IAS/Park City Mathematics Series
Volume 00, Pages 000–000
S 1079-5634(XX)0000-0

Introductory Lectures on Stochastic Optimization

John C. Duchi

Contents

1 Introduction 2
1.1 Scope, limitations, and other references 3
1.2 Notation 4
2 Basic Convex Analysis 5
2.1 Introduction and Definitions 5
2.2 Properties of Convex Sets 7
2.3 Continuity and Local Differentiability of Convex Functions 14
2.4 Subgradients and Optimality Conditions 16
2.5 Calculus rules with subgradients 21
3 Subgradient Methods 24
3.1 Introduction 24
3.2 The gradient and subgradient methods 25
3.3 Projected subgradient methods 31
3.4 Stochastic subgradient methods 35
4 The Choice of Metric in Subgradient Methods 43
4.1 Introduction 43
4.2 Mirror Descent Methods 44
4.3 Adaptive stepsizes and metrics 54
5 Optimality Guarantees 60
5.1 Introduction 60
5.2 Le Cam's Method 65
5.3 Multiple dimensions and Assouad's Method 70
A Technical Appendices 74
A.1 Continuity of Convex Functions 74
A.2 Probability background 76
A.3 Auxiliary results on divergences 78
B Questions and Exercises 80

2010 *Mathematics Subject Classification.* Primary 65Kxx; Secondary 90C15, 62C20.
Key words and phrases. Convexity, stochastic optimization, subgradients, mirror descent, minimax optimal.

1. Introduction

In this set of four lectures, we study the basic analytical tools and algorithms necessary for the solution of stochastic convex optimization problems, as well as for providing various optimality guarantees associated with the methods. As we proceed through the lectures, we will be more exact about the precise problem formulations, providing a number of examples, but roughly, by a stochastic optimization problem we mean a numerical optimization problem that arises from observing data from some (random) data-generating process. We focus almost exclusively on first-order methods for the solution of these types of problems, as they have proven quite successful in the large scale problems that have driven many advances throughout the early 2000s.

Our main goal in these lectures, as in the lectures by S. Wright in this volume, is to develop methods for the solution of optimization problems arising in large-scale data analysis. Our route will be somewhat circuitous, as we will build the necessary convex analytic and other background (see Lecture 2), but broadly, the problems we wish to solve are the problems arising in stochastic convex optimization. In these problems, we have samples S coming from a sample space $\mathcal{S}$, drawn from a distribution P, and we have some *decision vector* $x \in \mathbb{R}^n$ that we wish to choose to minimize the expected loss

$$f(x) := \mathbb{E}_P[F(x;S)] = \int_{\mathcal{S}} F(x;s)dP(s), \tag{1.0.1}$$

where F is convex in its first argument.

The methods we consider for minimizing problem (1.0.1) are typically simple methods that are slower to converge than more advanced methods—such as Newton or other second-order methods—for deterministic problems, but have the advantage that they are robust to noise in the optimization problem itself. Consequently, it is often relatively straightforward to derive generalization bounds for these procedures: if they produce an estimate $\widehat{x}$ exhibiting good performance on some sample $S_1, \ldots, S_m$ drawn from P, then they are likely to exhibit good performance (on average) for future data, that is, to have small objective $f(\widehat{x})$; see Lecture 3, and especially Theorem 3.4.11. It is of course often advantageous to take advantage of problem structure and geometric aspects of the problem, broadly defined, which is the goal of mirror descent and related methods, which we discuss in Lecture 4.

The last part of our lectures is perhaps the most unusual for material on optimization, which is to investigate optimality guarantees for stochastic optimization problems. In Lecture 5, we study the sample complexity of solving problems of the form (1.0.1). More precisely, we measure the performance of an optimization procedure given samples $S_1, \ldots, S_m$ drawn independently from the population distribution P, denoted by $\widehat{x} = \widehat{x}(S_{1:m})$, in a uniform sense: for a class of objective functions $\mathcal{F}$, a procedure's performance is its expected error—or risk—for the worst member of the class $\mathcal{F}$. We provide lower bounds on this maximum risk, showing that the first-order procedures we have developed satisfy certain notions of optimality.

We briefly outline the coming lectures. The first lecture provides definitions and the convex analytic tools necessary for the development of our algorithms and other ideas, developing separation properties of convex sets as well as other properties of convex functions from basic principles. The second two lectures investigate subgradient methods and their application to certain stochastic optimization problems, demonstrating a number of convergence results. The second lecture focuses on standard subgradient-type methods, while the third investigates more advanced material on mirror descent and adaptive methods, which require more care but can yield substantial practical performance benefits. The final lecture investigates optimality guarantees for the various methods we study, demonstrating two standard techniques for proving lower bounds on the ability of any algorithm to solve stochastic optimization problems.

1.1. Scope, limitations, and other references The lectures assume some limited familiarity with convex functions and convex optimization problems and their formulation, which will help appreciation of the techniques herein. All that is truly essential is a level of mathematical maturity that includes some real analysis, linear algebra, and introductory probability. In terms of real analysis, a typical undergraduate course, such as one based on Marsden and Hoffman's *Elementary Real Analysis* [37] or Rudin's *Principles of Mathematical Analysis* [50], are sufficient. Readers should not consider these lectures in any way a comprehensive view of convex analysis or stochastic optimization. These subjects are well-established, and there are numerous references.

Our lectures begin with convex analysis, whose study Rockafellar, influenced by Fenchel, launched in his 1970 book *Convex Analysis* [49]. We develop the basic ideas necessary for our treatment of first-order (gradient-based) methods for optimization, which includes separating and supporting hyperplane theorems, but we provide essentially no treatment of the important concepts of Lagrangian and Fenchel duality, support functions, or saddle point theory more broadly. For these and other important ideas, I have found the books of Rockafellar [49], Hiriart-Urruty and Lemaréchal [27, 28], Bertsekas [8], and Boyd and Vandenberghe [12] illuminating.

Convex optimization itself is a huge topic, with thousands of papers and numerous books on the subject. Because of our focus on solution methods for large-scale problems arising out of data collection, we are somewhat constrained in our views. Boyd and Vandenberghe [12] provide an excellent treatment of the possibilities of modeling engineering and scientific problems as convex optimization problems, as well as some important numerical methods. Polyak [47] provides a treatment of stochastic and non-stochastic methods for optimization from which ours borrows substantially. Nocedal and Wright [46] and Bertsekas [9] also describe more advanced methods for the solution of optimization problems, focusing on non-stochastic optimization problems for which there are many sophisticated methods.

Because of our goal to solve problems of the form (1.0.1), we develop first-order methods that are in some ways robust to many types of noise from sampling. There are other approaches to dealing with data uncertainty, and researchers in of robust optimization [6], who study and develop tractable (polynomial-time-solvable) formulations for a variety of data-based problems in engineering and the sciences. The book of Shapiro et al. [54] provides a more comprehensive picture of stochastic modeling problems and optimization algorithms than we have been able to in our lectures, as stochastic optimization is by itself a major field. Several recent surveys on online learning and online convex optimization provide complementary treatments to ours [26,52].

The last lecture traces its roots to seminal work in information-based-complexity by Nemirovski and Yudin in the early 1980s [41], who investigate the limits of "optimal" algorithms, where optimality is defined in a worst-case sense according to an oracle model of algorithms given access to function, gradient, or other types of local information about the problem at hand. Issues of optimal estimation in statistics are as old as the field itself, and the minimax formulation we use is originally due to Wald in the late 1930s [59,60]. We prove our results using information theoretic tools, which have broader applications across statistics, and that have been developed by many authors [31,33,61,62].

1.2. Notation We use mostly standard notation throughout these notes, but for completeness, we collect it here. We let $\mathbb{R}$ denote the typical field of real numbers, with $\mathbb{R}^n$ having its usual meaning as n-dimensional Euclidean space. Given vectors x and y, we let $\langle x, y\rangle$ denote the inner product between x and y. Given a norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ is defined as

$$\|z\|_* := \sup\left\{\langle z, x\rangle \mid \|x\| \leqslant 1\right\}.$$

Hölder's inequality (see Exercise 4) shows that the ℓ_p and ℓ_q norms, defined by

$$\|x\|_p = \left(\sum_{j=1}^n |x_j|^p\right)^{\frac{1}{p}}$$

(and as the limit $\|x\|_\infty = \max_j |x_j|$) are dual to one another, where $1/p + 1/q = 1$ and $p, q \in [1, \infty]$. Throughout, we will assume that $\|x\|_2 = \sqrt{\langle x, x\rangle}$ is the norm defined by the inner product $\langle \cdot, \cdot\rangle$.

We also require notation related to sets. For a sequence of vectors $v_1, v_2, v_3, \ldots$, we let (v_n) denote the entire sequence. Given sets A and B, we let $A \subset B$ to denote that A is a subset (possibly equal to) B, and $A \subsetneq B$ to mean that A is a strict subset of B. The notation $\mathop{\rm cl} A$ denotes the closure of A, while $\mathop{\rm int} A$ denotes the interior of the set A. For a function f, the set $\mathop{\rm dom} f$ is its domain. If $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex, we let $\mathop{\rm dom} f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$.

2. Basic Convex Analysis

Lecture Summary: In this lecture, we will outline several standard facts from convex analysis, the study of the mathematical properties of convex functions and sets. For the most part, our analysis and results will all be with the aim of setting the necessary background for understanding first-order convex optimization methods, though some of the results we state will be quite general.

2.1. Introduction and Definitions This set of lecture notes considers convex optimization problems, numerical optimization problems of the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in C, \end{aligned} \tag{2.1.1}$$

where f is a convex function and C is a convex set. While we will consider tools to solve these types of optimization problems presently, this first lecture is concerned most with the analytic tools and background that underlies solution methods for these problems.

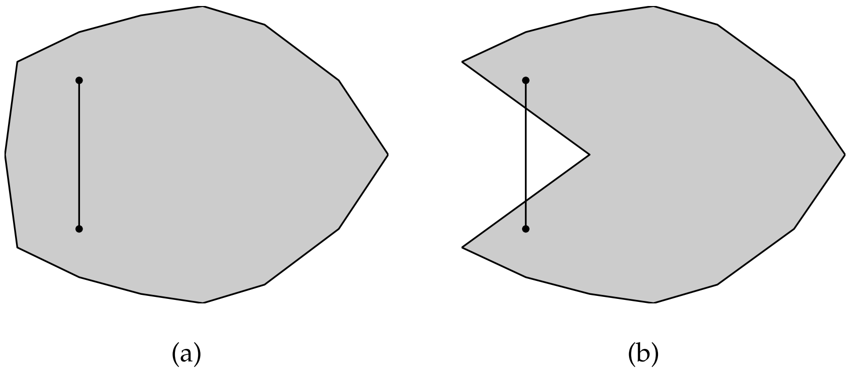

Figure 2.1.2. (a) A convex set (b) A non-convex set.

The starting point for any study of convex functions is the definition and study of convex sets, which are intimately related to convex functions. To that end, we recall that a set $C \subset \mathbb{R}^n$ is *convex* if for all $x, y \in C$,

$$\lambda x + (1 - \lambda)y \in C \text{ for } \lambda \in [0, 1].$$

See Figure 2.1.2.

A convex function is similarly defined: a function $f : \mathbb{R}^n \to (-\infty, \infty]$ is *convex* if for all $x, y \in \operatorname{dom} f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\}$

$$f(\lambda x + (1 - \lambda)y) \leqslant \lambda f(x) + (1 - \lambda)f(y) \text{ for } \lambda \in [0, 1].$$

The epigraph of a function is defined as

$$\operatorname{epi} f := \{(x, t) : f(x) \leqslant t\},$$

and by inspection, a function is convex if and only if its epigraph is a convex set. A convex function f is closed if its epigraph is a closed set; continuous convex functions are always closed. We will assume throughout that any convex function we deal with is closed. See Figure 2.1.3 for graphical representations of these ideas, which make clear that the epigraph is indeed a convex set.

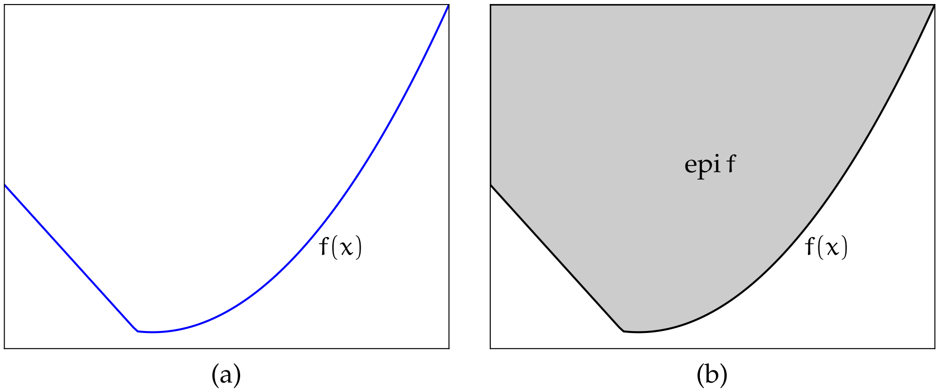

FIGURE 2.1.3. (a) The convex function $f(x) = \max\{x^2, -2x - .2\}$ and (b) its epigraph, which is a convex set.

One may ask why, precisely, we focus convex functions. In short, as Rockafellar [49] notes, convex optimization problems are the clearest dividing line between numerical problems that are efficiently solvable, often by iterative methods, and numerical problems for which we have no hope. We give one simple result in this direction first:

Observation. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and x be a local minimum of f (respectively a local minimum over a convex set C). Then x is a global minimum of f (resp. a global minimum of f over C).*

To see this, note that if x is a local minimum then for any $y \in C$, we have for small enough $t > 0$ that

$$f(x) \leqslant f(x + t(y - x)) \text{ or } 0 \leqslant \frac{f(x + t(y - x)) - f(x)}{t}.$$

We now use the *criterion of increasing slopes*, that is, for any convex function f the function

$$t \mapsto \frac{f(x + tu) - f(x)}{t} \tag{2.1.5}$$

is *increasing* in $t > 0$. (See Fig. 2.1.4.) Indeed, let $0 \leqslant t_1 \leqslant t_2$. Then

$$\begin{aligned}\frac{f(x + t_1 u) - f(x)}{t_1} &= \frac{t_2}{t_1} \frac{f(x + t_2(t_1/t_2)u) - f(x)}{t_2} \\ &= \frac{t_2}{t_1} \frac{f((1 - t_1/t_2)x + (t_1/t_2)(x + t_2 u)) - f(x)}{t_2}\end{aligned}$$

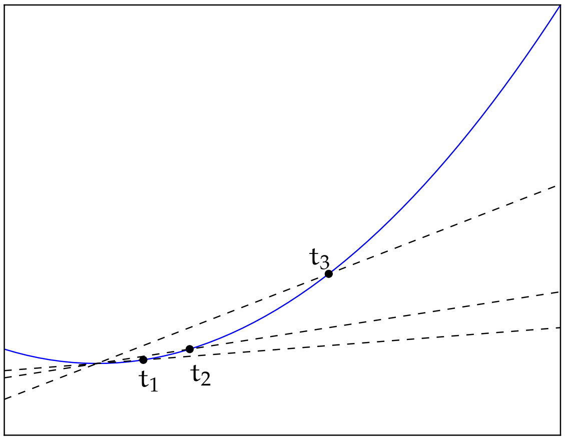

FIGURE 2.1.4. The slopes $\frac{f(x+t)-f(x)}{t}$ increase, with $t_1 < t_2 < t_3$.

$$\leqslant \frac{t_2}{t_1} \frac{(1 - t_1/t_2)f(x) + (t_1/t_2)f(x + t_2u) - f(x)}{t_2} = \frac{f(x + t_2u) - f(x)}{t_2}.$$

In particular, because $0 \leqslant f(x + t(y - x))$ for small enough $t > 0$, we see that for all $t > 0$ we have

$$0 \leqslant \frac{f(x + t(y - x)) - f(x)}{t} \text{ or } f(x) \leqslant \inf_{t \geqslant 0} f(x + t(y - x)) \leqslant f(y)$$

for all $y \in C$.

Most of the results herein apply in general Hilbert (complete inner product) spaces, and many of our proofs will not require anything particular about finite dimensional spaces, but for simplicity we use $\mathbb{R}^n$ as the underlying space on which all functions and sets are defined.[1] While we present all proofs in the chapter, we try to provide geometric intuition that will aid a working knowledge of the results, which we believe is the most important.

2.2. Properties of Convex Sets Convex sets enjoy a number of very nice properties that allow efficient and elegant descriptions of the sets, as well as providing a number of nice properties concerning their separation from one another. To that end, in this section, we give several fundamental properties on separating and supporting hyperplanes for convex sets. The results here begin by showing that there is a unique (Euclidean) projection to any convex set C, then use this fact to show that whenever a point is not contained in a set, it can be separated from the set by a hyperplane. This result can be extended to show separation of convex sets from one another and that points in the boundary of a convex set have a hyperplane tangent to the convex set running through them. We leverage these results in the sequel by making connections of supporting hyperplanes to

[1] The generality of Hilbert, or even Banach, spaces in convex analysis is seldom needed. Readers familiar with arguments in these spaces will, however, note that the proofs can generally be extended to infinite dimensional spaces in reasonably straightforward ways.

epigraphs and gradients, results that in turn find many applications in the design of optimization algorithms as well as optimality certificates.

A few basic properties We list a few simple properties that convex sets have, which are evident from their definitions. First, if C_α are convex sets for each $\alpha \in \mathcal{A}$, where $\mathcal{A}$ is an arbitrary index set, then the intersection

$$C = \bigcap_{\alpha \in \mathcal{A}} C_\alpha$$

is also convex. Additionally, convex sets are closed under scalar multiplication: if $\alpha \in \mathbb{R}$ and C is convex, then

$$\alpha C := \{\alpha x : x \in C\}$$

is evidently convex. The Minkowski sum of two convex sets is defined by

$$C_1 + C_2 := \{x_1 + x_2 : x_1 \in C_1, x_2 \in C_2\},$$

and is also convex. To see this, note that if $x_i, y_i \in C_i$, then

$$\lambda(x_1 + x_2) + (1-\lambda)(y_1 + y_2) = \underbrace{\lambda x_1 + (1-\lambda) y_1}_{\in C_1} + \underbrace{\lambda x_2 + (1-\lambda) y_2}_{\in C_2} \in C_1 + C_2.$$

In particular, convex sets are closed under all linear combination: if $\alpha \in \mathbb{R}^m$, then $C = \sum_{i=1}^m \alpha_i C_i$ is also convex.

We also define the convex hull of a set of points $x_1, \ldots, x_m \in \mathbb{R}^n$ by

$$\mathrm{Conv}\{x_1, \ldots, x_m\} = \left\{ \sum_{i=1}^m \lambda_i x_i : \lambda_i \geqslant 0, \sum_{i=1}^m \lambda_i = 1 \right\}.$$

This set is clearly a convex set.

Projections We now turn to a discussion of orthogonal projection onto a convex set, which will allow us to develop a number of separation properties and alternate characterizations of convex sets. See Figure 2.2.5 for a geometric view of projection. We begin by stating a classical result about the projection of zero onto a convex set.

Theorem 2.2.1 (Projection of zero)**.** *Let C be a closed convex set not containing the origin 0. Then there is a unique point $x_C \in C$ such that $\|x_C\|_2 = \inf_{x \in C} \|x\|_2$. Moreover, $\|x_C\|_2 = \inf_{x \in C} \|x\|_2$ if and only if*

$$\langle x_C, y - x_C \rangle \geqslant 0 \tag{2.2.2}$$

for all $y \in C$.

Proof. The key to the proof is the following parallelogram identity, which holds in any inner product space: for any x, y,

$$\frac{1}{2}\|x - y\|_2^2 + \frac{1}{2}\|x + y\|_2^2 = \|x\|_2^2 + \|y\|_2^2. \tag{2.2.3}$$

Define $M := \inf_{x \in C} \|x\|_2$. Now, let $(x_n) \subset C$ be a sequence of points in C such that $\|x_n\|_2 \to M$ as $n \to \infty$. By the parallelogram identity (2.2.3), for any $n, m \in \mathbb{N}$, we have

$$\frac{1}{2}\left\|x_n - x_m\right\|_2^2 = \left\|x_n\right\|_2^2 + \left\|x_m\right\|_2^2 - \frac{1}{2}\left\|x_n + x_m\right\|_2^2.$$

Fix $\epsilon > 0$, and choose $N \in \mathbb{N}$ such that $n \geqslant N$ implies that $\left\|x_n\right\|_2^2 \leqslant M^2 + \epsilon$. Then for any $m, n \geqslant N$, we have

$$\frac{1}{2}\left\|x_n - x_m\right\|_2^2 \leqslant 2M^2 + 2\epsilon - \frac{1}{2}\left\|x_n + x_m\right\|_2^2. \tag{2.2.4}$$

Now we use the convexity of the set C. We have $\frac{1}{2}x_n + \frac{1}{2}x_m \in C$ for any n, m, which implies

$$\frac{1}{2}\left\|x_n + x_m\right\|_2^2 = 2\left\|\frac{1}{2}x_n + \frac{1}{2}x_m\right\|_2^2 \geqslant 2M^2$$

by definition of M. Using the above inequality in the bound (2.2.4), we see that

$$\frac{1}{2}\left\|x_n - x_m\right\|_2^2 \leqslant 2M^2 + 2\epsilon - 2M^2 = 2\epsilon.$$

In particular, $\left\|x_n - x_m\right\|_2 \leqslant 2\sqrt{\epsilon}$; since ϵ was arbitrary, (x_n) forms a Cauchy sequence and so must converge to a point x_C. The continuity of the norm $\left\|\cdot\right\|_2$ implies that $\left\|x_C\right\|_2 = \inf_{x \in C}\left\|x\right\|_2$, and the fact that C is closed implies that $x_C \in C$.

Now we show the inequality (2.2.2) holds if and only if x_C is the projection of the origin 0 onto C. Suppose that inequality (2.2.2) holds. Then

$$\left\|x_C\right\|_2^2 = \left\langle x_C, x_C\right\rangle \leqslant \left\langle x_C, y\right\rangle \leqslant \left\|x_C\right\|_2\left\|y\right\|_2,$$

the last inequality following from the Cauchy-Schwartz inequality. Dividing each side by $\left\|x_C\right\|_2$ implies that $\left\|x_C\right\|_2 \leqslant \left\|y\right\|_2$ for all $y \in C$. For the converse, let x_C minimize $\left\|x\right\|_2$ over C. Then for any $t \in [0, 1]$ and any $y \in C$, we have

$$\left\|x_C\right\|_2^2 \leqslant \left\|(1-t)x_C + ty\right\|_2^2 = \left\|x_C + t(y - x_C)\right\|_2^2 = \left\|x_C\right\|_2^2 + 2t\left\langle x_C, y - x_C\right\rangle + t^2\left\|y - x_C\right\|_2^2.$$

Subtracting $\left\|x_C\right\|_2^2$ and $t^2\left\|y - x_C\right\|_2^2$ from both sides of the above inequality, we have

$$-t^2\left\|y - x_C\right\|_2^2 \leqslant 2t\left\langle x_C, y - x_C\right\rangle.$$

Dividing both sides of the above inequality by $2t$, we have

$$-\frac{t}{2}\left\|y - x_C\right\|_2^2 \leqslant \left\langle x_C, y - x_C\right\rangle$$

for all $t \in (0, 1]$. Letting $t \downarrow 0$ gives the desired inequality. □

With this theorem in place, a simple shift gives a characterization of more general projections onto convex sets.

Corollary 2.2.6 (Projection onto convex sets)**.** *Let* C *be a closed convex set and* $x \in \mathbb{R}^n$*. Then there is a unique point* $\pi_C(x)$*, called the* projection of x onto C, *such that* $\left\|x - \pi_C(x)\right\|_2 = \inf_{y \in C}\left\|x - y\right\|_2$*, that is,* $\pi_C(x) = \operatorname{argmin}_{y \in C}\left\|y - x\right\|_2^2$*. The projection is characterized by the inequality*

$$\left\langle \pi_C(x) - x, y - \pi_C(x)\right\rangle \geqslant 0 \tag{2.2.7}$$

for all $y \in C$.

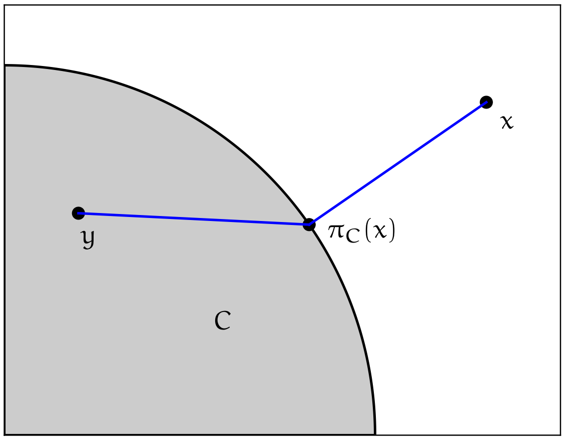

Figure 2.2.5. Projection of the point x onto the set C (with projection $\pi_C(x)$), exhibiting $\langle x - \pi_C(x), y - \pi_C(x)\rangle \leqslant 0$.

Proof. When $x \in C$, the statement is clear. For $x \notin C$, the corollary simply follows by considering the set $C' = C - x$, then using Theorem 2.2.1 applied to the recentered set. □

Corollary 2.2.8 (Non-expansive projections). *Projections onto convex sets are non-expansive, in particular,*

$$\|\pi_C(x) - y\|_2 \leqslant \|x - y\|_2$$

for any $x \in \mathbb{R}^n$ *and* $y \in C$.

Proof. When $x \in C$, the inequality is clear, so assume that $x \notin C$. Now use inequality (2.2.7) from the previous corollary. By adding and subtracting y in the inner product, we have

$$\begin{aligned} 0 &\leqslant \langle \pi_C(x) - x, y - \pi_C(x)\rangle \\ &= \langle \pi_C(x) - y + y - x, y - \pi_C(x)\rangle \\ &= -\|\pi_C(x) - y\|_2^2 + \langle y - x, y - \pi_C(x)\rangle \end{aligned}$$

We rearrange the above and then use the Cauchy-Schwartz or Hölder's inequality, which gives

$$\|\pi_C(x) - y\|_2^2 \leqslant \langle y - x, y - \pi_C(x)\rangle \leqslant \|y - x\|_2 \, \|y - \pi_C(x)\|_2 .$$

Now divide both sides by $\|\pi_C(x) - y\|_2$. □

Separation Properties Projections are important not just because of their existence, but because they also guarantee that convex sets can be described by halfplanes that contain them as well as that any two convex sets are separated by hyperplanes. Moreover, the separation can be strict if one of the sets is compact.

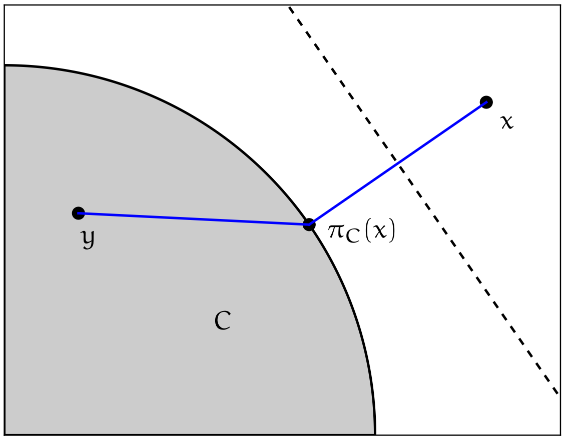

Figure 2.2.9. Separation of the point x from the set C by the vector $v = x - \pi_C(x)$.

Proposition 2.2.10 (Strict separation of points)**.** *Let C be a closed convex set. Given any point $x \notin C$, there is a vector v such that*

$$\langle v, x\rangle > \sup_{y \in C} \langle v, y\rangle \tag{2.2.11}$$

Moreover, we can take the vector $v = x - \pi_C(x)$, and $\langle v, x\rangle \geqslant \sup_{y \in C} \langle v, y\rangle + \|v\|_2^2$. See Figure 2.2.9.

Proof. Indeed, since $x \notin C$, we have $x - \pi_C(x) \neq 0$. By setting $v = x - \pi_C(x)$, we have from the characterization (2.2.7) that

$$0 \geqslant \langle v, y - \pi_C(x)\rangle = \langle v, y - x + x - \pi_C(x)\rangle = \langle v, y - x + v\rangle = \langle v, y - x\rangle + \|v\|_2^2 .$$

In particular, we see that $\langle v, x\rangle \geqslant \langle v, y\rangle + \|v\|^2$ for all $y \in C$. □

Proposition 2.2.12 (Strict separation of convex sets)**.** *Let C_1, C_2 be closed convex sets, with C_2 compact. Then there is a vector v such that*

$$\inf_{x \in C_1} \langle v, x\rangle > \sup_{x \in C_2} \langle v, x\rangle .$$

Proof. The set $C = C_1 - C_2$ is convex and closed.[2] Moreover, we have $0 \notin C$, so that there is a vector v such that $0 < \inf_{z \in C} \langle v, z\rangle$ by Proposition 2.2.10. Thus we have

$$0 < \inf_{z \in C_1 - C_2} \langle v, z\rangle = \inf_{x \in C_1} \langle v, x\rangle - \sup_{x \in C_2} \langle v, x\rangle ,$$

which is our desired result. □

[2] If C_1 is closed and C_2 is compact, then $C_1 + C_2$ is closed. Indeed, let $z_n = x_n + y_n$ be a convergent sequence of points (say $z_n \to z$) with $z_n \in C_1 + C_2$. We claim that $z \in C_1 + C_2$. Indeed, passing to a subsequence if necessary, we may assume $y_n \to y$. Then on the subsequence, we have $x_n = z_n - y_n \to z - y$, so that x_n is convergent and necessarily converges to a point $x \in C_1$.

We can also investigate the existence of hyperplanes that support the convex set C, meaning that they touch only its boundary and never enter its interior. Such hyperplanes—and the halfspaces associated with them—provide alternate descriptions of convex sets and functions. See Figure 2.2.13.

