposed a geometrically decaying stepsize which obtains a linear convergence rate under a condition equivalent to the function being WS. The stepsize depends on explicit knowledge of the error bound constant c, a bound on the subgradients, and the initial distance $d(x_1, \mathcal{X}_h)$. Goffin [85] extended the analysis of [84] to a slightly more general notion than HEB.[1] Rosenburg [86] extended these results to constrained problems. In [108], Polyak showed that the method still converges linearly when the subgradients are corrupted by bounded, deterministic noise.

The paper [87] also considers functions satisfying HEB with $\theta = 1$ with (deterministically) noisy subgradients. For constant stepsizes, they show convergence of $\liminf h(x_k)$ to h^* plus a tolerance level depending on noise. For diminishing stepsizes, $\liminf h(x_k)$ actually converges to h^* despite the noise. However [87] does not discuss *convergence rates*, which is the topic of

[1] Our analysis in this chapter also holds for Goffin's condition number; see Sec. 4.2.5.

the current chapter.

The authors of [109] introduced the *restarted subgradient method* RSG for when h satisfies HEB. The method implements a predetermined number of averaged subgradient iterations with a constant stepsize and then restarts the averaging and uses a new, smaller stepsize. The authors show that after $O(\epsilon^{2(\theta-1)}\log\frac{1}{\epsilon})$ iterations the method is guaranteed to find a point such that $h(x_k) - h^* \leq \epsilon$. For $\theta = 1$ this is a logarithmic iteration complexity. This improves the iteration complexity of the classical subgradient method which is $O(\epsilon^{-2})$. RSG has another advantage that the dependence of the iteration complexity on the initial distance to the solution set (or the initial objective function gap) is logarithmic.

The recent paper [93] extends RSG to stochastic optimization. In particular they provide a similar restart scheme that can also handle stochastic subgradient calls, and guarantees $h(x) - h^* \leq \epsilon$ with high probability. The iteration complexity is the same as for RSG, up to constants. However, this constant is large leading to a large number of inner iterations, making it difficult to implement the method in practice.

For WS functions, the paper [110] introduced a method similar to RSG except it does not require averaging at the end of each constant stepsize phase. The method also obtains a logarithmic iteration complexity in the $\theta = 1$ case.

The paper [116] is concerned with a two-person zero-sum game equilibrium problem with a linear payoff structure. The authors show that finding the solution to the equilibrium problem is equivalent to a WS minimization problem. Using this fact, they derive a method based on Nesterov's smoothing technique with logarithmic iteration complexity. This is superior to the $O(1/\epsilon)$ of standard Nesterov smoothing. Connections between our results and [116] are discussed in Section 4.5.1.

The work [117] studies stochastic subgradient descent under the assumption that the function satisfies WS locally and QG globally. They show a faster convergence rate of the iterates to a minimizer, both in expectation and with high probability, than is known under the classical analysis.

The work [118] proposes a new subgradient method for functions satisfying a similar condition to HEB but with h^* replaced by a strict lower bound on h^*. Like RSG, this algorithm has a logarithmic dependence on the initial distance to the solution set. However it still obtains an $O(1/\epsilon^2)$ iteration complexity, which is the same as the classical subgradient method.

In [119, 120] Renegar presented a framework for converting a convex conic program to a general convex problem with an affine constraint, to which projected subgradient methods can be applied. He further showed how this can be applied to general convex optimization problems, such as Prob. (4.1), by representing them as a conic problem. For the special case where the objective and constraint set is polyhedral, one of the subgradient methods proposed by Renegar has a logarithmic iteration complexity [119, Cor. 3.4]. The main drawback of this method is that it requires knowledge of the optimal value, h^*. It also requires a point in the interior of the constraint set. Similarly the stepsizes proposed in Thm. 2 of [27, Sec 5.3.] and [88, Prop. 2.11] depend on exact knowledge of h^* and also obtain a logarithmic iteration complexity under WS.

In recent times, convergence analyses for the subgradient method have focused on the objective function rather than the distance of the iterates from the optimal set. However in the early period of development, there were many works focusing on the distance (e.g. [88, 84, 108, 85]). The subgradient method is not a descent method with respect to function values, however it is with respect to the distances to the optimal set. Thus the distance is a natural metric to study for the subgradient method. Furthermore, for some applications, the distance to the solution set arguably matters more than the objective function value. For example in machine learning, the objective function is only a surrogate for the actual objective of interest – expected prediction error.

Without further assumptions, [27, p. 167–168] showed that the convergence rate of the distance of the iterates of the subgradient method to the optimal set can be made arbitrarily slow. This is true even for smooth convex problems. In this case, gradient descent with a constant stepsize obtains an $O(1/k)$ *objective function* convergence rate, however the iterates can be made to converge arbitrarily slowly to a minimizer. In this chapter, HEB allows us to derive less pessimistic convergence rates for the distance to the optimal set.

4.1.4 Chapter Contributions

Define the *standard subgradient method* as

$$x_{k+1} = P_{\mathcal{C}}(x_k - \alpha_k g_k) : \quad \forall k \geq 1, g_k \in \partial h(x_k),\ x_1 \in \mathcal{C}, \tag{4.4}$$

where the choice of the *stepsize* α_k is not specified. The projection onto $\mathcal{C}$ is defined as $P_{\mathcal{C}}$. Recall the definition of the subgradient of h at x [1, Def.

16.1]:

$$\partial h(x) \triangleq \{g \in \mathcal{H} : h(y) \geq h(x) + \langle g, y - x\rangle, \forall y \in \mathcal{H}\}.$$

Despite the long history of analysis of subgradient methods discussed in the previous section, the simplest stepsize choices for (4.4) have not been studied for objective functions satisfying HEB. These are the constant stepsize, $\alpha_k = \alpha$, and the nonsummable decaying stepsize, $\alpha_k = \alpha_1 k^{-p}$ for $p \in (0, 1]$. This brings us to our contributions in this chapter.

Firstly we determine the convergence rate of a constant stepsize choice in the subgradient method which previously had only been determined for the special case of $\theta = 1/2$ (see [88, Prop. 2.4]). Interestingly, *for any* $\theta \in (0, 1]$ the method obtains a linear convergence rate for $d(x_k, \mathcal{X}_h)$ up to a specific tolerance level of order $O(\alpha^\theta)$.

Secondly, we determine the convergence rate of decaying polynomial stepsize choices. Previously, these results had only been obtained for the case where $\theta = 1/2$. For $\theta = 1$ the paper [117] obtains an asymptotic convergence rate for $p = 1$ with an additional global QG assumption. The big advantage of the nonsummable stepsizes is that, for $\theta \geq 1/2$, they require no information about the problem's parameters in order to guarantee convergence. In contrast, we show that for $\theta > 1/2$ summable stepsizes can obtain much faster rates with additional information. For instance summable stepsizes require an upper bound on the initial distance to the solution set, otherwise convergence is impossible.

We frame our convergence rates in terms of $d(x_k, \mathcal{X}_h)$ because this quantity arises naturally in our analysis. If the rate of convergence of $h(\hat{x}_k)$ to h^* is known for some sequence $\hat{x}_k$, a naive estimate of the rate of convergence of $d(\hat{x}_k, \mathcal{X}_h)$ can be obtained via the HEB. For example, the classical analysis of the subgradient method leads to the rate

$$h(\hat{x}_k) - h^* = O(k^{-\frac{1}{2}}),$$

where $\hat{x}_k$ is a specific average of the previous iterates and $\alpha_k = O(1/\sqrt{k})$ [89]. Combining this with HEB yields

$$d(\hat{x}_k, \mathcal{X}_h) = O(k^{-\frac{\theta}{2}}).$$

This rate is slower than the result of our specialized analysis. For example, we show that with the proper choice of p and α_1, the subgradient method with decaying stepsize can obtain the convergence rate

$$d(x_k, \mathcal{X}) \leq O(k^{-\frac{\theta}{2(1-\theta)}}), \quad \forall \theta < 1.$$

It can be seen that the absolute value of the exponent is a factor $1/(1-\theta)$ larger in our analysis.

Our third major contribution is a new "descending staircase" stepsize choice for the subgradient method (DS-SG). The method achieves the same convergence rate as the best decaying stepsize for $\theta < 1$. In addition for the case $\theta = 1$ it achieves linear convergence. Unlike the methods of [119, 120] and [47, Exercise 6.3.3], our proposal does not require h^*. The methods of [110, 84, 85] have a similar complexity for $\theta = 1$ but cannot handle $\theta < 1$. The method RSG of [109] obtains the same iteration complexity but requires averaging. Averaging is disadvantageous in applications where the solution is sparse (or low rank) because it can spoil this property [121]. In Section 4.5.1 we discuss in more detail why averaging can be disadvantageous. The method retains the same iteration complexity even when the subgradients are corrupted, provided the noise is small relative to the sharpness constant c.

DS-SG and our proposed decaying stepsize require knowledge of the constant c in HEB which can be hard to estimate in practice. This motivates us to develop our final major contribition: a "doubling trick" for the descending staircase stepsize which does not require c and still obtains the same iteration complexity up to a small constant. The competing methods of [109, 110, 84, 85][2] all require knowledge of c.

In summary, our contributions under HEB are as follows:

1. We show that the subgradient method with a constant stepsize obtains linear convergence for $d(x_k, \mathcal{X}_h)$ to within a region of the optimal set for all $\theta \in (0, 1]$.

2. We compute nonasymptotic convergence rates for both nonsummable and summable decaying stepsizes under HEB for all $\theta \in (0, 1]$.

3. We develop a new "Descending Stairs" stepsize with iteration complexity $O(\epsilon^{1-\frac{1}{\theta}})$ when $\theta < 1$ and $\ln \frac{1}{\epsilon}$ when $\theta = 1$ for finding a point such that $d(x_k, \mathcal{X})^2 \leq \epsilon$. We also develop an adaptive variant which does not need c but retains the same iteration complexity up to a small constant.

[2] The authors of [109] proposed an adaptive method which does not require c, however the analysis only works for $\theta < 1$.

4.2 The Key Recursion

4.2.1 Optimality Condition and Assumptions

If 0 is in the strict relative interior of $\mathcal{C} - \mathrm{dom}(f)$ then the solution set $\mathcal{X}_h$ of Problem (4.1) is characterized by the optimality condition [1, Prop. 26.5]

$$\mathcal{X}_h = \{x : 0 \in \partial h(x) + N_{\mathcal{C}}(x)\},$$

where $N_{\mathcal{C}}(x)$ is the normal cone of $\mathcal{C}$ at x. Note that we don't explicitly use this optimality criterion anywhere in our analysis and we only include it for completeness.

For Prob. (4.1), throughout the chapter we will assume that $\mathcal{C} \subseteq \mathrm{dom}(\partial h)$, so that for any query point $x \in \mathcal{C}$ it is possible to find a $g \in \partial h(x)$. If h is convex and closed, the solution set $\mathcal{X}_h = \{x : h(x) = h^*\}$ is convex and closed [1]. Following are the precise assumptions we will use throughout the chapter.

Assumption 3. (Problem (4.1)). Assume $\mathcal{C}$ is convex, closed, and nonempty. Assume h is convex, closed, and satisfies HEB(c, θ). Assume $\mathcal{X}_h$ is nonempty. Assume $\mathcal{C} \subseteq \mathrm{dom}(\partial h)$. Assume there exists G such that $\|g\| \leq G$ for all $g \in \partial h(x)$ and $x \in \mathcal{C}$. Let $\kappa \triangleq G/c$.

4.2.2 The Recursion under HEB

In this section we derive the crucial recursion which describes the evolution of the error $d(x_k, \mathcal{X}_h)^2$ for the iterates of the standard subgradient method under HEB. The same recursion has been derived many times before for the special cases $\theta = \{1/2, 1\}$ (e.g. [85, 84, 88]). For the point x_k let x_k^* be the unique projection of x_k onto $\mathcal{X}_h$.

Proposition 21 *Suppose Assumption 3 holds. Then for all $k \geq 1$ for the iterates $\{x_k\}$ of (4.4)*

$$d(x_{k+1}, \mathcal{X}_h)^2 \quad \leq \quad d(x_k, \mathcal{X}_h)^2 - 2\alpha_k c(d(x_k, \mathcal{X}_h)^2)^{\frac{1}{2\theta}} + \alpha_k^2 G^2.$$

Proof For $k \geq 1$,

$$\begin{aligned}
d(x_{k+1}, \mathcal{X}_h)^2 &= \|x_{k+1} - x_{k+1}^*\|^2 \\
&\leq \|x_{k+1} - x_k^*\|^2 \\
&\leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k \langle g_k, x_k - x_k^* \rangle + \alpha_k^2 \|g_k\|^2 \\
&\leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k (h(x_k) - f_*) + \alpha_k^2 G^2 \\
&\leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k c (d(x_k, \mathcal{X}_h)^2)^{\frac{1}{2\theta}} + \alpha_k^2 G^2.
\end{aligned}$$

In the first inequality, we used the fact that x_{k+1}^* is the closest point to x_{k+1} in $\mathcal{X}_h$. In the second inequality, we used the nonexpansive properties of the projection operator. In the third, we used the convexity of h and in the final inequality we used the error bound.

Let $e_k \triangleq d(x_k, \mathcal{X}_h)^2$ and $\gamma = \frac{1}{2\theta} \in [\frac{1}{2}, +\infty)$ then for all $k \geq 1$

$$0 \leq e_{k+1} \leq e_k - 2\alpha_k c e_k^\gamma + \alpha_k^2 G^2. \tag{4.5}$$

The main effort of our analysis is in deriving convergence rates for this recursion for various stepsizes.

4.2.3 Deterministic Noise in the Subgradient when $\theta = 1$

For the weakly sharp case ($\theta = 1$), the subgradient method exhibits resilience to bounded noise. This has been observed in [87, 108]. Suppose that at each iteration we have access to a noisy subgradient:

$$\tilde{g}_k = g_k + r_k : g_k \in \partial h(x_k), \|r_k\| \leq R$$

and as before the method iterates for all $k \geq 0$

$$x_{k+1} = P_{\mathcal{C}}(x_k - \alpha_k \tilde{g}_k).$$

Repeating the analysis of Sec. 4.2.2

$$\begin{aligned}
d(x_{k+1}, \mathcal{X}_h)^2 &= \|x_{k+1} - x_{k+1}^*\|^2 \\
&\leq \|x_{k+1} - x_k^*\|^2 \\
&\leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k \langle \tilde{g}_k, x_k - x_k^* \rangle + \alpha_k^2 \|\tilde{g}_k\|^2 \\
&\leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k (h(x_k) - h^*) - 2\alpha_k \langle r_k, x_k - x_k^* \rangle \\
&\quad +2\alpha_k^2 (R^2 + G^2) \\
&\leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k d(x_k, \mathcal{X}_h)(c - R) + 2\alpha_k^2 (R^2 + G^2),
\end{aligned}$$

where in the third inequality we have used $\|\tilde{g}_k + r_k\|^2 \leq 2\|\tilde{g}_k\|^2 + 2\|r\|^2$. We see that this is exactly the same recursion as (4.5) with the error bound constant c replaced by $c-R$, and G^2 replaced by $2(G^2+R^2)$. Thus, if $R < c$, all of the results presented throughout for $\theta = 1$ hold with a new error bound constant $\tilde{c} = c - R$, and bound on the subgradients $\tilde{G}^2 = 2(G^2 + R^2)$. In particular this refers to Theorems 24, 25, 26, 27, and 30.

4.2.4 Incremental Subgradient Methods

Suppose $h(x) = \sum_{i=1}^m h_i(x)$. Such objective functions which are a finite sum of terms often arise in machine learning in the guise of *empirical risk minimization* [122]. For such problems the *incremental* subgradient method can be used [88]. This method proceeds by computing the subgradient with respect to each individual function h_i in a fixed order. More precisely the method proceeds for $k \geq 1$ with $x_1 \in \mathcal{C}$ as

$$\begin{aligned} x_{k+1} &= \psi_{m,k} && (4.6)\\ \psi_{i,k} &= P_{\mathcal{C}}(\psi_{i-1,k} - \alpha_k g_{i,k}), g_{i,k} \in \partial h_i(\psi_{i-1,k}), \;\; i = 1, \ldots, m && (4.7)\\ \psi_{0,k} &= x_k. && (4.8) \end{aligned}$$

This method has been analyzed extensively in [88].

Proposition 22 ([88]) *Suppose Assumption 3 holds. Then for all $k \geq 1$ for the iterates of (4.6)–(4.8)*

$$d(x_{k+1}, \mathcal{X})^2 \leq d(x_k, \mathcal{X})^2 - 2\alpha_k c d(x_k, \mathcal{X})^{\frac{1}{\theta}} + \alpha_k^2 m^2 G^2.$$

This is the same as the main recursion we analyze in (4.5) with G^2 replaced by m^2G^2. Thus all our results in the following sections apply to the incremental subgradient method (4.6)–(4.8) with this change in constants.

4.2.5 Goffin's Condition Number

Goffin [85] discussed a condition number for quantifying the convergence rate of subgradient methods. The condition number is a generalization of the ordinary notion defined for a smooth strongly convex function as the ratio of the Lipschitz constant of the gradient to the strong convexity parameter. In contrast Goffin's condition number requires neither smoothness or strong convexity. The condition number is also more general than Shor's eccentricity measure [84]. The condition number for a convex function h is defined as

$$\mu_h = \inf\left\{\frac{\langle u, x - x_p^*\rangle}{\|u\|\|x - x_p^*\|} : x \in \mathcal{C}\backslash\mathcal{X}_h, u \in \partial h(x), x_p^* = \text{proj}_{\mathcal{X}_h}(x)\right\}. \tag{4.9}$$

By convexity and the Cauchy-Schwarz inequality $0 \leq \mu_h \leq 1$. Goffin showed that if h satisfies HEB(c, θ) with $\theta = 1$ and $\|g\| \leq G$ for all $g \in \partial h(x), x \in \mathcal{C}$, then it satisfies (4.9) with

$$\mu_h \geq \frac{c}{G} = \frac{1}{\kappa}$$

which proves that functions satisfying (4.9) with $\mu_h > 0$ are more general than weakly sharp functions.

