Geometry and algorithms for signal recovery: from convex duality to non-convex formulations

Kellie J. MacPhee

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Dmitriy Drusvyatskiy, Chair

James V. Burke

Rekha R. Thomas

Program Authorized to Offer Degree:
Mathematics

University of Washington

Abstract

Geometry and algorithms for signal recovery:
from convex duality to non-convex formulations

Kellie J. MacPhee

Chair of the Supervisory Committee:
Dmitriy Drusvyatskiy
Department of Mathematics

Structured signal recovery is a central task in a variety of scientific applications, and naturally leads to non-linear and non-convex optimization problems that present many interesting mathematical and algorithmic challenges. In this work, we showcase results for three approaches to solving convex and non-convex optimization problems motivated by structured signal recovery.

First, we generalize and strengthen the theory of gauge duality in convex optimization. This includes developing a perturbation framework for gauge duality, establishing optimality conditions, and generalizing the gauge results to all nonnegative and convex functions. Second, we investigate a generalization of subgradient methods, originally designed for convex functions, to functions that are only weakly convex but enjoy the geometric advantage of sharpness. We see that subgradient methods on this class of functions converge at a local linear rate. Finally, we extend our work to the stochastic setting, presenting results for stochastic model-based minimization of functions with high-order growth. These results relax the traditional requirement of a global Lipschitz constant and allow for higher-order growth in the function to be minimized, using the tools of Legendre functions and Bregman divergences.

Throughout and wherever possible, we emphasize applications arising in the context of signal recovery, and provide numerical illustrations of our results.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The author wishes to express her sincere gratitude to the mentors and collaborators who have made this work possible, especially Dima Drusvyatskiy, Michael Friedlander, Sasha Aravkin, Jim Burke, Courtney Paquette, and Damek Davis. Thank you for sharing your wisdom and being excellent sources of knowledge and support over the past several years.

Thank you also to the University of Washington and the NSF TRIPODS funded Algorithmic Foundations of Data Science Institute (ADSI) for providing the financial and other resources necessary to complete this research.

Finally, a huge thank you to the peers, colleagues, friends, and family members who have provided various other kinds of support during the past several years. Your patience, kindness, encouragement, and willingness to help in whatever ways you can has been critical over the years that this research took place, and I am so grateful to have had you by my side.

Chapter 1

INTRODUCTION

Scientific applications, ranging from X-ray crystallography to image processing, often lead to noisy, nonlinear measurements from which a signal must be recovered. Often such recovery problems can be cast as the minimization of an appropriate penalty on the misfit between the predicted and the observed data, at which point mathematical optimization methods can aid in finding a solution. However, the natural optimization problems that arise may be intractable for a variety of reasons, with common challenges stemming from non-convexity and non-smoothness of the function to be minimized.

Broadly speaking, there are two approaches to dealing with non-convex optimization problems. The first is to apply an optimization method designed to work well in the convex setting directly to the non-convex problem, and hope for favorable convergence behavior to extend to this setting as well. By "favorable" we typically mean both convergence to a global minimizer (given an appropriate initialization), and a "fast" rate of convergence. When the objective function is smooth, for example, gradient descent coupled with an appropriate initialization is by far the most popular method in use, but must contend with the challenges of saddle points and local minimizers. In non-smooth optimization, there is no similarly universal "local search" procedure, and the appropriate method depends more heavily on context.

The second approach to solving these non-convex problems, which came into focus in the mid-2000's with the advent of compressed sensing, is the approach of *convex relaxations*. This method involves first approximating the original non-convex problem by a convex surrogate (e.g. by replacing a sparsity or rank constraint with a 1-norm or nuclear norm, or a general non-convex function with its convex envelope), and then solving this "relaxation" using

traditional methods for convex optimization. A series of influential works in the field of compressed sensing [4, 5, 11, 36, 39] spearheaded this approach, which is now widely used in applications with a sparse or low-rank structure, including phase retrieval and matrix completion [6, 7]. Some convex relaxations also involve a lifting procedure, which embeds the problem in a higher-dimensional space to allow the use of efficient convex optimization methods; one example is the celebrated max-cut formulation of Goemans and Williamson [21], another is the PhaseLift formulation of phase retrieval [6].

Convex relaxations are typically optimized over a space of very large dimension n, with the goal being to recover the signal based on a much smaller number of measurements m. Optimization problems over the primal space of dimension n are therefore computationally expensive, even in the convex setting. To avoid these inefficiencies, Friedlander and Macêdo [20] pioneered a method using the framework of *gauge duality* to solve convex relaxations of phase retrieval and blind deconvolution problems by optimization over a dual space of dimension $m << n$. In this work, we generalize the optimality conditions and primal-from-dual recovery framework derived in [20] to hold for general gauge optimization problems, significantly extending the utility of this approach. We also extend these results to hold for nonnegative convex functions which are not gauges, including common loss functions used in machine learning and statistics such as the Huber loss, hinge loss, and other piecewise linear-quadratic loss functions.

The convex relaxation approach is advantageous because it allows the use of well-developed tools from convex optimization, but this approach can be inefficient or even impossible in certain settings. Not all non-convex problems are well-approximated by a convex relaxation, and relaxations can be plagued by the above-mentioned issues of high-dimensionality. Due to these drawbacks, recently attention has returned to considering local search methods similar to gradient descent, applied directly to the non-convex problem.

For example, one important class of non-smooth, non-convex functions which have garnered recent attention is the class of composite functions $f = h \circ c$, given by the composition of a convex function h with a smooth map c [10, 13, 15]. In the composite optimization setting, the

prox-linear method has been shown to be very effective [12–15, 25]. The prox-linear approach has been applied, in particular, to a robust variant of the phase retrieval problem [10, 14, 16].

Composite functions fall within the more general category of *weakly convex* functions – those that become convex after the addition of a quadratic term $\alpha\|\cdot\|^2$. Weak convexity is a historically well-studied subject [8, 18, 31, 34], but only recently has the combination of weak convexity with a characteristic called *sharpness* been identified as leading to fast rates of convergence for simple non-convex optimization methods. In the context of phase retrieval, the works [16] and [14] recently showed that sharpness holds (for a particular variant of the problem), and Davis, Drusvyatskiy, and Paquette showed that a simple subgradient method converges linearly given a sufficiently close initialization [10]. In this work, we will see that these same convergence guarantees hold for general weakly convex and sharp problems, and that these two conditions combine to give a fast *linear rate* of convergence.

When the function to be minimized is only available through a stochastic sampling mechanism, which occurs for example in the setting of statistical learning, similar results have been shown as well [9, 15]. In particular, [9] proved that stochastic model-based minimization procedures, applied to non-smooth and weakly convex functions, converge at the rate $O(k^{-1/4})$. One of the major contributions of [9] was in first determining an appropriate way to measure the rate of convergence for these methods applied to non-smooth and non-convex objective functions. In this work, we generalize these results to allow for higher-order function growth by incorporating problem geometry via Bregman divergences, in an approach that significantly generalizes the mirror descent method which the reader may be familiar with.

1.1 Outline

The mathematical content of this dissertation is organized as follows. The remainder of Chapter 1 is devoted to a more detailed discussion of the background material and a summary of the mathematical contributions contained in later chapters. Chapter 2 focuses on optimization in the convex setting, developing a perturbation approach and optimality conditions for the gauge duality and perspective duality framework. Chapters 3 and 4 then

delve into the non-convex setting. Chapter 3 describes conditions under which subgradient-type methods applied to non-convex problems converge linearly, and includes numerical experiments illustrating this phenomenon. Chapter 4 extends this work, adapting these methods to the given problem geometry and allowing for higher-order function growth than is possible in the setting of Chapter 3. Chapters 2, 3, and 4 contain material from the following papers, respectively:

- A.Y. Aravkin, J.V. Burke, D. Drusvyatskiy, M.P. Friedlander, and K.J. MacPhee. Foundations of gauge and perspective duality. *SIAM Journal on Optimization,* 28(3), 2018.[1]

- D. Davis, D. Drusvyatskiy, K.J. MacPhee, and C. Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications,* 179(3), 2018.[2]

- D. Davis, D. Drusvyatskiy, and K.J. MacPhee. Stochastic model-based minimization under high-order growth. *arXiv:1807.00255,* 2018.

1.2 Duality theories for convex optimization

Duality theories, ubiquitous in convex optimization and more generally in mathematics, play a central role in transforming a difficult mathematical problem into a more tractable one. The idea of pairing a given object (whether a set, a function, or an optimization problem) with a second object of similar type allows us to approach problems from a complementary perspective, often opening up new routes to a solution. In this section, we consider the application of duality theories to solving convex signal recovery problems, such as the convex relaxations discussed previously. We will be specifically interested in optimization problems

involving gauge functions, a generalization of norms. This framework has been successfully applied in the context of phase retrieval and blind deconvolution [20], and our results, detailed in Chapter 2, allow extensions to a wider range of application settings.

Specifically, we will be interested in problems of the form

$$\underset{x}{\text{minimize}} \quad \kappa(x) \quad \text{subject to} \quad \rho(b - Ax) \leq \sigma, \tag{G_p}$$

where κ and ρ are non-negative and convex functions. Such problems arise when b is our observed data, A is a (linear) measurement operator, ρ is a penalty on the data misfit for a candidate solution x, and κ is a structure-inducing regularizer, for example the 1-norm for sparsity or nuclear norm for low-rank structure. For the sake of brevity, we will focus our introductory discussion on the case when κ and ρ are gauge functions (to be defined); the more general case is treated in Chapter 2.

1.2.1 Background and notation

Before continuing, we pause to fix notation and recall relevant definitions from convex analysis. We define $\overline{\mathbb{R}}$ to be the extended real line, $\mathbb{R} \cup \pm\infty$. We assume that all functions we encounter are proper, meaning they take a finite value at least at one point and never take the value $-\infty$. Recall that a function $f \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is convex if for any $x, y \in \mathbb{R}^n$ and $0 \leq \lambda \leq 1$, we have

$$f(\lambda x + (1 - \lambda y)) \leq \lambda f(x) + (1 - \lambda) f(y).$$

Equivalently, convexity can be characterized in terms of the *epigraph* of f, which is defined as

$$\operatorname{epi} f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq r\}.$$

Then f is convex if and only if $\operatorname{epi} f$ is a convex set, and similarly we say that f is *closed* if and only if $\operatorname{epi} f$ is a closed set.

For a convex function f the *subdifferential* $\partial f(x)$ at a point x is a set-valued generalization of the gradient, defined for nonsmooth functions. It is given by the set of affine underestimators of f based at the point x:

$$\partial f(x) := \{v | f(x) + \langle v, y - x \rangle \leq f(y), \, \forall y\}.$$

In particular, if f is smooth then we have $\partial f(x) = \{\nabla f(x)\}$. The subdifferential thus describes the variability of the function f, locally around the point x, and is critical to the variational approach that we will soon consider for deriving duality relationships in convex optimization.

Modern treatment of duality in the context of convex optimization problems derives from a perturbation or sensitivity analysis, which was elegantly formalized by Rockafellar and Wets [35]. The traditional and ubiquitous Fenchel duality fits into this framework, relying on the duality correspondence between convex functions $f\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and their *Fenchel conjugates* $f^*\colon \mathbb{R}^n \to \overline{\mathbb{R}}$. The latter is the function defined as

$$f^*(y) := \sup_x [\langle x, y\rangle - f(x)].$$

When f is closed and convex, we have the bi-conjugate relation $(f^*)^* = f$. In general, we will have $(f^*)^* \leq f$.

Gauge functions, which significantly generalize norms, have an additional duality correspondence with their polar gauges. A convex function $\kappa\colon \mathbb{R}^n \to [0, +\infty]$ is called a *gauge* if it is nonnegative, vanishes at the origin, and is positively homogeneous (i.e. $\kappa(\lambda x) = \lambda\kappa(x)$ for all $x \in \mathbb{R}^n$ and $\lambda > 0$). The geometric theory of polarity for convex cones suggests pairing a gauge κ with an associated *polar gauge*, which we denote as $\kappa^\circ\colon \mathbb{R}^n \to \overline{\mathbb{R}}_+$ and define by

$$\kappa^\circ(y) := \sup_{\kappa(x)\leq 1} \langle x, y\rangle.$$

Note that if κ is a norm, then κ° is precisely the corresponding dual norm. Equivalently, we could frame this in terms of polarity of cones by considering the epigraphs:

$$\operatorname{epi} \kappa^\circ = \{(y, -\lambda) | (y, \lambda) \in (\operatorname{epi} \kappa)^\circ\}.$$

Similar to the case for Fenchel conjugates, when κ is closed we have the symmetric relation $(\kappa^\circ)^\circ = \kappa$.

First investigated by Freund [19] in 1987, the theory of gauge duality exploits this relationship between gauges and their polars in a similar manner to the way in which Fenchel

conjugacy is exploited in Fenchel duality. The two operations yield significantly different dual problems, yet both can be derived via the general perturbation framework of [35].

With this notation in mind, the *gauge dual problem* paired with (G_p) is as follows:

$$\underset{y}{\text{minimize}} \quad \kappa^\circ(A^T y) \quad \text{subject to} \quad \langle b, y\rangle - \sigma\rho^\circ(y) \geq 1. \tag{G_d}$$

In contrast to the primal problem and the usual Fenchel or Lagrange dual, (G_d) has both small dimension $m \ll n$ and the linear operator appearing in the objective rather than the constraint. As a result, if κ° is simple we can compute subgradients of the objective function $\kappa^\circ \circ A^T$ using the chain rule, and thus projection-based algorithms are feasible for (G_d). This method is exploited in [20] for problems in phase retrieval and blind deconvolution.

1.2.2 Perturbation framework for duality

Consider the generic optimization problem

$$\inf_{x\in\mathbb{R}^n} \varphi(x) \tag{1.2.1}$$

where $\varphi\colon \mathbb{R}^n \to \overline{\mathbb{R}}$ is some convex function. Duality in convex optimization corresponds to defining a second optimization problem which encodes the sensitivity of the optimal value of (1.2.1) to perturbations in the objective function φ. More specifically, we introduce a perturbation parametrized by the variable $y \in \mathbb{R}^m$ and define the value function

$$p(y) = \inf_x f(x, y),$$

where $f(x, y)\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ is a convex function chosen to satisfy $f(x, 0) = \varphi(x)$. Then $p(0)$ corresponds to the optimal value of the original problem (1.2.1), and this naturally yields the primal-dual pair of optimization problems

$$p(0) = \inf_x\{\varphi(x) = f(x, 0)\} \quad \text{and} \quad p^{**}(0) = \sup_y\{\psi(y) = -f^*(0, y)\}. \tag{1.2.2}$$

Here p^{**} is the double-conjugate of the convex function p, and thus by solving the dual problem we obtain the lower bound $p^{**}(0) \leq p(0)$.

Under certain regularity conditions, the two optimal values are equal, and both are attained. That is, we can find some x^* and y^* such that

$$\varphi(x^*) = p(0) = p^{**}(0) = \psi(y^*).$$

Moreover, we obtain the following characterization of such optimal solutions in terms of the subdifferential of the perturbation function f and its conjugate.

Theorem 1.2.1. *Under appropriate regularity conditions, the following are equivalent.*

(a) x^* *is primal optimal and* y^* *is dual optimal,*

(b) $(0, y^*) \in \partial f(x^*, 0)$,

(c) $(x^*, 0) \in \partial f^*(0, y^*)$.

Under slightly stronger regularity conditions, we also have that

$$y^* \text{ is dual optimal} \quad \Longleftrightarrow \quad y^* \in \partial p(0),$$

which says precisely that the variability of the value function p near zero is governed by y^*. Thus, the optimal dual variables y^* measure sensitivity of the optimal value of the primal problem to the perturbation we have defined.

This general duality framework can be applied with different choices of the perturbation function f. Fenchel duality, Lagrange duality, and gauge duality can each be seen as particular examples. A major result contained in Chapter 2 is the derivation of this perturbation framework for gauge duality.

1.2.3 Gauge duality

Exploiting the fact that κ and ρ in (G_p) are gauges, we can instead define the following perturbation scheme. First, using a standard trick we rewrite (G_p) as the equivalent optimization problem

$$\inf_{\mu \geq 0,\, x} \quad \mu \quad \text{subject to} \quad \rho(b - Ax) \leq \sigma, \quad \kappa(x) \leq \mu. \tag{G_p}$$

We then define the perturbed optimization problem by

$$v_p(y) := \inf_{\mu \geq 0,\, x} \{\mu | \rho\,(b - Ax + \mu y) \leq \sigma,\ \kappa(x) \leq \mu\}\,.$$

The major difference between this gauge duality perturbation and the usual Fenchel duality perturbation is the multiplicative relationship between the dual variable y and the optimal value μ of the primal problem. A careful computation shows that the dual problem defined by this perturbation recovers the gauge dual problem (G_d).

1.2.4 Optimality conditions and primal-from-dual recovery

Of course, finding a minimizer y^* for the gauge dual problem (G_d) is usually not satisfactory on its own; what we really want is a minimizer x^* for the original problem (G_p). Thus, conditions relating the optimal solutions of (G_p) and (G_d) are critical. Such optimality conditions can be derived via Theorem 1.2.1 using our perturbation scheme, assuming a standard constraint qualification holds for the feasible regions of (G_p) and (G_d). We call this constraint qualification *strict feasibility*.

We show in the work of Chapter 2 that strict feasibility corresponds to zero lying in the interior of the domain of the value functions p and p^{**}, and thus from Theorem 1.2.1 we obtain the following optimality conditions for gauge duality.

Theorem 1.2.2. *If the pair (x^*, y^*) is primal-dual feasible for (G_p) and (G_d), then (x^*, y^*) is primal-dual optimal if and only if it satisfies the following four conditions:*

$$\rho(b - Ax^*) = \sigma \quad \text{or} \quad \rho^\circ(y^*) = 0 \qquad \textit{(primal activity)} \tag{1.2.3a}$$

$$\langle b, y^* \rangle - \sigma\rho^\circ(y^*) = 1 \qquad \textit{(dual activity)} \tag{1.2.3b}$$

$$\langle x^*, A^T y^* \rangle = \kappa(x^*) \cdot \kappa^\circ(A^T y^*) \qquad \textit{(objective alignment)} \tag{1.2.3c}$$

$$\langle b - Ax^*, y^* \rangle = \rho(b - Ax^*) \cdot \rho^\circ(y^*). \qquad \textit{(constraint alignment)} \tag{1.2.3d}$$

These conditions are analogous to the KKT conditions for Lagrange duality, but do not require smoothness of the objective and constraint functions. The alignment conditions

(1.2.3c)-(1.2.3d) correspond to equality holding in the Hölder-like inequalities

$$\langle x, A^T y\rangle \leq \kappa(x) \cdot \kappa^\circ(A^T y) \quad \text{and} \quad \langle b - Ax, y\rangle \leq \rho(b - Ax) \cdot \rho^\circ(y),$$

which hold for every x and y as an immediate consequence of the definition of the polar gauge. These conditions can be exploited for recovery of primal solutions x^* given an optimal y^*.

As a simple example, consider the setting of sparse recovery in which we have $\kappa = \|\cdot\|_1$. Then $\kappa^\circ = \|\cdot\|_\infty$, and the objective alignment condition (2.3.5c) corresponds to

$$\langle x^*, A^T y^*\rangle = \|x^*\|_1 \cdot \|A^T y^*\|_\infty,$$

which holds precisely when the the support of x^* aligns with the coordinates of $A^T y^*$ having maximal absolute value. Thus a gauge dual solution y^* immediately recovers the sparsity pattern in the coordinates of optimal solutions x^* for the gauge primal. This is interesting in its own right, as knowledge of the appropriate sparsity pattern allows us in principle to reduce the dimension in the primal problem (G_p) and solve the resulting low-dimensional problem directly.

1.2.5 Extension to non-gauge functions

In Chapter 2, we also extend the optimality conditions and gauge duality relationships detailed above to non-gauge functions. In particular, the gauge functions κ and ρ can be replaced by any nonnegative, nonconvex function. This possibility is achieved by "lifting" the non-gauge functions via the *perspective transform*, and applying the gauge duality results to the perspective functions. Pulling back the results gives us similar optimality conditions and duality relationships, which can be applied to common loss functions such as the piecewise linear-quadratic penalties.

1.3 Weak convexity and sharpness in non-convex optimization

Rather than solving a dual problem and subsequently recovering a primal-optimal solution, we can also work directly with the primal problem. In this section, we consider methods for

solving general constrained optimization problems of the form

$$\min_{x \in \mathcal{X}} \; g(x), \tag{1.3.1}$$

where $\mathcal{X}$ is a closed and convex set, and g is a potentially non-smooth and non-convex function. We will use the notation g^* to denote the value of the above minimum, and $\mathcal{X}^*$ to denote the set of minimizers, i.e. the set of points $x^* \in \mathcal{X}$ with $g(x^*) = g^*$.

1.3.1 Setting the scene with convexity

When g is convex, classical (projected) subgradient methods for solving (1.3.1) proceed as given in Algorithm 1. Here the notation $\text{proj}_{\mathcal{X}}(y)$ means the nearest point of $\mathcal{X}$ to y (i.e. the projection of y onto $\mathcal{X}$).

Algorithm 1: Subgradient method

Data: Initial point x_0 and stepsize sequence $\{\alpha_k \geq 0\}_{k \geq 0}$

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial g(x_k)$

if $\zeta_k \neq 0$ **then**

Set $x_{k+1} = \text{proj}_{\mathcal{X}} \left(x_k - \alpha_k \cdot \frac{\zeta_k}{\|\zeta_k\|} \right)$

else

Set $x_{k+1} = x_k$

end

Subgradient methods differ primarily in their choice of the stepsize sequence $\{\alpha_k\}$, and are perhaps the simplest approach to optimization for non-smooth convex problems. In general, however, the worst-case convergence rates for subgradient methods can be fairly slow. For convex problems with Lipschitz objective functions, after T iterations a subgradient method with appropriate stepsize sequence will find a point y_T satisfying $g(y_T) - g^* \leq O(1/\sqrt{T})$ [3]. This rate has been shown to be optimal over all black-box optimization algorithms for this problem class [28, 29]. In comparison, when the objective function g is smooth (differentiable with Lipschitz gradient), this rate improves to $O(1/T)$ [3].

Without further assumptions on the function g, we can only guarantee these rates of convergence in terms of function values, and we have no guarantees on the convergence rate of the sequence x_k. A superior mode of convergence is convergence at a *linear rate*, which occurs when the iterates satisfy

$$\text{dist}(x_{k+1}; \mathcal{X}^*) \leq q \cdot \text{dist}(x_k; \mathcal{X}^*), \qquad \text{for some } q \in [0, 1).$$

Here the distance function between a point z and a set C is defined as $\text{dist}(z; C) := \min_{x \in C} \|z - x\|$.

In order for subgradient methods to converge at a linear rate we must impose additional regularity conditions on the problem of interest. The appropriate condition here is *sharpness*, defined by the inequality

$$g(z) - \min_{x \in \mathcal{X}} g(x) \geq \mu \cdot \text{dist}(z; \mathcal{X}^*) \qquad \text{holds for all } z \in \mathcal{X},$$

where $\mu > 0$ is the sharpness constant. In words, $\mu-$sharpness means that the function g grows linearly away from its solution set.

For non-smooth convex functions which are L-Lipschitz, and μ-sharp, the subgradient method (with an appropriate choice of stepsizes) exhibits a linear rate, with $q = \sqrt{1 - (\mu/L)^2}$. Results of this type date back to the 60's and 70's [17, 22, 32, 33, 37], while some more recent approaches have appeared in [24, 38, 40].

1.3.2 Linear convergence for non-convex problems

A natural question to ask is whether the subgradient method, and the convergence results detailed above, can be extended to the setting of non-convex optimization. In short, we will see that for the class of sharp and weakly-convex objective functions g, the answer is yes. Before we can formally address this question, though, we must define what we mean by a subgradient in this non-convex setting.

To that end, we define the *Fréchet subdifferential* of g at x, denoted $\partial g(x)$, as the set of

all vectors $v \in \mathbb{R}^d$ satisfying

$$g(y) \geq g(x) + \langle v, y - x \rangle + o(\|y - x\|) \qquad \text{as } y \to x. \tag{1.3.2}$$

Here the term $o(t)$ represents any quantity satisfying $o(t)/t \to 0$ as $t \to 0^+$. When g is convex, this reduces to our earlier definition of the subdifferential. When g is weakly-convex, the Fréchet subdifferential coincides with other common definitions of the subdifferential. We say that g is ρ-weakly convex [30] if the perturbed function $x \mapsto g(x) + \frac{\rho}{2}\|\cdot\|^2$ is convex for some $\rho \geq 0$. In particular, g itself may not be convex.

We will say that a point $\bar{x} \in \mathcal{X}$ is *stationary* for the target problem (1.3.1) if

$$g(x) - g(\bar{x}) \geq o(\|x - \bar{x}\|) \qquad \text{as } x \to \bar{x} \text{ in } \mathcal{X}.$$

That is, $\bar{x}$ is *stationary* precisely when the zero vector is a subgradient of $g + \delta_{\mathcal{X}}$ at $\bar{x}$.

We note that for a $\rho-$weakly convex function g, its subgradients given in (1.3.2) automatically satisfy the much stronger property:

$$g(y) \geq g(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ v \in \partial g(x). \tag{1.3.3}$$

Weakly convex functions are widespread in applications, one source being the composite problem class:

$$F(x) := h(c(x)), \tag{1.3.4}$$

where $h\colon \mathbb{R}^m \to \mathbb{R}$ is convex and L-Lipschitz, and $c\colon \mathbb{R}^d \to \mathbb{R}^m$ is a C^1-smooth map with β-Lipschitz gradient. A straightforward argument shows that F is $\rho-$weakly convex with $\rho \leq L\beta$. The elements of the subdifferential $\partial F(x)$ are straightforward to compute through the chain rule [35, Theorem 10.6, Corollary 10.9]:

$$\partial F(x) = \nabla c(x)^* \partial h(c(x)) \qquad \text{for all } x \in \mathbb{R}^d.$$

In general, weak convexity coupled with sharpness of the function g gives us a tube around the solution set $\mathcal{X}^*$ that contains no extraneous stationary points:

$$\mathcal{T} := \left\{ x \in \mathcal{X} : \operatorname{dist}(x; \mathcal{X}^*) < \frac{2\mu}{\rho} \right\}.$$

Given an initial point that lies within a slight contraction of this tube, we show in Chapter 3 that the standard linearly convergent subgradient methods of Algorithm 1, originally designed for convex problems, achieve similar rates for weakly convex and sharp objective functions. That is, for $\gamma > 0$ we define the following tube

$$\mathcal{T}_\gamma := \left\{ x \in \mathcal{X} : \operatorname{dist}(x; \mathcal{X}^*) < \gamma \cdot \frac{\mu}{\rho} \right\}$$

and the constant

$$L := \sup \left\{ \|\zeta\| : \zeta \in \partial g(x),\, x \in \mathcal{T}_1 \right\}. \tag{1.3.5}$$

Observing that μ and L play reciprocal roles in characterizing the shape of the function g, we define the ratio $\tau := \mu/L$ as a measure of conditioning.

With this notation in mind, we define the weakly-convex and sharp extension of the subgradient method to be simply Algorithm 1 with the additional restriction that the initial point x_0 must lie in $\mathcal{T}_\gamma$ for some specified $\gamma \in (0, 1)$. We focus on three stepsize rules: Polyak stepsize [17, 32], geometrically decaying stepsize [22, 37], and constant stepsize [24, 38, 40]. The stepsize definitions and resulting convergence guarantees are summarized in Table 1.3.1. Note that the Polyak stepsize requires knowledge of the true minimum g^*, but no other information. For the geometrically decaying stepsize, we must fix a $\lambda > 0$ and $q \in (0, 1)$, which are required to satisfy some conditions in terms of γ, μ, and L. Finally, note that the constant stepsize method (with α sufficiently small) gives linear convergence to within a fixed distance of the solution set; this distance depends on the choice of α.

Chapter 3 contains proofs of these results, as well as illustrations of the convergence behavior for numerical examples of (real) phase retrieval and covariance matrix estimation.

1.4 Bregman functions and high-order growth

The subgradient methods previously considered can be seen as a special case of what we term *model-based minimization procedures*. Model-based minimization procedures, given a function f to be minimized and an iterate x_k, set the next iterate to be the minimizer of a local model

Table 1.3.1: A summary of the results contained in Chapter 3. We list the convergence rates of three variants of Algorithm 1 applied to sharp, weakly-convex problems.

Stepsize	Upper Bound on $\text{dist}^2(x_k; \mathcal{X}^*)$
Polyak: $\alpha_k = \frac{g(x_k)-g^*}{\|\zeta_k\|}$	$\left(1-(1-\gamma)\tau^2\right)^k \text{dist}^2(x_0; \mathcal{X}^*)$
Geometric Decay: $\alpha_k = \lambda \cdot q^k$	$\left(1-(1-\gamma)\tau^2\right)^k \cdot \max\left\{\frac{\lambda^2}{\tau^2}, \text{dist}(x_0; \mathcal{X}^*)^2\right\}$
Constant: $\alpha_k = \alpha$	$E^* + \max\{q^k(\text{dist}(x_0; \mathcal{X}^*)^2 - E^*), 2\alpha^2\}$

of the function f based at the point x_k. That is, these methods iterate the steps

$$x_{k+1} = \underset{y}{\text{argmin}} \left\{ f_{x_k}(y) + \eta \cdot d(y, x_k) \right\} \tag{1.4.1}$$

where f_{x_k} is a local model of the objective function f near the point x_k, $d(\cdot, \cdot)$ is a "distance" measure penalizing large deviations from the current iterate, and $\eta > 0$ is a step-size parameter. Gradient and subgradient methods take the models f_{x_k} to be linear, i.e.

$$f_{x_k}(y) = f(x_k) + \langle v_k, y - x_k \rangle, \qquad v_k \in \partial f(x_k) \tag{1.4.2}$$

and the distance measure to be $d(y, x) = \frac{1}{2}\|y - x\|^2$. When f is convex, the model (1.4.2) is a global under-estimator of the function f. When f is weakly-convex, (1.4.2) globally underestimates f up to a quadratic error term (by definition of the subdifferential). The results of Chapter 3 show that with an appropriate choice of the stepsize η, the subgradient method can overcome weak convexity in the original function f and obtain favorable convergence guarantees, under the additional assumption of sharpness.

In settings where computing an exact (sub)gradient v_k is expensive or impossible, for example in machine learning applications where the objective function f contains either a large finite sum or an expectation, *stochastic* optimization is a common alternative. Stochastic model-based methods, including the popular stochastic gradient descent, simply replace the models f_{x_k} given above by stochastic counterparts. That is, stochastic methods first

sample a random variable $\xi \sim P$, and then minimize a stochastic model $f_{x_k}(\cdot, \xi)$ such that $\mathbb{E}_\xi[f_{x_k}(y, \xi)] \approx f(y)$ for points y near x_k. These methods are often applied to problems of the form

$$\min_{x \in \mathbb{R}^d} \; F(x) := f(x) + r(x) \tag{1.4.3}$$

where we only have access to the function f via a stochastic sampling mechanism, and r is a regularization term used to prevent overfitting or induce a particular structure in the solution.

Assuming a global Lipschitz-type property for the stochastic models, Davis and Drusvyatskiy [9] show that the stochastic model-based minimization procedure, applied to weakly convex functions, drives a natural stationarity measure to zero at the rate $O(k^{-1/4})$. The stationarity measure they consider is the gradient of the Moreau envelope F_λ, defined by

$$F_\lambda(x) := \inf_y \left\{ F(y) + \tfrac{1}{2\lambda}\|y - x\|_2^2 \right\}. \tag{1.4.4}$$

The Moreau envelope F_λ can be thought of as a smoothed version of the original function F, with smoothing parameter $\lambda > 0$.

Global Lipschitz assumptions are widespread in the optimization literature, but recent literature has highlighted the fact that many problems of interest do not satisfy these conditions [1, 2, 23, 26, 27]. The aim of Chapter 4 is to relax this global Lipschitz assumption, using the tools of Legendre functions and Bregman divergences. Under mild technical conditions, we will show that the stochastic model-based minimization algorithm drives the gradient of the Bregman envelope (to be defined) to zero at the rate $O(k^{-1/4})$, where the size of the gradient is measured in the local norm induced by a Legendre function Φ. Much of this work is inspired by the ideas of relative smoothness and relative continuity presented in [26, 27].

1.4.1 Legendre functions and Bregman divergences

A *Legendre function* $\Phi \colon \mathbb{R}^d \to \overline{\mathbb{R}}$ is a function satisfying:

1. (Convexity) Φ is proper, closed, and strictly convex. That is, the epigraph of Φ is non-empty and closed, and for all $x, y \in \mathbb{R}^d$ with $x \neq y$, we have strict inequality in the convexity inequality:

$$\lambda\Phi(x) + (1-\lambda)\Phi(y) < \Phi(\lambda x + (1-\lambda)y), \quad \forall \lambda \in (0,1).$$

2. (Essential smoothness) The domain of Φ has nonempty interior $U := \text{int}(\text{dom}\,\Phi)$, Φ is differentiable on U, and for any sequence $\{x_k\} \subset U$ converging to a boundary point of $\text{dom}\,\Phi$, it must be the case that $\|\nabla\Phi(x_k)\| \to \infty$.

Each Legendre function Φ induces a distance-like measure D_Φ called a *Bregman divergence*, which is defined by

$$D_\Phi(y, x) := \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle, \quad \forall x, y.$$

Note that the strict convexity of Φ ensures $D_\Phi(y, x) \geq 0$ for all x and y, with equality holding if and only if $x = y$. In addition, we can define a notion of weak-convexity relative to the function Φ as follows: the function h is $\rho-$weakly convex relative to Φ if

$$h(y) \geq h(x) + \langle v, y - x\rangle - \rho D_\Phi(y, x), \quad \forall x, y.$$

1.4.2 Stationarity measure

We now introduce the stationarity measure that we will use to describe the convergence rate of the stochastic model-based minimization procedure with Bregman divergences. Analagous to the Euclidean setting, we define the Bregman envelope

$$F_\lambda^\Phi(x) := \inf_y \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

and the associated Φ-*proximal map*

$$\text{prox}_{\lambda f}^\Phi(x) := \underset{y}{\text{argmin}} \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\}.$$

Note that in the Euclidean setting with $\Phi = \frac{1}{2}\|\cdot\|^2$, these definitions coincide with the standard Moreau envelope and proximal mapping.

We will measure the convergence guarantees of our algorithm based on the rate at which the quantity

$$\mathbb{E}\left[D_\Phi\left(\text{prox}_{\lambda F}^\Phi(x), x\right)\right] \tag{1.4.5}$$

tends to zero for some fixed $\lambda > 0$.The meaning of this quantity becomes more apparent after making slightly stronger assumptions on the Legendre function Φ. Namely, if Φ is 1-strongly convex with respect to some norm $\|\cdot\|$ and twice differentiable at every point in U, then we have the following result.

Theorem 1.4.1 (Smoothness of the Φ-envelope)**.** *Suppose Φ is 1-strongly convex and twice-differentiable on U. Then for any small enough positive λ, the envelope F_λ^Φ is differentiable at any point $x \in U$ with gradient given by*

$$\nabla F_\lambda^\Phi(x) := \frac{1}{\lambda}\nabla^2\Phi(x)\left(x - \text{prox}_{\lambda F}^\Phi(x)\right).$$

For $x \in U$, we define the local norm $\|y\|_x := \left\|\nabla^2\Phi(x)y\right\|_*$ which has corresponding dual norm $\|v\|_x^* = \left\|\nabla^2\Phi(x)^{-1}v\right\|$. Then Theorem 1.4.1 tells us that for small positive λ and $x \in U$ we have

$$\sqrt{D_\Phi\left(\text{prox}_{\lambda F}^\Phi(x), x\right)} \geq \frac{\lambda}{\sqrt{2}}\|\nabla F_\lambda^\Phi(x)\|_x^*.$$

Thus the square root of the Bregman divergence, which we will show tends to zero along the iterate sequence at a controlled rate, bounds the local norm of the gradient ∇F_λ^Φ. Since the function F_λ^Φ is a smoothed version of the original function F, in the setting of Theorem 1.4.1 we will see that the gradient of a smooth approximation to F tends to zero at a given rate.

1.4.3 Stochastic model-based minimization with Bregman divergences

Given a Legendre function Φ and associated Bregman divergence D_Φ, which are well-adapted to the geometry of the problem (1.4.3) and the stochastic models $f_x(\cdot, \xi)$, the stochastic model-based minimization algorithm of Chapter 4 proceeds as in Algorithm 2. The constant

ρ is the weak-convexity constant of the stochastic models $f_x(\cdot, \xi) + r(\cdot)$, while the constant τ controls the growth of the stochastic models away from the true function values, measured in terms of the Bregman divergence:

$$\mathbb{E}_\xi[f_x(y, \xi) - f(y)] \leq \tau D_\Phi(y, x), \quad \forall x, y \in U \cap \operatorname{dom} r$$

We also require a Lipschitz-type property, relative to the Bregman divergence, for the stochastic models $f(\cdot, \xi)$. Note, however, that this is quite different from the common global Lipschitz assumption, as the Bregman divergence can account for higher-order function growth.

Algorithm 2: Stochastic Model Based Minimization

Data: $x_0 \in U \cap \operatorname{dom} r$, real $\lambda < (\tau + \rho)^{-1}$, a nonincreasing sequence $\{\eta_t\}_{t \geq 0} \subseteq (0, \lambda)$, and iteration count T.

Step $t = 0, \ldots, T$:

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \underset{x}{\operatorname{argmin}} \ \left\{ f_{x_t}(x, \xi_t) + r(x) + \frac{1}{\eta_t} D_\Phi(x, x_t) \right\} \end{array} \right\},$$

Sample $t^* \in \{0, \ldots, T\}$ according to the discrete probability distribution

$$\mathbb{P}(t^* = t) \propto \frac{\eta_t}{1 - \eta_t \rho}.$$

Return x_{t^*}

The convergence guarantee for Algorithm 2, although obtained for arbitrary nonincreasing stepsize sequences $\{\eta_t\}$, simplifies in the case that η_t is constant. In particular, we obtain the following result, which shows that our expected stationarity measure decreases at the rate $O(1/\sqrt{T})$. Similar results are detailed in Chapter 4 for general stepsize sequences.

Theorem 1.4.2 (Convergence rate for constant stepsize)**.** *Fix some $\alpha > 0$ and number of iterations T, and set the stepsize as $\eta_t = \frac{1}{\lambda^{-1}+\alpha^{-1}\sqrt{T+1}}$ for all indices $t = 1, \ldots, T$. Then the point x_{t^*} returned by Algorithm 2 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\hat{x}_{t^*}, x_{t^*}\right)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0) - \min F) + \frac{\lambda \mathsf{L}^2\alpha^2}{4} + \frac{\lambda((r(x_0)-\inf r))}{\lambda^{-1}-\rho+\alpha^{-1}}}{1-\lambda(\tau+\rho)} \cdot \left(\frac{\lambda^{-1}-\rho}{T+1} + \frac{1}{\alpha\sqrt{T+1}}\right)$$

for $\hat{x}_{t^*} = \operatorname{prox}_{\lambda F}^\Phi(x_{t^*})$.