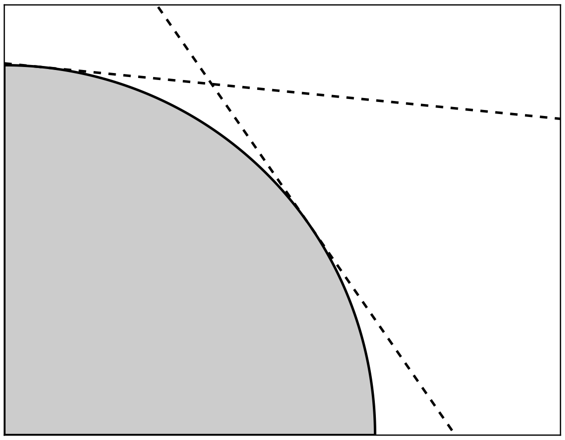

FIGURE 2.2.13. Supporting hyperplanes to a convex set.

Theorem 2.2.14 (Supporting hyperplanes)**.** *Let C be a closed convex set and $x \in \operatorname{bd} C$, the boundary of C. Then there exists a vector $v \neq 0$ supporting C at x, that is,*

$$\langle v, x\rangle \geqslant \langle v, y\rangle \text{ for all } y \in C. \tag{2.2.15}$$

Proof. Let (x_n) be a sequence of points approaching x from outside C, that is, $x_n \notin C$ for any n, but $x_n \to x$. For each n, we can take $s_n = x_n - \pi_C(x_n)$ and define $v_n = s_n / \|s_n\|_2$. Then (v_n) is a sequence satisfying $\langle v_n, x\rangle > \langle v_n, y\rangle$ for all $y \in C$, and since $\|v_n\|_2 = 1$, the sequence (v_n) belongs to the compact set $\{v : \|v\|_2 \leqslant 1\}$.[3] Passing to a subsequence if necessary, it is clear that there is a vector v such that $v_n \to v$, and we have $\langle v, x\rangle \geqslant \langle v, y\rangle$ for all $y \in C$. □

Theorem 2.2.16 (Halfspace intersections)**.** *Let $C \subsetneq \mathbb{R}^n$ be a closed convex set. Then C is the intersection of all the spaces containing it; moreover,*

$$C = \bigcap_{x \in \operatorname{bd} C} H_x \tag{2.2.17}$$

where H_x denotes the intersection of the halfspaces contained in hyperplanes supporting C at x.

Proof. It is clear that $C \subseteq \bigcap_{x \in \operatorname{bd} C} H_x$. Indeed, let $h_x \neq 0$ be a hyperplane supporting to C at $x \in \operatorname{bd} C$ and consider $H_x = \{y : \langle h_x, x\rangle \geqslant \langle h_x, y\rangle\}$. By Theorem 2.2.14 we see that $H_x \supseteq C$.

[3] In a general Hilbert space, this set is actually weakly compact by Alaoglu's theorem. However, in a weakly compact set, any sequence has a weakly convergent subsequence, that is, there exists a subsequence $n(m)$ and vector v such that $\langle v_{n(m)}, y\rangle \to \langle v, y\rangle$ for all y.

Now we show the other inclusion: $\bigcap_{x \in \mathrm{bd}\, C} H_x \subseteq C$. Suppose for the sake of contradiction that $z \in \bigcap_{x \in \mathrm{bd}\, C} H_x$ satisfies $z \notin C$. We will construct a hyperplane supporting C that separates z from C, which will be a contradiction to our supposition. Since C is closed, the projection of $\pi_C(z)$ of z onto C satisfies $\langle z - \pi_C(z), z \rangle > \sup_{y \in C} \langle z - \pi_C(z), y \rangle$ by Proposition 2.2.10. In particular, defining $v_z = z - \pi_C(z)$, the hyperplane $\{y : \langle v_z, y \rangle = \langle v_z, \pi_C(z) \rangle\}$ is supporting to C at the point $\pi_C(z)$ (Corollary 2.2.6) and the halfspace $\{y : \langle v_z, y \rangle \leqslant \langle v_z, \pi_C(z) \rangle\}$ does not contain z but does contain C. This contradicts the assumption that $z \in \bigcap_{x \in \mathrm{bd}\, C} H_x$. □

As a not too immediate consequence of Theorem 2.2.16 we obtain the following characterization of a convex function as the supremum of all affine functions that minorize the function (that is, affine functions that are everywhere less than or equal to the original function). This is intuitive: if f is a closed convex function, meaning that epi f is closed, then epi f is the intersection of all the halfspaces containing it. The challenge is showing that we may restrict this intersection to non-vertical halfspaces. See Figure 2.2.18.

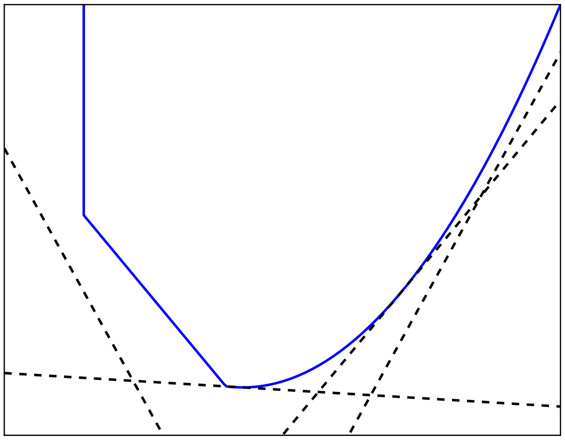

FIGURE 2.2.18. The function f (solid blue line) and affine underestimators (dotted lines).

Corollary 2.2.19. *Let* f *be a closed convex function that is not identically* $-\infty$. *Then*

$$f(x) = \sup_{v \in \mathbb{R}^n, b \in \mathbb{R}} \{\langle v, x \rangle + b : f(y) \geqslant b + \langle v, y \rangle \text{ for all } y \in \mathbb{R}^n\}.$$

Proof. First, we note that epi f is closed by definition. Moreover, we know that we can write

$$\mathrm{epi}\, f = \cap \{H : H \supset \mathrm{epi}\, f\},$$

where H denotes a halfspace. More specifically, we may index each halfspace by $(v, a, c) \in \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$, and we have $H_{v,a,c} = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \langle v, x \rangle + at \leqslant c\}$. Now, because $H \supset \mathrm{epi}\, f$, we must be able to take $t \to \infty$ so that $a \leqslant 0$. If $a < 0$, we may divide by $|a|$ and assume without loss of generality that $a = -1$, while otherwise $a = 0$. So if we let

$$\mathcal{H}_1 := \left\{(v, c) : H_{v,-1,c} \supset \operatorname{epi} f\right\} \text{ and } \mathcal{H}_0 := \{(v, c) : H_{v,0,c} \supset \operatorname{epi} f\}.$$

then

$$\operatorname{epi} f = \bigcap_{(v,c)\in\mathcal{H}_1} H_{v,-1,c} \cap \bigcap_{(v,c)\in\mathcal{H}_0} H_{v,0,c}.$$

We would like to show that $\operatorname{epi} f = \cap_{(v,c)\in\mathcal{H}_1} H_{v,-1,c}$, as the set $H_{v,0,c}$ is a vertical hyperplane separating the domain of f, dom f, from the rest of the space.

To that end, we show that for any $(v_1, c_1) \in \mathcal{H}_1$ and $(v_0, c_0) \in \mathcal{H}_0$, then

$$H := \bigcap_{\lambda\geqslant 0} H_{v_1+\lambda v_0,-1,c_1+\lambda c_1} = H_{v_1,-1,c_1} \cap H_{v_0,0,c_0}.$$

Indeed, suppose that $(x, t) \in H_{v_1,-1,c_1} \cap H_{v_0,0,c_0}$. Then

$$\langle v_1, x\rangle - t \leqslant c_1 \text{ and } \lambda \langle v_0, x\rangle \leqslant \lambda c_0 \text{ for all } \lambda \geqslant 0.$$

Summing these, we have

$$\langle v_1 + \lambda v_0, x\rangle - t \leqslant c_1 + \lambda c_0 \text{ for all } \lambda \geqslant 0, \tag{2.2.20}$$

or $(x, t) \in H$. Conversely, if $(x, t) \in H$ then inequality (2.2.20) holds, so that taking $\lambda \to \infty$ we have $\langle v_0, x\rangle \leqslant c_0$, while taking $\lambda = 0$ we have $\langle v_1, x\rangle - t \leqslant c_1$.

Noting that $H \in \{H_{v,-1,c} : (v, c) \in \mathcal{H}_1\}$, we see that

$$\operatorname{epi} f = \bigcap_{(v,c)\in\mathcal{H}_1} H_{v,-1,c} = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} : \langle v, x\rangle - t \leqslant c \text{ for all } (v, c) \in \mathcal{H}_1\}.$$

This is equivalent to the claim in the corollary. □

2.3. Continuity and Local Differentiability of Convex Functions Here we discuss several important results concerning convex functions in finite dimensions. We will see that assuming that a function f is convex is quite strong. In fact, we will see the (intuitive if one pictures a convex function) facts that f is continuous, has a directional derivatve everywhere, and in fact is locally Lipschitz. We prove the first two results here on continuity in Appendix A.1, as they are not fully necessary for our development.

We begin with the fact that if f is defined on a compact domain, then f has an upper bound. The first step in this direction is to argue that this holds for ℓ_1 balls, which can be proved by a simple argument with the definition of convexity.

Lemma 2.3.1. *Let* f *be convex and defined on the* ℓ_1 *ball in* n *dimensions:* $B_1 = \{x \in \mathbb{R}^n : \|x\|_1 \leqslant 1\}$. *Then there exist* $-\infty < m \leqslant M < \infty$ *such that* $m \leqslant f(x) \leqslant M$ *for all* $x \in B_1$.

We provide a proof of this lemma, as well as the coming theorem, in Appendix A.1, as they are not central to our development, relying on a few results in the sequel. The coming theorem makes use of the above lemma to show that on compact domains, convex functions are Lipschitz continuous. The proof of the theorem begins by showing that if a convex function is bounded in some set, then it is Lipschitz continuous in the set, then using Lemma 2.3.1 we can show that on compact sets f is indeed bounded.

Theorem 2.3.2. *Let f be convex and defined on a set C with non-empty interior. Let $B \subseteq \operatorname{int} C$ be compact. Then there is a constant L such that $|f(x) - f(y)| \leqslant L \|x - y\|$ on B, that is, f is L-Lipschitz continuous on B.*

The last result, which we make strong use of in the next section, concerns the existence of directional derivatives for convex functions.

Definition 2.3.3. The *directional derivative* of a function f at a point x in the direction u is

$$f'(x; u) := \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[f(x + \alpha u) - f(x)\right].$$

This definition makes sense by our earlier arguments that convex functions have increasing slopes (recall expression (2.1.5)). To see that the above definition makes sense, we restrict our attention to $x \in \operatorname{int} \operatorname{dom} f$, so that we can approach x from all directions. By taking $u = y - x$ for any $y \in \operatorname{dom} f$,

$$f(x + \alpha(y - x)) = f((1 - \alpha)x + \alpha y) \leqslant (1 - \alpha) f(x) + \alpha f(y)$$

so that

$$\frac{1}{\alpha} \left[f(x + \alpha(y - x)) - f(x)\right] \leqslant \frac{1}{\alpha} \left[\alpha f(y) - \alpha f(x)\right] = f(y) - f(x) = f(x + u) - f(x).$$

We also know from Theorem 2.3.2 that f is locally Lipschitz, so for small enough α there exists some L such that $f(x + \alpha u) \geqslant f(x) - L\alpha \|u\|$, and thus $f'(x; u) \geqslant -L \|u\|$. Further, an argument by convexity (the criterion (2.1.5) of increasing slopes) shows that the function

$$\alpha \mapsto \frac{1}{\alpha}[f(x + \alpha u) - f(x)]$$

is increasing, so we can replace the limit in the definition of $f'(x; u)$ with an infimum over $\alpha > 0$, that is, $f'(x; u) = \inf_{\alpha > 0} \frac{1}{\alpha}[f(x + \alpha u) - f(x)]$. Noting that if x is on the boundary of $\operatorname{dom} f$ and $x + \alpha u \notin \operatorname{dom} f$ for any $\alpha > 0$, then $f'(x; u) = +\infty$, we have proved the following theorem.

Theorem 2.3.4. *For convex f, at any point $x \in \operatorname{dom} f$ and for any u, the directional derivative $f'(x; u)$ exists and is*

$$f'(x; u) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} \left[f(x + \alpha u) - f(x)\right] = \inf_{\alpha > 0} \frac{1}{\alpha} \left[f(x + \alpha u) - f(x)\right].$$

If $x \in \operatorname{int} \operatorname{dom} f$, there exists a constant $L < \infty$ such that $|f'(x; u)| \leqslant L \|u\|$ for any $u \in \mathbb{R}^n$. If f is Lipschitz continuous with respect to the norm $\|\cdot\|$, we can take L to be the Lipschitz constant of f.

Lastly, we state a well-known condition that is equivalent to convexity. This is inuitive: if a function is bowl-shaped, it should have positive second derivatives.

Theorem 2.3.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then f is convex if and only if $\nabla^2 f(x) \succeq 0$ for all x, that is, $\nabla^2 f(x)$ is positive semidefinite.*

Proof. We may essentially reduce the argument to one-dimensional problems, because if f is twice continuously differentiable, then for each $v \in \mathbb{R}^n$ we may define $h_v : \mathbb{R} \to \mathbb{R}$ by

$$h_v(t) = f(x + tv),$$

and f is convex if and only if h_v is convex for each v (because convexity is a property only of lines, by definition). Moreover, we have

$$h_v''(0) = v^\top \nabla^2 f(x) v,$$

and $\nabla^2 f(x) \succeq 0$ if and only if $h_v''(0) \geqslant 0$ for all v.

Thus, with no loss of generality, we assume $n = 1$ and show that f is convex if and only if $f''(x) \geqslant 0$. First, suppose that $f''(x) \geqslant 0$ for all x. Then using that

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}(y - x)^2 f''(\widetilde{x})$$

for some $\widetilde{x}$ between x and y, we have that $f(y) \geqslant f(x) + f'(x)(y - x)$ for all x, y. Let $\lambda \in [0, 1]$. Then we have

$$\begin{aligned}
&f(y) \geqslant f(\lambda x + (1-\lambda)y) + \lambda f'(\lambda x + (1-\lambda)y)(y - x) \text{ and} \\
&f(x) \geqslant f(\lambda x + (1-\lambda)y) + (1-\lambda) f'(\lambda x + (1-\lambda)y)(x - y).
\end{aligned}$$

Multiplying the first equation by $1 - \lambda$ and the second by λ, then adding, we obtain

$$(1-\lambda)f(y) + \lambda f(x) \geqslant (1-\lambda)f(\lambda x + (1-\lambda)y) + \lambda f(\lambda x + (1-\lambda)y) = f(\lambda x + (1-\lambda)y),$$

that is, f is convex.

For the converse, let $\delta > 0$ and define $x_1 = x + \delta > x > x - \delta = x_0$. Then we have $x_1 - x_0 = 2\delta$, and

$$f(x_1) = f(x) + f'(x)\delta + 2\delta^2 f''(\widetilde{x}_1) \text{ and } f(x_0) = f(x) - f'(x)\delta + 2\delta^2 f''(\widetilde{x}_0)$$

for $\widetilde{x}_1, \widetilde{x}_0 \in [x - \delta, x + \delta]$. Adding these quantities and defining $c_\delta = f(x_1) + f(x_0) - 2f(x) \geqslant 0$ (the last inequality by convexity), we have

$$c_\delta = 2\delta^2 [f''(\widetilde{x}_1) + f''(\widetilde{x}_0)].$$

By continuity, we have $f''(\widetilde{x}_i) \to f''(x)$ as $\delta \to 0$, and as $c_\delta / 2\delta^2 \geqslant 0$ for all $\delta > 0$, we must have

$$2f''(x) = \limsup_{\delta \to 0} \{f''(\widetilde{x}_1) + f''(\widetilde{x}_0)\} = \limsup_{\delta \to 0} \frac{c_\delta}{2\delta^2} \geqslant 0.$$

This gives the result. □

2.4. Subgradients and Optimality Conditions The subgradient set of a function f at a point $x \in \operatorname{dom} f$ is defined as follows:

$$\partial f(x) := \{g : f(y) \geqslant f(x) + \langle g, y - x \rangle \text{ for all } y\}. \tag{2.4.1}$$

Intuitively, since a function is convex if and only if $\operatorname{epi} f$ is convex, the subgradient set ∂f should be non-empty and consist of supporting hyperplanes to $\operatorname{epi} f$. That is, f should always have global linear underestimators of itself. When a function f is convex, the subgradient generalizes the derivative of f (which is a global linear underestimator of f when f is differentiable), and is also intimately related to optimality conditions for convex minimization.

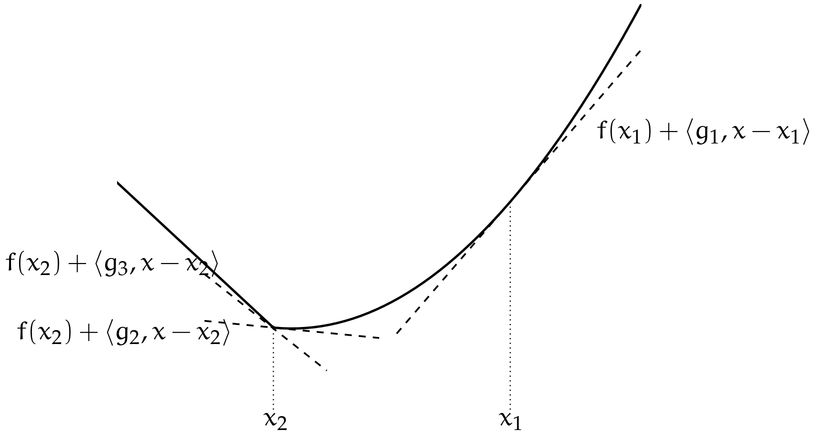
$f(x_1) + \langle g_1, x - x_1 \rangle$
$f(x_2) + \langle g_3, x - x_2 \rangle$
$f(x_2) + \langle g_2, x - x_2 \rangle$
x_2
x_1

FIGURE 2.4.2. Subgradients of a convex function. At the point x_1, the subgradient g_1 is the gradient. At the point x_2, there are multiple subgradients, because the function is non-differentiable. We show the linear functions given by $g_2, g_3 \in \partial f(x_2)$.

Existence and characterizations of subgradients Our first theorem guarantees that the subdifferential set is non-empty.

Theorem 2.4.3. *Let* $x \in \operatorname{int}\operatorname{dom} f$. *Then* $\partial f(x)$ *is nonempty, closed convex, and compact.*

Proof. The fact that $\partial f(x)$ is closed and convex is straightforward. Indeed, all we need to see this is to recognize that

$$\partial f(x) = \bigcap_z \{g : f(z) \geqslant f(x) + \langle g, z - x \rangle\}$$

which is an intersection of half-spaces, which are all closed and convex.

Now we need to show that $\partial f(x) \neq \emptyset$. This will essentially follow from the following fact: the set $\operatorname{epi} f$ has a supporting hyperplane at the point $(x, f(x))$. Indeed, from Theorem 2.2.14, we know that there exist a vector v and scalar b such that

$$\langle v, x \rangle + b f(x) \geqslant \langle v, y \rangle + bt$$

for all $(y, t) \in \operatorname{epi} f$ (that is, y and t such that $f(y) \leqslant t$). Rearranging slightly, we have

$$\langle v, x - y \rangle \geqslant b(t - f(x))$$

and setting $y = x$ shows that $b \leqslant 0$. This is close to what we desire, since if $b < 0$ we set $t = f(y)$ and see that

$$-bf(y) \geqslant -bf(x) + \langle v, y - x \rangle \text{ or } f(y) \geqslant f(x) - \left\langle \frac{v}{b}, y - x \right\rangle$$

for all y, by dividing both sides by $-b$. In particular, $-v/b$ is a subgradient. Thus, suppose for the sake of contradiction that $b = 0$. In this case, we have $\langle v, x - y \rangle \geqslant 0$ for all $y \in \operatorname{dom} f$, but we assumed that $x \in \operatorname{int} \operatorname{dom} f$, so for small enough $\epsilon > 0$, we can set $y = x + \epsilon v$. This would imply that $\langle v, x - y \rangle = -\epsilon \langle v, v \rangle = 0$, i.e. $v = 0$, contradicting the fact that at least one of v and b must be non-zero.

For the compactness of $\partial f(x)$, we use Lemma 2.3.1, which implies that f is bounded in an ℓ_1-ball around of x. As $x \in \operatorname{int} \operatorname{dom} f$ by assumption, there is some $\epsilon > 0$ such that $x + \epsilon B \subset \operatorname{int} \operatorname{dom} f$ for the ℓ_1-ball $B = \{v : \|v\|_1 \leqslant 1\}$. Lemma 2.3.1 implies that $\sup_{v \in B} f(x + \epsilon v) = M < \infty$ for some M, so we have $M \geqslant f(x + \epsilon v) \geqslant f(x) + \epsilon \langle g, v \rangle$ for all $v \in B$ and $g \in \partial f(x)$, or $\|g\|_\infty \leqslant (M - f(x))/\epsilon$. Thus $\partial f(x)$ is closed and bounded, hence compact. □

The next two results require a few auxiliary results related to the directional derivative of a convex function. The reason for this is that both require connecting the local properties of the convex function f with the sub-differential $\partial f(x)$, which is difficult in general since $\partial f(x)$ can consist of multiple vectors. However, by looking at directional derivatives, we can accomplish what we desire. The connection between a directional derivative and the subdifferential is contained in the next two lemmas.

Lemma 2.4.4. *An equivalent characterization of the subdifferential* $\partial f(x)$ *of* f *at* x *is*

$$\partial f(x) = \left\{ g : \langle g, u \rangle \leqslant f'(x; u) \text{ for all } u \right\}. \tag{2.4.5}$$

Proof. Denote the set on the right hand side of the equality (2.4.5) by $S = \{g : \langle g, u \rangle \leqslant f'(x; u)\}$, and let $g \in S$. By the increasing slopes condition, we have

$$\langle g, u \rangle \leqslant f'(x; u) \leqslant \frac{f(x + \alpha u) - f(x)}{\alpha}$$

for all u and $\alpha > 0$; in particular, by taking $\alpha = 1$ and $u = y - x$, we have the standard subgradient inequality that $f(x) + \langle g, y - x \rangle \leqslant f(y)$. So if $g \in S$, then $g \in \partial f(x)$. Conversely, for any $g \in \partial f(x)$, the definition of a subgradient implies that

$$f(x + \alpha u) \geqslant f(x) + \langle g, x + \alpha u - x \rangle = f(x) + \alpha \langle g, u \rangle.$$

Subtracting $f(x)$ from both sides and dividing by α gives that

$$\frac{1}{\alpha} [f(x + \alpha u) - f(x)] \geqslant \sup_{g \in \partial f(x)} \langle g, u \rangle$$

for all $\alpha > 0$; in particular, $g \in S$. □

The representation (2.4.5) gives another proof that $\partial f(x)$ is compact, as claimed in Theorem 2.4.3. Because we know that $f'(x; u)$ is finite for all u as $x \in \operatorname{int} \operatorname{dom} f$, and $g \in \partial f(x)$ satisfies

$$\|g\|_2 = \sup_{u : \|u\|_2 \leqslant 1} \langle g, u \rangle \leqslant \sup_{u : \|u\|_2 \leqslant 1} f'(x; u) < \infty.$$

Lemma 2.4.6. *Let* f *be closed convex and* $\partial f(x) \neq \emptyset$*. Then*

$$f'(x; u) = \sup_{g \in \partial f(x)} \langle g, u \rangle . \tag{2.4.7}$$

Proof. Certainly, Lemma 2.4.4 shows that $f'(x; u) \geqslant \sup_{g \in \partial f(x)} \langle g, u \rangle$. We must show the other direction. To that end, note that viewed as a function of u, we have $f'(x; u)$ is convex and positively homogeneous, meaning that $f'(x; tu) = tf'(x; u)$ for $t \geqslant 0$. Thus, we can always write (by Corollary 2.2.19)

$$f'(x; u) = \sup \left\{ \langle v, u \rangle + b : f'(x; w) \geqslant b + \langle v, w \rangle \text{ for all } w \in \mathbb{R}^n \right\} .$$

Using the positive homogeneity, we have $f'(x; 0) = 0$ and thus we must have $b = 0$, so that $u \mapsto f'(x; u)$ is characterized as the supremum of linear functions:

$$f'(x; u) = \sup \left\{ \langle v, u \rangle : f'(x; w) \geqslant \langle v, w \rangle \text{ for all } w \in \mathbb{R}^n \right\} .$$

But the set $\{v : \langle v, w \rangle \leqslant f'(x; w) \text{ for all } w\}$ is simply $\partial f(x)$ by Lemma 2.4.4. □

A relatively straightforward calculation using Lemma 2.4.4, which we give in the next proposition, shows that the subgradient is simply the gradient of differentiable convex functions. Note that as a consequence of this, we have the first-order inequality that $f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle$ for any differentiable convex function.

Proposition 2.4.8. *Let* f *be convex and differentiable at a point* x*. Then* $\partial f(x) = \{\nabla f(x)\}$*.*

Proof. If f is differentiable at a point x, then the chain rule implies that

$$f'(x; u) = \langle \nabla f(x), u \rangle \geqslant \langle g, u \rangle$$

for any $g \in \partial f(x)$, the inequality following from Lemma 2.4.4. By replacing u with $-u$, we have $f'(x; -u) = -\langle \nabla f(x), u \rangle \geqslant -\langle g, u \rangle$ as well, or $\langle g, u \rangle = \langle \nabla f(x), u \rangle$ for all u. Letting u vary in (for example) the set $\{u : \|u\|_2 \leqslant 1\}$ gives the result. □

Lastly, we have the following consequence of the previous lemmas, which relates the norms of subgradients $g \in \partial f(x)$ to the Lipschitzian properties of f. Recall that a function f is L-Lipschitz with respect to the norm $\|\cdot\|$ over a set C if

$$|f(x) - f(y)| \leqslant L \|x - y\|$$

for all $x, y \in C$. Then the following proposition is an immediate consequence of Lemma 2.4.6.

Proposition 2.4.9. *Suppose that* f *is* L*-Lipschitz with respect to the norm* $\|\cdot\|$ *over a set* C*, where* $C \subset \operatorname{int} \operatorname{dom} f$*. Then*

$$\sup\{\|g\|_* : g \in \partial f(x), x \in C\} \leqslant L.$$

Examples We can provide a number of examples of subgradients. A general rule of thumb is that, if it is possible to compute the function, it is possible to compute its subgradients. As a first example, we consider

$$f(x) = |x|.$$

Then by inspection, we have

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0. \end{cases}$$

A more complex example is given by any vector norm $\|\cdot\|$. In this case, we use the fact that the dual norm is defined by

$$\|y\|_* := \sup_{x : \|x\| \leqslant 1} \langle x, y \rangle .$$

Moreover, we have that $\|x\| = \sup_{y : \|y\|_* \leqslant 1} \langle y, x \rangle$. Fixing $x \in \mathbb{R}^n$, we thus see that if $\|g\|_* \leqslant 1$ and $\langle g, x \rangle = \|x\|$, then

$$\|x\| + \langle g, y - x \rangle = \|x\| - \|x\| + \langle g, y \rangle \leqslant \sup_{v : \|v\|_* \leqslant 1} \langle v, y \rangle = \|y\| .$$

It is possible to show a converse—we leave this as an exercise for the interested reader—and we claim that

$$\partial \|x\| = \{g \in \mathbb{R}^n : \|g\|_* \leqslant 1, \langle g, x \rangle = \|x\|\}.$$

For a more concrete example, we have

$$\partial \|x\|_2 = \begin{cases} x / \|x\|_2 & \text{if } x \neq 0 \\ \{u : \|u\|_2 \leqslant 1\} & \text{if } x = 0. \end{cases}$$

Optimality properties Subgradients also allows us to characterize solutions to convex optimization problems, giving similar characterizations as those we provided for projections. The next theorem, containing necessary and sufficient conditions for a point x to minimize a convex function f, generalizes the standard first-order optimality conditions for differentiable f (e.g., Section 4.2.3 in [12]). The intuition for Theorem 2.4.11 is that there is a vector g in the subgradient set $\partial f(x)$ such that $-g$ is a supporting hyperplane to the feasible set C at the point x. That is, the directions of decrease of the function f lie outside the optimization set C. Figure 2.4.10 shows this behavior.

Theorem 2.4.11. *Let f be convex. The point $x \in \operatorname{int} \operatorname{dom} f$ minimizes f over a convex set C if and only if there exists a subgradient $g \in \partial f(x)$ such that simultaneously for all $y \in C$,*

$$\langle g, y - x \rangle \geqslant 0. \tag{2.4.12}$$

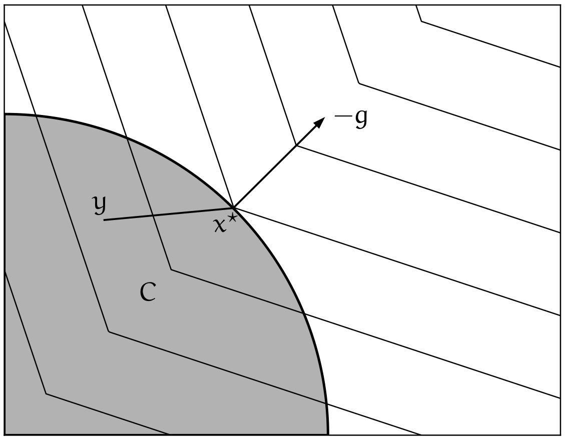

Figure 2.4.10. The point $x^\star$ minimizes f over C (the shown level curves) if and only if for some $g \in \partial f(x^\star)$, $\langle g, y - x^\star \rangle \geqslant 0$ for all $y \in C$. Note that not all subgradients satisfy this inequality.

