Carnegie Mellon University

MELLON COLLEGE OF SCIENCE

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

TITLE: First-Order Methods in Convex Optimization: Acceleration, Conditioning, and Rescaling

PRESENTED BY: David Gutman

ACCEPTED BY THE DEPARTMENT OF: Mathematical Sciences

Javier Peña	May 2019
MAJOR PROFESSOR	DATE
Thomas Bohman	May 2019
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca W. Doerge	May 2019
DEAN	DATE

First-Order Methods in Convex Optimization:

Acceleration, Conditioning, and Rescaling

David Huckleberry Gutman

Committee:
Javier Peña
Steven Shreve
Fatma Kilinç-Karzan
Alan Frieze
Robert Freund

A thesis presented for the degree of Doctor of Philosophy

Department of Mathematical Sciences
Carnegie Mellon University
May 16, 2019

Abstract

This thesis focuses on three themes related to the mathematical theory of first-order methods for convex minimization: acceleration, conditioning, and rescaling.

Chapters 1 and 2 explore the acceleration theme. In chapter 1, we give a novel proof of the $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ convergence rates of the proximal gradient and accelerated proximal gradient methods for composite convex minimization. The crux of the new proof is an upper bound constructed via the convex conjugate of the objective function.

Chapter 2 extends the approach of chapter 1 to the convergence analysis of Bregman proximal first-order algorithms for convex minimization. We provide novel proofs of the convergence rates of the Bregman proximal subgradient, Bregman proximal gradient, and a new accelerated Bregman proximal gradient algorithm under fairly general and mild assumptions. Our accelerated Bregman proximal gradient algorithm attains the best-known accelerated rate of convergence when suitable relative smoothness and triangle scaling assumptions hold. However, the algorithm requires no prior knowledge of any related smoothness or triangle scaling constants.

Chapter 3 explores the conditioning theme by proposing a condition number of a differentiable convex function relative to a reference set and distance function pair. This relative condition number is defined as the ratio of a relative smoothness constant to a relative strong convexity constant. We show that the relative condition number extends the main properties of the traditional condition number both in terms of its geometric insight and in terms of its role in characterizing the linear convergence of first-order methods for constrained convex minimization.

Chapter 4 explores the rescaling theme. In this chapter, we propose three enhanced versions of the projection and rescaling algorithm's basic procedures, using an efficient algorithmic implementation of Carathéodory's Theorem. Each of these enhanced procedures improves upon the order of complexity of its analogue in Peña and Soheili (Math Program 166(1):87111, 2017) when the dimension of the subspace is sufficiently smaller than the dimension of its ambient space.

Acknowledgments

I would like to thank my advisor, Prof. Javier Peña, for his guidance and insight over the course of my doctoral studies. Without his support, this thesis would not have been possible. Under his tutelage, I learned not only how to do quality research, but also how to adeptly navigate and contribute to the mathematical optimization community at large. Thank you to the remaining members of my thesis committee, Steven Shreve, Fatma Kilinç-Karzan, Alan Frieze, and Robert Freund, for the time and attention they dedicated to the review and defense of my thesis.

I would also like to thank Steven Shreve and the mathematical finance faculty for taking a risk by admitting an unorthodox candidate to Carnegie Mellon University's doctoral program.

My studies at Carnegie Mellon were greatly enhanced by the friendships and professional relationships I developed within the mathematics department and nearby departments at other universities. Many thanks go to Quinn Donahoe, Jackson Pfeiffer, Antoine Rémond-Tiedrez, and Kevin Ou. Without their kindness, support, and constant readiness to eat Mad Mex brownie sundaes, this work would not have been nearly as enjoyable.

Finally, I want to express my gratitude to my family: my parents, Evan and Loretta; my grandparents, Morton, Phyllis, Donald, and Mary Eleanor; and my stepfather, John. I am eternally indebted to them for their warmth, sustained encouragement, and unyielding belief in my ability to succeed at Carnegie Mellon and in life.

Contents

Introduction

This thesis focuses on the mathematical theory of first-order methods for convex minimization problems. The generic convex minimization problem is

$$\min_{x \in X} f(x)$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ denotes a convex objective function and $X \subseteq \mathbb{R}^n$ is a convex feasible set. Essentially, a first-order method is an iterative scheme that uses the subdifferential/gradient information of the objective function to generate each iterate. More formally, a first-order method generates iterates

$$x_{k+1} = x_k + t_k d_k$$

for $k = 0, 1, 2, ...$ where t_k and d_k depend on X, k, and $\{\nabla f(x_i)\}_{i=0}^k$ (or $\{\partial f(x_i)\}_{i=0}^k$ if f is non-differentiable). In the case that f is differentiable then $\nabla f(x)$ will denote the gradient of f evaluated at $x \in \text{dom}(f)$. The effective domain of f is the set

$$\text{dom}(f) := \{x \in \mathbb{R}^n : f(x) < \infty\}.$$

We will say that ∇f is Lipschitz or a Lipschitz gradient when it is Lipschitz continuous on $\text{dom}(f)$. When f is non-differentiable $\partial f(x)$ denotes the subdifferential of f evaluated at $x \in \text{dom}(f)$, that is

$$\partial f(x) := \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y \in \text{dom}(f)\}.$$

The new era of big data and machine learning spawned a resurgent interest in these classical methods. The typical machine learning algorithm depends on the successful, approximate solution of a large-scale optimization problem comprised of millions, if not more, decision variables. Thus, the low storage and iteration costs of first-order methods relative to more elaborate schemes make these methods particularly attractive for modern data science applications. This thesis consists of four chapters that explore three themes: acceleration, conditioning, and rescaling. The first two chapters explore the first theme while the final two chapters explore the latter themes.

Acceleration

The relatively slow convergence rates of first-order methods offset their relatively cheap iteration and storage costs. For a general convex objective $f : \mathbb{R}^n \to \mathbb{R}$, first-order

methods for the unconstrained minimization problem $\min_{x\in\mathbb{R}^n} f(x)$ typically converge in $\mathcal{O}\left(1/\sqrt{k}\right)$ iterates which we formally write as

$$f(x_k) - \min_{x\in X} f(x) \leq \mathcal{O}\left(1/\sqrt{k}\right)$$

where the quantity on the left is often referred to as the suboptimality gap. If we further assume that f has a Lipschitz gradient then the standard gradient descent algorithm for $\min_{x\in\mathbb{R}^n} f(x)$ converges in $\mathcal{O}(1/k)$ iterates. Moreover, it is well-known that the lower bound for the convergence rate for the minimization of a convex function with a black box first-order oracle and Lipschitz gradient is $\mathcal{O}(1/k^2)$ [53].

In his seminal paper, [52] Nesterov devised an accelerated first-order algorithm with the optimal $\mathcal{O}(1/k^2)$ rate of convergence in this setting via a modification of the standard gradient descent algorithm that includes *momentum* steps. A later breakthrough was the acceleration of the *proximal gradient method* independently developed by Beck and Teboulle [8] and by Nesterov [54]. The proximal gradient method, also known as the forward-backward splitting method [45], is an extension of the gradient descent method to solve the composite minimization problem

$$\min_{x\in\mathbb{R}^n} f(x) + \Psi(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is differentiable on $\operatorname{dom}(f)$ and $\Psi : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is a closed convex function such that $\operatorname{dom}(\Psi) \subseteq \operatorname{dom}(f)$ and such that for $L > 0$ the proximal map $\mathsf{Prox}_{\frac{1}{L}} : \mathbb{R}^n \to \operatorname{dom}(\Psi)$ defined by

$$\mathsf{Prox}_{\frac{1}{L}}(x) := \operatorname*{argmin}_{y\in\mathbb{R}^n} \left\{ \Psi(y) + \frac{L}{2}\|y - x\|^2 \right\} \tag{2}$$

is computable.

The class of *Bregman proximal first-order methods*, a more flexible and generalized class of proximal gradient methods, are based on the *Bregman proximal map*

$$(x, g) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \operatorname*{argmin}_{y\in\mathbb{R}^n} \{\langle g, y\rangle + \Psi(y) + LD_h(y, x)\} \tag{3}$$

where $D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x\rangle$ is the Bregman distance [16] generated by some *reference* convex function $h : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$. If the reference function in (3) is the squared Euclidean norm $h(x) := \frac{1}{2}\|x\|_2^2$ then we precisely recover the proximal map which defines the class of proximal gradient methods.

The mirror descent method [7, 47, 51] is a well-known instance of a Bregman proximal first-order method when $\Psi = I_C$, the indicator function of C, for some closed convex set $C \subseteq \mathbb{R}^n$. Some more recent instances of Bregman proximal methods include the *NoLips* algorithm introduced by Bauschke, Bolte, and Teboulle [4], which follows a *Bregman proximal gradient* template [3, 9, 69, 70]. This same algorithmic template

underlies the *relative gradient scheme* proposed by Lu, Freund, and Nesterov [4]. Both [4] and [48] establish convergence results for the Bregman proximal gradient method by relying on a *Lipschitz-like convexity condition* (LC) as defined in [4] or the equivalent *relative smoothness* condition as defined in [48]. Both papers derive a $\mathcal{O}(1/k)$ rate assuming that the relative smoothness condition holds. This relative smoothness condition is crucial to the analysis in [34]. The *Bregman proximal subgradient* method [11, 10, 23, 69] which allows for f to be non-differentiable and converges in $\mathcal{O}(1/\sqrt{k})$ iterates is also an instance of this class. Recently, Hanzely, Richtarik, and Xiao [34] achieved the acceleration of Bregman proximal gradient methods for relatively smooth functions. In particular, this accelerated scheme achieves a $\mathcal{O}(1/k^\gamma)$ rate where the constant $\gamma > 0$ depends on the underlying Bregman divergence.

The enormous significance of Nesterov's and Beck and Teboulle's original breakthroughs prompted interest in new explanations for how to achieve the acceleration of first-order methods [1, 17, 22, 26, 43, 59, 68]. Some of these approaches are based on geometric [17, 22], control [43], and differential equations [68] techniques. The recent article [59] relies on the convex conjugate to give a unified and succinct derivation of the known $\mathcal{O}(1/\sqrt{k}), \mathcal{O}(1/k)$, and $\mathcal{O}(1/k^2)$ convergence rates of the subgradient, gradient, and accelerated gradient methods for unconstrained smooth convex minimization. A natural question is whether this approach extends to the broader class of proximal gradient methods and its generalization, Bregman proximal gradient methods. Chapters 1 and 2, respectively based on the papers [33] and [31], offer an affirmative answer to this question. The crux of the approach is a generic upper bound, constructed via the convex conjugate of the objective function, of the iterates generated by the subgradient, gradient, and accelerated gradient algorithms from the proximal and Bregman proximal classes. The frameworks presented in these chapters not only provide general templates for the design of future accelerated methods, but also subsume a number of popular methods.

In chapter 1, which is based on the paper [33] written jointly with Prof. Javier Peña, we apply this new technique to recover the known rates for the proximal gradient and subgradient methods, and the accelerated proximal gradient method. Our treatment covers the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA), a popular variant of the accelerated proximal gradient method, when the step size is chosen by backtracking line search. Notably, we provide a unified derivation of the convergence rates of the three methods under weaker conditions than those previously assumed in the literature. Previously, authors obtained the convergence rates under the hypothesis that the objective function's smooth component possessed a Lipschitz gradient. We assume only that the iterates of each algorithm satisfy a decrease condition which is easily seen to hold if the smooth component has a Hölder continuous gradient and the step sizes are chosen judiciously.

In chapter 2, which is based on the paper [31] written jointly with Prof. Javier Peña, we extend the technique of chapter 1 to derive new and known rates for the class of Bregman proximal first-order methods. We provide a unified derivation of the convergence rates of the Bregman proximal gradient method, Bregman subgradient

method, and the new accelerated Bregman proximal gradient of Hanzely, Richtarik, and Xiao [34]. In particular, we show that our algorithmic template for accelerated Bregman proximal methods and its related analysis subsumes the new algorithm of [34]. As in chapter 1, we assume only that a decrease condition holds at each iterate. Furthermore, we provide periodic restart schemes for the accelerated Bregman proximal gradient template that ensure linear convergence assuming sufficient smoothness and error bound conditions.

Chapter 2 highlights the interplay of the selection of the momentum parameter and step size for Bregman proximal methods based on arbitrary reference functions. Accounting for this relationship facilitates the simultaneous treatment of the effect of continuity and geometric conditions for the objective and reference functions on convergence rates for accelerated and basic algorithms. However, certain algorithms, such as FISTA with backtracking, require selections that do not accord with this relationship. Chapter 1 recognizes that the simple structure of the squared Euclidean norm, which is the Bregman divergence used for the proximal gradient methods, enables the analysis of a broader class of step size and momentum parameter regimes for proximal methods. Consequently, as corollaries of chapter 1's main theorem, we are able to consider flexible step size selection for FISTA and derive the modern convergence rates for the projected subgradient algorithm from [29].

Conditioning

Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function. The *condition number* of f is the ratio L_f/μ_f where L_f and μ_f are respectively the *smoothness* and *strong convexity* constants of the function f measured with respect to $\|\cdot\|_2$. The condition number L_f/μ_f is closely tied to a number of fundamental properties of the function f. In the special case when f is a quadratic convex function the condition number has the following geometric interpretation. Suppose $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ where $A \in \mathbb{R}^{n\times n}$ is non-singular. Then the condition number of f is

$$\frac{L_f}{\mu_f} = \|A^\mathsf{T} A\| \cdot \|(A^\mathsf{T} A)^{-1}\| = (\|A\| \cdot \|A^{-1}\|)^2. \tag{4}$$

The latter quantity is the square of the aspect ratio of the ellipsoid $A(\mathbb{B}) := \{Ax : x \in \mathbb{R}^n, \|x\|_2 \leq 1\}$ since $\|A\|$ and $1/\|A^{-1}\|$ are respectively the radii of the smallest ball that contains $A(\mathbb{B})$ and the largest ball contained in $A(\mathbb{B})$.

The condition number L_f/μ_f also bounds the linear convergence rate of the gradient descent algorithm for the unconstrained minimization problem

$$\bar{f} = \min_{x\in\mathbb{R}^m} f(x).$$

More precisely, for a suitable choice of step sizes the iterates $x_k,\ k = 0, 1, \ldots$ generated

by the gradient descent algorithm satisfy

$$\|\bar{X} - x_k\|_2^2 \leq \left(1 - \frac{\mu_f}{L_f}\right)^k \|\bar{X} - x_0\|_2^2$$

and

$$f(x_k) - \bar{f} \leq \frac{L_f}{2}\left(1 - \frac{\mu_f}{L_f}\right)^k \|\bar{X} - x_0\|_2^2,$$

where $\bar{X} := \{x \in \mathbb{R}^n : f(x) = \bar{f}\}$ and $\|\bar{X} - x\|_2 = \inf_{y \in \bar{X}} \|y - x\|_2$. The articles [17, 22, 39, 49, 50, 53, 54], among others, discuss the above type of linear convergence and a number of interesting related developments. In particular, Necoara, Nesterov and Glineur [50] establish linear convergence properties for a wide class of first-order methods under assumptions that are relaxations of strong convexity.

However, the condition number of a convex function alone is unfit to describe the performance of some optimization algorithms including the away step variant of the Frank-Wolfe algorithm. At each iteration k, the Frank-Wolfe algorithm selects its search direction by minimizing the first-order Taylor approximation of the objective function over the feasible set. Formally, for $k = 0, 1, 2, ...$, the algorithm selects d_k according to the rule

$$\begin{aligned} v_k &= \min_{y \in X} \left[f(x_k) + \langle \nabla f(x_k), y - x_k \rangle\right], \\ d_k &= v_k - x_k. \end{aligned}$$

To ensure a linear convergence rate when the objective is strongly convex, the away step variant of the Frank-Wolfe algorithm is usually used in place of the basic algorithm. As [60, 6, 42] show, the linear convergence rates for the away step variant of Frank-Wolfe depend on constants informed not only by characteristics of the objective function, but also the geometry of the feasible set induced by the norm of choice.

This joint dependence on the function and the domain is the inspiration for our theme in chapter 3. The theme of this research line is *relative condition numbers*, condition numbers for constrained convex optimization that depend on the objective function, the underlying domain, and a distance-like function on the underlying domain. In chapter 3, which is based on the paper [32] written jointly with Prof. Javier Peña, we propose a relative smoothness constant $L_{f,X,D}$ and a relative strong convexity constant $\mu_{f,X,D}$ of the function f relative to the pair (X, D) where $X \subseteq \text{dom}(f)$ is a convex set, and $D : X \times X \to \mathbb{R}_+$ is a distance-like function, that is, $D(y, x) \geq 0$ and $D(x, x) = 0$ for all $x, y \in X$. Our main results highlight the tight connection between the relative constants and geometric features of the set X. In particular, we consider functions of the form $f = g \circ A$ for some matrix $A \in \mathbb{R}^{m \times n}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$. For functions of this form, we provide characterizations and bounds on $L_{f,X,D}$ and $\mu_{f,X,D}$ in terms of L_g and μ_g, and geometric properties of the pair (A, X).

The relative constants $L_{f,X,D}$ and $\mu_{f,X,D}$ are defined *globally.* In particular, they do not depend on any specific point in X. We consider several variants of relative

strong convexity following the constructions of Necoara, Nesterov and Glineur [50]. In particular, we define a relative *quasi-strong convexity constant* $\mu^{\star}_{f,X,D}$ and a *D-functional growth constant* $\mu^{\sharp}_{f,X,D}$. See Definition 5 and equation (3.9). Unlike $\mu_{f,X,D}$, the constants $\mu^{\star}_{f,X,D}$ and $\mu^{\sharp}_{f,X,D}$ depend on the set of minimizers $\bar{X}$ of f on X. We show that relative quasi-strong convexity is a relaxation of relative strong convexity. We also show that under suitable assumptions D-functional growth is a relaxation of quasi-strong convexity. Not surprisingly, there are classes of non-strongly convex functions for which the constant $\mu^{\sharp}_{f,X,D}$ is positive while $\mu_{f,X,D}$ and $\mu^{\star}_{f,X,D}$ may not be. (See Theorem 11.)

We show that the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ and some related quantities readily yield a linear convergence rate for the mirror descent algorithm for the constrained minimization problem

$$\bar{f} = \min_{x \in X} f(x). \tag{5}$$

Furthermore, we show that a similar relative condition number also yields a linear convergence rate for a version of the Frank-Wolfe algorithm with away steps when the set X is a polytope.

We should note that the linear convergence of the mirror descent algorithm, Frank-Wolfe algorithm with away steps, and other first-order methods have been previously established in [6, 42, 48, 50, 54, 60, 69] under various kinds of assumptions. Our approach based on the relative condition number shows that the linear convergence of the mirror descent algorithm (Propositions 8 and 9) holds for a sharper rate and under conditions that are weaker than those assumed in [48, 69]. Our approach based on the relative condition number yields a proof of linear convergence for the Frank-Wolfe algorithm with away steps that is significantly shorter, simpler, and more general than the ones previously presented in [6, 42, 60]. In particular, our most general statement of linear convergence for the Frank-Wolfe algorithm with away steps (Proposition 10) is at least as sharp or sharper than the linear convergence statements in [6, 42, 60].

Our work in chapter 3 draws on and connects several seemingly unrelated threads of research on first-order methods [4, 6, 42, 48, 50, 60, 69] and on condition measures for convex optimization [25, 24, 28, 27, 44, 56, 58, 63, 64]. Our construction of $L_{f,X,D}$ and $\mu_{f,X,D}$ is inspired by and closely related to the work of Lu, Freund, and Nesterov [48] and of Bauschke, Bolte, and Teboulle [4, 69]. Lu et al. [48] extend the concepts of smoothness and strong convexity constants by considering them *relative* to a *reference* function h, see [48, Definition 1.1 and 1.2]. Our construction is identical to theirs in the special case when the distance function is the Bregman distance function D_h associated to the reference function h and the function f is strictly convex. Bauschke, Bolte, and Teboulle [4] define a concept of *Lipschitz-like* condition that is equivalent to smoothness relative to a reference function. Our construction of $L_{f,X,D}$ and $\mu_{f,X,D}$ is also related to the *away curvature constant* and *geometric strong convexity constant* proposed by Lacoste-Julien and Jaggi in [42, Appendix C]. Our constructions of D-functional growth, and relative quasi strong convexity are natural extensions of analogous con-

cepts proposed by Necoara, Nesterov, and Glineur [50] to unveil relaxations of strong convexity that ensure the linear convergence of first-order methods. Our D-functional growth is in the same spirit as a quadratic growth approach used by Beck and Shtern [6] to established the linear convergence of a conditional gradient algorithm with away steps for non-strongly convex functions.

In contrast to the approaches in [6, 42, 48, 50, 60], our construction of the relative condition constants applies to any pair (X, D) of reference set and distance function. Our main results (Section 3.3 and Section 3.4) reveal some interesting insights when D is a squared norm. We establish a close connection between our relative conditioning approach and the conditioning of linear conic systems pioneered by Renegar [63, 64] and further developed by a number of authors [19, 25, 24, 28, 27, 44, 56, 58, 57]. We especially draw on ideas developed in the recent paper [57]. We note that consistent with our construction of the relative constants $L_{f,X,D}$, $\mu_{f,X,D}$, $\mu^{\star}_{f,X,D}$, $\mu^{\sharp}_{f,X,D}$, all of our results concerning them scale appropriately, that is, they scale by λ whenever the objective function f is replaced by $\tilde{f} = \lambda f$ for some constant $\lambda > 0$. In particular, the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ and all of our bounds on it are invariant under positive scaling of f. Due to the dependence of these constants on geometric properties of the pair (X, D), they are not invariant under rescalings of X.

Rescaling

In contrast to first-order methods, second-methods trade low cost iterates for fast convergence rates. Newton's method, the canonical second-order method, can be viewed as a gradient descent algorithm that incorporates a Hessian-based rescaling of the objective function's level sets. This straightforward enhancement yields local quadratic convergence for well-conditioned problem instances. In this light, it seems reasonable to ask if periodic rescaling or reconditioning can improve the convergence rate of first-order methods. Rescaling to improve the complexity of first-order methods for solving linear feasibility problems is the third theme of this thesis. We pay specific attention to the *Projection and Rescaling algorithm* of Peña and Soheili [61].

A *projection method* seeks to solve the linear feasibility problem

$$\text{Find } x \in L \cap \mathbb{R}^n_{++}, \tag{6}$$

where L is a subspace of $\mathbb{R}^n$. Typically this problem is recast as

$$\text{Find } x \text{ such that } P_L x > 0. \tag{7}$$

A projection method is a first-order method that iteratively reduces the value of $\|P_L x\|$ until a feasible point is reached or a certificate of infeasibility is discovered. The Von Neumann and Perceptron algorithms, which can be viewed as special cases of the Frank-Wolfe algorithm, are well-known instances of projection methods. Convergence analyses for projection methods usually rely upon condition numbers that measure how "deeply interior" points in L are to the cone $\mathbb{R}^n_+$ [19, 12, 38].

In the case of a poorly conditioned system, i.e. one for which points in L are not "deeply interior" to $\mathbb{R}^n_+$, it is desirable to recondition the system to improve the complexity of projection methods. *Rescaling* provides a popular and modern enhancement that preconditions problems for projection methods. A rescaling step rescales the ambient space such that the condition number improves, i.e. points in L become more deeply interior. Peña and Soheili [61] introduce the Projection and Rescaling algorithm, which extends an algorithm by Chubanov [20], to solve the linear feasibility problem. This is a two-step algorithm that alternates between the limited execution of a projection method, referred to in this context as the basic procedure, and a rescaling of the ambient space determined by the basic procedure's output.

Recent literature has attempted to improve or extend the projection and rescaling procedure in many ways. Some authors have attempted to extend the projection and rescaling algorithm to other settings and using different progress measures. The paper [40] extends the algorithm to the setting in which $\mathbb{R}^n_+$ is replaced with the second-order cone while [46] extends it to the setting where $\mathbb{R}^n_+$ is replaced with a general symmetric cone. Other authors have attempted to enhance the basic procedure to improve the complexity of the algorithm. In particular [66], produces a modified basic procedure that reduces the iteration bound for Chubanov's original projection and rescaling algorithm by a factor of 5.

In chapter 4, based on the forthcoming paper [30], we propose enhancements to three of the four Von Neumann/Perceptron basic procedures in [61] to improve the complexity of the basic procedures from $O(n^4m)$ to $O(n^2m^3)$ operations: a significant improvement when $m << n$ where m is the dimension of L. Our enhancements depend on an algorithmic implementation of Carathéodory's Theorem. Moreover, the implementation of this technique strongly resembles the revised Simplex method.

Chapter 1

Acceleration, Part I: Convergence Rates of Proximal Gradient Methods via the Convex Conjugate

1.1 Introduction

In this chapter, we give a unified derivation of the convergence rates of the proximal gradient, accelerated proximal gradient, and proximal subgradient algorithms for the composite convex minimization problem (1). The central results of this chapter (Theorems 1-3) are upper bounds on the iterates generated by the non-accelerated proximal gradient, accelerated proximal gradient, and proximal subgradient methods. The expressions in the three upper bounds (see (1.6), (1.9), and (1.14)) as well as their proofs (see section 1.4) are strikingly similar. They highlight the commonalities and differences of the three methods. The upper bounds are constructed via the convex conjugate of the objective function. Theorem 1 and Theorem 2 readily yield the widely known $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ convergence rates of the proximal gradient and accelerated proximal gradient algorithms for (1) when the smooth component f has Lipschitz gradient and the step sizes are chosen judiciously. The convex conjugate approach underlying Theorems 1 and 2 also extend to a *proximal subgradient algorithm* when the component f is merely convex but not necessarily smooth. (See Algorithm 2 and Theorem 3.) This extension automatically yields a novel derivation of both classical [53, Theorem 3.2.2] as well as modern convergence rates [29, Theorem 5] for the projected subgradient algorithm.

We should note that in contrast to the classical proofs of the iconic convergence rates $\mathcal{O}(1/k)$ for proximal gradient, $\mathcal{O}(1/k^2)$ for accelerated proximal gradient, and $\mathcal{O}(1/\sqrt{k})$ for projected subgradient algorithms, our central results, namely Theorems 1-3 require substantially weaker assumptions. More precisely, Theorems 1-3 hold under suitable assumptions on the step sizes and momentum steps but do not require any Lipschitz condition on the components of the objective function or on their gradients.

As a consequence, for the proximal gradient method Theorem 1 guarantees convergence of the iterates' objective values to optimality in the absence of Lipschitz continuity provided the step sizes are not summable. Similarly, for the accelerated proximal gradient Theorem 2 guarantees the same type of convergence under an even milder boundedness condition. Finally, Theorem 3 yields convergence results of similar flavor for the projected subgradient method provided the subgradient oracle satisfies a fairly mild and general steepness condition.

Throughout the chapter we assume that $\mathbb{R}^n$ is endowed with an inner product $\langle \cdot, \cdot \rangle$ and that $\| \cdot \|$ denotes the corresponding Euclidean norm.

1.2 Proximal Gradient and Accelerated Proximal Gradient Methods

Let $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function such that the proximal map (2) is computable and let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a differentiable convex function such that $\mathrm{dom}(\Psi) \subseteq \mathrm{dom}(f)$. Let $\phi := f + \Psi$ and consider the problem (1) that can be rewritten as

$$\min_{x \in \mathbb{R}^n} \phi(x). \tag{1.1}$$

Algorithm 1 describes a template of a proximal gradient algorithm for (1.1).

Algorithm 1 Template for proximal gradient method

1: **input:** $x_0 \in \mathrm{dom}(f)$
2: $y_0 := x_0$; $\theta_0 := 1$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4: pick $t_k > 0$
5: $x_{k+1} := \mathsf{Prox}_{t_k}(y_k - t_k \nabla f(y_k))$
6: pick $\theta_{k+1} \in (0, 1]$
7: $y_{k+1} := x_{k+1} + \frac{\theta_{k+1}(1-\theta_k)}{\theta_k}(x_{k+1} - x_k)$
8: **end for**

Step 7 of Algorithm 1 incorporates a momentum step. The non-accelerated proximal gradient method is obtained by choosing $\theta_{k+1} = 1$ in Step 6. In this case Step 7 simply sets $y_{k+1} = x_{k+1}$ and does not incorporate any momentum. Other choices of $\theta_{k+1} \in (0, 1]$ yield accelerated versions of the proximal gradient method. In particular, the FISTA algorithm in [8] is obtained by choosing $\theta_{k+1} \in (0, 1]$ via the rule $\theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$. In this case $\theta_k \in (0, 1)$ for $k \geq 1$ and there is a non-trivial momentum term in Step 7. Algorithm 1 implicitly assumes that the choice of θ_{k+1} in Step 6 is so that the point y_{k+1} in Step 7 satisfies $y_{k+1} \in \mathrm{dom}(f)$. This holds provided $\mathrm{dom}(f)$ is sufficiently larger than $\mathrm{dom}(\Psi)$.

The main results in this chapter are Theorem 1 and its variant Theorem 2 below which subsume the widely known convergence rates $\mathcal{O}(1/k)$ and $\mathcal{O}(1/k^2)$ of the prox-

imal gradient and accelerated proximal gradient algorithms under suitable choices of $t_k, \theta_k,\ k = 0, 1, \ldots$.

Theorem 1 relies on a suitably constructed sequence $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$. The construction of $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$ in turn is motivated by the identity (1.3) below.

Consider Step 5 in Algorithm 1, namely

$$x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla f(y_k)) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \Psi(x) + \frac{1}{2t_k} \|x - (y_k - t_k \nabla f(y_k))\|^2 \right\}. \quad (1.2)$$

The optimality conditions for (1.2) can be written as

$$g_k^{\Psi} + \frac{1}{t_k}(x_{k+1} - (y_k - t_k \nabla f(y_k))) = 0$$

for some $g_k^{\Psi} \in \partial \Psi(x_{k+1})$. These conditions imply that

$$x_{k+1} = y_k - t_k \cdot g_k$$

where $g_k := g_k^f + g_k^{\Psi}$ for $g_k^f := \nabla f(y_k)$ and for some $g_k^{\Psi} \in \partial \Psi(x_{k+1})$. Thus Step 5 and Step 7 of Algorithm 1 imply that for $k = 0, 1, \ldots$

$$\frac{y_{k+1} - (1 - \theta_{k+1})x_{k+1}}{\theta_{k+1}} = \frac{x_{k+1} - (1 - \theta_k)x_k}{\theta_k} = \frac{y_k - (1 - \theta_k)x_k}{\theta_k} - \frac{t_k}{\theta_k} g_k.$$

Since $\theta_0 = 1$ and $y_0 = x_0$, it follows that for $k = 1, 2, \ldots$

$$\frac{y_k - (1 - \theta_k)x_k}{\theta_k} = x_0 - \sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i \Leftrightarrow (1 - \theta_k)(y_k - x_k) = \theta_k \left(x_0 - y_k - \sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i \right). \quad (1.3)$$

As it is customary, we will assume that the step sizes t_k chosen at Step 4 in Algorithm 1 satisfy the following decrease condition

$$\begin{aligned} \phi(x_{k+1}) &\leq \min_{x \in \mathbb{R}^n} \left\{ f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \frac{1}{2t_k} \|x - y_k\|^2 + \Psi(x) \right\} \quad (1.4) \\ &= f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^{\Psi}, y_k - x_{k+1} \right\rangle - \frac{t_k}{2} \|g_k\|^2. \end{aligned}$$

The condition (1.4) holds in particular when ∇f is Lipschitz and $t_k,\ k = 0, 1, \ldots$ are chosen via a standard backtracking procedure. Observe that (1.4) implies $\phi(x_{k+1}) \leq \phi(y_k)$.

