SOME EXTENSIONS ON THE REACH OF FIRST-ORDER OPTIMIZATION THEORY

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Benjamin David Grimmer

August 2021

© 2021 Benjamin David Grimmer

ALL RIGHTS RESERVED

SOME EXTENSIONS ON THE REACH OF FIRST-ORDER OPTIMIZATION THEORY

Benjamin David Grimmer, Ph.D.

Cornell University 2021

This thesis concerns the foundations of first-order optimization theory. In recent years, these methods have found tremendous success in a wide range of domains due to their incredibly scalable nature. We aim to extend the reach of first-order optimization guarantees, lessening the gap between practice and theory, as well as enabling the design of new algorithms. We show that many of the typical assumptions in the theory literature are not fundamentally needed. For example, analysis in nonsmooth optimization typically relies on Lipschitz continuity, convexity, and carefully chosen stepsize parameters. Chapters 2-4 show that classic methods can be applied and analyzed without relying on these structures. Then Chapters 5-8 consider reformulations of optimization problems that further extend the reach of first-order methods. Constrained optimization can be reformulated to avoid orthogonal projections, generic optimization problems can be reformulated to possess Hölder growth/error bounds, and minimax optimization problems can be smoothed and convexified/concavified. Together these results challenge the limitations on which problems are historically considered tractable for analysis sake.

BIOGRAPHICAL SKETCH

Benjamin Grimmer was born in Dayton, Ohio, growing up in a myriad of different states. He graduated from the Oklahoma School of Science and Math in 2012, mentored by Professor J. Adrian Zimmer. From there, he completed coterminal undergraduate and Master's degrees in Computer Science at the Illinois Institute of Technology in 2016, with a focus on discrete optimization advised by Professor Gruia Calinescu. Ben then began his doctoral studies in the School of Operations Research & Information Engineering at Cornell University, where his focus changed to continuous optimization advised by Professors James Renegar and Damek Davis.

ACKNOWLEDGEMENTS

Cornell has provided me the opportunity to work with and be around so many incredible people. First, I am overwhelming grateful that I had the opportunity to be advised by Jim Renegar. His patient and caring attention has played a crucial role in my growth as a researcher, teacher, and person. He exemplifies the kind of mathematician and human I hope to be. Thank you, Jim.

I have been further gifted with a second invaluable adviser in Damek Davis. I can only hope to emulate his energy and focus in tackling impactful problems. Thank you, Damek for helping me learn to identify important problems and introducing me to Clarke. I likely cannot express the full extent of my admiration and respect for my advisors, so I hope to at least convey my fullest gratitude.

Cornell has further given me the opportunity to learn from the wide-range of ORIE faculty. I would specifically like to acknowledge Adrian Lewis and Madeleine Udell for their routine advice and insights. I further want to thank a pair of my fellow PhD students, Mateo Diaz and Lijun Ding, for a number of great group discussions and research endeavors. Outside Cornell, Sean Lu has been a fantastic colleague and friend, pointing me toward numerous interesting problems. I would like to specifically acknowledge Rob Freund and Pratik Worah whose expertise and attention I have had the privilege to receive.

Throughout my PhD, I have been supported by a number of great organizations. I am incredibly thankful to the NSF which supported me through the Graduate Research Fellowships Program under grant DGE-1650441, the Simons Institute for the Theory of Computing where I spent the Fall of 2017, and Google Research where I spent the Spring of 2020.

The warm community among Cornell's ORIE has been an essential part of my happiness over the past five years. My PhD cohort has been a phenomenally

supportive group, particularly thank you Amy Zhang, Angela Zhou, Chamsi Hssaine, Lijun Ding, and Woo-Hyung Cho. My seniors Andrew Daw and Sam Gutekunst were invaluable guides and friends as I navigated the PhD program. ORIE also provided me with ample people to learn (board) game theory from: In no particular order, thank you David Eckmann and David Lingenbrink for playing Gloomhaven, thank you Matthew Zalesak, Pamela Badian-Pessot, Alyf Janmohamed, and Sean Sinclair for playing Hanabi, thank you Jamol Pender, Andrew Daw, and numerous passersby for playing Liar's Dice, and thank you Sander Aarts and Varun Suriyanarayana for playing the above and many many other games. I have also had the great fortune to virtually play games with David Applegate, Barb Chubb, Aaron Archer, and Matthew Fahrbach each week throughout the pandemic.

Lastly, I want to thank my friends and family for their continual support. Thank you Dad for teaching me how to think. Thank you Mom for teaching me how to plan. Thank you Daniel Grimmer for continually giving me perspective. Thank you Krishen Blows for keeping my focus towards betterment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This dissertation considers first-order optimization methods, which have undergone a renaissance over the past decade due to their highly scalable nature. Second-order (quasi-)Newton methods and interior point methods can produce high accuracy solutions in only a few iterations, but at a modern scale, may fail to complete any iteration in a reasonable amount of time. This has led first-order optimization methods, which tend to produce modest quality solutions after many relatively cheap iterations, to take center-stage.

Many of the most important current applications in optimization come from machine learning and data science. These problems perfectly fit the mold for first-order methods. The size of datasets has been rapidly growing, easily exceeding many million in sizes: In 1999, the well-worn MNIST dataset, used for learning handwritten digits, was released containing 70,000 testing and training images. In 2009, the famous Netflix Prize for learning to predict user ratings for films concluded where algorithms were given approximately three million sample ratings and had to predict approximately one hundred million ratings. Another decade later, problem sizes have only continued to grow. These huge scales lend themselves to the low iteration cost of first-order optimization, as anything more risks becoming intractable. Stochastic first-order methods can be particularly efficient in this setting by only examining a small, randomly selected portion of the data at each iteration. Often this provides an unbiased estimation of the full objective while reducing the iteration cost from considering millions of data points to only hundreds (or even just a single data point).

Modern learning problems further align with first-order methods since high-

accuracy solutions offer them little benefit and risk statistical problems of overfitting. Hence the fact that first-order methods only offer modest accuracy guarantees (compared to higher-order methods) does not present a major issue to the effectiveness of these methods. Indeed stochastic gradient descent (and its variants) has become *the* foundational method used throughout machine learning. Stochastic subgradient methods form a core numerical subroutine for several widely used solvers, including Google's TensorFlow and the open-source PyTorch library.

The focus of this dissertation is on understanding what structure is fundamentally needed for us to design and deploy first-order optimization algorithms. The practical usage of these methods, especially in the machine learning literature, has shown they are experimentally effective for a wide range of problems. Despite this, the theory for first-order methods primarily applies in more narrow settings where properties like Lipschitz continuity or convexity hold or where operations like orthogonal projections are tractable. We will depart from relying on these traditional notions, which are cornerstones of the classic first-order method optimization theory. The essence of each chapter of this work can be understood by examining and questioning the analysis of perhaps the most fundamental first-order method, the *Subgradient Method*.

1.1 The Subgradient Method

Consider the problem of minimizing a function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ over a set $Q \subseteq \mathbb{R}^n$

$$\min_{x \in Q} f(x). \tag{1.1}$$

In case f is not differentiable, we consider the generalization of its gradient given by its subdifferential $\partial f(x) = \{g \in \mathbb{R}^n \mid f(x') \geq f(x) + \langle g, x' - x\rangle \forall x' \in \mathbb{R}^n\}$, capturing the first-order geometry of f. The elements of this differential are called subgradients. Following the idea of steepest descent, this problem can be solved by repeatedly moving in negative subgradient directions and then projecting back onto the feasible region

$$x_{k+1} = \text{proj}_Q(x_k - \alpha_k g_k) \quad \text{where} \quad g_k \in \partial f(x_k) \tag{1.2}$$

where α_k is some sequence of stepsizes and $\text{proj}_Q(x) = \text{argmin}\{\|x' - x\| \mid x' \in Q\}$ denotes orthogonal projection. The classic analysis of this method [135] bounds its convergence rate provided (i) the objective is uniformly Lipschitz continuous, (ii) the objective and constraints are convex, (iii) the stepsize α_k is carefully selected, and (iv) the subgradients and orthogonal projections can be computed.

Theorem 1.1.1. *For any M-Lipschitz continuous, convex function f and closed convex set Q, the subgradient method with Polyak stepsize $\alpha_k = (f(x_k) - \inf f)/\|g_k\|^2$ has*

$$\min_{k \leq T} f(x_k) - \inf f \leq \frac{M \, \mathbf{dist}(x_0, \text{argmin } f)}{\sqrt{T+1}}.$$

Additionally supposing that (v) the objective grows sufficiently fast near its minimizers (typically assumed in the form of a KL condition or growth/error bound), this rate can be further improved. Nesterov [126] derives matching complexity lower bounds, preventing nearly any first-order method from beating Theorem 1.1.1's rate or its improved rates under growth conditions.

Although these matching upper and lower bounds nicely wrap up this family of nonsmooth optimization problems, they rely on all of the nontrivial structures (i)-(v) above. In the following seven chapters of this dissertation, we develop first-order method machinery that bypasses these classic requirements.

1.1.1 Relaxing Lipschitz Continuity Assumptions Like (i)

In Chapter 2, we consider convergence guarantees for the classic subgradient method (1.2) outside the setting of Lipschitz continuous objective functions. The generalization we propose relies on the existence of a generic upper bound on the objective function of the form

$$f(x) - f(x^*) \leq \mathcal{D}(\|x - x^*\|)$$

for some minimizer x^*. For any such function, we derive a convergence rate guarantee of

$$\min_{k \leq T} f(x_k) - \inf f \leq \mathcal{D}\left(\frac{\mathbf{dist}(x_0, \operatorname{argmin} f)}{\sqrt{T+1}}\right).$$

Considering $\mathcal{D}(t) = Mt$ shows this generalizes M-Lipschitz continuity as $f(x) - f(x^*) \leq M\|x - x^*\|$ and recovers Theorem 1.1.1's rate of

$$\frac{M\mathbf{dist}(x_0, \operatorname{argmin} f)}{\sqrt{T+1}}.$$

An upper bound of this form holds for any problem that is locally Lipschitz around a minimizer, a much weaker requirement than global Lipschitz continuity. As a second example beyond the reach of Theorem 1.1.1, any smooth function will satisfy an upper bound of the form $\mathcal{D}(t) = Lt^2$, from which our theory recovers gradient descent's known smooth convergence rate of $O(L\mathbf{dist}(x_0, \operatorname{argmin} f)^2/T)$. Moreover, the sum of a smooth function and a nonsmooth Lipschitz function will have an upper bound $\mathcal{D}(t) = Lt^2 + Mt$ and consequently has a convergence guarantee given exactly by the sum of the classic smooth and nonsmooth rates when considered separately.

1.1.2 Relaxing Convexity Assumptions Like (ii)

In Chapter 3, we propose and analyze a subgradient method that converges without relying on convexity. The generalization we propose relies on the objective function having

$$x \mapsto f(x) + \frac{\rho}{2}\|x\|^2 \text{ be convex} \tag{1.3}$$

for some $\rho > 0$. For any such function, we introduce an algorithm inspired by the proximal point method [147] that mimics it using only stochastic subgradient evaluations. We show that this method produces nearly stationary points for nonsmooth nonconvex optimization at the same rate that stochastic gradient descent does for smooth nonconvex optimization. For example, this algorithm and our guarantees apply for any $f(x) = g(x) + h(x)$ given by the sum of a nonconvex smooth g and a convex nonsmooth function h. This form occurs in learning problems involving optimizing a smooth but nonconvex model g with a convex regularization term h, such as an ℓ_1 norm to induce sparsity.

1.1.3 Relaxing Stepsize Requirements Like (iii)

In Chapter 4, we address the limitation of the subgradient method in needing a carefully chosen step size (for example, the classic guarantee of Theorem 1.1.1 critically relies on taking $\alpha_k = (f(x_k) - \inf f)/\|g_k\|^2$). To do so, we revisit a classic subgradient method, the proximal bundle method, developed in 1975 [93, 170] that is known to converge in the limit for any configuration of its stepsize. We derive a full characterization of this subgradient method's convergence rate showing that with any configuration of its stepsize, the bundle

method adapts to converge faster in the presence of various continuity, smoothness, and growth conditions. Inspired by Renegar and Grimmer [142], we also propose a parallel bundle method that attains our strongest convergence rates under every combination of assumptions considered.

1.1.4 Removing the Need for Orthogonal Projections Like (iv)

In Chapters 5 and 6, we present an alternative development and generalization of the radial reformulation invented by Renegar [140]. In doing so, we devise a radial duality between nonnegative optimization problems that facilitates the development of new families of projection-free first-order methods applicable even in the presence of nonconvex objectives and constraint sets. Applying the subgradient method to the radially dual optimization problem gives a radial subgradient method that converges without requiring Lipschitz continuity or orthogonal projection, and for appropriate nonconvex problems, avoids the weakened convexity condition (1.3). This machinery further enables radial smoothing and accelerated methods, capable of scaling-up much more efficiently than their classic counterparts.

1.1.5 Growth Bounds Like (v) Hold Without Loss of Generality

In Chapter 7, we consider the improved performance of first-order methods under growth or error bound conditions. The subgradient method's analysis is typically done twice, once for the general case and again for the growth bounded case. We give meta-theorems for deriving general convergence rates

from those assuming a growth lower bound. Applying this simple but conceptually powerful tool to the subgradient method as well as the proximal point method, the proximal bundle method, gradient descent and universal accelerated methods immediately recovers their known convergence rates for general convex optimization problems from their specialized rates. Our results apply to lift any rate based on Hölder continuity of the objective's gradient and Hölder growth bounds to apply to any problem with a weaker growth bound or when no growth bound is assumed.

1.1.6 Characterizing the Landscape of the Proximal Point Method for Nonconvex-Nonconcave Optimization

In Chapter 8, we deviate from the minimization model (1.1) and instead consider minimax optimization problems of the form

$$\min_x \max_y L(x, y).$$

We consider the direct generalization of the subgradient method to this setting, descending in x and ascending in y. When $L(x, y)$ is convex in x and concave in y, the convergence of such gradient methods is well understood. Outside this convex-concave regime, algorithms can fall into nonconvergent or diverging trajectories. We show that a classic generalization of the Moreau envelope [113] by Attouch [11] provides key insights into explaining these varied behaviors in the nonconvex-nonconcave settings. Particularly, we find that the related envelope is a smoothing of the original objective L and can be convex-concave even when L lacks such structure. As a result, gradient descent ascent methods can be analyzed on this reformulated envelope even when the original objective

does not possess such nice structure.

1.2 Related Works

In recent years, there have been a number of major steps forward in the optimization community extending the reach of first-order methods and their theory. Before we begin developing the advancements outlined above, we place them in the context of these other recent advances in the analysis of classic first-order methods and powerful new tools like Bregman and restarting methods.

Bregman methods (see [166]) improve on first-order methods for convex optimization by replacing the Euclidean norm with a reference function tailored to the given problem's geometry. By doing so, classic notions of Lipschitz continuity and smoothness can be replaced by relative measures with respect to the chosen reference function. Such theory for smooth optimization beyond having a Lipschitz gradient was developed independently in [107] and [13]. Such theory for nonsmooth optimization beyond having a Lipschitz objective function was developed by Lu [105] and provides an alternative approach to the results we develop in Chapter 2.

A series of works by Peña and Gutman [116, 69, 68] presented unified convergence proofs for many convex optimization methods ranging from nonsmooth subgradient methods to smooth accelerated methods. Further, the analysis in [68] connects these method's guarantees with those of conjugate gradient and Bregman methods, identifying a common primal-dual gap underlying the convergence of all of these methods.

Early in the development of first-order convex optimization theory, Nemirovskii and Nesterov [122] observed that first-order methods could have their convergence rates improved by periodically restarting the algorithm at its current iterate. Numerous recent works [102, 173, 150, 142] have built on these restarting ideas to great practical success, improving convergence guarantees under KL conditions or related growth bounds. These works complement the theory we develop in Chapter 7 by showing general convergence theory (without relying on the existence of growth/error bounds) can be utilized to give specialized results in growth bounded settings.

Following the work presented in Chapter 3, Davis and Drusvyatskiy [32] derived several nonconvex convergence guarantees related to subgradient methods under the weakened convexity condition (1.3). Their guarantees apply to a wide range of methods fitting within a generic family of model-based methods and match the convergence results (in expectation) that we present. Further, Davis et al. [34] derives faster nonconvex rates in the presence of sharp function growth and [33] derives limiting guarantees for the sweeping class of Whitney stratifiable functions.

CHAPTER 2

NONLIPSCHITZ SUBGRADIENT METHOD CONVERGENCE RATES

2.1 Introduction

We consider the nonsmooth, convex optimization problems of the form (1.1) for some lower semicontinuous convex function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and closed convex feasible region Q. We assume Q lies in the domain of f and that this problem has a nonempty set of minimizers X^* (with minimum value denoted by f^*). Further, we assume orthogonal projection onto Q is computationally tractable (which we denote by $\mathrm{proj}_Q(\cdot)$).

Since f may be nondifferentiable, we weaken the notion of gradients to subgradients. Recall the set of all subgradients at some $x \in Q$ (referred to as the subdifferential) is denoted by

$$\partial f(x) = \{g \in \mathbb{R}^d \mid \left(\forall y \in \mathbb{R}^d\right)\ f(y) \geq f(x) + g^T(y - x)\}.$$

We consider solving this problem via a (potentially stochastic) projected subgradient method. These methods have received much attention lately due to their simplicity and scalability; see [20, 126], as well as [74, 89, 105, 118, 139] for a sample of more recent works.

Deterministic and stochastic subgradient methods differ in the type of oracle used to access the subdifferential of f. For deterministic methods, we consider an oracle $g(x)$, which returns an arbitrary subgradient at x. For stochastic methods, we utilize a weaker, random oracle $g(x;\xi)$, which is an unbiased estimator of a subgradient (i.e., $\mathbb{E}_{\xi\sim D}\ g(x;\xi) \in \partial f(x)$ for some easily sampled distribution D). One of the earliest works motivating stochastic gradient methods was

Robbins and Monro [143]. An overview of the early work analyzing stochastic subgradient methods in optimization is given by Shor [161, §2.4] and the references therein.

In this chapter, we analyze two classic subgradient methods, differing in their step size policy. Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$. Given a deterministic oracle, we consider the following normalized subgradient method

$$x_{k+1} := \operatorname{proj}_Q\left(x_k - \alpha_k \frac{g(x_k)}{\|g(x_k)\|}\right), \tag{2.1}$$

for some positive sequence $(\alpha_k)_{k=0}^T$. Note that since $\|g(x_k)\| = 0$ only if x_k minimizes f, this iteration is well-defined until a minimizer is found. Given a stochastic oracle, we consider the following method

$$x_{k+1} := \operatorname{proj}_Q(x_k - \alpha_k g(x_k; \xi_k)), \tag{2.2}$$

for some positive sequence $(\alpha_k)_{k=0}^T$ and i.i.d. sample sequence $\xi_k \sim D$.

The standard convergence bounds for these methods assume all $x \in Q$ satisfy $\|g(x)\| \leq L$ or $\mathbb{E}_\xi \|g(x;\xi)\|^2 \leq L^2$ for some constant $L > 0$. Then after $T > 0$ iterations, a point is found with objective gap (in expectation for (2.2)) bounded by

$$f(x) - f^* \leq O\left(\frac{L\|x_0 - x^*\|}{\sqrt{T}}\right), \tag{2.3}$$

for any $x^* \in X^*$ under reasonable selection of the step size sequence $(\alpha_k)_{k=0}^T$.

The bound $\|g(x)\| \leq L$ for all $x \in Q$ is implied by f being L-Lipschitz continuous on some open convex set U containing Q (which is often the assumption made). Uniformly bounding subgradients restricts the classic convergence rates to functions with at most linear growth (at rate L). When Q is bounded, one can invoke a compactness argument to produce a uniform Lipschitz constant.

However, such an approach may lead to a large constant heavily dependent on the size of Q (and frankly, lacks the elegance that such a fundamental method deserves).

In stark contrast to these limitations, early in the development of subgradient methods Shor [160] observed that the normalized subgradient method (2.1) enjoys some form of convergence guarantee for any convex function with a nonempty set of minimizers. Shor showed for any minimizer $x^* \in X^*$: there exists some iterate $k \leq T$ for which either $x_k \in X^*$ or

$$\left(\frac{g(x_k)}{\|g(x_k)\|}\right)^T (x_k - x^*) \leq O\left(\frac{\|x_0 - x^*\|}{\sqrt{T}}\right),$$

under reasonable selection of the step size sequence $(\alpha_k)_{k=0}^T$. Thus for any convex function, the subgradient method has convergence in terms of this inner product value (which convexity implies is always nonnegative). This quantity can be interpreted as the distance from the hyperplane $\{x \mid g(x_k)^T(x - x_k) = 0\}$ to x^*. By driving this distance to zero via proper selection of $(\alpha_k)_{k=0}^T$, Shor characterized the asymptotic convergence of (2.1).

There is a substantial discrepancy in generality between the standard convergence bound (2.3) and Shor's result. In this chapter, we address this for both deterministic and stochastic subgradient methods. The remainder of this section formally states our generalized convergence rate bounds. For the deterministic case, our bounds follow directly from Shor's result, while the stochastic case requires an alternative approach. Then Section 2.2 applies these bounds to a few common problem classes outside the scope of uniform Lipschitz continuity. Finally, our convergence analysis is presented in Section 2.3 and an extension of our model is discussed in Section 2.4.

2.1.1 Extended Deterministic Convergence Bounds

Shor's convergence guarantees for general convex functions will serve as the basis of our objective gap convergence rates for the subgradient method (2.1) without assuming uniform Lipschitz continuity. Formally, Shor [160] showed the following general guarantee for any sequence of step sizes $(\alpha_k)_{k=0}^T$ (for completeness, an elementary proof is provided in Section 2.3).

Theorem 2.1.1 (Shor's Hyperplane Distance Convergence)**.** *Consider any convex f and fix some $x^* \in X^*$. Then for any positive sequence $(\alpha_k)_{k=0}^T$, there exists some iterate $k \leq T$ of the iteration* (2.1) *for which either $x_k \in X^*$ or*

$$\left(\frac{g(x_k)}{\|g(x_k)\|}\right)^T (x_k - x^*) \leq \frac{\|x_0 - x^*\|^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}. \tag{2.4}$$

Two Simple Stepsize Selections. The classic objective gap convergence of the subgradient method follows as a simple consequence of this. Indeed, f being convex and L-Lipschitz continuous on an open set containing Q (which implies $\|g(x_k)\| \leq L$) together with (2.4) imply

$$\min_{k=0\ldots T}\left\{\frac{f(x_k) - f^*}{L}\right\} \leq \frac{\|x_0 - x^*\|^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}.$$

Given either an upper bound[1] $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$, a convergence rate follows for either of the two following choices of the stepsize sequence $(\alpha_k)_{k=0}^T$. Taking $\alpha_k = R/\sqrt{T+1}$ produces

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \frac{LR}{\sqrt{T+1}}.$$

Alternatively, taking $\alpha_k = \epsilon/L$ yields

$$T \geq \left(\frac{L\|x_0 - x^*\|}{\epsilon}\right)^2 \implies \min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \epsilon.$$

[1]Note that an upper bound R can often be produced when Q is simple and bounded or when f possesses some structural property like strong convexity.

Here Lipschitz continuity enabled us to convert a bound on "hyperplane distance to a minimizer" into a bound on the objective gap. Our extended convergence bounds for the deterministic subgradient method follow from observing that more general assumptions than uniform Lipschitz continuity suffice to provide such a conversion. In particular, we assume there is an upper bound on f of the form

$$f(x) - f^* \leq \mathcal{D}(\|x - x^*\|), \quad (\forall x \in \mathbb{R}^d) \tag{2.5}$$

for some fixed $x^* \in X^*$ and nondecreasing nonnegative function $\mathcal{D}\colon \mathbb{R}_+ \to \mathbb{R}_+ \cup \{\infty\}$. In this case, we show the following objective gap convergence guarantee.

Theorem 2.1.2 (Extended Deterministic Rate)**.** *Consider any convex f satisfying* (2.5)*. Then for any positive sequence $(\alpha_k)_{k=0}^T$, the iteration* (2.1) *satisfies*

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \mathcal{D}\left(\frac{\|x_0 - x^*\|^2 + \sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}\right).$$

Two Simple Stepsize Selections. Suppose either an upper bound $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$ is known. Under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration (2.1) satisfies

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \mathcal{D}\left(\frac{R}{\sqrt{T+1}}\right).$$

Under the constant step size $\alpha_k = \mathcal{D}^{-1}(\epsilon)$, the iteration (2.1) satisfies

$$T \geq \left(\frac{\|x_0 - x^*\|}{\mathcal{D}^{-1}(\epsilon)}\right)^2 \implies \min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \epsilon,$$

where $\mathcal{D}^{-1}(\epsilon) = \inf\{t \mid \mathcal{D}(t) \geq \epsilon\}$.

Note that any L-Lipschitz continuous function on $\mathbb{R}^d$ satisfies this growth bound with $\mathcal{D}(t) = Lt$. Thus we immediately recover the standard $L\|x_0 -$

$x^*\|/\sqrt{T}$ convergence rate for unconstrained problems. Similarly, any L-Lipschitz continuous function on an open neighborhood of Q satisfies this growth bound with

$$\mathcal{D}(t) = \begin{cases} Lt & \text{if } t \leq \delta \\ +\infty & \text{otherwise} \end{cases}$$

for some $\delta > 0$. From this, we recover the standard $L\|x_0 - x^*\|/\sqrt{T}$ rate for constrained problems provided $T \geq \|x_0 - x^*\|^2/\delta^2$.

Using growth bounds allows us to apply our convergence guarantees to many problems outside the scope of uniform Lipschitz continuity. Theorem 2.1.2 also implies the classic convergence rate for gradient descent on differentiable functions with an L-Lipschitz continuous gradient of $O(L\|x_0 - x^*\|^2/T)$ [126]. Any such function has growth bounded by $\mathcal{D}(t) = Lt^2/2$ on $Q = \mathbb{R}^d$ (see Lemma 2.2.1). Then a convergence rate immediately follows from Theorem 2.1.2 (for simplicity, we consider a constant step size given an upper bound $R \geq \|x_0 - x^*\|$).

Corollary 2.1.3 (Generalizing Gradient Descent's Convergence)**.** *Consider any convex function f satisfying* (2.5) *with $\mathcal{D}(t) = Lt^2/2$. Then under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration* (2.1) *satisfies*

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \frac{LR^2}{2(T+1)}.$$

Thus a convergence rate of $O(LR^2/T)$ can be attained without any mention of smoothness or differentiability. In Section 2.2, we provide a similar growth bound and thus objective gap convergence for any function with a Hölder continuous gradient, which also parallels the standard rate for gradient descent. In general, for any problem with $\lim_{t\to 0^+} \mathcal{D}(t)/t = 0$, Theorem 2.1.2 produces convergence at a rate of $o(1/\sqrt{T})$.

Suppose that $\mathcal{D}(t)/t$ is finite in some neighborhood of 0 (as is the case for any f that is locally Lipschitz around x^*). Then simple limiting arguments yield the following eventual convergence rate of (2.1) based on Theorem 2.1.2: for any $\epsilon > 0$, there exists $T_0 > 0$ such that all $T > T_0$ have

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \mathcal{D}\left(\frac{R}{\sqrt{T+1}}\right) \leq \left(\limsup_{t\to 0^+} \frac{\mathcal{D}(t)}{t} + \epsilon\right)\frac{R}{\sqrt{T+1}}.$$

As a result, the asymptotic convergence rate of (2.1) is determined entirely by the rate of growth of f around its minimizers, and conversely, steepness far from optimality plays no role in the asymptotic behavior.

2.1.2 Extended Stochastic Convergence Bounds

Now we turn our attention to giving more general convergence bounds for the stochastic subgradient method. This is harder as we can no longer leverage Shor's result since normalizing stochastic subgradients may introduce bias or may not be well-defined if $g(x_k;\xi) = 0$. As a consequence, we need a different approach to generalizing the standard stochastic assumptions.

We begin by reviewing the standard convergence results for this method. The following convergence guarantee is immediate from the analysis given in [161, §2.4] and well-known in the literature.

Theorem 2.1.4 (Classic Stochastic Rate)**.** *Consider any convex function f and stochastic subgradient oracle satisfying $\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq L^2$ for all $x \in Q$. Fix some $x^* \in X^*$. Then for any positive sequence $(\alpha_k)_{k=0}^T$, the iteration* (2.2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{\sum_{k=0}^T \alpha_k x_k}{\sum_{k=0}^T \alpha_k}\right) - f^*\right] \leq \frac{\|x_0 - x^*\|^2 + L^2\sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}.$$

Two Simple Stepsize Selections Similar to the deterministic setting, given either an upper bound $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$, simple constant stepsizes can be analyzed. Under the selection $\alpha_k = R/(L\sqrt{T+1})$, the iteration (2.2) satisfies

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^{T} x_k\right) - f^*\right] \leq \frac{LR}{\sqrt{T+1}}.$$

Under the selection $\alpha_k = \epsilon/L^2$, the iteration (2.2) satisfies

$$T \geq \left(\frac{L\|x_0 - x^*\|}{\epsilon}\right)^2 \implies \mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^{T} x_k\right) - f^*\right] \leq \epsilon.$$

We say f is μ-strongly convex on Q for some $\mu > 0$ if for every $x \in Q$ and $g \in \partial f(x)$,

$$f(y) \geq f(x) + g^T(y - x) + \frac{\mu}{2}\|y - x\|^2 \quad (\forall y \in Q).$$

Under this condition, the convergence of (2.2) can be improved to $O(1/T)$ [74, 89, 139]. Below, we present one such bound from [89].

Theorem 2.1.5 (Classic Strongly Convex Stochastic Rate)**.** *Consider any μ-strongly convex function f and stochastic subgradient oracle satisfying $\mathbb{E}_\xi\|g(x;\xi)\|^2 \leq L^2$ for all $x \in Q$. Then for the sequence of step sizes $\alpha_k = 2/\mu(k+2)$, the iteration* (2.2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^{T}(k+1)x_k\right) - f^*\right] \leq \frac{2L^2}{\mu(T+2)}.$$

We remark that Lipschitz continuity and strong convexity are fundamentally at odds. Lipschitz continuity allows at most linear growth while strong convexity requires quadratic growth. The only way both can occur is when Q is bounded.

The standard analysis assumes that $\mathbb{E}_\xi\|g(x;\xi)\|^2$ is uniformly bounded by some $L^2 > 0$. We generalize this by allowing the expectation to be larger when

the objective gap at x is large as well. In particular, we assume a bound of the form

$$\mathbb{E}_\xi \|g(x;\xi)\|^2 \leq L_0^2 + L_1(f(x) - f^*) \tag{2.6}$$

for some constants $L_0, L_1 \geq 0$. When L_1 equals zero, this is exactly the classic model. When L_1 is positive, this model allows functions with up to quadratic growth. (To see this, suppose the subgradient oracle is deterministic. Then (2.6) corresponds to a differential inequality of the form $\|\nabla f(x)\| \leq \sqrt{L_1(f(x) - f^*) + L_0^2}$, which has a simple quadratic solution. This interpretation is formalized in Section 2.2.4.)

The additional generality allowed by (2.6) is important for two reasons. First, it allows us to consider many classic problems which fundamentally have quadratic growth (for example, any quadratically regularized problem, like training a support vector machine, which is considered in Section 2.2.3). Secondly, this model allows us to avoid the inherent conflict in Theorem 2.1.5 between Lipschitz continuity and strong convexity since a function can globally satisfy both (2.6) and strong convexity.

Based on this generalization of Lipschitz continuity, we have the following guarantees for convex and strongly convex problems.

Theorem 2.1.6 (Extended Stochastic Rate)**.** *Consider any convex function f and stochastic subgradient oracle satisfying* (2.6)*. Fix some $x^* \in X^*$. Then for any positive sequence $(\alpha_k)_{k=0}^T$ with $L_1\alpha_k < 2$ for all $k = 0 \dots T$, the iteration* (2.2) *satisfies*

$$\mathbb{E}_{\xi_{0\dots T}}\left[f\left(\frac{\sum_{k=0}^T \alpha_k(2 - L_1\alpha_k)x_k}{\sum_{k=0}^T \alpha_k(2 - L_1\alpha_k)}\right) - f^*\right] \leq \frac{\|x_0 - x^*\|^2 + L_0^2 \sum_{k=0}^T \alpha_k^2}{\sum_{k=0}^T \alpha_k(2 - L_1\alpha_k)}.$$

Two Simple Stepsize Selections Given either an upper bound $R \geq \|x_0 - x^*\|$ or a target accuracy $\epsilon > 0$, we present bounds for two simple constant stepsizes.

Under the selection $\alpha_k = R/L_0\sqrt{T+1}$, the iteration (2.2) satisfies

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^{T} x_k\right) - f^*\right] \leq \frac{L_0 R}{\sqrt{T+1}} + \frac{L_1 R^2}{T+1},$$

provided $T \geq (RL_1/L_0)^2$. Under the selection $\alpha_k = \epsilon/(2L_0^2)$, the iteration (2.2) satisfies

$$T \geq \left(\frac{L_0\|x_0 - x^*\|}{\epsilon}\right)^2 \implies \mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{1}{T+1}\sum_{k=0}^{T} x_k\right) - f^*\right] \leq \epsilon,$$

provided $\epsilon \leq 2L_0^2/L_1$.

Theorem 2.1.7 (Extended Strongly Convex Stochastic Rate)**.** *Consider any μ-strongly convex function f and stochastic subgradient oracle satisfying* (2.6)*. Fix some $x^* \in X^*$. Then for the sequence of step sizes*

$$\alpha_k = \frac{2}{\mu(k+2) + \frac{L_1^2}{\mu(k+1)}},$$

the iteration (2.2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{\sum_{k=0}^{T}(k+1)(2 - L_1\alpha_k)x_k}{\sum_{k=0}^{T}(k+1)(2 - L_1\alpha_k)}\right) - f^*\right] \leq \frac{2L_0^2(T+1) + L_1^2\|x_0 - x^*\|^2/2}{\mu\sum_{k=0}^{T}(k+1)(2 - L_1\alpha_k)}.$$

The following simpler averaging yields a bound weakened roughly by a factor of two:

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{2}{(T+1)(T+2)}\sum_{k=0}^{T}(k+1)x_k\right) - f^*\right] \leq \frac{4L_0^2}{\mu(T+2)} + \frac{L_1^2\|x_0 - x^*\|^2}{\mu(T+1)(T+2)}.$$

We remark that one important insight given by Theorem 2.1.7 comes from its dependence on the initial point x_0. The rate only depends on the initial iterate in the second term above, which decays at a rate of $O(1/T^2)$. This shows the choice of the initial point has a relatively small impact on the asymptotic guarantees. Note the standard analysis of this method (see Theorem 2.1.5) does not give any insight into the dependence on x_0 and instead uses the implicit bound on $\|x_0 - x^*\|$ given by strong convexity and Lipschitz continuity.

2.1.3 Related Works

Recently, Renegar [140] introduced a novel framework that allows first-order methods to be applied to general (non-Lipschitz) convex optimization problems via a radial transformation. In Chapters 5 and 6, we will further develop this tool, showing a simple radial subgradient method has convergence paralleling the classic $O(1/\sqrt{T})$ rate without assuming Lipschitz continuity. This algorithm is applied to a transformed version of the original problem and replaces orthogonal projection by a line search at each iteration.

Lu [105] analyzes an interesting subgradient-type method (which is a variation of mirror descent) for non-Lipschitz problems that is customized for a particular problem via a reference function. This approach produces convergence guarantees for both deterministic and stochastic problems based on a relative-continuity constant instead of a uniform Lipschitz constant.

Although the works of Renegar [140], Chapters 5 and 6, and Lu [105] provide convergence rates for specialized subgradient methods without assuming Lipschitz continuity, objective gap guarantees for the classic subgradient methods (2.1) and (2.2), such as the ones presented here, have been missing prior to our work.

2.2 Applications of Our Extended Convergence Bounds

In this section, we apply our convergence bounds to a variety of problems outside the scope of the traditional theory based on uniform Lipschitz constants.

2.2.1 Smooth Optimization

The standard analysis of gradient descent in smooth optimization assumes the gradient of the objective function is uniformly Lipschitz continuous, or more generally, uniformly Hölder continuous. A differentiable function f has (L, v)-Hölder continuous gradient on $\mathbb{R}^d$ for some $L > 0$ and $v \in (0, 1]$ if for all $x, y \in \mathbb{R}^d$

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

Note this is exactly Lipschitz continuity of the gradient when $v = 1$. Below, we state a simple bound on the growth $\mathcal{D}(t)$ of any such function.

Lemma 2.2.1. *Consider any $f \in C^1$ with a (L, v)-Hölder continuous gradient on $\mathbb{R}^d$ and any minimizer $x^* \in X^*$. Then*

$$f(x) - f(x^*) \leq \frac{L}{v+1}\|x - x^*\|^{v+1} \quad (\forall x \in \mathbb{R}^d).$$

Proof. Since $\nabla f(x^*) = 0$, the bound follows directly as

$$\begin{aligned} f(x) &= f(x^*) + \int_0^1 \nabla f(x^* + t(x - x^*))^T (x - x^*)\, dt \\ &\leq f(x^*) + \nabla f(x^*)^T (x - x^*) + \int_0^1 Lt^v \|x - x^*\|^{v+1}\, dt \\ &= f(x^*) + \frac{L}{v+1}\|x - x^*\|^{v+1}. \end{aligned}$$

□

This lemma with $v = 1$ implies any function with an L-Lipschitz gradient has growth bounded by $\mathcal{D}(t) = Lt^2/2$. Then Theorem 2.1.2 produces our generalization of the classic gradient descent convergence rate claimed in Corollary 2.1.3. Further, for any function with a Hölderian gradient, Theorem 2.1.2 gives a $O(1/T^{(v+1)/2})$ convergence rate. The following Corollary generalizes this fact giving a convergence rate for any (potentially non-differentiable) function with upper bound $\mathcal{D}(t) = Lt^{v+1}/(v + 1)$.

Corollary 2.2.2 (Generalizing Hölderian Gradient Descent's Convergence)**.** *Consider any convex function f satisfying* (2.5) *with $\mathcal{D}(t) = Lt^{v+1}/(v+1)$. Then under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration* (2.1) *satisfies*

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \frac{LR^{v+1}}{(v+1)(T+1)^{(v+1)/2}}.$$

2.2.2 Additive Composite Optimization

Often problems arise where the objective is to minimize a sum of smooth and nonsmooth functions. We consider the following general formulation of this problem

$$\min_{x\in\mathbb{R}^d} f(x) := \Phi(x) + h(x),$$

for any differentiable convex function Φ with (L_Φ, v)-Hölderian gradient and any L_h-Lipschitz continuous convex function h. Such problems occur when regularizing smooth optimization problems, where h would be the sum of one or more nonsmooth regularizers (for example, $\|\cdot\|_1$ to induce sparsity).

Additive composite problems can be solved by prox-gradient or splitting methods, which solve a subproblem based on h at each iteration. However, this limits these methods to problems where h is relatively simple. The subgradient method avoids this limitation by only requiring the computation of a subgradient of f at each iteration, with the subdifferential being given by $\partial f(x) = \nabla\Phi(x) + \partial h(x)$. The classic convergence theory fails to provide any guarantees for this problem since f may be non-Lipschitz. In contrast, we show this problem class has a simple growth bound from which guarantees for the classic subgradient method directly follow.

Lemma 2.2.3. *Consider any $\Phi \in C^1$ with a (L_Φ, v)-Hölder continuous gradient on $\mathbb{R}^d$, any L_h-Lipschitz continuous h on $\mathbb{R}^d$, and any minimizer $x^* \in X^*$. Then*

$$f(x) - f(x^*) \leq \frac{L_\Phi}{v+1}\|x - x^*\|^{v+1} + 2L_h\|x - x^*\| \quad (\forall x \in \mathbb{R}^d).$$

Proof. From the optimality conditions of f, we know $g^* := -\nabla\Phi(x^*) \in \partial h(x^*)$. Define the following lower bound on $f(x)$

$$l(x) := \Phi(x) + h(x^*) + g^{*T}(x - x^*).$$

Notice that $f(x)$ and $l(x)$ both minimize at x^* with $f(x^*) = l(x^*)$. Since $l(x)$ has a (L_Φ, v)-Hölder continuous gradient, Lemma 2.2.1 implies for any $x \in \mathbb{R}^d$,

$$l(x) - l(x^*) \leq \frac{L_\Phi}{v+1}\|x - x^*\|^{v+1}.$$

The Lipschitz continuity of h implies

$$l(x) = \Phi(x) + h(x^*) + g^{*T}(x - x^*) \geq \Phi(x) + (h(x) - L_h\|x - x^*\|) - L_h\|x - x^*\|.$$

Combining these two inequalities completes the proof. □

Plugging $\mathcal{D}(t) = L_\Phi t^{v+1}/(v+1) + 2L_h t$ into Theorem 2.1.2 immediately results in the following $O(1/\sqrt{T})$ convergence rate (for simplicity, we state the bound for constant step size).

Corollary 2.2.4 (Additive Composite Convergence)**.** *Consider the deterministic subgradient oracle $\nabla\Phi(x) + g_h(x)$. Then under the constant step size $\alpha_k = R/\sqrt{T+1}$, the iteration* (2.1) *satisfies*

$$\min_{k=0\ldots T}\{f(x_k) - f^*\} \leq \frac{L_\Phi R^{v+1}}{(v+1)(T+1)^{(v+1)/2}} + \frac{2L_h R}{\sqrt{T+1}}.$$

The first term in this rate exactly matches the convergence rate on functions, such as Φ, with Hölderian gradient (see Corollary 2.2.2). Further, up to a factor

of two, the second term matches the convergence rate on Lipschitz continuous functions, such as h (see (2.3)). Thus the subgradient method on $\Phi(x) + h(x)$ has convergence guarantees no worse than those of the subgradient method on $\Phi(x)$ or $h(x)$ separately.

2.2.3 Quadratically Regularized Stochastic Optimization

Another common class of optimization problems results from adding in a quadratic regularization term $(\lambda/2)\|x\|^2$ to the objective function, for some parameter $\lambda > 0$. Consider solving

$$\min_{x \in \mathbb{R}^d} f(x) := h(x) + \frac{\lambda}{2}\|x\|^2$$

for any Lipschitz continuous convex function h. Suppose we have a stochastic subgradient oracle for h denoted by $g_h(x; \xi)$ for which $\mathbb{E}_\xi \; g_h(x; \xi) \in \partial h(x)$ and $\mathbb{E}_\xi \|g_h(x; \xi)\|^2 \leq L^2$. Although the function h and its stochastic oracle meet the necessary conditions for the classic theory to be applied, the addition of a quadratic term violates uniform Lipschitz continuity. Nonetheless, simple arguments yield a subgradient norm bound like (2.6) and the following $O(1/T)$ convergence rate.

Corollary 2.2.5 (Quadratically Regularized Convergence)**.** *Consider the step sizes*

$$\alpha_k = \frac{2}{\lambda(k+2) + \frac{36\lambda}{k+1}}$$

and stochastic subgradient oracle $g_h(x; \xi) + \lambda x$. Fix some $x^* \in X^*$. The iteration (2.2) *satisfies*

$$\mathbb{E}_{\xi_{0 \ldots T}} \left[f\left(\frac{2}{(T+1)(T+2)} \sum_{k=0}^{T} (k+1) x_k \right) - f^* \right] \leq \frac{24L^2}{\lambda(T+2)} + \frac{36\lambda\|x_0 - x^*\|^2}{(T+1)(T+2)}.$$

Proof. Consider any $x^* \in X^*$ and $g^* := -\lambda x^* \in \partial h(x^*)$ (this inclusion follows from the first-order optimality conditions for x^*). From the assumed stochastic subgradient norm bound $\mathbb{E}_\xi \|g_h(x;\xi)\|^2 \leq L^2$, all subgradients of h must have norm bounded by L. This follows since each differentiable point has $\|\nabla f(x)\|^2 \leq \mathbb{E}_\xi \|g_h(x;\xi)\|^2 \leq L^2$ and the subdifferential at nondifferentiable points is given by the convex hull of nearby gradients: $\partial f(x) = \operatorname{conv}\{\lim \nabla f(z_k) \mid \lim z_k \to x, z_k \in Q\}$ where Q is the set of differentiable points near x (see [27] for a proof of this characterization). Then the expected norm squared of the stochastic subgradient $g_h(x;\xi) + \lambda x$ is bounded by

$$\begin{aligned}
\mathbb{E}_\xi \|g_h(x;\xi) + \lambda x\|^2 &= \mathbb{E}_\xi \|g_h(x;\xi) - g^* + g^* + \lambda x\|^2 \\
&\leq 3\mathbb{E}_\xi \|g_h(x;\xi)\|^2 + 3\|g^*\|^2 + 3\|g^* + \lambda x\|^2 \\
&\leq 6L^2 + 3\|g^* + \lambda x\|^2 \\
&\leq 6L^2 + 6\lambda(f(x) - f(x^*)),
\end{aligned}$$

where the first inequality uses Jensen's inequality, the second inequality uses the subgradient norm bound, and the third inequality uses the λ-strong convexity of f. From this, our bound follows by Theorem 2.1.7. □

One common example of a problem of the form $h(x) + (\lambda/2)\|x\|^2$ is training a Support Vector Machine (SVM). Suppose one has n data points each with a feature vector $w_i \in \mathbb{R}^d$ and label $y_i \in \{-1, 1\}$. Then one trains a model $x \in \mathbb{R}^d$ for some parameter $\lambda > 0$ by solving

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i w_i^T x\} + \frac{\lambda}{2}\|x\|^2.$$

Here, a stochastic subgradient oracle can be given by selecting a summand $i \in$

$[n]$ uniformly at random and then setting

$$g_h(x;i) = \begin{cases} -y_i w_i & \text{if } 1 - y_i w_i^T x \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

which satisfies $\mathbb{E}_i \|g_h(x,i)\|^2 \leq \frac{1}{n} \sum_{i=1}^n \|w_i\|^2$.

Much work has previously been done solving problems of the form $h(x) + \frac{\lambda}{2}\|x\|^2$ and SVMs in particular. If one adds the constraint that x lies in some large ball Q (which will then be projected onto at each iteration), the classic strongly convex rate can be applied [158] as the objective function will be Lipschitz on Q. A similar approach utilized in [89] is to show that, in expectation, all of the iterates of a stochastic subgradient method lie in a large ball (provided the initial iterate does). We remark that the resulting guarantees only apply for $x_0 \in Q$ and utilize a constant L dependent on the size of Q. Corollary 2.2.5 avoids these issues, giving a convergence bound for any choice of $x_0 \in \mathbb{R}^d$.

The specialized mirror descent method proposed by Lu [105] produces convergence guarantees for SVMs at a rate of $O(1/\sqrt{T})$ without needing a bounding ball. Splitting methods and quasi-Newton methods capable of solving this problem are given in [43] and [175], respectively, which both avoid needing to assume subgradient bounds.

2.2.4 Interpreting (2.6) as a Quadratic Growth Upper Bound

Here we provide an alternative interpretation of bounding the size of subgradients by (2.6) on some convex open set $U \subseteq \mathbb{R}^d$ for deterministic subgradient

oracles. In particular, suppose all $x \in U$ have

$$\|g(x)\|^2 \leq L_0^2 + L_1(f(x) - \inf_{x' \in U} f(x')) \tag{2.7}$$

First consider the classic model where $L_1 = 0$. This is equivalent to f being L_0-Lipschitz continuous on U and can be restated as the following upper bound holding for each $x \in U$:

$$f(y) \leq f(x) + L_0\|y - x\| \quad (\forall y \in U).$$

This characterization shows the limitation to linear growth of the classic model. In the following proposition, we present an upper bound characterization when $L_1 > 0$, which can be viewed as allowing up to quadratic growth.

Proposition 2.2.6. *A convex function f satisfies* (2.7) *on some open convex $U \subseteq \mathbb{R}^d$ if and only if the following quadratic upper bound holds for each $x \in U$*

$$f(y) \leq f(x) + \frac{L_1}{4}\|y - x\|^2 + \|y - x\|\sqrt{L_1(f(x) - \inf_{x' \in U} f(x')) + L_0^2} \quad (\forall y \in U).$$

Proof. First we prove the forward direction. Consider any $x, y \in U$ and subgradient oracle $g(\cdot)$. Let $v = (y - x)/\|y - x\|$ denote the unit direction from x to y, and $h(t) = f(x + tv) - \inf_{x' \in U} f(x')$ denote the restriction of f to this line shifted to have nonnegative value. Notice that $h(0) = f(x) - \inf_{x' \in U} f(x')$ and $h(\|y - x\|) = f(y) - \inf_{x' \in U} f(x')$. The convexity of h implies it is differentiable almost everywhere in the interval $[0, \|y - x\|]$. Thus h satisfies the following, for almost every $t \in [0, \|y - x\|]$,

$$|h'(t)| = |v^T g(x + tv)| \leq \|g(x + tv)\|.$$

This produces the differential inequality of $|h'(t)| \leq \sqrt{L_1 h(t) + L_0^2}$. Note that the unique solution to the ordinary differential equation $y'(t) = \sqrt{L_1 y(t) + L_0^2}$

with initial condition $y(0) = h(0)$ is $y(t) = h(0) + \frac{L_1}{4}t^2 + t\sqrt{L_1 h(0) + L_0^2}$. Then the claimed bound will follow from showing $h(t) \leq y(t)$ at $t = \|y - x\|$. This inequality must be the case for all $t \geq 0$, as otherwise some $t \geq 0$ must have $h(t) = y(t)$ and $\limsup_{t' \to t} h'(t') > y'(t)$, which implies

$$\limsup_{t' \to t} h'(t') > \sqrt{L_1 y(t) + L_0^2} = \sqrt{L_1 h(t) + L_0^2},$$

contradicting our premise.

Now we prove the reverse direction. Denote the upper bound given by some $x \in U$ as

$$u_x(y) := f(x) + \frac{L_1}{4}\|y - x\|^2 + \|y - x\|\sqrt{L_1(f(x) - \inf_{x' \in U} f(x')) + L_0^2}.$$

Further, let D_v denote the directional derivative operator in some unit direction $v \in \mathbb{R}^d$. Then for any subgradient $g \in \partial f(x)$,

$$v^T g \leq D_v f(x) \leq D_v u_x(x),$$

where the first inequality uses the definition of D_v and the second uses the fact that u_x upper bounds f. A simple calculation shows $D_v u_x(x) \leq \sqrt{L_1(f(x) - \inf_{x' \in U} f(x')) + L_0^2}$. Then our subgradient bound follows by taking $v = g/\|g\|$. □

2.3 Convergence Proofs

Each of our extended convergence theorems follows from essentially the same proof as its classic counterpart. The central inequality in analyzing subgradient methods is the following.

Lemma 2.3.1. *Consider any convex function f. For any $x, y \in Q$ and $\alpha > 0$,*

$$\mathbb{E}_\xi \|\text{proj}_Q(x - \alpha g(x;\xi)) - y\|^2 \leq \|x - y\|^2 - 2\alpha(\mathbb{E}_\xi\, g(x;\xi))^T(x - y) + \alpha^2 \mathbb{E}_\xi \|g(x;\xi)\|^2.$$

Proof. Since orthogonal projection onto a convex set is nonexpansive, we have

$$\begin{aligned}\|\text{proj}_Q(x - \alpha g(x;\xi)) - y\|^2 &\leq \|x - \alpha g(x;\xi) - y\|^2 \\ &= \|x - y\|^2 - 2\alpha g(x;\xi)^T(x - y) + \alpha^2 \|g(x;\xi)\|^2.\end{aligned}$$

Taking the expectation over $\xi \sim D$ yields

$$\mathbb{E}_\xi \|\text{proj}_Q(x - \alpha g(x;\xi)) - y\|^2 \leq \|x - y\|^2 - 2\alpha(\mathbb{E}_\xi\, g(x;\xi))^T(x - y) + \alpha^2 \mathbb{E}_\xi \|g(x;\xi)\|^2.$$

□

Let $D_k^2 = \mathbb{E}_{\xi_{0\ldots T}} \|x_k - x^*\|^2$ denote the expected distance squared from each iterate to the minimizer x^*. Each of our proofs follows the same general outline: use Lemma 2.3.1 to set up a telescoping inequality on D_k^2, then sum the telescope. We begin by proving Shor's convergence result as its derivation is short and informative.

2.3.1 Proof of Shor's Theorem 2.1.1

From Lemma 2.3.1 with $x = x_k$, $y = x^*$, and $\alpha = \alpha_k / \|g(x_k)\|$, it follows that

$$D_{k+1}^2 \leq D_k^2 - \frac{2\alpha_k g(x_k)^T(x_k - x^*)}{\|g(x_k)\|} + \alpha_k^2,$$

Inductively applying this implies

$$0 \leq D_{T+1}^2 \leq D_0^2 - \sum_{k=0}^{T} 2\alpha_k \frac{g(x_k)^T(x_k - x^*)}{\|g(x_k)\|} + \sum_{k=0}^{T} \alpha_k^2.$$

Thus

$$\min_{k=0\ldots T}\left\{\frac{g(x_k)^T(x_k-x^*)}{\|g(x_k)\|}\right\} \leq \frac{D_0^2+\sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k},$$

completing the proof. □

2.3.2 Proof of Theorem 2.1.2

This follows directly from Theorem 2.1.1. Note the result trivially holds if some iterate $0 \leq k \leq T$ satisfies $x_k \in X^*$. Suppose x_k satisfies the inequality in Theorem 2.1.1. Let y be the closest point in $\{x \mid g(x_k)^T(x-x_k)=0\}$ to x^*. Then our assumed growth bound implies

$$f(y)-f^* \leq \mathcal{D}(\|y-x^*\|) = \mathcal{D}\left(\frac{g_k^T(x_k-x^*)}{\|g_k\|}\right) \leq \mathcal{D}\left(\frac{D_0^2+\sum_{k=0}^T \alpha_k^2}{2\sum_{k=0}^T \alpha_k}\right).$$

The convexity of f implies $f(x_k) \leq f(y)$ completing the proof.

2.3.3 Proof of Theorem 2.1.6

From Lemma 2.3.1 with $x = x_k$, $y = x^*$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\leq D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[2\alpha_k(\mathbb{E}_\xi g(x_k;\xi_k))^T(x_k-x^*)\right] + \alpha_k^2\mathbb{E}_{\xi_{0\ldots T}}\|g(x_k,\xi_k)\|^2 \\ &\leq D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[(2\alpha_k - L_1\alpha_k^2)(f(x_k)-f^*)\right] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the convexity of f and the bound on $\mathbb{E}_\xi\|g(x;\xi)\|^2$. Inductively applying this implies

$$0 \leq D_{T+1}^2 \leq D_0^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[\sum_{k=0}^T (2\alpha_k - L_1\alpha_k^2)(f(x_k)-f^*)\right] + L_0^2\sum_{k=0}^T \alpha_k^2.$$

The convexity of f yields

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{\sum_{k=0}^T \alpha_k(2-L_1\alpha_k)x_k}{\sum_{k=0}^T \alpha_k(2-L_1\alpha_k)}\right) - f^*\right] \leq \frac{D_0^2+L_0^2\sum_{k=0}^T \alpha_k^2}{\sum_{k=0}^T \alpha_k(2-L_1\alpha_k)},$$

completing the proof.

2.3.4 Proof of Theorem 2.1.7

Our proof follows the style of [89]. Observe that our choice of step size α_k satisfies the following pair of conditions: First, note that it is a solution to the recurrence

$$(k+1)\alpha_k^{-1} = (k+2)(\alpha_{k+1}^{-1} - \mu). \tag{2.8}$$

Second, note that $L_1\alpha_k \leq 1$ for all $k \geq 0$ since

$$L_1\alpha_k = \frac{2\mu(k+2)L_1}{(\mu(k+2))^2 + \frac{k+2}{k+1}L_1^2} \leq \frac{2\mu(k+2)L_1}{(\mu(k+2))^2 + L_1^2} \leq 1. \tag{2.9}$$

From Lemma 2.3.1 with $x = x_k$, $y = x^*$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\leq D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[2\alpha_k(\mathbb{E}_\xi g(x_k;\xi))^T(x_k - x^*)\right] + \alpha_k^2\mathbb{E}_{\xi_{0\ldots T}}\|g(x_k,\xi_k)\|^2 \\ &\leq (1-\mu\alpha_k)D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[(2\alpha_k - L_1\alpha_k^2)(f(x_k) - f^*)\right] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the strong convexity of f and the bound on $\mathbb{E}_\xi\|g(x;\xi)\|^2$. Multiplying by $(k+1)/\alpha_k$ and invoking (2.9) yields

$$\begin{aligned} (k+1)\alpha_k^{-1}D_{k+1}^2 \leq &(k+1)(\alpha_k^{-1} - \mu)D_k^2 \\ &- \mathbb{E}_{\xi_{0\ldots T}}[(k+1)(2 - L_1\alpha_k)(f(x_k) - f^*)] + L_0^2(k+1)\alpha_k. \end{aligned}$$

Notice that this inequality telescopes due to (2.8). Inductively applying this implies

$$0 \leq (\alpha_0^{-1} - \mu)D_0^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[\sum_{k=0}^{T}(k+1)(2 - L_1\alpha_k)(f(x_k) - f^*)\right] + L_0^2\sum_{k=0}^{T}(k+1)\alpha_k.$$

Since $\sum_{k=0}^{T}(k+1)\alpha_k \leq 2(T+1)/\mu$ and $\alpha_0^{-1} - \mu = L_1^2/2\mu$, we have

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\sum_{k=0}^{T}(k+1)(2-L_1\alpha_k)(f(x_k)-f^*)\right] \leq \frac{L_1^2 D_0^2}{2\mu} + \frac{2L_0^2(T+1)}{\mu}.$$

Observe that the coefficients of each $f(x_k) - f^*$ above are positive due to (2.9). Then the convexity of f yields our first convergence bound. From (2.9), we know $2 - L_1\alpha_k \geq 1$ for all $k \geq 0$. Then the previous inequality can be weakened to

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\sum_{k=0}^{T}(k+1)(f(x_k)-f^*)\right] \leq \frac{L_1^2 D_0^2}{2\mu} + \frac{2L_0^2(T+1)}{\mu}.$$

The convexity of f yields our second convergence bound.

2.4 Improved Convergence Without Strong Convexity

The idea of utilizing growth lower bounds to improve convergence guarantees is far from new. Nemirovskii and Nesterov [122] showed restarted variations of standard first-order methods can give optimal convergence rates for convex problems satisfying the following Hölder growth bound (referred to therein as a "strict minimum condition") for all $x \in \mathbb{R}^d$: $f(x) - f^* \geq \mu\|x - x^*\|^{v+1}$ where x^* is the unique minimizer of f. Later, the work of Burke and Ferris [23] studied (potentially nonconvex) problems with *weak sharp minima*, that is, for some $S \subseteq \mathbb{R}^d$ and $y \in S$, all x near S have $f(x) - f(y) \geq \mu\mathbf{dist}(x, S)$. They show this condition often holds and enables improved convergence guarantees, sometimes ensuring finite convergence.

Many recent works have considered weakening the assumption of strong convexity while maintaining the standard improvements in convergence rate for smooth optimization problems (for example, see [18, 40, 108, 117]). Instead,

the weaker condition of requiring quadratic growth away from the set of minimizers suffices. We demonstrate that this weakening of strong convexity is also sufficient for (2.2) to have a convergence rate of $O(1/T)$.

A function f has μ-quadratic growth for some $\mu > 0$ if all $x \in Q$ satisfy

$$f(x) \geq f^* + \frac{\mu}{2}\mathbf{dist}(x, X^*)^2.$$

The proof of Theorem 2.1.7 only uses strong convexity once for the following inequality:

$$g(x_k)^T(x_k - x^*) \geq f(x_k) - f^* + \frac{\mu}{2}\|x_k - x^*\|^2.$$

Having μ-quadratic growth suffices to produce a similar inequality, weakened by a factor of $1/2$:

$$g(x_k)^T(x_k - \mathrm{proj}_{X^*}(x_k)) \geq f(x_k) - f^* \geq \frac{1}{2}(f(x_k) - f^*) + \frac{\mu}{4}\mathbf{dist}(x_k, X^*)^2.$$

Then simple modifications of the proof of Theorem 2.1.7 yield the following convergence rate.

Theorem 2.4.1. *Consider any convex function f with μ-quadratic growth and stochastic subgradient oracle satisfying* (2.6)*. Then for the sequence of step sizes*

$$\alpha_k = \frac{4}{\mu(k+2) + \frac{4L_1^2}{\mu(k+1)}},$$

the iteration (2.2) *satisfies*

$$\mathbb{E}_{\xi_{0\ldots T}}\left[f\left(\frac{\sum_{k=0}^T (k+1)(1 - L_1\alpha_k)x_k}{\sum_{k=0}^T (k+1)(1 - L_1\alpha_k)}\right) - f^*\right] \leq \frac{4L_0^2(T+1) + L_1^2\mathbf{dist}(x_0, X^*)^2}{\mu\sum_{k=0}^T (k+1)(1 - L_1\alpha_k)}.$$

Proof. Observe that our choice of step size α_k satisfies the following pair of conditions. First, note that it is a solution to the recurrence

$$(k+1)\alpha_k^{-1} = (k+2)(\alpha_{k+1}^{-1} - \mu/2). \tag{2.10}$$

Second, note that $L_1\alpha_k < 1$ for all $k \geq 0$. This follows as

$$L_1\alpha_k = \frac{4\mu(k+2)L_1}{(\mu(k+2))^2 + 4\frac{k+2}{k+1}L_1^2} \leq \frac{4\mu(k+2)L_1}{(\mu(k+2))^2 + (2L_1)^2} \leq 1, \tag{2.11}$$

where the first inequality is strict if $L_1 > 0$ and the second inequality is strict if $L_1 = 0$.

Let $D_k^2 = \mathbb{E}_{\xi_{0\ldots T}}\mathbf{dist}(x_k, X^*)^2$ denote the expected distance squared from each iterate to the set of minimizers X^*. From Lemma 2.3.1 with $x = x_k$, $y = \mathrm{proj}_{X^*}(x_k)$, and $\alpha = \alpha_k$, it follows that

$$\begin{aligned} D_{k+1}^2 &\leq D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[2\alpha_k(\mathbb{E}_\xi g(x_k;\xi))^T(x_k - y)\right] + \alpha_k^2\mathbb{E}_{\xi_{0\ldots T}}\|g(x_k,\xi_k)\|^2 \\ &\leq (1 - \mu\alpha_k/2)D_k^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[(\alpha_k - L_1\alpha_k^2)(f(x_k) - f^*)\right] + L_0^2\alpha_k^2, \end{aligned}$$

where the second inequality uses the quadratic growth of f and the bound on $\mathbb{E}_\xi\|g(x;\xi)\|^2$. Multiplying by $(k+1)/\alpha_k$ and invoking (2.11) yields

$$\begin{aligned} (k+1)\alpha_k^{-1}D_{k+1}^2 \leq &(k+1)(\alpha_k^{-1} - \mu/2)D_k^2 \\ &- \mathbb{E}_{\xi_{0\ldots T}}[(k+1)(1 - L_1\alpha_k)(f(x_k) - f^*)] + L_0^2(k+1)\alpha_k. \end{aligned}$$

Notice that this inequality telescopes due to (2.10). Inductively applying this implies

$$0 \leq (\alpha_0^{-1} - \mu/2)D_0^2 - \mathbb{E}_{\xi_{0\ldots T}}\left[\sum_{k=0}^{T}(k+1)(1 - L_1\alpha_k)(f(x_k) - f^*)\right] + L_0^2\sum_{k=0}^{T}(k+1)\alpha_k.$$

Since $\sum_{k=0}^{T}(k+1)\alpha_k \leq 4(T+1)/\mu$ and $\alpha_0^{-1} - \mu/2 = L_1^2/\mu$, we have

$$\mathbb{E}_{\xi_{0\ldots T}}\left[\sum_{k=0}^{T}(k+1)(1 - L_1\alpha_k)(f(x_k) - f^*)\right] \leq \frac{L_1^2D_0^2}{\mu} + \frac{4L_0^2(T+1)}{\mu}.$$

Observe that the coefficients of each $f(x_k) - f^*$ above are positive due to (2.11). Then the convexity of f completes the proof. □

Observe that this convergence rate is on the order of $O(1/T)$. To see this, we need to show the sum $\sum_{k=0}^{T}(k+1)(1-L_1\alpha_k)$ is at least $\Omega(T^2)$, which follows as

$$\begin{aligned}
\sum_{k=0}^{T}(k+1)(1-L_1\alpha_k) &= \sum_{k=0}^{T}(k+1) - \sum_{k=0}^{T}(k+1)L_1\alpha_k \\
&\geq \frac{(T+1)(T+2)}{2} - \sum_{k=0}^{T}\frac{4L_1}{\mu} \\
&= \frac{(T+1)(T+2)}{2} - (T+1)\frac{4L_1}{\mu} \\
&= \frac{(T+1)(T+2-8L_1/\mu)}{2}.
\end{aligned}$$

CHAPTER 3

NONCONVEX SUBGRADIENT METHOD CONVERGENCE RATES

3.1 Introduction

Stochastic approximation methods iteratively minimize the expectation of a family of known loss functions with respect to an unknown probability distribution. Such methods are of fundamental importance in machine learning, signal processing, statistics, and data science more broadly. For example, in machine learning, one is often interested in designing a classifier that performs well on the entire population of samples, given only a finite list of correctly labeled pairs $z_1, \ldots, z_n$ obtained from a fixed, but otherwise unknown distribution $\mathbb{P}$. Such problems can be formulated in our recurring form (1.1) as *population risk minimization*:

$$\text{minimize } F(x) := \begin{cases} \mathbb{E}_{z\sim\mathbb{P}}[f(x,z)] & \text{if } x \in \mathcal{X}; \\ \infty & \text{otherwise,} \end{cases} \tag{3.1}$$

Here, $\mathcal{X} \subseteq \mathbb{R}^d$ denotes a constraint set, while $f(x, z)$ represents the loss of the decision rule parameterized by $x \in \mathcal{X}$ on the population data z.

Much algorithmic development has been inspired by (3.1). Robbins-Monro's pioneering work 1951 work [143] developed the first method for solving (3.1) when each $f(\cdot, z)$ is smooth and strongly convex and $\mathcal{X} = \mathbb{R}^d$. This and most later methods are variants of the stochastic projected (sub)gradient method, which iteratively constructs approximate solutions x_t of (3.1) through the re-

cursion

$$\text{Sample } z_t \sim \mathbb{P}$$

$$\text{Set } x_{t+1} = \text{proj}_{\mathcal{X}}(x_t - \alpha_t \nabla_x f(x_t, z_t)),$$

where $z_1, \ldots, z_t, \ldots$ are i.i.d. and α_t is an appropriate control sequence. For nonsmooth $f(\cdot, z_t)$, sample gradients are simply replaced by sample subgradients $v_t \in \partial f(x_t, z_t)$, where $\partial f(x_t, z_t)$ denotes the subdifferential in the sense of convex analysis [148].

The complexity of minimizing (3.1) is directly related to the regularity of $f(\cdot, z)$. For example, for convex functions $f(\cdot, z)$ the stochastic subgradient method attains expected functional accuracy ε after $O(\varepsilon^{-2})$ stochastic subgradient evaluations. For strongly convex losses, the number of stochastic subgradient evaluations drops to $O(\varepsilon^{-1})$. The interested reader may turn to the seminal work [123] for an in-depth investigation of these methods and for information-theoretic lower bounds showing such rates are unimprovable without further assumptions.

For convex functions, complexity theory does not favor smooth losses over nonsmooth losses. For nonconvex problems, the situation is less clear. In the smooth case, the seminal work of Ghadimi, Lan, and Zhang [58] develops a variant of the stochastic projected gradient method and establishes that the expected norm of the projected gradient

$$\mathbb{E}_{z_1,\ldots,z_t}\left[\|x_t - \text{proj}_{\mathcal{X}}(x_t - \nabla_x \mathbb{E}_{z\sim P}[f(x_t, z)])\|^2\right], \tag{3.2}$$

a natural measure of stationarity, tends to zero at a controlled rate. Namely, with $O(\varepsilon^{-2})$ stochastic gradient evaluations, the algorithm produces a point with expected projected gradient norm squared less than ε.

At the time of writing the original version of this manuscript [36], there was no similar rate of convergence in the nonsmooth nonconvex setting for any known subgradient-based algorithm. Part of the difficulty in establishing a complexity theory for nonsmooth nonconvex subgradient-based methods is that the "usual criteria," namely the objective error and the norm of the gradient, can be completely meaningless. Indeed, on the one hand, one cannot expect the objective error $F(x_t) - \inf F$ to tend to zero—even in the smooth setting. On the other hand, simple examples, e.g., $F(x) = |x|$, show that $\mathbf{dist}(0, \partial F(x_t))$ can be strictly bounded below by a fixed constant for all t.

In contrast to subgradient-based methods, the "usual criteria" is meaningful for the *proximal point method* [147], which constructs a sequence x_t of approximate minimizers through the recursion

$$x_{t+1} = \underset{x \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ F(x) + \frac{1}{2\gamma} \|x - x_t\|^2 \right\},$$

where γ is a control parameter. Namely, it is a simple exercise to show that under minimal assumptions on F, the subdifferential distance $\mathbf{dist}(0, \partial F(x_t))$ tends to zero. Of course, each step of the proximal point method is difficult, if not impossible to execute without further assumptions on F.