Our results for $\theta = 1$ throughout this chapter can be extended to functions satisfying (4.9) with $\mu_h > 0$ if we make the slight modification to the subgradient method.

Lemma 23 ([85]) *Let $\{x_k\}$ be a sequence satisfying*

$$x_{k+1} = P_{\mathcal{C}}\left(x_k - \alpha_k \frac{g_k}{\|g_k\|}\right) : \forall k \geq 1, g_k \in \partial h(x_k), x_1 \in \mathcal{C}. \tag{4.10}$$

If $\mathcal{X}_h$ is nonempty and h is CCP and satisfies (4.9) with $\mu_h > 0$ then for all $k \geq 1$

$$d(x_{k+1}, \mathcal{X}_h)^2 \leq d(x_k, \mathcal{X}_h)^2 - 2\alpha_k \mu_h d(x_k, \mathcal{X}_h) + \alpha_k^2.$$

This is the same recursion as (4.5) with $G = 1$, $\theta = 1$, and $c = \mu_h$. Thus all the results derived in this chapter for HEB with $\theta = 1$ can be derived for the scheme (4.10) applied to functions satisfying (4.9) so long as c is replaced by μ_h and $G = 1$. Also note that Lemma 23 does not require that the subgradients are uniformly bounded over $\mathcal{C}$.

4.3 Constant Stepsize

Consider the projected subgradient method with *constant*, or fixed, stepsize given in Algorithm FixedSG. This is often used in practice especially for stochastic problems. Previously it was shown that if $\theta = 1/2$ then this method achieves linear convergence to within a region of the solution set

Algorithm 2: (FixedSG)

```
Require: $K > 0$, $\alpha > 0$, $x_1 \in \mathcal{C}$
for $k = 1, 2, \ldots, K$ do
    $x_{k+1} = P_{\mathcal{C}}\left(x_k - \alpha_k g_k\right) : \quad g_k \in \partial h(x_k)$
end for
return $x_{k+1}$
```

[88, 100]. Rather suprisingly, we show in the next theorem that linear convergence to within a certain region of $\mathcal{X}_h$ occurs for any $\theta \in (0, 1]$.

Theorem 24 *Suppose Assumption 3 holds. Let* $e_* = \left(\frac{\alpha G^2}{2c}\right)^{2\theta}$.

1. *For all* $k \geq 1$ *the iterates of FixedSG satisfy*

$$d(x_k, \mathcal{X})^2 \leq \max\left\{d(x_1, \mathcal{X})^2, e_* + \alpha^2 G^2\right\}.$$

2. *If* $0 < \theta \leq \frac{1}{2}$ *then for all* $k \geq 2$ *the iterates of FixedSG satisfy*

$$d(x_k, \mathcal{X}_h)^2 - e_* \leq q_1^{k-1}(d(x_1, \mathcal{X}_h)^2 - e_*). \tag{4.11}$$

where

$$q_1 = \left(1 - \frac{1}{\theta}\alpha c e_*^{\frac{1-2\theta}{2\theta}}\right). \tag{4.12}$$

If additionally

$$0 < \alpha < 2^{\frac{1-2\theta}{2(1-\theta)}} G^{\frac{2\theta-1}{1-\theta}} c^{\frac{\theta}{\theta-1}} \tag{4.13}$$

then $q_1 \in (-1, 1)$.

3. *If* $d(x_k, \mathcal{X}_h)^2 \leq D$ *for all* k *for the iterates of FixedSG,* $\frac{1}{2} \leq \theta \leq 1$, *and*

$$0 < \alpha < \frac{2\theta D^{1-\frac{1}{2\theta}}}{c}, \tag{4.14}$$

then for all $k \geq 2$

$$d(x_k, \mathcal{X}_h)^2 - e_* \leq \max\{q_2^{k-1}(d(x_1, \mathcal{X}_h)^2 - e_*), e_* + \alpha^2 G^2\} \tag{4.15}$$

where

$$q_2 = 1 - \frac{\alpha c D^{\frac{1}{2\theta}-1}}{\theta} \in (-1, 1).$$

Proof Recall our notation $e_k = d(x_k, \mathcal{X}_h)^2$ and let $\gamma = \frac{1}{2\theta}$. Returning to the main recursion (4.5) derived in Prop. 21 and replacing the stepsize with a constant yields

$$0 \leq e_{k+1} \leq e_k - 2\alpha c e_k^{\gamma} + \alpha^2 G^2 \tag{4.16}$$

where $\gamma \geq \frac{1}{2}$. We would like to derive the convergence rate of $e_k - e_*$, where $e_* = \left(\frac{\alpha G^2}{2c}\right)^{\frac{1}{\gamma}}$ is the the only fixed point of this recursion, which is derived by setting $e_k = e_{k+1} = e_*$. The key is to write the recursion (4.5) as

$$e_{k+1} - e_* \leq e_k - e_* - 2\alpha c(e_k^{\gamma} - e_*^{\gamma}). \tag{4.17}$$

<u>**Boundedness:**</u>

We first prove e_k is bounded. Considering (4.17) we see that if $e_k \geq e_*$ then $e_{k+1} \leq e_k$. On the other hand, if $e_k \leq e_*$, then (4.16) yields $e_{k+1} \leq e_k + \alpha^2 G^2 \leq e_* + \alpha^2 G^2$. Therefore

$$e_{k+1} \leq \max\{e_k, e_* + \alpha^2 G^2\} \leq \max\{e_1, e_* + \alpha^2 G^2\}.$$

<u>Case 1</u>: $\boldsymbol{\theta \leq \frac{1}{2}}$.

For $\theta \leq \frac{1}{2}$, $\gamma \geq 1$ and by the convexity of t^{γ},

$$e_k^{\gamma} - e_*^{\gamma} \quad \geq \quad \gamma e_*^{\gamma-1}(e_k - e_*).$$

Therefore

$$e_{k+1} - e_* \leq (1 - 2\alpha c \gamma e_*^{\gamma-1})(e_k - e_*).$$

Thus so long as

$$-1 < 1 - 2\alpha c \gamma e_*^{\gamma-1} < 1, \tag{4.18}$$

linear convergence is guaranteed. Simplifying (4.18)

$$\begin{aligned} 2\alpha c \gamma e_*^{\gamma-1} \quad &< \quad 2 \\ &\implies \quad c\gamma\alpha\left(\frac{\alpha G^2}{2c}\right)^{\frac{\gamma-1}{\gamma}} < 1 \\ &\implies \quad \alpha < \left(\frac{1}{\gamma} G^{\frac{2(1-\gamma)}{\gamma}} 2^{\frac{\gamma-1}{\gamma}} c^{-\frac{1}{\gamma}}\right)^{\frac{\gamma}{2\gamma-1}} \end{aligned}$$

which implies (4.11), (4.12), and (4.13).

<u>Case 2</u>: $\boldsymbol{\theta \geq \frac{1}{2}}$.

For $\theta \in [\frac{1}{2}, 1]$, $\gamma \in [\frac{1}{2}, 1]$, which implies by concavity

$$e_*^\gamma - e_k^\gamma \le \gamma e_k^{\gamma-1}(e_* - e_k).$$

Therefore

$$e_k^\gamma - e_*^\gamma \ge \gamma e_k^{\gamma-1}(e_k - e_*).$$

Substituting this inequality into (4.17) yields

$$e_{k+1} - e_* \quad \le \quad e_k - e_* - 2\alpha c\gamma e_k^{\gamma-1}(e_k - e_*).$$

Now if $e_k \ge e_*$ then using $e_k \le D$ implies

$$e_{k+1} - e_* \quad \le \quad (1 - 2\alpha c\gamma D^{\gamma-1})(e_k - e_*) = q_2(e_k - e_*).$$

Thus so long as

$$1 > 1 - 2\alpha c\gamma D^{\gamma-1} > -1$$

(which is implied by (4.14)), we have $q_2 \in (-1, 1)$. On the other hand if $e_k \le e_*$ then $e_{k+1} \le e_* + \alpha^2 G^2$. Thus for all $k \ge 1$

$$e_{k+1} - e_* \le \max\left\{q_2(e_k - e_*), e_* + \alpha^2 G^2\right\}.$$

Iterating this recursion and using the fact that $q_2 \in (-1, 1)$ yields (4.15).

4.4 Iteration Complexity for Constant Stepsize

Using the results of the previous section we can derive the iteration complexity of a constant stepsize for finding a point such that $d(x_k, \mathcal{X}_h)^2 \le \epsilon$. Rather surprisingly, this section shows that restarting is not necessary for $\theta \le \frac{1}{2}$. This is because for $\theta \le \frac{1}{2}$ the iteration complexity for a constant stepsize is equal to the complexity of RSG derived in [109]. However, for $\theta > \frac{1}{2}$, restarting does improve the iteration complexity. In Section 4.5 we propose a new descending stairs stepsize which significantly accelerates the constant stepsize choice. For $\frac{1}{2} < \theta \le 1$ RSG also outperforms the constant stepsize.

The basic idea in the following theorem is to pick $\alpha = O(\epsilon^{\frac{1}{2\theta}})$, so that e_* defined in Theorem 24 is equal to ϵ. Then the iteration complexity can be determined from the linear convergence rate of $d(x_k, \mathcal{X}_h)^2$ to e_*.

Theorem 25 *Suppose Assumption 3 holds. Choose $\epsilon > 0$ and set*

$$\alpha = \frac{2c\epsilon^{\frac{1}{2\theta}}}{G^2}. \tag{4.19}$$

1. *If $0 < \theta \leq \frac{1}{2}$,*

$$\begin{aligned} 0 < \epsilon &\leq \left(\frac{\theta\kappa^2}{2}\right)^{\frac{\theta}{1-\theta}}, &(4.20)\\ K &\triangleq \frac{1}{2}\theta\kappa^2 \ln\left(\frac{d(x_1,\mathcal{X}_h)^2}{\epsilon}\right)\epsilon^{1-\frac{1}{\theta}}, \end{aligned}$$

then for the iterates of FixedSG, $d(x_{k+1},\mathcal{X}_h)^2 \leq 2\epsilon$ for all $k \geq K$.

2. *If $\frac{1}{2} < \theta \leq 1$,*

$$\begin{aligned} D &\geq 2\max\{d(x_1,\mathcal{X}_h)^2,\epsilon\} &(4.21)\\ 0 < \epsilon &\leq \min\left\{\left(\frac{\kappa^2}{4}\right)^{\frac{\theta}{1-\theta}}, \left(\frac{\theta\kappa^2}{2}\right)^{2\theta} D^{2\theta-1}\right\}, \textit{ and} &(4.22)\\ K &\triangleq \frac{1}{2}\theta\kappa^2 D^{1-\frac{1}{2\theta}} \ln\left(\frac{d(x_1,\mathcal{X}_h)^2}{\epsilon}\right)\epsilon^{-\frac{1}{2\theta}}, &(4.23) \end{aligned}$$

then for the iterates of FixedSG, $d(x_{k+1},\mathcal{X}_h)^2 \leq 3\epsilon$ for all $k \geq K$.

Proof We consider two cases: $\theta \leq 1/2$ and $\theta > 1/2$.
Case 1: $\theta \leq \frac{1}{2}$.
From Theorem 24, the convergence factor in the constant stepsize case is $q_1 = 1 - \frac{\alpha c}{\theta} e_*^{\frac{1}{2\theta}-1}$ where $e_* = \left(\frac{\alpha G^2}{2c}\right)^{2\theta}$. Recall the notation $e_k = d(x_k,\mathcal{X}_h)^2$. From Theorem 24 we know that for all $k \geq 1$

$$e_{k+1} - e_* \leq q_1^k(e_1 - e_*)$$

which implies

$$e_{k+1} - e_* \leq |q_1|^k e_1.$$

This means that

$$\ln(\max\{0, e_{k+1} - e_*\}) \leq k\ln|q_1| + \ln e_1.$$

Thus $e_{k+1} - e_* \leq \epsilon$ is implied by

$$k\ln|q_1| + \ln e_1 \leq \ln\epsilon \iff k \geq \frac{\ln\frac{e_1}{\epsilon}}{\ln\frac{1}{|q_1|}}.$$

so long as $|q_1| < 1$. Now we want $e_* = \epsilon$, which requires

$$\left(\frac{\alpha G^2}{2c}\right)^{2\theta} = \epsilon \iff \alpha = \frac{2c\epsilon^{\frac{1}{2\theta}}}{G^2}.$$

Now if ϵ satisfies (4.20) then $q_1 > 0$. Thus

$$\ln q_1 = \ln\left(1 - \frac{\alpha c}{\theta} e_*^{\frac{1}{2\theta}-1}\right) \leq -\frac{\alpha c}{\theta} e_*^{\frac{1}{2\theta}-1} \iff \ln\frac{1}{q_1} \geq \frac{\alpha c}{\theta} e_*^{\frac{1}{2\theta}-1}.$$

Therefore if $\alpha = \frac{2c\epsilon^{\frac{1}{2\theta}}}{G^2}$ and

$$k \geq \frac{\theta \ln\frac{e_1}{\epsilon}}{\alpha c e_*^{\frac{1}{2\theta}-1}} = \frac{\theta G^2 \ln\frac{e_1}{\epsilon}}{2c^2 \epsilon^{\frac{1}{\theta}-1}} = \frac{\theta G^2 \ln\frac{e_1}{\epsilon}}{2c^2} \epsilon^{1-\frac{1}{\theta}}$$

then

$$e_{k+1} \leq 2\epsilon.$$

<u>Case 2</u>: $\boldsymbol{\theta > \frac{1}{2}}$.

As before, $\alpha = \frac{2c\epsilon^{\frac{1}{2\theta}}}{G^2}$ which implies $e_* = \epsilon$. First note that by Part 1 of Theorem 24,

$$\begin{aligned} d(x_k, \mathcal{X}_h)^2 &\leq \max\{d(x_1, \mathcal{X}_h)^2, e_* + \alpha^2 G^2\} \\ &= \max\left\{d(x_1, \mathcal{X}_h)^2, \epsilon + \frac{4c^2}{G^2}\epsilon^{\frac{1}{\theta}}\right\} \\ &\leq \max\{d(x_1, \mathcal{X}_h)^2, 2\epsilon\} \\ &\leq D \end{aligned}$$

for all $k \geq 1$, where we used (4.22). Recalling (4.15) we see that for all $k \geq 1$

$$e_{k+1} \leq \max\{e_* + q_2^k(d(x_1, \mathcal{X}_h)^2 - e_*), 2e_* + \alpha^2 G^2\}. \tag{4.24}$$

Consider the first argument to the max in (4.24). This case is the same as Case 1 for $\theta \leq 1/2$, except for a different convergence factor. The convergence factor is

$$q_2 = 1 - \frac{\alpha c}{\theta} D^{\frac{1}{2\theta}-1}$$

which is greater than 0 (and less than 1) if ϵ satisfies (4.22). Thus

$$\ln q_2 = \ln\left(1 - \frac{\alpha c}{\theta} D^{\frac{1}{2\theta}-1}\right) \leq -\frac{\alpha c}{\theta} D^{\frac{1}{2\theta}-1} \implies \ln\frac{1}{q_2} \geq \frac{\alpha c}{\theta} D^{\frac{1}{2\theta}-1}.$$

Therefore if

$$k \geq \frac{\theta G^2 D^{1-\frac{1}{2\theta}}}{2c^2} \ln\left(\frac{e_1}{\epsilon}\right) \epsilon^{-\frac{1}{2\theta}}$$

then the first argument to max in (4.24) is upper bounded by 2ϵ.

Now consider the second argument to the max in (4.24), which is

$$2e_* + \alpha^2 G^2 = 2\epsilon + \alpha^2 G^2 = 2\epsilon + \frac{4c^2}{G^2}\epsilon^{1/\theta} \leq 3\epsilon.$$

where we have used again (4.20).

The upper bounds on ϵ given in (4.20) and (4.22) are typically mild in practice because the ratio G/c is at least equal to one, and we are interested in ϵ being small. Theorem 25 shows that, in terms of $d(x_k, \mathcal{X}_h)$, there is no theoretical advantage in restarting for $\theta \leq \frac{1}{2}$. This is because [109] showed that the restart method requires $O(\epsilon'^{2(\theta-1)})$ iterations (suppressing constants and a $\ln\frac{1}{\epsilon}$ factor) to achieve $h(x) - h^* \leq \epsilon'$. Now using the error bound in order to guarantee $d(x_k, \mathcal{X}_h)^2 \leq \epsilon$, we need $h(x) - h^* \leq \epsilon' = \epsilon^{\frac{1}{2\theta}}$. Using this in the iteration complexity from [109] yields an iteration complexity of $O(\epsilon^{1-\frac{1}{\theta}})$, which is the same as the constant stepsize for $\theta \leq 1/2$. For $\theta > \frac{1}{2}$ restarting has better dependence on ϵ, especially as $\theta \to 1$. However, for $\theta = 1/2$, the constant stepsize depends on $\ln d(x_1, \mathcal{X})$ and has the same dependence on ϵ. This remarkable property makes it preferable to the more sophisticated restart methods in this case.

The comparison with the classical result for the subgradient method is as follows. It is easy to show that for the subgradient method with a constant stepsize α

$$\frac{1}{k}\sum_{i=1}^{k}(h(x_i) - h^*) \leq \frac{d(x_1, \mathcal{X}_h)^2}{2\alpha k} + \frac{\alpha}{2}G^2.$$

Setting $\alpha = \epsilon^{1/2\theta}/(2G^2)$ and

$$k \geq \frac{G^2 d(x_1, \mathcal{X}_h)^2 \epsilon^{-1/\theta}}{2}$$

implies $h(x_k^{av}) - h^* \leq \epsilon^{1/2\theta}$ where $x_k^{av} = \frac{1}{k}\sum_{i=1}^{k} x_i$. Now using the error bound this yields $d(x_k, \mathcal{X}_h)^2 \leq \epsilon$. With respect to ϵ, this classical iteration complexity is clearly worse than the result of Theorem 24 for all $\theta \in (0, 1]$. Furthermore, the dependence on $d(x_1, \mathcal{X}_h)$ is worse. For $\theta \leq 1/2$, the fixed stepsize depends on $\ln d(x_1, \mathcal{X}_h)$, whereas the classical stepsize has iteration complexity which depends linearly on $d(x_1, \mathcal{X}_h)$.