Proof. One direction of the theorem is easy. Indeed, pick $y \in C$. Then certainly there exists $g \in \partial f(x)$ for which $\langle g, y - x \rangle \geqslant 0$. Then by definition,

$$f(y) \geqslant f(x) + \langle g, y - x \rangle \geqslant f(x).$$

This holds for any $y \in C$, so x is clearly optimal.

For the converse, suppose that x minimizes f over C. Then for any $y \in C$ and any $t \geqslant 0$ such that $x + t(y - x) \in C$, we have

$$f(x + t(y - x)) \geqslant f(x) \text{ or } 0 \leqslant \frac{f(x + t(y - x)) - f(x)}{t}.$$

Taking the limit as $t \to 0$, we have $f'(x; y - x) \geqslant 0$ for all $y \in C$. Now, let us suppose for the sake of contradiction that there exists a y such that for all $g \in \partial f(x)$, we have $\langle g, y - x \rangle < 0$. Because

$$\partial f(x) = \{g : \langle g, u \rangle \leqslant f'(x; u) \text{ for all } u \in \mathbb{R}^n\}$$

by Lemma 2.4.6, and $\partial f(x)$ is compact, we have that $\sup_{g \in \partial f(x)} \langle g, y - x \rangle$ is attained, which would imply

$$f'(x; y - x) < 0.$$

This is a contradiction. □

2.5. Calculus rules with subgradients We present a number of calculus rules that show how subgradients are, essentially, similar to derivatives, with a few exceptions (see also Ch. VII of [27]). When we develop methods for optimization problems based on subgradients, these basic calculus rules will prove useful.

Scaling. If we let $h(x) = \alpha f(x)$ for some $\alpha \geqslant 0$, then $\partial h(x) = \alpha \partial f(x)$.

Finite sums. Suppose that $f_1, \ldots, f_m$ are convex functions and let $f = \sum_{i=1}^m f_i$. Then

$$\partial f(x) = \sum_{i=1}^m \partial f_i(x),$$

where the addition is Minkowski addition. To see that $\sum_{i=1}^m \partial f_i(x) \subset \partial f(x)$, let $g_i \in \partial f_i(x)$ for each i, in which case it is clear that $f(y) = \sum_{i=1}^m f_i(y) \geqslant \sum_{i=1}^m f_i(x) + \langle g_i, y - x\rangle$, so that $\sum_{i=1}^m g_i \in \partial f(x)$. The converse is somewhat more technical and is a special case of the results to come.

Integrals. More generally, we can extend this summation result to integrals, assuming the integrals exist. These calculations are essential for our development of stochastic optimization schemes based on stochastic (sub)gradient information in the coming lectures. Indeed, for each $s \in \mathcal{S}$, where $\mathcal{S}$ is some set, let f_s be convex. Let μ be a positive measure on the set $\mathcal{S}$, and define the convex function $f(x) = \int f_s(x) d\mu(s)$. In the notation of the introduction (Eq. (1.0.1)) and the problems coming in Section 3.4, we take μ to be a probability distribution on a set $\mathcal{S}$, and if $F(\cdot; s)$ is convex in its first argument for all $s \in \mathcal{S}$, then we may take

$$f(x) = \mathbb{E}[F(x; S)]$$

and satisfy the conditions above. We shall see many such examples in the sequel.

Then if we let $g_s(x) \in \partial f_s(x)$ for each $s \in \mathcal{S}$, we have (assuming the integral exists and that the selections $g_s(x)$ are appropriately measurable)

$$\int g_s(x) d\mu(s) \in \partial f(x). \tag{2.5.1}$$

To see the inclusion, note that for any y we have

$$\begin{aligned}\left\langle \int g_s(x) d\mu(s), y - x \right\rangle &= \int \langle g_s(x), y - x\rangle \, d\mu(s) \\ &\leqslant \int (f_s(y) - f_s(x)) d\mu(s) = f(y) - f(x).\end{aligned}$$

So the inclusion (2.5.1) holds. Eliding a few technical details, one generally obtains the equality

$$\partial f(x) = \left\{ \int g_s(x) d\mu(s) : g_s(x) \in \partial f_s(x) \text{ for each } s \in \mathcal{S} \right\}.$$

Returning to our running example of stochastic optimization, if we have a collection of functions $F : \mathbb{R}^n \times \mathcal{S} \to \mathbb{R}$, where for each $s \in \mathcal{S}$ the function $F(\cdot; s)$ is convex, then $f(x) = \mathbb{E}[f(x; S)]$ is convex when we take expectations over S, and taking

$$g(x; s) \in \partial F(x; s)$$

gives a *stochastic gradient* with the property that $\mathbb{E}[g(x; S)] \in \partial f(x)$. For more on these calculations and conditions, see the classic paper of Bertsekas [7], which addresses the measurability issues.

Affine transformations. Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then $h : \mathbb{R}^n \to \mathbb{R}$ defined by $h(x) = f(Ax + b)$ is convex and has subdifferential

$$\partial h(x) = A^T \partial f(Ax + b).$$

Indeed, let $g \in \partial f(Ax + b)$, so that

$$h(y) = f(Ay + b) \geqslant f(Ax + b) + \langle g, (Ay + b) - (Ax + b) \rangle = h(x) + \left\langle A^T g, y - x \right\rangle,$$

giving the result.

Finite maxima. Let f_i, $i = 1, \ldots, m$, be convex functions, and $f(x) = \max_{i \leqslant m} f_i(x)$. Then we have

$$\operatorname{epi} f = \bigcap_{i \leqslant m} \operatorname{epi} f_i,$$

which is convex, and f is convex. Now, let i be any index such that $f_i(x) = f(x)$, and let $g_i \in \partial f_i(x)$. Then we have for any $y \in \mathbb{R}^n$ that

$$f(y) \geqslant f_i(y) \geqslant f_i(x) + \langle g_i, y - x \rangle = f(x) + \langle g_i, y - x \rangle.$$

So $g_i \in \partial f(x)$. More generally, we have the result that

$$\partial f(x) = \operatorname{Conv}\{\partial f_i(x) : f_i(x) = f(x)\}, \tag{2.5.2}$$

that is, the subgradient set of f is the convex hull of the subgradients of *active* functions at x, that is, those attaining the maximum. If there is only a single unique active function f_i, then $\partial f(x) = \partial f_i(x)$. See Figure 2.5.3 for a graphical representation.

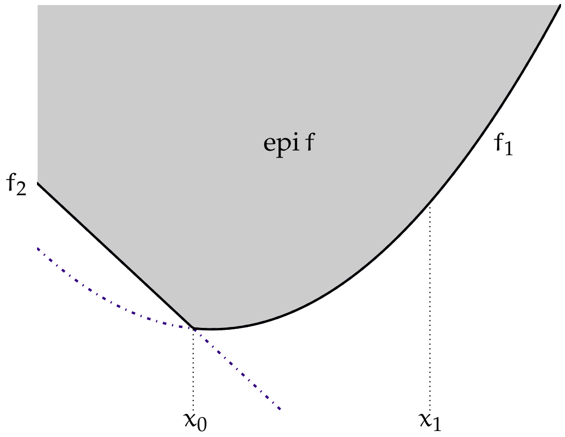

FIGURE 2.5.3. Subgradients of finite maxima. The function $f(x) = \max\{f_1(x), f_2(x)\}$ where $f_1(x) = x^2$ and $f_2(x) = -2x - \frac{1}{5}$, and f is differentiable everywhere except at $x_0 = -1 + \sqrt{4/5}$.

Uncountable maxima (supremum). Lastly, consider $f(x) = \sup_{\alpha \in \mathcal{S}} f_\alpha(x)$, where $\mathcal{A}$ is an arbitrary index set and f_α is convex for each α. First, let us assume that the supremum is attained at some $\alpha \in \mathcal{A}$. Then, identically to the above, we have that $\partial f_\alpha(x) \subset \partial f(x)$. More generally, we have

$$\partial f(x) \supset \mathrm{Conv}\{\partial f_\alpha(x) : f_\alpha(x) = f(x)\}.$$

Achieving equality in the preceding definition requires a number of conditions, and if the supremum is not attained, the function f *may* not be subdifferentiable.

Notes and further reading The study of convex analysis and optimization originates, essentially, with Rockafellar's 1970 book *Convex Analysis* [49]. Because of the limited focus of these lecture notes, we have only barely touched on many topics in convex analysis, developing only those we need. Two omissions are perhaps the most glaring: except tangentially, we have provided no discussion of conjugate functions and conjugacy, and we have not discussed Lagrangian duality, both of which are central to any study of convex analysis and optimization.

A number of books provide coverage of convex analysis in finite and infinite dimensional spaces and make excellent further reading. For broad coverage of convex optimization problems, theory, and algorithms, Boyd and Vandenberghe [12] is an excellent reference, also providing coverage of basic convex duality theory and conjugate functions. For deeper forays into convex analysis, personal favorites of mine include the books of Hiriart-Urruty and Lemarécahl [27, 28], as well as the shorter volume [29], and Bertsekas [8] also provides an elegant geometric picture of convex analysis and optimization. Our approach here follows Hiriart-Urruty and Lemaréchal's most closely. For a treatment of the issues of separation, convexity, duality, and optimization in infinite dimensional spaces, an excellent reference is the classic book by Luenberger [36].

3. Subgradient Methods

Lecture Summary: In this lecture, we discuss first order methods for the minimization of convex functions. We focus almost exclusively on subgradient-based methods, which are essentially universally applicable for convex optimization problems, because they rely very little on the structure of the problem being solved. This leads to effective but slow algorithms in classical optimization problems. In large scale problems arising out of machine learning and statistical tasks, however, subgradient methods enjoy a number of (theoretical) optimality properties and have excellent practical performance.

3.1. Introduction In this lecture, we explore a basic subgradient method, and a few variants thereof, for solving general convex optimization problems. Throughout, we will attack the problem

$$\underset{x}{\mathrm{minimize}} \ f(x) \quad \text{subject to } x \in C \tag{3.1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is convex (though it may take on the value $+\infty$ for $x \notin \mathrm{dom}\, f$) and C is a closed convex set. Certainly in this generality, finding a universally good method for solving the problem (3.1.1) is hopeless, though we will see that the subgradient method does essentially apply in this generality.

Convex programming methodologies developed in the last fifty years or so have given powerful methods for solving optimization problems. The performance of many methods for solving convex optimization problems is measured by the amount of time or number of iterations required of them to give an ϵ-optimal solution to the problem (3.1.1), roughly, how long it takes to find some $\widehat{x}$ such that $f(\widehat{x}) - f(x^\star) \leqslant \epsilon$ and $\mathrm{dist}(\widehat{x}, C) \leqslant \epsilon$ for an optimal $x^\star \in C$. Essentially any problem for which we can compute subgradients efficiently can be solved to accuracy ϵ in time polynomial in the dimension n of the problem and $\log \frac{1}{\epsilon}$ by the ellipsoid method (cf. [41, 45]). Moreover, for somewhat better structured (but still quite general) convex problems, interior point and second order methods [12, 45] are practically and theoretically quite efficient, sometimes requiring only $\mathcal{O}(\log \log \frac{1}{\epsilon})$ iterations to achieve optimization error ϵ. (See the lectures by S. Wright in this volume.) These methods use the Newton method as a basic solver, along with specialized representations of the constraint set C, and are quite powerful.

However, for large scale problems, the time complexity of standard interior point and Newton methods can be prohibitive. Indeed, for n-dimensional problems—that is, when $x \in \mathbb{R}^n$—interior point methods scale at best as $\mathcal{O}(n^3)$, and can be much worse. When n is large (where today, large may mean $n \approx 10^9$), this becomes highly non-trivial. In such large scale problems and problems arising from any type of data-collection process, it is reasonable to expect that our representation of problem data is inexact at best. In statistical machine learning problems, for example, this is often the case; generally, many applications do not require accuracy higher than, say $\epsilon = 10^{-2}$ or 10^{-3}, in which case faster but less exact methods become attractive.

It is with this motivation that we attack solving the problem (3.1.1) in this lecture, showing classical subgradient algorithms. These algorithms have the advantage that their per-iteration costs are low—$\mathcal{O}(n)$ or smaller for n-dimensional problems—but they achieve low accuracy solutions to (3.1.1) very quickly. Moreover, depending on problem structure, they can sometimes achieve convergence rates that are *independent* of problem dimension. More precisely, and as we will see later, the methods we study will guarantee convergence to an ϵ-optimal solution to problem (3.1.1) in $O(1/\epsilon^2)$ iterations, while methods that achieve better dependence on ϵ require at least $n \log \frac{1}{\epsilon}$ iterations.

3.2. The gradient and subgradient methods We begin by focusing on the unconstrained case, that is, when the set C in problem (3.1.1) is $C = \mathbb{R}^n$. That is, we wish to solve

$$\operatorname*{minimize}_{x \in \mathbb{R}^n} \; f(x).$$

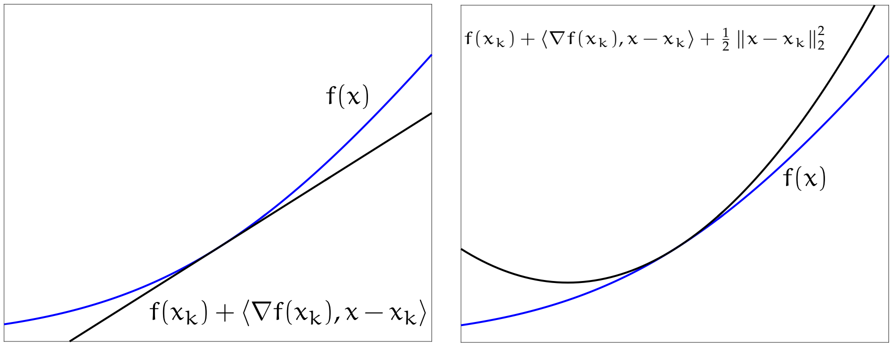

Figure 3.2.1. Left: linear approximation (in black) to the function $f(x) = \log(1 + e^x)$ (in blue) at the point $x_k = 0$. Right: linear plus quadratic upper bound for the function $f(x) = \log(1 + e^x)$ at the point $x_k = 0$. This is the upper-bound and approximation of the gradient method (3.2.3) with the choice $\alpha_k = 1$.

We first review the gradient descent method, using it as motivation for what follows. In the gradient descent method, minimize the objective (3.1.1) by iteratively updating

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \tag{3.2.2}$$

where $\alpha_k > 0$ is a positive sequence of stepsizes. The original motivations for this choice of update come from the fact that $x^\star$ minimizes a convex f if and only if $0 = \nabla f(x^\star)$; we believe a more compelling justification comes from the idea of modeling the convex function being minimized. Indeed, the update (3.2.2) is equivalent to

$$x_{k+1} = \operatorname*{argmin}_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}. \tag{3.2.3}$$

The interpretation is as follows: the linear functional $x \mapsto \{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle\}$ is the best linear approximation to the function f at the point x_k, and we would like to make progress minimizing x. So we minimize this linear approximation, but to make sure that it has fidelity to the function f, we add a quadratic $\|x - x_k\|_2^2$ to penalize moving too far from x_k, which would invalidate the linear approximation. See Figure 3.2.1. Assuming that f is continuously differentiable (often, one assumes the gradient $\nabla f(x)$ is Lipschitz), then gradient descent is a descent method if the stepsize $\alpha_k > 0$ is small enough—it monotonically decreases the objective $f(x_k)$. We spend no more time on the convergence of gradient-based methods, except to say that the choice of the stepsize α_k is often extremely important, and there is a body of research on carefully choosing directions as well as stepsize lengths; Nesterov [44] provides an excellent treatment of many of the basic issues.

Subgradient algorithms The subgradient method is a minor variant of the method (3.2.2), except that instead of using the gradient, we use a subgradient. The method can be written simply: for $k = 1, 2, \ldots$, we iterate

i. Choose any subgradient

$$g_k \in \partial f(x_k)$$

ii. Take the subgradient step

$$x_{k+1} = x_k - \alpha_k g_k. \tag{3.2.4}$$

Unfortunately, the subgradient method is not, in general, a descent method. For a simple example, take the function $f(x) = |x|$, and let $x_1 = 0$. Then except for the choice $g = 0$, all subgradients $g \in \partial f(0) = [-1, 1]$ are *ascent* directions. This is not just an artifact of 0 being optimal for f; in higher dimensions, this behavior is common. Consider, for example, $f(x) = \|x\|_1$ and let $x = e_1 \in \mathbb{R}^n$, the first standard basis vector. Then $\partial f(x) = e_1 + \sum_{i=2}^n t_i e_i$, where $t_i \in [-1, 1]$. Any vector $g = e_1 + \sum_{i=2}^n t_i e_i$ with $\sum_{i=2}^n |t_i| > 1$ is an ascent direction for f, meaning that $f(x - \alpha g) > f(x)$ for all $\alpha > 0$. If we were to pick a uniformly random $g \in \partial f(e_1)$, for example, then the probability that g is a descent direction is exponentially small in the dimension n.

In general, the characterization of the subgradient set $\partial f(x)$ as in Lemma 2.4.4, as $\{g : f'(x; u) \geqslant \langle g, u \rangle \text{ for all } u\}$ where $f'(x; u) = \lim_{t \to 0} \frac{f(x+tu)-f(x)}{t}$ is the directional derivative, and the fact that $f'(x; u) = \sup_{g \in \partial f(x)} \langle g, u \rangle$ guarantees that

$$\operatorname*{argmin}_{g \in \partial f(x)} \{\|g\|_2^2\}$$

is a descent direction, but we do not prove this here. Indeed, finding such a descent direction would require explicitly calculating the entire subgradient set $\partial f(x)$, which for a number of functions is non-trivial and breaks the simplicity of the subgradient method (3.2.4), which works with *any* subgradient.

It is the case, however, that so long as the point x does not minimize $f(x)$, then subgradients descend on a related quantity: the distance of x to *any* optimal point. Indeed, let $g \in \partial f(x)$, and let $x^\star \in \operatorname{argmin} f(x)$ (we assume such a point exists), which need not be unique. Then we have for any α that

$$\frac{1}{2} \|x - \alpha g - x^\star\|_2^2 = \frac{1}{2} \|x - x^\star\|_2^2 - \alpha \langle g, x - x^\star \rangle + \frac{\alpha^2}{2} \|g\|_2^2.$$

The key is that for small enough $\alpha > 0$, the quantity on the right is strictly smaller than $\frac{1}{2} \|x - x^\star\|_2^2$, as we now show. We use the defining inequality of the subgradient, that is, that $f(y) \geqslant f(x) + \langle g, y - x \rangle$ for all y, including $x^\star$. This gives $-\langle g, x - x^\star \rangle = \langle g, x^\star - x \rangle \leqslant f(x^\star) - f(x)$, and thus

$$\frac{1}{2} \|x - \alpha g - x^\star\|_2^2 \leqslant \frac{1}{2} \|x - x^\star\|_2^2 - \alpha (f(x) - f(x^\star)) + \frac{\alpha^2}{2} \|g\|_2^2. \tag{3.2.5}$$

From inequality (3.2.5), we see immediately that, no matter our choice $g \in \partial f(x)$, we have

$$0 < \alpha < \frac{2(f(x) - f(x^\star))}{\|g\|_2^2} \text{ implies } \|x - \alpha g - x^\star\|_2^2 < \|x - x^\star\|_2^2.$$

Summarizing, by noting that $f(x) - f(x^\star) > 0$, we have

Observation 3.2.6. *If $0 \notin \partial f(x)$, then for any $x^\star \in \operatorname{argmin}_x f(x)$ and any $g \in \partial f(x)$, there is a stepsize $\alpha > 0$ such that $\|x - \alpha g - x^\star\|_2^2 < \|x - x^\star\|_2^2$.*

This observation is the key to the analysis of subgradient methods.

Convergence guarantees Perhaps unsurprisingly, given the simplicity of the subgradient method, the analysis of convergence for the method is also quite simple. We begin by stating a general result on the convergence of subgradient methods; we provide a number of variants in the sequel. We make a few simplifying assumptions in stating our result, several of which are not completely necessary, but which considerably simplify the analysis. We enumerate them here:

i. There is at least one (possibly non-unique) minimizing point $x^\star \in \operatorname{argmin}_x f(x)$ with $f(x^\star) = \inf_x f(x) > -\infty$

ii. The subgradients are bounded: for all x and all $g \in \partial f(x)$, we have the subgradient bound $\|g\|_2 \leqslant M < \infty$ (independently of x).

Theorem 3.2.7. *Let $\alpha_k \geqslant 0$ be any non-negative sequence of stepsizes and the preceding assumptions hold. Let x_k be generated by the subgradient iteration (3.2.4). Then for all $K \geqslant 1$,*

$$\sum_{k=1}^{K} \alpha_k [f(x_k) - f(x^\star)] \leqslant \frac{1}{2} \|x_1 - x^\star\|_2^2 + \frac{1}{2} \sum_{k=1}^{K} \alpha_k^2 M^2.$$

Proof. The entire proof essentially amounts to writing down the distance $\|x_{k+1} - x^\star\|_2^2$ and expanding the square, which we do. By applying inequality (3.2.5), we have

$$\begin{aligned} \frac{1}{2} \|x_{k+1} - x^\star\|_2^2 &= \frac{1}{2} \|x_k - \alpha_k g_k - x^\star\|_2^2 \\ &\overset{(3.2.5)}{\leqslant} \frac{1}{2} \|x_k - x^\star\|_2^2 - \alpha_k (f(x_k) - f(x^\star)) + \frac{\alpha_k^2}{2} \|g_k\|_2^2. \end{aligned}$$

Rearranging this inequality and using that $\|g_k\|_2^2 \leqslant M^2$, we obtain

$$\begin{aligned} \alpha_k [f(x_k) - f(x^\star)] &\leqslant \frac{1}{2} \|x_k - x^\star\|_2^2 - \frac{1}{2} \|x_{k+1} - x^\star\|_2^2 + \frac{\alpha_k^2}{2} \|g_k\|_2^2 \\ &\leqslant \frac{1}{2} \|x_k - x^\star\|_2^2 - \frac{1}{2} \|x_{k+1} - x^\star\|_2^2 + \frac{\alpha_k^2}{2} M^2. \end{aligned}$$

By summing the preceding expression from $k = 1$ to $k = K$ and canceling the alternating $\pm \|x_k - x^\star\|_2^2$ terms, we obtain the theorem. □

Theorem 3.2.7 is the starting point from which we may derive a number of useful consquences. First, we use convexity to obtain the following immediate corollary (we assume that $\alpha_k > 0$ in the corollary).

Corollary 3.2.8. *Let* $A_k = \sum_{i=1}^{k} \alpha_i$ *and define* $\overline{x}_K = \frac{1}{A_K} \sum_{k=1}^{K} \alpha_k x_k$. *Then*

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{\|x_1 - x^\star\|_2^2 + \sum_{k=1}^{K} \alpha_k^2 M^2}{2 \sum_{k=1}^{K} \alpha_k}.$$

Proof. Noting that $A_K^{-1} \sum_{k=1}^{K} \alpha_k = 1$, we see by convexity that

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{1}{\sum_{k=1}^{K} \alpha_k} \sum_{k=1}^{K} \alpha_k f(x_k) - f(x^\star) = A_K^{-1} \left[\sum_{k=1}^{K} \alpha_k (f(x_k) - f(x^\star)) \right].$$

Applying Theorem 3.2.7 gives the result. □

Corollary 3.2.8 allows us to give a number of basic convergence guarantees based on our stepsize choices. For example, we see that whenever we have

$$\alpha_k \to 0 \text{ and } \sum_{k=1}^{\infty} \alpha_k = \infty,$$

then $\sum_{k=1}^{K} \alpha_k^2 / \sum_{k=1}^{K} \alpha_k \to 0$ and so

$$f(\overline{x}_K) - f(x^\star) \to 0 \text{ as } K \to \infty.$$

Moreover, we can give specific stepsize choices to optimize the bound. For example, let us assume for simplicity that $R^2 = \|x_1 - x^\star\|_2^2$ is our distance (radius) to optimality. Then choosing a fixed stepsize $\alpha_k = \alpha$, we have

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{R^2}{2K\alpha} + \frac{\alpha M^2}{2}. \tag{3.2.9}$$

Optimizing this bound by taking $\alpha = \frac{R}{M\sqrt{K}}$ gives

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{RM}{\sqrt{K}}.$$

Given that subgradient descent methods are not descent methods, it often makes sense, instead of tracking the (weighted) average of the points or using the final point, to use the best point observed thus far. Naturally, if we let

$$x_k^{\text{best}} = \operatorname*{argmin}_{x_i : i \leqslant k} f(x_i)$$

and define $f_k^{\text{best}} = f(x_k^{\text{best}})$, then we have the same convergence guarantees that

$$f(x_k^{\text{best}}) - f(x^\star) \leqslant \frac{R^2 + \sum_{k=1}^{K} \alpha_k^2 M^2}{2 \sum_{k=1}^{K} \alpha_k}.$$

A number of more careful stepsize choices are possible, though we refer to the notes at the end of this lecture for more on these choices and applications outside of those we consider, as our focus is naturally circumscribed.

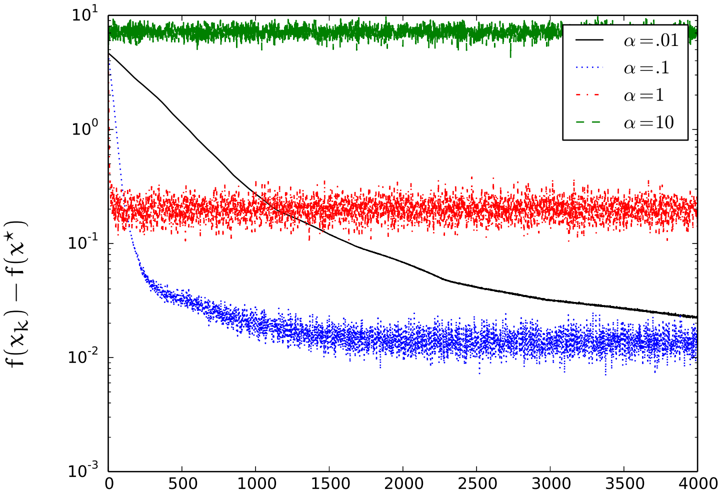

FIGURE 3.2.10. Subgradient method applied to the robust regression problem (3.2.12) with fixed stepsizes.

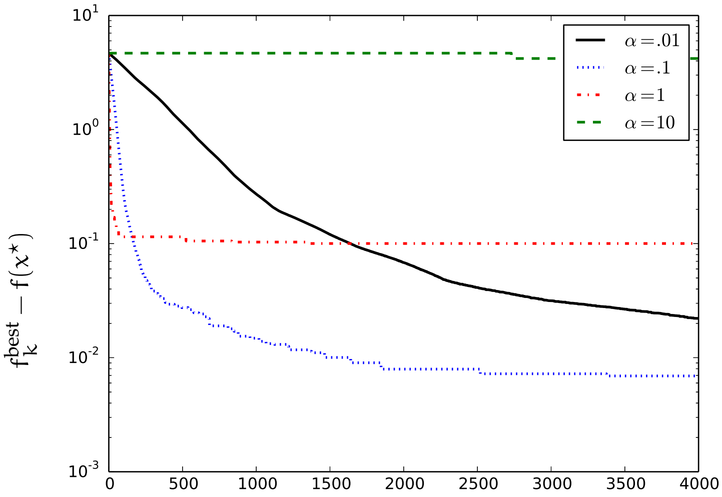

FIGURE 3.2.11. Subgradient method applied to the robust regression problem (3.2.12) with fixed stepsizes, showing performance of the best iterate $f_k^{\rm best} - f(x^\star)$.

Example We now present an example that has applications in robust statistics and other data fitting scenarios. As a motivating scenario, suppose we have a sequence of vectors $a_i \in \mathbb{R}^n$ and target responses $b_i \in \mathbb{R}$, and we would like to predict b_i via the inner product $\langle a_i, x \rangle$ for some vector x. If there are outliers or other data corruptions in the targets b_i, a natural objective for this task, given the data matrix $A = [a_1 \ \cdots \ a_m]^\top \in \mathbb{R}^{m \times n}$ and vector $b \in \mathbb{R}^m$, is the absolute error

$$f(x) = \frac{1}{m} \|Ax - b\|_1 = \frac{1}{m} \sum_{i=1}^m |\langle a_i, x\rangle - b_i|. \tag{3.2.12}$$

We perform subgradient descent on this objective, which has subgradient

$$g(x) = \frac{1}{m} A^\top \operatorname{sign}(Ax - b) = \frac{1}{m} \sum_{i=1}^m a_i \operatorname{sign}(\langle a_i, x\rangle - b_i) \in \partial f(x)$$

at the point x, for $K = 4000$ iterations with a fixed stepsize $\alpha_k \equiv \alpha$ for all k. We give the results in Figures 3.2.10 and 3.2.11, which exhibit much of the typical behavior of subgradient methods. From the plots, we see roughly a few phases of behavior: the method with stepsize $\alpha = 1$ makes progress very quickly initially, but then enters its “jamming” phase, where it essentially makes no more progress. (The largest stepsize, $\alpha = 10$, simply jams immediately.) The accuracy of the methods with different stepsizes varies greatly, as well—the smaller the stepsize, the better the (final) performance of the iterates x_k, but initial progress is much slower.

3.3. Projected subgradient methods It is often the case that we wish to solve problems not over $\mathbb{R}^n$ but over some constrained set, for example, in the Lasso [57] and in compressed sensing applications [20] one minimizes an objective such as $\|Ax - b\|_2^2$ subject to $\|x\|_1 \leqslant R$ for some constant $R < \infty$. Recalling the problem (3.1.1), we more generally wish to solve the problem

$$\text{minimize } f(x) \text{ subject to } x \in C \subset \mathbb{R}^n,$$

where C is a closed convex set, not necessarily $\mathbb{R}^n$. The projected subgradient method is close to the subgradient method, except that we replace the iteration with

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k) \tag{3.3.1}$$

where

$$\pi_C(x) = \operatorname*{argmin}_{y \in C} \{\|x - y\|_2\}$$

denotes the (Euclidean) projection onto C. As in the gradient case (3.2.3), we can reformulate the update as making a linear approximation, with quadratic damping, to f and minimizing this approximation: by algebraic manipulation, the update (3.3.1) is equivalent to

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k\rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}. \tag{3.3.2}$$

Figure 3.3.3 shows an example of the iterations of the projected gradient method applied to minimizing $f(x) = \|Ax - b\|_2^2$ subject to the ℓ_1-constraint $\|x\|_1 \leqslant 1$. Note that the method iterates between moving outside the ℓ_1-ball toward the minimum of f (the level curves) and projecting back onto the ℓ_1-ball.