The search for an appropriate class of functions F for which each proximal subproblem may be (approximately) executed naturally leads us to the deceptively simple, yet surprisingly broad class of *ρ-weakly convex* functions. Formally, a function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is *ρ-weakly convex* if

$$\text{the assignment } x \mapsto F(x) + \frac{\rho}{2}\|x\|^2 \text{ is convex.}$$

For example, any C^2 function on a compact, convex set becomes convex after adding the quadratic $\frac{|\lambda|}{2}\|\cdot\|^2$, where λ is the minimal eigenvalue of its Hessian

across all points in the set. In the nonsmooth setting, this class includes all *convex composite losses*

$$h(c(x))$$

where h is convex and L-Lipschitz and c is C^1 with β-Lipschitz Jacobian; such functions are known to be βL-weakly convex [39, Lemma 4.2]. The additive composite class is another widely used, much studied class of weakly convex functions, formed from all sums

$$g(x) + r(x)$$

where r is closed and convex and g is C^1 with β-Lipschitz gradient; such functions are known to be β-weakly convex. For further examples of weakly convex functions, see [32, Section 2.1], which includes formulations of robust phase retrieval, covariance matrix estimation, blind deconvolution, sparse dictionary learning, robust principal component analysis, and conditional value at risk. We provide several further examples in Section 3.2.1. It is important to note that none of these applications are covered by the seminal work of Ghadimi, Lan, and Zhang [58], which assumes an additive composite objective form.

Contributions. In this chapter, we develop the first known complexity guarantees for a subgradient-based method for a general class of nonsmooth nonconvex losses in stochastic optimization. The guarantees in this chapter apply to ρ-weakly convex losses F. Our algorithm, called the Proximally Guided stochastic Subgradient Method (PGSG) (Algorithm 2), follows an inner-outer loop strategy that may be compactly and informally summarized as

$$x_{t+1} = \varepsilon\text{-}\operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ F(x) + \rho \|x - x_t\|^2 \right\} \qquad \text{(in expectation)}. \tag{3.3}$$

The outer loop of PGSG is governed by the approximate proximal point method applied to the population risk F. Due to ρ-weak convexity of F, the inner loop

subproblem is a *strongly convex* stochastic optimization problem. Thus, by classical complexity theory, approximate solutions to the inner loop subproblems may be quickly found. We then turn our attention to establishing complexity guarantees.

As stated before, simple examples show that one cannot expect the iterates produced by a subgradient-based algorithm themselves to be ε-stationary because $\mathbf{dist}(0, \partial F(x_t))$ may be bounded below for all t. Instead, we introduce the following convergence measure: a (random) point $\bar{x}$ is an *ε-solution* if

$$\mathbb{E}\big[\mathbf{dist}(\bar{x}, \{x \mid \mathbf{dist}(0, \partial F(x))^2 \leq \varepsilon\})^2\big] \leq \varepsilon, \tag{3.4}$$

where ∂F is denotes the subdifferential of F in the sense of variational analysis [149]; see Section 3.3. We then show that when both inner and outer loops are coupled together appropriately, an outer-loop iterate chosen uniformly at random is an *ε-solution* after $O(\varepsilon^{-2})$ stochastic subgradient evaluations. The nearly stationary point nearby $\bar{x}$ is itself a solution to a *strongly convex* stochastic optimization problem. Thus, it is in principle obtainable to any desired degree of accuracy; see Remark 3.3.4.

Having established expectation guarantees, we turn our attention to probabilistic guarantees. Namely, following [57] (which considers the smooth case), we say a (random) point $\bar{x}$ is an *(ε, Λ)-solution* if

$$\mathbb{P}\big(\mathbf{dist}(\bar{x}, \{x \mid \mathbf{dist}(0, \partial F(x))^2 \leq \varepsilon\})^2 \leq \varepsilon\big) \geq 1 - \Lambda.$$

Markov's inequality shows that PGSG finds an (ε, Λ)-solution $\bar{x}$ after

$$O\left(\frac{1}{\Lambda^2 \varepsilon^2}\right)$$

stochastic subgradient evaluations. To improve this complexity, we introduce a

2-phase algorithm, called 2PGSG, which produces an (ε, Λ)-solution after

$$O\left(\frac{\log(1/\Lambda)}{\varepsilon^2} + \frac{\log(1/\Lambda)}{\Lambda\varepsilon}\right),$$

stochastic subgradient evaluations, substantially reducing the variance of our solution estimate. The technique for achieving this improvement is somewhat different than what [57] proposes in the smooth case. The challenge in establishing the result is that we no longer have unbiased estimates of subgradients at nearly stationary points. Indeed, the iterates produced by subgradient methods are only nearby nearly stationary points and are not nearly stationary themselves.

Finally, we turn our attention to a more practical variant of PGSG, which does not assume that the weak convexity constant ρ is known. In this setting, a simple idea—letting the outer loop stepsize tend to infinity—results in a point $\bar{x}$, which satisfies (3.4) after $O(\varepsilon^{2/(1-\beta)})$ stochastic subgradient evaluations, where $\beta \in (0, 1)$ is a user defined meta-parameter. We mention that the seminal work of Ghadimi, Lan, and Zhang [58] also assumes knowledge of the weak convexity constant ρ; in their setting ρ is simply the Lipschitz constant of the gradient.

We validate our results with some preliminary numerical experiments on the population objective of a robust real phase retrieval problem. We also discuss several more examples of weakly convex functions in Section 3.2.1.

3.1.1 Related Work

Stochastic Gradient Methods The convergence rates presented in [57] match known rates for the stochastic gradient method in nonconvex optimization.

There, the standard stochastic gradient method may be used without modification. Interestingly, recent work has developed methods which ensure $\mathbb{E}\|\nabla F\|^2 \leq \varepsilon$ after at most $O(\varepsilon^{-3/2})$ oracle calls [48]. This shows a surprising gap between smooth and nonsmooth nonconvex optimization not present in the convex case.

Stochastic Proximal-Gradient Methods For additive composite problems

$$\text{minimize}\{\mathbb{E}_z[f(x,z)] + r(x)\},$$

one often employs stochastic proximal-gradient methods, which require, at every iteration, a (potentially costly) evaluation of the mapping $\text{prox}_{,r}y = \text{argmin}\{r(x) + \frac{1}{2}\|x-y\|^2\}$. These methods achieve expected projected gradient norm ε, as in (3.2), after $O(\varepsilon^{-2})$ stochastic gradient evaluations [58]. These methods have also been extended to regularizers that are arbitrary closed prox-bounded functions r [171], a setting which we do not recover.

Evaluating the proximal mapping of r could be substantially more expensive than computing a subgradient. For example, if $r = \|\cdot\|_2$ is the spectral norm on $\mathbb{R}^{n\times n}$, then its proximal mapping requires a full singular value decomposition. In contrast, a subgradient may be computed from a single maximal eigenvector.

Another advantage of stochastic subgradient methods over stochastic proximal-gradient methods, is that multiple nonsmooth functions may be present in the objective function F. The same is not true for stochastic proximal-gradient methods: even if two functions r_1 and r_2 have simple proximal operators, the proximal operator of the sum $r = r_1 + r_2$ can be quite complex. Similarly, the proximal operator of an expectation $\mathbb{E}_z[r(x,z)]$ could be intractable.

Stochastic Methods for Convex Composite Recently [42] proposed a method for finding stationary points of the convex composite problem in which $f(x,z) = h(c(x,z),z)$. The first method adapts the prox-linear algorithm [22, 24, 21, 40, 97, 53] to the stochastic setting: given x_t, sample z_t and form x_{t+1} as the solution to the convex problem:

$$x_{t+1} = \operatorname*{argmin}_{x \in X} \left\{ h(c(x_t,z_t) + \nabla c(x_t,z_t)(x-x_t),z_t) + \frac{1}{2\gamma_t}\|x-x_t\|^2 \right\}, \qquad (3.5)$$

where $\gamma_t = \theta(1/\sqrt{t})$. The second proposed method is a straightforward application of the stochastic projected subgradient method [119]. Both methods are shown to almost surely converge to stationary points, but no rates of convergence are given.

We remark that the convergence proof presented in [42] is complex, being based on the highly nontrivial theory of nonconvex differential inclusions. We believe there is a benefit to having a simple proof of convergence, albeit for a slightly different subgradient method, which is what we provide in this chapter.

Further work on minimizing convex composite problems appears in [98, 168, 167]. This series of papers analyzes nested expectations: $F(x) = \mathbb{E}_v[h(\mathbb{E}_w[c(x,w) \mid v], v)]$. Although the stochastic structure considered in these papers is more general than what we consider in Problem 3.1, the assumptions made on F are much stronger than our assumptions on F. In particular, the authors prove rates under the assumption that (a) F is convex, (b) F is strongly convex, or (c) F is nonconvex, but *differentiable* with Lipschitz continuous gradient. For case (c), the authors propose an algorithm that finds an ε-stationary point of F after $O(\varepsilon^{-2.25})$ gradient evaluations [168] (in particular, they consider unconstrained problems).

Inexact Proximal Point Methods in Nonconvex Optimization The idea of using the inexact proximal point method to guide a nonconvex optimization algorithm to stationary points is not new. For example, Hare and Sagastizabal [71, 70] propose a method for computing inexact proximal points, which then enables the analysis of a nonconvex bundle method. The more recent work [133] exploits linearly convergent algorithms for solving the proximal subproblems. In contrast for the subproblems considered in this chapter, there are no linearly convergent stochastic subgradient algorithms capable of minimizing the proximal point step.

Subgradient Methods for Weakly Convex Problems This chapter is not the first to consider subgradient methods under weak convexity. For example, the early work [132] proves subsequential convergence of the (non projected) subgradient method for weakly convex *deterministic* problems. However, no rates were given in that work.

Almost Sure Convergence of Stochastic Subgradient Methods for Nonconvex Problems Convergence to stationary points of stochastic subgradient methods in nonsmooth, nonconvex optimization has previously been attained under several different scenarios, some of which are more general than the scenario considered in Problem (3.1) [154, 45, 46]. No rates of convergence were given in these works. In contrast, the novelty of the proposed approach lies in the attained rate of convergence, which matches the best known rates of convergence for smooth, nonconvex stochastic optimization [57].

Rates of Convergence in Stochastic Weakly Convex Optimization Since the first draft of these results appeared on arXiv in July 2017, several works appearing in 2018 have established convergence of the standard stochastic projected subgradient method under weak convexity [32, 31]. The obtained rates (in expectation) are essentially the same as those obtained in this chapter, namely they are of the form presented in equation (3.4). The authors of [32, 31] do not provide any probabilistic guarantees.

3.1.2 Outline

Section 3.2 presents notation and several basic results used in this chapter, as well as further examples of weakly convex functions. Section 3.3 presents our convergence analysis under the assumption that ρ is known. Section 3.3.2 presents our probabilistic guarantees. Section 3.3.3 presents our convergence analysis when ρ is unknown. Section 3.4 preliminary presents numerical results obtained on a robust phase retrieval problem.

3.2 Notation and Basic Results

Most of the notation and concepts we use in this chapter can be found in [149, 27]. Our main probabilistic assumption is that we work in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{R}^d$ is equipped with the Borel σ-algebra, which we use to define measurable mappings.

For a given function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, we let

$$\text{dom } F = \{x \in \mathbb{R}^d \mid F(x) < \infty\} \qquad \text{epi } F = \{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq t\}.$$

We say a function is *closed* if $\text{epi } F$ is a closed set. We say a function is *proper* if $\text{dom } F \neq \emptyset$.

Let $F \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed function. At any point $x \in \text{dom } F$, we let

$$\partial F(x) = \{v \in \mathbb{R}^d \mid \left(\forall y \in \mathbb{R}^d\right) F(y) \geq F(x) + \langle v, y - x\rangle + o(\|y - x\|)\}$$

denote the *Fréchet subdifferential* of F at x. On the other hand, if $x \notin \text{dom } F$ we let $\partial F(x) = \emptyset$. It is an easy exercise to show that at any local minimizer x of F, we have the inclusion $0 \in \partial F(x)$.

For the class of weakly convex functions, all elements of the subdifferential generate quadratic underestimators of the function F, as the following proposition shows. The equivalences are based on [29, Theorem 3.1].

Proposition 3.2.1 (Subgradients of Weakly Convex Functions)**.** *Suppose that $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a closed function. Then the following are equivalent*

1. *F is ρ-weakly convex. That is, $F + \frac{\rho}{2}\|\cdot\|^2$ is convex.*
2. *For any $x, y \in \mathbb{R}^d$ with $v \in \partial F(x)$, we have*

$$F(y) \geq F(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2. \tag{3.6}$$

3. *For all $x, y \in \mathbb{R}^d$ and $\alpha \in [0,1]$, we have*

$$F(\alpha x + (1-\alpha)y) \leq \alpha F(x) + (1-\alpha)f(y) + \frac{\rho\alpha(1-\alpha)}{2}\|x - y\|^2.$$

3.2.1 Examples of Weakly Convex Functions

As stated in the introduction, the class of weakly convex functions is broad. In the nonsmooth setting, this class includes all *convex composite losses*

$$h(c(x))$$

where h is convex and L-Lipschitz and c is C^1 with β-Lipschitz Jacobian; such functions are known to be βL-weakly convex [39, Lemma 4.2]. Several popular weakly convex formulations are presented in [32, Section 2.1]. We now discuss several further examples.

Example 3.2.2 (Censored Block Model)**.** *The censored block model [1] is a variant of the standard stochastic block model [2], which seeks to detect two communities in a partially observed graph. Mathematically, we encode such communities by forming the "community matrix" $M = \bar{\theta}\bar{\theta}^T \in \{-1, 1\}^d$, where $\bar{x} \in \{-1, 1\}^d$ is a membership vector in which $\bar{x}_i = 1$ if node i is in the first community, and $\bar{x}_i = -1$ otherwise. In the censored block model, we observe a randomly corrupted version $\hat{M}$ of the matrix M*

$$\hat{M} = \begin{cases} 0 & \text{with probability } 1 - p; \\ M_{ij} & \text{with probability } p(1 - \epsilon); \\ -M_{ij} & \text{with probability } p\epsilon. \end{cases}$$

Then our task is to recover M given only $\hat{M}$. We may formulate this problem in the following form convex, composite form:

$$F(x) = \sum_{ij|\hat{M}_{ij} \neq 0} |x_i x_j - \hat{M}_{ij}|.$$

Notice that absolute value function encourages the matrix to xx^T agree with $\hat{M}$ in most of its nonzero entries—the bulk of which are equal to M_{ij}—due to the sparsity promoting behavior of the nonsmooth absolute value function.

Example 3.2.3 (Robust Phase Retrieval)**.** *Phase retrieval is a common task in computational science; applications include imaging, X-ray crystallography, and speech processing. Given a set of tuples* $\{(a_i, b_i)\}_{i=1}^m \subset \mathbb{R}^d \times \mathbb{R}$*, the (real) phase retrieval problem seeks a vector* $x \in \mathbb{R}^d$ *satisfying* $(a_i^T x)^2 = b_i$ *for each index* $i = 1, \ldots, m$*. This problem is NP-hard [51]. Strictly speaking, phase retrieval is a feasibility problem. However, when the set of measurements* $\{b_i\}$ *is corrupted by gross outliers, one considers the following "robust" phase retrieval objective:*

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} |\langle a_i, x \rangle^2 - b_i|.$$

Notice that this nonsmooth objective is given in convex composite form, and therefore, it is weakly convex.

Example 3.2.4 (Nonsmooth Trimmed Estimation)**.** *Let* $f_1, \ldots, f_n$ *be Lipschitz continuous, convex loss functions on* $\mathbb{R}^d$*. The goal of trimmed estimation [152, 4, 110] is to fit a model while simultaneously detecting and removing "outlier" objectives* f_i*. Mathematically, we fix a number* $h \in \{1, \ldots, n\}$ *indicating the number of "inliers," and formulate the problem as follows:*

$$\begin{aligned} \text{minimize}_{x \in \mathbb{R}^d, w \in \mathbb{R}^n} \ & \sum_{i=1}^{n} w_i f_i(x) \\ \text{subject to: } & w_i \in [0, 1] \text{ and } \sum_{i=1}^{n} w_i = h. \end{aligned}$$

One can see that for fixed x*, the only objective values that contribute to the sum are those that are among the* h*-minimal elements of the set* $\{f_1(x), \ldots, f_n(x)\}$*. At the end of the chapter, we provide a short proof that this objective is weakly convex. Notice that it is in general nonconvex, despite each* f_i *begin convex.*

3.3 Proximally Guided Stochastic Subgradient Method

In this section, we formalize the proposed algorithm. First we slightly generalize the problem considered in the introduction, namely we assume that

$$\text{minimize}_{x\in\mathbb{R}^d}\ F(x) = \begin{cases} f(x) & \text{if } x \in \mathcal{X} \\ \infty & \text{otherwise.} \end{cases} \tag{3.7}$$

where f is a closed ρ-weakly convex function. Weak convexity of f implies that each of the proximal subproblems $\min_{x\in\mathbb{R}^d}\{F(x) + (1/2\gamma)\|x - x_t\|^2\}$ is

$$\mu := \gamma^{-1} - \rho$$

strongly convex. Next we introduce a stochastic subgradient oracle and a basic assumption on F.

Assumption 1. *Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and equip $\mathbb{R}^d$ with the Borel σ-algebra. Then we assume that*

(A1) *It is possible to generate IID realizations $z_1, z_2, \ldots$ from $\mathbb{P}$.*

(A2) *There is an open set $U \subseteq \mathbb{R}^d$ containing $\mathcal{X}$ and a measurable mapping $G : U \times \Omega \to \mathbb{R}^d$ such that $\mathbb{E}_z[G(x, z)] \in \partial f(x)$.*

(A3) *There is a constant $L \geq 0$ such that all $x \in U$ have $\mathbb{E}_z[\|G(x, z)\|^2] \leq L^2$.*

Assumption 1 is standard in the literature on stochastic subgradient methods. In particular, assumptions (A1) and (A2) are identical to assumptions (A1) and (A2) in [119], while assumption (A3) is identical to [119, Equation (2.5)]. A useful consequence of (A3) is that f itself is Lipschitz.

Lemma 3.3.1 (Lipschitz Continuity of f [32, Section 3.2])**.** *Suppose that assumption (A3) holds. Then f is L-Lipschitz continuous on U.*

The main workhorse of PGSG is a stochastic subgradient method for solving regularized subproblems $\min_{x\in\mathbb{R}^d}\{F(x) + (1/2\gamma)\|x - x_t\|^2\}$ induced by the proximal point method. We now state this method.

Algorithm 1 Projected Stochastic Subgradient Method for Proximal Point Subproblems $\mathrm{PSSM}(y_0, G, \gamma, \{\alpha_t\}, J)$

Require: $y_0 \in \mathcal{X}$, quadratic multiplier $\gamma > 0$, maximum iterations $J \in \mathbb{N}$, nonnegative stepsize sequence $\{\alpha_t\}$.

1: **for** $j = 0, \ldots, J-2$ **do**
2: Sample z_j and set $v_j = G(y_j, z_j) + \frac{1}{\gamma}(y_j - y_0)$
3: $y_{j+1} = \mathrm{proj}_{\mathcal{X}}(y_{t,j} - \alpha_j v_j)$
4: **end for**
5: **return** $\tilde{y} = \frac{2}{J(J+1)}\sum_{j=0}^{J-1}(j+1)y_j$.

Before introducing the Proximally Guided stochastic Subgradient (PGSG) method, we introduce two necessary algorithm parameters:

$$j_t \geq \frac{11}{\gamma^2\mu^2}; \tag{3.8}$$

$$\alpha_j = \frac{2}{\mu\left(j + 2 + \dfrac{36}{\gamma^4\mu^4(j+1)}\right)}. \tag{3.9}$$

The algorithm now follows.

Algorithm 2 Proximally Guided Stochastic Subgradient Method $\mathrm{PGSG}(y_0, G, \gamma, \{j_t\}, T)$

Require: $x_0 \in \mathcal{X}$, weak convexity constant $\rho > 0$, $\gamma \in (0, 1/\rho)$, maximum iterations $T \in \mathbb{N}$, maximum inner loop iteration $\{j_t\}$ satisfying (3.8).

1: Define the stepsize sequence $\{\alpha_j\}$ as in (3.9)
2: **for** $t = 0, \ldots, T-2$ **do**
3: $x_{t+1} = \mathrm{PSSM}(x_t, G, \gamma, \{\alpha_j\}, j_t)$
4: **end for**
5: **return** x_R, where R is sampled uniformly from $\{0, \ldots, T-1\}$.

As stated in the introduction PGSG employs an inner-outer loop strategy, which is shown in Algorithm 2. The outer loop executes $T - 1$ approximate

proximal point steps, resulting in the iterates $\{x_t\}$. The inner loop, shown in Algorithm 1, approximately solves the proximal point subproblem, which is now strongly convex, using a stochastic subgradient method for strongly convex optimization [89]. Beyond its use in governing the outer loop dynamics of PGSG, the proximal point subproblems also lead to a natural measure of stationarity.

Indeed, for all $t \in \mathbb{N}$, define the proximal point

$$\hat{x}_t := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ F(x) + \frac{1}{2\gamma} \|x - x_t\|^2 \right\}. \tag{3.10}$$

Note that $\hat{x}_t$ exists and is unique by the μ-strong convexity of the proximal subproblem. We stress that this point, although in principle obtainable via convex optimization, is never computed. Instead it is only used to formulate convergence guarantees. To that end, the following Lemma shows that the gap $\gamma^{-1}\|x_t - \hat{x}_t\|$ is a natural measure of stationarity.

Lemma 3.3.2 (Convergence Criteria)**.** *Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed function. Let $x \in \mathbb{R}^d$. If*

$$\hat{x} \in \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ F(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\},$$

then we have the bound

$$\mathbf{dist}(x, \{y \in \mathbb{R}^d \mid \mathbf{dist}(0, \partial F(y))^2 \leq \gamma^{-2}\|x - \hat{x}\|^2\}) \leq \|x - \hat{x}\|^2. \tag{3.11}$$

Proof. As $\hat{x}$ is a minimizer, we have

$$0 \in \partial \left[F(\cdot) + \frac{1}{2\gamma} \|\cdot - x\|^2 \right](y) = \partial F(y) + \frac{1}{\gamma}(y - x),$$

where the second equality follows by the sum rule for a smooth additive term $(2\gamma)^{-1}\|\cdot - x\|^2$ [149]. Thus, we have the inclusion $\hat{x} \in \{y \in \mathbb{R}^d \mid \mathbf{dist}(0, \partial F(y))^2 \leq \gamma^{-2}\|x - \hat{x}\|^2\}$, which leads to the desired conclusion. □

Based on this Lemma, the iterate x_t is ε-close to an ε-stationary point in expectation whenever

$$\mathbb{E}\|x_t - \hat{x}_t\|^2 \leq \min\{\varepsilon, \gamma^2\varepsilon\}.$$

Establishing this fact is the main technical goal of the following theorem.

Theorem 3.3.3 (Convergence of PGSG)**.** *Let $x_0 \in \mathcal{X}$, consider any $T \in \mathbb{N}$, and let $x_R = \mathrm{PGSG}(x_0, G, \gamma, \{j_t\}, T)$ Define the quantity*

$$\mathcal{B}_{T,\{j_t\}} := \frac{4}{T\mu}\left(F(x_0) - \inf F + \sum_{t=0}^{T-1} \frac{72L^2}{\mu(j_t+1)}\right).$$

Then $\mathbb{E}\|x_R - \hat{x}_R\|^2 \leq \mathcal{B}_{T,\{j_t\}}$. Consequently, we have the following bound:

$$\mathbb{E}\big[\mathbf{dist}(x_R, \{x \mid \mathbf{dist}(0, \partial F(x))^2 \leq \gamma^{-2}\mathcal{B}_{T,\{j_t\}}\})^2\big] \leq \mathcal{B}_{T,\{j_t\}}$$

In particular, given $\Delta \geq F(x_0) - \inf F$, and setting

$$j_t := \left\lceil \max\left(\frac{576L^2}{\mu^2 \min\{\varepsilon, \varepsilon\gamma^2\}}, \frac{11}{\gamma^2\mu^2}\right)\right\rceil \qquad \textit{and} \qquad T := \left\lceil \frac{4\Delta}{\mu \min\{\varepsilon, \varepsilon\gamma^2\}}\right\rceil,$$

we have

$$\mathbb{E}\big[\mathbf{dist}(x_R, \{x \mid \mathbf{dist}(0, \partial F(x))^2 \leq \varepsilon)^2\big] \leq \varepsilon.$$

The total number of stochastic oracle evaluations required to compute this point is bounded by $j_t \cdot T = O(\Delta L^2 \varepsilon^{-2})$.

Remark 3.3.4 (Obtaining a Nearly Stationary Point)**.** *As stated, the theorem indicates that x_R is nearby a nearly stationary point. The proof of Lemma 3.3.2 shows that one can in principle obtain the nearly stationary point $\hat{x}_R$ by solving the* strongly convex *stochastic optimization problem*

$$\hat{x}_R = \operatorname*{argmin}_{x \in \mathbb{R}^d}\Big\{F(x) + \tfrac{1}{2\gamma}\|x - x_R\|^2\Big\},$$

which is solvable to any desired degree of accuracy (in expectation). Furthermore, Lemma 3.3.2 shows that one can estimate the degree of stationarity of $\hat{x}_R$ by the bound

$\mathbf{dist}(0, \partial F(\hat{x}_R))^2 \leq \gamma^{-2}\|x_R - \hat{x}_R\|^2$. *In particular, given an estimate,* $\tilde{x}_R \approx \hat{x}_R$*, we have the bound* $\mathbf{dist}(0, \partial F(\hat{x}_R))^2 \leq 2\gamma^{-2}\|x_R - \tilde{x}_R\|^2 + 2\gamma^{-2}\|\tilde{x}_R - \hat{x}_R\|^2$*, which indicates that* $2\gamma^{-2}\|x_R - \tilde{x}_R\|^2$ *may serve as a bound on the true stationarity of* $\hat{x}_R$ *(up to tolerance* $2\gamma^{-2}\|\tilde{x}_R - \hat{x}_R\|^2$*).*

3.3.1 Proof of Theorem 3.3.3

Throughout the proof we will need the following bound on the proximal point step:

Lemma 3.3.5 (Bounded Steplengths)**.** *Let* $\gamma > 0$*,* $x \in \mathcal{X}$*, and suppose that*

$$\hat{x} \in \operatorname*{argmin}_{y \in \mathbb{R}^d} \Big\{ F(y) + \tfrac{1}{2\gamma}\|y - x\|^2 \Big\}.$$

Then $\gamma^{-1}\|x - \hat{x}\| \leq 2L$*.*

Proof. Note that

$$\frac{1}{2\gamma}\|x - \hat{x}\|^2 \leq F(x) - F(\hat{x}) \leq L\|x - \hat{x}\|,$$

where Lipschitz continuity follows from Lemma 3.3.1. Divide both sides of the inequality by $\frac{1}{2}\|x - \hat{x}\|$ to get the result. □

We now analyze one inner loop of Algorithm 2. This inner loop may be interpreted as a variant of the stochastic projected subgradient method applied to the strongly convex optimization problem,

$$\text{minimize}_{x \in \mathbb{R}^d}\ F_y(x) := F(x) + \tfrac{1}{2\gamma}\|x - y\|^2,$$

We note that the following proof is similar in outline to [89], but the results of that work are not sufficient for our purposes.

Proposition 3.3.6 (Analysis of PSSM). *Let $y \in \mathcal{X}$ and let $\hat{y}$ be the unique minimize of $F_y(x)$ over all $x \in \mathbb{R}^d$. Set $\tilde{y} = \mathrm{PSSM}(y, G, \gamma, \{\alpha_j\}, J)$. Then if $\gamma \in (0, 1/\rho)$ and $\{\alpha_j\}$ is chosen as in* (3.9)*, we have*

$$\begin{aligned}
\mathbb{E}[F_y(\tilde{y}) - F_y(\hat{y}))] &\leq \frac{72L^2}{\mu(J+1)} + \frac{30\|y-\hat{y}\|^2}{\gamma^4\mu^3 J(J+1)};\\
\mathbb{E}\big[\|\tilde{y}-\hat{y}\|^2\big] &\leq \frac{144L^2}{\mu^2(J+1)} + \frac{60\|y-\hat{y}\|^2}{\gamma^4\mu^4 J(J+1)};\\
\mathbb{E}\big[\|y-\tilde{y}\|^2\big] &\leq \frac{288L^2}{\mu^2(J+1)} + \left(2 + \frac{120}{\gamma^4\mu^4 J(J+1)}\right)\|y-\hat{y}\|^2.
\end{aligned}$$

On the other hand, if $0 < \alpha_j \leq 2\gamma$ for all j, but $\{\alpha_j\}$ and γ are otherwise unconstrained, we have

$$\mathbb{E}[\|y-\tilde{y}\|] \leq L\sum_{i=0}^{J-1}\alpha_i.$$

Proof. Since $\hat{y} \in \mathcal{X}$ and $\mathrm{proj}_{\mathcal{X}}$ is nonexpansive, we have

$$\begin{aligned}
\|y_{j+1}-\hat{y}\|^2 &\leq \|y_j - \alpha_j v_j - \hat{y}\|^2\\
&= \|y_j-\hat{y}\|^2 - 2\alpha_j\langle y_j-\hat{y}, v_j\rangle + \alpha_j^2\|v_j\|^2. \qquad (3.12)
\end{aligned}$$

To proceed further, we must bound $\|v_j\|^2$. To that end, recall that F_y is μ-strongly convex. Therefore, for any $x \in \mathcal{X}$,

$$\begin{aligned}
\mathbb{E}_z\left\|G(x,z) + \frac{1}{\gamma}(x-y)\right\|^2 &= \mathbb{E}_z\left\|G(x,z) - \frac{1}{\gamma}(\hat{y}-y) + \frac{1}{\gamma}(x-\hat{y})\right\|^2\\
&\leq \mathbb{E}_z 3\|G(x,z)\|^2 + 3\left\|\frac{1}{\gamma}(\hat{y}-y)\right\|^2 + 3\left\|\frac{1}{\gamma}(x-\hat{y})\right\|^2\\
&\leq 15L^2 + 3\left\|\frac{1}{\gamma}(x-\hat{y})\right\|^2\\
&\leq 15L^2 + \frac{6}{\gamma^2\mu}(F_y(x) - F_y(\hat{y})),
\end{aligned}$$

where the first inequality follows from Jensen's inequality, the second inequality uses (A3) twice and Lemma 3.3.5, and the third inequality follows from the strong convexity.

Returning to Equation (3.12), we let $\bar{v}_j = \mathbb{E}_j v_j \in \partial F_y(y_j)$, where $\mathbb{E}_j[\cdot]$ denotes the expectation conditioned on $y_1, \ldots, y_j$. Now, we take the conditional expectation of both sides of the equation, yielding

$$\begin{aligned}
\mathbb{E}_j\|y_{j+1} - \hat{y}\|^2 &\le \mathbb{E}_j\|y_j - \hat{y}\|^2 - 2\alpha_j\langle y_j - \hat{y}, \bar{v}_j\rangle + \alpha_j^2\mathbb{E}_j\|v_j\|^2 \\
&\le \mathbb{E}_j\|y_j - \hat{y}\|^2 + \alpha_j^2\left(15L^2 + \frac{6}{\gamma^2\mu}\mathbb{E}_j F_y(y_j) - F_y(\hat{y})\right) \\
&\quad - 2\alpha_j\left(\mathbb{E}_j F_y(y_j) - F_y(\hat{y}) + \frac{\mu}{2}\mathbb{E}_j\|y_j - \hat{y}\|^2\right) \\
&= (1 - \alpha_j\mu)\mathbb{E}_j\|y_j - \hat{y}\|^2 + 15\alpha_j^2L^2 - \left(2\alpha_j - \frac{6\alpha_j^2}{\gamma^2\mu}\right)(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})) \\
&\le (1 - \alpha_j\mu)\mathbb{E}_{t,j}\|y_j - \hat{y}\|^2 + 15\alpha_j^2L^2 - \alpha_j(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})),
\end{aligned}$$

where the second inequality uses our bound on $\mathbb{E}_z\|G(x,z) + \gamma^{-1}(x-y)\|^2$ and the strong convexity of F_y, and the third inequality is a consequence of the bound:

$$\frac{6\alpha_j}{\gamma^2\mu} = \frac{2\mu(j+2)(6/\gamma^2\mu)}{(\mu(j+2))^2 + \frac{36}{\gamma^4\mu^2}\frac{j+2}{j+1}} \le \frac{2\mu(j+2)(6/\gamma^2\mu)}{(\mu(j+2))^2 + (6/\gamma^2\mu)^2} \le 1.$$

Multiplying by $(j+1)/\alpha_j$, we find that

$$\begin{aligned}
(j+1)\alpha_j^{-1}\mathbb{E}_j\|y_{j+1} - \hat{y}\|^2 \le (j+1)\big(\alpha_j^{-1} - \mu\big)\mathbb{E}_j\|y_j - \hat{y}\|^2 + 15(j+1)\alpha_j L^2 \\
- (j+1)(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})).
\end{aligned}$$

By our choice of α_j, we have $(j+1)\alpha_j^{-1} = (j+2)(\alpha_{j+1}^{-1} - \mu)$. Therefore, summing the previous inequality, we have

$$0 \le \big(\alpha_0^{-1} - \mu\big)\|y - \hat{y}\|^2 + 15L^2\sum_{j=0}^{J-1}(j+1)\alpha_j - \sum_{j=0}^{J-1}(j+1)(\mathbb{E}_j F_y(y_j) - F_y(\hat{y})).$$

Therefore, noting that $\sum_{j=0}^{j_t-1}(j+1)\alpha_j \le 2j_t/\mu$ and $\alpha_0^{-1} - \mu = 18/(\gamma^4\mu^3)$, and using the convexity of F_y, we deduce

$$\mathbb{E}(F_y(\tilde{y}) - F_y(\hat{y})) \le \frac{36\|y - \hat{y}\|^2}{\gamma^4\mu^3 J(J+1)} + \frac{60L^2}{\mu(J+1)}.$$

The first distance bound then follows as a direct consequence of the strong convexity of F_y, while the second follows from the convexity of $\|\cdot\|^2$.

Finally, we now work in the case in which γ may be strictly greater than $1/\rho$. We claim that for all $j = 0, \ldots, J-1$, we have $\mathbb{E}[\|y_j - y_0\|] \leq L\sum_{i=0}^{j}\alpha_i$. Indeed, this is clearly true for $j = 0$. Inductively, we also have

$$\begin{aligned}\mathbb{E}_j\|y_{j+1} - x_t\| &\leq \mathbb{E}_j\|y_j - \alpha_j(G(y_j, z_j) + (y_j - x_t)/\gamma) - x_t\| \\ &\leq |1 - \alpha_j/\gamma| \cdot \mathbb{E}_j\|y_j - x_t\| + \alpha_j\mathbb{E}_{\Xi_0}\|G(y_j, z_j)\| \\ &\leq \mathbb{E}_j\|y_j - x_t\| + \alpha_j L,\end{aligned}$$

where the first inequality follows by nonexpansiveness of $\mathrm{proj}_{\mathcal{X}}$ and the third follow from the inequality $0 < \alpha_j \leq 2\gamma$. Applying the law of expectation completes the inductive step. Therefore, we have

$$\begin{aligned}\mathbb{E}[\|y - \tilde{y}\|] \leq \mathbb{E}\left[\frac{2}{J(J+1)}\sum_{j=0}^{J-1}(j+1)\|y_j - y\|\right] &\leq \frac{2}{J(J+1)}\sum_{j=0}^{J-1}(j+1)\left(L\sum_{i=0}^{j}\alpha_i\right) \\ &\leq L\sum_{i=0}^{J-1}\alpha_i,\end{aligned}$$

as desired. □

We now give the proof of Theorem 3.3.3.

Proof of Theorem 3.3.3. By the strong convexity of the proximal point subproblem, we have

$$F(\hat{x}_t) \leq F(x_t) - \left(\frac{1}{2\gamma} + \frac{\mu}{2}\right)\|\hat{x}_t - x_t\|^2.$$

Then by Proposition 3.3.6, we have the following bound:

$$\begin{aligned}\mathbb{E}_t[F(x_{t+1})] &\leq F(\hat{x}_t) + \frac{1}{2\gamma}\|\hat{x}_t - x_t\|^2 + \frac{72L^2}{\mu(j_t+1)} + \frac{30\|x_t - \hat{x}_t\|^2}{\gamma^4\mu^3 j_t(j_t+1)} \\ &\leq F(x_t) + \frac{72L^2}{\mu(j_t+1)} - \left(\frac{\mu}{2} - \frac{30}{\gamma^4\mu^3 j_t(j_t+1)}\right)\|x_t - \hat{x}_t\|^2,\end{aligned}$$

where $\mathbb{E}_t[\cdot]$ denotes the expectation conditioned on $x_1, \ldots, x_t$. Rearranging, using the lower bound on j_t (which makes the multiple of $\|\hat{x}_t - x_t\|^2$ larger than $\mu/4$ as $30/121 < 1/4$), applying the law of total expectation, and summing, we find that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|x_t - \hat{x}_t\|^2\big] \leq \frac{4}{T\mu}\left(F(x_0) - \inf F + \sum_{t=0}^{T-1}\frac{72L^2}{\mu(j_t+1)}\right),$$

as desired. To complete the proof, apply Lemma 3.3.2. □

3.3.2 Probabilistic Guarantees

In the previous section, we developed expected complexity results, which describe the average behavior of the PGSG over multiple runs. We are also interested in giving guarantees for a single run of an algorithm. Thus, in this section we recall the notion of an (ε, Λ)-solution given in the introduction: a random variable $\bar{x}$ is called an (ε, Λ)-*solution* if

$$\mathbb{P}\big(\mathbf{dist}(\bar{x}, \{x \mid \mathbf{dist}(0, \partial F(x))^2 \leq \varepsilon\})^2 \leq \varepsilon\big) \geq 1 - \Lambda.$$

Theorem 3.3.3 together with Markov's inequality implies that x_R, generated with

$$j_t := \left\lceil \max\left(\frac{576L^2}{\mu^2 \min\{\varepsilon\Lambda, \varepsilon\Lambda\gamma^2\}}, \frac{12}{\gamma^2\mu^2}\right)\right\rceil \quad \text{and} \quad T := \left\lceil \frac{4\Delta}{\mu \min\{\varepsilon\Lambda, \varepsilon\Lambda\gamma^2\}}\right\rceil,$$

where $\Delta \geq F(x_0) - \inf F$, is an (ε, Λ)-solution after

$$j_t \cdot T = O(\Delta L^2 (\varepsilon\Lambda)^{-2}) \tag{3.13}$$

stochastic oracle evaluations. In this section, we develop a two stage algorithm that significantly improves the dependence on Λ in this bound.

The method we propose proceeds in two phases. In the first phase, multiple independent copies of PGSG are called, resulting in candidates $x_{R^1}, \ldots, x_{R^S}$. For each of the candidates, we then compute an approximate proximal point $\tilde{x}_{R^s} \approx \hat{x}_{R^s}$. In the second phase, we select one of the candidates $x_{R^{\bar{s}}}$ based on the size of $\gamma^{-1}\|x_{R^s} - \tilde{x}_{R^s}\|$, a proxy for the true proximal step length. We will see that such a point is (ε, Λ)-solution, and the total number of stochastic oracle evaluations has a much better dependence on Λ.

Before we introduce the algorithm, let us define three parameters

$$j_t := \left\lceil \max\left\{ \frac{576L^2}{\mu^2 \min\{\varepsilon/24, \varepsilon\gamma^2/24\}}, \frac{11}{\gamma^2\mu^2} \right\} \right\rceil, \qquad T := \left\lceil \frac{4\Delta}{\mu \min\{\varepsilon/24, \varepsilon\gamma^2/24\}} \right\rceil, \tag{3.14}$$

and

$$J := \left\lceil \max\left\{ \frac{48L^2\sqrt{2}}{\mu \min\{\varepsilon, \varepsilon\gamma^2\}} \cdot \frac{S}{\Lambda}, \frac{11}{\gamma^2\mu^2} \cdot \sqrt{\frac{S}{\Lambda}} \right\} \right\rceil,$$

where $\Delta \geq F(x_0) - \inf F$. The algorithm now follows.

Algorithm 3 Two Phase Proximally Guided Stochastic Subgradient Method $2\mathrm{PGSG}(x_0, G, \gamma, J, S)$

Require: $x_0 \in \mathcal{X}$, weak convexity constant $\rho > 0$, $\gamma \in (0, 1/\rho)$, Stochastic Subgradient Iteration $J \in \mathbb{N}$, number of copies $S \in \mathbb{N}$.

1: Define the maximum iterations T as in (3.14)
2: Define the maximum inner loop iteration $\{j_t\}$ as in (3.14)
3: Define the stepsize sequence $\{\alpha_j\}$ as in (3.9)
4: **Optimization Phase**
5: **for** $s = 1, \ldots, S$ **do**
6: Set $x_{R^s} = \mathrm{PGSG}(x_0, G, \gamma, \{j_t\}, T)$.
7: Set $\tilde{x}_{R^s} = \mathrm{PSSM}(x_{R^s}, G, \gamma, \{\alpha_j\}, J)$.
8: **end for**
9: **Post-Optimization Phase**
10: Choose $x^* = x_{R^{\bar{s}}}$ from the candidate list $\{x_r^s\}_{s=1}^S$ such that

$$\bar{s} = \operatorname*{argmin}_{s=1,\ldots,S} \|x_{R^s} - \tilde{x}_{R^s}\|$$

11: **return** x^*

The analysis of this algorithm requires a bound on the expectation of $\|x_{R^s} - \tilde{x}_{R^s}\|^2$ and $\|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2$, which we now provide.

Lemma 3.3.7. *Let x_{R^s} be generated as in Algorithm 3. Then*

$$\mathbb{E}\big[\|x_{R^s} - \tilde{x}_{R^s}\|^2\big] \leq \frac{1}{4}\min\{\varepsilon, \gamma^2\varepsilon\};$$
$$\mathbb{E}\big[\|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2\big] \leq \frac{\Lambda}{4S}\min\{\varepsilon, \gamma^2\varepsilon\}$$

Proof. By Proposition 3.3.6 and Theorem 3.3.3, the bound holds:

$$\begin{aligned}\mathbb{E}\big[\|x_{R^s} - \tilde{x}_{R^s}\|^2\big] &\leq \frac{288L^2}{\mu^2(J+1)} + \left(2 + \frac{120}{\gamma^4\mu^4 J(J+1)}\right)\mathbb{E}\big[\|x_{R^s} - \hat{x}_{R^s}\|^2\big] \\ &\leq \frac{288L^2}{\mu^2(J+1)} + \left(2 + \frac{120}{\gamma^4\mu^4 J(J+1)}\right)\mathcal{B}_{T,\{j_t\}} \\ &\leq \frac{\Lambda}{8S}\min\{\varepsilon, \gamma^2\varepsilon\}/8 + \min\{\varepsilon, \gamma^2\varepsilon\}/8 \leq \min\{\varepsilon, \gamma^2\varepsilon\}/4,\end{aligned}$$

which proves the first bound.

On the other hand, Proposition 3.3.6 and Theorem 3.3.3 imply that

$$\begin{aligned}\mathbb{E}\big[\|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2\big] &\leq \frac{144L^2}{\mu^2(J+1)} + \frac{60}{\gamma^4\mu^4 J(J+1)}\mathbb{E}\big[\|x_{R^s} - \hat{x}_{R^s}\|^2\big] \\ &\leq \frac{144L^2}{\mu^2(J+1)} + \frac{60}{\gamma^4\mu^4 J(J+1)}\mathcal{B}_{T,\{j_t\}} \\ &\leq \frac{\Lambda}{8S}\min\{\varepsilon, \gamma^2\varepsilon\} + \frac{\Lambda}{8S}\min\{\varepsilon, \gamma^2\varepsilon\} = \frac{\Lambda}{4S}\min\{\varepsilon, \gamma^2\varepsilon\},\end{aligned}$$

which proves the second bound and completes the proof. □

We now state the convergence guarantees for Algorithm 3.

Theorem 3.3.8. *Let $x_0 \in \mathcal{X}$ and let $S = \log_2(2/\Lambda)$. Then $x^* = 2\text{PGSG}(x_0, G, \gamma, J, S)$ returned by Algorithm 3 is an (ε, Λ)-solution. The total number of stochastic oracle evaluations called by Algorithm 3 is equal to*

$$S \cdot (j_t \cdot T + J) = O\left(\frac{\log_2(1/\Lambda)\Delta L^2}{\varepsilon^2} + \frac{\log_2(1/\Lambda)L^2}{\varepsilon\Lambda}\right). \tag{3.15}$$

Proof. By Lemma 3.3.2, it suffices to show that

$$\mathbb{P}\big(\|x^* - \hat{x}^*\|^2 \leq \min\{\varepsilon, \gamma^2\varepsilon\}\big) \geq 1 - \Lambda.$$

To that end, note that

$$\begin{aligned}\|x^* - \hat{x}^*\|^2 &= \|(x_{R^{\bar{s}}} - \hat{x}_{R^{\bar{s}}})\|^2 \\ &\leq 2\|x_{R^{\bar{s}}} - \tilde{x}_{R^{\bar{s}}}\|^2 + 2\|\tilde{x}_{R^{\bar{s}}} - \hat{x}_{R^{\bar{s}}}\|^2 \\ &\leq 2 \min_{s=1,\ldots,S} \|x_{R^s} - \tilde{x}_{R^s}\|^2 + 2 \max_{s=1,\ldots,S} \|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{P}\big(\|x^* - \hat{x}^*\|^2 \geq \min\{\varepsilon, \gamma^2\varepsilon\}\big) \leq \mathbb{P}&\left\{ \min_{s=1,\ldots,S} \|x_{R^s} - \tilde{x}_{R^s}\|^2 \geq \frac{1}{2}\min\{\varepsilon, \gamma^2\varepsilon\} \right) \\ &+ \mathbb{P}\left(\max_{s=1,\ldots,S} \|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2 \geq \frac{1}{2}\min\{\varepsilon, \gamma^2\varepsilon\} \right).\end{aligned}$$

Notice that by Markov's inequality, independence, and Proposition 3.3.7, we have:

$$\mathbb{P}\left(2 \min_{s=1,\ldots,S} \|x_{R^s} - \tilde{x}_{R^s}\|^2 \geq \frac{1}{2}\min\{\varepsilon, \gamma^2\varepsilon\} \right) \leq 2^{-S} \leq \frac{\Lambda}{2}.$$

On the other hand, by Markov's inequality, a union bound, and Proposition 3.3.7, we have

$$\mathbb{P}\left(2 \max_{s=1,\ldots,S} \|\tilde{x}_{R^s} - \hat{x}_{R^s}\|^2 \geq \frac{1}{2}\min\{\varepsilon, \gamma^2\varepsilon\} \right) \leq \frac{\Lambda}{2},$$

which shows that x^* is an (ε, Λ)-solution. □

When the second term in (3.15) is dominating, the obtained bound (3.15) is $\log_2(2/\Lambda)/\varepsilon\Lambda$ times smaller than the bound (3.13) obtained by the PGSG algorithm.

3.3.3 PGSG with Unknown Weak Convexity Constant

Algorithm 2 requires that the parameters ε, L, ρ, and Δ are known. In practice, computing bounds on L, ρ, and Δ may be nontrivial. In this section we show that a simple strategy—letting j_t tend to infinity and γ_t tend to zero—results in a sublinear convergence rate without knowledge of any problem parameters. We formalize this procedure in Algorithm 4 using the following parameters: fix a hyper-parameter $0 < \beta < 1$, and define

$$\gamma_t := (t+1)^{-\beta}; \tag{3.16}$$

$$j_t := t + 44; \tag{3.17}$$

$$\alpha_{t,j} := \frac{4\gamma_t}{j + 1 + \frac{288}{j+1}}. \tag{3.18}$$

The algorithm now follows.

Algorithm 4 Parameter Free Proximally Guided Stochastic Subgradient Method $\mathrm{PFPGSG}(y_0, G, T, \beta)$

Require: $x_0 \in \mathcal{X}$, maximum iterations $T \in \mathbb{N}$, hyper-parameter $0 < \beta < 1$.

1: Define the sequence $\{\gamma_t\}$ as in (3.16)
2: Define the stepsize sequence $\{\alpha_{t,j}\}$ as in (3.18)
3: Define the maximum inner loop iteration $\{j_t\}$ as in (3.17)
4: **for** $t = 0, \dots, T-2$ **do**
5: $\quad x_{t+1} = \mathrm{PSSM}(x_t, G, \gamma_t, \{\alpha_{t,0}, \alpha_{t,1}, \alpha_{t,2}, \dots\}, j_t)$
6: **end for**
7: **return** x_R, where R is sampled with probability $\mathbb{P}(R = t) \propto \gamma_t$ from $\{0, \dots, T-1\}$.

In the following, we establish convergence guarantees for the parameter free variant of PGSG. The proof splits the analysis of PFPGSG into two parts. In the first part, $\gamma_t \geq 1/\rho$. In this setting, the analysis of the previous section does not apply. Thus, we show that that the iterates do not wander very far. In the second part, $\gamma_t \leq 1/\rho$, and an argument similar to the one presented in Theorem 3.3.3

applies. Combining these results then leads to the theorem. To that end, we address the first part now.

Lemma 3.3.9. *Let* $T_0 = \lceil (2\rho)^{1/\beta} \rceil$. *Then*

$$\mathbb{E}[F(x_{T_0})] \leq F(x_0) + L^2 T_0 \log(T_0 + 125).$$

Proof. By Proposition 3.3.6, as $\alpha_j < 2\gamma_t$, we have we have $\mathbb{E}_{T_0}\|x_{t+1} - x_t\| \leq L \sum_{j=0}^{j_t - 1} \alpha_j$ for all $t = 0, \ldots, T_0 - 1$.

$$\begin{aligned}
\mathbb{E}_{T_0} F(x_{T_0}) &\leq F(x_0) + L\mathbb{E}_{T_0}\|x_{T_0} - x_0\| \\
&\leq F(x_0) + L \sum_{t=0}^{T_0 - 1} \mathbb{E}_{T_0}\|x_{t+1} - x_t\| \\
&\leq F(x_0) + L^2 \sum_{t=0}^{T_0 - 1} \sum_{j=0}^{j_t - 1} \alpha_{t,j} \\
&\leq F(x_0) + L^2 \sum_{t=0}^{T_0 - 1} \sum_{j=0}^{j_t - 1} \frac{4}{j+1} \\
&\leq F(x_0) + L^2 T_0 \log(T_0 + 125), \qquad (3.19)
\end{aligned}$$

as desired. □

We now address the second part of the argument, and with it, deduce the following theorem. At first glance, the presented rate appears to be better than the rate obtained by Algorithm 2, which requires knowledge of ρ. However, it is not because the factor $\gamma_R^{-2} = (R+1)^{2\beta}$ is no longer a constant. Instead, the convergence rate of Algorithm 4 is on the order of $O(T^{1-\beta})$ in the worst case.

Theorem 3.3.10 (Convergence of Parameter Free PGSG)**.** *Let* $T_0 = \lceil (2\rho)^{1/\beta} \rceil$. *And consider any* $T \in \mathbb{N}$. *Let* $x_R = \mathrm{PFPGSG}(x_0, G, T, \beta)$. *Define the quantity*

$$\mathcal{C}_{T,\{j_t\}} := \frac{8(1+\beta)}{(T+1)^{1+\beta}} \big(F(x_0) - \inf F + (144C + T_0 \log(T_0 + 125) + \tfrac{T_0}{2}) L^2 \big),$$

where $C := \sum_{t=T_0}^{\infty} t^{-1-\beta} < \infty$. *Then* $\mathbb{E}\|x_R - \hat{x}_R\|^2 \le \mathcal{C}_{T,\{j_t\}}$. *Consequently, we have the following bound:*

$$\mathbb{E}\big[\mathbf{dist}(x_R, \{x \mid \mathbf{dist}(0, \partial F(x))^2 \le (R+1)^{2\beta}\mathcal{C}_{T,\{j_t\}}\})^2\big] \le \mathcal{C}_{T,\{j_t\}}.$$

Proof. Suppose that $t \ge T_0$ and notice this ensures $\gamma_t \in (0, 1/\rho)$. Following an argument nearly identical to the proof of Theorem 3.3.3, we find that for all $t \ge T_0$, we have

$$\mathbb{E}_t[F(x_{t+1})] \le F(x_t) + \frac{72L^2}{\mu_t(j_t+1)} - \left(\frac{\mu_t}{2} - \frac{30}{\gamma_t^4\mu_t^3 j_t(j_t+1)}\right)\|x_t - \hat{x}_t\|^2,$$

where $\mu_t = \gamma_t^{-1} - \rho$ and $\mathbb{E}_t[\cdot]$ denotes the expectation conditioned on $x_1, \ldots, x_t$. We now show that the coefficient of $-\|x_t - \hat{x}_t\|^2$ is greater than or equal to $\mu_t/4$. Indeed, it suffices to show that $j_t \ge 12/(1-\gamma_t\rho)^2$. To that end, note that

$$\gamma_t \le (\lceil (2\rho)^{1/\beta} \rceil + 1)^{-\beta} \le 1/(2\rho).$$

Therefore, $1 - \gamma_t\rho \ge 1/2$, which leads to the claimed inequality: $12/(1-\gamma_t\rho)^2 \le 44 \le j_t$.

Using the lower bound $\mu_t \ge 1/(2\gamma_t)$ (which follows because $t \ge T_0$), we thus find

$$\begin{aligned}\sum_{t=T_0}^{T-1} \frac{1}{8\gamma_t}\mathbb{E}\big[\|x_t - \hat{x}_t\|^2\big] &\le \mathbb{E}[F(x_{T_0}) - \inf F] + \sum_{t=T_0}^{T-1} \frac{144\gamma_t L^2}{(j_t+1)} \\ &\le F(x_0) - \inf F + \sum_{t=T_0}^{T-1} \frac{144\gamma_t L^2}{(j_t+1)} + L^2 T_0 \log(T_0 + 125).\end{aligned}$$

We would like to extend the sum on the left hand side of the previous inequality to all t between 0 and $T-1$. To that end, we bound the excess terms

$$\sum_{t=0}^{T_0-1} \frac{1}{8\gamma_t}\mathbb{E}\big[\|x_t - \hat{x}_t\|^2\big] \le \sum_{t=0}^{T_0-1} \frac{\gamma_t L^2}{2} \le \frac{T_0 L^2}{2}.$$

Therefore, using the bounds $\sum_{t=T_0}^{\infty} \gamma_t/(j_t+1) \leq \sum_{t=T_0}^{\infty} t^{-1-\beta} = C < \infty$ and $\sum_{t=0}^{T-1} \gamma_t^{-1} \geq \int_{-1}^{T-1} (t+1)^{\beta} dt = T^{1+\beta}/(1+\beta)$, we have,

$$\begin{aligned}
&\mathbb{E}\big[\|x_R - \hat{x}_R\|^2\big] \\
&= \frac{1}{\sum_{t=0}^{T-1} \gamma_t^{-1}} \sum_{t=0}^{T-1} \frac{1}{\gamma_t} \mathbb{E}\big[\|x_t - \hat{x}_t\|^2\big] \\
&\leq \frac{8(1+\beta)}{(T+1)^{1+\beta}} \bigg(F(x_0) - \inf F + 144CL^2 + L^2 T_0 \log(T_0 + 125) + \frac{T_0 L^2}{2} \bigg),
\end{aligned}$$

as desired. To complete the proof, apply Lemma 3.3.2. □

We remark that this convergence rate can be directly utilized to give a complexity bound on computing an ε-expected stationary point. Completing T outer iterations, requires $O(T^2)$ oracle evaluations since j_t is selected as (3.17). Then observing that $\mathcal{C}_{T,\{j_t\}} \leq \varepsilon$ if $T \geq O(\varepsilon^{-1/(1+\beta)})$, we must find an ε-expected stationary point after at most $O(\varepsilon^{-2/(1+\beta)})$ oracle evaluations.

3.4 Experimental Results

In this section we address the population version of the robust real phase retrieval problem: fix a vector $\bar{x} \in \mathbb{R}^d$ and define

$$F(x) := \mathbb{E}_{a,\delta,\xi}\big[|\langle a, x\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)|\big], \tag{3.20}$$

where $a, \delta,$ and ξ are independent random variables satisfying the following assumptions

(B1) a is a zero mean standard Gaussian random variable in $\mathbb{R}^d$;

(B2) δ is a $\{0,1\}$-random variable with $P(\delta = 1) = 0.25$;

(B3) ξ is a zero mean Laplace random variable with scale parameter 1.

In this setting, it is possible to show that the only minimizers of $F(x)$ are $\pm\bar{x}$ [35, Lemma B.8]. In Lemma 3.5.2, we show that this function is 2-weakly convex.

Implementation. Each step of PGSG and the stochastic subgradient method requires access to a subgradient of a random function of the form

$$f(x, a, \delta, \xi) = |\langle a, x\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)|.$$

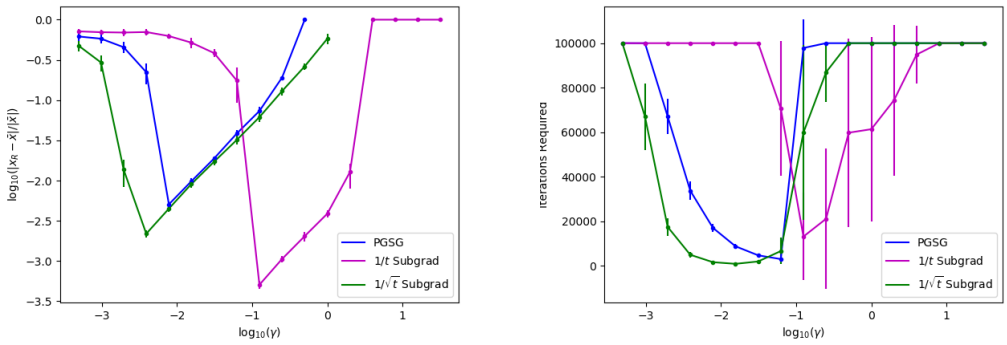
We choose the selection operator

$$G(x, a, \delta, \xi) = 2\langle a, x\rangle a \cdot \operatorname{sign}(\langle a, x\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)) \in \partial_x f(x, a, \delta, \xi).$$

It is a straightforward exercise to show that G satisfies assumption 1 on any bounded set $\mathcal{X}$. For our purposes we choose $\mathcal{X}$ to be a closed ball with a large radius, $r = 10^6$. In our experiments, we never had to explicitly enforce this constraint.

Experiment 1: Sensitivity to Stepsize. In the first experiment we compare the performance of PGSG to the stochastic subgradient method, which possessed no complexity guarantees at the time of writing this manuscript. In the stochastic subgradient method, we choose stepsizes of the form $\gamma/(t + 10)^\beta$ for varying $\gamma > 0$ and $\beta \in \{1/2, 1\}$. For PGSG, we chose varying values of $\gamma > 0$ and then set α_j by (3.9), $j_t = 250$, and $\mu = 1/2\gamma$. Figure 3.1 shows the result of running these two methods to solve robust real phase retrieval problems with $d = 50$.

Like the choice of γ and consequently α_j, the choice of j_t is also important for the practical performance of PGSG. The condition (3.8) used in the analysis of PGSG is often overly conservative and leads to worse performance in practice.



(a) (b)

Figure 3.1: Performance of PGSG and the subgradient method for values of γ averaged over 50 trials. Error bars are included to show one standard deviation. Plot (a) shows the relative distance to a minimizer after 25000 subgradient evaluations. Plot (b) shows the number of subgradient evaluations needed until the relative distance 0.05 to a minimizer.

In our experiments, we chose the constant $j_t = 250$ to balance the quality of the solution to the proximal subproblems with the total number of approximately solved subproblems.

Experiment 2: Mean and Variance of Solution Estimates. Unlike the subgradient method, PGSG provides an easily computed estimate of the of how close x_R is to a nearly stationary point; see the discussion surrounding Lemma 3.11 and Remark 3.3.4. For PGSG, 2PGSG, and PFPGSG, this is given by $\gamma^{-1}\|x_R - x_{R+1}\|$, $\gamma^{-1}\|x_{R^s} - \tilde{x}_{R^s}\|$, and $\gamma_R^{-1}\|x_R - x_{R+1}\|$ respectively. Proposition 3.3.6 shows these estimates are close to $\gamma^{-1}\|x_R - \hat{x}_R\|$ in expectation, which, according to Lemma 3.11, is a natural measure of stationarity. Using these stationarity measures, we analyze the numerical performance of the three algorithms proposed in this manuscript.

Based on the results of Experiment 1, we set $\gamma = 2^{-6}$ for the PGSG and 2-PGSG algorithms. We furthermore set α_j by (3.9) and let $\mu = 1/2\gamma$. For both

methods, we consider two different selections for the number of inner iterations $j_t \in \{10^3, 10^4\}$. These choices determine the level of stationarity reached by the algorithm. For 2PGSG, we fix $S = 5$ and $J = 5T$. For PFPGSG, we set $\beta = 1/2$, $\gamma_t = (t+1)^{-\beta}/10$ (which differs from (3.16) by a factor of ten), j_t as in (3.17), and α_j as in (3.18).

Table 3.1 lists the mean and variance of the stationarity measures averaged over 50 trials. Each sub column shows the performance of the target algorithm as the computational budget increases. We find that with $j_t = 10^3$, both PGSG and 2PGSG quickly converge to a region of stationarity and then do not improve. With $j_t = 10^4$, both of these methods reach a level of stationarity an order of magnitude smaller than with the choice $j_t = 10^3$. Under sufficiently large computational budget (2500000 stochastic subgradient evaluations), the variance of the stationarity reported by 2PGSG is consistently lower than that of PGSG as expected from Theorem 3.3.8. Finally, we note that the performance of PFPGSG is similar to PGSG in most regimes.

3.5 Addendum - Examples of Weak Convexity

3.5.1 Trimmed Estimation

Proposition 3.5.1. *Suppose that $f_1, \ldots, f_n$ are convex, L-Lipschitz continuous functions on $\mathbb{R}^d$. Then the objective*

$$F(w, x) = \begin{cases} \frac{1}{n} \sum_{i=1}^{n} w_i f_i(x) & \text{if } w_i \in [0,1] \text{ and } \sum_{i=1}^{n} w_i = h; \\ \infty & \text{otherwise.} \end{cases}$$

Calls		PGSG $j_t = 1000$	PGSG $j_t = 10000$	2PGSG $j_t = 1000$	2PGSG $j_t = 10000$	PFPGSG
		$d = 50$				
100000	mean	1.538	10.02	1.099	12.46	2.877
	var.	0.0380	1.683	0.0153	5.871	0.178
500000	mean	1.492	0.2043	1.024	8.406	1.615
	var.	0.0542	9.27e-4	0.0119	0.669	0.0421
2500000	mean	1.575	0.2083	1.034	0.1331	0.847
	var.	0.0600	7.53e-4	0.0152	2.562e-4	0.0128
		$d = 100$				
100000	mean	3.632	17.12	2.703	23.04	6.625
	var.	0.137	3.287	0.0544	6.117	0.361
500000	mean	3.579	3.678	2.534	11.83	3.815
	var.	0.145	22.35	0.0430	0.891	0.145
2500000	mean	3.622	0.540	2.564	0.365	2.121
	var.	0.127	2.71e-3	0.0468	1.01e-3	0.0380
		$d = 500$				
100000	mean	27.67	76.86	24.32	100.7	41.65
	var.	8.843	16.95	2.465	20.31	6.471
500000	mean	25.53	23.52	17.13	42.25	25.59
	var.	1.519	1.474	0.341	4.772	1.946
2500000	mean	25.64	4.759	17.10	3.519	15.09
	var.	1.236	0.0454	0.374	0.0118	0.452
		$d = 1000$				
100000	mean	64.73	156.5	34.36	199.5	59.25
	var.	14.37	49.09	2.388	53.92	167.2
500000	mean	55.97	40.99	33.61	86.97	54.48
	var.	3.426	2.091	0.890	9.233	71.38
2500000	mean	55.27	11.88	33.40	9.008	33.86
	var.	4.854	0.119	0.634	0.055	1.350

Table 3.1: Estimated stationarity level for each of the proposed algorithms averaged over 50 trails.

is L-weakly convex.

Proof. We argue using Proposition 3.2.1. Let $(w, x), (\tilde{w}, \tilde{x}) \in \text{dom } F$ and let $\lambda \in [0, 1]$. Then

$$
\begin{aligned}
&F((1-\lambda)(w,x) + \lambda(\tilde{w},\tilde{x})) \\
&= \frac{1}{n}\sum_{i=1}^{n}((1-\lambda)w_i + \lambda\tilde{w}_i) f_i((1-\lambda)x + \lambda\tilde{x}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)w_i f_i((1-\lambda)x + \lambda\tilde{x}) + \frac{1}{n}\sum_{i=1}^{n}\lambda\tilde{w}_i f_i((1-\lambda)x + \lambda\tilde{x}) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)w_i((1-\lambda)f_i(x) + \lambda f_i(\tilde{x})) + \frac{1}{n}\sum_{i=1}^{n}\lambda\tilde{w}_i((1-\lambda)f_i(x) + \lambda f_i(\tilde{x})) \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)w_i f_i(x) + \frac{1}{n}\sum_{i=1}^{n}\lambda(1-\lambda)w_i(f_i(\tilde{x}) - f_i(x)) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\lambda\tilde{w}_i f_i(\tilde{w}_i) + \frac{1}{n}\sum_{i=1}^{n}\lambda(1-\lambda)\tilde{w}_i((f_i(x) - f_i(\tilde{x})) \\
&= (1-\lambda)F(w,x) + \lambda F(\tilde{w},\tilde{x}) + \frac{1}{n}\lambda(1-\lambda)\sum_{i=1}^{n}(\tilde{w}_i - w_i)((f_i(x) - f_i(\tilde{x})) \\
&\leq (1-\lambda)F(w,x) + \lambda F(\tilde{w},\tilde{x}) + \frac{\lambda(1-\lambda)L}{2}\|w - \tilde{w}\|^2 + \frac{\lambda(1-\lambda)L}{2}\|x - \tilde{x}\|^2,
\end{aligned}
$$

as desired. □

3.5.2 Weak Convexity of Robust Phase Retrieval

Lemma 3.5.2. *The robust phase retrieval loss defined in* (3.20) *is* 2*-weakly convex.*

Proof. For all $x, y, a \in \mathbb{R}^d$, we have

$$\langle a, \lambda x + (1-\lambda)y\rangle^2 = \lambda\langle a, x\rangle^2 + (1-\lambda)\langle a, y\rangle^2 - \lambda(1-\lambda)\langle a, y - x\rangle^2.$$

Thus, we have

$$
\begin{aligned}
&F(\lambda x + (1-\lambda)y) \\
&= \mathbb{E}_{a,\delta,\xi}\big[|\langle a, \lambda x + (1-\lambda)y\rangle^2 - (\langle a, \bar{x}\rangle^2 + \delta \cdot \xi)|\big] \\
&\leq \lambda F(x) + (1-\lambda)F(y) + \lambda(1-\lambda)\mathbb{E}_a\big[\langle a, y - x\rangle^2\big] \\
&= \lambda F(x) + (1-\lambda)F(y) + \lambda(1-\lambda)\|x - y\|^2.
\end{aligned}
$$

Therefore, by Proposition 3.2.1, F is 2-weakly convex. □

CHAPTER 4

PROXIMAL BUNDLE METHOD CONVERGENCE RATES

4.1 Introduction

Adaptive optimization algorithms, those capable of adapting as they encounter problem structure, play an important role in the development of practical optimization methods. These methods often speed-up in the presence of prevalent structures like growth/error bounds or KL conditions without requiring prior knowledge of what structure will be present.

In particular, this chapter is interested in solving unconstrained convex minimization problems of our recurring form (1.1)

$$\min_{x \in \mathbb{R}^n} f(x) \tag{4.1}$$

that attain their minimum value at some $x^* \in \mathbb{R}^n$. We are interested in adapting to solve this problem under a variety of different assumptions on f, considering settings where f may be M-Lipschitz continuous

$$|f(x) - f(y)| \leq M\|x - y\| \,, \tag{4.2}$$

may be differentiable with an L-Lipschitz gradient (often referred to as L-smoothness)

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \,, \tag{4.3}$$

and may satisfy a Hölderian growth bound

$$f(x) - f^* \geq \mu\|x - x^*\|^p \,. \tag{4.4}$$

Particularly important cases are when $p = 1$ and $p = 2$, which correspond to sharp growth (μ-SG) [23] and quadratic growth (μ-QG), generalizing strong convexity, respectively.

The core finding of this chapter is showing that a very classic and time-tested first-order method, the proximal bundle method, is an adaptive algorithm, speeding up in the presence of either smoothness or Hölder growth. Bundle methods were first developed and proposed independently in [93] and [170]. Computationally cheaper bundle methods which aggregate cuts were analyzed in [82, 84]. The central result in the convergence theory of bundle methods is that (for convex f that attain their minimum value somewhere) its two sequences of iterates $\{z_k\}_{k=0}^{\infty}$ and $\{x_k\}_{k=0}^{\infty}$ both converge to a minimizer of f. That is, the bundle method when run with no stopping criteria has

$$\lim_{k\to\infty} x_k = \lim_{k\to\infty} z_k = x^* \in \operatorname{argmin} f. \tag{4.5}$$

Importantly, this holds for *any* constant configuration of its algorithmic parameters. See [84, Thm. 4.9], [78, Thm. XV.3.2.4], or [155, Thm. 7.16] for proofs of different variations of this result.