This concludes our overview of the results contained in this dissertation. Please see Chapters 2-4 for further details.

References

[1] H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[2] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *arXiv:1706.06461*, 2017.

[3] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends ® in Machine Learning*, 8(3-4):231–357, 2015.

[4] E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse Problems*, 23(3):969, 2007.

[5] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *arXiv:math/0409186*, 2004.

[6] E. J. Candes, T. Strohmer, and V. Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[7] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *arXiv:0903.1476*, 2009.

[8] F. H. Clarke, R. Stern, and P. Wolenski. Proximal smoothness and the lower-c2 property. *Journal of Convex Analysis*, 2(1-2):117–144, 1995.

[9] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[10] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *arXiv:1711.03247*, 2017.

[11] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[12] D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[13] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, pages 1–56, 2018.

[14] J. C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv:1705.02356*, 2017.

[15] J. C. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[16] Y. C. Eldar and S. Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[17] I. Eremin. The relaxation method of solving systems of inequalities with convex functions on the left-hand side. *Dokl. Akad. Nauk SSSR*, 160:994–996, 1965.

[18] H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3):418–491, 1959.

[19] R. M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.

[20] M. P. Friedlander and I. Macêdo. Low-rank spectral optimization via gauge duality. *SIAM Journal on Scientific Computing*, 28(3):A1616–A1638, 2016.

[21] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

[22] J. Goffin. On convergence rates of subgradient optimization methods. *Mathematical Programming*, 13(3):329–347, 1977.

[23] F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *arXiv:1803.07374*, 2017.

[24] P. Johnstone and P. Moulin. Faster subgradient methods for functions with Hölderian growth. *arXiv:1704.00196*, 2017.

[25] A. S. Lewis and S. J. Wright. A proximal method for composite minimization. *Mathematical Programming*, 158(1-2):501–546, 2016.

[26] H. Lu. Relative continuity for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv:1710.04718*, 2017.

[27] H. Lu, R. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *arXiv:1610.05708*, 2016.

[28] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley Interscience, 1983.

[29] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[30] E. A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[31] R. Poliquin and R. Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348(5):1805–1838, 1996.

[32] B. Poljak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9:14–29, 1969.

[33] B. Poljak. Subgradient methods: a survey of Soviet research. In *Nonsmooth optimization (Proc. IIASA Workshop, Laxenburg, 1977)*, volume 3 of *IIASA Proc. Ser.*, pages 5–29. Pergamon, Oxford-New York, 1978.

[34] R. T. Rockafellar. Favorable classes of lipschitz continuous functions in subgradient optimization. 1981.

[35] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media, 1998.

[36] M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In *2006 40th Annual Conference on Information Sciences and Systems*, pages 207–212. IEEE, 2006.

[37] N. Shor. The rate of convergence of the method of the generalized gradient descent with expansion of space. *Kibernetika (Kiev)*, 2:80–85, 1970.

[38] S. Supittayapornpong and M. Neely. Staggered time average algorithm for stochastic non-smooth optimization with $O(1/t)$ convergence. *arXiv:1607.02842*, 2016.

[39] J. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[40] T. Yang and Q. Lin. RSG: Beating subgradient method without smoothness and strong convexity. *arXiv:1512.03107*, 2016.

Chapter 2

FOUNDATIONS OF GAUGE AND PERSPECTIVE DUALITY

This chapter represents joint work with Aleksandr Y. Aravkin, James V. Burke, Dmitriy Drusvyatskiy, and Michael P. Friedlander. The contents appeared in *SIAM Journal on Optimization*, Aug 2018, 28-3 (p. 2406-2434).

Abstract: We revisit the foundations of gauge duality and demonstrate that it can be explained using a modern approach to duality based on a perturbation framework. We therefore put gauge duality and Fenchel-Rockafellar duality on equal footing, including explaining gauge dual variables as sensitivity measures, and showing how to recover primal solutions from those of the gauge dual. This vantage point allows a direct proof that optimal solutions of the Fenchel-Rockafellar dual of the gauge dual are precisely the primal solutions rescaled by the optimal value. We extend the gauge duality framework to the setting in which the functional components are general nonnegative convex functions, including problems with piecewise linear quadratic functions and constraints that arise from generalized linear models used in regression.

2.1 Introduction

Sensitivity of the optimal values and solutions of optimization problems, with respect to perturbations in the problem data, is a central concern of Fenchel-Rockafellar duality theory. Lagrange duality can be regarded as a special case of this theory, in which perturbations to the data are introduced in a particular manner. Gauge duality, on the other hand, as introduced in 1987 by Freund [13], was developed without any reference to sensitivity. It relies instead on a special polarity correspondence that exists for nonnegative, positively homogeneous convex functions that vanish at the origin; these are known as *gauge functions*.

In 2014, Friedlander, Macêdo, and Pong [15] made partial progress towards connecting gauge and Lagrange dualities. In the present work, we show that gauge duality may be regarded as a particular application of Fenchel-Rockafellar duality theory that is different than the one required for Lagrange duality. This connection provides a useful vantage point from which to develop new algorithms for an important class of convex optimization problems. We also describe how gauge duality theory can be extended beyond the optimization of gauge functions to the optimization of all convex functions that are bounded below. We call this extension *perspective duality*.

A convenient and fully general formulation for our approach is the problem

$$\underset{x}{\text{minimize}} \quad \kappa(x) \quad \text{subject to} \quad \rho(b - Ax) \leq \sigma, \tag{G_p}$$

where $A\colon \mathbb{R}^n \to \mathbb{R}^m$ is a linear map, b is an m-vector, and κ and ρ are closed gauge functions. For many applications, the function κ is used to regularize the problem in order to obtain solutions with certain desirable properties. For example, in statistical and machine-learning applications the regularizer κ is often a nonsmooth, structure-inducing function; e.g., the 1-norm, which is frequently used to encourage sparsity in the solution. The function ρ may be regarded as a penalty function, such as the 2-norm, that measures the degree of misfit between the data b and the linear model Ax, and may reflect a statistical model of the noise in the data b. The perspective duality extension enables us to consider optimization problems with a wider range of applications by allowing functions κ and ρ that are not positively homogeneous, including the Huber function used for robust regression [17], the elastic net used for group detection [28], and the logistic loss used for classification [1, 18].

The formulation (G_p) gives rise to two different "dual" problems:

$$\underset{y}{\text{maximize}} \quad \langle b, y\rangle - \sigma\rho^\circ(y) \quad \text{subject to} \quad \kappa^\circ(A^T y) \leq 1, \tag{L_d}$$

$$\underset{y}{\text{minimize}} \quad \kappa^\circ(A^T y) \quad \text{subject to} \quad \langle b, y\rangle - \sigma\rho^\circ(y) \geq 1. \tag{G_d}$$

Here ρ° and κ° are the polars of ρ and κ, which are also gauge functions; see section 2.2.1 for a precise definition. In the important case $\sigma = 0$, we interpret $\sigma\rho^\circ$ as the indicator function

of the closure of the domain of ρ° (see the discussion in section 2.2.3). The first problem (L_d) is the standard Lagrangian (or Fenchel-Rockafellar) dual, which is the dual problem typically considered in connection with convex optimization problems. Strong duality, reflected in the equality

$$\operatorname{val}(\mathrm{G}_p) = \operatorname{val}(\mathrm{L}_d),$$

and in the attainment of the optimal value of the Lagrange primal-dual pair, holds under mild interiority conditions often referred to as the Slater constraint qualification. The second problem (G_d) is the gauge dual and is less well-known. Under interiority conditions similar to those required by Lagrange duality, strong duality holds in the gauge duality setting; this is reflected in the analogous equality

$$1 = \operatorname{val}(\mathrm{G}_p) \cdot \operatorname{val}(\mathrm{G}_d),$$

and in the attainment of the optimal value of the gauge primal-dual pair.

In certain contexts, the gauge dual (G_d) can be preferable for computation to the primal (G_p) and the Lagrangian dual (L_d), particularly when the polar κ° has a special structure. Friedlander and Macêdo [14], for example, use gauge duality to derive an effective algorithm for an important class of low-rank spectral optimization problems that arise in signal-recovery applications, including phase recovery and blind deconvolution. Indeed, the effectiveness of numerous convex optimization algorithms—particularly first-order methods—relies on being able to project easily onto the constraint set. The appearance of the linear map A in the constraints of both (G_p) and (L_d) means that such methods may not be efficient, though some recent methods have been proposed that circumvent this difficulty [24]. In contrast, the map A appears in the gauge dual (G_d) only in the objective, and computing subgradients of this objective only requires subgradients of κ°, together with the ability to efficiently implement matrix-vector multiplication. Moreover, typical applications occur in the regime $m \ll n$. For example, m is often logarithmic in n [6, 7, 11, 26]. Because the dual variables y of (G_d) lie in the much smaller space $\mathbb{R}^m$, projections onto the feasible region may be computed efficiently, depending on the context. An example of how an interior method may be used for

this purpose is given in section 2.5.2.

2.1.1 Approach

This paper has two main goals. The first goal, addressed in section 2.3, is to show how the foundations of gauge duality can be derived via a perturbation framework pioneered by Rockafellar [20, 21], in which the optimal value and optimal solution depend on parameters to the problem. We follow Rockafellar and Wets [23, 11.H], who consider an arbitrary convex perturbation function F on $\mathbb{R}^n \times \mathbb{R}^m$ that determines how the parameters enter the problem, and define the value functions

$$p(u) := \inf_x \; F(x,u) \qquad \text{and} \qquad q(v) := \inf_y \; F^\star(v,y). \tag{2.1.1}$$

This set-up immediately yields the primal-dual pair

$$p(0) = \inf_x \; F(x,0) \qquad \text{and} \qquad p^{\star\star}(0) = \sup_y \; -F^\star(0,y) \equiv -q(0). \tag{2.1.2}$$

Fenchel-Rockafellar duality theory flows from an appropriate choice of F. We show that gauge duality fits equally well into this framework under a judicious choice of the perturbation function F, thereby putting Fenchel-Rockafellar and gauge duality theories on an equal footing. Strong duality, primal-dual optimality conditions, and an interpretation of the gauge dual solutions as sensitivity measures—i.e., subgradients of the value function—quickly follow; cf. section 2.3.2. These results, in particular, answer an open question posed by Freund in his original work [13], which asked for an interpretation of gauge dual variables for problems with nonlinear constraints. It also completes a partial analysis by Friedlander et al. [15] on the interpretation of gauge dual variables as sensitivity measures.

This viewpoint allows us to prove a striking relationship between optimal solutions of the primal and optimal solutions of the Lagrangian dual of the gauge dual: the two coincide up to scaling by the optimal value (section 2.3.5). Consequently, Lagrangian primal-dual methods applied to the gauge dual can be used to recover solutions of the original primal

problem. We illustrate this idea in section 2.7 with an application of Chambolle and Pock's primal-dual algorithm [8] to a specific problem instance.

The second goal of this paper is to extend the applicability of the gauge duality paradigm beyond gauges to capture more general convex problems. Section 2.4 extends gauge duality to problems involving convex functions that are merely nonnegative, and by an appropriate translation, functions that are bounded from below. The approach is based on using the perspective transform of a convex function [20, p. 35], which increases a function's domain from $\mathbb{R}^n$ to $\mathbb{R}^{n+1}$ and makes it positively homogeneous, enabling the property that is key to the application of gauge duality. We term the resulting dual problem the *perspective dual*. The perspective-polar transformation, needed to derive the perspective dual problem, is developed in section 2.4. Concrete illustrations of perspective duality for the family of piecewise linear-quadratic functions, which are often used in data-fitting applications, and for the setting of generalized linear models, are given in section 2.5. We further explore examples of optimality conditions and primal-from-dual recovery in section 2.6. Numerical illustrations for a case-study of perspective duals comprise section 2.7.

2.2 Notation and assumptions

The derivation of our results relies on standard notions from convex analysis. Unless otherwise specified, we generally follow Rockafellar [20] for standard definitions and notation, including domains and epigraphs, relative interiors, convex conjugate functions, subdifferentials, polar sets, etc. In this section we collect less well-known definitions and notation used throughout the paper, and establish blanket assumptions on the problem data.

Let $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ denote the extended real line, and $\overline{\mathbb{R}}_+ := \{\, x \in \overline{\mathbb{R}} \mid x \geq 0 \,\}$ denote the nonnegative extended reals. Let $f \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and $g \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ denote general closed convex functions. For a closed convex set $\mathcal{C} \subseteq \mathbb{R}^n$, its convex indicator $\delta_{\mathcal{C}}$ is the closed convex function whose value is zero on $\mathcal{C}$ and $+\infty$ otherwise. Let $\operatorname{cone} \mathcal{C} := \{\, \lambda x \mid \lambda \geq 0,\, x \in \mathcal{C} \,\}$ denote the cone generated by $\mathcal{C}$. We often abbreviate fractions such as $(1/(2\mu))$ to $(1/2\mu)$.

2.2.1 The perspective transform

For any convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, its *perspective* is the function on $\mathbb{R}^{n+1}$ whose epigraph is the cone generated by the set $(\operatorname{epi} f) \times \{1\}$. Because this transform is not necessarily closed—even when f is closed—we choose to work with its closure, and redefine the transform as

$$f^{\pi}(x, \lambda) := \begin{cases} \lambda f(\lambda^{-1} x) & \text{if } \lambda > 0 \\ f^{\infty}(x) & \text{if } \lambda = 0 \\ +\infty & \text{if } \lambda < 0, \end{cases} \tag{2.2.1}$$

where $f^{\infty}(x)$ is the *recession function* of f [20, Theorem 8.5]. A calculus for the perspective transform $f \mapsto f^{\pi}$ is described by Aravkin, Burke, and Friedlander [2, Section 3.3] and, for the infinite-dimensional case, by Combettes [9, 10], where properties of the perspective transform are described in detail. We often apply more than one transformation to a function, and in such cases, the multiple transformations are applied in the order that they appear; e.g., $f^{\pi\circ} := (f^{\pi})^{\circ}$.

2.2.2 Gauge functions

The following is only a brief description of gauge functions. A complete description is given by Rockafellar [20, Section 15].

A convex function $\kappa : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called a *gauge* if it is nonnegative, positively homogeneous, and vanishes at the origin. The symbols $\kappa \colon \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\rho \colon \mathbb{R}^m \to \overline{\mathbb{R}}$ will always denote closed gauges. The polar of a gauge κ is the function κ° defined by

$$\kappa^{\circ}(y) := \inf \{\, \mu > 0 \mid \langle x, y \rangle \leq \mu \kappa(x),\ \forall x \,\}, \tag{2.2.2}$$

which is also a gauge and satisfies $\kappa^{\circ\circ} = \kappa$ when κ is closed [20, Theorem 15.1]. For example, if κ is a norm then κ° is the corresponding dual norm. Note the identity

$$\operatorname{epi} \kappa^{\circ} = \{\, (y, -\lambda) \mid (y, \lambda) \in (\operatorname{epi} \kappa)^{\circ} \,\}. \tag{2.2.3}$$

It follows directly from (2.2.2) and positive homogeneity of a gauge function that its polar can be characterized as the support function to the unit level set, i.e.,

$$\kappa^\circ = \delta^*_{\mathcal{U}_\kappa} = \sup\{\,\langle u, \cdot\rangle \mid u \in \mathcal{U}_\kappa\,\} \quad \text{where} \quad \mathcal{U}_\kappa := \{\,u \mid \kappa(u) \leq 1\,\}. \tag{2.2.4}$$

Moreover, κ and κ° satisfy a Hölder-like inequality

$$\langle x, y\rangle \leq \kappa(x) \cdot \kappa^\circ(y) \quad \forall x \in \operatorname{dom}\kappa,\ \forall y \in \operatorname{dom}\kappa^\circ, \tag{2.2.5}$$

which we refer to as the *polar-gauge inequality*. The zero level set

$$\mathcal{H}_\kappa := \{\,u \mid \kappa(u) = 0\,\}$$

plays a key role when $\sigma = 0$. It is straightforward to show that

$$\mathcal{U}^\circ_\kappa = \mathcal{U}_{\kappa^\circ}, \quad \mathcal{U}^\infty_\kappa = \mathcal{H}_\kappa, \quad (\operatorname{dom}\kappa)^\circ = \mathcal{H}_{\kappa^\circ}, \quad \text{and} \quad \mathcal{H}^\circ_\kappa = \operatorname{cl}\operatorname{dom}\kappa^\circ \tag{2.2.6}$$

whenever κ is closed, where $\mathcal{U}^\infty_\kappa$ is the *recession cone* for $\mathcal{U}_\kappa$ [20, Section 8]. We include proofs of (2.2.6) in section 2.8.

2.2.3 Assumptions on the feasible region

Define the following primal and dual feasible sets:

$$\mathcal{F}_p := \{\,u \mid \rho(b-u) \leq \sigma\,\} \qquad \text{and} \qquad \mathcal{F}_d := \{\,y \mid \langle b, y\rangle - \sigma\rho^\circ(y) \geq 1\,\}. \tag{2.2.7}$$

The nonnegativity of ρ implies that the Slater condition can fail when $\sigma = 0$, and thus special attention is required. In this case, we make the replacement

$$(\rho,\ \sigma) \ \Rightarrow\ (\delta_{\mathcal{H}_\rho},\ 1) \quad \text{whenever} \quad \sigma = 0. \tag{2.2.8}$$

This replacement yields a gauge optimization problem whose solution set and optimal value coincide with those of (G_p). Observe that because $\mathcal{H}_\rho$ is a closed convex cone, $\delta_{\mathcal{H}_\rho} = \delta^*_{\mathcal{H}^\circ_\rho}$ is a closed gauge that satisfies, by virtue of (2.2.6), $\delta^\circ_{\mathcal{H}_\rho} = \delta_{\mathcal{H}^\circ_\rho} = \delta_{\operatorname{cl}\operatorname{dom}\rho^\circ}$. This motivates the convention made immediately following (G_d) that

$$\sigma\rho^\circ := \delta_{\operatorname{cl}\operatorname{dom}\rho^\circ} \equiv \delta_{\mathcal{H}^\circ_\rho} \quad \text{when} \quad \sigma = 0. \tag{2.2.9}$$

The replacement (2.2.8) allows us to make the useful assumption that $\sigma > \inf \rho$, which significantly streamlines our analysis. The convention (2.2.9) also makes sense from an epigraphical perspective, because the functions $\sigma\rho^\circ$ epigraphically converge to $\delta_{\operatorname{cl}\operatorname{dom}\rho^\circ}$ as $\sigma \downarrow 0$ [23, Proposition 7.4(c)].

The gauge primal (G_p) and dual (G_d) problems are said to be *feasible*, respectively, if the following intersections are nonempty:

$$A^{-1}\mathcal{F}_p \cap (\operatorname{dom}\kappa) \quad \text{and} \quad A^T\mathcal{F}_d \cap (\operatorname{dom}\kappa^\circ).$$

Similarly, the primal and dual problems are said to be *relatively strictly feasible*, respectively, if the following intersections are nonempty:

$$A^{-1}\left(\operatorname{ri}\mathcal{F}_p\right) \cap (\operatorname{ri}\operatorname{dom}\kappa) \quad \text{and} \quad A^T \operatorname{ri}\mathcal{F}_d \cap (\operatorname{ri}\operatorname{dom}\kappa^\circ).$$

If the intersections above are nonempty, with interior replacing relative interior, then we say that the problems are *strictly feasible*. We have

$$\operatorname{ri}\mathcal{F}_p = \begin{cases} \{\, u \mid b - u \in \operatorname{ri}\operatorname{dom}\rho,\ \rho(b-u) < \sigma \,\} & \text{if } \sigma > 0 \\ \{\, u \mid b - u \in \operatorname{ri}\mathcal{H}_\rho \,\} & \text{if } \sigma = 0, \end{cases}$$

$$\operatorname{ri}\mathcal{F}_d = \begin{cases} \{\, y \mid y \in \operatorname{ri}\operatorname{dom}\rho^\circ,\ \langle b, y\rangle - \sigma\rho^\circ(y) > 1 \,\} & \text{if } \sigma > 0 \\ \{\, y \mid y \in \operatorname{ri}\mathcal{H}_\rho^\circ,\ \langle b, y\rangle > 1 \,\} & \text{if } \sigma = 0, \end{cases}$$

which follows from Rockafellar [20, Theorem 7.6] when $\sigma > 0$, and from the convention (2.2.9) when $\sigma = 0$.

We assume throughout that $\rho(b) > \sigma$. Otherwise, $\mathcal{F}_p$ contains the origin, which is a trivial solution of (G_p). This assumption is consistent with classical applications in signal processing and machine learning, where the corresponding assumption is that the data b does not entirely consist of noise.

2.3 Perturbation analysis for gauge duality

Modern treatment of duality in convex optimization is based on an interpretation of multipliers as giving sensitivity information relative to perturbations in the problem data. No such

analysis, however, has existed for gauge duality. In this section we show that for a particular kind of perturbation, the gauge dual (G_d) can in fact be derived via such an approach.

2.3.1 General perturbation framework

Our analysis is based on a perturbation theory described by Rockafellar and Wets [23, 11.H]. In this section we summarize the main results from [23] that we need. Fix an arbitrary convex function $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$, and consider the value functions defined by (2.1.1)–(2.1.2). Observe the equality $q(0) = -p^{\star\star}(0)$. For example, Fenchel-Rockafellar duality for the problem

$$\underset{x}{\text{minimize}} \;\; f(Ax) + g(x), \tag{2.3.1}$$

is obtained from the general perturbation theory by setting $F(x,u) = f(Ax+u) + g(x)$. In that case, the primal-dual pair takes the familiar form

$$p(0) = \inf_x \left\{ f(Ax) + g(x) \right\} \quad \text{and} \quad p^{\star\star}(0) = \sup_y \left\{ -f^\star(-y) - g^\star(A^T y) \right\}.$$

Under certain conditions, described in the following theorem, strong duality holds, i.e. $p(0) = p^{\star\star}(0)$, and the optimal values are attained.

Theorem 2.3.1 (Multipliers and sensitivity [23, Theorem 11.39])**.** *Consider the primal-dual pair* (2.1.2)*, where* $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \overline{\mathbb{R}}$ *is proper, closed, and convex.*

(a) The inequality $p(0) \geq -q(0)$ *always holds.*

(b) If $0 \in \operatorname{ri}\operatorname{dom} p$*, then equality* $p(0) = -q(0)$ *holds and, if finite, the infimum* $q(0)$ *is attained with* $\partial p(0) = \operatorname{argmax}_y -F^\star(0,y)$*. Similarly, if* $0 \in \operatorname{ri}\operatorname{dom} q$*, then equality* $p(0) = -q(0)$ *holds and, if finite, the infimum* $p(0)$ *is attained with* $\partial q(0) = \operatorname{argmin}_x F(x,0)$*.*

(c) The set $\operatorname{argmax}_y -F^\star(0,y)$ *is nonempty and bounded if and only if* $p(0)$ *is finite and* $0 \in \operatorname{int}\operatorname{dom} p$*.*

(d) The set $\operatorname{argmin}_x F(x,0)$ *is nonempty and bounded if and only if* $q(0)$ *is finite and* $0 \in \operatorname{int}\operatorname{dom} q$.

(e) Optimal solutions are characterized jointly through the conditions

$$\left.\begin{array}{l} \bar{x} \in \operatorname{argmin}_x \ F(x,0) \\ \bar{y} \in \operatorname{argmax}_y \ -F^\star(0,y) \\ F(\bar{x},0) = -F^\star(0,\bar{y}) \end{array}\right\} \iff (0,\bar{y}) \in \partial F(\bar{x},0) \iff (\bar{x},0) \in \partial F^\star(0,\bar{y}).$$

Proof. The only difference between the statement of this theorem and that in [23, Theorem 11.39] is in part (b). Here, we make use of the relative interior rather than the interior. Thus, we only prove part (b). Suppose $0 \in \operatorname{ri}\operatorname{dom} p$. If $p(0) = -\infty$, then $p(0) = -q(0)$ follows by Part ((a)). Hence we can assume that $p(0)$ is finite, and conclude that p is proper. By [20, Theorem 23.4], $\partial p(0) \neq \emptyset$, and given $\phi \in \partial p(0)$,

$$p(0) \leq p(u) - \langle \phi, u \rangle = \inf_x \left\{ F(x,u) - \left\langle \begin{pmatrix} 0 \\ \phi \end{pmatrix}, \begin{pmatrix} x \\ u \end{pmatrix} \right\rangle \right\} \quad \forall u.$$

By taking the infimum over u and recognizing the right-hand side as $-F^\star(0,\phi)$, we deduce that $p(0) \leq -F^\star(0,\phi) \leq -q(0)$. Combining this with Part (a) yields $p(0) = -F^\star(0,\phi) = -q(0)$. Hence $\phi \in \operatorname{argmax}_y -F^\star(0,y) \neq \emptyset$. Conversely, given any $\phi \in \operatorname{argmax}_y -F^\star(0,y)$, we have

$$\begin{aligned} p(0) = -F^\star(0,\phi) &= \inf_{x,u} \left\{ F(x,u) - \left\langle \begin{pmatrix} 0 \\ \phi \end{pmatrix}, \begin{pmatrix} x \\ u \end{pmatrix} \right\rangle \right\} \\ &= \inf_u \{ p(u) - \langle \phi, u \rangle \} \leq p(v) - \langle \phi, v \rangle \qquad \forall\ v, \end{aligned}$$

and so $\phi \in \partial p(0)$. The case $0 \in \operatorname{ri}\operatorname{dom} q$ follows by an analogous argument. □

2.3.2 A perturbation for gauge duality

We now show that the problems (G_p) and (G_d) constitute a primal-dual pair under the framework set out by theorem 2.3.1. The key is to postulate the correct pairing function F.

In the derivation below, we show that the gauge primal-dual pair corresponds to the primal and dual value functions

$$v_p(u) := \inf_{\mu>0,\, x} \{\, \mu \mid \rho\,(b - Ax + \mu u) \le \sigma,\ \kappa(x) \le \mu \,\}\,, \tag{2.3.2a}$$

$$v_d(t,\theta) := \inf_{y} \{\, \kappa^\circ(A^T y + t) \mid \langle b, y\rangle - \sigma\rho^\circ(y) \ge 1 + \theta \,\}\,, \tag{2.3.2b}$$

where, as in (G_d), we use the convention described by (2.2.8) and (2.2.9). The parameters u and (t,θ) are perturbations to the primal and dual gauge problems, respectively. This perturbation scheme differs significantly from that used in Fenchel-Rockafellar duality—cf. (2.3.1)—because of the product μu.

We begin by observing that $v_p(0)$ is equal to the optimal value of the primal (G_p). Because u and μ appear as a product in this definition, it is convenient to reparameterize the problem by setting $\lambda := 1/\mu$ and $w := x/\mu$. The positive homogeneity of κ and ρ allows us to equivalently phrase the primal value function as

$$v_p(u) = \inf_{\lambda>0,\, w} \{\, 1/\lambda \mid \rho(\lambda b - Aw + u) \le \sigma\lambda,\ w \in \mathcal{U}_\kappa \,\}\,.$$

In particular, this reparameterization shows that the value function v_p is convex because it is the infimal projection of a convex function, and it is proper when the primal (G_p) is feasible.

We now construct the function F appearing in theorem 2.3.1 associated with this duality framework. In this construction, we assume that $\sigma > 0$, possibly making the replacement (2.2.8) if $\sigma = 0$. Note that minimizing $1/\lambda$ is equivalent to minimizing $-\lambda$ for $\lambda \ge 0$. Define the convex function $F\colon \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \to \overline{\mathbb{R}}$ by

$$F(w,\lambda,u) := -\lambda + \delta_{(\operatorname{epi}\rho)\times\mathcal{U}_\kappa}\left(W\begin{bmatrix} w \\ \lambda \\ u \end{bmatrix}\right), \quad \text{where} \quad W := \begin{bmatrix} -A & b & I_m \\ 0 & \sigma & 0 \\ I_n & 0 & 0 \end{bmatrix}.$$

Observe that the matrix W is nonsingular.

Because $(0,0,0) \in \operatorname{dom} F$, and κ and ρ are closed, the function F is closed and proper. This pairing function gives rise to the infimal projection problems

$$p(u) := \inf_{\lambda\ge 0, w} F(w,\lambda,u) \quad \text{and} \quad q(t,\theta) := \inf_{y} F^\star(t,\theta,y), \tag{2.3.3}$$

which correspond to the general definitions shown in (2.1.1). Note that the function p is the reciprocal of v_p, as formalized in the following lemma (stated without proof).

Lemma 1. *Equality $v_p(u) = -1/p(u)$ holds provided that $v_p(u)$ is nonzero and finite. Moreover, $v_p(u) = 0$ if and only if $p(u) = -\infty$, and $p(u) = 0$ if and only if $v_p(u) = +\infty$.*

We now compute the conjugate of F, which is needed to derive the dual value function q. By Rockafellar and Wets [23, Theorem 11.23(b)],

$$F^\star(t,\theta,y) = \operatorname{cl} \inf_{z,\beta,r} \left\{ \delta^\star_{(\operatorname{epi}\rho)\times\mathcal{U}_\kappa} \begin{pmatrix} z \\ \beta \\ r \end{pmatrix} \;\middle|\; W^T \begin{bmatrix} z \\ \beta \\ r \end{bmatrix} = \begin{bmatrix} t \\ \theta \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\},$$

where the closure operation cl is applied to the function on the right-hand side with respect to the argument (t,θ,y). Using the definition of W, the constraint in the description of $F^\star$ is precisely $(r - A^T z, \langle b, z\rangle + \sigma\beta, z) = (t, \theta+1, y)$, and the unique vector that satisfies these constraints is $(z,\beta,r) = (y,\ \sigma^{-1}(\theta + 1 - \langle b, y\rangle),\ t + A^T y)$. The closure operation is therefore superfluous, and we obtain

$$\begin{aligned} F^\star(t,\theta,y) &= \delta^\star_{(\operatorname{epi}\rho)\times\mathcal{U}_\kappa} \begin{pmatrix} y \\ \sigma^{-1}(1+\theta-\langle b,y\rangle) \\ t + A^T y \end{pmatrix} \\ &= \delta^\star_{\operatorname{epi}\rho} \begin{pmatrix} y \\ \sigma^{-1}(1+\theta-\langle b,y\rangle) \end{pmatrix} + \delta^\star_{\mathcal{U}_\kappa}(t + A^T y). \end{aligned}$$

Since $\delta^\star_{\operatorname{epi}\rho}(z_1,z_2) = \delta_{\operatorname{epi}\rho^\circ}(z_1,-z_2)$ and $\delta^\star_{\mathcal{U}_\kappa} = \kappa^\circ$ by (2.2.3) and (2.2.4), this reduces to

$$F^\star(t,\theta,y) = \delta_{\operatorname{epi}\rho^\circ} \begin{pmatrix} y \\ -\sigma^{-1}(1+\theta-\langle b,y\rangle) \end{pmatrix} + \kappa^\circ(t + A^T y).$$

The application of theorem 2.3.1 asks that we evaluate these conjugates at $(t,\theta) = (0,0)$, which yields the expression

$$F^\star(0,0,y) = \begin{cases} \kappa^\circ(A^T y) & \text{if } \ \langle b,y\rangle - \sigma\rho^\circ(y) \geq 1 \\ +\infty & \text{otherwise.} \end{cases}$$

Thus, the dual problem

$$-q(0,0) = -\inf_y F^\star(0,0,y) = \sup_y -F^\star(0,0,y)$$

recovers, up to a sign change, the required gauge dual problem (G_d) when $\sigma > 0$. When $\sigma = 0$, we also recover the gauge dual problem (G_d) by making the appropriate substitutions (2.2.8) under the convention (2.2.9).

This discussion justifies the definition of the dual perturbation function $v_d(t,\theta) := \inf_y F^\star(t,\theta,y)$, which is equivalent to the expression (2.3.2b). Note that $v_d(0,0)$ is the optimal value of (G_d). In summary, $(-1/v_p)$ and v_d, respectively, play the roles of p and q as defined in (2.3.3). In the application of theorem 2.3.1, we identify x with (w,λ), and v with (t,θ).

2.3.3 Proof of gauge duality

We now use the perturbation framework from section 2.3.2 to prove weak and strong duality results for the gauge duality setting. theorem 2.3.2 [15, section 5] is already known, but the proof via perturbation is new.

The following auxiliary result ties the feasibility of the gauge pair (G_p) and (G_d) to the domain of the value function. The proof of this result, which is largely an application of the calculus of relative interiors, is deferred to section 2.8.

Lemma 2 (Feasibility and domain of the value function)**.** *If the primal* (G_p) *is relatively strictly feasible, then* $0 \in \operatorname{ri}\operatorname{dom} p$. *If the dual* ($G_d$) *is relatively strictly feasible, then* $0 \in \operatorname{ri}\operatorname{dom} v_d$. *The analogous implications, where the* ri *operator is replaced by the* int *operator, hold under strict feasibility (not relative).*

The duality relations in the gauge framework follow analogous principles to Lagrange duality, except that instead of an additive relationship between the primal and dual optimal values the relationship is multiplicative. The following theorem summarizes weak and strong duality for gauge optimization.

Theorem 2.3.2 (Gauge duality [15])**.** *Set $\nu_p := v_p(0)$ and $\nu_d := v_d(0,0)$. Then the following relationships hold for the gauge primal-dual pair (G_p) and (G_d).*

(a) (Basic Inequalities) It is always the case that

$$(i) \quad (1/\nu_p) \leq \nu_d \quad \text{and} \quad (ii) \quad (1/\nu_d) \leq \nu_p.$$

In particular, if $\nu_p = 0$ (resp. $\nu_d = 0$), then (G_d) (resp. (G_p)) is infeasible.

(b) (Weak duality) If x and y are primal and dual feasible, then

$$1 \leq \nu_p \nu_d \leq \kappa(x) \cdot \kappa^\circ(A^T y).$$

(c) (Strong duality) If the dual (resp. primal) is feasible and the primal (resp. dual) is relatively strictly feasible, then $\nu_p \nu_d = 1$ and the gauge dual (resp. primal) attains its optimal value.

Proof. To simplify notation, in this proof we denote the optimal value of the primal value function by $p_0 \equiv p(0)$.

Part (a). We begin with the inequality (i). theorem 2.3.1 guarantees the inequality

$$p_0 \geq -\inf_y F^\star(0,0,y) = -\nu_d. \tag{2.3.4}$$

By lemma 1, whenever ν_p is nonzero and finite, equality $p_0 = -1/\nu_p$ holds, which together with (2.3.4) yields (i). If, on the other hand, $\nu_p = +\infty$, then (i) is trivial. Finally, if $\nu_p = 0$, lemma 1 yields $p_0 = -\infty$, and hence (2.3.4) implies $\nu_d = +\infty$, and (i) again holds. Thus, (i) holds always. To establish (ii), it suffices to consider the case $\nu_d = 0$. From (2.3.4) we conclude $p_0 \geq 0$, that is either $p_0 = 0$ or $p_0 = +\infty$. By lemma 1, the first case $p_0 = 0$ implies $\nu_p = +\infty$ and therefore (ii) holds. The second case $p_0 = +\infty$ implies that the primal problem is infeasible, that is $\nu_p = +\infty$, and again (ii) holds. Thus (ii) holds always, as required.

Part (b). Because the gauge primal and dual problems are both feasible, ν_p and ν_d are nonzero and finite so the result follows from part (a).

Part (c). Suppose the dual is feasible and the primal is relatively strictly feasible. In particular, both ν_p and ν_d are nonzero and finite by part (a). Hence $1 \leq \nu_p \nu_d = -\nu_d/p_0$. On the other hand, by lemma 2 the assumption that the primal is relatively strictly feasible implies $0 \in \operatorname{ri} \operatorname{dom} p$. This last inequality implies $p_0 = p(0)$ is finite, and hence $p(\cdot)$ is proper. theorem 2.3.1(b) tells us that $p_0 = -\nu_d$ and the infimum in the dual ν_d is attained. Thus we deduce $1 = \nu_p \nu_d$, as claimed.

Conversely, suppose that the primal is feasible and the dual is relatively strictly feasible. Then, by lemma 2, $0 \in \operatorname{ri} \operatorname{dom} q$. This in turn implies $p_0 = -\nu_d$ and that the infimum in $p(0)$ is attained. Since the primal is feasible, by lemma 1, p_0 is nonzero, and hence $1 = \nu_p \nu_d$ and the infimum in the primal is attained. □

2.3.4 Gauge optimality conditions

Our perturbation framework can be harnessed to develop optimality conditions for the gauge pair that relate the primal-dual solutions to subgradients of the corresponding value function. This yields a version of parts (b) and (d) in theorem 2.3.1 that are specialized to gauge duality.

Theorem 2.3.3 (Gauge multipliers and sensitivity)**.** *The following relationships hold for the gauge primal-dual pair* (G_p) *and* (G_d).

(a) If the primal is relatively strictly feasible and the dual is feasible, then the set of optimal solutions for the dual is nonempty and coincides with

$$\partial p(0) = \partial(-1/v_p)(0).$$

If it is further assumed that the primal is strictly feasible, then the set of optimal solutions to the dual is bounded.

(b) If the dual is relatively strictly feasible and the primal is feasible, then the set of optimal solutions for the primal is nonempty with solutions $x^* = w^*/\lambda^*$, where

$$(w^*, \lambda^*) \in \partial v_d(0,0) \quad \textit{and} \quad \lambda^* > 0.$$

If it is further assumed that the dual is strictly feasible, then the set of optimal solutions to the primal is bounded.

Proof. Part (a). Because (G_p) is relatively strictly feasible, it follows from lemma 2 that $0 \in \operatorname{ri}\operatorname{dom} p$, and because the dual is feasible, $p(0)$ is finite. theorem 2.3.1 and lemma 1 then imply the conclusion of Part (a). The statement on the boundedness of the set of the optimal solutions to the dual follows from theorem 2.3.1.

Part (b). Because (G_d) is relatively strictly feasible, it follows from lemma 2 that $0 \in \operatorname{ri}\operatorname{dom} v_d$, and because the primal is feasible, $v_d(0)$ is finite. theorem 2.3.1 then implies that the optimal primal set is nonempty, and $\operatorname{argmin}_{w,\lambda} F(w, \lambda, 0) = \partial v_d(0, 0)$. Because the primal and dual problems are feasible, any pair $(w^*, \lambda^*) \in \operatorname{argmin}_{w,\lambda} F(w, \lambda, 0)$ must satisfy $\lambda^* > 0$ by theorem 2.3.2 and lemma 1. Thus, this inclusion is equivalent to $x^* = w^*/\lambda^*$ being optimal for the primal problem, with optimal value $1/\lambda^*$. This proves Part (b). The statement on the boundedness of the set of the optimal solutions to the primal again follows from theorem 2.3.1. □

We use the sensitivity interpretation given by theorem 2.3.3 to develop a set of necessary and sufficient optimality conditions that mirror the more familiar KKT conditions from Lagrange duality. For a primal-dual optimal pair (x^*, y^*), the condition $\rho^\circ(y^*) = 0$ characterizes a degenerate case when $\sigma > 0$ because in that case the primal constraint is inactive at x^* (i.e., $\rho(b - Ax^*) < \sigma$). On the other hand, the dual constraint is always active at optimality because the positive homogeneity of the dual objective and the dual constraint imply $\langle b, y^* \rangle - \sigma\rho^\circ(y^*) = 1$. The full primal-dual optimality conditions for gauge duality are described in the following theorem.