We note that as $\theta \to 0$ the iteration complexity can be made arbitrarily large. This is not suprising, as it has been proved in [27, p. 167-168] that the convergence rate of $x_k \to x^*$ can be made arbitrarily bad for gradient methods. In fact it was shown there that for any decreasing sequence $\{\epsilon_k\}$, there exists a smooth convex function with domain in $\mathbb{R}$ such that for the iterates x_k of gradient descent $x_k \geq \epsilon_k$, for all k. Despite this, the convergence rate of the *function values*, $h(x_k) \to h^*$ is no worse than $O(1/k)$ for any smooth convex function.

4.5 A "Descending Stairs" Stepsize with Better Complexity for $\theta > 1/2$

In this section we propose a "descending stairs" stepsize for the subgradient method which obtains a better iteration complexity than the fixed stepsize for $\theta > 1/2$. In fact for $\theta = 1$ the iteration complexity is logarithmic, i.e. $O(\ln \frac{1}{\epsilon})$. The basic idea is to use a constant stepsize in the subgradient method and every K iterations reduce the stepsize by a factor of $\beta_{ds}^{\frac{1}{2\theta}} > 1$. Also the number of iterations K increases by a factor $\beta_{ds}^{\frac{1}{\theta}-1}$. Our analysis allows us to determine good choices for the initial stepsize and number of iterations which lead to an improved rate.

The algorithm is similar to RSG [109] and the algorithm proposed in [110, Sec. V]. However our method has some important advantages and a different analysis. Unlike RSG our method does not require averaging the iterates after every inner loop. This is beneficial on problems where a sparse or low-rank solution is desired as averaging spoils these properties. The main advantage of DS-SG over the scheme of [110, Sec V] is that it can handle $\theta < 1$.

We call our algorithm the "descending stairs subgradient method" (DS-SG). The method requires an upper bound on the distance of the starting point to the solution, i.e. $\Omega_1 \geq d(x_{\text{init}}, \mathcal{X}_h)^2$. If $\mathcal{C}$ is bounded then one can use the diameter of $\mathcal{C}$. If a lower bound on the optimal value is known, i.e. $h_l \leq h^*$, then by the error bound $d(x_1, \mathcal{X}_h) \leq c^{-1} \left(h(x_1) - h^*\right)^\theta \leq c^{-1} \left(h(x_1) - h_l\right)^\theta$ implies we can use $\Omega_1 = c^{-2} \left(h(x_1) - h_l\right)^{2\theta}$.

Theorem 26 *Suppose Assumption 3 holds and $\frac{1}{2} < \theta \leq 1$. Choose $x_{init} \in \mathcal{C}$*

Algorithm 3: (DS-SG) Descending Stairs Subgradient Method for $\theta > 1/2$

```
Require: $\beta_{ds}$, $M$, $x_{\text{init}}$, $\Omega_1$, $G$, $c$, $\theta$.
$K_1 = \left\lceil \frac{3^{\frac{1}{2\theta}}\theta G^2 \Omega_1^{1-\frac{1}{\theta}}}{2^{\frac{1}{2\theta}} c^2} \beta_{ds}^{\frac{1}{2\theta}} \ln(3\beta_{ds}) \right\rceil$
$\alpha(1) = \frac{2c}{3^{\frac{1}{2\theta}} G^2} (\beta_{ds}^{-1}\Omega_1)^{\frac{1}{2\theta}}$
$\hat{x}_0 = x_{\text{init}}$
for $m = 1, 2, \dots, M$ do
    $\hat{x}_m = \text{FixedSG}(K_m, \alpha(m), \hat{x}_{m-1})$
    $\alpha(m+1) = \beta_{ds}^{-\frac{1}{2\theta}} \alpha(m)$
    $K_{m+1} = \beta_{ds}^{\frac{1}{\theta}-1} K_m$
end for
return $\hat{x}_M$
```

and Ω_1 such that $d(x_{init}, \mathcal{X}_h)^2 \leq \Omega_1$. Choose $0 < C_\beta < 1$ and β_{ds} so that

$$\beta_{ds} \geq \frac{1}{1 - C_\beta}.$$

In addition, if $\theta < 1$ ensure that

$$\beta_{ds} \quad \geq \quad \frac{1}{3} \max\left\{ \left(\frac{\kappa^2}{4}\right)^{\frac{\theta}{\theta-1}} \Omega_1, 2\theta^{-2\theta}\kappa^{-4\theta}\Omega_1^{2(1-\theta)} \right\}. \tag{4.25}$$

Fix $\epsilon > 0$ and choose $M \geq \left\lceil \frac{\ln \frac{\Omega_1}{\epsilon}}{\ln \beta_{ds}} \right\rceil$. Then for $\hat{x}_M$ returned by Algorithm DS-SG, $d(\hat{x}_m, \mathcal{X}_h)^2 \leq \epsilon$.

1. *If $\theta = 1$ this requires at most*

$$\theta \left(\frac{3}{2}\right)^{\frac{1}{2\theta}} \kappa^2 \beta_{ds}^{\frac{1}{2\theta}} \ln(3\beta_{ds}) \left(\frac{\ln \frac{\Omega_1}{\epsilon}}{\ln \beta_{ds}} + 1 \right) \tag{4.26}$$

$$= \quad O\left(\kappa^2 \ln \frac{\Omega_1}{\epsilon}\right) \tag{4.27}$$

subgradient evaluations.

If $\theta < 1$ this requires at most

$$\frac{2\theta}{C_\beta} \left(\frac{3}{2}\right)^{\frac{1}{2\theta}} \kappa^2 \beta_{ds}^{\frac{1}{2\theta}} \ln(3\beta_{ds}) \epsilon^{1-\frac{1}{\theta}} \tag{4.28}$$

$$= \quad \tilde{O}\left(\max\left\{ \kappa^2, \kappa^{\frac{2\theta-1}{\theta-1}} \Omega_1^{\frac{1}{2\theta}}, \Omega_1^{\frac{1}{\theta}-1} \right\} \epsilon^{1-\frac{1}{\theta}} \right) \tag{4.29}$$

subgradient evaluations, where $\tilde{O}$ suppresses constants and terms which depend on $\log\kappa$ or $\log\Omega_1$.

Proof We need some new notation. For $\hat{x}_m$ defined in line 5 of DS-SG, let $\hat{e}_m = d(\hat{x}_m, \mathcal{X}_h)^2$. We will use a sequence of tolerances $\{\epsilon_m\}$ defined as $\epsilon_m = \beta_{ds}^{-m}\Omega_1$. Another sequence $\{D_m\}$ is chosen as $D_m = 2\beta_{ds}\epsilon_m$. The stepsize $\alpha(m)$ is equal to

$$\alpha(m) = \frac{2c}{G^2}\left(\frac{\epsilon_m}{3}\right)^{\frac{1}{2\theta}}$$

and the number of iterations K_m is chosen to satisfy

$$K_m = \left\lceil \frac{3^{\frac{1}{2\theta}}\theta G^2}{2^{\frac{1}{2\theta}}c^2}\beta_{ds}^{1-\frac{1}{2\theta}}\ln(3\beta_{ds})\,\epsilon_m^{1-\frac{1}{\theta}} \right\rceil. \tag{4.30}$$

Note that K_1, given in Line 1 of Algorithm DS-SG, can be written as (4.30) by substituting $\epsilon_1 = \beta_{ds}^{-1}\Omega_1$. Furthermore, for K_m defined in (4.30), note the recursive relationship:

$$K_m = \beta_{ds}^{\frac{1}{\theta}-1}K_{m-1},$$

which is implemented on Line 7 of Algorithm DS-SG. Altogether, this implies that K_m, written in Line 7 of Algorithm DS-SG, satisfies (4.30) for all $m \geq 1$. The set $\{\epsilon_m/3, D_m, K_m, \alpha(m)\}$ will be used in statement 2 of Theorem 25 in place of $\{\epsilon, D, K, \alpha\}$. This will show that $\hat{e}_m \leq \epsilon_m$.

We now show that $\{\epsilon_m/3, D_m, K_m, \alpha(m)\}$ satisfy (4.19), (4.21), (4.22), and (4.23). First we prove that condition (4.25) ensures that (4.22) is satisfied for all $m \geq 1$. The first argument to the min in (4.22) requires that

$$\frac{\epsilon_m}{3} = \frac{1}{3}\beta_{ds}^{-m}\Omega_1 \leq \left(\frac{\kappa^2}{4}\right)^{\frac{\theta}{1-\theta}}.$$

In order for this to be satisfied for all m, it must hold for $m = 1$. This is implied by (4.25). The second argument to the min in (4.22) requires

$$\frac{\epsilon_m}{3} \leq \left(\frac{\theta\kappa^2}{2}\right)^{2\theta} D_m^{2\theta-1} = \frac{1}{2}\left(\theta\kappa^2\right)^{2\theta}\beta_{ds}^{2\theta-1}\epsilon_m^{2\theta-1}.$$

Using $\epsilon_m = \beta_{ds}^{-m}\Omega_1$ and rearranging this yields

$$\beta_{ds}^{2m(1-\theta)+2\theta-1} \geq \frac{2}{3}\theta^{-2\theta}\kappa^{-4\theta}\Omega_1^{2(1-\theta)}.$$

In order to hold for all $m \geq 1$ it must hold for $m = 1$ which is implied by (4.25).

By definition, $\alpha(m)$ satisfies (4.19) for all $m \geq 1$. We prove (4.21) and (4.23) by induction. For $m = 1$, D_1 clearly satisfies (4.21). Also K_1, given in Line 1 of Algorithm DS-SG, satisfies (4.23). Altogether this implies $\hat{e}_1 \leq \epsilon_1$.

Next, assume it holds true at iteration $m - 1$, which implies by Theorem 24 $\hat{e}_{m-1} \leq \epsilon_{m-1}$. Since FixedSG is initialized at $\hat{x}_{m-1}$, and $d(\hat{x}_{m-1}, \mathcal{X})^2 \leq \epsilon_{m-1}$, then

$$D_m = 2\beta_{ds}\epsilon_m = 2\epsilon_{m-1}$$

satisfies (4.21). Next to satisfy (4.23) we require

$$K_m \quad \geq \quad \frac{3^{\frac{1}{2\theta}}\theta G^2}{2c^2} \ln\left(\frac{3d(\hat{x}_{m-1}, \mathcal{X}_h)^2}{\epsilon_m}\right) D_m^{1-\frac{1}{2\theta}} \epsilon_m^{-\frac{1}{2\theta}} \tag{4.31}$$

which is satisfied by K_m. This can be seen by substituting $D_m = 2\beta_{ds}\epsilon_m$ and $d(\hat{x}_{m-1}, \mathcal{X}_h)^2 \leq \epsilon_{m-1} = \beta_{ds}\epsilon_m$ into (4.31), and comparing with (4.30). Thus $\{\epsilon_m/3, D_m, K_m, \alpha(m)\}$ satisfies the requirements of Theorem 24 part 2 which implies $\hat{e}_m \leq 3(\epsilon_m/3) = \epsilon_m$.

Now the choice $M = \left\lceil \frac{\ln \frac{\Omega_1}{\epsilon}}{\ln \beta_{ds}} \right\rceil$ implies $\epsilon_m \leq \epsilon$. If $\theta = 1$, the total number of subgradient evaluations is

$$MK_1 \quad \leq \quad \theta \left(\frac{3}{2}\right)^{\frac{1}{2\theta}} \kappa^2 \beta_{ds}^{\frac{1}{2\theta}} \ln(3\beta_{ds}) \left(\frac{\ln \frac{\Omega_1}{\epsilon}}{\ln \beta_{ds}} + 1\right).$$

In the case $\theta = 1$ note that (4.25) reduces to

$$\beta_{ds} \geq \max\left\{\frac{1}{1 - C_\beta}, \frac{2}{3\kappa^4}\right\} \geq \max\left\{\frac{1}{1 - C_\beta}, \frac{2}{3}\right\} = \frac{1}{1 - C_\beta},$$

since $\kappa \geq 1$ when $\theta = 1$ (and typically $\kappa \gg 1$). Therefore β_{ds} can be treated as a constant, which implies (4.27).

If $\theta < 1$ the total number of subgradient evaluations is

$$\begin{aligned}
K_1 + K_2 + \ldots + K_M &= K_1 \left(1 + \beta_{ds}^{\frac{1}{\theta}-1} + (\beta_{ds}^{\frac{1}{\theta}-1})^2 + \ldots + (\beta_{ds}^{\frac{1}{\theta-1}})^{M-1} \right) \\
&= K_1 \frac{(\beta_{ds}^{\frac{1}{\theta}-1})^M - 1}{(\beta_{ds}^{\frac{1}{\theta}-1}) - 1} \\
&\leq K_1 \frac{(\beta_{ds}^{\frac{1}{\theta}-1})^M}{(\beta_{ds}^{\frac{1}{\theta}-1}) - 1} \\
&\leq \frac{1}{C_\beta} \frac{K_1}{\beta_{ds}^{\frac{1}{\theta}-1}} (\beta_{ds}^{\frac{1}{\theta}-1})^M. \qquad (4.32)
\end{aligned}$$

Now since

$$M \leq \frac{\ln \frac{\Omega_1}{\epsilon}}{\ln \beta_{ds}} + 1$$

it follows that

$$(\beta_{ds}^{\frac{1}{\theta}-1})^M \leq \beta_{ds}^{\frac{1}{\theta}-1} (\Omega_1/\epsilon)^{\frac{1}{\theta}-1}. \qquad (4.33)$$

Also

$$K_1 \leq 2\theta \left(\frac{3}{2}\right)^{\frac{1}{2\theta}} \kappa^2 \beta_{ds}^{\frac{1}{2\theta}} \Omega_1^{1-\frac{1}{\theta}} \ln(3\beta_{ds}). \qquad (4.34)$$

Using (4.33) and (4.34) in (4.32) yields

$$K_1 + K_2 + \ldots + K_M \leq \frac{2\theta}{C_\beta} \left(\frac{3}{2}\right)^{\frac{1}{2\theta}} \kappa^2 \beta_{ds}^{\frac{1}{2\theta}} \ln(3\beta_{ds}) \epsilon^{1-\frac{1}{\theta}}.$$

Now if β_{ds} satisfies (4.25) with equality then this reduces to (4.29).

4.5.1 Discussion

The optimal choice for β_{ds} can be found as follows. We wish to minimize the iteration complexity given in (4.26) for $\theta = 1$ and (4.28) for $\theta < 1$. For $\theta = 1$, (4.26) is a convex function in $\beta_{ds} > 1$. The optimal choice can be found by setting the derivative w.r.t. β_{ds} to 0 however the closed form expression is not particularly enlightening. Solving it numerically, we find the optimal choice for β_{ds} is typically between 2 and 2.5, depending on the value of $\ln \frac{\Omega_1}{\epsilon}$. For $\theta < 1$, the iteration complexity in (4.28) is increasing with β_{ds}, therefore the optimal choice is to set β_{ds} to equal (4.25) with equality.

The method of [110, Sec V] corresponds to the special case of our method when $\theta = 1$. However the analysis of [110] does not extend naturally to $\theta < 1$. With regards to RSG in [109], the iteration complexity is very similar to ours, even though the analysis is different. There are several things to note in comparing the two. First is that their error metric is $h(x) - h^*$. Now if h is convex with bounded subgradients then

$$\begin{aligned} h(x) - h^* &\leq |\langle g, x - x^* \rangle| \quad \forall g \in \partial h(x), x^* \in \mathcal{X}_h \\ &\leq \|g\|\|x - x^*\| \quad \forall g \in \partial h(x), x^* \in \mathcal{X}_h \\ &\leq G\|x - x^*\| \quad \forall x^* \in \mathcal{X}_h. \end{aligned}$$

In particular choosing x^* to be the projection of x onto $\mathcal{X}_h$ yields $h(x) - h^* \leq Gd(x, \mathcal{X}_h)$. Combining this with the error bound

$$cd(x, \mathcal{X}_h)^\theta \leq h(x) - h^* \leq Gd(x, \mathcal{X}_h).$$

On the other hand our error metric is $d(x_k, \mathcal{X}_h)^2$. Furthermore their iteration complexity is for finding $h(x) - h^* \leq 2\epsilon$. To do an apples-to-apples comparison, we can convert their error metric to $d(x_k, \mathcal{X}_h)^2$ by using $\epsilon' = \epsilon^{\frac{1}{2\theta}}/2$ in their iteration complexity. Recall their iteration complexity is $O(\epsilon'^{2(\theta-1)} \ln \frac{1}{\epsilon'})$. Thus, if we make the substitution, we see that their iteration complexity is the same as ours except they have an extra $\log \frac{1}{\epsilon}$ term. The dependence on $\kappa^2 = G^2/c^2$ is the same.

With respect to their algorithm implementation as given in [109, Algorithm 2], the major difference to DS-SG is that [109] requires averaging to be done after every inner loop. This may be undesirable on problems where nonergodic methods are preferable. For instance, in problems where $\mathcal{C}$ enforces sparsity or low-rank, the averaging phase spoils this property [121]. Indeed some matrix problems are intractable unless the iterates remain low rank [123]. Another situation in which averaging is undesirable is when learning with reproducing kernals [124]. In such problems, the variable is represented as a linear combination of a kernel evaluated at different points. After t iterations of the subgradient method, the solution is $\sum_{i=1}^{t-1} \alpha_i k(x_i, \cdot)$ where $k : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is the kernel function. Thus it is necessary to store the $t-1$ points $\{x_i\}$ after t iterations which is infeasible. The key to making the method practical is that for certain objectives the coefficients α_i decay geometrically and the early iterations can be safely ignored. Thus only a small fraction of the last t points are recorded. However, if averaging is used, the earlier coefficients are no longer negligible which compromises the feasibility of the method. Another advantage of our approach over [109] will arise in the next section, where we develop a method for adapting to unknown c.