The proof of theorem 1 relies on the convex conjugate function. The algorithm does not require the computation of this function. Recall that if $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex function then its *convex conjugate* $h^* : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as

$$h^*(z) = \sup_{x \in \mathbb{R}^n} \{\langle z, x \rangle - h(x)\}.$$

Theorem 1. *Suppose $\theta_k \in (0,1],\ k = 0,1,2,\ldots$ and the step sizes $t_k > 0,\ k = 0,1,2,\ldots$ are such that (1.4) holds. Let $x_k \in \mathbb{R}^n,\ k = 1,2,\ldots$ be the iterates generated by Algorithm 1. Let $z_k \in \mathbb{R}^n,\ k = 1,2\ldots$ be as follows*

$$z_k := \frac{\sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i}{\sum_{i=0}^{k-1} \frac{t_i}{\theta_i}}. \tag{1.5}$$

Then for $k = 1,2,\ldots$

$$\mathsf{LHS}_k \leq -\phi^*(z_k) + \langle z_k, x_0\rangle - \frac{\sum_{i=0}^{k-1} \frac{t_i}{\theta_i}}{2}\|z_k\|^2, \tag{1.6}$$

where LHS_k is as follows depending on the choice of $\theta_k \in (0,1]$ and $t_k > 0$.

(a) When $\theta_k = 1, k = 0,1,\ldots$ let

$$\mathsf{LHS}_k := \frac{\sum_{i=0}^{k} t_i \phi(x_{i+1})}{\sum_{i=0}^{k} t_i}.$$

(b) When $t_k = 1/L,\ k = 0,1,\ldots$ for some positive constant L and $\theta_k,\ k = 0,1,2,\ldots$ are chosen via $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1-\theta_{k+1}),\ k = 0,1,\ldots$ let

$$\mathsf{LHS}_k = \phi(x_k).$$

Theorem 1 readily implies that in both case (a) and case (b)

$$\begin{aligned}
\mathsf{LHS}_k &\leq \inf_{u\in\mathbb{R}^n}\{\phi(u) - \langle z_k, u\rangle\} + \min_{u\in\mathbb{R}^n}\left\{\langle z_k, u\rangle + \frac{1}{2\cdot\sum_{i=0}^{k-1}\frac{t_i}{\theta_i}}\|u - x_0\|^2\right\} \\
&\leq \inf_{u\in\mathbb{R}^n}\left\{\phi(u) + \frac{1}{2\cdot\sum_{i=0}^{k-1}\frac{t_i}{\theta_i}}\|u - x_0\|^2\right\} \\
&\leq \phi(x) + \frac{1}{2\cdot\sum_{i=0}^{k-1}\frac{t_i}{\theta_i}}\|x - x_0\|^2
\end{aligned}$$

for all $x \in \mathbb{R}^n$.

Let $\bar{\phi}$ and $\bar{X}$ respectively denote the optimal value and set of optimal solutions to (1.1). If $\bar{\phi}$ is finite and $\bar{X}$ is nonempty then in both case (a) and case (b) of Theorem 1 we get

$$\phi(x_k) - \bar{\phi} \leq \frac{\mathsf{dist}(x_0, \bar{X})^2}{2\cdot\sum_{i=0}^{k-1}\frac{t_i}{\theta_i}}. \tag{1.7}$$

Suppose $t_k \geq 1/L,\ k = 0, 1, 2, \ldots$ for some constant $L > 0$. This holds in particular for $L := \max\{L_0, L_f/\alpha\}$ if ∇f is L_f-Lipschitz and t_k is chosen via the following standard type of backtracking procedure: pick $t_k = 1/L_0$ for some $L_0 > 0$ and scale t_k by $\alpha \in (0, 1)$ until (1.4) holds for $x_{k+1} = \mathsf{Prox}_{t_k}(y_k - t_k \nabla f(y_k))$. Then inequality (1.7) yields the following known convergence bound for the proximal gradient method

$$\phi(x_k) - \bar{\phi} \leq \frac{L \cdot \mathsf{dist}(x_0, \bar{X})^2}{2k}.$$

On the other hand, suppose $t_k = 1/L,\ k = 0, 1, 2, \ldots$ for some constant $L > 0$ and $\theta_k,\ k = 0, 1, 2, \ldots$ are chosen via $\theta_0 = 1$ and $\theta_{k+1}^2 = \theta_k^2(1-\theta_{k+1})$. Then a straightforward induction shows that

$$\sum_{i=0}^{k-1} \frac{t_i}{\theta_i} = (1 - \theta_k) \sum_{i=0}^{k} \frac{t_i}{\theta_i} = \frac{1}{L\theta_{k-1}^2}. \tag{1.8}$$

The conditions $\theta_0 = 1,\ \theta_{k+1}^2 = \theta_k^2(1 - \theta_{k+1})$ and an additional induction show that $\theta_{k-1}^2 \leq 4/(k+1)^2,\ k = 1, 2, \ldots$. Thus Theorem 1(b), inequality (1.7), and equation (1.8) yield the following known convergence bound for the accelerated proximal gradient method

$$\phi(x_k) - \bar{\phi} \leq \frac{2L \cdot \mathsf{dist}(x_0, \bar{X})^2}{(k+1)^2}.$$

Although Theorem 1 yields the iconic $\mathcal{O}(1/k^2)$ convergence rate of the accelerated proximal gradient algorithm, it applies under the somewhat restrictive conditions as stated in case (b) above. In particular, case (b) does not cover the more general case when $t_k,\ k = 0, 1, \ldots$ are chosen via backtracking as in the FISTA with backtracking algorithm in [8]. The convergence rate in this case, namely [8, Theorem 4.4] is a consequence of Theorem 2 below. Theorem 2 is a variant of Theorem 1(b) that applies to more flexible choices of $t_k, \theta_k,\ k = 0, 1, \ldots$. In particular, Theorem 2 applies to the popular choice $\theta_k = \frac{2}{k+2},\ k = 0, 1, \ldots$.

Theorem 2. *Suppose $\bar{\phi} = \min_{x \in \mathbb{R}^n} \phi(x)$ is finite, $\theta_k \in (0, 1],\ k = 0, 1, 2, \ldots$ satisfy $\theta_0 = 1$ and $\theta_{k+1}^2 \geq \theta_k^2(1 - \theta_{k+1})$, and the step sizes $t_k > 0,\ k = 0, 1, 2, \ldots$ are non-increasing and such that (1.4) holds. Let $x_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$ be the iterates generated by Algorithm 1. Let $z_k \in \mathbb{R}^n,\ k = 1, 2, \ldots$ be as follows*

$$z_k = \frac{\theta_{k-1}^2}{t_{k-1}} \cdot \sum_{i=0}^{k-1} \frac{t_i}{\theta_i} g_i.$$

Then for $k = 1, 2, \ldots$

$$\phi(x_k) - \bar{\phi} \leq -(R_k \cdot (f - \bar{\phi}))^*(z_k) + \langle z_k, x_0 \rangle - \frac{t_{k-1}}{2\theta_{k-1}^2} \|z_k\|^2, \tag{1.9}$$

where $R_1 = 1$ *and* $R_{k+1} = \frac{t_{k-1}}{t_k} \cdot \frac{\theta_k^2}{\theta_{k-1}^2(1-\theta_k)} \cdot R_k \geq 1,\ k = 1, 2, \ldots$. *In particular, if* $\bar{X} = \{x \in \mathbb{R}^n : \phi(x) = \bar{\phi}\}$ *is nonempty then*

$$\phi(x_k) - \bar{\phi} \leq \inf_{u \in \mathbb{R}^n} \left\{ R_k \cdot (\phi(u) - \bar{\phi}) + \frac{\theta_{k-1}^2}{2t_{k-1}} \|u - x_0\|^2 \right\} \leq \frac{\theta_{k-1}^2 \cdot \mathsf{dist}(x_0, \bar{X})^2}{2t_{k-1}}.$$

Suppose the step sizes $t_k,\ k = 0, 1, 2, \ldots$ are non-increasing, satisfy (1.4), and $t_k \geq 1/L,\ k = 0, 1, 2, \ldots$ for some constant $L > 0$. This holds in particular when ∇f is Lipschitz and t_k is chosen via a suitable backtracking procedure as the one in [8]. If $\theta_0 = 1$ and $\theta_{k+1}^2 \geq \theta_k^2(1 - \theta_{k+1}),\ k = 0, 1, \ldots$ then Theorem 2 implies that

$$\phi(x_k) - \bar{\phi} \leq \frac{L\theta_{k-1}^2 \cdot \mathsf{dist}(x_0, \bar{X})^2}{2}.$$

If $\theta_{k+1}^2 = \theta_k^2(1-\theta_{k+1}),\ k = 0, 1, \ldots$ or $\theta_k = 2/(k+2),\ k = 0, 1, \ldots$ then $\theta_{k-1}^2 \leq 4/(k+1)^2$ and so

$$\phi(x_k) - \bar{\phi} \leq \frac{2L \cdot \mathsf{dist}(x_0, \bar{X})^2}{(k+1)^2}.$$

We conclude this section by noting other immediate and interesting consequences of Theorem 1 and Theorem 2. Observe that these two theorems rely only on some assumptions on the step sizes $t_k,\ k = 0, 1, 2, \ldots$ and on the momentum steps $\theta_k,\ k = 0, 1, 2, \ldots$. Unlike classical proofs of convergence for the proximal gradient and accelerated proximal gradient algorithms, Theorem 1 and Theorem 2 do not require ∇f to be Lipschitz continuous. As a consequence, the iterates generated by Algorithm 1 satisfy $\phi(x_k) \to \bar{\phi}$ for a broader class of functions. In particular, consider the special case when $\Psi = 0$ and ∇f satisfies the following type of Hölder continuity: there exist constants L and $v \in (0, 1]$ such that for all $x, y \in \mathbb{R}^n$

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|x - y\|^v.$$

In this case $\phi = f$ and some straightforward calculations show that for all $x, y \in \mathbb{R}^n$

$$\phi(y) \leq \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{L}{1+v}\|y - x\|^{1+v}.$$

Thus a standard backtracking procedure guarantees that the stepsize t_k at each main iteration of Algorithm 1 can be chosen so that (1.4) holds and

$$t_k \geq C \cdot \|\nabla\phi(y_k)\|^{\frac{1-v}{v}} \tag{1.10}$$

for some constant $C > 0$. When $\theta_k = 1,\ k = 1, 2, \ldots$ inequality (1.4) implies that the sequences $\phi(x_k),\ k = 0, 1, 2, \ldots$ and $\mathsf{dist}(x_k, \bar{X}),\ k = 0, 1, 2, \ldots$ are non-increasing. Thus in that case the convexity of f implies that

$$\phi(x_k) - \bar{\phi} \leq \|\nabla\phi(x_k)\| \cdot \mathsf{dist}(x_k, \bar{X}) \leq \|\nabla\phi(x_k)\| \cdot \mathsf{dist}(x_0, \bar{X}). \tag{1.11}$$

Combining equation (1.10), equation (1.11), Theorem 1, and the fact that $\phi(x_k),\ k = 0, 1, 2, \ldots$ is non-increasing, we see that $\phi(x_k) \to \bar{\phi}$ and $\nabla\phi(x_k) \to 0$ when $\theta_k = 1,\ k = 1, 2, \ldots$.

Theorem 1 also implies that the iterates generated by Algorithm 1 satisfy $\phi(x_k) \to \bar{\phi}$ when $\theta_k = 1,\ k = 1, 2, \ldots$ provided $\sum_{i=0}^{k} t_i \to \infty$. Similarly, Theorem 2 implies that the iterates generated by Algorithm 1 satisfy $\phi(x_k) \to \bar{\phi}$ when $\theta_{k+1}^2 \geq \theta_k^2(1-\theta_{k+1})$ provided $t_k/\theta_k^2 \to \infty$. We note that the condition $\sum_{i=0}^{k} t_i \to \infty$ is implied by and therefore weaker than the popular Lipschitz continuity assumption on ∇f. Likewise for the condition $t_k/\theta_k^2 \to \infty$ when $\theta_k = 2/(k+2),\ k = 0, 1, 2, \ldots$ or $\theta_{k+1}^2 = \theta_k^2(1-\theta_{k+1}),\ k = 0, 1, 2, \ldots$.

1.3 Proximal Subgradient Method

Algorithm 2 describes a variant of Algorithm 1 for the case when $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is merely convex.

Algorithm 2 Proximal subgradient method

1: **input:** $x_0 \in \mathrm{dom}(f)$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: pick $g_k^f \in \partial f(x_k)$ and $t_k > 0$
4: $x_{k+1} := \mathsf{Prox}_{t_k}(x_k - t_k g_k^f)$
5: **end for**

When Ψ is the indicator function I_C of a closed convex set $C \subseteq \mathrm{dom}(f)$, Step 4 in Algorithm 2 can be rewritten as $x_{k+1} = \underset{x \in C}{\mathrm{argmin}} \|x_k - t_k \cdot g_k^f - x\| = P_C(x_k - t_k \cdot g_k^f)$, where P_C is the projection onto the set C. Hence when $\Psi = I_C$ Algorithm 2 becomes the projected subgradient method for

$$\min_{x \in C} f(x). \tag{1.12}$$

The classical convergence rate for the projected subgradient method is an immediate consequence of Theorem 3 as we detail below. Observe that

$$x_{k+1} = \mathsf{Prox}_{t_k}(x_k - t_k g_k^f) \Leftrightarrow x_{k+1} = x_k - t_k \cdot g_k$$

where $g_k = g_k^f + g_k^\Psi$ for some $g_k^\Psi \in \partial\Psi(x_{k+1})$. Next, let $z_k \in \mathbb{R}^n,\ k = 0, 1, 2 \ldots$ be as follows

$$z_k = \frac{\sum_{i=0}^{k} t_i g_i}{\sum_{i=0}^{k} t_i}. \tag{1.13}$$

Theorem 3. *Let $x_k \in \mathbb{R}^n$, $k = 0, 1, 2, \ldots$ be the sequence of iterates generated by Algorithm 2 and let $z_k \in \mathbb{R}^n$, $k = 0, 1, 2 \ldots$ be defined by (1.13). Then for $k = 0, 1, 2, \ldots$*

$$\frac{\sum_{i=0}^k t_i(f(x_i) + \Psi(x_{i+1})) + \frac{1}{2}\sum_{i=0}^k t_i^2(\|g_i^\Psi\|^2 - \|g_i^f\|^2)}{\sum_{i=0}^k t_i} \leq -f^*(z_k) + \langle z_k, x_0\rangle - \frac{\sum_{i=0}^k t_i}{2}\|z_k\|^2. \quad (1.14)$$

Let $C \subseteq \mathbb{R}^n$ be a nonempty closed convex set and $\Psi = I_C$. As noted above, in this case Algorithm 2 becomes the projected subgradient algorithm for problem (1.12). We next show that in this case Theorem 3 yields the classical convergence rates (1.16) and (1.17), as well and the modern and more general one (1.18) recently established by Grimmer [29, Theorem 5].

Suppose $\bar{f} = \min_{x\in C} f(x)$ is finite and $\bar{X} := \{x \in C : f(x) = \bar{f}\}$ is nonempty. From Theorem 3 it follows that

$$\frac{\sum_{i=0}^k t_i f(x_i) + \frac{1}{2}\sum_{i=0}^k t_i^2(\|g_i^\Psi\|^2 - \|g_i^f\|^2)}{\sum_{i=0}^k t_i} \leq \inf_{u\in C}\{f(u) - \langle z_k, u\rangle\} + \min_u \left\{\langle z_k, u\rangle + \frac{1}{2\sum_{i=0}^k t_i}\|u - x_0\|^2\right\} \leq \bar{f} + \frac{\mathsf{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k t_i}.$$

Therefore,

$$\sum_{i=0}^k t_i(f(x_i) - \bar{f}) \leq \frac{\sum_{i=0}^k t_i^2(\|g_i^f\|^2 - \|g_i^\Psi\|^2) + \mathsf{dist}(x_0, \bar{X})^2}{2}. \quad (1.15)$$

In particular, if $\|g\| \leq L$ for all $x \in C$ and $g \in \partial f(x)$ then (1.15) implies

$$\min_{i=0,\ldots,k}(f(x_i) - \bar{f}) \leq \frac{\sum_{i=0}^k t_i^2 L^2 + \mathsf{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k t_i}. \quad (1.16)$$

Let $\alpha_i := t_i\|g_i^f\|$, $i = 0, 1, \ldots$. Then Step 4 in Algorithm 2 can be rewritten as $x_{k+1} = P_C\left(x_k - \alpha_k \cdot g_k^f/\|g_k^f\|\right)$ provided $\|g_k^f\| > 0$, which occurs as long as x_k is not an optimal solution to (1.12). If $\|g_i^f\| > 0$ for $i = 0, 1, \ldots, k$ and $\|g\| \leq L$ for all $x \in C$ and $g \in \partial f(x)$ then (1.15) implies

$$\min_{i=0,\ldots,k}(f(x_i) - \bar{f}) \leq L \cdot \frac{\sum_{i=0}^k \alpha_i^2 + \mathsf{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k \alpha_i}. \quad (1.17)$$

Let $\mathcal{L} : \mathbb{R}_+ \to \mathbb{R}_+$. According to the definition of Grimmer [29], the subgradient oracle for f is $\mathcal{L}$-steep on C if for all $x \in C$ and $g \in \partial f(x)$

$$\|g\| \leq \mathcal{L}(f(x) - \bar{f}).$$

As discussed by Grimmer [29], $\mathcal{L}$-steepness is a more general condition than the traditional bound $\|g\| \leq L$ for $x \in C$ and $g \in \partial f(x)$ used above. Indeed, the latter bound is precisely $\mathcal{L}$-steepness for the constant function $\mathcal{L}(t) = L$ and holds when f is L-Lipschitz on C. For another example of $\mathcal{L}$-steepness, consider the case when $C = \mathbb{R}^n$, and f is differentiable on $\mathbb{R}^n$ and such that ∇f is L-Lipschitz. In this case it readily follows that

$$\bar{f} \leq \min_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \right\} = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2.$$

Thus the subgradient oracle for f is $\mathcal{L}$-steep for $\mathcal{L}(t) = \sqrt{2Lt}$. More generally, if ∇f is Hölder-continuous, that is, if there exist L and $v > 0$ such that for all $x, y \in \mathbb{R}^n$

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|x - y\|^v,$$

then

$$\begin{aligned} \bar{f} &\leq \min_y \left\{ f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{1+v}\|y - x\|^{1+v} \right\} \\ &= f(x) - \frac{v}{1+v} \cdot \frac{1}{L^{\frac{1}{v}}} \|\nabla f(x)\|^{\frac{1+v}{v}}. \end{aligned}$$

Thus the subgradient oracle for f is $\mathcal{L}$-steep for $\mathcal{L}(t) = ((1+v)^v L t^v / v^v)^{1/(1+v)}$.

Suppose the subgradient oracle for f is $\mathcal{L}$-steep for some $\mathcal{L} : \mathbb{R}_+ \to \mathbb{R}_+$. If $\alpha_i := t_i \|g_i^f\| > 0$ for $i = 0, 1, \ldots, k$ then (1.15) implies

$$\sum_{i=0}^{k} \alpha_i \cdot \frac{f(x_i) - \bar{f}}{\mathcal{L}(f(x_i) - \bar{f})} \leq \frac{\sum_{i=0}^k \alpha_i^2 + \mathsf{dist}(x_0, \bar{X})^2}{2},$$

and thus

$$\min_{i=0,\ldots,k} (f(x_i) - \bar{f}) \leq \sup \left\{ t : \frac{t}{\mathcal{L}(t)} \leq \frac{\sum_{i=0}^k \alpha_i^2 + \mathsf{dist}(x_0, \bar{X})^2}{2\sum_{i=0}^k \alpha_i} \right\}. \tag{1.18}$$

For $\alpha_i = a/\sqrt{k+1},\ i = 0, \ldots, k$ with $a > 0$ inequality (1.18) yields

$$\min_{i=0,\ldots,k} (f(x_i) - \bar{f}) \leq \sup \left\{ t : \frac{t}{\mathcal{L}(t)} \leq \frac{1}{2\sqrt{k+1}} \left(a + \frac{\mathsf{dist}(x_0, \bar{X})^2}{a} \right) \right\}. \tag{1.19}$$

As we discussed above, when f is L-Lipschitz on C then the subgradient oracle is $\mathcal{L}$-steep for $\mathcal{L}(t) = L$. Hence inequality (1.19) yields the classical $\mathcal{O}(1/\sqrt{k})$ convergence rate of the projected subgradient method. Furthermore, when $C = \mathbb{R}^n$ and f is differentiable and ∇f is L-Lipschitz, then the subgradient oracle is $\mathcal{L}$-steep for $\mathcal{L}(t) = \sqrt{2Lt}$. Hence inequality (1.19) yields

$$\min_{i=0,\ldots,k} (f(x_i) - \bar{f}) = \mathcal{O}(1/k) \tag{1.20}$$

which matches the dependence on k of the classical convergence rate of the gradient method. As noted by Grimmer [29], it is striking that (1.20) holds for Algorithm 2 which relies only on the availability of a subgradient oracle for f. However, we should note that the $\mathcal{O}(1/k)$ rate attained by Algorithm 2 depends on the choice $\alpha_i = a/\sqrt{k+1},\ i = 0, \ldots, k$. In particular, (1.20) holds for a prescribed number k of iterations and the constant in the $\mathcal{O}(1/k)$ expression in (1.20) depends on how closely a approximates $\mathsf{dist}(x_0, \bar{X})$.

1.4 Proofs of Theorems 1, 2, and 3

Our convex conjugate-based analysis depends on standard convex analysis notation and results as presented in [5, 14, 35, 65]. We will use the following properties of the convex conjugate.

Suppose $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a convex function. Then

$$h^*(z) + h(x) \geq \langle z, x \rangle \tag{1.21}$$

for all $z, x \in \mathbb{R}^n$, and equality holds if $z \in \partial h(x)$.

Suppose $f, \phi, \Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are convex functions and $\phi = f + \Psi$. Then

$$\phi^*(z^f + z^\Psi) \leq f^*(z^f) + \Psi^*(z^\Psi) \ \text{ for all } z^f, z^\Psi \in \mathbb{R}^n. \tag{1.22}$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}_+ \cup \{\infty\}$ is a convex function and $R \geq 1$. Then

$$(R \cdot f)^*(Rz) = R \cdot (f^*(z)), \tag{1.23}$$

and

$$(R \cdot f)^*(z) \leq f^*(z). \tag{1.24}$$

Proof of Theorem 1

We prove (1.6) by induction. To ease notation, let $\mu_k := \frac{1}{\sum_{i=0}^{k-1} t_i/\theta_i}$ throughout this proof. For $k = 1$ we have

$$\begin{aligned}
\mathsf{LHS}_1 = \phi(x_1) &\leq f(x_0) + \Psi(x_1) + \left\langle g_0^\Psi, x_0 - x_1 \right\rangle - \frac{t_0}{2}\|g_0\|^2 \\
&= f(x_0) - \left\langle g_0^f, x_0 \right\rangle + \Psi(x_1) - \left\langle g_0^\Psi, x_1 \right\rangle + \langle g_0, x_0 \rangle - \frac{t_0}{2}\|g_0\|^2 \\
&= -f^*(g_0^f) - \Psi^*(g_0^\Psi) + \langle g_0, x_0 \rangle - \frac{t_0}{2}\|g_0\|^2 \\
&\leq -\phi^*(z_1) + \langle z_1, x_0 \rangle - \frac{\|z_1\|^2}{2\mu_1}.
\end{aligned}$$

The first step follows from (1.4). The third step follows from (1.21) and $g_0^f = \nabla f(x_0)$, $g_0^\Psi \in \partial\Psi(x_1)$. The last step follows from (1.22), the choice of $z_1 = g_0 = g_0^f + g_0^\Psi$, and $\mu_1 = 1/t_0$.

Suppose (1.6) holds for k and let $\gamma_k = \frac{t_k/\theta_k}{\sum_{i=0}^{k} t_i/\theta_i}$. The construction (1.5) implies that

$$\begin{aligned} z_{k+1} &= (1-\gamma_k)z_k + \gamma_k g_k \\ \mu_{k+1} &= (1-\gamma_k)\mu_k. \end{aligned}$$

Therefore,

$$\begin{aligned} \langle z_{k+1}, x_0\rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} = (1-\gamma_k)\left(\langle z_k, x_0\rangle - \frac{\|z_k\|^2}{2\mu_k}\right) \\ + \gamma_k\left(\left\langle g_k, x_0 - \frac{z_k}{\mu_k}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right). \end{aligned} \tag{1.25}$$

In addition, the convexity of ϕ^*, properties (1.21), (1.22), and $g_k^f = \nabla f(y_k)$, $g_k^\Psi \in \partial\Psi(x_{k+1})$, $g_k = g_k^f + g_k^\Psi$ imply

$$\begin{aligned} -\phi^*(z_{k+1}) &\geq -(1-\gamma_k)\phi^*(z_k) - \gamma_k f^*(g_k) \\ &\geq -(1-\gamma_k)\phi^*(z_k) - \gamma_k(f^*(g_k^f) + \Psi^*(g_k^\Psi)) \\ &= -(1-\gamma_k)\phi^*(z_k) - \gamma_k\left(\left\langle g_k^f, y_k\right\rangle - f(y_k) + \left\langle g_k^\Psi, x_{k+1}\right\rangle - \Psi(x_{k+1})\right). \end{aligned} \tag{1.26}$$

Let RHS_k denote the right-hand side in (1.6). From (1.25) and (1.26) it follows that

$$\mathsf{RHS}_{k+1} - (1-\gamma_k)\mathsf{RHS}_k \geq \gamma_k \cdot D_k$$

where

$$D_k := \left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k}\right\rangle + f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^\Psi, y_k - x_{k+1}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2.$$

Hence to complete the proof of (1.6) by induction it suffices to show that

$$\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \leq \gamma_k \cdot D_k. \tag{1.27}$$

To that end, we consider case (a) and case (b) separately.

Case (a). In this case $\gamma_k = \frac{t_k}{\sum_{i=0}^{k} t_i}$ and $y_k = x_k$. Thus $\mu_k = \frac{1}{\sum_{i=0}^{k-1} t_i}$, $\frac{\gamma_k}{(1-\gamma_k)\mu_k} = t_k$, and $x_0 - y_k - \frac{z_k}{\mu_k} = 0$. Therefore

$$\begin{aligned} &\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \\ &= \gamma_k \cdot \phi(x_{k+1}) \\ &\leq \gamma_k\left(f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^\Psi, y_k - x_{k+1}\right\rangle - \frac{t_k}{2}\|g_k\|^2\right) \\ &= \gamma_k\left(f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^\Psi, y_k - x_{k+1}\right\rangle - \frac{\gamma_k}{2(1-\gamma_k)\mu_k}\|g_k\|^2\right) \\ &= \gamma_k \cdot D_k. \end{aligned}$$

The second step follows from (1.4). The third and fourth steps follow from $\frac{\gamma_k}{(1-\gamma_k)\mu_k} = t_k$ and $x_0 - y_k - \frac{z_k}{\mu_k} = 0$ respectively. Thus (1.27) holds in case (a).

Case (b). In this case equation (1.8) yields $\gamma_k = \theta_k$ and $\frac{\gamma_k^2}{(1-\gamma_k)\mu_k} = t_k$. Therefore

$$\begin{aligned}
&\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k \\
&= \phi(x_{k+1}) - (1-\gamma_k)(f(x_k) + \Psi(x_k)) \\
&\leq f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^{\Psi}, y_k - x_{k+1} \right\rangle - \frac{t_k}{2}\|g_k\|^2 \\
&\quad - (1-\gamma_k)\left(f(y_k) + \left\langle g_k^{f}, x_k - y_k \right\rangle + \Psi(x_{k+1}) + \left\langle g_k^{\Psi}, x_k - x_{k+1} \right\rangle\right) \\
&= \gamma_k \left(f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^{\Psi}, y_k - x_{k+1} \right\rangle\right) + (1-\gamma_k)\left\langle g_k, y_k - x_k \right\rangle - \frac{t_k}{2}\|g_k\|^2 \\
&= \gamma_k \cdot D_k.
\end{aligned}$$

The second step follows from (1.4) and the convexity of f and Ψ. The last step follows from $\theta_k = \gamma_k$, equation (1.3), and $\frac{\gamma_k^2}{(1-\gamma_k)\mu_k} = t_k$. Thus (1.27) holds in case (b) as well.