Theorem 2.3.4 (Optimality conditions)**.** *Suppose both problems of the gauge dual pair* (G_p) *and* (G_d) *are relatively strictly feasible, and the pair* (x^*, y^*) *is primal-dual feasible. Then*

(x^*, y^*) is primal-dual optimal if and only if it satisfies the conditions

$$\rho(b - Ax^*) = \sigma \quad or \quad \rho^\circ(y^*) = 0 \qquad \textit{(primal activity)} \tag{2.3.5a}$$

$$\langle b, y^* \rangle - \sigma\rho^\circ(y^*) = 1 \qquad \textit{(dual activity)} \tag{2.3.5b}$$

$$\langle x^*, A^T y^* \rangle = \kappa(x^*) \cdot \kappa^\circ(A^T y^*) \qquad \textit{(objective alignment)} \tag{2.3.5c}$$

$$\langle b - Ax^*, y^* \rangle = \sigma\rho^\circ(y^*). \qquad \textit{(constraint alignment)} \tag{2.3.5d}$$

Proof. First suppose that $(\bar{x}, \bar{y})$ satisfies (2.3.5a)-(2.3.5d). By theorem 2.3.2, to show that $(\bar{x}, \bar{y})$ is primal-dual optimal it is sufficient to show that $\kappa(\bar{x}) \cdot \kappa^\circ(A^T\bar{y}) = 1$. Add (2.3.5c) and (2.3.5d) to obtain

$$\langle b, \bar{y} \rangle = \kappa(\bar{x}) \cdot \kappa^\circ(A^T\bar{y}) + \sigma\rho^\circ(\bar{y}).$$

By combining the above with (2.3.5b) we obtain $\kappa(\bar{x}) \cdot \kappa^\circ(A^T\bar{y}) = 1$, as desired.

Suppose now that (x^*, y^*) is primal-dual optimal. We begin by assuming that $\sigma > 0$ and obtain the case $\sigma = 0$ by applying the result for the $\sigma > 0$ case under the replacement (2.2.8). By the positive homogeneity of κ° and the optimality of y^*, (2.3.5b) holds. Also note that $\kappa(x^*)$ and $\kappa^\circ(A^T y^*)$ are both nonzero and finite because of the strong duality guaranteed by theorem 2.3.2.

Define $\lambda^* := 1/\kappa(x^*)$ and $w^* := \lambda^* x^*$, so that $\kappa(w^*) = 1$. By theorem 2.3.1(e) and theorem 2.3.3(b), we must have $(0, 0, y^*) \in \partial F(w^*, \lambda^*, 0)$. Since the primal problem is relatively strictly feasible, we can apply [20, Theorem 23.9] to deduce the characterization

$$\partial F(w, \lambda, 0) = -\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + W^T \mathcal{N}_{(\operatorname{epi}\rho)\times\mathcal{U}_k} \begin{pmatrix} \lambda b - Aw \\ \sigma\lambda \\ w \end{pmatrix}, \tag{2.3.6}$$

where $\mathcal{N}_\mathcal{C}(\cdot)$ denotes the normal cone to a set $\mathcal{C}$. We now consider two cases. First, suppose $\rho(\lambda^* b - Aw^*) = \lambda^*\sigma$. Then (2.3.5a) holds, and by straightforward computations involving

only (2.2.4) and the definitions of normal cones and subdifferentials, we have

$$\mathcal{N}_{(\operatorname{epi}\rho)\times\mathcal{U}_k}\begin{pmatrix}\lambda^*b - Aw^*\\ \sigma\lambda^*\\ w^*\end{pmatrix} = \mathcal{N}_{\operatorname{epi}\rho}\begin{pmatrix}\lambda^*b - Aw^*\\ \sigma\lambda^*\end{pmatrix} \times \mathcal{N}_{\mathcal{U}_k}(w^*),$$

where

$$\mathcal{N}_{\operatorname{epi}\rho}\begin{pmatrix}\lambda^*b - Aw^*\\ \sigma\lambda^*\end{pmatrix} = \operatorname{cone}\left(\partial\rho(\lambda^*b - Aw^*) \times \{-1\}\right)$$

and $\mathcal{N}_{\mathcal{U}_k}(w^*) = \{\, v \mid \kappa^\circ(v) \le \langle v, w^*\rangle \,\}$. Substitute these formulas into (2.3.6) to obtain

$$\partial F(w^*,\lambda^*,0) = \left\{ \begin{bmatrix} v - \mu A^T z\\ \mu(\langle b, z\rangle - \sigma) - 1\\ \mu z\end{bmatrix} \;\middle|\; \mu \ge 0,\ z \in \partial\rho(\lambda^*b - Aw^*),\ v \in \mathcal{N}_{\mathcal{U}_\kappa}(w^*) \right\}.$$

We deduce the existence of $z^* \in \partial\rho(\lambda^*b - Aw^*)$ and $\mu^* \ge 0$ such that

$$y^* = \mu^* z^* \tag{2.3.7a}$$

$$\mu^*(\langle b, z^*\rangle - \sigma) = 1 \tag{2.3.7b}$$

$$\kappa^\circ(\mu^* A^T z^*) \le \langle \mu^* A^T z^*, w^*\rangle. \tag{2.3.7c}$$

Note that $\mu^* = 0$ cannot satisfy (2.3.7b), hence (2.3.7c), together with the polar-gauge inequality and the fact that $\kappa(w^*) = 1$, implies

$$\kappa^\circ(A^T y^*)\cdot\kappa(w^*) = \kappa^\circ(A^T y^*) \le \langle A^T y^*, w^*\rangle \le \kappa^\circ(A^T y^*)\cdot\kappa(w^*).$$

Equality must hold in the above, and dividing through by $\lambda^* > 0$ we see that (2.3.5c) is satisfied. Finally, we aim to show that (2.3.5d) holds using the fact that $y^* \in \mu^*\partial\rho(\lambda^*b - Aw^*)$. From the characterization (2.2.4) of the polar, we have

$$\partial\rho(u) = \operatorname*{argmax}_y \{\, \langle y, u\rangle \mid \rho^\circ(y) \le 1 \,\}. \tag{2.3.8}$$

In particular, this characterization implies $\langle y^*/\mu^*, \lambda^*b - Aw^*\rangle \ge \langle 0, \lambda^*b - Aw^*\rangle = 0$. If $\rho(\lambda^*b - Aw^*) = 0$, then by the polar-gauge inequality (2.2.5) we have

$$0 \le \langle y^*, \lambda^*b - Aw^*\rangle \le \rho(\lambda^*b - Aw^*)\cdot\rho^\circ(y^*) = 0,$$

which gives condition (2.3.5d) after dividing through by λ^*. On the other hand, if $\rho(u) > 0$ then the set (2.3.8) is given by $\{\, y \mid \rho(u) = \langle y, u\rangle,\ \rho^\circ(y) = 1 \,\}$. Thus when $\rho(\lambda^* b - Aw^*) > 0$, we again have $\langle y^*/\mu^*, \lambda^* b - Aw^*\rangle = \rho^\circ(y^*/\mu^*) \cdot \rho(\lambda^* b - Aw^*)$, and multiplying through by μ^*/λ^* and applying (2.3.5a) gives (2.3.5d).

We have shown the forward implication of the theorem when $\rho(\lambda^* b - Aw^*) = \lambda^* \sigma$. The other case we need to consider is when $\rho(\lambda^* b - Aw^*) < \lambda^* \sigma$, or equivalently when $\rho(b - Ax^*) < \sigma$. An easy argument (e.g., see [12, Proposition 2.14(iv)]) shows

$$\mathcal{N}_{\operatorname{epi}\rho}(\lambda^* b - Aw^*, \lambda^* \sigma) = \mathcal{N}_{\operatorname{dom}\rho}(\lambda^* b - Aw^*) \times \{0\}.$$

Similar to the first case, we now have

$$\partial F(w^*, \lambda^*, 0) = \left\{ \begin{bmatrix} v - A^T z \\ \langle b, z\rangle - 1 \\ z \end{bmatrix} \;\middle|\; z \in \mathcal{N}_{\operatorname{dom}\rho}(\lambda^* b - Aw^*),\ v \in \mathcal{N}_{\mathcal{U}_\kappa}(w^*) \right\}.$$

We deduce that $y^* \in \mathcal{N}_{\operatorname{dom}\rho}(\lambda^* b - Aw^*)$ and also that $\langle b, y^*\rangle = 1$ and $\kappa^\circ(A^T y^*) \le \langle A^T y^*, w^*\rangle$. Again, because $\kappa(w^*) = 1$, the polar-gauge inequality implies (2.3.5c) holds.

We now show that $\rho^\circ(y^*) = 0$ and $\langle b - Ax^*, y^*\rangle = 0$, which, if true, establishes (2.3.5a) and (2.3.5d) are satisfied as well. First note that $y^* \in \mathcal{N}_{\operatorname{dom}\rho}(u)$ implies $y^* \in (\operatorname{dom}\rho)^\circ$, which implies $\rho^\circ(y^*) = 0$ by (2.2.6) . Thus, by (2.3.5b), (2.3.5c), and the fact that $\kappa(x^*) \cdot \kappa^\circ(A^T y^*) = 1$ from theorem 2.3.2, we have

$$\langle b - Ax^*, y^*\rangle = 1 - \kappa(x^*) \cdot \kappa^\circ(A^T y^*) = 0.$$

Thus if (x^*, y^*) is primal-dual optimal, then (2.3.5a)-(2.3.5d) hold, as claimed. This finishes the proof for $\sigma > 0$.

Let us now consider the case when $\sigma = 0$ and apply what we have just proved to the pair (G_p) and (G_d) under the replacement (2.2.8). Then (x^*, y^*) is primal-dual optimal if and only if the conditions (2.3.5a)-(2.3.5d) hold with $(\rho, \sigma) = (\delta_{\mathcal{H}_\rho}, 1)$, i.e.,

$$\begin{aligned} \delta_{\mathcal{H}_\rho^\circ}(y^*) &= 0, \qquad & \langle x^*, A^T y^*\rangle &= \kappa(x^*) \cdot \kappa^\circ(A^T y^*), \\ \langle b, y^*\rangle - \delta_{\mathcal{H}_\rho^\circ}(y^*) &= 1, \qquad & \langle b - Ax^*, y^*\rangle &= 0. \end{aligned}$$

If we combine this with primal feasibility, $\rho(b - Ax^*) = 0$, and use the identity (2.2.9) that $0 \cdot \rho^\circ = \delta_{\mathcal{H}_\rho^\circ}$, then these conditions are equivalent to (2.3.5a)-(2.3.5d) for $\sigma = 0$, ρ, and ρ° as written above. □

The following corollary describes a variation of the optimality conditions outlined by theorem 2.3.4. These conditions assume that a solution y^* of the dual problem is available, and gives conditions that can be used to determine a corresponding solution of the primal problem. An application of the following result appears in section 2.6.

Corollary 1 (Gauge primal-dual recovery)**.** *Suppose that the primal-dual pair* (G_p) *and* (G_d) *are each relatively strictly feasible. If y^* is optimal for* (G_d)*, then for any primal feasible x the following conditions are equivalent:*

(a) x is optimal for (G_p)*;*

(b) $\langle x, A^T y^* \rangle = \kappa(x) \cdot \kappa^\circ(A^T y^*)$ and $b - Ax \in \partial(\sigma\rho^\circ)(y^*)$;

(c) $A^T y^* \in \kappa^\circ(A^T y^*) \cdot \partial\kappa(x)$ and $b - Ax \in \partial(\sigma\rho^\circ)(y^*)$,

where, by convention, $\sigma\rho^\circ = \delta_{\operatorname{cl}\operatorname{dom}\rho^\circ}$ when $\sigma = 0$.

Proof. We use the optimality conditions given in theorem 2.3.4. As noted before, by the optimality of y^* we automatically have equality (2.3.5b) in the dual constraint.

We first show that (b) implies (a). Suppose (b) holds. Then (2.3.5c) holds automatically. From the characterization (2.2.4) of the polar, we have

$$\sigma\rho^\circ(y^*) = \left\{ \begin{array}{ll} \sigma \cdot \sup\limits_{\rho(z)\leq 1} \langle y^*, z\rangle & \text{if } \sigma > 0 \\ \\ \delta_{\operatorname{cl}\operatorname{dom}\rho^\circ}(y^*) & \text{if } \sigma = 0 \end{array} \right\} = \sup_{\rho(z)\leq\sigma} \langle y^*, z\rangle, \tag{2.3.9}$$

where the case $\sigma = 0$ uses the convention (2.2.9). Thus, $\partial(\sigma\rho^\circ)(y^*)$ is the set of maximizing elements in this supremum. Because $b - Ax \in \partial(\sigma\rho^\circ)(y^*)$, it holds that $\rho(b - Ax) \leq \sigma$. If we additionally use the polar-gauge inequality, we deduce that

$$\sigma\rho^\circ(y^*) = \langle y^*, b - Ax\rangle \leq \rho(b - Ax) \cdot \rho^\circ(y^*) \leq \sigma\rho^\circ(y^*),$$

and therefore the above inequalities are all tight. Thus conditions (2.3.5a) and (2.3.5d) hold, and by theorem 2.3.4, (x, y) is a primal-dual optimal pair.

We next show that (a) implies (b). Suppose that x is optimal for (G_p). Then the first condition of (b) holds by (2.3.5c), and (2.3.5a) and (2.3.5d) combine to give us

$$\sigma\rho^\circ(y^*) = \rho(b - Ax) \cdot \rho^\circ(y^*) = \langle b - Ax, y^* \rangle.$$

This implies that $z := b - Ax$ is a maximizing element of the supremum in (2.3.9), and thus $b - Ax \in \partial(\sigma\rho^\circ)(y^*)$.

Finally, to show the equivalence of (b) and (c), note that by the polar-gauge inequality, $\langle x, A^T y^* \rangle = \kappa(x) \cdot \kappa^\circ(A^T y^*)$ if and only if x minimizes the convex function $\kappa^\circ(A^T y^*)\,\kappa(\cdot) - \langle \cdot, A^T y^* \rangle$. This, in turn, is true if and only if $0 \in \kappa^\circ(A^T y^*)\,\partial\kappa(x) - A^T y^*$, or equivalently, $A^T y^* \in \kappa^\circ(A^T y^*) \cdot \partial\kappa(x)$. □

2.3.5 The relationship between Lagrange and gauge multipliers

We now use the perturbation framework for duality to establish a relationship between gauge dual and Lagrange dual variables. We begin with an auxiliary result that characterizes the subdifferential of the perspective function (2.2.1). Combettes [10, Prop. 2.3(v)] also describes an equivalent formula for the subdifferential, though the derivation and subsequent form of the expression are very different. The formula in lemma 3 is more suitable for our purposes.

Lemma 3 (Subdifferential of perspective function)**.** *Let $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a closed proper convex function. Then for $(x, \mu) \in \operatorname{dom} g^\pi$, equality holds:*

$$\partial g^\pi(x, \mu) = \begin{cases} \{\, (z, -g^\star(z)) \mid z \in \partial g(x/\mu) \,\} & \text{if } \mu > 0 \\ \{\, (z, \gamma) \mid (z, -\gamma) \in \operatorname{epi} g^\star,\ z \in \partial g^\infty(x) \,\} & \text{if } \mu = 0. \end{cases}$$

Proof. Recall that the subdifferential of the support function to any nonempty closed convex set $\mathcal{C}$ is given by $\partial\delta^\star_{\mathcal{C}}(x) = \operatorname{argmax}\{\, \langle z, x \rangle \mid z \in \mathcal{C} \,\}$ [20, Theorem 23.5 and Corollary 23.5.3]. By [20, Corollary 13.5.1], $g^\pi = \delta^\star_{\mathcal{C}}$, where $\mathcal{C} = \{\, (z, \gamma) \mid g^\star(z) \le -\gamma \,\}$ is a closed convex set. If

$(x, \mu) \in \operatorname{dom} g^\pi$, then $\mathcal{C}$ is nonempty and

$$\partial g^\pi(x, \mu) = \operatorname*{argmax}_{(z,\gamma)\in\mathcal{C}} \{ \langle (x, \mu), (z, \gamma) \rangle \} = \operatorname*{argmax}_{(z,\gamma)\in\mathcal{C}} \{ \langle z, x \rangle + \mu\gamma \}.$$

Suppose now that $\mu > 0$. Then

$$\sup_{(z,\gamma)\in\mathcal{C}} \{ \langle z, x \rangle + \mu\gamma \} = \sup_{z\in\operatorname{dom} g^\star} \{ \langle z, x \rangle - \mu g^\star(z) \} = \mu \cdot \sup_{z\in\operatorname{dom} g^\star} \{ \langle z, x/\mu \rangle - g^\star(z) \}. \tag{2.3.10}$$

Using the expression for the subdifferential of a support function, (z, γ) achieves the supremum of (2.3.10) if $z \in \partial g(x/\mu)$ and $-\gamma = g^\star(z)$. On the other hand, if $\mu = 0$ then

$$\sup_{(z,\gamma)\in\mathcal{C}} \langle z, x \rangle = \sup_{z\in\operatorname{dom} g^\star} \langle z, x \rangle = \delta^\star_{\operatorname{dom} g^\star}(x) = g^\infty(x).$$

Again using the expression for the subdifferential of a support function, (z, γ) achieves the supremum of (2.3.10) if and only if $z \in \partial g^\infty(x)$ and $(z, -\gamma) \in \operatorname{epi} g^\star$. □

We now state the main result relating the optimal solutions of (G_p) to the optimal solutions of the Lagrange dual of (G_d).

Theorem 2.3.5. *Suppose that the gauge dual (G_d) is relatively strictly feasible and the primal (G_p) is feasible. Let (L_p) denote the Lagrange dual of (G_d), and let ν_L denote its optimal value. Then*

$$z^* \text{ is optimal for } (L_p) \iff z^*/\nu_L \text{ is optimal for } (G_p).$$

Proof. We first note that (L_p) can be derived via the framework of theorem 2.3.1 through the Lagrangian value function

$$h(w) = \inf_y \left\{ \kappa^\circ(A^T y + w) + \delta_{\langle b,\cdot\rangle - \sigma\rho^\circ(\cdot)\geq 1}(y) \right\}.$$

Here h plays the role of p in theorem 2.3.1; cf. [23, Example 11.41]. Strong duality in theorem 2.3.2 guarantees that $h(0)$ is nonzero and finite, and by lemma 2,

$$(G_d) \text{ relatively strictly feasible} \implies (0, 0) \in \operatorname{ri}\operatorname{dom} v_d \implies 0 \in \operatorname{ri}\operatorname{dom} h.$$

Thus, it follows from theorem 2.3.1 that the optimal points z^* for (L_p) are characterized by $z^* \in \partial h(0)$. Note also that $h(0) = \nu_L$.

On the other hand, by theorem 2.3.3(b) the solutions to (G_p) are precisely the points w^*/λ^* such that $(w^*, \lambda^*) \in \partial v_d(0,0)$. Thus to relate the solution sets of (L_p) and (G_p), we must relate $\partial h(0)$ and $\partial v_d(0,0)$.

For θ in a neighborhood of zero and all t, by positive homogeneity of κ° and ρ° we have

$$v_d(t,\theta) = (1+\theta)h\left(\frac{t}{1+\theta}\right) = \inf_y \left\{ (1+\theta)\kappa^\circ\left(A^T y + \frac{t}{1+\theta}\right) + \delta_{\langle b,\cdot\rangle - \sigma\rho^\circ(\cdot)\geq 1}(y) \right\}.$$

Thus by lemma 3, $\partial v_d(0,0) = \{\, (z, -h^\star(z)) \mid z \in \partial h(0) \,\}$. However, for $z \in \partial h(0)$ the Fenchel-Young equality gives us

$$0 = \langle 0, z\rangle = h^\star(z) + h(0) = h^\star(z) + \nu_L.$$

Thus we obtain the convenient description

$$\partial v_d(0,0) = \partial h(0) \times \{h(0)\} = \partial h(0) \times \{\nu_L\}$$

and the set of optimal solutions for (G_p) is precisely $\frac{1}{\nu_L}\partial h(0)$. □

2.4 Perspective duality

We now move on to an extension of the gauge duality framework, which allows us to consider functions that are not necessarily positively homogeneous, but continue to be nonnegative and convex. (The same framework applies to functions that are bounded below because these can be made nonnegative by translation.) For the remainder of the paper, consider functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}_+$ and $g : \mathbb{R}^m \to \overline{\mathbb{R}}_+$, that are closed, convex and nonnegative over their domains. In this section we derive and analyze the *perspective-dual* pair

$$\underset{x}{\text{minimize}} \quad f(x) \qquad \text{subject to} \quad g(b - Ax) \leq \sigma, \tag{N_p}$$

$$\underset{y,\ \alpha,\ \mu}{\text{minimize}} \quad f^\sharp(A^T y, \alpha) \quad \text{subject to} \quad \langle b, y\rangle - \sigma g^\sharp(y,\mu) \geq 1 - (\alpha + \mu). \tag{N_d}$$

The functions $f^\sharp$ and $g^\sharp$ are the polars of the perspective transforms of f and g. This transform is a key operation needed to derive perspective duality. In the next section we describe properties of that transform and its application to the derivation of the perspective-dual pair. Throughout this section, we assume that $\sigma > \inf_u g(u) \geq 0$.

2.4.1 Perspective-polar transform

Given a closed proper convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}_+$, define the perspective-polar transform by $f^\sharp := (f^\pi)^\circ$.

An explicit characterization of the perspective-polar transform is given by

$$f^\sharp(z, -\xi) = \inf \{ \mu > 0 \mid \langle z, x \rangle \leq \xi + \mu f(x), \forall x \} . \tag{2.4.1}$$

This representation can be obtained by applying the definition of the gauge polar (2.2.2) to the perspective transform as follows:

$$\begin{aligned} f^\sharp(z, -\xi) &= \inf \{ \mu > 0 \mid \langle z, x \rangle - \xi\lambda \leq \mu f^\pi(x, \lambda), \forall x, \forall \lambda \} \\ &= \inf \{ \mu > 0 \mid \langle z, x \rangle - \xi\lambda \leq \mu\lambda f(x/\lambda), \forall x, \forall \lambda > 0 \} \\ &= \inf \{ \mu > 0 \mid \langle z, \lambda x \rangle - \xi\lambda \leq \mu\lambda f(x), \forall x, \forall \lambda > 0 \} , \end{aligned}$$

which yields (2.4.1) after dividing through by λ. Rockafellar's extension [20, p.136] of the polar gauge transform to nonnegative convex functions that vanish at the origin coincides with $f^\sharp(z, -1)$.

The following theorem provides an alternative characterization of the perspective-polar transform in terms of the more familiar Fenchel conjugate $f^\star$. It also provides an expression for the perspective-polar of f in terms of the Minkowski function generated by the epigraph of the conjugate of f, i.e.,

$$\gamma_{\operatorname{epi} f^\star}(x, \tau) := \inf \{ \lambda > 0 \mid (x, \tau) \in \lambda \operatorname{epi} f^\star \} ,$$

which is a gauge. Nonnegativity of f is not required for the first part of this result.

Theorem 2.4.1. *For any closed proper convex function f with $0 \in \operatorname{dom} f$, we have $f^{\pi\star}(z,-\xi) = \delta_{\operatorname{epi} f^\star}(z,\xi)$. If, in addition, f is nonnegative, $f^\sharp(z,-\xi) = \gamma_{\operatorname{epi} f^\star}(z,\xi)$.*

Proof. Because of the assumptions on f, we have $f^\pi(x,0) = \liminf_{\lambda\to 0^+} f^\pi(x,\lambda)$ for each $x \in \mathbb{R}^n$ [20, Corollary 8.5.2]. Thus we obtain the following chain of equalities:

$$\begin{aligned} f^{\pi\star}(z,-\xi) &= \sup\{\, \langle z,x\rangle - \lambda\xi - f^\pi(x,\lambda) \mid x\in\mathbb{R}^n,\ \lambda\in\mathbb{R}\,\} \\ &= \sup\{\, \langle z,x\rangle - \lambda\xi - \lambda f(\lambda^{-1}x) \mid x\in\mathbb{R}^n,\ \lambda>0\,\} \\ &= \sup\{\, \langle z,\lambda y\rangle - \lambda\xi - \lambda f(y) \mid y\in\mathbb{R}^n,\ \lambda>0\,\} \\ &= \sup\{\, \lambda\cdot\sup_y\{\,\langle z,y\rangle - \xi - f(y)\,\} \mid \lambda>0\,\} \\ &= \sup\{\, \lambda(f^\star(z)-\xi) \mid \lambda>0\,\} = \delta_{\operatorname{epi} f^\star}(z,\xi). \end{aligned}$$

This proves the first statement. Now additionally suppose that f is nonnegative. Because f^π is closed, it is identical to its biconjugate, and so $f^\pi(x,\lambda) = \delta^\star_{\operatorname{epi} f^\star}(x,-\lambda)$. Also, $\operatorname{epi} f^\star$ is closed and convex, and contains the origin because f is nonnegative. Therefore, it follows from [20, Corollary 15.1.2] that

$$f^\sharp(z,-\xi) \equiv f^{\pi\circ}(z,-\xi) = \delta^{\star\circ}_{\operatorname{epi} f^\star}(z,\xi) = \gamma_{\operatorname{epi} f^\star}(z,\xi).$$

□

The following result relates the level sets of the perspective-polar transform to the level sets of the conjugate perspective. This result is useful in deriving the constraint sets for certain perspective-dual problems for which there is no closed form for the perspective polar; cf. theorem 2.5.4.

Theorem 2.4.2 (Level-set equivalence)**.** *Let $f:\mathbb{R}^n\to\overline{\mathbb{R}}_+$ be a nonnegative, closed proper convex function with $0\in\operatorname{dom} f$. Then, for any $(z,\xi,\mu)\in\mathbb{R}^n\times\mathbb{R}\times\mathbb{R}$,*

$$f^\sharp(z,\xi)\le\mu \quad\iff\quad [\ 0\le\mu \ \text{ and } \ f^{\star\pi}(z,\mu)\le -\xi\ ].$$

Proof. The following chain of equivalences follows from theorem 2.4.1:

$$\begin{aligned} f^\sharp(z,\xi) \le \mu &\iff \gamma_{\operatorname{epi} f^\star}(z,-\xi) \le \mu \\ &\iff \inf\{\,\lambda>0 \mid (z,-\xi) \in \lambda \operatorname{epi} f^\star\,\} \le \mu \\ &\iff \inf\{\,\lambda>0 \mid f^\star(z/\lambda) \le -\xi/\lambda\,\} \le \mu \\ &\iff \inf\{\,\lambda>0 \mid f^{\star\pi}(z,\lambda) \le -\xi\,\} \le \mu. \end{aligned} \tag{2.4.2}$$

Define $\alpha = \inf\{\,\lambda>0 \mid f^{\star\pi}(z,\lambda) \le -\xi\,\}$.

We first show that $f^\sharp(z,\xi) \le \mu$ implies $0 \le \mu$ and $f^{\star\pi}(z,\mu) \le -\xi$. By (2.4.2), $0 \le \alpha \le \mu$. If $\alpha < \mu$, there exists λ with $0 < \lambda < \mu$ such that $f^{\star\pi}(z,\lambda) \le -\xi$. Because f is nonnegative, $\mu f \ge \lambda f$, and thus $(\mu f)^\star \le (\lambda f)^\star$. In particular,

$$f^{\star\pi}(z,\mu) = (\mu f)^\star(z) \le (\lambda f)^\star(z) = f^{\star\pi}(z,\lambda) \le -\xi.$$

On the other hand, if $\alpha = \mu$, there exists a sequence $\lambda_k \to \mu$ such that $f^{\star\pi}(z,\lambda_k) \le -\xi$ for each k. Now by the lower semi-continuity of $f^{\star\pi}$, we obtain

$$f^{\star\pi}(z,\mu) \le \liminf_{k\to\infty} f^{\star\pi}(z,\lambda_k) \le -\xi.$$

This establishes the forward implication of the theorem.

For the reverse implication, suppose $0 \le \mu$ and $f^{\star\pi}(z,\mu) \le -\xi$. If $0 < \mu$, it follows from (2.4.2) that $f^\sharp(z,\xi) \le \mu$. Now suppose otherwise that $\mu = 0$. We want to show $f^\sharp(z,\xi) \le 0$. By hypothesis, $(z,0,-\xi) \in \operatorname{epi} f^{\star\pi}$. Thus there exists a sequence (z_k,μ_k,r_k) with $\lim_{k\to\infty}(z_k,\mu_k,r_k) = (z,0,-\xi)$ and $f^{\star\pi}(z_k,\mu_k) \le r_k$ for all k. With no loss in generality, we can assume that $\mu_k > 0$ for all k. Then for each k, we have $\mu_k f^\star(z_k/\mu_k) \le r_k$ for which we have the following equivalences:

$$\begin{aligned} \mu_k f^\star(z_k/\mu_k) \le r_k &\iff \sup_w \{\,\langle w, z_k\rangle - \mu_k f(w)\,\} \le r_k \\ &\iff \langle w, z_k\rangle \le r_k + \mu_k f(w),\ \forall w \in \mathbb{R}^n \\ &\iff \mu_k \ge \inf\{\,\lambda>0 \mid \langle w, z_k\rangle \le r_k + \lambda f(w),\ \forall w \in \mathbb{R}^n\,\} \\ &\underset{(2.4.1)}{\iff} \mu_k \ge f^\sharp(z_k,-r_k), \end{aligned}$$

which gives $f^\sharp(z,\xi) \le 0 = \mu$ in the limit, since $f^\sharp$ is closed. □

Calculus rules

Two useful calculus rules are now developed that govern the perspective-polar transform when applied to gauge functions and separable sums.

Example 2.4.3 (Gauge functions)**.** *Suppose that f is a closed proper gauge. Then*

$$f^{\sharp}(z,\xi) = f^{\circ}(z) + \delta_{\mathbb{R}_-}(\xi).$$

Use expression (2.4.1) for this derivation. When $\xi > 0$, take $x = 0$ in the infimum in (2.4.1) to deduce that $f^{\sharp}(z,\xi) = +\infty$. On the other hand, when $\xi \leq 0$, the positive homogeneity of f implies that $f^{\sharp}(z,\xi) = f^{\circ}(z)$. We leave the details to the reader. More generally, if f vanishes at the origin, then $f^{\sharp}(z,\xi) = +\infty$ for all $\xi > 0$.

Example 2.4.4 (Separable sums)**.** *Suppose that $f(x) := \sum_{i=1}^{n} f_i(x_i)$, where each convex function $f_i : \mathbb{R}^{n_i} \to \overline{\mathbb{R}}_+$ is nonnegative. Then a straightforward computation shows that $f^{\pi}(x,\lambda) = \sum_{i=1}^{n} f_i^{\pi}(x_i,\lambda)$. Furthermore, taking into account [15, Proposition 2.4], which expresses the polar of a separable sum of gauges, we deduce*

$$f^{\sharp}(z,\xi) = \max_{i=1,\ldots,n} f_i^{\sharp}(z_i,\xi).$$

2.4.2 Derivation of the perspective dual via lifting

We now derive the relationship between the primal and dual problems (N_p) and (N_d) by lifting (N_p) to an equivalent gauge optimization problem, and then recognizing (N_d) as its gauge dual.

Theorem 2.4.5 (Gauge lifting of the primal)**.** *A point x^* is optimal for (N_p) if and only if $(x^*, 1)$ is optimal for the gauge problem*

$$\underset{x,\lambda}{\text{minimize}} \quad f^{\pi}(x,\lambda) \quad \text{subject to} \quad \rho\left(\begin{bmatrix} b \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} A & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix}\right) \leq \sigma, \tag{2.4.3}$$

where $\rho(z,\mu,\tau) := g^{\pi}(z,\tau) + \delta_{\{0\}}(\mu)$ is a gauge function.

Proof. By definition of f^π, x^* is optimal for (N_p) if and only if the pair $(x^*, 1)$ is optimal for

$$\underset{x,\lambda}{\text{minimize}} \quad f^\pi(x,\ \lambda) \quad \text{subject to} \quad \lambda = 1,\ g^\pi(b - Ax, 1) \le \sigma.$$

The following equivalence follows from the definition of ρ:

$$[\,\lambda = 1 \quad \text{and} \quad g^\pi(b - Ax, 1) \le \sigma\,] \quad \Longleftrightarrow \quad \rho(b - Ax, 1 - \lambda, 1) \le \sigma.$$

Thus we arrive at the constraint expressed in (2.4.3). □

Corollary 2 (Gauge dual)**.** *Problem* (N_d) *is the gauge dual of* (2.4.3)*.*

Proof. It follows from the canonical dual pairing (G_p) and (G_d) that the gauge dual of (2.4.3) is

$$\begin{aligned} \underset{y,\alpha,\mu}{\text{minimize}} \quad & f^{\pi\circ}\left(\begin{bmatrix} A^T & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} y \\ \alpha \\ \mu \end{bmatrix}\right) \\ \text{subject to} \quad & \langle (y,\alpha,\mu), (b,1,1)\rangle - \sigma\rho^\circ(y,\alpha,\mu) \ge 1. \end{aligned} \tag{2.4.4}$$

Because ρ is separable in (z, μ) and β, it follows from [15, Proposition 2.4] that

$$\rho^\circ(y,\alpha,\mu) = \max\{\, g^{\pi\circ}(y,\mu),\ \delta^\circ_{\{0\}}(\alpha)\,\}.$$

Since $\delta^\circ_{\{0\}}(\alpha)$ is identically zero, the result follows. □

The next result generalizes theorem 2.3.2 to the case where f and g are convex and nonnegative but not necessarily gauges. We use a construction parallel to the one in (2.2.7), and for this section only redefine the feasible sets by

$$\begin{aligned} \mathcal{F}_p &:= \{\, u \mid g(b - u) \le \sigma \,\} \\ \mathcal{F}_d &:= \{\, (y,\alpha,\mu) \mid \langle b, y\rangle - \sigma g^\sharp(y,\mu) \ge 1 - (\alpha + \mu) \,\}. \end{aligned}$$

Thus, (N_p) is *relatively strictly feasible* if

$$A^{-1}\,\mathrm{ri}\,\mathcal{F}_p \cap (\mathrm{ri}\,\mathrm{dom}\, f) \ne \emptyset.$$

Similarly, (N_d) is *relatively strictly feasible* if there exists a triple (y, α, μ) such that

$$(A^T y, \alpha) \in \operatorname{ri} \operatorname{dom} f^\sharp \quad \text{and} \quad (y, \alpha, \mu) \in \operatorname{ri} \mathcal{F}_d.$$

Strict feasibility follows the same definitions, where the operation ri is replaced by int.

Theorem 2.4.6 (Perspective duality)**.** *Let ν_p and ν_d, respectively, denote the optimal values of the pair (N_p) and (N_d). Then the following relationships hold for the perspective dual pair (N_p) and (N_d).*

(a) (Basic Inequalities) It is always the case that

$$\textit{(i)}\ (1/\nu_p) \leq \nu_d \quad \textit{and} \quad \textit{(ii)}\ (1/\nu_d) \leq \nu_p.$$

Thus, $\nu_p = 0$ and $\nu_d = 0$, respectively, imply that (N_d) and (N_p) are infeasible.

(b) (Weak duality) If x and (y, α, μ) are primal and dual feasible, then

$$1 \leq \nu_p \nu_d \leq f(x) \cdot f^\sharp(A^T y, \alpha).$$

(c) (Strong duality) If the dual (resp. primal) is feasible and the primal (resp. dual) is relatively strictly feasible, then $\nu_p \nu_d = 1$ and the perspective dual (resp. primal) attains its optimal value.

Proof. Parts (a) and (b) follow immediately from the analogous result in theorem 2.3.2, together with theorem 2.4.5 and corollary 2.

Next we demonstrate that (N_p) is relatively strictly feasible if and only if (2.4.3) is relatively strictly feasible. By the description of relative interiors of sublevel sets given in [20, Theorem 7.6], (2.4.3) is relatively strictly feasible if and only if there exists a point $(x, 1) \in \operatorname{ri} \operatorname{dom} f^\pi$ such that

$$(b - Ax, 0, 1) \in \operatorname{ri} \operatorname{dom} \rho \quad \text{and} \quad \rho(b - Ax, 0, 1) = g(b - Ax) < \sigma.$$

We now seek a description of $\operatorname{ri}\operatorname{dom} f^\pi$. We have

$$\begin{aligned}\operatorname{dom} f^\pi &= \{\,(x,\mu) \mid f^\pi(x,\mu) < \infty\,\} \\ &= \operatorname{cl}\left(\{0\} \cup \{\,(x,\mu) \mid \mu > 0,\ f(x/\mu) < \infty\,\}\right) = \operatorname{cl}\operatorname{cone}\left(\operatorname{dom} f \times \{1\}\right).\end{aligned}$$

By [20, Corollary 6.8.1], the above description yields

$$\operatorname{ri}\operatorname{dom} f^\pi = \{\,(x,\mu) \mid \mu > 0,\ x \in \mu \operatorname{ri}\operatorname{dom} f\,\}.$$

Thus $(x,1) \in \operatorname{ri}\operatorname{dom} f^\pi$ if and only if $x \in \operatorname{ri}\operatorname{dom} f$. Similarly,

$$\operatorname{dom}\rho = \{\,(y,0,\mu) \mid (y,\mu) \in \operatorname{dom} g^\pi\,\},$$

and so

$$\operatorname{ri}\operatorname{dom}\rho = \{\,(y,0,\mu) \mid (y,\mu) \in \operatorname{ri}\operatorname{dom} g^\pi\,\} = \{\,(y,0,\mu) \mid \mu > 0, y \in \mu \operatorname{ri}\operatorname{dom} g\,\}.$$

In particular, the condition $(b - Ax, 0, 1) \in \operatorname{ri}\operatorname{dom}\rho$ is equivalent to $b - Ax \in \operatorname{ri}\operatorname{dom} g$. Thus the conditions for relative strict feasibility of (2.4.3) and (N_p) are identical.

A similar argument verifies that (N_d) is relatively strictly feasible if and only if (2.4.4) is relatively strictly feasible. Strong duality then follows from relative interiority, corollary 2, theorem 2.4.5, and the analogous strong-duality result in theorem 2.3.2. □

2.4.3 Optimality conditions

The following result generalizes theorem 2.3.3 to include the perspective-dual pair.