An advantage of RSG over DS-SG is that RSG only requires the error bound to be satisfied on a local region such that $h(x) - h^* \leq \epsilon$, where ϵ is the target accuracy. However if the function satisfies HEB with $\theta = 1$ on a local region, then it is automatically satisfied on the entire space. This quite intuitive observation can be shown by considering the equivalent subgradient characterization of WS functions given in [94]. We note that for $\theta < 1$, the iteration complexity of DS-SG has worse dependency on Ω_1 than RSG.

We also mention Algorithm 3 of [118] which is a new subgradient method for functions satisfying a similar condition HEB with $\theta = 1$, but with h^* replaced by a strict lower bound on h^*. Like DS-SG and RSG, this algorithm has a logarithmic dependence on the initial distance to the solution set. However it obtains an $O(1/\epsilon^2)$ iteration complexity which is worse than the $O(\ln \frac{1}{\epsilon})$ rate obtained by DS-SG and RSG in the weakly sharp case.

The argument in the proof of Theorem 26 for the case $\theta = 1$ is similar to [116, Thm. 2] (see also [119, Cor. 3.4]). Both theorems take a base algorithm and create a meta-algorithm with faster overall convergence. In [116] the problem of interest is a linear min-max saddlepoint problem and the base algorithm is Nesterov's smoothing. In Thm. 26 the base algorithm is the constant stepsize subgradient method. Finding a unifying theory would be an interesting topic for future research.

4.6 Double Descending Stairs Stepsize Method for Unknown c

In our method DS-SG, the initial number of inner iterations is

$$K_1 = \left\lceil \frac{3^{\frac{1}{2\theta}} \theta G^2 \Omega_1^{1-\frac{1}{\theta}}}{2^{\frac{1}{2\theta}} c^2} \beta_{ds}^{\frac{1}{2\theta}} \ln(3\beta_{ds}) \right\rceil . \tag{4.35}$$

If a lower bound for c is known, then using this value in (4.35) ensures convergence. However in many problems c is unknown. Further if c is greatly underestimated than this will lead to many more inner iterations than necessary. For the case where no accurate lower bound for c is known, we propose the following "doubling trick" which still guarantees an overall logarithmic iteration complexity. The analysis only holds when $\mathcal{C}$ is bounded. Let the diameter of $\mathcal{C}$ be $\Omega_{\mathcal{C}} = \max_{x,x' \in \mathcal{C}} \|x - x'\|^2$. The basic idea is to repeat

DS-SG with a new c which is $1/2$ the old estimate. In this way it takes only a logarithmic number of trial choices for c until it lower bounds the true constant. Furthermore, if the initial estimate c_1 is much larger than the true c, then the number of inner iterations is relatively small, which is why the overall iteration complexity comes out to be only a factor of $(4/3)$ times larger than that of DS-SG. This means it is advantageous to use a large over-estimate of c. Following the naming convention of [109] we call the method the "Descending Stairs Squared" subgradient method (DS2-SG).

Algorithm 4: Double Descending Stairs subgradient method for $\theta = 1$, unknown c (DS2-SG)

```
Require: $\beta_{ds}$, $G$, $M$, $c_1$, $\Omega_{\mathcal{C}}, x_1$, stopping criterion
l = 1
while stopping criterion not satisfied do
    $\tilde{x}_l$ =DS-SG($\beta_{ds}, M, \tilde{x}_{l-1}, \Omega_{\mathcal{C}}, G_l, c_l, \theta, \epsilon$)
    $c_{l+1} = c_l/2$
    $l = l + 1$
end while
return $\tilde{x}_{l-1}$
```

Theorem 27 *Suppose Assumption 3 holds and $\theta > 1/2$. Suppose $\mathcal{C}$ is bounded with diameter $\Omega_{\mathcal{C}}$. Choose $C_\beta \in (0, 1)$, $\beta_{ds} > 0$ and $c_1 > 0$ so that*

$$\beta_{ds} \geq \frac{1}{1 - C_\beta}. \tag{4.36}$$

In addition, if $\theta < 1$ ensure that

$$\beta_{ds} \quad \geq \quad \frac{1}{3} \max\left\{ \left(\frac{\kappa_1^2}{4}\right)^{\frac{\theta}{\theta-1}} \Omega_{\mathcal{C}}, 2\theta^{-2\theta}\kappa_1^{-4\theta}\Omega_{\mathcal{C}}^{2(1-\theta)} \right\}, \tag{4.37}$$

where $\kappa_1 = G/c_1$. Fix $\epsilon > 0$ and choose

$$M \geq \left\lceil \frac{\ln \frac{\Omega_{\mathcal{C}}}{\epsilon}}{\ln \beta_{ds}} \right\rceil.$$

For the output of Algorithm DS2-SG, if $l \geq L = \max\{0, \lceil \log_2 c_1/c \rceil\} + 1$, then $d(\tilde{x}_l, \mathcal{X}_h)^2 \leq \epsilon$. This requires the following number of subgradient evaluations:

$$O\left(\overline{\kappa}^2 \ln \frac{\Omega_{\mathcal{C}}}{\epsilon}\right) \textit{ if } \theta = 1, \tag{4.38}$$

and

$$\tilde{O}\left(\max\left\{\overline{\kappa}^2, \overline{\kappa}^2\kappa_1^{\frac{1}{\theta-1}}\Omega_{\mathcal{C}}^{\frac{1}{2\theta}}, \left(\frac{\overline{\kappa}}{\kappa_1}\right)^2\Omega_{\mathcal{C}}^{\frac{1}{\theta}-1}\right\}\epsilon^{1-\frac{1}{\theta}}\right) \quad \text{if } \theta < 1, \qquad (4.39)$$

where $\overline{\kappa} = \max\{\kappa, \kappa_1\}$ *and* $\kappa_1 = G/c_1$. *If* $c_1 = G\Omega_{\mathcal{C}}^{1-\frac{1}{\theta}}$, $\kappa_1 \leq \kappa$ *and* $\overline{\kappa} = \kappa$.

Proof If $c_l \leq c$, for any $l \leq L$ then, $d(\tilde{x}_l, \mathcal{X}_h)^2 \leq \epsilon$ by Theorem 26. So we assume $c_l > c$ for $l = 1, 2\ldots, L-1$. For $l < L$ it is clear that since the iterates remain in the constraint set $\mathcal{C}$, $d(\tilde{x}_l, \mathcal{X}_h)^2 \leq \Omega_{\mathcal{C}}$. Now by the choice of L, $c_l \leq c$ for all $l \geq L$. Therefore we can apply Theorem 26 to the iterations within the while loop when $l \geq L$, which implies $d(\tilde{x}_l, \mathcal{X}_h)^2 \leq \epsilon$ for $l \geq L$.

We now determine the overall iteration complexity. let K_j^l for $l = 1, 2, \ldots, L$ and $j = 1, 2, \ldots M$ be the number of iterations passed to FixedSG within the jth call to FixedSG in DS-SG, during the lth loop in DS2-SG. The total number of subgradient calls of DS2-SG is

$$\begin{aligned} & (K_1^1 + K_2^1 + \ldots K_M^1) + (K_1^2 + K_2^2 + \ldots K_M^2) + \ldots (K_1^L + K_2^L + \ldots + K_M^L) \\ = \; & (K_1^1 + K_2^1 + \ldots K_M^1)\left(1 + 4 + 16 + \ldots + 4^{L-1}\right) \\ = \; & \frac{1}{3}(K_1^1 + K_2^1 + \ldots K_M^1)(4^L - 1) \\ = \; & \frac{4}{3}(K_1^1 + K_2^1 + \ldots K_M^1)\max\left\{\left(\frac{c_1}{c}\right)^2, 1\right\}. \end{aligned}$$

which reduces to the iteration complexity given in (4.38)–(4.39).

Now

$$c d(x, \mathcal{X})^{\frac{1}{\theta}} \leq h(x) - h^* \leq \|g\|\|x - x^*\|$$

for all $x \in \mathcal{C}$, $g \in \partial h(x)$. Therefore, let $x^* = \text{proj}_{\mathcal{X}}(x)$ then

$$c d(x, \mathcal{X})^{\frac{1}{\theta}} \leq G d(x, \mathcal{X}) \implies c \leq G d(x, \mathcal{X})^{1-\frac{1}{\theta}} \quad \forall x.$$

Minimizing the R.H.S. yields $c \leq G\Omega_{\mathcal{C}}^{1-\frac{1}{\theta}}$. Therefore the choice $c_1 = G\Omega_{\mathcal{C}}^{1-\frac{1}{\theta}}$ guarantees $\kappa_1 \leq \kappa$.

The competing methods for $\theta = 1$ which also obtain a $O(\log\frac{1}{\epsilon})$ complexity cannot handle unknown c. This is a major advantage of DS2-SG. The authors of RSG [109] proposed a variant which also uses exponentially increasing number of inner iterations, however the initial stepsize remains constant. An advantage of that method is it does not require the constraint set to be bounded. However their analysis is only valid for $\theta < 1$, which excludes important problems such as polyhedral convex optimization.

A drawback of DS2-SG is it does not have an explicit stopping rule. In particular, the number of "wrapper" iterations, L, depends on the true error bound constant c, which is unknown. This is also the main drawback for the variant restart scheme of [109] (along with the fact it cannot be applied when $\theta = 1$). As was suggested in [109], we suggest using an independent stopping criterion. For example on a machine learning problem, one could use the error on a validation set as an indication the algorithm has converged. If a lower bound $h_{LB} \leq h^*$ is known, then $\frac{1}{c^\theta}\left(h(x_k) - h_{LB}\right)^\theta$ can be used as a stopping criterion. This is because $d(x_k, \mathcal{X}_h) \leq \frac{1}{c^\theta}\left(h(x_k) - h_{LB}\right)^\theta$. Furthermore since, $d(x_k, \mathcal{X})^{\frac{1}{\theta}-1} \leq \|g\|$ for $g \in \partial h(x)$, the norm of the subgradient could be used as a stopping criterion for $\theta < 1$.

In practice we often observe an increase in the objective function value occurs at the beginning of each new iteration inside the while loop. This occurs because the stepize is reduced by $1/2$ which breaks the algorithm away from its current fixed point. It is therefore a good strategy to keep track of the iterate $\tilde{x}_l$ with the smallest objective function value so far, and use this as the output. Thus the modified algorithm returns $\arg\min_{l=0,1,\ldots,L} h(\tilde{x}_l)$. This does not change the overall iteration complexity.

4.7 Convergence Rates for Nonsummable Stepsizes

We now turn our attention to nonsummable but square summable stepsize sequences for the subgradient method under HEB. These stepsizes are used frequently for the stochastic and deterministic subgradient method, however their behavior under HEB has not been studied in detail with the exception of [117, 110]. We will see that these nonsummable stepsizes are slower than the "descending staircase" stepsizes and summable stepsizes when $\theta > 1/2$. However for $\theta \geq 1/2$ the nonsummable stepsizes have the advantage that they do not require G, c, and an upper bound for $d(x_1, \mathcal{X})^2$. We will first state and discuss our results. The proofs are in Section 4.10.

4.7.1 Results for $\theta \in (0, \frac{1}{2})$

Theorem 28 *Suppose Assumption 3 holds and $0 < \theta < 1/2$. Let $\alpha_k = \alpha_1 k^{-p}$. Let*

$$\begin{aligned} C_1 &\triangleq 2^{2p\theta+1}\left(\left(\frac{\alpha_1 G^2}{c}\right)^{2\theta} + \alpha_1^2 G^2\right) \qquad (4.40) \\ C_2 &\triangleq \left(\frac{\alpha_1(1-2\theta)}{2\theta(1-p)}\right)^{\frac{2\theta}{2\theta-1}}. \end{aligned}$$

Then if

$$\frac{1}{2(1-\theta)} \leq p \leq 1 \qquad (4.41)$$

and α_1 is chosen so that

$$\begin{aligned} C_1 &\leq \left(\frac{2\theta(1-p)}{\alpha_1(1-2\theta)}\right)^{\frac{2\theta}{1-2\theta}} (k_0+1)^{\frac{2\theta(2p(1-\theta)-1)}{1-2\theta}}, \qquad (4.42) \\ \alpha_1 &\leq \frac{2\theta(1-p)d(x_1, \mathcal{X}_h)^{\frac{2\theta-1}{\theta}}}{1-2\theta}, \qquad (4.43) \end{aligned}$$

then for all $k \geq k_0$

$$d(x_k, \mathcal{X}_h)^2 \leq \max\{C_1, C_2\} \max\left\{k^{-2p\theta}, k^{\frac{2\theta(1-p)}{2\theta-1}}\right\}. \qquad (4.44)$$

In the following corollary we give the optimal choice for p that makes the two arguments to the max function in (4.44) equal.

Corollary 29 *In the setting of Theorem 28 with $0 < \theta < \frac{1}{2}$ and C_1 defined in (4.40), if $p = \frac{1}{2(1-\theta)}$, and α_1 is chosen so that (4.43) holds and*

$$\alpha_1^{\frac{2\theta}{1-2\theta}} C_1 \leq \left(\frac{\theta}{1-\theta}\right)^{\frac{2\theta}{1-2\theta}} \qquad (4.45)$$

then for all $k \geq 1$

$$d(x_k, \mathcal{X}_h)^2 \leq \alpha_1^{\frac{2\theta}{2\theta-1}} \left(\frac{\theta}{1-\theta}\right)^{\frac{2\theta}{1-2\theta}} k^{\frac{-\theta}{1-\theta}}.$$

If α_1 is chosen so that (4.45) is satisfied with equality, then

$$d(x_k, \mathcal{X}_h)^2 \leq C_1 k^{\frac{-\theta}{1-\theta}}.$$

We note that our derived convergence rate $O(k^{\frac{-\theta}{1-\theta}})$ is faster than the naive application of the classical result, which is $d(\hat{x}_k, \mathcal{X}_h)^2 = O(k^{-\theta})$ at the averaged point $\hat{x}_k = \sum \alpha_k x_k / \sum \alpha_k$. Furthermore our result is nonergodic (no averaging is required).

Thus for $\theta < 1/2$ decaying polynomial stepsize sequences can achieve the same convergence rate as RSG of [109] and the constant stepsize we derived in Theorem 25.

4.7.2 Results for $\theta \in [\frac{1}{2}, 1]$

We now consider nonsummable stepsizes for $\theta \geq 1/2$. The primary advantage of the following theorem is that the stepsize does not require knowledge of G, c, or $d(x_1, \mathcal{X})^2$.

Theorem 30 *Suppose Assumption 3 holds and $1/2 \leq \theta \leq 1$. Suppose $\alpha_k = \alpha_1 k^{-p}$ for some $p \in (0,1)$ and $\alpha_1 > 0$. Let C_1 be as defined in (4.40),*

$$\begin{aligned} C_3 &\triangleq C_1^{\frac{1+2p(\theta-1)}{1-p}} \left(\frac{\alpha_1(1-2^{p-1})ce}{4p\theta}\right)^{-\frac{2p\theta}{1-p}} \\ C_4 &\triangleq 16\left(\frac{8\theta C_1}{\alpha_1 ce}\right)^{2\theta} \\ C_5 &\triangleq d(x_1, \mathcal{X}_h)^{\frac{2+4p(\theta-1)}{1-p}} \left(\frac{\alpha_1 ce}{4p\theta}\right)^{-\frac{2p\theta}{1-p}}. \end{aligned}$$

Then for all $k \geq 4$

$$d(x_k, \mathcal{X}_h)^2 \leq 4 \max\{C_1, C_3, C_4, C_5\} k^{-2p\theta}. \tag{4.46}$$

Once again this improves on the known classical *ergodic* convergence rate of $O(k^{-\theta})$. As $p \to 1$ the method can get arbitrarily close to the best rate $O(k^{-2\theta})$, however $p = 1$ is not covered by our analysis other than the special case $\theta = \frac{1}{2}$ discussed in Theorem 31 and Proposition 32 below. The decaying stepsize does not require knowledge of θ, c, G, h^*, or $d(x_1, \mathcal{X}_h)$ to set the parameters α_1 and p. The result holds for arbitrary $\alpha_1 > 0$ and $p \in (0,1)$. Nevertheless, the constants are affected by the choice of α_1 and p as well as practical performance.

The convergence rate for the decaying stepsizes is much slower than DS-SG, the summable stepsizes in Sec. 4.8, and RSG [109]. These methods obtain the rate $O\left(k^{\frac{\theta}{\theta-1}}\right)$ for $\theta > 1/2$. On the other hand Theorems 28 and 25 imply restarting is unnecessary for $\theta \leq 1/2$ as either the constant choice or the decaying polynomial choice have the same convergence rate as RSG.

The case $\theta = 1$ in Theorem 30 can be compared with the main result of [117] which also proves $O(1/k^2)$ rate of convergence for $d(x_k, \mathcal{X}_h)^2$. A difference is their result only holds for sufficiently large k. They also assume the function satisfies the quadratic growth condition (i.e. $\theta = 1/2$ error bound) globally. For problems where $\mathcal{C}$ is compact, this does not matter, since QG is implied by WS on a compact set. An advantage of [117] is that it holds for stochastic gradient descent.

4.7.3 Results for $\theta = \frac{1}{2}$

For the special case of $\theta = \frac{1}{2}$ our analysis extends to the choice $p = 1$.