This stands in harsh contrast to other first-order methods: for example, gradient descent and its accelerated variants rely on selecting a stepsize inversely proportional to the level of smoothness, and in nonsmooth optimization, subgradient methods rely on carefully controlled decreasing stepsize sequences. These simpler algorithms may fail to converge if these rules are not followed.

In 2000, Kiwiel [86] gave the first convergence rate for the proximal bundle method, showing that an ϵ-minimizer x_k is found within

$$k \leq O\left(\frac{\|x_0 - x^*\|^4}{\epsilon^3}\right)$$

iterations. More recently, Du and Ruszczyński [41] gave the first analysis of bundle methods when applied to problems satisfying a quadratic growth bound. In this case, an ϵ-minimizer is found within $\widetilde{O}(1/\epsilon)$ iterations. Despite having weaker convergence rate guarantees than simple alternatives like the subgradient method, bundle methods have persisted as a method of choice for nonsmooth convex optimization. In practice, bundle methods have proven to be efficient methods for solving many nonsmooth problems (see [94, 156, 157, 76] for further discussion). Extensions that apply to nonconvex problems have been considered in [3, 72, 83, 111] and an extension to problems where only an inexact first-order oracle was recently given in [73].

Stronger convergence rates have been established for related level bundle methods [95], which share many core elements with proximal bundle methods. Further variations of level bundle methods were studied in [85] and [90]. The results of Lan [90] are particularly impressive as their proposed method has optimal convergence rates for both smooth and nonsmooth problems while requiring very little input.

Our Contributions. We show that the most classic version of the bundle method (the proximal bundle method) is an adaptive algorithm, converging faster in the presence of smoothness or Hölder growth. Our analysis technique applies to every combination of continuity/smoothness assumption (4.2) or (4.3) and growth assumption (4.4) and to constant stepsize selections as well as adaptively chosen ones. In Table 4.1, we show the leading term in $O(\epsilon^{-1})$ of our convergence rates for each of these varied settings. Full theorem statements are given in Section 4.3 and apply for any Hölder growth exponent (rather than just the cases of $p = 1$ and $p = 2$ shown in the table).

Assumptions		Rate for generic ρ	Rate for tuned ρ	Rate for adaptive ρ_k
M-Lipschitz	No Growth	$O\left(\frac{M^2\|x_0 - x^*\|^4}{\rho\epsilon^3}\right)$	$O\left(\frac{M^2\|x_0 - x^*\|^2}{\epsilon^2}\right)$	$O\left(\frac{M^2\|x_0 - x^*\|^2}{\epsilon^2}\right)$
	μ-QG	$O\left(\frac{M^2}{\min\{\mu,\rho\}\epsilon}\right)$	$O\left(\frac{M^2}{\mu\epsilon}\right)$	$O\left(\frac{M^2}{\mu\epsilon}\right)$
	μ-SG	$O\left(\frac{M^2}{\rho\epsilon}\right)$	$O\left(\frac{M^2}{\mu^2}\sqrt{\frac{f(x_0) - f^*}{\epsilon}}\right)$	$O\left(\frac{M^2}{\mu^2}\log\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right)$
L-Smooth	No Growth	$O\left(\frac{L^3\|x_0 - x^*\|^2}{\rho^2\epsilon}\right)$	$O\left(\frac{L\|x_0 - x^*\|^2}{\epsilon}\right)$	$O\left(\frac{L\|x_0 - x^*\|^2}{\epsilon}\right)$
	μ-QG	$O\left(\frac{L^3}{\rho^2\mu}\log\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{f(x_0) - f^*}{\epsilon}\right)\right)$

Table 4.1: The first column applies for any choice of the algorithmic parameter ρ, showing progressively faster convergence as more structure is introduced. The second column shows the rate after optimizing the choice of ρ. The third column further improves these by allowing nonconstant stepsizes ρ_k.

The existing convergence theory for the proximal bundle method applies to settings that are comparable to the first two rows of our table. Kiwiel [86] derived a $O(\epsilon^{-3})$ convergence rate for Lipschitz problems, which agrees with our theory. Du and Ruszczynski [41] showed a $O(\log(1/\epsilon)/\epsilon)$ convergence rate for Lipschitz, strongly convex problems which we improve on by removing the extra logarithmic term and thus achieve the optimal convergence rate for this setting of $O(1/\epsilon)$. To our knowledge, the rest of our convergence results apply to wholly new settings for the proximal bundle method. In all of the M-Lipschitz settings considered, we show that using a nonconstant stepsize, the bundle method attains the optimal nonsmooth convergence rate. In the L-smooth settings considered, we find that the bundle method converges at the same rate as gradient descent (although, unlike gradient descent, our convergence theory applies to any configuration of its algorithmic parameters).

Finally, we propose a parallelizable variant of the bundle method that avoids the reliance on tuning a stepsize or sequence of stepsizes based on potentially unrealistic knowledge of underlying problem constants. Applying our analysis

technique to this parallel method shows it attains the optimal nonsmooth convergence rates for Lipschitz problems with any level of Hölder growth (up to the cost of running a logarithmic number of instances of the bundle method in parallel).

4.2 Bundle Methods

Here we formally define the family of proximal bundle methods to which our theory applies as well as our improved parallel bundle method. We conclude this section with an outline of our convergence analysis technique.

4.2.1 The Proximal Bundle Methods

Proximal Bundle Methods work by maintaining a model function $f_k \colon \mathbb{R}^n \to \mathbb{R}$ at each iteration k. Each iteration has a current iterate x_k and computes a candidate for the next iterate as

$$z_{k+1} = \operatorname*{argmin}_{z \in \mathcal{X}} f_k(z) + \frac{\rho_k}{2}\|z - x_k\|^2.$$

Note the first-order optimality condition for this subproblem defines a subgradient

$$s_{k+1} = \rho_k(z_{k+1} - x_k) \in \partial f_k(z_{k+1}).$$

If the candidate z_{k+1} has at least $\beta \in (0, 1)$ fraction of the decrease in objective value that our model $f_k(\cdot)$ predicts, then the method takes $x_{k+1} = z_{k+1}$ as the next iterate, called a *Descent Step*. Otherwise the method keeps the iterate the same $x_{k+1} = x_k$, called a *Null Step*. Regardless of which type of step was taken,

the model function is improved based on z_{k+1} and a subgradient $g_{k+1} \in \partial f(z_{k+1})$ satisfying the following three properties for all $x \in \mathbb{R}^d$:
it must always be uniformly upper bounded by the original objective

$$f_{k+1}(x) \leq f(x) \,, \tag{4.6}$$

the subgradient of f at z_{k+1} must give an affine lower bound of

$$f_{k+1}(x) \geq f(z_{k+1}) + \langle g_{k+1}, x - z_{k+1} \rangle \,, \tag{4.7}$$

and after any null step k, the subgradient of f_k at z_{k+1} must give an affine lower bound of

$$f_{k+1}(x) \geq f_k(z_{k+1}) + \langle s_{k+1}, x - z_{k+1} \rangle \,. \tag{4.8}$$

Lastly, we require that after any null step k, the proximal parameter does not change

$$\rho_{k+1} = \rho_k \,. \tag{4.9}$$

The proximal bundle method is stated fully in Algorithm 5.

Algorithm 5 Proximal Bundle Method

Require: $z_0 = x_0 \in \mathbb{R}^n$, $f_0(z) = f(x_0) + \langle g_0, z - x_0 \rangle$
1: **for** $k \geq 0$ **do**
2: Compute candidate iterate $z_{k+1} \leftarrow \operatorname{argmin}_{z \in \mathcal{X}} f_k(z) + \frac{\rho_k}{2}\|z - x_k\|^2$.
3: **if** $\beta(f(x_k) - f_k(z_{k+1})) \leq f(x_k) - f(z_{k+1})$ (Descent step)
4: set $x_{k+1} \leftarrow z_{k+1}$,
5: **else** (Null step)
6: set $x_{k+1} \leftarrow x_k$.
7: Update f_{k+1} and ρ_{k+1} satisfying (4.6), (4.7), (4.8), and (4.9).
8: **end for**

Bundle Method Model Function Choices

Several different methods for constructing the model functions f_k satisfying (4.6)-(4.8) have been considered. In practice, the main consideration lies

in weighing potentially greater per iteration gains from having more complex models and lower iteration costs from having simpler models. Recently, Nesterov and Florea [129] proposed a method for efficiently solving piecewise linear subproblems similar to those considered here in the context of smooth optimization.

Full-Memory Proximal Bundle Method The earliest proposed bundle methods [93, 170] rely on using all of the past subgradient evaluations to construct the models as

$$f_{k+1}(x) = \max_{j=0\ldots k+1} \{f(z_j) + \langle g_j, x - z_j \rangle\}. \tag{4.10}$$

In this case, solving the quadratically regularized subproblem at each iteration amounts to solving a quadratic program.

Finite Memory Proximal Bundle Method Alternatively using the cut-aggregation approach of [82, 84], the collection of $k+1$ lower bounds used by (4.10) can be simplified down to just two linear lower bounds. The only two necessary lower bounds are exactly those required by (4.7) and (4.8). Namely, one could construct the model functions as

$$f_{k+1}(x) = \max\{f_k(z_{k+1}) + \langle \rho_k(z_{k+1} - x_k), x - z_{k+1} \rangle,\ f(z_{k+1}) + \langle g_{k+1}, x - z_{k+1} \rangle\}. \tag{4.11}$$

Then the subproblem that needs to be solved at each iteration can be done in closed form, see (4.17). Hence the iteration cost using this model is limited primarily by the cost of one subgradient evaluation.

Bundle Method Stepsize Choices

The simplest stepsize choice is to select $\rho_k = \rho \in \mathbb{R}$ as a constant value. In Section 4.3.1, we present convergence theory for the bundle method using a generic constant, giving the first column of Table 4.1. Tuning this constant to minimize the resulting convergence guarantee gives the second column of Table 4.1.

Further improvements in convergence guarantees follow from allowing nonconstant stepsizes. If the optimal objective $f(x^*)$ is known, Section 4.3.2 derives guarantees following from selecting the proximal stepsize parameter as

$$\rho_k = (f(x_k) - f(x^*))/D^2 \tag{4.12}$$

where $D^2 \geq \sup\{\|x - x^*\|^2 \mid f(x) \leq f(x_0)\}$ for generic problems and as

$$\rho_k = \mu^{2/p}(f(x_k) - f(x^*))^{1-2/p} \tag{4.13}$$

when Hölder growth (4.4) holds, matching the third column of Table 4.1. These choices are motivated by mimicking the following idealistic (and impractical) stepsize rules that naturally arises from our theory

$$\rho_k = \frac{f(x_k) - f(x^*)}{\|x_k - x^*\|^2} \, . \tag{4.14}$$

Other interesting nonconstant stepsizes could also be considered. Stepsizes that decrease over time, mirroring those employed for subgradient methods, could be taken to decrease ρ_k over time based on the number of descent steps taken so far (using the number of descent steps rather than total steps to satisfy (4.9)). The analysis of such schemes is beyond the scope of this chapter.

4.2.2 The Parallel Bundle Method

Here we give a practical scheme for applying the bundle method that attains the same complexity as our optimally tuned nonconstant stepsizes without any knowledge of problem structure (i.e., the presence of smoothness or growth bounds or the associated constants). We do this by employing a logarithmic number of instances of the bundle method with different constant stepsizes in parallel that continually share their progress with each other (inspired by the ideas of [142]). By doing so, we recover our optimal rates, up to the cost of running a logarithmic number of algorithms which can be mitigated through parallelization.

The core observation behind our parallel method is that our nonconstant stepsize rules (4.12) and (4.13) are always in the interval

$$\rho_k \in \left[\Omega(\epsilon), O(\epsilon^{-1})\right]$$

before an ϵ-minimizer is found. As input, we only assume the following are given: a lower bound $\bar{\rho}$ and an upper bound $2^J\bar{\rho}$ on the range of stepsizes to consider. Provided our stepsize rules (4.12) and (4.13) lie in this interval,

$$\rho_k \in \left[\bar{\rho}, 2^J\bar{\rho}\right] ,$$

we are able to recover our optimal convergence rates. Notice that the interval $[\bar{\rho}, 2^J\bar{\rho}]$ can span the whole range of stepsizes needed for our Hölder growth analysis by setting $\bar{\rho} = O(\epsilon)$ and $J = O(\log(1/\epsilon^2))$. Our resulting convergence guarantees only depend logarithmically on the size of this interval (a cost which can be mitigated through parallelization), so $\bar{\rho}$ and $2^J\bar{\rho}$ can bet set generously at little cost.

Defining the Parallel Bundle Method

We propose running J copies of the bundle method in parallel, which share their progress with each other as described below. Each bundle method $j \in \{0, \dots J-1\}$ uses a constant stepsize $\rho^{(j)} = 2^j \bar{\rho}$. Denote the iterates of bundle method j by $x_k^{(j)}$ and its model objectives by $f_k^{(j)}$. Each bundle method j proceeds as normal with the only modification being that after it takes a descent step, the algorithm checks if any other bundle method j' has an iterate with an even lower objective value $f(x_k^{(j')}) < f(x_{k+1}^{(j)})$. If such an improvement exists, the bundle method instead descends to the best such iterate, setting

$$\begin{cases} x_{k+1}^{(j)} & \leftarrow x_k^{(j')} \\ f_{k+1}^{(j)}(z) & \leftarrow f(x_k^{(j')}) + \langle g_k^{(j')}, z - x_k^{(j')} \rangle \end{cases}$$

and then proceeds.

For the sake of analysis, we assume that each parallel instance of the bundle method operates synchronously, with every instance completing one iteration before any instance completes a second iteration. This process can be implemented sequentially by cycling through the bundle method instances, computing one iteration for each before repeating. An asynchronous variant of this procedure could be analyzed as well (using ideas from the asynchronous restarting analysis of [142]) but is beyond the focus of this chapter.

Analysis Overview and Proof Sketch

Each iteration of the bundle method can be viewed as attempting to mimic the proximal point method, using the model f_k instead of the true objective function

f. At each iteration k, denote the objective gap of the proximal subproblem (called the *proximal gap*) by

$$\Delta_k = f(x_k) - \Big(f(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1} - x_k\|^2\Big)$$

where $\bar{x}_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^d}\big\{f(x) + \frac{\rho_k}{2}\|x - x_k\|^2\big\}$.

Regardless of which continuity, smoothness and growth assumptions are made, our analysis works by relating the proximal steps computed by the bundle method on the models f_k to proximal steps on f. The following pair of observations show that the behavior on both descent steps and null steps is controlled by the proximal gap Δ_k.

(i) **Descent steps attain decrease proportional to the proximal gap.**

Lemma 4.2.1. *Every descent step k has*

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

(ii) **The number of consecutive null steps is bounded by the proximal gap.**

Lemma 4.2.2. *A descent step k followed by T consecutive null steps has at most*

$$T \leq \frac{2G_k^2}{(1-\beta)^2\rho_k\Delta_{k+1}}$$

where $G_k = \sup\{\|g_{t+1}\| \mid k \leq t \leq k+T\}$. For Lipschitz or smooth objectives, this simplifies to

$$T \leq \begin{cases} \dfrac{2M^2}{(1-\beta)^2\rho_k\Delta_{k+1}} & \text{if } f \text{ is } M\text{-Lipschitz} \\ \dfrac{4(L+\rho_k)^3}{(1-\beta)^2\rho_k^3} & \text{if } f \text{ is } L\text{-smooth .} \end{cases}$$

With these two observations in hand, convergence guarantees for the bundle method follow from specifying any choice of the parameter ρ_k. Standard analysis [155] of the proximal gap shows the following bound for any minimizer x^*.

Lemma 4.2.3. *For any $x_k \in \mathbb{R}^n$, the proximal gap is lower bounded by*

$$\Delta_k \geq \begin{cases} \frac{1}{2\rho_k}\left(\frac{f(x_k) - f(x^*)}{\|x_k - x^*\|}\right)^2 & \text{if } f(x_k) - f(x^*) \leq \rho_k \|x_k - x^*\|^2 \\ f(x_k) - f(x^*) - \frac{\rho_k}{2}\|x_k - x^*\|^2 & \text{otherwise.} \end{cases} \tag{4.15}$$

All of our analysis (for any combination of continuity/smoothness condition (4.2) or (4.3) and potential growth condition (4.4)) follows directly from applying these core lemmas. We bound the number of descent steps by combining Lemmas 4.2.1 and 4.2.3 to give a recurrence relation describing the decrease in the objective gap. Then Lemmas 4.2.2 and 4.2.3 together allow us to bound the number of consecutive null steps between each of these descent steps, which can then be summed up to bound the total number of iterations required.

4.3 Formal Statement of Convergence Guarantees

We present our convergence guarantees in order of increasing sophistication in the stepsize policy. First we give guarantees for any constant configuration of the proximal bundle method, then we consider two nonconstant stepsize policies, and finally, we analyze our parallel bundle method.

4.3.1 Convergence Rates from Constant Stepsize Choice

First we formalize our convergence theory for the proximal bundle method using any constant choice of the stepsize parameter $\rho_k = \rho$ and any $\beta \in (0, 1)$. These guarantees all match those claimed in the first column of Table 4.1. After each theorem, we remark on the tuned choice of ρ that gives rise to the claimed rate in the second column of Table 4.1. First we consider the setting where only Lipschitz continuity is assumed.

Theorem 4.3.1. *For any M-Lipschitz objective function f, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{2\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{\beta} \right\rceil_+$$

and the number of null steps is at most

$$\frac{12\rho M^2 D^4}{\beta(1-\beta)^2\epsilon^3} + \frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2}$$

where $D^2 = \sup_k \|x_k - x^*\|^2 < \infty$.

Remark 4.3.2. *It follows from [155, (7.64)] that $D^2 \leq \|x_0 - x^*\|^2 + \frac{2(1-\beta)(f(x_0)-f^*)}{\beta\rho}$. Alternatively, if the level sets of f are bounded, the fact that $f(x_k)$ is non-increasing ensures $D^2 \leq \sup\{\|x - x^*\|^2 \mid f(x) \leq f(x_0)\}$.*

Remark 4.3.3. *Carefully selecting the proximal parameter $\rho > 0$ reduces the number of gradient oracle queries required to find an ϵ-minimizer. Selecting $\rho = \epsilon/D^2$ gives an overall complexity bound of*

$$O\left(\frac{M^2 D^2}{\epsilon^2}\right)$$

and matches the optimal rate for nonsmooth, Lipschitz optimization, plus an additive log term.

If instead of Lipschitz continuity of the objective, we assume the objective has Lipschitz gradient, the bundle method adapts to give the following faster rate.

Theorem 4.3.4. *For any L-smooth objective function f, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{2\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{\beta} \right\rceil_+$$

and the number of null steps is at most

$$\frac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left(\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{2\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{\beta} \right\rceil_+ + 1\right)$$

where $D^2 = \sup_k \|x_k - x^*\|^2 < \infty$.

Remark 4.3.5. *Carefully selecting the proximal parameter $\rho > 0$ can improve the dependency on L in the convergence rate. Namely selecting $\rho = L$ gives an overall complexity bound of*

$$\frac{16LD^2}{\beta(1-\beta)^2\epsilon}.$$

This matches the standard convergence rate for gradient descent.

Next we reconsider the settings of Lipschitz continuity and smoothness with additional structure in the form of a Hölder growth bound. We find that the convergence guarantees divide into three regions depending on the growth exponent p, whether it is large, equal to, or smaller than 2. Here $p = 2$ is the critical exponent value since the proximal subproblem is adding in quadratic regularization. Regardless, as p decreases, the bundle method converges faster.

Theorem 4.3.6. *For any M-Lipschitz objective function f satisfying the Hölder growth condition* (4.4)*, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\begin{cases} \dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{2\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{\beta} \right\rceil_+ & \text{if } p > 2 \\ \left\lceil \dfrac{2\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{\beta\min\{\mu/\rho, 1\}} \right\rceil & \text{if } p = 2 \\ \left\lceil \dfrac{2\log\left(\frac{(\rho/\mu^{2/p})^{1/(1-2/p)}}{\epsilon}\right)}{\beta} \right\rceil_+ + \dfrac{2\rho(f(x_0)-f^*)^{2/p-1}}{(1-2^{1-2/p})\beta\mu^{2/p}} & \text{if } 1 \leq p < 2 \end{cases}$$

and the number of null steps is at most

$$\begin{cases} \dfrac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} + \dfrac{8M^2}{\beta(1-\beta)^2(\rho/\mu^{2/p})^{1/(1-2/p)}} & \text{if } p > 2 \\ \dfrac{4M^2}{\beta(1-\beta)^2\min\{\mu/\rho, 1\}\rho\epsilon} & \text{if } p = 2 \\ \dfrac{4M^2}{\beta(1-\beta)^2\rho\epsilon} + \dfrac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}C & \text{if } 1 \leq p < 2 \end{cases}$$

where $C = \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\}\min\left\{\frac{1}{1-2^{-|4/p-3|}}, \left\lceil\log_2\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)\right\rceil\right\}$.

Remark 4.3.7. *Carefully selecting the proximal parameter $\rho > 0$ can improve the dependency on ϵ and μ in the convergence rate. When $p = 2$, selecting $\rho = \mu$ gives an overall complexity bound of $O(M^2/\mu\epsilon)$. This matches the optimal rate, plus an additive log term. When $p = 1$, selecting $\rho = O(1/\sqrt{\epsilon})$ minimizes this bound, but the resulting sublinear $O(1/\sqrt{\epsilon})$ rate falls short of the best possible rate (linear convergence) for sharp, Lipschitz optimization. In the next section where we consider nonconstant stepsizes, this disconnect will be remedied and a linear convergence guarantee will follow.*

Theorem 4.3.8. *For any L-smooth objective function f satisfying the Hölder growth condition* (4.4)*, consider applying the bundle method using a constant stepsize $\rho_k = \rho$.*

Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most

$$\begin{cases} \dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{2\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{\beta} \right\rceil_+ & \text{if } p > 2 \\ \left\lceil \dfrac{2\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{\beta\min\{\mu/\rho, 1\})} \right\rceil & \text{if } p = 2 \end{cases}$$

and the number of null steps is at most

$$\begin{cases} \dfrac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left(\dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{2\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{\beta} \right\rceil_+ + 1\right) & \text{if } p > 2 \\ \dfrac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left\lceil \dfrac{2\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{\beta\min\{\mu/\rho, 1\}} \right\rceil & \text{if } p = 2\,. \end{cases}$$

Remark 4.3.9. *Selecting $\rho = L$ gives a complexity bound matching gradient descent.*

4.3.2 Convergence Rates from Improved Stepsize Choice

Selecting ρ_k to vary over the course of the bundle method's application allows for stronger convergence guarantees. These rates are formalized in the following pair of theorems that consider settings with and without Hölder growth. In the latter case, we find that our stepsize choice (4.13) removes the need for piecewise guarantees around growth exponent $p = 2$, which notably simplifies the statement of our guarantees.

Theorem 4.3.10. *For any M-Lipschitz objective function f, consider applying the bundle method using the stepsize policy* (4.12) *with any choice of $D^2 \geq \sup\{\|x - x^*\|^2 \mid f(x) \leq f(x_0)\}$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps*

before an ϵ-minimizer is found is at most

$$\left\lceil \frac{2\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{\beta} \right\rceil$$

and the number of null steps is at most

$$\left(\frac{1}{1-(1-\beta/2)^2}\right)\frac{2M^2D^2}{(1-\beta)^2\epsilon^2} \ .$$

Theorem 4.3.11. *For any M-Lipschitz objective function f satisfying the Hölder growth condition* (4.4)*, consider applying the bundle method using the stepsize policy* (4.13)*. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\left\lceil \frac{2\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{\beta} \right\rceil$$

and the number of null steps is at most

$$\begin{cases} \left(\dfrac{1}{1-(1-\beta/2)^{2-2/p}}\right)\dfrac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} & \text{if } p > 1 \\ \dfrac{4M^2}{(1-\beta)^2\mu^2}\left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{\beta} \right\rceil & \text{if } p = 1 \ . \end{cases}$$

4.3.3 Convergence Rates for the Parallel Bundle Method

First, we remark that all of our previous convergence theory for constant stepsizes (Theorems 4.3.1, 4.3.4, 4.3.6, and 4.3.8) immediately apply to the Parallel Bundle Method fixing $\rho = 2^j\bar{\rho}$ for any $j \in \{0, \dots J-1\}$. This follows by observing that our proofs just rely on a decrease in objective value at descent steps via Lemma 4.2.1. Then the key fact is that Lemma 4.2.1 still holds even in the new case of a bundle method restarting at another method's lower objective value iterate. Hence any individual instance of the bundle method with $\rho^{(j)} = 2^j\bar{\rho}$ in

our parallel scheme will converge at least as fast as Theorems 4.3.1, 4.3.4, 4.3.6, and 4.3.8 guarantee it would converge on its own.

Further and more importantly, when our nonconstant stepsize rules (4.12) and (4.13) lie in the interval $[\bar{\rho}, 2^J\bar{\rho}]$, we find that their convergence theory (Theorem 4.3.10 and 4.3.11) also extends to our parallel algorithm. This is formalized as follows.

Theorem 4.3.12. *For any M-Lipschitz objective function f that satisfies the Hölder growth condition* (4.4)*, consider applying the Parallel Bundle Method with stepsizes $\rho = 2^j\bar{\rho}$ for $j \in \{0, \ldots, J-1\}$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, if*

$$\bar{\rho} \leq \frac{1}{4}\mu^{2/p} \min\{\epsilon^{1-2/p}, (f(x_0) - f(x^*))^{1-2/p}\}$$

and

$$J \geq \log_2\left(\frac{\mu^{2/p}(\max\{\epsilon^{1-2/p}, (f(x_0) - f(x^*))^{1-2/p}\}}{4\bar{\rho}}\right),$$

then one of our J bundle methods will find an ϵ-minimizer within its first

$$\begin{cases} \left(\dfrac{2}{1-(1-\beta/2)^{2-2/p}}\right)\dfrac{16M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} + 2\left\lceil\dfrac{2\log(\frac{f(x_0)-f^*}{\epsilon})}{\beta}\right\rceil & \text{if } p > 1 \\ 2\left(\dfrac{16M^2}{(1-\beta)^2\mu^2} + 1\right)\left\lceil\dfrac{2\log(\frac{f(x_0)-f^*}{\epsilon})}{\beta}\right\rceil & \text{if } p = 1 \end{cases}$$

iterations.

Remark 4.3.13. *These rates match the optimal lower bounds for nonsmooth Lipschitz optimization, up to small constants (and an additive logarithmic term when $p > 1$). For example, under quadratic growth when $p = 2$, we have*

$$\bar{\rho} \leq \mu/4 \text{ and } J \geq \log_2(\mu/4\bar{\rho}) \implies \text{a rate of } O\left(\frac{M^2}{\mu\epsilon}\right).$$

Under sharp growth $p = 1$, we have

$$\bar{\rho} \leq \frac{\mu^2}{4(f(x_0) - f(x^*))} \text{ and } J \geq \log_2(\mu^2/4\bar{\rho}\epsilon) \implies \text{a rate of } O\left(\frac{M^2}{\mu^2}\log(1/\epsilon)\right).$$

Critically, these convergence rates only depend on $\bar{\rho}$ and J through the (parallelizable) cost of updating the J bundle method instances at each iteration.

4.4 Convergence Analysis

Our convergence analysis is built on Lemmas 4.2.1 and 4.2.2, which relate descent steps and null steps to the proximal gap Δ_k. After proving these two essential lemmas, all of our subsequent analysis for Theorems 4.3.1 through 4.3.12 follow from using a recurrence relation stemming from Lemma 4.2.1 to bound the number of descent steps, and from using Lemma 4.2.2 to bound the number of null steps.

4.4.1 Proof of the Descent Step Lemma 4.2.1

Let $\bar{x}_{k+1} = \operatorname{argmin}\{f(\cdot) + \frac{\rho_k}{2}\|\cdot - x_k\|^2\}$. From (4.6), we have

$$\begin{aligned} f_k(x_{k+1}) &\leq f_k(x_{k+1}) + \frac{\rho_k}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f_k(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1} - x_k\|^2 \\ &\leq f(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1} - x_k\|^2 \,. \end{aligned}$$

Hence $f(x_k) - f_k(x_{k+1}) \geq \Delta_k$. Since we have assumed that iteration k was a descent step, this implies $(f(x_k) - f(x_{k+1}))/\beta \geq \Delta_k$.

4.4.2 Proof of the Null Step Lemma 4.2.2

Consider some descent step k followed by T consecutive null steps. Denote the proximal subproblem gap at iteration $k < t \leq k + T$ on the model f_t by

$$\widetilde{\Delta}_t = f(x_{k+1}) - \Big(f_t(z_{t+1}) + \frac{\rho_k}{2}\|z_{t+1} - x_{k+1}\|^2\Big).$$

The core of this null step bound relies on the following recurrence showing that every null step t decreases this quantity as

$$\widetilde{\Delta}_{t+1} \leq \widetilde{\Delta}_t - \frac{\rho_k(1-\beta)^2\widetilde{\Delta}_t^2}{2G_k^2}\,. \tag{4.16}$$

Before deriving this inequality, we show how it completes the proof of this lemma: After T consecutive null steps, the fact that $f_{k+T} \leq f$ ensures $\widetilde{\Delta}_{k+T} \geq \Delta_{k+T} = \Delta_{k+1}$. Then solving this recurrence relation (see Lemma 4.5.1 at the end of the chapter with $\epsilon = \Delta_{k+1}$), we conclude the number of consecutive null steps is at most

$$T \leq \frac{2G_k^2}{(1-\beta)^2\rho_k\Delta_{k+1}}\,.$$

Now all that remains is to derive the recurrence (4.16).

Consider some null step $k < t \leq k + T$ in the sequence of consecutive null steps. Denote the necessary lower bound on f_{t+1} given by (4.7) and (4.8) as

$$\tilde{f}_{t+1}(\cdot) := \max\{f_t(z_{t+1}) + \langle s_{t+1}, \cdot - z_{t+1}\rangle,\ f(z_{t+1}) + \langle g_{t+1}, \cdot - z_{t+1}\rangle\} \leq f_{t+1}(\cdot)\,.$$

Denote the result of a proximal step on $\tilde{f}_{t+1}$ by $y_{t+2} = \operatorname{argmin}\Big\{\tilde{f}_{t+1}(\cdot) + \frac{\rho_k}{2}\|\cdot - x_{k+1}\|^2\Big\}$. A simple computation gives an explicit form for the minimizer of this problem

$$\begin{aligned}\theta_{t+1} &= \min\left\{1, \frac{\rho_k(f(z_{t+1}) - f_t(z_{t+1}))}{\|g_{t+1} - s_{t+1}\|^2}\right\}\\ y_{t+2} &= x_k - \frac{1}{\rho_k}(\theta_{t+1}g_{t+1} + (1-\theta_{t+1})s_{t+1}).\end{aligned} \tag{4.17}$$

Hence the objective of the proximal subproblem at iteration $t+1$ is lower bounded by

$$
\begin{aligned}
&f_{t+1}(z_{t+2}) + \frac{\rho_k}{2}\|z_{t+2} - x_{k+1}\|^2 \\
&\geq \tilde{f}_{t+1}(y_{t+2}) + \frac{\rho_k}{2}\|y_{t+2} - x_{k+1}\|^2 \\
&\geq \theta_{t+1}(f(z_{t+1}) + \langle g_{t+1}, y_{t+2} - z_{t+1}\rangle) + (1-\theta_{t+1})(f_t(z_{t+1}) + \langle s_{t+1}, y_{t+2} - z_{t+1}\rangle) \\
&\qquad\qquad + \frac{\rho_k}{2}\|y_{t+2} - x_{k+1}\|^2 \\
&= f_t(z_{t+1}) + \theta_{t+1}\big(f(z_{t+1}) - f^t(z_{t+1})\big) + \langle \theta_{t+1} g_{t+1} + (1-\theta_{t+1}) s_{t+1}, y_{t+2} - z_{t+1}\rangle \\
&\qquad\qquad + \frac{\rho_k}{2}\|y_{t+2} - x_{k+1}\|^2 \\
&= f_t(z_{t+1}) + \theta_{t+1}\big(f(z_{t+1}) - f^t(z_{t+1})\big) + \theta_{t+1}^2\|g_{t+1} - s_{t+1}\|^2/\rho_k + \frac{\rho_k}{2}\|z_{t+1} - x_{k+1}\|^2 ,
\end{aligned}
$$

where the first inequality uses that $f_{t+1} \geq \tilde{f}_{t+1}$, the second inequality takes a convex combination of the two affine functions defining $\tilde{f}_{t+1}$, and the second equality uses the definition of y_{t+2}. Thus we have

$$
\widetilde{\Delta}_{t+1} \leq \widetilde{\Delta}_t - \theta_{t+1}(f(z_{t+1}) - f_t(z_{t+1})) + \theta_{t+1}^2\|g_{t+1} - s_{t+1}\|^2/\rho_k .
$$

The amount of decrease guaranteed above can be lower bounded as follows

$$
\begin{aligned}
&\theta_{t+1}(f(z_{t+1}) - f_t(z_{t+1})) + \theta_{t+1}^2\|g_{t+1} - s_{t+1}\|^2/\rho_k \\
&\geq \min\left\{ f(z_{t+1}) - f_t(z_{t+1}),\ \frac{2\rho_k(f(z_{t+1}) - f_t(z_{t+1}))^2}{\|g_{t+1} - s_{t+1}\|^2} \right\} \\
&\geq \min\left\{ (1-\beta)\widetilde{\Delta}_t,\ \frac{2\rho_k(1-\beta)^2\widetilde{\Delta}_t^2}{\|g_{t+1} - s_{t+1}\|^2} \right\} \\
&\geq \min\left\{ (1-\beta)\widetilde{\Delta}_t,\ \frac{\rho_k(1-\beta)^2\widetilde{\Delta}_t^2}{\|g_{t+1}\|^2 + \|s_{t+1}\|^2} \right\} \\
&\geq \min\left\{ \frac{2\rho_k(1-\beta)\widetilde{\Delta}_t^2}{G_k^2},\ \frac{\rho_k(1-\beta)^2\widetilde{\Delta}_t^2}{2G_k^2} \right\} \\
&\geq \frac{\rho_k(1-\beta)^2\widetilde{\Delta}_t^2}{2G_k^2}
\end{aligned}
$$

where the first inequality uses the definition of θ_{t+1} and drops a norm squared term, the second inequality uses the definition of a null step, and the fourth inequality uses that $2\widetilde{\Delta}_t \leq G_k^2/\rho_k$, $\|g_{t+1}\|^2 \leq G_k^2$, and $\|s_{t+1}\|^2 \leq 2\rho_k\widetilde{\Delta}_t \leq G_k^2$. This verifies (4.16) and completes the proof of our general bound.

For any M-Lipschitz objective, our specialized result follows from observing that $G_k \leq M$ as subgradients everywhere are uniformly bounded in norm by the Lipschitz constant. For any L-smooth objective, the following three inequalities hold for any null step t in the sequence of consecutive null steps following a descent step $k < t$:

$$\|g_{t+1}\| \leq \|g_{k+1}\| + L\|z_{t+1} - x_{k+1}\| \tag{4.18}$$

$$\|z_{t+1} - x_{k+1}\| \leq \|g_{k+1}\|/\rho_k \tag{4.19}$$

$$\|g_{k+1}\| \leq \sqrt{2(L+\rho_k)\Delta_{k+1}}\,. \tag{4.20}$$

Before proving these three inequalities, we note that combined they give the claimed bound as

$$\begin{aligned} G_k = \sup_t\{\|g_{t+1}\|\} &\leq \sup_t\{\|g_{k+1}\| + L\|z_{t+1} - x_{k+1}\|\} \\ &\leq (1 + L/\rho_k)\|g_{k+1}\| \\ &\leq (1 + L/\rho_k)\sqrt{2(L+\rho_k)\Delta_k} \end{aligned}$$

and thus $G_k^2 \leq 2(L+\rho_k)^3\Delta_k/\rho_k^2$. First (4.18) follows directly from the gradient being L-Lipschitz continuous. Second (4.19) follows from considering the ρ_k-strongly convex model proximal subproblem $f_t(z) + \frac{\rho_k}{2}\|z - x_{k+1}\|^2$. Since z_{t+1} uniquely minimizes this,

$$\begin{aligned} \frac{\rho_k}{2}\|z_{t+1} - x_{k+1}\|^2 &\leq f_t(x_{k+1}) - \left(f_t(z_{t+1}) + \frac{\rho_k}{2}\|z_{t+1} - x_{k+1}\|^2\right) \\ &\leq \widetilde{\Delta}_t \leq \widetilde{\Delta}_{k+1} \leq \|g_{k+1}\|^2/\rho_k \end{aligned}$$

where the last inequality uses (4.7). Third (4.20) follows from the L-smoothness of f and considering the full proximal subproblem $f(z) + \frac{\rho_k}{2}\|z - x_{k+1}\|^2$ since

$$\begin{aligned}\Delta_{k+1} &= f(x_{k+1}) - \min_z \left\{ f(z) + \frac{\rho_k}{2}\|z - x_{k+1}\|^2 \right\} \\ &\geq f(x_{k+1}) - \min_z \left\{ f(x_{k+1}) + \langle g_{k+1}, z - x_{k+1} \rangle + \frac{L + \rho_k}{2}\|z - x_{k+1}\|^2 \right\} \\ &= \frac{\|g_{k+1}\|^2}{2(L + \rho_k)} \, .\end{aligned}$$

4.4.3 Proof of Theorem 4.3.1

For a constant stepsize $\rho_k = \rho$, we can simplify the lower bound (4.15) to only depend on x_k through a simple threshold on $f(x_k) - f^*$ as

$$\Delta_k \geq \begin{cases} \dfrac{1}{2\rho}\left(\dfrac{f(x_k) - f^*}{D}\right)^2 & \text{if } f(x_k) - f^* \leq \rho D^2 \\ \dfrac{1}{2}(f(x_k) - f^*) & \text{otherwise.} \end{cases} \tag{4.21}$$

Combining this with Lemma 4.2.1 gives a recurrence relation describing the decrease in the objective gap $\delta_k = f(x_k) - f^*$ on any descent step k of

$$\delta_{k+1} \leq \begin{cases} \delta_k - \dfrac{\beta \delta_k^2}{2\rho D^2} & \text{if } \delta_k \leq \rho D^2 \\ (1 - \beta/2)\delta_k & \text{if } \delta_k > \rho D^2 \, . \end{cases}$$

Our analysis of the bundle method then proceeds considering these two cases separately. In each case, solving the given recurrence relation bounds the number of descent steps possible and applying Lemma 4.2.2 bounds the number of null steps possible.

Bounding steps with $\delta_k > \rho D^2$.

First we show that the number of descent steps with $\delta_k > \rho D^2$ is bounded by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{\rho D^2}\right)}{-\log(1-\beta/2)} \right\rceil_+ \tag{4.22}$$

and the number of null steps with $\delta_k > \rho D^2$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2} \ . \tag{4.23}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (4.22). Index the descent steps before a ρD^2-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ρD^2. Define $k_0 = -1$. Then for each $i = 0 \ldots n-1$,

$$f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\rho D^2 \ .$$

It follows from (4.15) that $\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\rho D^2/2$. Plugging this into Lemma 4.2.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i}\frac{4M^2}{(1-\beta)^2\rho^2 D^2} \ .$$

Summing this over $i = 0 \ldots n-1$ bounds the total number of null steps before a ρD^2-minimizer is found by (4.23) as

$$\sum_{i=0}^{n-1}(1-\beta/2)^{(n-1)-i}\frac{4M^2}{(1-\beta)^2\rho^2 D^2} \leq \frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2} \ .$$

Bounding steps with $\rho D^2 \geq \delta_k > \epsilon$.

Now we complete our proof of Theorem 4.3.1 by bounding the number of descent steps with $\rho D^2 \geq \delta_k > \epsilon$ by

$$\frac{2\rho D^2}{\beta\epsilon} \tag{4.24}$$

and the number of null steps with $\rho D^2 \geq \delta_k > \epsilon$ by

$$\frac{12\rho D^4 M^2}{(1-\beta)^2\epsilon^3}\,. \tag{4.25}$$

After the bundle method has passed objective value ρD^2, the recurrence relation becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\delta_k^2}{2\rho D^2}.$$

Solving this recurrence (see Lemma 4.5.1) implies $\delta_k > \epsilon$ holds for at most (4.24) descent steps. Then we can bound the number of null steps between these descent steps by noting (4.21) implies $\Delta_k \geq (f(x_k) - f^*)^2/2\rho D^2 \geq \epsilon^2/2\rho D^2$. Then Lemma 4.2.2 upper bounds the number of consecutive null steps by $4D^2M^2/(1-\beta)^2\epsilon^2$. Then multiplying this by our bound on the number of descent steps gives (4.25) as

$$\left(\frac{2\rho D^2}{\beta\epsilon} + 1\right)\frac{4D^2M^2}{(1-\beta)^2\epsilon^2} \leq \frac{12\rho D^4 M^2}{\beta(1-\beta)^2\epsilon^3}\,.$$

4.4.4 Proof of Theorem 4.3.4

Our bound on the number of descent steps comes directly from Theorem 4.3.1. Our claimed bound on the total number of null steps follows by multiplying this by the constant bound on the number of consecutive null steps from Lemma 4.2.2.

4.4.5 Proof of Theorem 4.3.6

Assuming Hölder growth (7.3) holds and fixing $\rho_k = \rho$, the lower bound (4.15) simplifies to only depend on a simple threshold with $f(x_k) - f^*$ as

$$\Delta_k \geq \begin{cases} \dfrac{\mu^{2/p}(f(x_k) - f^*)^{2-2/p}}{2\rho} & \text{if } (f(x_k) - f^*)^{1-2/p} \leq \rho/\mu^{2/p} \\ \dfrac{1}{2}(f(x_k) - f^*) & \text{otherwise .} \end{cases} \tag{4.26}$$

From this, we arrive at a recurrence relation on the objective gap $\delta_k = f(x_k) - f^*$ decrease at each descent step k by plugging this lower bound into Lemma 4.2.1 of

$$\delta_{k+1} \leq \begin{cases} \delta_k - \dfrac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho} & \text{if } \delta_k^{1-2/p} \leq \rho/\mu^{2/p} \\ (1 - \beta/2)\delta_k & \text{if } \delta_k^{1-2/p} > \rho/\mu^{2/p} \text{ .} \end{cases}$$

Our analysis proceeds by considering the two cases of this recurrence and the three cases of $p > 2$, $p = 2$, and $1 \leq p < 2$ separately. In each case, solving the given recurrence relation bounds the number of descent steps possible and applying Lemma 4.2.2 bounds the number of null steps possible.

Given $p > 2$, bounding steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$.

First we show that the number of descent steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is bounded by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ \tag{4.27}$$

and the number of null steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2(\rho/\mu^{2/p})^{1/(1-2/p)}} \text{ .} \tag{4.28}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (4.27). Index the descent steps before a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than $(\rho/\mu^{2/p})^{1/(1-2/p)}$. Define $k_0 = -1$. Then for each $i = 0 \dots n-1$,

$$f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}(\rho/\mu^{2/p})^{1/(1-2/p)}.$$

It follows from (4.15) that

$$\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)}(\rho/\mu^{2/p})^{1/(1-2/p)}/2.$$

Plugging this into Lemma 4.2.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i}\frac{4M^2}{(1-\beta)^2(\rho/\mu^{2/p})^{1/(1-2/p)}}\ .$$

Summing this over $i = 0 \dots n-1$ bounds the total number of null steps before a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found by (4.28) as

$$\sum_{i=0}^{n-1}(1-\beta/2)^{(n-1)-i}\frac{4M^2}{(1-\beta)^2(\rho/\mu^{2/p})^{1/(1-2/p)}} \leq \frac{8M^2}{\beta(1-\beta)^2(\rho/\mu^{2/p})^{1/(1-2/p)}}\ .$$

Given $p > 2$, bounding steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$.

Next we show that the total number of descent steps with$(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is bounded by

$$\frac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} \tag{4.29}$$

and the number of null steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is at most

$$\frac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}}\ . \tag{4.30}$$

In this case, the recurrence relation on objective value decrease becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta \mu^{2/p} \delta_k^{2-2/p}}{2\rho}.$$

Applying Lemma 4.5.1 gives our bound on the number of descent steps with $\delta_k > \epsilon$ in (4.29). Plugging the lower bound $\Delta_k \geq \mu^{2/p}(f(x_k) - f^*)^{2-2/p}/2\rho \geq \mu^{2/p}\epsilon^{2-2/p}/2\rho$ into Lemma 4.2.2, the number of consecutive null steps after a descent step is at most

$$\frac{4M^2}{(1-\beta)^2 \mu^{2/p} \epsilon^{2-2/p}}.$$

Then multiplying our limit on consecutive null steps by the number of descent steps between finding a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer and finding an ϵ-minimizer gives the bound (4.30) as

$$\left(\frac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + 1\right) \frac{4M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} \leq \frac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}}.$$

Given $p = 2$, bounding steps with $\delta_k > \epsilon$.

Here both cases of our recurrence relation have a similar form, and so we directly bound the total number of descent steps with $\delta_k > \epsilon$ by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil \tag{4.31}$$

and the number of null steps with $\delta_k > \epsilon$ by

$$\frac{2M^2}{\beta(1-\beta)^2 \min\{\mu/2\rho, 1/2\}\rho\epsilon}. \tag{4.32}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta\min\{\mu/2\rho, 1/2\})\delta_k$. This immediately bounds the number of descent steps by (4.31). Index the descent steps before an ϵ-minimizer

is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ϵ. Define $k_0 = -1$. Then for each $i = 0 \dots n-1$,

$$f(x_{k_i+1}) - f^* \geq (1 - \beta \min\{\mu/2\rho, 1/2\})^{i-(n-1)}\epsilon \, .$$

It follows from (4.15) that $\Delta_{k_i+1} \geq (1 - \beta \min\{\mu/2\rho, 1/2\})^{i-(n-1)}\epsilon/2$. Plugging this into Lemma 4.2.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1 - \beta \min\{\mu/2\rho, 1/2\})^{(n-1)-i} \frac{2M^2}{(1-\beta)^2\rho\epsilon} \, .$$

Summing this over $i = 0 \dots n-1$ bounds the total number of null steps before an ϵ-minimizer is found by

$$\sum_{i=0}^{n-1} (1 - \beta \min\{\mu/2\rho, 1/2\})^{(n-1)-i} \frac{2M^2}{(1-\beta)^2\rho\epsilon} \leq \frac{2M^2}{\min\{\mu/2\rho, 1/2\}\beta(1-\beta)^2\rho\epsilon} \, .$$

Given $1 \leq p < 2$, bounding steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$.

Now we show that the number of descent steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is bounded by

$$\frac{2\rho(f(x_0) - f^*)^{2/p-1}}{(1 - 2^{1-2/p})\beta\mu^{2/p}} \tag{4.33}$$

and the number of null steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} C \tag{4.34}$$

where $C = \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\} \min\left\{\frac{1}{1-2^{-|4/p-3|}}, \left\lceil \log_2\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)\right\rceil\right\}$. Notice that since $p < 2$, the power $1 - 2/p$ of δ_k in the threshold condition of our recurrence is negative. In this case, the recurrence relation on objective value decrease becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho} \, .$$

As an intermediate step, for any $i \geq 0$, we first bound the number of descent and null steps with

$$2^{i+1}(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > 2^i(\rho/\mu^{2/p})^{1/(1-2/p)} .$$

Since descent steps decreases the objective gap by at least $\beta\mu^{2/p}\delta_k^{2-2/p}/2\rho$, there are at most

$$\frac{2\rho(2^i(\rho/\mu^{2/p})^{1/(1-2/p)})^{2/p-1}}{\beta\mu^{2/p}} = \frac{2^{(2/p-1)i+1}}{\beta}$$

descent steps in this interval. Further, noting that in this interval

$$\Delta_k \geq \frac{\mu^{2/p}(2^i(\rho/\mu^{2/p})^{1/(1-2/p)})^{2-2/p}}{2\rho} = 2^{(2-2/p)i-1}(\rho/\mu^{2/p})^{1/(1-2/p)} ,$$

we can bound the number of consecutive null steps following any of these descent steps via Lemma 4.2.2. Hence there are at most

$$\frac{2^{(4/p-3)i+3}M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}$$

null steps in this interval.

The bundle method halves its objective value at most $N = \lceil \log_2((f(x_0) - f^*)/(\rho/\mu^{2/p})^{1/(1-2/p)}) \rceil$ times before an $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found. Then summing up these bounds on the descent and null steps in each interval limits the number of descent steps needed to find a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer by (4.33) as

$$\begin{aligned}
\sum_{i=0}^{N-1} \frac{2^{(2/p-1)i+1}}{\beta} &\leq \frac{2}{\beta}\sum_{i=0}^{N-1} 2^{(2/p-1)i} \\
&\leq \frac{2^{(2/p-1)(N-1)+1}}{(1-2^{1-2/p})\beta} \\
&\leq \frac{2\rho(f(x_0)-f^*)^{2/p-1}}{(1-2^{1-2/p})\beta\mu^{2/p}}
\end{aligned}$$

and similarly, the number of null steps needed by (4.34) as

$$\sum_{i=0}^{N-1} \frac{2^{(4/p-3)i+3}M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}$$
$$\leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \sum_{i=0}^{N-1} 2^{(4/p-3)i}$$
$$\leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\} \min\left\{\frac{1}{1-2^{-|4/p-3|}}, N\right\}$$

where the last inequality bounds the geometric sum regardless of the sign of the exponent $4/p-3$.

Given $1 \leq p < 2$, bounding steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$.

Finally, we bound the number of descent steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ by

$$\left\lceil \frac{\log\left(\frac{(\rho/\mu^{2/p})^{1/(1-2/p)}}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil \tag{4.35}$$

and the number of null steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is at most

$$\frac{4M^2}{\beta(1-\beta)^2\rho\epsilon}\,. \tag{4.36}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (4.35). Index the descent steps after a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer but before an ϵ-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ϵ. Then for each $i = 0 \ldots n-1$,

$$f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\epsilon\,.$$

It follows from (4.15) that $\Delta_{k_i+1} \geq (f(x_{k_i+1})-f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\epsilon/2$. Plugging this into Lemma 4.2.2 upper bounds the number of consecutive null steps

after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1 - \beta/2)^{(n-1)-i} \frac{2M^2}{(1-\beta)^2 \rho \epsilon} .$$

Summing this over $i = 0 \dots n-1$ bounds the additional number of null steps before an ϵ-minimizer is found by (4.36) as

$$\sum_{i=0}^{n-1} (1 - \beta/2)^{(n-1)-i} \frac{2M^2}{(1-\beta)^2 \rho \epsilon} \leq \frac{4M^2}{\beta(1-\beta)^2 \rho \epsilon} .$$

4.4.6 Proof of Theorem 4.3.8

Our bound on the number of descent steps comes directly from Theorem 4.3.6. Our claimed bound on the total number of null steps follows by multiplying this by the constant bound on the number of consecutive null steps from Lemma 4.2.2.

4.4.7 Proof of Theorem 4.3.10

Combining the lower bound $\Delta_k \geq \frac{1}{2}(f(x_k) - f^*)$ with Lemma 4.2.1 shows linear decrease in the objective every descent step

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\beta}{2}\right)(f(x_k - f^*).$$

Our bound on the number of descent steps follows immediately from this. Combining the lower bound $\Delta_k \geq \frac{1}{2}(f(x_k) - f^*)$ with Lemma 4.2.2 shows that at most

$$\frac{2M^2 D^2}{(1-\beta)^2 (f(x_k) - f^*)^2}$$

null steps occur between each descent step. Denote the sequence of descent steps taken by the bundle method by $k_1, k_2, k_3 \dots$ and as a base case define $k_0 =$

-1. Let k_n be the first descent step finding an ϵ-minimizer, which must have $n \leq \lceil \log_{(1-\beta/2)}(\frac{\epsilon}{f(x_0)-f^*}) \rceil_+$. From our linear decrease condition, we know for any $i = 0, 1, 2, 3, \ldots n-1$

$$f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\epsilon$$

and from our null step bound, we know for any $i = 0, 1, 2, \ldots n-1$

$$k_{i+1} - k_i - 1 \leq \frac{2M^2D^2}{(1-\beta)^2(f(x_{k_i+1}) - f^*)^2} \leq (1-\beta/2)^{2(i-(n-1))}\frac{2M^2D^2}{(1-\beta)^2\epsilon^2} \ .$$

Then summing up our null step bounds ensures

$$k_n - n \leq \sum_{i=1}^{n}(1-\beta/2)^{2(i-1-(n-1))}\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}.$$

Bounding this geometric series shows us that the bundle method finds an ϵ-minimizer with the number of null steps bounded by

$$\left(\frac{1}{1-(1-\beta/2)^2}\right)\frac{2M^2D^2}{(1-\beta)^2\epsilon^2} \ .$$

4.4.8 Proof of Theorem 4.3.11

Our bound on the number of descent steps follows from Theorem 4.3.10. Our proof of the null step bound follows the same approach as Theorem 4.3.10 with only minor differences. Applying Lemma 4.2.2 with our stepsize choice (4.13) bounds the number of consecutive null steps after some descent step k by

$$\frac{2M^2}{(1-\beta)^2\mu^{2/p}(f(x_k) - f^*)^{2-2/p}} \ .$$

Denote the descent steps $-1 = k_0 < k_1 < k_2 < \ldots$ and suppose the x_{k_n+1} is the first ϵ-minimizer. Then

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(2-2/p)(i-(n-1))}\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}$$

since $f(x_{k_i+1}) - f^* \geq \left(1 - \frac{\beta}{2}\right)^{i-(n-1)}\epsilon$. Summing this up gives

$$k_n - n \leq \sum_{i=1}^{n}(1-\beta/2)^{(2-2/p)(i-1-(n-1))}\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} .$$

When $p > 1$, this geometric series shows us that the bundle method finds an ϵ-minimizer with the number of null steps bounded by

$$\left(\frac{1}{1-(1-\beta/2)^{2-2/p}}\right)\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}.$$

When $p = 1$, we have a constant upper bound on the number of null steps following a descent step. Hence the number of null steps is bounded by

$$\frac{2M^2}{(1-\beta)^2\mu^2}\left\lceil\frac{\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)}{-\log(1-\beta/2)}\right\rceil .$$

4.4.9 Proof of Theorem 4.3.12

Let $\delta_k = \min_{j\in\{0,\ldots,J-1\}}\{f(x_k^{(j)}) - f^*\}$ denote the lowest objective gap among all of our J instances of the bundle method after they have taken k synchronous steps. Then the core of our convergence proof is bounding the number of iterations where this lowest objective gap is in the interval

$$(1-\beta/2)^{-n}\epsilon \leq \delta_k \leq (1-\beta/2)^{-(n+1)}\epsilon .$$

for any integer $0 \leq n < N := \left\lceil\frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil$. Within this interval, we focus on the instance

$$j = \left\lceil\log_2\left(\frac{\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p}}{4\bar{\rho}}\right)\right\rceil.$$

This instance of the bundle method's constant stepsize $\rho^{(j)} = 2^j\bar{\rho}$ approximates the stepsize (4.13) as

$$\frac{1}{4}\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p} \leq \rho^{(j)} \leq \frac{1}{2}\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p} .$$

Then (4.26) bounds this method's proximal gap before an $(1-\beta/2)^{-n}\epsilon$-minimizer is found by

$$\Delta_k^{(j)} \geq \frac{1}{2}(f(x_k^{(j)}) - f^*) \geq (1-\beta/2)^{-n}\epsilon/2 \,.$$

Letting $\delta_k^{(j)} = f(x_k^{(j)}) - f^*$, each descent step k improves method j's objective gap according to the recurrence

$$\delta_{k+1}^{(j)} \leq \min\{(1-\beta/2)\delta_k^{(j)}, \delta_k\}$$

where the first term in the minimum comes from Lemma 4.2.1 and the second term comes from method j taking any further improvement from the other bundle methods. By assumption, we have $\delta_k \leq (1-\beta/2)^{-(n+1)}\epsilon$, and so after one descent step $k' > k$ we must have $\delta_{k'+1}^{(j)} \leq (1-\beta/2)^{-(n+1)}\epsilon$. Thus after a second descent step $k'' > k'$, our intermediate target accuracy is met as $\delta_{k''+1} \leq \delta_{k''+1}^{(j)} \leq (1-\beta/2)^{-n}\epsilon$.

Applying Lemma 4.2.2 bounds the number of null steps between these descent steps by

$$\frac{2M^2}{(1-\beta)^2\rho^{(j)}\Delta_k^{(j)}} \leq \frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} \,.$$

Hence the total number of steps before $\delta_k^{(j)} < 2^n\epsilon$ (and consequently $\delta_k < 2^n\epsilon$) is at most

$$2\left(\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} + 1\right) .$$

Summing over this bound completes our proof. When $p > 1$, this gives a geo-

metric sum as

$$\begin{aligned}
&\sum_{n=0}^{N-1} 2\left(\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}}+1\right)\\
&= 2\sum_{n=0}^{N-1}\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} + 2\left\lceil\frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil\\
&\leq \left(\frac{2}{1-(1-\beta/2)^{2-2/p}}\right)\frac{16M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} + 2\left\lceil\frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil.
\end{aligned}$$

When $p = 1$, the number of steps in each of our intervals is constant. Consequently, the total number of iterations before an ϵ minimizer is found is at most

$$\sum_{n=0}^{N-1} 2\left(\frac{16M^2}{(1-\beta)^2\mu^2}+1\right) = 2\left(\frac{16M^2}{(1-\beta)^2\mu^2}+1\right)\left\lceil\frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil.$$

4.5 Addendum - Solutions to Recurrence Relations

Throughout our analysis, we frequently encounter recurrence relations of the form $\delta_{k+1} \leq \delta_k - \alpha\delta_k^q$ for some $\alpha > 0$ and $q > 1$. The following lemma bounds the number of steps of such a recurrence to reach a desired level of accuracy $\delta_k \leq \epsilon$.

Lemma 4.5.1. *For any $\epsilon > 0$, the recurrence $\delta_{k+1} \leq \delta_k - \alpha\delta_k^q$ has $\delta_k \leq \epsilon$ satisfied by some*

$$k \leq \left\lceil\frac{1}{(q-1)\alpha\epsilon^{q-1}}\right\rceil.$$

Proof. It suffices to show the following upper bound on δ_k as a function of k

$$\delta_k \leq \left(\frac{1}{(q-1)\alpha k}\right)^{1/(q-1)}.$$

First we show this bound holds with $k = 1$. This follows as

$$\delta_1 \leq \delta_0 - \alpha\delta_0^q \leq \max_{\delta\in\mathbb{R}}\{\delta - \alpha\delta^q\} \leq \left(\frac{1}{q\alpha}\right)^{1/(q-1)}.$$

Then we complete our proof by induction using the following *weighted arithmetic-geometric mean (AM-GM) inequality*, which ensures for any $a, \alpha, b, \beta > 0$,

$$a^\alpha b^\beta \leq \left(\frac{\alpha a + \beta b}{\alpha + \beta}\right)^{\alpha+\beta}.$$

For any $k \geq 1$, the fact that $(k - (q-1)^{-1})(k+1)^{1/(q-1)} \leq k^{q/(q-1)}$, which is the AM-GM inequality with $a = k - (q-1)^{-1}$, $\alpha = 1$, $b = k+1$, $\beta = 1/(q-1)$, completes our proof

$$\begin{aligned}
\delta_{k+1} \leq \delta_k - \alpha\delta_k^q &\leq \left(\frac{1}{(q-1)\alpha k}\right)^{1/(q-1)} - \alpha\left(\frac{1}{(q-1)\alpha k}\right)^{q/(q-1)} \\
&= \left(\frac{1}{(q-1)\alpha}\right)^{1/(q-1)}\left(\frac{k}{k^{q/(q-1)}} - \frac{1}{(q-1)k^{q/(q-1)}}\right) \\
&= \left(\frac{1}{(q-1)\alpha}\right)^{1/(q-1)}\frac{k - (q-1)^{-1}}{k^{q/(q-1)}} \\
&\leq \left(\frac{1}{(q-1)\alpha(k+1)}\right)^{1/(q-1)}.
\end{aligned}$$

□

CHAPTER 5

RADIAL DUALITY: FOUNDATIONS

5.1 Introduction

Renegar [140] introduced a framework for conic programming (and by reduction, convex optimization), which turns such problems into uniformly Lipschitz optimization. After being radially transformed, a simple subgradient method can be applied and analyzed. Notably, even for constrained problems, such an algorithm maintains a feasible solution at each iteration while avoiding the use of orthogonal projections, which can often be a bottleneck for first-order methods. Subsequently, Grimmer [60] showed that in the case of convex optimization, a simplified radial subgradient method can be applied with simpler and stronger convergence guarantees. In [141], Renegar further showed that the transformation of hyperbolic cones is amenable to the application of smoothing techniques, offering notable improvements over radial subgradient methods.

In this chapter, we provide a wholly different development and generalization of the ideas behind Renegar's framework, which avoids relying on convex cones or functions as the central object. Instead, our approach is based on the following simple projective transformation, which we dub the *radial point transformation*, given by

$$\Gamma(x, u) = (x, 1)/u$$

for any $(x, u) \in \mathcal{E} \times \mathbb{R}_{++}$, where $\mathcal{E}$ is some finite dimensional Euclidean space and $\mathbb{R}_{++}$ is the set of positive real numbers. Applying this elementwise to a set

$S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ gives the *radial set transformation*, denoted by

$$\Gamma S = \{\Gamma(x, u) \mid (x, u) \in S\}.$$

To motivate the nomenclature of calling these transformations radial, consider the transformation of a vertical line in $\mathcal{E} \times \mathbb{R}_{++}$: for any $x \in \mathcal{E}$,

$$\Gamma\{(x, \lambda) \mid \lambda \in \mathbb{R}_{++}\} = \{\gamma(x, 1) \mid \gamma \in \mathbb{R}_{++}\}.$$

We see that this transformation maps vertical lines into rays extending from the origin (and rays into vertical lines since the point transformation is dual, $\Gamma\Gamma(x, u) = (x, u)$).

To extend this set operation to apply to functions, we consider functions $f : \mathcal{E} \to \mathbb{R}_{++} \cup \{0, \infty\}$ mapping into the extended positive reals. Then we define the *upper and lower radial function transformations* of f as[1]

$$f^{\Gamma}(y) = \sup\{v > 0 \mid (y, v) \in \Gamma(\text{epi } f)\},$$

$$f_{\Gamma}(y) = \inf\{v > 0 \mid (y, v) \in \Gamma(\text{hypo } f)\}$$

where $\text{epi } f = \{(x, u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \leq u\}$ denotes the epigraph of f and $\text{hypo } f = \{(x, u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \geq u\}$ denotes the hypograph of f. The upper transformation has the interpretation of reshaping the epigraph of f via the radial set transformation and returning the smallest function whose hypograph contains that set. Likewise the lower transformation aims to turn the hypograph of f into the epigraph of a new function.

Connections To Prior Works. Noting $f^{\Gamma}(y) = \sup\{v > 0 \mid v \cdot f(y/v) \leq 1\}$, this relates to the transformation used by Grimmer [60] and those of Renegar [140, 141]. The upper and lower transformations coincide in the convex

[1]Since we are considering functions mapping into $\mathbb{R}_{++} \cup \{0, \infty\}$, if no $v > 0$ satisfies $(y, v) \in \Gamma(\text{epi } f)$, we have the supremum defining $f^{\Gamma}(y)$ equal zero.

settings of these previous works but may diverge in the general setting considered herein. For our analysis, we primarily focus on the upper transformation, but equivalent results always hold for the lower transform. Artstein-Avidan and Milman [8] (and the subsequent [7]) consider the same underlying projective point transformation Γ, but consider the similar but quite different function transformation $\inf\{v > 0 \mid (y, v) \in \Gamma(\text{epi } f)\}$. Considering this transformation limits their theory to the restrictive setting of nonnegative convex functions that minimize to value zero at the origin. As a result, their theory does not provide an interesting duality between optimization problems.

We remark on one other way to view the radial function transformation. Denote the *Minkowski gauge* of a set $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ at some $(y, v) \in \mathcal{E} \times \mathbb{R}_{++}$ by $\gamma_S((y, v)) = \inf\{\lambda > 0 \mid (y, v) \in \lambda S\}$. Then the lower radial transformation can be restated as the restriction of this gauge to $v = 1$

$$f_\Gamma(y) = \gamma_{\text{hypo } f}((y, 1)).$$

This relationship to gauges motivates our notation of Γ to denote the radial transformation. From this point of view, a connection can be made between this radial framework and the perspective duality considered by Aravkin et al [5]. They extend the theory of gauge duality developed by Freund [55], which then applies to nonnegative convex functions by considering perspective functions. The resulting perspective dual optimization problem minimizes the function $\gamma_{\text{hypo } f^*}((y, v))$ where f^* is the Fenchel conjugate of f. Thus their perspective duality can be viewed as a combination of applying Fenchel duality and our radial machinery.

From this connection, we point out a key difference between this radial duality and these previous dualities. The classic theories of Lagrange and gauge

duality are based on a conjugate or polar defined as a supremum over the dual vector space. In contrast, the radial dual and the Minkowski gauge are defined by a one-dimensional problem. This difference allows the radial dual to be efficiently computed numerically for generic problems using a linesearch or bisection, whereas evaluating the Fenchel conjugate of a function is as hard as optimizing over it.

Our Contributions. This work serves to establish the foundations of radial transformations as a new addition to the optimization toolbox. The second part of this work leverages this machinery to develop new radial optimization algorithms. Our development establishes this tool in the following three ways: (i) The radial transformation is dual and enjoys rich structure stemming from this. (ii) The radial transformation produces a new duality between nonnegative optimization problems. For example, constraints are dually transformed into gauges, which allow algorithms to replace orthogonal projections with potentially cheaper, one-dimensional linesearches. We refer any numerically or algorithmically inclined reader to the motivating example of quadratic programming at the start of the next chapter to see this machinery fully in action. (iii) The radial transformation is the unique operation of its kind.

Duality of the radial transformation. We precisely characterize the family of functions where the duality $f^{\Gamma\Gamma} = f$ holds through the star-convexity of their hypograph. Moreover, when this duality holds, we find that a number of important classes of functions are dual to each other or self-dual under the radial

transformations. Namely,

$$\begin{aligned} f \text{ is upper semicontinuous} &\iff f^\Gamma \text{ is lower semicontinuous,} \\ f \text{ is continuous} &\iff f^\Gamma \text{ is continuous,} \\ f \text{ is concave} &\iff f^\Gamma \text{ is convex,} \\ f \text{ is quasiconcave} &\iff f^\Gamma \text{ is quasiconvex,} \\ f \text{ is } k \text{ times differentiable} &\iff f^\Gamma \text{ is } k \text{ times differentiable,} \\ f \text{ is analytic} &\iff f^\Gamma \text{ is analytic} \end{aligned}$$

under appropriate regularity conditions. We also derive a calculus for the radial transformations, providing formulas for the (sub)gradients and Hessians of f^Γ based on those of f.

Radial duality between optimization problems. For a wide range of functions, $\Gamma(\text{epi } f)$ is the hypograph of another function, and so $\text{hypo } f^\Gamma = \Gamma(\text{epi } f)$. As a result, for such functions, points in $\text{hypo } f$ and $\text{epi } f^\Gamma$ can be directly related by the bijection Γ and its inverse (which is also Γ). This relation also applies to the maximizers of f and the minimizers of f^Γ. Namely for any function $f\colon \mathcal{E} \to \mathbb{R}_{++} \cup \{0, \infty\}$, consider the primal problem

$$p^* = \max_{x \in \mathcal{E}} f(x). \tag{5.1}$$

Then the radially dual problem is given by

$$d^* = \min_{y \in \mathcal{E}} f^\Gamma(y) \tag{5.2}$$

and has $\text{argmax } f \times \{p^*\} = \Gamma\big(\text{argmin } f^\Gamma \times \{d^*\}\big)$ under certain regularity conditions. Hence maximizing f is equivalent to minimizing f^Γ.

The radially dual problem (5.2) can exhibit very different behavior than the original problem (5.1). For example, consider the function $f(x) = \sqrt{1 - \|x\|^2}$

defined on the unit ball, which has arbitrarily large gradients and Hessians as x approaches the boundary of the ball. Despite this function's poor behavior, $f^\Gamma(y) = \sqrt{1 + \|y\|^2}$ has full domain with gradients and Hessians bounded in norm by one everywhere. This structure is very appealing for the analysis of first-order optimization methods which tend to heavily rely on these quantities being bounded. The second part of this work utilizes such structure arising from the radial duality developed herein to propose and analyze projection-free radial optimization methods.

Uniqueness of the radial transformation. From our construction of the radial transformation, it is natural to ask if other interesting transformations of optimization problems can be given by reshaping the epigraph of a function. Under some basic assumptions (primarily that the reshaping is invertible and convexity preserving), there are only two transformations of this form, up to affine transformations: the trivial duality between maximizing a function and minimizing its negative and the nontrivial duality given by the radial transformation. These results are similar in spirit to the characterization of order isomorphisms by [7].

Outline Section 5.2 develops theory for the radial point and set transformations on $\mathcal{E} \times \mathbb{R}_{++}$. Informed by this, Section 5.3 derives the core theory establishing our radial function transformations. Then Section 5.4 develops the calculus and optimality relationships between the primal (5.1) and radial dual (5.2). Lastly, Section 5.5 shows that this radial framework is the unique transformation of nonnegative-valued optimization problems of its type.