Theorem 2.4.7 (Perspective optimality)**.** *Suppose* (N_p) *is strictly feasible. Then the tuple* $(x^*, y^*, \alpha^*, \mu^*)$ *is perspective primal-dual optimal if and only if*

$$\begin{aligned}
g(b - Ax^*) = \sigma \quad \text{or} \quad g^\sharp(y^*, \mu^*) = 0 \qquad & \textit{(primal activity)} \\
\langle b, y^*\rangle - \sigma g^\sharp(y^*, \mu^*) = 1 - (\alpha^* + \mu^*) \qquad & \textit{(dual activity)} \\
\langle x^*, A^T y^*\rangle + \alpha^* = f(x^*) \cdot f^\sharp(A^T y^*, \alpha^*) \qquad & \textit{(objective alignment)} \\
\langle b - Ax^*, y^*\rangle + \mu^* = g(b - Ax^*) \cdot g^\sharp(y^*, \mu^*). \qquad & \textit{(constraint alignment)}
\end{aligned}$$

Proof. By construction, x^* is optimal for (N_p) if and only if $(x^*, 1)$ is optimal for its gauge reformulation (2.4.3). Apply theorem 2.3.3 to (2.4.3) and the corresponding gauge dual (N_d) to obtain the required conditions. □

The following result mirrors corollary 1 for the perspective-duality case.

Corollary 3 (Perspective primal-dual recovery)**.** *Suppose that the primal* (N_p) *is strictly feasible. If* (y^*, α^*, μ^*) *is optimal for* (N_d)*, then for any primal feasible* $x \in \mathbb{R}^n$*, the following conditions are equivalent:*

(a) x *is optimal for* (N_p)*;*

(b) $\langle x, A^T y^* \rangle + \alpha^* = f(x) \cdot f^\sharp(A^T y^*, \alpha^*)$ *and* $(b - Ax, 1) \in \sigma \partial g^\sharp(y^*, \mu^*)$;

(c) $A^T y^* \in f^\sharp(A^T y^*, \alpha^*) \cdot \partial f(x)$ *and* $(b - Ax, 1) \in \sigma \partial g^\sharp(y^*, \mu^*)$.

Proof. By construction, x is optimal for (N_p) if and only if $(x, 1)$ is optimal for its gauge reformulation (2.4.3). Apply corollary 1 to (2.4.3) and its gauge dual (N_d) to obtain the equivalence of (a) and (b). To show the equivalence of (b) and (c), note that by the polar-gauge inequality, $\langle (x, 1), (A^T y^*, \alpha^*) \rangle \leq f^\pi(x, 1) \cdot f^\sharp(A^T y^*, \alpha^*)$ for all x, or equivalently,

$$\langle x, A^T y^* \rangle + \alpha^* \leq f(x) \cdot f^\sharp(A^T y^*, \alpha^*), \quad \forall x.$$

The inequality is tight for a fixed x if and only if x minimizes the function $h := f^\sharp(A^T y^*, \alpha^*) f(\cdot) - \langle \cdot, A^T y^* \rangle - \alpha^*$. This in turn is equivalent to $0 \in \partial h(x)$, or

$$A^T y^* \in f^\sharp(A^T y^*, \alpha^*) \cdot \partial f(x).$$

This shows the equivalence of (b) and (c) and completes the proof. □

Section 2.6 illustrates an application of corollary 3 for recovering primal optimal solutions from perspective-dual optimal solutions.

2.4.4 Reformulations of the perspective dual

Two reformulations of the perspective dual (N_d) may be useful depending on the functions f and g involved in (N_p). First, an important simplification of the perspective dual occurs when one or both of these functions are gauges.

Corollary 4 (Simplification for gauges)**.** *If f is a gauge, then a triple (y^*, α^*, μ^*) is optimal for (N_d) if and only if $\alpha^* \leq 0$ and (y^*, μ^*) is optimal for*

$$\underset{y,\alpha}{\text{minimize}} \;\; f^\circ\big(A^T y\big) \;\; \text{subject to} \;\; \langle b, y\rangle - \sigma g^\sharp(y, \mu) \geq 1 - \mu.$$

If, in addition, g is a gauge, then a triple (y^*, α^*, μ^*) is optimal for (N_d) if and only if $\alpha^* \leq 0$, $\mu^* \leq 0$, and y^* solves (G_d).

Proof. Follows from the formulas for $f^\sharp$ and $g^\sharp$ established in section 2.4.1. □

Theorem 2.4.2 also allows us to express the level sets of $g^\sharp$ in terms of its conjugate polar as in the following corollary.

Corollary 5. *The point (y^*, α^*, μ^*) is optimal for (N_d) if and only if there exists a scalar ξ^* such that $(y^*, \alpha^*, \mu^*, \xi^*)$ is optimal for the problem*

$$\begin{aligned} \underset{y,\alpha,\mu,\xi}{\text{minimize}} \quad & f^\sharp(A^T y, \alpha) \\ \text{subject to} \quad & \langle b, y\rangle - \sigma\xi = 1 - (\alpha + \mu), \quad g^{\star\pi}(y, \xi) \leq -\mu, \;\; \xi \geq 0. \end{aligned}$$

Proof. By introducing the variable $\xi := (\langle b, y\rangle + \alpha + \mu - 1)/\sigma$ in (N_d), the result follows from theorem 2.4.2. □

2.5 Examples: piecewise linear-quadratic and GLM constraints

From a computational standpoint, the perspective-dual formulation may be an attractive alternative to the original primal problem. The efficiency of this approach requires that the dual constraints are in some sense more tractable than those of the primal. For example, we may consider the dual feasible set "easy" if it admits an efficient procedure for projecting

onto that set. In this section, we examine two special cases that admit tractable dual problems in this sense. The first case is the family of piecewise linear quadratic (PLQ) functions, introduced by Rockafellar [22] and subsequently examined by Rockafellar and Wets [23, p.440], and Aravkin, Burke, and Pillonetto [3]. The second case is when g is a Bregman divergence arising from a maximum likelihood estimation problem over a family of exponentially distributed random variables.

For this section only, we will assume for the sake of simplicity that the objective f is a gauge, so that the perspective dual in each of this cases simplifies as in corollary 4. The more general case still applies.

2.5.1 PLQ constraints

The family of PLQ functions is a large class of convex functions that includes such commonly used penalties as the Huber function, the Vapnik ϵ-loss, and the hinge loss. The last two are used in support-vector regression and classification [3]. PLQ functions take the form

$$g(y) = \sup_{u \in U} \{ \langle u, By + b \rangle - \tfrac{1}{2}\|Lu\|_2^2 \}, \quad \mathcal{U} := \{ u \in \mathbb{R}^\ell \mid Wu \leq w \}, \tag{2.5.1}$$

where g is defined by linear operators $L \in \mathbb{R}^{\ell \times \ell}$ and $W \in \mathbb{R}^{k \times \ell}$, a vector $w \in \mathbb{R}^k$, and an injective affine transformation $B(\cdot) + b$ from $\mathbb{R}^k$ to $\mathbb{R}^\ell$. We may assume without loss of generality that $B(\cdot) + b$ is the identity transformation, since the primal problem (N_p) already allows for composition of the constraint function g with an affine transformation. We also assume that $\mathcal{U}$ contains the origin, which implies that g is nonnegative and thus can be interpreted as a penalty function. Aravkin, Burke, and Pillonetto [3] describe a range of PLQ functions that often appear in applications.

The conjugate representation of g, given by

$$g^\star(y) = \delta_{\mathcal{U}}(y) + \tfrac{1}{2}\|Ly\|^2,$$

is useful for deriving its polar perspective $g^\sharp$. In the following discussion, it is convenient to interpret the quadratic function $-(1/2\mu)\|Ly\|^2$ as a closed convex function of $\mu \in \mathbb{R}_-$, and thus when $\mu = 0$, we make the definition $-(1/2\mu)\|Ly\|^2 = \delta_{\{0\}}(y)$.

Theorem 2.5.1. *If g is a PLQ function, then*

$$\begin{aligned} g^\sharp(y,\mu) &= \delta_{\mathbb{R}_-}(\mu) + \max\left\{\gamma_{\mathcal{U}}(y),\ -(1/2\mu)\|Ly\|^2\right\} \\ &= \delta_{\mathbb{R}_-}(\mu) + \max\left\{-(1/2\mu)\|Ly\|^2,\ \max_{i=1,\dots,k}\{\, W_i^T y/w_i \,\}\right\}, \end{aligned}$$

where $W_1^T, \dots, W_k^T$ are the rows of W that define $\mathcal{U}$ in (2.5.1).

Proof. First observe that when g is PLQ, $\operatorname{epi} g^\star = \{\, (y,\tau) \mid y \in \mathcal{U},\ \frac{1}{2}\|Ly\|^2 \le \tau \,\}$. Apply theorem 2.4.1 and simplify to obtain the chain of equalities

$$\begin{aligned} g^\sharp(y,\mu) = \gamma_{\operatorname{epi} g^\star}(y,-\mu) &= \inf\{\, \lambda > 0 \mid (y,-\mu) \in \lambda \operatorname{epi} g^\star \,\} \\ &= \inf\{\, \lambda > 0 \mid y/\lambda \in \mathcal{U},\ (1/2\lambda^2)\|Ly\|^2 \le -\mu/\lambda \,\} \\ &= \delta_{\mathbb{R}_-}(\mu) + \max\left\{\gamma_{\mathcal{U}}(y),\ -(1/2\mu)\|Ly\|^2\right\}. \end{aligned}$$

Because $\mathcal{U}$ is polyhedral, we can make the explicit description

$$\begin{aligned} \gamma_{\mathcal{U}}(y) &= \inf\{\, \lambda > 0 \mid y \in \lambda\mathcal{U} \,\} \\ &= \inf\{\, \lambda > 0 \mid W(y/\lambda) \le w \,\} = \max\left\{0,\ \max_{i=1,\dots,k}\{W_i^T y/w_i\}\right\}. \end{aligned}$$

This follows from considering cases on the signs of the $W_i^T y$, and noting that $w \ge 0$ because $\mathcal{U}$ contains the origin. Combining the above results, the theorem is proved. □

The next example illustrates how theorem 2.5.1 can be applied to compute the perspective-polar transform of the Huber function.

Example 2.5.2 (Huber function)**.** *The Huber function [17], which is a smooth approximation to the absolute value function, is also its Moreau envelope of order η. Thus it can be stated in conjugate form as*

$$h_\eta(x) = \sup_{u\in[-\eta,\eta]}\left\{\, ux - (\eta/2)u^2 \,\right\} = \sup_u\left\{\, ux - [\delta_{[-\eta,\eta]}(u) + (\eta/2)u^2] \,\right\},$$

which reveals $h_\eta^\star(y) = \delta_{[-\eta,\eta]}(y) + (\eta/2)y^2$. *We then apply theorem 2.4.1 to obtain*

$$\begin{aligned} h_\eta^\sharp(z,\xi) &= \gamma_{h_\eta^\star}(z,-\xi) \\ &= \inf\{\,\lambda > 0 \mid (z,-\xi) \in \lambda \operatorname{epi} h_\eta^\star\,\} \\ &= \inf\{\,\lambda > 0 \mid |z|/\lambda \le \eta,\ (\eta/2\lambda)z^2 \le -\xi\,\} \\ &= \delta_{\mathbb{R}_-}(\xi) + \max\left\{|z|/\eta,\ -(\eta/2\xi)z^2\right\}. \end{aligned}$$

Note that this can easily be extended beyond the univariate case to a separable sum by applying the result of Example 2.4.4.

We can now write down an explicit formulation of the perspective dual problem (N_d) when the primal problem (N_p) has a PLQ-constrained feasible region (i.e., g is PLQ) and a gauge objective (i.e., f is a closed gauge). The constraint set of (N_d) simplifies significantly so that, for example, a first-order projection method might be applied to solve the problem. Apply theorem 2.5.1 and introduce a scalar variable ξ to rephrase the dual problem (N_d) as

$$\begin{aligned} \underset{y,\,\mu,\,\xi}{\text{minimize}} \quad & f^\circ\big(A^T y\big) \\ \text{subject to} \quad & \langle b, y\rangle + \mu - \sigma\xi = 1,\ \mu \le 0,\ \xi \ge 0, \\ & Wy \le \xi w,\ -(1/2\mu)\|Ly\|^2 \le \xi. \end{aligned} \tag{2.5.2}$$

We can further simplify the constraint set using the fact that

$$\left[\ \|Ly\|^2 \le -2\mu\xi \text{ and } \mu \le 0,\ \xi \ge 0\ \right] \iff \left\| \begin{bmatrix} 2Ly \\ \xi + 2\mu \end{bmatrix} \right\|_2 \le \xi - 2\mu, \tag{2.5.3}$$

Thus, projecting a point $\overline{y}$ onto the feasible set of (2.5.2) is equivalent to solving a second-order cone program (SOCP). In many important cases, the operator L is extremely sparse. For example, when g is a sum of separable Huber functions, we have $L = \sqrt{\eta}I$. Hence in many practical cases, particularly when $m \ll n$ and the dual variables are low-dimensional, this projection problem could be solved efficiently using SOCP solvers that take advantage of sparsity, e.g., Gurobi [16].

2.5.2 Generalized linear models and the Bregman divergence

Suppose we are given a data set $\{(a_i, b_i)\}_{i=1}^m \subseteq \mathbb{R}^{n+1}$, where each vector a_i describes features associated with observations b_i. Assume that the vector b of observations is distributed according to an exponential density $p(y \mid \theta) = \exp[\langle \theta, y\rangle - \phi^\star(\theta) - p_0(y)]$, where the conjugate of $\phi : \mathbb{R}^n \to \mathbb{R}$ is the cumulant generating function of the distribution and $p_0 : \mathbb{R}^n \to \mathbb{R}$ serves to normalize the distribution. We assume that ϕ is a closed convex function of the Legendre type [20, p.258]. The maximum likelihood estimate (MLE) can be obtained as the maximizer of the log-likelihood function $\log p(y \mid \theta)$.

In applications that impose an *a priori* distribution on the parameters, the goal is to find an approximation to the MLE estimate that penalizes a regularization function f (a surrogate for the prior). We assume a linear dependence between the parameters and feature vectors, and thus set $\theta = Ax$, where the matrix A has rows a_i. A regularized MLE estimate could be obtained by solving the constrained problem

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad d_{\phi^\star}(Ax; \nabla\phi(b)) \leq \sigma,$$

where $d_\phi(v; w) := \phi(v) - \phi(w) - \langle \nabla\phi(w), v - w\rangle$ is the *Bregman divergence* function, and σ is a positive parameter that controls the divergence between the linear model Ax and the first-moment $\nabla\phi(b)$ relative to the density defined by ϕ [4].

We use corollary 5 to derive the perspective dual, which requires the computation of the conjugate of $g(z) := d_{\phi^\star}(z; \nabla\phi(b))$:

$$\begin{aligned} g^\star(y) &= \sup_z \{\, \langle z, y\rangle - d_{\phi^\star}(z; \nabla\phi(b)) \,\} \\ &= \sup_z \{\, \langle z, y\rangle - \phi^\star(z) + \phi^\star(\nabla\phi(b)) + \langle b, z - \nabla\phi(b)\rangle \,\} \\ &= \phi^\star(\nabla\phi(b)) - \langle b, \nabla\phi(b)\rangle + \phi(y + b), \end{aligned}$$

where we simplify the expression using the inverse relationship between the gradients of ϕ and its conjugate. Assume for simplicity that f is a gauge, which is typical when it serves as

a regularization function. In that case, the perspective dual reduces to

$$\begin{array}{ll} \underset{y,\mu,\xi}{\text{minimize}} & f^\circ(A^T y) \\ \text{subject to} & \phi^\pi(y+\xi b,\xi) \le \xi[\langle b, \nabla\phi(b)\rangle - \phi^\star(\nabla\phi(b)) - \sigma] - 1,\ \xi \ge 0; \end{array} \tag{2.5.4}$$

cf. Corollaries 4 and 5.

Example 2.5.3 (Gaussian distribution)**.** *As a first example, consider the case where the b_i are distributed as independent Gaussian variables with unit variance. In this case, $\phi := \frac{1}{2}\|\cdot\|^2$ and the above constraints specialize to*

$$\tfrac{1}{2\xi}\|y\|^2 + \langle b, y\rangle \le -(1+\sigma\xi),\ \xi \ge 0.$$

This is an example of a PLQ constraint, which falls into the category of problems described in section 2.5.1.

Example 2.5.4 (Poisson distribution)**.** *Consider the case where the observations b_i are independent Poisson observations, which corresponds to $\phi(\theta) = \theta\log\theta - \theta$ and $\phi^\star(y) = e^y$. Straightforward calculations show that the perspective dual constraints for the Poisson case reduce to*

$$\sum_{i=1}^m z_i \log(z_i/\xi) \le \beta\xi + \sum_{i=1}^m z_i - (1+\sigma\xi), \quad z = y + \xi b, \quad \xi \ge 0,$$

where $\beta = \sum_{i=1}^m (b_i + b_i \log b_i)$ is a constant. By introducing new variables, this can be further simplified to require only affine constraints and m relative-entropy constraints. To solve projection subproblems onto a constraint set of this form, we note that

$$F(x,y,r) = 400(-\log(x/y) - \log(\log(x/y) - r/y) - 4\log(y))$$

is a self-concordant barrier for the set $\{\,(x,y,r) \mid y > 0,\ y\log(y/x) \le r\,\}$, which is the epigraph of the relative entropy function; see Nesterov and Nemirovski [19, Proposition 5.1.4] and Boyd and Vandenberghe [5, Example 9.8]. Standard interior methods can therefore be used to project onto the constraint set.

Example 2.5.5 (Bernoulli distribution)**.** *When the observations b_i are independent Bernoulli observations, which corresponds to $\phi(\theta) = \theta \log \theta + (1-\theta)\log(1-\theta)$ and $\phi^\star(y) = \log(1+e^y)$, the perspective dual constraints in (2.5.4) reduce to*

$$\sum_{i=1}^{m} [z_i \log(z_i/\xi) + (\xi - z_i)\log((\xi - z_i)/\xi)] \leq \beta\xi - (1+\sigma\xi), \quad z = y + \xi b, \quad \xi \geq 0,$$

where $\beta = \sum_{i=1}^{m}(b_i \log b_i + (1-b_i)\log(1-b_i))$ is a constant. By introducing new variables, this can be rewritten with only affine constraints and $2m$ relative-entropy constraints. Thus the projection subproblems can be solved as in the Poisson case.

2.6 Examples: recovering primal solutions

Once we have solved the gauge or perspective dual problems, we have two available approaches for recovering a corresponding primal optimal solution. If we applied a (Lagrange) primal-dual algorithm (e.g., the algorithm of Chambolle and Pock [8]) to solve the dual, then theorem 2.3.5 gives a direct recipe for constructing a primal solution from the algorithm's output. On the other hand, if we applied a primal-only algorithm to solve the dual, we must instead rely on corollary 1 or corollary 3 to recover a primal solution. Interestingly, the alignment conditions in these theorems can provide insight into the structure of the primal optimal solution, as illustrated by the following examples.

2.6.1 Recovery for basis pursuit denoising

Our first example illustrates how corollary 1 can be used to recover primal optimal solutions from dual optimal solutions for a simple gauge problem. Consider the gauge dual pair

$$\underset{x}{\text{minimize}} \ \|x\|_1 \qquad \text{subject to} \qquad \|b - Ax\|_2 \leq \sigma \tag{2.6.1a}$$

$$\underset{y}{\text{minimize}} \ \|A^T y\|_\infty \ \text{ subject to } \ \langle b, y\rangle - \sigma\|y\|_2 \geq 1, \tag{2.6.1b}$$

which corresponds to the basis pursuit denoising problem. The 1-norm in the primal objective encourages sparsity in x, while the constraint enforces a maximum deviation between a forward model Ax and observations b.

Let y^* be optimal for the dual problem (2.6.1b), and set $z = A^T y^*$. Define the active set

$$I(z) = \{\, i \mid |z_i| = \|z\|_\infty \,\}$$

as the set of indices of z that achieve the optimal objective value of the gauge dual. We use corollary 1 to determine properties of a primal solution x^*. In particular, the first part of corollary 1(b) holds if and only if $x_i^* = 0$ for all $i \notin I(z)$, and $\operatorname{sign}(x_i^*) = \operatorname{sign}(z_i)$ for all $i \in I(z)$. Thus, the maximal-in-modulus elements of $A^T y^*$ determine the support for any primal optimal solution x^*. The second condition in corollary 1(b) holds if and only if $b - Ax = \sigma y^*/\|y^*\|_2$. In order to satisfy this last condition, we solve the least-squares problem restricted to the support of the solution:

$$\underset{x}{\text{minimize}} \quad \|b - Ax - \sigma\,(y^*/\|y^*\|_2)\|_2^2 \quad \text{subject to} \quad x_i = 0 \;\; \forall i \notin I(z).$$

(Note that $y^* \neq 0$, otherwise the primal problem is infeasible.) The efficiency of this least-squares solve depends on the number of elements in $I(z)$. For many applications of basis pursuit denoising, for example, we expect the support to be small relative to the length of x, and in that case, the least-squares recovery problem is expected to be a relatively inexpensive subproblem. We may interpret the role of the dual problem as that of determining the optimal support of the primal, and the role of the above least-squares problem as recovering the actual values of the support.

2.6.2 Sparse recovery with Huber misfit

For an example where the constraint is not a gauge function, consider the variant of (2.6.1a)

$$\underset{x}{\text{minimize}} \;\; \|x\|_1 \;\; \text{subject to} \;\; h(b - Ax) \leq \sigma, \quad \text{with} \quad h(r) = \sum_{i=1}^{m} h_\eta(r_i), \tag{2.6.2}$$

where h_η is the Huber function; cf. theorem 2.5.2. This problem corresponds to (N_p) with $f(x) = \|x\|_1$ and $g = h$. Suppose that the tuple (y, α, μ), with $\mu < 0$, is optimal for the perspective dual, and that (N_p) attains its optimal value. Because f is a gauge, corollary 4

asserts that $\alpha = 0$, and thus corollary 3(b) reduces to the conditions

$$\langle x, A^T y \rangle = f(x) \cdot f^\circ(A^T y) \tag{2.6.3a}$$

$$(b - Ax, 1) \in \sigma \partial h^\sharp(y, \mu). \tag{2.6.3b}$$

As we did for the related example in section 2.6.1, we use (2.6.3a) to deduce the support of the optimal primal solution. It follows from theorem 2.5.1 that because g is PLQ,

$$h^\sharp(y, \mu) = \delta_{\mathbb{R}_-}(\mu) + \max\left(\max_{i=1,\ldots,k} \{W_i^T y / w_i\},\ -(1/2\mu)\|Ly\|^2 \right).$$

In particular, because h is a separable sum of Huber functions, $W = [I\ -I]^T$, w is the constant vector of all ones, and $L = \sqrt{\eta} I$. Since $\mu < 0$, it follows that

$$\partial h^\sharp(y, \mu) = \partial \left(\max \left\{ \|y\|_\infty, -(\eta/2\mu)\|y\|^2 \right\} \right)(y, \mu).$$

For the set $\{ v_1, \ldots, v_{2m+1} \} := \left\{ y_1, \ldots, y_m, -y_1, \ldots, -y_m, -\frac{\eta}{2\mu}\|y\|^2 \right\}$, let $J(y, \mu) := \{ j \mid |v_j| = \max_{i=1,\ldots,2m+}$ be the set of maximizing indices. Then

$$\partial h^\sharp(y, \mu) = \operatorname{conv} \{ \nabla v_j \mid j \in J(y, \mu) \},$$

where conv denotes the convex hull operation. More concretely, precisely the following terms are contained in the convex hull above:

- $\left(-\frac{\eta}{\mu} y,\ \frac{\eta}{2\mu^2}\|y\|_2^2 \right)$ if $-\frac{\eta}{2\mu}\|y\|^2 \geq \|y\|_\infty$;
- $(\operatorname{sign}(y_i) \cdot e_i,\ 0)$ if $i \in [m]$ and $|y_i| = \|y\|_\infty \geq -\frac{\eta}{2\mu}\|y\|^2$,

where e_i is the ith standard basis vector. Note that if an optimal solution to (N_p) exists, then corollary 3 tells us that $(-(\eta/\mu)y,\ (\eta/2\mu^2)\|y\|^2)$ must be included in this convex hull, otherwise it is impossible to have $(b - Ax, 1) \in \partial h^\sharp(y, \mu)$.

In summary, corollary 3 tells us that to find an optimal solution x for (N_p), we need to solve a linear program to ensure that $(b - Ax, 1) \in \operatorname{conv} \{ \nabla v_j \mid j \in J(y, \mu) \}$ subject to the optimal support of x, as determined by (2.6.3a). In cases where the size of the support is expected to be small (as might be expected with a 1-norm objective), this required linear program can be solved efficiently.

2.7 Numerical experiment: sparse robust regression

To illustrate the usefulness of the primal-from-dual recovery procedure implied by theorem 2.3.5, we continue to examine the sparse robust regression problem (2.6.2), considered by Aravkin et al. [2]. The aim is to find a sparse signal (e.g., a spike train) from measurements contaminated by outliers. These experiments have been performed with the following data: $m = 120$, $n = 512$, $\sigma = 0.2$, $\eta = 1$, and A is a Gaussian matrix. The true solution $\overline{x} \in \{-1, 0, 1\}$ is a spike train which has been constructed to have 20 nonzero entries, and the true noise $b - A\overline{x}$ has been constructed to have 5 outliers.

We compare two approaches for solving problem (2.6.2). In both, we use Chambolle and Pock's (CP) algorithm [8], which is primal-dual (in the sense of Lagrange duality) and can be adapted to solve both the primal problem (2.6.2) and its perspective dual (2.5.2). Other numerical methods could certainly be applied to either of these problems, such as Shefi and Teboulle's dual moving-ball method [25]. We note that a primal-only method, for example, applied to (2.5.2), would require us to use the methods of section 2.6 rather than theorem 2.3.5 for the recovery of a primal solution.

The CP method applied to problem (2.3.1) at each iteration k computes

$$\begin{aligned} y^{k+1} &:= \operatorname{prox}_{\alpha_y f^\star}\left(y^k + \alpha_y A[2x^k - x^{k-1}]\right) \\ x^{k+1} &:= \operatorname{prox}_{\alpha_x g}(x^k - \alpha_x A^T y^{k+1}), \end{aligned}$$

where $\operatorname{prox}_{\alpha f}(x) := \operatorname{argmin}_y\{f(y) + \frac{1}{2\alpha}\|x - y\|_2^2\}$. The positive scalars α_x and α_y are chosen to satisfy $\alpha_x \alpha_y \|A\|^2 < 1$. Setting $f = \delta_{h(b-\cdot)\leq\sigma}$, and $g = \|\cdot\|_1$ yields the primal problem (2.6.2). In this case, the proximal operators $\operatorname{prox}_{\alpha f^\star}$ and $\operatorname{prox}_{\alpha g}$ can be computed using the Moreau identity, i.e.,

$$\begin{aligned} \operatorname{prox}_{\alpha f^\star}(x) &= x - \operatorname{prox}_{(\alpha f^\star)^\star}(x) = x - \alpha \Pi_f(x/\alpha) \\ \operatorname{prox}_{\alpha g}(y) &= y - \operatorname{prox}_{(\alpha g)^\star}(y) = y - \Pi_{\alpha \mathbb{B}_\infty}(y/\alpha), \end{aligned}$$

where Π_f is the projection onto the sublevel set in the definition of f and $\Pi_{\alpha\mathbb{B}_\infty}$ is the

projection onto the infinity-norm ball of radius α. We implement Π_f using the `Convex.jl` [27] and `Gurobi` [16] software packages.

On the other hand, to apply CP to the perspective dual problem (2.5.2), one instead takes $f = (\|\cdot\|)^\circ = \|\cdot\|_\infty$ and $g = \delta_{\mathcal{Q}}$, where $\mathcal{Q}$ is the constraint set for (2.5.2), and take A to be the corresponding adjoint to the operator in (2.6.2). To compute $\text{prox}_{\alpha_y g}$, which is the projection onto $\mathcal{Q}$, we solve the SOCP (2.5.3) using Gurobi. To evaluate $\text{prox}_{\alpha f^\star}$, we again use the Moreau identity and project onto level sets of $\|\cdot\|_1$.

fig. 2.7.1 compares the outcomes of running CP on the primal and perspective dual problems. This experiment exhibited similar behavior when run 500 times with different realizations of the random data, and so here we report on a single problem instance. Note that performing an iteration of CP on the perspective dual is significantly faster than performing an iteration of CP on the primal because $\Pi_{\mathcal{Q}}$ can be computed much more efficiently than Π_f (see the discussion in section 2.5.1). This also appears to make convergence of CP on the perspective dual more stable, as seen in fig. 2.7.1(a). fig. 2.7.1(c)-(d) illustrate the sparsity patterns of the iterates x_k relative to those $\overline{x}$. Notably, we recover the correct sparsity patterns using theorem 2.3.5. The recovery procedure outlined in section 2.6.2 also recovers the correct sparsity pattern, when applied to the final perspective dual iterate.

2.8 Discussion

Gauge duality is fascinating in part because it shares many symmetric properties with Lagrange duality, and yet Freund's 1987 development of the concept flows from an entirely different principle based on polarity of the sets that define the gauge functions. On the other hand, Lagrange duality proceeds from a perturbation argument, which yields as one of its hallmarks a sensitivity interpretation of the dual variables. The discussion in section 2.3 reveals that both duality notions can be derived from the same Fenchel-Rockafellar perturbation framework. The derivation of gauge duality using this framework appears to be its first application to a perturbation that does not lead to Lagrange duality. This new link between gauge duality and the perturbation framework establishes a sensitivity interpretation for

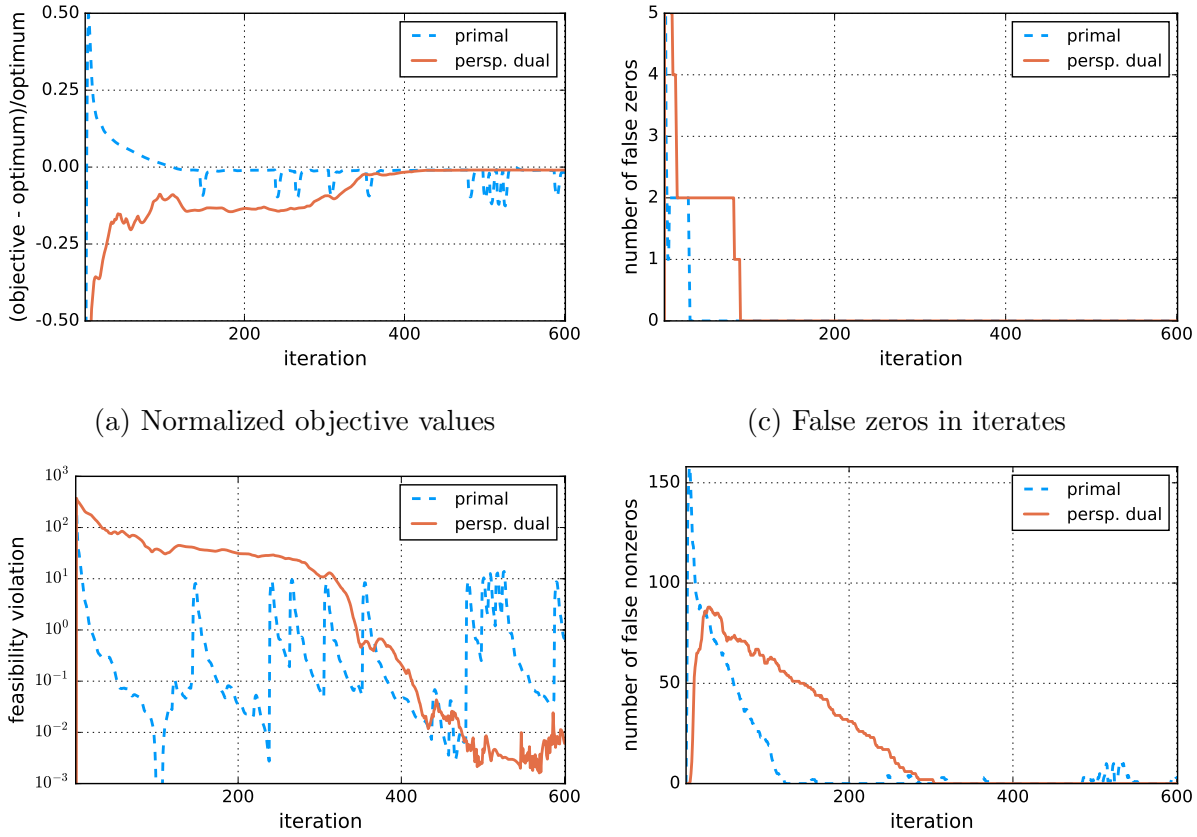
gauge dual variables, which has not been available until now.

One motivation for this work is to explore alternative formulations of optimization problems that might be computationally advantageous for certain problem classes. The phase-retrieval problem, based on an SDP formulation, was a first application of ideas from gauge duality for developing large-scale solvers [14]. That approach, however, was limited in its flexibility because it required gauge functions. The discussions of section 2.4 pave the way to new extensions, such as different models of the measurement process, as described in section 2.5.2.

Another implication of this work is that it establishes the foundation for exploring a new breed of primal-dual algorithms based on perspective duality. Our own application of Chambolle and Pock's primal-dual algorithm [8] to the perspective-dual problem, together with a procedure for extracting a primal estimate, is a first exploratory step towards developing variations of such methods. Future directions of research include the development of such algorithms, along with their attendant convergence properties and an understanding of the classes of problems for which they are practicable.

Acknowledgments

We are grateful to Patrick Combettes for pointing us to recent comprehensive work on properties of the perspective function and its applications [9, 10]. Our sincere thanks to two anonymous referees who provided an extensive list of corrections and suggestions that helped us to arrive at several strengthened results and to streamline our presentation.



(a) Normalized objective values

(c) False zeros in iterates

(b) Feasibility violations for iterates

(d) False nonzeros in iterates

Figure 2.7.1: The CP algorithm applied to sparse robust regression (section 2.7). Dashed lines indicate CP applied to the primal problem (2.6.2), and solid lines indicate CP applied to its perspective dual (2.5.2) where the primal solution is recovered via the method of theorem 2.3.5. Plots show (a) normalized deviation of objective value $\|x^k\|_1$ from optimal value $\|\overline{x}\|_1$; (b) infeasibility measure $\max(h(b - Ax^k) - \sigma, 0)$ for iterate x^k; (c) number false zeros in iterate x^k relative to $\overline{x}$; (d) number of false nonzeros in iterate x^k relative to $\overline{x}$.

References

[1] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy. Level-set methods for convex optimization. *arXiv:1602.01506*, 2016.

[2] A. Y. Aravkin, J. V. Burke, and M. P. Friedlander. Variational properties of value functions. *SIAM Journal on Optimization*, 23(3):1689–1717, 2013.

[3] A. Y. Aravkin, J. V. Burke, and G. Pillonetto. Linear system identification using stable spline kernels and PLQ penalties. In *52nd IEEE Decision and Control Proceedings*, pages 5168–5173, Dec 2013.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(Oct):1705–1749, 2005.

[5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[6] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, Feb 2006.

[7] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

[8] A. Chambolle and T. 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.

[9] P. L. Combettes. Perspective functions: Properties, constructions, and examples. *Set-Valued Variational Analysis*, pages 1–18, 2017.

[10] P. L. Combettes and C. L. Müler. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications*, 2016.

[11] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[12] D. Drusvyatskiy, A. Ioffe, and A. Lewis. Clarke subgradients of directionally Lipschitzian stratifiable functions. *Mathematics of Operations Research*, 40(2):328–349, 2015.

[13] R. M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.

[14] M. P. Friedlander and I. Macêdo. Low-rank spectral optimization via gauge duality. *SIAM Journal on Scientific Computing*, 28(3):A1616–A1638, 2016.

[15] M. P. Friedlander, I. Macedo, and T. K. Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

[16] I. Gurobi Optimization. Gurobi optimizer reference manual, 2015.

[17] P. Huber. *Robust Statistics*. Wiley, 1981.

[18] J. A. Nelder and R. J. Baker. Generalized linear models. *Encyclopedia of Statistical Sciences*, 1972.

[19] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. SIAM, 1994.

[20] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1972.

[21] R. T. Rockafellar. *Conjugate Duality and Optimization*. SIAM, 1974.

[22] R. T. Rockafellar. First- and second-order epi-differentiability in nonlinear programming. *Transactions of the American Mathematical Society*, 307(1):75–108, May 1988.

[23] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.

[24] R. Shefi and M. Teboulle. A dual method for minimizing a nonsmooth objective over one smooth inequality constraint. *Mathematical Programming*, 159(1-2):137–164, 2016.

[25] R. Shefi and M. Teboulle. On the rate of convergence of the proximal alternating linearized minimization algorithm for convex problems. *EURO Journal on Computational Optimization*, 4(1):27–46, 2016.

[26] J. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[27] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd. Convex optimization in Julia. *SC14 Workshop on High Performance Technical Computing in Dynamic Languages*, 2014.

[28] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67(2):301–320, 2005.

Appendix

2.A.1 Proof of (2.2.6)

We prove each fact in succession.

1. $(\mathcal{U}_\kappa^\circ = \mathcal{U}_{\kappa^\circ})$. By definition of the polar gauge and the polar cone, we have $y \in \mathcal{U}_{\kappa^\circ}$ if and only if
$$\sup\{\, \langle x, y\rangle \mid \kappa(x) \le 1 \,\} \le 1 \iff y \in \mathcal{U}_\kappa^\circ.$$

2. $(\mathcal{U}_\kappa^\infty = \mathcal{H}_\kappa)$. Suppose $x \in \mathcal{H}_\kappa$. Then for any $u \in \mathcal{U}_\kappa$ and $\lambda > 0$, by sublinearity of κ we have $\kappa(u + \lambda x) \le \kappa(u) + \lambda\kappa(x) \le 1 + \lambda \cdot 0 = 1$. Thus $x \in \mathcal{U}_\kappa^\infty$, and $\mathcal{H}_\kappa \subseteq \mathcal{U}_\kappa^\infty$. Suppose now that $y \in \mathcal{U}_\kappa^\infty \setminus \mathcal{H}_\kappa$. Then in particular, $\kappa\left(y/\kappa(y) + \lambda y\right) \le 1$ for all $\lambda > 0$. But then by positive homogeneity, $\left(1/\kappa(y) + \lambda\right)\kappa(y) \le 1$, for all $\lambda > 0$. This is a contradiction since $\kappa(y) > 0$, so we conclude that $\mathcal{H}_\kappa = \mathcal{U}_\kappa^\infty$.