Theorem 31 *Suppose Assumption 3 holds and $\theta = 1/2$. Suppose $\alpha_k = \alpha_1 k^{-1}$ and*

$$\alpha_1 \leq \frac{1}{c}.$$

Then for $k \geq 1$

$$d(x_k, \mathcal{X}_h)^2 \leq \max\left\{\frac{2\alpha_1 G^2}{c}, d(x_1, \mathcal{X}_h)^2\right\} k^{-c\alpha_1}. \tag{4.47}$$

Strongly convex functions with strong convexity parameter μ_{sc} satisfy the error bound with $\theta = \frac{1}{2}$ and $c = \frac{\mu_{sc}}{2}$. In this case $C_1 = \frac{8G^2}{c^2}$. Thus, for the choice $\alpha_1 = \frac{2}{\mu_{sc}}$ we have proved that

$$d(x_k, \mathcal{X}_h)^2 \leq \frac{\max\left\{d(x_1, \mathcal{X}_h)^2, \frac{32G^2}{\mu_{sc}^2}\right\}}{k}.$$

This result can be compared with several papers. The result [125, Theorem 6.2] finds an $O(1/k)$ convergence rate for $h(\hat{x}_k) - h^*$ for a particular averaged point $\hat{x}_k$ under strong convexity. This, combined with HEB, implies an $O(1/k)$ rate for $d(\hat{x}_k, \mathcal{X}_h)^2$. The work [126, Thm 1] obtained a nonergodic $O(1/k)$ rate for $d(x_k, \mathcal{X}_h)^2$ in stochastic mirror descent under strong convexity for a similar stepsize sequence to Theorem 31. The result [88, Prop. 2.8] provides convergence rates for the (incremental) subgradient method with stepsize $\alpha_k = \alpha_1 k^{-1}$ for all values of α_1 under QG. This is more general than Theorem 31 as they cover the case where $\alpha_1 > 1/c$. However, for $\alpha_1 = 1/c$, [88, Prop. 2.8] only proves $O(\log k/k)$ convergence whereas Theorem 31 implies $O(1/k)$ convergence. The result of [89, Eq. (2.9)] says that for strongly convex functions with parameter μ_{sc}, the subgradient method achieves a nonergodic $O(1/k)$ convergence so long as $\alpha_1 > \frac{1}{2\mu_{sc}}$. In contrast we do not require strong convexity but only the weaker error bound. The result can also be compared to [100, Thm. 4] which proved an $O(1/k)$ rate for the objective function gap under QG. However they additionally require Lipschitz smoothness. Both [89] and [100] considered the stochastic subgradient method.

We also provide another choice of stepsize which guarantees a convergence rate of $O(1/k)$ for $d(x_k, \mathcal{X}_h)^2$ in the case where $\theta = \frac{1}{2}$. This proof is a direct adaptation of [100, Thm. 4]. Unlike [100, Thm. 4], it does not require smoothness of the objective.

Proposition 32 *In the setting of Theorem 31, consider the subgradient method with*

$$\alpha_k = \frac{2k+1}{2c(k+1)^2}.$$

Then for all k

$$d(x_{k+1}, \mathcal{X}_h)^2 \leq \frac{d(x_1, \mathcal{X}_h)^2}{(k+1)^2} + \frac{G^2}{c^2(k+1)}.$$

Note that the stepsizes of Theorem 31 and Proposition 32 both require exact knowledge of c to achieve the $O(1/k)$ rate.

4.7.4 Local Error Bounds

So far we have assumed that the error bound is satisfied for all $x \in \mathcal{C}$. As discussed in Sec. 4.5.1 in the case where $\theta = 1$, if the bound is satisfied on a local region then it is also satisfied on the entire set $\mathcal{C}$. However for other problems (particularly when $\theta = 1/2$) it may be that the error bound is satisfied on any compact set but with a different value of the error bound constant c depending on the set. Enlarging the set necessarily leads to a smaller constant. For example this is the case with ℓ_1 regularized least-squares [69, Lemma 10] and logistic regression [100, Sec. 2.3]. It has been shown to be true for a general class of convex functions [103, Theorem 3.3].

For square summable stepsize sequences in the subgradient method it is trivial to prove that $d(x_k, \mathcal{X}_h)$ is bounded. Thus if $\mathcal{X}_h$ is bounded than this implies that x_k is bounded. Therefore our results are applicable to a wider range of problems.

Corollary 33 *Assume $\mathcal{X}_h$ is nonempty and bounded, h is CCP, and $\mathcal{C} \subseteq dom(\partial h)$. Fix $\theta \in (0,1]$. Suppose that for any closed and compact set $\mathcal{C}'$ there exists $c(\mathcal{C}')$ and $G(\mathcal{C}')$ such that for all $x \in \mathcal{C}'$ h satisfies HEB with the exponent θ and constant $c(\mathcal{C}')$, and if $g \in \partial h(x)$, then $\|g\| \leq G(\mathcal{C}')$. Then the conclusions of Theorem 28, 30, and 31 hold.*

We exclude Corollary 29 and Proposition 32 as the stepsizes in these results depend on explicit knowledge of c.

4.8 Faster Rates for Decaying Stepsizes for $\frac{1}{2} \leq \theta < 1$

If $\frac{1}{2} \leq \theta < 1$, the constraint set is compact, an upper bound for G is known, and a lower bound for c is known, then it is possible to obtain the same iteration complexity as DS-SG using decaying stepsizes.

Theorem 34 *Suppose Assumption 3 holds and $\frac{1}{2} \leq \theta < 1$. Suppose $\|x - y\|^2 \leq \Omega_{\mathcal{C}}$ for all $x, y \in \mathcal{C}$. Choose c small enough (or G large enough) so that*

$$\kappa \geq \sqrt{3}\Omega_{\mathcal{C}}^{\frac{1-\theta}{2\theta}}.$$

For the iterates of the subgradient method (4.4), let $\alpha_k = \alpha_1 k^{-p}$ where

$$p = \frac{1}{2(1-\theta)}$$

and

$$\alpha_1 = \frac{c}{G^2}\left(\frac{\theta\kappa^2}{1-\theta}\right)^p. \tag{4.48}$$

Then, for all $k \geq \lceil \frac{2\theta}{1-\theta} \rceil$

$$d(x_k, \mathcal{X})^2 \leq \left(\frac{\theta}{1-\theta}\right)^{\frac{\theta}{1-\theta}} \left(\frac{k}{\kappa^2}\right)^{\frac{\theta}{\theta-1}}. \tag{4.49}$$

Proof The recursion describing the subgradient method is, for $k \geq 1$,

$$e_{k+1} \leq e_k - 2\alpha_k c e_k^{\gamma} + \alpha_k^2 G^2, \tag{4.50}$$

where $e_k = d(x_k, \mathcal{X})^2$ and $\gamma = \frac{1}{2\theta}$. Let $\alpha_k = \alpha_1 k^{-p}$. We wish to prove that if

$$p = \frac{\gamma}{2\gamma - 1}$$

and the constant α_1 is chosen as in (4.48), then $e_k \leq C_e k^{-b}$ where

$$b \triangleq \frac{1}{2\gamma - 1},$$

for all $k \geq k_0 \triangleq \lceil 2b \rceil$, and C_e is defined in (4.49). Note that $p = \gamma b$. This will be proved by induction. The initial condition is

$$e_{k_0} \leq C_e k_0^{-b}$$

which is implied by

$$\Omega_{\mathcal{C}} \leq C_e k_0^{-b} \iff C_e \geq \Omega_{\mathcal{C}} k_0^b. \tag{4.51}$$

Next, assume it is true for some $k \geq k_0$. That is $e_k = aC_e k^{-b}$ where $0 \leq a \leq 1$. We wish to prove $e_{k+1} \leq C_e(k+1)^{-b}$. Substitute $e_k = aC_e k^{-b}$ into the right hand side of (4.50) yields the inequality

$$\begin{aligned} & aC_e k^{-b} - 2\alpha_1 c a^\gamma C_e^\gamma k^{-(p+\gamma b)} + \alpha_1^2 G^2 k^{-2p} \\ = \quad & aC_e k^{-b} + \left(\alpha_1^2 G^2 - 2\alpha_1 c a^\gamma C_e^\gamma\right) k^{-2p} \leq C_e(k+1)^{-b} \end{aligned} \tag{4.52}$$

using the fact that $p + \gamma b = 2p$. We need (4.52) to hold for all $a \in [0,1]$. Since $\frac{1}{2} \leq \theta < 1$, $\frac{1}{2} < \gamma \leq 1$, therefore the L.H.S. is a convex function of a. Therefore if the inequality holds for $a = 0$ and $a = 1$, then it holds for all $a \in [0,1]$. Consider first, $a = 0$. The condition is

$$\alpha_1^2 G^2 k^{-2\gamma b} \leq C_e(k+1)^{-b}.$$

This is equivalent to

$$\alpha_1 \leq G^{-1} C_e^{\frac{1}{2}} k^{\gamma b} (k+1)^{-\frac{b}{2}}. \tag{4.53}$$

We will verify this condition later for the specific α_1 chosen in (4.48).

Next consider $a = 1$. For this case we simplify (4.52) using

$$C_e(k+1)^{-b} = C_e k^{-b}(1 + k^{-1})^{-b} \geq C_e k^{-b} - bC_e k^{-(b+1)},$$

where we used convexity of t^{-b}. Therefore in the case $a = 1$, (4.52) is true if

$$\left(\alpha_1^2 G^2 - 2\alpha_1 c C_e^\gamma\right) k^{-2p} \leq -bC_e k^{-(b+1)} \tag{4.54}$$

Now $2p = b+1$, therefore (4.54) holds if

$$\alpha_1^2 G^2 - 2\alpha_1 c C_e^\gamma \leq -bC_e,$$

which is a positive-definite quadratic in α_1. Solving it yields the two solutions

$$\frac{2cC_e^\gamma \pm \sqrt{4c^2 C_e^{2\gamma} - 4G^2 b C_e}}{2G^2}.$$

The quadratic has a real solution if

$$4c^2 C_e^{2\gamma} - 4G^2 bC_e \geq 0 \implies C_e \geq \left(\frac{G^2 b}{c^2}\right)^{\frac{1}{2\gamma-1}} = \left(\frac{G^2}{(2\gamma-1)c^2}\right)^{\frac{1}{2\gamma-1}}. \quad (4.55)$$

We will choose $C_e = (\kappa^2 b)^b$ and then the only valid choice for α_1 is

$$\alpha_1 = \frac{cC_e^\gamma}{G^2}$$

which corresponds to (4.48).

We now verify that this choice of α_1 satisfies (4.53) for all $k \geq k_0 = \lceil 2b \rceil$. Plugging α_1 into (4.53) yields

$$\frac{c}{G^2} C_e^\gamma \leq G^{-1} C_e^{\frac{1}{2}} k^{\gamma b}(k+1)^{-\frac{b}{2}}$$

which can be rearranged to

$$G \geq c C_e^{\gamma - \frac{1}{2}} k^{-\gamma b}(k+1)^{\frac{b}{2}}. \quad (4.56)$$

Then

$$C_e^{\frac{2\gamma-1}{2}} = \kappa\sqrt{b}.$$

Plugging this into (4.56) yields

$$k^{\gamma b}(k+1)^{-\frac{b}{2}} \geq \sqrt{b}. \quad (4.57)$$

Now

$$\begin{aligned} (k+1)^{-\frac{b}{2}} &= k^{-b/2}(1+k^{-1})^{-b/2} \\ &\geq k^{-\frac{b}{2}}\left(1 - \frac{b}{2}k^{-1}\right) \\ &= k^{-\frac{b}{2}} - \frac{b}{2}k^{-\frac{b}{2}-1}. \end{aligned}$$

Therefore (4.57) is implied by

$$k^{b(\gamma-\frac{1}{2})} - \frac{b}{2}k^{b(\gamma-\frac{1}{2})-1} \geq \sqrt{b}.$$

Now substituting $b = (2\gamma - 1)^{-1}$ into the two exponents yields

$$k^{\frac{1}{2}} - \frac{b}{2}k^{-\frac{1}{2}} \geq \sqrt{b}$$

which is equivalent to

$$t^2 - \sqrt{b}t - \frac{b}{2} \geq 0$$

with the substitution $t = \sqrt{k}$. Thus we require

$$t \geq \frac{1+\sqrt{3}}{2}\sqrt{b}$$

which is implied by $k \geq 2b$.

Finally, we verify that C_e satisfies the initial condition (4.51). Thus

$$C_e = \left(b\kappa^2\right)^b \geq \Omega_{\mathcal{C}} k_0^b.$$

Since $k_0 = \lceil 2b \rceil \leq 2b + 1 \leq 3b$, this is implied by

$$\left(b\kappa^2\right)^b \geq \Omega_{\mathcal{C}}(3b)^b.$$

Diving by b^b this yields

$$\kappa^2 \geq 3\Omega_{\mathcal{C}}^{\frac{1}{b}}$$

which completes the proof.

The convergence rate given in (4.49) yields the following iteration complexity: The subgradient method with this stepsize yields a point such that $d(x_k, \mathcal{X})^2 \leq \epsilon$ for all

$$k \geq \frac{2\theta}{1-\theta}\max\{\kappa^2, 3\Omega_{\mathcal{C}}^{\frac{1}{\theta}-1}\}\epsilon^{1-\frac{1}{\theta}}.$$

This is equal (up to constants) to the iteration complexity derived for DS-SG in Theorem 26. The main drawback versus DS-SG is that the analysis only holds for a bounded constraint set. It is also trivial to embed this stepsize into the "doubling" framework used in DS2-SG so that one does not need a lower bound for c. Since the analysis is the same as given in Theorem 27, we omit the details. The proof of Theorem 34 is inspired by [85] which considered geometrically decaying stepsizes when $\theta = 1$. It could be considered a natural extension of [85] to $\theta < 1$.

We can obtain the same rate for this choice of α_1 and p when $\theta < 1/2$. In this case, the constraint set does not need to be bounded and the rate holds for all $k \geq 1$.

Theorem 35 *Suppose Assumption 3 holds and* $0 < \theta \leq \frac{1}{2}$. *Suppose* $d(x_1, \mathcal{X})^2 \leq \Omega_1$. *Choose* c *small enough (or* G *large enough) so that*

$$\kappa^2 \geq \max\left\{1, \frac{1-\theta}{\theta}\Omega_1^{\frac{1-\theta}{\theta}}\right\}. \tag{4.58}$$

For the iterates of the subgradient method (4.4), let $\alpha_k = \alpha_1 k^{-p}$ *where*

$$p = \frac{1}{2(1-\theta)}$$

and α_1 *be defined as in (4.48) Then, for all* $k \geq 1$, $d(x_k, \mathcal{X})^2$ *satisfies (4.49).*

Proof Recall $\gamma = 1/(2\theta)$ and note that $\gamma \geq 1$ since $\theta \leq 1/2$. Recall

$$b = \frac{1}{2\gamma - 1} \leq 1 \text{ and } p = \gamma b.$$

As with the proof of Theorem 34, this will be a proof by induction. We wish to prove that $e_k \leq C_e k^{-b}$ for all $k \geq 1$ for the constant C_e defined as $C_e = \left(\kappa^2 b\right)^b$. The initial condition is $e_1 \leq C_e$ which is implied by $C_e \geq \Omega_1$. This in turn is implied by (4.58).

Now we assume $e_k = aC_e k^{-b}$ for some $k \geq 1$ and $a \in [0,1]$ and will show that $e_{k+1} \leq C_e(k+1)^{-b}$. Using the inductive assumption in the main recursion (4.50) yields

$$aC_e k^{-b} + \left(\alpha_1^2 G^2 - 2\alpha_1 c a^\gamma C_e^\gamma\right) k^{-2p} \leq C_e(k+1)^{-b}, \tag{4.59}$$

where we used the fact that $p + \gamma b = 2p$. We need this to hold for all $a \in [0,1]$. Since the L.H.S. is concave in a for $\gamma \geq 1$, we compute the maximizer as follows. Let $D_1 = \alpha_1^2 G^2 k^{-2p}$, $D_2 = C_e k^{-b}$, and $D_3 = 2\alpha_1 c C_e^\gamma k^{-2\gamma b}$. Then let

$$f(a) = D_1 + D_2 a - D_3 a^\gamma.$$

Let a_* be the solution to $0 = f'(a_*) = D_2 - \gamma D_3 a_*^{\gamma-1}$ which implies

$$\begin{aligned} a_* &= \left(\frac{D_2}{\gamma D_3}\right)^{\frac{1}{\gamma-1}} \\ &= C_e^{-1}(2\alpha_1\gamma c)^{\frac{1}{1-\gamma}} k^{\frac{1}{\gamma-1}} = C_e^{-1} D_4 \alpha^{\frac{1}{1-\gamma}} k^{\frac{1}{\gamma-1}}, \end{aligned}$$

where $D_4 = (2\gamma c)^{\frac{1}{1-\gamma}}$. But recall that $a \in [0,1]$, therefore the maximizer of $f(a)$ in $[0,1]$ is given by

$$\min\{1, C_e^{-1} D_4 \alpha^{\frac{1}{1-\gamma}} k^{\frac{1}{\gamma-1}}\}.$$

Thus if

$$k \geq (C_e D_4^{-1})^{\gamma-1} \alpha_1 \tag{4.60}$$

then the maximizer is equal to 1.

The analysis with $a = 1$ is the same as for this case where $\theta \geq 1/2$ given in Theorem 34. Recall from that proof that the choice of stepsize and constant, $\alpha_1 = \frac{c}{G^2} C_e^\gamma$ and $C_e = (\kappa^2 b)^b$, implies that the inequality (4.59) is satisfied for all $k \geq 1$. Substituting these values into (4.60) yields

$$k \quad \geq \quad (C_e D_4^{-1})^{\gamma-1} \frac{c}{G^2} C_e^\gamma = \frac{2\gamma}{2\gamma - 1}.$$

Since $\gamma \geq 1$ this is implied by $k \geq 2$.