5.1.1 Notation

We primarily consider sets in $\mathcal{E} \times \mathbb{R}_{++}$, which inherits the standard Euclidean inner product and norm from $\mathcal{E} \times \mathbb{R}$. Denote the ball of radius $r > 0$ around a point $(x, u) \in \mathcal{E} \times \mathbb{R}$ by

$$B((x, u), r) := \{(x', u') \in \mathcal{E} \times \mathbb{R} \mid \|(x', u') - (x, u)\| \leq r\}.$$

Further, denote orthogonal projection onto a closed set $S \subseteq \mathcal{E} \times \mathbb{R}$ by

$$\mathrm{proj}_S((x, u)) := \mathrm{argmin}\{\|(x', u') - (x, u)\| \mid (x', u') \in S\}.$$

Note proj_S is set valued and may not be a singleton if S is not convex.

We consider functions $f \colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, where $\overline{\mathbb{R}}_{++} = \mathbb{R}_{++} \cup \{0, +\infty\}$ denotes the "extended positive reals". Here 0 and $+\infty$ are the limit objects of $\mathbb{R}_{++}$, mirroring the roles of $-\infty$ and $+\infty$ in the extended reals. The effective domain of such a function is denoted by

$$\mathrm{dom}\ f := \{x \in \mathcal{E} \mid f(x) \in \mathbb{R}_{++}\}.$$

Such functions relate to $\mathcal{E} \times \mathbb{R}_{++}$ through their graphs, epigraphs, and hypographs

$$\begin{aligned} \mathrm{graph}\ f &:= \{(x, u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) = u\}, \\ \mathrm{epi}\ f &:= \{(x, u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \leq u\}, \\ \mathrm{hypo}\ f &:= \{(x, u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \geq u\}. \end{aligned}$$

We say a function $f \colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is upper (lower) semicontinuous if $\mathrm{hypo}\ f$ ($\mathrm{epi}\ f$) is closed with respect to $\mathcal{E} \times \mathbb{R}_{++}$. Equivalently, a function is upper semicontinuous if for all $x \in \mathcal{E}$, $f(x) = \limsup_{x' \to x} f(x')$ and lower semicontinuous if $f(x) = \liminf_{x' \to x} f(x')$.

We say a function $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is concave (convex) if $\text{hypo } f$ ($\text{epi } f$) is convex. The set of *convex normal vectors* of a set $S \subseteq \mathcal{E} \times \mathbb{R}$ at some $(x, u) \in S$ is denoted by

$$N_S^C((x,u)) := \{(\zeta, \delta) \mid (\zeta, \delta)^T((x,u) - (x',u')) \geq 0 \; \forall (x',u') \in S\}.$$

Then the *convex subdifferential* of a function f at some $x \in \text{dom } f$ is denoted by

$$\partial_C f(x) := \{\zeta \mid (\zeta, -1) \in N^C_{\text{epi } f}((x, f(x)))\}.$$

Likewise, the *convex supdifferential* of a function f at some $x \in \text{dom } f$ is denoted by

$$\partial^C f(x) := \{\zeta \mid (-\zeta, 1) \in N^C_{\text{hypo } f}((x, f(x)))\}.$$

The elements of these differentials are referred to as convex subgradients or supgradients.

For sets and functions that are not convex, we consider the generalization given by proximal normals and differentials. The set of *proximal normal vectors* of a set $S \subseteq \mathcal{E} \times \mathbb{R}$ at some $(x, u) \in S$ is denoted by

$$N_S^P((x,u)) := \{(\zeta, \delta) \mid (x,u) \in \text{proj}_S((x,u) + \epsilon(\zeta, \delta)) \text{ for some } \epsilon > 0\}.$$

Then the *proximal subdifferential* of a function f at some $x \in \text{dom } f$ is denoted by

$$\partial_P f(x) := \{\zeta \mid (\zeta, -1) \in N^P_{\text{epi } f}((x, f(x)))\}.$$

Likewise, the *proximal supdifferential* of a function f at some $x \in \text{dom } f$ is denoted by

$$\partial^P f(x) := \{\zeta \mid (-\zeta, 1) \in N^P_{\text{hypo } f}((x, f(x)))\}.$$

The elements of these differentials are referred to as proximal subgradients or supgradients.

5.2 The Radial Set Transformation

We begin by observing a number of properties of the radial point and set transformations. Section 5.2.1 uses these to characterize the convex and proximal normal vectors of a radially transformed set. Then Section 5.2.2 concludes with a number of examples and pictures illustrating the radial set transformation. A careful understanding of this operation on sets forms the foundation for understanding the radial function transformation.

One can easily check the point transformation is a continuous analytic bijection on $\mathcal{E} \times \mathbb{R}_{++}$. Further, both the point and set transformations are dual since

$$\Gamma\Gamma(x,u) = \Gamma\frac{(x,1)}{u} = \frac{(x/u,1)}{1/u} = (x,u). \tag{5.3}$$

Now we observe a few basic properties of the set transformation on any pair of sets $S, T \subseteq \mathcal{E} \times \mathbb{R}_{++}$. First, since the point transformation is invertible (in fact, it is its own inverse), the set transformation preserves inclusions between sets, giving

$$S \subseteq T \iff \Gamma S \subseteq \Gamma T. \tag{5.4}$$

Furthermore, the radial set transformation distributes over unions and intersections, giving

$$\Gamma(S \cap T) = \Gamma S \cap \Gamma T, \tag{5.5}$$

$$\Gamma(S \cup T) = \Gamma S \cup \Gamma T. \tag{5.6}$$

Since the radial point transformation is a projective transformation, convex sets, halfspaces, and ellipsoids map into convex sets, halfspaces, and ellipsoids, respectively. For completeness sake, we give direct proofs of these results at

the end of the chapter yielding simple formulas for radially dual halfspaces and ellipsoids in the latter two cases.

Proposition 5.2.1. *A set $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ is convex if and only if ΓS is convex.*

In particular, consider the radial transformation of any halfspace in $\mathcal{E} \times \mathbb{R}_{++}$. Direct manipulation of its definition shows that the radial transformation of a halfspace is another halfspace.

Proposition 5.2.2. *A set $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ is a halfspace if and only if ΓS is a halfspace. In particular, for any halfspace defined by*

$$S = \left\{ (x', u') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix}^T \begin{bmatrix} x' - x \\ u' - u \end{bmatrix} \leq 0 \right\},$$

letting $(y, v) = \Gamma(x, u)$, ΓS is the following halfspace

$$\Gamma S = \left\{ (y', v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ -(\zeta, \delta)^T (x, u) \end{bmatrix}^T \begin{bmatrix} y' - y \\ v' - v \end{bmatrix} \leq 0 \right\}.$$

We say that a set is *polyhedral* if it is the intersection of finitely many halfspaces and $\mathcal{E} \times \mathbb{R}_{++}$. Then as an immediate consequence of Proposition 5.2.2 and (5.5), being polyhedral is preserved under the radial set transformation.

Corollary 5.2.3. *A set $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ is polyhedral if and only if ΓS is polyhedral.*

Lastly, we consider the radial transformation of ellipsoids. A set $S \subseteq \mathcal{E} \times \mathbb{R}$ is an ellipsoid if for some center (x, u) and positive definite linear mapping H,

$$S = \left\{ (x', u') \in \mathcal{E} \times \mathbb{R} \mid \begin{bmatrix} x' - x \\ u' - u \end{bmatrix}^T H \begin{bmatrix} x' - x \\ u' - u \end{bmatrix} \leq 1 \right\}. \tag{5.7}$$

Similar to halfspaces, the radial transformation of such an ellipsoid in $\mathcal{E} \times \mathbb{R}_{++}$ is an ellipsoid in $\mathcal{E} \times \mathbb{R}_{++}$. Curiously, the center of ΓS is not $\Gamma(x, u)$ (as one might expect), but rather the depends on H.

Proposition 5.2.4. *A set S is an ellipsoid if and only if ΓS is an ellipsoid.*

5.2.1 Normal Vectors Under the Radial Set Transformation

Now we consider how the normal vectors of a set relate to those of its radial transformation. Proposition 5.2.2's description of transformed halfspaces characterizes convex normal vectors under the transformation. Combining this result with Proposition 5.2.4's description of transformed ellipsoids gives a characterization for proximal normal vectors.

Proposition 5.2.5. *For any $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$, all $(y, v) \in \Gamma S$ have*

$$N_{\Gamma S}^C((y, v)) = \left\{ \begin{bmatrix} \zeta \\ -(\zeta, \delta)^T(x, u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N_S^C((x, u)) \right\}$$

where $(x, u) = \Gamma(y, v)$.

Proof. For any $(x, u) \in S$, $(\zeta, \delta) \in N_S^C((x, u))$ if and only if

$$S \subseteq \left\{ (x', u') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix}^T \begin{bmatrix} x' - x \\ u' - u \end{bmatrix} \leq 0 \right\}.$$

Letting $(y, v) = \Gamma(x, u)$, Proposition 5.2.2 and (5.4) imply

$$\Gamma S \subseteq \left\{ (y', v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ -(\zeta, \delta)^T(x, u) \end{bmatrix}^T \begin{bmatrix} y' - y \\ v' - v \end{bmatrix} \leq 0 \right\}.$$

Thus $\big(\zeta,\ -(\zeta,\delta)^T(x,u)\big) \in N^C_{\Gamma S}((y,v))$. This gives the containment

$$N^C_{\Gamma S}((y,v)) \supseteq \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_S((x,u)) \right\}$$

and repeating the argument, replacing S by ΓS, gives the radially dual containment

$$N^C_S((x,u)) = N^C_{\Gamma\Gamma S}((x,u)) \supseteq \left\{ \begin{bmatrix} \zeta' \\ -(\zeta',\delta')^T(y,v) \end{bmatrix} \mid \begin{bmatrix} \zeta' \\ \delta' \end{bmatrix} \in N^C_{\Gamma S}((y,v)) \right\}.$$

Applying these two containments in succession shows

$$\begin{aligned} N^C_{\Gamma S}((y,v)) &\supseteq \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_S((x,u)) \right\} \\ &\supseteq \left\{ \begin{bmatrix} \zeta' \\ -(\zeta', -(\zeta',\delta')^T(y,v))^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta' \\ \delta' \end{bmatrix} \in N^C_{\Gamma S}((y,v)) \right\} \\ &= N^C_{\Gamma S}((y,v)) \end{aligned}$$

yielding the claimed formula. □

Proposition 5.2.6. *For any $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$, all $(y,v) \in \Gamma S$ have*

$$N^P_{\Gamma S}((y,v)) = \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^P_S((x,u)) \right\}$$

where $(x,u) = \Gamma(y,v)$.

Proof. Consider any $(x,u) \in S$ and $(\zeta,\delta) \in N^P_S((x,u))$. Then for some $\epsilon > 0$, the ball

$$E = B\left(\begin{bmatrix} x \\ u \end{bmatrix} + \epsilon \begin{bmatrix} \zeta \\ \delta \end{bmatrix},\ \epsilon \left\| \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \right\| \right) \subset \mathcal{E} \times \mathbb{R}_{++}$$

has $E \cap S = \{(x,u)\}$. Recall from Proposition 5.2.4 that ΓE is an ellipsoid. Applying (5.5) implies $\Gamma E \cap \Gamma S = \{(y,v)\}$ where $(y,v) = \Gamma(x,u)$. Since $-(\zeta,\delta) \in N_E^C((x,u))$, Proposition 5.2.5 implies $\left(-\zeta,\ (\zeta,\delta)^T(x,u)\right) \in N_{\Gamma E}^C((y,v))$. Then for sufficiently small $\epsilon' > 0$, the ball

$$E' = B\left(\begin{bmatrix} y \\ v \end{bmatrix} + \epsilon' \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix},\ \epsilon' \left\| \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \right\|\right)$$

lies in ΓE, and hence has $E' \cap \Gamma S = \{(y,v)\}$. Thus $\left(\zeta,\ -(\zeta,\delta)^T(x,u)\right) \in N_{\Gamma S}^P((y,v))$, and so

$$N_{\Gamma S}^P((y,v)) \supseteq \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N_S^P((x,u)) \right\}.$$

As shown in the proof of Proposition 5.2.5, the claimed formula follows from this containment and the dual containment given by replacing S by ΓS. □

5.2.2 Examples and Pictures

In Figures 5.1 through 5.10, we give five examples of pairs of sets in $\mathbb{R} \times \mathbb{R}_{++}$ radially dual to each other. Each figure includes the horizontal line $L = \{(x,1) \mid x \in \mathbb{R}\}$ as a black dashed line. Observe that L is exactly the set of fixed points of the radial point transformation. Further, points above L always map into points below L (and vice versa).

The first two example pairs given in Figures 5.1 and 5.2 and Figures 5.3 and 5.4 show the radial transformation of a halfspace and a polyhedron (which must be a halfspace and a polyhedron by Proposition 5.2.2 and Corollary 5.2.3). Examining the transformation of the horizontal and vertical faces of the square

in Figure 5.3 demonstrates two simple properties of the radial set transformation: (i) horizontal lines map into horizontal lines and (ii) vertical lines map into rays extending away from the origin (and vice versa).

Figures 5.5 and 5.6 show the radial transformation of an ellipsoid (which must be an ellipsoid by Proposition 5.2.4). Figures 5.7 and 5.8 consider the radial set transformation of a parabola, which is nearly an ellipsoid in $\mathbb{R} \times \mathbb{R}_{++}$ but it approaches height 0 at the origin.

Our last pair of examples in Figures 5.9 and 5.10 show the radial set transformation of a sine wave. Notice that the resulting set is not the graph of any function. As we now transition to discussing our radial function transformations, considering how graphs, epigraphs, and hypographs behave under the set transformation provides key intuitions. The fact that the epigraph of our example parabola does not transform into the hypograph of another function and that the graph of our example sine wave does not transform into the graph of another function (as we will see) corresponds to our radial duality not holding for these function.

5.3 The Radial Function Transformation

Recall that we defined two radial function transformations based on the radial set transformation. The *upper radial function transformation* of some $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is defined by

$$\begin{aligned} f^{\Gamma}(y) &= \sup\{v > 0 \mid (y, v) \in \Gamma(\operatorname{epi} f)\} \\ &= \sup\{v > 0 \mid v \cdot f(y/v) \leq 1\}. \end{aligned} \tag{5.8}$$

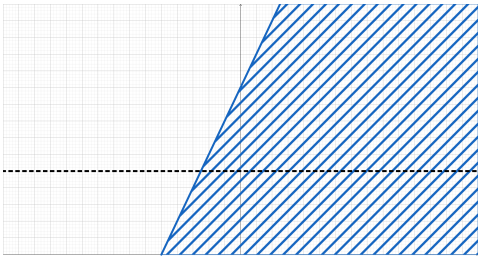


Figure 5.1: A halfspace.

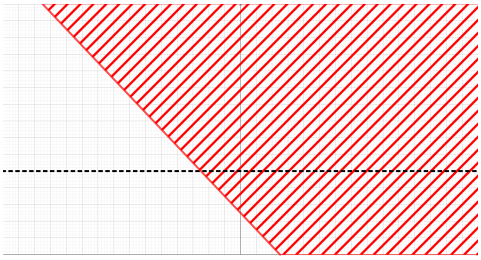


Figure 5.2: Dual halfspace.

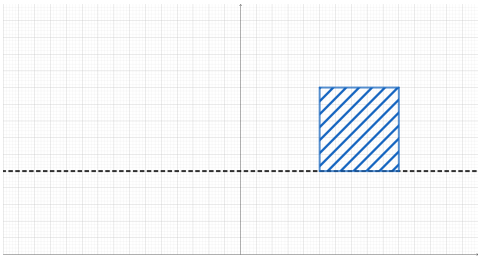


Figure 5.3: A polyhedron.

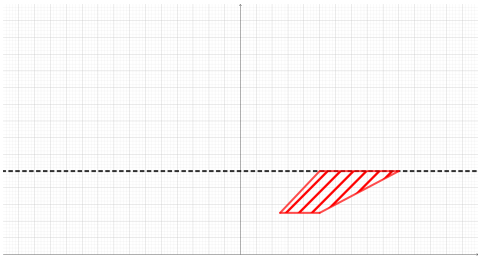


Figure 5.4: Dual polyhedron.

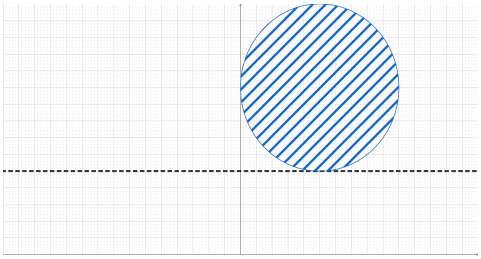


Figure 5.5: An ellipsoid.

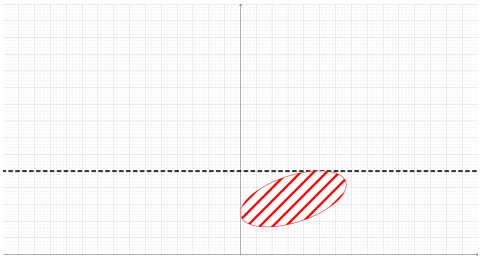


Figure 5.6: Dual ellipsoid.

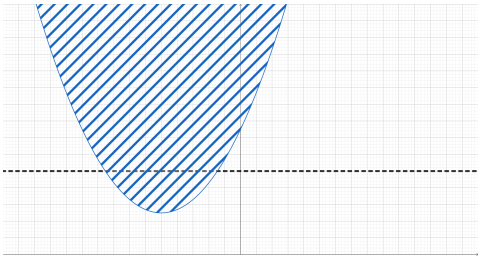


Figure 5.7: A quadratic.

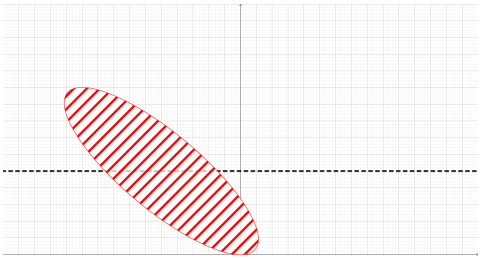


Figure 5.8: Dual of a quadratic.

Figure 5.9: A sine wave.

Figure 5.10: Dual of sine wave.

This transformation essentially applies Γ to the epigraph of f and then interprets $\Gamma(\text{epi } f)$ as the hypograph of a new function. Alternatively, interpreting $\Gamma(\text{hypo } f)$ as the epigraph of a new function gives the *lower radial function transformation* defined by

$$\begin{aligned} f_\Gamma(y) &= \inf\{v > 0 \mid (y, v) \in \Gamma(\text{hypo } f)\} \\ &= \inf\{v > 0 \mid v \cdot f(y/v) \geq 1\}. \end{aligned} \tag{5.9}$$

Based on (5.8) and (5.9), these transformations can alternatively be defined using the *perspective function* of f, which we denote by $f^p(y, v) = v \cdot f(y/v)$. It is immediate from this viewpoint that

$$f^\Gamma = f_\Gamma \iff f^p(y, \cdot) \text{ is nondecreasing and strictly increasing on its domain.} \tag{5.10}$$

Having nondecreasing $f^p(y, \cdot)$ can be understood in terms of the intersection of rays with the epigraph or hypograph of f. The following lemma shows that if $f^p(y, \cdot)$ is nondecreasing, the ray $\{\lambda(y, 1) \mid \lambda > 0\}$ has $\lambda(y, 1)$ lie in the hypograph for all $\lambda < \lambda_0$ and lie in the epigraph for all $\lambda > \lambda_0$ for some $\lambda_0 \in \overline{\mathbb{R}}_{++}$. Thus the hypograph of any such function is star-shaped with respect to the origin.

Lemma 5.3.1. *The following three conditions are equivalent:*
(i) all $y \in \mathcal{E}$ have $f^p(y, \cdot)$ nondecreasing,
(ii) all $(y, v) \in \text{epi } f$ and $t \leq 1$ have $(y, v)/t \in \text{epi } f$,
(iii) all $(y, v) \in \text{hypo } f$ and $t \geq 1$ have $(y, v)/t \in \text{hypo } f$.

Proof. First suppose $f^p(y, \cdot)$ is nondecreasing and consider any $(y, v) \in \text{epi } f$ and $t \leq 1$. Then $t \cdot f(y/t) \leq f(y) \leq v$, and so $(y, v)/t \in \text{epi } f$. Inversely, suppose some $t < t'$ has $f^p(y, t) > f^p(y, t')$. Then every $f(y/t') < \alpha < (t/t') \cdot f(y/t)$

must have $(y/t', \alpha) \in \operatorname{epi} f$. However, dividing this point by $t/t' \leq 1$ gives $(y/t, (t'/t)\alpha) \notin \operatorname{epi} f$. Hence $(i) \iff (ii)$. Symmetric arguments show the equivalent hypograph condition as $(i) \iff (iii)$. □

We say a function f is *upper (lower) radial* whenever $f^p(y, \cdot)$ is nondecreasing and upper (lower) semicontinuous for all $y \in \mathcal{E}$. If in addition $f^p(y, \cdot)$ is strictly increasing on its domain, we say f is *strictly upper (lower) radial*. The following theorem shows being upper (lower) radial is exactly the condition for the duality of the point and set transformations (5.3) to carry over to the upper (lower) radial function transformation.

Theorem 5.3.2. *A function f is upper radial if and only if $f^{\Gamma\Gamma} = f$.*
Likewise[2]*, a function f is lower radial if and only if $f_{\Gamma\Gamma} = f$.*

Proof. Observe that $(f^\Gamma)^p(x, \cdot)$ is nondecreasing since it can be written as

$$\begin{aligned} u \cdot f^\Gamma(x/u) &= u \cdot \sup\{v > 0 \mid v \cdot f(x/vu) \leq 1\} \\ &= \sup\{w > 0 \mid w \cdot f(x/w) \leq u\}. \end{aligned}$$

Then the twice radially transformed function equals the following infimum

$$\begin{aligned} f^{\Gamma\Gamma}(x) &= \inf\{u > 0 \mid u \cdot f^\Gamma(x/u) > 1\} \\ &= \inf\{u > 0 \mid \sup\{w > 0 \mid w \cdot f(x/w) \leq u\} > 1\} \\ &= \inf\{u > 0 \mid \exists w > 1 \text{ s.t. } w \cdot f(x/w) \leq u\} \\ &= \inf\{w \cdot f(x/w) \mid w > 1\}. \end{aligned}$$

The claimed duality follows as $f(x) = \inf\{w \cdot f(x/w) \mid w > 1\}$ if and only if $w \mapsto w \cdot f(x/w)$ is nondecreasing and upper semicontinuous on $w > 0$ for all $x \in \mathcal{E}$. □

[2]Throughout this manuscript, we claim a mirrored results for the lower radial transformation in our theorems and propositions. We omit the proofs for these as they parallel those for the upper radial case.

The radial duality among upper (or lower) radial functions is central to understanding our radial function transformations. In Section 5.3.1, we begin by characterizing when important classes of functions are upper or lower radial (i.e., semicontinuous, differentiable, convex, and concave functions). Then Section 5.3.2 shows being radial is preserved under many standard operations (i.e., conic combinations, linear compositions, minimums, and maximums). Sections 5.3.3, 5.3.4, 5.3.5, and 5.3.6 consider the radial transformations of semicontinuous, piecewise linear, concave/convex, and quasiconcave/quasiconvex functions, respectively. We conclude this section by giving several examples of radial function transformations in Section 5.3.7

5.3.1 Characterizing Radial Functions

Radial Semicontinuous Functions

Here we consider when an upper (or lower) semicontinuous function is upper (or lower) radial, and thus when our duality holds. Unlike Lemma 5.3.1 which focuses on rays from the origin, we give a necessary and sufficient condition based on proximal normal vectors of the function's hypograph (or epigraph). We find it suffices to consider whether the origin lies below the hyperplane induced by each proximal normal vector.

Proposition 5.3.3. *An upper semicontinuous f is upper radial if and only if all $(x, u) \in \operatorname{hypo} f$ and $(\zeta, \delta) \in N^P_{\operatorname{hypo} f}((x, u))$ satisfy*

$$(\zeta, \delta)^T (x, u) \geq 0.$$

Likewise, a lower semicontinuous f is lower radial if and only if all $(x, u) \in \operatorname{epi} f$ and

$(\zeta, \delta) \in N^P_{\text{epi } f}((x,u))$ *satisfy*

$$(\zeta, \delta)^T(x,u) \leq 0.$$

Proof. First suppose f is upper radial and consider any $(x,u) \in \text{hypo } f$ and $(\zeta,\delta) \in N^P_{\text{hypo } f}((x,u))$. By Lemma 5.3.1, $(x,u)/t \in \text{hypo } f$ for all $t \geq 1$. Since $(x,u) \in \text{proj}_{\text{hypo } f}((x,u) + \epsilon(\zeta,\delta))$ for some $\epsilon > 0$, all $t \geq 1$ satisfy

$$\|(x,u) + \epsilon(\zeta,\delta) - (x,u)\|^2 \leq \|(x,u) + \epsilon(\zeta,\delta) - (x,u)/t\|^2.$$

Simplifying this gives

$$0 \leq (1 - 1/t)^2\|(x,u)\|^2 + 2\epsilon(1-1/t)(\zeta,\delta)^T(x,u),$$

and so taking $t \to 1$ verifies $(\zeta,\delta)^T(x,u) \geq 0$.

Note that $f^p(y,\cdot)$ is upper semicontinuous by assumption. Now suppose $f^p(y,\cdot)$ is not nondecreasing. Then Lemma 5.3.1 guarantees some $(x,u) \in \text{hypo } f$ and $\gamma > 1$ has $(x,u)/\gamma \notin \text{hypo } f$. The assumed upper semicontinuity guarantees $\text{hypo } f$ is closed, and thus for some $\epsilon > 0$,

$$B((x,u)/\gamma, \epsilon) \cap \text{hypo } f = \emptyset.$$

Hence the following supremum is well defined

$$\gamma' := \sup\{1 < t \leq \gamma \mid B((x,u)/t, \epsilon/2) \cap \text{hypo } f \neq \emptyset\}.$$

Notice that $1 < \gamma' < \gamma$. Further, $\text{int } B((x,u)/\gamma', \epsilon/2) \cap \text{hypo } f = \emptyset$. Moreover, since $\text{hypo } f$ is closed, some $(x',u') \in \text{hypo } f$ lies on the boundary of this ball – that is, $\|(x,u)/\gamma' - (x',u')\| = \epsilon/2$. Then $\text{hypo } f$ at (x',u') has the following proximal normal vector

$$(\zeta',\delta') := (x,u)/\gamma' - (x',u') \in N^P_{\text{hypo } f}((x',u')).$$

Since all $\gamma' < t < \gamma$ have $(x', u') \notin B((x,u)/t, \epsilon/2)$,

$$\begin{aligned}
\left(\frac{\epsilon}{2}\right)^2 &\leq \left\| \frac{(x,u)}{t} - (x',u') \right\|^2 \\
&= \left\| \frac{\gamma'}{t}(\zeta',\delta') - \left(1 - \frac{\gamma'}{t}\right)(x',u') \right\|^2 \\
&= \left\| \frac{\gamma'}{t}(\zeta',\delta') \right\|^2 - 2\frac{\gamma'}{t}\left(1 - \frac{\gamma'}{t}\right)(\zeta',\delta')^T(x',u') + \left\| \left(1 - \frac{\gamma'}{t}\right)(x',u') \right\|^2 \\
&= \left(\frac{\gamma'}{t}\frac{\epsilon}{2}\right)^2 - 2\frac{\gamma'}{t}\left(1 - \frac{\gamma'}{t}\right)(\zeta',\delta')^T(x',u') + \left\| \left(1 - \frac{\gamma'}{t}\right)(x',u') \right\|^2.
\end{aligned}$$

Rearrangement of this inequality gives

$$2(\zeta',\delta')^T(x',u') \leq -\left(\frac{t}{\gamma'} + 1\right)\left(\frac{\epsilon}{2}\right)^2 + \left(\frac{t}{\gamma'} - 1\right)\|(x',u')\|^2.$$

Taking $t \to \gamma'$ shows $(\zeta',\delta')^T(x',u') \leq -\epsilon^2/2 < 0$. □

Radial Differentiable Functions

Now we specialize the previous result for semicontinuous functions to differentiable functions. We want to allow functions like the previously considered example $f(x) = \sqrt{1 - \|x\|^2}$ (with value 0 whenever $\|x\| \geq 1$) in our theory here. To this end, we say a function is *continuously differentiable* if f is continuous on $\mathcal{E}$ and $\nabla f(x)$ exists and is continuous on its effective domain $\operatorname{dom} f = \{x \in \mathcal{E} \mid f(x) \in \mathbb{R}_{++}\}$.

For any such function f and $x \in \operatorname{dom} f$, if some nonzero $(\zeta,\delta) \in N^P_{\operatorname{epi} f}((x,u))$ exists, then $u = f(x)$ and $(\zeta,\delta) = \lambda(\nabla f(x), -1)$ for some $\lambda \geq 0$. Further, for a dense subset of $\operatorname{dom} f$, the converse holds (which follows from the density theorem of proximal calculus [27, Theorem 1.3.1]). Then the continuity of ∇f and Proposition 5.3.3 imply the following condition is necessary and sufficient

for f to be radial[3].

Proposition 5.3.4. *A continuously differentiable f is radial if and only if for all $x \in \operatorname{dom} f$,*

$$(\nabla f(x), -1)^T (x, f(x)) \leq 0.$$

This characterization can be alternatively derived by considering when the partial derivative $\frac{\partial}{\partial v} f^p(y, v) = f(y/v) - \nabla f(y/v)^T (y/v)$ is nonnegative. Based on this observation, having a positive derivative is sufficient to ensure $f^p(y, \cdot)$ is strictly increasing on its domain, and thus $f^\Gamma = f_\Gamma$ by (5.10).

Proposition 5.3.5. *A continuously differentiable f is strictly radial if for all $x \in \operatorname{dom} f$,*

$$(\nabla f(x), -1)^T (x, f(x)) < 0.$$

Radial Convex and Concave Functions

Lastly we consider conditions for convex or concave functions to be upper or lower radial. For convex functions, the proximal subdifferential and convex subdifferential are equal giving the following characterization.

Proposition 5.3.6. *A lower semicontinuous convex f is lower radial if and only if all $(x, u) \in \operatorname{epi} f$ and $(\zeta, \delta) \in N^C_{\operatorname{epi} f}((x, u))$ have*

$$(\zeta, \delta)^T (x, u) \leq 0.$$

Now we consider concave functions, finding that it suffices to have points arbitrarily close to the origin with nonzero function value. As a result, every

[3] A continuous functions is (strictly) upper radial if and only if it is (strictly) lower radial. In such cases, we simply say the function is (strictly) radial as a shorthand.

upper semicontinuous concave function can be translated to become upper radial.

Proposition 5.3.7. *An upper semicontinuous concave f has $f^p(y,\cdot)$ nondecreasing if and only if $0 \in \operatorname{cl}\{x \mid f(x) > 0\}$ or $f = 0$.*

Proof. Trivially $f = 0$ has $f^p(y,\cdot)$ nondecreasing and so we assume some x' has $f(x') > 0$. Then $f^p(y,\cdot)$ being nondecreasing implies all $t \geq 1$ have $f(x'/t) \geq f(x')/t > 0$. Taking $t \to \infty$ gives a sequence of points verifying $0 \in \operatorname{cl}\{y \mid f(y) > 0\}$.

Conversely, suppose $0 \in \operatorname{cl}\{x \mid f(x) > 0\}$ and consider any $(x,u) \in \operatorname{hypo} f$. Since $\operatorname{hypo} f$ is closed and convex and $(0,0) \in \operatorname{cl}\operatorname{hypo} f$, the line segment $((0,0),(x,u)]$ must lie in $\operatorname{hypo} f$. This is equivalent to $f^p(y,\cdot)$ being nondecreasing by Lemma 5.3.1. □

Furthermore, if the origin lies in the interior of $\{x \mid f(x) > 0\}$, we find that $f^p(x,\cdot)$ is strictly increasing on its domain, and thus $f^\Gamma = f_\Gamma$ by (5.10). This condition can easily be attained for any concave f whenever a point in the interior of the function's domain is known by translating it to the origin. This directly corresponds to the setting assumed by Grimmer [60] and is equivalent to the conic setting assumed by Renegar [140].

Proposition 5.3.8. *A concave f has $f^p(y,\cdot)$ strictly increasing on its domain if $0 \in \operatorname{int}\{x \mid f(x) > 0\}$.*

Proof. Consider any $y \in \mathcal{E}$ and $0 < v < v'$ with $v \cdot f(y/v) \in \mathbb{R}_{++}$. Since $(y/v, f(y/v)) \in \operatorname{hypo} f$, the convexity of f ensures that the line segment $((0,0),(y/v,f(y/v)))$ lies in the interior of the hypograph of f. In particular,

$(v/v') \cdot (y/v, f(y/v)) = (y/v', (v/v') \cdot f(y/v)) \in \operatorname{int} \operatorname{hypo} f$. Therefore $f(y/v') > (v/v')f(y/v)$ and so $f^p(y, v) < f^p(y, v')$. □

5.3.2 Closure of Radial Functions Under Common Operations

Building on our characterizations of when important classes of functions are upper or lower radial, here we show that this structure is preserved under many common operations. The following result shows this is the case for any conic combination (that is, $\sum_{i=1}^{k} \lambda_i f_i(x)$ with each $\lambda_i > 0$), composition with linear maps, and taking finite minimums and maximums.

Proposition 5.3.9. *For any pair of (strictly) upper radial functions f_1, f_2, the following functions are also (strictly) upper radial functions:*
(i) Positive Rescaling by $\lambda > 0$: $\lambda \cdot f_1$,
(ii) Composition with a linear map $A \colon \mathcal{E}' \to \mathcal{E}$: $f_1 \circ A$,
(iii) Addition: $f_1 + f_2$,
(iv) Minimums: $\min\{f_1, f_2\}$,
(v) Maximums: $\max\{f_1, f_2\}$.
Likewise, these operations all preserve being (strictly) lower radial.

Proof. Each of these operations preserves upper and lower semicontinuity and being nondecreasing (or strictly increasing). Consequently, they preserve being (strictly) upper or lower radial. □

Note that (i) and (iii) above together give the claimed result for conic combinations. For all of these operations except addition, we can give simple formulas for the resulting radial transformation, formalized in the following proposition.

Proposition 5.3.10. *For any pair of functions f_1, f_2, the following identities hold for any $\lambda > 0$ and linear $A\colon \mathcal{E}' \to \mathcal{E}$*

$$\begin{aligned}(\lambda \cdot f_1)^\Gamma(y) &= f_1^\Gamma(\lambda y)/\lambda,\\ (f_1 \circ A)^\Gamma &= f_1^\Gamma \circ A,\\ \min\{f_1, f_2\}^\Gamma &= \max\{f_1^\Gamma, f_2^\Gamma\}.\end{aligned}$$

Further, if f_1, f_2 are upper radial, then

$$\max\{f_1, f_2\}^\Gamma = \min\{f_1^\Gamma, f_2^\Gamma\}.$$

Likewise, similar identities hold for the lower radial transformation.

Proof. The results for positive rescaling by some $\lambda > 0$, for composition with a linear map $A\colon \mathcal{E}' \to \mathcal{E}$, and for minimums follow immediately from the definition of our radial transformation as

$$\begin{aligned}(\lambda \cdot f)^\Gamma(y) &= \sup\{v > 0 \mid \lambda v \cdot f(y/v) \leq 1\}\\ &= \sup\{w > 0 \mid w \cdot f(\lambda y/w) \leq 1\}/\lambda\\ &= f^\Gamma(\lambda y)/\lambda,\\ (f \circ A)^\Gamma(y) &= \sup\{v > 0 \mid v \cdot f(Ay/v) \leq 1\}\\ &= f^\Gamma(Ay),\\ \min\{f_1, f_2\}^\Gamma(y) &= \sup\{v > 0 \mid v \cdot \min\{f_1(y/v), f_2(y/v)\} \leq 1\}\\ &= \max_{i=1,2}\{\sup\{v > 0 \mid v \cdot f_i(y/v) \leq 1\}\}\\ &= \max\{f_1^\Gamma(y), f_2^\Gamma(y)\}.\end{aligned}$$

The claimed formula for maximums follows as

$$\begin{aligned}\max\{f_1, f_2\}^\Gamma(y) &= \sup\{v > 0 \mid v \cdot \max\{f_1(y/v), f_2(y/v)\} \leq 1\} \\ &= \min_{i=1,2}\{\sup\{v > 0 \mid v \cdot f_i(y/v) \leq 1\}\} \\ &= \min\{f_1^\Gamma(y), f_2^\Gamma(y)\},\end{aligned}$$

where the second equality above relies on each $v \cdot f_i(y/v)$ being nondecreasing. □

From these simple operations, we can build up to more complex functions that preserve being upper radial. For example, consider the operations taking the kth largest or smallest element out of a set of n elements

$$\begin{aligned}k\text{-}\min\{x_1, \ldots, x_n\} &:= x_{i_k} \text{ where } x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_n} \\ k\text{-}\max\{x_1, \ldots, x_n\} &:= x_{i_{n-k+1}} \text{ where } x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_n}.\end{aligned}$$

and averaging the k largest or smallest elements

$$\begin{aligned}k\text{-minavg}\{x_1, \ldots, x_n\} &:= \frac{1}{k}\sum_{j=1}^{k} x_{i_j} \text{ where } x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_n} \\ k\text{-maxavg}\{x_1, \ldots, x_n\} &:= \frac{1}{k}\sum_{j=1}^{k} x_{i_{n-j+1}} \text{ where } x_{i_1} \leq x_{i_2} \leq \cdots \leq x_{i_n}.\end{aligned}$$

Corollary 5.3.11. *For any (strictly) upper radial functions $f_1, \ldots, f_n$, the functions $k\text{-}\min\{f_i(x)\}$, $k\text{-}\max\{f_i(x)\}$, k-minavg$\{f_i(x)\}$, and k-maxavg$\{f_i(x)\}$ are all (strictly) upper radial with*

$$(k\text{-}\min\{f_i(x)\})^\Gamma(y) = k\text{-}\max\{f_i^\Gamma(y)\}.$$

Proof. This follows immediately from Propositions 5.3.9 and 5.3.10 since these operations can be described as combinations of minimums, maximums, positive

rescaling, and addition

$$k\text{-}\min\{f_1(x),\dots,f_n(x)\} = \min\{\max\{f_i(x) \mid i \in S\} \mid S \subseteq \{1,\dots,n\}, |S| = k\},$$
$$k\text{-}\max\{f_1(x),\dots,f_n(x)\} = \max\{\min\{f_i(x) \mid i \in S\} \mid S \subseteq \{1,\dots,n\}, |S| = k\},$$
$$k\text{-minavg}\{f_1(x),\dots,f_n(x)\} = \min\left\{\frac{1}{k}\sum_{i\in S} f_i(x) \mid S \subseteq \{1,\dots,n\}, |S| = k\right\},$$
$$k\text{-maxavg}\{f_1(x),\dots,f_n(x)\} = \max\left\{\frac{1}{k}\sum_{i\in S} f_i(x) \mid S \subseteq \{1,\dots,n\}, |S| = k\right\}. \qquad \square$$

5.3.3 Radial Transformation of Semicontinuous Functions

Now we turn our focus to understanding how various families of functions behave under the radial transformation. Considering the transformation of upper and lower semicontinuous functions shows that these become lower and upper semicontinuous, respectively, whenever f is appropriately radial.

Proposition 5.3.12. *For any lower semicontinuous, lower radial f, f^Γ is upper semicontinuous. Likewise, for any upper semicontinuous, upper radial f, f_Γ is lower semicontinuous.*

Proof. Consider any $y \in \mathcal{E}$. Upper semicontinuity trivially holds at y if $f^\Gamma(y) = \infty$. Now assume $f^\Gamma(y) < \infty$ and consider any $\gamma > f^\Gamma(y)$. Then $\gamma \cdot f(y/\gamma) > 1$. From the lower semicontinuity of f, for some $\epsilon > 0$, all $y' \in B(y,\epsilon)$ satisfy $\gamma \cdot f(y'/\gamma) > 1$. Therefore $f^\Gamma(y') \leq \gamma$. Taking the limit as γ approaches $f^\Gamma(y)$ shows $f^\Gamma(y) = \limsup_{y'\to y} f^\Gamma(y')$. $\square$

These results with upper and lower semicontinuity reversed do not hold in general. However, whenever $f^\Gamma = f_\Gamma$, the reversed propositions immediately

hold. Thus, for strictly upper (lower) radial functions, upper semicontinuity and lower semicontinuity are dual to each other under the upper (lower) transformation.

Proposition 5.3.13. *For any f with $f^p(y,\cdot)$ strictly increasing on its domain for all $y \in \mathcal{E}$,*

$$\begin{aligned} f \text{ upper semicontinuous} &\implies f^\Gamma = f_\Gamma \text{ lower semicontinuous}, \\ f \text{ lower semicontinuous} &\implies f^\Gamma = f_\Gamma \text{ upper semicontinuous}, \\ f \text{ continuous} &\implies f^\Gamma = f_\Gamma \text{ continuous}. \end{aligned}$$

Proof. Since $f^\Gamma = f_\Gamma$ by (5.10), both directions of Proposition 5.3.12 apply. □

5.3.4 Radial Transformation of Piecewise Linear Functions

We say a function f is *convex polyhedral* if $\operatorname{epi} f$ is the intersection of finitely many halfspaces and $\mathcal{E} \times \mathbb{R}_{++}$. Likewise, f is *concave polyhedral* if $\operatorname{hypo} f$ is the intersection of finitely many halfspaces and $\mathcal{E} \times \mathbb{R}_{++}$. Recall Corollary 5.2.3 ensures polyhedral sets map to polyhedral sets under the radial set transformation. The following proposition shows how this property is mirrored by the radial function transformation on polyhedral functions.

Proposition 5.3.14. *If f is convex polyhedral then f^Γ is concave polyhedral.*
Likewise, if f is concave polyhedral then f_Γ is convex polyhedral.

Proof. If $\operatorname{epi} f$ is polyhedral, then Corollary 5.2.3 implies $\Gamma(\operatorname{epi} f)$ is also polyhedral. Since $\Gamma(\operatorname{epi} f)$ is closed with respect to $\mathcal{E} \times \mathbb{R}_{++}$, the hypograph of f^Γ can

be written as

$$\text{hypo } f^\Gamma = \{(y, v) \mid \exists v' \geq v,\ (y, v') \in \Gamma(\text{epi } f)\}.$$

Then Fourier-Motzkin ensures $\text{hypo } f^\Gamma$ is polyhedral. □

Like the previous results on semicontinuity, the converses do not hold in general. However, whenever $f^p(y, \cdot)$ is strictly increasing on its domain for all $y \in \mathcal{E}$, they immediately hold as $f^\Gamma = f_\Gamma$.

5.3.5 Radial Transformation of Concave/Convex Functions

Recall from Proposition 5.2.1 that the radial set transformation preserves convexity. This structure carries over to the function setting where convex functions become concave and vice versa.

Proposition 5.3.15. *If f is concave then f^Γ and f_Γ are convex.*
Likewise, if f is convex then f^Γ and f_Γ are concave.

Proof. Note that the perspective function $f^p(y, v)$ is concave (convex) whenever f is concave (convex) [19]. Supposing f is concave. Consider any $(y, v), (y', v') \in \text{epi } f^\Gamma$ and $0 \leq \lambda \leq 1$. Note all $t > v$ and $t' > v'$ have $t \cdot f(y/t) > 1$ and $t' \cdot f(y'/t') > 1$. Then the concavity of f^p implies

$$(\lambda t + (1-\lambda)t') \cdot f\left(\frac{\lambda y + (1-\lambda)y'}{\lambda t + (1-\lambda)t'}\right) > 1.$$

Thus $f^\Gamma(\lambda y + (1-\lambda)y') \leq \lambda v + (1-\lambda)v'$ since this holds for all $t > v$ and $t' > v'$.

Now consider any $(y, v), (y', v') \in \text{epi } f_\Gamma$ and $0 \leq \lambda \leq 1$. Note there must exist t, t' near $f_\Gamma(y), f_\Gamma(y')$ with $t \cdot f(y/t) \geq 1$ and $t' \cdot f(y'/t') \geq 1$. Then the concavity

of f^p implies

$$(\lambda t + (1-\lambda)t') \cdot f\left(\frac{\lambda y + (1-\lambda)y'}{\lambda t + (1-\lambda)t'}\right) \geq 1.$$

Thus $f_\Gamma(\lambda y + (1-\lambda)y') \leq \lambda f_\Gamma(y) + (1-\lambda) f_\Gamma(y') \leq \lambda v + (1-\lambda)v'$. □

Thus the family of upper radial concave functions is dual to the family of upper radial convex functions. This is particularly interesting because these families of functions are very different due to the symmetry breaking nature of working with the extended positive reals. Proposition 5.3.7 shows that any concave function can be translated to become radial, whereas no similar operation exists for convex functions. This is a critical algorithmic insight since it allows us to take generic concave maximization problems and transform them into minimization problem that is both convex and upper radial. In the second part of this work, we will see that radially dual minimization problems are very structured, often being globally uniformly Lipschitz continuous, despite us starting with a quite generic maximization problem.

5.3.6 Radial Transformation of Quasi-concave/-convex Functions

Lastly we consider the generalization of concavity and convexity given by quasiconcavity and quasiconvexity. We say a function is *quasiconcave (quasiconvex)* if its superlevel sets $\{x \in \mathcal{E} \mid f(x) \geq z\}$ (sublevel sets $\{x \in \mathcal{E} \mid f(x) \leq z\}$) are convex for all $z > 0$. Similar to the previous section's results, we find that quasiconcave functions are dual to quasiconvex functions (although the additional

condition that $f^p(y,\cdot)$ is nondecreasing is needed for our fullest version of this result to hold).

Proposition 5.3.16. *If f is quasiconcave, f^Γ is quasiconvex. If in addition $f^p(y,\cdot)$ is nondecreasing, f_Γ is quasiconvex. Likewise, if f is quasiconvex, f_Γ is quasiconcave. If in addition $f^p(y,\cdot)$ is nondecreasing, f^Γ is quasiconcave.*

Proof. Suppose f is quasiconcave and fix any level $z \in \mathbb{R}_{++}$. Consider any $0 \leq \lambda \leq 1$ and $y, y' \in \mathcal{E}$ with $f^\Gamma(y) \leq z$ and $f^\Gamma(y') \leq z$. First, we consider the upper radial transformation. Note all $\gamma > z$ have $\gamma \cdot f(y/\gamma) > 1$ and $\gamma \cdot f(y'/\gamma) > 1$. Then the quasiconcavity of f implies

$$\begin{aligned}\gamma \cdot f\left(\frac{\lambda y + (1-\lambda)y'}{\gamma}\right) &\geq \gamma \cdot \min\{f(y/\gamma), f(y'/\gamma)\} \\ &= \min\{\gamma \cdot f(y/\gamma), \gamma \cdot f(y'/\gamma)\} > 1.\end{aligned}$$

Thus $f^\Gamma(\lambda x + (1-\lambda)y) \leq z$ since this holds for every $\gamma > z$.

Now we consider the lower radial transformation and further assume $f^p(y,\cdot)$ is nondecreasing. Note all $\gamma > z$ must have $\gamma \cdot f(y/\gamma) \geq 1$ and $\gamma \cdot f(y'/\gamma) \geq 1$. Then the quasiconvexity of f implies

$$\begin{aligned}\gamma \cdot f\left(\frac{\lambda y + (1-\lambda)y'}{\gamma}\right) &\geq \gamma \cdot \min\{f(y/\gamma), f(y'/\gamma)\} \\ &= \min\{\gamma \cdot f(y/\gamma), \gamma \cdot f(y'/\gamma)\} \geq 1.\end{aligned}$$

Thus $f_\Gamma(\lambda y + (1-\lambda)y') \leq z$ since this holds for every $\gamma > z$. □

5.3.7 Examples and Pictures

In Figures 5.11 through 5.20, we give a number of examples of radial function transformations. As done in our illustrations of the radial set transformation,

each figure includes the horizontal line $\{(x,1) \mid x \in \mathbb{R}\}$ as a black dashed line for reference.

Figures 5.11, 5.12, and 5.13 show the absolute value function and its upper and lower radial transformations respectively. A simple calculation shows

$$|\cdot|^{\Gamma}(y) = \begin{cases} +\infty & \text{if } -1 \leq y \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$|\cdot|_{\Gamma}(y) = \begin{cases} +\infty & \text{if } -1 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note $|x|$ is both upper and lower radial, but not strictly. As a result, although $|\cdot|^{\Gamma}$ and $|\cdot|_{\Gamma}$ are not equal, both are dual to $|x|$ under their respective transformations.

Figures 5.14 and 5.15 show the strictly radial, concave function $\sqrt{1-x^2}$ (with the value at all $x^2 > 1$ set to 0) and its transformation $\sqrt{1+y^2}$. Notice the transformed function is convex as guaranteed by Proposition 5.3.15. Moreover, it is uniformly Lipschitz and smooth even though the original function possesses neither of these properties.

Figures 5.16 and 5.17 show the strictly radial function $e^{-|x|} + 1/2$ and its transformation. Even though this function is neither concave nor convex, our transformation can still be directly applied. Moreover, this function is quasiconcave and so, as guaranteed by Proposition 5.3.16, its radial transformation is quasiconvex.

Lastly, Figures 5.18, 5.19, and 5.20 continue the example of transforming the quadratic $(x+1)^2 + 1/2$ which was used in Figures 5.7 and 5.8 to illustrate the radial set transformation. Notice that this quadratic is not upper radial as its

epigraph transforms into an ellipsoid-like shape rather than the hypograph of another function. Hence its upper radial transformation is not dual to the original quadratic. Figure 5.20 shows the result of applying the upper radial transformation to this quadratic twice. In this case, the functions in Figures 5.19 and 5.20 are upper radial and thus radially dual.

5.4 Optimization Based on Radial Transformations

Here we develop the necessary machinery to propose and analyze optimization methods based on the radial transformation. One powerful facet of the duality between (5.1) and (5.2) lies in how constraints are transformed and the subsequent algorithmic gains. Consider maximizing $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$ over some $S \subseteq \mathcal{E}$. To move the constraints into the objective, we define the indicator function of a set $S \subseteq \mathcal{E}$ as

$$\iota_S(x) = \begin{cases} +\infty & \text{if } x \in S \\ 0 & \text{if } x \notin S. \end{cases}$$

Then it is immediate that the initial problem (5.1) can be written as

$$\max_{x \in S} f(x) = \max_{x \in \mathcal{E}} \min\{f(x), \iota_S(x)\}.$$

Problems of this form have particularly nice radial transformations since it distributes over the minimum by Proposition 5.3.10. Thus the radially dual problem (5.2) is

$$\min_{y \in \mathcal{E}} \max\{f^\Gamma(y), \iota_S^\Gamma(y)\}.$$

For convex S with $0 \in S$, $\iota_S^\Gamma(y)$ is precisely its gauge $\gamma_S(y) = \inf\{\lambda \geq 0 \mid y \in \lambda S\}$. As a result, we find that constraints in the original problem become gauges in

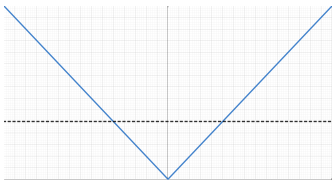


Figure 5.11: $|x|$

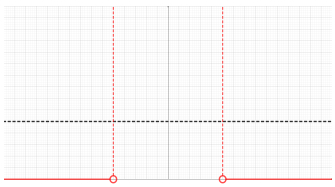


Figure 5.12: $|\cdot|^{\Gamma}(y)$

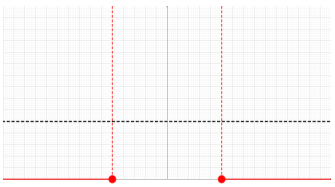


Figure 5.13: $|\cdot|_{\Gamma}(y)$

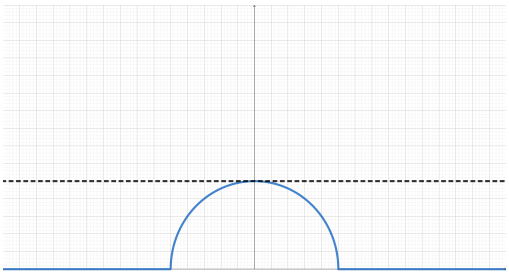


Figure 5.14: $f(x) = \sqrt{1-x^2}$

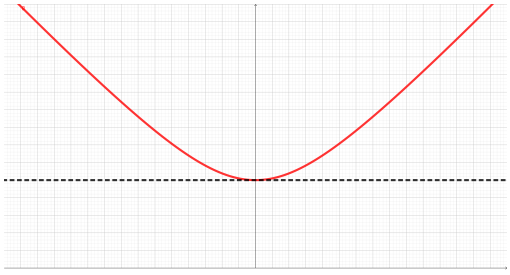


Figure 5.15: $f^{\Gamma} = f_{\Gamma} = \sqrt{1+y^2}$

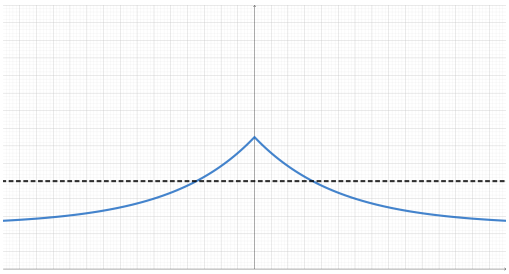


Figure 5.16: $g(x) = e^{-|x|} + 1/2$

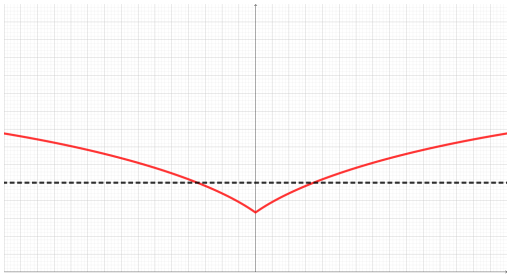


Figure 5.17: $g^{\Gamma}(y) = g_{\Gamma}(y)$

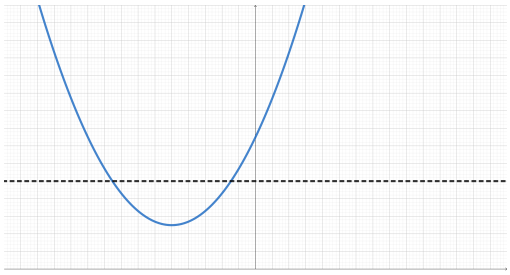


Figure 5.18: $h(x) = (x+1)^2 + 1/2$

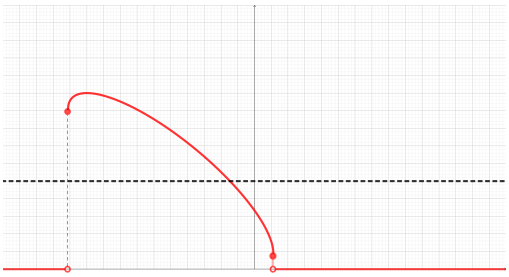


Figure 5.19: $h^{\Gamma}(y)$

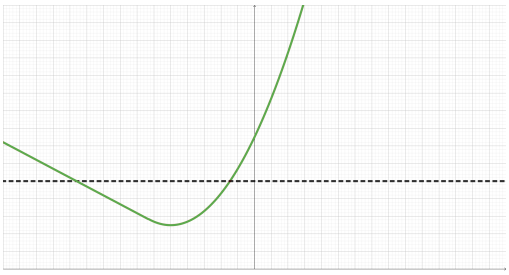


Figure 5.20: $h^{\Gamma\Gamma}(x)$

the radially dual problem:

$$\begin{aligned}\iota_S^{\Gamma}(y) &= \sup\{v > 0 \mid v \cdot \iota_S(y/v) \leq 1\} \\ &= \sup\{v > 0 \mid y/v \notin S\} \\ &= \inf\{\lambda > 0 \mid y \in \lambda S\} \\ &= \gamma_S(y).\end{aligned}$$

This observation is very useful for designing first-order methods. Typically constrained optimization problems require orthogonal projections onto the feasible region at each iteration (which may be substantially more expensive than computing a single subgradient, often dominating an algorithm's runtime). Evaluating the gauge of a set and computing one of its subgradients can be far cheaper than orthogonal projection as it requires at most a one-dimensional line search and computing a single normal vector of the constraint set. Thus the radially dual problem effectively replaces the need for orthogonal projections with much simpler operations. Observing and taking advantage of this structure in the context of conic programming was a central contribution of [140]. Discussing the fuller implications of transforming constraints into their related gauge function on the design of algorithms is deferred to the next chapter.

In the remainder of this section, we develop a calculus for the radially dual optimization problem. For any appropriately radial function, formulas for the convex and proximal subdifferentials and supdifferentials of its radial transformation are given in Section 5.4.1. Further, assuming f is sufficiently differentiable, Section 5.4.2 characterizes the gradients and Hessians of its radial transformations. In Section 5.4.3, we relate the optimal points (minimizers and maximizers) and stationary points of a function and its radial transformations. These calculus and optimality relations form the foundations of relating the pair

of radially dual optimization problems (5.1) and (5.2).

5.4.1 Convex and Proximal Subgradients and Supgradients

To understand the convex and proximal subdifferentials under radial function transformations, we leverage Propositions 5.2.5 and 5.2.6 which described normal vectors under the radial set transformation. The following lemma relates the epigraph and hypograph of radially transformed functions to those of the original function.

Lemma 5.4.1. *For any upper radial f, $\operatorname{epi} f^\Gamma \subseteq \Gamma(\operatorname{hypo} f)$.*
Likewise, for any lower radial f, $\operatorname{hypo} f_\Gamma \subseteq \Gamma(\operatorname{epi} f)$.
If $f^p(y,\cdot)$ is strictly increasing on its domain, equality holds in both cases.

Proof. Noting that $f^\Gamma(y) \leq v \implies v \cdot f(y/v) \geq 1$ for upper radial f, this follows directly as

$$\begin{aligned}\Gamma(\operatorname{hypo} f) &= \left\{ \frac{(x,1)}{u} \mid f(x) \geq u \right\} \\ &= \{(y,v) \mid f(y/v) \geq 1/v\} \\ &= \{(y,v) \mid v \cdot f(y/v) \geq 1\} \\ &\supseteq \left\{(y,v) \mid f^\Gamma(y) \geq v\right\} = \operatorname{epi} f^\Gamma.\end{aligned}$$

When f is strictly upper radial, $f^\Gamma(y) \leq v \iff v \cdot f(y/v) \geq 1$, and so equality holds. □

In light of Lemma 5.4.1, we can immediately apply our results on normal vectors under the radial set transformation to understand differentials under

the function transformation. The following pair of propositions do this for the convex and proximal subdifferential and supdifferential.

Proposition 5.4.2. *For any strictly upper radial f,*

$$\partial_C f^\Gamma(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_{\text{hypo } f}((x,u)),\ (\zeta,\delta)^T(x,u) > 0 \right\}$$

where $(x,u) = \Gamma(y, f^\Gamma(y))$. Likewise, for any strictly lower radial f,

$$\partial^C f_\Gamma(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_{\text{epi } f}((x,u)),\ (\zeta,\delta)^T(x,u) < 0 \right\}.$$

Proof. Recall that Propositions 5.2.5 characterized the convex normal vectors of the radial transformation of a set in terms of the original set. Since the assumed strict increase ensures equality holds in Lemma 5.4.1, this applies to the epigraph and hypograph of f^Γ and f_Γ, respectively. Thus when f is strictly upper radial

$$N^C_{\text{epi } f^\Gamma}((y,v)) = \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_{\text{hypo } f}((x,u)) \right\} \tag{5.11}$$

and when f is strictly lower radial

$$N^C_{\text{hypo } f_\Gamma}((y,v)) = \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_{\text{epi } f}((x,u)) \right\}. \tag{5.12}$$

Then the claimed subgradient and supgradient formulas follow by definition. □

Proposition 5.4.3. *For any strictly upper radial f,*

$$\partial_P f^\Gamma(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^P_{\text{hypo } f}((x,u)),\ (\zeta,\delta)^T(x,u) > 0 \right\}$$

where $(x,u) = \Gamma(y, f^\Gamma(y))$*. Likewise, for any strictly lower radial* f*,*

$$\partial^P f_\Gamma(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^P_{\text{epi } f}((x,u)),\ (\zeta,\delta)^T(x,u) < 0 \right\}.$$

Proof. Recall that Propositions 5.2.6 characterized the proximal normal vectors of the radial transformation of a set in terms of the original set. Since the assumed strict increase ensures equality holds in Lemma 5.4.1, this applies to the epigraph and hypograph of f^Γ and f_Γ, respectively. Thus when f is strictly upper radial

$$N^P_{\text{epi } f^\Gamma}((y,v)) = \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^P_{\text{hypo } f}((x,u)) \right\} \tag{5.13}$$

and when f is strictly lower radial

$$N^P_{\text{hypo } f_\Gamma}((y,v)) = \left\{ \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^P_{\text{epi } f}((x,u)) \right\}. \tag{5.14}$$

Then the claimed subgradient and supgradient formulas follow by definition. □

5.4.2 Gradients and Hessians for Differentiable Functions

Now we narrow our focus to consider differentiable functions under the radial transformation. Whenever f^Γ is differentiable, a formula for its gradient follows from the subgradient formula in Proposition 5.4.3. To establish when f^Γ is differentiable, we show that being k times continuously differentiable (or analytic) is preserved under the radial transformation for appropriate functions. Lastly, we give a formula for the Hessian of the radial transformation of any appropriate twice differentiable function.

As a first step, we give a simple bijection between the graphs (and thus domains) of a function f and its radial transformation f^Γ whenever f is continuous and strictly radial.

Lemma 5.4.4. *For any continuous, strictly radial f,*

$$\text{graph } f^\Gamma = \Gamma(\text{graph } f).$$

Hence, if $y \in \text{dom } f^\Gamma$ then $y/f^\Gamma(y) \in \text{dom } f$.

Proof. Noting that $\text{graph } f = \text{epi } f \cap \text{hypo } f$, we have

$$\begin{aligned}\text{graph } f^\Gamma = \text{epi } f^\Gamma \cap \text{hypo } f^\Gamma &= \Gamma(\text{epi } f) \cap \Gamma(\text{hypo } f) \\ &= \Gamma(\text{epi } f \cap \text{hypo } f) \\ &= \Gamma(\text{graph } f)\end{aligned}$$

where the second equality follows from Lemma 5.4.1 and the third follows from (5.5). □

This lemma lets us view the graph of the radial transformation as the relation $\Gamma(\text{graph } f)$. Applying the implicit function theorem to this relation shows differentiability is preserved under the transformation for appropriate functions. Then leveraging the previous section's results on the proximal subdifferential gives a formula for the gradient of the radial transformation.

Proposition 5.4.5. *Consider any strictly upper radial f and $x, y \in \mathcal{E}$ with $(x, f(x)) = \Gamma(y, f^\Gamma(y))$. Then f is k times continuously differentiable (or analytic) around x with*

$$(\nabla f(x), -1)^T (x, f(x)) < 0$$

if and only if $f^\Gamma = f_\Gamma$ is k times continuously differentiable (or analytic) around y with

$$(\nabla f^\Gamma(y), -1)^T (y, f^\Gamma(y)) < 0$$

where

$$\nabla f^{\Gamma}(y) = \frac{\nabla f(x)}{(\nabla f(x), -1)^T (x, f(x))}.$$

Proof. It suffices to only show the forward direction as the duality of the radial function transformation (Theorem 5.3.2) will then imply the reverse direction. Define the following k times continuously differentiable (or analytic) function

$$F(y', v') = v' \cdot f(y'/v') - 1.$$

Then from Lemma 5.4.4, we know $\operatorname{graph} f^{\Gamma} = \{(y', v') \in \mathcal{E} \times \mathbb{R}_{++} \mid F(y', v') = 0\}$. Noting

$$\frac{\partial}{\partial v} F(y, f^{\Gamma}(y)) = f(y/f^{\Gamma}(y)) - \nabla f(y/f^{\Gamma}(y))^T (y/f^{\Gamma}(y)),$$

we find $\frac{\partial}{\partial v} F(y, f^{\Gamma}(y)) = f(x) - \nabla f(x)^T x > 0$. Thus the implicit function theorem can be applied to produce a k times continuously differentiable (or analytic) function $g \colon U \to \mathbb{R}_{++}$ for some open neighborhood U of y such that

$$\operatorname{graph} f^{\Gamma} \cap (U \times \mathbb{R}_{++}) = \{(y, g(y)) \mid y \in U\}.$$

As a result, f^{Γ} must equal g near y, and hence is also k times continuously differentiable (or analytic) near y.

Now all that remains is to derive our gradient formula and show it satisfies the claimed inequality. Consider any $y \in \operatorname{dom} f^{\Gamma}$ and set $x = y/f^{\Gamma}(y) \in \operatorname{dom} f$. The density theorem of proximal calculus [27, Theorem 1.3.1] guarantees a sequence $y_i \to y$ exists with all $\partial_P f^{\Gamma}(y_i) \neq \emptyset$. Then letting $x_i = y_i/f^{\Gamma}(y_i)$,

$$\nabla f^{\Gamma}(y_i) = \frac{\nabla f(x_i)}{(\nabla f(x_i), -1)^T (x_i, f(x_i))}$$

since $N^P_{\operatorname{epi} f}(x_i, f(x_i)) \subseteq \{\lambda(\nabla f(x_i), -1) \mid \lambda \geq 0\}$. Since $x_i \to x$, the continuous differentiability of f and f^{Γ} ensures

$$\nabla f^{\Gamma}(y) = \frac{\nabla f(x)}{(\nabla f(x), -1)^T (x, f(x))}.$$

From this, its immediate that

$$\begin{aligned}(\nabla f^\Gamma(y), -1)^T(y, f^\Gamma(y)) &= \left(\frac{\nabla f(x)}{(\nabla f(x), -1)^T(x, f(x))}, -1\right)^T \left(\frac{x}{f(x)}, \frac{1}{f(x)}\right) \\ &= \frac{1}{f(x)}\left(\frac{\nabla f(x)^T x}{(\nabla f(x), -1)^T(x, f(x))} - 1\right) \\ &= \frac{1}{(\nabla f(x), -1)^T(x, f(x))} < 0. \end{aligned}$$

□

We remark that this result does not capture all functions for which the radial transformation is differentiable. For example, consider the strictly upper radial function

$$f(x) = \begin{cases} 1 + \sqrt{1 - x^2} & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

This function is not differentiable everywhere in its domain (namely, it fails at $x = \pm 1$). However, its upper radial transformation is differentiable everywhere as it equals

$$f^\Gamma(y) = \begin{cases} (y^2 + 1)/2 & \text{if } -1 \leq y \leq 1 \\ |y| & \text{otherwise.} \end{cases}$$

Differentiating the gradient formula of Proposition 5.4.5 directly gives a Hessian formula for the radial transformation of a function.

Proposition 5.4.6. *Consider strictly upper radial f and $x, y \in \mathcal{E}$ satisfying $(x, f(x)) = \Gamma(y, f^\Gamma(y))$. If f is twice continuously differentiable around x with*

$$(\nabla f(x), -1)^T(x, f(x)) < 0,$$

the Hessian of $f^\Gamma = f_\Gamma$ at y is given by

$$\nabla^2 f^\Gamma(y) = \frac{f(x)}{(\nabla f(x), -1)^T(x, f(x))} \cdot J \nabla^2 f(x) J^T$$

where $J = I - \frac{\nabla f(x) x^T}{(\nabla f(x), -1)^T(x, f(x))}$.

Proof. Denote the bijection relating the domains of f^Γ and f by $\pi(y) = y/f^\Gamma(y)$ (as shown by Lemma 5.4.4). Then the gradient of the radial transformation is

$$\nabla f^\Gamma(y) = \frac{\nabla f(\pi(y))}{(\nabla f(\pi(y)), -1)^T(\pi(y), f(\pi(y)))}.$$

Thus the Jacobian of π is given by

$$\begin{aligned}
\nabla\pi(y) &= I/f^\Gamma(y) - y\nabla f^\Gamma(y)^T/f^\Gamma(y)^2 \\
&= \frac{1}{f^\Gamma(y)}\left(I - \frac{y\nabla f(\pi(y))^T}{f^\Gamma(y)(\nabla f(\pi(y)), -1)^T(\pi(y), f(\pi(y)))}\right) \\
&= f(\pi(y))\left(I - \frac{\pi(y)\nabla f(\pi(y))^T}{(\nabla f(\pi(y)), -1)^T(\pi(y), f(\pi(y)))}\right)
\end{aligned}$$

where the third equality uses that $y/f^\Gamma(y) = \pi(y)$ and $1/f^\Gamma(y) = f(\pi(y))$ by Lemma 5.4.4. Let $g(x) = \nabla f(x)/(\nabla f(x), -1)^T(x, f(x))$ denote $\nabla f^\Gamma \circ \pi^{-1}$. Noting that the gradient of $(\nabla f(x), -1)^T(x, f(x))$ is $\nabla^2 f(x)^T x$, the Jacobian of g is given by

$$\begin{aligned}
\nabla g(x) &= \frac{\nabla^2 f(x)}{(\nabla f(x), -1)^T(x, f(x))} - \frac{\nabla f(x)x^T\nabla f^2(x)}{(\nabla f(x), -1)^T(x, f(x))^2} \\
&= \frac{1}{(\nabla f(x), -1)^T(x, f(x))}\left(I - \frac{\nabla f(x)x^T}{(\nabla f(x), -1)^T(x, f(x))}\right)\nabla^2 f(x)
\end{aligned}$$

Since $\nabla f^\Gamma(y) = g(\pi(y))$, the Hessian of f^Γ is given by $\nabla g(\pi(y))\nabla\pi(y)$ which is exactly the claimed formula. □

5.4.3 Optimality Under the Radial Transformation

Before addressing optimality under our radial duality, we observe that inequalities between functions are reversed by applying either radial function transformation. This mirrors (5.4), where we saw the radial set transformation preserves inclusions between sets. We say $f \leq g$ if $f(x) \leq g(x)$ for all $x \in \mathcal{E}$.