3. $((\operatorname{dom}\kappa)^\circ = \mathcal{H}_{\kappa^\circ})$. By positive homogeneity of κ and the definition of the polar gauge, $y \in \mathcal{H}_{\kappa^\circ}$ if and only if
$$\sup_{\kappa(x)\le 1} \langle x, y\rangle = 0 \iff \sup_{\kappa(x)<\infty} \langle x, y\rangle = 0 \iff y \in (\operatorname{dom}\kappa)^\circ.$$

4. $(\mathcal{H}_\kappa^\circ = \operatorname{cl}\operatorname{dom}\kappa^\circ)$. Apply the third equality, replacing κ by κ°, and then take polars on both sides. This concludes the proof.

□

2.A.2 Proof of Lemma 2

With no loss in generality, we can assume that $\sigma > 0$, because if $\sigma = 0$, we use the convention (2.2.8) and its implication (2.2.9).

First suppose that the primal (G_p) is relatively strictly feasible. A point u lies in the domain of p if and only if the system

$$(u,0,0) \in M \begin{bmatrix} w \\ \lambda \end{bmatrix} + (\operatorname{epi}\rho) \times \mathcal{U}_\kappa, \quad \text{where} \quad M := \left[\begin{smallmatrix} A & -b \\ 0 & -\sigma \\ -I & 0 \end{smallmatrix}\right],$$

is solvable for (w,λ). Thus the set $(\operatorname{dom} p) \times \{0\} \times \{0\}$ coincides with

$$L \bigcap (\operatorname{range} M + (\operatorname{epi}\rho) \times \mathcal{U}_\kappa), \tag{2.8.1}$$

where $L := \{\, (a,b,c) \mid b = 0,\ c = 0 \,\}$ is a linear subspace. We aim to show $(0,0,0)$ is in the relative interior of (2.8.1), which will show $0 \in \operatorname{ri}\operatorname{dom} p$. Use [20, Lemma 7.3] and [20, Theorem 7.6] to obtain

$$\begin{aligned} \operatorname{ri}\operatorname{epi}\rho &= \{\, (z,r) \in (\operatorname{ri}\operatorname{dom}\rho) \times \mathbb{R} \mid \rho(z) < r \,\} \\ \operatorname{ri}\mathcal{U}_\kappa &= \{\, x \in \operatorname{ri}\operatorname{dom}\kappa \mid \kappa(x) < 1 \,\}. \end{aligned}$$

From relative strict feasibility of (G_p), the fact that $\sigma > 0$, and again [20, Theorem 7.6], we deduce existence of an $x \in \operatorname{ri}\operatorname{dom}\kappa$ with $b - Ax \in \operatorname{ri}\operatorname{dom}\rho$ and $\rho(b-Ax) < \sigma$. Fix a constant $r > \kappa(x)$ and define the pair $(w,\lambda) := (x/r, 1/r)$. Then we immediately have $(b\lambda - Aw, \sigma\lambda) \in \operatorname{ri}\operatorname{epi}\rho$ and $\kappa(w) < 1$. It follows that the vector $-M\left[\begin{smallmatrix} w \\ \lambda \end{smallmatrix}\right]$ lies in $(\operatorname{ri}\operatorname{epi}\rho) \times \operatorname{ri}\mathcal{U}_\kappa$. Thus $(0,0,0)$ lies in the intersection

$$L \bigcap (\operatorname{range} M + [(\operatorname{ri}\operatorname{epi}\rho) \times \operatorname{ri}\mathcal{U}_\kappa]). \tag{2.8.2}$$

Use [20, Theorem 6.5, Corollary 6.6.2] to deduce that (2.8.2) is the relative interior of the intersection (2.8.1). Thus $y = 0$ lies in the relative interior of $\operatorname{dom} p$ as claimed.

Next, suppose that the gauge dual (G_d) is strictly feasible. By definition of $F^\star$, the tuple (w,λ) lies in the domain of v_d if and only if

$$(w,0,-\lambda) \in (\operatorname{dom}\kappa^\circ \times \operatorname{epi}(\sigma\rho^\circ - \langle b, \cdot\rangle + 1)) - \operatorname{range} B, \quad \text{with} \quad B := \left[\begin{smallmatrix} A^T \\ I \\ 0 \end{smallmatrix}\right].$$

Thus $\operatorname{dom} v_d$ is linearly isomorphic to the intersection

$$L' \bigcap ((\operatorname{dom}\kappa^\circ \times \operatorname{epi}(\sigma\rho^\circ - \langle b, \cdot\rangle + 1)) - \operatorname{range} B), \tag{2.8.3}$$

where L' is the linear subspace $L' := \{(a, b, c) \mid b = 0\}$.

However, by [20, Lemma 7.3], relative strict feasibility of the dual (G_d) amounts to the inclusion

$$(0, 0, 0) \in (\operatorname{ri}\operatorname{dom}\kappa^\circ \times \operatorname{ri}\operatorname{epi}(\sigma\rho^\circ - \langle b, \cdot\rangle + 1)) - \operatorname{range} B.$$

Strict feasibility of (G_d) implies, via [20, Corollary 6.5.1, Corollary 6.6.2], that $(0, 0, 0)$ is in the relative interior of the intersection (2.8.3), and thus $0 \in \operatorname{ri}\operatorname{dom} v_d$, as claimed.

Finally, the exact same arguments, but with relative interiors replaced by interiors, will prove the claims relating strict feasibility and interiority. This concludes the proof. □

Chapter 3

SUBGRADIENT METHODS FOR SHARP WEAKLY CONVEX FUNCTIONS

This chapter represents joint work with Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The contents appeared in *Journal of Optimization Theory and Applications*, Dec 2018, 179-3 (p. 962-982).

Abstract: Subgradient methods converge linearly on a convex function that grows sharply away from its solution set. In this work, we show that the same is true for sharp functions that are only weakly convex, provided that the subgradient methods are initialized within a fixed tube around the solution set. A variety of statistical and signal processing tasks come equipped with good initialization, and provably lead to formulations that are both weakly convex and sharp. Therefore, in such settings, subgradient methods can serve as inexpensive local search procedures. We illustrate the proposed techniques on phase retrieval and covariance estimation problems.

3.1 Introduction

Typical methods for statistics and signal processing tasks follow the two-step strategy: (i) find a moderately accurate solution at a low sample complexity cost (e.g., using spectral initialization), and (ii) refine the approximate solution by an iterative "local search algorithm" that converges rapidly under natural statistical assumptions. For smooth problem formulations, the term "local search" almost universally refers to gradient descent or a close variant thereof; see e.g. [1, 2, 4, 7, 21, 22, 24, 38]. For nonsmooth and nonconvex problems, the meaning of local search is much less clear. In this work, we ask the following question.

> Is there a generic gradient-based local search procedure for nonsmooth and non-convex problems, which converges linearly under standard regularity conditions?

We answer this question positively for the class of weakly convex functions, i.e. those that become convex after a sufficiently large quadratic perturbation. We show that standard subgradient methods, originally designed for sharp convex problems, locally converge linearly when applied to sharp functions that are only weakly convex. We focus on three stepsize rules: Polyak stepsize [18,30], geometrically decaying stepsize [19,34], and constant stepsize [23,36,39]. As a proof of concept, we illustrate the resulting algorithms on phase retrieval and covariance estimation problems.

3.2 Discussion

Our approach is rooted in subgradient methods for convex optimization. To motivate the discussion, consider the constrained optimization problem

$$\min_{x\in\mathcal{X}} \; g(x), \tag{3.2.1}$$

where g is an L-Lipschitz convex function on $\mathbb{R}^d$ and $\mathcal{X}$ is a closed and convex set. Given a current iterate x_k, subgradient methods proceed as follows:

$$\left\{ \begin{array}{l} \text{Choose any } \zeta_k \in \partial g(x_k) \\ \text{Set } x_{k+1} = \text{proj}_{\mathcal{X}} \left(x_k - \alpha_k \cdot \dfrac{\zeta_k}{\|\zeta_k\|} \right) \end{array} \right\}.$$

Here, the symbol $\text{proj}_{\mathcal{X}}(y)$ denotes the nearest point of $\mathcal{X}$ to y and $\{\alpha_k\}$ is a specified stepsize sequence. The choice of the sequence $\{\alpha_k\}$ determines the behavior of the scheme, and is the main distinguishing feature among subgradient methods. In this work, we will only be interested in subgradient methods that are linearly convergent. As usual, linear rates of convergence of iterative methods require some regularity conditions to hold. Here, the appropriate regularity condition is *sharpness* [3, 32] (or equivalently a global error bound): there exists a real $\mu > 0$ satisfying

$$g(z) - \min_{x\in\mathcal{X}} g(x) \geq \mu \cdot \text{dist}(z; \mathcal{X}^*) \qquad \text{for all } z \in \mathcal{X},$$

where $\mathcal{X}^*$ denotes the set of minimizers of (3.2.1), and $\operatorname{dist}(z;\mathcal{X}^*)$ denotes the distance of z to $\mathcal{X}^*$. Assuming sharpness holds, subgradient methods with a judicious choice of $\{\alpha_k\}$ produce iterates that converge to $\mathcal{X}^*$ at the linear rate $\sqrt{1-(\mu/L)^2}$. Results of this type date back to the 60's and 70's [18, 19, 30, 31, 34], while some more recent approaches have appeared in [23, 36, 39].

Various contemporary problems lead to formulations that are indeed sharp, but are only *weakly convex* and *locally Lipschitz*. Recall that a function g is ρ-weakly convex [26] if the perturbed function $x \mapsto g(x) + \frac{\rho}{2}\|\cdot\|^2$ is convex for some $\rho \geq 0$.[1] Note that weakly convex functions need not be smooth nor convex. A quick computation (Lemma 3.3.1) shows that, if g is μ-sharp and ρ-weakly convex, then there is a tube around the solution set $\mathcal{X}^*$ that contains no extraneous stationary points:

$$\mathcal{T} := \left\{ x \in \mathcal{X} : \operatorname{dist}(x;\mathcal{X}^*) < \frac{2\mu}{\rho} \right\}.$$

In this work, we show that the standard linearly convergent subgradient methods, originally designed for convex problems, apply in this much greater generality, provided they are initialized within a slight contraction of the tube $\mathcal{T}$. The methods exhibit essentially the same linear rate of convergence as in the convex case, while the weak convexity constant ρ only determines the validity of the initialization. We focus on three stepsize rules: Polyak stepsize [18, 30], geometrically decaying stepsize [19, 34], and constant stepsize [23, 36, 39]. As a proof of concept, we illustrate the resulting algorithms on phase retrieval and covariance estimation problems.

Our current work sits within the broader scope of analyzing subgradient and proximal methods for weakly convex problems [9, 11, 13–16, 26, 27]; see also the recent survey [12]. In particular, the paper [9] proves a global sublinear rate of convergence, in terms of a

[1] To the best of our knowledge, the class of weakly convex functions was introduced in 1972 in a local publication of the Institute of Cybernetics in Kiev, and then used in [26]. These are the functions that are lower bounded by a linear function up to a little-o term that is uniform in the base point. The term ρ-weak convexity, as we use it here, corresponds to the setting when the error term is a quadratic. A number of other names are used in the literature instead of ρ-weak convexity, such as semi-convex, prox-regular, and lower-C^2.

natural stationarity measure, of a (stochastic) subgradient method on any weakly convex function. In contrast, here we are interested in subgradient methods that are locally linearly convergent under the additional sharpness assumption. The arguments we present are all quick modification of the proofs already available in the convex setting. Nonetheless, we believe that the drawn conclusions are interesting and powerful, opening the door to generic local search procedures for nonsmooth and nonconvex problems.

3.3 Notation

Throughout, we consider the Euclidean space $\mathbb{R}^d$, equipped with an inner-product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| := \sqrt{\langle x, x \rangle}$. The *distance* and the *projection* of any point $y \in \mathbb{R}^d$ onto a set $\mathcal{X}$, are defined by

$$\operatorname{dist}(y; \mathcal{X}) := \inf_{x \in \mathcal{X}} \|y - x\| \qquad \text{and} \qquad \operatorname{proj}_{\mathcal{X}}(y) := \operatorname*{argmin}_{x \in \mathcal{X}} \|y - x\|,$$

respectively. Note that $\operatorname{proj}_{\mathcal{X}}(y)$ is nonempty as long as $\mathcal{X}$ is a closed set. The *indicator function* of a set $\mathcal{X}$, denoted by $\delta_{\mathcal{X}}$, is defined to be zero on $\mathcal{X}$ and $+\infty$ off it.

3.3.1 Weakly Convex Functions

Our main focus is on those functions that are convex up to an additive quadratic perturbation. Namely, a function $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is called *ρ-weakly convex* (with $\rho \geq 0$), if the assignment $x \mapsto g(x) + \frac{\rho}{2}\|x\|^2$ is a convex function. The algorithms we consider will all use generalized derivative constructions. Variational analytic literature highlights a number of distinct subdifferentials (e.g. [20, 25, 28, 33]); for weakly convex functions, all of these constructions coincide. Consider a ρ-weakly convex function g. The *subdifferential* of g at x, denoted $\partial g(x)$, is the set of all vectors $v \in \mathbb{R}^d$ satisfying

$$g(y) \geq g(x) + \langle v, y - x \rangle + o(\|y - x\|) \qquad \text{as } y \to x. \tag{3.3.1}$$

Though the condition (3.3.1) appears to lack uniformity with respect to the basepoint x, the subgradients of g automatically satisfy the much stronger property:

$$g(y) \geq g(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2, \qquad \forall x, y \in \mathbb{R}^d,\ v \in \partial g(x). \tag{3.3.2}$$

Indeed, this follows quickly by applying the subgradient inequality to the convex function $g + \frac{\rho}{2}\|\cdot\|^2$. Thus we may use the two conditions, (3.3.1) and (3.3.2), interchangeably for weakly convex functions. We note in passing that localizing condition (3.3.2) leads to so-called prox-regular functions, introduced in [29].

Weakly convex functions are widespread in applications and are typically easy to recognize. One common source is the composite problem class:

$$\min_x\ F(x) := h(c(x)), \tag{3.3.3}$$

where $h\colon \mathbb{R}^m \to \mathbb{R}$ is convex and L-Lipschitz, and $c\colon \mathbb{R}^d \to \mathbb{R}^m$ is a C^1-smooth map with β-Lipschitz gradient. An easy argument shows that F is $L\beta$-weakly convex. This is a worst case estimate. In concrete circumstances, the composite function F may have a much more favorable weak convexity constant ρ. The elements of the subdifferential $\partial F(x)$ are straightforward to compute through the chain rule [33, Theorem 10.6, Corollary 10.9]:

$$\partial F(x) = \nabla c(x)^* \partial h(c(x)) \qquad \text{for all } x \in \mathbb{R}^d.$$

For a discussion of some recent uses of weakly convex functions in optimization, see the short survey [12]. Throughout the paper, we will use the following two running examples to illustrate our results.

Example 3.3.1 (Phase retrieval)**.** Phase retrieval is a common computational problem, with applications in diverse areas such as imaging, X-ray crystallography, and speech processing. For simplicity, we will focus on the version of the problem over the reals. The (real) phase retrieval problem seeks to determine a point x satisfying the magnitude conditions,

$$|\langle a_i, x\rangle|^2 \approx b_i \quad \text{for } i = 1, \ldots, m,$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are given. Note that we can only recover the optimal x up to a universal sign change, since $|\langle a_i, x\rangle| = |\langle a_i, -x\rangle|$. In this work, we will focus on the following optimization formulation of the problem [10, 15, 17]:

$$\min_x \ \frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle^2 - b_i|.$$

Clearly, this is an instance of (3.3.3). Indeed, under mild statistical assumptions on the way a_i are generated, the formulation is ρ-weakly convex, for some numerical constant ρ independent of d and m [15, Corollary 3.2]. Moreover, under an appropriate model of the noise in the measurements, the problem is sharp [15, Propostion 3]. It is worthwhile to mention that numerous other approaches to phase retrieval exist, based on different problem formulations; see for example [4, 5, 35, 37].

Experiment set-up: All of the experiments on phase retrieval will be generated according to the following procedure. In the *exact set-up*, we generate standard Gaussian measurements $a_i \sim N(0, I_{d\times d})$, for $i = 1, \ldots, m$, and generate the target signal $\bar{x} \sim N(0, I_{d\times d})$. We then set $b_i = \langle a_i, \bar{x}\rangle^2$ for each $i = 1, \ldots, m$. In the *corrupted set-up*, we generate a_i and $\bar{x}$ as in the exact case. We then corrupt a proportion of the measurements with outliers. Namely, we set $b_i = (1 - z_i)\langle a_i, \bar{x}\rangle^2 + z_i|\zeta_i|$, where $z_i \sim$ Bernoulli(0.1) and $\zeta_i \sim \mathcal{N}(0, 100)$.

Example 3.3.2 (Covariance matrix estimation)**.** The problem of covariance estimation from quadratic measurements, introduced in [6], is a higher rank variant of phase retrieval. Let $a_1, \ldots, a_m \in \mathbb{R}^d$ be measurement vectors. The goal is to recover a low rank decomposition of a covariance matrix $\bar{X}\bar{X}^T$, with $\bar{X} \in \mathbb{R}^{d\times r}$ for a given $0 \le r \le d$, from quadratic measurements

$$b_i \approx a_i^T \bar{X}\bar{X}^T a_i = \mathrm{trace}(\bar{X}\bar{X}^T a_i a_i^T).$$

Note that we can only recover $\bar{X}$ up to multiplication by an orthogonal matrix. This problem arises in a variety of contexts, such as covariance sketching for data streams and spectrum estimation of stochastic processes. We refer the reader to [6] for details. In our examples, we

will assume m is even and will focus on the potential function[2]

$$\min_{X\in\mathbb{R}^{d\times r}} \frac{1}{m}\sum_{i=1}^{m}\left|\left\langle XX^T, a_{2i}a_{2i}^T - a_{2i-1}a_{2i-1}^T\right\rangle - (b_{2i}-b_{2i-1})\right|. \tag{3.3.4}$$

Under exact measurements, i.e., $b_i = a_i^T\bar{X}\bar{X}^Ta_i$ and under appropriate statistical assumptions on how a_i are generated, the formulation (3.3.4) is ρ-weakly convex for a numerical constant ρ, independent of d or m, and is sharp. Indeed, this is a simple consequence of two results, namely [6, Corollary 1] and [38, Lemma 5.4]. It is possible to show that the objective is also sharp when the measurements are corrupted by gross outliers. This guarantee is beyond the scope of our current work, and will appear in a different paper.

Experiment set-up: All of the experiments on covariance matrix estimation will be generated according to the following procedure. In the *exact set-up*, we generate standard Gaussian measurements $a_i \sim N(0, I_{d\times d})$ for $i = 1,\ldots,m$, and generate the target matrix $\bar{X}\in\mathbb{R}^{d\times r}$ as a standard Gaussian. We then set $b_i = \|\bar{X}^Ta_i\|_F^2$ for each $i = 1,\ldots,m$. In the *corrupted set-up*, we generate a_i and $\bar{X}$ as in the exact case. We then corrupt a proportion of the measurements with outliers. Namely, we set $b_i = (1-z_i)\|\bar{X}^Ta_i\|_F^2 + z_i|\zeta_i|$, where $\zeta_i \sim \mathcal{N}(0,100)$ and $z_i \sim$ Bernoulli(0.1). All plots will show iteration counter k versus the scaled Procrustes distance $\mathrm{dist}(X_k,\mathcal{X}^*)/\|\bar{X}\| = \min_{\Omega^T\Omega=I}\|\Omega X_k - \bar{X}\|_F/\|\bar{X}\|_F$.

3.3.2 Setting of the Paper

Throughout the manuscript, we make the following assumption.

Assumption A. Consider the optimization problem

$$\min_{x\in\mathcal{X}} g(x), \tag{3.3.5}$$

satisfying the following properties for some real $\rho \geq 0$ and $\mu > 0$.

[2]Note this potential function is different from the direct generalization of the one used in phase retrieval, $\min_x \sum_i \left|\|X^Ta_i\|^2 - b_i\right|$. The reason for the more exotic formulation is that the differences $a_{2i}a_{2i}^T - a_{2i-1}a_{2i-1}^T$ in (3.3.4) allow the authors of [6] to prove sharpness of the objective function by appealing to an ℓ_2/ℓ_1 restricted isometry property of the resulting linear map. See [6, Section III.B] for details. It is not clear if the naive objective function is also sharp with high probability under reasonable statistical assumptions.

1. **(Weak-convexity)** The function $g\colon \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex, and the set $\mathcal{X} \subset \mathbb{R}^d$ is closed and convex. The set of minimizers $\mathcal{X}^* := \operatorname*{argmin}_{x\in\mathcal{X}} g(x)$ is nonempty.

2. **(Sharpness)** The inequality

$$g(z) - \min_{x\in\mathcal{X}} g(x) \geq \mu \cdot \operatorname{dist}(z; \mathcal{X}^*) \qquad \text{holds for all } z \in \mathcal{X}.$$

We will say that a point $\bar{x} \in \mathcal{X}$ is *stationary* for the target problem (3.3.5) if

$$g(x) - g(\bar{x}) \geq o(\|x - \bar{x}\|) \qquad \text{as } x \to \bar{x} \text{ in } \mathcal{X}.$$

That is, $\bar{x}$ is *stationary* precisely when the zero vector is a subgradient of $g + \delta_{\mathcal{X}}$ at $\bar{x}$.

Shortly, we will discuss subgradient methods that converge linearly to $\mathcal{X}^*$ under appropriate initialization. As a first step, therefore, we must identify a neighborhood of $\mathcal{X}^*$ that is devoid of extraneous stationary points of (3.3.5). This is the content of the following lemma.

Lemma 3.3.1 (Neighborhood with no stationary points)**.**
The problem (3.3.5) *has no stationary points x satisfying*

$$0 < \operatorname{dist}(x; \mathcal{X}^*) < \frac{2\mu}{\rho}. \tag{3.3.6}$$

Proof. Fix a stationary point $x \in \mathcal{X} \setminus \mathcal{X}^*$ of (3.3.5). Choosing an arbitrary $\bar{x} \in \operatorname{proj}_{\mathcal{X}^*}(x)$, observe

$$\mu \cdot \operatorname{dist}(x; \mathcal{X}^*) \leq g(x) - g(\bar{x}) \leq \frac{\rho}{2}\|x - \bar{x}\|^2 = \frac{\rho}{2} \cdot \operatorname{dist}^2(x; \mathcal{X}^*).$$

Dividing through by $\operatorname{dist}(x; \mathcal{X}^*)$, the result follows. □

In light of Lemma 3.3.1, for any $\gamma > 0$ we define the following tube

$$\mathcal{T}_\gamma := \left\{ x \in \mathcal{X} : \operatorname{dist}(x; \mathcal{X}^*) < \gamma \cdot \frac{\mu}{\rho} \right\}$$

and the constant

$$L := \sup \left\{ \|\zeta\| : \zeta \in \partial g(x),\ x \in \mathcal{T}_1 \right\}. \tag{3.3.7}$$

Lemma 3.3.1 guarantees that the tubes $\mathcal{T}_\gamma$ contain no extraneous stationary points of the problem for any $0 < \gamma \leq 2$. Moreover, observe that μ and L play reciprocal roles; consequently, the ratio $\tau := \mu/L$ should serve as a measure of conditioning. The following lemma verifies the inclusion $\tau \in [0, 1]$.

Lemma 3.3.2 (Condition measure)**.** *The inclusion* $\tau \in [0, 1]$ *holds.*

Proof. Consider an arbitrary point $x \in \mathcal{T}_1 \setminus \mathcal{X}^*$ and choose any $\bar{x} \in \mathrm{proj}_{\mathcal{X}^*}(x)$. By Lebourg's mean value theorem [8, Theorem 2.3.7], there exists a point z in the open segment $(x, \bar{x})$ and a vector $\zeta \in \partial g(z)$ satisfying

$$g(x) - g(\bar{x}) = \langle \zeta, x - \bar{x} \rangle. \tag{3.3.8}$$

Trivially z lies in $\mathcal{T}_1$, and therefore $\|\zeta\| \leq L$. Using this estimate and sharpness in (3.3.8) yields the guarantee

$$\mu \cdot \mathrm{dist}(x; \mathcal{X}^*) \leq g(x) - g(\bar{x}) \leq \|\zeta\| \cdot \|x - \bar{x}\| \leq L \cdot \mathrm{dist}(x; \mathcal{X}^*).$$

The result follows. □

To summarize, we will use the following symbols to describe the parameters of the problem class (3.3.5): ρ is the weak convexity constant of g, μ is the sharpness constant of g, L is the maximal subgradient norm at points in the tube $\mathcal{T}_1$, and τ is the condition measure $\tau = \mu/L \in [0, 1]$.

3.4 Polyak subgradient method

In this section, we consider the Polyak subgradient method for the problem (3.3.5). A preliminary version of this material applied to the phase retrieval problem appeared in [10]; we present the arguments here for the sake of completeness.

The Polyak subgradient method is summarized in Algorithm 3. The method requires knowing the optimal value $\min_{x \in \mathcal{X}} g(x)$. In a number of circumstances, this indeed is

reasonable (e.g. exact penalty approach for solving nonlinear equations). The latter sections explore subgradient methods that do not require a known optimal value.

Algorithm 3: Polyak Subgradient Method

Data: Initial point $x_0 \in \mathbb{R}^d$

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial g(x_k)$

if $\zeta_k \neq 0$ **then**

Set $x_{k+1} = \mathrm{proj}_{\mathcal{X}} \left(x_k - \frac{g(x_k) - \min_{x \in \mathcal{X}} g(x)}{\|\zeta_k\|^2} \zeta_k \right)$

else

Set $x_{k+1} = x_k$

end

The following theorem shows that Algorithm 3, originally proposed for convex problems, enjoys the same linear convergence guarantees for functions that are only weakly convex, provided it is initialized within a certain tube of the optimal solution set.

Theorem 3.4.1 (Linear rate)**.** *Fix a real* $0 < \gamma < 1$*. Then Algorithm 3 initialized at any point* $x_0 \in \mathcal{T}_\gamma$ *produces iterates that converge Q-linearly to* $\mathcal{X}^*$*. That is, for each index* $k \geq 0$ *we have*

$$\mathrm{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \left(1 - (1-\gamma)\tau^2\right) \mathrm{dist}^2(x_k; \mathcal{X}^*). \tag{3.4.1}$$

Proof. We proceed by induction. Suppose that the theorem holds up to iteration $k-1$. We will prove the inequality (3.4.1). To this end, observe that the inductive hypothesis implies, $\mathrm{dist}(x_k; \mathcal{X}^*) \leq \mathrm{dist}(x_0; \mathcal{X}^*)$, and therefore x_k lies in $\mathcal{T}_\gamma$. Observe that in the case $\zeta_k = 0$, we have $x_{k+1} = x_k$ by the definition of the subgradient method and $x_k \in \mathcal{X}^*$ by Lemma 3.3.1; hence (3.4.1) holds trivially. Therefore, we may suppose $\zeta_k \neq 0$ for the rest of the proof. Choose now a point $\bar{x} \in \mathrm{proj}_{\mathcal{X}^*}(x_k)$. Taking into account that $\mathrm{proj}_{\mathcal{X}}$ is nonexpansive, we

successively deduce

$$\begin{aligned}
\|x_{k+1} - \bar{x}\|^2 &\leq \left\| (x_k - \bar{x}) - \tfrac{g(x_k)-g(\bar{x})}{\|\zeta_k\|^2}\zeta_k \right\|^2 \\
&= \|x_k - \bar{x}\|^2 + \frac{2(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2} \cdot \langle \zeta_k, \bar{x} - x_k \rangle + \frac{(g(x_k) - g(\bar{x}))^2}{\|\zeta_k\|^2} \\
&\leq \|x_k - \bar{x}\|^2 + \frac{2(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2} \left(g(\bar{x}) - g(x_k) + \frac{\rho}{2}\|x_k - \bar{x}\|^2 \right) \\
&\quad + \frac{(g(x_k) - g(\bar{x}))^2}{\|\zeta_k\|^2} \\
&= \|x_k - \bar{x}\|^2 + \frac{(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2} \left(\rho\|x_k - \bar{x}\|^2 - (g(x_k) - g(\bar{x})) \right) \\
&\leq \|x_k - \bar{x}\|^2 + \frac{(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2} \left(\rho\|x_k - \bar{x}\|^2 - \mu\|x_k - \bar{x}\| \right) \\
&= \|x_k - \bar{x}\|^2 + \frac{\rho(g(x_k) - g(\bar{x}))}{\|\zeta_k\|^2} \left(\|x_k - \bar{x}\| - \frac{\mu}{\rho} \right) \|x_k - \bar{x}\|.
\end{aligned}$$

Combining the inclusion $x_k \in \mathcal{T}_\gamma$ with sharpness, we therefore deduce

$$\operatorname{dist}^2(x_{k+1}; \mathcal{X}^*) \leq \|x_{k+1} - \bar{x}\|^2 \leq \left(1 - \frac{(1-\gamma)\mu^2}{\|\zeta_k\|^2} \right) \|x_k - \bar{x}\|^2.$$

The result follows. □

As a numerical illustration, we apply the Polyak subgradient method (Figure 3.4.1) to our two running examples, phase retrieval and covariance matrix estimation. Notice that a linear rate of convergence is observed in all experiments except for two, with the rate improving monotonically with an increasing number of measurements m. In the two exceptional experiments, the number of measurements m is too small to guarantee that the initial point x_0 is within the basin of attraction, and the subgradient methods stagnates.

3.5 Subgradient method with constant stepsize

Recall that the Polyak subgradient method (Algorithm 3) crucially relies on knowing the minimal value of the optimization problem (3.3.5). Henceforth, all the subgradient methods we consider are agnostic to this value. That being said, they will require some estimates on the

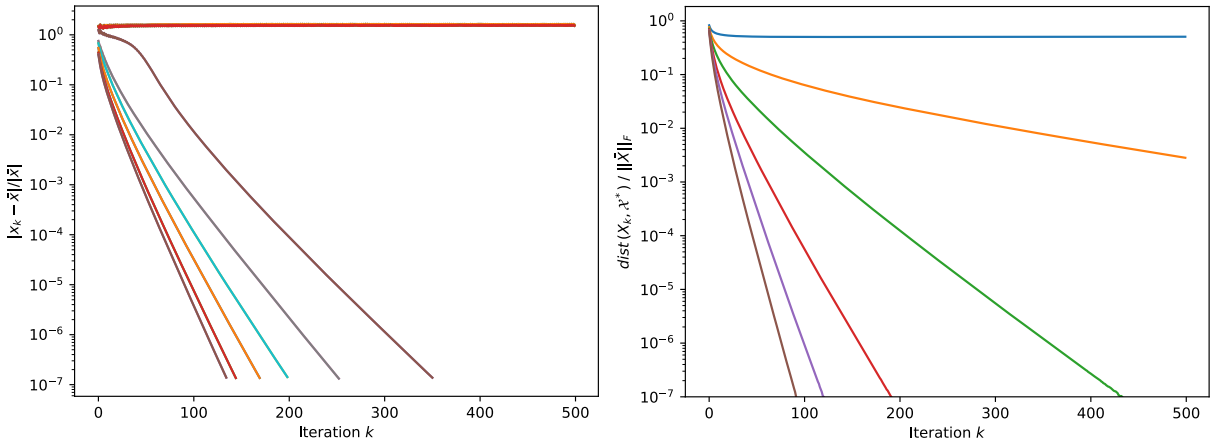


Figure 3.4.1: Polyak subgradient method. (Left) Phase retrieval with the exact set-up; $d = 5000$ and $m \in \{11000, 12225, 13500, 14750, 16000, 17250, 18500\}$. (Right) Covariance matrix estimation with the exact set-up; $d = 1000$, $r = 3$, $m \in \{5000, 8000, 11000, 14000, 17000, 20000\}$. In both experiments, convergence rates uniformly improve with increasing m.

problem parameters (μ, ρ, L). We begin by analyzing a subgradient method with a constant stepsize (Algorithm 4). Constant-stepsize schemes are often methods of choice in practice. We will show that when properly initialized, the subgradient method with constant stepsize generates iterates x_k such that $\text{dist}(x_k; \mathcal{X}^*)$ converges linearly up to a certain threshold.

Algorithm 4: Subgradient method with constant stepsize

Data: Initial point $x_0 \in \mathbb{R}^d$ and stepsize $\alpha > 0$

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial g(x_k)$

if $\zeta_k \neq 0$ **then**

Set $x_{k+1} = \text{proj}_{\mathcal{X}} \left(x_k - \alpha \cdot \frac{\zeta_k}{\|\zeta_k\|} \right)$

else

Set $x_{k+1} = x_k$

end

The analysis we present fundamentally relies on the following estimate, often used in the analysis of subgradient methods. To simplify notation, for any point $x \in \mathbb{R}^d$, we set

$$E(x) := \operatorname{dist}^2(x; \mathcal{X}^*).$$

Whenever x has an index k as a subscript, we will set $E_k := E(x_k)$. The following lemma will feature in both the constant and geometrically decaying stepsize schemes.

Lemma 3.5.1 (Basic recurrence)**.** *Consider a point $x \in \mathcal{T}_1$ and a nonzero subgradient $\zeta \in \partial g(x)$, and define $x^+ := \operatorname{proj}_{\mathcal{X}}\left(x - \alpha\frac{\zeta}{\|\zeta\|}\right)$ for some $\alpha > 0$. Then the estimate holds:*

$$E(x^+) \leq \left(1 + \tfrac{\rho\alpha}{L}\right) E(x) - 2\alpha\tau\sqrt{E(x)} + \alpha^2. \tag{3.5.1}$$

Proof. Choose an arbitrary point $\bar{x} \in \operatorname{proj}_{\mathcal{X}^*}(x)$. Observe

$$\begin{aligned}
\|x^+ - \bar{x}\|^2 &\leq \left\|(x - \bar{x}) - \alpha\tfrac{\zeta}{\|\zeta\|}\right\| = \|x - \bar{x}\|^2 + \tfrac{2\alpha}{\|\zeta\|} \cdot \langle \zeta, \bar{x} - x\rangle + \alpha^2 \\
&\leq \|x - \bar{x}\|^2 + \tfrac{2\alpha}{\|\zeta\|} \cdot \left(g(\bar{x}) - g(x) + \frac{\rho}{2}\|x - \bar{x}\|^2\right) + \alpha^2 \\
&\leq \left(1 + \frac{\alpha\rho}{\|\zeta\|}\right)\|x - \bar{x}\|^2 - \frac{2\alpha\mu}{\|\zeta\|} \cdot \|x - \bar{x}\| + \alpha^2.
\end{aligned}$$

Thus the inequality holds:

$$E(x^+) \leq \left(1 + \tfrac{\alpha\rho}{\|\zeta\|}\right) E(x) - \tfrac{2\alpha\mu}{\|\zeta\|} \cdot \sqrt{E(x)} + \alpha^2,$$

and consequently taking into account $\alpha/\|\zeta\| \geq \alpha/L$ we have

$$E(x^+) \leq \sup_{t \geq \alpha/L} \left\{(1 + \rho t)\, E(x) - 2\mu t \cdot \sqrt{E(x)} + \alpha^2\right\}.$$

Notice, the function inside the supremum is linear in t with slope given by $s := \rho E(x) - 2\mu\sqrt{E(x)}$. The inclusion $x \in \mathcal{T}_1$ directly implies $s \leq 0$. Therefore the supremum on the right-hand-side is attained at $t = \frac{\alpha}{L}$, yielding the claimed estimate (3.5.1). □

In light of Lemma 3.5.1, we can now prove that the quantities $E_k = E(x_k)$ converge linearly below a certain fixed threshold. The proof is a modification of that in [23, Section 4].

Lemma 3.5.2 (Contraction inequality)**.** *Fix a constant* $0 < \alpha < \frac{\tau\mu}{\rho}$ *and let* $\{x_k\}_{k\geq 0}$ *be the iterates generated by Algorithm 4. Define the quantity*

$$E^* := \left(\frac{\alpha L}{\mu + \sqrt{\mu^2 - \alpha\rho L}}\right)^2. \tag{3.5.2}$$

Then whenever an iterate x_k *lies in* $\mathcal{T}_1$*, the estimate holds:*

$$E_{k+1} - E^* \leq q_k(E_k - E^*),$$

where $q_k := 1 + \frac{\alpha}{L}\left(\rho - \frac{2\mu}{\sqrt{E_k}+\sqrt{E^*}}\right)$ *satisfies* $q_k < 1$.

Proof. Looking back at the estimate (3.5.1), consider the following equation in the variable e:

$$e = \left(1 + \tfrac{\alpha\rho}{L}\right) e - 2\alpha\tau \cdot \sqrt{e} + \alpha^2. \tag{3.5.3}$$

An easy computation shows that the minimal positive solution to (3.5.3) is exactly E^*, defined in (3.5.2). Note that E^* is well-defined by the inequality $\alpha \leq \tau \cdot \frac{\mu}{\rho}$.

Subtracting (3.5.3) from (3.5.1) yields the estimate

$$\begin{aligned} E_{k+1} - E^* &\leq (1 + \tfrac{\alpha\rho}{L})(E_k - E^*) - 2\alpha\tau(\sqrt{E_k} - \sqrt{E^*}) \\ &= \left(1 + \tfrac{\alpha}{L}\left(\rho - \tfrac{2\mu}{\sqrt{E_k}+\sqrt{E^*}}\right)\right)(E_k - E^*). \end{aligned}$$

Finally, noting

$$\sqrt{E^*} = \frac{\alpha L}{\mu + \sqrt{\mu^2 - \alpha\rho L}} = \frac{\mu - \sqrt{\mu^2 - \alpha\rho L}}{\rho} \leq \frac{\mu}{\rho},$$

we obtain

$$\rho - \frac{2\mu}{\sqrt{E_k} + \sqrt{E^*}} < \rho - \frac{2\mu}{2\mu/\rho} = 0.$$

This completes the proof of the lemma. □

Iterating Lemma 3.5.2, we see that the quantities E_k decrease to a value at most E^* at a linear rate. Figure 3.5.1 illustrates this behavior on our two running examples. It is also clear from the figure that the linear rate of convergence improves as E_k tends to E^*. An

explanation is immediate from the expression for q_k in Lemma 3.5.2. Indeed, as E_k decreases, so do the contraction factors q_k, and for $E_k \approx E^*$, we have $q_k \approx (1-\tau^2) + \frac{\alpha\rho}{L}$. Thus as the stepsize tends to zero, the limiting linear rate coincides with an ideal rate of $\approx 1-\tau^2$.

Even though E_k could be smaller than E^* and q could be negative, it is apparent from Figure 3.5.1 that even after E_k becomes smaller than E^*, all the following values E_k stay close to E^*. This is the content of the following theorem. The convex version of this theorem appears in [23, Theorem 1].