Proof of Theorem 2

The proof of Theorem 2 is a modification of the proof of Theorem 1. Without loss of generality assume $\bar{\phi} = 0$ as otherwise we can work with $\phi - \bar{\phi}$ in place of ϕ. Again we prove (1.9) by induction. To ease notation, let $\mu_k := \theta_{k-1}^2/t_{k-1}$ throughout this proof. For $k = 1$ inequality (1.9) is identical to (1.6) since $R_1 = 1$ and $\theta_0 = 1$. Hence this case follows from the proof of Theorem 1 for $k = 1$. Suppose (1.9) holds for k. Observe that

$$\begin{aligned}
z_{k+1} &= \rho_k(1-\theta_k)z_k + \theta_k g_k \\
\mu_{k+1} &= \rho_k(1-\theta_k)\mu_k
\end{aligned}$$

for $\rho_k := \frac{R_{k+1}}{R_k} = \frac{t_{k-1}}{t_k} \cdot \frac{\theta_k^2}{\theta_{k-1}^2(1-\theta_k)} = \frac{\mu_{k+1}}{\mu_k(1-\theta_k)} \geq 1$. Next, proceed as in the proof of Theorem 1. First,

$$\begin{aligned}
&\langle z_{k+1}, x_0 \rangle - \frac{\|z_{k+1}\|^2}{2\mu_{k+1}} \qquad (1.28) \\
&= \rho_k(1-\theta_k)\left(\langle z_k, x_0 \rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \theta_k \cdot \left\langle g_k, x_0 - \frac{z_k}{\mu_k} \right\rangle - \frac{\theta_k^2}{2\mu_{k+1}}\|g_k\|^2 \\
&= \rho_k(1-\theta_k)\left(\langle z_k, x_0 \rangle - \frac{\|z_k\|^2}{2\mu_k}\right) + \theta_k \cdot \left\langle g_k, x_0 - \frac{z_k}{\mu_k} \right\rangle - \frac{t_k}{2}\|g_k\|^2.
\end{aligned}$$

Second, the convexity of ϕ^* and the fact that $\phi \geq \bar{\phi} = 0$ imply

$$\begin{aligned}
-(R_{k+1} \cdot \phi)^*(z_{k+1}) &\geq -(1-\theta_k)(R_{k+1} \cdot \phi)^*(\rho_k \cdot z_k) - \theta_k (R_{k+1} \cdot \phi)^*(g_k) \\
&\geq -(1-\theta_k)(\rho_k \cdot R_k \cdot \phi)^*(\rho_k \cdot z_k) - \theta_k \cdot \phi^*(g_k) \\
&\geq -\rho_k(1-\theta_k)(R_k \cdot \phi)^*(z_k) - \theta_k (f^*(g_k^f) + \Psi^*(g_k^\Psi)) \\
&= -\rho_k(1-\theta_k)(R_k \cdot \phi)^*(z_k) \\
&\quad - \theta_k \left(\left\langle g_k^f, y_k \right\rangle - f(y_k) + \left\langle g_k^\Psi, x_{k+1} \right\rangle - \Psi(x_{k+1}) \right).
\end{aligned} \tag{1.29}$$

The first step follows from the convexity of ϕ^*. The second step follows from (1.24). The third step follows from (1.22) and (1.23). The last step follows from (1.21) and $g_k^f = \nabla f(y_k),\ g_k^\Psi \in \partial \Psi(x_{k+1})$.

Let RHS_k denote the right-hand side in (1.9). The induction hypothesis implies that $\mathsf{RHS}_k \geq \phi(x_k) \geq 0$. Thus from (1.28), (1.29), and $\rho_k \geq 1$ it follows that

$$\begin{aligned}
&\mathsf{RHS}_{k+1} - (1-\theta_k)\mathsf{RHS}_k \\
&\geq \mathsf{RHS}_{k+1} - \rho_k(1-\theta_k)\mathsf{RHS}_k \\
&\geq \theta_k \left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k} \right\rangle + f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^\Psi, y_k - x_{k+1} \right\rangle \right) - \frac{t_k}{2}\|g_k\|^2.
\end{aligned} \tag{1.30}$$

Finally, proceeding exactly as in case (b) in the proof of Theorem 1 we get

$$\begin{aligned}
&\phi(x_{k+1}) - (1-\theta_k)\phi(x_k) \\
&\leq \theta_k \left(f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^\Psi, y_k - x_{k+1} \right\rangle \right) + (1-\theta_k) \left\langle g_k, y_k - x_k \right\rangle - \frac{t_k}{2}\|g_k\|^2 \\
&= \theta_k \left(\left\langle g_k, x_0 - y_k - \frac{z_k}{\mu_k} \right\rangle + f(y_k) + \Psi(x_{k+1}) + \left\langle g_k^\Psi, y_k - x_{k+1} \right\rangle \right) - \frac{t_k}{2}\|g_k\|^2 \\
&\leq \mathsf{RHS}_{k+1} - (1-\theta_k)\mathsf{RHS}_k.
\end{aligned}$$

The second step follows from (1.3). The third step follows from (1.30). This completes the proof by induction.

Proof of Theorem 3

Let LHS_k and RHS_k denote respectively the left-hand and right-hand sides in (1.14). We proceed by induction. For $k = 0$ we have

$$\begin{aligned}
\mathsf{LHS}_0 &= f(x_0) + \Psi(x_1) + \frac{t_0(\|g_0^\Psi\|^2 - \|g_0^f\|^2)}{2} \\
&= -f^*(g_0^f) + \left\langle g_0^f, x_0 \right\rangle - \Psi^*(g_0^\Psi) + \left\langle g_0^\Psi, x_1 \right\rangle + \frac{t_0(\|g_0^\Psi\|^2 - \|g_0^f\|^2)}{2} \\
&\leq -\phi^*(g_0) + \langle g_0, x_0 \rangle - \frac{t_0\|g_0\|^2}{2} \\
&= \mathsf{RHS}_0.
\end{aligned}$$

The second step follows from (1.21) and $g_0^f \in \partial f(x_0),\ g_0^\Psi \in \partial \Psi(x_1)$. The third step follows from (1.22) and $g_0 = g_0^f + g_0^\Psi,\ x_1 = x_0 - t_0 \cdot g_0$.

Next we show the main inductive step k to $k+1$. Observe that $z_{k+1} = (1-\gamma_k)z_k + \gamma_k g_{k+1}$ for $k = 0, 1, \ldots$ where $\gamma_k = \frac{t_{k+1}}{\sum_{i=0}^{k+1} t_i} \in (0,1)$. Proceeding exactly as in the proof of Theorem 1 we get

$$\begin{aligned}
&\mathsf{RHS}_{k+1} - (1-\gamma_k)\mathsf{RHS}_k \\
&\qquad \geq \gamma_k \left(f(x_{k+1}) + \Psi(x_{k+2}) + \left\langle g_{k+1}^\Psi, x_{k+1} - x_{k+2} \right\rangle - \frac{t_{k+1}\|g_{k+1}\|^2}{2} \right) \\
&\qquad = \gamma_k \left(f(x_{k+1}) + \Psi(x_{k+2}) + \frac{t_{k+1}\|g_{k+1}^\Psi\|^2}{2} - \frac{t_{k+1}\|g_{k+1}^f\|^2}{2} \right).
\end{aligned}$$

The second step follows because $g_{k+1} = g_{k+1}^f + g_{k+1}^\Psi$ and $x_{k+2} = x_{k+1} - t_{k+1} \cdot g_{k+1}$. The proof is thus completed by observing that

$$\mathsf{LHS}_{k+1} - (1-\gamma_k)\mathsf{LHS}_k = \gamma_k \left(f(x_{k+1}) + \Psi(x_{k+2}) + \frac{t_{k+1}\|g_{k+1}^\Psi\|^2}{2} - \frac{t_{k+1}\|g_{k+1}^f\|^2}{2} \right).$$

Chapter 2

Acceleration, Part II: A Unified Framework for Bregman Proximal Methods: Subgradient, Gradient, and Accelerated Gradient Schemes

2.1 Introduction

The central contribution of this chapter is a framework to analyze the convergence of *Bregman proximal first-order methods* for the convex composite minimization problem

$$\min_{x\in\mathbb{R}^n} \phi(x) := f(x) + \Psi(x). \tag{2.1}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ are closed convex functions. The class of Bregman proximal first-order methods, a flexible generalization of the proximal method class of chapter 1, are based on the *Bregman proximal map*

$$(x, g) \mapsto \operatorname*{argmin}_{y\in\mathbb{R}^n} \{\langle g, y\rangle + \Psi(y) + LD_h(y, x)\} \tag{2.2}$$

where $D_h(y, x) := h(y) - h(x) - \langle \nabla h(x), y - x\rangle$ is the Bregman distance [16] generated by some *reference* convex function $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$. Naturally, Bregman proximal methods rely on the critical assumption that the problem (2.2) is well-posed and has a computable solution.

Our framework hinges on the *convex conjugate* introduced in chapter 1. The framework can be seen as a natural extension of the approach that was introduced in [59, 31] and discussed in chapter 1, which was restricted to the Euclidean setting. We rely on standard convex analysis notation and results as presented in [5, 14, 35, 65]. By construction the convex conjugate function F^* is convex and satisfies the following *Fenchel inequality:* For all $x \in \mathbb{R}^n$, $u \in \mathbb{R}^n$ we have $F(x) + F^*(u) \geq \langle u, x\rangle$ and $F(x) + F^*(u) = \langle u, x\rangle$ if and only if $u \in \partial F(x)$.

Our convex conjugate framework automatically yields new derivations of convergence rates for the Bregman proximal subgradient method and for the Bregman proximal gradient method (Sections 2.2 and 2.3). In addition, and perhaps most interesting, our convex conjugate framework also applies to a new accelerated Bregman proximal gradient method (Section 2.4). The gist of our convex conjugate approach can be summarized as follows. Suppose $y \in \operatorname{dom}(\phi)$ and a convex distance function $D : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ satisfies

$$\phi(y) \leq -\phi^*(u) - D^*(u). \tag{2.3}$$

for an appropriately selected u. From (2.3) it immediately follows that $\phi(y) - \phi(x) \leq D(x)$ for all $x \in \operatorname{dom}(\phi)$ since (2.3) implies

$$\phi(y) \leq \inf_{w \in \mathbb{R}^n} \{\phi(w) + \langle \nabla D(z), w \rangle\} + \inf_{w \in \mathbb{R}^n} \{D(w) - \langle \nabla D(z), w \rangle\} \leq \phi(x) + D(x).$$

In the main sections of the chapter we show that three classes of Bregman proximal methods (subgradient, gradient, accelerated gradient) generate sequences $x_k, z_k \in \operatorname{dom}(\phi) \cap \mathsf{relint}(\operatorname{dom}(h)),\ k = 0, 1, 2, \ldots$ such that (2.3), or a slight modification of it, holds for $y = x_k,\ z = z_k$, and $D(\cdot) = C_k D_h(\cdot, x_0)$ for some nondecreasing sequence $C_k \in \mathbb{R}_+,\ k = 0, 1, 2, \ldots$. More precisely, Theorem 6 shows that (2.3) holds for the accelerated Bregman proximal gradient method iterates, see (2.11). Theorem 5 shows that an inequality stronger than (2.3) holds for the Bregman proximal gradient method iterates, see (2.7). Theorem 4 shows that a slight variation of (2.3) holds for the Bregman proximal subgradient method iterates, see (2.4). In particular, for the Bregman proximal gradient and accelerated Bregman proximal gradient methods Theorem 5 and Theorem 6 yield

$$\phi(x_k) - \phi(x) \leq C_k D_h(x, x_0)$$

for all $x \in \operatorname{dom}(\phi)$. We also get a similar inequality for the Bregman proximal subgradient method. In each case it will be easy to see that the sequence $C_k,\ k = 0, 1, \ldots$ goes to zero under fairly mild and general assumptions. In particular, we show that the sequence C_k is as follows under suitable assumptions on f, ϕ, and h:

- For the Bregman proximal subgradient method $C_k = \mathcal{O}(1/\sqrt{k})$ if the pair (ϕ, h) satisfies the $W[\phi, h]$ boundedness condition as defined in [69]. See Corollary 2.
- For the Bregman proximal gradient method $C_k = \mathcal{O}(1/k)$ if f is *smooth relative to* h as defined by [4, 48]. See Corollary 3.
- For the accelerated Bregman proximal gradient method $C_k = \mathcal{O}(1/k^\gamma)$ if f is *smooth relative to* h and D_h has a *triangle scaling exponent* $\gamma > 0$ as defined in [34]. See Theorem 7.

The above results yield new derivations of known convergence rates via our convex conjugate approach. However, our main results, namely Theorem 4, Theorem 5, and

Theorem 6 hold more broadly. In particular, Theorem 4 only requires the Bregman steps to be *admissible* as defined below. Theorem 5 and Theorem 6 only require the Bregman steps to be admissible and to satisfy a suitable *decrease condition*. None of these three main results requires any further assumptions like Lipschitz continuity or relative smoothness.

The main sections of the chapter are organized as follows. Sections 2.2 through 2.4 develop our convex conjugate approach in the contexts of the Bregman proximal subgradient, Bregman proximal gradient, and accelerated Bregman proximal gradient templates. In the latter case we discuss the connection between our work and the recent work of Hanzely, Richtarik and Xiao [34]. Section 2.5 shows that a variant of our accelerated Bregman proximal gradient template that includes periodic restart has linear convergence provided that suitable smoothness and functional growth conditions hold. Finally, Section 2.6 summarizes some numerical experiments on the D-optimal design problem and on the Poisson linear inverse problem. Consistent with the numerical evidence reported in [34], we observe that the accelerated Bregman proximal gradient method converges approximately at a rate $\mathcal{O}(1/k^2)$. Furthermore, our computational experiments provide interesting new numerical evidence that explains this behavior.

2.1.1 Technical Assumptions

We aim to present our developments in as much generality as possible. To that end, throughout the chapter we make the blanket Assumption 1 below. We should note that the *admissibility condition* (A.3) is primarily a technicality. This condition is concerned with the choice of $L > 0$ that guarantees the well-posedness of problem (2.2). As Example 1(b,c) below illustrates, in many cases problem (2.2) is readily well-posed and thus the admissibility condition (A.3.i) automatically holds for all $L > 0, g \in \mathbb{R}^n, x \in \mathsf{relint}(\mathrm{dom}(h))$. However, Example 1(a) also illustrates that in some cases the well-posedness of problem (2.2) may require a more careful choice of $L > 0$.

Assumption 1. The functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$, and $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ satisfy the following conditions.

(A.1) The functions f and Ψ are closed and convex. Throughout the sequel, we let $\phi := f + \Psi$.

(A.2) The reference function h is convex and differentiable on $\mathsf{relint}(\mathrm{dom}(h))$ and satisfies $\mathrm{dom}(\Psi) \subseteq \overline{\mathrm{dom}(h)}$ and $\emptyset \neq \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi) \subseteq \mathsf{relint}(\mathrm{dom}(f))$.

(A.3) The pair of functions (h, Ψ) satisfies the following *admissibility conditions:*

  (i) For all $g \in \mathbb{R}^n$ and $x \in \mathsf{relint}(\mathrm{dom}(h))$ there exists $L > 0$ such that the Bregman proximal map (2.2) has a unique solution in $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$. When this holds we shall say that L is *admissible* for g at x.

(ii) There is an oracle that takes as input $g \in \mathbb{R}^n, x \in \mathsf{relint}(\mathrm{dom}(h)), L > 0$ and yields as output either a certificate that L is not admissible for g at x or the unique solution to (2.2) in $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$. Observe that in the latter case the solution to (2.2) is the unique point $y \in \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$ that satisfies the optimality conditions

$$g + g^{\Psi} + L(\nabla h(y) - \nabla h(x)) = 0,\ g^{\Psi} \in \partial\Psi(y).$$

Observe that a constraint of the form $x \in C$ for a closed convex set $C \subseteq \mathbb{R}^n$ can be easily incorporated in the above setting by adding the indicator function I_C to Ψ. The admissibility condition (A.3.i) can be ensured under suitable assumptions on Ψ and h. In particular, as detailed in [4, 69], condition (A.3.i) holds when h is a Legendre function [65] and Ψ is bounded below and satisfies $\mathsf{relint}(\mathrm{dom}(\Psi)) \subseteq \mathsf{relint}(\mathrm{dom}(h))$, see [69, Lemma 2.3]. Furthermore, in concrete applications it is often easy to verify directly the admissibility conditions (A.3.i) and (A.3.ii) as Example 1 shows. For simplicity, Example 1 assumes that $\Psi = 0$. The admissibility properties in Example 1 can be extended to popular choices of regularization functions Ψ such as $\Psi(x) = \lambda\|x\|_2^2/2$ or $\Psi(x) = \lambda\|x\|_1$ for $\lambda > 0$. They can also be extended to popular choices of indicator functions such as $\Psi = \delta_{\Delta_{n-1}}$ for $\Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$.

Example 1. Suppose $\Psi = 0$. The admissibility conditions (A.3.i) and (A.3.ii) hold for the following reference functions $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$.

(a) The Burg entropy function $h(x) := -\sum_{i=1}^n \log(x_i)$. In this case $L > 0$ is admissible for $g \in \mathbb{R}^n$ at $x \in \mathbb{R}^n_{++} = \mathsf{relint}(\mathrm{dom}(h))$ if and only if $-\nabla h(x) + g/L \in \mathbb{R}^n_{++}$ and in this case the solution to (2.2) is the vector $y \in \mathbb{R}^n_{++}$ defined componentwise as $y_i = 1/(1/x_i + g_i/L),\ i = 1, \ldots, n$.

(b) The Boltzmann-Shannon entropy function $h(x) := \sum_{i=1}^n x_i \log(x_i)$. In this case any $L > 0$ is admissible for any $g \in \mathbb{R}^n$ at any $x \in \mathbb{R}^n_{++} = \mathsf{relint}(\mathrm{dom}(h))$ and the solution to (2.2) is the vector $y \in \mathbb{R}^n_+$ defined componentwise as $y_i = e^{\log(x_i) - g_i/L},\ i = 1, \ldots, n$.

(c) The squared Euclidean function $h(x) := \|x\|_2^2/2$. In this case any $L > 0$ is admissible for any $g \in \mathbb{R}^n$ at any $x \in \mathbb{R}^n = \mathsf{relint}(\mathrm{dom}(h))$ and the solution to (2.2) is the vector $y = x - g/L$.

To sharpen some of our results, sometimes we will assume that the pair (h, Ψ) satisfies the *sufficient admissibility condition* defined below. Observe that this condition is satisfied by the three reference functions h in Example 1 and the popular choices of Ψ mentioned above. By [69, Lemma 2.3], the sufficient admissibility condition also holds when h is a Legendre function and $\mathsf{relint}(\mathrm{dom}(\Psi)) \subseteq \mathsf{relint}(\mathrm{dom}(h))$.

Definition 1. Let $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function differentiable on $\mathsf{relint}(\mathrm{dom}(h))$ and let $\Psi : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a closed convex function with $\mathrm{dom}(\Psi) \subseteq \overline{\mathrm{dom}(h)}$ and

$\operatorname{dom}(\Psi) \cap \mathsf{relint}(\operatorname{dom}(h)) \neq \emptyset$. The pair (h, Ψ) satisfies the *sufficient admissibility condition* if $L > 0$ is admissible for $g \in \mathbb{R}^n$ at $x \in \mathsf{relint}(\operatorname{dom}(h))$ whenever the function

$$y \mapsto \langle g, y \rangle + \Psi(y) + LD_h(y, x)$$

is bounded below.

We will rely on properties of the convex conjugate [5, 14, 35, 65] and on the following *three-point property* [18, Lemma 3.1] of the Bregman distance induced by h. For all $a \in \operatorname{dom}(h)$ and $b, c \in \mathsf{relint}(\operatorname{dom}(h))$

$$D_h(a, b) + D_h(b, c) = D_h(a, c) - \langle \nabla h(b) - \nabla h(c), a - b \rangle .$$

2.2 Bregman Proximal Subgradient

We first consider the case when f is convex and we only have a subgradient oracle for f. Algorithm 3 describes a Bregman proximal subgradient template for (2.1). This algorithmic template has been discussed in [23, 69]. Observe that Step 1 and Step 4 in Algorithm 3 automatically guarantee that $x_k \in \mathsf{relint}(\operatorname{dom}(f))$, $k = 0, 1, \ldots$ by conditions (A.2) and (A.3) in Assumption 1.

Algorithm 3 Bregman proximal subgradient template

1: **input:** $x_0 \in \mathsf{relint}(\operatorname{dom}(h)) \cap \operatorname{dom}(\Psi)$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: pick $g_k \in \partial f(x_k)$ and $t_k > 0$ so that $1/t_k$ is admissible for g_k at x_k
4: $x_{k+1} := \operatorname{argmin}_{x \in \mathbb{R}^n} \{t_k(\langle g_k, x \rangle + \Psi(x)) + D_h(x, x_k)\}$
5: **end for**

Theorem 4. *For $k = 0, 1, 2, \ldots$ and $u_k := \frac{1}{\sum_{i=0}^k t_i}(\nabla h(x_0) - \nabla h(x_{k+1}))$ the iterates generated by Algorithm 3 satisfy*

$$\begin{aligned} &\frac{\sum_{i=0}^k t_i(f(x_i) + \Psi(x_{i+1}) + \langle g_i, x_{i+1} - x_i \rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \\ &\quad \leq -\phi^*(u_k) + \langle u_k, x_{k+1} \rangle + \frac{1}{\sum_{i=0}^k t_i} D_h(x_{k+1}, x_0) \\ &\quad = -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^k t_i} D_h(\cdot, x_0) \right)^* (-u_k). \end{aligned} \tag{2.4}$$

Proof. The optimality conditions for Step 4 in Algorithm 3 can be written as

$$t_k(g_k + g_k^{\Psi}) + \nabla h(x_{k+1}) - \nabla h(x_k) = 0$$

for some $g_k^\Psi \in \partial\Psi(x_{k+1})$. Therefore $u_k = \frac{1}{\sum_{i=0}^k t_i} \sum_{i=0}^k t_i(g_i + g_i^\Psi)$. On the other hand, since $g_k \in \partial f(x_k)$ and $g_k^\Psi \in \partial\Psi(x_{k+1})$ for each $k = 0, 1, \ldots$ and $\phi = f + \Psi$, it follows that

$$\begin{aligned} f(x_k) + \Psi(x_{k+1}) + \langle g_k, x_{k+1} - x_k\rangle &= -f^*(g_k) - \Psi^*(g_k^\Psi) + \left\langle g_k + g_k^\Psi, x_{k+1}\right\rangle \\ &\le -\phi^*(g_k + g_k^\Psi) + \left\langle g_k + g_k^\Psi, x_{k+1}\right\rangle. \end{aligned} \tag{2.5}$$

Now we prove (2.4) by induction on k. The case $k = 0$ readily follows from (2.5) for $k = 0$. Suppose (2.4) holds for k and let $\theta_k := \frac{t_{k+1}}{\sum_{i=0}^{k+1} t_i}$. Observe that $u_{k+1} = (1 - \theta_k)\, u_k + \theta_k(g_k + g_k^\Psi)$. Therefore (2.4), (2.5), the convexity of ϕ^*, and the three-point property of D_h yield

$$\begin{aligned} &\frac{\sum_{i=0}^{k+1} t_i(f(x_i) + \Psi(x_{i+1}) + \langle g_i, x_{i+1} - x_i\rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^{k+1} t_i} \\ &\quad\le -\phi^*(u_{k+1}) + \langle u_{k+1}, x_{k+1}\rangle + \frac{1}{\sum_{i=0}^{k+1} t_i}(D_h(x_{k+1}, x_0) + D_h(x_{k+2}, x_{k+1})) \\ &\quad= -\phi^*(u_{k+1}) + \langle u_{k+1}, x_{k+2}\rangle + \frac{1}{\sum_{i=0}^{k+1} t_i} D_h(x_{k+2}, x_0). \end{aligned}$$

□

Theorem 4 implies the convergence of $\min_{i=0,1,\ldots,k} \phi(x_i)$ to $\min_x \phi(x)$ under fairly mild and general conditions as detailed in Corollary 1 and Corollary 2 below. To that end, we will rely on the following type of boundedness condition discussed by Teboulle [69].

Definition 2. The pair (f, h) satisfies the condition $W[f, h]$ on $C \subseteq \operatorname{dom}(f) \cap \operatorname{dom}(h)$ if there exists some $G > 0$ such that for all $x, u \in C, g \in \partial f(x)$, and $t > 0$ the following inequality holds

$$\langle tg, u - x\rangle - D_h(u, x) \le \frac{G^2 t^2}{2}.$$

As noted by Teboulle [69], the condition $W[f, h]$ holds for $G = L/\sigma$ whenever f is L-Lipschitz and h is σ-strongly convex for some norm on $\mathbb{R}^n$. It is also easy to see that the condition $W[f, h]$ holds if f is G-continuous relative to h as defined by Lu [47].

The following result concerns the special case when $\Psi = I_C$ for some closed convex set $C \subseteq \operatorname{dom}(f) \cap \operatorname{dom}(h)$. In this case Algorithm 3 is the mirror-descent method for the problem

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

Corollary 1. *Suppose $\Psi = I_C$ for some closed convex set $C \subseteq \operatorname{dom}(f) \cap \operatorname{dom}(h)$ and the pair (f, h) satisfies the $W[f, h]$ condition for some $G > 0$ on C. Then the iterates generated by Algorithm 3 satisfy*

$$\min_{i=0,\ldots,k}(f(x_i) - f(x)) \le \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2/2}{\sum_{i=0}^k t_i}.$$

for all $x \in C$.

Proof. In this case $\Psi(x) = 0$ for all $x \in C$ and thus Theorem 4 implies that

$$\begin{aligned}&\frac{\sum_{i=0}^k t_i(f(x_i) + \langle g_i, x_{i+1} - x_i\rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i}\\ &\quad\leq \min_{x\in C}\{f(x) - \langle u_k, x\rangle\} + \min_x \left\{\frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0) + \langle u_k, x\rangle\right\}.\end{aligned}$$

Therefore for all $x \in C$

$$\frac{\sum_{i=0}^k t_i(f(x_i) + \langle g_i, x_{i+1} - x_i\rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i} \leq f(x) + \frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0).$$

Since each $g_i \in \partial f(x_i)$, the convexity of f and $W[f, h]$ condition imply that

$$\begin{aligned}\min_{i=0,\ldots,k}(f(x_i) - f(x)) &\leq \frac{D_h(x, x_0) + \sum_{i=0}^k \langle t g_i, x_i - x_{i+1}\rangle - D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i}\\ &\leq \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2/2}{\sum_{i=0}^k t_i}.\end{aligned}$$

□

For general Ψ, we have the following result discussed in [69]. This result is also closely related to some results by Bello-Cruz [10] on the proximal subgradient method.

Corollary 2. *Suppose the pair* (ϕ, h) *satisfies the* $W[\phi, h]$ *condition for some* $G > 0$ *on* $\mathrm{dom}(\phi)$. *Then the iterates generated by Algorithm 3 satisfy*

$$\min_{i=0,\ldots,k}(\phi(x_i) - \phi(x)) \leq \frac{D_h(x, x_0) + \sum_{i=0}^k t_i^2 G^2/2}{\sum_{i=0}^k t_i}$$

for all $x \in \mathrm{dom}(\phi)$.

Proof. The convexity of Ψ and Theorem 4 imply that

$$\begin{aligned}&\frac{\sum_{i=0}^k t_i(\phi(x_i) + \langle g_i + \tilde{g}_i^\Psi, x_{i+1} - x_i\rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^k t_i}\\ &\quad\leq \min_x\{\phi(x) - \langle u_k, x\rangle\} + \min_x \left\{\frac{1}{\sum_{i=0}^k t_i} D_h(x, x_0) + \langle u_k, x\rangle\right\}\end{aligned}$$

for any $\tilde{g}_i^{\Psi} \in \partial\Psi(x_i)$. Hence for all $x \in \mathrm{dom}(\phi)$

$$\frac{\sum_{i=0}^{k} t_i(\phi(x_i) + \langle g_i + \tilde{g}_i, x_{i+1} - x_i\rangle) + D_h(x_{i+1}, x_i)}{\sum_{i=0}^{k} t_i} \leq f(x) + \frac{1}{\sum_{i=0}^{k} t_i} D_h(x, x_0).$$

Since each $g_i + \tilde{g}_i \in \partial\phi(x_i)$, the convexity of ϕ and $W[\phi, h]$ condition imply that

$$\begin{aligned}\min_{i=0,\ldots,k}(\phi(x_i) - \phi(x)) &\leq \frac{D_h(x, x_0) + \sum_{i=0}^{k} \langle t(g_i + \tilde{g}_i), x_i - x_{i+1}\rangle - D_h(x_{i+1}, x_i)}{\sum_{i=0}^{k} t_i} \\ &\leq \frac{D_h(x, x_0) + \sum_{i=0}^{k} t_i^2 G^2/2}{\sum_{i=0}^{k} t_i}.\end{aligned}$$

□

In both Corollary 1 and Corollary 2 it is easy to see that if $t_i = 1/\sqrt{k+1},\ i = 0, 1, \ldots, k$ are admissible then for this choice of $t_i,\ i = 0, 1, \ldots, k$ we have

$$\min_{i=0,\ldots,k}(\phi(x_i) - \phi(x)) \leq \frac{D_h(x, x_0) + G^2/2}{\sqrt{k+1}}.$$

A closer look at the proof of Corollary 2 also reveals that if $t_i := 1/(i+1),\ i = 0, 1, \ldots$ are admissible then for this choice of $t_i,\ i = 0, 1, \ldots$ we have $\min_{i=0,\ldots,k} \phi(x_i) \to \inf_{x\in\mathbb{R}^n} \phi(x)$ provided the following weaker version of $W[\phi, h]$ holds: there exist $\gamma > 1$ and $G > 0$ such that for all $x, u \in \mathrm{dom}(\phi) \cap \mathrm{dom}(h)$ and $g \in \partial\phi(x)$

$$\langle tg, u - x\rangle - D_h(u, x) \leq (Gt)^{\gamma}.$$

Likewise for Corollary 1.

2.3 Bregman Proximal Gradient

Next, we consider the case when f is differentiable on $\mathsf{relint}(\mathrm{dom}(f))$ and we have a gradient oracle for f. Algorithm 4 describes a Bregman proximal gradient template for (2.1). This template has been discussed in [3, 4, 48, 69]. Observe that Step 1 and Step 4 in Algorithm 4 automatically guarantee that $x_k \in \mathsf{relint}(\mathrm{dom}(f)),\ k = 0, 1, \ldots$ by conditions (A.2) and (A.3) in Assumption 1.

The bound (2.7) in Theorem 5 below is similar to the bound (2.4) in Theorem 4. The similarity is more salient if we let $t_k := 1/L_k,\ k = 0, 1, \ldots$.