On the other hand, if $k = 1$, then (4.59) becomes

$$a C_e + \alpha_1^2 G^2 - 2\alpha_1 c C_e^\gamma a^\gamma \leq C_e 2^{-b} \quad \forall a \in [0,1]. \tag{4.61}$$

The maximizer of the L.H.S. is $a_* = 1 - \frac{1}{2\gamma}$. The L.H.S. of (4.61) is a convex quadratic in α_1. Solving it yields an upper bound and a lower bound on α_1. We will now verify that our choice for α_1 given in (4.48) satisfies the two inequalities. Recall the choice for α_1:

$$\alpha_1 = \frac{c C_e^\gamma}{G^2}. \tag{4.62}$$

First the upper bound:

$$\alpha_1 \quad \leq \quad \frac{c C_e^\gamma a_*^\gamma}{G^2} \left(1 + \sqrt{C_e^{2\gamma} - \kappa^2 C_e a_*^{-2\gamma}(a_* - 2^{-b})}\right).$$

Simplifying further

$$\begin{aligned}\alpha_1 &\leq \frac{cC_e^\gamma a_*^\gamma}{G^2}\left(1 + C_e^\gamma\sqrt{1+\kappa^2 C_e^{1-2\gamma}a_*^{-2\gamma}(2^{-b}-a_*)}\right)\\ &= \frac{cC_e^\gamma a_*^\gamma}{G^2}\left(1 + C_e^\gamma\sqrt{1+b^{-2}a_*^{-2\gamma}(2^{-b}-a_*)}\right).\end{aligned}$$

It can be verified that $a_* = (1-\frac{1}{2\gamma}) \leq 2^{-b} = 2^{\frac{1}{1-2\gamma}}$ for all $\gamma \geq 1$. Therefore the term inside the square-root is greater than or equal to 1. Thus our choice of α_1 in (4.62) is viable if

$$a_*^\gamma(1+C_e^\gamma) \geq 1 \iff C_e^\gamma \geq a_*^{-\gamma} - 1.$$

Simplifying yields

$$\begin{aligned}(\kappa^2 b)^{\gamma b} \geq a_*^{-\gamma} - 1 \iff \kappa^2 &\geq \frac{1}{b}\left(a_*^{-\gamma b} - 1\right)^{\frac{1}{\gamma b}}\\ &= (2\gamma-1)\left(\left(1-\frac{1}{2\gamma}\right)^{\frac{\gamma}{1-2\gamma}} - 1\right)^{2-\frac{1}{\gamma}}.\end{aligned}$$

It can be confirmed numerically that for $\gamma \geq 1$ this is implied by $\kappa \geq 1$. Finally, the lower bound on α_1 is

$$\alpha_1 \geq \frac{cC_e^\gamma a_*^\gamma}{G^2}\left(1 - \sqrt{C_e^{2\gamma} - \kappa^2 C_e a_*^{-2\gamma}(a_* - 2^{-b})}\right).$$

Since $a_* \leq 1$ and the term in parantheses is less than or equal to 1, our choice for α_1 in (4.62) satisfies this inequality.

4.9 Numerical Experiment

In this section we present the results of a simulation to demonstrate some of the theoretical findings in this chapter. We consider an example satisfying HEB(c,θ) with $\theta = 1$ to test our proposed descending stairs stepsize choice in DS-SG and our "double descending stairs" method for unknown c, DS2-SG. Consider the following problem:

$$\min_x \|Ex - b\|_1 : \quad \|x\|_1 \leq \tau. \tag{4.63}$$

This objective function is used in regression problems where one observes $b = Ex + \eta$ and would like to recover x, given that η is some unknown noise term. If η is Gaussian, then the maximum likelihood estimator is the least-squares minimizer. However, if the noise is known to contain several outliers, or equivalently is sampled from a distribution with "heavier tails" then the Gaussian, the least absolute deviation loss is a more robust choice as the resulting estimator is less sensitive to outliers [127]. The ℓ_1 box constraint is used to encourage a sparse solution x. In the context of regression, enforcing sparsity makes sense when only a small subset of the features is actually correlated with the target variable [122]. The statistical estimation properties of (4.63) were discussed in [128, 129, 130].

Besides the subgradient techniques consider in this chapter, there are a few other methods which can tackle Prob. (4.63). The problem can be written as a linear program and solved via any LP solver. A popular option is an interior point method. These are second-order methods that rely on computing second-order information and solving potentially large linear systems at each iteration. Unfortunately they are not competitive with subgradient methods on large scale problems. Simplex methods are another option [131]. While their typical performance is good, these methods have exponential computational complexity in the worst case. The alternating direction method of multipliers (ADMM) is another approach to solving Prob. (4.63), however it involves solving a quadratic program at each iteration, placing it in the same computational regime as the interior point methods [16]. The primal-dual splitting method of [80] is a first-order method which can tackle Prob. (4.63). The main drawback of the method is that one must know the largest singular value of E in order to choose the stepsizes correctly. As such, it is not directly comparable with the subgradient methods developed in this chapter which do not require this information. The paper [128] introduces a method for solving Prob. (4.63) which is similar to the LARS method for solving the LASSO [132]. The method solves Prob. (4.63) for an increasing sequence of τ. At every iteration it solves a linear system, using the previous solution in a smart way. However, as far as we are aware, the iteration complexity of this method is unknown. Edgeworth's algorithm is a coordinate descent method for Prob. (4.63) which has shown promising empirical performance [133]. However unlike the subgradient methods considered here, the method is not guaranteed to converge to a minimizer. In fact specific examples exist where Edgeworth's algorithm converges to a non-optimal point [134].

Problem (4.63) is a polyhedral optimization problem therefore HEB(c, θ) is satisfied for all x with $\theta = 1$ [109]. However, it is not easy to compute c. Note that the constraint set is compact thus DS2-SG is applicable. Projection onto the ℓ_1 ball can be done in linear time in expectation via the method of [135].

To test the subgradient methods we consider a small synthetic instance of Problem (4.63). We set $m = 100$ and $n = 50$ and construct E of size $m \times n$ with i.i.d. $\mathcal{N}(0, 1)$ entries. We construct b of size $m \times 1$ with i.i.d. $\mathcal{N}(0, 1)$ entries. We set $\tau = 1$, which was chosen to obtain a fairly sparse solution with only 10% of its entries not equal to 0. All tested algorithms were randomly initialized to the same point. The purpose of this small experiment is to test some of the theoretical findings made in this chapter.

To start we test the convergence rates predicted by Theorem 30 for decaying stepsizes. We consider two stepsizes $\alpha_k^i = \alpha_{0,i} k^{-p_i}$ for $i = 1, 2$. These are $(\alpha_{0,1}, p_1) = (0.1, 0.99)$ and $(\alpha_{0,2}, p_2) = (0.01, 0.5)$, where the constant was tuned to achieve good performance. In Fig. 4.1 we plot the log of $d(x_k, \mathcal{X}_h)^2$ versus $\log_{10} k$, where k is the number of iterations. An optimal solution x^* is estimated by running DS-SG until it converges to within numerical precision. Looking at the figure it appears that for $k > 100$ the convergence rates are as predicted in Theorem 30. Specifically for the first parameter choice, $d(x_k, \mathcal{X}_h)^2 \approx O(k^{-1.98})$ and for the second $d(x_k, \mathcal{X}_h)^2 \approx O(k^{-1})$.

The figure confirms that DS-SG has a linear convergence rate, verifying Theorem 26. Its performance is very similar to Shor's method. While RSG does appear to obtain linear convergence, its rate is slower than DS-SG and Shor's method. Also observe that for the first 15000 iterations, the diminishing stepsize with $\alpha_k = O(k^{-1})$ is the best performing method. This is because the three linearly convergent methods are all highly sensitive to the condition number G/c, which can be large. This suggests that diminishing stepsize rules can still play a role on highly ill-conditioned polyhedral optimization problems.

As was mentioned we had to tune c to get good performance of DS-SG, RSG, and Shor's method. We now compare these three methods with our proposed 'doubling trick' variant DS2-SG, which does not need the value of c. We also compare with the method R^{2}SG proposed in [109]. Note that this method only works for $\theta < 1$ so following the advice of [109], we use the approximate value of $\theta = 0.8$. We initialize DS2-SG with the same parameters as DS-SG but with $c_1 = G = 160$. To demonstrate the effect of poorly chosen c in DS-SG, RSG, and Shor's method, we set $c = 100$ for all these methods (recall the tuned values were smaller). The results are given in Fig. 4.3. We compare function values and for each algorithm we keep track of the iterate with the smallest function value so far. This is because for R^{2}SG and DS2-SG, a large increase in objective function value often occurs every time a smaller estimate of c is tried. All the rates we

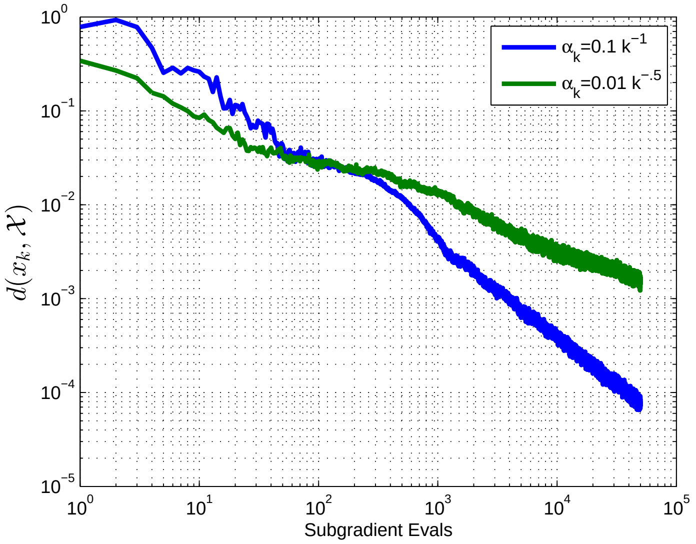

Figure 4.1: Problem (4.63): Log of square distance to the (unique) solution vs log of number of subgradient evaluations for decaying stepsizes with $(\alpha_1, p) = (0.1, 0.99)$ and $(\alpha_1, p) = (0.01, 0.5)$.

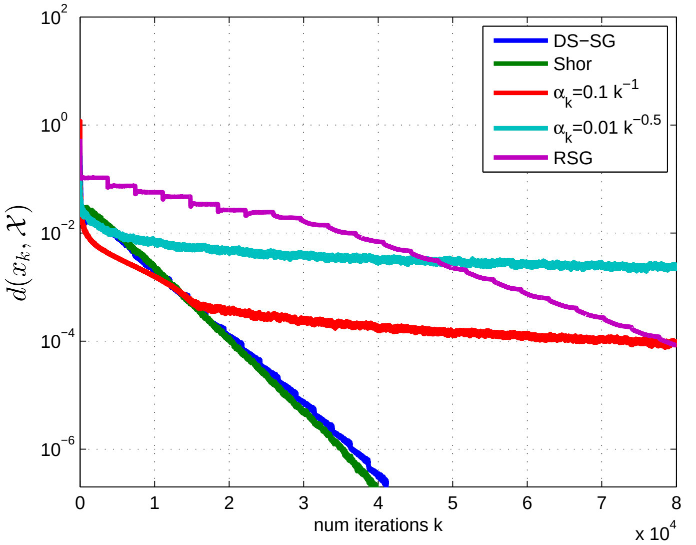

Figure 4.2: Problem (4.63): Log of square distance to the (unique) solution vs number of subgradient evaluations for DS-SG, RSG, and decaying stepsizes with $(\alpha_1, p) = (0.1, 0.99)$ and $(\alpha_1, p) = (0.01, 0.5)$.

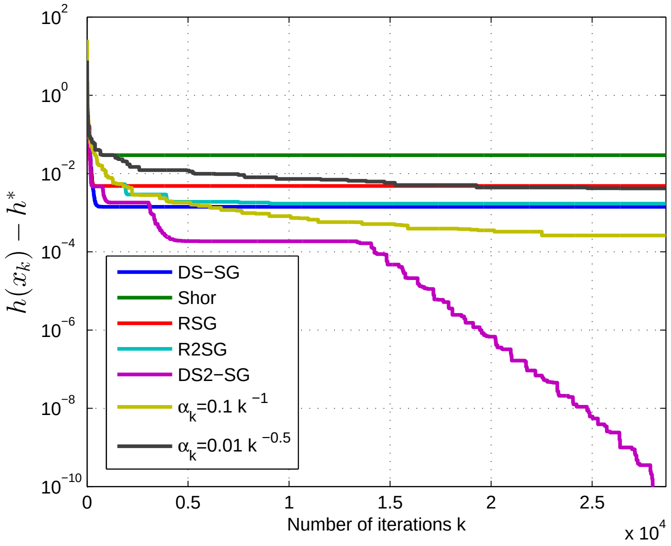

Figure 4.3: Problem (4.63): Log of $h(x) - h^*$ vs number of subgradient evaluations for DS-SG, RSG, and Shor's method with $c = 100$, R^{2}SG with and DS2-SG with the initial $c_1 = G = 160$.

derived for the last iterate also hold trivially for the best iterate. We see that DS-SG, RSG, and Shor's method converge to suboptimal solutions due to the incorrect value of c. However DS2-SG finds the correct solution to within the specified tolerance. This is even better than the performance of DS-SG and Shor's method with the parameter c tuned. R^{2}SG has slower convergence, which is not surprising since it is not guaranteed to obtain linear convergence on this problem. It is also encouraging that DS2-SG is faster than the summable decaying stepsize $\alpha_k = 0.1k^{-0.99}$, since this choice also does not require knowledge of c.

4.10 Proof of Theorems 28, 30, and 31

4.10.1 Preliminaries

In order to determine the convergence rate of the recursion (4.5) derived in Prop. 21 under nonsummable stepsizes, we need two Lemmas. We start with a result from [27] which considers (4.5) when $\theta < \frac{1}{2}$ without the nuisance term $\alpha_k^2 G^2$.

Lemma 36 *Suppose*

$$0 \leq u_{k+1} \leq u_k - \gamma_k u_k^{1+q}$$

for $k = 0, 1, \ldots$ where $\gamma_k \geq 0$ and $q > 0$. Then

$$u_k \leq u_0 \left(1 + q u_0^q \sum_{i=0}^{k-1} \gamma_i \right)^{-\frac{1}{q}}.$$

Proof [27, Lemma 6 pp. 46].

We will also use the following estimates for the sum of stepsizes $\sum_{i=k_0}^{k} \alpha_i$.

Lemma 37 *Let $k \geq k_0 \geq 1$.*

1. *If $p \in (0, 1)$*

$$\sum_{i=k_0}^{k} i^{-p} \geq \frac{(k+1)^{1-p} - k_0^{1-p}}{1-p}.$$

2. *If $p = 1$*

$$\sum_{i=k_0}^{k} i^{-p} \geq \ln \frac{k+1}{k_0}.$$

Proof A straightforward integral test.

4.10.2 Main Proof for Theorems 28 and 30

Continuing with the main analysis, the goal is to derive convergence rates for a sequence e_k satisfying (4.5). To this end, let

$$I = \{k : \alpha_k G^2 \geq c e_k^{\gamma}\}. \tag{4.64}$$

Recall the notation $\gamma = 1/(2\theta)$. We will consider three types of iterates and bound the convergence rate in each case. First, for those iterates $k \in I$ it is easy to derive the convergence rate. Second, we will bound the rate for an iterate in I^c when the previous iterate is in I. Finally we will consider s consecutive iterates in I^c, for which we can use the inequality in (4.64) to simplify recursion (4.5). Note that s can be arbitrarily large. In particular when I is finite there are an unbounded number of consecutive iterates in I^c. Together these three cases cover all possible iterates.

First for, $k \in I$ and $\alpha_k > 0$

$$\alpha_k c e_k^{\gamma} \leq \alpha_k^2 G^2 \implies e_k \leq \left(\frac{\alpha_k G^2}{c}\right)^{\frac{1}{\gamma}}.$$

Thus the rate of e_k is $O\left(\alpha_k^{\frac{1}{\gamma}}\right)$ for $k \in I$. In particular since $\alpha_k = \alpha_1 k^{-p}$, then for $k \in I$ and $\alpha_1 > 0$

$$e_k \leq \left(\frac{\alpha_1 G^2}{c}\right)^{2\theta} k^{-2p\theta}. \tag{4.65}$$

Now assume $k \in I$ and $k+1 \in I^c$. Then

$$e_{k+1} \leq e_k + \alpha_k^2 G^2 \leq \left(\frac{\alpha_k G^2}{c}\right)^{\frac{1}{\gamma}} + \alpha_k^2 G^2. \tag{4.66}$$

Now since $\frac{1}{\gamma} = 2\theta \in (0,2)$, for $k \geq 1$

$$k^{-2p\theta} \geq k^{-2p}.$$

Therefore (4.66) implies that for $k \in I$, $k+1 \in I^c$, and $k \geq 1$,

$$e_{k+1} \quad \leq \quad C_1 (k+1)^{-2p\theta}, \tag{4.67}$$

where

$$C_1 = 2^{2p\theta}\left(\left(\frac{\alpha_1 G^2}{c}\right)^{\frac{1}{\gamma}} + \alpha_1^2 G^2\right).$$

Next assume $k \in I$, $k+1 \in I^c$, and $k+i \in I^c$ for $i = 2, \ldots s$ for some $s \geq 2$. Then for $i = 2, \ldots s$

$$e_{k+i} < e_{k+i-1} - \alpha_k c e_{k+i-1}^{\gamma}. \tag{4.68}$$

To analyze the recursion (4.68) we consider $\theta < \frac{1}{2}$ and $\theta \geq \frac{1}{2}$ separately.

<u>Case 1</u>: $\boldsymbol{\theta < \frac{1}{2}}$.