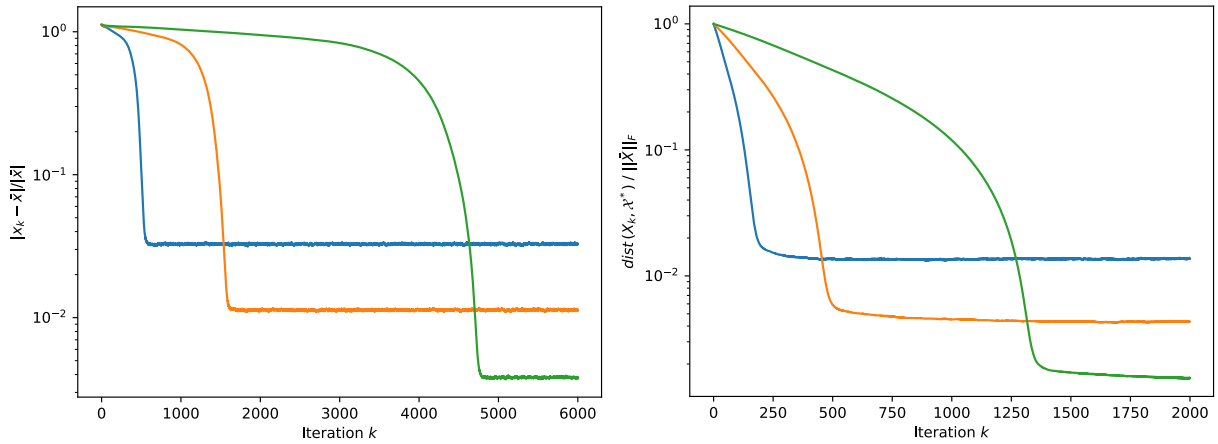


Figure 3.5.1: Constant stepsize subgradient method. (Left) Phase retrieval with the corrupted set-up; $d = 1000$, $m = 3000$, and $\alpha \in \{1, 1/3, 1/9\}$. (Right) Covariance matrix estimation with the corrupted set-up; $d = 1000$, $r = 3$, $m = 10000$, and $\alpha \in \{1, 1/3, 1/9\}$. The lower curves correspond to smaller stepsizes in both experiments.

Theorem 3.5.1 (Convergence of fixed stepsize subgradient method)**.**
Fix constants $0 < \gamma < 1$ *and* α *satisfying*

$$0 < \alpha < \frac{\gamma\tau}{\sqrt{1+2\tau^2}} \cdot \frac{\mu}{\rho}. \tag{3.5.4}$$

Let x_k *be the iterates generated by Algorithm 4 with stepsize* α *and initial point* $x_0 \in \mathcal{T}_\gamma$.

Define the constants

$$E^* := \left(\frac{\alpha L}{\mu + \sqrt{\mu^2 - \alpha\rho L}}\right)^2 \qquad \text{and} \qquad D := \sqrt{\max\{E_0, 2\alpha^2 + E^*\}}.$$

Then for each index k, the estimates hold:

$$\sqrt{E_k} \leq D < \tfrac{\gamma\mu}{\rho} \qquad \text{and} \qquad E_k - E^* \leq \max\left\{q^k(E_0 - E^*), 2\alpha^2\right\},$$

where the coefficient $q := 1 + \frac{\alpha}{L}\left(\rho - \frac{\mu}{D}\right)$ satisfies $0 < q < 1$.

Proof. We first verify the claims that are independent of the iteration counter. To this end, observe that (3.5.4) directly implies $\alpha < \tau \cdot \frac{\mu}{\rho}$, and therefore E^* is well defined. Next, we show $D < \frac{\gamma\mu}{\rho}$. Indeed, noting $\sqrt{E^*} < \alpha\tau^{-1}$ and using (3.5.4), we deduce

$$D^2 = \max\{E_0, 2\alpha^2 + E^*\} \leq \max\{E_0, \alpha^2(2 + \tau^{-2})\} < \left(\frac{\gamma\mu}{\rho}\right)^2.$$

Next, we show the inequalities $0 < q < 1$. To this end, observe

$$-1 \leq -\frac{\tau\alpha}{D} < \tfrac{\alpha}{L}\left(\rho - \tfrac{\mu}{D}\right) \leq (1 - \gamma^{-1}) \cdot \frac{\rho\alpha}{L} < 0,$$

where the first inequality follows from the inequality, $\alpha \leq D$, and the third follows from the inequality, $D < \frac{\gamma\mu}{\rho}$. Thus we conclude $0 < q < 1$, as claimed.

We now proceed by induction. Fix an index k and suppose as inductive hypothesis that for each index $i = 0, 1, \ldots, k$, the estimates hold:

$$\sqrt{E_i} \leq D \qquad \text{and} \qquad E_i - E^* \leq \max\left\{q^i(E_0 - E^*), 2\alpha^2\right\}.$$

Let us consider two cases. Suppose first $E_k \geq E^*$. Then by applying Lemma 3.5.2, we deduce

$$\begin{aligned} E_{k+1} - E^* &\leq \left(1 + \tfrac{\alpha}{L}\left(\rho - \tfrac{2\mu}{\sqrt{E_k} + \sqrt{E^*}}\right)\right)(E_k - E^*) \\ &\leq \left(1 + \tfrac{\alpha}{L}\left(\rho - \tfrac{\mu}{D}\right)\right)(E_k - E^*) \\ &= q(E_k - E^*). \end{aligned}$$

Suppose now that the second case, $E_k < E^*$, holds. Then Lemma 3.5.2 implies

$$\begin{aligned} E_{k+1} &\leq E_k + \tfrac{\alpha\rho}{L}(E_k - E^*) - 2\alpha\tau(\sqrt{E_k} - \sqrt{E^*}) \\ &\leq \max_{E \in [0, E^*]} \{E + \tfrac{\alpha\rho}{L}(E - E^*) - 2\alpha\tau(\sqrt{E} - \sqrt{E^*})\} \\ &= \max\{E^*, \tfrac{\alpha}{L}(2\mu\sqrt{E^*} - \rho E^*)\}. \end{aligned}$$

Subtracting E^*, we conclude

$$E_{k+1} - E^* \leq \max\{0, 2\tau\alpha\sqrt{E^*}\} \leq 2\alpha^2.$$

Thus in both cases, we have the estimate

$$E_{k+1} - E^* \leq \max\left\{q(E_k - E^*), 2\alpha^2\right\}.$$

In particular, we immediately deduce $\sqrt{E_{k+1}} \leq D$. Applying the inductive hypothesis, we conclude

$$\begin{aligned} E_{k+1} - E^* &\leq \max\left\{q \cdot \max\{q^k(E_0 - E^*), 2\alpha^2\}, 2\alpha^2\right\} \\ &= \max\{q^{k+1}(E_0 - E^*), 2\alpha^2\}. \end{aligned}$$

The theorem is proved. □

3.6 Geometrically decaying stepsize

In the previous section, we showed linear convergence of the constant stepsize scheme up to a fixed tolerance E^*. To obtain a linearly convergent method to the true solution set, we will allow the stepsize to decrease geometrically. The analogous strategy in the convex setting goes back to Goffin [19], and our argument is a direct generalization of [19] to the weakly convex setting. The intuition for why one may expect linear convergence under such stepsizes may be gleaned from the Polyak method under the optimal stepsize

$$\alpha_k = \frac{g(x_k) - \min_{x \in \mathcal{X}} g(x)}{\|\zeta_k\|}.$$

It is easy to verify that since the E_k tend to zero Q-linearly, the stepsizes α_k tend to zero R-linearly. We implement such a geometrically decaying stepsize in Algorithm 5 and prove linear convergence of the method in Theorem 3.6.1. In the proof, we use the notation $E_k := \operatorname{dist}^2(x_k; \mathcal{X}^*)$ as in the previous section.

Algorithm 5: Subgradient method with geometrically decaying stepsize

Data: Constants $\lambda > 0$ and $0 < q < 1$.

Step k**:** $(k \geq 0)$

Choose $\zeta_k \in \partial g(x_k)$

Set stepsize $\alpha_k = \lambda \cdot q^k$

if $\zeta_k \neq 0$ **then**

Set $x_{k+1} = \operatorname{proj}_{\mathcal{X}} \left(x_k - \alpha_k \frac{\zeta_k}{\|\zeta_k\|} \right)$

else

Set $x_{k+1} = x_k$

end

Theorem 3.6.1. *Fix a real* $0 < \gamma < 1$ *and suppose* $\tau \leq \sqrt{\frac{1}{2-\gamma}}$. *Set*

$$0 < \lambda \leq \frac{\gamma \mu^2}{\rho L} \quad \text{and} \quad q := \sqrt{1 - (1-\gamma)\tau^2},$$

where λ *is assumed finite. Then the iterates* x_k *generated by Algorithm 5, initialized at some point* x_0 *with* $\operatorname{dist}(x_0; \mathcal{X}^*) \leq \frac{\lambda}{\tau - \sqrt{\tau^2 - (1 - \rho\lambda L^{-1} - q^2)}}$, *satisfy:*

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \max\left\{ \tfrac{\lambda^2}{\tau^2}, \operatorname{dist}(x_0; \mathcal{X}^*)^2 \right\} \left(1 - (1-\gamma)\tau^2\right)^k. \tag{3.6.1}$$

Proof. We will prove the result by induction. To this end, suppose the bound (3.6.1) holds for all $i = 0, \ldots, k$. Appealing to Lemma 3.5.1. and using the relation $\alpha_k = \lambda q^k$, we obtain

$$E_{k+1} \leq \left(1 + \frac{\rho \lambda q^k}{L}\right) E_k - 2\lambda \tau q^k \sqrt{E_k} + \lambda^2 q^{2k}. \tag{3.6.2}$$

Define the constant $M := \max\left\{ \frac{\lambda}{\tau}, \operatorname{dist}(x_0; \mathcal{X}^*) \right\}$. Recall the induction assumption guarantees $\sqrt{E_k} \leq M q^k$. Let us therefore fix some value $R \in [0, M]$ satisfying $\sqrt{E_k} = R q^k$. Inequality

(3.6.2) then implies

$$E_{k+1} \leq \max_{R\in[0,M]} \left\{ R^2 q^{2k} + \frac{\rho\lambda R^2}{L} q^{3k} - 2\lambda\tau R q^{2k} + \lambda^2 q^{2k} \right\}.$$

Note that the expression inside the maximum is a convex quadratic in R and therefore the maximum must occur either at $R = 0$ or $R = M$. We therefore deduce

$$E_{k+1} \leq q^{2k} \cdot \max\left\{ \lambda^2,\ M^2 + \frac{\rho\lambda}{L} M^2 q^k - 2\lambda\tau M + \lambda^2 \right\}. \tag{3.6.3}$$

To complete the induction, it is therefore sufficient to show

$$\lambda^2 \leq M^2 q^2 \qquad \text{and} \qquad M^2 + \frac{\rho\lambda}{L} M^2 q^k - 2\lambda\tau M + \lambda^2 \leq M^2 q^2. \tag{3.6.4}$$

We begin by verifying the first property. Note $M \geq \frac{\lambda}{\tau}$. Hence, it suffices to show that $\tau \leq q$. Observe that the assumption $\tau \leq \sqrt{\frac{1}{2-\gamma}}$ directly implies

$$\tau^2 + (1-\gamma)\tau^2 \leq 1.$$

Rearranging yields $\tau^2 \leq 1 - (1-\gamma)\tau^2 = q^2$. Hence, the first condition in (3.6.4) holds.

Next we show that M satisfies the second property in (3.6.4). Rearranging the expression, we must establish

$$\left(1 + \frac{\rho\lambda}{L} q^k - q^2\right) M^2 - 2\lambda\tau M + \lambda^2 \leq 0. \tag{3.6.5}$$

We will show that the quadratic on the left-hand-side in M has two real positive roots. To this end, a quick computation shows that the two roots are

$$\frac{\lambda\tau \pm \sqrt{\lambda^2\tau^2 - \lambda^2\left(1 + \frac{\rho\lambda}{L} q^k - q^2\right)}}{1 + \frac{\rho\lambda}{L} q^k - q^2} = \frac{\lambda}{\tau \mp \sqrt{\tau^2 - \left(1 + \frac{\rho\lambda}{L} q^k - q^2\right)}}.$$

To see that the discriminant is non-negative, observe

$$\tau^2 - \left(1 + \frac{\rho\lambda}{L} q^k - q^2\right) \geq \tau^2 - \left(1 + \frac{\rho\lambda}{L} - q^2\right) \geq \tau^2 - (1 + \gamma\tau^2 - q^2) = 0.$$

Thus the convex quadratic in (3.6.5) has two real roots. Observe $M \geq \lambda/\tau$ and therefore M is greater than the minimal root. It suffices therefore to argue that M is smaller than the

largest root. If $M = \lambda/\tau$, then this is clearly true. Otherwise, it must be that $M = \sqrt{E_0}$. Our assumption on the initial distance $\sqrt{E_0}$ then guarantees that $\sqrt{E_0}$ is smaller than the largest positive root of (3.6.5), as claimed. Hence the condition (3.6.5) holds, and the induction is complete. □

In the convex setting, $\rho = 0$, Theorem 3.6.1 recovers the guarantees in Goffin [19]. Theorem 3.6.1 demonstrates a precise relationship between the initial distance, the constant λ which controls the rate, and the constants τ and ρ. In particular, the larger the initial distance, the larger λ needs to be to guarantee convergence. It is now straightforward to see how one should set the parameters q and λ to guarantee linear convergence within the tube $\mathcal{T}_\gamma$.

Corollary 3.6.1. *Fix a real $0 < \gamma < 1$ and suppose $\tau \leq \sqrt{\frac{1}{2-\gamma}}$ and $\rho > 0$. Set*

$$\lambda := \frac{\gamma\mu^2}{\rho L} \quad \text{and} \quad q := \sqrt{1-(1-\gamma)\tau^2}.$$

Then the iterates x_k generated by Algorithm 5, initialized at any point $x_0 \in \mathcal{T}_\gamma$, satisfy:

$$\operatorname{dist}^2(x_k; \mathcal{X}^*) \leq \frac{\gamma^2\mu^2}{\rho^2}\left(1-(1-\gamma)\tau^2\right)^k. \tag{3.6.6}$$

Proof. Looking back at Theorem 3.6.1, we see that $\operatorname{dist}(x_0; \mathcal{X}^*)$ satisfies the assumed upper bound:

$$\operatorname{dist}(x_0; \mathcal{X}^*) < \frac{\gamma\mu}{\rho} = \frac{\lambda}{\tau} \leq \frac{\lambda}{\tau - \sqrt{\tau^2 - (1-\rho\lambda L^{-1} - q^2)}}.$$

Since $\frac{\lambda}{\tau} = \frac{\gamma\mu}{\rho}$, we conclude that $\max\left\{\operatorname{dist}(x_0; \mathcal{X}^*), \frac{\lambda}{\tau}\right\} \leq \frac{\gamma\mu}{\rho}$. The result immediately follows after applying Theorem 3.6.1. □

We now illustrate the performance of Algorithm 5 on our two running examples in Figure 3.6.1. Empirically, we observed that $\lambda > 0$ and $0 < q < 1$ must be tuned for performance, which is what we did in the experiments. We observe linear convergence in all cases, and the convergence rate of the method improves monotonically as the chosen rate q is decreased. The Polyak scheme pictured in Figure 3.4.1 clearly outperforms all other methods,

while the geometrically decaying stepsize scheme performs much better than the constant stepsize scheme in Figure 3.5.1. That being said, we should stress that the Polyak subgradient method requires knowledge of the optimal value of the problem, which is often unavailable. We also note that the work [23] provides schemes which can indeed automatically adapt to the unknown parameter μ for convex problems. We leave extensions of such algorithms to weakly convex problems for future work.

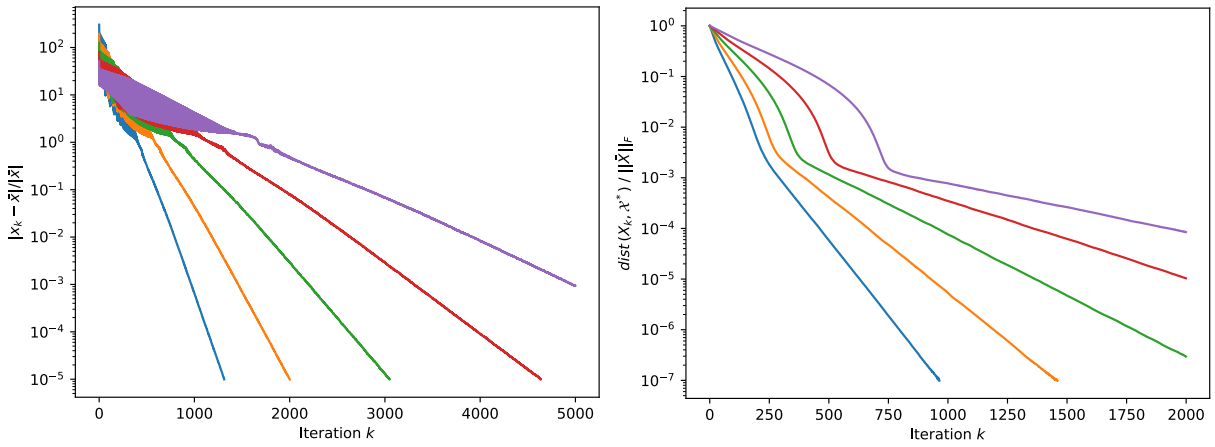


Figure 3.6.1: Geometrically decaying stepsize. (Left) Phase retrieval with the corrupted set-up; $d = 1000$, $m = 3000$, $q \in \{0.983, 0.989, 0.993, 0.996, 0.997\}$. (Right) Covariance matrix estimation with the corrupted set-up; $d = 1000$, $r = 3$, $m = 10000$, $q \in \{0.986, 0.991, 0.994, 0.996, 0.998\}$. The depicted rates uniformly improve with lower values of q, in both figures.

Acknowledgements. Research of Drusvyatskiy and MacPhee was partially supported by the AFOSR YIP award FA9550-15-1-0237 and by the NSF DMS 1651851 and CCF 1740551 awards. Research of Paquette was supported by NSF CCF 1740796.

References

[1] N. Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

[2] A. Brutzkus and A. Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv:1702.07966*, 2017.

[3] J. Burke and M. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[4] E. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[5] Y. Chen and E. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on Pure and Applied Mathematics*, 70(5):822–883, 2017.

[6] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[7] Y. Chen and M. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv:1509.03025*, 2015.

[8] F. Clarke. *Optimization and Nonsmooth Analysis*. Wiley Interscience, NY, 1983.

[9] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[10] D. Davis, D. Drusvyatskiy, and C. Paquette. The nonsmooth landscape of phase retrieval. *arXiv:1711.03247*, 2017.

[11] D. Davis and B. Grimmer. Proximally guided stochastic method for nonsmooth, nonconvex problems. *arXiv:1707.03505*, 2017.

[12] D. Drusvyatskiy. The proximal point method revisited. *To appear in SIAG/OPT Views and News, arXiv:1712.06038*, 2018.

[13] D. Drusvyatskiy and A. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Mathematics of Operations Research, arXiv:1602.06661*, 2016.

[14] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *arXiv:1605.00125*, 2016.

[15] J. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *arXiv:1705.02356*, 2017.

[16] J. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *arXiv:1703.08570*, 2017.

[17] Y. Eldar and S. Mendelson. Phase retrieval: stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[18] I. Eremin. The relaxation method of solving systems of inequalities with convex functions on the left-hand side. *Dokl. Akad. Nauk SSSR*, 160:994–996, 1965.

[19] J. Goffin. On convergence rates of subgradient optimization methods. *Mathematical Programming*, 13(3):329–347, 1977.

[20] A. Ioffe. *Variational Analysis of Regular Mappings*. Springer Monographs in Mathematics. Springer, Cham, 2017. Theory and applications.

[21] P. Jain, C. Jin, S. Kakade, and P. Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 479–488, 2017.

[22] P. Jane and P. Netrapalli. Fast exact matrix completion with finite samples. In P. Grnwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1007–1034, Paris, France, 03–06 Jul 2015. PMLR.

[23] P. Johnstone and P. Moulin. Faster subgradient methods for functions with Hölderian growth. *arXiv:1704.00196*, 2017.

[24] R. Meka, P. Jain, and I. Dhillon. Guaranteed rank minimization via singular value projection. *arXiv:0909.5457*, 2009.

[25] B. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006.

[26] E. A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[27] E. A. Nurminskii. Minimization of nondifferentiable functions in the presence of noise. *Cybernetics*, 10(4):619–621, Jul 1974.

[28] J.-P. Penot. *Calculus without Derivatives*, volume 266 of *Graduate Texts in Mathematics*. Springer, New York, 2013.

[29] R. Poliquin and R. Rockafellar. Prox-regular functions in variational analysis. *Transactions of the American Mathematical Society*, 348:1805–1838, 1996.

[30] B. Poljak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9:14–29, 1969.

[31] B. Poljak. Subgradient methods: a survey of Soviet research. In *Nonsmooth optimization (Proc. IIASA Workshop, Laxenburg, 1977)*, volume 3 of *IIASA Proc. Ser.*, pages 5–29. Pergamon, Oxford-New York, 1978.

[32] B. 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*, 3:369–380, 1979.

[33] R. Rockafellar and R.-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[34] N. Shor. The rate of convergence of the method of the generalized gradient descent with expansion of space. *Kibernetika (Kiev)*, 2:80–85, 1970.

[35] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *To appear in Foundations of Computational Mathematics, arXiv:1602.06664*, 2017.

[36] S. Supittayapornpong and M. Neely. Staggered time average algorithm for stochastic non-smooth optimization with $O(1/t)$ convergence. *arXiv:1607.02842*, 2016.

[37] Y. Tan and R. Vershynin. Phase retreival via randomized kaczmarz: Theoretical guarantees. *arXiv:1605.08285*, 2017.

[38] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 964–973. JMLR.org, 2016.

[39] T. Yang and Q. Lin. RSG: Beating subgradient method without smoothness and strong convexity. *arXiv:1512.03107*, 2016.

Chapter 4

STOCHASTIC MODEL-BASED MINIMIZATION UNDER HIGH-ORDER GROWTH

This chapter represents joint work with Damek Davis and Dmitriy Drusvyatskiy.

Abstract: Given a nonsmooth, nonconvex minimization problem, we consider algorithms that iteratively sample and minimize stochastic convex models of the objective function. Assuming that the one-sided approximation quality and the variation of the models is controlled by a Bregman divergence, we show that the scheme drives a natural stationarity measure to zero at the rate $O(k^{-1/4})$. Under additional convexity and relative strong convexity assumptions, the function values converge to the minimum at the rate of $O(k^{-1/2})$ and $\widetilde{O}(k^{-1})$, respectively. We discuss consequences for stochastic proximal point, mirror descent, regularized Gauss-Newton, and saddle point algorithms.

4.1 Introduction

Common stochastic optimization algorithms proceed as follows. Given an iterate x_t, the method samples a model of the objective function formed at x_t and declares the next iterate to be a minimizer of the model regularized by a proximal term. Stochastic proximal point, proximal subgradient, and Gauss-Newton type methods are common examples. Let us formalize this viewpoint, following [15]. Namely, consider the optimization problem

$$\min_{x\in\mathbb{R}^d} \; F(x) := f(x) + r(x). \tag{4.1.1}$$

where the function $r\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is closed and convex and the only access to $f\colon \mathbb{R}^d \to \mathbb{R}$ is by sampling a *stochastic one-sided model*. That is, for every point x, there exists a family of

models $f_x(\cdot, \xi)$ of f, indexed by a random variable $\xi \sim P$. This setup immediately motivates the following algorithm, analyzed in [15]:

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P, \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \left\{ f_{x_t}(x, \xi_t) + r(x) + \frac{1}{2\eta_t}\|x - x_t\|_2^2 \right\} \end{array} \right\}, \tag{4.1.2}$$

where $\eta_t > 0$ is an appropriate control sequence that governs the step-size of the algorithm.

Some thought shows that convergence guarantees of the method (4.1.2) should rely at least on two factors: (i) control over the approximation quality, $f_x(\cdot, \xi) - f(\cdot)$, and (ii) growth/stability properties of the individual models $f_x(\cdot, \xi)$. With this in mind, the paper [15] isolates the following assumptions:

$$\mathbb{E}_\xi[f_x(x, \xi)] = f(x) \qquad \text{and} \qquad \mathbb{E}_\xi[f_x(y, \xi) - f(y)] \leq \frac{\tau}{2}\|y - x\|_2^2 \qquad \forall x, y, \tag{4.1.3}$$

and there exists a square integrable function $L(\cdot)$ satisfying

$$f_x(x, \xi) - f_x(y, \xi) \leq L(\xi)\|x - y\|_2 \qquad \forall x, y. \tag{4.1.4}$$

Condition (4.1.3) simply says that in expectation, the model $f_x(\cdot, \xi)$ must globally lower bound $f(\cdot)$ up to a quadratic error, while agreeing with f at the base point x; when (4.1.3) holds, the paper [15] calls the assignment $(x, y, \xi) \mapsto f_x(y, \xi)$ a stochastic one-sided model of f. Property (4.1.4), in contrast, asserts a Lipschitz type property of the individual models $f_x(\cdot, \xi)$.[1] The main result of [15] shows that under these assumption, the scheme (4.1.2) drives a natural stationarity measure of the problem to zero at the rate $O(k^{-1/4})$. Indeed, the stationarity measure is simply the gradient of the Moreau envelope

$$F_\lambda(x) := \inf_y \left\{ F(y) + \tfrac{1}{2\lambda}\|y - x\|_2^2 \right\}, \tag{4.1.5}$$

where $\lambda > 0$ is a smoothing parameter on the order of τ.

The assumptions (4.1.3) and (4.1.4) are perfectly aligned with existing literature. Indeed, common first-order algorithms rely on global Lipschitz continuity of the objective function or of

[1] The stated assumption (A4) in [15] is stronger than (4.1.4); however, a quick look at the arguments shows that property (4.1.4) suffices to obtain essentially the same convergence guarantees.

its gradient; see for example the monographs [5,31,33]. Recent work [2,8,26,29,30], in contrast, has emphasized that global Lipschitz assumptions can easily fail for well-structured problems. Nonetheless, these papers show that it is indeed possible to develop efficient algorithms even without the global Lipschitz assumption. The key idea, originating in [2, 29, 30], is to model errors in approximation by a Bregman divergence, instead of a norm. The ability to deal with problems that are not globally Lipschitz is especially important in stochastic nonconvex settings, where line-search strategies that exploit local Lipschitz continuity are not well-developed.

Motivated by the recent work on relative continuity/smoothness [2, 29, 30], we extend the results of [15] to non-globally Lipschitzian settings. Formally, we simply replace the squared norm $\frac{1}{2}\|\cdot\|^2$ in the displayed equations (4.1.2)-(4.1.5) by a Bregman divergence

$$D_\Phi(y, x) = \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle,$$

generated by a Legendre function Φ. With this modification and under mild technical conditions, we will show that algorithm (4.1.2) drives the gradient of the Bregman envelope (4.1.5) to zero at the rate $O(k^{-1/4})$, where the size of the gradient is measured in the local norm induced by Φ. As a consequence, we obtain new convergence guarantees for stochastic proximal point, mirror descent[2], and regularized Gauss-Newton methods, as well as for an elementary algorithm for stochastic saddle point problems. Perhaps the most important application arena is when the functional components of the problem grow at a polynomial rate. In this setting, we present a simple Legendre function Φ that satisfies the necessary assumptions for the convergence guarantees to take hold. We also note that the stochastic mirror descent algorithm that we present here does not require mini-batching the gradients, in contrast to the previous seminal work [24].

When the stochastic models $f_x(\cdot, \xi)$ are themselves convex and globally under-estimate f in expectation, we prove that the scheme drives the expected functional error to zero at

[2]This work appears on arXiv a month after a preprint of Zhang and He [42], who provide similar convergence guarantees specifically for the stochastic mirror descent algorithm. The results of the two papers were obtained independently and are complementary to each other.

the rate $O(k^{-1/2})$. The rate improves to $\widetilde{O}(k^{-1})$ when the regularizer r is μ-strongly convex relative to Φ in the sense of [30]. In the special case of mirror descent, these guarantees extend the results for convex unconstrained problems in [29] to the proximal setting. Even specializing to the proximal subgradient method, the convergence guarantees appear to be different from those available in the literature. Namely, previous complexity estimates [7, 21] depend on the largest norms of the subgradients of r along the iterate sequence, whereas Theorems 4.7.1 and 4.7.2 replace this dependence only by the initial error $r(x_0) - \inf r$.

The outline of the manuscript is as follows. Section 4.2 reviews the relevant concepts of convex analysis, focusing on Legendre functions and the Bregman divergence. Section 4.3 introduces the problem class and the algorithmic framework. This section also interprets the assumptions made for the stochastic proximal point, mirror descent, and regularized Gauss-Newton methods, as well as for a stochastic approximation algorithm for saddle point problems. Section 4.4 discusses the stationarity measure we use to quantify the rate of convergence. Section 4.5 contains the complete convergence analysis of the stochastic model-based algorithm. Section 4.6 presents a specialized analysis for the mirror descent algorithm when f is smooth and the stochastic gradient oracle has finite variance. Finally, in Section 4.7 we prove convergence rates in terms of function values for stochastic model-based algorithms under (relative strong) convexity assumptions.

4.2 Legendre functions and the Bregman divergence

Throughout, we follow standard notation from convex analysis, as set out for example by Rockafellar [37]. The symbol $\mathbb{R}^d$ will denote an Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. For any set $Q \subset \mathbb{R}^d$, we let $\operatorname{int} Q$ and $\operatorname{cl} Q$ denote the interior and closure of Q, respectively. Whenever Q is convex, the set $\operatorname{ri} Q$ is the interior of Q relative to its affine hull. The effective domain of any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, denoted by $\operatorname{dom} f$, consists of all points where f is finite. Abusing notation slightly, we will use the symbol $\operatorname{dom}(\nabla f)$ to denote the set of all points where f is differentiable.

This work analyzes stochastic model-based minimization algorithms, where the "errors"

are controlled by a Bregman divergence. For wider uses of the Bregman divergence in first-order methods, we refer the interested reader to the expository articles of Bubeck [10], Juditsky-Nemirovski [27], and Teboulle [40].

Henceforth, we fix a *Legendre function* $\Phi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, meaning:

1. (Convexity) Φ is proper, closed, and strictly convex.

2. (Essential smoothness) The domain of Φ has nonempty interior, Φ is differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$, and for any sequence $\{x_k\} \subset \operatorname{int}(\operatorname{dom}\Phi)$ converging to a boundary point of $\operatorname{dom}\Phi$, it must be the case that $\|\nabla\Phi(x_k)\| \to \infty$.

Typical examples of Legendre functions are the squared Euclidean norm $\Phi(x) = \frac{1}{2}\|x\|_2^2$, the Shannon entropy $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$ with $\operatorname{dom}\Phi = \mathbb{R}^d_+$, and the Burge function $\Phi(x) = -\sum_{i=1}^d \log(x_i)$ with $\operatorname{dom}\Phi = \mathbb{R}^d_{++}$. For more examples, we refer the reader to the articles [1, 3, 22, 39] and the recent survey [40].

We will often use the observation that the subdifferential of a Legendre function Φ is empty on the boundary of its domain [37, Theorem 26.1]:

$$\partial\Phi(x) = \emptyset \qquad \text{for all } x \notin \operatorname{int}(\operatorname{dom}\Phi).$$

The Legendre function Φ induces the *Bregman divergence*

$$D_\Phi(y, x) := \Phi(y) - \Phi(x) - \langle \nabla\Phi(x), y - x\rangle,$$

for all $x \in \operatorname{int}(\operatorname{dom}\Phi),\ y \in \operatorname{dom}\Phi$. Notice that since Φ is strictly convex, equality $D_\Phi(y, x) = 0$ holds for some $x, y \in \operatorname{int}(\operatorname{dom}\Phi)$ if and only if $y = x$. Analysis of algorithms based on the Bregman divergence typically relies on the following three point inequality; see e.g. [41, Property 1].

Lemma 4 (Three point inequality)**.** *Consider a closed convex function $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ satisfying $\operatorname{ri}(\operatorname{dom} g) \subset \operatorname{int}(\operatorname{dom}\Phi)$. Then for any point $z \in \operatorname{int}(\operatorname{dom}\Phi)$, any minimizer z^+ of the problem*

$$\min_x \ g(x) + D_\Phi(x, z),$$

lies in $\operatorname{int}(\operatorname{dom}\Phi)$*, is unique, and satisfies the inequality:*

$$g(x) + D_\Phi(x, z) \geq g(z_+) + D_\Phi(z_+, z) + D_\Phi(x, z_+) \qquad \forall x \in \operatorname{dom}\Phi.$$

Recall that a function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called ρ-weakly convex if the perturbed function $f + \frac{\rho}{2}\|\cdot\|_2^2$ is convex [34]. By analogy, we will say that f is *ρ-weakly convex relative to* Φ if the perturbed function $f + \rho\Phi$ is convex. This notion is closely related to the relative smoothness condition introduced in [2, 30].

Relative weak convexity, like its classical counterpart, can be caracterized through generalized derivatives. Recall that the *Fréchet subdifferential* of a function f at a point $x \in \operatorname{dom} f$, denoted $\hat{\partial} f(x)$, consists of all vectors $v \in \mathbb{R}^d$ satisfying

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

The *limiting subdifferential* of f at x, denoted $\partial f(x)$, consists of all vectors $v \in \mathbb{R}^d$ such that there exist sequences $x_k \in \mathbb{R}^d$ and $v_k \in \hat{\partial} f(x_k)$ satisfying $(x_k, f(x_k), v_k) \to (x, f(x), v)$.

Lemma 5 (Subdifferential characterization)**.**
The following are equivalent for any locally Lipschitz function $f\colon \mathbb{R}^d \to \mathbb{R}$*.*

1. *The function* f *is* ρ*-weakly convex relative to* Φ*.*

2. *For any* $x \in \operatorname{int}(\operatorname{dom}\Phi), y \in \operatorname{dom}\Phi$ *and any* $v \in \hat{\partial} f(x)$*, the inequality holds:*

$$f(y) \geq f(x) + \langle v, y - x\rangle - \rho D_\Phi(y, x). \tag{4.2.1}$$

3. *For any* $x \in \operatorname{int}(\operatorname{dom}\Phi) \cap \operatorname{dom}(\nabla f)$*, and any* $y \in \operatorname{dom}\Phi$*, the inequality holds:*

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle - \rho D_\Phi(y, x). \tag{4.2.2}$$

If f *and* Φ *are* C^2*-smooth on* $\operatorname{int}(\operatorname{dom}\Phi)$*, then the three properties above are all equivalent to*

$$\nabla^2 f(x) \succeq -\rho\nabla^2\Phi(x) \qquad \forall x \in \operatorname{int}(\operatorname{dom}\Phi). \tag{4.2.3}$$

Proof. Define the perturbed function $g := f + \rho\Phi$. We prove the implications 1 $\Rightarrow$ 2 $\Rightarrow$ 3 $\Rightarrow$ 1 in order. To this end, suppose 1 holds. Since g is convex, the subgradient inequality holds:

$$g(y) \geq g(x) + \langle w, y - x\rangle \qquad \text{for all } x, y \in \mathbb{R}^d, w \in \partial g(x). \tag{4.2.4}$$

Taking into account that Φ is differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$, we deduce $\hat{\partial} g(x) = \hat{\partial} f(x) + \rho\nabla\Phi(x)$ for all $x \in \operatorname{int}(\operatorname{dom}\Phi)$; see e.g. [38, Exercise 8.8]. Rewriting (4.2.4) with this in mind immediately yields 2. The implication 2 $\Rightarrow$ 3 is immediate since $\hat{\partial} f(x) = \{\nabla f(x)\}$, whenever f is differentiable at x.

Suppose 3 holds. Fix an arbitrary point $x \in \operatorname{int}(\operatorname{dom}\Phi) \cap \operatorname{dom}(\nabla f)$. Algebraic manipulation of inequality (4.2.2) yields the equivalent description

$$g(y) \geq g(x) + \langle \nabla f(x) + \rho\nabla\Phi(x), y - x\rangle \qquad \text{for all } y \in \operatorname{dom}\Phi. \tag{4.2.5}$$

It follows that the vector $\nabla f(x) + \rho\nabla\Phi(x)$ lies in the convex subdifferential of g at x. Since f is locally Lipschitz continuous, Rademacher's theorem shows that $\operatorname{dom}(\nabla f)$ has full measure in $\mathbb{R}^d$. In particular, we deduce from (4.2.5) that the convex subdifferential of g is nonempty on a dense subset of $\operatorname{int}(\operatorname{dom} g)$. Taking limits, it quickly follows that the convex subdifferential of g is nonempty at every point $x \in \operatorname{int}(\operatorname{dom} g)$ Using [9, Exercise 3.1.12(a)], we conclude that g is convex on $\operatorname{int}(\operatorname{dom} g)$. Moreover, appealing to the sum rule [38, Exercise 10.10], we deduce that $\partial g(x) = \emptyset$ for all $x \notin \operatorname{int}(\operatorname{dom}\Phi)$, since $\partial\Phi(x) = \emptyset$ for all $x \notin \operatorname{int}(\operatorname{dom}\Phi)$. Therefore ∂g is a globally monotone map globally. Appealing to [38, Theorem 12.17], we conclude that g is a convex function. Thus item 1 holds. This completes the proof of the equivalences 1 $\Leftrightarrow$ 2 $\Leftrightarrow$ 3.

Finally suppose that f and Φ are C^2-smooth on $\operatorname{int}(\operatorname{dom}\Phi)$. Clearly, if f is ρ-weakly convex relative to Φ, then second-order characterization of convexity of the function $g = f + \rho\Phi$ directly implies (4.2.3). Conversely, (4.2.3) immediately implies that g is convex on the interior of its domain. The same argument using [38, Theorem 12.17], as in the implication 3 $\Rightarrow$ 1, shows that g is convex on all of $\mathbb{R}^d$. □

Notice that the setup so far has not relied on any predefined norm. Let us for the moment

make the common assumption that Φ is 1-strongly convex relative to some norm $\|\cdot\|$ on $\mathbb{R}^d$, which implies

$$D_\Phi(y, x) \geq \tfrac{1}{2}\|y - x\|^2. \tag{4.2.6}$$

Then using Lemma 5, we deduce that to check that f is ρ-weakly convex relative to Φ, it suffices to verify the inequality

$$f(y) \geq f(x) + \langle v, y - x\rangle - \frac{\rho}{2}\|y - x\|^2 \qquad \text{for all } x, y \in \operatorname{dom}\Phi, v \in \partial f(x).$$

Recall that a function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called ρ-smooth if it satisfies:

$$\|\nabla f(y) - \nabla f(x)\|_* \leq \rho\|y - x\| \qquad \text{for all } x, y \in \mathbb{R}^d,$$

where $\|\cdot\|_*$ is the dual norm. Thus any ρ-smooth function f is automatically ρ-weakly convex relative to Φ. Our main result will not require Φ to be 1-strongly convex; however, we will impose this assumption in Section 4.6 where we augment our guarantees for the stochastic mirror descent algorithm under a differentiability assumption.