Theorem 5. *Suppose $L_k,\ k = 0, 1, \ldots$ in Step 3 of Algorithm 4 are chosen so that the following decrease condition holds*

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + L_k D_h(x_{k+1}, x_k). \tag{2.6}$$

Algorithm 4 Bregman proximal gradient template

input: $x_0 \in \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$
for $k = 0, 1, 2, \ldots$ **do**
  pick $L_k > 0$ admissible for $\nabla f(x_k)$ at x_k
  $x_{k+1} := \mathrm{argmin}_{x \in \mathbb{R}^n} \{\langle \nabla f(x_k), x\rangle + L_k D_h(x, x_k) + \Psi(x)\}$
end for

Then for $k = 1, 2, \ldots$ *and* $u_k := \frac{1}{\sum_{i=0}^{k-1} 1/L_i}(\nabla h(x_0) - \nabla h(x_k))$ *the iterates generated by Algorithm 4 satisfy*

$$\begin{aligned} \frac{\sum_{i=0}^{k-1} \phi(x_{i+1})/L_i}{\sum_{i=0}^{k-1} 1/L_i} &\leq -\phi^*(u_k) + \langle u_k, x_k\rangle + \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x_k, x_0) \\ &= -\phi^*(u_k) - \left(\frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(\cdot, x_0)\right)^* (-u_k). \end{aligned} \tag{2.7}$$

Proof. The optimality conditions for Step 4 in Algorithm 4 can be written as

$$\nabla f(x_k) + g_k^\Psi + L_k(\nabla h(x_{k+1}) - \nabla h(x_k)) = 0 \tag{2.8}$$

for some $g_k^\Psi \in \partial\Psi(x_{k+1})$. In particular, $u_1 = \nabla f(x_0) + g_0^\Psi$. On the other hand, inequality (2.6), the fact that $g_k^\Psi \in \partial\Psi(x_{k+1})$, and $\phi = f + \Psi$ imply that

$$\begin{aligned} \phi(x_{k+1}) &\leq -f^*(\nabla f(x_k)) - \Psi^*(g_k^\Psi) + \left\langle \nabla f(x_k) + g_k^\Psi, x_{k+1}\right\rangle + L_k D_h(x_{k+1}, x_k) \\ &\leq -\phi^*(\nabla f(x_k) + g_k^\Psi) + \left\langle \nabla f(x_k) + g_k^\Psi, x_{k+1}\right\rangle + L_k D_h(x_{k+1}, x_k). \end{aligned} \tag{2.9}$$

We now prove (2.7) by induction on k. The case $k = 1$ readily follows from (2.9) and the fact that $u_1 = \nabla f(x_0) + g_0^\Psi$ noted above. Suppose (2.7) holds for k. Let $\theta_k := \frac{1/L_k}{\sum_{i=0}^{k} 1/L_i}$. From (2.8) it follows that $u_{k+1} = (1 - \theta_k)\, u_k + \theta_k(\nabla f(x_k) + g_k^\Psi)$. Therefore (2.7), (2.9), the convexity of ϕ^*, and the three-point property of D_h yield

$$\begin{aligned} &\frac{\sum_{i=0}^{k} \phi(x_{i+1})/L_i}{\sum_{i=0}^{k} 1/L_i} \\ &\leq -\phi^*(u_{k+1}) + \langle u_{k+1}, x_{k+1}\rangle + \langle u_k, x_k - x_{k+1}\rangle + \frac{1}{\sum_{i=0}^{k} 1/L_i}(D_h(x_k, x_0) + D_h(x_{k+1}, x_k)) \\ &= -\phi^*(u_{k+1}) + \langle u_{k+1}, x_{k+1}\rangle + \frac{1}{\sum_{i=0}^{k} 1/L_i} D_h(x_{k+1}, x_0). \end{aligned}$$

□

Corollary 3. *If the assumptions of Theorem 5 hold then*

$$\phi(x_k) - \phi(x) \leq \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0)$$

for all $x \in \mathrm{dom}(\phi)$.

Proof. From Theorem 5 and the convexity of ϕ it follows that $\phi(x_{k+1}) \leq \phi(x_k)$ and

$$\frac{\sum_{i=0}^{k-1} \phi(x_{i+1})/L_i}{\sum_{i=0}^{k-1} 1/L_i} \leq \min_x\{\phi(x) - \langle u_k, x\rangle\} + \min_x \left\{ \langle u_k, x\rangle + \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0) \right\}.$$

In particular,

$$\phi(x_k) \leq \frac{\sum_{i=0}^{k-1} \phi(x_{i+1})/L_i}{\sum_{i=0}^{k-1} 1/L_i} \leq \phi(x) - \frac{1}{\sum_{i=0}^{k-1} 1/L_i} D_h(x, x_0)$$

for all $x \in \mathrm{dom}(\phi)$. □

Consider the case when f is L_f-*smooth* relative to h on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$ as defined in [4, 48]. This means that $h - L_f f$ is convex on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$ for some constant $L_f > 0$ and as a consequence [4, Lemma 1]

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + L_f D_h(y, x)$$

for all $x, y \in \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$. Suppose that in addition $\phi = f + \Psi$ is bounded below and the pair (h, Ψ) satisfies the sufficient admissibility condition (see Definition 1). Thus to ensure (2.6) we can choose $L_k = L := L_f$ if L_f is known, or more generally $L_k \leq L := \max\{\bar{L}, \alpha L_f\}$ for some $\alpha > 1$ and some initial guess $\bar{L}$ via a standard backtracking procedure. In this case Corollary 3 thus yields the following convergence rate previously established in [4, 48]: for all $x \in \mathrm{dom}(\phi)$

$$\phi(x_k) - \phi(x) \leq \frac{L D_h(x, x_0)}{k}.$$

2.4 Accelerated Bregman Proximal Gradient

The interesting challenge of devising an accelerated version of Algorithm 4 was posed as an open problem in both [69] and [48]. A solution to this challenge was recently given by Hanzely, Richtarik, and Xiao in [34]. We develop a new accelerated Bregman proximal gradient template as described in Algorithm 5. This algorithmic template shares some similarities with Algorithm ABPG in [34] but there are also some key differences. In particular, Algorithm 5 relies only on the decrease condition (2.10) at each iteration. The algorithm does not require explicit knowledge of relative smooth or triangle scaling constants. Like Steps 1, 2, and 3 in [34, Algorithm ABPG], the updates of the sequences x_k, y_k, z_k in Steps 6, 8, and 9 of Algorithm 5 follow the same pattern used in the *Improved Interior Gradient Algorithm* (IGA) in [3].

As in Algorithm ABPG in [34] and in Algorithm IGA in [3], the gist of achieving acceleration in Algorithm 5 is to generate different sequences for the main iterates, the gradients, and the reference points used in the Bregman proximal gradient steps. (See steps 6, 8, and 9.) This is in sharp contrast to Algorithm 4 that generates a single

Algorithm 5 Accelerated Bregman proximal gradient template

input: $x_0 \in \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$; $\theta_0 := 1$ $z_0 := x_0$; $y_0 := x_0$
pick $L_0 > 0$ admissible for $\nabla f(x_0)$ at x_0
$x_1 := z_1 := \mathrm{argmin}_{z\in\mathbb{R}^n}\{\langle \nabla f(x_0), z\rangle + L_0 D_h(z, x_0) + \Psi(z)\}$
for $k = 1, 2, \ldots$ **do**
    pick $\theta_k \in (0, 1)$ so that L_k is admissible for $\nabla f(y_k)$ at z_k for L_k and y_k as below
    $y_k := (1 - \theta_k)x_k + \theta_k z_k$
    $L_k := L_{k-1}\theta_{k-1}(1 - \theta_k)/\theta_k$
    $z_{k+1} := \mathrm{argmin}_{z\in\mathbb{R}^n}\{\langle \nabla f(y_k), z\rangle + L_k D_h(z, z_k) + \Psi(z)\}$
    $x_{k+1} := (1 - \theta_k)x_k + \theta_k z_{k+1}$
end for

sequence. The idea of generating different sequences can be traced back to Nesterov's seminal accelerated gradient algorithm [52] and underlies a number of other accelerated first-order algorithms [8, 22, 52, 53, 54].

The bound (2.11) in Theorem 6 below has a similar format to the bounds (2.4) and (2.7).

Theorem 6. *Suppose $L_0 > 0$ and $\theta_k \in (0, 1]$, $k = 0, 1, 2, \ldots$ are chosen so that each L_k is admissible for $\nabla f(y_k)$ at z_k and the following decrease condition holds for $k = 0, 1, 2, \ldots$*

$$\phi(x_{k+1}) \leq (1 - \theta_k)\phi(x_k) + \theta_k(f(y_k) + \langle \nabla f(y_k), z_{k+1} - y_k\rangle) + L_k D_h(z_{k+1}, z_k) + \Psi(z_{k+1})). \tag{2.10}$$

Then for $k = 1, 2, \ldots$ and $u_k := \theta_{k-1}L_{k-1}(\nabla h(x_0) - \nabla h(z_k))$ the iterates generated by Algorithm 5 satisfy

$$\begin{aligned} \phi(x_k) &\leq -\phi^*(u_k) + \langle u_k, z_k\rangle + \theta_{k-1}L_{k-1}D_h(z_k, x_0) \\ &= -\phi^*(u_k) - (\theta_{k-1}L_{k-1}D_h(\cdot, x_0))^*(-u_k). \end{aligned} \tag{2.11}$$

Proof. The optimality conditions for Step 8 of Algorithm 5 can be written as

$$\nabla f(y_k) + g_k^\Psi + L_k(\nabla h(z_{k+1}) - \nabla h(z_k)) = 0 \tag{2.12}$$

for some $g_k^\Psi \in \partial\Psi(z_{k+1})$. In particular, $u_1 = \nabla f(x_0) + g_0^\Psi$.

We next prove (2.11) by induction. The case $k = 1$ follows from (2.10). Indeed, since $\theta_0 = 1$, $y_0 = x_0, z_1 = x_1$, and $u_1 = \nabla f(x_0) + g_0^\Psi$ with $g_0^\Psi \in \partial\Psi(z_1)$, inequality (2.10) yields

$$\begin{aligned} \phi(x_1) &\leq f(x_0) + \langle \nabla f(x_0), z_1 - x_0\rangle + \theta_0 L_0 D_h(z_1, z_0) + \Psi(z_1) \\ &= -f^*(\nabla f(x_0)) - \Psi^*(g_0^\Psi) + \left\langle \nabla f(x_0) + g_0^\Psi, z_1\right\rangle + \theta_0 L_0 D_h(z_1, z_0) \\ &\leq -\phi^*(u_1) + \langle u_1, z_1\rangle + \theta_0 L_0 D_h(z_1, z_0). \end{aligned}$$

Suppose (2.11) holds for k. Since $\phi = f + \Psi$ and $g_k^\Psi \in \partial\Psi(z_{k+1})$ we have

$$\begin{aligned} -\phi^*(\nabla f(y_k) + g_k^\Psi) &\geq -f^*(\nabla f(y_k)) - \Psi^*(g_k^\Psi) \\ &= f(y_k) - \langle \nabla f(y_k), y_k \rangle + \Psi(z_{k+1}) - \left\langle g_k^\Psi, z_{k+1} \right\rangle. \end{aligned} \tag{2.13}$$

From (2.12) it follows that $u_{k+1} = (1-\theta_k)u_k + \theta_k(\nabla f(y_k) + g_k^\Psi)$. Thus the identity $\theta_k L_k = (1-\theta_k)\theta_{k-1}L_{k-1}$, convexity of ϕ^*, three-point property of D_h, and inequality (2.13) yield

$$\begin{aligned} &-\phi^*(u_{k+1}) + \langle u_{k+1}, z_{k+1} \rangle + \theta_k L_k D_h(z_{k+1}, x_0) \\ &\quad \geq (1-\theta_k)\left(-\phi^*(u_k) + \langle u_k, z_{k+1} \rangle + \theta_{k-1}L_{k-1}D_h(z_{k+1}, x_0)\right) \\ &\qquad + \theta_k(-\phi^*(\nabla f(y_k) + g_k^\Psi) + \left\langle \nabla f(y_k) + g_k^\Psi, z_{k+1} \right\rangle) \\ &\quad \geq (1-\theta_k)\left(-\phi^*(u_k) + \langle u_k, z_k \rangle + \theta_{k-1}L_{k-1}D_h(z, x_0)\right) \\ &\qquad + \theta_k(f(y_k) + \langle \nabla f(y_k), z_{k+1} - y_k \rangle + L_k D_h(z_{k+1}, z_k) + \Psi(z_{k+1})) \end{aligned}$$

Therefore the induction hypothesis and (2.10) imply that

$$\begin{aligned} &-\phi^*(u_{k+1}) + \langle u_{k+1}, z_{k+1} \rangle + \theta_k L_k D_h(z_{k+1}, x_0) \\ &\quad \geq (1-\theta_k)\phi(x_k) + \theta_k(f(y_k) + \langle \nabla f(y_k), z_{k+1} - y_k \rangle + L_k D_h(z_{k+1}, z_k) + \Psi(z_{k+1})) \\ &\quad \geq \phi(x_{k+1}). \end{aligned}$$

□

Corollary 4. *If the assumptions of Theorem 6 hold then for $k = 1, 2, \ldots$*

$$\phi(x_k) - \phi(x) \leq \theta_{k-1}L_{k-1}D_h(x, x_0)$$

for all $x \in \operatorname{dom}(\phi)$.

Proof. From Theorem 6 it follows that

$$\phi(x_k) \leq \min_x\{\phi(x) - \langle u_k, x \rangle\} + \min_x\{\langle u_k, x \rangle + \theta_{k-1}L_{k-1}D_h(x, x_0)\}$$

Thus $\phi(x_k) \leq \phi(x) + \theta_{k-1}L_{k-1}D_h(x, x_0)$ for all $x \in \operatorname{dom}(\phi)$. □

Suppose f is L_f-smooth relative to h on $\mathsf{relint}(\operatorname{dom}(h)) \cap \operatorname{dom}(\Psi)$ and let $\gamma > 0$ be a *triangle scaling exponent* of D_h introduced by Hanzely, Richtarik, and Xiao [34]. That is, for all $x, z, \tilde{z} \in \operatorname{dom}(h)$ and $\theta \in [0, 1]$ the Bregman distance D_h satisfies the following *triangle scaling property*

$$D_h((1-\theta)x + \theta\tilde{z}, (1-\theta)x + \theta z) \leq \theta^\gamma D_h(\tilde{z}, z).$$

The next proposition shows that if both constants L_f and γ are known then the admissibility of L_k and condition (2.10) for $k = 0, 1, \ldots$ can be ensured by choosing of $L_0 := L_f$ and θ_k via $\theta_0 = 1$, $\theta_k^\gamma = (1-\theta_k)\theta_{k-1}^\gamma$. In this case Algorithm 5 is identical to Algorithm APGM in [34] and to Algorithm APDA in [34] when $x_0 = \operatorname{argmin}_{x \in \mathbb{R}^n} h(x)$. Furthermore, Theorem 6 yields the convergence rate established in [34] as the next proposition shows.

Proposition 1. *Suppose f is L_f-smooth relative to h on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$ and $\gamma \geq 1$ is a triangle scaling exponent of D_h. Suppose also that ϕ is bounded below and the pair (h, Ψ) satisfies the sufficient admissibility condition. If Algorithm 5 chooses $L_0 = L_f$ and θ_k via $\theta_0 = 1$ and $\theta_k^\gamma = (1-\theta_k)\theta_{k-1}^\gamma, k = 1, 2, \ldots$ then each L_k is admissible for $\nabla f(y_k)$ at z_k and (2.10) holds for $k = 0, 1, \ldots$. Furthermore, in this case*

$$\phi(x_k) - \phi(x) \leq \left(\frac{\gamma}{k-1+\gamma}\right)^\gamma L_f D_h(x, x_0) \tag{2.14}$$

for all $x \in \mathrm{dom}(\phi)$.

Proof. For this choice L_0 and θ_k we have $\theta_k L_k = \theta_k^\gamma L_f,\ k = 0, 1, \ldots$. To show that L_k is admissible, we establish a lower bound on the function $z \mapsto \langle \nabla f(y_k), z\rangle + L_k D_h(z, z_k)$ and use the sufficient admissibility condition of (h, Ψ). To that end, observe that the convexity of f and Ψ, the triangle scaling property of D_h, and the L_f-smoothness of f imply that for all $k = 0, 1, \ldots,\ z \in \mathsf{relint}(\mathrm{dom}(h))$, and $x := (1-\theta_k)x_k + \theta_k z$

$$\begin{aligned}
&(1-\theta_k)\phi(x_k) + \theta_k(f(y_k) + \langle \nabla f(y_k), z - y_k\rangle) + L_k D_h(z, z_k) + \Psi(z))\\
&\geq f(y_k) + \langle \nabla f(y_k), x - y_k\rangle + \theta_k^\gamma L_f D_h(z, z_k) + \Psi(x)\\
&\geq f(y_k) + \langle \nabla f(y_k), x - y_k\rangle + L_f D_h(x, y_k) + \Psi(x)\\
&\geq \phi(x).
\end{aligned}$$

The boundedness of ϕ and the sufficient admissibility condition of (h, Ψ) thus imply that L_k is admissible for $\nabla f(y_k)$ at z_k for $k = 0, 1, 2, \ldots$. Furthermore, (2.10) also follows by taking $z := z_{k+1}$. To finish, observe that as shown in [34, Lemma 4], the sequence $\theta_k,\ k = 0, 1, \ldots$ defined via $\theta_0 = 1$ and $\theta_k^\gamma = (1-\theta_k)\theta_{k-1}^\gamma, k = 1, 2, \ldots$. satisfies

$$\theta_k \leq \frac{\gamma}{k+\gamma}.$$

Since $\theta_{k-1}L_{k-1} = \theta_{k-1}^\gamma L_f$, Corollary 4 yields (2.14). □

Our next result shows that the same, or a possibly faster, rate of convergence as (2.14) is attained by Algorithm 5 if $L_0 = L_f$ and $\theta_k \in (0, 2/3]$ is chosen as large as possible. This choice of θ_k is motivated by the following considerations. Observe that the iterates of Algorithm 5 satisfy $\theta_k L_k = (1-\theta_k)\theta_{k-1}L_{k-1},\ k = 1, 2, \ldots$. Therefore, the *larger* the $\theta_k,\ k = 1, 2, \ldots$, the tighter the bound in Corollary 4. In Theorem 7 we make the ideal assumptions that $L_0 = L_f$ and that $\theta_k \in (0, 2/3]$ is chosen as large as possible simply for ease of exposition. As we detail below, the slightly weaker rate (2.15) holds for more realistic and easily implementable line-search procedures that choose L_0 and θ_k.

Theorem 7. *Suppose f is L_f-smooth relative to h on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$ and $\gamma > 0$ is a triangle scaling exponent of D_h. Suppose also that ϕ is bounded below and the pair (h, Ψ) satisfies the sufficient admissibility condition. If Algorithm 5 chooses*

$L_0 = L_f$ in Step 2 and θ_k, $k = 1, 2, \ldots$ in Step 5 as the largest $\theta_k \in (0, 2/3]$ such that L_k is admissible for $\nabla f(y_k)$ at z_k and (2.10) holds then for $k = 1, 2, \ldots$

$$\phi(x_k) - \phi(x) \leq \left(\frac{\gamma}{k - 1 + \gamma}\right)^{\gamma} L_f D_h(x, x_0)$$

for all $x \in \mathrm{dom}(\phi)$.

Proof. By Corollary 4, it suffices to show that

$$\theta_{k-1} L_{k-1} \leq \left(\frac{\gamma}{k - 1 + \gamma}\right)^{\gamma} L_f.$$

We proceed by induction. The case $k = 1$ is an immediate consequence of the L_f-smoothness of f relative to h, the boundedness of ϕ, and the sufficient admissibility condition of (h, Ψ). For the main inductive step, we follow a similar reasoning to that in the proof of Proposition 1. Observe that the convexity of f and Ψ, the construction of L_k, and the triangle scaling property of D_h imply that for all $z \in \mathsf{relint}(\mathrm{dom}(h))$ and $x := (1 - \theta_k)x_k + \theta_k z$

$$\begin{aligned}
&(1 - \theta_k)\phi(x_k) + \theta_k(f(y_k) + \langle \nabla f(y_k), z - y_k \rangle) + L_k D_h(z, z_k) + \Psi(z)) \\
&\geq f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \frac{(1 - \theta_k)\theta_{k-1} L_{k-1}}{\theta_k^{\gamma}} D_h(x, y_k) + \Psi(x).
\end{aligned}$$

Hence the L_f-smoothness of f, boundedness of ϕ, and sufficient admissibility condition of (f, Ψ) imply that L_k is admissible for $\nabla f(y_k)$ at z_k and (2.10) holds provided $\theta_k \in (0, 2/3]$ is such that

$$\frac{(1 - \theta_k)\theta_{k-1} L_{k-1}}{\theta_k^{\gamma}} \leq L_f.$$

The induction hypothesis implies that

$$\frac{1 - \theta_k}{\theta_k^{\gamma}} \leq \frac{L_f}{\theta_{k-1} L_{k-1}} = \left(\frac{\hat{k} - 1 + \gamma}{\gamma}\right)^{\gamma}$$

for some $\hat{k} \geq k$ not necessarily integral. Thus $\theta_k \geq \hat{\theta}$ where $\hat{\theta} \in (0, 2/3]$ is the root of

$$\frac{1 - \hat{\theta}}{\hat{\theta}^{\gamma}} = \frac{L_f}{\theta_{k-1} L_{k-1}} = \left(\frac{\hat{k} - 1 + \gamma}{\gamma}\right)^{\gamma}.$$

As shown in [34, Lemma 3], the arithmetic mean geometric mean inequality implies that $\hat{\theta} \leq \gamma/(\hat{k} + \gamma) \leq 2/3$. Therefore,

$$\theta_k L_k = (1 - \theta_k)\theta_{k-1} L_{k-1} \leq (1 - \hat{\theta})\theta_{k-1} L_{k-1} = \hat{\theta}^{\gamma} L_f \leq \left(\frac{\gamma}{\hat{k} + \gamma}\right)^{\gamma} L_f \leq \left(\frac{\gamma}{k + \gamma}\right)^{\gamma} L_f.$$

□

Consider the following more realistic line-search procedures. Suppose we choose L_0 via the following standard binary search procedure: Start with an initial guess $L_0 > 0$ for L_f and repeatedly scale L_0 (up or down) by $\alpha > 1$ until (2.10) just holds for $k = 0$. This kind of procedure will choose $L_0 \leq \alpha L_f$. Suppose $\theta_k \in (0, 2/3],\ k = 1, 2, \ldots$ is chosen via the following binary search procedure which is a variant of the approach used in [34, Algorithm ABPG-e]: Set $\theta_k := \gamma_k/(k + \gamma_k)$ for some initial guess $\gamma_k > 0$ and repeatedly increase or decrease γ_k by some sufficiently small $\delta > 0$ until (2.10) just holds. These two procedures and a straightforward modification of the proof of Theorem 7 imply that for some $\tilde{\gamma} \geq \gamma - \delta$ the iterates generated by Algorithm 5 satisfy

$$\phi(x_k) - \phi(x) \leq \left(\frac{\tilde{\gamma}}{k - 1 + \tilde{\gamma}}\right)^{\tilde{\gamma}} \alpha L_f D_h(x, x_0) \tag{2.15}$$

for all $x \in \mathrm{dom}(\phi)$.

We note that although (2.15) is theoretically weaker than (2.14), the above line-search procedures could choose $L_0 < L_f$ and $\theta_k > \gamma/(k + \gamma)$ thereby yielding a faster rate of convergence. Our numerical experiments in Section 2.6 shed some light into this phenomenon.

2.5 Linear Convergence of Accelerated Bregman Proximal Gradient

We next show that some variants of Algorithm 5 that include restart attain an accelerated linear rate of convergence provided that some suitable relative smoothness and functional growth conditions hold. The algorithmic schemes and proofs in this section follow in a fairly straightforward fashion from the same ideas used in known restart schemes such as those in [50, 54, 55, 67]. We should note than unlike the previous algorithms in the chapter, Algorithm 6 and Algorithm 7 below require some additional knowledge about the problem.

Throughout this section assume that $\bar{\phi} := \min_x \phi(x) < \infty$ and $\bar{X} := \{x \in \mathrm{dom}(\phi) : \phi(x) = \bar{\phi}\} \neq \emptyset$. Let for $x \in \mathrm{dom}(\phi)$ let $D_h(\bar{X}, x) := \inf_{\bar{x} \in \bar{X}} D_h(\bar{x}, x)$. Suppose f is both L_f-smooth relative to h and μ_f-strongly convex relative to h on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$. That is, both $L_f h - f$ and $f - \mu_f h$ are convex on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$. As discussed in [69] and [48], under these conditions the iterates generated by Algorithm 4 satisfy

$$D_h(\bar{X}, x_k) \leq \left(1 - \frac{\mu_f}{L_f}\right)^k D_h(\bar{X}, x_0)$$

and

$$\phi(x_k) - \bar{\phi} \leq L_f \left(1 - \frac{\mu_f}{L_f}\right)^k D_h(\bar{X}, x_0)$$

provided $L_k = L_f,\ k = 0, 1, \ldots$. A straightforward modification of the argument in [69] shows that these inequalities also hold with L_f replaced with $\max\{\bar{L}, \alpha L_f\}$ if L_k

is instead chosen via a backtracking procedure that starts with an initial guess $\bar{L}$ for L_f and repeatedly scales it up by $\alpha > 1$ until condition (2.6) holds.

The above bounds imply that Algorithm 4 yields $x_k \in \mathrm{dom}(\phi)$ with $\phi(x_k) - \bar{\phi} < \epsilon$ in at most

$$k = \mathcal{O}\left(\frac{L_f}{\mu_f} \cdot \log\left(\frac{L_f D_h(\bar{X}, x_0)}{\epsilon}\right)\right)$$

iterations. We next show that under the same relative smoothness assumption and a relative *functional growth* assumption, two variants of Algorithm 5 that include restart achieve a faster linear rate when $\gamma > 1$. Note that Algorithm 6 requires knowledge of the optimal value $\bar{\phi}$. On the other hand, Algorithm 7 requires knowledge of a certain *condition number* L_f/κ_ϕ of ϕ and of the triangle scaling exponent γ of D_h.

Following [50], define the *functional growth constant* κ_ϕ of ϕ relative to h as follows

$$\kappa_\phi := \inf_{x \in \mathbb{R}^n \setminus \bar{X}} \frac{\phi(x) - \bar{\phi}}{D_h(\bar{X}, x)}.$$

Algorithm 6 Accelerated Bregman proximal gradient with restart (version 1)

Pick $w_0 \in \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$
for $\ell = 0, 1, \ldots$ **do**
  let $x_0 := w_\ell$ and run Algorithm 5 until

$$\phi(x_k) - \bar{\phi} \leq \frac{\phi(x_0) - \bar{\phi}}{2}$$

  let $w_{\ell+1} := x_k$
end for

Algorithm 7 Accelerated Bregman proximal gradient with restart (version 2)

Pick $w_0 \in \mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$
for $\ell = 0, 1, \ldots$ **do**
  let $x_0 := w_\ell$ and run Algorithm 5 until

$$k = \gamma \left(\frac{2L_f}{\kappa_\phi}\right)^{1/\gamma}$$

  let $w_{\ell+1} := x_k$
end for

Proposition 2. *Suppose f is L_f-smooth relative to h on $\mathsf{relint}(\mathrm{dom}(h)) \cap \mathrm{dom}(\Psi)$, ϕ has positive functional growth constant κ_ϕ relative to h, and D_h has triangle scaling*

exponent $\gamma \geq 1$. Then each call to Algorithm 5 in Algorithm 6 halts after at most

$$k = \gamma \left(\frac{2L_f}{\kappa_\phi} \right)^{1/\gamma} \tag{2.16}$$

iterations. On the other hand, the sequence of outer iterates $\{w_\ell : \ell = 0, 1, \dots\}$ generated by Algorithm 7 satisfies

$$\phi(w_{\ell+1}) - \bar{\phi} \leq \frac{\phi(w_\ell) - \bar{\phi}}{2}. \tag{2.17}$$

and

$$D_h(\bar{X}, w_{\ell+1}) \leq \frac{D_h(\bar{X}, w_\ell)}{2}.$$

In particular, either Algorithm 6 or Algorithm 7 yields $x_K \in \mathrm{dom}(\phi)$ such that $\phi(x_K) - \bar{\phi} < \epsilon$ after at most

$$K = \mathcal{O}\left(\left(\frac{L_f}{\kappa_\phi} \right)^{1/\gamma} \log \left(\frac{L_f D_h(\bar{X}, x_0)}{\epsilon} \right) \right)$$

accelerated Bregman proximal gradient iterations.

Proof. Theorem 7 implies that the iterates generated by Algorithm 5 satisfy

$$\phi(x_k) - \bar{\phi} \leq \left(\frac{\gamma}{k - 1 + \gamma} \right)^\gamma L_f D_h(\bar{X}, x_0) \leq \frac{L_f}{\kappa_\phi} \left(\frac{\gamma}{k - 1 + \gamma} \right)^\gamma (\phi(x_0) - \bar{\phi}).$$

Thus both (2.16) and (2.17) follow. In addition, for $\ell = 0, 1, \dots$ the outer iterates generated by Algorithm 7 satisfy

$$\phi(w_{\ell+1}) - \bar{\phi} \leq \left(\frac{\gamma}{k - 1 + \gamma} \right)^\gamma L_f D_h(\bar{X}, w_\ell) \leq \frac{\kappa_\phi D_h(\bar{X}, w_\ell)}{2}$$

and so

$$D_h(\bar{X}, w_{\ell+1}) \leq \frac{\phi(w_{\ell+1}) - \bar{\phi}}{\kappa_\phi} \leq \frac{D_h(\bar{X}, w_\ell)}{2}.$$

□

2.6 Numerical Experiments

We implemented a Python version of Algorithm 4 with line-search to choose L_k. Following the convention in [34], we will refer to this implementation as Algorithm BPG-LS. We also implemented two Python versions of Algorithm 5. The first one sets $L_0 := L_f$ and θ_k via $\theta_0 = 1$ and $\theta_k^\gamma = (1 - \theta_k)\theta_{k-1}^\gamma$, $k = 1, 2, \dots$ assuming that L_f and γ are known. As indicated in Section 2.4, this version is identical to Algorithm ABPG in

[34]. We also implemented a second version of Algorithm 5 with the line-search procedures to choose L_0 and θ_k sketched at the end of Section 2.4 for $\alpha = 2$ and $\delta = 0.1$. In particular, our implementation sets $\theta_k = \frac{\gamma_k}{k+\gamma_k},\ k = 1, 2, \ldots$ were $\gamma_k > 0$ is chosen via line-search so that (2.10) holds. We refer to this version as Algorithm ABPG-LS.

We next report results on some numerical experiments on random instances of two problems that provide interesting tests for Bregman proximal methods. The first one is the D-optimal design problem [2, 48]

$$\min_{x \in \Delta_{n-1}} \ -\log(\det(HXH^{\mathsf{T}}))$$

where $X = \mathsf{diag}(x)$ and $H \in \mathbb{R}^{m \times n}$ with $m < n$ and $\Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$.