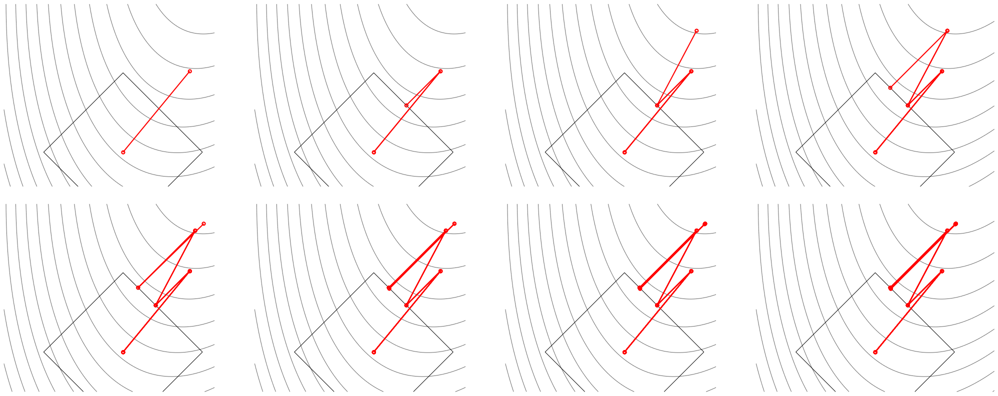

FIGURE 3.3.3. Example execution of the projected gradient method (3.3.1), on minimizing $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ subject to $\|x\|_1 \leqslant 1$.

It is very important in the projected subgradient method that the projection mapping π_C be efficiently computable—the method is effective essentially only in problems where this is true. In many situations, this is the case, but some care is necessary if the objective f is simple while the set C is complex. In such scenarios, projecting onto the set C may be as complex as solving the original optimization problem (3.1.1). For example, a general linear programming problem is described by

$$\underset{x}{\text{minimize}} \ \langle c, x\rangle \ \text{subject to} \ Ax = b, \ Cx \preceq d.$$

Then computing the projection onto the set $\{x : Ax = b, Cx \preceq d\}$ is at least as difficult as solving the original problem.

Examples of projections As noted above, it is important that projections π_C be efficiently calculable, and often a method's effectiveness is governed by how quickly one can compute the projection onto the constraint set C. With that in mind, we now provide two examples exhibiting convex sets C onto which projection is reasonably straightforward and for which we can write explicit, concrete projected subgradient updates.

Example 3.1: Suppose that C is an affine set, represented by $C = \{x \in \mathbb{R}^n : Ax = b\}$ for $A \in \mathbb{R}^{m \times n}$, $m \leqslant n$, where A is full rank. (So that A is a short and fat matrix and $AA^T \succ 0$.) Then the projection of x onto C is

$$\pi_C(x) = (I - A^T(AA^T)^{-1}A)x + A^T(AA^T)^{-1}b,$$

and if we begin the iterates from a point $x_k \in C$, i.e. with $Ax_k = b$, then

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k) = x_k - \alpha_k(I - A^T(AA^T)^{-1}A)g_k,$$

that is, we simply project g_k onto the nullspace of A and iterate. $\Diamond$

Example 3.2 (Some norm balls)**:** Let us consider updates when $C = \{x : \|x\|_p \leqslant 1\}$ for $p \in \{1, 2, \infty\}$, each of which is reasonably simple, though the projections are no longer affine. First, for $p = \infty$, we consider each coordinate $j = 1, 2, \ldots, n$ in turn, giving

$$[\pi_C(x)]_j = \min\{1, \max\{x_j, -1\}\},$$

that is, we simply truncate the coordinates of x to be in the range $[-1, 1]$. For $p = 2$, we have a similarly simple to describe update:

$$\pi_C(x) = \begin{cases} x & \text{if } \|x\|_2 \leqslant 1 \\ x / \|x\|_2 & \text{otherwise.} \end{cases}$$

When $p = 1$, that is, $C = \{x : \|x\|_1 \leqslant 1\}$, the update is somewhat more complex. If $\|x\|_1 \leqslant 1$, then $\pi_C(x) = x$. Otherwise, we find the (unique) $t \geqslant 0$ such that

$$\sum_{j=1}^{n} \left[|x_j| - t\right]_+ = 1,$$

and then set the coordinates j via

$$[\pi_C(x)]_j = \text{sign}(x_j) \left[|x_j| - t\right]_+ .$$

There are numerous efficient algorithms for finding this t (e.g. [14, 23]). $\diamondsuit$

Convergence results We prove the convergence of the projected subgradient using an argument similar to our proof of convergence for the classic (unconstrained) subgradient method. We assume that the set C is contained in the interior of the domain of the function f, which (as noted in the lecture on convex analysis) guarantees that f is Lipschitz continuous and subdifferentiable, so that there exists $M < \infty$ with $\|g\|_2 \leqslant M$ for all $g \in \partial f$. We make the following assumptions in the next theorem.

i. The set $C \subset \mathbb{R}^n$ is compact and convex, and $\|x - x^\star\|_2 \leqslant R < \infty$ for all $x \in C$.
ii. There exists $M < \infty$ such that $\|g\|_2 \leqslant M$ for all $g \in \partial f(x)$ and $x \in C$.

We make the compactness assumption to allow for a slightly different result than Theorem 3.2.7.

Theorem 3.3.4. *Let x_k be generated by the projected subgradient iteration* (3.3.1)*, where the stepsizes $\alpha_k > 0$ are non-increasing. Then*

$$\sum_{k=1}^{K} [f(x_k) - f(x^\star)] \leqslant \frac{R^2}{2\alpha_K} + \frac{1}{2} \sum_{k=1}^{K} \alpha_k M^2.$$

Proof. The starting point of the proof is the same basic inequality as we have been using, that is, the distance $\|x_{k+1} - x^\star\|_2^2$. In this case, we note that projections can never increase distances to points $x^\star \in C$, so that

$$\|x_{k+1} - x^\star\|_2^2 = \|\pi_C(x_k - \alpha_k g_k) - x^\star\|_2^2 \leqslant \|x_k - \alpha_k g_k - x^\star\|_2^2 .$$

Now, as in our earlier derivation, we apply inequality (3.2.5) to obtain

$$\frac{1}{2}\left\|x_{k+1}-x^\star\right\|_2^2 \leqslant \frac{1}{2}\left\|x_k-x^\star\right\|_2^2 - \alpha_k[f(x_k)-f(x^\star)] + \frac{\alpha_k^2}{2}\left\|g_k\right\|_2^2.$$

Rearranging this slightly by dividing by α_k, we find that

$$f(x_k)-f(x^\star) \leqslant \frac{1}{2\alpha_k}\left[\left\|x_k-x^\star\right\|_2^2 - \left\|x_{k+1}-x^\star\right\|_2^2\right] + \frac{\alpha_k}{2}\left\|g_k\right\|_2^2.$$

Now, using a variant of the telescoping sum in the proof of Theorem 3.2.7 we have

(3.3.5)
$$\sum_{k=1}^K [f(x_k)-f(x^\star)] \leqslant \sum_{k=1}^K \frac{1}{2\alpha_k}\left[\left\|x_k-x^\star\right\|_2^2 - \left\|x_{k+1}-x^\star\right\|_2^2\right] + \sum_{k=1}^K \frac{\alpha_k}{2}\left\|g_k\right\|_2^2.$$

We rearrange the middle sum in expression (3.3.5), obtaining

$$\begin{aligned}
&\sum_{k=1}^K \frac{1}{2\alpha_k}\left[\left\|x_k-x^\star\right\|_2^2 - \left\|x_{k+1}-x^\star\right\|_2^2\right] \\
&= \sum_{k=2}^K \left(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}}\right)\left\|x_k-x^\star\right\|_2^2 + \frac{1}{2\alpha_1}\left\|x_1-x^\star\right\|_2^2 - \frac{1}{2\alpha_K}\left\|x_K-x^\star\right\|_2^2 \\
&\leqslant \sum_{k=2}^K \left(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}}\right)R^2 + \frac{1}{2\alpha_1}R^2
\end{aligned}$$

because $\alpha_k \leqslant \alpha_{k-1}$. Noting that this last sum telescopes and that $\|g_k\|_2^2 \leqslant M^2$ in inequality (3.3.5) gives the result. □

One application of this result is when we use a decreasing stepsize of $\alpha_k = \alpha/\sqrt{k}$, which allows nearly as strong of a convergence rate as in the fixed stepsize case when the number of iterations K is known, but the algorithm provides a guarantee for all iterations k. Here, we have that

$$\sum_{k=1}^K \frac{1}{\sqrt{k}} \leqslant \int_0^K t^{-\frac{1}{2}}dt = 2\sqrt{K},$$

and so by taking $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$ we obtain the following corollary.

Corollary 3.3.6. *In addition to the conditions of the preceding paragraph, let the conditions of Theorem 3.3.4 hold. Then*

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{R^2}{2\alpha\sqrt{K}} + \frac{M^2\alpha}{\sqrt{K}}.$$

So we see that convergence is guaranteed, at the "best" rate $1/\sqrt{K}$, for all iterations. Here, we say "best" because this rate is unimprovable—there are worst case functions for which no method can achieve a rate of convergence faster than $RM/\sqrt{K}$—but in practice, one would hope to attain better behavior by leveraging problem structure.

3.4. Stochastic subgradient methods The real power of subgradient methods, which has become evident in the last ten or fifteen years, is in their applicability to large scale optimization problems. Indeed, while subgradient methods guarantee only slow convergence—requiring $1/\epsilon^2$ iterations to achieve ϵ-accuracy—their simplicity provides the benefit that they are robust to a number of errors. In fact, subgradient methods achieve unimprovable rates of convergence for a number of optimization problems with noise, and they often do so very computationally efficiently.

Stochastic optimization problems The basic building block for stochastic (sub)gradient methods is the *stochastic (sub)gradient*, often called the stochastic (sub)gradient oracle. Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be a convex function, and fix $x \in \operatorname{dom} f$. (We will typically omit the sub- qualifier in what follows.) Then a random vector g is a stochastic gradient for f at the point x if $\mathbb{E}[g] \in \partial f(x)$, or

$$f(y) \geqslant f(x) + \langle \mathbb{E}[g], y - x \rangle \text{ for all } y.$$

Said somewhat more formally, we make the following definition.

Definition 3.4.1. A *stochastic gradient oracle* for the function f consists of a triple $(g, \mathcal{S}, P)$, where $\mathcal{S}$ is a sample space, P is a probability distribution, and $g : \mathbb{R}^n \times \mathcal{S} \to \mathbb{R}^n$ is a mapping that for each $x \in \operatorname{dom} f$ satisfies

$$\mathbb{E}_P[g(x, S)] = \int g(x, s) dP(s) \in \partial f(x),$$

where $S \in \mathcal{S}$ is a sample drawn from P.

Often, with some abuse of notation, we will use g or $g(x)$ for shorthand of the random vector $g(x, S)$ when this does not cause confusion.

A standard example for these types of problems is *stochastic programming*, where we wish to solve the convex optimization problem

$$\begin{aligned} &\text{minimize } f(x) := \mathbb{E}_P[F(x; S)] \\ &\text{subject to } x \in C. \end{aligned} \tag{3.4.2}$$

Here S is a random variable on the space $\mathcal{S}$ with distribution P (so the expectation $\mathbb{E}_P[F(x; S)]$ is taken according to P), and for each $s \in \mathcal{S}$, the function $x \mapsto F(x; s)$ is convex. Then we immediately see that if we let

$$g(x, s) \in \partial_x F(x; s),$$

then g is a stochastic gradient when we draw $S \sim P$ and set $g = g(x, S)$, as in Lecture 2 (recall expression (2.5.1)). Recalling this calculation, we have

$$f(y) = \mathbb{E}_P[F(y; S)] \geqslant \mathbb{E}_P[F(x; S) + \langle g(x, S), y - x \rangle] = f(x) + \langle \mathbb{E}_P[g(x, S)], y - x \rangle$$

so that $\mathbb{E}_P[g(x, S)]$ is a stochastic subgradient.

To make the setting (3.4.2) more concrete, consider the robust regression problem (3.2.12), which uses

$$f(x) = \frac{1}{m} \|Ax - b\|_1 = \frac{1}{m} \sum_{i=1}^{m} |\langle a_i, x\rangle - b_i|.$$

Then a natural stochastic gradient, which requires time only $\mathcal{O}(n)$ to compute (as opposed to $\mathcal{O}(m \cdot n)$ to compute $Ax - b$), is to uniformly at random draw an index $i \in [m]$, then return

$$g = a_i \operatorname{sign}(\langle a_i, x\rangle - b_i).$$

More generally, given any problem in which one has a large dataset $\{s_1, \ldots, s_m\}$, and we wish to minimize the sum

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} F(x; s_i),$$

then drawing an index $i \in \{1, \ldots, m\}$ uniformly at random and using $g \in \partial_x F(x; s_i)$ is a stochastic gradient. Computing this stochastic gradient requires only the time necessary for computing some element of the subgradient set $\partial_x F(x; s_i)$, while the standard subgradient method applied to these problems is m-times more expensive in each iteration.

More generally, the expectation $\mathbb{E}[F(x; S)]$ is generally intractable to compute, especially if S is a high-dimensional distribution. In statistical and machine learning applications, we may not even know the distribution P, but we can observe samples $S_i \overset{\text{iid}}{\sim} P$. In these cases, it may be impossible to even implement the calculation of a subgradient $f'(x) \in \partial f(x)$, but sampling from P is possible, allowing us to compute *stochastic* subgradients.

Stochastic subgradient method With this motivation in place, we can describe the (projected) stochastic subgradient method. Simply, the method iterates as follows:

(1) Compute a stochastic subgradient g_k at the point x_k, where $\mathbb{E}[g_k \mid x_k] \in \partial f(x)$

(2) Perform the projected subgradient step

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k).$$

This is essentially identical to the projected gradient method (3.3.1), except that we replace the true subgradient with a stochastic gradient.

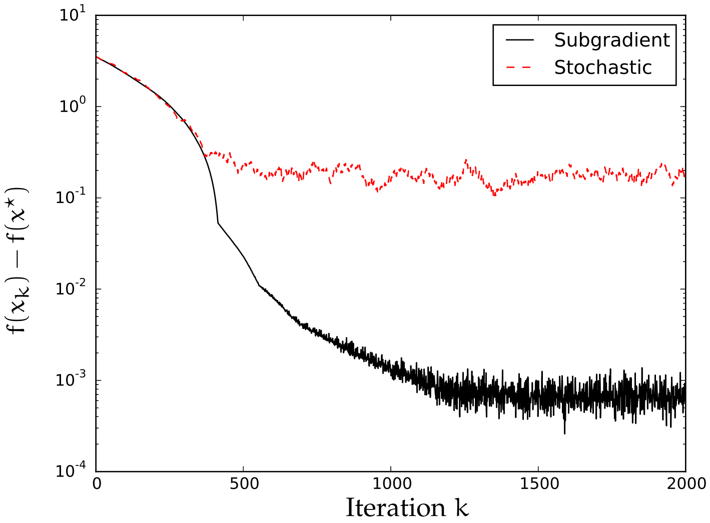

FIGURE 3.4.3. Stochastic subgradient method versus non-stochastic subgradient method performance on problem (3.4.4).

In the next section, we analyze the convergence of the procedure, but here we give two examples example here that exhibit some of the typical behavior of these methods.

Example 3.3 (Robust regression)**:** We consider the robust regression problem (3.2.12), solving

$$\text{(3.4.4)} \qquad \underset{x}{\text{minimize}} \;\; f(x) = \frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle - b_i| \;\; \text{subject to} \;\; \|x\|_2 \le R,$$

using the random sample $g = a_i \,\text{sign}(\langle a_i, x\rangle - b_i)$ as our stochastic gradient. We generate $A = [a_1 \;\cdots\; a_m]^\top$ by drawing $a_i \overset{\text{iid}}{\sim} N(0, I_{n\times n})$ and $b_i = \langle a_i, u\rangle + \varepsilon_i|\varepsilon_i|^3$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$ and u is a Gaussian random variable with identity covariance. We use $n = 50$, $m = 100$, and $R = 4$ for this experiment.

We plot the results of running the stochastic gradient iteration versus standard projected subgradient descent in Figure 3.4.3; both methods run with the fixed stepsize $\alpha = R/M\sqrt{K}$ for $M^2 = \frac{1}{m}\|A\|_{\text{Fr}}^2$, which optimizes the convergence guarantees for the methods. We see in the figure the typical performance of a stochastic gradient method: the initial progress in improving the objective is quite fast, but the method eventually stops making progress once it achieves some low accuracy (in this case, 10^{-1}). In this figure we should make clear, however, that each iteration of the stochastic gradient method requires time $\mathcal{O}(n)$, while each iteration of the (non-noisy) projected gradient method requires times $\mathcal{O}(n \cdot m)$, a factor of approximately 100 times slower. $\Diamond$

Example 3.4 (Multiclass support vector machine)**:** Our second example is somewhat more complex. We are given a collection of 16×16 grayscale images of

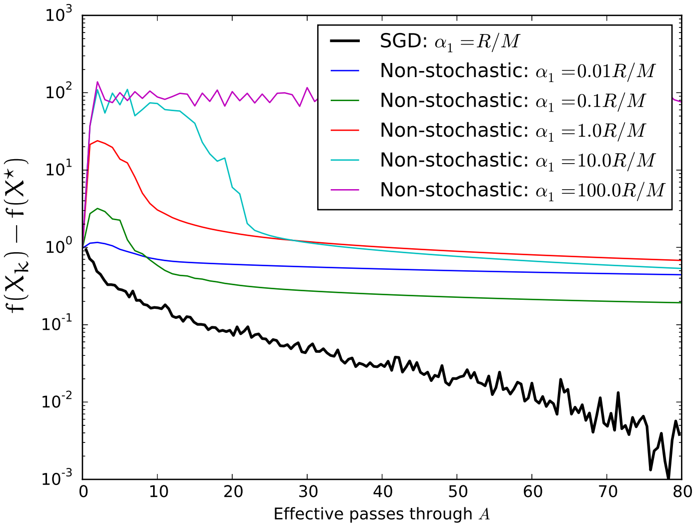

Figure 3.4.5. Comparison of stochastic versus non-stochastic methods for the average hinge-loss minimization problem (3.4.6). The horizontal axis is a measure of the time used by each method, represented as the number of times the matrix-vector product $X^T a_i$ is computed. Stochastic gradient descent vastly outperforms the non-stochastic methods.

handwritten digits $\{0, 1, \ldots, 9\}$, and wish to classify images, represented as vectors $a \in \mathbb{R}^{256}$, as one of the 10 digits. In a general k-class classification problem, we represent the multiclass classifier using the matrix

$$X = [x_1 \; x_2 \; \cdots \; x_k] \in \mathbb{R}^{n \times k},$$

where $k = 10$ for the digit classification problem. Given a data vector $a \in \mathbb{R}^n$, the "score" associated with class l is then $\langle x_l, a \rangle$, and the goal (given image data) is to find a matrix X assigning high scores to the correct image labels. (In machine learning, the typical notation is to use weight vectors $w_1, \ldots, w_k \in \mathbb{R}^n$ instead of $x_1, \ldots, x_k$, but we use X to remain consistent with our optimization focus.) The predicted class for a data vector $a \in \mathbb{R}^n$ is then

$$\operatorname*{argmax}_{l \in [k]} \langle a, x_l \rangle = \operatorname*{argmax}_{l \in [k]} \{[X^T a]_l\}.$$

We represent single training examples as pairs $(a, b) \in \mathbb{R}^n \times \{1, \ldots k\}$, and as a convex surrogate for a misclassification error that the matrix X makes on the pair (a, b), we use the *multiclass hinge* loss function

$$F(X; (a, b)) = \max_{l \neq b} \, [1 + \langle a, x_l - x_b \rangle]_+$$

where $[t]_+ = \max\{t, 0\}$ denotes the positive part. Then F is convex in X, and for a pair (a, b) we have $F(X; (a, b)) = 0$ if and only if the classifer represented by X

has a *large margin*, meaning that

$$\langle a, x_b \rangle \geqslant \langle a, x_l \rangle + 1 \text{ for all } l \neq b.$$

In this example, we have a sample of $N = 7291$ digits $(a_i, b_i) \in \mathbb{R}^n \times \{1, \ldots, k\}$, and we compare the performance of stochastic subgradient descent to standard subgradient descent for solving the problem

$$\text{minimize } f(X) = \frac{1}{N} \sum_{i=1}^{N} F(X; (a_i, b_i)) \text{ subject to } \|X\|_{\text{Fr}} \leqslant R \tag{3.4.6}$$

where $R = 40$. We perform stochastic gradient descent using stepsizes $\alpha_k = \alpha_1/\sqrt{k}$, where $\alpha_1 = R/M$ and $M^2 = \frac{1}{N}\sum_{i=1}^{N} \|a_i\|_2^2$ (this is an approximation to the Lipschitz constant of f). For our stochastic gradient oracle, we select an index $i \in \{1, \ldots, N\}$ uniformly at random, then take $g \in \partial_X F(X; (a_i, b_i))$. For the standard subgradient method, we also perform projected subgradient descent, where we compute subgradients by taking $g_i \in \partial F(X; (a_i, b_i))$ and setting $g = \frac{1}{N}\sum_{i=1}^{N} g_i \in \partial f(X)$. We use an identical stepsize strategy of setting $\alpha_k = \alpha_1/\sqrt{k}$, but use the five stepsizes $\alpha_1 = 10^{-j} R/M$ for $j \in \{-2, -1, \ldots, 2\}$. We plot the results of this experiment in Figure 3.4.5, showing the optimality gap (vertical axis) plotted against the number of matrix-vector products $X^\top a$ computed, normalized by $N = 7291$. The plot makes clear that computing the entire subgradient $\partial f(X)$ is wasteful: the non-stochastic methods' convergence, in terms of iteration count, is potentially faster than that for the stochastic method, but the large (7291×) per-iteration speedup the stochastic method enjoys because of its random sampling yields substantially better performance. Though we do not demonstrate this in the figure, this benefit remains typically true even across a range of stepsize choices, suggesting the benefits of stochastic gradient methods in stochastic programming problems such as problem (3.4.6). ◊

Convergence guarantees We now turn to guarantees of convergence for the stochastic subgradient method. As in our analysis of the projected subgradient method, we assume that C is compact and there is some $R < \infty$ such that $\|x^\star - x\|_2 \leqslant R$ for all $x \in C$, that projections π_C are efficiently computable, and that for all $x \in C$ we have the bound $\mathbb{E}[\|g(x, S)\|_2^2] \leqslant M^2$ for our stochastic oracle g. (The oracle's noise S may depend on the previous iterates, but we always have the unbiased condition $\mathbb{E}[g(x, S)] \in \partial f(x)$.)

Theorem 3.4.7. *Let the conditions of the preceding paragraph hold and let $\alpha_k > 0$ be a non-increasing sequence of stepsizes. Let $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$. Then*

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leqslant \frac{R^2}{2K\alpha_K} + \frac{1}{2K}\sum_{k=1}^{K} \alpha_k M^2.$$

Proof. The analysis is quite similar to our previous analyses, in that we simply expand the error $\|x_{k+1} - x^\star\|_2^2$. Let use define $f'(x) := \mathbb{E}[g(x, S)] \in \partial f(x)$ to be the expected subgradient returned by the stochastic gradient oracle, and let $\xi_k = g_k - f'(x_k)$ be the error in the kth subgradient. Then

$$\begin{aligned}\frac{1}{2}\left\|x_{k+1}-x^\star\right\|_2^2 &= \frac{1}{2}\left\|\pi_C(x_k-\alpha_k g_k)-x^\star\right\|_2^2\\ &\leqslant \frac{1}{2}\left\|x_k-\alpha_k g_k-x^\star\right\|_2^2\\ &= \frac{1}{2}\left\|x_k-x^\star\right\|_2^2-\alpha_k\left\langle g_k, x_k-x^\star\right\rangle+\frac{\alpha_k^2}{2}\left\|g_k\right\|_2^2,\end{aligned}$$

as in the proof of Theorems 3.2.7 and 3.3.4. Now, we add and subtract $\alpha_k \langle f'(x_k), x_k - x^\star\rangle$, which gives

$$\begin{aligned}\frac{1}{2}\left\|x_{k+1}-x^\star\right\|_2^2 &\leqslant \frac{1}{2}\left\|x_k-x^\star\right\|_2^2-\alpha_k\left\langle f'(x_k), x_k-x^\star\right\rangle+\frac{\alpha_k^2}{2}\left\|g_k\right\|_2^2-\alpha_k\left\langle \xi_k, x_k-x^\star\right\rangle\\ &\leqslant \frac{1}{2}\left\|x_k-x^\star\right\|_2^2-\alpha_k[f(x_k)-f(x^\star)]+\frac{\alpha_k^2}{2}\left\|g_k\right\|_2^2-\alpha_k\left\langle \xi_k, x_k-x^\star\right\rangle,\end{aligned}$$

where we have used the standard first-order convexity inequality.

Except for the error term $\langle \xi_k, x_k - x^\star\rangle$, the proof is completely identical to that of Theorem 3.3.4. Indeed, dividing each side of the preceding display by α_k and rearranging, we have

$$f(x_k)-f(x^\star) \leqslant \frac{1}{2\alpha_k}\left(\left\|x_k-x^\star\right\|_2^2-\left\|x_{k+1}-x^\star\right\|_2^2\right)+\frac{\alpha_k}{2}\left\|g_k\right\|_2^2-\left\langle \xi_k, x_k-x^\star\right\rangle.$$

Summing this inequality, as in the proof of Theorem 3.3.4 following inequality (3.3.5), yields that

$$\sum_{k=1}^{K}[f(x_k)-f(x^\star)] \leqslant \frac{R^2}{2\alpha_K}+\frac{1}{2}\sum_{k=1}^{K}\alpha_k\left\|g_k\right\|_2^2-\sum_{k=1}^{K}\left\langle \xi_k, x_k-x^\star\right\rangle. \tag{3.4.8}$$

The inequality (3.4.8) is the basic inequality from which all our subsequent convergence guarantees follow.

For this theorem, we need only take expectations, realizing that

$$\begin{aligned}\mathbb{E}[\langle \xi_k, x_k-x^\star\rangle] &= \mathbb{E}\left[\mathbb{E}[\left\langle g(x_k)-f'(x_k), x_k-x^\star\right\rangle \mid x_k]\right]\\ &= \mathbb{E}\Big[\langle\underbrace{\mathbb{E}[g(x_k)\mid x_k]}_{=f'(x_k)}-f'(x_k), x_k-x\rangle\Big]=0.\end{aligned}$$

Thus we obtain

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k)-f(x^\star))\right] \leqslant \frac{R^2}{2\alpha_K}+\frac{1}{2}\sum_{k=1}^{K}\alpha_k M^2$$

once we realize that $\mathbb{E}[\|g_k\|_2^2] \leqslant M^2$, which gives the desired result. □

Theorem 3.4.7 makes it clear that, in expectation, we can achieve the same convergence guarantees as in the non-noisy case. This does not mean that stochastic subgradient methods are always as good as non-stochastic methods, but it does show the robustness of the subgradient method even to substantial noise. So while the subgradient method is *very* slow, its slowness comes with the benefit that it can handle large amounts of noise.

We now provide a few corollaries on the convergence of stochastic gradient descent. For background on probabilistic modes of convergence, see Appendix A.2.

Corollary 3.4.9. *Let the conditions of Theorem 3.4.7 hold, and let* $\alpha_k = R/M\sqrt{k}$ *for each* k. *Then*

$$\mathbb{E}[f(\overline{x}_K)] - f(x^\star) \leqslant \frac{3RM}{2\sqrt{K}}$$

for all $K \in \mathbb{N}$.

The proof of the corollary is identical to that of Corollary 3.3.6 for the projected gradient method, once we substitute $\alpha = R/M$ in the bound. We can also obtain convergence in probability of the iterates more generally.

Corollary 3.4.10. *Let* α_k *be non-summable but convergent to zero, that is,* $\sum_{k=1}^{\infty} \alpha_k = \infty$ *and* $\alpha_k \to 0$. *Then* $f(\overline{x}_K) - f(x^\star) \overset{p}{\to} 0$ *as* $K \to \infty$, *that is, for all* $\epsilon > 0$ *we have*

$$\limsup_{k \to \infty} \mathbb{P}\left(f(\overline{x}_k) - f(x^\star) \geqslant \epsilon\right) = 0.$$

The above corollaries guarantee convergence of the iterates in expectation and with high probability, but sometimes it is advantageous to give finite sample guarantees of convergence with high probability. We can do this under somewhat stronger conditions on the subgradient noise sequence and using the Azuma-Hoeffding inequality (Theorem A.2.5 in Appendix A.2), which we present now.

Theorem 3.4.11. *In addition to the conditions of Theorem 3.4.7, assume that* $\|g\|_2 \leqslant M$ *for all stochastic subgradients* g. *Then for any* $\epsilon > 0$,

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{R^2}{2K\alpha_K} + \sum_{k=1}^{K} \frac{\alpha_k}{2} M^2 + \frac{RM}{\sqrt{K}}\epsilon$$

with probability at least $1 - e^{-\frac{1}{2}\epsilon^2}$.