4.3 The problem class and the algorithm

We are now ready to introduce the problem class considered in this paper. We will be interested in the optimization problem

$$\min_x \; F(x) := f(x) + r(x) \tag{4.3.1}$$

where

- $f\colon \mathbb{R}^d \to \mathbb{R}$ is a locally Lipschitz function,
- $r\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a closed function having a convex domain,
- $\Phi\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is some Legendre function satisfying the compatibility conditions:

$$\operatorname{ri}(\operatorname{dom} r) \subseteq \operatorname{int}(\operatorname{dom}\Phi) \qquad \text{and} \qquad \partial(r + \Phi)(x) = \emptyset \text{ for all } x \notin \operatorname{int}(\operatorname{dom}\Phi). \tag{4.3.2}$$

The first two items are standard and mild. The third stipulates that r must be compatible with Φ. In particular, the inclusion $\operatorname{ri}(\operatorname{dom} r) \subseteq \operatorname{int}(\operatorname{dom} \Phi)$ automatically implies (4.3.2), whenever r is convex [37, Theorem 23.8], or more generally whenever a standard qualification condition holds.[3] To simplify notation, henceforth set $U := \operatorname{int}(\operatorname{dom} \Phi)$.

4.3.1 Assumptions and the Algorithm

We now specify the model-based algorithms we will analyze. Fix a probability space $(\Omega, \mathcal{F}, P)$ and equip $\mathbb{R}^d$ with the Borel $\sigma-$algebra. To each point $x \in \operatorname{dom} f$ and each random element $\xi \in \Omega$, we associate a stochastic one-sided model $f_x(\cdot, \xi)$ of the function f. Namely, we assume that there exist $\tau, \rho, \mathsf{L} > 0$ satisfying the following properties.

(A1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1, \ldots, \xi_T \sim P$

(A2) **(One-sided accuracy)** There is a measurable function $(x, y, \xi) \mapsto f_x(y, \xi)$ defined on $U \times U \times \Omega$ satisfying both

$$\mathbb{E}_\xi \left[f_x(x, \xi) \right] = f(x), \qquad \forall x \in U \cap \operatorname{dom} r$$

and

$$\mathbb{E}_\xi \left[f_x(y, \xi) - f(y) \right] \leq \tau D_\Phi(y, x), \qquad \forall x, y \in U \cap \operatorname{dom} r. \tag{4.3.3}$$

(A3) **(Weak convexity of the models)** The functions $f_x(\cdot, \xi) + r(\cdot)$ are ρ-weakly convex relative to Φ for all $x \in U \cap \operatorname{dom} r$, and a.e. $\xi \in \Omega$.

(A4) **(Lipschitzian property)** There exists a square integrable function $L \colon \Omega \to \mathbb{R}_+$ such that for all $x, y \in U \cap \operatorname{dom} r$, the following inequalities hold:

$$\begin{aligned} f_x(x, \xi) - f_x(y, \xi) &\leq L(\xi) \sqrt{D_\Phi(y, x)}, \\ \sqrt{\mathbb{E}_\xi \left[L(\xi)^2 \right]} &\leq \mathsf{L}. \end{aligned} \tag{4.3.4}$$

[3] Qualification condition: $\partial^\infty r(x) \cap -N_{\operatorname{dom}\Phi}(x) = \{0\}$, for all $x \in \operatorname{dom} r \cap \operatorname{dom} \Phi$; see [38, Proposition 8.12, Corollary 10.9].

Some comments are in order. Assumption (A1) is standard and is necessary for all sampling based algorithms. Assumption (A2) specifies the accuracy of the models. That is, we require the model in expectation to agree with f at the basepoint, and to globally lower-bound f up to an error controlled by the Bregman divergence. Assumption (A3) is very mild, since in most practical circumstances the function $f_x(\cdot, \xi) + r(\cdot)$ is convex, i.e. $\rho = 0$. The final Assumption (A4) controls the order of growth of the individual models $f_x(y, x)$ as the argument y moves away from x.

Notice that the assumptions (A1)-(A4) do not involve any norm on $\mathbb{R}^d$. However, when Φ is 1-strongly convex relative to some norm, the properties (4.3.3) and (4.3.4) are implied by standard assumptions. Namely (4.3.3) holds if the error in the model approximation satisfies

$$\mathbb{E}_\xi \left[f_x(y, \xi) - f(y) \right] \leq \frac{\tau}{2} \|y - x\|^2, \qquad \forall x, y \in U.$$

Similarly (4.3.4) will hold as long as for every $x \in U \cap \operatorname{dom} r$ and a.e. $\xi \in \Omega$ the models $f_x(\cdot, \xi)$ are $L(\xi)$-Lipschitz continuous on U in the norm $\| \cdot \|$. The use of the Bregman divergence allows for much greater flexibility as it can, for example, model higher order growth of the functions in question. To illustrate, let us look at the following example where the Lipschitz constant $L(\xi)$ of the models $f_x(\cdot, \xi)$ is bounded by a polynomial.

Example 4.3.1 (Bregman divergence under polynomial growth)**.** Consider a degree n univariate polynomial

$$p(u) = \sum_{i=0}^{n} a_i u^i,$$

with coefficients $a_i \geq 0$. Suppose now that the one-sided Lipschitz constants of the models satisfy the growth property:

$$\frac{f_x(x, \xi) - f_x(y, \xi)}{\|x - y\|_2} \leq L(\xi) \sqrt{\frac{p(\|x\|_2) + p(\|y\|_2)}{2}} \qquad \text{for all distinct } x, y \in \mathbb{R}^d.$$

Motivated by [29, Proposition 5.1], the following proposition constructs a Bregman divergence that is well-adapted to the polynomial $p(\cdot)$. We defer its proof to Appendix 4.7.1. In particular, with the choice of the Legendre function Φ in (4.3.5), the required estimate (4.3.4) holds.

Proposition 1. *Define the convex function*

$$\Phi(x) = \sum_{i=0}^{n} a_i \left(\frac{3i+7}{i+2}\right) \|x\|_2^{i+2}. \tag{4.3.5}$$

Then for all $x, y \in \mathbb{R}^d$, we have

$$D_\Phi(y, x) \geq \frac{p(\|x\|_2) + p(\|y\|_2)}{2} \cdot \|x - y\|_2^2,$$

and therefore the estimate (4.3.4) holds.

The final ingredient we need before stating the algorithm is an estimate on the weak convexity constant of F. The following simple lemma shows that Assumptions (A2) and (A3) imply that F itself is $(\tau + \rho)$-weakly convex relative to Φ.

Lemma 6. *The function F is $(\tau + \rho)$-weakly convex relative to Φ.*

Proof. We first show that the function $g := F + (\rho+\tau)\Phi$ is convex on $\operatorname{ri}(\operatorname{dom} F)$. To this end, fix arbitrary points $x, y \in \operatorname{ri}(\operatorname{dom} g)$, and note the equality $\operatorname{ri}(\operatorname{dom} g) = U \cap \operatorname{ri}(\operatorname{dom} r)$ [37, Theorem 6.5]. Choose $\lambda \in (0, 1)$ and set $\bar{x} = \lambda x + (1 - \lambda)y$. Taking into account (A3), we deduce

$$\begin{aligned}
g(\bar{x}) &= f(\bar{x}) + r(\bar{x}) + (\rho + \tau)\Phi(\bar{x}) \\
&= \mathbb{E}_\xi[f_{\bar{x}}(\bar{x}, \xi) + r(\bar{x}) + \rho\Phi(\bar{x})] + \tau\Phi(\bar{x}) \\
&\leq \mathbb{E}_\xi[\lambda(f_{\bar{x}}(x, \xi) + r(x) + \rho\Phi(x)) + (1 - \lambda)(f_{\bar{x}}(y, \xi) + r(y) + \rho\Phi(y))] + \tau\Phi(\bar{x}) \\
&= \lambda\mathbb{E}_\xi[f_{\bar{x}}(x, \xi) + r(x)] + (1 - \lambda)\mathbb{E}_\xi[f_{\bar{x}}(y, \xi) + r(y)] + \tau\Phi(\bar{x}) + \lambda\rho\Phi(x) + (1 - \lambda)\rho\Phi(y) \\
&= \lambda\mathbb{E}_\xi[f_{\bar{x}}(x, \xi) + r(x) - \tau D_\Phi(x, \bar{x})] + (1 - \lambda)\mathbb{E}_\xi[f_{\bar{x}}(y, \xi) + r(y) - \tau D_\Phi(y, \bar{x})] \\
&\quad + \lambda\tau(\Phi(\bar{x}) + D_\Phi(x, \bar{x})) + (1 - \lambda)\tau(\Phi(\bar{x}) + D_\Phi(y, \bar{x})) + \lambda\rho\Phi(x) + (1 - \lambda)\rho\Phi(y).
\end{aligned} \tag{4.3.6}$$

Now observe

$$\Phi(\bar{x}) + D_\Phi(x, \bar{x}) = \Phi(x) - (1 - \lambda)\langle \nabla\Phi(\bar{x}), x - y\rangle,$$

and similarly

$$\Phi(\bar{x}) + D_\Phi(y, \bar{x}) = \Phi(y) - \lambda\langle \nabla\Phi(\bar{x}), y - x\rangle.$$

Hence algebraic manipulation of the two equalities above yields the expression

$$\lambda\tau(\Phi(\bar{x}) + D_\Phi(x, \bar{x})) + (1-\lambda)\tau(\Phi(\bar{x}) + D_\Phi(y, \bar{x})) = \lambda\tau\Phi(x) + (1-\lambda)\tau\Phi(y).$$

Continuing with (4.3.6), we obtain

$$\begin{aligned} g(\bar{x}) &\leq \lambda f(x) + r(x) + (1-\lambda)(f(y) + r(y)) \\ &\qquad + \lambda\tau\Phi(x) + (1-\lambda)\tau\Phi(y) + \lambda\rho\Phi(x) + (1-\lambda)\rho\Phi(y) \\ &= \lambda[f(x) + r(x) + (\tau+\rho)\Phi(x)] + (1-\lambda)[f(y) + r(y) + (\tau+\rho)\Phi(y)] \\ &\leq \lambda g(x) + (1-\lambda)g(y). \end{aligned}$$

We have thus verified that g is convex on $\operatorname{ri}(\operatorname{dom} g)$. Appealing to (4.3.2) and the sum rule [38, Exercise 10.10], we deduce that the subdifferential $\partial g(x)$ is empty at every point in $x \notin \operatorname{ri}(\operatorname{dom} g)$, and therefore ∂g is a globally monotone map. Using [38, Theorem 12.17], we conclude that g is a convex function, as needed. □

In light of Lemma 6, we also make the following additional assumption on the solvability of the Bregman proximal subproblems.

(A5) **(Solvability)** The convex problems

$$\min_y \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\} \qquad \text{and} \qquad \min_y \left\{ f_x(y, \xi) + r(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

admit a minimizer for any $\lambda < (\tau+\rho)^{-1}$, any $x \in U$, and a.e. $\xi \in \Omega$.[4] The minimizers vary measurably in $(x, \xi) \in U \times \Omega$.

Assumption (A5) is very mild. In particular, it holds automatically if (i) Φ is strongly convex with respect to some norm, or if (ii) the functions $f_x(\cdot, \xi) + r(\cdot) + \rho D_\Phi(\cdot, x)$ and $F + (\tau+\rho)\Phi$ are bounded from below and Φ has bounded sublevel sets [40, Lemma 2.3].

We are now ready to state the stochastic model-based algorithm we analyze—Algorithm 6.

[4] Note the minimizers are automatically unique by Lemma 4

Algorithm 6: Stochastic Model Based Minimization

Data: $x_0 \in U \cap \operatorname{dom} r$, real $\lambda < (\tau+\rho)^{-1}$, a nonincreasing sequence $\{\eta_t\}_{t\geq 0} \subseteq (0,\lambda)$, and iteration count T.

Step $t = 0, \ldots, T$:

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \left\{ f_{x_t}(x,\xi_t) + r(x) + \frac{1}{\eta_t} D_\Phi(x, x_t) \right\} \end{array} \right\},$$

Sample $t^* \in \{0, \ldots, T\}$ according to the discrete probability distribution

$$\mathbb{P}(t^* = t) \propto \frac{\eta_t}{1 - \eta_t \rho}.$$

Return x_{t^*}

4.3.2 Examples

Before delving into the convergence analysis of Algorithm 6, in this section we illustrate the algorithmic framework on four examples. In all cases, assumptions (A1) and (A5) are self-explanatory. Therefore, we only focus on verifying (A2)-(A4). For simplicity, we also assume that $r(\cdot)$ is convex in all examples.

Stochastic Bregman-proximal point. Suppose that the models $(x, y, \xi) \mapsto f_x(y, \xi)$ satisfy

$$\mathbb{E}_\xi[f_x(y,\xi)] = f(y) \qquad \forall x, y \in U \cap \operatorname{dom} r.$$

With this choice of the models, Algorithm 6 becomes the stochastic Bregman-proximal point method. Analysis of the deterministic version of the method for convex problems goes back to [13, 14, 22]. Observe that Assumption (A2) holds trivially. Assumption (A3) and Assumption (A4) should be verified in particular circumstances, depending on how the models are generated. In particular, one can verify Assumption (A4) under polynomial growth of the Lipschitz constant, by appealing to Example 4.3.1.

Stochastic mirror descent. Suppose that the models $(x, y, \xi) \mapsto f_x(y, \xi)$ are given by

$$f_x(y, \xi) = f(x) + \langle G(x, \xi), y - x \rangle,$$

for some measurable mapping $G\colon U \times \Omega \to \mathbb{R}^d$ satisfying $\mathbb{E}_\xi[G(x, \xi)] \in \partial f(x)$ for all $x \in U \cap \operatorname{dom} r$. Algorithm 6 then becomes the stochastic mirror descent algorithm, classically studied in [6,31] in the convex setting and more recently analyzed in [2,29,30] under convexity and relative continuity assumptions. Assumption (A2) simply says that f is τ-weakly convex relative to Φ, while Assumption (A3) holds trivially with $\rho = 0$. Assumption (A4) is directly implied by the relative continuity condition of Lu [29]. Namely it suffices to assume that there is a square integrable function $L\colon \Omega \to \mathbb{R}_{++}$ satisfying

$$\|G(x, \xi)\|_* \leq L(\xi) \frac{\sqrt{D(y, x)}}{\|y - x\|} \qquad \forall x, y \in U,$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^d$, and $\|\cdot\|_*$ is the dual norm. We refer to [29] for more details on this condition and examples.

Gauss-Newton method with Bregman regularization. In the next example, suppose that f has the composite form

$$f(x) = \mathbb{E}_\xi[h(c(x, \xi), \xi)],$$

for some measurable function $h(y, \xi)$ that is convex in y for a.e. $\xi \in \Omega$ and a measurable map $c(x, \xi)$ that is C^1-smooth in x for a.e. $\xi \in \Omega$. We may then use the convex models

$$f_x(y, \xi) := h\left(c(x, \xi) + \nabla c(x, \xi)(y - x), \xi\right), \xi\right),$$

which automatically satisfy (A3) with $\rho = 0$. Algorithm 6 then becomes a stochastic Gauss-Newton method with Bregman regularization.

In the Euclidean case $\Phi = \frac{1}{2}\|\cdot\|^2$, the method reduces to the stochastic prox-linear algorithm, introduced in [20] and further analyzed in [15]. The deterministic prox-linear method has classical roots, going back at least to [11,23,36], while a more modern complexity

theoretic perspective appears in [12, 18, 19, 28, 32]. Even in the deterministic setting, to make progress, one typically assumes that h and ∇c are globally Lipschitz. More generally and in line with our current work, one may introduce a different Legendre function Φ. For example, in the case of polynomial growth, the following propositions construct Legendre functions that are compatible with Assumptions (A2) and (A4). We defer their proofs to Appendix 4.7.3. In the two propositions, we assume that the outer functions $h(\cdot, \xi)$ are globally Lipschitz, while the inner maps $c(\cdot, \xi)$ may have a high order of growth. It is possible to also analyze the setting when $h(\cdot, \xi)$ has polynomial growth, but the resulting statements and assumptions become much more cumbersome; we therefore omit that discussion.

Proposition 2 (Satisfying (A2))**.** *Suppose there are square integrable functions $L_1, L_2 \colon \Omega \to \mathbb{R}_+$ and a univariate polynomial $p(u) = \sum_{i=0}^n a_i u^i$ with nonnegative coefficients satisfying*

$$\begin{aligned} \frac{|h(v,\xi) - h(w,\xi)|}{\|v - w\|_2} &\le L_1(\xi) \qquad \forall v \neq w, \\ \frac{\|\nabla c(x,\xi) - \nabla c(y,\xi)\|_{\mathrm{op}}}{\|x - y\|_2} &\le L_2(\xi)(p(\|x\|_2) + p(\|y\|_2)) \qquad \forall x \neq y. \end{aligned}$$

Define the Legendre function $\Phi(x) := \sum_{i=0}^n \frac{a_i(3i+7)}{i+2}\|x\|_2^{i+2}$. Then assumption (A2) holds with $\tau := \frac{4}{3}\mathbb{E}\left[L_1(\xi)L_2(\xi)\right]$.

Proposition 3 (Satisfying (A4))**.** *Suppose there are square integrable functions $L_1, L_2 \colon \Omega \to \mathbb{R}_+$ and a univariate polynomial $q(u) = \sum_{i=0}^n b_i u^i$ with nonnegative coefficients satisfying*

$$\begin{aligned} \frac{|h(v,\xi) - h(w,\xi)|}{\|v - w\|_2} &\le L_1(\xi) \qquad \forall v \neq w, \\ \|\nabla c(x,\xi)\|_{\mathrm{op}} &\le L_2(\xi) \cdot \sqrt{q(\|x\|_2)} \qquad \forall x, \xi. \end{aligned}$$

Then with the Legendre function $\Phi(x) = \sum_{i=0}^n \frac{b_i}{i+2}\|x\|_2^{i+2}$, assumption (A4) holds with $L(\xi) = \sqrt{2}L_1(\xi)L_2(\xi)$.

To construct a Bregman function compatible with both (A2) and (A4) simultaneously, one may simply add the two Legendre functions constructed in Propositons 2 and 3.

Stochastic saddle point problems. As the final example, suppose that f is given in the stochastic conjugate form

$$f(x) = \mathbb{E}\left[\sup_{w \in W} g(x, w, \xi)\right],$$

where W is some auxiliary set and $g\colon \mathbb{R}^d \times W \times \Omega \to \mathbb{R}$ is some function. Thus we are interested in solving the stochastic saddle-point problem

$$\inf_x \ \mathbb{E}\left[\sup_{w \in W} g(x, w, \xi)\right] + r(x). \tag{4.3.7}$$

Such problems appear often in data science, where the variation of w in the "uncertainty set" W makes the loss function robust. One popular example is adversarial training [25]. In this setting, we have $g(x, w, \xi) = \mathcal{L}(x + w, y, \xi)$, where $\mathcal{L}(\cdot, \cdot)$ is a loss function, y encodes the observed data, and w varies over some uncertainty set W, such as an ℓ_p-ball.

In order to apply our algorithmic framework, we must have access to stochastic one-sided models $f_x(\cdot, \xi)$ of f. It is quite natural to construct such models by using one-sided stochastic models $g_x(\cdot, w, \xi)$ of g. Indeed, it is appealing to simply set

$$f_x(y, \xi) = g_x(y, \widehat{w}(x, \xi), \xi) \qquad \text{for any} \qquad \widehat{w}(x, \xi) \in \operatorname*{argmax}_w \ g_x(x, w, \xi). \tag{4.3.8}$$

All of the model types in the previous examples could now serve as the models $g_x(\cdot, w, \xi)$, provided they meet the conditions outlined below.

Formally, to ensure that (A1)-(A5) hold for the models $f_x(y, \xi)$, we must make the following assumptions:

1. The mapping $(x, \xi) \to \sup_{w \in W} g(x, w, \xi)$ is measurable and has finite first moment for every fixed $x \in U \cap \operatorname{dom} r$.

2. The function $g_x(\cdot, w, \xi)$ is ρ-weakly convex relative to Φ, for every fixed $x \in U \cap \operatorname{dom} r$, $w \in W$, and a.e. $\xi \in \Omega$.

3. There exists a mapping $\widehat{w}\colon U \times \Omega \to \mathbb{R}^m$ satisfying

$$\widehat{w}(x, \xi) \in \operatorname*{argmax}_w g_x(x, w, \xi),$$

for all $x \in U \cap \operatorname{dom} r$ and a.e. $\xi \in \Omega$ with the property that the functions $(x, y, \xi) \mapsto g_x(y, \widehat{w}(x, \xi), \xi)$ and $(x, y, \xi) \mapsto g(y, \widehat{w}(x, \xi), \xi)$ are measurable.

4. For all $x, y \in U \cap \operatorname{dom} r$, we have

$$\mathbb{E}_\xi \left[g_x(x, \widehat{w}(x, \xi), \xi) \right] = \mathbb{E}_\xi \left[g(x, \widehat{w}(x, \xi), \xi) \right]$$

and

$$\mathbb{E} \left[g_x(y, \widehat{w}(x, \xi), \xi) - g(y, \widehat{w}(x, \xi), \xi) \right] \leq \tau D_\Phi(y, x).$$

5. There exists a square integrable function $L \colon \Omega \to \mathbb{R}_+$ such that

$$g_x(x, \widehat{w}(x, \xi), \xi) - g_x(y, \widehat{w}(x, \xi), \xi) \leq L(\xi) \sqrt{D_\Phi(y, x)}, \qquad \text{for all } x, y \in U \cap \operatorname{dom} r.$$

Given these assumptions, let us define $f_x(y, \xi)$ as in (4.3.8) We now verify properties (A2)-(A4). Property (A2) follows from Property 4, which implies that $\mathbb{E}\left[f_x(x, \xi)\right] = f(x)$ and

$$\begin{aligned} \mathbb{E}_\xi \left[f_x(y, \xi) - f(y) \right] &= \mathbb{E}_\xi \left[g_x(y, \widehat{w}(x, \xi), \xi) - \sup_{w \in W} g(y, w, \xi) \right] \\ &\leq \mathbb{E}_\xi \left[g_x(y, \widehat{w}(x, \xi), \xi) - g(y, \widehat{w}(x, \xi), \xi) \right] \\ &\leq \tau D_\Phi(y, x). \end{aligned}$$

Property (A3) follows directly from Property 2. Finally, (A4) follows from Property 5.

4.4 Stationarity measure

In this section, we introduce a natural stationarity measure that we will use to describe the convergence rate of Algorithm 6. The stationarity measure is simply the size of the gradient of an appropriate smooth approximation of the problem (4.3.1). This idea is completely analogous to the Euclidean setting [15, 16]. Setting the stage, for any $\lambda > 0$, define the Φ-*envelope*

$$F_\lambda^\Phi(x) := \inf_y \left\{ f(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

and the associated Φ-*proximal map*

$$\operatorname{prox}_{\lambda f}^{\Phi}(x) := \operatorname*{argmin}_{y} \left\{ F(y) + \frac{1}{\lambda} D_{\Phi}(y, x) \right\}.$$

Note that in the Euclidean setting $\Phi = \frac{1}{2}\|\cdot\|^2$, these two constructions reduce to the standard Moreau envelope and the proximity map; see for example the monographs [35, 38] or the note [17] for recent perspectives.

We will measure the convergence guarantees of Algorithm 6 based on the rate at which the quantity

$$\mathbb{E}[D_{\Phi}\left(\operatorname{prox}_{\lambda F}^{\Phi}(x_{t^*}), x_{t^*}\right)] \tag{4.4.1}$$

tends to zero for some fixed $\lambda > 0$. The significance of this quantity becomes apparent after making slightly stronger assumptions on the Legendre function Φ. In this section only, suppose that $\Phi\colon \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is 1-strongly convex with respect to some norm $\|\cdot\|$ and that Φ is twice differentiable at every point in $\operatorname{int}(\operatorname{dom}\Phi)$. With these assumptions, the following result shows that the Φ-envelope is differentiable, with a meaningful gradient. Indeed, this result follows quickly from [4]. For the sake of completeness, we present a self-contained argument in Appendix 4.7.4.

Theorem 4.4.1 (Smoothness of the Φ-envelope)**.** *For any positive $\lambda < (\tau+\rho)^{-1}$, the envelope F_{λ}^{Φ} is differentiable at any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$ with gradient given by*

$$\nabla F_{\lambda}^{\Phi}(x) := \frac{1}{\lambda}\nabla^2\Phi(x)\left(x - \operatorname{prox}_{\lambda F}^{\Phi}(x)\right).$$

In light of Theorem 4.4.1, for any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$, we may define the local norm

$$\|y\|_x := \left\|\nabla^2\Phi(x)y\right\|_*.$$

Then a quick computation shows that the dual norm is given by

$$\|v\|_x^* = \left\|\nabla^2\Phi(x)^{-1}v\right\|.$$

Therefore appealing to Theorem 4.4.1, for any positive $\lambda < (\tau+\rho)^{-1}$ and $x \in \operatorname{int}(\operatorname{dom}\Phi)$ we obtain the estimate

$$\sqrt{D_{\Phi}\left(\operatorname{prox}_{\lambda F}^{\Phi}(x), x\right)} \geq \frac{\lambda}{\sqrt{2}}\|\nabla F_{\lambda}^{\Phi}(x)\|_x^*.$$

Thus the square root of the Bregman divergence, which we will show tends to zero along the iterate sequence at a controlled rate, bounds the local norm of the gradient ∇F_λ^Φ.

4.5 Convergence analysis

We now present convergence analysis of Algorithm 6 under Assumptions (A1)-(A5). Henceforth, let $\{x_t\}_{t\geq 0}$ be the iterates generated by Algorithm 6 and let $\{\xi_t\}_{t\geq 0}$ be the corresponding samples used. For each index $t \geq 0$, define the Bregman-proximal point

$$\hat{x}_t = \operatorname{prox}_{\lambda F}^{\Phi}(x_t).$$

To simplify notation, we will use the symbol $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$. The entire argument of Theorem 4.5.1—our main result—relies on the following lemma.

Lemma 7. *For each iteration $t \geq 0$, the iterates of Algorithm 6 satisfy*

$$\mathbb{E}_t\left[D_\Phi(\hat{x}_t, x_{t+1})\right] \leq \tfrac{1+\eta_t\tau-\eta_t/\lambda}{1-\eta_t\rho} D_\Phi(\hat{x}_t, x_t) + \tfrac{(\mathsf{L}\eta_t)^2}{4(1-\eta_t\rho)} + \tfrac{\eta_t}{1-\eta_t\rho}\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right].$$

Proof. Taking into account assumption (A3), we may apply the three point inequality in Lemma 4 with the convex function $g = f_{x_t}(\cdot, \xi_t) + r(\cdot) + \rho D_\Phi(\cdot, x_t)$ and with $(\frac{1}{\eta_t} - \rho)D_\Phi(\cdot, x_t)$ replacing the Bregman divergence. Thus for any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$, we obtain the estimate

$$f_{x_t}(x,\xi_t)+r(x)+\frac{1}{\eta_t}D_\Phi(x,x_t) \geq f_{x_t}(x_{t+1},\xi_t)+r(x_{t+1})+\frac{1}{\eta_t}D_\Phi(x_{t+1},x_t)+\left(\frac{1}{\eta_t}-\rho\right)D_\Phi(x,x_{t+1}). \tag{4.5.1}$$

Setting $x = \hat{x}_t$, rearranging terms, and taking expectations, we deduce

$$\begin{aligned}\mathbb{E}_\xi[f_{x_t}(\hat{x}_t,\xi_t)+r(\hat{x}_t) - f_{x_t}(x_{t+1},\xi_t) - r(x_{t+1})] \\ \geq \frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1}) - D_\Phi(\hat{x}_t,x_t) + D_\Phi(x_{t+1},x_t)\right].\end{aligned} \tag{4.5.2}$$

We seek to upper bound the left-hand-side of (4.5.2). Using assumptions (A2) and (A4), we

obtain:

$$\begin{aligned}
&\mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t)-f_{x_t}(x_{t+1},\xi_t)\right]\\
&\quad\le \mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t)-f_{x_t}(x_t,\xi_t)+L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right]\\
&\quad= \mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t)-f(\hat{x}_t)\right]+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right]-f(x_t)+f(\hat{x}_t)\\
&\quad\le \tau D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right]-f(x_t)+f(\hat{x}_t).
\end{aligned}\tag{4.5.3}$$

By the definition of $\hat{x}_t$ as the Bregman-proximal point, we have

$$f(\hat{x}_t)+r(\hat{x}_t)+\frac{1}{\lambda}D_\Phi(\hat{x}_t,x_t)\le f(x_t)+r(x_t).\tag{4.5.4}$$

The right hand side of (4.5.2) is thus upper bounded by

$$\begin{aligned}
&\tau D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}-f(x_t)-r(x_{t+1})\right]+f(\hat{x}_t)+r(\hat{x}_t)\\
&\le \tau D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}+(r(x_t)-r(x_{t+1}))\right]+f(\hat{x}_t)+r(\hat{x}_t)-f(x_t)-r(x_t)\\
&\le \left(\tau-\frac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}+(r(x_t)-r(x_{t+1}))\right]
\end{aligned}$$

where the last inequality follows from (4.5.4). Combining this estimate with (4.5.2), we obtain

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})-D_\Phi(\hat{x}_t,x_t)+D_\Phi(x_{t+1},x_t)\right]\\
&\qquad\le\left(\tau-\frac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}+(r(x_t)-r(x_{t+1}))\right].
\end{aligned}$$

Multiplying through by η_t and rearranging yields

$$\begin{aligned}
&(1-\eta_t\rho)\mathbb{E}_t\left[D_\Phi(\hat{x}_t,x_{t+1})\right]\\
&\quad\le\left(1+\eta_t\tau-\frac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\mathbb{E}_t\left[\eta_t L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}-D_\Phi(x_{t+1},x_t)\right]\\
&\qquad+\eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].
\end{aligned}\tag{4.5.5}$$

Now define $\gamma:=\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}$. Note that Cauchy-Schwarz implies

$$\mathbb{E}_t\left[\eta_t L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right]\le\eta_t\mathsf{L}\gamma.$$

Using this estimate in (4.5.5), we obtain

$$(1-\eta_t\rho)\mathbb{E}_t\left[D_\Phi(\hat{x}_t,x_{t+1})\right] \leq \left(1+\eta_t\tau-\frac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\eta_t\mathsf{L}\gamma-\gamma^2 + \eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].$$

Maximizing the right hand side in γ (i.e. taking $\gamma=\frac{\mathsf{L}\eta_t}{2}$), yields the guarantee

$$(1-\eta_t\rho)\mathbb{E}_t\left[D_\Phi(\hat{x}_t,x_{t+1})\right] \leq \left(1+\eta_t\tau-\frac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\mathsf{L}\eta_t)^2}{4}+\eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].$$

Dividing through by $1-\eta_t\rho$ completes the proof. □

We can now prove our main theorem.

Theorem 4.5.1 (Convergence rate)**.** *The point x_{t^*} returned by Algorithm 6 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}^\Phi_{\lambda F}(x_{t^*}),x_{t^*}\right)\right] \leq \frac{\lambda^2}{1-\lambda(\tau+\rho)}\left(\frac{F^\Phi_\lambda(x_0)-\min F}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}+\frac{\mathsf{L}^2\sum_{t=0}^T\frac{\eta_t^2}{4\lambda(1-\eta_t\rho)}}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}+\frac{\frac{\eta_0}{\lambda(1-\eta_0\rho)}(r(x_0)-\inf r)}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}\right).$$

Proof. Using the definitions of x_{t+1} and $\hat{x}_t$ along with Lemma 7, we obtain

$$\begin{aligned}
\mathbb{E}_t\left[F^\Phi_\lambda(x_{t+1})\right] &\leq \mathbb{E}_t\left[F(\hat{x}_t)+\frac{1}{\lambda}D_\Phi(\hat{x}_t,x_{t+1})\right]\\
&\leq \mathbb{E}_t\left[F(\hat{x}_t)+\frac{1}{\lambda(1-\eta_t\rho)}\left(\left(1+\eta_t\left(\tau-\frac{1}{\lambda}\right)\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\mathsf{L}\eta_t)^2}{4}\right)\right]\\
&\quad+\frac{\eta_t}{\lambda(1-\eta_t\rho)}\mathbb{E}_t\left[(r(x_t)-r(x_{t+1}))\right]\\
&= F^\Phi_\lambda(x_t)+\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\mathsf{L}\eta_t)^2}{4\lambda(1-\eta_t\rho)}\\
&\quad+\frac{\eta_t}{\lambda(1-\eta_t\rho)}\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right].
\end{aligned}$$

Recursing and applying the tower rule for expectations, we obtain

$$\begin{aligned}
\mathbb{E}\left[F^\Phi_\lambda(x_{T+1})\right] \leq F^\Phi_\lambda(x_0)&+\sum_{t=0}^T\left(\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)\mathbb{E}[D_\Phi(\hat{x}_t,x_t)]+\frac{(\mathsf{L}\eta_t)^2}{4\lambda(1-\eta_t\rho)}\right)\\
&+\sum_{t=0}^T\frac{\eta_t}{\lambda(1-\eta_t\rho)}\mathbb{E}\left[r(x_t)-r(x_{t+1})\right].
\end{aligned} \tag{4.5.6}$$

Taking into account that η_t is nonincreasing yields the inequality

$$\sum_{t=0}^{T} \tfrac{\eta_t}{\lambda(1-\eta_t\rho)}(r(x_t) - r(x_{t+1})) \leq \tfrac{\eta_0}{\lambda(1-\eta_0\rho)}(r(x_0) - \inf r).$$

See the auxiliary Lemma 10 for a verification. Combining this bound with (4.5.6), using the inequality $\mathbb{E}\left[F_\lambda(x_{T+1})\right] \geq \min F$, and rearranging, we conclude

$$\frac{1}{\lambda}\left(\frac{1}{\lambda} - \tau - \rho\right)\sum_{t=0}^{T} \frac{\eta_t}{1-\eta_t\rho}\mathbb{E}[D_\Phi(\hat{x}_t, x_t)] \leq F_\lambda^\Phi(x_0) - \min F + \mathsf{L}^2 \sum_{t=0}^{T} \frac{\eta_t^2}{4\lambda(1-\eta_t\rho)} + \frac{\eta_0}{\lambda(1-\eta_0\rho)}(r(x_0) - \inf r),$$

or equivalently

$$\sum_{t=0}^{T} \frac{\eta_t}{1-\eta_t\rho}\mathbb{E}\left[D_\Phi(\hat{x}_t, x_t)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0) - \min F)}{1-\lambda(\tau+\rho)} + \frac{\lambda^2\mathsf{L}^2}{1-\lambda(\tau+\rho)} \sum_{t=0}^{T} \frac{\eta_t^2}{4\lambda(1-\eta_t\rho)} + \frac{\lambda^2\eta_0}{\lambda(1-\lambda(\tau+\rho))(1-\eta_0\rho)}(r(x_0) - \inf r).$$

Dividing through by $\sum_{t=0}^{T} \frac{\eta_t}{1-\eta_t\rho}$ and recognizing the left-hand-side as $\mathbb{E}[D_\Phi(\hat{x}_{t^*}, x_{t^*})]$, the result follows. □

As an immediate corollary of Theorem 4.5.1, we have the following rate of convergence when the stepsize η_t is constant.

Corollary 6 (Convergence rate for constant stepsize)**.** *For some $\alpha > 0$, set $\eta_t = \frac{1}{\lambda^{-1}+\alpha^{-1}\sqrt{T+1}}$ for all indices $t = 1, \ldots, T$. Then the point x_{t^*} returned by Algorithm 7 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}_{\lambda F}^\Phi(x_{t^*}), x_{t^*}\right)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0) - \min F) + \frac{\lambda\mathsf{L}^2\alpha^2}{4} + \frac{\lambda((r(x_0)-\inf r))}{\lambda^{-1}-\rho+\alpha^{-1}}}{1-\lambda(\tau+\rho)} \cdot \left(\frac{\lambda^{-1}-\rho}{T+1} + \frac{1}{\alpha\sqrt{T+1}}\right).$$

4.6 Mirror descent: smoothness and finite variance

Assumptions (A1)-(A5) are reasonable for the examples described in Section 4.3.2, being in line with standard conditions in the literature. However, in the special case that f is smooth and we apply stochastic mirror descent, Assumption (A4) is nonstandard. Ideally, one would

like to replace this assumption with a bound on the variance of the stochastic estimator of the gradient. In this section, we show that this is indeed possible by slightly modifying the argument in Section 4.5.

Henceforth, let Φ be a Legendre function and set $U := \operatorname{int}(\operatorname{dom}\Phi)$. In this section, we make the following assumptions:

(B1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1, \ldots, \xi_T \sim P$

(B2) **(Stochastic gradient)** There is a measurable mapping $G : U \times \Omega \to \mathbb{R}^d$ satisfying

$$\mathbb{E}_\xi\left[G(x,\xi)\right] = \nabla f(x), \qquad \forall x \in U \cap \operatorname{dom} r.$$

(B3) **(Relative Smoothness)** There exist real $\tau, M \geq 0$, such that

$$-\tau D_\Phi(y,x) \leq f(y) - f(x) - \langle \nabla f(x)\rangle y - x \leq M D_\Phi(y,x) \qquad \forall x, y \in U \cap \operatorname{dom} r.$$

(B4) **(Relative convexity)** The function r is ρ-weakly convex relative to Φ.

(B5) **(Strong convexity of Φ)** The Legendre function Φ is 1-strongly convex with respect to some norm $\|\cdot\|$.

(B6) **(Finite variance)** The following variance is finite:

$$\mathbb{E}_\xi\left[\|G(x,\xi) - \nabla f(x)\|_*^2\right] \leq \frac{\sigma^2}{2} < \infty.$$

Henceforth, we denote by $f_x(\cdot,\xi)$ the linear models

$$f_x(y,\xi) := f(x) + \langle G(x,\xi)\rangle y - x,$$

which are built from the stochastic gradient estimator G. With this notation in hand, let us compare Assumptions (B1)-(B5) with Assumptions (A1)-(A4). Evidently, Assumptions (B1) and (A1) are identical. Upon taking expectations, Assumptions (B2) and (B3) imply the

stochastic one-sided accuracy property (A2) for the linear models $f_x(\cdot, \xi)$, while (B4) directly implies (A3). Assumptions (B5) and (B6) replace the Lipschitzian property (A4).

Finally, we reiterate that the relative smoothness property in (B3) was recently introduced in [2, 30] for smooth convex minimization, and extended to smooth nonconvex problems in [8] and to nonsmooth stochastic problems in [26, 29]. This property allows for higher order growth than the standard Lipschitz gradient assumptions, commonly analyzed in the literature. We refer the reader to [2, 30] for various examples of Bregman functions that arise in applications.

For the sake of clarity, Algorithm 7 instantiates Algorithm 6 in our setting.