Lemma 5.4.7. *For any functions f, g, if $f \leq g$, then $g^\Gamma \leq f^\Gamma$ and $g_\Gamma \leq f_\Gamma$.*

Proof. Notice that $f \leq g$ is equivalent to $\operatorname{epi} g \subseteq \operatorname{epi} f$. Then (5.4) gives $\Gamma(\operatorname{epi} g) \subseteq \Gamma(\operatorname{epi} f)$. Therefore $f^\Gamma(y) \geq g^\Gamma(y)$ for all $y \in \mathcal{E}$. □

Now we consider how the extreme values and points of a function and its radial transformations relate. First, we show for radial functions, the supremum value of f equals the reciprocal of the infimum value of f^Γ in Proposition 5.4.8. Then Proposition 5.4.9 shows the maximizers of f are related to minimizers of f^Γ by the radial point transformation.

Proposition 5.4.8. *For any function f, $(\inf f) \cdot \left(\sup f^\Gamma\right) = 1$ where we let $\infty \cdot 0 = 0 \cdot \infty = 1$. Further, if f is upper radial, $(\sup f) \cdot \left(\inf f^\Gamma\right) = 1$.*
Likewise, $(\sup f) \cdot (\inf f_\Gamma) = 1$ and if f is lower radial, $(\inf f) \cdot (\sup f_\Gamma) = 1$.

Proof. Observe that $f \geq \inf f$ and so applying Lemma 5.4.7 implies

$$f^\Gamma \leq (\inf f)^\Gamma = 1/\inf f.$$

Now we show the $\geq$ inequality (which is trivial if $\inf f(x) = \infty$). Suppose $\inf f < \infty$. Let x_i be a sequence with $\lim f(x_i) = \inf f$. Then fix any $\epsilon > 0$ and set $y_i = x_i/(f(x_i)+\epsilon)$. Observe that $v \cdot f(y_i/v) < 1$ when $v = 1/(f(x_i)+\epsilon)$. Therefore $f^\Gamma(y_i) \geq 1/(f(x_i)+\epsilon)$ and so taking the limit as $\epsilon \to 0$ implies $\sup f^\Gamma \geq 1/\inf f$. Lastly, supposing f is upper radial, the upper radial transformation is dual by Theorem 5.3.2, and so

$$1 = \left(\inf f^\Gamma\right) \cdot \left(\sup f^{\Gamma\Gamma}\right) = \left(\inf f^\Gamma\right) \cdot (\sup f).$$ □

Proposition 5.4.9. *For any upper radial f with $\sup f \in \mathbb{R}_{++}$,*

$$(\operatorname{argmin} f^\Gamma) \times \{\inf f^\Gamma\} \subseteq \Gamma((\operatorname{argmax} f) \times \{\sup f\}).$$

Likewise for any lower radial f with $\inf f \in \mathbb{R}_{++}$,

$$(\operatorname{argmax} f_\Gamma) \times \{\sup f_\Gamma\} \subseteq \Gamma((\operatorname{argmin} f) \times \{\inf f\}).$$

If $f^p(y,\cdot)$ is strictly increasing on its domain, equality holds in both cases.

Proof. Consider any $(y,v) \in (\operatorname{argmin} f^\Gamma) \times \{\inf f^\Gamma\}$ and set $(x,u) = \Gamma(y,v)$. Then $(y,v) \in \operatorname{epi} f^\Gamma$, and so Lemma 5.4.1 ensures $(x,u) \in \operatorname{hypo} f$. Therefore x attains the maximum value of f by Proposition 5.4.8. Hence $(x,u) \in (\operatorname{argmax} f) \times \{\sup f\}$.

When $f^p(y,\cdot)$ is strictly increasing, equality holds in Lemma 5.4.1 and the above argument can be repeated in reverse. □

For nonconvex optimization problems, finding global solutions is often intractable and so the focus of many optimization methods is on finding stationary points (that is, points with a zero sub(sup)gradient in their sub(sup)differential). Just as optimal solutions were related between the primal and radial dual problems, stationary points are also directly related by the radial point transformation.

Proposition 5.4.10. *For any strictly upper radial f,*

$$\{(y, f^\Gamma(y)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial_P f^\Gamma(y)\} = \Gamma\{(x, f(x)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial^P f(x)\}.$$

Likewise for any strictly lower radial f,

$$\{(y, f_\Gamma(y)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial^P f_\Gamma(y)\} = \Gamma\{(x, f(x)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial_P f(x)\}.$$

Proof. Follows from Proposition 5.4.3. □

5.5 Characterizing Epigraph Reshaping Transformations

In this section, we consider a broader class of transformations given by reshaping a function's epigraph via some mapping G. In particular, given a generic optimization problem

$$p^* = \max_{x \in \mathcal{E}} f(x), \tag{5.15}$$

we transform f by reshaping its epigraph into the hypograph of a new function

$$f^G(y) = \sup\{v \mid (y, v) \in G(\text{epi } f)\}.$$

If $G(\text{epi } f)$ is indeed the a function's hypograph, the identity $\text{hypo } f^G = G(\text{epi } f)$ holds. Then the transformed optimization problem is defined as

$$d^* = \min_{y \in \mathcal{E}} f^G(y). \tag{5.16}$$

Paralleling the development of the radial function transformation f^Γ, we would like to relate minimizers of f^G to maximizers of f and vice versa through the mapping G. To this end, we assume **invertibility**:

$$G \text{ is a bijection.} \tag{A1}$$

Whenever the given function f is concave, we want to preserve this structure by having f^G be convex. To ensure this, we assume G is **convexity preserving**: for any $S \subseteq \text{dom } G$,

$$S \text{ is convex } \implies GS \text{ is convex.} \tag{A2}$$

Lastly, we need a relationship between minimizers of f and maximizers of f^G. This follows by assuming G is **height reversing**: for any pairs $(x, u) = G^{-1}(y, v) \in \text{dom } G$ and $(x', u') = G^{-1}(y', v') \in \text{dom } G$,

$$u \geq u' \implies v \leq v'. \tag{A3}$$

Under these three assumptions, any function f satisfying the identity $\text{hypo } f^G = G(\text{epi } f)$ must have $\text{argmax } f \times \{p^*\} = G^{-1}(\text{argmin } f^G \times \{d^*\})$. Hence the problems (5.15) and (5.16) are equivalent and converting points between these problems only requires evaluating G or its inverse.

First as an example, we consider transforming functions $f : \mathcal{E} \to \mathbb{R} \cup \{\pm\infty\}$ mapping into the extended reals. Then G must map $\mathcal{E} \times \mathbb{R}$ into $\mathcal{E} \times \mathbb{R}$. We may additionally want to impose a condition requiring the function transformation f^G satisfies $\text{hypo } f^G = G(\text{epi } f)$ for a reasonably large class a functions. Namely, we assume the function transformation is **well-defined**: for all linear $f : \mathcal{E} \to \mathbb{R}$,

$$\text{hypo } f^G = G(\text{epi } f). \tag{A4}$$

The Fundamental Theorem of Affine Geometry (stated below) gives us an immediate way to characterize what possible transformations satisfy these four assumptions. See [6, 137] as references.

Theorem 5.5.1. *For $n \geq 2$, if $F : \mathbb{R}^n \to \mathbb{R}^n$ is a bijective, convexity preserving map, then F is an affine transformation.*

Utilizing this, we find that any transformation satisfying these four assumptions must be producing an affinely shifted version of the original problem (5.16). Essentially, any duality of this form amounts to the trivial duality between maximizing a function and minimizing its negative. This is proven in Section 5.5.1.

Theorem 5.5.2. *Consider any map $G : \mathcal{E} \times \mathbb{R} \to \mathcal{E} \times \mathbb{R}$ satisfying* (A1), (A2), *and* (A3). *Then G is an affine map*

$$G(x, u) = \begin{bmatrix} A & \alpha \\ 0 & c \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} + \begin{bmatrix} b \\ d \end{bmatrix}$$

with $c < 0$. *Furthermore, if* (A4) *holds, then* $\alpha = 0$ *and*

$$f^G(y) = cf(A^{-1}(y - b)) + d.$$

Thus there are notable limitations on what a transformation satisfying these four assumptions can accomplish. However, the following theorem provides an alternative to using the Fundamental Theorem of Affine Geometry and facilitates studying more general transformations. See [7] for a reference or [159] for the original version of this result, which takes a more general perspective based in projective spaces.

Theorem 5.5.3. *For* $n \geq 2$, *if for some convex set* $K \subseteq \mathbb{R}^n$ *with nonempty interior,* $F \colon K \to \mathbb{R}^n$ *is an injective, convexity preserving map, then* F *is a fractional linear map.*

This indicates that there is more potential for interesting transformations if we can restrict our assumptions to a convex subset of $\mathcal{E} \times \mathbb{R}$ (in our case, $\mathcal{E} \times \mathbb{R}_{++}$). To this end, we now consider transforming functions $f \colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$ mapping into the extended positive reals.

We also suppose the transformed function $f^G \colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$ maps into the extend positive reals, and so G maps $\mathcal{E} \times \mathbb{R}_{++}$ into $\mathcal{E} \times \mathbb{R}_{++}$. Our first three assumptions (namely, invertibility (A1), convexity preserving (A2), and height reversing (A3)) extend directly to G having restricted domain and codomain. We consider the following assumption paralleling (A4) to ensure the function transformation is **well-defined** for a basic class of linear-like functions: for all linear $f \colon \mathcal{E} \to \mathbb{R}$,

$$\text{hypo } (f_+)^G = G(\text{epi } f_+) \tag{B4}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ denotes nonnegative thresholding. Under these four assumptions, we find that any such transformation of nonnegative-valued optimization problems must produce an affinely shifted version of the upper radial function transformation. This is proven in Section 5.5.2.

Theorem 5.5.4. *Consider any map $G\colon \mathcal{E} \times \mathbb{R}_{++} \to \mathcal{E} \times \mathbb{R}_{++}$ satisfying* (A1), (A2), *and* (A3). *Then G is a fractional linear map*

$$G(x, u) = \frac{(Ax + \alpha u + b, d)}{u}$$

with $d > 0$. Furthermore, if (B4) *holds, then $b = 0$ and the function transformation is given by*

$$f^G(y) = \frac{1}{d} f^\Gamma(A^{-1}(y - \alpha)).$$

This result is similar in spirit to [8, Theorem 5], avoiding their reliance on nonnegative convex functions with value 0 at the origin. Thus the radial set transformation Γ and function transformation f^Γ provide the unique mechanism for deriving equivalent nonnegative-valued optimization problems.

5.5.1 Proof of Theorem 5.5.2

From assumptions (A1) and (A2), Theorem 5.5.1 immediately implies that

$$G(x, u) = \begin{bmatrix} A & \alpha \\ \beta^T & c \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} + \begin{bmatrix} b \\ d \end{bmatrix}.$$

Further, (A3) becomes for all $(x, u), (x', u') \in \mathcal{E} \times \mathbb{R}$,

$$u \leq u' \implies \beta^T x + cu \geq \beta^T x' + cu'.$$

Hence, we must have $\beta = 0$ and $c \leq 0$. Moreover, since G is a bijection, $c < 0$.

Now additionally assume (A4), that this transformation is well-defined for all linear functions. Observe that the inverse of G is given by

$$G^{-1}(y,v) = \begin{bmatrix} A & \alpha \\ 0 & c \end{bmatrix}^{-1} \begin{bmatrix} y-b \\ v-d \end{bmatrix} = \begin{bmatrix} A^{-1}(y-b-\alpha(v-d)/c) \\ (v-d)/c \end{bmatrix}.$$

Suppose for contradiction that $\alpha \neq 0$. Then the linear function $f(x) = -2\alpha^T Ax/\|\alpha\|^2$ has

$$\begin{aligned} G(\text{epi } f) &= \{(y,v) \mid G^{-1}(y,v) \in \text{epi } f\} \\ &= \{(y,v) \mid -2\alpha^T(y-b-\alpha(v-d)/c)/\|\alpha\|^2 \leq (v-d)/c\} \\ &= \{(y,v) \mid -2\alpha^T(y-b) \leq -\|\alpha\|^2(v-d)/c\}. \end{aligned}$$

However, this is not the hypograph of any function as $(b,d) \in G(\text{epi } f)$, but $(b,d-1) \notin G(\text{epi } f)$, which contradicts (A4). Thus we must have $\alpha = 0$. From this, we have $G^{-1}(y,v) = (A^{-1}(y-b),(v-d)/c)$. Then $f^G(y)$ equals $cf(A^{-1}(y-b)) + d$ since

$$\text{hypo } f^G = \{(y,v) \mid f(A^{-1}(y-b)) \leq (v-d)/c\} = \{(y,v) \mid G^{-1}(y,v) \in \text{epi } f\} = G(\text{epi } f).$$

5.5.2 Proof of Theorem 5.5.4

From assumptions (A1) and (A2), Theorem 5.5.3 immediately implies that

$$G(x,u) = \frac{\begin{bmatrix} A & \alpha \\ \beta^T & c \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} + \begin{bmatrix} b \\ d \end{bmatrix}}{\eta^T x + gu + h}.$$

Since G is a bijection, all $(x,u) \in \mathcal{E} \times \mathbb{R}_{++}$ must have $G(x,u) \in \mathcal{E} \times \mathbb{R}_{++}$, and so

$$\frac{\beta^T x + cu + d}{\eta^T x + gu + h} > 0.$$

Thus $\beta = \eta = 0$. Then the mapping $\sigma(u) = (cu + d)/(gu + h)$ must be a bijection from $\mathbb{R}_{++}$ to $\mathbb{R}_{++}$. Therefore σ must be in one of the following two forms: either $g = 0$ and so $\sigma(u) = cu/h$ with $c/h > 0$, or $g \neq 0$ and so $\sigma(u) = d/(gu)$ with $d/g > 0$.

From (A3), the latter of these two possibilities must be the case. Thus $g \neq 0$ and so without loss of generality, we can suppose $g = 1$ and $d > 0$. Hence,

$$G(x, u) = \frac{(Ax + \alpha u + b, d)}{u}.$$

Now additionally assume that this transformation is well-defined for all linear functions (after thresholding to nonnegative values), namely (B4). Observe that the inverse of G is given by

$$G^{-1}(y, v) = \frac{(A^{-1}(y - bv - \alpha), 1)}{dv}.$$

Suppose for contradiction that $b \neq 0$. Then the function $f(x) = (-2b^T Ax/\|b\|^2)_+$ has

$$\begin{aligned}
G(\text{epi } f) &= \{(y, v) \in \mathcal{E} \times \mathbb{R}_{++} \mid G^{-1}(y, v) \in \text{epi } f\} \\
&= \{(y, v) \in \mathcal{E} \times \mathbb{R}_{++} \mid (-2b^T(y - bv - \alpha)/(dv\|b\|^2))_+ \leq 1/(dv)\} \\
&= \{(y, v) \in \mathcal{E} \times \mathbb{R}_{++} \mid -2b^T(y - bv - \alpha)/(dv\|b\|^2) \leq 1/(dv)\} \\
&= \{(y, v) \in \mathcal{E} \times \mathbb{R}_{++} \mid -2b^T(y - \alpha) \leq -\|b\|^2 v\}.
\end{aligned}$$

However, this is not the hypograph of any function as $(b - \alpha, 2)) \in G(\text{epi } f)$, but $(b - \alpha, 1) \notin G(\text{epi } f)$, which contradicts (B4). Thus we must have $b = 0$. From this, we have $G^{-1}(y, v) = (A^{-1}(y - \alpha), 1)/(dv)$. Therefore f^G must be an affine

translation of f^Γ since

$$\begin{aligned} f^G(y) &= \sup\{v > 0 \mid G^{-1}(y, v) \in \text{epi } f\} \\ &= \sup\{v > 0 \mid f(A^{-1}(y - \alpha)/(dv)) \leq 1/(dv)\} \\ &= \frac{1}{d} \sup\{w > 0 \mid f(A^{-1}(y - \alpha)/w) \leq 1/w\} \\ &= \frac{1}{d} f^\Gamma(A^{-1}(y - \alpha)). \end{aligned}$$

5.6 Addendum - Computing Some Radial Set Transformations

Here we present direct proofs of our claimed Propositions 5.2.1, 5.2.2, and 5.2.4 characterizing the result of applying the projective transformation Γ to convex sets, halfspaces, and ellipsoids, respectively.

5.6.1 Proof of Proposition 5.2.1

If suffices to show S being convex implies ΓS is convex, since the duality of the radial set transformation (5.3) will then imply the reverse direction. Consider any $(y, v), (y', v') \in \Gamma S$. Let $(x, u) = \Gamma(y, v)$ and $(x', u') = \Gamma(y', v')$. Then since $(x, u), (x', u') \in S$, all $0 \leq \lambda \leq 1$ have

$$\lambda(x, u) + (1 - \lambda)(x', u') \in S.$$

Therefore the line segment between (y, v) and (y', v') lies in ΓS as

$$\begin{aligned} \Gamma S \ni \frac{(\lambda x + (1 - \lambda)x', 1)}{\lambda u + (1 - \lambda)u'} \\ = \frac{(\lambda y/v + (1 - \lambda)y'/v', 1)}{\lambda/v + (1 - \lambda)/v'} \\ = \frac{\lambda/v}{\lambda/v + (1 - \lambda)/v'}(y, v) + \frac{(1 - \lambda)/v'}{\lambda/v + (1 - \lambda)/v'}(y', v'). \end{aligned}$$

5.6.2 Proof of Proposition 5.2.2

If suffices to show S being a halfspace implies ΓS is a halfspace, since the duality of the radial set transformation (5.3) will then imply the reverse direction. Applying Γ to the definition of S yields

$$\begin{aligned}
\Gamma S &= \left\{ \frac{(x',1)}{u'} \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix}^T \begin{bmatrix} x'-x \\ u'-u \end{bmatrix} \leq 0 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix}^T \begin{bmatrix} y'/v'-x \\ 1/v'-u \end{bmatrix} \leq 0 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix}^T \begin{bmatrix} y'-v'x \\ 1-v'u \end{bmatrix} \leq 0 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix}^T \begin{bmatrix} y' \\ v' \end{bmatrix} + \delta \leq 0 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} \zeta \\ -(\zeta,\delta)^T(x,u) \end{bmatrix}^T \begin{bmatrix} y'-y \\ v'-v \end{bmatrix} \leq 0 \right\}.
\end{aligned}$$

5.6.3 Proof of Proposition 5.2.4

If suffices to show S being an ellipsoid in $\mathcal{E} \times \mathbb{R}_{++}$ implies ΓS is an ellipsoid $\mathcal{E} \times \mathbb{R}_{++}$, since the duality of the radial set transformation (5.3) will then imply the reverse direction. Denote the blocks of H by $\begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix}$ and define the

following matrix

$$G = \begin{bmatrix} H_{11} & -\begin{bmatrix} H_{11} \\ H_{12} \end{bmatrix}^T \begin{bmatrix} x \\ u \end{bmatrix} \\ -\begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} H_{11} \\ H_{12} \end{bmatrix} & \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} - 1 \end{bmatrix},$$

related to the radially dual ellipsoid. In particular, for any ellipsoid $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ defined by (5.7), ΓS is the following ellipsoid in $\mathcal{E} \times \mathbb{R}_{++}$ with center $\begin{bmatrix} y \\ v \end{bmatrix} = G^{-1} \begin{bmatrix} -H_{12} \\ H_{12}^T x + H_{22} u \end{bmatrix}$

$$\Gamma S = \left\{ (y', v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} y' - y \\ v' - v \end{bmatrix}^T \left(\frac{G}{\begin{bmatrix} y \\ v \end{bmatrix}^T G \begin{bmatrix} y \\ v \end{bmatrix} - H_{22}} \right) \begin{bmatrix} y' - y \\ v' - v \end{bmatrix} \leq 1 \right\}.$$

First, we observe that G is indeed positive definite. Since H is postive definite, considering its Schur complements ensures H_{11} is positive definite and $H_{22} - H_{12}^T H_{11}^{-1} H_{12} > 0$. Likewise, G is positive definite if H_{11} is positive definite and

$$\left(\begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} - 1 \right) - \left(\begin{bmatrix} H_{11} & H_{12} \end{bmatrix}^T \begin{bmatrix} x \\ u \end{bmatrix} \right)^T H_{11}^{-1} \left(\begin{bmatrix} H_{11} & H_{12} \end{bmatrix}^T \begin{bmatrix} x \\ u \end{bmatrix} \right)$$

is positive. Simplifying this condition for G to be positive definite yields the equivalent inequality $u^2(H_{22} - H_{12}^T H_{11}^{-1} H_{12}) > 1$. This must be the case since $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$ and so $H_{22} - H_{12}^T H_{11}^{-1} H_{12} > 1/u^2$.

Applying Γ to the definition of S and completing the square gives the claim as

$$\begin{aligned}
\Gamma S &= \left\{ \frac{(x',1)}{u'} \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} x'-x \\ u'-u \end{bmatrix}^T \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \begin{bmatrix} x'-x \\ u'-u \end{bmatrix} \leq 1 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} y'/v'-x \\ 1/v'-u \end{bmatrix}^T \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \begin{bmatrix} y'/v'-x \\ 1/v'-u \end{bmatrix} \leq 1 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} y'-v'x \\ 1-v'u \end{bmatrix}^T \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix} \begin{bmatrix} y'-v'x \\ 1-v'u \end{bmatrix} \leq v'^2 \right\} \\
&= \left\{ (y',v') \in \mathcal{E} \times \mathbb{R}_{++} \mid \begin{bmatrix} y' \\ v' \end{bmatrix}^T G \begin{bmatrix} y' \\ v' \end{bmatrix} + 2H_{12}^T y' - 2(H_{12}^T x + H_{22}u)v' \leq -H_{22} \right\}.
\end{aligned}$$

CHAPTER 6

RADIAL DUALITY: APPLICATIONS AND ALGORITHMS

6.1 Introduction

The previous chapter established a theory of radial duality relating nonnegative optimization problems through a projective transformation, extending the ideas of Renegar [140] from their origins in conic programming. We give a minimal overview here of our radial duality theory needed to begin algorithmically benefiting from it and then a fuller but terse summary in Section 6.2.3 of the core results necessary to derive our radial optimization guarantees.

For a finite dimensional Euclidean space $\mathcal{E}$, our three transformations of interest are the radial point transformation, radial set transformation, and upper radial function transformation, which are denoted by

$$\Gamma(x, u) = (x, 1)/u,$$

$$\Gamma S = \{\Gamma(x, u) \mid (x, u) \in S\},$$

$$f^\Gamma(y) = \sup\{v > 0 \mid (y, v) \in \Gamma(\text{epi } f)\}$$

for any point $(x, u) \in \mathcal{E} \times \mathbb{R}_{++}$, set $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$, and function $f \colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, respectively. Here $\overline{\mathbb{R}}_{++}$ denotes the extended positive reals $\mathbb{R}_{++} \cup \{0, +\infty\}$. It is immediate that the point and set transformations are dual since

$$\Gamma\Gamma(x, u) = \Gamma\frac{(x, 1)}{u} = \frac{(x/u, 1)}{1/u} = (x, u).$$

Central to establishing our theory of radial duality is the characterization of exactly when this duality carries over to the function transformation. We say a

function f is *upper radial* if the perspective function $f^p(y, v) = v \cdot f(y/v)$ is upper semicontinuous and nondecreasing in $v \in \mathbb{R}_{++}$. Moreover, it is *strictly upper radial* if it is strictly increasing in v whenever $f^p(y, v) \in \mathbb{R}_{++}$. The cornerstone theorem of our radial duality Theorem 5.3.2 is that

$$f = f^{\Gamma\Gamma} \iff f \text{ is upper radial.} \tag{6.1}$$

The duality of the radial function transformation provides a duality between optimization problems. For any strictly upper radial function $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, consider the primal problem

$$p^* = \max_{x \in \mathcal{E}} f(x). \tag{6.2}$$

Then the radially dual problem is given by

$$d^* = \min_{y \in \mathcal{E}} f^\Gamma(y) \tag{6.3}$$

and has $(\operatorname{argmax} f) \times \{p^*\} = \Gamma\big((\operatorname{argmin} f^\Gamma) \times \{d^*\}\big)$. Thus maximizing f is equivalent to minimizing f^Γ and solutions can be converted between these problems by applying the radial point transformation Γ or its inverse (which is also Γ by duality).

Importantly, the two nonnegative optimization problems (6.2) and (6.3) can exhibit very different structural properties. For example, consider maximizing $f(x) = \sqrt{1 - \|x\|_{2+}^2}$ which takes value zero outside the unit ball and has arbitrarily large gradients and Hessians as x approaches the boundary of this ball. Its radial dual $f^\Gamma(y) = \sqrt{1 + \|y\|_2^2}$ has full domain with gradients and Hessians bounded in norm by one everywhere. Thus our radial duality theory poses an opportunity to extend the reach of many standard optimization algorithms reliant on such structure. The previous works of Renegar [140] and

Grimmer [60] analyzing subgradient methods and Renegar [141] employing accelerated smoothing techniques on a radial reformulation of the objective critically rely on the reformulation being uniformly Lipschitz continuous, which always occurs in the special cases of the radial dual that they consider.

Our Contributions. This chapter leverages our radial duality theory to present and analyze projection-free radial optimization algorithms in this newfound, wider context than previous works were able to. Finding that a mild condition ensures the radial dual is uniformly Lipschitz continuous, we analyze a radial subgradient method for a broad range of non-Lipschitz primal problems with or without concavity. Observing that constraints radially transform into related gauges, we propose a radial smoothing method that takes advantage of this structure for concave maximization. Further, we find that our radial transformation extends smoothness on a level set of the primal to hold globally in the radial dual, which prompts our analysis of a radial accelerated method. Of greater importance than these particular algorithms, this chapter aims to demonstrate the breadth of applications and algorithms that can be approached using our radial duality theory.

Outline. We begin with a motivating example of the computational benefits and scalability that follow from designing algorithms based on the radial dual (6.3) in Section 6.2. Then Section 6.3 formally establishes algorithmically useful properties of our radial dual, namely Lipschitz continuity, smoothness, and growth conditions. Finally, Section 6.4 addresses the convergence of our radial algorithms for concave maximization and Section 6.5 addresses applications and guarantees in nonconcave maximization.

6.2 A Motivating Setting of Polyhedral Constraints

We begin by motivating the algorithmic usefulness of our radial duality by considering optimization with polyhedral constraints. Consider any maximization problem with upper semicontinuous objective $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ and m inequality constraints $a_i^T x \leq b_i$ given by

$$\begin{cases} \max_x & f(x) \\ \text{s.t.} & Ax \leq b. \end{cases} \tag{6.4}$$

We assume this problem is feasible. Then without loss of generality, we have $0 \in \operatorname{int}(\{x \mid Ax \leq b\} \cap \{x \mid f(x) > 0\})$. This can be achieved by computing any point x_0 in the relative interior of $\{x \mid Ax \leq b\} \cap \{x \mid f(x) \in \mathbb{R}\}$ and then (i) translating the problem to place x_0 at the origin, (ii) adding a constant to the objective to ensure $f(0) > 0$, and (iii) if needed, re-parameterizing the problem[1] to only consider the smallest subspace containing $\{x \mid Ax \leq b\} \cap \{x \mid f(x) > 0\}$. Note that doing this translation suffices to guarantee that any concave f will have $f_+(x) := \max\{f(x), 0\}$ be strictly upper radial by Proposition 5.3.8. We will only make the weaker assumption here that f_+ is strictly upper radial rather than the narrower case of it being concave. Then this problem can be reformulated as the following nonnegative optimization problem of our primal form (6.2)

$$\begin{cases} \max_x & f_+(x) \\ \text{s.t.} & Ax \leq b \end{cases} = \max_x \min_i \Big\{ f_+(x),\ \iota_{a_i^T x \leq b_i}(x) \Big\}$$

[1]Instead of using a re-parameterization, one can explicitly include equality constraints in our model. The details of this approach are given in Section 6.2.2, where we see that equality constraints are unaffected by the radial dual.

where $\iota_{a_i^T x \leq b_i}(x) = \begin{cases} +\infty & \text{if } a_i^T x \leq b_i \\ 0 & \text{if } a_i^T x > b_i \end{cases}$ is an indicator function for each inequality constraint. Note that each $\iota_{a_i^T x \leq b_i}$ is strictly upper radial since 0 is strictly feasible and so applying Proposition 5.3.9 ensures the primal objective $\min_i \Big\{ f_+(x),\ \iota_{a_i^T x \leq b_i}(x) \Big\}$ is strictly upper radial. Then we can compute the radially dual optimization problem (6.3) using Proposition 5.3.10 as

$$\min_y \max_i \big\{ f_+^\Gamma(y),\ a_i^T y / b_i \big\} \tag{6.5}$$

since the radial transformation of each indicator functions is linear

$$\begin{aligned} \iota^\Gamma_{a_i^T x \leq b_i}(y) &= \sup \Big\{ v > 0 \mid v \cdot \iota_{a_i^T x \leq b_i}(y/v) \leq 1 \Big\} \\ &= \sup \big\{ v > 0 \mid a_i^T (y/v) > b_i \big\} \\ &= (a_i^T y / b_i)_+ . \end{aligned}$$

We drop the nonnegative thresholding on each $a_i^T y / b_i$ above since $f_+^\Gamma(y)$ is nonnegative.

Importantly, the dual formulation (6.5) is unconstrained, unlike the primal, since the primal inequality constraints have transformed into simple linear lower bounds on the radially dual objective. This dual further profits from the structure of its objective function as it is often globally Lipschitz continuous (a common property among radial duals that we will show in Proposition 6.3.1) and has the simple form of a finite maximum. This radially dual structure gives us an algorithmic angle of attack not available in the primal problem.

6.2.1 Quadratic Programming

To make these benefits concrete, consider solving a generic quadratic program of the following form

$$\begin{cases} \max_x & 1 - \frac{1}{2}x^T Q x - c^T x \\ \text{s.t.} & Ax \leq b \end{cases} \tag{6.6}$$

for some $Q \in \mathbb{R}^{n \times n}$, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m_{++}$. We reformulate this problem as the following nonnegative optimization problem of the form (6.2)

$$\begin{cases} \max_x & 1 - \frac{1}{2}x^T Q x - c^T x \\ \text{s.t.} & Ax \leq b \end{cases} = \max_x \min_i \left\{ (1 - \frac{1}{2}x^T Q x - c^T x)_+,\ \iota_{a_i^T x \leq b_i}(x) \right\}.$$

Whenever this primal objective is strictly upper radial, the radial dual of our quadratic program is[2]

$$\min_y \max_i \left\{ \left(\frac{c^T y + 1 + \sqrt{(c^T y + 1)^2 + 2y^T Q y}}{2} \right)_+,\ a_i^T y / b_i \right\} \tag{6.7}$$

where the first term in our maximum is set to zero if $(c^T y + 1)^2 + 2y^T Q y < 0$ as can occur for nonconcave primal objectives. We find that our radial duality holds here whenever $\frac{1}{2}x^T Q x > -1$ for all $Ax \leq b$. This captures two natural settings: (i) when the primal objective is concave (as Q is positive semidefinite) or (ii) when the primal objective is nonconcave but has a compact feasible region (since we can rescale the objective to be $1 - \lambda x^T Q x/2 - \lambda c^T x$ without changing the

[2]Our calculation of the radial dual of the quadratic objective $(1 - \frac{1}{2}x^T Q x - c^T x)_+$ follows by definition as

$$\begin{aligned} (1 - \frac{1}{2}x^T Q x - c^T x)_+^\Gamma(y) &= \sup\left\{ v > 0 \mid v\left(1 - \frac{y^T Q y}{2v^2} - \frac{c^T y}{v}\right) \leq 1 \right\} \\ &= \sup\{ v > 0 \mid v^2 - \frac{1}{2}y^T Q y - (c^T y + 1)v \leq 0 \} \\ &= \left(\frac{c^T y + 1 + \sqrt{(c^T y + 1)^2 + 2y^T Q y}}{2} \right)_+. \end{aligned}$$

set of maximizers but ensuring $\frac{\lambda}{2}x^TQx > -1$ everywhere). Section 6.5.1 shows more generally that any differentiable objective with compact constraints can be rescaled to apply our radial duality theory.

We verify that our primal objective is strictly upper radial (and so our radial duality holds) for this upper semicontinuous objective by checking when $f^p(y,\cdot)$ is strictly increasing on its domain. The partial derivative with respect to v of the perspective function

$$\begin{aligned} & v \cdot \min_i \left\{ (1 - \frac{1}{2}(y/v)^T Q(y/v) - c^T(y/v))_+,\ \iota_{a_i^T x \leq b_i}(y/v) \right\} \\ &= \begin{cases} v\left(1 - \frac{y^TQy}{2v^2} - \frac{c^Ty}{v}\right) & \text{if } A(y/v) \leq b \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

is $1 + \frac{y^TQy}{2v^2}$ at every feasible y/v. This is always positive (and hence the perspective function is increasing in v) exactly when every $x = y/v$ with $Ax \leq b$ has $\frac{1}{2}x^TQx > -1$.

Quadratic Programming Numerics

As previously noted, the radially dual formulation (6.7) is unconstrained and Lipschitz continuous despite the primal possessing neither of these properties. This differs from the structure found from taking a Lagrange dual [37] or gauge dual [55]. As a result, our radial dual is well set up for the application of a subgradient method. We consider the following *radial subgradient method* with stepsizes $\alpha_k > 0$ defined by Algorithm 6.

Algorithm 6 The Radial Subgradient Method

Require: $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, $x_0 \in \operatorname{dom} f$, $T \geq 0$

$(y_0, v_0) = \Gamma(x_0, f(x_0))$ *Transform into the radial dual*

for $k = 0 \dots T-1$ **do**

  $y_{k+1} = y_k - \alpha_k \zeta'_k$, where $\zeta'_k \in \partial_P f^\Gamma(y_k)$ *Run the subgradient method*

end for

$(x_T, u_T) = \Gamma(y_T, f^\Gamma(y_T))$ *Transform back to the primal*

return x_T

Further noting that the radially dual problem is a finite maximum of simple smooth Lipschitz functions, we can apply the smoothing ideas of Nesterov [127]. Perhaps the most clear description of these techniques is given by Beck and Teboulle [15]. In particular, for any fixed $\eta > 0$, we consider the smooth function given by taking a "soft-max"

$$g_\eta(y) = \eta \log\left(\exp\left(\frac{c^T y + 1 + \sqrt{(c^T y + 1)^2 + 2y^T Q y}}{2\eta}\right) + \sum_{i=1}^{m} \exp\left(\frac{a_i^T y}{b_i \eta}\right)\right) \quad (6.8)$$

which approaches our radially dual objective as $\eta \to 0$. Then we can minimize the radial dual up to accuracy $O(\eta)$ by minimizing this smoothed objective. Doing so with Nesterov's accelerated method gives the following *radial smoothing method* defined by Algorithm 7 (a similar radial algorithm was employed by Renegar [141] showing that the transformation of any hyperbolic programming problem also admits a smoothing that can be efficiently minimized).

Algorithm 7 The Radial Smoothing Method

Require: $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, $x_0 \in \operatorname{dom} f$, $\eta > 0$, $L_\eta > 0$, $T \geq 0$

$(y_0, v_0) = \Gamma(x_0, f(x_0))$ and $\tilde{y}_0 = y_0$ *Transform into the radial dual*

Let $g_\eta(y)$ denote an η-smoothing of $f^\Gamma(y)$

for $k = 0 \dots T-1$ **do**

  $\tilde{y}_{k+1} = y_k - \nabla g_\eta(y_k)/L_\eta$ *Run the accelerated method*

  $y_{k+1} = \tilde{y}_{k+1} + \frac{k-1}{k+2}(\tilde{y}_{k+1} - \tilde{y}_k)$

end for

$(x_T, u_T) = \Gamma(y_T, f^\Gamma(y_T))$ *Transform back to the primal*

return x_T

The per iteration cost of these radial methods is controlled by the cost of evaluating one subgradient of the radially dual objective (6.7) or one gradient of our smoothing of the radially dual objective (6.8). Both of these can be done efficiently in closed form in terms of the two matrix-vector products Ay and Qy. Despite this low iteration cost, a feasible primal solution $(x_k, u_k) = \Gamma(y_k, f^\Gamma(y_k))$ is known at every iteration. Convergence guarantees for the radial subgradient and smoothing methods for concave maximization are given later in Sections 6.4.1 and 6.4.2.

Classic optimization algorithms that preserve feasibility at every iteration tend to have much higher iteration costs. Here we compare with three of the most standard first-order methods that enforce feasibility: projected gradient descent (or rather, projected gradient ascent)

$$x_{k+1} = \text{proj}_{\{x|Ax\leq b\}}(x + \nabla f(x)/L),$$

an accelerated projected gradient method

$$\begin{cases} \tilde{x}_{k+1} & = \text{proj}_{\{x|Ax\leq b\}}(x_k + \nabla f(x_k)/L) \\ x_{k+1} & = \tilde{x}_{k+1} + \frac{k-1}{k+2}(\tilde{x}_{k+1} - \tilde{x}_k), \end{cases}$$

and the Frank-Wolfe method[3] with stepsize sequence $\beta_k > 0$

$$\begin{cases} \tilde{x}_{k+1} & \in \text{argmax}_x \big\{ \nabla f(x_k)^T x \mid Ax \leq b \big\} \\ x_{k+1} & = x_k + \beta_k(\tilde{x}_{k+1} - x_k). \end{cases}$$

All three of these methods require solving a subproblem at each iteration. The projected gradient and accelerated gradient methods require repeated projection onto the polyhedron $\{x \mid Ax \leq b\}$, which is itself an instance of (6.6)

[3]Quadratic programming was, in fact, the original motivating setting for the Frank-Wolfe algorithm [54].

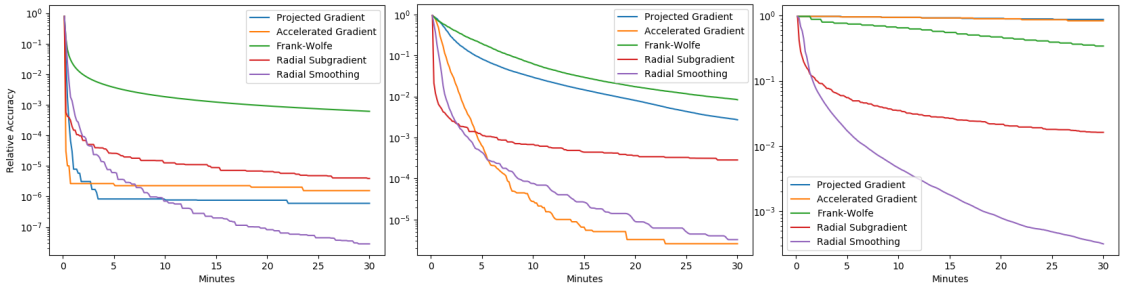
specialized to $Q = I$. The Frank-Wolfe method requires repeatedly solving a linear program over this polyhedron. Both of these operations are far more expensive than the matrix-vector products required by the radial subgradient and smoothing methods but may allow them to have a greater improvement in objective value per iteration[4].

To weigh this tradeoff, we consider running these five algorithms on synthetic quadratic programs given by drawing two matrices $A \in \mathbb{R}^{m \times n}$ and $P \in \mathbb{R}^{n \times 100}$ and a vector $c \in \mathbb{R}^n$ with i.i.d. Guassian entries and setting $Q = PP^T$ and all $b_i = 1$. Then we run each algorithm for 30 minutes on instances of size $(n, m) \in \{(100, 400), (400, 1600), (1600, 6400)\}$. Our numerical experiments are conducted on a four-core Intel i7-6700 CPU using Julia 1.4.1 and Gurobi 9.0.3 to solve any subproblems[5]. For each method, we set $x_0 = 0$ and use the following choice of stepsizes: the projected and accelerated gradient methods use $L = \lambda_{max}(Q)$, the Frank-Wolfe method uses an exact linesearch $\beta_k = \min\left(\frac{\nabla f(x_k)^T(\tilde{x}_{k+1} - x_k)}{\|P^T(\tilde{x}_{k+1} - x_k)\|^2}, 1\right)$, the radial subgradient method uses the Polyak stepsize $\alpha_k = \frac{f^\Gamma(y_k) - d^*}{\|\zeta_k'\|^2}$, and the radial smoothing method fixes $L_\eta = 0.1 \max\{\|a_i/b_i\|^2\}/\eta$ and $\eta \in \{10^{-8}, 10^{-8}, 10^{-6}\}$ for each of our three problem sizes.

The best primal objective value seen by each method is shown in realtime in Figure 6.1. First, we remark on the total number of iterations completed by each method in the allotted half hour, shown in the following table.

[4]There are other QP solvers like OSQP [162] that also only rely on cheap matrix operations, but such operator splitting methods do not maintain a feasible solution at each iteration. Hence they cannot be compared as directly.

[5]The source code is available at github.com/bgrimmer/Radial-Duality-QP-Example



(a) $(n, m) = (100, 400)$ (b) $(n, m) = (400, 1600)$ (c) $(n, m) = (1600, 6400)$

Figure 6.1: The minimum relative accuracy $\frac{p^* - f(x_k)}{p^*}$ of (6.6) seen by the projected gradient, accelerated gradient, Frank-Wolfe, radial subgradient and radial smoothing methods over 30 minutes.

(n, m)	$(100, 400)$	$(400, 1600)$	$(1600, 6400)$
Projected Grad	17,788 iter.	487 iter.	7 iter.
Accelerated Grad	18,412 iter.	506 iter.	7 iter.
Frank-Wolfe	24,137 iter.	333 iter.	26 iter.
Radial Subgrad	8,950,726 iter.	6,835,355 iter.	213,381 iter.
Radial Smoothing	3,827,988 iter.	757,829 iter.	39,005 iter.

In our largest problem setting $(n, m) = (1600, 6400)$, which has approximately ten million nonzeros, the projected gradient, accelerated gradient, and Frank-Wolfe methods only compute a couple dozen steps within our time budget whereas our radial methods take tens or hundreds of thousands of steps. For any larger problem instances, these classic methods may not even complete a single step and so the radial subgradient and smoothing methods vacuously outperform them.

At every scale of problem size, the radial smoothing method is competitive. For our smallest instance $(n, m) = (100, 400)$, the accelerated and projected gradient methods quickly reach an accuracy around 10^{-6}, which is the default tol-

erance of Gurobi, followed shortly afterward by the radial smoothing method. However, for our moderate-sized instance $(n, m) = (400, 1600)$, the classic methods begin to fall off with the radial smoothing method and accelerated method performing comparably. For our largest instance $(n, m) = (1600, 6400)$, the methods relying on orthogonal projection make essentially no progress due to their high iteration cost, and the Frank-Wolfe method only makes minor amounts of progress relative to our radial methods. Our algorithms based on radial duality appear to provide a far more scalable approach.

Throughout our experiments, the radial smoothing method outperforms the radial subgradient method by a couple orders of magnitude. This agrees with our convergence theory showing that the radial subgradient method converges at a $O(1/\epsilon^2)$ rate while the smoothing technique enables $O(1/\epsilon)$ convergence, presented in Sections 6.4.1 and 6.4.2, respectively.

6.2.2 Broader Computational Advantages of Considering Radially Dual Problems

We conclude this motivating section with a high-level discussion of the computational advantages we see in optimizing over radially dual problem formulations. These benefits all extend beyond the particular radial optimization algorithms considered herein.

Maintaining Primal Feasible Iterates Without Costly Projections

Here we generalize the setting of polyhedral constraints considered by (6.4). After a translation, any convex constraints can be expressed as the intersection of a convex set $S \subseteq \mathcal{E}$ with $0 \in \operatorname{int} S$ and a subspace $T = \{x \in \mathcal{E} \mid Ax = 0\}$. Consider any primal problem with strictly upper radial objective f given by

$$\begin{cases} \max & f(x) \\ \text{s.t.} & x \in S \\ & Ax = 0 \end{cases} = \max_{x \in \mathcal{E}} \min\{f(x), \iota_S(x), \iota_T(x)\}$$

where $\iota_S(x) = \begin{cases} +\infty & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases}$ Then the radially dual problem is

$$\min_{y \in \mathcal{E}} \max\{f^\Gamma(y), \gamma_S(y), \gamma_T(y)\} = \begin{cases} \min & \max\{f^\Gamma(y), \gamma_S(y)\} \\ \text{s.t.} & Ay = 0 \end{cases}$$

where $\gamma_S(y) = \inf\{\lambda \geq 0 \mid y \in \lambda S\}$ denotes the Minkowski gauge since

$$\iota_S^\Gamma(y) = \sup\{v > 0 \mid v \cdot \iota_S(y/v) \leq 1\} = \inf\{\lambda > 0 \mid y \in \lambda S\} = \gamma_S(y).$$

Having multiple set constraints $S_1 \dots S_n$ in the primal $\max_{x \in S_1 \cap \dots \cap S_n} f(x)$ simply adds more terms to the radially dual finite maximum of $\min_{y \in \mathcal{E}} \max\{f^\Gamma(y), \gamma_{S_i}(y)\}$.

This formulation allows algorithms to maintain a feasible primal solution at each iteration without requiring costly subproblems relating to S. Instead, a primal feasible solution can be recovered from any radial dual solution $y \in \mathcal{E}$ with $Ay = 0$ as $x = y/\max\{f^\Gamma(y), \gamma_S(y)\} \in S \cap T$. Algorithmically, this replaces the need for orthogonal projections onto the feasible region $S \cap T$ with the cheaper operations of orthogonally projecting onto the subspace T and evaluating the

gauge of S. This computational gain was one of the key contributions identified by [140] and was central to the motivation of [141, 60] as well as being a motivation of this work.

Handling Nonconcave Objectives and Nonconvex Constraints

Our calculation of the radial dual for quadratic programming did not fundamentally rely on concavity as it also applies to nonconcave problems with a bounded feasible region. Indeed one of the key insights from the first part of this work was divorcing the idea of radial transformations from relying on notions of convexity or concavity. In Section 6.5.1, we discuss several nonconcave primal maximization problems where radial duality holds, generalizing the above reasoning to star-convex constraints and covering important areas like nonconvex regularization and optimization with outliers.

Efficiently Evaluating Generic Radial Duals

A remark on the efficiency of computing the upper radial function transformation $f^\Gamma(y)$: In general, we do not have a closed-form as we found in our quadratic programming example. However, numerically evaluating $f^\Gamma(y)$ is a one-dimensional subproblem that can be solved by bisection whenever f is upper radial (since $v \mapsto vf(y/v)$ is then nondecreasing). Even if f is not upper radial, $f^\Gamma(y)$ may still be tractable to compute. For example, any polynomial f has evaluation of f^Γ amount to polynomial root finding. Once $f^\Gamma(y)$ has been computed, its gradients and Hessian follow from (6.25) and (6.26).

Improving Conditioning and Problem Structure

As a final motivating example of the structural advantages of taking the radial dual, consider the following Poisson inverse problem. Given linear measurements with Poisson distribution noise $b_i \sim \texttt{Poisson}(a_i^T x)$, the maximum likelihood estimator is given by maximizing

$$\mathcal{L}(x) := \begin{cases} \sum_i b_i \log(a_i^T x) - a_i^T x & \text{if all } a_i^T x > 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Then given any convex regularizer $r(x)$ and constraint set $S \subseteq \mathbb{R}^n$, we formulate a Poisson inverse problem as

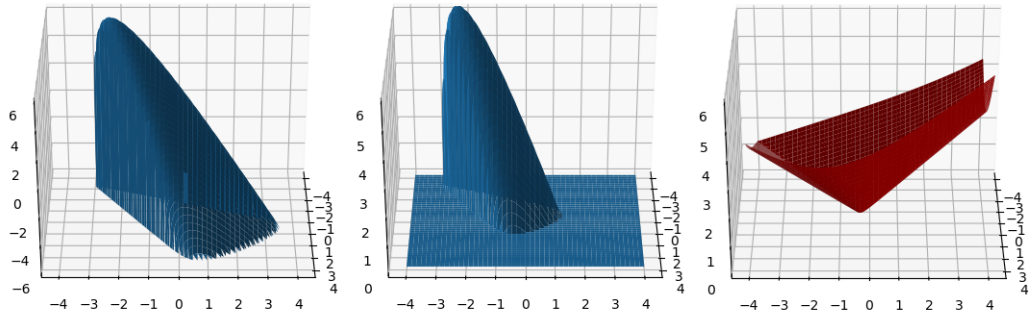
$$\max_{x \in S} \mathcal{L}(x) - r(x) \tag{6.9}$$

This type of problem arises in image processing (see [16] for a survey of applications from astronomy to medical imaging) as well as in network diffusion and time series modeling (see the many references in [75]). Although this problem is concave, the blow-up from the logarithmic terms prevents standard first-order methods from being applied. Provided the regularization r and constraints S are sufficiently simple, customized primal-dual [75] or Bregman methods [13, 115] provide a powerful tactic for solving this problem.

For generic S and r, our radial duality can be applied. Given any $x_0 \in \operatorname{int} \operatorname{dom} \mathcal{L} \cap S$ and $u_0 < \mathcal{L}(x_0) - r(x_0)$, we can reformulate this objective function to be strictly upper radial via a simple translation and truncation. We consider the equivalent problem of

$$\max_{x \in \mathbb{R}^n} \min\{(\mathcal{L}(x + x_0) - r(x + x_0) - u_0)_+, \iota_S(x + x_0)\}.$$

Then we can employ our radial duality machinery using [62, Proposition 3.8] since our translated and truncated objective is concave with 0 strictly in its do-



(a) $\mathcal{L}(x + x_0) - u_0$ (b) $(\mathcal{L}(x + x_0) - u_0)_+$ (c) $(\mathcal{L}(x + x_0) - u_0)_+^\Gamma(y)$

Figure 6.2: Example (a) translating, (b) truncating, and then (c) taking the radial dual of (6.9).

main. The radial dual here is defined everywhere $\operatorname{dom} f^\Gamma = \mathbb{R}^n$, is globally uniformly Lipschitz continuous (see Proposition 6.3.1) and if $S = \mathbb{R}^n$ and $r(x)$ is twice continuously differentiable, has globally Lipschitz continuous gradient (see Corollary 6.3.3). The primal formulation is none of these. Note that different translations of the objective (here corresponding to a different choice of (x_0, u_0)) produce different radial duals, which in turn can have very different global Lipschitz and smoothness constants.

Figure 6.2 shows the steps of taking the radial dual of a two-dimensional likelihood maximization problem with $\{a_1, a_2, a_3\} = \{(2, -1), (1, 1), (-1, 2)\}$, $b_i = 1$, $S = \mathbb{R}^2$, and $r = 0$: (a) shows the translated objective with $x_0 = (3, 3)$ and $u_0 = -10$, (b) shows the truncated strictly upper radial nonnegative optimization problem, and (c) shows the well behaved radially dual objective.

6.2.3 Notation and Review

We consider functions $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, where $\overline{\mathbb{R}}_{++} = \mathbb{R}_{++} \cup \{0, +\infty\}$ denotes the "extended positive reals". Here 0 and $+\infty$ are the limit objects of $\mathbb{R}_{++}$, mirroring

the roles of $-\infty$ and $+\infty$ in the extended reals. The effective domain, graph, epigraph, and hypograph of such a function are

$$\begin{aligned}
\operatorname{dom} f &:= \{x \in \mathcal{E} \mid f(x) \in \mathbb{R}_{++}\},\\
\operatorname{graph} f &:= \{(x,u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) = u\},\\
\operatorname{epi} f &:= \{(x,u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \leq u\},\\
\operatorname{hypo} f &:= \{(x,u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \geq u\}.
\end{aligned}$$

We say a function $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is upper (lower) semicontinuous if $\operatorname{hypo} f$ ($\operatorname{epi} f$) is closed with respect to $\mathcal{E} \times \mathbb{R}_{++}$. Equivalently, a function is upper semicontinuous if for all $x \in \mathcal{E}$, $f(x) = \limsup_{x' \to x} f(x')$ and lower semicontinuous if $f(x) = \liminf_{x' \to x} f(x')$. We say a function $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is concave (convex) if $\operatorname{hypo} f$ ($\operatorname{epi} f$) is convex. The set of *convex normal vectors* of a set $S \subseteq \mathcal{E} \times \mathbb{R}$ at some $(x,u) \in S$ is denoted by

$$N_S^C((x,u)) := \{(\zeta,\delta) \mid (\zeta,\delta)^T((x,u) - (x',u')) \geq 0 \ \forall (x',u') \in S\}.$$

Then the *convex subdifferential* and *convex supdifferential* of a function f is denoted by

$$\begin{aligned}
\partial_C f(x) &:= \{\zeta \mid (\zeta,-1) \in N^C_{\operatorname{epi} f}((x,f(x)))\},\\
\partial^C f(x) &:= \{\zeta \mid (-\zeta,1) \in N^C_{\operatorname{hypo} f}((x,f(x)))\}.
\end{aligned}$$

We also consider the generalization given by proximal normal vectors and sub/supdifferentials of

$$\begin{aligned}
N_S^P((x,u)) &:= \{(\zeta,\delta) \mid (x,u) \in \operatorname{proj}_S((x,u) + \epsilon(\zeta,\delta)) \text{ for some } \epsilon > 0\},\\
\partial_P f(x) &:= \{\zeta \mid (\zeta,-1) \in N^P_{\operatorname{epi} f}((x,f(x)))\},\\
\partial^P f(x) &:= \{\zeta \mid (-\zeta,1) \in N^P_{\operatorname{hypo} f}((x,f(x)))\}.
\end{aligned}$$

Dual Families of Functions Most of our theory characterizing the radial transformation relies on the given function being (strictly) upper radial. Recall that Proposition 5.3.3 shows an upper semicontinuous function f is upper radial (that is, our radial duality $f^{\Gamma\Gamma} = f$ holds) if and only if all $(x, u) \in \operatorname{hypo} f$ and $(\zeta, \delta) \in N^P_{\operatorname{hypo} f}((x, u))$ satisfy

$$(\zeta, \delta)^T (x, u) \geq 0. \tag{6.10}$$

Geometrically, this corresponds to the origin lying below all of the hyperplanes induced by proximal normal vectors of the hypograph. Similarly, Proposition 5.3.5 ensures a continuously differentiable function f is strictly upper radial if all $x \in \operatorname{dom} f$ satisfy

$$(\nabla f(x), -1)^T (x, u) < 0. \tag{6.11}$$

For concave functions, being upper radial corresponds to the origin lying in the function's domain. In particular, Proposition 5.3.8 ensures an upper semicontinuous concave function f is strictly upper radial if

$$0 \in \operatorname{int} \{x \mid f(x) > 0\}. \tag{6.12}$$

Assuming strict upper radiality holds, the following families of functions are radially dual

$$f \text{ is upper semicontinuous} \iff f^\Gamma \text{ is lower semicontinuous,} \tag{6.13}$$

$$f \text{ is continuous} \iff f^\Gamma \text{ is continuous,} \tag{6.14}$$

$$f \text{ is concave} \iff f^\Gamma \text{ is convex,} \tag{6.15}$$

where these follow from Propositions 5.3.13 and 5.3.15. For differentiable functions satisfying (6.11), Proposition 5.4.5 shows

$$f \text{ is } k \text{ times differentiable} \iff f^\Gamma \text{ is } k \text{ times differentiable,} \tag{6.16}$$

$$f \text{ is analytic} \iff f^\Gamma \text{ is analytic.} \tag{6.17}$$

Relating Extreme Points, (Sub)Gradients, and Hessians We recall a few bijections relating functions and their radial transformations. For any strictly upper radial f, Lemma 5.4.1 ensures

$$\operatorname{epi} f^{\Gamma} = \Gamma(\operatorname{hypo} f). \tag{6.18}$$

Further, Lemma 5.4.4 shows for any continuous strictly upper radial function, the following pair of bijections between graphs and domains hold

$$\operatorname{graph} f^{\Gamma} = \Gamma(\operatorname{graph} f), \tag{6.19}$$

$$y \in \operatorname{dom} f^{\Gamma} \iff y/f^{\Gamma}(y) \in \operatorname{dom} f. \tag{6.20}$$

Then Propositions 5.4.9 and 5.4.10 shows that the radial point transformation relates the maximizers of strictly upper radial functions f to minimizers of f^{Γ} as well as relates their stationary points

$$\operatorname{argmin} f^{\Gamma} \times \{\inf f^{\Gamma}\} = \Gamma(\operatorname{argmax} f \times \{\sup f\}), \tag{6.21}$$

$$\{(y, f^{\Gamma}(y)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial_P f^{\Gamma}(y)\} = \Gamma\{(x, f(x)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial^P f(x)\}. \tag{6.22}$$

In particular, for any upper semicontinuous, strictly upper radial f, the convex and proximal subgradients of its upper radial transformation are given by Propositions 5.4.2 and 5.4.3 as

$$\partial_C f^{\Gamma}(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^C_{\operatorname{hypo} f}((x,u)),\ (\zeta,\delta)^T(x,u) > 0 \right\} \tag{6.23}$$

$$\partial_P f^{\Gamma}(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix} \zeta \\ \delta \end{bmatrix} \in N^P_{\operatorname{hypo} f}((x,u)),\ (\zeta,\delta)^T(x,u) > 0 \right\} \tag{6.24}$$

where $(x,u) = \Gamma(y, f^{\Gamma}(y))$. Further, if f is continuously differentiable and satisfies (6.11), Proposition 5.4.5 shows the gradient of the upper radial transforma-

tion at $y = x/f(x)$ is

$$\nabla f^{\Gamma}(y) = \frac{\nabla f(x)}{(\nabla f(x), -1)^T (x, f(x))}. \tag{6.25}$$

If in addition we suppose f is twice continuously differentiable around x, Proposition 5.4.6 shows the Hessian of the upper radial transformation is

$$\nabla^2 f^{\Gamma}(y) = \frac{f(x)}{(\nabla f(x), -1)^T (x, f(x))} \cdot J \nabla^2 f(x) J^T \tag{6.26}$$

where $J = I - \frac{\nabla f(x) x^T}{(\nabla f(x), -1)^T (x, f(x))}$.

6.3 Conditioning of the Radially Dual Problem

As we have seen, the radial dual often enjoys very favorable structural properties. In the following three subsections, we characterize the Lipschitz continuity, smoothness, and growth conditions of the radially dual problem. Historically, these properties are all of great importance to the development of first-order optimization algorithms.

6.3.1 Lipschitz Continuity of the Radially Dual Problem

We say a function f is uniformly M-Lipschitz continuous if for all $x, x' \in \mathcal{E}$,

$$|f(x) - f(x')| \leq M \|x - x'\|.$$

For any lower semicontinuous function $f \colon \mathcal{E} \to \mathbb{R}_{++} \cup \{\infty\}$, M-Lipschitz continuity is equivalent to all proximal subgradients $\zeta \in \partial_P f(x)$ having norm bounded by M [27, Theorem 1.7.3].

Lipschitz continuity plays an important role in the analysis of many first-order methods for nonsmooth optimization. Recalling that the previous works [140, 141, 60] critically rely on their radially reformulated objective being uniformly Lipschitz, here we present a general characterization of when the radial transformation of a function is uniformly Lipschitz. To take advantage of the second characterization of Lipschitz continuity above, we need to ensure f^Γ maps into $\mathbb{R}_{++} \cup \{\infty\}$. The following simple assumption is equivalent to this (by the definition of the upper radial transformation): for all $y \in \mathcal{E}$

$$\lim_{v \to 0} v \cdot f(y/v) = 0.$$

This condition is always the case when f is bounded above as will typically be the case for our primal maximization problem. Under this condition, we find that the Lipschitz continuity of f^Γ is controlled by the distance (measured in $\mathcal{E}$) from the origin to each hyperplane defined by a proximal normal vector:

$$R(f) = \inf\{\|x'\| \mid (\zeta, \delta) \in N^P_{\text{hypo } f}(x, u),\ (\zeta, \delta)^T((x', 0) - (x, u)) = 0\}.$$

The following proposition gives the exact Lipschitz constant in terms of $R(f)$.

Proposition 6.3.1. *Consider any upper semicontinuous, strictly upper radial f where all $y \in \mathcal{E}$ have $\lim_{v \to 0} v \cdot f(y/v) = 0$. Then f^Γ is $1/R(f)$-Lipschitz continuous.*

Proof. The key observation here is that for any $(x, u) \in \operatorname{hypo} f$ and $(\zeta, \delta) \in N^P_{\text{hypo } f}((x, u))$,

$$\begin{aligned}
(\zeta, \delta)^T(x, u) &= \inf\big\{\zeta^T x' \mid (\zeta, \delta)^T((x', 0) - (x, u)) = 0\big\} \\
&= \|\zeta\| \inf\big\{\|x'\| \mid (\zeta, \delta)^T((x', 0) - (x, u)) = 0\big\} \\
&\geq \|\zeta\| R(f)
\end{aligned}$$

where the first equality is trivial and the second uses that the minimum norm point in this hyperplane will be a multiple of ζ. Then the subgradient formula (6.24) ensures any $\zeta' \in \partial_P f^\Gamma(y)$ must have

$$\|\zeta'\| = \frac{\|\zeta\|}{(\zeta,\delta)^T(x,u)} \leq 1/R(f)$$

for $(x,u) = \Gamma(y, f^\Gamma(y))$ and some $(\zeta,\delta) \in N^P_{\text{hypo } f}((x,u))$. Since every radially dual subgradients is uniformly bounded, f^Γ is uniformly Lipschitz. Considering a sequence of $(\zeta,\delta) \in N^P_{\text{hypo } f}((x,u))$ approaching attainment of $R(f)$ makes this argument tight. □

The condition $(x,u)^T(\zeta,\delta) \geq R(f)\|\zeta\|$ can be viewed as a natural way to quantify how radial f is by strengthening (6.10). When f is concave, $R(f)$ can be simplified to

$$R(f) = \inf\{\|x\| \mid f(x) = 0\}, \tag{6.27}$$

which matches the Lipschitz constants used in the previous works [140, 141, 60]. This gives a natural way to measure the extent of radiality of a concave function by strengthening (6.12). From this, we see any concave maximization problem (with a known point in the interior of its domain) can be translated and transformed into a convex minimization problem that is uniformly Lipschitz continuous with constant depending on how interior the known point is to the function's domain.

6.3.2 Smoothness of the Radially Dual Problem

We say a continuously differentiable function f is uniformly L-smooth if its gradient is L-Lipschitz continuous: for all $x, x' \in \operatorname{dom} f$

$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|.$$

As an example, consider the radial dual of the continuously differentiable function $f(x) = \sqrt{1 - x^T Q x}_+$, which is upper radial for any matrix Q. This radially transforms into the similarly shaped function

$$\begin{aligned} f^\Gamma(y) &= \sup\{v > 0 \mid v\sqrt{1 - y^T Q y/v^2} \leq 1\} \\ &= \sup\{v > 0 \mid v^2 - y^T Q y \leq 1\} \\ &= \sqrt{1 + y^T Q y}_+. \end{aligned}$$

Supposing Q is positive semidefinite and nonzero, our primal is concave and differentiable on its domain but fails to have a Lipschitz gradient since $\nabla f(x)$ blows up at the boundary of its domain. However, in this case, the radially dual f^Γ is well behaved, being convex and $\lambda_{max}(Q)$-smooth.

For generic functions, we cannot hope to find smoothness out of thin air (like we do in the above example or quite generically with Lipschitz continuity in the previous section). This is due to (6.16) which establishes differentiability is preserved under the radial transformation. In line with this equivalence, we find that when f is L-smooth, f^Γ is $O(L)$-smooth, provided the domain of f is bounded. Let $D(f) = \sup\{\|x\| \mid x \in \operatorname{dom} f\}$ denote the norm of the largest point in the domain of f. Note that since we are primarily taking the radial dual of maximization problems that are bounded above and truncated below to be nonnegative optimization, $D(f)$ can be viewed as bounding the level set $\operatorname{dom} f = \{x \mid f(x) > 0\}$.

The following proposition shows the operator norm of the radial transformation's Hessian is controlled by the ratio between $D(f)$ and $R(f)$ and the norm of the primal Hessian. From this, we conclude for twice differentiable L-smooth functions, the radial dual is also $O(L)$-smooth.

Proposition 6.3.2. *Consider any upper radial f with $D(f) < \infty$ and $R(f) > 0$ and $x, y \in \mathcal{E}$ satisfying $(x, f(x)) = \Gamma(y, f^\Gamma(y))$. If f is twice continuously differentiable around x, then*

$$\|\nabla^2 f^\Gamma(y)\| \leq \left(1 + \frac{D(f)}{R(f)}\right)^3 \|\nabla^2 f(x)\|.$$

Proof. First we verify that $(\nabla f(x), -1)^T(x, f(x)) < 0$ holds for all $x \in \operatorname{dom} f$ and so the Hessian formula (6.26) can be applied: if $\nabla f(x) = 0$, $(\nabla f(x), -1)^T(x, f(x)) = -f(x) < 0$ and if $\nabla f(x) \neq 0$, $(\nabla f(x), -1)^T(x, f(x)) \leq -\|\nabla f(x)\| R(f) < 0$. Then our bound on the Hessian of f^Γ follows from the following pair of inequalities. First, we have

$$\begin{aligned}
\frac{f(x)}{(\nabla f(x), -1)^T(x, f(x))} &= 1 + \frac{\nabla f(x)^T x}{(\nabla f(x), -1)^T(x, f(x))} \\
&\leq 1 + \frac{\|\nabla f(x)\| \|x\|}{|(\nabla f(x), -1)^T(x, f(x))|} \\
&\leq 1 + \frac{\|x\|}{R(f)}.
\end{aligned}$$

Second, the matrix $J = I - \frac{\nabla f(x) x^T}{(\nabla f(x), -1)^T(x, f(x))}$ has operator norm bounded by

$$\|J\| \leq 1 + \frac{\|\nabla f(x)\| \|x\|}{|(\nabla f(x), -1)^T(x, f(x))|} \leq 1 + \frac{\|x\|}{R(f)}.$$

Then applying our bounds to each term in the Hessian formula (6.26) gives the claimed result. □

Corollary 6.3.3. *Consider any upper radial, twice continuously differentiable f with $D(f) < \infty$ and $R(f) > 0$. If f is L-smooth, then f^Γ is $\left(1 + \frac{D(f)}{R(f)}\right)^3$ L-smooth.*

Proof. For a twice continuously differentiable function, having L-Lipschitz gradient is equivalent to having Hessian bounded in operator norm by L. Noting that $R(f) > 0$ implies f is strictly upper radial by (6.11), we have a bijection between the domains of f and f^Γ from (6.20). Hence the Hessian of f^Γ is uniformly bounded by $\left(1 + \frac{D(f)}{R(f)}\right)^3 L$. □

Although this result requires smoothness of the primal objective f to be maximized, it still provides an algorithmically valuable tool due to the symmetry-breaking nature of considering functions on the extended positive reals $\overline{\mathbb{R}}_{++}$. Supposing f is bounded above, this result allows us to extend smoothness of f on a level set $\operatorname{dom} f = \{x \mid f(x) > 0\}$ to global smoothness of the dual f^Γ on $\operatorname{dom} f^\Gamma = \mathcal{E}$.

For example, consider an unconstrained $S = \mathbb{R}^d$ instance of our previous motivating example of the Poisson likelihood problem (6.9) which is not defined everywhere (only on $\{x \mid a_i^T x > 0\}$) with gradients blowing up as x approaches the boundary of this domain. However, provided the measurements $\{a_i\}$ span $\mathbb{R}^n$, this objective has bounded level sets. Consequently, for any twice continuously differentiable $r(x)$, our radial duality provides a reformulation that extends the smoothness on the level set $\{x \mid \mathcal{L}(x) - r(x) > 0\}$ to hold globally.

More broadly, (6.16) ensures being k-times differentiable is preserved by the radial transformation. We expect similar bounds can be derived showing higher-order smoothness bounds (that is, showing Lipschitz continuity of higher-order derivatives) on the primal level set $\{x \mid f(x) > 0\}$ extend to the radial dual as global higher-order smoothness bounds. Carefully conducting such analysis would likely enable the study of radial second-order methods, like radial Newton methods.

6.3.3 Growth Conditions in the Radially Dual Problem

For a lower semicontinuous function $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, we say the Łojasiewicz condition holds at a local minimum x^* if for some constants $r > 0$, $C > 0$ and exponent $\theta \in [0, 1)$, all nearby $x \in B(x^*, r)$ have

$$\mathbf{dist}(0, \partial_P f(x)) \geq C(f(x) - f(x^*))^\theta. \tag{6.28}$$

For an upper semicontinuous function f with local maximum x^*, we instead require all nearby $x \in B(x^*, r)$ have

$$\mathbf{dist}(0, \partial^P f(x)) \geq C(f(x^*) - f(x))^\theta. \tag{6.29}$$

These conditions are widespread, holding for generic subanalytic functions [104, 103] and nonsmooth subanalytic convex functions [17]. These properties are closely related to the Kurdyka-Łojasiewicz condition [88] and Hölderian growth/error bounds used by [18, 173, 150, 142], which are known to speed up the convergence of many first-order methods.

Under mild conditions, the Łojasiewicz condition is preserved by our radial transformation. Consequently, optimization algorithms based on solving the radially dual problem can enjoy the same improved convergence historically expected in the primal from such conditions.