Written differently, we see that by taking $\alpha_k = \frac{R}{\sqrt{k}M}$ and setting $\delta = e^{-\frac{1}{2}\epsilon^2}$, we have

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{3MR}{\sqrt{K}} + \frac{MR\sqrt{2\log\frac{1}{\delta}}}{\sqrt{K}}$$

with probability at least $1 - \delta$. That is, we have convergence of $\mathcal{O}(MR/\sqrt{K})$ with high probability.

Before providing the proof proper, we discuss two examples in which the boundedness condition holds. Recall from Lecture 2 that a convex function f is M-Lipschitz if and only if $\|g\|_2 \leqslant M$ for all $g \in \partial f(x)$ and $x \in \mathbb{R}^n$, so Theorem 3.4.11 requires that the random functions $F(\cdot; S)$ are Lipschitz over the domain C. Our robust regression and multiclass support vector machine examples both satisfy the conditions of the theorem so long as the data is bounded. More precisely, for the robust regression problem (3.2.12) with loss $F(x; (a, b)) = |\langle a, x\rangle - b|$, we have $\partial F(x;(a,b)) = a\,\mathrm{sign}(\langle a, x\rangle - b)$ so that the condition $\|g\|_2 \leqslant M$ holds if and only if $\|a\|_2 \leqslant M$. For the multiclass hinge loss problem (3.4.6), with $F(X;(a,b)) = \sum_{l \neq b} [1 + \langle a, x_l - x_b\rangle]_+$, Exercise 5 develops the subgradient calculations, but again, we have the boundedness of $\partial_X F(X;(a,b))$ if and only if $a \in \mathbb{R}^n$ is bounded.

Proof. We begin with the basic inequality of Theorem 3.4.7, inequality (3.4.8). We see that we would like to bound the probability that

$$\sum_{k=1}^{K} \langle \xi_k, x^\star - x_k\rangle$$

is large. First, we note that the iterate x_k is a function of $\xi_1, \ldots, \xi_{k-1}$, and we have the conditional expectation

$$\mathbb{E}[\xi_k \mid \xi_1, \ldots, \xi_{k-1}] = \mathbb{E}[\xi_k \mid x_k] = 0.$$

Moreover, using the boundedness assumption that $\|g\|_2 \leqslant M$, we have $\|\xi_k\|_2 = \|g_k - f'(x_k)\|_2 \leqslant 2M$ and

$$|\langle \xi_k, x_k - x^\star\rangle| \leqslant \|\xi_k\|_2 \, \|x_k - x^\star\|_2 \leqslant 2MR.$$

Thus, the sequence $\sum_{k=1}^{K} \langle \xi_k, x_k - x^\star\rangle$ is a bounded difference martingale sequence, and we may apply Azuma's inequality (Theorem A.2.5), which gurantees

$$\mathbb{P}\left(\sum_{k=1}^{K} \langle \xi_k, x^\star - x_k\rangle \geqslant t\right) \leqslant \exp\left(-\frac{t^2}{2KM^2R^2}\right)$$

for all $t \geqslant 0$. Substituting $t = MR\sqrt{K}\epsilon$, we obtain that

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} \langle \xi_k, x^\star - x_k\rangle \geqslant \frac{\epsilon MR}{\sqrt{K}}\right) \leqslant \exp\left(-\frac{\epsilon^2}{2}\right),$$

as desired. □

Summarizing the results of this section, we see a number of consequences. First, stochastic gradient methods guarantee that after $\mathcal{O}(1/\epsilon^2)$ iterations, we have error at most $f(x) - f(x^\star) = \mathcal{O}(\epsilon)$. Secondly, this convergence is (at least to the order in ϵ) the same as in the non-noisy case; that is, stochastic gradient methods are robust enough to noise that their convergence is hardly affected by it. In addition to this, they are often applicable in situations in which we cannot even evaluate the objective f, whether for computational reasons or because we do not have access to it, as in statistical problems. This robustness to noise and good performance has led to wide adoption of subgradient-like methods as the *de facto* choice for many large-scale data-based optimization problems. In the coming sections, we give further discussion of the optimality of stochastic gradient methods, showing that—roughly—when we have access only to noisy data, it is impossible to solve (certain) problems to accuracy better than ϵ given $1/\epsilon^2$ data points;

thus, using more expensive but accurate optimization methods may have limited benefit (though there may still be some benefit practically!).

Notes and further reading Our treatment in this chapter borrows from a number of resources. The two heaviest are the lecture notes for Stephen Boyd's Stanford's EE364b course [10, 11] and Polyak's *Introduction to Optimization* [47]. Our guarantees of high probability convergence are similar to those originally developed by Cesa-Bianchi et al. [16] in the context of online learning, which Nemirovski et al. [40] more fully develop. More references on subgradient methods include the lecture notes of Nemirovski [43] and Nesterov [44].

A number of extensions of (stochastic) subgradient methods are possible, including to online scenarios in which we observe streaming sequences of functions [25, 63]; our analysis in this section follows closely that of Zinkevich [63]. The classic paper of Polyak and Juditsky [48] shows that stochastic gradient descent methods, coupled with averaging, can achieve asymptotically optimal rates of convergence even to constant factors. Recent work in machine learning by a number of authors [18, 32, 53] has shown how to leverage the structure of optimization problems based on *finite sums*, that is, when $f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$, to develop methods that achieve convergence rates similar to those of interior point methods but with iteration complexity close to stochastic gradient methods.

4. The Choice of Metric in Subgradient Methods

Lecture Summary: Standard subgradient and projected subgradient methods are inherently Euclidean—they rely on measuring distances using Euclidean norms, and their updates are based on Euclidean steps. In this lecture, we study methods for more carefully choosing the metric, giving rise to mirror descent, also known as non-Euclidean subgradient descent, as well as methods for adapting the updates performed to the problem at hand. By more carefully studying the geometry of the optimization problem being solved, we show how faster convergence guarantees are possible.

4.1. Introduction In the previous lecture, we studied projected subgradient methods for solving the problem (2.1.1) by iteratively updating $x_{k+1} = \pi_C(x_k - \alpha_k g_k)$, where π_C denotes Euclidean projection. The convergence of these methods, as exemplified by Corollaries 3.2.8 and 3.4.9, scales as

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{MR}{\sqrt{K}} = \mathcal{O}(1)\frac{\mathrm{diam}(C)\mathrm{Lip}(f)}{\sqrt{K}}, \tag{4.1.1}$$

where $R = \sup_{x \in C} \|x - x^\star\|_2$ and M is the Lipschitz constant of f over the set C with respect to the ℓ_2-norm,

$$M = \sup_{x \in C} \sup_{g \in \partial f(x)} \left\{ \|g\|_2 = \left(\sum_{j=1}^{n} g_j^2 \right)^{\frac{1}{2}} \right\}.$$

The convergence guarantee (4.1.1) reposes on Euclidean measures of scale—the diameter of C and norm of the subgradients g are both measured in ℓ_2-norm. It is thus natural to ask if we can develop methods whose convergence rates depend on other measures of scale of f and C, obtainining better problem-dependent behavior and geometry. With that in mind, in this lecture we derive a number of methods that use either non-Euclidean or adaptive updates to better reflect the geometry of the underlying optimization problem.

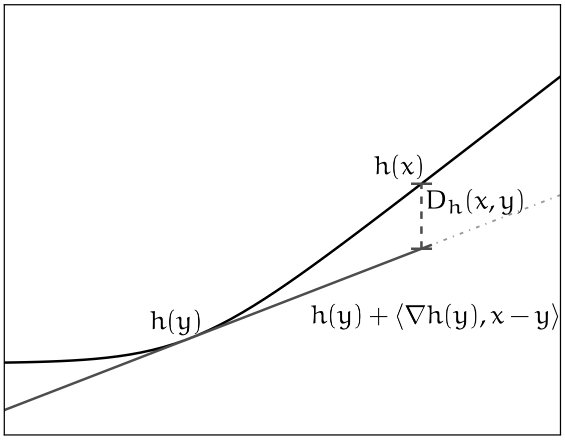

FIGURE 4.2.1. Bregman divergence $D_h(x, y)$. The bottom upper function is $h(x) = \log(1 + e^x)$, the lower (linear) is the linear approximation $x \mapsto h(y) + \langle \nabla h(y), x - y \rangle$ to h at y.

4.2. Mirror Descent Methods Our first set of results focuses on mirror descent methods, which modify the basic subgradient update to use a different distance-measuring function rather than the squared ℓ_2-term. Before presenting these methods, we give a few definitions. Let h be a differentiable convex function, differentiable on C. The *Bregman divergence* associated with h is defined as

$$D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle . \tag{4.2.2}$$

The divergence D_h is always nonnegative, by the standard first-order inequality for convex functions, and measures the gap between the linear approximation $h(y) + \langle \nabla h(y), x - y \rangle$ for $h(x)$ taken from the point y and the value $h(x)$ at x. See Figure 4.2.1. As one standard example, if we take $h(x) = \frac{1}{2} \|x\|_2^2$, then $D_h(x, y) = \frac{1}{2} \|x - y\|_2^2$. A second common example follows by taking the entropy functional $h(x) = \sum_{j=1}^n x_j \log x_j$, restricting x to the probability simplex (i.e. $x \succeq 0$ and $\sum_j x_j = 1$). We then have $D_h(x, y) = \sum_{j=1}^n x_j \log \frac{x_j}{y_j}$, the entropic or Kullback-Leibler divergence.

Because the quantity (4.2.2) is always non-negative and convex in its first argument, it is natural to treat it as a distance-like function in the development of optimization procedures. Indeed, by recalling the updates (3.2.3) and (3.3.2), by analogy we consider the method

i. Compute subgradient $g_k \in \partial f(x_k)$

ii. Perform update

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in C} \left\{ f(x_k) + \langle g_k, x - x_k \rangle + \frac{1}{\alpha_k} D_h(x, x_k) \right\} \\ &= \operatorname*{argmin}_{x \in C} \left\{ \langle g_k, x \rangle + \frac{1}{\alpha_k} D_h(x, x_k) \right\}. \end{aligned} \tag{4.2.3}$$

This scheme is the *mirror descent method*. Thus, each differentiable convex function h gives a new optimization scheme, where we often attempt to choose h to better match the geometry of the underlying constraint set C.

To this point, we have been vague about the "geometry" of the constraint set, so we attempt to be somewhat more concrete. We say that h is λ-*strongly convex* over C with respect to the norm $\|\cdot\|$ if

$$h(y) \geqslant h(x) + \langle \nabla h(x), y - x \rangle + \frac{\lambda}{2} \|x - y\|^2 \text{ for all } x, y \in C.$$

Importantly, this norm need not be the typical ℓ_2 or Euclidean norm. Then our goal is, roughly, to choose a strongly convex function h so that the diameter of C is small in the norm $\|\cdot\|$ with respect to which h is strongly convex (as we see presently, an analogue of the bound (4.1.1) holds). In the standard updates (3.2.3) and (3.3.2), we use the squared Euclidean norm to trade between making progress on the linear approximation $x \mapsto f(x_k) + \langle g_k, x - x_k \rangle$ and making sure the approximation is reasonable—we *regularize* progress. Thus it is natural to ask that the function h we use provide a similar type of regularization, and consequently, we will require that the function h be 1-strongly convex (usually shortened to the unqualified strongly convex) with respect to some norm $\|\cdot\|$ over the constraint set C in the mirror descent method (4.2.3).[4] Note that strong convexity of h is equivalent to

$$D_h(x, y) \geqslant \frac{1}{2} \|x - y\|^2 \text{ for all } x, y \in C.$$

Examples of mirror descent Before analyzing the method (4.2.3), we present a few examples, showing the updates that are possible as well as verifying that the associated divergence is appropriately strongly convex. One of the nice consequences of allowing different divergence measures D_h, as opposed to only the Euclidean divergence, is that they often yield cleaner or simpler updates.

Example 4.1 (Gradient descent is mirror descent)**:** Let $h(x) = \frac{1}{2}\|x\|_2^2$. Then $\nabla h(y) = y$, and

$$D_h(x, y) = \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\|y\|_2^2 - \langle y, x - y \rangle = \frac{1}{2}\|x\|_2^2 + \frac{1}{2}\|y\|_2^2 - \langle x, y \rangle = \frac{1}{2}\|x - y\|_2^2.$$

[4] This is not strictly a requirement, and sometimes it is analytitcally convenient to avoid this, but our analysis is simpler when h is strongly convex.

Thus, substituting into the update (4.2.3), we see the choice $h(x) = \frac{1}{2}\|x\|_2^2$ recovers the standard (stochastic sub)gradient method

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ \langle g_k, x \rangle + \frac{1}{2\alpha_k} \|x - x_k\|_2^2 \right\}.$$

It is evident that h is strongly convex with respect to the ℓ_2-norm for any constraint set C. ♢

Example 4.2 (Solving problems on the simplex with exponentiated gradient methods)**:** Suppose that our constraint set $C = \{x \in \mathbb{R}_+^n : \langle \mathbf{1}, x \rangle = 1\}$ is the probability simplex in $\mathbb{R}^n$. Then updates with the standard Euclidean distance are somewhat challenging—though there are efficient implementations [14, 23]—and it is natural to ask for a simpler method.

With that in mind, let $h(x) = \sum_{j=1}^n x_j \log x_j$ be the negative entropy, which is convex because it is the sum of convex functions. (The derivatives of $f(t) = t \log t$ are $f'(t) = \log t + 1$ and $f''(t) = 1/t > 0$ for $t \geqslant 0$.) Then we have

$$\begin{aligned} D_h(x, y) &= \sum_{j=1}^n \left[x_j \log x_j - y_j \log y_j - (\log y_j + 1)(x_j - y_j) \right] \\ &= \sum_{j=1}^n x_j \log \frac{x_j}{y_j} + \langle \mathbf{1}, y - x \rangle = D_{\mathrm{kl}}\left(x | y\right), \end{aligned}$$

the KL-divergence between x and y (when extended to $\mathbb{R}_+^n$, though over C we have $\langle \mathbf{1}, x - y \rangle = 0$). This gives us the form of the update (4.2.3).

Let us consider the update (4.2.3). Simplifying notation, we would like to solve

$$\text{minimize } \langle g, x \rangle + \sum_{j=1}^n x_j \log \frac{x_j}{y_j} \text{ subject to } \langle \mathbf{1}, x \rangle = 1, \; x \succeq 0.$$

We assume that the $y_j > 0$, though this is not strictly necessary. Though we have not discussed this, we write the Lagrangian for this problem by introducing Lagrange multipliers $\tau \in \mathbb{R}$ for the equality constraint $\langle \mathbf{1}, x \rangle = 1$ and $\lambda \in \mathbb{R}_+^n$ for the inequality $x \succeq 0$. Then we obtain Lagrangian

$$\mathcal{L}(x, \tau, \lambda) = \langle g, x \rangle + \sum_{j=1}^n \left[x_j \log \frac{x_j}{y_j} + \tau x_j - \lambda_j x_j \right] - \tau.$$

Minimizing out x to find the appropriate form for the solution, we take derivatives with respect to x and set them to zero to find

$$0 = \frac{\partial}{\partial x_j} \mathcal{L}(x, \tau, \lambda) = g_j + \log x_j + 1 - \log y_j + \tau - \lambda_j,$$

or

$$x_j(\tau, \lambda) = y_j \exp(-g_j - 1 - \tau + \lambda_j).$$

We may take $\lambda_j = 0$, as the latter expression yields all positive x_j, and to satisfy the constraint that $\sum_j x_j = 1$, we set $\tau = \log(\sum_j y_j e^{-g_j}) - 1$. Thus we have the update

$$x_i = \frac{y_i \exp(-g_i)}{\sum_{j=1}^n y_j \exp(-g_j)}.$$

Rewriting this in terms of the precise update at time k for the mirror descent method, we have for each coordinate i of iterate $k+1$ of the method that

$$x_{k+1,i} = \frac{x_{k,i} \exp(-\alpha_k g_{k,i})}{\sum_{j=1}^n x_{k,j} \exp(-\alpha_k g_{k,j})}. \tag{4.2.4}$$

This is the so-called *exponentiated gradient* update, also known as *entropic mirror descent*.

Later, after stating and proving our main convergence theorems, we will show that the negative entropy is strongly convex with respect to the ℓ_1-norm, meaning that our coming convergence guarantees apply. $\Diamond$

Example 4.3 (Using ℓ_p-norms)**:** As a final example, we consider using squared ℓ_p-norms for our distance-generating function h. These have nice robustness properties, and are also finite on any compact set (unlike the KL-divergence of Example 4.2). Indeed, let $p \in (1, 2]$, and define $h(x) = \frac{1}{2(p-1)} \|x\|_p^2$. We claim without proof that h is strongly convex with respect to the ℓ_p-norm, that is,

$$D_h(x, y) \geqslant \frac{1}{2} \|x - y\|_p^2 .$$

(See, for example, the thesis of Shalev-Shwartz [51] and Question 9 in the exercises. This inequality fails for powers other than 2 as well as for $p > 2$.)

We do not address the constrained case here, assuming instead that $C = \mathbb{R}^n$. In this case, we have

$$\nabla h(x) = \frac{1}{p-1} \|x\|_p^{2-p} \left[\text{sign}(x_1)|x_1|^{p-1} \ \cdots \ \text{sign}(x_n)|x_n|^{p-1}\right]^\top .$$

Now, if we define the function $\phi(x) = (p-1)\nabla h(x)$, then a calculation verifies that the function $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ defined coordinate-wise by

$$\phi_j(x) = \|x\|_p^{2-p} \,\text{sign}(x_j)|x_j|^{p-1} \text{ and } \varphi_j(y) = \|y\|_q^{2-q} \,\text{sign}(y_j)|y_j|^{q-1},$$

where $\frac{1}{p} + \frac{1}{q} = 1$, satisfies $\varphi(\phi(x)) = x$, that is, $\varphi = \phi^{-1}$ (and similarly $\phi = \varphi^{-1}$). Thus, the mirror descent update (4.2.3) when $C = \mathbb{R}^n$ becomes the somewhat more complex

$$x_{k+1} = \varphi(\phi(x_k) - \alpha_k(p-1)g_k) = (\nabla h)^{-1}(\nabla h(x_k) - \alpha_k g_k). \tag{4.2.5}$$

The second form of the update (4.2.5), that is, that involving the inverse of the gradient mapping $(\nabla h)^{-1}$, holds more generally, that is, for any strictly convex and differentiable h. This is the original form of the mirror descent update (4.2.3), and it justifies the name *mirror* descent, as the gradient is "mirrored" through the distance-generating function h and back again. Nonetheless, we find the modeling perspective of (4.2.3) somewhat easier to explain.

We remark in passing that while constrained updates are somewhat more challenging for this case, a few are efficiently solvable. For example, suppose that

$C = \{x \in \mathbb{R}^n_+ : \langle \mathbf{1}, x\rangle = 1\}$, the probability simplex. In this case, the update with ℓ_p-norms becomes a problem of solving

$$\operatorname*{minimize}_x \ \langle v, x\rangle + \frac{1}{2}\left\|x\right\|_p^2 \ \text{ subject to } \langle \mathbf{1}, x\rangle = 1, \ x \succeq 0,$$

where $v = \alpha_k(p-1)g_k - \phi(x_k)$, and φ and ϕ are defined as above. An analysis of the Karush-Kuhn-Tucker conditions for this problem (omitted) yields that the solution to the problem is given by finding the $t^\star \in \mathbb{R}$ such that

$$\sum_{j=1}^n \varphi_j(\left[-v_j + t^\star\right]_+) = 1 \text{ and setting } x_j = \varphi(\left[-v_j + t^\star\right]_+).$$

Because φ is increasing in its argument with $\varphi(0) = 0$, this $t^\star$ can be found to accuracy ϵ in time $\mathcal{O}(n \log \frac{1}{\epsilon})$ by binary search. $\diamond$

Convergence guarantees With the mirror descent method described, we now provide an analysis of its convergence behavior. In this case, the analysis is somewhat more complex than that for the subgradient, projected subgradient, and stochastic subgradient methods, as we cannot simply expand the distance $\left\|x_{k+1} - x^\star\right\|_2^2$. Thus, we give a variant proof that relies on the optimality conditions for convex optimization problems, as well as a few tricks involving norms and their dual norms. Recall that we assume that the function h is strongly convex with respect to some norm $\left\|\cdot\right\|$, and that the associated dual norm $\left\|\cdot\right\|_*$ is defined by

$$\left\|y\right\|_* := \sup_{x:\left\|x\right\| \leqslant 1} \langle y, x\rangle .$$

Theorem 4.2.6. *Let $\alpha_k > 0$ be any sequence of non-increasing stepsizes and the above assumptions hold. Let x_k be generated by the mirror descent iteration (4.2.3). If $D_h(x, x^\star) \leqslant R^2$ for all $x \in C$, then for all $K \in \mathbb{N}$*

$$\sum_{k=1}^K [f(x_k) - f(x^\star)] \leqslant \frac{1}{\alpha_K} R^2 + \sum_{k=1}^K \frac{\alpha_k}{2} \left\|g_k\right\|_*^2 .$$

If $\alpha_k \equiv \alpha$ is constant, then for all $K \in \mathbb{N}$

$$\sum_{k=1}^K [f(x_k) - f(x^\star)] \leqslant \frac{1}{\alpha} D_h(x^\star, x_1) + \frac{\alpha}{2} \sum_{k=1}^K \left\|g_k\right\|_*^2 .$$

As an immediate consequence of this theorem, we see that if $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$ or $\overline{x}_K = \operatorname{argmin}_{x_k} f(x_k)$ and we have the gradient bound $\left\|g\right\|_* \leqslant M$ for all $g \in \partial f(x)$ for $x \in C$, then (say, in the second case) convexity implies

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{1}{K\alpha} D_h(x^\star, x_1) + \frac{\alpha}{2} M^2. \tag{4.2.7}$$

By comparing with the bound (3.2.9), we see that the mirror descent (non-Euclidean gradient descent) method gives roughly the same type of convergence guarantees as standard subgradient descent. Roughly we expect the following type of behavior with a fixed stepsize: a rate of convergence of roughly $1/\alpha K$ until we are within a radius α of the optimum, after which mirror descent and subgradient descent essentially jam—they just jump back and forth near the optimum.

Proof. We begin by considering the progress made in a single update of x_k, but whereas our previous proofs all began with a Lyapunov function for the distance $\|x_k - x^\star\|_2$, we use function value gaps instead of the distance to optimality. Using the first order convexity inequality—i.e. the definition of a subgradient—we have

$$f(x_k) - f(x^\star) \leqslant \langle g_k, x_k - x^\star \rangle .$$

The idea is to show that replacing x_k with x_{k+1} makes the term $\langle g_k, x_k - x^\star \rangle$ small because of the definition of x_{k+1}, but x_k and x_{k+1} are close together so that this is not much of a difference.

First, we add and subtract $\langle g_k, x_{k+1} \rangle$ to obtain

$$f(x_k) - f(x^\star) \leqslant \langle g_k, x_{k+1} - x^\star \rangle + \langle g_k, x_k - x_{k+1} \rangle . \tag{4.2.8}$$

Now, we use the the first-order necessary and sufficient conditions for optimality of convex optimization problems given by Theorem 2.4.11. Because x_{k+1} solves problem (4.2.3), we have

$$\left\langle g_k + \alpha_k^{-1} \left(\nabla h(x_{k+1}) - \nabla h(x_k) \right), x - x_{k+1} \right\rangle \geqslant 0 \text{ for all } x \in C.$$

In particular, this inequality holds for $x = x^\star$, and substituting into expression (4.2.8) yields

$$f(x_k) - f(x^\star) \leqslant \frac{1}{\alpha_k} \langle \nabla h(x_{k+1}) - \nabla h(x_k), x^\star - x_{k+1} \rangle + \langle g_k, x_k - x_{k+1} \rangle .$$

We now use two tricks: an algebraic identity involving D_h and the Fenchel-Young inequality. By algebraic manipulations, we have that

$$\langle \nabla h(x_{k+1}) - \nabla h(x_k), x^\star - x_{k+1} \rangle = D_h(x^\star, x_k) - D_h(x^\star, x_{k+1}) - D_h(x_{k+1}, x_k).$$

Substituting into the preceding display, we have

(4.2.9)

$$f(x_k) - f(x^\star) \leqslant \frac{1}{\alpha_k} \left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1}) - D_h(x_{k+1}, x_k) \right] + \langle g_k, x_k - x_{k+1} \rangle .$$

The second insight is that the subtraction of $D_h(x_{k+1}, x_k)$ allows us to cancel some of $\langle g_k, x_k - x_{k+1} \rangle$. To see this, recall the Fenchel-Young inequality, which states that

$$\langle x, y \rangle \leqslant \frac{\eta}{2} \|x\|^2 + \frac{1}{2\eta} \|y\|_*^2$$

for any pair of dual norms $(\|\cdot\|, \|\cdot\|_*)$ and any $\eta > 0$. To see this, note that by definition of the dual norm, we have $\langle x, y \rangle \leqslant \|x\| \|y\|_*$, and for any constants $a, b \in \mathbb{R}$ and $\eta > 0$, we have $0 \leqslant \frac{1}{2}(\eta^{\frac{1}{2}} a - \eta^{-\frac{1}{2}} b)^2 = \frac{\eta}{2} a^2 + \frac{1}{2\eta} b^2 - ab$, so that $\|x\| \|y\|_* \leqslant \frac{\eta}{2} \|x\|^2 + \frac{1}{2\eta} \|y\|_*^2$. In particular, we have

$$\langle g_k, x_k - x_{k+1} \rangle \leqslant \frac{\alpha_k}{2} \|g_k\|_*^2 + \frac{1}{2\alpha_k} \|x_k - x_{k+1}\|^2 .$$

The strong convexity assumption on h guarantees $D_h(x_k, x_{k+1}) \geqslant \frac{1}{2}\|x_k - x_{k+1}\|^2$, or that

$$-\frac{1}{\alpha_k} D_h(x_{k+1}, x_k) + \langle g_k, x_k - x_{k+1}\rangle \leqslant \frac{\alpha_k}{2}\|g_k\|_*^2.$$

Substituting this into inequality (4.2.9), we have

$$f(x_k) - f(x^\star) \leqslant \frac{1}{\alpha_k}\left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1})\right] + \frac{\alpha_k}{2}\|g_k\|_*^2. \tag{4.2.10}$$

This inequality should look similar to inequality (3.3.5) in the proof of Theorem 3.3.4 on the projected subgradient method in Lecture 3. Indeed, using that $D_h(x^\star, x_k) \leqslant R^2$ by assumption, an identical derivation to that in Theorem 3.3.4 gives the first result of this theorem. For the second when the stepsize is fixed, note that

$$\begin{aligned}\sum_{k=1}^K [f(x_k) - f(x^\star)] &\leqslant \sum_{k=1}^K \frac{1}{\alpha}\left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1})\right] + \sum_{k=1}^K \frac{\alpha}{2}\|g_k\|_*^2 \\ &= \frac{1}{\alpha}\left[D_h(x^\star, x_1) - D_h(x^\star, x_{K+1})\right] + \sum_{k=1}^K \frac{\alpha}{2}\|g_k\|_*^2,\end{aligned}$$

which is the second result. □

We briefly provide a few remarks before moving on. As a first remark, all of the preceding analysis carries through in an almost completely identical fashion in the stochastic case. We state the most basic result, as the extension from Section 3.4 is essentially straightforward.

Corollary 4.2.11. *Let the conditions of Theorem 4.2.6 hold, except that instead of receiving a vector $g_k \in \partial f(x_k)$ at iteration k, the vector g_k is a stochastic subgradient satisfying $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$. Then for any non-increasing stepsize sequence α_k (where α_k may be chosen dependent on $g_1, \ldots, g_k$),*

$$\mathbb{E}\left[\sum_{k=1}^K (f(x_k) - f(x^\star))\right] \leqslant \mathbb{E}\left[\frac{R^2}{\alpha_K} + \sum_{k=1}^K \frac{\alpha_k}{2}\|g_k\|_*^2\right].$$

Proof. We sketch the result. The proof is identical to that for Theorem 4.2.6, except that we replace g_k with the particular vector $f'(x_k)$ satisfying $\mathbb{E}[g_k \mid x_k] = f'(x_k) \in \partial f(x_k)$. Then

$$f(x_k) - f(x^\star) \leqslant \left\langle f'(x_k), x_k - x^\star\right\rangle = \langle g_k, x_k - x^\star\rangle + \left\langle f'(x_k) - g_k, x_k - x^\star\right\rangle,$$

and an identical derivation yields the following analogue of inequality (4.2.10):

$$f(x_k) - f(x^\star) \leqslant \frac{1}{\alpha_k}\left[D_h(x^\star, x_k) - D_h(x^\star, x_{k+1})\right] + \frac{\alpha_k}{2}\|g_k\|_*^2 + \left\langle f'(x_k) - g_k, x_k - x^\star\right\rangle.$$

This inequality holds regardless of how we choose α_k. Moreover, by iterating expectations, we have

$$\mathbb{E}[\left\langle f'(x_k) - g_k, x_k - x^\star\right\rangle] = \mathbb{E}[\left\langle f'(x_k) - \mathbb{E}[g_k \mid x_k], x_k - x^\star\right\rangle] = 0,$$

which gives the corollary once we follow an identical derivation to Theorem 4.2.6. □

Thus, if we have the bound $\mathbb{E}[\|g\|_*^2] \leqslant M^2$ for all stochastic subgradients, then taking $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$ and $\alpha_k = R/M\sqrt{k}$, then

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leqslant \frac{RM}{\sqrt{K}} + \frac{R \max_k \mathbb{E}[\|g_k\|_*^2]}{M} \sum_{k=1}^K \frac{1}{2\sqrt{k}} \leqslant 3\frac{RM}{\sqrt{K}} \tag{4.2.12}$$

where we have used that $\mathbb{E}[\|g\|_*^2] \leqslant M^2$ and $\sum_{k=1}^K k^{-\frac{1}{2}} \leqslant 2\sqrt{K}$.

In addition, we can provide concrete convergence guarantees for a few methods, revisiting our earlier examples. We begin with Example 4.2, exponentiated gradient descent.