Algorithm 7: Mirror descent for smooth minimization

Data: $x_0 \in U \cap \operatorname{dom} r$, positive $\lambda < (\tau + \rho)^{-1}$, a sequence $\{\eta_t\}_{t\geq 0} \subseteq \left(0, \frac{\lambda}{1+\lambda M}\right)$, and iteration count T

Step $t = 0, \ldots, T$:

$$\left\{ \begin{array}{l} \text{Sample } \xi_t \sim P \\ \text{Set } x_{t+1} = \operatorname*{argmin}_x \left\{ \langle G(x_t, \xi_t) \rangle x + r(x) + \frac{1}{\eta_t} D_\Phi(x, x_t) \right\} \end{array} \right\},$$

Sample $t^* \in \{0, \ldots, T\}$ according to the discrete probability distribution

$$\mathbb{P}(t^* = t) \propto \frac{\eta_t}{1 - \eta_t \rho}.$$

Return x_{t^*}

As in Section 4.5, the convergence analysis relies on the following key lemma. We let $\{x_t\}_{t\geq 0}$ be the iterates generated by Algorithm 7 and let $\{\xi_t\}_{t\geq 0}$ be the corresponding samples used. For each index $t \geq 0$, we continue to use the notation $\hat{x}_t = \operatorname{prox}^\Phi_{\lambda F}(x)$ and let $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$.

Lemma 8. *For each iteration $t \geq 0$, the iterates of Algorithm 7 satisfy*

$$\mathbb{E}_t \left[D_\Phi(\hat{x}_t, x_{t+1}) \right] \leq \frac{1 + \eta_t \tau - \eta_t/\lambda}{(1 - \eta_t \rho)} \cdot D_\Phi(\hat{x}_t, x_t) + \frac{1}{4} \cdot \frac{(\sigma \eta_t)^2}{(1 - \eta_t(M + \frac{1}{\lambda}))(1 - \eta_t \rho)}.$$

Proof. Following the initial steps of the proof of Lemma 7, we arrive at the estimate (4.5.2),

namely

$$\frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})-D_\Phi(\hat{x}_t,x_t)+D_\Phi(x_{t+1},x_t)\right]$$
$$\leq \mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t)+r(\hat{x}_t)-f_{x_t}(x_{t+1},\xi_t)-r(x_{t+1})\right]. \tag{4.6.1}$$

We now seek to bound the right-hand side of (4.6.1) using (B3)-(B6). To that end, the following bound will be useful:

$$\begin{aligned} f_{x_t}(x_{t+1},\xi_t) &= f(x_t,\xi_t)+\langle G(x_t,\xi_t),x_{t+1}-x_t\rangle \\ &\geq f(x_t,\xi_t)+\langle \nabla f(x_t),x_{t+1}-x_t\rangle - \|G(x_t,\xi_t)-\nabla f(x_t)\|_*\|x_{t+1}-x_t\|. \end{aligned}$$

Taking expectations of both sides and applying Cauchy-Schwarz and (B3)-(B6), we obtain

$$\begin{aligned} \mathbb{E}_t\left[f_{x_t}(x_{t+1},\xi_t)\right] &\geq \mathbb{E}_t\left[f(x_t)+\langle\nabla f(x_t),x_{t+1}-x_t\rangle\right]-\mathbb{E}_t\left[\|G(x_t,\xi_t)-\nabla f(x_t)\|_*\|x_{t+1}-x_t\|\right] \\ &\geq \mathbb{E}_t\left[f(x_{t+1})-MD_\Phi(x_{t+1},x_t)\right]-\sqrt{\mathbb{E}_t\left[\|G(x_t,\xi_t)-\nabla f(x_t)\|_*^2\right]}\sqrt{\mathbb{E}_t\left[\|x_{t+1}-x_t\|^2\right]} \\ &\geq \mathbb{E}_t\left[f(x_{t+1})-MD_\Phi(x_{t+1},x_t)\right]-\sigma\sqrt{\mathbb{E}_t\left[\tfrac{1}{2}\|x_{t+1}-x_t\|^2\right]} \\ &\geq \mathbb{E}_t\left[f(x_{t+1})-MD_\Phi(x_{t+1},x_t)\right]-\sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}. \end{aligned} \tag{4.6.2}$$

Continuing, add $f_{x_t}(\hat{x}_t,\xi_t)$ to both sides of (4.6.2), rearrange, and apply (B3) to obtain

$$\begin{aligned} &\mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t)-f_{x_t}(x_{t+1},\xi_t)\right] \\ &\quad\leq \mathbb{E}_t\left[f_{x_t}(\hat{x}_t,\xi_t)-f(x_{t+1})+MD_\Phi(x_{t+1},x_t)\right]+\sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]} \\ &\quad\leq \mathbb{E}_t\left[f(\hat{x}_t)-f(x_{t+1})+\tau D_\Phi(\hat{x}_t,x_t)+MD_\Phi(x_{t+1},x_t)\right]+\sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}. \end{aligned} \tag{4.6.3}$$

On the other hand, by the definition of $\hat{x}_t$ we have

$$f(\hat{x}_t)+r(\hat{x}_t)+\frac{1}{\lambda}D_\Phi(\hat{x}_t,x_t)\leq f(x_{t+1})+r(x_{t+1})+\frac{1}{\lambda}D_\Phi(x_{t+1},x_t).$$

Inserting this equation into (4.6.3), we obtain

$$\begin{aligned} &\mathbb{E}_t\left[f(\hat{x}_t)+r(\hat{x}_t)-f(x_{t+1})-r(x_{t+1})\right] \\ &\leq \mathbb{E}_t\left[\left(M+\tfrac{1}{\lambda}\right)D_\Phi(x_{t+1},x_t)+\left(\tau-\tfrac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)\right]+\sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}. \end{aligned} \tag{4.6.4}$$

Combining (4.6.4) with (4.6.1) gives the estimate

$$\begin{aligned}\frac{1}{\eta_t}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})-D_\Phi(\hat{x}_t,x_t)+D_\Phi(x_{t+1},x_t)\right]\\ \leq \left(M+\tfrac{1}{\lambda}\right)\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]+\left(\tau-\tfrac{1}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\sigma\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]},\end{aligned}$$

Multiplying through by η_t and rearranging, we obtain

$$\begin{aligned}\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})+\left(1-\eta_t\left(M+\tfrac{1}{\lambda}\right)\right)D_\Phi(x_{t+1},x_t)\right]\\ \leq \left(1+\eta_t(\tau-\tfrac{1}{\lambda})\right)D_\Phi(\hat{x}_t,x_t)+\sigma\eta_t\sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}.\end{aligned}$$

Now define $\gamma := \sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}$, and rewrite the above as

$$\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})\right] \leq \left(1+\eta_t\tau-\tfrac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\sigma\eta_t\gamma-\left(1-\eta_t(M+\tfrac{1}{\lambda})\right)\gamma^2.$$

Maximizing the right hand side in γ, i.e. taking $\gamma = \frac{\sigma\eta_t}{2\left(1-\eta_t\left(M+\frac{1}{\lambda}\right)\right)}$, we conclude

$$\mathbb{E}_t\left[(1-\eta_t\rho)D_\Phi(\hat{x}_t,x_{t+1})\right] \leq \left(1+\eta_t\tau-\tfrac{\eta_t}{\lambda}\right)D_\Phi(\hat{x}_t,x_t)+\frac{1}{4}\cdot\frac{(\sigma\eta_t)^2}{1-\eta_t(M+\frac{1}{\lambda})},$$

as desired. □

With Lemma 8 at hand, we can now establish a convergence rate of Algorithm 7.

Theorem 4.6.1. *The point x_{t^*} returned by Algorithm 7 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}^\Phi_{\lambda F}(x_{t^*}),x_{t^*}\right)\right] \leq \frac{\lambda}{(1-(\tau+\rho)\lambda)}\left(\frac{\lambda(F^\Phi_\lambda(x_0)-\min F)}{\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}+\frac{\sigma^2\sum_{t=0}^T\frac{\eta_t^2}{(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}}{4\sum_{t=0}^T\frac{\eta_t}{1-\eta_t\rho}}\right).$$

Proof. Using Lemma 8, we obtain

$$\begin{aligned}\mathbb{E}_t\left[F^\Phi_\lambda(x_{t+1})\right] &\leq \mathbb{E}_t\left[f(\hat{x}_t)+\frac{1}{\lambda}D_\Phi(\hat{x}_t,x_{t+1})\right]\\ &\leq \mathbb{E}_t\left[f(\hat{x}_t)+\frac{1}{\lambda(1-\eta_t\rho)}\left(\left(1+\eta_t\tau-\eta_t/\lambda\right)D_\Phi(\hat{x}_t,x_t)+\frac{1}{4}\cdot\frac{(\sigma\eta_t)^2}{1-\eta_t(M+1/\lambda)}\right)\right]\\ &= F^\Phi_\lambda(x_t)+\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)D_\Phi(\hat{x}_t,x_t)+\frac{(\sigma\eta_t)^2}{4\lambda(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}\end{aligned}$$

Recursing and applying the tower rule for expectations, we obtain

$$\mathbb{E}\left[F_\lambda^\Phi(x_{T+1})\right] \leq F_\lambda^\Phi(x_0) + \sum_{t=0}^{T}\left(\frac{\eta_t}{\lambda}\left(\frac{\tau+\rho-1/\lambda}{1-\eta_t\rho}\right)\mathbb{E}\left[D_\Phi(\hat{x}_t, x_t)\right] + \frac{(\sigma\eta_t)^2}{4\lambda(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}\right)$$

Rearranging and using the fact that $\mathbb{E}\left[F_\lambda(x_{T+1})\right] \geq \min F$, we obtain

$$\sum_{t=0}^{T}\frac{\eta_t}{\lambda}\left(\frac{1/\lambda-\tau-\rho}{1-\eta_t\rho}\right)\mathbb{E}\left[D_\Phi(\hat{x}_t, x_t)\right] \leq F_\lambda^\Phi(x_0) - \min F + \frac{\sigma^2}{4\lambda}\sum_{t=0}^{T}\frac{\eta_t^2}{(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}$$

or equivalently

$$\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}\mathbb{E}\left[D_\Phi(\hat{x}_t, x_t)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0)-\min F)}{1-(\tau+\rho)\lambda} + \frac{\lambda\sigma^2}{4(1-(\tau+\rho)\lambda)}\sum_{t=0}^{T}\frac{\eta_t^2}{(1-\eta_t(M+1/\lambda))(1-\eta_t\rho)}.$$

Dividing through by $\sum_{t=0}^{T}\frac{\eta_t}{1-\eta_t\rho}$ and recognizing the left-hand-side as $\mathbb{E}[D_\Phi(\hat{x}_{t^*}, x_{t^*})]$, the result follows. □

As an immediate corollary, we obtain a convergence rate for Algorithm 7 with a constant stepsize.

Corollary 7. *For some $\alpha > 0$, set $\eta_t = \frac{1}{M+\lambda^{-1}+\alpha^{-1}\sqrt{T+1}}$ for all indices $t = 1, \ldots, T$. Then the point x_{t^*} returned by Algorithm 7 satisfies:*

$$\mathbb{E}\left[D_\Phi\left(\mathrm{prox}_{\lambda F}^\Phi(x_{t^*}), x_{t^*}\right)\right] \leq \frac{\lambda^2(F_\lambda^\Phi(x_0)-\min F) + \lambda(\frac{\sigma\alpha}{2})^2}{(1-(\tau+\rho)\lambda)} \cdot \left(\frac{M+\lambda^{-1}-\rho}{T+1} + \frac{1}{\alpha\sqrt{T+1}}\right).$$

4.7 Rates in function value for convex problems

In this final section, we examine convergence rates for stochastic model based minimization under convexity assumptions and prove rates of converge on function values. To this end, we will use the following definition from [30]. A function $g\colon \mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$ is *μ-strongly convex relative to Φ* if the function $g - \mu\Phi$ is convex. Notice that $\mu = 0$ corresponds to plain convexity of g.

In this section, we make the following assumptions:

(C1) **(Sampling)** It is possible to generate i.i.d. realizations $\xi_1, \ldots, \xi_T \sim P$

(C2) **(One-sided accuracy)** There is a measurable function $(x, y, \xi) \mapsto f_x(y, \xi)$ defined on $U \times U \times \Omega$ satisfying both

$$\mathbb{E}_\xi [f_x(x, \xi)] = f(x), \qquad \forall x \in U \cap \operatorname{dom} r$$

and

$$\mathbb{E}_\xi [f_x(y, \xi)] \leq f(y), \qquad \forall x, y \in U \cap \operatorname{dom} r. \tag{4.7.1}$$

(C3) **(Convexity of the models)** The exists some $\mu \geq 0$ such that the functions $f_x(\cdot, \xi)+r(\cdot)$ are μ-strongly convex relative to Φ for all $x \in U \cap \operatorname{dom} r$ and a.e. $\xi \in \Omega$.

(C4) **(Lipschitz property)** There exists a square integrable function $L \colon \Omega \to \mathbb{R}_+$ such that for all $x, y \in U \cap \operatorname{dom} r$, the following inequalities holds:

$$\begin{aligned} f_x(x, \xi) - f_x(y, \xi) &\leq L(\xi)\sqrt{D_\Phi(y, x)}, \\ \sqrt{\mathbb{E}_\xi \left[L(\xi)^2\right]} &\leq \mathsf{L}. \end{aligned} \tag{4.7.2}$$

(C5) **(Solvability)** The convex problems

$$\min_y \left\{ F(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\} \qquad \text{and} \qquad \min_y \left\{ f_x(y, \xi) + r(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

admit a minimizer for any $\lambda > 0$, any $x \in U$, and a.e. $\xi \in \Omega$. The minimizers vary measurably in $(x, \xi) \in U \times \Omega$.

Thus the only difference between assumptions (C1)-(C5) and (A1)-(A5) is that in expectation the stochastic models $f(\cdot, \xi)$ are global under-estimators (C2) and the functions $f(\cdot, \xi) + r(\cdot)$ are relatively strongly convex, instead of weakly convex (C3). Note that under assumptions (C1)-(C5), the objective function F is μ-strongly convex relative to Φ; the argument is completely analogous to that of Lemma 6.

Henceforth, we let $\{x_t\}_{t \geq 0}$ be the iterates generated by Algorithm 6 (with $\tau = \rho = 0$) and let $\{\xi_t\}_{t \geq 0}$ be the corresponding samples used. For each index $t \geq 0$, we continue to use the notation $\hat{x}_t = \operatorname{prox}^\Phi_{\lambda F}(x)$ and let $\mathbb{E}_t[\cdot]$ to denote the expectation conditioned on all the

realizations $\xi_0, \xi_1, \ldots, \xi_{t-1}$. We need the following key lemma, which identifies the Bregman divergence $D_\Phi(x^*, x_t)$, between the iterates and an optimal solution, as a useful potential function. Notice that this is in contrast to the nonconvex setting, where it was the envelope $F_\lambda^\Phi(x_t)$ that served as an appropriate potential function.

Lemma 9. *For each iteration $t \geq 0$, the iterates of Algorithm 6 satisfy*

$$\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*, x_{t+1})\right] \leq D_\Phi(x^*, x_t) + \frac{(\mathsf{L}\eta_t)^2}{4} + \eta_t\mathbb{E}_t\left[r(x_t) - r(x_{t+1})\right] - \eta_t(F(x_t) - F(x^*)),$$

where x^* is any minimizer of F.

Proof. Appealing to the three point inequality in Lemma 4 and (C3), we deduce that all points $x \in \operatorname{dom} r$ satisfy

$$f_{x_t}(x,\xi_t) + r(x) + \frac{1}{\eta_t}D_\Phi(x, x_t) \geq f_{x_t}(x_{t+1},\xi_t) + r(x_{t+1}) + \frac{1}{\eta_t}D_\Phi(x_{t+1}, x_t) + \frac{(1+\eta_t\mu)}{\eta_t}D_\Phi(x, x_{t+1}). \tag{4.7.3}$$

Setting $x = x^*$, rearranging terms, and taking expectations, we deduce

$$\begin{aligned}\frac{1}{\eta_t}&\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*, x_{t+1}) - D_\Phi(x^*, x_t) + D_\Phi(x_{t+1}, x_t)\right] \\ &\leq \mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) + r(x^*) - f_{x_t}(x_{t+1},\xi_t) - r(x_{t+1})\right].\end{aligned} \tag{4.7.4}$$

We seek to upper bound the right-hand-side of (4.7.4). Assumptions (C2) and (C4) imply:

$$\begin{aligned}\mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) - f_{x_t}(x_{t+1},\xi_t)\right] &\leq \mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) - f_{x_t}(x_t,\xi_t) + L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)}\right] \\ &= \mathbb{E}_t\left[f_{x_t}(x^*,\xi_t) - f(x^*)\right] + \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)}\right] - f(x_t) + f(x^*) \\ &\leq \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)}\right] - f(x_t) + f(x^*).\end{aligned}$$

The left hand side of (4.7.4) is therefore upper bounded by

$$\begin{aligned}&\mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)} - f(x_t) - r(x_{t+1})\right] + f(x^*) + r(x^*) \\ &= \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1}, x_t)} + (r(x_t) - r(x_{t+1}))\right] - (F(x_t) - F(x^*)).\end{aligned}$$

Putting everything together, we arrive at

$$\frac{1}{\eta_t}\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1}) - D_\Phi(x^*,x_t) + D_\Phi(x_{t+1},x_t)\right] \\ \leq \mathbb{E}_t\left[L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)} + (r(x_t)-r(x_{t+1}))\right] - (F(x_t)-F(x^*))$$

Multiplying through by η_t and rearranging yields

$$\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1})\right] \leq D_\Phi(x^*,x_t) + \mathbb{E}_t\left[\eta_t L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)} - D_\Phi(x_{t+1},x_t)\right] \\ + \eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right] - \eta_t(F(x_t)-F(x^*)).$$

Now define $\gamma := \sqrt{\mathbb{E}_t\left[D_\Phi(x_{t+1},x_t)\right]}$. By Cauchy-Schwarz, we have that $\mathbb{E}_t\left[\eta_t L(\xi)\sqrt{D_\Phi(x_{t+1},x_t)}\right] \leq \eta_t \mathsf{L}\gamma$. Thus we obtain

$$\mathbb{E}_t\left[(1+\eta_t\mu)D_\Phi(x^*,x_{t+1})\right] \leq D_\Phi(x^*,x_t) + \eta_t\mathsf{L}\gamma - \gamma^2 + \eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right] - \eta_t(F(x_t)-F(x^*)) \\ \leq D_\Phi(x^*,x_t) + \frac{(\mathsf{L}\eta_t)^2}{4} + \eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right] - \eta_t(F(x_t)-F(x^*)),$$

where the last inequality follows by maximizing the right-hand-side in γ. □

We are now ready to prove convergence guarantees in the case that $\mu = 0$.

Theorem 4.7.1 (Convergence rate under convexity)**.** *For all $T > 0$, we have*

$$\mathbb{E}\left[F\left(\frac{1}{\sum_{t=0}^T \eta_t}\sum_{t=0}^T \eta_t x_t\right) - F(x^*)\right] \leq \frac{D_\Phi(x^*,x_0) + \sum_{t=0}^t \frac{(\eta_t\mathsf{L})^2}{4} + \eta_0(r(x_0)-\inf r)}{\sum_{t=0}^T \eta_t},$$

where x^* is any minimizer of F.

Proof. Lower-bounding the left-hand-side of Lemma 9 by zero $D_\Phi(x^*,x_{t+1})$, we deduce

$$\eta_t\left[F(x_t)-F(x^*)\right] \leq \frac{(\mathsf{L}\eta_t)^2}{4} + \eta_t\mathbb{E}_t\left[r(x_t)-r(x_{t+1})\right] + \mathbb{E}_t[D_\Phi(x^*,x_t) - D_\Phi(x^*,x_{t+1})]$$

Applying the tower rule for expectations yields

$$\sum_{t=0}^T \eta_t\mathbb{E}\left[F(x_t)-F(x^*)\right] \\ \leq \sum_{t=0}^T \frac{(\eta_t\mathsf{L})^2}{4} + \mathbb{E}\left[\sum_{t=0}^T \eta_t(r(x_t)-r(x_{t+1}))\right] + \mathbb{E}\left[\sum_{t=0}^T (D_\Phi(x^*,x_t) - D_\Phi(x^*,x_{t+1}))\right].$$

Using Jensen's inequality, telescoping and using the auxiliary Lemma 10, we conclude

$$\mathbb{E}\left[F\left(\frac{1}{\sum_{t=0}^{T}}\sum_{t=0}^{T}\eta_t x_t\right) - F(x^*)\right] \leq \frac{D_\Phi(x^*, x_0) + \sum_{t=0}^{t}\frac{(\eta_t \mathsf{L})^2}{4} + \eta_0(r(x_0) - \inf r)}{\sum_{t=0}^{T}\eta_t},$$

as claimed. □

As an immediate corollary of Theorem 4.5.1, we have the following rate of convergence when the stepsize η_t is constant.

Corollary 8 (Convergence rate under convexity for constant stepsize)**.** *For any $\alpha > 0$ and corresponding constant stepsize $\eta_t = \frac{\alpha}{\sqrt{T+1}}$, we have*

$$\mathbb{E}\left[F\left(\frac{1}{T+1}\sum_{t=0}^{T}x_t\right) - F(x^*)\right] \leq \frac{D_\Phi(x^*, x_0) + \frac{(\alpha \mathsf{L})^2}{4} + \alpha(r(x_0) - \inf r)}{\alpha\sqrt{T+1}},$$

where x^* is any minimizer of F.

The final result of this section proves that Algorithm 6, with an appropriate choice of stepsize, drives the expected error in function values to zero at the rate $\widetilde{O}(\frac{1}{k})$, whenever $\mu > 0$.

Theorem 4.7.2 (Convergence rate strongly convex case)**.** *Suppose that $\eta_t = \frac{1}{\mu(t+1)}$ for all $t \geq 0$. Then for all $T > 0$, we have*

$$\mathbb{E}\left[F\left(\frac{1}{T+1}\sum_{t=0}^{T}x_t\right) - F(x^*) + \mu D_\Phi(x^*, x_{T+1})\right] \leq \frac{\frac{\mathsf{L}^2(1+\log(T+1))}{4\mu} + r(x_0) - \inf r + \mu D_\Phi(x^*, x_0)}{T+1}$$

where x^* is any minimizer of F.

Proof. Using Lemma 9 and the law of total expectation, we have

$$\mathbb{E}\left[F(x_t) - F(x^*)\right] \leq \frac{\eta_t \mathsf{L}^2}{4} + \mathbb{E}\left[(r(x_t) - r(x_{t+1})) + \frac{1}{\eta_t}D_\Phi(x^*, x_t) - \frac{(1+\eta_t\mu)}{\eta_t}D_\Phi(x^*, x_{t+1})\right]$$

Setting $\eta_t = \frac{1}{\mu(t+1)}$, averaging, and applying Jensen's inequality yields

$$\begin{aligned}\mathbb{E}\left[F\left(\frac{1}{T+1}\sum_{t=0}^{T} x_t\right) - F(x^*)\right] &\leq \frac{1}{T+1}\sum_{t=0}^{T}\frac{\mathsf{L}^2}{4\mu(t+1)} + \frac{\mathbb{E}\left[r(x_0) - r(x_{T+1})\right]}{T+1} \\ &\quad + \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left[\mu(t+1)D_\Phi(x^*, x_t) - \mu(t+2)D_\Phi(x^*, x_{t+1})\right] \\ &\leq \frac{\frac{\mathsf{L}^2(1+\log(T+1))}{4\mu} + r(x_0) - \inf r + \mu D_\Phi(x^*, x_0)}{T+1} \\ &\quad - \mathbb{E}\left[\frac{(T+2)}{(T+1)}\mu D_\Phi(x^*, x_{T+1})\right],\end{aligned}$$

where the last inequality follows from telescoping the terms in the sum and using the lower bound $r(x_{T+1}) \geq \inf r$. This completes the proof. □

References

[1] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[2] H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematical of Operations Research*, 42(2):330–348, 2017.

[3] H. Bauschke and J. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

[4] H. Bauschke, M. Dao, and S. Lindstrom. Regularizing with Bregman-Moreau envelopes. *arXiv:1705.06019*, 2017.

[5] A. Beck. *First-order methods in optimization*, volume 25 of *MOS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, 2017.

[6] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] J. Bello Cruz. On proximal subgradient splitting method for minimizing the sum of two nonsmooth convex functions. *Set-Valued and Variational Analysis*, 25(2):245–263, Jun 2017.

[8] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *arXiv:1706.06461*, 2017.

[9] J. Borwein and A. Lewis. *Convex analysis and nonlinear optimization*, volume 3 of *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC*. Springer, New York, second edition, 2006. Theory and examples.

[10] S. Bubeck. *Convex Optimization: Algorithms and Complexity*. Foundations and Trends in Machine Learning. Now Publishers, 2015.

[11] J. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33(3):260–279, 1985.

[12] C. Cartis, N. Gould, and P. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

[13] Y. Censor and S. Zenios. Proximal minimization algorithm with D-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464, 1992.

[14] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

[15] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *arXiv:1803.06523*, 2018.

[16] D. Davis and D. Drusvyatskiy. Stochastic subgradient method converges at the rate $O(k^{-1/4})$ on weakly convex functions. *arXiv:1802.02988*, 2018.

[17] D. Drusvyatskiy. Proximal algorithms. *SIAG/OPT Views and News*, 26(1):1–8, Jan. 2018.

[18] D. Drusvyatskiy and A. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *To appear in Mathematics of Operations Research, arXiv:1602.06661, version 2*, 2016.

[19] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *To appear in Mathematical Programming, arXiv:1605.00125*, 2018.

[20] J. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *arXiv:1703.08570*, 2017.

[21] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10(Dec):2899–2934, 2009.

[22] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

[23] R. Fletcher. A model algorithm for composite nondifferentiable optimization problems. *Mathematical Programming Study*, (17):67–76, 1982. Nondifferential and variational techniques in optimization (Lexington, Ky., 1980).

[24] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2, Ser. A):267–305, 2016.

[25] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[26] F. Hanzely and P. Richtárik. Fastest rates for stochastic mirror descent methods. *arXiv:1803.07374*, 2017.

[27] A. Juditsky and A. Nemirovski. First order methods for nonsmooth convex large-scale optimization, II: Utilizing problem's structure. In S. J. W. Suvrit Sra, Sebastian Nowozin, editor, *Optimization for Machine Learning*, pages 29–63. MIT Press, Aug. 2010.

[28] A. Lewis and S. Wright. A proximal method for composite minimization. *Mathematical Programming*, pages 1–46, 2015.

[29] H. Lu. Relative continuity for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv:1710.04718*, 2017.

[30] H. Lu, R. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *arXiv:1610.05708*, 2016.

[31] A. Nemirovsky and D. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

[32] Y. Nesterov. Modified Gauss-Newton scheme with worst case guarantees for global performance. *Optimization Methods & Software*, 22(3):469–483, 2007.

[33] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[34] E. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[35] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, Jan. 2014.

[36] M. Powell. On the global convergence of trust region algorithms for unconstrained minimization. *Mathematical Programming*, 29(3):297–303, 1984.

[37] R. Rockafellar. *Convex Analysis*. Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[38] R. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, third edition, 2009.

[39] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):670–690, 1992.

[40] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, May 2018.

[41] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. http://www.mit.edu/ dimitrib/PTseng/papers/apgm.pdf, May 2008.

[42] S. Zhang and N. He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. arXiv:1806.04781, June 2018.

Appendix: proofs of auxilliary results

4.7.1 Proof of Proposition 1

Let us write

$$\Phi = \widehat{\Phi} + \widetilde{\Phi},$$

for the two functions

$$\widehat{\Phi}(x) := \sum_{i=0}^{n} \frac{a_i}{i+2}\|x\|_2^{i+2} \qquad \text{and} \qquad \widetilde{\Phi}(x) := \sum_{i=0}^{n} 3a_i\|x\|_2^{i+2}.$$

The result [29, Equation (25)] yields the estimate

$$D_{\widehat{\Phi}}(y,x) \geq \frac{1}{2}\sum_{i=0}^{n} a_i\|x\|_2^i \cdot \|x-y\|_2^2 \qquad \forall x, y.$$

Thus the proof will be complete once we establish the inequality,

$$D_{\widetilde{\Phi}}(y,x) \geq \frac{1}{2}\sum_{i=0}^{n} a_i\|y\|_2^i \cdot \|x-y\|_2^2 \qquad \forall x, y. \tag{4.7.5}$$

To this end, fix an index i, and set $\eta := 3(i+2)$ and $\widetilde{\Phi}_i(x) := 3a_i\|x\|_2^{i+2}$. We will show

$$D_{\widetilde{\Phi}_i}(y,x) \geq \frac{a_i}{2}\|y\|_2^i \cdot \|x-y\|_2^2,$$

which together with the identity, $D_{\widetilde{\Phi}}(y,x) = \sum_{i=0}^{n} D_{\widetilde{\Phi}_i}(y,x)$, completes the proof of (4.7.5).

A quick computation shows that

$$D_{\widetilde{\Phi}_i}(y,x) = 3a_i\left(\|y\|_2^{i+2} + (i+1)\|x\|_2^{i+2} - (i+2)\|x\|_2^i\langle x, y\rangle\right).$$

Let us consider two cases. First suppose that $\eta^{1/i}\|x\|_2 \geq \|y\|_2$. In this case, [29, Proposition 5.1] implies

$$D_{\widetilde{\Phi}_i}(y,x) \geq \frac{a_i\eta}{2}\|x\|_2^i \cdot \|x-y\|_2^2 \geq \frac{a_i}{2}\|y\|_2^i \cdot \|x-y\|_2^2,$$

as desired.

Now suppose that $\|y\|_2 \geq \eta^{1/i}\|x\|_2$. We will show that $D_{\widetilde{\Phi}_i}(y,x) \geq \eta^{-1} D_{\widetilde{\Phi}_i}(x,y)$, which will complete the proof since

$$\eta^{-1} D_{\widetilde{\Phi}_i}(x,y) \geq \frac{a_i}{2}\|y\|^i \cdot \|x-y\|_2^2,$$

by [29, Proposition 5.1]. To that end, we compute

$$\begin{aligned}
&D_{\widetilde{\Phi}_i}(y,x) = 3a_i\left(\|y\|_2^{i+2} + (i+1)\|x\|_2^{i+2} - (i+2)\|x\|_2^i\langle x,y\rangle\right)\\
&\geq \eta^{-1} D_{\widetilde{\Phi}_i}(x,y) = \tfrac{a_i}{i+2}\left(\|x\|_2^{i+2} + (i+1)\|y\|_2^{i+2} - (i+2)\|y\|_2^i\langle x,y\rangle\right)\\
\iff &(1-\eta^{-1}(i+1))\|y\|_2^{i+2} + \eta^{-1}(i+2)\|y\|_2^i\langle x,y\rangle \geq (\eta^{-1}-(i+1))\|x\|_2^{i+2} + (i+2)\|x\|_2^i\langle x,y\rangle\\
\impliedby &(1-\eta^{-1}(i+1))\|y\|_2^i\left(\|y\|^2 + \frac{\eta^{-1}(i+2)}{(1-\eta^{-1}(i+1))}\langle x,y\rangle\right) \geq (i+2)\|x\|_2^i\langle x,y\rangle.
\end{aligned}$$

Let us show that the last inequality is true: First, we upper bound the right hand side

$$(i+2)\|x\|_2^i\langle x,y\rangle \leq \frac{(i+2)}{\eta^{(1+i)/i}}\|y\|^{i+2}.$$

Next, we lower bound the left hand side:

$$\begin{aligned}
&(1-\eta^{-1}(i+1))\|y\|_2^i\left(\|y\|^2 + \frac{\eta^{-1}(i+2)}{(1-\eta^{-1}(i+1))}\langle x,y\rangle\right)\\
&\geq (1-\eta^{-1}(i+1))\left(1 - \frac{\eta^{-1}(i+2)}{\eta^{1/i}(1-\eta^{-1}(i+1))}\right)\|y\|_2^{i+2}\\
&= \left(1-\eta^{-1}(i+1+\tfrac{(i+2)}{\eta^{1/i}})\right)\|y\|_2^{i+2}.
\end{aligned}$$

Therefore, we need only verify that η satisfies

$$\begin{aligned}
&\frac{(i+2)}{\eta^{(1+i)/i}} \leq \left(1-\eta^{-1}(i+1+\frac{(i+2)}{\eta^{1/i}})\right)\\
\iff &(i+2) \quad \leq \eta^{(1+i)/i} - \eta^{1/i}(i+1) - (i+2)\\
\iff &2(i+2) \leq \eta^{1/i}\left(\eta - (i+1)\right),
\end{aligned}$$

which holds by the definition of η. Thus the result is proved.

4.7.2 An auxiliary lemma on sequences.

Lemma 10. *Consider any nonincreasing sequence $\{a_t\}_{t\geq 0} \subset \mathbb{R}_{++}$ and any sequence $\{b_t\}_{t\geq 0} \subset \mathbb{R}$. Then for any index $T \in \mathbb{N}$, we have*

$$\sum_{t=0}^{T} a_t(b_t - b_{t+1}) \leq a_0(b_0 - b^*),$$

where we set $b^* = \inf_{t\geq 0} b_t$.

Proof. We successively deduce

$$\begin{aligned}\sum_{t=0}^{T} a_t(b_t - b_{t+1}) &= \sum_{t=0}^{T} a_t[(b_t - b^*) - (b_{t+1} - b^*)] \\ &= a_0(b_0 - b^*) - a_T(b_{T+1} - b^*) + \sum_{t=0}^{T-1} (a_{t+1} - a_t)\,(b_{t+1} - b^*) \\ &\leq a_0(b_0 - b^*),\end{aligned}$$

as claimed. □

4.7.3 Proofs of Propositions 2 and 3

Proof of Proposition 2. Using the fundamental theorem of calculus and convexity of the function $x \mapsto p(\|x\|_2)$ we compute

$$\begin{aligned}&\|c(x,\xi) + \nabla c(x,\xi)(y-x) - c(y,\xi)\|_2 \\ &= \left\| \int_0^1 (\nabla c(x + t(y-x),\xi) - \nabla c(x,\xi))\,(y-x)\,dt \right\|_2 \\ &\leq \int_0^1 \|\nabla c(x + t(y-x),\xi) - \nabla c(x,\xi)\|_{\mathrm{op}}\,\|y-x\|_2\,dt \\ &\leq L_2(\xi)\|y-x\|_2^2 \int_0^1 (p(\|x + t(y-x)\|_2) + p(\|x\|_2))\,t\,dt \\ &\leq L_2(\xi)\|y-x\|_2^2 \int_0^1 ((1-t)p(\|x\|_2) + tp(\|y\|_2)) + p(\|x\|_2))\,t\,dt \\ &\leq \frac{2L_2(\xi)}{3}\|y-x\|_2^2 \cdot (p(\|x\|_2) + p(\|y\|_2)).\end{aligned}$$

Hence, we deduce

$$\begin{aligned} h\left(c(x,\xi)+\nabla c(x,\xi)(y-x),\xi\right)-h\left(c(y,\xi),\xi\right) &\leq L_1(\xi)\cdot\|c(x,\xi)+\nabla c(x,\xi)(y-x)-c(y,\xi)\|_2 \\ &\leq \tfrac{2}{3}L_1(\xi)L_2(\xi)\|y-x\|_2^2\cdot\left(p(\|x\|_2)+p(\|y\|_2)\right) \\ &\leq \tfrac{4}{3}L_1(\xi)L_2(\xi)\cdot D_\Phi(y,x), \end{aligned}$$

where the last inequality follows form Proposition 1. Taking expectations yields the claimed guarantee. □

Proof of Proposition 3. We successively compute

$$\begin{aligned} h\left(c(x,\xi),\xi\right)-h\left(c(x,\xi)+\nabla c(x,\xi)(y-x),\xi\right) &= L_1(\xi)\|\nabla c(x,\xi)(y-x)\|_2 \\ &\leq L_1(\xi)L_3(\xi)\cdot\sqrt{q(\|x\|_2)}\|y-x\|_2 \\ &\leq \sqrt{2}L_1(\xi)L_3(\xi)\cdot\sqrt{D_\Phi(y,x)}, \end{aligned}$$

where the last line follows from [29, Equation (25)]. The result follows. □

4.7.4 Proof of Theorem 4.4.1

First we rewrite F_λ^Φ, using the definition of the Bregman divergence, as

$$\begin{aligned} F_\lambda^\Phi(x) &= \inf_y\left\{F(y)+\frac{1}{\lambda}\Phi(y)-\frac{1}{\lambda}\langle\nabla\Phi(x)\rangle y\right\}-\frac{1}{\lambda}\Phi(x)+\frac{1}{\lambda}\langle\nabla\Phi(x)\rangle x \\ &= -\sup_y\left\{\langle\frac{1}{\lambda}\nabla\Phi(x)\rangle y-\left(F+\frac{1}{\lambda}\Phi\right)(y)\right\}-\frac{1}{\lambda}\Phi(x)+\frac{1}{\lambda}\langle\nabla\Phi(x)\rangle x \\ &= -\left(F+\frac{1}{\lambda}\Phi\right)^\star\left(\frac{1}{\lambda}\nabla\Phi(x)\right)-\frac{1}{\lambda}\Phi(x)+\frac{1}{\lambda}\langle\nabla\Phi(x)\rangle x. \end{aligned}$$

Note that $F+\frac{1}{\lambda}\Phi$ is closed and $\left(\frac{1}{\lambda}-(\rho+\tau)\right)$-strongly convex. Thus the conjugate $(F+\frac{1}{\lambda}\Phi)^\star$ is differentiable. By the chain and sum rules for differentiation, we have

$$\begin{aligned} \nabla F_\lambda^\Phi(x) &= -\frac{1}{\lambda}\nabla^2\Phi(x)\left[\nabla\left(F+\frac{1}{\lambda}\Phi\right)^\star\right]\left(\frac{1}{\lambda}\nabla\Phi(x)\right)+\frac{1}{\lambda}\nabla^2\Phi(x)x \\ &= \frac{1}{\lambda}\nabla^2\Phi(x)\left(x-\left[\nabla\left(F+\frac{1}{\lambda}\Phi\right)^\star\right]\left(\frac{1}{\lambda}\nabla\Phi(x)\right)\right) \end{aligned}$$

The (sub)gradient of a convex conjugate function is simply the set of maximizers in the supremum defining the conjugate, so that

$$\begin{aligned}\left[\nabla\left(F+\frac{1}{\lambda}\Phi\right)^{\star}\right]\left(\frac{1}{\lambda}\nabla\Phi(x)\right) &= \operatorname*{argmax}_{y}\left\{\langle\frac{1}{\lambda}\nabla\Phi(x)\rangle y-\left(F+\frac{1}{\lambda}\Phi\right)(y)\right\}\\ &= \operatorname*{argmin}_{y}\left\{F(y)+\frac{1}{\lambda}D_{\Phi}(y,x)\right\}\\ &= \operatorname{prox}_{\lambda F}^{\Phi}(x).\end{aligned}$$

Putting everything together, we obtain, $\nabla F_{\lambda}^{\Phi}(x)=\frac{1}{\lambda}\nabla^{2}\Phi(x)\,(x-\hat{x})$, as desired.