The second one is the Poisson linear inverse problem [4]

$$\min_{x \in \mathbb{R}^n_+} \ D_{KL}(b, Ax)$$

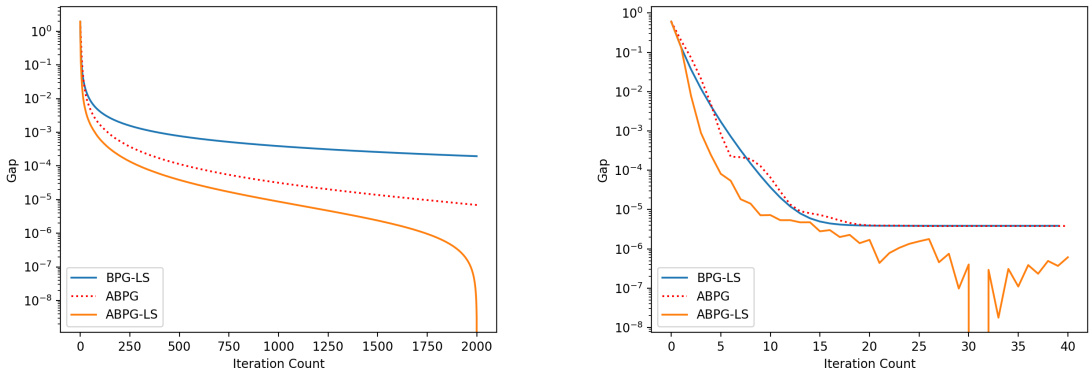
where $b \in \mathbb{R}^n_{++}$ and $A \in \mathbb{R}^{m \times n}_+$ with $m > n$ and $D_{KL}(\cdot, \cdot)$ is the Kullback-Leibler divergence, that is, the Bregman distance associated to the Boltzmann-Shannon entropy function $x \mapsto \sum_{i=1}^n x_i \log(x_i)$.

It was shown in [48] that the function $f(x) = -\log(\det(HXH^{\mathsf{T}}))$ is 1-smooth relative to the Burg entropy $h(x) = -\sum_{i=1}^n \log(x_i)$. On the other hand, it was shown in [4] that the function $x \mapsto D_{KL}(b, Ax)$ is $\|b\|_1$-smooth relative to $h(x) = -\sum_{i=1}^n \log(x_i)$. Thus we use the Burg entropy $h(x) = -\sum_{i=1}^n \log(x_i)$ as reference function for both problems. The implementation of Algorithm ABPG requires values of L_f and γ as input. We used the values $L_f = 1$ for the D-optimal design problem and $L_f = \|b\|_1$ for the Poisson linear inverse problem which are "safe" as per the above relative smoothness results. For γ, we used the default value $\gamma = 2$. This value is attractive because it yields the accelerated rate $\mathcal{O}(1/k^2)$ but is not safe because as discussed in [34], the Bregman distance for the Burg entropy has a smaller uniform triangle scaling exponent. Nonetheless, as in the experiments reported in [34], the choice of $\gamma = 2$ worked fine in our numerical experiments.

Figure 2.1 depicts the convergence of Algorithms BPG-LS, ABPG, and ABPG-LS on two typical random instances $H \in \mathbb{R}^{100 \times 250}$ and $H \in \mathbb{R}^{200 \times 300}$ for the D-optimal design problem. The suboptimality gap is measured relative to the smallest objective value attained by the three algorithms, which was ABPG-LS in all cases. The entries of the instances H for this problem are independent draws from the standard normal distribution.

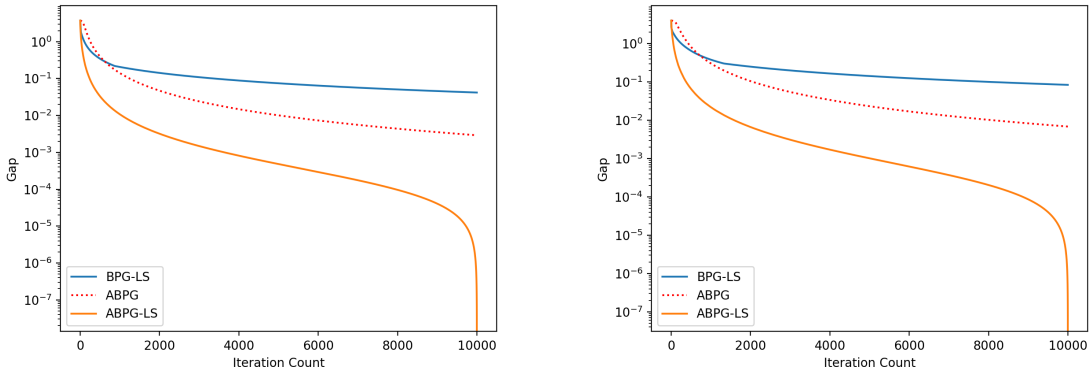
Figure 2.2 depicts similar convergence results on typical random instances $A \in \mathbb{R}^{250 \times 100},\ b \in \mathbb{R}^{250}$ and $A \in \mathbb{R}^{300 \times 200},\ b \in \mathbb{R}^{300}$ for the Poisson linear inverse problem. In this case the entries of A and of b are independent draws from the uniform distribution on $[0, 1]$.

The numerical experiments demonstrate that the convergence rates of the algorithms BPG-LS, ABPG, and ABPG-LS usually follow the pattern one would expect: In most cases Algorithm BPG-LS is the slowest while ABPG-LS is the fastest. An



(a) $m = 100, n = 250$ (b) $m = 200, n = 300$

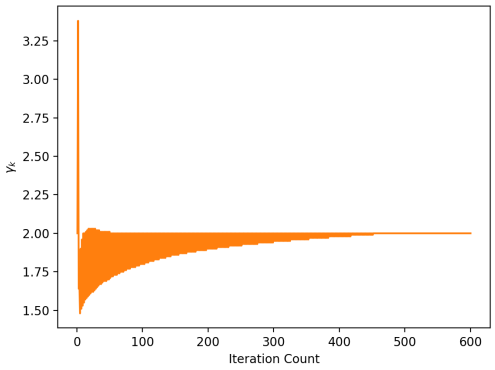
Figure 2.1: Suboptimality gap on typical instances of the D-optimal design problem.



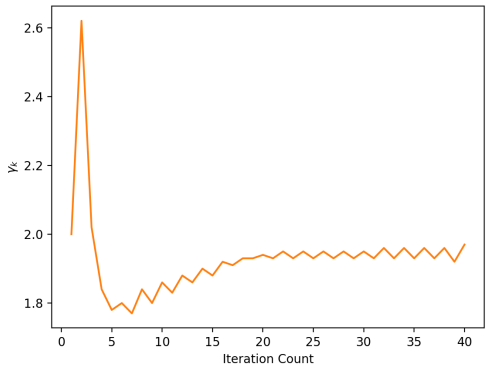
(a) $m = 250, n = 100$ (b) $m = 300, n = 200$

Figure 2.2: Suboptimality gap on typical instances of the Poisson linear inverse problem.

exception occurs in the easier 200×300 D-optimal design instances where BPG-LS performs performs as well as ABPG or better. As noted in [34] this can be attributed to the better conditioning of these instances and the linear convergence property of Algorithm 4. Figure 2.3 and Figure 2.4 depict an interesting phenomenon that we observed in our experiments. These figures display plots of the values of γ_k throughout the execution of Algorithm ABPG-LS in the four instances discussed above. In all of these cases it is evident that γ_k hovers near 2. Since the algorithm sets $\theta_k = \gamma_k/(k+\gamma_k)$, these values of γ_k imply that Algorithm ABPG-LS approximately attains the iconic $\mathcal{O}(1/k^2)$ convergence rate of accelerated gradient methods. This numerical evidence is striking and consistent with the results reported in [34].

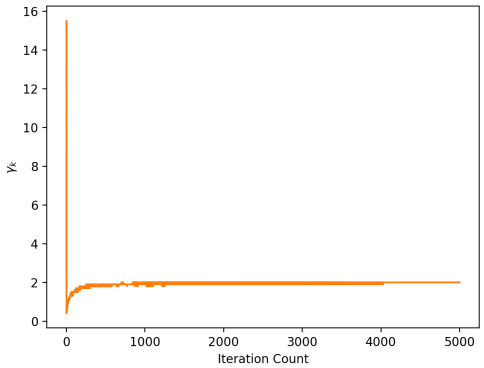


(a) $m = 100, n = 250$

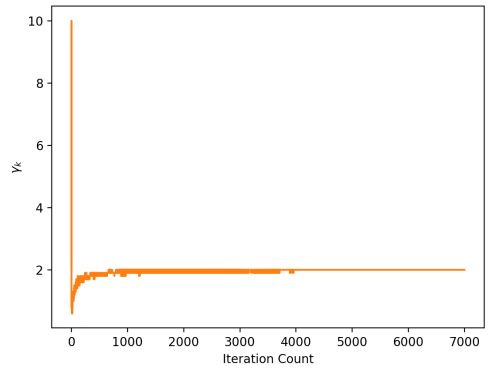


(b) $m = 200, n = 300$

Figure 2.3: Sequence $\{\gamma_k : k = 1, 2, \dots\}$ in ABPG-LS for typical instances of D-design optimal problem.



(a) $m = 250, n = 100$



(b) $m = 300, n = 200$

Figure 2.4: Sequence $\{\gamma_k : k = 1, 2, \dots\}$ in ABPG-LS for typical instances of Poisson linear inverse problem.

Chapter 3

Conditioning: The Condition Number of a Function Relative to a Set

3.1 Introduction

In this chapter, we propose a relative smoothness constant $L_{f,X,D}$ and a relative strong convexity constant $\mu_{f,X,D}$ of the function f relative to the pair (X, D) where $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is a convex differentiable function, $X \subseteq \mathrm{dom}(f)$ is a convex set, and $D : X \times X \to \mathbb{R}_+$ a distance-like function, that is, $D(y, x) \geq 0$ and $D(x, x) = 0$ for all $x, y \in X$. See Definition 4 and equation (3.5) below for details.

The main sections of the chapter are organized as follows. Section 3.2 presents our central construction, namely relative smoothness and relative strong convexity. This section also introduces relative quasi strong convexity and D-functional growth, both of which are variants of relative strong convexity. Section 3.3 and Section 3.4 present the main technical results of the chapter when D is a squared norm. Section 3.3 develops several properties of the constants $L_{f,X,D}$ and $\mu_{f,X,D}$. More precisely, Proposition 4 gives an upper bound on $L_{f,X,D}$ when f is of the form $g \circ A$ for some $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$. The more involved Theorem 8 and Theorem 9 give lower bounds on $\mu_{f,X,D}$ when f is of the form $g \circ A$ and X is a convex cone or a polyhedron. These bounds readily imply that for $f = g \circ A$ the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ can be bounded in terms of the product of the classical condition number L_g/μ_g and a condition number of the pair (A, X). See equation (3.14) and equation (3.16). Section 3.4 develops properties analogous to those in Section 3.3 but for the constants $\mu^\star_{f,X,D}$ and $\mu^\sharp_{f,X,D}$. Finally Section 3.5 presents linear convergence results for the mirror descent algorithm and for the Frank-Wolfe algorithm with away steps in terms of the relative constants $L_{f,X,D}$ and $\mu^\star_{f,X,D}, \mu^\sharp_{f,X,D}$.

3.2 Conditioning Relative to a Reference Set and Distance Function Pair

This section presents the central ideas of this chapter. We introduce the concepts of relative smoothness and relative strong convexity of a function relative to a reference set and distance function pair. We also introduce some variants of relative strong convexity that are natural extensions of the approach developed by Necoara, Nesterov and Glineur [50].

Throughout the entire chapter we will typically make the following blanket assumption about the triple (f, X, D).

Assumption 2. The function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable. The set $X \subseteq \mathrm{dom}(f)$ is convex. The function $D : X \times X \to \mathbb{R}_+$ is a reference *distance-like* function, that is, $D(y, x) \geq 0$ for all $x, y \in X$ and $D(x, x) = 0$ for all $x \in X$.

Throughout our development we will consider mainly the following two classes of distance-like functions:

- The *Bregman distance* associated to a reference convex differentiable function $h : X \to \mathbb{R}$, that is,

$$D(y, x) := D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle .$$

- The square of a (non-necessarily Euclidean) norm $\|\cdot\|$ in $\mathbb{R}^n$, that is,

$$D(y, x) := \frac{1}{2}\|y - x\|^2.$$

Our main construction is based on bounding the behavior of the Bregman distance associated to f in terms of the reference distance function D. The following object provides a key building block for our construction. For $y \in X$ let $Z_{f,X}(y) \subseteq X$ denote the set

$$Z_{f,X}(y) := \{x \in X : f(x) = f(y) \ \text{ and } \ \langle \nabla f(x) - \nabla f(y), x - y \rangle = 0\}.$$

Observe that if f is strictly convex then $Z_{f,X}(y) = \{y\}$ for all $y \in X$.

3.2.1 Relative Smoothness and Relative Strong Convexity

To motivate our main construction we first recall the classical notion of smoothness and strong convexity constants. We recall these classical concepts in a format that we subsequently use for our main construction. Recall that for a convex differentiable function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $x, y \in \mathrm{dom}(f)$ the Bregman distance $D_f(y, x)$ is

$$D_f(y, x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle .$$

Definition 3. Suppose (f, X, D) satisfy Assumption 2 and $D(y,x) = \frac{1}{2}\|y-x\|^2$ for some norm $\|\cdot\|$ in $\mathbb{R}^n$.

(a) The function f is smooth on X if there exists a constant $L > 0$ such that

$$D_f(y,x) \leq LD(y,x) \text{ for all } x, y \in X. \tag{3.1}$$

(b) The function f is strongly convex on X if there exists a constant $\mu > 0$ such that

$$D_f(y,x) \geq \mu D(y,x) \text{ for all } x, y \in X. \tag{3.2}$$

Next, we present our main construction. In Definition 4 and throughout the chapter we will use the following notational convention. For a nonempty $S \subseteq X$ and $x \in X$ let $D_f(S,x)$ and $D(S,x)$ denote $\inf_{y\in S} D_f(y,x)$ and $\inf_{y\in S} D(y,x)$ respectively.

Definition 4. Let (f, X, D) satisfy Assumption 2.

(a) We say that f is *smooth relative* to (X, D) if there exists a constant $L > 0$ such that

$$D_f(y,x) \leq LD(y,x) \text{ for all } x, y \in X. \tag{3.3}$$

(b) We say that f is *strongly convex relative* to (X, D) if there exists a constant $\mu > 0$ such that

$$D_f(Z_{f,X}(y), x) \geq \mu D(Z_{f,X}(y), x) \text{ for all } x, y \in X. \tag{3.4}$$

When $D = D_h$ for some convex differentiable function $h : X \to \mathbb{R}$, the above relative smoothness concept is identical to the smoothness of f relative to h on X as defined in [48]. The latter in turn is equivalent to the *Lipschitz-like condition* defined in [4]. Furthermore, when $D = D_h$ and f is strictly convex, the above relative strong convexity concept is identical to the strong convexity of f relative to h on X as defined in [48].

We will use the following notation throughout the rest of the chapter. Suppose (f, X, D) satisfies Assumption 2. Let $L_{f,X,D}$ and $\mu_{f,X,D}$ be the following relative smoothness and strong convexity constants

$$L_{f,X,D} := \inf\{L > 0 : (3.3) \text{ holds}\},\ \mu_{f,X,D} := \sup\{\mu > 0 : (3.4) \text{ holds}\}. \tag{3.5}$$

In addition, suppose (f, X, D) satisfies Assumption 2 and $D(x,y) = \frac{1}{2}\|x-y\|^2$ for some norm $\|\cdot\|$ in $\mathbb{R}^n$. Let L_f and μ_f be the following classical smoothness and strong convexity constants

$$L_f := \inf\{L > 0 : (3.1) \text{ holds}\},\ \mu_f := \sup\{\mu > 0 : (3.2) \text{ holds}\}. \tag{3.6}$$

The following example illustrates the values of the relative smoothness and convexity constants $L_{f,X,D}$ and $\mu_{f,X,D}$ of a convex quadratic function relative to (X, D) for some canonical choices of X and D. It also lays the ground for the main properties that we develop in Section 3.3.

Example 2. Let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$ with $A \neq 0$ and $\mathbb{R}^n$ and $\mathbb{R}^n$ be endowed with the Euclidean norm. Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $D(y, x) := \frac{1}{2}\|y - x\|_2^2$. Then f has the following smoothness and strong convexity constants $L_{f,X,D}$ and $\mu_{f,X,D}$ relative to (X, D) for two particular choices of X.

(a) For $X = \mathbb{R}^n$ we have $L_{f,X,D} = \sigma_{\max}(A^\mathsf{T}A) = \sigma_{\max}(A)^2$ and $\mu_{f,X,D} = \sigma^+_{\min}(A^\mathsf{T}A) = \sigma^+_{\min}(A)^2 > 0$, where $\sigma^+_{\min}(\cdot)$ denotes the smallest *positive* singular value. Observe that in this case $L_f = L_{f,X,D}$ but $\mu_f = \mu_{f,X,D}$ only when A is full column rank.

(b) Suppose $X \subseteq \mathbb{R}^n$ is a linear subspace such that the mapping $A|X : X \to \mathbb{R}^m$ defined via $x \in X \mapsto Ax \in \mathbb{R}^m$ is nonzero. Then $L_{f,X,D} = \sigma_{\max}(A|X)^2$ and $\mu_{f,X,D} = \sigma^+_{\min}(A|X)^2$. Observe that in this case $L_{f,X,D} \leq L_f$ and $L_{f,X,D}$ can be quite a bit smaller. Likewise, $\mu_{f,X,D} \geq \mu_f$ and $\mu_{f,X,D}$ could be quite a bit larger.

The statements (a) and (b) in Example 2 can be verified directly but they also follow from the more general Proposition 4 and Corollary 5 in Section 3.3 below.

3.2.2 Relative Quasi-strong Convexity and D-functional Growth

Following [50], we next consider two variants of relative strong convexity that are natural extensions of the *quasi-strong convexity* and *quadratic functional growth* concepts defined in [50]. For that purpose, we will rely on the following strengthening of Assumption 2.

Assumption 3. Suppose (f, X, D) satisfy Assumption 2, $\bar{f} := \min_{x\in X} f(x)$ is finite, $\bar{X} := \{x \in X : f(x) = \bar{f}\} \neq \emptyset$, and the map $x \mapsto \bar{x} := \operatorname{argmin}_{y\in\bar{X}} D(y, x)$ is well defined for all $x \in X$.

Definition 5. Suppose (f, X, D) satisfies Assumption 3.

(a) We say that f is *quasi-strongly-convex relative* to (X, D) if there exists a constant $\mu > 0$ such that

$$D_f(\bar{x}, x) \geq \mu D(\bar{x}, x) \;\text{ for all }\; x \in X. \tag{3.7}$$

(b) We say that f has *D-functional growth* on X if there exists a constant $\mu > 0$ such that

$$f(x) - \bar{f} \geq \mu D(\bar{x}, x) \;\text{ for all }\; x \in X. \tag{3.8}$$

Throughout the sequel we will use the following notation analogous to (3.5). Suppose (f, X, D) satisfies Assumption 3. Let $\mu^\star_{f,X,D}$ and $\mu^\sharp_{f,X,D}$ be as follows

$$\mu^\star_{f,X,D} := \sup\{\mu > 0 : (3.7) \text{ holds}\},\ \mu^\sharp_{f,X,D} := \sup\{\mu > 0 : (3.8) \text{ holds}\}. \tag{3.9}$$

The next proposition shows that, as one may intuitively expect, relative quasi-strong convexity is a relaxation of relative strong convexity. In other words, $\mu_{f,X,D} \leq \mu^\star_{f,X,D}$ whenever (f, X, D) satisfies Assumption 3.

Proposition 3. *Suppose* (f, X, D) *satisfy Assumption 3. If* $\mu > 0$ *is such that* (f, X, D, μ) *satisfies* (3.4) *then* (f, X, D, μ) *satisfies* (3.7).

Proof. The construction of $Z_{f,X}(y)$ implies that $Z_{f,X}(y) \subseteq \bar{X}$ for all $y \in \bar{X}$. Therefore, if (f, X, D, μ) satisfies (3.4) then by taking $y = \bar{x}$ it follows that

$$D_f(\bar{x}, x) \geq D_f(Z_{f,X}(\bar{x}), x) \geq \mu D(Z_{f,X}(\bar{x}), x) = \mu D(\bar{x}, x) \text{ for all } x \in X.$$

□

The following simple example shows that, perhaps contrary to what one might intuitively expect, D-functional growth is not necessarily a relaxation of strong relative convexity unless some additional assumptions are made about f, X, or D.

Example 3. Let $a > 0$ and $f : \mathbb{R} \to \mathbb{R}$ be the function $f(x) = e^{ax}$. For $X := \mathbb{R}_+$ we have $\bar{X} = \{0\}$. Thus for $D := D_f$ and $\mu = 1$ the tuple (f, X, D, μ) satisfies (3.4). However, observe that for all $\hat{\mu} > 0$ and $x \geq 1/(\hat{\mu}a)$

$$f(x) - \bar{f} = e^{ax} - 1 < \hat{\mu}(1 + axe^{ax}) = \hat{\mu}(\bar{f} - f'(x)(0 - x)) = \hat{\mu} D(\bar{X}, x).$$

In particular, $(f, X, D, \hat{\mu})$ does not satisfy (3.8) for any $\hat{\mu} > 0$.

It can be shown that under additional assumptions on f, X, or D the D-functional growth condition is a relaxation of the relative strong convexity condition. For instance, it is easy to see that this is the case if D_f is symmetric, that is, $D_f(y, x) = D_f(x, y)$ for $x, y \in X$. In addition, D-functional growth is a relaxation of relative strong convexity when D is a squared norm and X as we discuss in Section 3.4 below.

3.3 Properties of $L_{f,X,D}$ and $\mu_{f,X,D}$ when D is a Squared Norm and f is of the Form $g \circ A$

This section develops some properties of the relative constants $L_{f,X,D}$ and $\mu_{f,X,D}$ when f is of the form $f := g \circ A$ for $A \in \mathbb{R}^{m \times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$, and D is of the form $D(x, y) = \frac{1}{2}\|x - y\|^2$ for some norm in $\mathbb{R}^n$. The main results of this section are Theorem 8 and Theorem 9. These results provide lower bounds on $\mu_{f,X,D}$ in terms of μ_g and the norms of some canonical set-valued maps that depend on A and X. In a similar vein, Proposition 4 gives an upper bound on $L_{f,X,D}$ in terms of L_g and the norm of a canonical map associated to A and X.

We will rely on the objects $Z_{A,X}(\cdot)$ and $A|C, (A|C)^{-1}$ defined next. For $A \in \mathbb{R}^{m \times n}$, $X \subseteq \mathbb{R}^n$ nonempty and $y \in X$ let $Z_{A,X}(y) := \{x \in X : Ax = Ay\}$. Observe that the set-valued mapping $Z_{A,X} : X \rightrightarrows X$ can be seen as an extension of the set-valued mapping $Z_{f,X} : X \rightrightarrows X$ introduced in Section 3.2.1.

For $A \in \mathbb{R}^{m\times n}$ and a convex cone $C \subseteq \mathbb{R}^n$ let $A|C : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be the set-valued mapping defined via

$$x \mapsto (A|C)(x) := \begin{cases} \{Ax\} & \text{if } x \in C \\ \emptyset & \text{otherwise.} \end{cases}$$

And let $(A|C)^{-1} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ be its inverse, that is,

$$v \mapsto (A|C)^{-1}(v) := \{x \in C : Ax = v\}.$$

Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Define the norms of $A|C$ and of $(A|C)^{-1}$ as follows

$$\|A|C\| := \max_{\substack{x\in C\\ \|x\|\le 1}} \|Ax\|, \quad \|(A|C)^{-1}\| := \max_{\substack{v\in A(C)\\ \|v\|\le 1}} \min_{\substack{x\in C\\ Ax=v}} \|x\|.$$

Observe that if $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ is a nonempty convex set such that $A(X)$ contains more than one point then

$$\|A|\operatorname{span}(X - X)\| = \sup_{\substack{y,x\in X\\ x\neq y}} \frac{\|Ay - Ax\|}{\|y - x\|}. \tag{3.10}$$

In particular, the following property of the relative smoothness constant readily follows.

Proposition 4. *Suppose $\mathbb{R}^m, \mathbb{R}^n$ are endowed with norms and $D(y, x) := \frac{1}{2}\|y - x\|^2$. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a nonempty convex set such that $A(X)$ contains more than one point.*

(a) *If $\mathbb{R}^m$ is endowed with the Euclidean norm and $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $b \in \mathbb{R}^m$ then*

$$L_{f,X,D} = \|A|\operatorname{span}(X - X)\|^2.$$

(b) *If $f = g \circ A$ where g is L_g smooth $A(X)$ for the norm in $\mathbb{R}^m$ then*

$$L_{f,X,D} \le L_g\|A|\operatorname{span}(X - X)\|^2.$$

Proof. **(a)** This follows from (3.10) and $D_f(y, x) = \frac{1}{2}\|Ay - Ax\|_2^2$.

(b) This follows from (3.10) and $D_f(y, x) = D_g(Ay, Ax) \le \frac{L_g}{2}\|Ay - Ax\|^2$. The latter inequality follows from the L_g smoothness of g.

□

We next discuss far more interesting results that either characterize or lower bound the relative strong convexity constant $\mu_{f,X,D}$.

3.3.1 Lower Bound on $\mu_{f,X,D}$ when X is a Convex Cone and $A(X)$ is a Linear Subspace

In this subsection we will consider the special case when $X \subseteq \mathbb{R}^n$ is a convex cone and $A \in \mathbb{R}^{m\times n}$ is such that $A(X)$ is a linear subspace of $\mathbb{R}^m$. The latter condition is equivalent to the following *Slater condition:* there exists $x \in \mathsf{relint}(X)$ such that $Ax = 0$. When this is the case, the norms $\|A|X\|$ and $\|(A|X)^{-1}\|$ have the following geometric interpretation. Let $\mathbb{B}^m$ and $\mathbb{B}^n$ denote the unit balls in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. It is easy to see that if X is a convex cone and $A(X)$ is a linear subspace then

$$\|A|X\| = \inf\{r : A(X \cap \mathbb{B}^n) \subseteq r\mathbb{B}^m \cap A(X)\} \tag{3.11}$$

and

$$\frac{1}{\|(A|X)^{-1}\|} = \sup\{r : r\mathbb{B}^m \cap A(X) \subseteq A(X \cap \mathbb{B}^n)\}. \tag{3.12}$$

In other words, $\|A|X\|$ is the radius of the *smallest* ball in $A(X)$ centered at the origin *containing* $A(X\cap\mathbb{B}^n)$ and $1/\|(A|X)^{-1}\|$ is the radius of the *largest* ball in $A(X)$ centered at the origin and *contained* in $A(X\cap\mathbb{B}^n)$. The above norms, especially $\|(A|X)^{-1}\|$ and other related quantities, have been extensively studied in the literature on condition measures for convex optimization [19, 24, 27, 58, 64, 63]. They have been further extended to the broader variational analysis context [44, 21]. In particular, when $A(X) = \mathbb{R}^m$ the family of conic systems $Ax = b, x \in X$ is *well-posed.* That is, for all $b \in \mathbb{R}^m$ the conic system $Ax = b, x \in X$ is feasible and remains so for sufficiently small perturbations of (A, b). In this case it follows from [64] that the quantity $1/\|(A|X)^{-1}\|$ is precisely the *distance to ill-posedness* introduced by Renegar [63, 64], that is, the size of the smallest perturbation ΔA on A so that the conic system $(A+\Delta A)x = b, x \in X$ is infeasible for some $b \in \mathbb{R}^m$. A similar identity holds for the *distance to non-surjectivity* of closed sublinear set-valued mappings [44]. The latter in turn extends to a far more general identity for the radius of metric regularity [21].

Observe that if $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ is a linear subspace then $A(X)$ is automatically a linear subspace. If in addition $\mathbb{R}^n$ and $\mathbb{R}^m$ are each endowed with Euclidean norms, then (3.11) and (3.12) yield

$$\|A|X\| = \sigma_{\max}(A|X) \;\text{ and }\; \frac{1}{\|(A|X)^{-1}\|} = \sigma^+_{\min}(A|X).$$

Corollary 5 and Theorem 8 below show that there is a tight connection between the relative strong convexity constant $\mu_{f,X,D}$ and the norm $\|(A|X)^{-1}\|$ when f is of the form $g \circ A$. Both of these results rely on the following proposition that characterizes a certain type of *Hoffman* constant [37]. Proposition 5 is closely related to developments in [57, 62].

Proposition 5. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a convex cone such that $A(X)$ contains more than one point. If $A(X)$ is a*

linear subspace then

$$\frac{1}{\|(A|X)^{-1}\|} = \inf_{\substack{x\in X\\ y\in X\setminus Z_{A,X}(x)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}. \tag{3.13}$$

Proof. Fix $y \in X$ and $x \in X \setminus Z_{A,X}(y)$. Since $A(X)$ is a linear subspace, it follows that $Ay - Ax \in A(X)$ and thus $Ay - Ax = Au$ for some $u \in X$ with $\|u\| \leq \|(A|X)^{-1}\| \cdot \|Ay-Ax\|$. Hence $x+u \in Z_{A,X}(y)$ and $\|Z_{A,X}(y)-x\| \leq \|u\| \leq \|(A|X)^{-1}\| \cdot \|Ay-Ax\|$. Since this holds for arbitrary $y \in X$ and $x \in X \setminus Z_{A,X}(y)$ we conclude that

$$\frac{1}{\|(A|X)^{-1}\|} \leq \inf_{\substack{y\in X\\ x\in X\setminus Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}.$$

To prove the reverse inequality, let $v \in A(X)$ be such that $\|v\| = 1$ and $\|y\| \geq \|(A|X)^{-1}\|$ for all $y \in X$ with $Ay = v$. Pick $\hat{y} \in X$ with $A\hat{y} = v$. Then $\|z\| \geq \|(A|X)^{-1}\|$ for all $z \in Z_{A,X}(\hat{y})$. Thus for $\hat{x} := 0 \in X \setminus Z_{A,X}(\hat{y})$ and

$$\frac{1}{\|(A|X)^{-1}\|} \geq \frac{\|A\hat{y}-A\hat{x}\|}{\|Z_{A,X}(\hat{y})-\hat{x}\|} \geq \inf_{\substack{y\in X\\ x\in X\setminus Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|}.$$

□

Proposition 5 readily yields the following result that generalizes Example 2.

Corollary 5. *Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm $\|\cdot\|_2$, $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$, and $D(x,y) = \frac{1}{2}\|x-y\|^2$. If $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, $X \subseteq \mathbb{R}^n$ is a convex cone, and $A(X)$ is a linear subspace that contains more than one point then*

$$\mu_{f,X,D} = \frac{1}{\|(A|X)^{-1}\|^2}.$$

Proof. This follows from Proposition 5 and the observation that for this choice of f and X we have $Z_{f,X}(y) = Z_{A,X}(y)$ and $f(y) - f(x) - \langle \nabla f(x), y-x\rangle = \frac{1}{2}\|Ay-Ax\|_2^2$. □

The following result extends Corollary 5 to a broader class of functions.