Corollary 4.2.13. *Let* $C = \{x \in \mathbb{R}_+^n : \langle \mathbf{1}, x\rangle = 1\}$, *and take* $h(x) = \sum_{j=1}^n x_j \log x_j$, *the negative entropy. Let* $x_1 = \frac{1}{n}\mathbf{1}$, *the vector whose entries are each* $1/n$. *Then if* $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$, *the exponentiated gradient method* (4.2.4) *with fixed stepsize* α *guarantees*

$$f(\overline{x}_K) - f(x^\star) \leqslant \frac{\log n}{K\alpha} + \frac{\alpha}{2K}\sum_{k=1}^K \|g_k\|_\infty^2 .$$

Proof. To apply Theorem 4.2.6, we must show that the negative entropy h is strongly convex with respect to the ℓ_1-norm, whose dual norm is the ℓ_∞-norm. By a Taylor expansion, we know that for any $x, y \in C$, we have

$$h(x) = h(y) + \langle \nabla h(y), x - y\rangle + \frac{1}{2}(x-y)^\top \nabla^2 h(\widetilde{x})(x-y)$$

for some $\widetilde{x}$ between x and y, that is, $\widetilde{x} = tx + (1-t)y$ for some $t \in [0,1]$. Calculating these quantities, this is equivalent to

$$\begin{aligned} D_{\rm kl}(x|y) = D_h(x,y) &= \frac{1}{2}(x-y)^\top \operatorname{diag}\left(\frac{1}{\widetilde{x}_1}, \dots, \frac{1}{\widetilde{x}_n}\right)(x-y) \\ &= \frac{1}{2}\sum_{j=1}^n \frac{(x_j - y_j)^2}{\widetilde{x}_j}. \end{aligned}$$

Using the Cauchy-Schwarz inequality and the fact that $\widetilde{x} \in C$, we have

$$\|x-y\|_1 = \sum_{j=1}^n |x_j - y_j| = \sum_{j=1}^n \sqrt{\widetilde{x}_j}\frac{|x_j - y_j|}{\sqrt{\widetilde{x}_j}} \leqslant \underbrace{\left(\sum_{j=1}^n \widetilde{x}_j\right)^{\frac{1}{2}}}_{=1} \left(\sum_{j=1}^n \frac{(x_j-y_j)^2}{\widetilde{x}_j}\right)^{\frac{1}{2}}.$$

That is, we have $D_{\rm kl}(x|y) = D_h(x,y) \geqslant \frac{1}{2}\|x-y\|_1^2$, and h is strongly convex with respect to the ℓ_1-norm over C.

With this strong convexity result in hand, we may apply second result of Theorem 4.2.6, achieving

$$\sum_{k=1}^K [f(x_k) - f(x^\star)] \leqslant \frac{D_{\rm kl}(x^\star|x_1)}{\alpha} + \frac{\alpha}{2}\sum_{k=1}^K \|g_k\|_\infty^2 .$$

If $x_1 = \frac{1}{n}\mathbf{1}$, then $D_{\rm kl}(x|x_1) = h(x) + \log n \leqslant \log n$, as $h(x) \leqslant 0$ for $x \in C$. Thus, dividing by K and using that $f(\overline{x}_K) \leqslant \frac{1}{K}\sum_{k=1}^K f(x_k)$ gives the corollary. □

Inspecting the guarantee Corollary 4.2.13 provides versus that guaranteed by the standard (non-stochastic) projected subgradient method (i.e. using $h(x) = \frac{1}{2}\left\|x\right\|_2^2$ as in Theorem 3.3.4) is instructive. In the case of projected subgradient descent, we have $D_h(x^\star, x) = \frac{1}{2}\left\|x^\star - x\right\|_2^2 \leqslant 1$ for all $x, x^\star \in C = \{x \in \mathbb{R}^n_+ : \langle \mathbf{1}, x\rangle = 1\}$ (and this distance is achieved). However, the dual norm to the ℓ_2-norm is ℓ_2, meaning we measure the size of the gradient terms $\left\|g_k\right\|$ in ℓ_2-norm. As $\left\|g_k\right\|_\infty \leqslant \left\|g_k\right\|_2 \leqslant \sqrt{n}\left\|g_k\right\|_\infty$, supposing that $\left\|g_k\right\|_\infty \leqslant 1$ for all k, the convergence guarantee

$$\mathcal{O}(1)\sqrt{\frac{\log n}{K}}$$

may be up to $\sqrt{n/\log n}$-times better than that guaranteed by the standard (Euclidean) projected gradient method.

Lastly, we provide a final convergence guarantee for the mirror descent method using ℓ_p-norms, where $p \in (1, 2]$. Using such norms has the benefit that D_h is bounded whenever the set C is compact—distinct from the relative entropy $D_h(x, y) = \sum_j x_j \log \frac{x_j}{y_j}$—and thus providing a nicer guarantee of convergence. Indeed, for $h(x) = \frac{1}{2}\left\|x\right\|_p^2$ we always have that

$$\begin{aligned} D_h(x,y) &= \frac{1}{2}\left\|x\right\|_p^2 - \frac{1}{2}\left\|y\right\|_p^2 - \sum_{j=1}^n \left\|y\right\|_p^{2-p} \operatorname{sign}(y_j)|y_j|^{p-1}(x_j - y_j) \\ &= \frac{1}{2}\left\|x\right\|_p^2 + \frac{1}{2}\left\|y\right\|_p^2 \underbrace{- \left\|y\right\|_p^{2-p} \sum_{j=1}^n |y_j|^{p-1} \operatorname{sign}(y_j) x_j}_{\leqslant \frac{1}{2}\left\|x\right\|_p^2 + \frac{1}{2}\left\|y\right\|_p^2} \leqslant \left\|x\right\|_p^2 + \left\|y\right\|_p^2, \end{aligned} \tag{4.2.14}$$

where the inequality uses that $q(p-1) = p$ and

$$\begin{aligned} \sum_{j=1}^n \left\|y\right\|_p^{2-p} |y_j|^{p-1}|x_j| &\leqslant \left\|y\right\|_p^{2-p}\left(\sum_{j=1}^n |y_j|^{q(p-1)}\right)^{\frac{1}{q}}\left(\sum_{j=1}^n |x_j|^p\right)^{\frac{1}{p}} \\ &= \left\|y\right\|_p^{2-p}\left\|y\right\|_p^{\frac{p}{q}}\left\|x\right\|_p = \left\|y\right\|_p\left\|x\right\|_p \leqslant \frac{1}{2}\left\|y\right\|_p^2 + \frac{1}{2}\left\|x\right\|_p^2. \end{aligned}$$

More generally, with $h(x) = \frac{1}{2}\left\|x - x_0\right\|_p^2$, we have $D_h(x, y) \leqslant \left\|x - x_0\right\|_p^2 + \left\|y - x_0\right\|_p^2$. As one example, we obtain the following corollary.

Corollary 4.2.15. *Let* $h(x) = \frac{1}{2(p-1)}\left\|x\right\|_p^2$, *where* $p = 1 + \frac{1}{\log(2n)}$, *and assume that* $C \subset \{x \in \mathbb{R}^n : \left\|x\right\|_1 \leqslant R_1\}$. *Then*

$$\sum_{k=1}^K [f(x_k) - f(x^\star)] \leqslant \frac{2R_1^2 \log(2n)}{\alpha_K} + \frac{e^2}{2}\sum_{k=1}^K \alpha_k \left\|g_k\right\|_\infty^2.$$

In particular, taking $\alpha_k = R_1\sqrt{\log(2n)/k}/e$ *and* $\overline{x}_K = \frac{1}{K}\sum_{k=1}^K x_k$ *gives*

$$f(\overline{x}_K) - f(x^\star) \leqslant 3e\frac{R_1\sqrt{\log(2n)}}{\sqrt{K}}.$$

Proof. First, we note that $h(x) = \frac{1}{2(p-1)} \|x\|_p^2$ is strongly convex with respect to the ℓ_p-norm, where $1 < p \leqslant 2$. (Recall Example 4.3 and see Exercise 9.) Moreover, we know that the dual to the ℓ_p-norm is the conjugate ℓ_q-norm with $1/p + 1/q = 1$, and thus Theorem 4.2.6 implies that

$$\sum_{k=1}^{K} [f(x_k) - f(x^\star)] \leqslant \frac{1}{\alpha_K} \sup_{x \in C} D_h(x, x^\star) + \sum_{k=1}^{K} \frac{\alpha_k}{2} \|g_k\|_q^2 .$$

Now, we use that if C is contained in the ℓ_1-ball of radius R_1, then $(p-1)D_h(x,y) \leqslant \|x\|_p^2 + \|y\|_p^2 \leqslant \|x\|_1^2 + \|y\|_1^2 \leqslant 2R_1^2$. Moreover, because $p = 1 + \frac{1}{\log(2n)}$, we have $q = 1 + \log(2n)$, and

$$\|v\|_q \leqslant \|\mathbf{1}\|_q \|v\|_\infty = n^{\frac{1}{q}} \|v\|_\infty = n^{\frac{1}{\log(2n)}} \|v\|_\infty \leqslant e \|v\|_\infty .$$

Substituting this into the previous display and noting that $1/(p-1) = \log(2n)$ gives the first result. Integrating $\sum_{k=1}^{K} k^{-\frac{1}{2}}$ and using convexity gives the second. □

So we see that, in more general cases than the simple simplex constraint afforded by the entropic mirror descent (exponentiated gradient) updates, we have convergence guarantees of order $\sqrt{\log n}/\sqrt{K}$, which may be substantially faster than that guaranteed by the standard projected gradient methods.

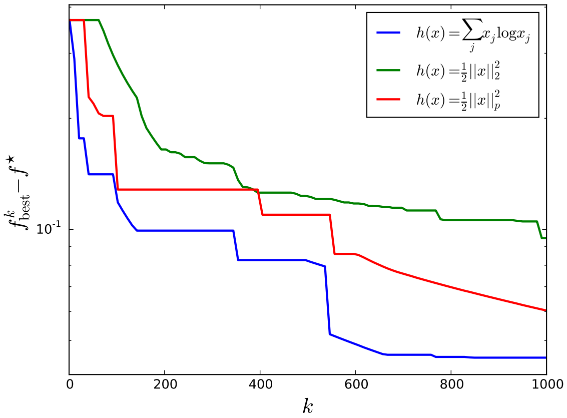

Figure 4.2.16. Convergence of mirror descent (entropic gradient method) versus projected gradient method.

A simulated mirror-descent example With our convergence theorems given, we provide a (simulation-based) example of the convergence behavior for an optimization problem for which it is natural to use non-Euclidean norms. We consider a robust regression problem of the following form: we let $A \in \mathbb{R}^{m \times n}$ have entries drawn i.i.d. $N(0,1)$ with rows $a_1^\top, \ldots, a_m^\top$. We let $b_i = \frac{1}{2}(a_{i,1} + a_{i,2}) + \varepsilon_i$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, 10^{-2})$, and $m = 20$ and the dimension $n = 3000$. Then we define

$$f(x) := \|Ax - b\|_1 = \sum_{i=1}^{m} |\langle a_i, x\rangle - b_i|,$$

which has subgradients $A^\top \operatorname{sign}(Ax - b)$. We minimize f over the simplex $C = \{x \in \mathbb{R}^n_+ : \langle \mathbf{1}, x\rangle = 1\}$; this is the same robust regression problem (3.2.12), except with a particular choice of C.

We compare the subgradient method to exponentiated gradient descent for this problem, noting that the Euclidean projection of a vector $v \in \mathbb{R}^n$ to the set C has coordinates $x_j = [v_j - t]_+$, where $t \in \mathbb{R}$ is chosen so that

$$\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} [v_j - t]_+ = 1.$$

(See the papers [14, 23] for a full derivation of this expression.) We use stepsizes $\alpha_k = \alpha_0/\sqrt{k}$, where the initial stepsize α_0 is chosen to optimize the convergence guarantee for each of the methods (see the coming section). In Figure 4.2.16, we plot the results of performing the projected gradient method versus the exponentiated gradient (entropic mirror decent) method and a method using distance generating functions $h(x) = \frac{1}{2}\|x\|_p^2$ for $p = 1 + 1/\log(2n)$, which can also be shown to be optimal, showing the optimality gap versus iteration count. All three methods are sensitive to initial stepsize, the mirror descent method (4.2.4) enjoys faster convergence than the standard gradient-based method.

4.3. Adaptive stepsizes and metrics In our discussion of mirror descent methods, we assumed we knew enough about the geometry of the problem at hand—or at least the constraint set—to choose an appropriate metric and associated distance-generating function h. In other situations, however, it may be advantageous to *adapt* the metric being used, or at least the stepsizes, to achieve faster convergence guarantees. We begin by describing a simple scheme for choosing stepsizes to optimize bounds on convergence, which means one does not need to know the Lipschitz constants of gradients ahead of time, and then move on to somewhat more involved schemes that use a distance-generating function of the type $h(x) = \frac{1}{2}x^\top Ax$ for some matrix A, which may change depending on information observed during solution of the problem. We leave proofs of the major results in these sections to exercises at the end of the lectures.

Adaptive stepsizes Let us begin by recalling the convergence guarantees for mirror descent in the stochastic case, given by Corollary 4.2.11, which assumes the stepsize α_k used to calculate x_{k+1} is chosen based on the observed gradients $g_1, \ldots, g_k$ (it may be specified ahead of time). In this case, taking $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$, we have by Corollary 4.2.11 that as long as $D_h(x, x^\star) \leqslant R^2$ for all $x \in C$, then

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leqslant \mathbb{E}\left[\frac{R^2}{K\alpha_K} + \frac{1}{K}\sum_{k=1}^{K}\frac{\alpha_k}{2}\|g_k\|_*^2\right]. \tag{4.3.1}$$

Now, if we were to use a fixed stepsize $\alpha_k = \alpha$ for all k, we see that the choice of stepsize minimizing

$$\frac{R^2}{K\alpha} + \frac{\alpha}{2K}\sum_{k=1}^{K}\|g_k\|_*^2$$

is

$$\alpha^\star = \sqrt{2}R\left(\sum_{k=1}^{K}\|g_k\|_*^2\right)^{-\frac{1}{2}},$$

which, when substituted into the bound (4.3.1) yields

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leqslant \sqrt{2}\frac{R}{K}\mathbb{E}\left[\left(\sum_{k=1}^{K}\|g_k\|_*^2\right)^{\frac{1}{2}}\right]. \tag{4.3.2}$$

While the stepsize choice $\alpha^\star$ and the resulting bound are not strictly possible, as we do not know the magnitudes of the gradients $\|g_k\|_*$ before the procedure executes, in Exercise 8, we prove the following corollary, which uses the "up to now" optimal choice of stepsize α_k.

Corollary 4.3.3. *Let the conditions of Corollary 4.2.11 hold. Let* $\alpha_k = R/\sqrt{\sum_{i=1}^{k}\|g_i\|_*^2}$. *Then*

$$\mathbb{E}[f(\overline{x}_K) - f(x^\star)] \leqslant 3\frac{R}{K}\mathbb{E}\left[\left(\sum_{k=1}^{K}\|g_k\|_*^2\right)^{\frac{1}{2}}\right],$$

where $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$.

When comparing Corollary 4.3.3 to Corollary 4.2.11, we see by Jensen's inequality that, if $\mathbb{E}[\|g_k\|_*^2] \leqslant M^2$ for all k, then

$$\mathbb{E}\left[\left(\sum_{k=1}^{K}\|g_k\|_*^2\right)^{\frac{1}{2}}\right] \leqslant \mathbb{E}\left[\sum_{k=1}^{K}\|g_k\|_*^2\right]^{\frac{1}{2}} \leqslant \sqrt{M^2K} = M\sqrt{K}.$$

Thus, ignoring the $\sqrt{2}$ versus 3 multiplier, the bound of Corollary 4.3.3 is always tighter than that provided by Corollary 4.2.11 and its immediate consequence (4.2.12). We do not explore these particular stepsize choices further, but turn to more sophisticated adaptation strategies.

Variable metric methods and the adaptive gradient method In variable metric methods, the idea is to adjust the metric with which one constructs updates to better reflect local (or non-local) problem structure. The basic framework is very similar to the standard subgradient method (or the mirror descent method), and proceeds as follows.

(i) Receive subgradient $g_k \in \partial f(x_k)$ (or stochastic subgradient g_k satisfying $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$)

(ii) Update positive semidefinite matrix $H_k \in \mathbb{R}^{n \times n}$

(iii) Compute update

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ \langle g_k, x \rangle + \frac{1}{2} \langle x, H_k x \rangle \right\}. \tag{4.3.4}$$

The method (4.3.4) subsumes a number of standard and less standard optimization methods. If $H_k = \frac{1}{\alpha_k} I_{n \times n}$, a scaled identity matrix, we recover the (stochastic) subgradient method (3.2.4) when $C = \mathbb{R}^n$ (or (3.3.2) generally). If f is twice differentiable and $C = \mathbb{R}^n$, then taking $H_k = \nabla^2 f(x_k)$ to be the Hessian of f at x_k gives the (undamped) Newton method, and using $H_k = \nabla^2 f(x_k)$ even when $C \neq \mathbb{R}^n$ gives a constrained Newton method. More general choices of H_k can even give the ellipsoid method and other classical convex optimization methods [56].

In our case, we specialize the iterations above to focus on diagonal matrices H_k, and we do not assume the function f is smooth (not even differentiable). This, of course, renders unusable standard methods using second order information in the matrix H_k (as it does not exist), but we may still develop useful algorithms. It is possible to consider more general matrices [22], but their additional computational cost generally renders them impractical in large scale and stochastic settings. With that in mind, let us develop a general framework for algorithms and provide their analysis.

We begin with a general convergence guarantee.

Theorem 4.3.5. *Let H_k be a sequence of positive definite matrices, where H_k is a function of $g_1, \ldots, g_k$ (and potentially some additional randomness). Let g_k be (stochastic) subgradients with $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$. Then*

$$\mathbb{E}\left[\sum_{k=1}^{K} (f(x_k) - f(x^\star))\right] \leqslant \frac{1}{2}\mathbb{E}\left[\sum_{k=2}^{K} \left(\|x_k - x^\star\|_{H_k}^2 - \|x_k - x^\star\|_{H_{k-1}}^2\right) + \|x_1 - x^\star\|_{H_1}^2\right] + \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K} \|g_k\|_{H_k^{-1}}^2\right].$$

Proof. In contrast to mirror descent methods, in this proof we return to our classic Lyapunov-based style of proof for standard subgradient methods, looking at the distance $\|x_k - x^\star\|$. Let $\|x\|_A^2 = \langle x, Ax \rangle$ for any positive semidefinite matrix. We claim that

$$\|x_{k+1} - x^\star\|_{H_k}^2 \leqslant \left\|x_k - H_k^{-1} g_k - x^\star\right\|_{H_k}^2, \tag{4.3.6}$$

the analogue of the fact that projections are non-expansive. This is an immediate consequence of the update (4.3.4): we have that

$$x_{k+1} = \operatorname*{argmin}_{x \in C} \left\{ \left\| x - (x_k - H_k^{-1} g_k) \right\|_{H_k}^2 \right\},$$

which is a Euclidean projection of $x_k - H_k^{-1} g_k$ into C (in the norm $\|\cdot\|_{H_k}$). Then the standard result that projections are non-expansive (Corollary 2.2.8) gives inequality (4.3.6).

Inequality (4.3.6) is the key to our analysis, as previously. Expanding the square on the right side of the inequality, we obtain

$$\begin{aligned}\frac{1}{2}\left\|x_{k+1} - x^\star\right\|_{H_k}^2 &\leqslant \frac{1}{2}\left\|x_k - H_k^{-1} g_k - x^\star\right\|_{H_k}^2 \\ &= \frac{1}{2}\left\|x_k - x^\star\right\|_{H_k}^2 - \langle g_k, x_k - x^\star\rangle + \frac{1}{2}\left\|g_k\right\|_{H_k^{-1}}^2,\end{aligned}$$

and taking expectations we have $\mathbb{E}[\langle g_k, x_k - x^\star\rangle \mid x_k] \geqslant f(x_k) - f(x^\star)$ by convexity and that $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$. Thus

$$\frac{1}{2}\mathbb{E}\left[\left\|x_{k+1} - x^\star\right\|_{H_k}^2\right] \leqslant \mathbb{E}\left[\frac{1}{2}\left\|x_k - x^\star\right\|_{H_k}^2 - [f(x_k) - f(x^\star)] + \frac{1}{2}\left\|g_k\right\|_{H_k^{-1}}^2\right].$$

Rearranging, we have

$$\mathbb{E}[f(x_k) - f(x^\star)] \leqslant \mathbb{E}\left[\frac{1}{2}\left\|x_k - x^\star\right\|_{H_k}^2 - \frac{1}{2}\left\|x_{k+1} - x^\star\right\|_{H_k}^2 + \frac{1}{2}\left\|g_k\right\|_{H_k^{-1}}^2\right].$$

Summing this inequality from $k = 1$ to K gives the theorem. □

We may specialize the theorem in a number of ways to develop particular algorithms. One specialization, which is convenient because the computational overhead is fairly small, is to use diagonal matrices H_k. In particular, the AdaGrad method sets

$$H_k = \frac{1}{\alpha}\operatorname{diag}\left(\sum_{i=1}^{k} g_i g_i^\top\right)^{\frac{1}{2}}, \tag{4.3.7}$$

where $\alpha > 0$ is a pre-specified constant (stepsize). In this case, the following corollary to Theorem 4.3.5 follows. Exercise 10 sketches the proof of the corollary, which is similar to that of Corollary 4.3.3. In the corollary, recall that $\operatorname{tr}(A) = \sum_{j=1}^{n} A_{jj}$ is the trace of a matrix.

Corollary 4.3.8 (AdaGrad convergence). *Let $R_\infty := \sup_{x\in C}\|x - x^\star\|_\infty$ be the ℓ_∞ radius of the set C and let the conditions of Theorem 4.3.5 hold. Then with the choice (4.3.7) in the variable metric method, we have*

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^\star))\right] \leqslant \frac{1}{2\alpha}R_\infty^2 \mathbb{E}[\operatorname{tr}(H_K)] + \alpha\mathbb{E}[\operatorname{tr}(H_K)].$$

Inspecting Corollary 4.3.8, we see a few consequences. First, by choosing $\alpha = R_\infty$, we obtain the expected convergence guarantee $\frac{3}{2}R_\infty\mathbb{E}[\operatorname{tr}(H_K)]$. If we let $\overline{x}_K = \frac{1}{K}\sum_{k=1}^{K} x_k$ as usual, and let $g_{k,j}$ denote the jth component of the kth gradient observed by the method, then we immediately obtain the convergence guarantee

$$\mathbb{E}\left[f(\overline{x}_K) - f(x^\star)\right] \leqslant \frac{3}{2K}R_\infty\mathbb{E}[\operatorname{tr}(H_K)] = \frac{3}{2K}R_\infty\sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{k=1}^{K} g_{k,j}^2\right)^{\frac{1}{2}}\right]. \tag{4.3.9}$$

In addition to proving the bound (4.3.9), Exercise 10 also shows that, if $C = \{x \in \mathbb{R}^n : \|x\|_\infty \leqslant 1\}$, then the bound (4.3.9) is always better than the bounds (e.g. Corollary 3.4.9) guaranteed by standard stochastic gradient methods. In addition, the bound (4.3.9) is unimprovable—there are stochastic optimization problems for which no algorithm can achieve a faster convergence rate. These types of problems generally involve data in which the gradients g have highly varying components (or components that are often zero, i.e. the gradients g are sparse), as for such problems geometric aspects are quite important.

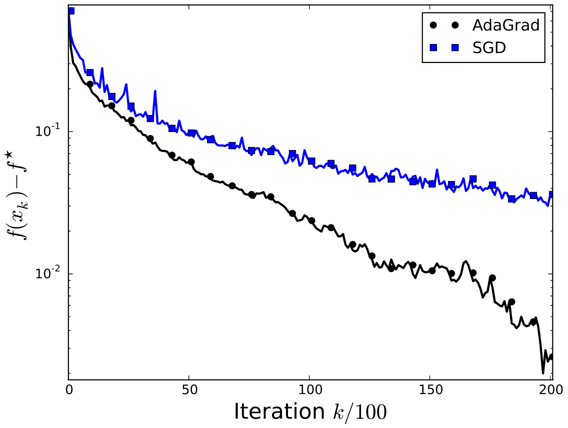

Figure 4.3.10. A comparison of the convergence of AdaGrad and SGD on the problem (4.3.11) for the best initial stepsize α for each method.

We now give an example application of the AdaGrad method, showing its performance on a simulated example. We consider solving the problem

$$\text{minimize } f(x) = \frac{1}{m}\sum_{i=1}^{m} [1 - b_i \langle a_i, x\rangle]_+ \text{ subject to } \|x\|_\infty \leqslant 1, \tag{4.3.11}$$

where the vectors $a_i \in \{-1, 0, 1\}^n$ with $m = 5000$ and $n = 1000$. This is the objective common to hinge loss (support vector machine) classification problems. For each coordinate $j \in \{1, \ldots, n\}$, we set $a_{i,j} \in \{\pm 1\}$ to have a random sign with probability $1/j$, and $a_{i,j} = 0$ otherwise. Letting $u \in \{-1, 1\}^n$ uniformly at random, we set $b_i = \text{sign}(\langle a_i, u\rangle)$ with probability .95 and $b_i = -\text{sign}(\langle a_i, u\rangle)$ otherwise. For this problem, the coordinates of a_i (and hence subgradients or stochastic subgradients of f) naturally have substantial variability, making it a natural problem for adaptation of the metric H_k.

In Figure 4.3.10, we show the convergence behavior of AdaGrad versus stochastic gradient descent (SGD) on one realization of this problem, where at each iteration we choose a stochastic gradient by selecting $i \in \{1, \ldots, m\}$ uniformly at random, then setting $g_k \in \partial [1 - b_i \langle a_i, x_k\rangle]_+$. For SGD, we use stepsizes

$\alpha_k = \alpha/\sqrt{k}$, where α is the best stepsize of several choices (based on the eventual convergence of the method), while AdaGrad uses the matrix (4.3.7), with α similarly chosen based on the best eventual convergence. The plot shows the typical behavior of AdaGrad with respect to stochastic gradient methods, at least for problems with appropriate geometry: with good initial stepsize choice, AdaGrad often outperforms stochastic gradient descent. (We have been vague about the "right" geometry for problems in which we expect AdaGrad to perform well. Roughly, problems for which the domain C is well-approximated by a box $\{x \in \mathbb{R}^n : \|x\|_\infty \leqslant c\}$ are those for which we expect AdaGrad to succeed, and otherwise, it may exhibit worse performance than standard subgradient methods. As in any problem, some care is needed in the choice of methods.) Figure 4.3.12 shows this somewhat more broadly, plotting the convergence $f(x_k) - f(x^\star)$ versus iteration k for a number of initial stepsize choices for both stochastic gradient descent and AdaGrad on the problem (4.3.11). Roughly, we see that both methods are sensitive to initial stepsize choice, but the best choice for AdaGrad often outperforms the best choice for SGD.

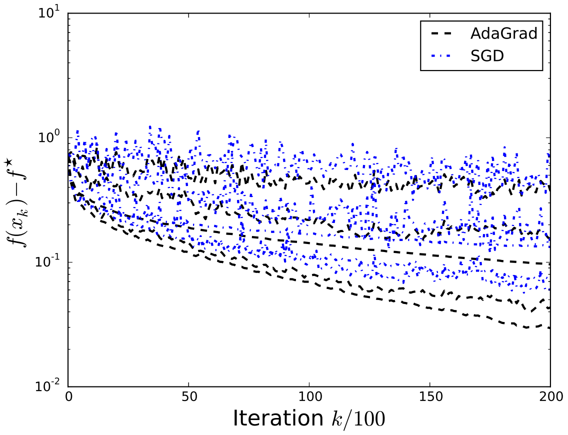

Figure 4.3.12. A comparison of the convergence of AdaGrad and SGD on the problem (4.3.11) for various initial stepsize choices $\alpha \in \{10^{-i/2}, i = -2, \ldots, 2\} = \{.1, .316, 1, 3.16, 10\}$. Both methods are sensitive to the initial stepsize choice α, though for each initial stepsize choice, AdaGrad has better convergence than the subgradient method.

Notes and further reading The mirror descent method was originally developed by Nemirovski and Yuding [41] in order to more carefully control the norms of gradients, and associated dual spaces, in first-order optimization methods. Since their original development, a number of researchers have explored variants and extensions of their methods. Beck and Teboulle [5] give an analysis of mirror descent as a non-Euclidean gradient method, which is the approach we take in this lecture. Nemirovski et al. [40] study mirror descent methods in stochastic settings, giving high-probability convergence guarantees similar to those we gave in the previous lecture. Bubeck and Cesa-Bianchi [15] explore the use of mirror descent methods in the context of *bandit* optimization problems, where instead of observing stochastic gradients one observes only random function values $f(x) + \varepsilon$, where ε is mean-zero noise.

Variable metric methods have a similarly long history. Our simple results with stepsize selection follow the more advanced techniques of Auer et al. [3] (see especially their Lemma 3.5), and the AdaGrad method (and our development) is due to Duchi, Hazan, and Singer [22] and McMahan and Streeter [38]. More general metric methods include Shor's space dilation methods (of which the ellipsoid method is a celebrated special case), which develop matrices H_k that make new directions of descent somewhat less correlated with previous directions, allowing faster convergence in directions toward $x^\star$; see the books of Shor [55,56] as well as the thesis of Nedić [39]. Newton methods, which we do not discuss, use scaled multiples of $\nabla^2 f(x_k)$ for H_k, while Quasi-Newton methods approximate $\nabla^2 f(x_k)$ with H_k while using only gradient-based information; for more on these and other more advanced methods for smooth optimization problems, see the books of Nocedal and Wright [46] and Boyd and Vandenberghe [12].

5. Optimality Guarantees

Lecture Summary: In this lecture, we provide a framework for demonstrating the optimality of a number of algorithms for solving stochastic optimization problems. In particular, we introduce minimax lower bounds, showing how techniques for reducing estimation problems to statistical testing problems allow us to prove lower bounds on optimization.