Proposition 6.3.4. *Consider any upper semicontinuous, strictly upper radial function f with $R(f) > 0$. If f satisfies the Łojasiewicz condition* (6.29) *at some stationary point $x^* \in \operatorname{dom} f$ with exponent θ, then f^Γ at $y^* = x^*/f(x^*) \in \operatorname{dom} f^\Gamma$ satisfies the Łojasiewicz condition* (6.28) *with the same exponent θ.*

Proof. Note that (6.13) ensures f^Γ is lower semicontinuous. Let r, C, θ satisfy the Łojasiewicz condition of f at some stationary point x^* and denote the radially

dual stationary point from (6.22) as $y^* = x^*/f(x^*)$. Since f^Γ is $1/R(f)$-Lipschitz continuous, every $0 < r' < f^\Gamma(y^*)R(f)$ will have all $y \in B(y^*, r')$ map to $x = y/f^\Gamma(y)$ with

$$\begin{aligned}
\|x - x^*\| &= \left\| \frac{y}{f^\Gamma(y)} - \frac{y^*}{f^\Gamma(y^*)} \right\| \\
&\leq \left\| \frac{y - y*}{f^\Gamma(y)} \right\| + \left\| \frac{y^*}{f^\Gamma(y)} - \frac{y^*}{f^\Gamma(y^*)} \right\| \\
&= \frac{\|y - y^*\|}{f^\Gamma(y)} + \frac{\|y^*\||f^\Gamma(y) - f^\Gamma(y^*)|}{f^\Gamma(y)f^\Gamma(y^*)} \\
&\leq \frac{r'}{f^\Gamma(y)} + \frac{\|y^*\|r'}{R(f)f^\Gamma(y)f^\Gamma(y^*)} \\
&\leq \frac{r'}{f^\Gamma(y^*) - r'/R(f)} + \frac{\|y^*\|r'}{R(f)(f^\Gamma(y^*) - r'/R(f))f^\Gamma(y^*)}.
\end{aligned}$$

Therefore selecting small enough r' guarantees that all of the dual points near y^* map back to primal points $x = y/f^\Gamma(y)$ in the ball $B(x^*, r)$ where the Łojasiewicz condition holds. Further the Lipschitz continuity of the radial dual allows us to guarantee that all of these primal points have $f(x)$ bounded below by nearly $f(x^*)$ as

$$f(x) = f^{\Gamma\Gamma}(x) \geq 1/f^\Gamma(y) \geq 1/(f^\Gamma(y^*) - r'/R(f)) = (f(x^*)^{-1} + (R(f)/r')^{-1})^{-1}.$$

Combining this with the assumed upper semincontinuity of f, we have $f(x) \to f(x^*)$ as $y \to y^*$ (despite not assuming continuity of the primal function f).

Then all that remains is to show the Łojasiewicz supgradient norm lower bound from the primal extends to lower bound the norm of the radially dual subgradients. For every $y \in B(y^*, r')$, the formula (6.24) ensures every $\zeta' \in \partial_P f^\Gamma(y)$ has $\zeta' = \zeta/(\zeta, \delta)^T(x, u)$ where $(x, u) = \Gamma(y, f^\Gamma(y))$ and $(\zeta, \delta) \in N^P_{\text{hypo } f}((x, u))$. First, suppose $\delta \neq 0$. Then $u = f(x)$ and $-\zeta/\delta \in \partial^P f(x)$ is a primal supgradient. Consequently, we can bound the size of our radially dual

subgradient as

$$\begin{aligned}\|\zeta'\| &= \frac{\|\zeta/\delta\|}{(\zeta/\delta, 1)^T(x, f(x))} \\ &\geq \frac{\|\zeta/\delta\|}{\|\zeta/\delta\|\|x\| + f(x)} \\ &\geq \frac{C(f(x^*) - f(x))^\theta}{C(f(x^*) - f(x))^\theta\|x\| + f(x)} \\ &= \frac{Cf^\theta(x)f^\theta(x^*)}{C(f(x^*) - f(x))^\theta\|x\| + f(x)}\left(f^\Gamma(y) - f^\Gamma(y^*)\right)^\theta\end{aligned}$$

where the final inequality uses that $f(x) \geq 1/f^\Gamma(y)$ and $f(x^*) = 1/f^\Gamma(y^*)$. Recalling that as $y \to y^*$, the related primal point $x = y/f^\Gamma(y) \to x^*$ and $f(x) \to f(x^*)$, the coefficient above must converge to a positive constant

$$\frac{Cf^\theta(x)f^\theta(x^*)}{C(f(x^*) - f(x))^\theta\|x\| + f(x)} \to \frac{Cf^{2\theta}(x^*)}{C0^\theta\|x^*\| + f(x^*)}.$$

The boundary case of horizontal normal vectors with $\delta = 0$ follows from the same argument above by passing to a sequence of points $(x_i, f(x_i)) \to (x, f(x))$ and proximal normal vectors $(\zeta_i, \delta_i) \in N^P_{\text{hypo } f}((x_i, f(x_i)))$ with $(\zeta_i, \delta_i) \to (\zeta, \delta)$ and $\delta_i \neq 0$. The existence of such a sequence is guaranteed by the Horizontal Approximation Theorem [27, Page 67]. □

The case of $\theta = 0$ above is an important special case known as sharpness. If this condition holds globally, (6.28) and (6.29) correspond to the following global error bounds holding for all $x \in \mathcal{E}$

$$f(x) \geq f(x^*) + C\|x - x^*\| \tag{6.30}$$

and

$$f(x) \leq (f(x^*) - C\|x - x^*\|)_+ \tag{6.31}$$

respectively. This condition has a long history in nonsmooth optimization (see Burke and Ferris [23] as a classic reference establishing the prevalence of sharp

minima). The two global sharp error bounds (6.31) and (6.30) are dually related as follows.

Proposition 6.3.5. *For any upper semicontinuous, strictly upper radial f satisfying* (6.31) *at $x^* \in \mathcal{E}$ with constant C, then f^Γ satisfies* (6.30) *at $y^* = x^*/f(x^*)$ with constant*

$$\frac{C}{C\|x^*\| + f(x^*)}.$$

Proof. Denote the assumed upper bound on f from sharpness as $h(x) := f(x^*) - C\|x - x^*\|$. Then h_+ must be strictly upper radial due to (6.12) since h is concave with $h(0) > 0$ as

$$h(0) = 2h(x^*/2) - h(x^*) \geq 2f(x^*/2) - f(x^*) > 0$$

where the first equality uses that h is linear on the segment $[0, x^*]$, the inequality uses that $h(x^*/2) \geq f(x^*/2)$, and the strict inequality uses that f is strictly upper radial. The upper radial transformation h_+^Γ is lower bounded by our claimed sharpness lower bound for any $y \in \mathcal{E}$

$$h_+^\Gamma(y) \geq \frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|} = f^\Gamma(y^*) + \frac{C\|y - y^*\|}{f(x^*) + C\|x^*\|}$$

since $h_+^p(y, v)$ at $v = \frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|}$ has

$$\begin{aligned}
h_+^p(y, v) &= \left(\frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|}\right) f(x^*) - C\left\|y - \left(\frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|}\right) x^*\right\| \\
&\leq 1 + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|} f(x^*) - C\|y - x^*/f(x^*)\| + \frac{C^2\|y - x^*/f(x^*)\|\|x^*\|}{f(x^*) + C\|x^*\|} \\
&= 1 + C\|y - x^*/f(x^*)\|\left(\frac{f(x^*)}{f(x^*) + C\|x^*\|} - 1 + \frac{C\|x^*\|}{f(x^*) + C\|x^*\|}\right) \\
&= 1
\end{aligned}$$

where the single inequality above uses the reverse triangle inequality. Using Lemma 5.4.7, $f \leq h_+$ implies $f^\Gamma \geq h_+^\Gamma$, competing our proof. □

6.4 Radial Algorithms for Concave Maximization

Now we turn our attention to understanding the primal convergence guarantees that follow from algorithms minimizing the radial dual. In this section, we consider concave maximization problems where being strictly upper radial and having $R(f) > 0$ hold without loss of generality via a simple translation. Then the following section tackles nonconcave maximization problems where more care must be taken to ensure our duality holds.

We first remark on the natural measure of optimality in the primal that arise from considering the radial dual. Recall the set of fixed points of Γ are exactly the horizontal line at height one $\{(y, 1) \mid x \in \mathcal{E}\} = \Gamma\{(x, 1) \mid x \in \mathcal{E}\}$. Consequently, a natural way to relate nearly optimal solutions between then primal and radial dual comes from considering when $\sup f = \inf f^\Gamma = 1$. In this case, finding a dual point with accuracy

$$f^\Gamma(y_k) - \inf f^\Gamma \leq \epsilon$$

is equivalent to the relative accuracy primal guarantee of

$$\frac{\sup f - f(x_k)}{f(x_k)} \leq \epsilon$$

using that $1/f^\Gamma(y_k) \leq f^{\Gamma\Gamma}(x_k) = f(x_k)$ for $x_k = y_k/f^\Gamma(y_k)$ on any upper radial f. Following from this, we state all of our radial algorithm convergence guarantees in relative terms.

Secondly, we remark on the meaning of finding a radially dual solution minimized all the way to zero objective value $f^\Gamma(y) = 0$. In this case, y certifies that the primal maximization is unbounded as the ray $(y, 1)/v \in \operatorname{epi} f$ for all $v > 0$. Note the converse of this is not true: for example, the strictly radial function $f(x) = \sqrt{x+1}_+$ is unbounded above, but has $f^\Gamma(y) > 0$ everywhere.

6.4.1 Radial Subgradient Method

We begin by considering the radial subgradient method previously defined in Algorithm 6. This method simply takes the radial dual, applies the classic subgradient method to the resulting minimization problem, and then takes the radial dual again to return a primal solution. Importantly this method is projection-free since any primal constraint set S appears in the radial dual objective through its gauge γ_S. This method is very similar to those considered in [140, 60] which also apply a subgradient method to a radial reformulation. However, those methods include additional steps periodically rescaling their radial objective. Our algorithm omits such steps while matching the improved convergence guarantees of [60].

The standard subgradient method analysis shows the radial subgradient iterates y_k converge in terms of radial dual optimality at a rate controlled by the radially dual Lipschitz constant. Recall that translating a point in the interior of $\operatorname{hypo} f$ to the origin ensures $R(f) > 0$ and so the radial dual is Lipschitz continuous by Proposition 6.3.1). Consequently, no structure needs to be assumed beyond concavity to analyze the radial subgradient method.

Theorem 6.4.1. *Consider any upper semicontinuous, concave f with $R(f) > 0$ and $p^* = \sup f \in \mathbb{R}_{++}$ attained on some nonempty set $X^* \subseteq \mathcal{E}$. Then the radial subgradient method (Algorithm 6) with stepsizes α_k has primal solutions $x_k = y_k/f^\Gamma(y_k)$ satisfy*

$$\min_{k<T}\left\{\frac{p^* - f(x_k)}{f(x_k)}\right\} \leq \frac{\mathbf{dist}(p^* y_0, X^*)^2 + \sum_{k=0}^{T-1}(p^*\alpha_k/R(f))^2}{2\sum_{k=0}^{T-1} p^*\alpha_k}.$$

Selecting $x_0 = 0$ and $\alpha_k = \epsilon f^\Gamma(y_k)/\|\zeta_k'\|^2$ for any $\epsilon > 0$ ensures

$$T \geq \frac{\mathbf{dist}(x_0, X^*)^2}{R(f)^2\epsilon^2} \quad \implies \quad \frac{1}{T}\sum_{k=0}^{T-1}\frac{p^* - f(x_k)}{p^*} \leq \epsilon.$$

Proof. Having $R(f) > 0$ ensures f is strictly upper radial by (6.12). Then f^Γ is convex by (6.15) and has minimum value $d^* = 1/p^*$ attained on $Y^* := X^*/p^*$ by (6.21). The classic convex convergence analysis of subgradient methods follows from the fact that: for any $y^* \in Y^*$,

$$\begin{aligned}\|y_{k+1} - y^*\|^2 &= \|y_k - y^*\|^2 - 2\alpha_k \zeta_k'^T (y_k - y^*) + \alpha_k^2 \|\zeta_k'\|^2 \\ &\leq \|y_k - y^*\|^2 - 2\alpha_k (f^\Gamma(y_k) - d^*) + \alpha_k^2 \|\zeta_k'\|^2\end{aligned}$$

and so inductively,

$$\sum_{k=0}^{T-1} \alpha_k (f^\Gamma(y_k) - d^*) \leq \frac{\|y_0 - y^*\|^2 + \sum_{k=0}^{T-1} \alpha_k^2 \|\zeta_k'\|^2}{2}. \tag{6.32}$$

Noting $(x_k, u_k) = \Gamma(y_k, f^\Gamma(y_k))$, the primal iterates have $f(x_k) \geq 1/f^\Gamma(y_k)$. Then multiplying through by $(1/d^*)^2$, which equals $(p^*)^2$, we arrive at the relative primal convergence rate of

$$\begin{aligned}\sum_{k=0}^{T-1} \frac{\alpha_k}{d^*} \left(\frac{p^* - f(x_k)}{f(x_k)} \right) &= \sum_{k=0}^{T-1} \frac{\alpha_k}{(d^*)^2} \left(\frac{1}{f(x_k)} - \frac{1}{p^*} \right) \\ &\leq \frac{\|y_0/d^* - y^*/d^*\|^2 + \sum_{k=0}^{T-1} (\alpha_k/d^*)^2 \|\zeta_k'\|^2}{2}.\end{aligned}$$

Since f^Γ is $1/R(f)$-Lipschitz (by Proposition 6.3.1), every radially dual subgradient is uniformly bounded by $\|\zeta_k'\| \leq 1/R(f)$. Then selecting $y^* = \text{proj}_{Y^*}(y_0)$ gives our claimed primal convergence rate. Observe that setting $x_0 = 0$ sets $y_0 = x_0/f(x_0) = 0$ as well. Then plugging $\alpha_k = \epsilon f^\Gamma(y_k)/\|\zeta_k'\|^2$ into (6.32) yields

$$\begin{aligned}\frac{\mathbf{dist}(x_0, X^*)^2}{2} = \frac{\mathbf{dist}(y_0/d^*, X^*)^2}{2} &\geq \sum_{k=0}^{T-1} \frac{\alpha_k}{d^*} \left(\frac{f^\Gamma(y_k) - d^*}{d^*} - \frac{1}{2} \left(\frac{\alpha_k}{d^*} \right) \|\zeta_k'\|^2 \right) \\ &\geq \sum_{k=0}^{T-1} \epsilon \left(\frac{f^\Gamma(y_k)}{d^* \|\zeta_k'\|} \right)^2 \left(\frac{p^* - f(x_k)}{p^*} - \frac{\epsilon}{2} \right) \\ &\geq \sum_{k=0}^{T-1} \epsilon R(f)^2 \left(\frac{p^* - f(x_k)}{p^*} - \frac{\epsilon}{2} \right).\end{aligned}$$

Rearranging this completes our proof as the primal convergence guarantee becomes

$$\frac{1}{T}\sum_{k=0}^{T-1}\frac{p^*-f(x_k)}{p^*} \leq \frac{\mathbf{dist}(x_0,X^*)^2}{2R(f)^2\epsilon T}+\frac{\epsilon}{2}. \qquad \square$$

Recall for concave f the formula for $R(f)$ can be simplified to $\inf\{\|x\| \mid f(x)=0\}$, which quantifies how interior the origin is to the set $\{x \mid f(x)>0\}$. In this light, the constants in this rate agree with those in the guarantees of [60], up to small constants.

The classic convergence rates of the subgradient method improve in the presence of growth conditions like (6.28) or (6.30). For example growth with exponent $\theta = 1/2$ corresponds to the case of quadratic growth (generalizing strong convexity) and leads to faster $O(1/\epsilon)$ convergence, see [89] as a simple example. When $\theta = 0$, sharp growth enables the classic subgradient method to converge linearly, as shown by Polyak [135, 136] more than 50 years ago. Recalling that these quantities are preserved from primal to radial dual (Propositions 6.3.4 and 6.3.5), we find the same improvements to hold for our radial subgradient method. The following two theorems establish this speed up when $\theta = 0$ and $\theta > 0$, using the radially dual Polyak stepsize $\alpha_k = (f^\Gamma(y_k)-d^*)/\|\zeta_k'\|^2$.

Theorem 6.4.2. *Consider any upper semicontinuous, concave f with $R(f) > 0$ and $p^* = \sup f \in \mathbb{R}_{++}$ attained at $x^* \in \mathcal{E}$. Fix $x_0 = 0$ and $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta_k'\|^2$. If f satisfies the sharp growth condition* (6.31)*, then the radial subgradient method (Algorithm 6) has $x_k = y_k/f^\Gamma(y_k)$ satisfy*

$$T \geq 4\left(\frac{p^*+C\mathbf{dist}(x_0,X^*)}{CR(f)}\right)^2 \log_2\left(\frac{p^*-f(x_0)}{f(x_0)\epsilon}\right) \implies \min_{k<T}\left\{\frac{p^*-f(x_k)}{f(x_k)}\right\} \leq \epsilon.$$

Proof. Plugging the stepsize choice $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta_k'\|^2$ into (6.32) implies

$$\sum_{k=0}^{T-1} \frac{(f^\Gamma(y_k) - d^*)^2}{2} \leq \frac{\|y_0 - y^*\|^2}{2R(f)^2} \tag{6.33}$$

where $y^* = x^*/p^*$ and Proposition 6.3.1 is used to bound $\|\zeta_k'\| \leq 1/R(f)$. Then the radially dual sharpness bound from Proposition 6.3.5 guarantees $\|y_0 - y^*\| \leq \frac{p^* + C\|x^*\|}{C}(f^\Gamma(y_0) - d^*)$. Hence

$$\frac{1}{T}\sum_{k=0}^{T-1}(f^\Gamma(y_k) - d^*)^2 \leq \frac{(p^* + C\|x^*\|)^2(f^\Gamma(y_0) - d^*)^2}{C^2 R(f)^2 T}.$$

Therefore some $k \leq 4\Big(\frac{p^* + C\|x^*\|}{CR(f)}\Big)^2$ has halved the dual objective gap, $f^\Gamma(y_k) - d^* \leq (f^\Gamma(y_0) - d^*)/2$. Repeatedly applying this, we conclude that for any $\epsilon' > 0$,

$$T \geq 4\left(\frac{p^* + C\mathbf{dist}(x_0, X^*)}{CR(f)}\right)^2 \log_2\left(\frac{f^\Gamma(y_0) - d^*}{\epsilon'}\right) \implies \min_{k<T}\big\{f^\Gamma(y_k) - d^*\big\} \leq \epsilon'.$$

Considering $\epsilon' = \epsilon/p^*$ gives the claimed linear convergence rate. □

This generalizes the linear convergence results shown by [140] for linear programming. To the best of our knowledge, this is the first first-order method linear convergence guarantee for generic non-Lipschitz, sharp convex optimization. holds with $\theta > 0$ around a maximizer x^*, the radial subgradient method enjoys improved convergence once it enters this local region. This result mirrors the tradition subgradient method's speed up, although the existing results all additionally require Lipschitz continuity.

Theorem 6.4.3. *Consider any upper semicontinuous, concave f with $R(f) > 0$ and $p^* = \sup f \in \mathbb{R}_{++}$ attained at $x^* \in \mathcal{E}$. Fix $x_0 = 0$ and $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta_k'\|^2$. If f satisfies the Łojasiewicz condition* (6.29) *with exponent $\theta > 0$, then the radial subgradient method (Algorithm 6) has $x_k = y_k/f^\Gamma(y_k)$ satisfy*

$$T \geq O\big(1/\epsilon^{2\theta}\big) \quad \implies \quad \min_{k<T}\left\{\frac{p^* - f(x_k)}{f(x_k)}\right\} \leq \epsilon.$$

Proof. By Proposition 6.3.4, the Łojasiewicz condition (6.28) holds at the dual minimizer $y^* = x^*/p^*$ for some constants r', C' with the same exponent θ. Integrating this condition (as done in [18, Theorem 5]) ensures every $y \in B(y^*, r')$ has the following local error bound

$$f^\Gamma(y) - d^* \geq (C'(1-\theta)\|y - y^*\|)^{1/(1-\theta)}. \tag{6.34}$$

The subgradient method must have some y_{k_0} in the ball $B(y^*, r')$ with

$$k_0 \leq \left(\frac{\|y_0 - y^*\|}{(C'(1-\theta)r')^{1/(1-\theta)} R(f)} \right)^2$$

since (6.33) ensures the average iterate has objective gap squared at most $(C'(1-\theta)r')^{2/(1-\theta)}$. Notice that the Polyak stepsize ensures the distance from the iterates y_k to y^* is nonincreasing as

$$\begin{aligned}
\|y_{k+1} - y^*\|^2 &= \|y_k - y^*\|^2 - 2\alpha_k \zeta_k'^T (y_k - y^*) + \alpha_k^2 \|\zeta_k'\|^2 \\
&\leq \|y_k - y^*\|^2 - 2\alpha_k (f^\Gamma(y_k) - d^*) + \alpha_k^2 \|\zeta_k'\|^2 \\
&\leq \|y_k - y^*\|^2 - \frac{(f^\Gamma(y_k) - d^*)^2}{\|\zeta_k'\|^2} \leq \|y_k - y^*\|^2.
\end{aligned}$$

Hence all $k \geq k_0$ have $y_k \in B(y^*, r')$ as well. Then our claimed convergence rate follows by bounding the number of iterations required to ensure the objective gap halves $f^\Gamma(y_{k_0+k}) - d^* \leq (f^\Gamma(y_{k_0}) - d^*)/2$. Applying the local error bound (6.34) to (6.33) initialized at y_{k_0} implies

$$\frac{1}{T} \sum_{k=0}^{T-1} (f^\Gamma(y_{k_0+k}) - d^*)^2 \leq \frac{(C'(1-\theta))^2 (f^\Gamma(y_{k_0}) - d^*)^{2(1-\theta)}}{C^2 R(f)^2 T}.$$

Therefore some $k \leq 4\left(\frac{C'(1-\theta)}{R(f)}\right)^2 / (f^\Gamma(y_{k_0}) - d^*)^{2\theta}$ iterations after k_0, the radially dual objective gap must have halved. Repeatedly applying this gives the following geometric sum limiting the number of iterations required to reach any $\epsilon' > 0$ level of radial dual accuracy

$$T \geq k_0 + \sum_{i=1}^{\infty} 4\left(\frac{C'(1-\theta)}{R(f)}\right)^2 \frac{1}{(2^i \epsilon')^{2\theta}} \implies \min_{k<T} \left\{ f^\Gamma(y_k) - d^* \right\} \leq \epsilon'.$$

Selecting $\epsilon' = \epsilon/p^*$ gives the claimed result as

$$T \geq k_0 + \frac{4}{1-2^{2\theta}}\left(\frac{C'(1-\theta)}{R(f)}\right)^2\left(\frac{2p^*}{\epsilon}\right)^{2\theta} \implies \min_{k<T}\left\{\frac{p^* - f(x_k)}{f(x_k)}\right\} \leq \epsilon. \qquad \square$$

The previous pair of convergence theorems relied on using a Polyak stepsize, which requires the often impractical knowledge of d^*. This can be remedied by replacing the simple subgradient method in Algorithm 6 with a more sophisticated stepping scheme like [81] or restarting scheme like [173, 150, 142] which all attain similar convergence guarantees.

6.4.2 Radial Smoothing Method

Now we turn our attention to the radial smoothing method previously defined as Algorithm 7 in the context of smoothing the radial dual of our quadratic program. More generally, we consider primal problems maximizing a minimum of smooth functions over polyhedral constraints

$$p^* = \begin{cases} \max_x & \min\{f_j(x) \mid j = 1, \ldots, m_1\} \\ \text{s.t.} & a_i^T x \leq b_i \quad \text{for } i = 1, \ldots, m_2. \end{cases} \tag{6.35}$$

where each f_j is twice continuously differentiable and concave with $R(f_j) \geq R > 0$ and $D(f_j) \leq D < \infty$ and each $b_i > 0$. Note that having $R(f_j) > 0$ and $b_i > 0$ can be attained without loss of generality by translating a strictly feasible point in the domain of each f_j to the origin. Further, assuming $D(f_j) < \infty$ implies each f_j has bounded level sets and so each f_j is L-smooth on the level set $\{x \mid f_j(x) > 0\}$ for some $\sup\{\|\nabla^2 f_j(x)\| \mid f(x) > 0\} \leq L < \infty$. This objective is strictly upper radial and its radial dual is

$$d^* = \min_{y \in \mathcal{E}} \max\left\{f_j^\Gamma(x), (a_i/b_i)^T y \mid j \in \{1, \ldots, m_1\}, i \in \{1, \ldots, m_2\}\right\}. \tag{6.36}$$

Then we consider the smoothing of this objective for any $\eta > 0$ given by

$$g_\eta(y) = \eta \log\left(\sum_{j=1}^{m_1} \exp\left(\frac{f_j^\Gamma(y)}{\eta}\right) + \sum_{i=1}^{m_2} \exp\left(\frac{a_i^T y}{b_i \eta}\right)\right).$$

Our radial smoothing method (Algorithm 7) proceeds by minimizing this smoothing with Nesterov's accelerated method to produce a radially dual solution with accuracy $O(\eta)$. Nearly any other fast iterative method could be employed here instead, which could then avoid needing knowledge of problem constants. Converting this radial dual guarantee back to the primal problem gives the following primal convergence theorem.

Theorem 6.4.4. *Consider any problem of the form* (6.35)*. Fixing* $L_\eta = (1+D/R)^3 L + \frac{\max\{1/R^2, \|a_i/b_i\|\}}{\eta}$ *and* $x_0 = 0$*, the radial smoothing method (Algorithm 7) has* $x_k = y_k / \max\{f_j^\Gamma(x), (a_i/b_i)^T y\}$ *feasible with*

$$\frac{p^* - \min\{f_j(x_k)\}}{\min\{f_j(x_k)\}} \leq \frac{2L_\eta(1 + \eta p^* \log(m_1+m_2))^2 D^2}{p^*(k+1)^2} + \eta p^* \log(m_1+m_2).$$

In particular, setting $\eta = \epsilon / 2\log(m_1+m_2)$*, this ensures the following* $O(1/\epsilon)$ *convergence rate*

$$\begin{aligned} k+1 &\geq 2(1+p^*\epsilon/2)D\sqrt{\frac{(1+D/R)^3 L}{p^*\epsilon} + \frac{2\max\{1/R^2, \|a_i/b_i\|^2\}\log(m_1+m_2)}{p^*\epsilon^2}} \\ \implies \quad & \frac{p^* - \min\{f_j(x_k)\}}{\min\{f_j(x_k)\}} \leq p^*\epsilon. \end{aligned}$$

Proof. Observe that all of the m_1+m_2 functions defining g_η are convex (by (6.15)), $\max\{1/R, \|a_i/b_i\|\}$-Lipschitz continuous (by Proposition 6.3.1) and $(1+D/R)^3 L$-smooth (by Corollary 6.3.3). Then [15, Proposition 4.1] ensures g_η is convex, is $(1+D/R)^3 L + \frac{\max\{1/R^2, \|a_i/b_i\|\}}{\eta}$-smooth, and closely follows the radially dual objective with every $y \in \mathcal{E}$ satisfying

$$0 \leq g_\eta(y) - \max\left\{f_j^\Gamma(y), (a_i/b_i)^T y\right\} \leq \eta \log(m_1+m_2). \tag{6.37}$$

Note that for any $s > 0$, the corresponding primal objective super-level set is bounded by

$$\sup\{\|x\| \mid f_j(x) \geq s,\ a_i^T x \leq b_i\} \leq D.$$

Then the bijection $\operatorname{epi} f^\Gamma = \Gamma(\operatorname{hypo} f)$ from (6.18) bounds every sub-level set of the dual with

$$\sup\{\|y\| \mid f_j^\Gamma(y) \leq 1/s,\ (a_i/b_i)^T y \leq 1/s\} \leq D/s.$$

In particular considering $s = p^* = 1/d^*$ shows every radial dual minimizer has norm bounded by d^*D. Then the upper bound from (6.37) ensures the $d^* + \eta \log(m_1 + m_2)$ sub-level set of g_η is nonempty and the lower bound from (6.37) allows us to bound this level set by

$$\sup\{\|y\| \mid g_\eta(y) \leq d^* + \eta \log(m_1 + m_2)\} \leq (d^* + \eta \log(m_1 + m_2))D.$$

Therefore the distance from $y_0 = 0$ to a minimizer of g_η is at most $(d^* + \eta \log(m_1 + m_2))D$.

Since g_η is smooth and has a minimizer, applying the standard accelerated method convergence guarantee [125] guarantees the iterates of our radial smoothing method have

$$g_\eta(y_k) - \inf g_\eta \leq \frac{2L_\eta(d^* + \eta \log(m_1 + m_2))^2 D^2}{(k+1)^2}.$$

Converting this guarantee to be in terms of our radially dual objective, (6.37) ensures

$$\max\left\{f_j^\Gamma(y_k), (a_i/b_i)^T y_k\right\} - d^* \leq \frac{2L_\eta(d^* + \eta \log(m_1 + m_2))^2 D^2}{(k+1)^2} + \eta \log(m_1 + m_2).$$

Finally, stating this to be in terms of the primal solution $x_k = y_k / \max\{f_j^\Gamma(x), (a_i/b_i)^T y\}$ yields

$$\frac{p^* - \min\{f_j(x_k)\}}{\min\{f_j(x_k)\}} \leq \frac{2L_\eta(1 + \eta p^* \log(m_1 + m_2))^2 D^2}{p^*(k+1)^2} + \eta p^* \log(m_1 + m_2). \quad \square$$

Renegar [141] uses the same general technique to give accelerated convergence guarantees for solving the broad family of hyperbolic programming problems (which includes semidefinite programming) where the radial dual also admits a natural smoothing. The restarting schemes of [150] and [142] both explicitly consider restarting smoothing methods to attain improved convergence when growth conditions like the Łojasiewicz condition (6.28) hold. Due to Proposition 6.3.4, applying these more sophisticated methods to solve the radially dual problem will give rise to radial algorithms that enjoy the same improved convergence. The analysis of such a method should follow similarly to that of Theorem 6.4.3.

6.4.3 Radial Accelerated Method

Motivated by our example transforming the Poisson likelihood problem (6.9), algorithms can be designed to take advantage of the radial transformation extending smoothness on a level set to hold globally. Consider maximizing any twice differentiable concave function $f\colon \mathcal{E} \to \mathbb{R} \cup \{-\infty\}$ with bounded level sets. Then, without loss of generality, we have $0 \in \operatorname{int}\{x \mid f(x) > 0\}$ and so f_+ is strictly upper radial. Letting $L = \sup\{\|\nabla^2 f(x)\| \mid f(x) > 0\}$, Corollary 6.3.3 ensures f_+^Γ is $(1 + D(f)/R(f))^3 L$-smooth on all of $\mathcal{E}$. Hence f_+^Γ can be minimized directly using Nesterov's accelerated method, giving the following *radial accelerated method* defined by Algorithm 8.

Algorithm 8 The Radial Accelerated Method

Require: $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, $x_0 \in \operatorname{dom} f$, $L > 0$, $T \geq 0$

1: $(y_0, v_0) = \Gamma(x_0, f(x_0))$ and $\tilde{y}_0 = y_0$ *Transform into the radial dual*
2: **for** $k = 0 \dots T - 1$ **do**
3: $\quad \tilde{y}_{k+1} = y_k - \nabla f^\Gamma(y)/(1 + D(f)/R(f))^3 L$ *Run the accelerated method*
4: $\quad y_{k+1} = \tilde{y}_{k+1} + \frac{k-1}{k+2}(\tilde{y}_{k+1} - \tilde{y}_k)$
5: **end for**
6: $(x_T, u_T) = \Gamma(y_T, f^\Gamma(y_T))$ *Transform back to the primal*
7: **return** x_T

This radial accelerated algorithm inherits the primal accelerated method's $O(\sqrt{L\mathbf{dist}(x_0, X^*)^2/\epsilon})$ rate, only requiring L-smoothness on the level set $\{x \mid f(x) > 0\}$ as follows.

Theorem 6.4.5. *Consider any twice differentiable, concave f with $R(f) > 0$, $D(f) < \infty$, $L = \sup\{\|\nabla^2 f(x)\| \mid f(x) > 0\}$, and $p^* = \sup f \in \mathbb{R}_{++}$ attained on some set $X^* \subseteq \mathcal{E}$. Fixing $x_0 = 0$, the radial accelerated method (Algorithm 8) has for any $\epsilon > 0$,*

$$k + 1 \geq (1 + D(f)/R(f))^{3/2} \sqrt{\frac{2L\mathbf{dist}(x_0, X^*)^2}{p^*\epsilon}} \quad \implies \quad \frac{p^* - f(x_k)}{f(x_k)} \leq \epsilon.$$

Proof. Recall the f^Γ is convex by (6.15) and is $(1 + D(f)/R(f))^3 L$-smooth by Corollary 6.3.3. Then Nesterov's classic analysis [125] ensures our radially dual iterates converge with

$$f^\Gamma(y_k) - d^* \leq \frac{2(1 + D(f)/R(f))^3 L\mathbf{dist}(y_0, Y^*)^2}{(k+1)^2}$$

where $Y^* = X^*/p^*$. Letting $(x_k, u_k) = \Gamma(y_k, v_k)$ yields primal iterates with $f(x_k) \geq 1/f^\Gamma(y_k)$. Then multiplying this bound through by $1/d^* = p^*$ produces the primal guarantee

$$\frac{p^* - f(x_k)}{f(x_k)} \leq \frac{2(1 + D(f)/R(f))^3 L\mathbf{dist}(y_0/d^*, X^*)^2}{p^*(k+1)^2}.$$

Noting that $y_0/d^* = x_0 = 0$, this gives the claimed convergence guarantee. □

A few remarks on this convergence result. The additional coefficient of $(1 + D(f)/R(f))^{3/2}$ is quite pessimistic as many of the examples we have considered have radial dual smoother than the primal, but Corollary 6.3.3 fails to capture this potential upside in its $O(L)$ bound. For particular applications, we expect much tighter bounds on the radially dual smoothness are possible. The proposed radial accelerated method unrealistically relies on knowledge of our smoothness constant upper bound $(1 + D(f)/R(f))^3 L$. However, this can be remedied by including a linesearch/backtracking as done in [14, 128].

Under growth conditions, the convergence of accelerated methods also improves. For example, applying the adaptive accelerated gradient method of [102] to solve the radially dual problem would give a radial method that speeds up in the presence of primal growth conditions by Proposition 6.3.4. The analysis of such a method should follow similarly to that of Theorem 6.4.3.

6.5 Radial Algorithms for Nonconcave Maximization

Our radial duality theory applies beyond the concave maximization problems that have been considered so far. The foundational theorem (6.1) establishes that our radial duality applies to the broader family of upper radial functions.

6.5.1 Examples of Radial Duality with Nonconvex Objectives or Constraints

Geometrically, upper radial functions all have a star-convex hypograph with respect to the origin Lemma 5.3.1, meaning that all $(y, v) \in \operatorname{hypo} f$ have $(y, v)/t \in \operatorname{hypo} f$ for all $t \geq 1$. Star-convexity has been considered throughout the optimization literature. The structure of optimizing over star-convex constraint sets has been considered as early as [153]. Efficient global optimization of star-convex objectives is possible if star-convexity holds with respect to a global optimizer (see [130, 66, 92, 67, 77]). However, in general, even linear optimization over star-convex bodies is NP-hard [26]. Our model of star-convexity w.r.t. the origin captures this NP-hard case.

Star-Convex Constraints

We say that a set $S \subseteq \mathcal{E}$ is *star-convex with respect to the origin* if every $x \in S$ has the line segment $\lambda x \in S$ for all $0 \leq \lambda \leq 1$. This is exactly the condition needed to ensure the indicator function $\iota_S(x) = \begin{cases} +\infty & \text{if } x \in S \\ 0 & \text{if } x \notin Ss \end{cases}$ is strictly upper radial[6]. Then the radial dual of such a star-convex set's indicator is given by the gauge

$$\iota_S^\Gamma(y) = \sup\{v > 0 \mid v \cdot \iota_S(y/v) \leq 1\} = \inf\{\lambda > 0 \mid y \in \lambda S\} = \gamma_S(y).$$

Importantly, the gauge $\gamma_S(y)$ is convex if and only if S is convex. As a result, algorithms utilizing the radial dual of star-convex constraints avoid needing

[6]This is essentially by definition as $v \cdot \iota_S(y/v)$ is nondecreasing in v if and only if S is star-convex w.r.t. the origin. Then its simple to check this function is upper semicontinuous and is vacuously strictly increasing on its effective domain $\operatorname{dom} \iota_S = \emptyset$, which is empty.

difficult nonconvex orthogonal projections, replacing them with evaluating a nonconvex gauge function appearing in the objective.

One important example where star-convex sets arises comes from considering chance constraints [87, 121, 176]. Given some distribution over potential constraint sets $S_\xi \subseteq \mathcal{E}$, a robust problem formulation may want to ensure that the constraint is satisfied with probability at least $\Lambda \in [0, 1]$. Then the chance-constrained feasible region is $S = \{x \mid \mathbb{P}(x \in S_\xi) \geq \Lambda\}$. If each potential constraint set is convex with $0 \in S_\xi$, then the chance-constrained set S is star-convex w.r.t. the origin.

Optimization over Compact Sets

Now we generalize our previous example from Section 6.2 where we saw that any nonconcave quadratic program with a compact polyhedral feasible region could be rescaled for our radial duality to apply. Consider maximizing any continuously differentiable function f over a compact set S that is star-convex w.r.t. the origin. Supposing $f(0) > 0$, this is equivalent to the following maximization problem of the primal form (6.2)

$$\max_{y \in \mathcal{E}} \min\{(1 + \lambda f(x))_+, \iota_S(x)\}$$

for any $\lambda > 0$. We can check when this objective is strictly upper radial (and so our duality holds) by considering whether its perspective function is strictly increasing on its domain:

$$v \cdot \min_i\{(1 + \lambda f(y/v))_+,\ \iota_S(y/v)\} = \begin{cases} (v + \lambda v f(y/v))_+ & \text{if } y/v \in S \\ 0 & \text{otherwise.} \end{cases}$$

The partial derivative of this with respect to v at any $y/v \in S \cap \operatorname{dom}(1+\lambda f)_+$ is

$$1 - \lambda(\nabla f(y/v), -1)^T(y/v, f(y/v)).$$

Noting that $(\nabla f(x), -1)^T(x, f(x))$ is continuous on the compact set $S \cap \operatorname{dom}(1+\lambda f)_+$, we can select $\lambda > 0$ small enough to always have $1 - \lambda(\nabla f(y/v), -1)^T(y/v, f(y/v)) > 0$. Doing so makes our objective strictly upper radial and hence our radial duality applies.

Nonconvex Regularization

Many optimization tasks take the additive composite form

$$\max_{y \in \mathcal{E}} f(x) - r(x)$$

where f is an upper semicontinuous, concave function with $f(0) > 0$ and $r(x)$ is an added (or rather subtracted since we are maximizing) regularization term. Many sparsity inducing regularization penalties decompose as a sum over the x's coordinates $r(x) = \sum_{i=1}^n \sigma(x_i)$ for some simple nonconvex function $\sigma : \mathbb{R} \to \mathbb{R}$. For example, ℓ_q-regularization sets $\sigma(t) = \lambda|t|^q$ for some $0 < q < 1$, bridging the gap between ℓ_0 and ℓ_1-regularization, and smoothly clipped absolute deviation (SCAD) regularization [47] sets

$$\sigma(t) = \begin{cases} \lambda|t| & \text{if } |t| \leq \lambda \\ (-|t|^2 + 2a\lambda|t| - \lambda^2)/2(a-1) & \text{if } \lambda < |t| \leq a\lambda \\ (1+a)\lambda^2/2 & \text{if } |t| > a\lambda \end{cases}$$

for some constants $a > 2$ and $\lambda > 0$. Many more regularizers are of this form, like MCP [177] and firm thresholding [56]. See [169] for a survey of numerous other important nonconvex regularization formulations and their usage in practice.

These regularizers are all continuous and have $r(y/v)$ nonincreasing in v. These two simple properties suffice to guarantee subtracting r from f will not break its upper radiality since

$$v(f(y/v) - r(y/v))_+ = \max\{vf(y/v) - vr(y/v), 0\}$$

is a sum of two upper semicontinuous, nondecreasing functions in v. As a result, our radial duality applies to the potentially nonconcave primal objective $(f(x) - r(x))_+$.

Optimization with Outliers

Many learning problems take the form of minimizing a stochastic loss function $\mathbb{E}_\xi[f(x, \xi)]$ using a finite sample approximation. Given i.i.d. samples $\xi_1, \ldots, \xi_s$, this problem can be formulated as the following maximization

$$\max_{x \in \mathcal{E}} \frac{1}{s} \sum_{i=1}^{s} -f(x, \xi_i).$$

If each $-f(\cdot, \xi_i)$ is concave, a translation will ensure every $-f(\cdot, \xi_i)$ is upper radial and our radial duality can be applied. In the presence of t outliers in the s samples $\xi_1, \ldots, \xi_s$, this finite sample approximation could be improved to only consider the loss function on the best $s - t$ samples

$$\max_{x \in \mathcal{E}} \max \left\{ \frac{1}{s-t} \sum_{i \in S} -f(x, \xi_i) \mid S \subseteq \{1...s\}, |S| = s - t \right\}.$$

Provided each $-f(\cdot, \xi_i)$ is upper radial, this whole objective will be upper radial by Corollary 5.3.11 and so our radial duality applies. The minimax formulation of [174] exactly corresponds to this problem formulation at its equilibrium. By the same corollary, our radial duality also applies to maximizing the $(s - t)$th largest element of $\{-f(x, \xi_i)\}_{i=1}^s$. Such an optimization problem captures the classic idea of least median of squares regression [151].

6.5.2 Example Nonconcave Guarantee for the Radial Subgradient Method

In this concluding section, we demonstrate the style of results possible from applying our radial duality to nonconcave maximization. In particular, we consider the nonconcave, nonsmooth primal problem of maximizing the minimum of a set of twice continuously differentiable, strictly upper radial f_j over some convex set $S \subseteq \mathcal{E}$

$$p^* = \begin{cases} \max_x & \min\{f_j(x) \mid j = 1, \dots, m\} \\ \text{s.t.} & x \in S. \end{cases} = \max_{x \in \mathcal{E}} \min\{f_j(x), \iota_S(x)\} \tag{6.38}$$

where each f_j has $R(f_j) \geq R > 0$ and bounded level sets $D(f_j) \leq D < \infty$ and the origin lies in the interior of the constraint set $B(0, R) \subseteq S$. Let $L \geq \sup\{\|\nabla^2 f_j(x)\| \mid f_j(x) > 0, x \in S\}$ bound the smoothness of each f_j on this compact level set.

This primal is strictly upper radial since each function defining the minimum is strictly upper radial and so our radial duality applies. The radial dual of this problem is

$$d^* = \min_{y \in \mathcal{E}} \max\{f_j^\Gamma(y), \gamma_S(y)\}. \tag{6.39}$$

Note each $f_j^\Gamma(y)$ is convex if and only if f_j is concave by (6.15). Hence if our primal (6.38) is nonconcave, our radial dual (6.39) will be nonconvex. Regardless, our previously proposed radial subgradient method (Algorithm 6) can still be applied and analyzed.

Recently, convergence theory for subgradient methods without convexity has been developed, following the ideas we presented in Chapter 3. Particularly, consider minimizing a nonconvex, nonsmooth function $g \colon \mathcal{E} \to \mathbb{R}$ that is

bounded below. Then [32, Theorem 3.1] ensures that provided g is uniformly M-Lipschitz and ρ-weakly convex (defined as $g + \frac{\rho}{2}\|\cdot\|^2$ being convex), the subgradient method $y_{k+1} = y_k - \epsilon\zeta_k/\|\zeta_k\|^2$ for $\zeta_k \in \partial_P g(y_k)$ will have some y_k be nearly stationary on the Moreau envelope of g. In particular, this implies some y_k will have a nearby y that is nearly stationary

$$\begin{aligned} T &\geq \frac{\rho M^2(g(y_0) - \inf g)}{\epsilon^4} \\ \implies &\quad \min_{k<T}\{\|y - y_k\|\} \leq \frac{\epsilon}{2\sqrt{\rho}} \text{ with } \mathbf{dist}(0, \partial_P g(y)) \leq \sqrt{\rho}\epsilon. \end{aligned} \tag{6.40}$$

Applying this machinery on the radial dual allows us to ensure a nearly stationary point y near a dual iterate y_k exists. Then converting this guarantee back to the primal gives the following primal convergence guarantee, preserving the above $O(1/\epsilon^4)$ rate despite not assuming the primal (6.38) is either Lipschitz or weakly convex.

Theorem 6.5.1. *Consider any problem of the form* (6.38) *with* $p^* \in \mathbb{R}_{++}$. *Fixing* $x_0 = 0$ *and* $\alpha_k = \epsilon/\|\zeta_k'\|^2$, *the radial subgradient method (Algorithm 6) has* $x_k = y_k/\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}$ *satisfy*

$$\begin{aligned} T &\geq \frac{(1 + D/R)^3 L(\min\{f_j(x_0)\} - p^*)}{R^2 \min\{f_j(x_0)\} p^* \epsilon^4} \\ &\implies \min_{k<T}\{\|x - x_k\|\} \leq \frac{p^*\epsilon}{2\sqrt{(1 + D/R)L}} \\ &\qquad \textit{with } \mathbf{dist}(0, \partial^P \min\{f_j, \iota_{a_i^T x \leq b_i}\}(x)) \leq \frac{p^*\sqrt{(1 + D/R)^3 L}\epsilon}{1 - \sqrt{(1 + D/R)^3 L}\epsilon D} \end{aligned}$$

for some nearby $x \in \mathcal{E}$ *provided* $0 < \epsilon < 1/\sqrt{(1 + D/R)^3 L}D$.

Proof. Observe that each function in the maximum defining the radial dual (6.36) is $1/R$-Lipschitz (by Proposition 6.3.1) and each f_j^Γ is $(1 + D/R)^3 L$-smooth (by Corollary 6.3.3). Then the whole radially dual objective

$\max\{f_j^\Gamma(y), \gamma_S(y)\}$ is $1/R$-Lipschitz and $(1 + D/R)^3 L$-weakly convex. Hence even though our primal is not assumed to be either Lipschitz or weakly convex, these two properties occur in the radial dual due to each f_i having $R(f_i) > 0$ and smoothness on the level set $\{x \mid f_j(x) > 0\}$ respectively. Then we can apply (6.40) implying a nearby dual solution y has

$$\begin{aligned} T &\geq \frac{(1 + D/R)^3 L(\min\{f_j^\Gamma(y_0)\} - d^*)}{R^2 \epsilon^4} \\ &\implies \min_{k<T}\{\|y - y_k\|\} \leq \frac{\epsilon}{2\sqrt{(1 + D/R)^3 L}} \\ &\qquad \text{with } \mathbf{dist}(0, \partial_P \max\{f_j^\Gamma, \gamma_S\}(y)) \leq \sqrt{(1 + D/R)^3 L}\, \epsilon. \end{aligned}$$

Relating this guarantee to the primal is done in the following two steps. First, we show the nearby radial dual solution y corresponds to a primal solution $x = y/\max\{f_j^\Gamma(y), \gamma_S(y)\}$ that is also near the primal iterates $x_k = y_k/\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}$. Then relating the dual stationarity of y to the primal stationarity of x completes our proof, showing it is a nearby, nearly stationary primal solution.

Observe that having $\|y - y_k\| \leq \epsilon/2\sqrt{(1 + D/R)^3 L}$ ensures the distance $\|x - x_k\|$ is at most

$$\begin{aligned} \|x - x_k\| &= \left\| \frac{y}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} - \frac{y_k}{\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}} \right\| \\ &\leq \frac{\|y - y_k\|}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} + \left\| \frac{y_k}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} - \frac{y_k}{\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}} \right\| \\ &= \frac{\|y - y_k\|}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} + \|x_k\| \left| \frac{\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} - 1 \right| \\ &\leq \frac{\|y - y_k\|}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} + \frac{D\|y - y_k\|/R}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} \\ &\leq \frac{1 + D/R}{d^*} \|y - y_k\| \leq \frac{\epsilon}{2d^*\sqrt{(1 + D/R)L}} \end{aligned}$$

where the first inequality uses the triangle inequality, the second uses the bounded primal level sets and the radially dual $1/R$-Lipschitz continuity, and the third uses that $d^* = 1/p^* \in \mathbb{R}_{++}$.

Lastly, let $v = \max\{f_j^\Gamma(y), \gamma_S(y)\}$, $u = 1/v$ and $\zeta' \in \partial_P \max\{f_j^\Gamma, \gamma_S\}(y)$ denote a radially dual subgradient with $\|\zeta'\| \leq \sqrt{(1 + D/R)^3 L}\, \epsilon$. Then we can bound

$$\begin{aligned}(\zeta', -1)^T(y, v) &\leq \|\zeta'\|\|y\| - v \\ &\leq \sqrt{(1 + D/R)^3 L}\epsilon\|x\|/u - 1/u \\ &\leq -(1 - \sqrt{(1 + D/R)^3 L}\epsilon D)/p^* < 0.\end{aligned}$$

Noting that $(\zeta', -1) \in N^P_{\text{epi } \min\{f_j^\Gamma, \gamma_S\}}(y, v)$, the primal then has a supgradient

$$\zeta := \frac{\zeta'}{(\zeta', -1)^T(y, v)} \in \partial^P \min\{f_j, \iota_S\}(x)$$

by (6.24) applied to the radial dual. This primal subgradient has norm at most $O(\epsilon)$ since

$$\|\zeta\| = \left\|\frac{\zeta'}{(\zeta', -1)^T(y, v)}\right\| = \frac{\|\zeta'\|}{|(\zeta', -1)^T(y, v)|} \leq \frac{p^*\sqrt{(1 + D/R)^3 L}\epsilon}{1 - \sqrt{(1 + D/R)^3 L}\epsilon D}. \qquad \square$$

CHAPTER 7

LIFTING CONVERGENCE RATES ASSUMING HÖLDER GROWTH

7.1 Introduction

We consider minimizing an unconstrained problem of our recurring form (1.1)

$$\min_{x \in \mathbb{R}^n} f(x) \tag{7.1}$$

that attains its minimum value at some x_*. Typically first-order optimization methods suppose f possesses some continuity or smoothness structure. These assumptions are broadly captured by assuming f is (L, η)-Hölder smooth, defined for any $L \geq 0$ and $0 \leq \eta \leq 1$ as

$$\|g - g'\| \leq L\|x - x'\|^\eta \text{ for all } x, x' \in \mathbb{R}^n, g \in \partial f(x), g' \in \partial f(x') \tag{7.2}$$

where $\partial f(x) = \{g \in \mathbb{R}^n \mid f(x') \geq f(x) + \langle g, x' - x \rangle \; \forall x' \in \mathbb{R}^n\}$ is the subdifferential of f at x. When $\eta = 1$, this corresponds to the common assumption of L-smoothness (i.e., having $\nabla f(x)$ be L-Lipschitz). When $\eta = 0$, this captures the standard nonsmooth optimization model of having $f(x)$ itself be L-Lipschitz.

Throughout the first-order optimization literature, improved convergence rates typically follow from assuming the given objective function satisfies a growth/error bound or Kurdyka-Łojasiewicz condition (see [104, 103, 88, 17, 18] for a sample of works developing these ideas). One common formalization of this comes from assuming (α, p)-Hölderian growth holds, defined by

$$f(x) \geq f(x_*) + \alpha\|x - x_*\|^p \text{ for all } x \in \mathbb{R}^n \,. \tag{7.3}$$

The two most important cases of this bound are Quadratic Growth given by $p = 2$ (generalizing strong convexity [117]) and Sharp Growth given by $p = 1$ (which occurs broadly in nonsmooth optimization [23]).

Typically different convergence proofs are employed for the cases of minimizing f with and without assuming the existence of a given growth lower bound. Table 7.1 summarizes the number of iterations required to guarantee ϵ-accuracy for several well-known first-order methods: the Proximal Point Method with any stepsize $\rho > 0$ defined by

$$x_{k+1} = \text{prox}_{\rho,f}(x_k) := \text{argmin}\{f(x) + \frac{1}{2\rho}\|x - x_k\|^2\}\,, \tag{7.4}$$

the Polyak Subgradient Method defined by

$$x_{k+1} = x_k - \rho_k g_k \text{ for some } g_k \in \partial f(x_k) \tag{7.5}$$

with stepsize $\rho_k = (f(x_k) - f(x_*))/\|g_k\|^2$, and Gradient Descent

$$x_{k+1} = x_k - \rho_k \nabla f(x_k) \tag{7.6}$$

with stepsize $\rho_k = \|\nabla f(x_k)\|^{(1-\eta)/\eta}/L^{1/\eta}$ as well as the more sophisticated Proximal Bundle Method considered previously in Chapter 4 of [93, 170] and a restarted variant of the Universal Gradient Method of [128].

Our Contribution. This chapter presents a pair of meta-theorems for deriving general convergence rates from rates that assume the existence of a growth lower bound. In terms of Table 7.1, we show that each convergence rate implies all of the convergence rates to its left: the quadratic growth column's rates imply all of the general setting's rates and each sharp growth rate implies that method's general and quadratic growth rate. More generally, our results show that any convergence rate assuming growth with exponent p implies rates for any growth exponent $q > p$ and for the general setting. An early version of these results [61] only considered the case of nonsmooth Lipschitz continuous optimization.

	General	Quadratic Growth	Sharp Growth
Proximal Point Method [147, 50]	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\frac{1}{\alpha}\log\left(\frac{f(x_0)-f(x_*)}{\epsilon}\right)\right)$	$O\left(\frac{f(x_0)-f(x_*)-\epsilon}{\alpha^2}\right)$
Polyak Subgradient Method [135, 136]	$O\left(\frac{1}{\epsilon^2}\right)$	$O\left(\frac{1}{\epsilon\alpha}\right)$	$O\left(\frac{1}{\alpha^2}\log\left(\frac{f(x_0)-f(x_*)}{\epsilon}\right)\right)$
Proximal Bundle Method ([86, 41] and Chapter 4)	$O\left(\frac{1}{\epsilon^3}\right)$	$O\left(\frac{1}{\epsilon\alpha^2}\right)$	-
Gradient Descent [126]	$O\left(\frac{1}{\epsilon}\right)$	$O\left(\frac{1}{\alpha}\log\left(\frac{f(x_0)-f(x_*)}{\epsilon}\right)\right)$	-
Restarted Universal Method [150, 142]	$O\left(\frac{1}{\sqrt{\epsilon}}\right)$	$O\left(\frac{1}{\sqrt{\alpha}}\log\left(\frac{f(x_0)-f(x_*)}{\epsilon}\right)\right)$	-

Table 7.1: Known convergence rates for several methods. The proximal point method makes no smoothness or continuity assumptions. The subgradient and bundle method rates assume Lipschitz continuity ($\eta = 0$) and the gradient descent and universal method rates assume Lipschitz gradient ($\eta = 1$), although these methods can be analyzed for generic η.

A natural question is whether the reverse implications hold (whether each column of Table 7.1 implies the rates to its right). If one is willing to modify the given first-order method, then the literature already provides a partial answer through restarting schemes [122, 102, 173, 150, 142]. Such schemes repeatedly run a given first-order method until some criteria is met and then restart the method at the current iterate. For example, Renegar and Grimmer [142] show how a general convergence rate can be extended to yield a convergence rate under Hölder growth. Combining this with our rate lifting theory allows any rate assuming growth with exponent p to give a convergence rate applying under any growth exponent $q < p$.

7.2 Rate Lifting Theorems

Now we formalize our model for a generic first-order method `fom`. Note that for it to be meaningful to lift a convergence rate to apply to general problems, the inputs to `fom` need to be independent of the existence of a growth bound (7.3), but may depend on the Hölder smoothness constants (L, η) or the optimal objective value $f(x_*)$. We make the following three assumptions about `fom`:

(A1) The method `fom` computes a sequence of iterates $\{x_k\}_{k=0}^{\infty}$. The next iterate x_{k+1} is determined by the first-order oracle values $\{(f(x_j), g_j)\}_{j=0}^{k+1}$ where $g_j \in \partial f(x_j)$.

(A2) The distance from any x_k to some fixed $x_* \in \operatorname{argmin} f$ is at most some constant $D > 0$.

(A3) For some $p \geq 1$, there exists a function $K : \mathbb{R}_{++}^3 \to \mathbb{R}$ such that if f is (L, η)-Hölder smooth and possesses (α, p)-Hölder growth on $B(x_*, D)$, then for any $\epsilon > 0$, `fom` finds an ϵ-minimizer x_k with

$$k \leq K(f(x_0) - f(x_*), \epsilon, \alpha).$$

On the Generality of (A1)-(A3). Note that (A1) allows the computation of x_{k+1} to depend on the function and subgradient values at x_{k+1}. Hence the proximal point method is included in our model since it is equivalent to $x_{k+1} = x_k - \rho_k g_{k+1}$, where $g_{k+1} \in \partial f(x_{k+1})$. We remark that (A2) holds for all of the previously mentioned algorithms under proper selection of their stepsize parameters. In fact, many common first-order methods are nonexpansive, giving $D = \|x_0 - x_*\|$. Lastly, note that the convergence bound $K(f(x_0) - f(x_*), \epsilon, \alpha)$ in (A3) can depend on L, η, p even though our notation does not enumerate this. If

one wants to make no continuity or smoothness assumptions about f, (L, η) can be set as $(\infty, 0)$ (the proximal point method is one such example as it converges independent of such structure).

The following pair of convergence rate lifting theorems show that these assumptions suffice to give general convergence guarantees without Hölder growth and guarantees for Hölder growth with any exponent $q > p$.

Theorem 7.2.1. *Consider any method* `fom` *satisfying (A1)-(A3) and any (L, η)-Hölder smooth function f. For any $\epsilon > 0$,* `fom` *will find $f(x_k - \gamma_k g_k) - f(x_*) \leq \epsilon$ by iteration*

$$k \leq K(f(x_0) - f(x_*), \epsilon, \epsilon/D^p)$$

$$\textit{where } \gamma_k = \begin{cases} \|g_k\|^{(1-\eta)/\eta}/L^{1/\eta} & \textit{if } \eta > 0 \\ 0 & \textit{if } \eta = 0. \end{cases}$$

Theorem 7.2.2. *Consider any method* `fom` *satisfying (A1)-(A3) and any (L, η)-Hölder smooth function f possessing (α, q)-Hölder growth with $q > p$. For any $\epsilon > 0$,* `fom` *will find $f(x_k - \gamma_k g_k) - f(x_*) \leq \epsilon$ by iteration*

$$k \leq K(f(x_0) - f(x_*), \epsilon, \alpha^{p/q}\epsilon^{1-p/q})$$

$$\textit{where } \gamma_k = \begin{cases} \|g_k\|^{(1-\eta)/\eta}/L^{1/\eta} & \textit{if } \eta > 0 \\ 0 & \textit{if } \eta = 0. \end{cases}$$

Applying these rate lifting theorems amounts to simply substituting α with ϵ/D^p or $\alpha^{p/q}\epsilon^{1-p/q}$ in any guarantee depending on (α, p)-Hölder growth. For example, the Polyak subgradient method satisfies (A2) with $D = \|x_0 - x_*\|$ and converges for any L-Lipschitz objective with quadratic growth $p = 2$ at rate

$$K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{8L^2}{\alpha\epsilon}$$

establishing (A3) (a proof of this fact is given at the end of the chapter for completeness). Then applying Theorem 7.2.1 recovers the method's classic convergence rate as

$$\implies K(f(x_0) - f(x_*), \epsilon, \epsilon/D^2) = \frac{8L^2\|x_0 - x_*\|^2}{\epsilon^2}\,.$$

For convergence rates that depend on $f(x_0) - f(x_*)$, this simple substitution falls short of recovering the method's known rates, often off by a log term. Again taking the subgradient method as an example, under sharp growth $p = 1$, convergence occurs at a rate of

$$K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{4L^2}{\alpha^2}\log_2\left(\frac{f(x_0) - f(x_*)}{\epsilon}\right).$$

Then Theorem 7.2.1 ensures the following weaker general rate

$$\implies K(f(x_0) - f(x_*), \epsilon, \epsilon/D) = \frac{4L^2\|x_0 - x_*\|^2}{\epsilon^2}\log_2\left(\frac{f(x_0) - f(x_*)}{\epsilon}\right)$$

and Theorem 7.2.2 ensures the weaker quadratic growth rate

$$\implies K(f(x_0) - f(x_*), \epsilon, \epsilon/D) = \frac{4L^2}{\alpha\epsilon}\log_2\left(\frac{f(x_0) - f(x_*)}{\epsilon}\right).$$

In the following section, we provide simple corollaries that remedy this issue.

7.2.1 Improving Rate Lifting via Restarting

The ideas from restarting schemes can also be applied to our rate lifting theory. We consider the following conceptual restarting method `restart-fom` that repeatedly halves the objective gap: Set an initial target accuracy of $\tilde{\epsilon} = 2^{N-1}\epsilon$ with $N = \lceil \log_2((f(x_0) - f(x_*))/\epsilon)\rceil$. Iteratively run `fom` until an $\tilde{\epsilon}$-optimal solution is found, satisfying

$$f(x_k - \gamma_k g_k) - f(x_*) \leq \tilde{\epsilon}$$

with $\gamma_k = \begin{cases} \|g_k\|^{(1-\eta)/\eta}/L^{1/\eta} & \text{if } \eta > 0 \\ 0 & \text{if } \eta = 0 \end{cases}$ and then restart `fom` at $x_0 \leftarrow x_k - \gamma_k g_k$ with new target accuracy $\tilde{\epsilon} \leftarrow \tilde{\epsilon}/2$.

Note for the proximal point method (7.4), Polyak subgradient method (7.5), and gradient descent (7.6), this restarting will not change the algorithm's trajectory: this follows as (i) all three of these methods have the iterates $\{y_k\}$ produced by `fom` initialized at $y_0 = x_T$ satisfy $y_k = x_{k+T}$ and (ii) the proximal point method and subgradient method both have $\gamma_T = 0$ and gradient descent has $\gamma_T = \rho_T$ equal to its stepsize. As a result, the following corollaries of our lifting theorems apply directly to these three methods without restarting.

Corollary 7.2.3. *Consider any method* `fom` *satisfying (A1)-(A3) and any* (L, η)*-Hölder smooth function* f*. For any* $\epsilon > 0$*,* `restart-fom` *will find* $f(x_k - \gamma_k g_k) - f(x_*) \leq \epsilon$ *after at most*

$$\sum_{n=0}^{N-1} K(2^{n+1}\epsilon, 2^n\epsilon, 2^n\epsilon/D^p)$$

total iterations.

Proof. Observe that `restart-fom` must have found an ϵ-minimizer after the Nth restart. Our corollary then follows from bounding the number of iterations required for each of these N restarts. Run $i \in \{1 \dots N\}$ of `fom` has initial objective gap at most $2^{N+1-i}\epsilon$, and so Theorem 7.2.1 ensures an $2^{N-i}\epsilon$-minimizer is found after at most

$$K(2^{N+1-i}\epsilon, 2^{N-i}\epsilon, 2^{N-i}\epsilon/D^p)$$

iterations. Summing this bound over all i gives the claimed result. □

Corollary 7.2.4. *Consider any method* `fom` *satisfying (A1)-(A3) and any* (L, η)*-Hölder smooth function* f *possessing* (α, q)*-Hölder growth with* $q > p$*. For any* $\epsilon > 0$*,*

`restart-fom` *will find* $f(x_k - \gamma_k g_k) - f(x_*) \leq \epsilon$ *after at most*

$$\sum_{n=0}^{N-1} K(2^{n+1}\epsilon, 2^n\epsilon, \alpha^{p/q}(2^n\epsilon)^{1-p/q})$$

total iterations.

Proof. Along the same lines as the proof of Corollary 7.2.3, this corollary follows from bounding the number of iterations required for each of `restart-fom`'s N restarts. Run $i \in \{1 \dots N\}$ of `fom` has initial objective gap at most $2^{N+1-i}\epsilon$, and so Theorem 7.2.2 ensures an $2^{N-i}\epsilon$-minimizer is found after at most

$$K(2^{N+1-i}\epsilon, 2^{N-i}\epsilon, \alpha^{p/q}(2^{N-i}\epsilon)^{1-p/q}/)$$

iterations. Summing this bound over all i gives the claimed result. □

Applying these corollaries to every entry in Table 7.1 verifies our claim that each column implies those to its left (as well as all of the omitted columns with $p \in (1,2) \cup (2,\infty)$). For example, the proximal point method has (A2) hold with $D = \|x_0 - x_*\|$ and (A3) hold with

$$K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{f(x_0) - f(x_*) - \epsilon}{\rho\alpha^2}$$

(a proof of this fact is given in the appendix for completeness). Then Corollary 7.2.3 recovers the proximal point method's general convergence rate of

$$\sum_{n=0}^{N-1} K(2^{n+1}\epsilon, 2^n\epsilon, 2^n\epsilon/D) = \sum_{n=0}^{N-1} \frac{2^{n+1}\epsilon - 2^n\epsilon}{\rho(2^n\epsilon/D)^2} = \sum_{n=0}^{N-1} \frac{D^2}{\rho 2^n\epsilon} = \frac{2\|x_0 - x_*\|^2}{\rho\epsilon}$$

and Corollary 7.2.4 recovers its linear convergence rate under quadratic growth $q = 2$ of

$$\sum_{n=0}^{N-1} K(2^{n+1}\epsilon, 2^n\epsilon, \alpha^{1/2}(2^n\epsilon)^{1/2}) = \sum_{n=0}^{N-1} \frac{2^{n+1}\epsilon - 2^n\epsilon}{\rho(2^{n/2}\alpha^{1/2}\epsilon^{1/2}))^2} = \sum_{n=0}^{N-1} \frac{1}{\rho\alpha} = \frac{N}{\rho\alpha}\,.$$

Likewise, the $O(\sqrt{L/\alpha}\log(f(x_0) - x(x_*))/\epsilon)$ rate of the universal accelerated method [128] for L-smooth optimization under quadratic growth $p = 2$ recovers Nesterov's classic accelerated convergence rate as

$$\sum_{n=0}^{N} K(2^{n+1}\epsilon, 2^n\epsilon, 2^n\epsilon/D^2) = \sum_{n=0}^{N} O\left(\sqrt{\frac{LD^2}{2^n\epsilon}}\log(2)\right) = O\left(\sqrt{\frac{LD^2}{\epsilon}}\right).$$

7.2.2 Recovering Lower Bounds on Oracle Complexity

The contrapositive of our rate lifting theorems immediately lifts complexity lower bounds from the general case to apply to the specialized case of Hölder growth. Well-known simple examples [126] give complexity lower bounds when no growth bound is assumed for first-order methods satisfying

(A4) For every $k \geq 0$, x_{k+1} lies in the span of $\{g_i\}_{i=0}^{k}$.

Optimal lower bounds under Hölder growth were claimed by Nemirovski and Nesterov [122, page 26], although no proof is given. Here we note that the simple examples from the general case suffice to give lower bounds on the possible convergence rates under growth conditions.

For M-Lipschitz nonsmooth optimization, a simple example shows any method satisfying (A4) cannot guarantee finding an ϵ-minimizer in fewer than $M^2\|x_0 - x_*\|^2/16\epsilon^2$ subgradient evaluations. Consequently, applying Theorem 7.2.1, any method satisfying (A1)-(A4) for M-Lipschitz optimization with (α, p)-Hölder growth must have its rate $K(f(x_0) - f(x_*), \epsilon, \alpha)$ bounded by

$$K(f(x_0) - f(x_*), \epsilon, \epsilon/D^p) \geq \frac{M^2\|x_0 - x^*\|^2}{16\epsilon^2}\,.$$

For example, this ensures the Polyak subgradient method's $O(1/\alpha\epsilon)$ rate under quadratic growth cannot be improved by more than small constants. The exponent on α cannot decrease since that would beat the general cases' lower bound dependence on $\|x_0 - x^*\|^2$ (as $D = \|x_0 - x^*\|$ here) and similarly the exponent of ϵ cannot be improved due to the lower bound dependence of $1/\epsilon^2$.

Likewise, for L-smooth optimization, a simple example shows no method satisfying (A4) can guarantee computing an ϵ-minimizer in fewer than $\sqrt{3L\|x_0 - x^*\|^2/32\epsilon}$ iterations. Applying Theorem 7.2.1, we conclude any method satisfying (A1)-(A4) for smooth optimization with (α, p)-Hölder growth has its rate $K(f(x_0) - f(x_*), \epsilon, \alpha)$ bounded by

$$K(f(x_0) - f(x_*), \epsilon, \epsilon/D^p) \geq \sqrt{\frac{3L\|x_0 - x^*\|^2}{32\epsilon}} \,.$$

7.3 Proofs of the Rate Lifting Theorems 7.2.1 and 7.2.2

The proofs of our two main results (Theorems 7.2.1 and 7.2.2) rely on several properties of Fenchel conjugates, which we will first review. The Fenchel conjugate of a function $f\colon \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is

$$f^*(g) := \sup_x \{\langle g, x\rangle - f(x)\} \,.$$

We say that $f \geq h$ if for all $x \in \mathbb{R}^n$, $f(x) \geq h(x)$. The conjugate reverses this partial ordering, having $f \geq h \implies h^* \geq f^*$. Applying the conjugate twice f^{**} gives the largest closed convex function majorized by f (that is, the "convex envelope" of f) and consequently, for closed convex functions, $f^{**} = f$.