Now since $\gamma > 1$ we can apply Lemma 36 to (4.68) and derive for $i = 2, \ldots, s$

$$e_{k+i} \leq e_{k+1}\left[1 + \frac{1-2\theta}{2\theta} e_{k+1}^{\frac{1-2\theta}{2\theta}} \sum_{j=1}^{i-1} \alpha_{k+j}\right]^{\frac{2\theta}{2\theta-1}}.$$

We then use Lemma 37 to derive

$$\begin{aligned} & \left[1 + \frac{1-2\theta}{2\theta} e_{k+1}^{\frac{1-2\theta}{2\theta}} \sum_{j=1}^{i-1} \alpha_{k+j}\right]^{\frac{2\theta}{2\theta-1}} \\ \leq \quad & \left[1 + \frac{\alpha_1(1-2\theta)}{2\theta(1-p)} e_{k+1}^{\frac{1-2\theta}{2\theta}} \left((k+i)^{1-p} - (k+1)^{1-p}\right)\right]^{\frac{2\theta}{2\theta-1}}. \end{aligned} \tag{4.69}$$

Now consider the condition given in (4.42). Note that since p satisfies (4.41), if (4.42) holds for $k = k_0$, it holds for all $k > k_0$. In particular if it holds for $k = 0$, then it holds for all k. Continuing, if (4.42) holds then for all $k > k_0$

$$1 - \frac{\alpha_1(1-2\theta)}{2\theta(1-p)} e_{k+1}^{\frac{1-2\theta}{2\theta}} (k+1)^{1-p} \geq 0, \tag{4.70}$$

where we have used the fact that $k+1 \in I^c$. Therefore since (4.70) holds we can simplify (4.69) to say that for $k \in I$ and $k+i \in I^c$ for $i = 2, 3, \ldots, s$, and $k > k_0$,

$$\begin{aligned} e_{k+i} &\leq e_{k+1} \left[\frac{\alpha_1(1-2\theta)}{2\theta(1-p)} e_{k+1}^{\frac{1-2\theta}{2\theta}} (k+i)^{1-p} \right]^{\frac{2\theta}{2\theta-1}} \\ &\leq \left(\frac{\alpha_1(1-2\theta)}{2\theta(1-p)} \right)^{\frac{2\theta}{2\theta-1}} (k+i)^{\frac{2\theta(1-p)}{2\theta-1}}. \end{aligned} \tag{4.71}$$

The final case to consider is when $i = 1, 2, \ldots, s$ are in I^c. In this case, the same bound (4.69) can be derived but with e_1 replacing e_{k+1}. Thus for $i = 2, 3, \ldots s$ in I

$$e_i \leq e_1 \left[1 + \frac{\alpha_1(1-2\theta)}{2\theta(1-p)} e_1^{\frac{1-2\theta}{2\theta}} \left(i^{1-p} - 1 \right) \right]^{\frac{2\theta}{2\theta-1}}.$$

Thus if α_1 is chosen to satisfy (4.43) then

$$e_i \leq \left(\frac{\alpha_1(1-2\theta)}{2\theta(1-p)} \right)^{\frac{2\theta}{2\theta-1}} i^{\frac{2\theta(1-p)}{2\theta-1}}. \tag{4.72}$$

Combining (4.65), (4.67), (4.71), and (4.72) establishes (4.44) and concludes the proof of Theorem 28.

<u>Case 2</u>: $\boldsymbol{\theta \geq \frac{1}{2}}$

Next we consider the case where $\frac{1}{2} \leq \theta \leq 1$ which will finish the proof of Theorem 30. Before commencing we introduce the following Lemma which allows us to bound a decaying exponential by an appropriately scaled decaying polynomial of any degree.

Lemma 38 *Suppose $\delta > 0$, then if $C_\delta \geq e^{-\delta}\delta^\delta$,*

$$\exp(-x) \leq C_\delta x^{-\delta} \quad \forall x > 0. \tag{4.73}$$

Proof Taking logs of both sides of (4.73) yields

$$-x \leq -\delta \ln x + \beta_\delta \quad \forall x > 0,$$

where $\beta_\delta = \ln C_\delta$. Therefore

$$\beta_\delta \quad \geq \quad \delta \ln x - x \quad \forall x > 0$$

which implies

$$\beta_\delta \quad \geq \quad \max_{x>0}\{\delta \ln x - x\}.$$

The right hand side is a smooth concave coercive maximization problem which therefore has a unique solution given by $x^* = \delta$. Therefore

$$\beta_\delta \geq \delta \ln \delta - \delta$$

which implies the Lemma.

Continuing, we consider $k \in I$, $k+1 \in I^c$, and $k+i \in I^c$ for $i = 2\ldots, s$ in the case where $\theta \geq \frac{1}{2}$, so $\gamma \leq 1$. Then since $k+i \in I^c$ for $i = 2, \ldots s$,

$$0 \leq \frac{e_{k+i-1}}{e_{k+1}} \leq 1 \implies \left(\frac{e_{k+i-1}}{e_{k+1}}\right)^\gamma \geq \frac{e_{k+i-1}}{e_{k+1}} \implies e_{k+i-1}^\gamma \geq e_{k+1}^{\gamma-1} e_{k+i-1}.$$

Thus for $k \in I$, $k+1 \in I^c$, and $k+i \in I^c$ for $i = 2, \ldots, s$ for some $s \geq 2$

$$\begin{aligned} e_{k+i} &\leq e_{k+i-1} - \alpha_{k+i-1} c e_{k+i-1}^\gamma \\ &\leq e_{k+i-1} - \alpha_{k+i-1} e_{k+1}^{\gamma-1} c e_{k+i-1}. \end{aligned} \tag{4.74}$$

Now taking logs and using $\log(1-x) \leq -x$,

$$\begin{aligned} \ln e_{k+i} &\leq \ln e_{k+i-1} + \ln(1 - e_{k+1}^{\gamma-1} c \alpha_{k+i-1}) \\ &\leq \ln e_{k+i-1} - e_{k+1}^{\gamma-1} c \alpha_{k+i-1}. \end{aligned}$$

Now summing and using Lemma 37

$$\begin{aligned} \ln e_{k+i} &\leq \ln e_{k+1} - \alpha_1 e_{k+1}^{\gamma-1} c \sum_{i=k+1}^{k+i-1} i^{-p} \\ &\leq \ln e_{k+1} - \frac{\alpha_1 e_{k+1}^{\gamma-1} c}{1-p} \left((k+i)^{1-p} - (k+1)^{1-p}\right). \end{aligned}$$

This leads to

$$\begin{aligned} e_{k+i} &\leq e_{k+1} \exp\left\{ -\frac{\alpha_1 e_{k+1}^{\gamma-1} c}{1-p} \left((k+i)^{1-p} - (k+1)^{1-p} \right) \right\} \qquad (4.75) \\ &= \exp\left\{ -\frac{\alpha_1 e_{k+1}^{\gamma-1} c (k+i)^{1-p}}{1-p} \left(1 - \left(\frac{k+1}{k+i} \right)^{1-p} \right) \right\}. \end{aligned}$$

We further consider two possible cases. If $i \geq k$, then

$$\frac{k+1}{k+i} \leq \frac{k+1}{k+k} = \frac{1}{2} + \frac{1}{2k},$$

therefore by concavity of t^{1-p}

$$\left(\frac{k+1}{k+i} \right)^{1-p} \leq 2^{p-1} \left[1 + \frac{1-p}{k} \right].$$

Take $k > 3$ so that

$$\frac{2^{p-1}(1-p)}{k} \leq \frac{1-2^{p-1}}{2}.$$

Hence

$$1 - \left(\frac{k+1}{k+i} \right)^{1-p} \geq 1 - 2^{p-1} \left[1 + \frac{1-p}{k} \right] \geq 1 - 2^{p-1} - \frac{2^{p-1}(1-p)}{k} \geq \frac{1-2^{p-1}}{2}.$$

Hence if $3 < k \leq i$ then

$$e_{k+i} \leq e_{k+1} \exp\left(-\frac{(1-2^{p-1}) \alpha_1 e_{k+1}^{\gamma-1} c}{2(1-p)} (k+i)^{1-p} \right).$$

Now by Lemma 38 for any $\delta_1 > 0$,

$$\begin{aligned} &\exp\left\{ -\frac{\alpha_1 (1-2^{p-1}) c e_{k+1}^{\gamma-1}}{2(1-p)} (k+i)^{1-p} \right\} \\ &\leq \delta_1^{\delta_1} e^{-\delta_1} e_{k+1}^{1+\delta_1(1-\gamma)} \left(\frac{\alpha_1 (1-2^{p-1}) c}{2(1-p)} (k+i)^{1-p} \right)^{-\delta_1}. \end{aligned}$$

Therefore using (4.67) for any $k \leq i$ and $k > 3$

$$e_{k+i} \leq \delta_1^{\delta_1} C_1^{1+\delta_1(1-\gamma)} \left(\frac{\alpha_1 (1-2^{p-1}) c e}{2(1-p)} \right)^{-\delta_1} (k+i)^{-\delta_1(1-p)}. \qquad (4.76)$$

Taking $\delta_1 = \frac{2p\theta}{1-p}$ and simplifying (4.76) yields

$$e_{k+i} \leq C_1^{\frac{1+2p(\theta-1)}{1-p}} \left(\frac{\alpha_1(1-2^{p-1})ce}{4p\theta} \right)^{-\frac{2p\theta}{1-p}} (k+i)^{-2p\theta}. \tag{4.77}$$

Next consider $k \geq i > 1$. Now

$$\begin{aligned}
(k+i)^{1-p} - (k+1)^{1-p} &= (k+i)^{1-p} \left(1 - \left(\frac{k+1}{k+i} \right)^{1-p} \right) \\
&= (k+i)^{1-p} \left(1 - \left(1 - \frac{i-1}{k+i} \right)^{1-p} \right) \\
&\geq (k+i)^{1-p} \left(1 - \left(1 - \frac{i-1}{2k} \right)^{1-p} \right) \\
&\geq \frac{(1-p)(k+i)^{1-p}(i-1)}{2k} \qquad (4.78) \\
&\geq \frac{1-p}{2} k^{-p}(i-1),
\end{aligned}$$

where in (4.78) we used the concavity of t^{1-p}. Thus plugging this into (4.75) implies for $k \geq i$

$$e_{k+i} \leq e_{k+1} \exp \left(\frac{-\alpha_1 e_{k+1}^{\gamma-1} c(i-1)}{2k^p} \right).$$

Therefore for all $\delta_2 \geq 0$ it follows Lemma 38 that

$$\begin{aligned}
e_{k+i} &\leq e_{k+1} \exp \left(\frac{-\alpha_1 e_{k+1}^{\gamma-1} c(i-1)}{2k^p} \right) \\
&\leq \delta_2^{\delta_2} e_{k+1} \left(\frac{\alpha_1 e_{k+1}^{\gamma-1} c(i-1)e}{2k^p} \right)^{-\delta_2} \\
&\leq C_1^{1+\delta_2(1-\gamma)} \left(\frac{4\delta_2}{c\alpha_1 e} \right)^{\delta_2} k^{-2p\theta(1+\delta_2(1-\gamma))} k^{p\delta_2} i^{-\delta_2}, \qquad (4.79)
\end{aligned}$$

where we used $e_{k+1} \leq C_1 k^{-2p\theta}$ and $(i-1)^{-\delta_2} \leq 2^{\delta_2} i^{-\delta_2}$. Now if we choose

$$\delta_2 = 2\theta$$

then (4.79) implies

$$e_{k+i} \leq C_4 i^{-2\theta}, \tag{4.80}$$

where

$$C_4 = \left(\frac{8\theta C_1}{c\alpha_1 e} \right)^{2\theta}.$$

Thus combining $e_{k+i} \leq e_{k+1} \leq C_1 k^{-2p\theta}$ and (4.80) implies that for $i \leq k$

$$e_{k+i} \leq \max\{C_1, C_4\} \min\{k^{-2p\theta}, i^{-2\theta}\}.$$

Now since $-2\theta < -2p\theta$,

$$e_{k+i} \leq \max\{C_1, C_4\} \min\{k^{-2p\theta}, i^{-2p\theta}\} \leq \frac{\max\{C_1, C_4\}}{\max\{k^{2p\theta}, i^{2p\theta}\}}.$$

If $2p\theta \geq 1$ then by convexity of $t^{2p\theta}$

$$\max\{k^{2p\theta}, i^{2p\theta}\} \geq \frac{1}{2} \left(k^{2p\theta} + i^{2p\theta} \right) \geq 2^{-2p\theta} (k+i)^{2p\theta}. \tag{4.81}$$

On the other hand if $2p\theta < 1$ then because $t^{2p\theta}$ is subadditive

$$\max\{k^{2p\theta}, i^{2p\theta}\} \geq \frac{1}{2} \left(k^{2p\theta} + i^{2p\theta} \right) \geq \frac{1}{2} (k+i)^{2p\theta}. \tag{4.82}$$

Combining (4.81) and (4.82) gives

$$e_{k+i} \leq 4 \max\{C_1, C_4\} (k+i)^{-2p\theta}. \tag{4.83}$$

Finally we consider the case where the first s iterates belong to I^c. Therefore, using (4.75), for $i = 1, 2, \ldots, s$

$$e_i \quad \leq \quad e_1 \exp\left\{ -\frac{\alpha_1 e_1^{\gamma-1} c}{1-p} \left(i^{1-p} - 1 \right) \right\}.$$

Now since for $x \geq 1$, $x - 1 \geq \frac{x}{2}$, this implies that

$$e_i \quad \leq \quad e_1 \exp\left\{ -\frac{\alpha_1 e_1^{\gamma-1} c}{2(1-p)} i^{1-p} \right\}.$$

Using Lemma 38 this implies that for any $\delta_3 > 0$

$$e_i \quad \leq \quad e^{-\delta_3} \delta_3^{\delta_3} e_1 \left(\frac{\alpha_1 e_1^{\gamma-1} c}{2(1-p)} i^{1-p} \right)^{-\delta_3}, \tag{4.84}$$

and we will use $\delta_3 = \frac{2p\theta}{1-p}$.

Combining (4.65), (4.67), (4.77), (4.83), and (4.84) yields the desired result (4.46) and concludes the proof of Theorem 30.

4.10.3 Proof of Theorem 31

The format of the proof is identical to Theorems 28 and 30. As before it is based on the set I defined in (4.64) and we consider three types of iterates. First we bound the convergence rate for iterates in I, second for iterates in I^c when the previous iterate is in I. And finally for s consecutive iterates in I^c where s may be unbounded.

If $k \in I$ then repeating (4.67) yields

$$e_k \leq \frac{\alpha_1 G^2}{c} k^{-1}. \tag{4.85}$$

Similarly for $k \in I$ and $k+1 \in I^c$,

$$e_{k+1} \leq \frac{2\alpha_1 G^2}{c}(k+1)^{-1}. \tag{4.86}$$

Finally for $k \in I$, $k+1 \in I^c$, and $k+i \in I^c$, for $i = 2, \ldots, s$, then repeating (4.74) but with $\gamma = 1$ this time,

$$e_{k+i} \leq e_{k+i-1}(1 - c\alpha_{k+i-1}).$$

Taking logs, using $\log(1-x) \leq -x$ and summing yields

$$\begin{aligned} \log e_{k+i} &\leq \log e_{k+1} - c\alpha_1 \sum_{j=k+1}^{k+i-1} j^{-1} \\ &\leq \log e_{k+1} - c\alpha_1 \left(\log(k+i) - \log(k+1)\right), \end{aligned}$$

where we applied Lemma 37 in the second inequality. This yields for all $k \in I$ and $k+i \in I^c$ for $i = 2, 3, \ldots, s$ for some $s \in \mathbb{N}$

$$e_{k+i} \leq e_{k+1} \left(\frac{k+i}{k+1}\right)^{-c\alpha_1}. \tag{4.87}$$

Using (4.86) yields

$$\begin{aligned} e_{k+i} &\leq \frac{2\alpha_1 G^2}{c}(k+1)^{-1}(k+1)^{c\alpha_1}(k+i)^{-c\alpha_1} \\ &\leq \frac{2\alpha_1 G^2}{c}(k+i)^{-c\alpha_1}. \end{aligned} \tag{4.88}$$

Finally we consider the case where the initial iterates $i = 1, 2, \ldots, s$ are in I^c. Therefore repeating (4.87) with $k = 0$ gives

$$e_i \leq e_1 i^{-c\alpha_1}. \tag{4.89}$$

Combining (4.85), (4.86), (4.88), and (4.89) yields (4.47) and concludes the proof of Theorem 31.

4.10.4 Proof of Proposition 32

As previously mentioned, this argument is a direct extension of [100, Thm. 4]. For $\theta = \frac{1}{2}$, (4.5) reads as

$$e_{k+1} \leq (1 - 2\alpha_k c)e_k + \alpha_k^2 G^2.$$

We consider the choice $\alpha_k = \frac{2k+1}{2c(k+1)^2}$. Then

$$e_{k+1} \leq \left(1 - \frac{2k+1}{(k+1)^2}\right) e_k + \frac{G^2(2k+1)^2}{4c^2(k+1)^4}.$$

Multiplying both sides by $(k+1)^2$ yields

$$\begin{aligned}(k+1)^2 e_{k+1} &\leq k^2 e_k + \frac{G^2(2k+1)^2}{4c^2(k+1)^2} \\ &\leq k^2 e_k + \frac{G^2}{c^2} \\ &\leq e_1 + \frac{G^2}{c^2}k.\end{aligned}$$

Therefore

$$e_{k+1} \leq \frac{e_1}{(k+1)^2} + \frac{G^2}{c^2(k+1)}.$$

REFERENCES

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer Science & Business Media, 2011.

[2] F. Bach, R. Jenatton, J. Mairal, G. Obozinski et al., "Convex optimization with sparsity-inducing norms," *Optimization for Machine Learning*, pp. 19–53, 2011.

[3] E. T. Hale, W. Yin, and Y. Zhang, "Fixed-point continuation for ℓ_1-minimization: methodology and convergence," *SIAM J. on Optimization*, vol. 19, no. 3, pp. 1107–1130, Oct. 2008.

[4] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Img. Sci.*, vol. 2, no. 1, pp. 183–202, Mar. 2009. [Online]. Available: http://dx.doi.org/10.1137/080716542

[5] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering.* Springer, 2011, pp. 185–212.

[6] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," *Signal Processing Magazine, IEEE*, vol. 31, no. 5, pp. 32–43, 2014.

[7] R. J. Tibshirani, "The lasso problem and uniqueness," *Electronic Journal of Statistics*, vol. 7, pp. 1456–1490, 2013.

[8] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *Image Processing, IEEE Transactions on*, vol. 19, no. 9, pp. 2345–2356, 2010.

[9] K. Choi, J. Wang, L. Zhu, T.-S. Suh, S. Boyd, and L. Xing, "Compressed sensing based cone-beam computed tomography reconstruction with a first-order method," *Medical Physics*, vol. 37, no. 9, pp. 5113–5125, 2010.

[10] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[11] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 1, no. 4, pp. 606–617, 2007.

[12] A. Chambolle and C. Dossal, "On the convergence of the iterates of the fast iterative shrinkage/thresholding algorithm," *Journal of Optimization Theory and Applications*, vol. 166, no. 3, pp. 968–982.