5.1. Introduction The procedures and algorithms we have presented thus far enjoy good performance on a number of statistical, machine learning, and stochastic optimization tasks, and we have provided theoretical guarantees on their performance. It is interesting to ask whether it is possible to improve the algorithms, or in what ways it may be possible to improve them. With that in mind, in this lecture we develop a number of tools for showing optimality—according to certain metrics—of optimization methods for stochastic problems.

Minimax rates We provide optimality guarantees in the *minimax* framework for optimality, which proceeds roughly as follows: we have a collection of possible problems and an error measure for the performance of a procedure, and we measure a procedure's performance by its behavior on the *hardest* (most difficult) member of the problem class. We then ask for the best procedure under this worst-case error measure. Let us describe this more formally in the context of our stochastic optimization problems, where the goal is to understand the difficulty of minimizing a convex function f subject to constraints $x \in C$ while observing only stochastic gradient (or other noisy) information about f. Our bounds build on three objects:

(i) A collection $\mathcal{F}$ of convex functions $f : \mathbb{R}^n \to \mathbb{R}$

(ii) A closed convex set $C \subset \mathbb{R}^n$ over which we optimize

(iii) A *stochastic gradient oracle*, which consists of a sample space $\mathcal{S}$, a gradient mapping

$$g : \mathbb{R}^n \times \mathcal{S} \times \mathcal{F} \to \mathbb{R}^n,$$

and (implicitly) a probability distributions P on $\mathcal{S}$. The stochastic gradient oracle may be *queried* at a point x, and when queried, draws $S \sim P$ with the property that

$$\mathbb{E}[g(x, S, f)] \in \partial f(x). \tag{5.1.1}$$

Depending on the scenario of the problem, the optimization procedure may be given access either to S or simply the value of the stochastic gradient $g = g(x, S, f)$, and the goal is to use the sequence of observations $g(x_k, S_k, f)$, for $k = 1, 2, \ldots$, to optimize f.

A simple example of the setting (i)–(iii) is as follows. Let $A \in \mathbb{R}^{n \times n}$ be a fixed positive definite matrix, and let $\mathcal{F}$ be the collection of convex functions of the form $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ for all $b \in \mathbb{R}^n$. Then C may be any convex set, and—for the sake of proving lower bounds, not for real applicability in solving problems—we might take the stochastic gradient

$$g = \nabla f(x) + \xi = Ax - b + \xi \text{ for } \xi \overset{\text{iid}}{\sim} N(0, I_{n \times n}).$$

A somewhat more complex example, but with more fidelity to real problems, comes from the stochastic programming problem (3.4.2) from Lecture 3 on subgradient methods. In this case, there is a known convex function $F : \mathbb{R}^n \times \mathcal{S} \to \mathbb{R}$, which is the instantaneous loss function $F(x; s)$. The problem is then to optimize

$$f_P(x) := \mathbb{E}_P[F(x; S)]$$

where the distribution P on the random variable S is unknown to the method *a priori*; there is then a correspondence between distributions P and functions $f \in \mathcal{F}$. Generally, an optimization is given access to a sample $S_1, \ldots, S_K$ drawn i.i.d. according to the distribution P (in this case, there is no selection of points x_i by the optimization procedure, as the sample $S_1, \ldots, S_K$ contains even more information than the stochastic gradients). A similar variant with a natural stochastic gradient oracle is to set $g(x, s, F) \in \partial F(x; s)$ instead of providing the sample $S = s$.

We focus in this note on the case when the optimization procedure may view only the sequence of subgradients $g_1, g_2, \ldots$ at the points it queries. We note in passing, however, that for many problems we can reconstruct S from a gradient $g \in \partial F(x; S)$. For example, consider a logistic regression problem with data $s =$

$(a, b) \in \{0, 1\}^n \times \{-1, 1\}$, a typical data case. Then

$$F(x; s) = \log(1 + e^{-b\langle a, x\rangle}), \text{ and } \nabla_x F(x; s) = -\frac{1}{1 + e^{b\langle a, x\rangle}} ba,$$

so that (a, b) is identifiable from any $g \in \partial F(x; s)$. More generally, classical linear models in statistics have gradients that are scaled multiples of the data, so that the sample s is typically identifiable from $g \in \partial F(x; s)$.

Now, given function f and stochastic gradient oracle g, an optimization procedure chooses query points $x_1, x_2, \ldots, x_K$ and observes stochastic subgradients g_k with $\mathbb{E}[g_k] \in \partial f(x_k)$. Based on these stochastic gradients, the optimization procedure outputs $\widehat{x}_K$, and we assess the quality of the procedure in terms of the excess loss

$$\mathbb{E}\left[f\left(\widehat{x}_K(g_1, \ldots, g_K)\right) - \inf_{x^\star \in C} f(x^\star)\right],$$

where the expectation is taken over the subgradients $g(x_i, S_i, f)$ returned by the stochastic oracle and any randomness in the chosen iterates, or query points, $x_1, \ldots, x_K$ of the optimization method. Of course, if we only consider this excess objective value for a fixed function f, then a trivial optimization procedure achieves excess risk 0: simply return some $x \in \operatorname{argmin}_{x \in C} f(x)$. It is thus important to ask for a more uniform notion of risk: we would like the procedure to have good performance *uniformly* across all functions $f \in \mathcal{F}$, leading us to measure the performance of a procedure by its worst-case risk

$$\sup_{f \in \mathcal{F}} \mathbb{E}\left[f(\widehat{x}(g_1, \ldots, g_k)) - \inf_{x \in C} f(x^\star)\right],$$

where the supremum is taken over functions $f \in \mathcal{F}$ (the subgradient oracle g then implicitly depends on f). An optimal estimator for this metric then gives the *minimax risk* for optimizing the family of stochastic optimization problems $\{f\}_{f \in \mathcal{F}}$ over $x \in C \subset \mathbb{R}^n$, which is

$$\mathfrak{M}_K(C, \mathcal{F}) := \inf_{\widehat{x}_K} \sup_{f \in \mathcal{F}} \mathbb{E}\left[f(\widehat{x}_K(g_1, \ldots, g_K)) - \inf_{x^\star \in C} f(x^\star)\right]. \tag{5.1.2}$$

We take the supremum (worst-case) over distributions $f \in \mathcal{F}$ and the infimum over all possible optimization schemes $\widehat{x}_K$ using K stochastic gradient samples.

A criticism of the framework (5.1.2) is that it is too pessimistic: by taking a worst-case over distributions of functions $f \in \mathcal{F}$, one is making the family of problems too challenging. We will not address these challenges except to say that one response is to develop *adaptive* procedures $\widehat{x}$, which are simultaneously optimal for a variety of collections of problems $\mathcal{F}$.

The basic approach There are a variety of techniques for providing lower bounds on the minimax risk (5.1.2). Each of them transforms the maximum risk by lower bounding it via a Bayesian problem (e.g. [31,33,34]), then proving a lower bound on the performance of all possible estimators for the Bayesian problem. In particular, let $\{f_v\} \subset \mathcal{F}$ be a collection of functions in $\mathcal{F}$ indexed by some (finite or countable) set $\mathcal{V}$ and π be any probability mass function over $\mathcal{V}$. Let $f^\star = \inf_{x \in C} f(x)$.

Then for any procedure $\widehat{x}$, the maximum risk has lower bound

$$\sup_{f \in \mathcal{F}} \mathbb{E}\left[f(\widehat{x}) - f^\star\right] \geqslant \sum_{v} \pi(v) \mathbb{E}\left[f_v(\widehat{x}) - f_v^\star\right].$$

While trivial, this lower bound serves as the departure point for each of the subsequent techniques for lower bounding the minimax risk. The lower bound also allows us to assume that the procedure $\widehat{x}$ is *deterministic*. Indeed, assume that $\widehat{x}$ is non-deterministic, which we can represent generally as depending on some auxiliary random variable U independent of the observed subgradients. Then we certainly have

$$\mathbb{E}\left[\sum_{v} \pi(v) \mathbb{E}\left[f_v(\widehat{x}) - f_v^\star \mid U\right]\right] \geqslant \inf_{u} \sum_{v} \pi(v) \mathbb{E}\left[f_v(\widehat{x}) - f_v^\star \mid U = u\right],$$

that is, there is some realization of the auxiliary randomness that is at least as good as the average realization. We can simply incorporate this into our minimax optimal procedures $\widehat{x}$, and thus we assume from this point onward that all our optimization procedures are deterministic when proving our lower bounds.

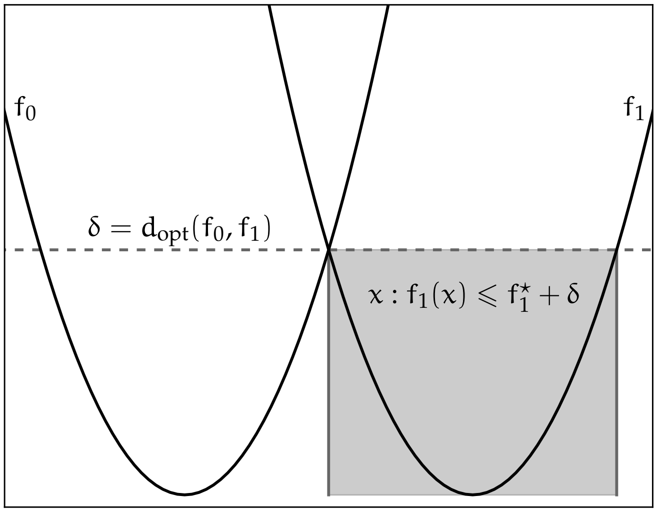
f_0
f_1
$\delta = d_{\rm opt}(f_0, f_1)$
$x : f_1(x) \leqslant f_1^\star + \delta$

Figure 5.1.3. Separation of optimizers of f_0 and f_1. Optimizing one function to accuracy better than $\delta = d_{\rm opt}(f_0, f_1)$ implies we optimize the other poorly; the gap $f(x) - f^\star$ is at least δ.

The second step in proving minimax bounds is to reduce the optimization problem to a type of statistical test [58, 61, 62]. To perform this reduction, we define a distance-like quantity between functions such that, if we have optimized a function f_v to better than the distance, we cannot have optimized other functions well. In particular, consider two convex functions f_0 and f_1. Let $f_v^\star = \inf_{x \in C} f_v(x)$ for $v \in \{0, 1\}$. We let the *optimization separation* between functions f_0 and f_1 over the set C be

$$d_{\mathrm{opt}}(f_0, f_1; C) :=$$

$$\sup\left\{\delta \geqslant 0 : \begin{array}{l} f_1(x) \leqslant f_1^\star + \delta \text{ implies } f_0(x) \geqslant f_0^\star + \delta \\ f_0(x) \leqslant f_0^\star + \delta \text{ implies } f_1(x) \geqslant f_1^\star + \delta \end{array} \text{ for any } x \in C\right\}. \tag{5.1.4}$$

That is, if we have any point x such that $f_v(x) - f_v^\star \leqslant d_{\mathrm{opt}}(f_0, f_1)$, then x cannot optimize f_{1-v} well, i.e. we can only optimize one of the two functions f_0 and f_1 to accuracy $d_{\mathrm{opt}}(f_0, f_1)$. See Figure 5.1.3 for an illustration of this quantity. For example, if $f_1(x) = (x + c)^2$ and $f_0(x) = (x - c)^2$ for a constant $c \neq 0$, then we have $d_{\mathrm{opt}}(f_1, f_0) = c^2$.

This separation d_{opt} allows us to give a reduction from optimization to testing via the *canonical hypothesis testing problem*, which is as defined as follows:

1. Nature chooses an index $V \in \mathcal{V}$ uniformly at random
2. Conditional on the choice $V = v$, the procedure observes stochastic subgradients for the function f_v according to the oracle $g(x_k, S_k, f_v)$ for i.i.d. S_k.

Then, given the observed subgradients, the goal is to test which of the random indices v nature chose. Intuitively, if we can optimize f_v well—to better than the separation $d_{\mathrm{opt}}(f_v, f_{v'})$—then we can identify the index v. If we can show this, then we can adapt classical statistical results on optimal hypothesis testing to lower bound the probability of error in testing whether the data was generated conditional on $V = v$.

More formally, we have the following key lower bound. In the lower bound, we say that a collection of functions $\{f_v\}_{v\in\mathcal{V}}$ is δ-*separated*, where $\delta \geqslant 0$, if

$$d_{\mathrm{opt}}(f_v, f_{v'}; C) \geqslant \delta \text{ for each } v, v' \in \mathcal{V} \text{ with } v \neq v'. \tag{5.1.5}$$

Then we have the next proposition.

Proposition 5.1.6. *Let S be drawn uniformly from $\mathcal{V}$, where $|\mathcal{V}| < \infty$, and assume the collection $\{f_v\}_{v\in\mathcal{V}}$ is δ-separated. Then for any optimization procedure $\widehat{x}$ based on the observed subgradients,*

$$\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}[f_v(\widehat{x}) - f_v^\star] \geqslant \delta \cdot \inf_{\widehat{v}} \mathbb{P}(\widehat{v} \neq V),$$

where the distribution $\mathbb{P}$ is the joint distribution over the random index V and the observed gradients $g_1, \ldots, g_K$ and the infimum is taken over all testing procedures $\widehat{v}$ based on the observed data.

Proof. We let P_v denote the distribution of the subgradients conditional on the choice $V = v$, meaning that $\mathbb{E}[g_k \mid V = v] \in \partial f_v(x_k)$. We observe that for any v, we have

$$\mathbb{E}[f_v(\widehat{x}) - f_v^\star] \geqslant \delta\mathbb{E}[\mathbf{1}\{f_v(\widehat{x}) \geqslant f_v^\star + \delta\}] = \delta P_v(f_v(\widehat{x}) \geqslant f_v^\star + \delta).$$

Now, define the hypothesis test $\widehat{\nu}$, which is a function of $\widehat{x}$, by

$$\widehat{\nu} = \begin{cases} \nu & \text{if } f_\nu(\widehat{x}) \leqslant f_\nu^\star + \delta \\ \text{arbitrary in } \mathcal{V} & \text{otherwise.} \end{cases}$$

This is a well-defined mapping, as by the condition that $d_{\rm opt}(f_\nu, f_{\nu'}) \geqslant \delta$, there can be only a *single* index ν such that $f_\nu(x) \leqslant f_\nu^\star + \delta$. We then note the following implication:

$$\widehat{\nu} \neq \nu \text{ implies } f_\nu(\widehat{x}) \geqslant f_\nu^\star + \delta.$$

Thus we have

$$P_\nu(\widehat{\nu} \neq \nu) \leqslant P_\nu(f_\nu(\widehat{x}) \geqslant f_\nu^\star + \delta),$$

or, summarizing, we have

$$\frac{1}{|\mathcal{V}|} \sum_{\nu \in \mathcal{V}} \mathbb{E}[f_\nu(\widehat{x}) - f_\nu^\star] \geqslant \delta \cdot \frac{1}{|\mathcal{V}|} \sum_{\nu \in \mathcal{V}} P_\nu(\widehat{\nu} \neq \nu).$$

But by definition of the distribution $\mathbb{P}$, we have $\frac{1}{|\mathcal{V}|} \sum_{\nu \in \mathcal{V}} P_\nu(\widehat{\nu} \neq \nu) = \mathbb{P}(\widehat{\nu} \neq V)$, and taking the best possible test $\widehat{\nu}$ gives the result of the proposition. □

Proposition 5.1.6 allows us to then bring in the tools of optimal testing in statistics and information theory, which we can use to prove lower bounds. To leverage Proposition 5.1.6, we follow a two phase strategy: we construct a well-separated function collection, and then we show that it is difficult to test which of the functions we observe data from. There is a natural tension in the proposition, as it is easier to distinguish functions that are far apart (i.e. large δ), while hard-to-distinguish functions (i.e. large $\mathbb{P}(\widehat{\nu} \neq V)$) often have smaller separation. Thus we trade these against one another carefully in constructing our lower bounds on the minimax risk. We also present a variant lower bound in Section 5.3 based on a similar reduction, except that we use multiple binary hypothesis tests.

5.2. Le Cam's Method Our first set of lower bounds is based on Le Cam's method [33], which uses optimality guarantees for simple binary hypothesis tests to provide lower bounds for optimization problems. That is, we let $\mathcal{V} = \{-1, 1\}$ and will construct only pairs of functions and distributions P_1, P_{-1} generating data. In this section, we show how to use these binary hypothesis tests to prove lower bounds on the family of stochastic optimization problems characterized by the following conditions: the domain $C \subset \mathbb{R}^n$ contains an ℓ_2-ball of radius R and the subgradients g_k satisfy the second moment bound

$$\mathbb{E}[\|g_k\|_2^2] \leqslant M^2$$

for all k. We assume that $\mathcal{F}$ consists of M-Lipschitz continuous convex functions.

With the definition (5.1.4) of the separation in terms of optimization value, we can provide a lower bound on optimization in terms of distances between distributions P_1 and P_{-1}. Before we continue, we require a few definitions about distances between distributions.

Definition 5.2.1. Let P and Q be distributions on a space $\mathcal{S}$, and assume that they are both absolutely continuous with respect to a measure μ on $\mathcal{S}$. The *variation distance* between P and Q is

$$\|P - Q\|_{\mathrm{TV}} := \sup_{A \subset \mathcal{S}} |P(A) - Q(A)| = \frac{1}{2}\int_{\mathcal{S}} |p(s) - q(s)| d\mu(s).$$

The *Kullback-Leibler divergence* between P and Q is

$$D_{\mathrm{kl}}(P|Q) := \int_{\mathcal{S}} p(s) \log \frac{p(s)}{q(s)} d\mu(s).$$

We can connect the variation distance to binary hypothesis tests via the following lemma, due to Le Cam. The lemma states that testing between two distributions is hard precisely when they are close in variation distance.

Lemma 5.2.2. *Let P_1 and P_{-1} be any distributions. Then*

$$\inf_{\widehat{v}} \{P_1(\widehat{v} \neq 1) + P_{-1}(\widehat{v} \neq -1)\} = 1 - \|P_1 - P_{-1}\|_{\mathrm{TV}}.$$

Proof. Any testing procedure $\widehat{v} : \mathcal{S} \to \{-1, 1\}$ maps one region of the sample space, call it A, to 1 and the complement A^c to -1. Thus, we have

$$P_1(\widehat{v} \neq 1) + P_{-1}(\widehat{v} \neq -1) = P_1(A^c) + P_{-1}(A) = 1 - P_1(A) + P_{-1}(A).$$

Optimizing over $\widehat{v}$ is then equivalent to optimizing over sets A, yielding

$$\begin{aligned}\inf_{\widehat{v}}\{P_1(\widehat{v} \neq 1) + P_{-1}(\widehat{v} \neq -1)\} &= \inf_{A}\{1 - P_1(A) + P_{-1}(A)\} \\ &= 1 - \sup_{A}\{P_1(A) - P_{-1}(A)\} = 1 - \|P_1 - P_{-1}\|_{\mathrm{TV}}\end{aligned}$$

as desired. □

As an immediate consequence of Lemma 5.2.2, we obtain the standard minimax lower bound based on binary hypothesis testing. In particular, let f_1 and f_{-1} be δ-separated and belong to $\mathcal{F}$, and assume that the method $\widehat{x}$ receives data (in this case, the data is the K subgradients) from P_v^K when f_v is the true function. Then we immediately have

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \inf_{\widehat{x}_K} \max_{v \in \{-1,1\}} \{\mathbb{E}_{P_v}[f_v(\widehat{x}_K) - f_v^\star]\} \geqslant \frac{1}{2}\delta \cdot \left[1 - \left\|P_1^K - P_{-1}^K\right\|_{\mathrm{TV}}\right]. \tag{5.2.3}$$

Inequality (5.2.3) gives a quantitative guarantee on an intuitive fact: if we observe data from one of two distributions P_1 and P_{-1} that are close, while the optimizers of the functions f_1 and f_{-1} associated with P_1 and P_{-1} differ, it is difficult to optimize well. Moreover, there is a natural tradeoff—the farther apart the functions f_1 and f_{-1} are (i.e. $\delta = d_{\mathrm{opt}}(f_1, f_{-1})$ is large), the bigger the penalty for optimizing one well, but conversely, this usually forces the distributions P_1 and P_{-1} to be quite different, as they provide subgradient information on f_1 and f_{-1}, respectively.

It is challenging to compute quantities—especially with multiple samples—involving the variation distance, so we now convert our bounds to ones involving the KL-divergence, which is computationally easier when dealing with multiple samples. First, we use Pinsker's inequality (see Appendix A.3, Proposition A.3.2 for a proof): for any distributions P and Q,

$$\|P - Q\|_{\mathrm{TV}}^2 \leqslant \frac{1}{2} D_{\mathrm{kl}}\left(P\|Q\right).$$

As we see presently, the KL-divergence *tensorizes* when we have multiple observations from different distributions (see Lemma 5.2.8 to come), allowing substantially easier computation of individual divergence terms. Then we have the following theorem.

Theorem 5.2.4. *Let $\mathcal{F}$ be a collection of convex functions, and let $f_1, f_{-1} \in \mathcal{F}$. Assume that when function f_ν is to be optimized, we observe K subgradients according to P_ν^K. Then*

$$\mathfrak{M}_K(C, \mathcal{P}) \geqslant \frac{d_{\mathrm{opt}}(f_{-1}, f_1; C)}{2}\left[1 - \sqrt{\frac{1}{2} D_{\mathrm{kl}}\left(P_1^K \| P_{-1}^K\right)}\right].$$

What remains to give a concrete lower bound, then, is (1) to construct a family of well-separated functions f_1, f_{-1}, and (2) to construct a stochastic gradient oracle for which we give a small *upper* bound on the KL-divergence between the distributions P_1 and P_{-1} associated with the functions, which means that testing between P_1 and P_{-1} is hard.

Constructing well-separated functions Our first goal is to construct a family of well-separated functions and an associated first-order subgradient oracle that makes the functions hard to distinguish. We parameterize our functions—of which we construct only 2—by a parameter $\delta > 0$ governing their separation. Our construction applies in dimension $n = 1$: let us assume that C contains the interval $[-R, R]$ (this is no loss of generality, as we may simply shift the interval). Then define the M-Lipschitz continuous functions

$$f_1(x) = M\delta|x - R| \text{ and } f_{-1}(x) = M\delta|x + R|. \tag{5.2.5}$$

See Figure 5.2.6 for an example of these functions, which makes clear that their separation (5.1.4) is

$$d_{\mathrm{opt}}(f_1, f_{-1}) = \delta M R.$$

We also consider the stochastic oracle for this problem, recalling that we must construct subgradients satisfying $\mathbb{E}[\|g\|_2^2] \leqslant M^2$. We will do slightly more: we will guarantee that $|g| \leqslant M$ always. With this in mind, we assume that $\delta \leqslant 1$, and define the stochastic gradient oracle for the distribution P_ν, $\nu \in \{-1, 1\}$ at the point x to be

$$g_\nu(x) = \begin{cases} M\,\mathrm{sign}(x - \nu R) & \text{with probability } \frac{1+\delta}{2} \\ -M\,\mathrm{sign}(x - \nu R) & \text{with probability } \frac{1-\delta}{2}. \end{cases} \tag{5.2.7}$$

At $x = \nu R$ the oracle simply returns a random sign. Then by inspection, we see that

$$\mathbb{E}[g_\nu(x)] = \frac{M\delta}{2}\,\mathrm{sign}(x - \nu R) - \frac{M\delta}{2}(-\,\mathrm{sign}(x - \nu R)) = M\delta\,\mathrm{sign}(x - \nu R) \in \partial f_\nu(x)$$

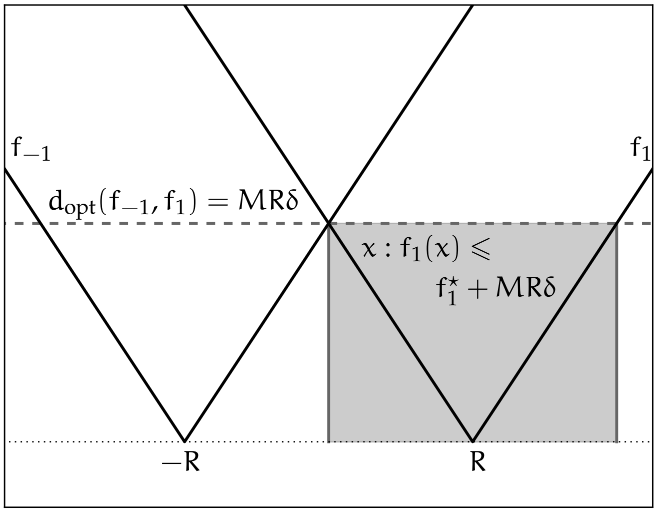

FIGURE 5.2.6. The function construction (5.2.5) with separation $d_{\text{opt}}(f_1, f_{-1}) = MR\delta$.

for $\nu = -1, 1$. Thus, the combination of the functions (5.2.5) and the stochastic gradient (5.2.7) give us a valid subgradient and well-separated pair of functions.

Bounding the distance between distributions The second step in proving our minimax lower bound is to upper bound the distance between the distributions that generate the subgradients our methods observe. This means that testing which of the functions we are optimizing is challenging, giving us a strong lower bound. At a high level, building off of Theorem 5.2.4, we hope to show an upper bound of the form

$$D_{\text{kl}}\left(P_1^K | P_{-1}^K\right) \leqslant \kappa\delta^2$$

for some κ. This is a local condition, allowing us to scale our problems with δ to achieve minimax bounds. If we have such a quadratic, we may simply choose $\delta^2 = 1/2\kappa$, giving the constant probability of error

$$1 - \left\|P_1^K - P_{-1}^K\right\|_{\text{TV}} \geqslant 1 - \sqrt{\frac{1}{2} D_{\text{kl}}\left(P_1^K | P_{-1}^K\right)/2} \geqslant 1 - \sqrt{\frac{\kappa\delta^2}{2}} \geqslant \frac{1}{2}.$$

To this end, we begin with a standard lemma (the chain rule for KL divergence), which applies when we have K potentially dependent observations from a distribution. The result is an immediate consequence of Bayes' rule.

Lemma 5.2.8. *Let $P(\cdot \mid g_1, \ldots, g_{k-1})$ denote the conditional distribution of g_k given $g_1, \ldots, g_{k-1}$. For each $k \in \mathbb{N}$ let P_1^k and P_{-1}^k be distributions on the K subgradients $g_1, \ldots, g_k$. Then*

$$D_{\text{kl}}\left(P_1^K | P_{-1}^K\right) = \sum_{k=1}^K \mathbb{E}_{P_1^{k-1}}\left[D_{\text{kl}}\left(P_1(\cdot \mid g_1, \ldots, g_{k-1}) | P_{-1}(\cdot \mid g_1, \ldots, g_{k-1})\right)\right].$$

Using Lemma 5.2.8, we have the following upper bound on the KL-divergence between P_1^K and P_{-1}^K for the stochastic gradient (5.2.7).

Lemma 5.2.9. *Let the* K *observations under distribution* P_v *come from the stochastic gradient oracle* (5.2.7). *Then for* $\delta \leqslant \frac{4}{5}$,

$$D_{kl}\left(P_1^K | P_{-1}^K\right) \leqslant 3K\delta^2.$$

Proof. We use the chain-rule for KL-divergence, whence we must only provide an upper bound on the individual terms. We first note that x_k is a function of $g_1, \ldots, g_{k-1}$ (because we may assume w.l.o.g. that x_k is deterministic) so that $P_v(\cdot \mid g_1, \ldots, g_{k-1})$ is the distribution of a Bernoulli random variable with distribution (5.2.7), i.e. with probabilities $\frac{1 \pm \delta}{2}$. Thus we have

$$\begin{aligned} D_{kl}\left(P_1(\cdot \mid g_1, \ldots, g_{k-1}) | P_{-1}(\cdot \mid g_1, \ldots, g_{k-1})\right) &\leqslant D_{kl}\left(\frac{1+\delta}{2} | \frac{1-\delta}{2}\right) \\ &= \frac{1+\delta}{2} \log \frac{1+\delta}{1-\delta} + \frac{1-\delta}{2} \log \frac{1-\delta}{1+\delta} \\ &= \delta \log \frac{1+\delta}{1-\delta}. \end{aligned}$$

By a Taylor expansion, we have that

$$\delta \log \frac{1+\delta}{1-\delta} = \delta\left(\delta - \frac{1}{2}\delta^2 + O(\delta^3)\right) - \delta\left(-\delta - \frac{1}{2}\delta^2 + O(\delta^3)\right) = 2\delta^2 + O(\delta^4) \leqslant 3\delta^2$$

for $\delta \leqslant \frac{4}{5}$, or

$$D_{kl}\left(P_1(\cdot \mid g_1, \ldots, g_{k-1}) | P_{-1}(\cdot \mid g_1, \ldots, g_{k-1})\right) \leqslant 3\delta^2$$

for $\delta \leqslant \frac{4}{5}$. Summing over k completes the proof. □

Putting it all together: a minimax lower bound With Lemma 5.2.9 in place along with our construction (5.2.5) of well-separated functions, we can now give a theorem on the best possible convergence guarantees for a broad family of problems.

Theorem 5.2.10. *Let* $C \subset \mathbb{R}^n$ *be a convex set containing an* ℓ_2 *ball of radius* R, *and let* $\mathcal{P}$ *denote the collection of distributions generating stochastic subgradients with* $\|g\|_2 \leqslant M$ *with probability* 1. *Then*

$$\mathfrak{M}_K(C, \mathcal{P}) \geqslant \frac{RM}{4\sqrt{6}\sqrt{K}}$$

for all $K \in \mathbb{N}$.