Theorem 8. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(x,y) = \frac{1}{2}\|x-y\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function, and $X \subseteq \mathbb{R}^n$ be a convex cone such that $A(X)$ is a linear subspace that contains more than one point. If g is μ_g strongly convex on $A(X)$ for the norm $\|\cdot\|$ in $\mathbb{R}^m$ then the function $f = g \circ A$ satisfies*

$$\mu_{f,X,D} \geq \frac{\mu_g}{\|(A|X)^{-1}\|^2}.$$

Proof. Observe that $D_f(y,x) = g(Ay) - g(Ax) - \langle g(Ax), A(y-x)\rangle$ for all $y, x \in X$ Since g is μ_g strongly convex, it follows that $D_f(y,x) \geq \mu_g \|Ay - Ax\|^2/2$ for all $y, x \in X$ and $Z_{f,X}(y) = \{x \in X : Ax = Ay\} = Z_{A,X}(y)$ for all $y \in X$. Therefore Proposition 5 implies that

$$\mu_{f,X,D} = \inf_{\substack{y \in X \\ x \in X \setminus \in Z_{A,X}(y)}} \frac{D_f(y,x)}{\|Z_{A,X}(y) - x\|^2/2} \geq \inf_{\substack{y \in X \\ x \in X \setminus \in Z_{A,X}(y)}} \frac{\mu_g \|Ay - Ax\|^2}{\|Z_{A,X}(y) - x\|^2} = \frac{\mu_g}{\|(A|X)^{-1}\|^2}.$$

□

If f, X, D are as in Corollary 5 then by Proposition 4 the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ is

$$\frac{L_{f,X,D}}{\mu_{f,X,D}} = \left(\|A|\operatorname{span}(X)\| \cdot \|(A|X)^{-1}\|\right)^2$$

which has a striking resemblance to the classical condition number of $f(x) = \frac{1}{2}\|Ax - b\|_2^2$. More generally, if f, X, D are as in Theorem 8 and g is also L_g Lipschitz then by Proposition 4 we obtain the following bound on the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ in terms of the condition number of g and a condition number of the pair (A, X):

$$\frac{L_{f,X,D}}{\mu_{f,X,D}} \leq \frac{L_g}{\mu_g} \cdot \left(\|A|\operatorname{span}(X)\| \cdot \|(A|X)^{-1}\|\right)^2. \tag{3.14}$$

3.3.2 Lower Bound on $\mu_{f,X,D}$ when X is a Polyhedron

The results in Section 3.3.1 require X to be a convex cone and $A(X)$ to be a linear subspace. We next provide some results of similar flavor that relax these assumptions in exchange for the assumption that X is a polyhedron. The main ideas and results that we next develop are inspired by the recent work of Peña, Vera, and Zuluaga [57]. For a nonempty polyhedron $X \subseteq \mathbb{R}^n$ let $\mathcal{T}(X) := \{T_X(y) : y \in X\}$, where $T_X(y)$ is tangent cone of X at y, that is,

$$T_X(y) := \{d \in \mathbb{R}^n : y + td \in X \ \text{ for some } \ t > 0\}.$$

Observe that $\mathcal{T}(X)$ is finite since X is polyhedral. The paper [57] shows that Proposition 5 extends as below in Proposition 6. From this, Corollary 5, and Theorem 8 extend naturally.

Proposition 6. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. Then*

$$\min_{C \in \mathcal{S}(X)} \frac{1}{\|(A|C)^{-1}\|} = \min_{C \in \mathcal{T}(X)} \frac{1}{\|(A|C)^{-1}\|} = \inf_{\substack{y \in X \\ x \in X \setminus Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|}. \tag{3.15}$$

where $\mathcal{S}(X) := \{T \in \mathcal{T}(X) : A(T)$ is a subspace$\}$.

Corollary 6. *Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm $\|\cdot\|_2$, $\mathbb{R}^n$ is endowed with a norm $\|\cdot\|$, and $D(x,y)=\frac{1}{2}\|x-y\|^2$. If $f(x)=\frac{1}{2}\|Ax-b\|_2^2$ for some $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$, and $X\subseteq\mathbb{R}^n$ is a polyhedron such that $A(X)$ contains more than one point then*

$$\mu_{f,X,D}=\min_{C\in\mathcal{S}(X)}\frac{1}{\|(A|C)^{-1}\|^2}.$$

Proof. Proceed exactly as in the proof of Corollary 5 but apply Proposition 6 instead of Proposition 5. □

Theorem 9. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(x,y)=\frac{1}{2}\|x-y\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A\in\mathbb{R}^{m\times n}$, $g:\mathbb{R}^m\to\mathbb{R}\cup\{\infty\}$ be a convex differentiable function, and $X\subseteq\mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. If g is μ_g strongly convex on $A(X)$ for the norm in $\mathbb{R}^m$ then the function $f=g\circ A$ satisfies*

$$\mu_{f,X,D}\geq\min_{C\in\mathcal{S}(X)}\frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

Proof. Proceeding exactly as in the proof of Theorem 8 but applying Proposition 6 instead of Proposition 5 we get

$$\mu_{f,X,D}=\inf_{\substack{y\in X\\ x\notin Z_{A,X}(y)}}\frac{D_f(y,x)}{\|Z_{A,X}(y)-x\|^2/2}\geq\inf_{\substack{y\in X\\ x\notin Z_{A,X}(y)}}\frac{\mu_g\|Ay-Ax\|^2}{\|Z_{A,X}(y)-x\|^2}=\min_{C\in\mathcal{S}(X)}\frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

□

Observe that if X is polyhedral then $\operatorname{span}(X-X)\in\mathcal{S}(X)$ and

$$\|A|\operatorname{span}(X-X)\|=\max_{C\in\mathcal{S}(X)}\|A|C\|.$$

Thus Proposition 4 implies that for f,X,D as in Corollary 6, the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ has the following expression, which is again strikingly similar to the classical condition number of $f(x)=\frac{1}{2}\|Ax-b\|_2^2$:

$$\frac{L_{f,X,D}}{\mu_{f,X,D}}=\left(\max_{C\in\mathcal{S}(X)}\|A|C\|\cdot\max_{C\in\mathcal{S}(X)}\|(A|C)^{-1}\|\right)^2.$$

Proposition 4 also implies that if f,X,D are as in Theorem 9 and g is L_g smooth then the relative condition number $L_{f,X,D}/\mu_{f,X,D}$ can be bounded in terms of the condition number of g and a condition number of the pair (A,X) as follows:

$$\frac{L_{f,X,D}}{\mu_{f,X,D}}\leq\frac{L_g}{\mu_g}\cdot\left(\max_{C\in\mathcal{S}(X)}\|A|C\|\cdot\max_{C\in\mathcal{S}(X)}\|(A|C)^{-1}\|\right)^2. \tag{3.16}$$

We next place some of the developments by Peña and Rodríguez [60] in the context of this chapter. To that end, consider the special case when X is the standard simplex

$\Delta_{n-1} := \{x \in \mathbb{R}^n_+ : \|x\|_1 = 1\}$ in $\mathbb{R}^n$. For $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix} \in \mathbb{R}^{m\times n}$ let $\mathsf{conv}(A) := \mathsf{conv}(\{a_1, \ldots, a_n\}) = \{Ax : x \in \Delta_{n-1}\}$ and let $\mathsf{faces}(\mathsf{conv}(A))$ denote the set of faces of $\mathsf{conv}(A)$. Furthermore, for $F \in \mathsf{faces}(\mathsf{conv}(A))$ let $A \setminus F$ denote the set of columns of A that do not belong to F. Suppose $\mathbb{R}^m$ is endowed with a norm and for $F, G \subseteq \mathbb{R}^m$ let $\mathsf{dist}(F, G) := \inf_{u\in F, v\in G} \|u - v\|$. Following [60] define the *facial distance* $\Phi(A)$ of A as follows

$$\Phi(A) := \min_{\substack{F\in\mathsf{faces}(\mathsf{conv}(A))\\ \emptyset\neq F\neq\mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A \setminus F)).$$

Let $\mathsf{diam}(A)$ denote the *diameter* of the set of columns of A defined as follows

$$\mathsf{diam}(A) := \max_{i,j\in\{1,\ldots,n\}} \|a_i - a_j\|.$$

In the special case when $X = \Delta_{n-1}$ it follows from [60, Theorem 1] that (3.15) in Proposition 6 has the following geometric characterization

$$\min_{\substack{y\in\Delta_{n-1}\\ x\in\Delta_{n-1}\setminus Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|_1} = \frac{\Phi(A)}{2}. \tag{3.17}$$

Furthermore, in this same special case when $X = \Delta_{n-1}$ it is easy to see that (3.10) has the following geometric characterization

$$\max_{\substack{x,y\in\Delta_{n-1}\\ x\neq y}} \frac{\|Ay - Ax\|}{\|y - x\|_1} = \frac{\mathsf{diam}(A)}{2}. \tag{3.18}$$

Example 4 below, a special case of Corollary 6, shows that for $f(x) = \frac{1}{2}\|Ax - b\|_2^2$, $X = \Delta_{n-1}$, and $D(y, x) = \frac{1}{2}\|y - x\|_1^2$ the relative condition number $L_{f,\Delta_{n-1},D}/\mu_{f,\Delta_{n-1},D}$ is the square of $\mathsf{diam}(A)/\Phi(A)$, which has a flavor of an aspect ratio of $\mathsf{conv}(A)$. This gives an interesting analogy to the classical condition number of f.

Example 4. Suppose $\mathbb{R}^n$ is endowed with the one-norm, $\mathbb{R}^m$ is endowed with the Euclidean norm, and $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ with at least two different columns and $b \in \mathbb{R}^m$. Then for $D(y, x) := \frac{1}{2}\|y - x\|_1^2$ Corollary 6 and identities (3.18) and (3.17) yield

$$L_{f,\Delta_{n-1},D} = \frac{\mathsf{diam}(A)^2}{4} \quad \text{and} \quad \mu_{f,\Delta_{n-1},D} = \frac{\Phi(A)^2}{4}.$$

In particular,

$$\frac{L_{f,\Delta_{n-1},D}}{\mu_{f,\Delta_{n-1},D}} = \left(\frac{\mathsf{diam}(A)}{\Phi(A)}\right)^2.$$

3.4 Properties of $\mu^{\star}_{f,X,D}$, and $\mu^{\sharp}_{f,X,D}$ when D is a Squared Norm

We next provide bounds on $\mu^{\star}_{f,X,D}$ and $\mu^{\sharp}_{f,X,D}$ analogous to those developed in Section 3.3 for $\mu_{f,X,D}$. Proposition 3 already established $\mu^{\star}_{f,X,D} \geq \mu_{f,X,D} \geq 0$. It is intuitively clear that $\mu^{\star}_{f,X,D}$ could be a lot larger. When D is a squared norm, the techniques developed in [50] show that $\mu^{\sharp}_{f,X,D} \geq \mu^{\star}_{f,X,D}$. Indeed, when D is a squared norm, the relationship among other variants of strong convexity introduced in [50] extend to our context in a straightforward fashion as we next explain.

Definition 6. Suppose (f, X, D) satisfy Assumption 3.

(a) We say that f has D-*under approximation* on X if there exists a constant $\mu > 0$ such that

$$D_f(x, \bar{x}) \geq \mu D(\bar{x}, x) \text{ for all } x \in X. \tag{3.19}$$

(b) We say that f has D-*gradient growth* on X if there exists a constant $\mu > 0$ such that

$$\langle \nabla f(x) - \nabla f(\bar{x}), x - \bar{x} \rangle \geq \mu D(\bar{x}, x) \text{ for all } x \in X. \tag{3.20}$$

Suppose (f, X, D) satisfies Assumption 3 and D is a squared norm. Then for $\mu > 0$ [50, Theorem 4] yields the following chain of implications for (f, X, D, μ):

$$(3.4) \Rightarrow (3.7) \Rightarrow (3.19) \Rightarrow (3.20) \Rightarrow (3.8).$$

We note that [50, Theorem 4] is stated and proven for the Euclidean norm but the same statement and proof hold for any norm.

From the above chain of implications it follows that if (f, X, D) satisfies Assumption 3 and D is a squared norm then $\mu_{f,X,D} \leq \mu^{\star}_{f,X,D} \leq \mu^{\sharp}_{f,X,D}$. In particular, any lower bound on $\mu_{f,X,D}$, such as those in Theorem 8 or Theorem 9, is also a lower bound on $\mu^{\star}_{f,X,D}$ and on $\mu^{\sharp}_{f,X,D}$ when D is a squared norm. We next show that the ideas in Section 3.3 can be extended to obtain sharper bounds on these two constants.

3.4.1 A Sharper Lower Bound on $\mu^{\star}_{f,X,D}$

Suppose $A \in \mathbb{R}^{m \times n}$ and $X \subseteq \mathbb{R}^n$ is a polyhedron such that $A(X)$ contains more than one point, and $S \subseteq X$ is nonempty. Proposition 6 readily implies

$$\inf_{\substack{y \in S \\ x \in X \setminus Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|} \geq \min_{C \in \mathcal{S}(X)} \frac{1}{\|(A|C)^{-1}\|} > 0. \tag{3.21}$$

Proposition 7 below, which extends Proposition 6, gives a sharper version of (3.21). Suppose $A \in \mathbb{R}^{m \times n}$, $X \subseteq \mathbb{R}^n$ is a polyhedron, and $S \subseteq X$ is nonempty. Let

$$\mathcal{T}(X; S, A) := \{T_X(y; S, A) : y \in X\}$$

where

$$T_X(y;S,A) := \{d \in \mathbb{R}^n : y + td \in X \text{ and } A(y+td) \in \overline{\mathsf{conv}}(A(S)) \text{ for some } t > 0\}.$$

Proposition 7. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms. Let $A \in \mathbb{R}^{m\times n}$ and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. Then for all nonempty $S \subseteq X$*

$$\inf_{\substack{y\in S\\ x\in X\setminus Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|} \geq \inf_{C\in\mathcal{T}(X;S,A)} \frac{1}{\|(A|C)^{-1}\|} \geq \min_{C\in\mathcal{T}(X)} \frac{1}{\|(A|C)^{-1}\|}.$$

Furthermore, if $A(S)$ is closed and convex then

$$\inf_{\substack{y\in S\\ x\in X\setminus Z_{A,X}(y)}} \frac{\|Ay-Ax\|}{\|Z_{A,X}(y)-x\|} = \inf_{C\in\mathcal{T}(X;S,A)} \frac{1}{\|(A|C)^{-1}\|}.$$

Proof. The construction of $T_X(y;S,A)$ readily implies that $T_X(y;S,A) \subseteq T_X(y)$ and $\|(A|T_X(y;A,S))^{-1}\| \leq \|(A|T_X(y))^{-1}\|$ for all $y \in X$. Hence

$$\sup_{C\in\mathcal{T}(X;S,A)} \|(A|C)^{-1}\| \leq \max_{C\in\mathcal{T}(X)} \|(A|C)^{-1}\|.$$

The rest of the proof follows from a straightforward modification of the proof in [57] of Proposition 6 by using $\sup_{C\in\mathcal{T}(X;S,A)} \|(A|C)^{-1}\|$ in lieu of $\max_{C\in\mathcal{T}(X)} \|(A|C)^{-1}\|$. □

The following theorem gives a lower bound on $\mu^\star_{f,X,D}$ analogous to the one on $\mu_{f,X,D}$ in Theorem 9. In light of Proposition 7, the lower bound on $\mu^\star_{f,X,D}$ in Theorem 10 is at least as large, and possibly much larger, than the one on $\mu_{f,X,D}$ in Theorem 9.

Theorem 10. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(y,x) = \frac{1}{2}\|y-x\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ has more than one point. If g is μ_g-strongly convex on $A(X)$ for the norm in $\mathbb{R}^m$ then the function $f = g\circ A$ satisfies*

$$\mu^\star_{f,X,D} \geq \inf_{\substack{y\in \bar{X}\\ x\in X\setminus Z_{A,X}(y)}} \frac{\mu_g\|Ay-Ax\|^2}{\|Z_{A,X}(y)-x\|^2} = \inf_{C\in\mathcal{T}(X;\bar{X},A)} \frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

Proof. Observe that for all $y \in \bar{X}$ and $x \in X$

$$D_f(y,x) = g(Ay) - g(Ax) - \langle g(Ax), A(y-x)\rangle .$$

Since g is μ_g strongly convex on $A(X)$, it follows that $D_f(y,x) \geq \mu_g\|Ay-Ax\|^2/2$ for all $y \in \bar{X}$ and $x \in X$, and it also follows that $Z_{A,X}(y) = \{x \in X : Ax = Ay\} = \bar{X}$ for all $y \in \bar{X}$. Therefore

$$\mu^\star_{f,X,D} = \inf_{x\in X\setminus\bar{X}} \frac{D_f(\bar{x},x)}{\|\bar{x}-x\|^2/2} \geq \inf_{\substack{y\in\bar{X}\\ x\in X\setminus\bar{X}}} \frac{D_f(y,x)}{\|y-x\|^2/2} \geq \inf_{\substack{y\in\bar{X}\\ x\in X\setminus Z_{A,X}(y)}} \frac{\mu_g\|Ay-Ax\|^2}{\|Z_{A,X}(y)-x\|^2}.$$

To finish, observe that $A(\bar{X})$ is closed and convex and apply Proposition 7. □

Once again there is an interesting connection with the developments in [60] when $X = \Delta_{n-1}$. Consider the special case when $X = \Delta_{n-1}$, $A \in \mathbb{R}^{m\times n}$ has at least two different columns, $S \subseteq \Delta_{n-1}$ is nonempty, and $G \in \mathsf{faces}(\mathsf{conv}(A))$ is the smallest face of $\mathsf{conv}(A)$ that contains $A(S)$. From [60, Theorem 3] it follows that if $\mathbb{R}^n$ is endowed with the one-norm then

$$\inf_{\substack{y\in S\\ x\in X\setminus Z_{A,X}(y)}} \frac{\|Ay - Ax\|}{\|Z_{A,X}(y) - x\|_1} \geq \min_{\substack{F\in\mathsf{faces}(G)\\ \emptyset\neq F\neq\mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A \setminus F)). \tag{3.22}$$

The following example illustrates the difference between $\mu_{f,X,D}$ and $\mu^\star_{f,X,D}$.

Example 5. Suppose $\mathbb{R}^n$ is endowed with the one-norm and $D(y,x) := \frac{1}{2}\|y-x\|_1^2$. Suppose $\mathbb{R}^m$ is endowed with the Euclidean norm, and $f(x) = \frac{1}{2}\|Ax-b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ with at least two different columns and $b \in \mathbb{R}^m$. As noted in Example 4, in this case

$$\mu_{f,\Delta_{n-1},D} = \frac{\Phi(A)^2}{4} = \frac{1}{4}\left(\min_{\substack{F\in\mathsf{faces}(\mathsf{conv}(A))\\ \emptyset\neq F\neq\mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A\setminus F))\right)^2.$$

This relative strong convexity constant depends only on A but not on b. On the other hand, the smallest face of $\mathsf{conv}(A)$ containing $\bar{X}$ is

$$G(b) := \operatorname*{argmin}_{G\in\mathsf{faces}(\mathsf{conv}(A))} \mathsf{dist}(G, b),$$

which evidently depends on both A and b. Theorem 10 and (3.22) yield

$$\mu^\star_{f,\Delta_{n-1},D} \geq \frac{1}{4}\left(\min_{\substack{F\in\mathsf{faces}(G(b))\\ \emptyset\neq F\neq\mathsf{conv}(A)}} \mathsf{dist}(F, \mathsf{conv}(A\setminus F))\right)^2.$$

3.4.2 A Sharper Lower Bound on $\mu^\sharp_{f,X,D}$

Suppose $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is defined as $f(x) = g(Ax)+\langle c, x\rangle$ where $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ is a strongly convex function, $A \in \mathbb{R}^{m\times n}$ and $c \in \mathbb{R}^n$. Theorem 10 does not apply to this kind of function due to the extra linear term $\langle c, x\rangle$. Indeed for a function of this form the constant $\mu^\star_{f,X,D}$ may be zero, see Example 6 below. On the other hand, the next result shows that for a function of this form and for a polyhedral set X it is always the case that $\mu^\sharp_{f,X,D} > 0$ provided a suitable linear cut is added to X.

Theorem 11. *Suppose $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with norms and $D(x,y) = \frac{1}{2}\|x-y\|^2$ for the norm $\|\cdot\|$ in $\mathbb{R}^n$. Let $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$, and $X \subseteq \mathbb{R}^n$ be a polyhedron such that $A(X)$ contains more than one point. Suppose $g : \mathbb{R}^m \to \mathbb{R}\cup\{\infty\}$ is μ_g-strongly convex on $A(X)$ for the norm in $\mathbb{R}^m$ and $f : \mathbb{R}^n \to \mathbb{R}\cup\{\infty\}$ is defined via $f(x) = g(Ax)+\langle c, x\rangle$. Then the vector $v := 2\nabla f(y)$ is the same for all $y \in \bar{X}$ and satisfies $\langle v, x-y\rangle \geq 0$ for all $x \in X$, $y \in \bar{X}$. Furthermore, one of the following two possible cases applies depending on the range of values of $\langle v, x-y\rangle$ for $x \in X, y \in \bar{X}$.*

Case 1: *For all $x \in X, y \in \bar{X}$ we have $\langle v, x - y\rangle = 0$. In this case*

$$\mu^{\sharp}_{f,X,D} \geq \inf_{\substack{y\in\bar{X}\\ x\in X\setminus Z_{A,X}(y)}} \frac{\mu_g\|Ay - Ax\|^2}{\|Z_{A,X}(y) - x\|^2} = \inf_{C\in\mathcal{T}(X;\bar{X},A)} \frac{\mu_g}{\|(A|C)^{-1}\|^2}.$$

Case 2: *For some $x \in X, y \in \bar{X}$ we have $\langle v, x - y\rangle > 0$. In this case for all $\delta > 0$*

$$\begin{aligned}\mu^{\sharp}_{f,X_\delta,D} \geq \inf_{\substack{y\in\bar{X}\\ x\in X_\delta\setminus Z_{M,X}(y)}} \frac{\|Ay - Ax\|^2 + \langle v, y - x\rangle}{\|Z_{M,X}(y) - x\|^2} &\geq \inf_{\substack{y\in\bar{X}\\ x\in X_\delta\setminus Z_{M,X}(y)}} \frac{\|My - Mx\|^2}{\|Z_{M,X}(y) - x\|^2}\\ &= \inf_{C\in\mathcal{T}(X_\delta;\bar{X},M)} \frac{1}{\|(M|C)^{-1}\|^2},\end{aligned}$$

for the polyhedron $X_\delta := \{x \in X : \langle v, x - y\rangle \leq \delta \text{ for all } y \in \bar{X}\} \supseteq \bar{X}$, the matrix $M \in \mathbb{R}^{(m+1)\times n}$, and the norm $\|\cdot\|$ in $\mathbb{R}^{m+1}$ defined as follows

$$M := \begin{bmatrix}\sqrt{\mu_g}\cdot A\\ \frac{1}{\sqrt{\delta}}\cdot v^{\mathsf{T}}\end{bmatrix} \quad \text{and} \quad \left\|\begin{bmatrix}y\\ y_{m+1}\end{bmatrix}\right\| := \sqrt{\|y\|^2 + y_{m+1}^2}.$$

Proof. The optimality conditions for $\min_{x\in X} f(x)$ imply that

$$\langle \nabla f(y), x - y\rangle = \left\langle A^{\mathsf{T}}\nabla g(Ay) + c, x - y\right\rangle \geq 0 \text{ for all } x \in X, y \in \bar{X}. \tag{3.23}$$

Thus for all $y, y' \in \bar{X}$ the strong convexity of g and (3.23) imply

$$\mu_g\|Ay - Ay'\|^2 \leq \langle \nabla g(Ay) - \nabla g(Ay'), Ay - Ay'\rangle = \langle \nabla f(y) - \nabla f(y'), y - y'\rangle \leq 0.$$

Hence $Ay = Ay'$ whenever $y, y' \in \bar{X}$. In particular, $v = 2\nabla f(y) = 2(A^{\mathsf{T}}\nabla g(Ay) + c)$ is the same for all $y \in \bar{X}$. Furthermore, the optimality conditions for $\min_{x\in X} f(x)$ imply that $\langle v, x - y\rangle \geq 0$ for all $x \in X, y \in Y^\star$. In particular, $\langle v, y\rangle = \min_{x\in X}\langle v, x\rangle$ for all $y \in \bar{X}$.

Next, the strong convexity of g on $A(X)$ implies that for all $x \in X, y \in \bar{X}$

$$\begin{aligned}f(x) - \bar{f} &= g(Ax) - g(Ay) + \langle c, x - y\rangle\\ &\geq \frac{\mu_g}{2}\|Ax - Ay\|^2 + \langle \nabla g(Ay), Ax - Ay\rangle + \langle c, x - y\rangle\\ &= \frac{1}{2}\left(\mu_g\|Ax - Ay\|^2 + \langle v, x - y\rangle\right).\end{aligned}$$

If $\langle v, x - y\rangle = 0$ for all $x \in X, y \in \bar{X}$ then Case 1 applies. In this case $Z_{A,X}(y) = \{x \in X : Ax = Ay\} = \bar{X}$ for all $y \in \bar{X}$ and thus

$$\mu^{\sharp}_{f,X,D} = \inf_{\substack{y\in\bar{X}\\ x\in X\setminus\bar{X}}} \frac{f(x) - \bar{f}}{\|y - x\|^2/2} \geq \inf_{\substack{y\in\bar{X}\\ x\in X\setminus Z_{A,X}(y)}} \frac{\mu_g\|Ay - Ax\|^2}{\|Z_{A,X}(y) - x\|^2}.$$

If $\langle v, x - y\rangle > 0$ for some $x \in X, y \in \bar{X}$ then Case 2 applies. In this case $Z_{M,X}(y) = \{x \in X : Ax = Ay, \langle v, x\rangle = \langle v, y\rangle\} = \bar{X}$ for all $y \in \bar{X}$ and thus

$$\mu^{\sharp}_{f,X_\delta,D} = \inf_{\substack{y\in\bar{X}\\ x\in X_\delta\setminus\bar{X}}} \frac{f(x)-\bar{f}}{\|y-x\|^2/2} \geq \inf_{\substack{y\in\bar{X}\\ x\in X_\delta\setminus Z_{M,X}(y)}} \frac{\mu_g\|Ay-Ax\|^2+\langle v, y-x\rangle}{\|Z_{M,X}(y)-x\|^2}.$$

Next, observe that for $y \in \bar{X}$ and $x \in X_\delta$

$$\mu_g\|Ay-Ax\|^2+\langle v, y-x\rangle \geq \mu_g\|Ay-Ax\|^2+\frac{\langle v, y-x\rangle^2}{\delta} = \|My-Mx\|^2.$$

To finish, apply Proposition 7 after observing that $A(\bar{X})$ is closed and convex in Case 1 and likewise for $M(\bar{X})$ in Case 2. □

Observe that if X in Theorem 11 is bounded then Case 2 gives a lower bound on $\mu^{\sharp}_{f,X,D}$ by taking $\delta := \max_{x\in X, y\in\bar{X}} \langle v, x - y\rangle$ because $X = X_\delta$ for this choice of δ.

We conclude this section with a simple example showing that $\mu^{\sharp}_{f,X,D} > \mu^{\star}_{f,X,D} = 0$ can occur. The example also shows that the additional bound on X_δ in Theorem 11, Case 2 cannot simply be dropped without making some additional assumptions.

Example 6. Let $\mathbb{R}^3$ be endowed with the one-norm and let $D(y, x) := \frac{1}{2}\|y - x\|_1^2$. Suppose $f : \mathbb{R}^3 \to \mathbb{R}$ is as follows

$$f(x) = \frac{1}{2}(x_1 - x_2)^2 + x_3.$$

If $X := \Delta_2 \subseteq \mathbb{R}^3$ then $\bar{X} = \{\begin{bmatrix}1/2 & 1/2 & 0\end{bmatrix}^{\mathsf{T}}\}$. For $x = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^{\mathsf{T}}$ we have $f(\bar{x}) - f(x) - \langle\nabla f(x), \bar{x} - x\rangle = 0$ and $\|\bar{x} - x\|_1 = 2$. Hence $\mu^{\star}_{f,X,D} = 0$. On the other hand, Theorem 11 implies that $\mu^{\sharp}_{f,X,D} > 0$. A more careful calculation shows that in this case $\mu^{\sharp}_{f,X,D} = 1/2$.

On the other hand, if $X = \mathbb{R}^3_+$ then $\bar{X} = \{\begin{bmatrix}t & t & 0\end{bmatrix}^{\mathsf{T}} : t \geq 0\}$. For $t > 0$ and $x = \begin{bmatrix}0 & 0 & t\end{bmatrix}^{\mathsf{T}}$ we have $f(x) - \bar{f} = t$ and $\|\bar{X} - x\|_1 = t$. Therefore $\mu^{\sharp}_{f,X,D} = 0$. Furthermore, in the context of Theorem 11 we have $v = \begin{bmatrix}0 & 0 & 1\end{bmatrix}^{\mathsf{T}}$. Thus for all $\delta > 0$ we have $X_\delta := \{x \in X : x_3 \leq \delta\}$ and $\mu^{\sharp}_{f,X_\delta,D} = 2/\delta > 0$.

3.5 Convergence of First-order Methods

In our statements in this section sometimes it will be convenient to use the following notation adapted from [50] about some functional classes. Given a convex set $X \subseteq \mathbb{R}^n$, a distance-like function $D : X \times X \to \mathbb{R}$, and positive constants L and μ, let $r\mathcal{S}_{L,\mu}(X, D)$, $q\mathcal{S}_{L,\mu}(X, D)$, and $\mathcal{F}_{L,\mu}(X, D)$ be defined as follows.

$$\begin{aligned}
r\mathcal{S}_{L,\mu}(X, D) &:= \{f : (f, X, D) \text{ satisfy Assumption } 3 \text{ and both (3.3) and (3.4) hold}\},\\
q\mathcal{S}_{L,\mu}(X, D) &:= \{f : (f, X, D) \text{ satisfy Assumption } 3 \text{ and both (3.3) and (3.7) hold}\},\\
\mathcal{F}_{L,\mu}(X, D) &:= \{f : (f, X, D) \text{ satisfy Assumption } 3 \text{ and both (3.3) and (3.8) hold}\}.
\end{aligned}$$

Observe that $r\mathcal{S}_{L,\mu}(X,D) \subseteq q\mathcal{S}_{L,\mu}(X,D)$ by Proposition 3. In addition, $q\mathcal{S}_{L,\mu}(X,D) \subseteq \mathcal{F}_{L,\mu}(X,D)$ when D is a squared norm as we noted in Section 3.4.