[13] J. Liang, J. Fadili, and G. Peyré, "Local linear convergence of forward–backward under partial smoothness," in *Advances in Neural Information Processing Systems*, 2014, pp. 1970–1978.

[14] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.

[15] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, 2013.

[16] J. Eckstein and W. Yao, "Augmented Lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results," *RUTCOR Research Reports*, vol. 32, 2012.

[17] G. B. Passty, "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space," *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2, pp. 383–390, 1979.

[18] D. A. Lorenz and T. Pock, "An inertial forward-backward algorithm for monotone inclusions," *Journal of Mathematical Imaging and Vision*, vol. 51, no. 2, pp. 311–325, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10851-014-0523-2

[19] D. Davis and W. Yin, *Convergence Rate Analysis of Several Splitting Schemes.* Cham: Springer International Publishing, 2016, pp. 115–163. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-41589-5_4

[20] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1168–1200, 2005.

[21] J. Y. Bello Cruz and T. T. Nghia, "On the convergence of the forward–backward splitting method with linesearches," *Optimization Methods and Software*, vol. 31, no. 6, pp. 1209–1238, 2016.

[22] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010.

[23] S. Shalev-Shwartz and A. Tewari, “Stochastic methods for ℓ_1-regularized loss minimization,” *The Journal of Machine Learning Research*, vol. 12, pp. 1865–1892, 2011.

[24] F. Alvarez, “On the minimizing property of a second order dissipative system in Hilbert spaces,” *SIAM Journal on Control and Optimization*, vol. 38, no. 4, pp. 1102–1119, 2000.

[25] H. Attouch, J. Peypouquet, and P. Redont, “A dynamical approach to an inertial forward-backward algorithm for convex minimization,” *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 232–256, 2014.

[26] W. Su, S. Boyd, and E. Candès, “A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2510–2518.

[27] B. T. Polyak, *Introduction to Optimization.* Optimization Software Inc., 1987.

[28] B. T. Polyak, “Some methods of speeding up the convergence of iteration methods,” *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[29] Y. Nesterov, “A method of solving a convex programming problem with convergence rate $O(1/k^2)$,” in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

[30] O. Güler, “New proximal point algorithms for convex minimization,” *SIAM Journal on Optimization*, vol. 2, no. 4, pp. 649–664, 1992.

[31] P. Tseng, “On accelerated proximal gradient methods for convex-concave optimization,” *submitted to SIAM Journal on Optimization*, 2008.

[32] Y. Nesterov, *Introductory Lectures on Convex Optimization: a Basic Course.* Springer, 2004.

[33] A. Moudafi and M. Oliny, “Convergence of a splitting inertial proximal method for monotone operators,” *Journal of Computational and Applied Mathematics*, vol. 155, no. 2, pp. 447 – 454, 2003.

[34] P.-E. Maingé, “Convergence theorems for inertial KM-type algorithms,” *Journal of Computational and Applied Mathematics*, vol. 219, no. 1, pp. 223–236, 2008.

[35] A. Chambolle and T. Pock, “On the ergodic convergence rates of a first-order primal–dual algorithm,” *Mathematical Programming*, vol. 159, no. 1, pp. 253–287, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10107-015-0957-3

[36] R. I. Boţ, E. R. Csetnek, and C. Hendrich, "Inertial Douglas–Rachford splitting for monotone inclusion problems," *Applied Mathematics and Computation*, vol. 256, pp. 472–487, 2015.

[37] J. Liang, J. Fadili, and G. Peyré, "Activity identification and local linear convergence of inertial forward-backward splitting," *arXiv preprint arXiv:1503.03703*, 2015.

[38] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont, "Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity," *Mathematical Programming*, pp. 1–53, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10107-016-0992-8

[39] K. Bredies and D. A. Lorenz, "Linear convergence of iterative soft-thresholding," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 813–837, 2008.

[40] A. Agarwal, S. Negahban, and M. J. Wainwright, "Fast global convergence of gradient methods for high-dimensional statistical recovery," *Ann. Statist.*, vol. 40, no. 5, pp. 2452–2482, 10 2012. [Online]. Available: http://dx.doi.org/10.1214/12-AOS1032

[41] J. Liang, J. Fadili, and G. Peyré, "Convergence rates with inexact non-expansive operators," *Mathematical Programming*, vol. 159, no. 1, pp. 403–434, 2016. [Online]. Available: http://dx.doi.org/10.1007/s10107-015-0964-4

[42] S. Tao, D. Boley, and S. Zhang, "Local Linear Convergence of ISTA and FISTA on the LASSO Problem," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 313–336, 2016.

[43] W. Hare and A. S. Lewis, "Identifying active constraints via partial smoothness and prox-regularity," *Journal of Convex Analysis*, vol. 11, no. 2, pp. 251–266, 2004.

[44] H. Attouch and J. Peypouquet, "The rate of convergence of nesterov's accelerated forward-backward method is actually faster than $1/k^2$," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1824–1834, 2016.

[45] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Foundations of Computational Mathematics*, pp. 1–18, 2012.

[46] R. D. Monteiro, C. Ortiz, and B. F. Svaiter, "An adaptive accelerated first-order method for convex optimization," *Computational Optimization and Applications*, vol. 64, no. 1, pp. 31–73, 2016.

[47] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena Scientific, 1999.

[48] R. S. Burachik and B. Svaiter, "ε-enlargements of maximal monotone operators in banach spaces," *Set-Valued Analysis*, vol. 7, no. 2, pp. 117–132, 1999.

[49] Z. Wen, W. Yin, H. Zhang, and D. Goldfarb, "On the convergence of an active-set method for ℓ_1 minimization," *Optimization Methods and Software*, vol. 27, no. 6, pp. 1127–1146, 2012.

[50] H. Zhang, W. Yin, and L. Cheng, "Necessary and sufficient conditions of solution uniqueness in 1-norm minimization," *Journal of Optimization Theory and Applications*, vol. 164, no. 1, pp. 109–122, 2015.

[51] Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.

[52] Z. Opial, "Weak convergence of the sequence of successive approximations for nonexpansive mappings," *Bulletin of the American Mathematical Society*, vol. 73, no. 4, pp. 591–597, 1967.

[53] 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," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[54] T. Blumensath and M. E. Davies, "Iterative thresholding for Sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 629–654, Dec. 2008. [Online]. Available: http://dx.doi.org/10.1007/s00041-008-9035-z

[55] D. Lazzaro, "A nonconvex approach to low-rank matrix completion using convex optimization," *Numerical Linear Algebra with Applications*, vol. 23, no. 5, pp. 801–824, 2016.

[56] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[57] P. Chen and D. Suter, "Recovering the missing components in a large noisy low-rank matrix: Application to SFM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 8, pp. 1051–1063, 2004.

[58] H. H. Zhang, J. Ahn, X. Lin, and C. Park, "Gene selection using support vector machines with non-convex penalty," *Bioinformatics*, vol. 22, no. 1, pp. 88–95, 2006.

[59] E. J. Candes, Y. C. Eldar, T. Strohmer, and V. Voroninski, "Phase retrieval via matrix completion," *SIAM Review*, vol. 57, no. 2, pp. 225–251, 2015.

[60] H. Ji, C. Liu, Z. Shen, and Y. Xu, "Robust video denoising using low rank matrix completion," in *CVPR*. Citeseer, 2010, pp. 1791–1798.

[61] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[62] R. I. Boţ, E. R. Csetnek, and S. C. László, "An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions," *EURO Journal on Computational Optimization*, vol. 4, no. 1, pp. 3–25, 2016.

[63] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1139–1147.

[64] K. Kurdyka, "On gradients of functions definable in o-minimal structures," in *Annales de l'institut Fourier*, vol. 48, no. 3, 1998, pp. 769–783.

[65] S. Lojasiewicz, "Une propriété topologique des sous-ensembles analytiques réels," *Les Équations aux Dérivées Partielles*, pp. 87–89, 1963.

[66] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 438–457, 2010.

[67] H. Attouch and J. Bolte, "On the convergence of the proximal algorithm for nonsmooth functions involving analytic features," *Mathematical Programming*, vol. 116, no. 1-2, pp. 5–16, 2009.

[68] P. Frankel, G. Garrigos, and J. Peypouquet, "Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates," *Journal of Optimization Theory and Applications*, vol. 165, no. 3, pp. 874–900, 2015.

[69] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, "From error bounds to the complexity of first-order descent methods for convex functions," *Mathematical Programming*, pp. 1–37, 2015.

[70] G. Li and T. K. Pong, "Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods," *arXiv preprint arXiv:1602.02915*, 2016.

[71] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[72] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[73] J. Liang, J. Fadili, and G. Peyré, “A multi-step inertial forward–backward splitting method for non-convex optimization,” *arXiv preprint arXiv:1606.02118*, 2016.

[74] P. R. Johnstone and P. Moulin, “Local and global convergence of a general inertial proximal splitting scheme for minimizing composite functions,” *Computational Optimization and Applications*, pp. 1–34, 2017. [Online]. Available: http://dx.doi.org/10.1007/s10589-017-9896-7

[75] P. Ochs, Y. Chen, T. Brox, and T. Pock, “iPiano: Inertial proximal algorithm for nonconvex optimization,” *SIAM Journal on Imaging Sciences*, vol. 7, no. 2, pp. 1388–1419, 2014.

[76] S. Zavriev and F. Kostyuk, “Heavy-ball method in nonconvex optimization problems,” *Computational Mathematics and Modeling*, vol. 4, no. 4, pp. 336–341, 1993.

[77] Y. Xu and W. Yin, “A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion,” *SIAM Journal on imaging sciences*, vol. 6, no. 3, pp. 1758–1789, 2013.

[78] J. Bolte, S. Sabach, and M. Teboulle, “Proximal alternating linearized minimization for nonconvex and nonsmooth problems,” *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[79] A. Beck and M. Teboulle, “Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems,” *Image Processing, IEEE Transactions on*, vol. 18, no. 11, pp. 2419–2434, 2009.

[80] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[81] D. L. Donoho, “Compressed sensing,” *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.

[82] R. Tibshirani, “Regression shrinkage and selection via the LASSO,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[83] A. Chambolle, “An algorithm for total variation minimization and applications,” *Journal of Mathematical Imaging and Vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

[84] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions.* Springer Science & Business Media, 2012, vol. 3.

[85] J.-L. Goffin, “On convergence rates of subgradient optimization methods,” *Mathematical Programming*, vol. 13, no. 1, pp. 329–347, 1977.

[86] E. Rosenberg, “A geometrically convergent subgradient optimization method for nonlinearly constrained convex programs,” *Mathematics of Operations Research*, vol. 13, no. 3, pp. 512–523, 1988.

[87] A. Nedić and D. P. Bertsekas, “The effect of deterministic noise in subgradient methods,” *Mathematical Programming*, vol. 125, no. 1, pp. 75–99, 2010.

[88] A. Nedić and D. Bertsekas, “Convergence rate of incremental subgradient algorithms,” in *Stochastic Optimization: Algorithms and Applications.* Springer, 2001, pp. 223–264.

[89] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[90] G. Li, “Global error bounds for piecewise convex polynomials,” *Mathematical Programming*, vol. 137, no. 1-2, pp. 37–64, 2013.

[91] P. Tseng, “Approximation accuracy, gradient methods, and error bound for structured convex optimization,” *Mathematical Programming*, vol. 125, no. 2, pp. 263–295, 2010.

[92] Z. Zhou and A. M.-C. So, “A unified approach to error bounds for structured convex optimization problems,” *Mathematical Programming*, pp. 1–40, 2017. [Online]. Available: http://dx.doi.org/10.1007/s10107-016-1100-9

[93] Y. Xu, Q. Lin, and T. Yang, “Accelerate stochastic subgradient method by leveraging local error bound,” *arXiv preprint arXiv:1607.01027*, 2016.

[94] J. Burke and M. C. Ferris, “Weak sharp minima in mathematical programming,” *SIAM Journal on Control and Optimization*, vol. 31, no. 5, pp. 1340–1359, 1993.

[95] H. Zhang and W. Yin, “Gradient methods for convex minimization: better rates under weaker conditions,” *arXiv preprint arXiv:1303.4645*, 2013.

[96] J.-S. Pang, “Error bounds in mathematical programming,” *Mathematical Programming*, vol. 79, no. 1-3, pp. 299–332, 1997.

[97] Z.-Q. Luo and P. Tseng, “Error bounds and convergence analysis of feasible descent methods: a general approach,” *Annals of Operations Research*, vol. 46, no. 1, pp. 157–178, 1993.

[98] M. C. Ferris, “Finite termination of the proximal point algorithm,” *Mathematical Programming*, vol. 50, no. 1, pp. 359–366, 1991.

[99] J. Burke and S. Deng, “Weak sharp minima revisited part i: basic theory,” *Control and Cybernetics*, vol. 31, pp. 439–469, 2002.

[100] H. Karimi, J. Nutini, and M. Schmidt, “Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2016, pp. 795–811.

[101] A. Beck and S. Shtern, “Linearly convergent away-step conditional gradient for non-strongly convex functions,” *Mathematical Programming*, pp. 1–27, 2015.

[102] J. M. Borwein, G. Li, and L. Yao, “Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets,” *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 498–527, 2014.

[103] J. Bolte, A. Daniilidis, and A. Lewis, “The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems,” *SIAM Journal on Optimization*, vol. 17, no. 4, pp. 1205–1223, 2007.

[104] H. Zhang, “New analysis of linear convergence of gradient-type methods via unifying error bound conditions,” *arXiv preprint arXiv:1606.00269*, 2016.

[105] I. Necoara, Y. Nesterov, and F. Glineur, “Linear convergence of first order methods for non-strongly convex optimization,” *arXiv preprint arXiv:1504.06298*, 2015.

[106] K. Hou, Z. Zhou, A. M.-C. So, and Z.-Q. Luo, “On the linear convergence of the proximal gradient method for trace norm regularization,” in *Advances in Neural Information Processing Systems*, 2013, pp. 710–718.

[107] D. Drusvyatskiy and A. S. Lewis, “Error bounds, quadratic growth, and linear convergence of proximal methods,” *arXiv preprint arXiv:1602.06661*, 2016.

[108] B. Poljak, “Nonlinear programming methods in the presence of noise,” *Mathematical Programming*, vol. 14, no. 1, pp. 87–97, 1978.

[109] T. Yang and Q. Lin, “RSG: beating subgradient method without smoothness and strong convexity,” *arXiv preprint arXiv:1512.03107*, 2015.

[110] S. Supittayapornpong and M. J. Neely, “Staggered time average algorithm for stochastic non-smooth optimization with $O(1/T)$ convergence,” *arXiv preprint arXiv:1607.02842*, 2016.

[111] A. Iouditski and Y. Nesterov, “Primal-dual subgradient methods for minimizing uniformly convex functions,” *arXiv preprint arXiv:1401.1792*, 2014.

[112] A. Money, J. Affleck-Graves, M. Hart, and G. Barr, "The linear regression model: Lp norm estimation and the choice of p," *Communications in Statistics-Simulation and Computation*, vol. 11, no. 1, pp. 89–109, 1982.

[113] H. Nyquist, "The optimal L_p norm estimator in linear regression models," *Communications in Statistics-Theory and Methods*, vol. 12, no. 21, pp. 2511–2524, 1983.

[114] G. Agrb, "Maximum likelihood and ℓ_p-norm estimators," *Statistics Applicata*, vol. 4, no. 1, p. 7, 1992.

[115] F. Bach et al., "Learning with submodular functions: A convex optimization perspective," *Foundations and Trends® in Machine Learning*, vol. 6, no. 2-3, pp. 145–373, 2013.

[116] A. Gilpin, J. Pena, and T. Sandholm, "First-order algorithm with $O(\ln(1/\epsilon))$ convergence for ϵ-equilibrium in two-person zero-sum games," *Mathematical Programming*, vol. 133, no. 1-2, pp. 279–298, 2012.

[117] E. Lim, "On the convergence rate for stochastic approximation in the nonsmooth setting," *Mathematics of Operations Research*, vol. 36, no. 3, pp. 527–537, 2011.

[118] R. M. Freund and H. Lu, "New computational guarantees for solving convex optimization problems with first order methods, via a function growth condition measure," *arXiv preprint arXiv:1511.02974*, 2015.

[119] J. Renegar, "A framework for applying subgradient methods to conic optimization problems," *arXiv preprint arXiv:1503.02611*, 2015.

[120] J. Renegar, ""efficient" subgradient methods for general convex optimization," *SIAM Journal on Optimization*, vol. 26, no. 4, pp. 2649–2676, 2016.

[121] D. Davis and W. Yin, "A three-operator splitting scheme and its optimization applications," *arXiv preprint arXiv:1504.01032*, 2015.

[122] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning.* Springer, 2009.

[123] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[124] J. Kivinen, A. J. Smola, and R. C. Williamson, "Online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2165–2176, 2004.

[125] S. Bubeck et al., "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[126] A. Nedic and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.

[127] S. Portnoy, R. Koenker et al., "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators," *Statistical Science*, vol. 12, no. 4, pp. 279–300, 1997.

[128] L. Wang, M. D. Gordon, and J. Zhu, "Regularized least absolute deviations regression and an efficient algorithm for parameter tuning," in *Data Mining, 2006. ICDM'06. Sixth International Conference on.* IEEE, 2006, pp. 690–700.

[129] L. Wang, "The ℓ_1 penalized LAD estimator for high dimensional linear regression," *Journal of Multivariate Analysis*, vol. 120, pp. 135–151, 2013.

[130] X. Gao and J. Huang, "Asymptotic analysis of high-dimensional lad regression with lasso," *Statistica Sinica*, pp. 1485–1506, 2010.

[131] I. Barrodale and F. D. Roberts, "An improved algorithm for discrete l_1 linear approximation," *SIAM Journal on Numerical Analysis*, vol. 10, no. 5, pp. 839–848, 1973.

[132] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani et al., "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[133] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, pp. 224–244, 2008.

[134] Y. Li and G. R. Arce, "A maximum likelihood approach to least absolute deviation regression," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 12, p. 948982, 2004.

[135] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning.* ACM, 2008, pp. 272–279.