The (L, q)-Hölder smoothness condition (7.2) is equivalent to having upper

bounds of the following form hold for every $x \in \mathbb{R}^n$ and $g \in \partial f(x)$

$$f(x') \leq f(x) + \langle g, x' - x \rangle + \frac{L}{\eta + 1} \|x' - x\|^{\eta+1} . \tag{7.7}$$

Taking the conjugate of this convex upper bound gives an equivalent dual condition: a function f is (L, η)-Hölder smooth if and only if its Fenchel conjugate has the following lower bound for every $g \in \mathbb{R}^n$ and $x \in \partial f^*(g)$

$$f^*(g') \geq f^*(g) + \langle x, g' - g \rangle + \begin{cases} \frac{\eta}{(\eta+1)L^{1/\eta}} \|g' - g\|^{(\eta+1)/\eta} & \text{if } \eta > 0 \\ \delta_{\|g'-g\| \leq L}(g') & \text{if } \eta = 0 \end{cases} \tag{7.8}$$

where $\delta_{\|g'-g\| \leq L}(g') = \begin{cases} 0 & \text{if } \|g' - g\| \leq L \\ \infty & \text{otherwise} \end{cases}$ is an indicator function.

7.3.1 Proof of Theorem 7.2.1

Suppose that no iteration $k \leq T$ has $x_k - \gamma_k g_k$ as an ϵ-minimizer of f. We consider the following convex auxiliary functions given by (L, η)-Hölder smoothness at each x_k and $g_k \in \partial f(x_k)$

$$h_k(x) := f(x_k) + \langle g_k, x - x_k \rangle + \frac{L}{\eta + 1} \|x - x_k\|^{\eta+1}$$

and at the minimizer x_* with zero subgradient $0 = g_* \in \partial f(x_*)$

$$h_*(x) := f(x_*) + \langle g_*, x - x_* \rangle + \frac{L}{\eta + 1} \|x - x_*\|^{\eta+1} .$$

Then we focus on the convex envelope surrounding these models

$$h(x) := (\min\{h_k(\cdot) \mid k \in \{0, \ldots, T, *\}\})^{**}(x) . \tag{7.9}$$

We consider the auxiliary minimization problem of $\min h(x)$, which shares and improves on the structure of f as described in the following three lemmas.

Lemma 7.3.1. *The objectives f and h have $f(x_i) = h(x_i)$ and $g_i \in \partial h(x_i)$ for each $i \in \{0, \ldots, T, *\}$.*

Lemma 7.3.2. *The objective h is (L, η)-Hölder smooth.*

Lemma 7.3.3. *The objective h has $(\epsilon/D^p, p)$-Hölder growth on $B(x_*, D)$.*

Before proving these three results, we show that they suffice to complete our proof. Lemma 7.3.1 and assumption (A1) together ensures that applying `fom` to either f or h produces the same sequence of iterates up to iteration T. Since f and h both minimize at x_* (as $g_* = 0 \in \partial f(x_*) \cap \partial h(x_*)$), no x_k with $k \leq T$ is an ϵ-minimizer of h. Hence applying the structural conditions from Lemmas 7.3.2 and 7.3.3 with assumption (A3) on h completes our rate lifting argument as

$$T < K(h(x_0) - h(x_*), \epsilon, \epsilon/D^p) = K(f(x_0) - f(x_*), \epsilon, \epsilon/D^p) \, .$$

Proof of Lemma 7.3.1

By definition, each g_i provides a linear lower bound on f as

$$f(x) \geq f(x_i) + \langle g_i, x - x_i \rangle \, .$$

Noting that all $k \in \{0, \ldots, T, *\}$ have $h_k \geq f$, it follows that

$$\min\{h_k(x) \mid k \in \{0, \ldots, T, *\}\} \geq f(x_i) + \langle g_i, x - x_i \rangle \, .$$

Since this linear lower bound is convex, its also a lower bound on the convex envelope h. Setting $x = x_i$, equality holds with this linear lower bound. Thus $f(x_i) = h(x_i)$ and g_i is also a subgradient of h at x_i.

Proof of Lemma 7.3.2

Here our proof relies on the dual perspective of (L, η)-Hölder smoothness given by (7.8). Since each h_k is (L, η)-Hölder smooth, each h_k^* satisfies the dual lower bounding condition (7.8). Obverse that

$$\begin{aligned} h^*(y) &= \left(\min_{k \in \{0,\dots,T,*\}} \{h_k(\cdot)\} \right)^* (y) \\ &= \sup_x \left\{ \langle y, x \rangle - \min_{k \in \{0,\dots,T,*\}} \{h_k(x)\} \right\} \\ &= \max_{k \in \{0,\dots,T,*\}} \{h_k^*(y)\} \ . \end{aligned}$$

Consequently, h^* also satisfies the dual condition (7.8) since it is the maximum of functions satisfying this lower bound. Therefore $h^{**} = h$ retains the (L, η)-Hölder smoothness of each of the models h_k and the original objective f.

Proof of Lemma 7.3.3

For any $x \in B(x_*, D)$ and $k \in \{0, \dots, T\}$, the function G_k lies above our claimed growth bound as

$$h_k(x) \geq h_k(x_k - \gamma_k g_k) \geq f(x_*) + \epsilon \geq f(x_*) + \frac{\epsilon}{D^p} \|x - x_*\|^p$$

where the first inequality uses that $x_k - \gamma_k g_k$ minimizes h_k, the second uses $h_k \geq f$ and no ϵ-minimizer has been found, and the last inequality that $x \in B(x_*, D)$. Similarly, our growth bound holds for h_* as

$$\begin{aligned} h_*(x) = h(x_*) + \frac{L}{\eta + 1} \|x - x_*\|^{\eta+1} &\geq h(x_*) + \frac{L}{(\eta+1) D^{p-(\eta+1)}} \|x - x_*\|^p \\ &\geq h(x_*) + \frac{\epsilon}{D^p} \|x - x_*\|^p \end{aligned}$$

where the last inequality uses that $\frac{LD^{\eta+1}}{\eta+1} \geq f(x_0) - f(x_*) \geq \epsilon$. Hence $\min_{k \in \{0,\dots,T,*\}} \{h_k(x)\}$ satisfies our claimed Hölder growth lower bound. Since

this lower bound is convex, the convex envelope h must also satisfy the Hölder bound (7.3).

7.3.2 Proof of Theorem 7.2.2

Suppose that no iteration $k \leq T$ has $x_k - \gamma_k g_k$ as an ϵ-minimizer of f. Consider the auxiliary objective $h(x)$ defined by (7.9). Then Lemmas 7.3.1 and 7.3.2 show h agrees with f everywhere `fom` visits and is (L, η)-Hölder smooth. From this, (A1) ensures that applying `fom` to either f or h produces the same sequence of iterates up to iteration T. Since f and h both minimize at x_*, no x_k with $k \leq T$ is an ϵ-minimizer of h. As before, h further satisfies a Hölder growth lower bound.

Lemma 7.3.4. *The objective h has $(\alpha^{p/q}\epsilon^{1-p/q}, p)$-Hölder growth on $B(x_*, D)$.*

Hence h is Hölder smooth and has Hölder growth with exponent p. Then (A3) ensures

$$T < K(h(x_0) - h(x_*), \epsilon, \alpha^{p/q}\epsilon^{1-p/q}) = K(f(f(x_0) - f(x_*), \epsilon, \alpha^{p/q}\epsilon^{1-p/q}) \, .$$

Proof of Lemma 7.3.4

This proof follows the same general approach used in proving Lemma 7.3.3. For any $x \in B(x_*, D)$ and $k \in \{0, \ldots, T, *\}$, h_k lies above our claimed growth bound: Any x with $\|x - x_*\| > (\epsilon/\alpha)^{1/q}$ has

$$h_k(x) \geq f(x) \geq f(x_*) + \alpha\|x - x_*\|^q \geq f(x_*) + \alpha^{p/q}\epsilon^{1-p/q}\|x - x_*\|^p \, .$$

Any x with $\|x - x_*\| \leq (\epsilon/\alpha)^{1/q}$ has $k \in \{0, \ldots, T\}$ satisfy

$$h_k(x) \geq h_k(x_k - \gamma_k g_k) \geq f(x_*) + \epsilon \geq f(x_*) + \alpha^{p/q}\epsilon^{1-p/q}\|x - x_*\|^p$$

and $k = *$ has

$$h_*(x) \geq f(x) \geq h(x_*) + \alpha\|x - x_*\|^q \geq h(x_*) + \alpha\frac{\|x - x_*\|^p}{(\epsilon/\alpha)^{(p-q)/q}}$$
$$= h(x_*) + \alpha^{p/q}\epsilon^{1-p/q}\|x - x_*\|^p \,.$$

Hence $\min_{k\in\{0,\ldots,T,*\}}\{h_k(x)\}$ satisfies our claimed Hölder growth lower bound. Since this lower bound is convex, the convex envelope h must also satisfy the Hölder bound (7.3).

7.4 Addendum - Example Rates Under Hölder Growth

7.4.1 Proximal Point Method Convergence Guarantees

First, we verify (A2) holds for the proximal point method.

Lemma 7.4.1. *For any minimizer x_*, (A2) holds with $D = \|x_0 - x_*\|$.*

Proof. The proximal operator $\mathrm{prox}_{\rho,f}(\cdot)$ is nonexpansive [134]. Then since any minimizer x_* of f is a fixed point of $\mathrm{prox}_{\rho,f}(\cdot)$, the distance from each iterate to x_* must be nonincreasing. □

Assuming sharpness (Hölder growth with $p = 1$) facilitates a finite termination bound on the number of iterations before an exact minimizer is found (see [50] for a more general proof and discussion of this finite result). Below we compute the resulting function $K(f(x_0) - f(x_*), \alpha, \epsilon)$ that satisfies (A3).

Lemma 7.4.2. *Consider any convex f satisfying* (7.3) *with $p = 1$. Then for any $\epsilon > 0$, the proximal point method with stepsize $\rho > 0$ will find an ϵ-minimizer x_k with*

$$k \leq K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{f(x_0) - f(x_*) - \epsilon}{\rho\alpha^2}.$$

Proof. The optimality condition of the proximal subproblem is $(x_{k+1} - x_k)/\rho \in \partial f(x_{k+1})$. Hence convexity ensures

$$\|x_{k+1} - x_k\|\|x_{k+1} - x_*\|/\rho \geq \langle (x_{k+1} - x_k)/\rho, x_{k+1} - x_* \rangle \geq f(x_{k+1}) - f(x_*).$$

Then supposing that $x_{k+1} \neq x_*$, the sharp growth bound ensures

$$\|x_{k+1} - x_k\| \geq \rho(f(x_{k+1}) - f(x_*))/\|x_{k+1} - x_*\| \geq \rho\alpha.$$

Noting the proximal subproblem is ρ-strongly convex, until a minimizer is found, the objective has constant decrease at each iteration

$$f(x_{k+1}) \leq f(x_k) - \frac{\|x_{k+1} - x_k\|^2}{\rho} \leq f(x_k) - \rho\alpha^2. \qquad \square$$

7.4.2 Polyak Subgradient Method Convergence Guarantees

Much like the proximal point method, the distance from each iterate of the subgradient method to a minimizer is nonincreasing when the Polyak stepsize is used.

Lemma 7.4.3. *For any minimizer x_*, (A2) holds with $D = \|x_0 - x_*\|$.*

Proof. The convergence of the subgradient method is governed by the following

inequality

$$\begin{aligned}\|x_{k+1} - x_*\|^2 &= \|x_k - x_*\|^2 - 2\langle \rho_k g_k, x_k - x_*\rangle + \rho_k^2 \|g_k\|^2 \\ &\leq \|x_k - x_*\|^2 - 2\rho_k (f(x_k) - f(x_*)) + \rho_k^2 \|g_k\|^2 \\ &= \|x_k - x_*\|^2 - \frac{(f(x_k) - f(x_*))^2}{\|g_k\|^2} \\ &\leq \|x_k - x_*\|^2 - \frac{(f(x_k) - f(x_*))^2}{L^2}, \end{aligned} \tag{7.10}$$

where the first inequality uses the convexity of f and the second uses our assumed subgradient bound. Thus the distance from each iterate to x_* is nonincreasing. □

Below we give a simple proof that the subgradient method under L-Lipschitz continuity and quadratic growth finds an ϵ-minimizer within $O(L^2/\alpha\epsilon)$ iterations.

Lemma 7.4.4. *Consider any convex f satisfying* (7.3) *with $p = 2$. Then for any $\epsilon > 0$, the subgradient method with the Polyak stepsize will find an ϵ-minimizer x_k with*

$$k \leq K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{8L^2}{\alpha\epsilon}.$$

Proof. This convergence rate follows by noting (7.10) implies the objective gap will halve $f(x_k) - f(x_*) \leq (f(x_0) - f(x_*))/2$ after at most

$$\frac{4L^2\|x_0 - x_*\|^2}{(f(x_0) - f(x_*))^2} \leq \frac{4L^2}{\alpha(f(x_0) - f(x_*))}$$

iterations. Iterating this argument, a $2^{-n}(f(x_0) - f(x_*))$-minimizer is found with

$$k \leq \sum_{i=0}^{n-1} \frac{4L^2}{2^{-i}\alpha(f(x_0) - f(x_*))} \leq \frac{8L^2}{2^{-n}\alpha(f(x_0) - f(x_*))}\ .$$

Considering $n = \lceil \log_2((f(x_0) - f(x_*))/\epsilon) \rceil$, gives our claimed bound on the number of iterations needed to find an ϵ-minimizer. □

Rather than relying on quadratic growth, assuming sharpness (Hölder growth with $p = 1$) allows us to derive a linear convergence guarantee. Below we compute the resulting function $K(f(x_0) - f(x_*), \alpha, \epsilon)$ that satisfies (A3).

Lemma 7.4.5. *Consider any convex f satisfying* (7.3) *with $p = 1$. Then for any $\epsilon > 0$, the subgradient method with the Polyak stepsize will find an ϵ-minimizer x_k with*

$$k \leq K(f(x_0) - f(x_*), \epsilon, \alpha) = \frac{4L^2}{\alpha^2} \log_2 \left(\frac{f(x_0) - f(x_*)}{\epsilon} \right).$$

Proof. Again, this convergence rate follows by noting (7.10) implies the objective gap will halve $f(x_k) - f(x_*) \leq (f(x_0) - f(x_*))/2$ after at most

$$\frac{4L^2 \|x_0 - x_*\|^2}{(f(x_0) - f(x_*))^2} \leq \frac{4L^2}{\alpha^2}$$

iterations, which immediately establishes the claimed rate. □

CHAPTER 8

NONCONVEX-NONCONCAVE MINIMAX OPTIMIZATION GUARANTEES

8.1 Introduction

Minimax optimization has become a central tool for modern machine learning, recently receiving increasing attention in optimization and machine learning communities. The problem of interest is the following saddle point optimization problem:

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^m} L(x,y)\ , \tag{8.1}$$

where $L(x,y)$ is a differentiable function in x and y. Many important problems in modern machine learning can be formulated as a minimax optimization problem with the form (8.1), and often the objective $L(x,y)$ is neither convex in x nor concave in y. For example,

- **(GANs).** Generative adversarial networks (GANs) [59] learn the distribution of observed samples through a two-player zero-sum game. While the generative network (parameterized by G) generates samples minimizing their difference from the true data distribution, the discriminative network (parameterized by D) maximizes its ability to distinguish between these distributions. This gives rise to the minimax formulation

  $$\min_G\max_D \mathbb{E}_{s\sim p_{data}}[\log D(s)] + \mathbb{E}_{e\sim p_{latent}}[\log(1-D(G(e)))]\ ,$$

  where p_{data} is the data distribution, and p_{latent} is the latent distribution.

- **(Robust Training).** Minimax optimization has a long history in robust optimization. Recently, it has found usage with neural networks, which have shown great success in machine learning tasks but are vulnerable to adversarial attack. Robust training [109] aims to overcome such issues by solving the minimax problem

$$\min_x \mathbb{E}_{(u,v)} \left[\max_{y \in S} \ell(u + y, v, x) \right],$$

where u is a feature vector, v is its label, x is the model parameters being trained, y is an adversarial modification, and S is the set of possible corruptions.

- **(Reinforcement Learning).** In reinforcement learning, the solution to Bellman equations can be obtained by solving a primal-dual minimax formulation. Such an approach can be viewed as having a dual critic seeking a solution satisfying the Bellman equation and a primal actor seeking state-action pairs to break this satisfaction [163, 28].

The Proximal Point Method (PPM) may be the most classic first-order method for solving minimax problems. It was first studied in the seminal work by Rockafellar in [147], and many practical algorithms for minimax optimization developed later on turn out to be approximations of PPM, such as Extragradient Method (EGM) [165, 120] and Optimistic Gradient Descent Ascent [30]. The update rule of PPM with step-size η is given by the proximal operator:

$$(x_{k+1}, y_{k+1}) = \text{prox}_{,\eta}(x_k, y_k) := \arg\min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x_k\|^2 - \frac{\eta}{2}\|v - y_k\|^2. \tag{8.2}$$

For convex-concave minimax problems, PPM is guaranteed to converge to an optimal solution. However, the dynamics of PPM for nonconvex-nonconcave

minimax problems are much more complicated. For example, consider the special case of minimax optimization problem with bilinear interaction defined as

$$\min_x \max_y L(x, y) = f(x) + x^T A y - g(y). \tag{8.3}$$

Figure 8.1 presents the sample paths of PPM from different initial solutions solving a simple two-dimensional nonconvex-nonconcave minimax problem (8.3) with $f(x) = g(x) = (x-3)(x-1)(x+1)(x+3)$ and different interaction terms A. This example may be the simplest non-trivial example of a nonconvex-nonconcave minimax problem. It turns out the behaviors of PPM heavily relies on the scale of the interaction term A: when the interaction term is small, PPM converges to local stationary solutions, as the interaction term increases, PPM may fall into a limit cycle indefinitely, and eventually when the interaction term is large enough, PPM converges globally to a stationary solution. Similar behaviors also happen in other classic algorithms for nonconvex-nonconcave minimax problems, in particular, EGM, which is known as one of the most effective algorithms for minimax problems. See Figure 8.2 in Appendix 8.6.1 for their trajectories for solving this simple two-dimension example (the study of these other algorithms is beyond the scope of this chapter). In practice, it is also well-known that classic first-order methods may fail to converge to a stable solution for minimax problems, such as GANs [49].

The goal of this chapter is to understand these varied behaviors of PPM when solving nonconvex-nonconcave minimax problems. We identify that the following *saddle envelope*, originating from Attouch and Wets [11], provides key insights

$$L_\eta(x, y) := \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2. \tag{8.4}$$

This generalizes the Moreau envelope, but differs in key ways. Most outstand-

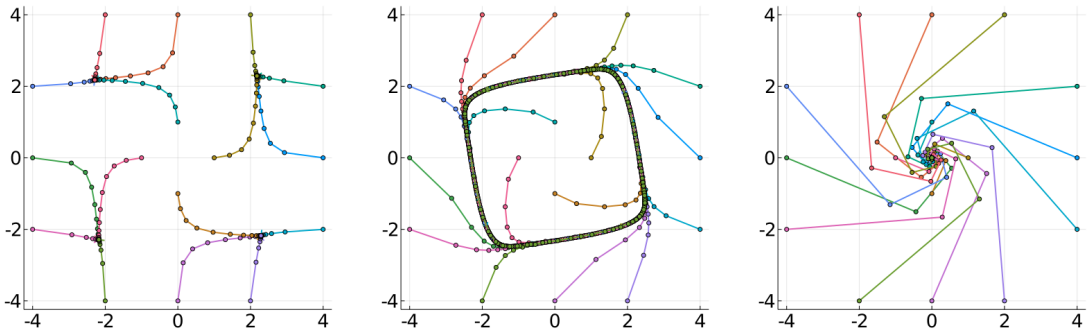


Figure 8.1: Sample paths of PPM from different initial solutions applied to (8.3) with $f(x) = (x+3)(x+1)(x-1)(x-3)$ and $g(y) = (y+3)(y+1)(y-1)(y-3)$ and different scalars A. As $A \geq 0$ increases, the solution path transitions from having four locally attractive stationary points, to a globally attractive cycle, and finally to a globally attractive stationary point.

ingly, we show that the saddle envelope not only smooths the objective but also can convexify and concavify nonconvex-nonconcave problems when $\nabla^2_{xy} L$ is sufficiently large (which can be interpreted as having a high level of the interaction between x and y). Understanding this envelope in our nonconvex-nonconcave setting turns out to be the cornerstone of explaining the above varied behaviors of PPM. Utilizing this machinery, we find that the three regions shown in the simple two-dimensional example (Figure 8.1) happen with generality for solving (8.1). Informally speaking,

1. When the interaction between x and y is dominant, PPM has global linear convergence to a stationary point of $L(x, y)$ (Figure 8.1 (c)). This argument utilizes the fact that, in this case, the closely related saddle envelope becomes convex-concave thanks to the high interaction terms, even though the original function $L(x, y)$ is nonconvex-nonconcave.

2. When the interaction between x and y is weak, properly initializing PPM yields local linear convergence to a nearby stationary point of $L(x, y)$ (Fig-

ure 8.1 (a)). The intuition is that due to the low interaction we do not lose much by ignoring the interaction and decomposing the minimax problems to a nonconvex minimization problem and a nonconcave maximization problem (where the local convergence of PPM is typical).

3. Between these interaction dominant and weak regimes, PPM may fail to converge at all and fall into cycling (Figure 8.1 (b)) or divergence (see the example in Section 8.5.1). In this scenario, we construct a "Lyapunov"-type function that characterizes how fast PPM may diverge and show that the resulting diverging bound is tight for PPM by constructing a worst-case example.

Furthermore, we believe a careful understanding of the saddle envelope of nonconvex-nonconcave functions will be broadly impactful outside its use herein analyzing the proximal point method. In Section 8.2, we develop the saddle envelope's calculus for nonconvex-nonconcave functions generalizing the convex-concave results of [9, 12]. As a byproduct of our analysis of the saddle envelope, we clearly see that the interaction term helps the convergence of PPM for minimax problems. This may not be the case for other algorithms, such as gradient descent ascent (GDA) and alternating gradient descent ascent (AGDA) (see Figure 8.2 in Appendix 8.6.1 for some examples and [64] for theoretical analysis).

We comment on the meaning of stationary points $\nabla L(z) = 0$ for nonconvex-nonconcave problems. By viewing the problem (8.1) as a simultaneous zero-sum game between a player selecting x and a player selecting y, a stationary point can be thought of as a first-order Nash Equilibrium. That is, neither player tends to deviate from their position based on their first-order information. One

can instead view the minimax problem as a sequential zero-sum game (where the minimizing player selects x and then the maximizing player exploits that choice in choosing y). Unlike the convex-concave case, the solutions between these two types of games no longer coincide and the optimal (sequential) minimax solution need not be a stationary point. In this case, a different asymmetric measure of optimality may be called for [30, 80, 49]. However such approaches are beyond the scope of this chapter as the limit points of the proximal point method are all stationary points.

In the rest of this section, we discuss the assumptions, related literature, and preliminaries that will be used later on. In Section 8.2, we develop our expanded theory for the saddle envelope. In particular, we introduce the interaction dominance condition (Definition 8.2.5) that naturally comes out as a condition for convexity-concavity of the saddle envelope. In Section 8.3, we present the global linear convergence of PPM for solving interaction dominant minimax problems. In Section 8.4, we discuss the behaviors of PPM in the interaction weak case and show that with a good initialization, PPM converges to a local stationary point. In Section 8.5, we show that PPM may diverge when our interaction dominance condition is slightly violated, showing the tightness of our global convergence theory. Further, we propose a natural "Lyapunov"-type function that applies to generic minimax problems, providing an upper bound on how quickly problems can divergence in the difficult interaction moderate setting. This bound on PPM's divergence is tight under our basic assumptions as we provide a worst-case example.

8.1.1 Assumptions and Algorithms

Basic definitions and assumptions. We say a function $M(x, y)$ is β-smooth if its gradient is uniformly β-Lipschitz

$$\|\nabla M(z) - \nabla M(z')\| \le \beta \|z - z'\|$$

or equivalently for twice differentiable functions, if $\|\nabla^2 M(z)\| \le \beta$. Further, we say a twice differentiable $M(x, y)$ is μ-strongly convex-strongly concave for some $\mu \ge 0$ if

$$\nabla^2_{xx} M(z) \succeq \mu I \ , \quad -\nabla^2_{yy} M(z) \succeq \mu I \ .$$

When $\mu = 0$, this corresponds to M being convex with respect to x and concave with respect to y.

Throughout this chapter, we are primarily interested in the weakening of this convexity condition to allow negative curvature given by ρ-weak convexity and ρ-weak concavity (recall these notions were introduced in Chapter 3): we assume that L is twice differentiable, and for any $z = (x, y) \in \mathbb{R}^n \times \mathbb{R}^m$ that

$$\nabla^2_{xx} L(z) \succeq -\rho I \ , \quad -\nabla^2_{yy} L(z) \succeq -\rho I \ . \tag{8.5}$$

Notice that the objective $L(x, y)$ is convex-concave when $\rho = 0$, and strongly convex-strongly concave when $\rho < 0$. Here our primary interest is in the regime where $\rho > 0$ is positive, quantifying how nonconvex-nonconcave the given problem instance is.

Algorithms for minimax problems. Besides PPM, Gradient Descent Ascent (GDA) is another classic algorithm for minimax problem (8.1). The update rule

is given by

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - s \begin{bmatrix} \nabla_x L(x_k, y_k) \\ -\nabla_y L(x_k, y_k) \end{bmatrix}, \tag{8.6}$$

with stepsize parameter $s > 0$. However, GDA is known to work only for strongly convex-strongly concave minimax problems, and it may diverge even for simple convex-concave problems [30, 106].

In this chapter, we study a more generalized algorithm, damped PPM, with damping parameter $\lambda \in (0, 1]$ and proximal parameter $\eta > 0$. The damped proximal point method updates by

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = (1 - \lambda) \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda \operatorname{prox}_{,\eta}(x_k, y_k) \,. \tag{8.7}$$

In particular, when $\lambda = 1$, we recover the traditional PPM (8.2). Interestingly, we find through our theory that some nonconvex-nonconcave problems only have the proximal point method converge when damping is employed (that is, $\lambda < 1$ strictly).

8.1.2 Related Literature.

Guarantees for Convex Minimax Optimization There is a long history of research into convex-concave minimax optimization. Rockafellar [147] studies PPM for solving monotone variational inequalities, and shows that, as a special case, PPM converges to the stationary point linearly when $L(x, y)$ is strongly convex-strongly concave or when $L(x, y)$ is bilinear. Later on, Tseng [165] shows that EGM converges linearly to a stationary point under similar conditions. Nemirovski [120] shows that EGM approximates PPM and presents the sublinear

rate of EGM. Recently, minimax problems have gained the attention of the machine learning community, perhaps due to the thriving of research on GANs. Daskalakis and Panageas [30] present an Optimistic Gradient Descent Ascent algorithm (OGDA) and shows that it converges linearly to the saddle-point when $L(x, y)$ is bilinear. Mokhtari et al. [112] show that OGDA is a different approximation to PPM. Lu [106] presents an ODE approach, which leads to unified conditions under which each algorithm converges, including a class of nonconvex-nonconcave problems.

There are also extensive studies on convex-concave minimax problems when the interaction term is bilinear (similar to our setting (8.1)). Some influential algorithms include Nesterov's smoothing [124], Douglas-Rachford splitting (a special case is Alternating Direction Method of Multipliers (ADMM)) [38, 44] and Primal-Dual Hybrid Gradient Method (PDHG) [25].

Guarantees for Nonconvex Minimax Optimization Recently, a number of works have been undertaken considering nonconvex-concave minimax problems. The basic technique is to first turn the minimax problem (8.1) to a minimization problem on $\Phi(x) = \max_y L(x, y)$, which is well-defined since $L(x, y)$ is concave in y, and then utilize the recent developments in nonconvex optimization [100, 101, 138, 164].

Unfortunately, the above technique cannot be extended to nonconvex-nonconcave setting, because $\Phi(x)$ is no longer tractable to compute (even approximately) as it is a nonconcave maximization problem itself. Indeed, the current understanding of nonconvex-nonconcave minimax problems is fairly limited, in particular compared with the growing literature on non-

convex optimization. The recent research on nonconvex-nonconcave minimax problems mostly relies on some form of convex-concave-like assumptions, such as Minty's Variational Inequality [99] and Polyak-Lojasiewicz conditions [131, 172], which are strong in general and successfully bypass the inherent difficulty in the nonconvex-nonconcave setting. Such theory, unfortunately, presupposes the existence of a globally attractive solution. As such, fundamental nonconvex-nonconcave structures like local solutions and cycling are prohibited.

In an early version of this work [63], we presented preliminary results for analyzing nonconvex-nonconcave bilinear problem (8.3). Simultaneous to (or after) the early version, [96] presents examples of nonconvex-nonconcave minimax problems where a reasonably large class of algorithms do not converge; [79] presents an ODE analysis for the limiting behaviors of different algorithms with shrinking step-size (equivalently it studies the ODE when step-size of an algorithm goes to 0) and shows the possibility to converge to an attractive cycle; [178] utilizes tools from discrete-time dynamic systems to study the behaviors of algorithms around a local stationary solution, which involves the non-transparent complex eigenvalues of the Jacobian matrix at a stationary solution; [64] studies higher-order resolution ODEs of different algorithms for nonconvex-nonconcave minimax problems, which presents more transparent conditions for when a stationary solution is locally attractive, and characterizes the threshold of phase transitions between limit cycles and limit points. Compared to these recent works, we identify theoretical machinery in the saddle envelope that facilitates directly analyzing nonconvex-nonconcave minimax problems. This enables us to obtain a global understanding of the PPM's trajectory, as well as more transparent conditions under which it converges/diverges.

Properties and Convergence of Saddle Envelopes A notion of epi/hypo-convergence of saddle functions and in particular the saddle envelope is developed by Attouch and Wets [11, 10]. These notions of convergence facilitate asymptotic studies of penalty methods [52] and approximate saddle points [65]. Rockafellar [145, 146] further builds generalized second derivatives for saddle functions, which may provide an avenue to relax our assumptions of twice differentiability here. Assuming the given function L is convex-concave, continuity/differentiability properties and relationships between saddle points are developed in [9, 12], which facilitate asymptotic convergence analysis of proximal point methods like [114]. In Section 8.2, we build on these results, giving a calculus for the saddle envelope of nonconvex-nonconcave functions.

Nonconvex Moreau Envelopes The idea to utilize a generalization of the Moreau envelope for nonconvex-nonconcave minimax problems is well motivated by the nonconvex optimization literature. In recent years, the Moreau envelope has found great success as an analysis tool in nonsmooth nonconvex optimization as shown in Chapter 3 and [32, 179] and in nonconvex-concave optimization [138]. There, the Moreau envelope provides an angle of attack for describing stationarity in settings where gradients need not converge to zero even as first-order methods converge. Although we identify a different primary barrier (PPM may not converge at all as cycling and divergence arise from reasonable instances), we still find the Moreau envelope provides the key insight. In our setting, the critical finding is that the saddle envelope can convexify and concavify nonconvex-nonconcave problems, which has no parallel for the classic Moreau envelope.

8.1.3 Preliminaries

Review of convex-concave saddle point optimization. Strongly convex-strongly concave minimax optimization problems $\min_x \max_y M(x,y)$ are well understood. The following lemma is key to the convergence of gradient descent ascent on these problems. In the language of monotone operators, this lemma corresponds to showing $F(x,y) := (\nabla_x M(x,y), -\nabla_y M(x,y))$ is locally strongly monotone (or coercive). From this, the subsequent theorem below shows that gradient descent ascent contracts towards a stationary point when strong convexity-strong concavity and smoothness hold in a region around it. Proofs of these two standard results are given in Appendix 8.6.2 for completeness.

Lemma 8.1.1. *Suppose $M(x,y)$ is μ-strongly convex-strongly concave on a convex set $S = S_x \times S_y$, then it holds for any $(x,y),(x',y') \in S$ that*

$$\mu \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|^2 \leq \left(\begin{bmatrix} \nabla_x M(x,y) \\ -\nabla_y M(x,y) \end{bmatrix} - \begin{bmatrix} \nabla_x M(x',y') \\ -\nabla_y M(x',y') \end{bmatrix} \right)^T \begin{bmatrix} x - x' \\ y - y' \end{bmatrix}.$$

In particular, when $\nabla M(x',y') = 0$, the distance to this stationary point is bounded by

$$\left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\| \leq \frac{\|\nabla M(x,y)\|}{\mu}.$$

Theorem 8.1.2. *Consider any minimax problem $\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} M(x,y)$ where $M(x,y)$ is β-smooth and μ-strongly convex-strongly concave on a set $B(x_0, r) \times B(y_0, r)$ with $r \geq 2\|\nabla M(x_0,y_0)\|/\mu$. Then GDA (8.6) with initial solution (x_0,y_0) and step-size $s \in (0, 2\mu/\beta^2)$ linearly converges to a stationary point $(x^*,y^*) \in B((x_0,y_0), r/2)$ with*

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - 2\mu s + \beta^2 s^2\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2.$$

Review of the Moreau envelope's properties. Denote the Moreau envelope of a function f with proximal parameter $\eta > 0$ by

$$e_\eta\{f\}(x) = \min_u \; f(u) + \frac{\eta}{2}\|u - x\|^2 \,. \tag{8.8}$$

The Moreau envelope of a function provides a lower bound on it everywhere as all $x \in \mathbb{R}^n$ have

$$e_\eta\{f\}(x) \le f(x) \tag{8.9}$$

and if f is ρ-weakly convex and $\eta > \rho$, these functions are equal at the stationary points of f

$$e_\eta\{f\}(x^*) = f(x^*) \quad \iff \quad \nabla f(x^*) = 0 \,. \tag{8.10}$$

Moreover, for ρ-weakly convex functions, there is a nice calculus for the Moreau envelope. Its gradient at some $x \in \mathbb{R}^n$ is determined by the proximal step $x_+ = \operatorname{argmin}_u \, f(u) + \frac{\eta}{2}\|u - x\|^2$ having

$$\nabla e_\eta\{f\}(x) = \eta(x - x_+) = \nabla f(x_+) \,. \tag{8.11}$$

For twice differentiable f, the Moreau envelope is twice differentiable as well with Hessian

$$\nabla^2 e_\eta\{f\}(x) = \eta I - (\eta I + \nabla^2 f(x_+))^{-1} \,. \tag{8.12}$$

From this formula, we can extract the following bounds related to smoothness and convexity

$$(\eta^{-1} - \rho^{-1})^{-1} I \preceq \nabla^2 e_\eta\{f\}(x) \preceq \eta I \,. \tag{8.13}$$

These bounds ensure the Moreau envelope is has a $\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient, which simplifies for convex f (that is, $\rho \le 0$) to have an η-Lipschitz gradient. Noting that $(\eta^{-1} - \rho^{-1})^{-1}$ always has the same sign as $-\rho$, we see that the Moreau envelope is (strongly/weakly) convex exactly when the given function f is (strongly/weakly) convex.

8.2 The Saddle Envelope

In this section, we consider the saddle envelope first developed by Attouch and Wets [11] and characterize its structure for nonconvex-nonconcave optimization. Recall for any proximal parameter $\eta > 0$, the saddle envelope (also referred to as an upper Yosida approximate and a mixed Moreau envelope) is defined as

$$L_\eta(x, y) := \min_{u \in \mathbb{R}^n} \max_{v \in \mathbb{R}^m} L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2.$$

We require that the parameter η is selected with $\eta > \rho$, which ensures the minimax problem in (8.4) is strongly convex-strongly concave. As a result, the saddle envelope is well-defined (as its subproblem has a unique minimax point) and often can be efficiently approximated.

The saddle envelope generalizes the Moreau envelope from the minimization literature to minimax problems. To see this reduction, taking any objective $L(x, y) = g(x)$ (that is, one constant with respect to y) gives

$$L_\eta(x, y) = \min_u \max_v \; g(u) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2 = e_\eta\{g\}(x) \; .$$

We take careful note throughout our theory of similarities and differences from the simpler case of Moreau envelopes. We begin by considering how the value of the saddle envelope L_η relates to the original objective L. Unlike the Moreau envelope in (8.9), the saddle envelope fails to provide a lower bound. If the objective function is constant with respect to y, having $L(x, y) = g(x)$, the saddle envelope becomes a Moreau envelope and provides a lower bound for every (x, y),

$$L_\eta(x, y) = e_\eta\{g\}(x) \leq g(x) = L(x, y) \; .$$

Conversely, if $L(x, y) = h(y)$, then the saddle envelope provides an upper bound. In generic settings between these extremes, the saddle envelope L_η need

not provide any kind of bound on L. The only generic relationship we can establish between $L(z)$ and $L_\eta(z)$ everywhere is that as $\eta \to \infty$, they approach each other. This result is formalized by [11] through epi-hypo convergence.

In the following pair of subsections, we build on the classic results of [11, 9, 12] by deriving a calculus of the saddle envelope of nonconvex-nonconcave functions (in Section 8.2.1) and characterizing the smoothing and convexifying effects of this operation (in Section 8.2.2).

8.2.1 Calculus for the Saddle Envelope L_η

Here we develop a calculus for the saddle envelope L_η, giving formulas for its gradient and Hessian in terms of the original objective L and the proximal operator. These results immediately give algorithmic insights into the proximal point method. First, we show that a generalization of the Moreau gradient formula (8.11) and the convex-concave formula of [9, Theorem 5.1 (d)] holds.

Lemma 8.2.1. *The gradient of the saddle envelope $L_\eta(x, y)$ at $z = (x, y)$ is*

$$\begin{bmatrix} \nabla_x L_\eta(z) \\ \nabla_y L_\eta(z) \end{bmatrix} = \begin{bmatrix} \eta(x - x_+) \\ \eta(y_+ - y) \end{bmatrix} = \begin{bmatrix} \nabla_x L(z_+) \\ \nabla_y L(z_+) \end{bmatrix}$$

where $z_+ = (x_+, y_+) = \operatorname{prox}_{,\eta}(z)$ is given by the proximal operator.

Proof. Notice that the saddle envelope is a composition of Moreau envelopes

$$\begin{aligned} L_\eta(x, y) &= \min_u \Big(\max_v L(u, v) - \frac{\eta}{2}\|v - y\|^2 \Big) + \frac{\eta}{2}\|u - x\|^2 \\ &= \min_u \; -e_\eta\{-L(u, \cdot)\}(y) + \frac{\eta}{2}\|u - x\|^2 = e_\eta\{g(\cdot, y)\}(x) \end{aligned}$$

where $g(u,y) = -e_\eta\{-L(u,\cdot)\}(y)$. Applying the gradient formula (8.11) gives our first claimed gradient formula in x of $\nabla_x L_\eta(x,y) = \eta(x - x_+)$ since x_+ is the unique minimizer of $u \mapsto g(u,y) + \frac{\eta}{2}\|u - x\|^2$. Symmetric reasoning gives our first claimed formula $\nabla_y L_\eta(x,y) = \eta(y_+ - y)$ in y. The second claimed equality is precisely the first-order optimality condition for (8.2). □

Corollary 8.2.2. *The stationary points of L_η are exactly the same as those of L.*

Proof. First consider any stationary point $z = (x,y)$ of L. Denote $z_+ = \mathrm{prox}_{\cdot,\eta}(z)$ and the objective function defining the proximal operator (8.2) as $M(u,v) = L(u,v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2$. Then observing that $\nabla M(z) = 0$, z must be the unique minimax point of M (that is, $z = z_+ = \mathrm{prox}_{\cdot,\eta}(z)$). Hence z must be a stationary point of L_η as well since $\nabla L_\eta(z) = \nabla L(z_+) = \nabla L(z) = 0$.

Conversely consider a stationary point $z = (x,y)$ of L_η. Then $\eta(x - x_+) = \nabla_x L_\eta(z) = 0$ and $\eta(y_+ - y) = \nabla_y L_\eta(z) = 0$. Hence we again find that $z = z_+ = \mathrm{prox}_{\cdot,\eta}(z)$ and consequently, this point must be a stationary point of L as well since $\nabla L(z) = \nabla L(z_+) = \nabla L_\eta(z) = 0$. □

Corollary 8.2.3. *One step of the (damped) PPM* (8.7) *on the original objective L is equivalent to one step of GDA* (8.6) *on the saddle envelope L_η with $s = \lambda/\eta$.*

Proof. Let $(x_k^+, y_k^+) = \mathrm{prox}_{\cdot,\eta}(x_k, y_k)$ and let (x_{k+1}, y_{k+1}) be a step of GDA on $L_\eta(x,y)$ from (x_k, y_k) with step-size $s = \lambda/\eta$. Then

$$\begin{aligned}\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - s \begin{bmatrix} \nabla_x L_\eta(z_k) \\ -\nabla_y L_\eta(z_k) \end{bmatrix} &= \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \frac{\lambda}{\eta} \begin{bmatrix} \eta(x_k - x_k^+) \\ -\eta(y_k^+ - y_k) \end{bmatrix} \\ &= (1-\lambda) \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda \begin{bmatrix} x_k^+ \\ y_k^+ \end{bmatrix}\end{aligned}$$

follows from Lemma 8.2.1. □

Similar to our previous lemma, the Hessian of the saddle envelope at some z is determined by the Hessian of L at $z_+ = \operatorname{prox}_{\cdot,\eta}(z)$. This formula generalizes the Moreau envelope's formula (8.12) whenever L is constant with respect to y.

Lemma 8.2.4. *The Hessian of the saddle envelope $L_\eta(z)$ is*

$$\begin{bmatrix} \nabla^2_{xx} L_\eta(z) & \nabla^2_{xy} L_\eta(z) \\ -\nabla^2_{yx} L_\eta(z) & -\nabla^2_{yy} L_\eta(z) \end{bmatrix} = \eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1}$$

where $z_+ = \operatorname{prox}_{\cdot,\eta}(z)$. Since $\eta > \rho$, we have

$$\nabla^2_{xx} L_\eta(z) = \eta I - \eta^2 \big(\eta I + \nabla_{xx} L(z_+) + \nabla^2_{xy} L(z_+)(\eta I - \nabla^2_{yy} L(z_+))^{-1} \nabla^2_{yx} L(z_+) \big)^{-1},$$

$$\nabla^2_{yy} L_\eta(z) = -\eta I + \eta^2 \big(\eta I + \nabla_{yy} L(z_+) + \nabla^2_{yx} L(z_+)(\eta I + \nabla^2_{xx} L(z_+))^{-1} \nabla^2_{xy} L(z_+) \big)^{-1}.$$

Proof. Consider some $z = (x, y)$ and a nearby point $z^\Delta = z + \Delta$. Denote one proximal step from each of these points by $z_+ = (x_+, y_+) = \operatorname{prox}_{\cdot,\eta}(z)$ and $z_+^\Delta = (x_+^\Delta, y_+^\Delta) = \operatorname{prox}_{\cdot,\eta}(z^\Delta)$. Then our claimed Hessian formula amounts to showing

$$\begin{aligned} &\begin{bmatrix} \nabla_x L_\eta(z^\Delta) \\ -\nabla_y L_\eta(z^\Delta) \end{bmatrix} - \begin{bmatrix} \nabla_x L_\eta(z) \\ -\nabla_y L_\eta(z) \end{bmatrix} \\ &= \left(\eta I - \eta^2 \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1} \right) \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} + o(\|\Delta\|). \end{aligned}$$

Recall Lemma 8.2.1 showed the gradient of the saddle envelope is given by $\nabla_x L_\eta(z) = \eta(x - x_+)$ and $\nabla_y L_\eta(z) = \eta(y_+ - y)$. Applying this at z and z_+ and dividing by η, our claimed Hessian formula becomes

$$\begin{bmatrix} x_+^\Delta - x_+ \\ y_+^\Delta - y_+ \end{bmatrix} = \eta \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1} \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} + o(\|\Delta\|)\,. \quad (8.14)$$

Our proof shows this in two steps: first considering a proximal step on the second-order Taylor approximation of L at z_+ and then showing this closely matches the result of a proximal step on L.

First, consider the following quadratic model of the objective around z_+:

$$\widetilde{L}(z) = L(z_+) + \nabla L(z_+)^T(z - z_+) + \frac{1}{2}(z - z_+)^T \nabla^2 L(z_+)(z - z_+) \, .$$

Denote the result of one proximal step on $\widetilde{L}$ from z^Δ by $\widetilde{z}_+^\Delta = (\widetilde{x}_+^\Delta, \widetilde{y}_+^\Delta)$. Since the proximal subproblem is strongly convex-strongly concave, this solution is uniquely determined by

$$\begin{bmatrix} \nabla_x \widetilde{L}(\widetilde{x}_+^\Delta, \widetilde{y}_+^\Delta) \\ -\nabla_y \widetilde{L}(\widetilde{x}_+^\Delta, \widetilde{y}_+^\Delta) \end{bmatrix} + \begin{bmatrix} \eta(\widetilde{x}_+^\Delta - x^\Delta) \\ \eta(\widetilde{y}_+^\Delta - y^\Delta) \end{bmatrix} = 0 \, .$$

Plugging in the definition of our quadratic model $\widetilde{L}$ yields

$$\begin{bmatrix} \nabla_x L(z_+) \\ -\nabla_y L(z_+) \end{bmatrix} + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \begin{bmatrix} \widetilde{x}_+^\Delta - x_+ \\ \widetilde{y}_+^\Delta - y_+ \end{bmatrix} + \begin{bmatrix} \eta(\widetilde{x}_+^\Delta - x^\Delta) \\ \eta(\widetilde{y}_+^\Delta - y^\Delta) \end{bmatrix} = 0 \, .$$

Hence

$$\begin{aligned} &\left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right) \begin{bmatrix} \widetilde{x}_+^\Delta - x_+ \\ \widetilde{y}_+^\Delta - y_+ \end{bmatrix} = \eta \begin{bmatrix} x^\Delta - x_+ - \eta^{-1} \nabla_x L(z_+) \\ y^\Delta - y_+ + \eta^{-1} \nabla_y L(z_+) \end{bmatrix} \\ \implies &\left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right) \begin{bmatrix} \widetilde{x}_+^\Delta - x_+ \\ \widetilde{y}_+^\Delta - y_+ \end{bmatrix} = \eta \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} \\ \implies &\begin{bmatrix} \widetilde{x}_+^\Delta - x_+ \\ \widetilde{y}_+^\Delta - y_+ \end{bmatrix} = \eta \left(\eta I + \begin{bmatrix} \nabla^2_{xx} L(z_+) & \nabla^2_{xy} L(z_+) \\ -\nabla^2_{yx} L(z_+) & -\nabla^2_{yy} L(z_+) \end{bmatrix} \right)^{-1} \begin{bmatrix} \Delta_x \\ \Delta_y \end{bmatrix} . \end{aligned}$$

This is nearly our target condition (8.14). All that remains is to show our second-order approximation satisfies $\|z_+^\Delta - \widetilde{z}_+^\Delta\| = o(\|\Delta\|)$. Denote the proximal subproblem objective by $M^\Delta(u, v) = L(u, v) + \frac{\eta}{2}\|u - x^\Delta\|^2 - \frac{\eta}{2}\|v - y^\Delta\|^2$ and

its approximation by $\widetilde{M}^{\Delta}(u,v) = \widetilde{L}(u,v) + \frac{\eta}{2}\|u - x^{\Delta}\|^2 - \frac{\eta}{2}\|v - y^{\Delta}\|^2$. Noting that $\|\nabla \widetilde{M}^{\Delta}(x_+, y_+)\| = \eta\|\Delta\|$, we can apply Lemma 8.1.1 to the $(\eta - \rho)$-strongly convex-strongly concave function $\widetilde{M}^{\Delta}$ to bound the distance to its minimax point as

$$\|z_+ - \widetilde{z}_+^{\Delta}\| \leq \frac{\eta}{\eta - \rho}\|\Delta\| \,.$$

Consequently, we can bound difference in gradients between L and its quadratic model $\widetilde{L}$ at $\widetilde{z}_+^{\Delta}$ by $\|\nabla L(\widetilde{z}_+^{\Delta}) - \nabla \widetilde{L}(\widetilde{z}_+^{\Delta})\| = o(\|\Delta\|)$. Therefore $\|\nabla M^{\Delta}(\widetilde{z}_+^{\Delta})\| = o(\|\Delta\|)$ and so applying Lemma 8.1.1 to the strongly convex-strongly concave function M^{Δ} bounds the distance to its minimax point as $\|z_+^{\Delta} - \widetilde{z}_+^{\Delta}\| = o(\|\Delta\|)$, which completes our proof. □

A careful understanding of the saddle envelope's Hessian allows us to describe its smoothness and when it is convex-concave. This is carried out in the following section and forms the crucial step in enabling our convergence analysis for nonconvex-nonconcave problems.

8.2.2 Smoothing and Convexifing from the Saddle Envelope

Recall that the Moreau envelope $e_\eta\{f\}$ serves as a smoothing of any ρ-weakly convex function since its Hessian has uniform bounds above and below (8.13). The lower bound on the Moreau envelope's Hessian guarantees it is convex exactly when the given function f is convex (that is, $\rho = 0$), and strongly convex if and only if f is strongly convex.

In the convex-concave case [12, Proposition 2.2] established the saddle envelope has $1/\eta$-Lipschitz gradient. Our Hessian formula in Lemma 8.2.4 allows us

to quantify the envelope's smoothness for nonconvex-nonconcave objectives. Remarkably, we find that the minimax extension of this result is much more powerful than its Moreau counterpart. The saddle envelope will be convex-concave not just when L is convex-concave, but whenever the following interaction dominance condition holds with a nonnegative parameter α.

Definition 8.2.5. *A function L is α-interaction dominant with respect to x if*

$$\nabla^2_{xx} L(z) + \nabla^2_{xy} L(z)(\eta I - \nabla^2_{yy} L(z))^{-1} \nabla^2_{yx} L(z) \succeq \alpha I \tag{8.15}$$

and α-interaction dominant with respect to y if

$$-\nabla^2_{yy} L(z) + \nabla^2_{yx} L(z)(\eta I + \nabla^2_{xx} L(z))^{-1} \nabla^2_{xy} L(z) \succeq \alpha I \,. \tag{8.16}$$

For any ρ-weakly convex-weakly concave function L, interaction dominance holds with $\alpha = -\rho$ since the second term in these definitions is always positive semidefinite. As a consequence, any convex-concave function is $\alpha \geq 0$-interaction dominant with respect to both x and y. Further, nonconvex-nonconcave functions are interaction dominant with $\alpha \geq 0$ when the second terms above are sufficiently positive definite (hence the name "interaction dominant" as the interaction term of the Hessian $\nabla^2_{xy} L(z)$ is dominating any negative curvature in Hessians $\nabla^2_{xx} L(z)$ and $-\nabla^2_{yy} L(z)$). For example, any problem with β-Lipschitz gradient in y has interaction dominance in x hold with non-negative parameter whenever

$$\frac{\nabla^2_{xy} L(z) \nabla^2_{yx} L(z)}{\eta + \beta} \succeq -\nabla^2_{xx} L(z)$$

since $\eta I - \nabla^2_{yy} L(z) \preceq (\eta + \beta) I$. Similarly, any problem with β-Lipschitz gradient in x has interaction dominance in y with a non-negative parameter whenever

$$\frac{\nabla^2_{yx} L(z) \nabla^2_{xy} L(z)}{\eta + \beta} \succeq \nabla^2_{yy} L(z) \,.$$

The following proposition derives bounds on the Hessian of the saddle envelope showing it is convex in x (concave in y) whenever $\alpha \geq 0$-interaction dominance holds in x (in y). Further, its Hessian lower bounds ensure that L_η is $(\eta^{-1} + \alpha^{-1})^{-1}$-strongly convex in x (strongly concave in y) whenever $\alpha > 0$-interaction dominance holds in x (in y).

Proposition 8.2.6. *If the x-interaction dominance* (8.15) *holds with $\alpha \in \mathbb{R}$, the saddle envelope is smooth and weakly convex with respect to x*

$$(\eta^{-1} + \alpha^{-1})^{-1} I \preceq \nabla_{xx}^2 L_\eta(z) \preceq \eta I \ ,$$

and if the y-interaction dominance condition (8.16) *holds with $\alpha \in \mathbb{R}$, the saddle envelope is smooth and weakly concave with respect to y*

$$(\eta^{-1} + \alpha^{-1})^{-1} I \preceq -\nabla_{yy}^2 L_\eta(z) \preceq \eta I \ .$$

Proof. Recall the formula for the x component of the Hessian given by Lemma 8.2.4. Then the interaction dominance condition (8.15) can lower bound this Hessian by

$$\begin{aligned}
\nabla_{xx}^2 L_\eta(z) &= \eta I - \eta^2 \big(\eta I + \nabla_{xx} L(z_+) + \nabla_{xy}^2 L(z_+)(\eta I - \nabla_{yy}^2 L(z_+))^{-1} \nabla_{yx}^2 L(z_+)\big)^{-1} \\
&\succeq \eta I - \eta^2 (\eta I + \alpha I)^{-1} \\
&= (\eta - \eta^2/(\eta + \alpha)) I \\
&= (\eta^{-1} + \alpha^{-1})^{-1} I \ .
\end{aligned}$$

Note that $\eta I + \nabla_{xx}^2 L(z_+)$ is positive definite (since $\eta > \rho$) and $\nabla_{xy}^2 L(z_+)(\eta I + \nabla_{yy}^2 L(z_+))^{-1} \nabla_{yx}^2 L(z_+)$ is positive semidefinite (since its written as a square). Then the inverse of their sum must also be positive definite and consequently $\nabla_{xx}^2 L_\eta(z)$ is upper bounded by

$$\eta I - \eta^2 \big(\eta I + \nabla_{xx} L(z_+) + \nabla_{xy}^2 L(z_+)(\eta I + \nabla_{yy}^2 L(z_+))^{-1} \nabla_{yx}^2 L(z_+)\big)^{-1} \preceq \eta I \ .$$

Symmetric reasoning applies to give bounds on $\nabla^2_{yy} L_\eta(z)$. $\square$

Remark 8.2.7. *Note that our definition of interaction dominance depends on the choice of the proximal parameter $\eta > \rho$. In our convergence theory, we will show that interaction dominance with nonnegative $\alpha > 0$ captures when the proximal point method with the same parameter η converges.*

Remark 8.2.8. *Proposition 8.2.6 generalizes the Hessian bounds for the Moreau envelope* (8.13) *since for any $L(x, y)$ that is constant in y, the α-interaction dominance condition in x simplifies to simply be ρ-weak convexity $\nabla^2_{xx} L(z) + \nabla^2_{xy} L(z)(\eta I - \nabla^2_{yy} L(z))^{-1} \nabla^2_{yx} L(z) = \nabla^2_{xx} L(z) \succeq \alpha I$. Hence this special case has $\alpha = -\rho$.*

In addition to bounding the Hessians of the x and y variables separately, we can also bound the overall smoothness of the saddle envelope. Our next result shows that the saddle envelope maintains the same $\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-smoothing effect as the Moreau envelope (8.13).

Proposition 8.2.9. *L_η has $\max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient.*

Proof. Consider two points $z = (x, y)$ and $\bar{z} = (\bar{x}, \bar{y})$ and denote one proximal step from each of them by $z_+ = (x_+, y_+) = \operatorname{prox}_{\cdot,\eta}(z)$ and $\bar{z}_+ = (\bar{x}_+, \bar{y}_+) = \operatorname{prox}_{\cdot,\eta}(\bar{z})$. Define the $(\eta - \rho)$-strongly convex-strongly concave function underlying the computation of the saddle envelope at z as

$$M(u, v) = L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2.$$

First we compute the gradient of M at $\bar{z}_+$ which is given by

$$\begin{bmatrix} \nabla_x M(\bar{z}_+) \\ \nabla_y M(\bar{z}_+) \end{bmatrix} = \begin{bmatrix} \nabla_x L(\bar{z}_+) + \eta(\bar{x}_+ - x) \\ \nabla_y L(\bar{z}_+) - \eta(\bar{y}_+ - y) \end{bmatrix} = \eta \begin{bmatrix} \bar{x} - x \\ y - \bar{y} \end{bmatrix}.$$

Applying Lemma 8.1.1, and noting that $z_+ = \text{prox}_{,\eta}(z)$ has $\nabla M(z_+) = 0$ yields

$$\left\| \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} \right\|^2 \leq \frac{\eta}{\eta - \rho} \begin{bmatrix} \bar{x} - x \\ \bar{y} - y \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix}.$$

Recalling the saddle envelope's gradient formula from Lemma 8.2.1, we can upper bound the difference between its gradients at z and $\bar{z}$ by

$$\begin{aligned} \frac{1}{\eta^2}\|\nabla L_\eta(z) - \nabla L_\eta(\bar{z})\|^2 &= \left\| \begin{bmatrix} x - x_+ \\ y_+ - y \end{bmatrix} - \begin{bmatrix} \bar{x} - \bar{x}_+ \\ \bar{y}_+ - \bar{y} \end{bmatrix} \right\|^2 \\ &= \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 + 2 \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} + \left\| \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} \right\|^2 \\ &\leq \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 + \left(\frac{\eta}{\eta - \rho} - 2 \right) \begin{bmatrix} \bar{x} - x \\ \bar{y} - y \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix}. \end{aligned}$$

Notice that $\begin{bmatrix} \bar{x} - x \\ \bar{y} - y \end{bmatrix}^T \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix}$ is non-negative but the sign of $\left(\frac{\eta}{\eta - \rho} - 2 \right)$ may be positive or negative. If this coefficient is negative, we can upperbound the second term above by zero, giving $\|\nabla L_\eta(z) - \nabla L_\eta(\bar{z})\|^2 \leq \eta^2 \|z - \bar{z}\|^2$. If instead $\left(\frac{\eta}{\eta - \rho} - 2 \right) \geq 0$, then we have smoothness constant $|\eta^{-1} - \rho^{-1}|^{-1}$ as

$$\begin{aligned} \|\nabla L_\eta(z) - \nabla L_\eta(\bar{z})\|^2 &\leq \eta^2 \left(1 + \left(\frac{\eta}{\eta - \rho} - 2 \right) \frac{\eta}{\eta - \rho} \right) \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 \\ &= \eta^2 \left(\frac{\eta}{\eta - \rho} - 1 \right)^2 \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 \\ &= \left(\frac{\eta\rho}{\eta - \rho} \right)^2 \left\| \begin{bmatrix} x - \bar{x} \\ y - \bar{y} \end{bmatrix} \right\|^2 \end{aligned}$$

by Cauchy Schwarz and using that $\left\| \begin{bmatrix} \bar{x}_+ - x_+ \\ \bar{y}_+ - y_+ \end{bmatrix} \right\| \leq \frac{\eta}{\eta - \rho} \left\| \begin{bmatrix} \bar{x} - x \\ \bar{y} - y \end{bmatrix} \right\|$. □

The setting of taking the Moreau envelope of a convex function gives a simpler smoothness bound of η since having $\rho \leq 0$ implies $\eta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$. The same simplification holds when applying our saddle envelope machinery to convex-concave problems: the saddle envelope of any convex-concave L is η-smooth, matching the results of [12, Proposition 2.1].

8.3 Interaction Dominant Regime

Our theory for the saddle envelope $L_\eta(z)$ shows it is much more structured than the original objective function $L(z)$. Proposition 8.2.6 established that for x and y interaction dominant problems, the saddle envelope is strongly convex-strongly concave. Proposition 8.2.9 established that the saddle envelope is always smooth (has a uniformly Lipschitz gradient). Both of these results hold despite us not assuming convexity, concavity, or smoothness of the original objective. Historically these two conditions are the key to linear convergence (see Theorem 8.1.2) and indeed we find interaction dominance causes the proximal point method to linearly converge. The proof of this result is deferred to the end of the section.

Theorem 8.3.1. *For any objective L that is ρ-weakly convex-weakly concave and $\alpha > 0$-interaction dominant in both x and y, the damped PPM* (8.7) *with η and λ satisfying*

$$\lambda \leq 2\frac{\min\{1, (\eta/\rho - 1)^2\}}{\eta/\alpha + 1}$$

linearly converges to the unique stationary point (x^*, y^*) of (8.1) *with*

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{2\lambda}{\eta/\alpha + 1} + \frac{\lambda^2}{\min\{1, (\eta/\rho - 1)^2\}}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2.$$

For example, setting $\eta = 2\rho$ and $\lambda = \frac{1}{1+\eta/\alpha}$, our convergence rate simplifies to

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{1}{(2\rho/\alpha + 1)^2}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2.$$

Remark 8.3.2. *Theorem 8.3.1 is valid even if $\alpha > 0$-interaction dominance only holds locally. That is, as long as α-interaction dominance holds within an l_2-ball around a local stationary point, and the initial point is sufficiently within this ball, then PPM converges linearly to this local stationary point.*

Remark 8.3.3. *For μ-strongly convex-strongly concave problems, this theorem recovers the standard proximal point convergence rate for any choice of $\eta > 0$. In this case, we have $\rho = -\mu$, $\alpha = \mu$, and can set $\lambda = \frac{1}{\eta/\mu+1}$, giving a $O(\eta^2/\mu^2 \log(1/\varepsilon))$ convergence rate matching [144].*

Remark 8.3.4. *The $\alpha > 0$-interaction dominance condition is tight for obtaining global linear convergence. A nonconvex-nonconcave quadratic example illustrating the sharpness of this boundary is presented in Section 8.5.1. Moreover, our example shows that it is sometimes necessary to utilize the damping parameter (that is, selecting $\lambda < 1$) for PPM to converge.*

If we only have $\alpha > 0$-interaction dominance with y, then the saddle envelope L_η is still much more structured than the original objective L. In this case, $L_\eta(x, y)$ may still be nonconvex in x, but Proposition 8.2.9 ensures it is strongly concave in y. Then our theory allows us to extend existing convergence guarantees for nonconvex-concave problems to this larger class of y interaction dominant problems. For example, Lin et al. [100] recently showed that GDA with different, carefully chosen stepsize parameters for x and y will converge to a stationary point at a rate of $O(\varepsilon^{-2})$. We find that running the following damped

proximal point method is equivalent to running their variant of GDA on the saddle envelope

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} \lambda x_k^+ + (1-\lambda)x_k \\ \gamma y_k^+ + (1-\gamma)y_k \end{bmatrix} \text{ where } \begin{bmatrix} x_k^+ \\ y_k^+ \end{bmatrix} = \operatorname{prox}_{\cdot,\eta}(x_k, y_k) \tag{8.17}$$

for proper choice of the parameters $\lambda, \gamma \in [0,1]$. From this, we derive the following sublinear convergence rate for nonconvex-nonconcave problems whenever y interaction dominance holds, proven at the end of the section.

Theorem 8.3.5. *For any objective L that is ρ-weakly convex-weakly concave and $\alpha > 0$-interaction dominant in y, consider the PPM variant* (8.17) *with damping constants $\lambda = \Theta\left(\frac{\min\{1, |\eta/\rho - 1|^3\}}{(1+\eta/\alpha)^2}\right)$ and $\gamma = \Theta(\min\{1, |\eta/\rho - 1|\})$. If the sequence y_k is bounded*[1]*, then a stationary point $\|\nabla L(x_T^+, y_T^+)\| \leq \varepsilon$ will be found by iteration $T \leq O(\varepsilon^{-2})$.*

Remark 8.3.6. *Symmetrically, we can guarantee sublinear convergence assuming only x-interaction dominance. Considering the problem of $\max_y \min_x L(x,y) = -\min_y \max_x -L(x,y)$, which is now interaction dominant with respect to the inner maximization variable, we can apply Theorem 8.3.5. This reduction works since although the original minimax problem and this maximin problem need not have the same solutions, they always have the same stationary points.*

8.3.1 Proof of Theorem 8.3.1

Propositions 8.2.6 and 8.2.9 show that L_η is $\mu = (\eta^{-1} + \alpha^{-1})^{-1}$-strongly convex-strongly concave and has a $\beta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient. Having strong convexity and strong concavity ensures L_η has a unique stationary

[1] We do not believe this boundedness condition is fundamentally needed, but we make it to leverage the results of [100] which utilize compactness.

point (x^*, y^*), which in turn must be the unique stationary point of L by Corollary 8.2.2. Recall Corollary 8.2.3 showed that the damped PPM (8.7) on L is equivalent to GDA (8.6) with $s = \lambda/\eta$ on L_η. Then provided

$$\lambda \leq 2\frac{\min\{1, (\eta/\rho - 1)^2\}}{\eta/\alpha + 1} = \frac{2(\eta^{-1} + \alpha^{-1})^{-1}}{\max\{\eta^2, (\eta^{-1} - \rho^{-1})^{-2}\}},$$

we have $s = \lambda/\eta \in (0, 2\mu/\beta^2)$. Hence applying Theorem 8.1.2 shows the iterations of GDA (and consequently PPM) linearly converge to this unique stationary point as

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{2\lambda}{\eta/\alpha + 1} + \frac{\lambda^2}{\min\{1, (\eta/\rho - 1)^2\}}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2.$$

8.3.2 Proof of Theorem 8.3.5

Proposition 8.2.6 shows that whenever interaction dominance holds for y the saddle envelope is $\mu = (\eta^{-1} + \alpha^{-1})^{-1}$-strongly concave in y and Proposition 8.2.9 ensures the saddle envelope has a $\beta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-Lipschitz gradient. Recently, Lin et al. [100] considered such nonconvex-strongly concave problems with a compact constraint $y \in D$. They analyzed the following variant of GDA

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \text{proj}_{\mathbb{R}^n \times D}\left(\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} -\nabla_x L(x_k, y_k)/\eta_x \\ \nabla_y L(x_k, y_k)/\eta_y \end{bmatrix}\right) \tag{8.18}$$

which projects onto the feasible region $\mathbb{R}^n \times D$ each iteration and has different stepsize parameters η_x and η_y for x and y. Lin et al. prove the following theorem showing a sublinear guarantee.

Theorem 8.3.7 (Theorem 4.4 of [100])**.** *For any β-smooth, nonconvex-μ-strongly concave L, let $\kappa = \beta/\mu$ be the condition number for y. Then for any $\varepsilon > 0$, GDA*

with stepsizes $\eta_x^{-1} = \Theta(1/\kappa^2\beta)$ *and* $\eta_y^{-1} = \Theta(1/\beta)$ *will find a point satisfying* $\|\nabla L(x_T, y_T)\| \le \varepsilon$ *by iteration*

$$T \le O\left(\frac{\kappa^2\beta + \kappa\beta^2}{\varepsilon^2}\right).$$

Assuming that the sequence y_k above stays bounded, this projected gradient method is equivalent to running GDA on our unconstrained problem by setting the domain of y as a sufficiently large compact set to contain all the iterates. Consider setting the averaging parameters as $\lambda = \Theta(\eta/\kappa^2\beta) = \Theta\left(\frac{\min\{1,|\eta/\rho-1|^3\}}{(1+\eta/\alpha)^2}\right)$ and $\gamma = \Theta(\eta/\beta) = \Theta(\min\{1, |\eta/\rho - 1|\})$. Then using the gradient formula from Lemma 8.2.1, we see that the damped proximal point method (8.17) is equivalent to running GDA on the saddle envelope with $\eta_x = \eta/\lambda$ and $\eta_y = \eta/\gamma$:

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} -\nabla_x L_\eta(x_k, y_k)/\eta_x \\ \nabla_y L_\eta(x_k, y_k)/\eta_y \end{bmatrix} = \begin{bmatrix} \lambda x_k^+ + (1-\lambda)x_k \\ \gamma y_k^+ + (1-\gamma)y_k \end{bmatrix}.$$

Then the above theorem guarantees that running this variant of the proximal point method on L (or equivalently, applying the GDA variant (8.18) to the saddle envelope) will converge to a stationary point with $\|\nabla L_\eta(z_T)\| \le \varepsilon$ within $T \le O(\varepsilon^{-2})$ iterations. It immediate follows from the gradient formula that $z_T^+ = \mathrm{prox}_{\cdot,\eta}(z_T)$ is approximately stationary for L as $\|\nabla L(z_T^+)\| = \|\nabla L_\eta(z_T)\| \le \varepsilon$.

8.4 Interaction Weak Regime

Our previous theory showed that when the interaction between x and y is sufficiently strong, global linear convergence occurs. Now we consider when there is limited interaction between x and y. At the extreme of having no interaction,

nonconvex-nonconcave minimax optimization separates into nonconvex minimization and nonconcave maximization. On these separate problems, local convergence of the proximal point method is well-understood.Here we show that under reasonable smoothness and initialization assumptions, this local convergence behavior extends to minimax problems with weak, but nonzero, interaction between x and y.

To formalize this, we make the following regularity assumptions

$$\|\nabla^2 L(z)\| \leq \beta \ , \quad \text{for all } z \in \mathbb{R}^n \times \mathbb{R}^m \tag{8.19}$$

$$\|\nabla^2 L(z) - \nabla^2 L(\bar{z})\| \leq H\|z - \bar{z}\| \ , \quad \text{for all } z, \bar{z} \in \mathbb{R}^n \times \mathbb{R}^m \tag{8.20}$$

and quantify how weak the interaction is by assuming

$$\|\nabla^2_{xy} L(z)\| \leq \delta \quad , \quad \text{for all } z \in \mathbb{R}^n \times \mathbb{R}^m \tag{8.21}$$

$$\begin{cases} \|\nabla^2_{xx} L(x,y) - \nabla^2_{xx} L(x,\bar{y})\| \leq \xi\|y - \bar{y}\| \\ \|\nabla^2_{yy} L(x,y) - \nabla^2_{yy} L(\bar{x},y)\| \leq \xi\|x - \bar{x}\| \end{cases} , \quad \text{for all } (x,y), (\bar{x},\bar{y}) \in \mathbb{R}^n \times \mathbb{R}^m \tag{8.22}$$

for some constants $\beta, H, \delta, \xi \geq 0$. Here we are particularly interested in problems where δ and ξ are sufficiently small. For example, the bilinear setting of (8.3) satisfies this with $(\delta, \xi) = (\lambda_{max}(A), 0)$ and so we are considering small interaction matrices A.