3.5.1 Mirror Descent Algorithm

Suppose $h : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and differentiable on $X \subseteq \mathbb{R}^n$ and the *Bregman proximal map*

$$g \mapsto \operatorname*{argmin}_{y \in X} \{\langle g, y\rangle + LD_h(y,x)\}$$

is computable for $x \in X$ and $L > 0$. The *mirror descent algorithm* for problem (5) is based on the following update for $x \in X$:

$$x_+ := \operatorname*{argmin}_{y \in X} \{\langle \nabla f(x), y\rangle + LD_h(y,x)\}$$

Algorithm 8 gives a description of the mirror descent algorithm for (5).

Algorithm 8 Mirror descent algorithm

1: Pick $x_0 \in X$;
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: choose $L_k > 0$
4: $x_{k+1} = \operatorname*{argmin}_{y \in X} \{\langle \nabla f(x_k), y\rangle + L_k D_h(y, x_k)\}$
5: **end for**

Proposition 8 and Proposition 9 show the linear convergence of Algorithm 8 when $f \in q\mathcal{S}_{L,\mu}(X, D_h)$ and when $f \in \mathcal{F}_{L,\mu}(X, D_h)$ respectively. We should note that Proposition 8 and its proof are straightforward modifications of the linear convergence results in [48] and [69]. The latter results rely on some relative smoothness and strong convexity assumptions. Proposition 8 shows that the linear convergence of Algorithm 8 holds with a sharper rate and under the weaker quasi strong convexity assumption. The following lemma, which is a straightforward extension of results presented in [69], provides the crux of the proof of Proposition 8.

Lemma 1. *Suppose* $f \in q\mathcal{S}_{L,\mu}(X, D_h), x \in X,$ *and*

$$x_+ = \operatorname*{argmin}_{y \in X} \{f(x) + \langle \nabla f(x), y - x\rangle + LD_h(y,x)\}. \tag{3.24}$$

Then

$$f(x_+) - \bar{f} \leq (L - \mu)D_h(\bar{x}, x) - LD_h(\bar{x}, x_+). \tag{3.25}$$

Proof. Since $f \in q\mathcal{S}_{L,\mu}(X, D_h)$ we have

$$f(x_+) \leq f(x) + \langle \nabla f(x), x_+ - x\rangle + LD_h(x_+, x). \tag{3.26}$$

and

$$f(x) \leq \bar{f} + \langle \nabla f(x), x - \bar{x} \rangle - \mu D_h(\bar{x}, x). \tag{3.27}$$

In addition, the three-point property of D_h [18, Lemma 3.1] yields

$$D_h(x_+, x) = D_h(\bar{x}, x) - D_h(\bar{x}, x_+) + \langle \nabla h(x_+) - \nabla h(x), x_+ - \bar{x} \rangle . \tag{3.28}$$

By putting together (3.26), (3.27), and (3.28) we get

$$\begin{aligned} f(x_+) \leq \bar{f} &+ (L - \mu) D_h(\bar{x}, x) - L D_h(\bar{x}, x_+) \\ &+ \langle \nabla f(x) + L(\nabla h(x_+) - \nabla h(x)), x_+ - \bar{x} \rangle . \end{aligned}$$

We get (3.25) by observing that the optimality conditions for (3.24) imply

$$\langle \nabla f(x) + L(\nabla h(x_+) - \nabla h(x)), x_+ - \bar{x} \rangle \leq 0.$$

□

Proposition 8. *Suppose $f \in q\mathcal{S}_{L,\mu}(X, D_h)$. If $L_k = L,\ k = 0, 1, \ldots$ in Algorithm 8 then the iterates generated by Algorithm 8 satisfy*

$$D_h(\bar{X}, x_k) \leq \left(1 - \frac{\mu}{L}\right)^k D_h(\bar{X}, x_0) \quad \textit{for} \quad k = 0, 1, \ldots \tag{3.29}$$

and

$$f(x_k) - \bar{f} \leq L \left(1 - \frac{\mu}{L}\right)^k D_h(\bar{X}, x_0) \quad \textit{for} \quad k = 1, 2, \ldots .$$

Proof. Lemma 1 applied to $x = x_k$ implies that

$$(L - \mu) D_h(\bar{x}_k, x_k) - L D_h(\bar{x}_k, x_{k+1}) \geq f(x_{k+1}) - \bar{f} \geq 0 \ \text{ for } \ k = 0, 1, \ldots . \tag{3.30}$$

Therefore

$$D_h(\bar{X}, x_{k+1}) \leq D_h(\bar{x}_k, x_{k+1}) \leq \left(1 - \frac{\mu}{L}\right) D_h(\bar{x}_k, x_k) \ \text{ for } \ k = 0, 1, \ldots .$$

Thus (3.29) readily follows. Inequality (3.30) also yields

$$f(x_k) - \bar{f} \leq L \left(1 - \frac{\mu}{L}\right) D_h(\bar{X}, x_{k-1}) \leq L \left(1 - \frac{\mu}{L}\right)^k D_h(\bar{X}, x_0) \ \text{ for } \ k = 1, 2, \ldots .$$

□

Proposition 8 implies that if $f \in q\mathcal{S}_{L,\mu}(X, D_h)$ then Algorithm 8 yields $x_k \in X$ such that $f(x_k) - \bar{f} < \epsilon$ in at most

$$\mathcal{O}\left(\frac{L}{\mu} \log\left(\frac{L D_h(\bar{X}, x_0)}{\epsilon}\right)\right)$$

iterations.

Proposition 9 below shows that the same kind of iteration bound holds when $f \in \mathcal{F}_{L,\mu}(X, D_h)$. Observe that neither Proposition 8 nor Proposition 9 implies the other since neither $q\mathcal{S}_{L,\mu}(X, D_h)$ nor $\mathcal{F}_{L,\mu}(X, D_h)$ necessarily includes the other. (See Example 3 and Example 6.)

Proposition 9. *Suppose $f \in \mathcal{F}_{L,\mu}(X, D_h)$. If $L_k = L,\ k = 0, 1, \ldots$ in Algorithm 8 then for $K = \lceil 2L/\mu \rceil$ the iterates generated by Algorithm 8 satisfy*

$$D_h(\bar{X}, x_{k+K}) \leq \frac{D_h(\bar{X}, x_k)}{2} \quad \textit{for} \quad k = 0, 1, 2, \ldots. \tag{3.31}$$

In addition, Algorithm 8 yields $x_k \in X$ such that $f(x_k) - \bar{f} < \epsilon$ in at most

$$\mathcal{O}\left(\frac{L}{\mu} \log\left(\frac{L D_h(\bar{X}, x_0)}{\epsilon}\right)\right) \tag{3.32}$$

iterations.

Proof. Since $f \in \mathcal{F}_{L,\mu}(X, D_h)$ and $L_k = L$, it follows from [48, Theorem 3.1] that the $(k + K)$-th iterate generated by Algorithm 8 satisfies

$$D_h(\bar{X}, x_{k+K}) \leq \frac{1}{\mu}(f(x_{k+K}) - \bar{f}) \leq \frac{L}{\mu K} D_h(\bar{X}, x_k) \leq \frac{D_h(\bar{X}, x_k)}{2}.$$

Thus (3.31) follows. It also follows that $k = mK,\ m = 1, 2, \ldots$

$$f(x_{mK}) - \bar{f} \leq \frac{L D_h(\bar{X}, x_{(m-1)k})}{K} \leq \frac{L D_h(\bar{X}, x_0)}{2^{m-1}}$$

and thus (3.32) follows as well. □

To ease our exposition we assumed $L_k = L$ in Proposition 8 and Proposition 9. However, it is easy to see that these two results also hold if the assumption $L_k = L$ is relaxed to the assumption $L_k \leq L$ and $f(x_{k+1}) \leq \min_{y \in X} \{\langle \nabla f(x_k), y \rangle + L_k D_h(y, x_k)\}$. The latter condition is easier to implement via standard backtracking.

In the recent paper [31], we showed that if $f \in \mathcal{F}_{L,\mu}(X, D_h)$ then an accelerated version of Algorithm 8 with periodic restart has a linear accelerated rate similar to (3.32) but with L/μ replaced by $(L/\mu)^{1/\gamma}$ where $\gamma > 0$ is the *triangle scaling constant* of the Bregman distance D_h as defined in [34].

3.5.2 Frank-Wolfe Algorithm with Away Steps

Suppose $X \subseteq \mathbb{R}^n$ is a polytope and a *vertex linear oracle* for X is available, that is, the map

$$g \mapsto \operatorname*{argmin}_{y \in X} \langle g, x \rangle$$

is computable and outputs a vertex of X for all $g \in \mathbb{R}^n$.

The Frank-Wolfe algorithm, also known as the conditional gradient algorithm, for (5) is based on the following update for $x \in X$

$$u := \operatorname*{argmin}_{y \in X} \langle \nabla f(x), y \rangle$$
$$x_+ := x + \alpha(u - x) \quad \text{for some} \quad \alpha \in [0, 1].$$

Each step of the Frank-Wolfe algorithm *adds weight* to some vertex u. The basic idea of the Frank-Wolfe algorithm with away steps is to combine the above *regular steps* with *away steps* that *reduce weight* from some vertex a. To that end, the algorithm requires an additional *vertex representation* of $x \in X$. More precisely, let $S(x) \subseteq \mathsf{vertices}(X)$ and $\lambda(x) \in \Delta(S(x)) := \{z \in \mathbb{R}_+^{S(x)} : \|z\|_1 = 1\}$ be such that

$$x = \sum_{s \in S(x)} \lambda_s(x) s \quad \text{and} \quad \lambda(x) > 0.$$

Algorithm 9 describes a Frank-Wolfe algorithm with away steps. We should highlight that although the set $\mathsf{vertices}(X)$ could be immense, the algorithm does not require it explicitly. Instead the algorithm only maintains $S(x)$ and $\lambda(x)$ that are far more manageable. Indeed, by using the IRR procedure in [6] or its modification described in [30], Step 10 in Algorithm 9 can guarantee that the sets $S(x_k)$ have size at most $n+1$ for $k = 0, 1, \ldots$.

Algorithm 9 Frank-Wolfe algorithm with away steps

1: Pick $x_0 \in \mathsf{vertices}(X)$; $S(x_0) := \{x_0\}$; $\lambda(x_0) = 1$
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: $\quad u := \operatorname{argmin}_{y \in X} \langle \nabla f(x_k), y \rangle$; $\quad a := \operatorname{argmax}_{y \in S(x_k)} \langle \nabla f(x_k), y \rangle$
4: $\quad$ **if** $\langle \nabla f(x_k), u - x_k \rangle < \langle \nabla f(x_k), x_k - a \rangle$ **then** (regular step)
5: $\quad\quad v := u - x_k$; $\alpha_{\max} = 1$;
6: $\quad$ **else** (away step)
7: $\quad\quad v := x_k - a$; $\alpha_{\max} = \frac{\lambda_a(x_k)}{1 - \lambda_a(x_k)}$;
8: $\quad$ **end if**
9: $\quad x_{k+1} := x_k + \alpha_k v$ for some $\alpha_k \in [0, \alpha_{\max}]$
10: $\quad$ update $S(x_{k+1})$ and $\lambda(x_{k+1})$
11: **end for**

Proposition 10 below establishes the linear convergence of Algorithm 9 under suitable assumptions on the chosen direction v at each iteration. This proposition can be readily inferred from the ideas and results introduced in [42] and further developed in [6, 60]. The rate of convergence in Proposition 10 stated in terms of the ratio $\sigma/\mathcal{L}$ is at least as sharp as the sharpest of the linear rates established in [6, 42, 60]. To provide a full picture of this linear convergence result, the section 3.6 gives a proof of Proposition 10 that replicates the main ideas in [6, 42, 60]. We discuss only the linear convergence of Algorithm 9 but the same techniques yield similar results for other variants of the Frank-Wolfe algorithm as those discussed in [42] or the one recently developed in [15].

Proposition 10. *Suppose $X \subseteq \mathbb{R}^n$ is a polytope and there exist positive constants σ and $\mathcal{L}$ such that the following conditions hold at each iteration k of Algorithm 9. First,*

$$\langle \nabla f(x_k), v \rangle^2 \geq \sigma(f(x_k) - \bar{f}). \tag{3.33}$$

Second, the stepsize α_k in Step 9 of Algorithm 9 is chosen so that

$$f(x_{k+1}) = f(x_k + \alpha_k v) \leq \min_{\alpha \in [0, \alpha_{\max}]} \left\{ f(x_k) + \alpha \langle \nabla f(x_k), v \rangle + \frac{\mathcal{L}\alpha^2}{2} \right\}. \tag{3.34}$$

Then the iterates generated by Algorithm 9 satisfy

$$f(x_k) - \bar{f} \leq \left(1 - \min\left\{\frac{1}{2}, \frac{\sigma}{2\mathcal{L}}\right\}\right)^{k/2} (f(x_0) - \bar{f}) \ \ \textit{for} \ \ k = 1, 2, \ldots.$$

We next give bounds on the constants $\mathcal{L}$ and σ in Proposition 10 in terms of the relative smoothness and quasi strong convexity or D-functional growth constants of a canonical function $\tilde{f}$ associated to f and X. To that end, let $A \in \mathbb{R}^{n \times N}$ denote the matrix whose columns are $\mathsf{vertices}(X)$ and let $\tilde{f} : \mathbb{R}^{N-1} \to \mathbb{R} \cup \{\infty\}$ be defined via $\tilde{f} := f \circ A$. Thus $X = \mathsf{conv}(A) = \{Az : z \in \Delta_{N-1}\}$ and for all $x = Az \in X$ we have $f(x) = \tilde{f}(z)$. Once again, we observe that although N could be immense, both A and $\tilde{f}$ are used only in our analysis but Algorithm 9 does not require them explicitly. Let

$$\Delta^\star_{N-1} := \{z \in \Delta_{N-1} : Az \in \bar{X}\} = \left\{ z \in \Delta_{N-1} : \tilde{f}(z) = \min_{w \in \Delta_{N-1}} \tilde{f}(w) \right\}.$$

Consider the following distance-like function in $\mathbb{R}^N$:

$$D(z, w) := \frac{1}{2}\|z - w\|_1^2.$$

Observe that $\tilde{f}$ is strongly convex relative to (Δ_{N-1}, D) if f is strongly convex on X. Indeed, Proposition 4, Theorem 9, and identities (3.18) and (3.17) imply that if f is L_f Lipschitz and μ_f strongly convex on X then

$$L_{\tilde{f}, \Delta_{N-1}, D} \leq \frac{L_f \cdot \mathsf{diam}(A)^2}{4} \ \ \text{and} \ \ \mu^\star_{\tilde{f}, \Delta_{N-1}, D} \geq \mu_{\tilde{f}, \Delta_{N-1}, D} \geq \frac{\mu_f \cdot \Phi(A)^2}{4}.$$

Similarly, Theorem 11 implies that $\mu^\sharp_{\tilde{f}, \Delta_{N-1}, D} > 0$ if f is of the form $f(x) = g(Ex) + \langle c, x \rangle$ for g strongly convex.

The following result provides an interesting link between the conditioning of $\tilde{f}$ relative to (Δ_{N-1}, D) and the constants $\mathcal{L}$ and σ in Proposition 10. It also shows an interesting identity between the relative smoothness constant $L_{\tilde{f}, \Delta_{N-1}, D}$ and the following *away curvature constant* C_f^A defined by Lacoste-Julien and Jaggi [42]:

$$C_f^A := \sup_{\substack{x, u, w \in X, \ \alpha > 0 \\ x + \alpha(u - w) \in X}} \frac{2}{\alpha^2}(f(x + \alpha(u - w)) - f(x) - \alpha \langle \nabla f(x), u - w \rangle). \tag{3.35}$$

We note that this expression for C_f^A is equivalent to but not identical to the definition of C_f^A in [42].

Proposition 11. *Suppose $f, X, A, \tilde{f}$, and D are as above.*

(a) *Inequality* (3.33) *holds for* $\sigma = 2\mu^\star_{\tilde{f},\Delta_{N-1},D}$.

(b) *Inequality* (3.33) *holds for* $\sigma = \mu^\sharp_{\tilde{f},\Delta_{N-1},D}/2$.

(c) $4L_{\tilde{f},\Delta_{N-1},D} = C^A_f$. *In particular, inequality* (3.34) *holds for* $\mathcal{L} = 4L_{\tilde{f},\Delta_{N-1},D}$ *and*

$$\alpha_k := \operatorname*{argmin}_{\alpha\in[0,\alpha_{\max}]} \left\{ f(x_k) + \alpha \left\langle \nabla f(x_k), v \right\rangle + \frac{\mathcal{L}\alpha^2}{2} \right\} = \min\left\{ -\frac{\left\langle \nabla f(x_k), v\right\rangle}{\mathcal{L}}, \alpha_{\max} \right\}. \tag{3.36}$$

Proof. Suppose that we are at iteration k of Algorithm 9. To ease notation, let $x := x_k$ and $S := S(x_k)$. Then $u = \operatorname{argmin}_{y\in X} \left\langle \nabla f(x), y\right\rangle$, $a := \operatorname{argmax}_{s\in S} \left\langle \nabla f(x), s\right\rangle$, and

$$-\left\langle \nabla f(x), v\right\rangle \geq \frac{\left\langle \nabla f(x), a-u\right\rangle}{2}. \tag{3.37}$$

In addition, $x = \sum_{s\in S} \lambda_s(x) s$ for some $\lambda(x) \in \Delta(S)$. By setting all components outside S to zero, we can take $\lambda(x) \in \Delta_{N-1}$ and write $x = A\lambda(x)$. From the construction of A and D it follows that $x - \bar{x} = \delta(w-y)/2$ for $\delta = \|\Delta^\star_{N-1} - \lambda(x)\|_1$ and for some $w = Az \in \mathsf{conv}(S)$ and $y \in X$. Then

$$\begin{aligned}
\left\langle \nabla f(x), x - \bar{x}\right\rangle &= \frac{\delta}{2}\left\langle \nabla f(x), w - y\right\rangle \\
&\leq \frac{\delta}{2}\left(\max_{w\in\mathsf{conv}(S)} \left\langle \nabla f(x), w\right\rangle - \min_{y\in X} \left\langle \nabla f(x), y\right\rangle \right) \\
&= \frac{\delta}{2}\left\langle \nabla f(x), a-u\right\rangle.
\end{aligned} \tag{3.38}$$

We next consider each part separately.

(a) For $\mu^\star := \mu^\star_{\tilde{f},\Delta_{N-1},D}$ inequality (3.38) implies

$$\frac{\mu^\star\delta^2}{2} \leq \bar{f} - f(x) + \left\langle \nabla f(x), x-\bar{x}\right\rangle \leq \bar{f} - f(x) + \frac{\delta}{2}\left\langle \nabla f(x), a-u\right\rangle.$$

Rearranging and applying the arithmetic-mean geometric-mean inequality we get

$$\left\langle \nabla f(x), a-u\right\rangle \geq \mu^\star\delta + 2(f(x)-\bar{f}) \geq 2\sqrt{2\mu^\star(f(x)-\bar{f})}.$$

Thus (3.37) implies that (3.33) holds for $\sigma = 2\mu^\star$.

(b) For $\mu^\sharp := \mu^\sharp_{\tilde{f},\Delta_{N-1},D}$ the convexity of f and inequality (3.38) imply

$$\frac{\mu^\sharp\delta^2}{2} \leq f(x) - \bar{f} \leq \left\langle \nabla f(x), x-\bar{x}\right\rangle \leq \frac{\delta}{2}\left\langle \nabla f(x), a-u\right\rangle.$$

Hence

$$\langle \nabla f(x), a - u \rangle \geq \sqrt{2\mu^\sharp (f(x) - \bar{f})}.$$

Again (3.37) implies that (3.33) holds for $\sigma = \mu^\sharp/2$.

(c) We first show $4L_{\tilde{f},\Delta_{N-1},D} \leq C_f^A$. To that end, suppose $\tilde{y}, \tilde{x} \in \Delta_{N-1}$ and $\alpha := 2\|\tilde{y} - \tilde{x}\|_1 > 0$. Then $A\tilde{y} - A\tilde{x} = \alpha A(\tilde{u} - \tilde{w})$ for some $\tilde{u}, \tilde{w} \in \Delta_{N-1}$. For $x := A\tilde{x}, y := A\tilde{y}, u := A\tilde{u}, w := A\tilde{w}$ we have $y = x + \alpha(u - w) \in X$ and thus

$$D_{\tilde{f}}(\tilde{y}, \tilde{x}) = f(y) - f(x) - \alpha \langle \nabla f(x), u - w \rangle \leq \frac{C_f^A}{2}\alpha^2 = \frac{C_f^A}{4} D(\tilde{y}, \tilde{x}).$$

Since this holds for all $\tilde{y}, \tilde{x} \in \Delta_{N-1}$ with $\|\tilde{y} - \tilde{x}\|_1 > 0$ it follows that $L_{\tilde{f},\Delta_{N-1},D} \leq C_f^A/4$.

We next show the reverse inequality $4L_{\tilde{f},\Delta_{N-1},D} \geq C_f^A$ via a similar argument. Suppose $x, u, w \in X$ and $\alpha > 0$ are such that $y := x + \alpha(u - w) \in X$. Then there exist $\tilde{u}, \tilde{w}, \tilde{x}, \tilde{y} \in \Delta_{N-1}$ such that $u = A\tilde{u}, w = A\tilde{w}, x = A\tilde{x}, y = A\tilde{y}$ and $\|\tilde{y} - \tilde{x}\|_1 = \alpha\|\tilde{u} - \tilde{w}\|_1$. Therefore

$$\begin{aligned} f(x + \alpha(u - w)) - f(x) - \alpha \langle \nabla f(x), u - w \rangle &= D_{\tilde{f}}(\tilde{y}, \tilde{x}) \\ &\leq L_{\tilde{f},\Delta_{N-1},D} \cdot D(\tilde{y}, \tilde{x}) = L_{\tilde{f},\Delta_{N-1},D} \cdot \frac{\alpha^2 \|\tilde{u} - \tilde{w}\|_1^2}{2} \leq 4L_{\tilde{f},\Delta_{N-1},D} \cdot \frac{\alpha^2}{2}. \end{aligned}$$

The last step holds because $\|\tilde{u} - \tilde{w}\|_1 \leq 2$ for all $\tilde{u}, \tilde{w} \in \Delta_{N-1}$. Since the above inequality holds for all $x, u, w \in X$ and $\alpha > 0$ such that $y := x + \alpha(u - w) \in X$, it follows that $C_f^A \leq 4L_{\tilde{f},\Delta_{N-1},D}$.

The identity $4L_{\tilde{f},\Delta_{N-1},D} = C_f^A$ and (3.35) readily imply that for $\mathcal{L} = 4L_{\tilde{f},\Delta_{N-1},D}$ at iteration k of Algorithm 9 we have

$$f(x_k + \alpha v) \leq f(x_k) + \alpha \langle \nabla f(x_k), v \rangle + \frac{\mathcal{L}\alpha^2}{2} \quad \text{for all} \quad \alpha \in [0, \alpha_{\max}].$$

Thus (3.34) holds for $\mathcal{L} = 4L$ and α_k chosen as in (3.36).

□

3.6 Proof of Proposition 10

We consider separately the three possible cases that can occur at iteration k, namely $\alpha_k < \alpha_{\max}$, $\alpha_k = \alpha_{\max} \geq 1$, and $\alpha_k = \alpha_{\max} < 1$.

Case 1: $\alpha_k < \alpha_{\max}$. In this case $|S(x_{k+1})| \leq |S(x_k)| + 1$. In addition, inequalities (3.33) and (3.34) imply that

$$f(x_{k+1}) - f(x_k) \leq -\frac{\langle \nabla f(x_k), v \rangle^2}{2\mathcal{L}} \leq -\frac{\sigma}{2\mathcal{L}}(f(x_k) - \bar{f}). \tag{3.39}$$

Case 2: $\alpha_k = \alpha_{\max} \geq 1$. In this case $|S(x_{k+1})| \leq |S(x_k)|$. In addition, inequality (3.34), the choice of v, and the convexity of f imply that

$$f(x_{k+1}) - f(x_k) \leq \frac{1}{2} \langle \nabla f(x_k), v \rangle \leq \frac{1}{2} \langle \nabla f(x_k), \bar{x}_k - x_k \rangle \leq -\frac{1}{2}(f(x_k) - \bar{f}). \quad (3.40)$$

Case 3: $\alpha_k = \alpha_{\max} < 1$. In this case $|S(x_{k+1})| \leq |S(x_k)| - 1$. In addition, (3.34) implies that

$$f(x_{k+1}) - f(x_k) \leq 0.$$

We next show that in the first k iterations Case 3 can occur at most $k/2$ times by using the argument introduced by Lacoste-Julien and Jaggi in [42]. Since $|S(x_0)| = 1$ and $|S(x_i)| \geq 1$ for $i = 1, 2, \ldots$, it follows that for each iteration when Case 3 occurred there must have been at least one previous iteration when Case 1 occurred. Hence in the first k iterations Case 3 could occur at most $k/2$ times.

To finish the proof, observe that at every iteration k when Case 1 or Case 2 occur inequalities (3.39) and (3.40) yield

$$f(x_{k+1}) - \bar{f} = f(x_k) - \bar{f} + f(x_{k+1}) - f(x_k) \leq \left(1 - \min\left\{\frac{1}{2}, \frac{\sigma}{2\mathcal{L}}\right\}\right)(f(x_k) - \bar{f}).$$

We note that the minimum in the last expression is necessary because $\sigma/\mathcal{L} > 1$ may indeed occur. For a concrete example, see [60, Example 6].

Chapter 4

Rescaling: Enhanced Basic Procedures for the Projection and Rescaling Algorithm

4.1 Introduction

Peña and Soheili [61] propose a two-step *projection and rescaling algorithm*, which extends an algorithm by Chubanov [20], to solve the conic feasibility problem

$$\text{Find } x \in L \cap \mathbb{R}^n_{++} \tag{4.1}$$

where L subspace of $\mathbb{R}^n$ [20, 61]. Assuming the projection matrix, P_L, for L is available we can rewrite (4.1) as

$$\text{Find } x \text{ such that } P_L x > 0. \tag{4.2}$$

The projection and rescaling algorithm consists of two subprocedures:

1. *Basic Procedure (Projection)*: This procedure uses P_L to find a point in $L \cap \mathbb{R}^n_{++}$ provided this cone contains a deeply interior point. This is implemented via one of four schemes based on the Von Neumann/Perceptron algorithm.

2. *Rescaling*: Using the final iterate from the basic procedure, this step rescales $L \cap \mathbb{R}^n_{++}$ such that its interior points - provided $L \cap \mathbb{R}^n_{++} \neq \emptyset$ - become "deeper" in the interior of $\mathbb{R}^n_+$.

In this chapter, we propose enhancements to three of the four Von Neumann/Perceptron basic procedures in [61]. The key feature of these three procedures that allows for our enhancements is that the number of non-zero entries in x grows by at most one in each iteration. Consequently, our enhancements can efficiently and iteratively apply a technique used to prove Carathéodory's Theorem. It remains an interesting open question how fast the Smooth Perceptron, the fourth basic procedure in [61], grows the number of non-zero entries. These enhanced procedures improve the complexity of

the basic procedure from $O(n^4m)$ to $O(n^2m^3)$ operations when L has dimension m: a significant improvement when $m << n$.

Fundamentally, the basic procedure of Peña and Soheili adapts the Von Neumann and Perceptron procedures to iteratively reduce $\|P_Lx\|_2$ on the standard n dimensional simplex $\Delta_{n-1} := \{x \in \mathbb{R}^n : \|x\|_1 = 1, x \geq 0\}$ until either $P_Lx > 0$ or $\|P_Lx\|_2 \leq \frac{1}{3\sqrt{n}}\|x\|_\infty$. Thus, the basic procedure intends to approximately solve the subproblem

$$\min_{x\in\Delta_{n-1}} \|P_Lx\|_2^2. \tag{4.3}$$

The convergence proofs in [61] depends on the observation that $\|x\|_\infty \geq \frac{1}{n}$ for all $x \in \Delta_{n-1}$. As such, the reasoning in [61] yields a faster rate provided we restrict the iterates of the Von Neumann/Perceptron schemes to proper faces of Δ_{n-1}. If $Q \in \mathbb{R}^{n\times m}$ is an orthonormal basis for L then $P_L = QQ^T$ and we may rephrase (4.3) as

$$\min_{x\in\Delta_{n-1}} \|Q^Tx\|_2^2 = \min_{z\in\mathsf{conv}(Q^T)} \|z\|_2^2 \tag{4.4}$$

where $\mathsf{conv}(Q^T)$ is the convex hull of the columns of Q^T. By Carathéodory's Theorem, any point in $\mathsf{conv}(Q^T)$ can be written as convex combination of at most $m+1$ columns of Q^T. Our proposed enhancements apply this observation to ensure that each of the iterates is a convex combinations of no more than $m + 1$ columns of Q^T.

Our enhanced basic procedures iteratively reduce the objective (4.4) using a Von Neumann/Perceptron scheme which applies a modified version of the Incremental Representation Reduction (IRR) procedure of [6] at each iteration. When provided a point $z \in \mathsf{conv}(Q^T)$, the IRR outputs a new affinely independent, convex representation of the point z provided it already contains a sufficiently large set of affinely independent vectors in its support. Whereas the IRR operates in $O(m^3)$ time, our version operates in $O(m^2)$ time by allowing for vectors in the representation of x that have zero support. We call this new version, the Modified Incremental Representation Reduction procedure (mIRR).

This chapter is organized as follows. Section 2 describes the mIRR, and proves important properties of it including its $O(m^2)$ complexity. Section 3 describes the limited support Von Neumann and Perceptron algorithms.

4.2 Modified Incremental Representation Reduction

The heart of our improved basic procedure is a modified version of the Incremental Representation Reduction Procedure of [6]. This subprocedure iteratively applies the main technique used in standard proofs of Carathéodory's Theorem [36]. To simplify notation, given a matrix $A \in \mathbb{R}^{m\times n}$ we define $\tilde{A}$ as the augmented matrix $\begin{bmatrix} 1\dots 1 \\ A \end{bmatrix}$. If $B = [B(1), ..., B(k)] \subseteq \{1, ..., n\}$ is an ordered set of indices and $x \in \mathbb{R}^n$ we let

$A_B = [A_{B(1)}...A_{B(k)}] \in \mathbb{R}^{m\times k}$ where $A_{B(i)}$ denotes the $B(i)$-th column of A and $x_B = (x_{B(1)}...x_{B(k)}) \in \mathbb{R}^k$ and $x_{B(i)}$ denotes the $B(i)$-th entry of x. Similarly, we regard $B^c = [B^c(1), ..., B^c(n-k)] = \{1, ..., n\} \setminus B$ as an ordered set of indices and define $x_{B^c} = (x_{B^c(1)}...x_{B^c(n-k)}) \in \mathbb{R}^{n-k}$ and $A_{B^c} = [A_{B^c(1)}...A_{B^c(n-k)}] \in \mathbb{R}^{m\times(n-k)}$. Given a full column rank matrix $M \in \mathbb{R}^{k\times l}$, we let $M^\dagger$ denote its unique pseudoinverse, $(M^TM)^{-1}M^T$. $I_{k\times k}$ will denote the identity matrix of dimensions $k \times k$ and $0_{k\times\ell}$ will denote the zero matrix of dimensions $k \times \ell$. For $i \in \{1, .., n\}$, let e_i denote the i-th coordinate vector. This notation strongly mimics the notation used in [13] to present the revised Simplex method. The resemblance is entirely intentional; our method strongly resembles the revised Simplex method.

Theorem 12. *Suppose $B \subseteq \{1, ..., n\}$ is an ordered set of indices such that A_B consists of affinely independent columns and $\tilde{A}_B^\dagger$ is known. If $z = Ax = A_Bx_B + A_jx_j$ for some $x \in \Delta_{n-1}$ and $j \in \{1, ..., n\}\backslash B$ then we can find $x^+ \in \Delta_{n-1}$, an ordered set of indices $B^+ \subseteq B' := \begin{bmatrix} B & j \end{bmatrix}$, and $\tilde{A}_{B^+}^\dagger$ such that $z = A_{B^+}x_{B^+}^+$ and A_{B^+} consists of affinely independent columns in $O(m^2)$ operations.*

Proof. There are two cases we must tackle:

1. $\tilde{A}_j \neq \tilde{A}_B\tilde{A}_B^\dagger\tilde{A}_j$: A_j is affinely *independent* of the columns of A_B, i.e. the matrix $\begin{bmatrix} A_B & A_j \end{bmatrix}$ has affinely independent columns.

2. $\tilde{A}_j = \tilde{A}_B\tilde{A}_B^\dagger\tilde{A}_j$: A_j is affinely *dependent* on the columns of A_B, i.e. the matrix $\begin{bmatrix} A_B & A_j \end{bmatrix}$ has affinely dependent columns.

Determining the equality of $\tilde{A}_j$ and $\tilde{A}_B\tilde{A}_B^\dagger\tilde{A}_j$ requires vector-matrix multiplication, an $O(m^2)$ operation. To simplify notation, we let $k := |B|$.

Case 1 ($\tilde{A}_j \neq \tilde{A}_B\tilde{A}_B^\dagger\tilde{A}_j$): In this case, let $B^+ = B'$ and $x^+ = x$. We claim that $\tilde{A}_{B^+}^\dagger$ is given by the formula

$$\tilde{A}_{B^+}^\dagger = \begin{bmatrix} \tilde{A}_B^\dagger \\ 0_{1\times(m+1)} \end{bmatrix} - \begin{bmatrix} \tilde{A}_B^\dagger\tilde{A}_j \\ -1 \end{bmatrix} \frac{\tilde{A}_j^T(I_{(m+1)\times(m+1)} - \tilde{A}_B\tilde{A}_B^\dagger)}{\tilde{A}_j^T(I_{(m+1)\times(m+1)} - \tilde{A}_B\tilde{A}_B^\dagger)\tilde{A}_j}. \tag{4.5}$$

This may be seen as an application of the Sherman-Morrison formula for the rank-one update of a matrix inverse. The quantity

$$\tilde{A}_j^T(I_{(m+1)\times(m+1)} - \tilde{A}_B\tilde{A}_B^\dagger)\tilde{A}_j = \|\tilde{A}_j - \tilde{A}_B\tilde{A}_B^\dagger\tilde{A}_j\|^2$$

is non-zero by hypothesis and thus the expression on the right hand side of (4.5) is well defined. It suffices to verify that right multiplication of the right hand side of (4.5) by $\tilde{A}_{B^+} = \begin{bmatrix} \tilde{A}_B & \tilde{A}_j \end{bmatrix}$ yields the identity matrix. We compute

$$\begin{bmatrix} \tilde{A}_B^\dagger \\ 0_{1\times(m+1)} \end{bmatrix} \begin{bmatrix} \tilde{A}_B & \tilde{A}_j \end{bmatrix} = \begin{bmatrix} \tilde{A}_B^\dagger\tilde{A}_B & \tilde{A}_B^\dagger\tilde{A}_j \\ 0_{1\times k} & 0 \end{bmatrix} = \begin{bmatrix} I_{k\times k} & \tilde{A}_B^\dagger\tilde{A}_j \\ 0_{1\times k} & 0 \end{bmatrix} \tag{4.6}$$

$$\tilde{A}_j^T(I_{(m+1)\times(m+1)} - \tilde{A}_B\tilde{A}_B^\dagger) \begin{bmatrix} \tilde{A}_B & \tilde{A}_j \end{bmatrix} = \tilde{A}_j^T(I_{(m+1)\times(m+1)} - \tilde{A}_B\tilde{A}_B^\dagger)\tilde{A}_j \begin{bmatrix} 0_{1\times k} & 1 \end{bmatrix}. \tag{4.7}$$

Observe that the right hand side of (4.7) is non-zero since $\tilde{A}_j \neq \tilde{A}_B \tilde{A}_B^\dagger \tilde{A}_j$. Combining equations (4.6) and (4.7) yields

$$\left(\begin{bmatrix} \tilde{A}_B^\dagger \\ 0_{1\times(m+1)} \end{bmatrix} - \begin{bmatrix} \tilde{A}_B^\dagger \tilde{A}_j \\ -1 \end{bmatrix} \frac{\tilde{A}_j^T (I_{(m+1)\times(m+1)} - \tilde{A}_B \tilde{A}_B^\dagger)}{\tilde{A}_j^T (I_{(m+1)\times(m+1)} - \tilde{A}_B \tilde{A}_B^\dagger) \tilde{A}_j} \right) \begin{bmatrix} \tilde{A}_B & \tilde{A}_j \end{bmatrix}$$
$$= \begin{bmatrix} I_{k\times k} & \tilde{A}_B^\dagger \tilde{A}_j \\ 0_{1\times k} & 0 \end{bmatrix} - \begin{bmatrix} 0 & \tilde{A}_B^\dagger \tilde{A}_j \\ 0_{1\times k} & -1 \end{bmatrix} = I_{(k+1)\times(k+1)} \quad (4.8)$$

thus verifying our formula for $\tilde{A}_{B^+}^\dagger$. The formula (4.5) uses matrix addition and vector-matrix multiplication so it takes at most $O(m^2)$ operations.

Case 2 ($\tilde{A}_j = \tilde{A}_B \tilde{A}_B^\dagger \tilde{A}_j$): Let $u = \begin{bmatrix} \tilde{A}_B^\dagger \tilde{A}_j \\ -1 \end{bmatrix}$, $\theta^* = \min_{i:u_i<0} \left(-\frac{x_{B'(i)}}{u_i} \right)$, $x_{B'}^+ = x_{B'} + \theta^* u$, and $x_{(B')^c}^+ = 0$. We must show that $x^+ \in \Delta_{n-1}$. By hypothesis, u is the solution to the system

$$\begin{bmatrix} \tilde{A}_B & \tilde{A}_j \end{bmatrix} u = 0_{(m+1)\times 1}$$

because $\tilde{A}_j = \tilde{A}_B \tilde{A}_B^\dagger \tilde{A}_j$. Hence, $\sum_{i=1}^{k+1} u_i = 0$ which implies

$$\sum_{i=1}^{k+1} x_{B'(i)}^+ = \sum_{i=1}^{k+1} x_{B'(i)} = 1.$$

Moreover, the definition of θ^* ensures $x^+ \geq 0$ completing our proof that $x^+ \in \Delta_{n-1}$.

Next, we construct B^+ and $\tilde{A}_{B^+}^\dagger$. Let i^* denote the smallest index such that $-\frac{x_{B'(i^*)}}{u_{i^*}} = \theta^*$ and $u_{i^*} < 0$. By construction, θ^* ensures $x_{B'(i^*)}^+ = 0$. We now have two subcases: $B(i^*) = j$ and $B(i^*) \neq j$. In the first case, let $B^+ = B$ and $\tilde{A}_{B^+}^\dagger = \tilde{A}_B^\dagger$. By hypothesis, $A_{B^+} = A_B$ consists of affinely independent columns. In the second case, let $B^+(i) = B(i)$ for $i \neq i^*$ and $B^+(i^*) = j$. We must show that A_{B^+} consists of affinely independent columns. Assume for the sake of contradiction that it does not. Then there must exist some $w \in \mathbb{R}^{k+1}$ such that $w_{i^*} = 0$ and

$$0_{(m+1)\times 1} = \tilde{A}_{B^+} w - \tilde{A}_j = \tilde{A}_B w - \tilde{A}_j$$

but this implies that

$$\tilde{A}_B \left(w - \tilde{A}_B^\dagger \tilde{A}_j \right) = \tilde{A}_j - \tilde{A}_j = 0_{(m+1)\times 1}.$$

since we assume $\tilde{A}_j = \tilde{A}_B \tilde{A}_B^\dagger \tilde{A}_j$. By affine independence of the columns of A_B we determine that

$$w - \tilde{A}_B^\dagger \tilde{A}_j = 0_{(m+1)\times 1} \Leftrightarrow w = \tilde{A}_B^\dagger \tilde{A}_j$$

so that the i^*-th entry of $\tilde{A}_B^\dagger \tilde{A}_j$, which is precisely u_{i^*}, is zero: a contradiction. Thus, the columns of A_{B^+} are affinely independent. Finally, we prove that it is possible

to derive A_{B^+} in $O(m^2)$ operations in this second case. Form the augmented matrix $\left[\begin{array}{c|c} \tilde{A}_B^\dagger & \tilde{A}_B^\dagger \tilde{A}_j \end{array}\right]$. Add to each row a multiple of the i^*-th row to make the last column equal to the coordinate vector e_{i^*}. The first $|B|$ columns of the resultant matrix are $\tilde{A}_{B^+}^\dagger$. This requires no more than $O(m^2)$ operations since at most m row operations are required. □

The proof of this theorem immediately yields our core algorithm as well as an easy corollary.

Algorithm 10 Modified Incremental Representation Reduction Procedure (mIRR)

1: Input: An ordered set of indices $B = [B(1), ..., B(k)] \subseteq \{1, ..., n\}$ such that A_B is a matrix with affinely independent columns, $\tilde{A}_B^\dagger$, $j \notin B$, and $z = Ax = A_B x_B + A_j x_j$ for some $x \in \Delta_{n-1}$ with $x_{[B,j]^c} = 0$.

2: Compute $u' = \tilde{A}_B^\dagger \tilde{A}_j$ and $u = \begin{bmatrix}(u')^T & -1\end{bmatrix}$. If $\tilde{A}_B u' \neq \tilde{A}_j$ then the columns of $\begin{bmatrix}\tilde{A}_B & \tilde{A}_j\end{bmatrix}$ are affinely independent. In this case, output $x^+ = x$, $B^+ = B \cup \{j\}$, and

$$\tilde{A}_{B^+}^\dagger = \begin{bmatrix} \tilde{A}_B^\dagger \\ 0_{1\times(m+1)} \end{bmatrix} - \begin{bmatrix} \tilde{A}_B^\dagger \tilde{A}_j \\ 1 \end{bmatrix} \frac{\tilde{A}_j^T (I_{(m+1)\times(m+1)} - \tilde{A}_B \tilde{A}_B^\dagger)}{\tilde{A}_j^T (I_{(m+1)\times(m+1)} - \tilde{A}_B \tilde{A}_B^\dagger) \tilde{A}_j},$$

to complete the procedure. Otherwise, proceed to the next step.

3: Let $\theta^* = \min_{i:u_i<0} \left(-\frac{x_i}{u_i}\right)$, i^* be the smallest index for which θ^* is achieved, and

$$x^+_{[B,j]} = x_{[B,j]} + \theta^* u, \quad x^+_{[B,j]^c} = 0_{(n-(k+1))\times 1}.$$

If $i^* = j$ then output x^+, $B^+ = B$, and $\tilde{A}_{B^+} = \tilde{A}_B$ to complete the procedure. Otherwise, proceed to the next step.

4: Let $B^+ = [B(1), ..., B(i^*-1), j, B(i^*), ..., B(k)]$. Form the $|B| \times (|B|+1)$ matrix $\begin{bmatrix}\tilde{A}_B^\dagger & u'\end{bmatrix}$. Add to each row a multiple of the i^*-th row to make the last column equal to the unit vector e_{i^*}. The first $|B|$ columns of the resultant matrix are $\tilde{A}_{B^+}^\dagger$. Output B^+, x^+ and $\tilde{A}_{B^+}^\dagger$.

Corollary 7. *The mIRR produces an affinely independent representation, $z = Ax = A_{B^+} x_{B^+}$ with $x \in \Delta_{n-1}$ and $x_{(B^+)^c} = 0_{(n-|B^+|)\times 1}$, of the input point $z = Ax$ and the pseudoinverse $\tilde{A}_{B^+}^\dagger$ in $O(m^2)$ operations.*

4.3 Limited Support Basic Procedures

In this section we propose each of our modified basic procedures. Recall that we assume the availability of an orthonormal basis for L and that $Q \in \mathbb{R}^{n\times m}$ is the matrix whose columns are these basis vectors. The convergence results of [61] for the original basic procedures depend upon the maximum size of the support of the iterates. If it were

possible to ensure that the iterates maintained affinely independent support then the support would have maximum size $m+1$. This is the crux of our enhanced procedures and the mIRR enables us to do this. Our enhanced procedures start with a single column of the matrix Q^T. Then, until the stopping condition is reached, they take a Von-Neumann/Perceptron-like step - which may increase the size of the support by at most one - followed by an application of mIRR to ensure the support remains affinely independent. We will let $\{q_i\}_1^n$ denote the columns of Q^T and we use P in place of P_L. Given $z \in \mathbb{R}^\ell$ for some $\ell \in \mathbb{N}$, we let $z^+ = (\max\{z_1, 0\}, ..., \max\{z_\ell, 0\})$.

The first two schemes, the Limited Support Von Neumann and Limited Support Perceptron, are subsumed in the following framework which we call the *Limited Support Scheme* (LSS). Each of these procedures is an enhancement of those found in [61] using mIRR. Our modified schemes exchange the roles played by x_t and z_t in [61][Algorithm 5] to accord with our notation in the mIRR. Namely, we want to ensure each x_t is an element of the simplex.

Algorithm 11 Limited Support Scheme

$x_0 = e_1$, $z_0 = Q^T x_0 = q_1$, $B_0 = \{1\}$, $\tilde{Q}^\dagger_{B_0} = \frac{1}{\|\tilde{q}_1\|}\tilde{q}_1^T$, $t = 0$
while $Px_t \not> 0$ and $\|(Px_t)^+\| \geq \frac{1}{3\sqrt{n}}\|x_t\|_\infty$ **do**
  Let $j = \text{argmin}_{i\in[n]} \langle q_i, z_t\rangle$
  $x'_{t+1} = x_t + \theta_t(e_j - x_t)$
  $z_{t+1} = Q^T x'_{t+1}$
  if $j \notin B_t$ **then**
    $(x_{t+1}, B_{t+1}, \tilde{Q}^\dagger_{B_{t+1}}) = mIRR(B_t, \tilde{Q}^\dagger_{B_t}, j, x'_{t+1}, Q_j)$
  else
    $x_{t+1} = x'_{t+1}$, $B_{t+1} = B_t$, $\tilde{Q}^\dagger_{B_{t+1}} = \tilde{Q}^\dagger_{B_t}$
  end if
  $t = t + 1$
end while

If $\theta_t = \frac{1}{t+1}$ then the resulting procedure is referred to as the *Limited Support Perceptron Scheme* (LSP). If θ_t is determine by an exact line search then the resulting procedure is referred to as the *Limited Support Von Neumann Scheme* (LSVN).

Proposition 12. *The following hold for algorithms LSP and LSVN:*

1. *For all $t \geq 0$ such that LSP or LSVN have not halted, $\|z_t\|^2 \leq \frac{1}{t}$.*
2. *The stopping condition, $Px_t > 0$ or $\|(Px_t)^+\| \leq \frac{\|x_t\|_\infty}{3\sqrt{n}}$, occurs in at most $9(m+1)^2 n$ iterations.*
3. *LSP and LSVN require $O(m^3n^2)$ arithmetic operations.*

Proof.

1. This part of our proposition is known from [61].

2. If z_t has affinely independent columns throughout the algorithm then

$$|\{i \in \{1, ..., n\} : x_i > 0\}| \leq |B_t| \leq m + 1.$$

In this case, $\|x_t\|_\infty \geq \frac{1}{m+1}$ since $x_t \in \Delta_m$. This implies that $\frac{1}{3\sqrt{n}}\|x\|_\infty \geq \frac{1}{3(m+1)\sqrt{n}}$. As $\|(Px)^+\| \leq \|Px\|$, we conclude from 1 that the one of the two stopping conditions occurs by $t = 9(m + 1)^2 n$.

We proceed by induction to show that z_t has affinely independent columns in its support and concurrently that LSS maintains the pseudoinverse of the matrix formed by the augmented columns in z_t's support. This is readily seen to be true for z_0 and Q_{B_0}. Suppose z_t has affinely independent columns and $\tilde{Q}^\dagger_{B_t}$ is available. Then x'_{t+1} has at most one additional non-zero entry x_j. Indeed, as we highlighted in the introduction, this is they key feature that permits our application of the mIRR. By corollary 7, the mIRR will generate x_{t+1} and $\tilde{Q}^\dagger_{B_{t+1}}$ such that $z_{t+1} = Q^T x_{t+1}$ and the non-zero entries of x_{t+1}, which are given by B_{t+1}, correspond to affinely independent columns of Q^T.

3. By part 1, LSP and LSVN terminate in at most $t = 9(m + 1)^2 n$ main iterations. Each of the operations besides mIRR requires at most nm operations while corollary 7 states mIRR requires $O(m^2)$ operations. Since $m^2 \leq nm$, we conclude each iteration has computational cost $O(nm)$. Thus, the number of required arithmetic operations for LSP and LSVN is $O(m^3 n^2)$.

□

4.3.1 Limited Support Von Neumann with Away Steps Scheme

Here we propose a limited support variation of the Von Neumann with Away Steps scheme proposed in [61]. This procedure is essentially the same as above except that it allows for "away" directions.

Algorithm 12 Limited Support Von Neumann with Away Steps Scheme (LSVN)

$x_0 = e_1$, $z_0 = Q^T x_0 = q_1$, $B_0 = \{1\}$, $\tilde{Q}^{\dagger}_{B_0} = \frac{1}{\|\tilde{q}_1\|}\tilde{q}_1^T$, $t = 0$
while $Px_t \not> 0$ and $\|(Px_t)^+\| \geq \frac{1}{3\sqrt{n}}\|x_t\|_\infty$ **do**
  Let $j = \text{argmin}_{i\in[n]} \langle q_i, z_t\rangle$,$k = \text{argmax}_{i\in[n]} \langle q_i, z_t\rangle$
  if $\|z_t\|^2 - \langle q_j, z_t\rangle > \langle q_k, z_t\rangle - \|z_t\|^2$ **then**
    (Regular Step) $a := e_j - x_t; \theta_{max} = 1$
  else
    (Away Step) $a := x_t - e_k; \theta_{max} = \frac{(x_t)_j}{1-(x_t)_j}$
  end if
  $\theta_t = \text{argmin}_{\theta\in[0,\theta_{max}]} \|P(x_t + \theta a)\|^2 = \min\left\{\theta_{max}, -\frac{\langle x_t, Pa\rangle}{\|Pa\|^2}\right\}$
  $x'_{t+1} = x_t + \theta_t a$
  $z_{t+1} = Q^T x'_{t+1}$
  if $j \notin B_t$ and a regular step is taken **then**
    $(x_{t+1}, B_{t+1}, \tilde{Q}^{\dagger}_{B_{t+1}}) = mIRR(B_t, \tilde{Q}^{\dagger}_{B_t}, j, x'_{t+1}, Q_j)$
  else
    $x_{t+1} = x'_{t+1}$, $B_{t+1} = B_t$, $\tilde{Q}^{\dagger}_{B_{t+1}} = \tilde{Q}^{\dagger}_{B_t}$
  end if
  $t = t + 1$
end while

Proposition 13. *The following hold for algorithm LSVNA:*

1. *For all $t \geq 0$ such that LSVNA has not halted, $\|z_t\|^2 \leq \frac{1}{t}$.*
2. *The stopping condition, $Px_t > 0$ or $\|(Px_t)^+\| \leq \frac{\|x_t\|_\infty}{3\sqrt{n}}$, occurs in at most $9(m+1)^2 n$ iterations.*
3. *LSVNA requires $O(m^3n^2)$ arithmetic operations.*

Proof. The proof of the first part of the proposition is known from [61]. The remaining parts follow from similar reasoning to that in the proof of proposition 12. □

Conclusion

Summary

In this section, we summarize our findings in the four chapters that comprise this thesis.

- In chapter 1, we applied the analysis of [59] to the class of proximal gradient methods. This method yielded known, modern convergence rates for the proximal gradient, accelerated proximal gradient, and proximal subgradient methods under assumptions weaker than those found in the literature. This chapter is based on our published paper [33].

- In chapter 2, we extended the analysis of chapter 1 and [59] to the class of Bregman proximal first-order methods, a class of first-order methods that is both more general and flexible than proximal gradient methods. Again, we derived convergence rates for this class of methods under weaker than previously employed conditions. This chapter is based on our paper [31] which is currently under review.

- In chapter 3, we proposed condition numbers for a differentiable convex function relative to a domain and a distance-like function on the domain. We demonstrated that these condition numbers naturally arise in the convergence analyses of first-order methods and that they retain much of the geometric flavor of the standard condition number of a convex function. This chapter is based on our paper [32] which is currently under review.

- In chapter 4, we proposed three enhanced versions of the basic procedures for the Projection and Rescaling Algorithm of [61]. These enhancements yield a substantially improved convergence rate when the subspace L has dimension sufficiently smaller than that of the ambient Euclidean space. This chapter is based on our published paper [30].

Extensions and Future Work

Acceleration

An alternative and intriguing framework for explaining acceleration examines first-order methods as discretizations of ordinary differential equations. Prominent examples of this approach include [68, 41]. However, these techniques have yet to be applied to derive the convergence rates of the primal gradient schemes of [48] or their accelerated variants in [34]. A future interest is the analysis of these methods using a differential equations-based framework.

Conditioning and Frank-Wolfe

Seemingly all of the research on the Frank-Wolfe algorithm assumes that the feasible set is bounded. For polyhedrally constrained convex optimization problems, such as the non-negative least squares problem

$$\min_{x \in \mathbb{R}^n_+} \frac{1}{2} \|Ax - b\|_2^2$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, this condition is not satisfied. The compactness hypothesis precludes Frank-Wolfe's defining linear subproblem from being unbounded and thus producing undefined search directions. However, for unbounded linear problems some well-known solvers such as the Simplex return an extreme ray along which the objective descends to $-\infty$. This hints at one possible path for extending the Frank-Wolfe algorithm to unbounded domains. This potential extension of the Frank-Wolfe algorithm is one of our topics of interest.

Convergence Analysis with Approximately Computable Bregman Proximal Maps

The literature on Bregman proximal gradient methods assumes the computability of the Bregman proximal map

$$(x, g) \mapsto \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ \langle g, y \rangle + \Psi(y) + L D_h(y, x) \right\}. \tag{4.9}$$

Yet in practice this map may only be approximately computable. Thus, in light of the practical relevance of approximate computablity of this map, one direction of interest is the extension of our convergence analysis to this case.

Bibliography

[1] Z. Allen-Zhu and L. Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[2] C. Atwood. Optimal and efficient designs of experiments. *The Annals of Mathematical Statistics*, pages 1570–1602, 1969.

[3] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[4] H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[5] H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[6] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1):1–27, 2017.

[7] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[8] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[9] S. Becker, E. Candès, and M. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011.

[10] J. Bello-Cruz. On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions. *Set-Valued and Variational Analysis*, 25(2):245–263, 2017.

[11] J. Bello-Cruz and T. Nghia. On the convergence of the forward–backward splitting method with line-searches. *Optimization Methods and Software*, 31(6):1209–1238, 2016.

[12] Alexandre Belloni, Robert M Freund, and Santosh Vempala. An efficient rescaled perceptron algorithm for conic systems. *Mathematics of Operations Research*, 34(3):621–641, 2009.

[13] D. Bertsimas and J. Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[14] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization*. Springer, New York, 2000.

[15] G. Braun, S. Pokutta, D. Tu, and S. Wright. Blended conditional gradients: the unconditioning of conditional gradients. *arXiv preprint arXiv:1805.07311*, 2018.

[16] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[17] S. Bubeck, Y. Lee, and M. Singh. A geometric alternative to Nesterov's accelerated gradient descent. *arXiv preprint arXiv:1506.08187*, 2015.

[18] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[19] D. Cheung and F. Cucker. A new condition number for linear programming. *Mathematical Programming*, 91(2):163–174, 2001.

[20] S. Chubanov. A polynomial projection algorithm for linear feasibility problems. *Mathematical Programming*, 153(2):687–713, 2015.

[21] A. L. Dontchev, A. S. Lewis, and R. T. Rockafellar. The radius of metric regularity. *Trans. Amer. Math. Soc.*, 355(2):493–517 (electronic), 2003.

[22] D. Drusvyatskiy, M. Fazel, and S. Roy. An optimal first order method based on optimal quadratic averaging. *SIAM Journal on Optimization*, 28(1):251–271, 2018.

[23] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

[24] M. Epelman and R. Freund. A new condition measure, preconditioners, and relations between different measures of conditioning for conic linear systems. *SIAM Journal on Optimization*, 12(3):627–655, 2002.

[25] M. Epelman and R. M. Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Mathematical Programming*, 88(3):451–485, 2000.

[26] N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *COLT*, pages 658–695, 2015.

[27] R. Freund. Complexity of convex optimization using geometry-based measures and a reference point. *Mathematical Programming*, 99(2):197–221, 2004.

[28] R. Freund and J. Vera. Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. *SIAM Journal on Optimization*, 10(1):155–176, 1999.

[29] B. Grimmer. Convergence rates for deterministic and stochastic subgradient methods without Lipschitz continuity. *arXiv preprint arXiv:1712.04104*, 2017.

[30] D. Gutman. Enhanced basic procedures for the projection and rescaling algorithm. *To Appear in Optimization Letters*, 2018.

[31] D. Gutman and J. Peña. A unified framework for Bregman proximal methods: subgradient, gradient, and accelerated gradient schemes. *arXiv preprint arXiv:1812.10198*, 2018.

[32] D. Gutman and J. Peña. The condition number of a function relative to a set. *arXiv preprint arXiv:1901.08359*, 2019.

[33] D. Gutman and J. Peña. Convergence rates of proximal gradient methods via the convex conjugate. *SIAM Journal on Optimization*, 29(1):162–174, 2019.

[34] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated Bregman proximal gradient methods for relatively smooth convex optimization. *arXiv preprint arXiv:1808.03045*, 2018.

[35] J. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms*. Springer–Verlag, Berlin, 1993.

[36] J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[37] A. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49(4):263–265, 1952.

[38] D. Rodriguez J. Peña and N. Soheili. On the von Neumann and Frank-Wolfe algorithms with away steps. *SIAM Journal on Optimization*, 26(1):499–512, 2016.

[39] S. Karimi and S. Vavasis. A single potential governing convergence of conjugate gradient, accelerated gradient and geometric descent. *arXiv preprint arXiv:1712.09498*, 2017.

[40] T. Kitahara and T. Tsuchiya. An extension of Chubanov's polynomial-time linear programming algorithm to second-order cone programming. *Optimization Methods and Software*, 33(1):1–25, 2018.

[41] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems 28*, 2015.

[42] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

[43] L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[44] A. Lewis. Ill-conditioned convex processes and conic linear systems. *Mathematics of Operations Research*, 24(4):829–834, 1999.

[45] P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[46] B. Lourenço, T. Kitahara, M. Muramatsu, and T. Tsuchiya. An extension of Chubanov's algorithm to symmetric cones. *Mathematical Programming*, 173(1):117–149, 2019.

[47] H. Lu. "Relative-continuity" for non-Lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv preprint arXiv:1710.04718*, 2017.

[48] H. Lu, R. Freund, and Y. Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[49] C. Ma, N. Gudapati, M. Jahani, R. Tappenden, and M. Takáč. Underestimate sequences via quadratic averaging. *arXiv preprint arXiv:1710.03695*, 2017.

[50] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2018.

[51] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley, 1983.

[52] Y. Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. Doklady AN SSSR (in Russian). *(English translation. Soviet Math. Dokl.)*, 269:543–547, 1983.

[53] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Kluwer Academic Publishers, 2004.

[54] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[55] B. O'Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732, 2015.

[56] F. Ordóñez and R. Freund. Computational experience and the explanatory value of condition measures for linear optimization. *SIAM Journal on Optimization*, 14(2):307–333, 2003.

[57] J. Peña, J. Vera, and L. Zuluaga. An algorithm to compute the Hoffman constant of a system of linear constraints. *arXiv preprint arXiv:1804.08418*, 2018.

[58] J. Peña. Understanding the geometry on infeasible perturbations of a conic linear system. *SIAM Journal on Optimization*, 10(2):534–550, 2000.

[59] J. Peña. Convergence of first-order methods via the convex conjugate. *Operations Research Letters*, 45:561–564, 2017.

[60] J. Peña and D. Rodríguez. Polytope conditioning and linear convergence of the Frank-Wolfe algorithm. *Mathematics of Operations Research*, 44(1):1–18, 2018.

[61] J. Peña and N. Soheili. Solving conic systems via projection and rescaling. *Mathematical Programming*, 166(1):87–111, 2017.

[62] A. Ramdas and J. Peña. Towards a deeper geometric, analytic and algorithmic understanding of margins. *Optimization Methods and Software*, 31(2):377–391, 2016.

[63] J. Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM Journal on Optimization*, 5(3):506–524, 1995.

[64] J. Renegar. Linear programming, complexity theory and elementary functional analysis. *Mathematical Programming*, 70(3):279–351, 1995.

[65] T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[66] K. Roos. An improved version of Chubanov's method for solving a homogeneous feasibility problem. *Optimization Methods and Software*, 33(1):26–44, 2018.

[67] V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems*, pages 1119–1129, 2017.

[68] 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*, pages 2510–2518, 2014.

[69] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, pages 1–30, 2018.

[70] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.