For such problems, we consider an initialization for the proximal point method based on our motivating intuition that when there is no interaction, we can find local minimizers and maximizers with respect to x and y. For a

fixed point $z' = (x', y')$, we compute our PPM initialization $z_0 = (x_0, y_0)$ as

$$\begin{cases} x_0 = \text{a local minimizer of } \min_u L(u, y') \ , \\ y_0 = \text{a local maximizer of } \max_v L(x', v) \ . \end{cases} \tag{8.23}$$

These subproblems amount to smooth nonconvex minimization, which is well-studied (see for example [91]), and so we take them as a blackbox.

The critical observation explaining why this is a good initialization is that provided δ and ξ are small enough, we have (i) that the interaction dominance conditions (8.15) and (8.16) hold at z_0 with a nearly positive $\alpha = \alpha_0$, often with $\alpha_0 > 0$ and (ii) that z_0 is a nearly stationary point of L. Below we formalize each of these properties and arrive at conditions quantifying how small we need ξ and δ to be for our local convergence theory to apply.

(i) First, we observe that the interaction dominance conditions (8.15) and (8.16) hold at z_0 with a nearly positive coefficient α_0. Since x_0 and y_0 are local optimum, for some $\mu \geq 0$, we must have

$$\nabla^2_{xx} L(x_0, y') \succeq \mu I \quad \text{and} \quad -\nabla^2_{yy} L(x', y_0) \succeq \mu I \ .$$

Then the Hessians at z_0 must be similarly bounded since the amount they can change is limited by (8.22). Hence

$$\nabla^2_{xx} L(z_0) \succeq (\mu - \xi \|y_0 - y'\|) I \quad \text{and} \quad -\nabla^2_{yy} L(z_0) \succeq (\mu - \xi \|x_0 - x'\|) I \ .$$

Adding a positive semidefinite term onto these (as is done in the definition of interaction dominance) can only increase the righthand-side above. In particular, we can bound the second term added in the interaction domi-

nance conditions (8.15) and (8.16) as

$$\begin{aligned}\nabla^2_{xy}L(z_0)(\eta I - \nabla^2_{yy}L(z_0))^{-1}\nabla^2_{yx}L(z_0) &\succeq \frac{\nabla^2_{xy}L(z_0)\nabla^2_{yx}L(z_0)}{\eta+\beta} \\ &\succeq \frac{\lambda_{min}(\nabla^2_{xy}L(z_0)\nabla^2_{yx}L(z_0))}{\eta+\beta} I \geq 0, \\ \nabla^2_{yx}L(z_0)(\eta I + \nabla^2_{xx}L(z_0))^{-1}\nabla^2_{xy}L(z_0) &\succeq \frac{\nabla^2_{yx}L(z_0)\nabla^2_{xy}L(z_0)}{\eta+\beta} \\ &\succeq \frac{\lambda_{min}(\nabla^2_{yx}L(z_0)\nabla^2_{xy}L(z_0))}{\eta+\beta} I \geq 0.\end{aligned}$$

Hence interaction dominance holds at z_0 in both x and y with

$$\begin{aligned}&\nabla^2_{xx}L(z_0) + \nabla^2_{xy}L(z_0)(\eta I - \nabla^2_{yy}L(z_0))^{-1}\nabla^2_{yx}L(z_0) \\ &\qquad\succeq \left(\mu + \frac{\lambda_{min}(\nabla^2_{xy}L(z_0)\nabla^2_{yx}L(z_0))}{\eta+\beta} - \xi\|y_0 - y'\|\right) I\ , \\ &-\nabla^2_{yy}L(z_0) + \nabla^2_{yx}L(z_0)(\eta I + \nabla^2_{xx}L(z_0))^{-1}\nabla^2_{xy}L(z_0) \\ &\qquad\succeq \left(\mu + \frac{\lambda_{min}(\nabla^2_{yx}L(z_0)\nabla^2_{xy}L(z_0))}{\eta+\beta} - \xi\|x_0 - x'\|\right) I\ .\end{aligned}$$

For our local linear convergence theory to apply, we need this to hold with positive coefficient. It suffices to have ξ sufficiently small, satisfying

$$\begin{cases}\xi\|y_0 - y'\| < \mu + \dfrac{\lambda_{min}(\nabla^2_{xy}L(z_0)\nabla^2_{yx}L(z_0))}{\eta+\beta} \\ \xi\|x_0 - x'\| < \mu + \dfrac{\lambda_{min}(\nabla^2_{yx}L(z_0)\nabla^2_{xy}L(z_0))}{\eta+\beta}\ .\end{cases} \tag{8.24}$$

Note this is trivially the case for problems with bilinear interaction (8.3) as $\xi = 0$. It is also worth noting that even if $\mu = 0$, the right-hand-sides above are still strictly positive if $\nabla_{xy}L(z_0)$ is full rank and the variable dimensions n and m of x and y are equal[2].

(ii) Next, we observe that z_0 is nearly stationary by applying (8.21) and using

[2]This works since having full rank square $\nabla^2_{xy}L(z_0)$ implies that both of its squares $\nabla^2_{xy}L(z_0)\nabla^2_{yx}L(z_0)$ and $\nabla^2_{yx}L(z_0)\nabla^2_{xy}L(z_0)$ are full rank as well. Hence these squares must be strictly positive definite and as a result, have strictly positive minimum eigenvalues.

the first-order optimality conditions of the subproblems (8.23):

$$\|\nabla L(z_0)\| \le \left\| \begin{bmatrix} \nabla_x L(x_0, y') \\ \nabla_y L(x', y_0) \end{bmatrix} \right\| + \delta \|z_0 - z'\| = \delta \|z_0 - z'\|.$$

For our convergence theory, this gradient needs to be sufficiently small

$$\delta \|z_0 - z'\| \le \frac{\alpha_0(\eta - \rho)}{2\left(1 + \frac{4\sqrt{2}(\eta+\alpha_0/2)}{\alpha_0} + \frac{4\sqrt{2}\beta(\eta+\alpha_0/2)}{\alpha_0(\eta-\rho)}\right) H\left(1 + \frac{2\delta}{\eta-\rho} + \frac{\delta^2}{(\eta-\rho)^2}\right)}. \quad (8.25)$$

Under these conditions, we have the following linear convergence guarantee.

Theorem 8.4.1. *For any objective L satisfying weak convexity-concavity* (8.5)*, the smoothness conditions* (8.19) *and* (8.20)*, and the interaction bounds* (8.21) *and* (8.22)*, consider the damped PPM* (8.7) *with initialization (x_0, y_0) given by* (8.23) *and η and λ satisfying*

$$\lambda \le 2\frac{\min\{1, (\eta/\rho - 1)^2\}}{2\eta/\alpha_0 + 1}.$$

Then PPM linearly converges to a nearby stationary point (x^*, y^*) of (8.1) *with*

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \le \left(1 - \frac{2\lambda}{2\eta/\alpha_0 + 1} + \frac{\lambda^2}{\min\{1, (\eta/\rho - 1)^2\}}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2$$

provided δ and ξ are small enough to satisfy (8.24) *and* (8.25)*.*

8.4.1 Proof of Theorem 8.4.1

Our proof of this local convergence guarantee considers two sets centered at (x_0, y_0): An inner region $B_{\text{inner}} = B(x_0, r) \times B(y_0, r)$ with radius

$$r := \frac{4(\eta + \alpha_0/2)}{\alpha_0} \frac{\|\nabla L(z_0)\|}{\eta - \rho}$$

and an outer ball $B_{\text{outer}} = B((x_0, y_0), R)$ with radius

$$R := \left(1 + \frac{4\sqrt{2}(\eta + \alpha_0/2)}{\alpha_0} + \frac{4\sqrt{2}\beta(\eta + \alpha_0/2)}{\alpha_0(\eta - \rho)}\right)\frac{\|\nabla L(z_0)\|}{\eta - \rho} \geq \sqrt{2}r \ .$$

Thus $B_{\text{inner}} \subseteq B_{\text{outer}}$. The following lemma shows that the $\alpha_0 > 0$-interaction dominance at z_0 (following from our initialization procedure) extends to give $\alpha_0/2$-interaction dominance on the whole outer ball B_{outer}.

Lemma 8.4.2. *On B_{outer}, $\alpha_0/2$-iteration dominance holds in both x and y.*

Proof. First, observe that the functions defining the interaction dominance conditions (8.15) and (8.16)

$$\nabla^2_{xx}L(z) + \nabla^2_{xy}L(z)(\eta I - \nabla^2_{yy}L(z))^{-1}\nabla^2_{yx}L(z),$$
$$-\nabla^2_{yy}L(z) + \nabla^2_{yx}L(z)(\eta I + \nabla^2_{xx}L(z))^{-1}\nabla^2_{xy}L(z)$$

are both uniformly Lipschitz with constant[3]

$$H\left(1 + \frac{2\delta}{\eta - \rho} + \frac{\delta^2}{(\eta - \rho)^2}\right).$$

Then our Lipschitz constant follows by observing the component functions defining it satisfy the following: $\nabla^2_{xx}L(z)$ and $\nabla^2_{yy}L(z)$ are H-Lipschitz, $\nabla^2_{xy}L(z)$ and its transpose $\nabla^2_{yx}L(z)$ are both H-Lipschitz and bounded in norm by δ, and $(\eta I + \nabla^2_{xx}L(z))^{-1}$ and $(\eta I - \nabla^2_{yy}L(z))^{-1}$ are both $H/(\eta-\rho)^2$-Lipschitz and bounded in norm by $(\eta - \rho)^{-1}$.

[3]This constant follows from multiple applications of the "product rule"-style formula that $A(z)B(z)$ is uniformly $(a'b + ab')$-Lipschitz provided $A(z)$ is bounded by a and a'-Lipschitz and $B(z)$ is bounded by b and b'-Lipschitz: any two points z, z' have

$$\begin{aligned}\|A(z)B(z) - A(z')B(z')\| &\leq \|A(z)B(z) - A(z')B(z)\| + \|A(z')B(z) - A(z')B(z')\| \\ &\leq (a'b + b'a)\|z - z'\|.\end{aligned}$$

It follows that every $z \in B_{\text{outer}}$ has $\alpha_0/2$-interaction dominance in x as

$$\begin{aligned}
&\nabla^2_{xx} L(z) + \nabla^2_{xy} L(z)(\eta I - \nabla^2_{yy} L(z))^{-1} \nabla^2_{yx} L(z) \\
&\succeq \nabla^2_{xx} L(z_0) + \nabla^2_{xy} L(z_0)(\eta I - \nabla^2_{yy} L(z_0))^{-1} \nabla^2_{yx} L(z_0) - H\left(1 + \frac{2\delta}{\eta - \rho} + \frac{\delta^2}{(\eta-\rho)^2}\right) RI \\
&\succeq \nabla^2_{xx} L(z_0) + \nabla^2_{xy} L(z_0)(\eta I - \nabla^2_{yy} L(z_0))^{-1} \nabla^2_{yx} L(z_0) - \alpha_0/2 I \\
&\succeq \alpha_0 I - \alpha_0/2 I = \alpha_0/2 I
\end{aligned}$$

where the first inequality uses Lipschitz continuity, the second inequality uses our assumed condition (8.25) of $H\left(1 + \frac{2\delta}{\eta-\rho} + \frac{\delta^2}{(\eta-\rho)^2}\right) R \leq \alpha_0/2$, and the third inequality uses the α_0-interaction dominance at z_0. Symmetric reasoning shows $\alpha_0/2$-interaction dominance in y holds for each $z \in B_{\text{outer}}$ as well. □

From this, interaction dominance on the outer ball suffices to ensure the saddle envelope is strongly convex-strongly concave on the inner region.

Lemma 8.4.3. *The saddle envelope is $(\eta^{-1} + (\alpha_0/2)^{-1})^{-1}$-strongly convex-strongly concave on B_{inner}.*

Proof. Given $\alpha_0/2$-interaction dominance holds on B_{outer}, it suffices to show that for any $z = (x, y) \in B_{\text{inner}}$, the proximal step $z_+ = \text{prox}_{\cdot,\eta}(z) \in B_{\text{outer}}$ as we can then apply the Hessian bounds from Proposition 8.2.6 to show strong convexity and strong concavity.

Define the function underlying the computation of the proximal step at (x, y) as

$$M(u, v) = L(u, v) + \frac{\eta}{2}\|u - x\|^2 - \frac{\eta}{2}\|v - y\|^2.$$

Our choice of $\eta > \rho$ ensures that M is $(\eta - \rho)$-strongly convex-strongly concave. Thus applying Lemma 8.1.1 and then the β-Lipschitz continuity of $\nabla L(z)$

implies

$$\left\| \begin{bmatrix} x - x_+ \\ y - y_+ \end{bmatrix} \right\| \leq \frac{\|\nabla M(x,y)\|}{\eta - \rho} = \frac{\|\nabla L(x,y)\|}{\eta - \rho} \leq \frac{\|\nabla L(x_0,y_0)\| + \beta\sqrt{2}r}{\eta - \rho}.$$

Hence $\|z_0 - z_+\| \leq \|z_0 - z\| + \|z - z_+\| \leq \sqrt{2}r + \frac{\|\nabla L(z_0)\| + \beta\sqrt{2}r}{\eta - \rho} = R.$ □

Armed with the knowledge that interaction dominance holds on B_{inner}, we return to the proof of Theorem 8.4.1. Observe that the gradient of the saddle envelope at $z_0 = (x_0, y_0)$ is bounded by Lemma 8.2.1 and Lemma 8.1.1 as

$$\|\nabla L_\eta(z_0)\| = \|\eta(z_0 - z_0^+)\| \leq \frac{\eta}{\eta - \rho}\|\nabla M_0(z_0)\| = \frac{\eta}{\eta - \rho}\|\nabla L(z_0)\|$$

where $z_0^+ = \operatorname{prox}_{\cdot,\eta}(z_0)$ and $M_0(u,v) = L(u,v) + \frac{\eta}{2}\|u - x_0\|^2 - \frac{\eta}{2}\|v - y_0\|^2$ is the $\eta - \rho$-strongly convex-strongly concave function defining it. Now we have shown all of the conditions necessary to apply Theorem 8.1.2 on the square $B(x_0, r) \times B(y_0, r)$ with

$$r = \frac{4(\eta + \alpha_0/2)\|\nabla L(z_0)\|}{\alpha_0(\eta - \rho)} = \frac{2\|\nabla L_\eta(z_0)\|}{\mu}$$

upon which the saddle envelope is $\mu = (\eta^{-1} + (\alpha_0/2)^{-1})^{-1}$-strongly convex-strongly concave and $\beta = \max\{\eta, |\eta^{-1} - \rho^{-1}|^{-1}\}$-smooth. Hence applying GDA with $s = \lambda/\eta$ to the saddle envelope produces iterates (x_k, y_k) converging to a stationary point (x^*, y^*) with

$$\left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \leq \left(1 - \frac{2\lambda}{\eta(\eta^{-1} + (\alpha_0/2)^{-1})} + \frac{\lambda^2}{\eta^2(\eta^{-1} - \rho^{-1})^2}\right)^k \left\| \begin{bmatrix} x_0 - x^* \\ y_0 - y^* \end{bmatrix} \right\|^2.$$

By Corollary 8.2.2, (x^*, y^*) must also be a stationary point of L. Further, by Corollary 8.2.3, this sequence (x_k, y_k) is the same as the sequence generated by running the damped PPM on (8.1).

8.5 Interaction Moderate Regime

Between the interaction dominant and interaction weak regimes, the proximal point method may diverge or cycle indefinitely (recall our introductory example in Figure 8.1 where convergence fails in this middle regime). We begin by considering the behavior of the proximal point method when applied to a nonconvex-nonconcave quadratic example. From this, our interaction dominance condition is tight, exactly describing when our example converges.

8.5.1 Tightness of the Interaction Dominance Regime

Consider the following nonconvex-nonconcave quadratic minimax problem of

$$\min_{x\in\mathbb{R}^n}\max_{y\in\mathbb{R}^n} L(x,y) = \frac{-\rho}{2}\|x\|^2 + ax^Ty - \frac{-\rho}{2}\|y\|^2 \tag{8.26}$$

where $a \in \mathbb{R}$ controls the size of the interaction between x and y and $\rho \geq 0$ controls how weakly convex-weakly concave the problem is. Notice this problem has a stationary point at the origin. Even though this problem is nonconvex-nonconcave, PPM will still converge to the origin for some selections of a, ρ, and η. Examining our interaction dominance conditions (8.15) and (8.16), this example is $\alpha = -\rho + a^2/(\eta - \rho)$-interaction dominant in both x and y.

For quadratic problems, PPM always corresponds to the matrix multiplica-

tion. In the case of (8.26), the damped PPM iteration is given by

$$\begin{aligned}
\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} &= (1-\lambda)\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \lambda \begin{bmatrix} (1-\rho/\eta)I & aI/\eta \\ -aI/\eta & (1-\rho/\eta)I \end{bmatrix}^{-1} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \\
&= (1-\lambda)\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \frac{\lambda\eta}{\eta-\rho}\left(\begin{bmatrix} I & aI/(\eta-\rho) \\ -aI/(\eta-\rho) & I \end{bmatrix} \right)^{-1} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \\
&= (1-\lambda)\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \frac{\lambda\eta}{a^2/(\eta-\rho)+\eta-\rho} \begin{bmatrix} I & -aI/(\eta-\rho) \\ aI/(\eta-\rho) & I \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \\
&= \begin{bmatrix} CI & -DI \\ DI & CI \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix}
\end{aligned}$$

for constants $C = 1 - \dfrac{\lambda\alpha}{\eta+\alpha}$ and $D = \dfrac{\lambda\eta a}{(\eta+\alpha)(\eta-\rho)}$. Notice that these constants are well-defined since $\eta - \rho > 0$ and $\eta + \alpha > 0$ (even if α is negative) since $\eta > \rho$ and $\alpha \geq -\rho$. Matrix multiplication of this special final form has the following nice property for any z,

$$\left\| \begin{bmatrix} CI & -DI \\ DI & CI \end{bmatrix} z \right\|^2 = (C^2 + D^2)\|z\|^2. \tag{8.27}$$

Hence this iteration will globally converge to the origin exactly when

$$\left(1 - \frac{\lambda\alpha}{\eta+\alpha}\right)^2 + \left(\frac{\lambda\eta a}{(\eta+\alpha)(\eta-\rho)}\right)^2 < 1\,.$$

Likewise, the damped proximal point method will cycle indefinitely when this holds with equality and diverges when it is strictly violated. As a result, violating $\alpha > 0$-interaction dominance (that is, having $\alpha \leq 0$) leads to divergence in (8.26) for any choice of the averaging parameter $\lambda \in (0, 1]$ since this forces $C \geq 1$ (and so $C^2 + D^2 > 1$). Hence our interaction dominance boundary is tight.

Further, this example shows that considering the damped proximal point method (as opposed to fixing $\lambda = 1$) is necessary to fully capture the convergence for interaction dominant problems. For example, setting $\rho = 1, a = 2, \eta = 3$ has $\alpha = 1$-interaction dominance in x and y and converges exactly when

$$(1 - \lambda/4)^2 + (3\lambda/4)^2 < 1$$

which is satisfied when $\lambda \in (0, 0.8)$, but not by the undamped proximal point method with $\lambda = 1$. Our theory from Theorem 8.3.1 is slightly more conservative, guaranteeing convergence whenever $\lambda \leq 0.5 = 2\min\{1, (\eta/\rho - 1)^2\}/(\eta/\alpha + 1)$.

8.5.2 A Lyapunov for Interaction Moderate Problems

The standard analysis of gradient descent on nonconvex optimization relies on the fact that the function value monotonically decays every iteration. However, such properties fail to hold in the nonconvex-nonconcave minimax setting: the objective is neither monotonically decreasing nor increasing while PPM runs. Worse yet, since we know the proximal point method may cycle indefinitely with gradients bounded away from zero (for example, recall the interaction moderate regime trajectories in Figure 8.1), no "Lyapunov"-type quantity can monotonically decrease along the iterates of the proximal point method.

In order to obtain a similar analysis as the standard nonconvex optimization approach, we propose to study the following "Lyapunov" function, which captures the difference between smoothing over y and smoothing over x using the classic Moreau envelope,

$$\mathcal{L}(x, y) := - e_\eta\{-L(x, \cdot)\}(y) - e_\eta\{L(\cdot, y)\}(x) \, . \tag{8.28}$$

The following proposition establishes structural properties supporting our consideration of $\mathcal{L}(x, y)$.

Theorem 8.5.1. *The Lyapnuov $\mathcal{L}(x, y)$ has the following structural properties:*

1. $\mathcal{L}(x, y) \geq 0$,
2. *When $\eta > \rho$, $\mathcal{L}(x, y) = 0$ if and only if (x, y) is a stationary point to $L(x, y)$,*
3. *When $\eta = 0$, $\mathcal{L}(x, y)$ recovers the well-known primal-dual gap of $L(x, y)$*

$$\mathcal{L}(x, y) = \max_v L(x, v) - \min_u L(u, y).$$

Proof. Recall that a Moreau envelope $e_\eta\{f(\cdot)\}(x)$ provides a lower bound (8.9) on f everywhere. Hence $e_\eta\{-L(x, \cdot)\}(y) \leq -L(x, y)$ and $e_\eta\{L(\cdot, y)\}(x) \leq L(x, y)$, and so our proposed Lyapunov is always nonnegative since

$$\mathcal{L}(x, y) = -e_\eta\{-L(x, \cdot)\}(y) - e_\eta\{L(\cdot, y)\}(x) \geq L(x, y) - L(x, y) = 0\ .$$

Further, it follows from (8.11) that for any ρ-weakly convex function f, selecting $\eta > \rho$ ensures the Moreau envelope equals the given function precisely at its stationary point. Then the preceding nonnegativity argument holds with equality if and only if

$$\nabla_y - L(x, \cdot)(y) = 0 \quad \text{and} \quad \nabla_x L(\cdot, y)(x) = 0\ .$$

Hence we have $\mathcal{L}(x, y) = 0 \iff \nabla L(x, y) = 0$. Lastly, when $\eta = 0$, we have

$$\begin{aligned}\mathcal{L}(x, y) &= -\min_v \Big\{-L(x, v) + \frac{\eta}{2}\|v - y\|^2\Big\} - \min_u \Big\{-L(u, y) + \frac{\eta}{2}\|u - x\|^2\Big\} \\ &= \max_v L(x, v) - \min_u L(u, y)\ ,\end{aligned}$$

recovering the primal-dual gap for $L(x, y)$. □

For example, computing the Moreau envelopes defining $\mathcal{L}(z)$ for (8.26) gives

$$e_\eta\{L(\cdot, y)\}(x) = \frac{1}{2}(\eta^{-1} - \rho^{-1})^{-1}\|x\|^2 + \frac{\eta a}{\eta - \rho}x^T y - \frac{\alpha}{2}\|y\|^2 \tag{8.29}$$

$$e_\eta\{-L(x, \cdot)\}(y) = -\frac{\alpha}{2}\|x\|^2 - \frac{\eta a}{\eta - \rho}x^T y + \frac{1}{2}(\eta^{-1} - \rho^{-1})^{-1}\|y\|^2 \tag{8.30}$$

where $\alpha = -\rho + a^2/(\eta - \rho)$ is this problem's interaction dominance. Hence

$$\mathcal{L}(z) = \frac{1}{2}\big(\alpha - (\eta^{-1} - \rho^{-1})^{-1}\big)\|z\|^2 .$$

Noting that $\alpha \geq -\rho$ and $-(\eta^{-1} - \rho^{-1})^{-1} > -\rho$, we see that the origin is the unique minimizer of $\mathcal{L}(z)$ and consequently the unique stationary point of L. In this case, minimizing $\mathcal{L}(z)$ is simple convex optimization.

Future works could identify further tractable nonconvex-nonconcave problem settings where algorithms can minimize $\mathcal{L}(x, y)$ instead as all of its global minimums are stationary points of the original objective. Since this problem is purely one of minimization, cycling can be ruled out directly. As previously observed, the proximal point method is not such an algorithm since it may fall into a cycle and fail to monotonically decrease $\mathcal{L}(z)$. Instead, we find the following weakened descent condition for $\mathcal{L}(z)$, relating its change to our α-interaction dominance conditions. Note that this result holds regardless of whether the interaction dominance parameter α is positive or negative.

Theorem 8.5.2. *For any ρ-weakly convex-weakly concave, $\alpha \in \mathbb{R}$-interaction dominant in x and y problem, any $z \in \mathbb{R}^n \times \mathbb{R}^m$ has $z_+ = \operatorname{prox}_{,\eta}(z)$ satisfy*

$$\mathcal{L}(z_+) \leq \mathcal{L}(z) - \frac{1}{2}\big(\alpha + (\eta^{-1} - \rho^{-1})^{-1}\big)\|z_+ - z\|^2 .$$

Remark 8.5.3. *This upper bound is attained by our example diverging problem* (8.26)*. This is example attains our bound since the proof of Theorem 8.5.2 only introduces*

inequalities by using the following four Hessian bounds for every (u, v)

$$\nabla^2_{xx} - e_\eta\{-L(u,\cdot)\}(v) \succeq \alpha I \ , \ \nabla^2_{yy} - e_\eta\{-L(u,\cdot)\}(v) \preceq -(\eta^{-1} - \rho^{-1})^{-1} I \ ,$$
$$\nabla^2_{yy} - e_\eta\{L(\cdot,v)\}(u) \succeq \alpha I \ , \ \nabla^2_{xx} - e_\eta\{L(\cdot,v)\}(u) \preceq -(\eta^{-1} - \rho^{-1})^{-1} I \ .$$

Observing that all four of these bounds hold with equality everywhere in (8.29) *and* (8.30) *shows our recurrence holds with equality.*

Remark 8.5.4. *For generic minimax problems, Theorem 8.5.2 bounds how quickly PPM can diverge. For any objective* L *that is* l*-Lipschitz and nearly convex-concave, satisfying weak convexity-weak concavity* (8.5) *with some* $\rho = \epsilon$*. Then since* $\alpha \geq -\rho = -\epsilon$*, the Lyapanov increases by at most* $O(\epsilon)$ *as*

$$\mathcal{L}(z_+) - \mathcal{L}(z) \leq -\frac{1}{2}\left(\alpha - \frac{\eta\rho}{\eta - \rho}\right)\|\nabla L(z_+)/\eta\|^2 \leq \frac{\epsilon l^2}{2\eta^2}\left(1 + \frac{\eta}{\eta - \epsilon}\right) \approx \frac{\epsilon l^2}{\eta^2} \ .$$

8.5.3 Proof of Theorem 8.5.2

First, we bound the Hessians of the functions defining our Lyapunov $\mathcal{L}(z)$.

Lemma 8.5.5. *If the* x*-interaction dominance* (8.15) *holds with* $\alpha \in \mathbb{R}$*, the function* $e_\eta\{L(\cdot, y)\}(x)$ *has Hessians in* x *and* y *bounded by*

$$(\eta^{-1} - \rho^{-1})^{-1} I \preceq \nabla^2_{xx} e_\eta\{L(\cdot,y)\}(x) \preceq \eta I \ \text{ and } \ \nabla^2_{yy} e_\eta\{L(\cdot,y)\}(x) \preceq -\alpha I \ .$$

Symmetrically, if the y*-interaction dominance* (8.16) *holds with* $\alpha \in \mathbb{R}$*,*

$$\nabla^2_{xx} e_\eta\{-L(x,\cdot)\}(y) \preceq -\alpha I \ \text{ and } \ (\eta^{-1} - \rho^{-1})^{-1} I \preceq \nabla^2_{yy} e_\eta\{-L(x,\cdot)\}(y) \preceq \eta I.$$

Proof. For the Hessian bound in the x variable, this follows directly from the Moreau envelope Hessian bounds (8.13). Considering $e_\eta\{L(\cdot, y)\}(x)$ as a function of y, we find that its gradient is given by $\nabla_y e_\eta\{L(\cdot,y)\}(x) = \nabla_y L(x_+, y)$ and

Hessian $\nabla^2_{yy} e_\eta\{L(\cdot, y)\}(x)$ is given by

$$\nabla^2_{yy} L(x_+, y) - \nabla^2_{yx} L(x_+, y)(\eta I + \nabla^2_{xx} L(x_+, y))^{-1} \nabla^2_{xy} L(x_+, y)$$

where $x_+ = \operatorname{argmin}_u L(u, y) + \frac{\eta}{2}\|u - x\|^2$. Noting that this Hessian matches the α-interaction dominance condition (8.16) gives our bound on $-\nabla^2_{yy} e_\eta\{L(\cdot, y)\}(x)$.

All that remains is to derive our claimed gradient and Hessian formulas in y. Consider a nearby point $y^\Delta = y + \Delta$ and denote $x^\Delta_+ = \operatorname{argmin}_u L(u, y^\Delta) + \frac{\eta}{2}\|u - x\|^2$. Consider the second-order Taylor model of the objective L around (x_+, y) denoted by $\widetilde{L}(u, v)$ with value

$$\begin{aligned} L(x_+, y) + &\begin{bmatrix} \nabla_x L(x_+, y) \\ \nabla_y L(x_+, y) \end{bmatrix}^T \begin{bmatrix} u - x_+ \\ v - y \end{bmatrix} \\ &+ \frac{1}{2} \begin{bmatrix} u - x_+ \\ v - y \end{bmatrix}^T \begin{bmatrix} \nabla^2_{xx} L(x_+, y) & \nabla^2_{xy} L(x_+, y) \\ \nabla^2_{yx} L(x_+, y) & \nabla^2_{yy} L(x_+, y) \end{bmatrix} \begin{bmatrix} u - x_+ \\ v - y \end{bmatrix}. \end{aligned}$$

Denote the $\widetilde{x}^\Delta_+ = \operatorname{argmin}_u \widetilde{L}(u, y^\Delta) + \frac{\eta}{2}\|u - x\|^2$. Noting this point is uniquely defined by its first-order optimality conditions, we have

$$\begin{aligned} &\nabla_x L(x_+, y) + \nabla^2_{xx} L(x_+, y)(\widetilde{x}^\Delta_+ - x_+) + \nabla^2_{xy} L(x_+, y)\Delta + \eta(\widetilde{x}^\Delta_+ - x) = 0 \,, \\ &\implies (\eta I + \nabla^2_{xx} L(x_+, y))(\widetilde{x}^\Delta_+ - x_+) = -\nabla^2_{xy} L(x_+, y)\Delta \,, \\ &\implies \widetilde{x}^\Delta_+ - x_+ = -(\eta I + \nabla^2_{xx} L(x_+, y))^{-1} \nabla^2_{xy} L(x_+, y)\Delta \,. \end{aligned}$$

Denote the proximal subproblem objective by $M^\Delta(u, v) = L(u, y^\Delta) + \frac{\eta}{2}\|u - x\|^2$ and its approximation by $\widetilde{M}^\Delta(u, v) = \widetilde{L}(u, y^\Delta) + \frac{\eta}{2}\|u - x\|^2$. Noting that $\|\nabla_x \widetilde{M}^\Delta(x_+, y^\Delta)\| = \|\nabla^2_{xy} L(x_+, y)\Delta\|$, the $(\eta - \rho)$-strongly convexity of $\widetilde{M}^\Delta$ bounds the distance to its minimizer by

$$\|x_+ - \widetilde{x}^\Delta_+\| \le \frac{\|\nabla^2_{xy} L(x_+, y)\Delta\|}{\eta - \rho} = O(\|\Delta\|) \,.$$

Consequently, we can bound difference in gradients between L and its model $\widetilde{L}$ at $\widetilde{x}_+^\Delta$ by $\|\nabla L(\widetilde{x}_+^\Delta, y^\Delta) - \nabla \widetilde{L}(\widetilde{x}_+^\Delta, y^\Delta)\| = o(\|\Delta\|)$. Therefore $\|\nabla M^\Delta(\widetilde{x}_+^\Delta, y^\Delta)\| = o(\|\Delta\|)$. Then using the strong convexity of M^Δ with this gradient bound, we conclude the distance from $\widetilde{x}_+^\Delta$ to the minimizer x_+^Δ is bounded by $\|\widetilde{x}_+^\Delta - x_+^\Delta\| = o(\|\Delta\|)$. Then our claimed gradient formula follows as

$$\begin{aligned}
&e_\eta\{L(\cdot, y^\Delta)\}(x) - e_\eta\{L(\cdot, y)\}(x) \\
&\quad = L(x_+^\Delta, y^\Delta) + \frac{\eta}{2}\|x_+^\Delta - x\|^2 - L(x_+, y) - \frac{\eta}{2}\|x_+ - x\|^2 \\
&\quad = \begin{bmatrix} \nabla_x L(x_+, y) + \eta(x_+ - x) \\ \nabla_y L(x_+, y) \end{bmatrix}^T \begin{bmatrix} x_+^\Delta - x_+ \\ \Delta \end{bmatrix} + o(\|\Delta\|) \\
&\quad = \nabla_y L(x_+, y)^T \Delta + o(\|\Delta\|) \, .
\end{aligned}$$

Moreover, our claimed Hessian formula follows as

$$\begin{aligned}
&\nabla_y e_\eta\{L(\cdot, y^\Delta)\}(x) - \nabla_y e_\eta\{L(\cdot, y)\}(x) \\
&\quad = \nabla_y L(x_+^\Delta, y^\Delta) - \nabla_y L(x_+, y) \\
&\quad = \nabla_y \widetilde{L}(\widetilde{x}_+^\Delta, y^\Delta) - \nabla_y L(x_+, y) + o(\|\Delta\|) \\
&\quad = \begin{bmatrix} \nabla_{xy}^2 L(x_+, y) \\ \nabla_{yy}^2 L(x_+, y) \end{bmatrix}^T \begin{bmatrix} -(\eta I + \nabla_{xx}^2 L(x_+, y))^{-1} \nabla_{xy}^2 L(x_+, y)\Delta \\ \Delta \end{bmatrix} + o(\|\Delta\|) \, .
\end{aligned}$$

□

Notice that $-e_\eta\{-L(u, \cdot)\}(y)$ has gradient at x_+ of $\nabla_x - e_\eta\{-L(x_+, \cdot)\}(y) = \nabla_x L(z_+) = \eta(x - x_+)$ and from Lemma 8.5.5 that its Hessian in x is uniformly lower bounded by αI. As a result, we have the following decrease in $-e_\eta\{-L(u, \cdot)\}(y)$ when moving from x to x_+

$$\begin{aligned}
-e_\eta\{-L(x_+, \cdot)\}(y) &\leq -e_\eta\{-L(x, \cdot)\}(y) + \nabla_x L(z_+)^T(x_+ - x) - \frac{\alpha}{2}\|x_+ - x\|^2 \\
&= -e_\eta\{-L(x, \cdot)\}(y) - \Big(\eta + \frac{\alpha}{2}\Big)\|x_+ - x\|^2 \, .
\end{aligned}$$

From the gradient formula (8.11), we know that $\nabla_y - e_\eta\{-L(x_+,\cdot)\}(y) = \nabla_y L(z_+) = \eta(y_+ - y)$ and from Lemma 8.5.5 that its Hessian in y is uniformly bounded above by $-(\eta^{-1} - \rho^{-1})^{-1} I$. Then we can upper bound the change in $-e_\eta\{-L(x_+,\cdot)\}(v)$ when moving from y to y_+ as

$$\begin{aligned}
&- e_\eta\{-L(x_+,\cdot)\}(y_+) + e_\eta\{-L(x_+,\cdot)\}(y) \\
&\quad \le \nabla_y L(z_+)^T (y_+ - y) + \frac{-(\eta^{-1} - \rho^{-1})^{-1}}{2} \|y_+ - y\|^2 \\
&\quad = \left(\eta - \frac{(\eta^{-1} - \rho^{-1})^{-1}}{2}\right) \|y_+ - y\|^2 .
\end{aligned}$$

Summing these two inequalities yields

$$\begin{aligned}
&- e_\eta\{-L(x_+,\cdot)\}(y_+) + e_\eta\{-L(x,\cdot)\}(y) \\
&\quad \le \left(\eta - \frac{(\eta^{-1} - \rho^{-1})^{-1}}{2}\right) \|y_+ - y\|^2 - \left(\eta + \frac{\alpha}{2}\right) \|x_+ - x\|^2 .
\end{aligned}$$

Symmetrically, the change in $-e_\eta\{L(\cdot, y)\}(x)$ from z to z_+ is

$$\begin{aligned}
&- e_\eta\{L(\cdot, y_+)\}(x_+) + e_\eta\{L(\cdot, y)\}(x) \\
&\quad \le \left(\eta - \frac{(\eta^{-1} - \rho^{-1})^{-1}}{2}\right) \|x_+ - x\|^2 - \left(\eta + \frac{\alpha}{2}\right) \|y_+ - y\|^2 .
\end{aligned}$$

Summing these two symmetric results gives the claimed bound.

8.6 Addendum - Deferred Figures and Proofs

8.6.1 Sample Paths From Other First-Order Methods

Figure 8.2 plots the solution paths of four common first-order methods for minimax problem for solving a two-dimensional nonconvex-nonconcave minimax

problem:

$$\min_x \max_y L(x,y) = (x+3)(x+1)(x-1)(x-3) + Axy - (y+3)(y+1)(y-1)(y-3), \tag{8.31}$$

with four different levels of interaction term, $A = 1, 10, 100, 1000$. This problem is globally $\rho = 20$-weakly convex and $\beta = 172$-smooth on the box $[-4,4]\times[-4,4]$.

Each plot in Figure 8.2 shows the sample paths generated by running 100 iterations of the given method from the twelve different initial solutions around the boundary of the plot $(4,0)$, $(0,4)$, $(-4,0)$, $(0,-4)$, $(4,2)$, $(2,4)$, $(4,-2)$, $(2,-4)$, $(-4,2)$, $(-2,4)$, $(-4,-2)$, $(-2,-4)$ and four initial solutions towards the center of the plot $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$.

Plots (a)-(d) show the behavior of the Proximal Point Method (PPM) (8.7) with $\eta = 2\rho = 40$ and $\lambda = 1$. These figures match the landscape described by our theory: $A = 1$ is small enough to have local convergence to four different stationary points (each around $\{\pm 2\} \times \{\pm 2\}$), $A = 10$ has moderate size and every sample path is attracted into a limit cycle, and finally $A = 100$ and $A = 1000$ give a large enough interaction term to create a globally attractive stationary point (moreover, comparing plots (c) and (d) shows as A becomes larger the rate of convergence increases).

Plots (e)-(h) show the behavior of the Extragradient Method (EG), which is defined by

$$\begin{aligned} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} &= \begin{bmatrix} x_k \\ y_k \end{bmatrix} + s \begin{bmatrix} -\nabla_x L(x_k, y_k) \\ \nabla_y L(x_k, y_k) \end{bmatrix} \\ \begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} &= \begin{bmatrix} x_k \\ y_k \end{bmatrix} + s \begin{bmatrix} -\nabla_x L(\tilde{x}, \tilde{y}) \\ \nabla_y L(\tilde{x}, \tilde{y}) \end{bmatrix} \end{aligned} \tag{8.32}$$

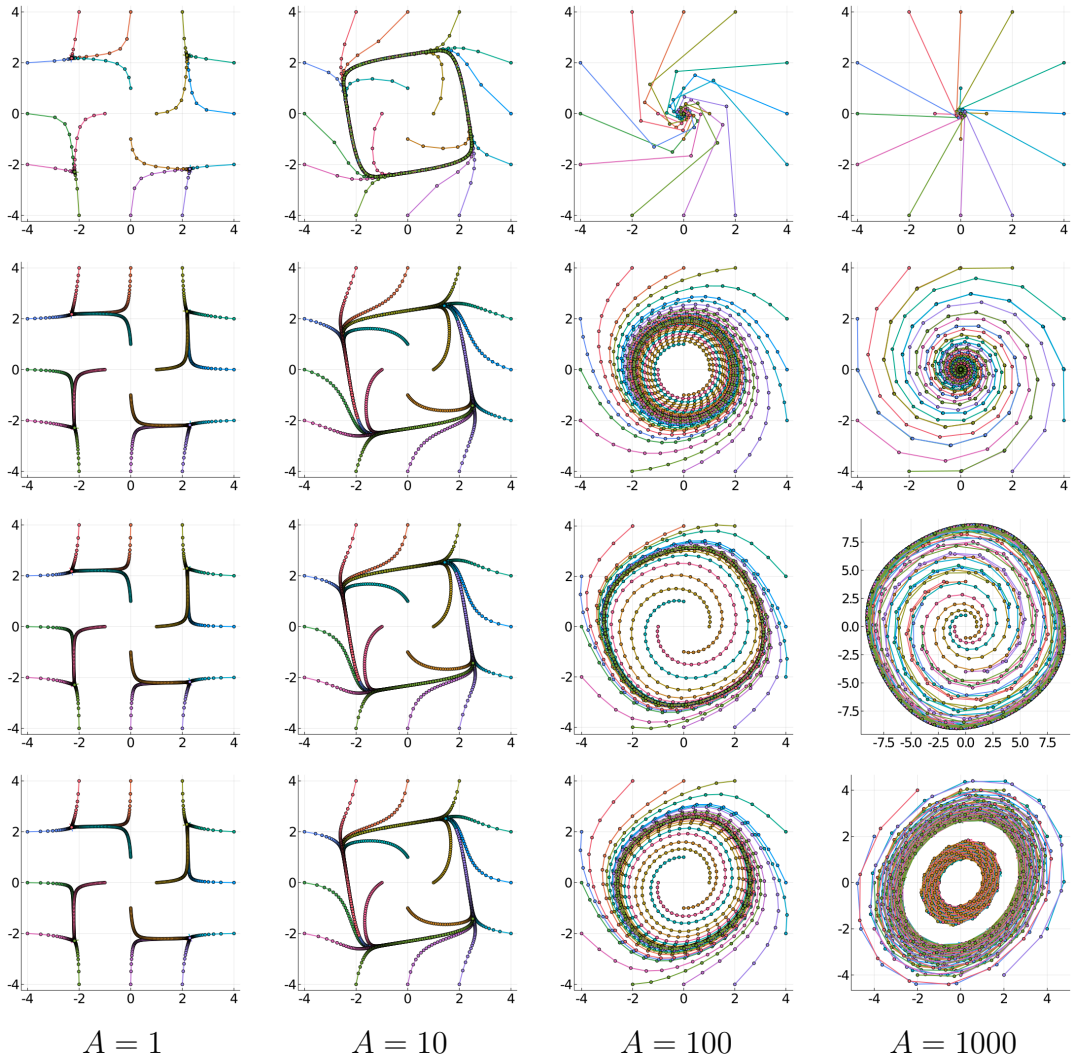


Figure 8.2: Sample paths of PPM, EGM, GDA, and AGDA extending Figure 8.1.

with stepsize chosen as $s = 1/2(\beta + A) = 1/(344 + 2A)$. This stepsize was chosen since the objective function has a $\beta + A$-Lipschitz gradient. These figures show that the extragradient method follows the same general trajectory as described by our theory for the proximal point method. For small $A = 1$, local convergence occurs. For moderate sized $A = 10$ and $A = 100$, the algorithm falls into an attractive limit cycle, never converging. For large enough $A = 1000$, the method globally converges to a stationary point. The extragradient method only differs from the proximal point method's landscape in that it requires a larger A to

transition into the interaction dominant regime.

Plots (i)-(l) show the behavior of Gradient Descent Ascent (GDA) (8.6) with $s = 1/2(\beta + A) = 1/(344 + 2A)$. This method is known to be unstable and diverge even for convex-concave problems. The same behavior carries over to our nonconvex-nonconcave example. For small A, we still see local convergence. However for $A = 10, 100, 1000$, we find that GDA falls into a limit cycle with increasingly large radius as A grows.

Lastly, plots (m)-(p) show the behavior of Alternating Gradient Descent Ascent (AGDA), defined by

$$\begin{cases} x_{k+1} &= x_k - s\nabla_x L(x_k, y_k) \\ y_{k+1} &= y_k + s\nabla_y L(x_{k+1}, y_k) \end{cases} \tag{8.33}$$

with $s = 1/2(\beta + A) = 1/(344 + 2A)$. Again for small A, we still see local convergence, but for larger $A = 10, 100, 1000$, AGDA always falls into a limit cycle.

8.6.2 Convex-Concave Optimization Analysis

Proof of Lemma 8.1.1

Observe that

$$\begin{aligned} M(x', y') &\leq M(x, y') - \nabla_x M(x', y')^T (x - x') - \frac{\mu}{2}\|x - x'\|^2 \\ &\leq M(x, y) + \nabla_y M(x, y)^T (y' - y) - \nabla_x M(x', y')^T (x - x') \\ &\quad - \frac{\mu}{2}\|y - y'\|^2 - \frac{\mu}{2}\|x - x'\|^2 \end{aligned}$$

where the first inequality uses strong convexity of M in x and the second uses strong concavity in y. Symmetrically,

$$\begin{aligned} M(x', y') &\geq M(x', y) - \nabla_y M(x', y')^T (y - y') + \frac{\mu}{2}\|y - y^*\|^2 \\ &\geq M(x, y) + \nabla_x M(x, y)^T (x' - x) - \nabla_y M(x', y')^T (y - y') \\ &\quad + \frac{\mu}{2}\|x - x'\|^2 + \frac{\mu}{2}\|y - y'\|^2. \end{aligned}$$

Combining the above two inequalities gives the first claimed inequality

$$\mu \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|^2 \leq \left(\begin{bmatrix} \nabla_x M(x, y) \\ -\nabla_y M(x, y) \end{bmatrix} - \begin{bmatrix} \nabla_x M(x', y') \\ -\nabla_y M(x', y') \end{bmatrix} \right)^T \begin{bmatrix} x - x' \\ y - y' \end{bmatrix}.$$

Furthermore, when $\nabla M(x', y') = 0$, we have

$$\mu \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|^2 \leq \|\nabla M(x, y)\| \left\| \begin{bmatrix} x - x' \\ y - y' \end{bmatrix} \right\|,$$

which finishes the proof of the second inequality.

Proof of Theorem 8.1.2

First we use Lemma 8.1.1 to conclude that if the set S is large enough, M must have a stationary point in S. Now define $B(z, r) = \{z' | \|z - z'\| \leq r\}$ as the closed Euclidean ball centered as a with radius r.

Lemma 8.6.1. *Suppose M is μ-strongly convex-strongly concave in a set $B(x, r) \times B(y, r)$ for some fixed (x, y) and $r \geq 2\|\nabla M(x, y)\|/\mu$, then there exists a stationary point of M in $B((x, y), r/2)$.*

Proof. Consider the constrained problem $\min_{x' \in B(x,r)} \max_{y' \in B(y,r)} M(x, y)$. Since $M(x, y)$ is strongly convex-strongly concave, it must have a unique solution

(x^*, y^*). The first-order optimality condition for (x^*, y^*) ensures

$$\nabla_x M(x^*, y^*) = -\lambda(x^* - x) \quad \text{and} \quad -\nabla_y M(x^*, y^*) = -\gamma(y^* - y)$$

for some constants $\lambda, \gamma \geq 0$ that are nonzero only if x^* or y^* are on the boundary of $B(x, r)$ and $B(y, r)$ respectively. Taking an inner product with $(x^* - x, y^* - y)$ gives

$$\begin{bmatrix} \nabla_x M(x^*, y^*) \\ -\nabla_y M(x^*, y^*) \end{bmatrix}^T \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} = -\left\| \begin{bmatrix} \sqrt{\lambda}(x^* - x) \\ \sqrt{\gamma}(y^* - y) \end{bmatrix} \right\|^2 \leq 0. \tag{8.34}$$

Applying Lemma 8.1.1 and utilizing (8.34), we conclude that

$$\mu \left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\|^2 + \begin{bmatrix} \nabla_x M(x, y) \\ -\nabla_y M(x, y) \end{bmatrix}^T \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \leq 0. \tag{8.35}$$

Hence

$$\left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\|^2 \leq \frac{1}{\mu} \left\| \begin{bmatrix} \nabla_x M(x, y) \\ -\nabla_y M(x, y) \end{bmatrix} \right\| \left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\|,$$

whereby

$$\left\| \begin{bmatrix} x^* - x \\ y^* - y \end{bmatrix} \right\| \leq \frac{1}{\mu} \left\| \begin{bmatrix} \nabla_x M(x, y) \\ -\nabla_y M(x, y) \end{bmatrix} \right\| < r/2 \,,$$

where the last inequality utilize the condition on r. Since (x^*, y^*) lies strictly inside the ball $B((x, y), r/2)$, the first-order optimality condition implies (x^*, y^*) is a stationary point of M. □

Lemma 8.6.1 ensures the existence of a nearby stationary point (x^*, y^*). Then the standard proof of strongly monotone (from Lemma 8.1.1) and Lipschitz op-

erators gives a contraction whenever $s \in (0, 2\mu/\beta^2)$:

$$\begin{aligned}
\left\| \begin{bmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{bmatrix} \right\|^2 &= \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 - 2s \begin{bmatrix} \nabla_x M(x_k, y_k) \\ -\nabla_y M(x_k, y_k) \end{bmatrix}^T \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \\
&\quad + s^2 \left\| \begin{bmatrix} \nabla_x M(x_k, y_k) \\ -\nabla_y M(x_k, y_k) \end{bmatrix} \right\|^2 \\
&\leq \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 - 2\mu s \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 + \beta^2 s^2 \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2 \\
&= \left(1 - 2\mu s + \beta^2 s^2\right) \left\| \begin{bmatrix} x_k - x^* \\ y_k - y^* \end{bmatrix} \right\|^2,
\end{aligned}$$

where the inequality utilizes (8.35) at $(x, y) = (x_k, y_k)$ and the smoothness of $M(x, y)$.

BIBLIOGRAPHY

[1] Emmanuel Abbe, Afonso S. Bandeira, Annina Bracher, and Amit Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

[2] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016.

[3] Pierre Apkarian, Dominikus Noll, and Olivier Prot. A Proximity Control Algorithm to Minimize Nonsmooth and Nonconvex Semi-infinite Maximum Eigenvalue Functions. *J. Convex Anal.*, 16(3-4):641–666, 2009.

[4] Aleksandr Aravkin and Damek Davis. A smart stochastic algorithm for nonconvex optimization with applications to robust machine learning. *arXiv preprint arXiv:1610.01101*, 2016.

[5] Aleksandr Y Aravkin, James V Burke, Dmitriy Drusvyatskiy, Michael P Friedlander, and Kellie J MacPhee. Foundations of gauge and perspective duality. *SIAM Journal on Optimization*, 28(3):2406–2434, 2018.

[6] Emil Artin. *Geometric Algebra*. Interscience. Wiley, 1988.

[7] Shiri Artstein-Avidan, Dan Florentin, and Vitali Milman. *Order Isomorphisms on Convex Functions in Windows*, pages 61–122. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[8] Shiri Artstein-Avidan and Vitali Milman. Hidden structures in the class of convex functions and a new duality transform. *Journal of the European Mathematical Society*, 013(4):975–1004, 2011.

[9] Hedy Attouch, Dominique Aze, and Roger J.-B. Wets. On continuity properties of the partial legendre-fenchel transform: Convergence of sequences of augmented lagrangian functions, moreau-yosida approximates and subdifferential operators. In J.-B. Hiriart-Urruty, editor, *Fermat Days 85: Mathematics for Optimization*, volume 129 of *North-Holland Mathematics Studies*, pages 1–42. North-Holland, 1986.

[10] Hedy Attouch and Roger J.-B. Wets. A convergence for bivariate functions aimed at the convergence of saddle values. 1983.

[11] Hedy Attouch and Roger J.-B. Wets. A convergence theory for saddle functions. *Transactions of the American Mathematical Society*, 280(1):1–41, 1983.

[12] Dominique Aze. Rate of convergence for the saddle points of convex-concave functions. *International Series of Numerical Mathematic*, 84:1–23, 1988.

[13] Heinz H. Bauschke, Jerome Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[14] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[15] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22:557–580, 2012.

[16] Mario Bertero, Patrizia Boccacci, Gabriele Desiderà, and Giuseppe Vicidomini. Image deblurring with poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, nov 2009.

[17] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[18] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, Oct 2017.

[19] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. p.89.

[20] Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Found. Trends Mach. Learn.*, 8(3-4):231–357, November 2015.

[21] James V Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33(3):260–279, 1985.

[22] James V. Burke. Second order necessary and sufficient conditions for convex composite ndo. *Mathematical Programming*, 38(3):287–302, 1987.

[23] James V. Burke and Michael C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[24] James V. Burke and Michael C. Ferris. A gaussnewton method for convex composite optimization. *Math. Program.*, 71:179194, 1995.

[25] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[26] Karthekeyan Chandrasekaran, Daniel Dadush, and Santosh Vempala. *Thin Partitions: Isoperimetric Inequalities and a Sampling Algorithm for Star Shaped Bodies*, pages 1630–1645.

[27] Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth Analysis and Control Theory*. Springer-Verlag, Berlin, Heidelberg, 1998.

[28] Bo Dai, Albert Shaw, Niao He, Lihong Li, and Le Song. Boosting the actor with dual critic. In *ICLR 2018*.

[29] Aris Daniilidis and Jérôme Malick. Filling the gap between lower-c1 and lower-c2 functions. *Journal of Convex Analysis*, 12(2):315–329, 2005.

[30] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9236–9246. Curran Associates, Inc., 2018.

[31] Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

[32] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[33] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D. Lee. Stochastic subgradient method converges on tame functions. *Foundions of Computational Mathematics*, 20:119–154, 2020.

[34] Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *J Optim Theory Appl*, 179:962982, 2018.

[35] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[36] Damek Davis and Benjamin Grimmer. Proximally Guided Stochastic Subgradient Method for Nonsmooth, Nonconvex Problems. *ArXiv e-prints*, 1707.03505, July 2017.

[37] William S. Dorn. Duality in quadratic programming. *Quarterly of Applied Mathematics*, 18(2):155–162, 1960.

[38] Jim Douglas and Henry H Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American mathematical Society*, 82(2):421–439, 1956.

[39] Dmitriy Drusvyatskiy and Courtney Kempton. An accelerated algorithm for minimizing convex compositions. *arXiv preprint arXiv:1605.00125*, 2016.

[40] Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[41] Yu Du and Andrzej Ruszczyński. Rate of Convergence of the Bundle Method. *J. Optim. Theory Appl.*, 173(3):908–922, June 2017.

[42] John Duchi and Feng Ruan. Stochastic methods for composite optimization problems. *arXiv preprint arXiv:1703.08570*, 2017.

[43] John Duchi and Yoram Singer. Efficient Online and Batch Learning Using Forward Backward Splitting. *J. Mach. Learn. Res.*, 10:2899–2934, December 2009.

[44] Jonathan Eckstein and Dimitri P Bertsekas. On the douglasrachford split-

ting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[45] Yu. M. Ermol'ev and V. I. Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2):196–215, Mar 1998.

[46] Yu. M. Ermoliev and V. I. Norkin. Solution of nonconvex nonsmooth stochastic optimization problems. *Cybernetics and Systems Analysis*, 39(5):701–715, Sep 2003.

[47] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[48] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

[49] Farzan Farnia and Asuman Ozdaglar. Do gans always have nash equilibria? In *International Conference on Machine Learning*, pages 3029–3039. PMLR, 2020.

[50] Michael Ferris. Finite Termination Of The Proximal Point Algorithm. *Math. Program.*, 50:359–366, 03 1991.

[51] M. Fickus, D.G. Mixon, A.A. Nelson, and Y. Wang. Phase retrieval from very few measurements. *Linear Algebra Appl.*, 449:475–499, 2014.

[52] Sjur Didrik Flam. On penalty methods for minimax problems. *Zeitschrift für Operations Research*, 30:A209A222, 1986.

[53] R. Fletcher. *A model algorithm for composite nondifferentiable optimization problems*, pages 67–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 1982.

[54] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(12):95–110, 1956.

[55] Robert M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Math. Program.*, 38:47–67, 1987.

[56] Hong-Ye Gao and Andrew G. Bruce. Wave shrink with firm shrinkage. *Statistica Sinica*, 7(4):855 – 874, 1997.

[57] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[58] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1):267–305, Jan 2016.

[59] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS14, page 26722680, Cambridge, MA, USA, 2014. MIT Press.

[60] Benjamin Grimmer. Radial subgradient method. *SIAM Journal on Optimization*, 28(1):459–469, 2018.

[61] Benjamin Grimmer. General Convergence Rates Follow From Specialized Rates Assuming Growth Bounds. *ArXiv:1905.06275*, May 2019.

[62] Benjamin Grimmer. Radial Duality Part I: Foundations. *arXiv: , 2021*, March 2021.

[63] Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. The landscape of nonconvex-nonconcave minimax optimization. *arXiv preprint arXiv:2006.08667*, 2020.

[64] Benjamin Grimmer, Haihao Lu, Pratik Worah, and Vahab Mirrokni. Limiting behaviors of nonconvex-nonconcave minimax optimization via continuous-time systems. *arXiv preprint arXiv:2010.10628*, 2020.

[65] Jean Guillerme. Convergence of approximate saddle points. *Journal of Mathematical Analysis and Applications*, 137(2):297–311, 1989.

[66] Sergey Guminov and Alexander Gasnikov. Accelerated Methods for α-Weakly-Quasi-Convex Problems. *arXiv e-prints*, page arXiv:1710.00797, October 2017.

[67] Sergey Guminov, Yurii Nesterov, Pavel Dvurechensky, and Alexander

Gasnikov. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. *Dokl. Math.*, 99:125–128, 2019.

[68] David H. Gutman and Javier F. Pe na. Perturbed fenchel duality and first-order methods, 2020.

[69] David H. Gutman and Javier F. Peña. Convergence rates of proximal gradient methods via the convex conjugate. *SIAM Journal on Optimization*, 29(1):162–174, 2019.

[70] W. Hare and C. Sagastizbal. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.

[71] Warren Hare and Claudia Sagastizábal. Computing proximal points of nonconvex functions. *Mathematical Programming*, 116(1):221–258, Jan 2009.

[72] Warren Hare and Claudia Sagastizábal. A Redistributed Proximal Bundle Method for Nonconvex Optimization. *SIAM J. Optim.*, 20(5):2442–2473, 2010.

[73] Warren Hare, Claudia Sagastizábal, and Mikhail Solodov. A Proximal Bundle Method for Nonsmooth Nonconvex Functions with Inexact Information. *Computational Optimization and Applications*, 63(1):1–28, Jan 2016.

[74] Elad Hazan and Satyen Kale. Beyond the Regret Minimization Barrier: Optimal Algorithms for Stochastic Strongly-convex Optimization. *J. Mach. Learn. Res.*, 15(1):2489–2512, January 2014.

[75] Niao He, Zaïd Harchaoui, Yichen Wang, and Le Song. Fast and simple optimization for poisson likelihood models. *CoRR*, abs/1608.01264, 2016.

[76] C. Helmberg and F. Rendl. A spectral bundle method for semidefinite programming. *SIAM Journal on Optimization*, 10(3):673–696, 2000.

[77] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 09–12 Jul 2020.

[78] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Acceleration of the Cutting-Plane Algorithm: Primal Forms of Bundle Methods*, pages 275–330. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

[79] Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: convergence to spurious non-critical sets. *arXiv preprint arXiv:2006.09065*, 2020.

[80] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

[81] Patrick R. Johnstone and Pierre Moulin. Faster Subgradient Methods for Functions with Hölderian Growth. *ArXiv e-prints*, April 2017.

[82] Krzysztof C. Kiwiel. An Aggregate Subgradient Method for Nonsmooth Convex Minimization. *Math. Program.*, 27(3):320–341, October 1983.

[83] Krzysztof C. Kiwiel. A Linearization Algorithm for Nonsmooth Minimization. *Mathematics of Operations Research*, 10(2):185–194, 1985.

[84] Krzysztof C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization.* Springer, Berlin, 1985.

[85] Krzysztof C. Kiwiel. Proximal Level Bundle Methods for Convex Nondifferentiable Optimization, Saddle-point Problems and Variational Inequalities. *Math. Program.*, 69(1-3):89–109, July 1995.

[86] Krzysztof C. Kiwiel. Efficiency of Proximal Bundle Methods. *Journal of Optimization Theory and Applications*, 104(3):589–603, Mar 2000.

[87] Willem K. Klein Haneveld, Maarten H. van der Vlerk, and Ward Romeijnders. *Chance Constraints*, pages 115–138. Springer International Publishing, Cham, 2020.

[88] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[89] Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[90] Guanghui Lan. Bundle-Level Type Methods Uniformly Optimal For Smooth And Nonsmooth Convex Optimization. *Mathematical Programming*, 149(1):1–45, Feb 2015.

[91] Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

[92] Jasper C.H. Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614, 2016.

[93] Claude Lemarechal. *An Extension of Davidon Methods to Nondifferentiable Problems*, pages 95–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975.

[94] Claude Lemaréchal. *Lagrangian Relaxation*, pages 112–156. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[95] Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New Variants of Bundle Methods. *Math. Program.*, 69(1-3):111–147, July 1995.

[96] Alistair Letcher. On the impossibility of global convergence in multi-loss optimization. *arXiv preprint arXiv:2005.12649*, 2020.

[97] Adrian Lewis and Stephen Wright. A proximal method for composite minimization. *Math. Program.*, 158:501546, 2016.

[98] Xiangru Lian, Mengdi Wang, and Ji Liu. Finite-sum Composition Optimization via Variance Reduced Gradient Descent. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1159–1167. PMLR, 20–22 Apr 2017.

[99] Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *arXiv preprint arXiv:1810.10207*, 2018.

[100] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

[101] Tianyi Lin, Chi Jin, and Michael I. Jordan. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020.

[102] Mingrui Liu and Tianbao Yang. Adaptive accelerated gradient converging method under holderian error bound condition. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[103] Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

[104] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[105] Haihao Lu. "Relative-Continuity" for Non-Lipschitz Non-Smooth Convex Optimization using Stochastic (or Deterministic) Mirror Descent. *ArXiv e-prints*, 1710.04718, October 2017.

[106] Haihao Lu. An $o(s^r)$-resolution ode framework for discrete-time optimization algorithms and applications to convex-concave saddle-point problems. *arXiv preprint arXiv:2001.08826*, 2020.

[107] Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

[108] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, Mar 1993.

[109] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[110] Matt Menickelly and Stefan M Wild. Robust learning of trimmed estimators via manifold sampling. *arXiv preprint arXiv:1807.02736*, 2018.

[111] Robert Mifflin. *A Modification and an Extension of Lemarechal's Algorithm for Nonsmooth Minimization*, pages 77–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 1982.

[112] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified anal-

ysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

[113] Jean Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 255:2897–2899, 1962.

[114] K. Mouallif. Variational convergence and perturbed proximal method for saddle point problems. In Szymon Dolecki, editor, *Optimization*, pages 115–140, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg.

[115] Mahesh Chandra Mukkamala, Jalal Fadili, and Peter Ochs. Global convergence of model function based bregman proximal minimization algorithms, arXiv:2012.13161, 2020.

[116] Javier Pe na. Convergence of first-order methods via the convex conjugate. *Operations Research Letters*, 45(6):561–564, 2017.

[117] Ion Necoara, Yurii Nesterov, and Francois Glineur. Linear Convergence Of First Order Methods For Non-strongly Convex Optimization. *Mathematical Programming*, 175(1):69–107, May 2019.

[118] Angelia Nedić and Soomin Lee. On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization*, 24(1):84–107, 2014.

[119] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[120] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[121] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

[122] Arkadii Nemirovskii and Yurii Nesterov. Optimal Methods of Smooth Convex Minimization. *USSR Comput. Math. Math. Phys.*, 25(3-4):21–30, July 1986.

[123] A.S. Nemirovsky and D.B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983.

[124] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[125] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[126] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2004.

[127] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127152, May 2005.

[128] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Math. Program.*, 152(1-2):381–404, August 2015.

[129] Yurii Nesterov and Mihai I. Florea. Gradient methods with memory. *Optimization Methods and Software*, 0(0):1–18, 2021.

[130] Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108:177–205, 08 2006.

[131] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems 32*, pages 14934–14942. Curran Associates, Inc., 2019.

[132] E. A. Nurminski. *On ε-subgradient methods of non-differentiable optimization*, pages 187–195. Springer Berlin Heidelberg, Berlin, Heidelberg, 1979.

[133] Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for gradient-based non-convex optimization. *arXiv preprint arXiv:1703.10993*, 2017.

[134] Neal Parikh and Stephen Boyd. Proximal Algorithms. *Found. Trends Optim.*, 1(3):127–239, January 2014.

[135] Boris T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[136] Boris T. Polyak. Sharp minima. *Institute of Control Sciences Lecture Notes,Moscow, USSR. Presented at the IIASA Workshop on Generalized Lagrangians and Their Applications, IIASA, Laxenburg, Austria.*, 1979.

[137] Viktor Prasolov and Vladimir Tikhomirov. *Geometry*. Providence, R.I. : American Mathematical Society ; Oxford University Press, 2001.

[138] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[139] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1571–1578, USA, 2012. Omnipress.

[140] James Renegar. "Efficient" Subgradient Methods for General Convex Optimization. *SIAM Journal on Optimization*, 26(4):2649–2676, 2016.

[141] James Renegar. Accelerated first-order methods for hyperbolic programming. *Math. Program.*, 173(1-2):1–35, 2019.

[142] James Renegar and Benjamin Grimmer. A Simple Nearly-Optimal Restart Scheme For Speeding-Up First Order Methods. *To appear in Foundations of Computational Mathematics*, 2021.

[143] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[144] R. Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976.

[145] R. Tyrrell Rockafellar. Maximal monotone relations and the second derivatives of nonsmooth functions. *Annales de l'I.H.P. Analyse non linéaire*, 2(3):167–184, 1985.

[146] R. Tyrrell Rockafellar. Generalized second derivatives of convex functions and saddle functions. *Transactions of the American Mathematical Society*, 322(1):51–77, 1990.

[147] Ralph Tyrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[148] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.

[149] Ralph Tyrell Rockafellar and Roger J B Wets. *Variational Analysis*, volume 317. Springer, 1998.

[150] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

[151] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[152] Peter J Rousseeuw. Multivariate Estimation with High Breakdown Point. *Mathematical statistics and applications*, 8:283–297, 1985.

[153] A.M. Rubinov and A.A. Yagubov. The space of star-shaped sets and its applications in nonsmooth optimization. *Mathematical Programming Studies*, 29, 1986.

[154] Andrzej Ruszczynski. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

[155] Andrzej Ruszczynski. *Nonlinear Optimization*. Princeton University Press, Princeton, NJ, USA, 2006.

[156] Claudia Sagastizábal. Divide to Conquer: Decomposition Methods for Energy Optimization. *Mathematical Programming*, 134(1):187–222, Aug 2012.

[157] Claudia Sagastizábal and Mikhail Solodov. An Infeasible Bundle Method for Nonsmooth Convex Constrained Optimization without a Penalty Function or a Filter. *SIAM Journal on Optimization*, 16(1):146–169, 2005.

[158] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal Estimated Sub-gradient Solver for SVM. *Mathematical Programming*, 127(1):3–30, Mar 2011.

[159] Bernard Shiffman. Synthetic Projective Geometry and Poincaré's Theorem

on Automorphisms of the Ball. *L'Enseignement Mathmatique.*, 41:201–215, 1995.

[160] Naum Zuselevich Shor. *Minimization Methods for Non-Differentiable Functions*, page 23. Springer Berlin Heidelberg, Berlin, Heidelberg, 1985.

[161] Naum Zuselevich Shor. *Subgradient and ϵ-Subgradient Methods*, pages 35–70. Springer US, Boston, MA, 1998.

[162] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[163] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[164] Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems 32*, pages 12680–12691. Curran Associates, Inc., 2019.

[165] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[166] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Tech. report*, 2008.

[167] Mengdi Wang, Ethan X. Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1):419–449, Jan 2017.

[168] Mengdi Wang, Ji Liu, and Ethan Fang. Accelerating stochastic composition optimization. In *Advances in Neural Information Processing Systems 29*, pages 1714–1722. 2016.

[169] Fei Wen, Lei Chu, Peilin Liu, and Robert C. Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

[170] Philip Wolfe. *A Method of Conjugate Subgradients for Minimizing Nondif-*

ferentiable Functions, pages 145–173. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975.

[171] Yangyang Xu and Wotao Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *SIAM Journal on Optimization*, 25(3):1686–1716, 2015.

[172] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

[173] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6):1–33, 2018.

[174] Jin Yu, Anders Eriksson, Tat-Jun Chin, and David Suter. An adversarial optimization approach to efficient outlier removal. *J Math Imaging Vis*, 48:451466, 2014.

[175] Jin Yu, S.V.N. Vishwanathan, Simon Günter, and Nicol N. Schraudolph. A Quasi-Newton Approach to Nonsmooth Convex Optimization Problems in Machine Learning. *J. Mach. Learn. Res.*, 11:1145–1200, March 2010.

[176] Yuan Yuan, Zukui Li, and Biao Huang. Robust optimization approximation for joint chance constrained optimization problem. *J Glob Optim*, 67:805–827, 2017.

[177] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.

[178] Guojun Zhang, Pascal Poupart, and Yaoliang Yu. Optimality and stability in non-convex-non-concave min-max optimization. *arXiv preprint arXiv:2002.11875*, 2020.

[179] Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.