Proof. We combine Le Cam's method, Lemma 5.2.2 (and the subsequent Theorem 5.2.4) with our construction (5.2.5) and their stochastic subgradients (5.2.7). Certainly, the class of n-dimensional optimization problems is at least as challenging as a 1-dimensional problem (we may always restrict our functions to depend only on a single coordinate), so that for any $\delta \geqslant 0$ we have

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{\delta MR}{2}\left(1 - \sqrt{\frac{1}{2} D_{kl}\left(P_1^K | P_{-1}^K\right)}\right).$$

Now we use Lemma 5.2.9, which guarantees the further lower bound

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{\delta MR}{2}\left(1 - \sqrt{\frac{3K\delta^2}{2}}\right),$$

valid for all $\delta \leqslant \frac{4}{5}$. Choosing $\delta^2 = \frac{1}{6K} < \frac{4}{5}$, we have that $D_{kl}\left(P_1^K | P_{-1}^K\right) \leqslant \frac{1}{2}$, and

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{\delta MR}{4}.$$

Substituting our choice of δ into this expression gives the theorem. □

In short, Theorem 5.2.10 gives a guarantee that matches the upper bounds of the previous lectures to within a numerical constant factor of 10. A more careful inspection of our analysis allows us to prove a lower bound, at least as $K \to \infty$, of $1/8\sqrt{K}$. In particular, by using Theorem 3.4.7 of our lecture on subgradient methods, we find that if the set C contains an ℓ_2-ball of radius R_{inner} and is contained in an ℓ_2-ball of radius R_{outer}, we have

$$\frac{1}{\sqrt{96}} \frac{MR_{\text{inner}}}{\sqrt{K}} \leqslant \mathfrak{M}_K(C, \mathcal{F}) \leqslant \frac{MR_{\text{outer}}}{\sqrt{K}} \tag{5.2.11}$$

for all $K \in \mathbb{N}$, where the upper bound is attained by the stochastic projected subgradient method.

5.3. Multiple dimensions and Assouad's Method The results in Section 5.2 provide guarantees for problems where we can embed much of the difficulty of our family $\mathcal{F}$ in optimizing a pair of only two functions—something reminiscent of problems in classical statistics on the "hardest one-dimensional subproblem" (see, for example, the work of Donoho, Liu, and MacGibbon [19]). In many stochastic optimization problems, the higher-dimension n yields increased difficulty, so that we would like to derive bounds that incorporate dimension more directly. With that in mind, we develop a family of lower bounds, based on Assouad's method [2], that reduce optimization to a collection of binary hypothesis tests, one for each of the n dimensions of the problem.

More precisely, we let $\mathcal{V} = \{-1, 1\}^n$ be the n-dimensional binary hypercube, and for each $v \in \mathcal{V}$, we assume we have a function $f_v \in \mathcal{F}$ where $f_v : \mathbb{R}^n \to \mathbb{R}$. Without loss of generality, we will assume that our constraint set C has the point 0 in its interior. Let $\delta \in \mathbb{R}_+^n$ be an n-dimensional nonnegative vector. Then we say that the functions $\{f_v\}$ induce a *δ-separation in the Hamming metric* if for any $x \in C \subset \mathbb{R}^n$ we have

$$f_v(x) - f_v^\star \geqslant \sum_{j=1}^n \delta_j \mathbf{1}\left\{\text{sign}(x_j) \neq v_j\right\}, \tag{5.3.1}$$

where the subscript j denotes the jth coordinate. For example, if we define the function $f_v(x) = \delta \|x - v\|_1$ for each $v \in \mathcal{V}$, then certainly $\{f_v\}$ is $\delta\mathbf{1}$-separated in the Hamming metric; more generally, $f_v(x) = \sum_{j=1}^n \delta_j |x_j - v_j|$ is δ-separated. With this definition, we have the following lemma, providing a lower bound for functions $f : \mathbb{R}^n \to \mathbb{R}$.

Lemma 5.3.2 (Generalized Assouad)**.** *Let $\delta \in \mathbb{R}_+^n$ and let $\{f_v\}$, where $v \in \mathcal{V} = \{-1,1\}^n$, be δ-separated in Hamming metric. Let $\widehat{x}$ be any optimization algorithm, and let P_v be the distribution of (all) the subgradients $g_1, \ldots, g_K$ the procedure $\widehat{x}$ observes when optimizing f_v. Define*

$$P_{+j} = \frac{1}{2^{n-1}} \sum_{v : v_j = 1} P_v \text{ and } P_{-j} = \frac{1}{2^{n-1}} \sum_{v : v_j = -1} P_v.$$

Then

$$\frac{1}{2^n} \sum_{v \in \{-1,1\}^n} \mathbb{E}[f_v(\widehat{x}) - f_v^\star] \geqslant \frac{1}{2} \sum_{j=1}^n \delta_j (1 - \left\|P_{+j} - P_{-j}\right\|_{\mathrm{TV}}).$$

Proof. By using the separation condition, we immediately see that

$$\mathbb{E}[f_v(\widehat{x}) - f_v^\star] \geqslant \sum_{j=1}^d \delta_j P_v(\mathrm{sign}(\widehat{x}_j) \neq v_j)$$

for any $v \in \mathcal{V}$. Averaging over the vectors $v \in \mathcal{V}$, we obtain

$$\begin{aligned} \frac{1}{2^n} \sum_{v \in \mathcal{V}} \mathbb{E}[f_v(\widehat{x}) - f_v^\star] &\geqslant \sum_{j=1}^d \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \delta_j P_v(\mathrm{sign}(\widehat{x}_j) \neq v_j) \\ &= \sum_{j=1}^d \delta_j \frac{1}{|\mathcal{V}|} \left[\sum_{v : v_j = 1} P_v(\mathrm{sign}(\widehat{x}_j) \neq 1) + \sum_{v : v_j = -1} P_v(\mathrm{sign}(\widehat{x}_j) \neq -1) \right] \\ &= \sum_{j=1}^d \frac{\delta_j}{2} \left[P_{+j}(\mathrm{sign}(\widehat{x}_j) \neq 1) + P_{-j}(\mathrm{sign}(\widehat{x}_j) \neq -1) \right]. \end{aligned}$$

Now we use Le Cam's lemma (Lemma 5.2.2) on optimal binary hypothesis tests to see that

$$P_{+j}(\mathrm{sign}(\widehat{x}_j) \neq 1) + P_{-j}(\mathrm{sign}(\widehat{x}_j) \neq -1) \geqslant 1 - \left\|P_{+j} - P_{-j}\right\|_{\mathrm{TV}}$$

which gives the desired result. □

As a nearly immediate consequence of Lemma 5.3.2, we see that if the separation is a constant $\delta > 0$ for each coordinate, we have the following lower bound on the minimax risk.

Proposition 5.3.3. *Let the collection $\{f_v\}_{v \in \mathcal{V}} \subset \mathcal{F}$, where $\mathcal{V} = \{-1,1\}^n$, be δ-separated in Hamming metric for some $\delta \in \mathbb{R}_+$, and let the conditions of Lemma 5.3.2 hold. Then*

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{n}{2}\delta\left(1 - \sqrt{\frac{1}{2n}\sum_{j=1}^n D_{\mathrm{kl}}\left(P_{+j} | P_{-j}\right)}\right).$$

Proof. Lemma 5.3.2 guarantees that

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{\delta}{2} \sum_{j=1}^n (1 - \left\|P_{+j} - P_{-j}\right\|_{\mathrm{TV}}).$$

Applying the Cauchy-Schwarz inequality, we have

$$\sum_{j=1}^{n} \left\|P_{+j} - P_{-j}\right\|_{TV} \leqslant \sqrt{n \sum_{j=1}^{n} \left\|P_{+j} - P_{-j}\right\|_{TV}^{2}} \leqslant \sqrt{\frac{n}{2} \sum_{j=1}^{n} D_{kl}\left(P_{+j} | P_{-j}\right)}$$

by Pinsker's inequality. Substituting this into the previous bound gives the desired result. □

With this proposition, we can give a number of minimax lower bounds. We focus on two concrete cases, which show that the stochastic gradient procedures we have developed are optimal for a variety of problems. We give one result, deferring others to the exercises associated with the lecture notes. For our main result using Assouad's method, we consider optimization problems for which the set $C \subset \mathbb{R}^n$ contains an ℓ_∞ ball of radius R. We also assume that the stochastic gradient oracle satisfies the ℓ_1-bound condition

$$\mathbb{E}[\|g(x, S, f)\|_1^2] \leqslant M^2.$$

This means that all the functions $f \in \mathcal{F}$ are M-Lipschitz continuous with respect to the ℓ_∞-norm, that is, $|f(x) - f(y)| \leqslant M \|x - y\|_\infty$.

Theorem 5.3.4. *Let $\mathcal{F}$ and the stochastic gradient oracle be as above, and assume $C \supset [-R, R]^n$. Then*

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant RM \min\left\{\frac{1}{5}, \frac{1}{\sqrt{96}} \frac{\sqrt{n}}{\sqrt{K}}\right\}.$$

Proof. Our proof is similar to our construction of our earlier lower bounds, except that now we must construct functions defined on $\mathbb{R}^n$ so that our minimax lower bound on convergence rate grows with the dimension. Let $\delta > 0$ be fixed for now. For each $v \in \mathcal{V} = \{-1, 1\}^n$, define the function

$$f_v(x) := \frac{M\delta}{n} \|x - Rv\|_1.$$

Then by inspection, the collection $\{f_v\}$ is $\frac{MR\delta}{n}$-separated in Hamming metric, as

$$f_v(x) = \frac{M\delta}{n} \sum_{j=1}^{n} |x_j - Rv_j| \geqslant \frac{M\delta}{n} \sum_{j=1}^{n} R\mathbf{1}\left\{\text{sign}(x_j) \neq v_j\right\}.$$

Now, we must (as before) construct a stochastic subgradient oracle. Let $e_1, \ldots, e_n$ be the n standard basis vectors. For each $v \in \mathcal{V}$, we define the stochastic subgradient as

$$g(x, f_v) = \begin{cases} Me_j \,\text{sign}(x_j - Rv_j) & \text{with probability } \frac{1+\delta}{2n} \\ -Me_j \,\text{sign}(x_j - Rv_j) & \text{with probability } \frac{1-\delta}{2n}. \end{cases} \tag{5.3.5}$$

That is, the oracle randomly chooses a coordinate $j \in \{1, \ldots, n\}$, then conditional on this choice, flips a biased coin and with probability $\frac{1+\delta}{2}$ returns the correctly signed jth coordinate of the subgradient, $Me_j \,\text{sign}(x_j - Rv_j)$, and otherwise returns the negative. Letting $\text{sign}(x)$ denote the vector of signs of x, we then have the equality

$$\mathbb{E}[g(x, f_\nu)] = M \sum_{j=1}^{n} e_j \left[\frac{1+\delta}{n} \operatorname{sign}(x_j - R\nu_j) - \frac{1-\delta}{n} \operatorname{sign}(x_j - R\nu_j)\right] = \frac{M\delta}{n} \operatorname{sign}(x - R\nu).$$

That is, $\mathbb{E}[g(x, f_\nu)] \in \partial f_\nu(x)$ as desired.

Now, we apply Proposition 5.3.3, which guarantees that

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{MR\delta}{2} \left(1 - \sqrt{\frac{1}{2n} \sum_{j=1}^{n} D_{kl}\left(P_{+j} | P_{-j}\right)}\right). \tag{5.3.6}$$

It remains to upper bound the KL-divergence terms. Let P_ν^K denote the distribution of the K subgradients the method observes for the function f_ν, and let $\nu_{(\pm j)}$ denote the vector ν except that its jth entry is forced to be ± 1. Then, we may use the convexity of the KL-divergence to obtain that

$$D_{kl}\left(P_{+j} | P_{-j}\right) \leqslant \frac{1}{2^n} \sum_{\nu \in \mathcal{V}} D_{kl}\left(P^K_{\nu_{(+j)}} | P^K_{\nu_{(-j)}}\right).$$

Let us thus bound $D_{kl}\left(P_\nu^K | P_{\nu'}^K\right)$ when ν and ν' differ in only a single coordinate (we let it be the first coordinate with no loss of generality). Let us assume for notational simplicity $M = 1$ for the next calculation, as this only changes the support of the subgradient distribution (5.3.5) but not any divergences. Applying the chain rule (Lemma 5.2.8), we have

$$D_{kl}\left(P_\nu^K | P_{\nu'}^K\right) = \sum_{k=1}^{K} \mathbb{E}_{P_\nu}\left[D_{kl}\left(P_\nu(\cdot \mid g_{1:k-1}) | P_{\nu'}(\cdot \mid g_{1:k-1})\right)\right].$$

We consider one of the terms, noting that the kth query x_k is a function of $g_1, \ldots, g_{k-1}$. We have

$$D_{kl}\left(P_\nu(\cdot \mid x_k) | P_{\nu'}(\cdot \mid x_k)\right) = P_\nu(g = e_1 \mid x_k) \log \frac{P_\nu(g = e_1 \mid x_k)}{P_{\nu'}(g = e_1 \mid x_k)} + P_\nu(g = -e_1 \mid x_k) \log \frac{P_\nu(g = -e_1 \mid x_k)}{P_{\nu'}(g = -e_1 \mid x_k)},$$

because P_ν and $P_{\nu'}$ assign the same probability to all subgradients except when $g \in \{\pm e_1\}$. Continuing the derivation, we obtain

$$D_{kl}\left(P_\nu(\cdot \mid x_k) | P_{\nu'}(\cdot \mid x_k)\right) = \frac{1+\delta}{2n} \log \frac{1+\delta}{1-\delta} + \frac{1-\delta}{2n} \log \frac{1-\delta}{1+\delta} = \frac{\delta}{n} \log \frac{1+\delta}{1-\delta}.$$

Noting that this final quantity is bounded by $\frac{3\delta^2}{n}$ for $\delta \leqslant \frac{4}{5}$ gives that

$$D_{kl}\left(P_\nu^K | P_{\nu'}^K\right) \leqslant \frac{3K\delta^2}{n} \text{ if } \delta \leqslant \frac{4}{5}.$$

Substituting the preceding calculation into the lower bound (5.3.6), we obtain

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant \frac{MR\delta}{2} \left(1 - \sqrt{\frac{1}{2n} \sum_{j=1}^{n} \frac{3K\delta^2}{n}}\right) = \frac{MR\delta}{2} \left(1 - \sqrt{\frac{3K\delta^2}{2n}}\right).$$

Choosing $\delta^2 = \min\{16/25, \frac{n}{4K}\}$ gives the result of the theorem. □

A few remarks are in order about the theorem. First, we see that it recovers the 1-dimensional result of Theorem 5.2.10, as we may simply take $n = 1$ in the theorem statement. Second, we see that if we wish to optimize over a set larger than the ℓ_2-ball, then there must necessarily be some dimension-dependent penalty, at least in the worst case. Lastly, the result again is sharp. By using Theorem 3.4.7, we obtain the following corollary.

Corollary 5.3.7. *In addition to the conditions of Theorem 5.3.4, let* $C \subset \mathbb{R}^n$ *contain an* ℓ_∞ *box of radius* $R_{\rm inner}$ *and be contained in an* ℓ_∞ *box of radius* $R_{\rm outer}$. *Then*

$$R_{\rm inner} M \min\left\{\frac{1}{5}, \frac{1}{\sqrt{96}} \frac{\sqrt{n}}{\sqrt{K}}\right\} \leqslant \mathfrak{M}_K(C, \mathcal{F}) \leqslant R_{\rm outer} M \min\left\{1, \frac{\sqrt{n}}{\sqrt{K}}\right\}.$$

Notes and further reading The minimax criterion for measuring optimality of optimization and estimation procedures has a long history, dating back at least to Wald [59] in 1939. The information-theoretic approach to optimality guarantees was extensively developed by Ibragimov and Has'minskii [31], and this is our approach. Our treatment in this chapter is specifically based off of that by Agarwal et al. [1] for proving lower bounds for stochastic optimization problems, though our results appear to have slightly sharper constants. Notably missing in our treatment is the use of Fano's inequality for lower bounds, which is commonly used to prove converse statements to achievability results in information theory [17,62]. Recent treatments of various techniques for proving lower bounds in statistics can be found in the book of Tsybakov [58] or the lecture notes [21].

Our focus on stochastic optimization problems allows reasonably straightforward reductions from optimization to statistical testing problems, for which information theoretic and statistical tools give elegant solutions. It is possible to give lower bounds for non-stochastic problems, where the classical reference is the book of Nemirovski and Yudin [41] (who also provide optimality guarantees for stochastic problems). The basic idea is to provide lower bounds for the *oracle model* of convex optimization, where we consider optimality in terms of the number of queries to an oracle giving true first- or second-order information (as opposed to the stochastic oracle studied here). More recent work, including the lecture notes [42] and the book [44] provide a somewhat easier guide to such results, while the recent paper of Braun et al. [13] shows how to leverage information-theoretic tools to prove optimality guarantees even for non-stochastic optimization problems.

A. Technical Appendices

A.1. Continuity of Convex Functions In this appendix, we provide proofs of the basic continuity results for convex functions. Our arguments are based on those of Hiriart-Urruty and Lemaréchal [27].

Proof of Lemma 2.3.1 We can write $x \in B_1$ as $x = \sum_{i=1}^n x_i e_i$, where e_i are the standard basis vectors and $\sum_{i=1}^n |x_i| \leqslant 1$. Thus, we have

$$\begin{aligned} f(x) = f\left(\sum_{i=1}^n e_i x_i\right) &= f\left(\sum_{i=1}^n |x_i| \operatorname{sign}(x_i) e_i + (1 - \|x\|_1) 0\right) \\ &\leqslant \sum_{i=1}^n |x_i| f(\operatorname{sign}(x_i) e_i) + (1 - \|x\|_1) f(0) \\ &\leqslant \max\{f(e_1), f(-e_1), f(e_2), f(-e_2), \ldots, f(e_n), f(-e_n), f(0)\}. \end{aligned}$$

The first inequality uses the fact that the $|x_i|$ and $(1 - \|x\|_1)$ form a convex combination, since $x \in B_1$, as does the second.

For the lower bound, note by the fact that $x \in \operatorname{int} B_1$ satisfies $x \in \operatorname{int} \operatorname{dom} f$, we have $\partial f(x) \neq \emptyset$ by Theorem 2.4.3. In particular, there is a vector g such that $f(y) \geqslant f(x) + \langle g, y - x \rangle$ for all y, and even more,

$$f(y) \geqslant f(x) + \inf_{y \in B_1} \langle g, y - x \rangle \geqslant f(x) - 2 \|g\|_\infty$$

for all $y \in B_1$.

Proof of Theorem 2.3.2 First, let us suppose that for each point $x_0 \in C$, there exists an open ball $B \subset \operatorname{int} \operatorname{dom} f$ such that

$$|f(x) - f(x')| \leqslant L \left\|x - x'\right\|_2 \text{ for all } x, x' \in B. \tag{A.1.1}$$

The collection of such balls B covers C, and as C is compact, there exists a finite subcover $B_1, \ldots, B_k$ with associated Lipschitz constants $L_1, \ldots, L_k$. Take $L = \max_i L_i$ to obtain the result. It thus remains to show that we can construct balls satisfying the Lipschitz condition (A.1.1) at each point $x_0 \in C$.

With that in mind, we use Lemma 2.3.1, which shows that for each point x_0, there is some $\epsilon > 0$ and $-\infty < m \leqslant M < \infty$ such that

$$-\infty < m \leqslant \inf_{v : \|v\|_2 \leqslant 2\epsilon} f(x + v) \leqslant \sup_{v : \|v\|_2 \leqslant 2\epsilon} f(x + v) \leqslant M < \infty.$$

We make the following claim, from which the condition (A.1.1) evidently follows based on the preceding display.

Lemma A.1.2. *Let $\epsilon > 0$, f be convex, and $B = \{v : \|v\|_2 \leqslant 1\}$. Suppose that $f(x) \in [m, M]$ for all $x \in x_0 + 2\epsilon B$. Then*

$$|f(x) - f(x')| \leqslant \frac{M - m}{\epsilon} \left\|x - x'\right\|_2 \text{ for all } x, x' \in x_0 + \epsilon B.$$

Proof. Let $x, x' \in x_0 + \epsilon B$. Let

$$x'' = x' + \epsilon \frac{x' - x}{\|x' - x\|_2} \in x_0 + 2\epsilon B,$$

as $(x - x')/\left\|x - x'\right\|_2 \in B$. By construction, we have that $x' \in \{tx + (1-t)x'', t \in [0,1]\}$, the segment between x and x''; explicitly,

$$\left(1 + \frac{\epsilon}{\|x' - x\|_2}\right) x' = x'' + \frac{\epsilon}{\|x' - x\|_2} x \text{ or } x' = \frac{\|x' - x\|_2}{\|x' - x\|_2 + \epsilon} x'' + \frac{\epsilon}{\|x' - x\|_2 + \epsilon} x.$$

Then we find that

$$f(x') \leqslant \frac{\|x - x'\|_2}{\|x - x'\|_2 + \epsilon} f(x'') + \frac{\epsilon}{\|x - x'\|_2 + \epsilon} f(x),$$

or

$$\begin{aligned} f(x') - f(x) &\leqslant \frac{\|x - x'\|_2}{\|x - x'\|_2 + \epsilon} \left[f(x'') - f(x)\right] \leqslant \frac{\|x - x'\|_2}{\|x - x'\|_2 + \epsilon} [M - m] \\ &\leqslant \frac{M - m}{\epsilon} \left\|x - x'\right\|_2. \end{aligned}$$

Swapping the roles of x and x' gives the result. □

A.2. Probability background In this section, we very tersely review a few of the necessary definitions and results that we employ here. We provide a non measure-theoretic treatment, as it is not essential for the basic uses we have.

Definition A.2.1. A sequence $X_1, X_2, \ldots$ of random vectors *converges in probability* to a random vector X_∞ if for all $\epsilon > 0$, we have

$$\limsup_{n \to \infty} \mathbb{P}(\|X_n - X_\infty\| > \epsilon) = 0.$$

Definition A.2.2. A sequence $X_1, X_2, \ldots$ of random vectors is a *martingale* if there is a sequence of random variables $Z_1, Z_2, \ldots$ (which may contain all the information about $X_1, X_2, \ldots$) such that for each n, (i) X_n is a function of Z_n, (ii) Z_{n-1} is a function of Z_n, and (iii) we have the conditional expectation condition

$$\mathbb{E}[X_n \mid Z_{n-1}] = X_{n-1}.$$

When condition (i) is satisfied, we say that X_n is *adapted to* Z. We say that a sequence $X_1, X_2, \ldots$ is a *martingale difference sequence* if $S_n = \sum_{i=1}^n X_i$ is a martingale, or, equivalently, if $\mathbb{E}[X_n \mid Z_{n-1}] = 0$.

We now provide a self-contained proof of the Azuma-Hoeffding inequality. Our first result is an important intermediate result.

Lemma A.2.3 (Hoeffding's Lemma [30]). *Let X be a random variable with $a \leqslant X \leqslant b$. Then*

$$\mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X]))\right] \leqslant \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right) \text{ for all } \lambda \in \mathbb{R}.$$

Proof. First, we note that if Y is any random variable with $Y \in [c_1, c_2]$, then $\mathrm{Var}(Y) \leqslant \frac{(c_2 - c_1)^2}{4}$. Indeed, we have that $\mathbb{E}[Y]$ minimizes $\mathbb{E}[(Y-t)^2]$ over $t \in \mathbb{R}$, so that

(A.2.4)
$$\mathrm{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2] \leqslant \mathbb{E}\left[\left(Y - \frac{c_2 + c_1}{2}\right)^2\right] \leqslant \left(c_2 - \frac{c_2 + c_1}{2}\right)^2 = \frac{(c_2 - c_1)^2}{4}.$$

Without loss of generality, we assume $\mathbb{E}[X] = 0$ and $0 \in [a, b]$. Let $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$. Then

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \text{ and } \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[Xe^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2}.$$

Note that $\psi'(0) = \mathbb{E}[X] = 0$. Let P denote the distribution of X, and assume without loss of generality that X has a density p.[5] Define the random variable Y to have the shifted density f defined by

$$f(y) = \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda X}]} p(y)$$

for $y \in \mathbb{R}$, where $p(y) = 0$ for $y \notin [a, b]$. Then $\mathbb{E}[Y] = \psi'(\lambda)$ and $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 = \psi''(\lambda)$. But of course, we know that $Y \in [a, b]$ because the distribution P of X is supported on $[a, b]$, so that

$$\psi''(\lambda) = \mathrm{Var}(Y) \leqslant \frac{(b-a)^2}{4}$$

by inequality (A.2.4). Using Taylor's theorem, we have that

$$\psi(\lambda) = \psi(0) + \underbrace{\psi'(0)}_{=0}\lambda + \frac{\lambda^2}{2}\psi''(\widetilde{\lambda})\psi(\lambda) = \psi(0) + \frac{\lambda^2}{2}\psi''(\widetilde{\lambda})$$

for some $\widetilde{\lambda}$ between 0 and λ. But $\psi''(\widetilde{\lambda}) \leqslant \frac{(b-a)^2}{4}$, so that $\psi(\lambda) \leqslant \frac{\lambda^2}{2}\frac{(b-a)^2}{4}$ as desired. □

Theorem A.2.5 (Azuma-Hoeffding Inequality [4]). *Let $X_1, X_2, \ldots$ be a martingale difference sequence with $|X_i| \leqslant B$ for all $i = 1, 2, \ldots$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{nB^2}\right)$$

and

$$\mathbb{P}\left(\sum_{i=1}^n X_i \leqslant -t\right) \leqslant \exp\left(-\frac{2t^2}{nB^2}\right)$$

for all $t \geqslant 0$.

Proof. We prove the upper tail, as the lower tail is similar. The proof is a nearly immediate consequence of Hoeffding's lemma (Lemma A.2.3) and the Chernoff bound technique. Indeed, we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geqslant t\right) \leqslant \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] \exp(-\lambda t)$$

[5] We may assume there is a dominating base measure μ with respect to which P has a density p.

for all $\lambda \geqslant 0$. Now, letting Z_i be the sequence to which the X_i are adapted, we iterate conditional expectations. We have

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\exp(\lambda X_n) \mid Z_{n-1}\right]\right] \\
&= \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right)\mathbb{E}[\exp(\lambda X_n) \mid Z_{n-1}]\right] \\
&\leqslant \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n-1} X_i\right) e^{\frac{\lambda^2 B^2}{8}}\right]
\end{aligned}$$

because $X_1, \ldots, X_{n-1}$ are functions of Z_{n-1}. By iteratively applying this calculation, we arrive at

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] \leqslant \exp\left(\frac{\lambda^2 n B^2}{8}\right). \tag{A.2.6}$$

Now we optimize by choosing $\lambda \geqslant 0$ to minimize the upper bound that inequality (A.2.6) provides, namely

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geqslant t\right) \leqslant \inf_{\lambda \geqslant 0} \exp\left(\frac{\lambda^2 n B^2}{8} - \lambda t\right) = \exp\left(-\frac{2t^2}{nB^2}\right)$$

by taking $\lambda = \frac{4t}{Bn}$. □

A.3. Auxiliary results on divergences We present a few standard results on divergences without proof, referring to standard references (e.g. the book of Cover and Thomas [17] or the extensive paper on divergence measures by Liese and Vajda [35]). Nonetheless, we state and prove a few results. The first is known as the *data processing inequality*, and it says that processing a random variable (even adding noise to it) can only make distributions closer together. See Cover and Thomas [17] or Theorem 14 of Liese and Vajda [35] for a proof.

Proposition A.3.1 (Data processing)**.** *Let P_0 and P_1 be distributions on a random variable $S \in \mathcal{S}$, and let $Q(\cdot \mid s)$ denote any conditional probability distribution conditioned on s, and define*

$$Q_v(A) = \int Q(A \mid s) dP_v(s)$$

for $v = 0, 1$ and all sets A. Then

$$\|Q_0 - Q_1\|_{TV} \leqslant \|P_0 - P_1\|_{TV} \text{ and } D_{kl}(Q_0|Q_1) \leqslant D_{kl}(P_0|P_1).$$

This proposition is perhaps somewhat intuitive: it says that if we do any processing on a random variable $S \sim P$, then there is less "information" about the initial distribution of P than if we did no further processing.

A consequence of this result is Pinsker's inequality.

Proposition A.3.2 (Pinsker's inequality)**.** *Let* P *and* Q *be arbitrary distributions. Then*

$$\|P - Q\|_{TV}^2 \leqslant \frac{1}{2} D_{kl}\left(P|Q\right).$$

Proof. First, we note that if we show the result assuming that the sample space $\mathcal{S}$ on which P and Q are defined is finite, we have the general result. Indeed, suppose that $A \subset \mathcal{S}$ achieves the supremum

$$\|P - Q\|_{TV} = \sup_{A \subset \mathcal{S}} |P(A) - Q(A)|.$$

(We may assume without loss of generality that such a set exists.) Then if we define $\widetilde{P}$ and $\widetilde{Q}$ to be the binary distributions with $\widetilde{P}(0) = P(A)$ and $\widetilde{P}(1) = 1 - P(A)$, and similarly for $\widetilde{Q}$, we have $\|P - Q\|_{TV} = \|\widetilde{P} - \widetilde{Q}\|_{TV}$, and Proposition A.3.1 immediately guarantees that

$$D_{kl}(\widetilde{P}|\widetilde{Q}) \leqslant D_{kl}\left(P|Q\right).$$

Let us assume then that $|\mathcal{S}| < \infty$.

In this case, Pinsker's inequality is an immediate consequence of the strong convexity of the negative entropy functional $h(p) = \sum_{i=1}^n p_i \log p_i$ with respect to the ℓ_1-norm over the probability simplex. For completeness, let us prove this. Let p and $q \in \mathbb{R}_+^n$ satisfy $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i = 1$. Then Taylor's theorem guarantees that

$$h(q) = h(p) + \langle \nabla h(p), q - p \rangle + \frac{1}{2}(q - p)^\top \nabla^2 h(\widetilde{q})(q - p),$$

where $\widetilde{q} = \lambda p + (1 - \lambda) q$ for some $\lambda \in [0, 1]$. Now, we note that

$$\nabla^2 h(p) = \text{diag}(1/p_1, \ldots, 1/p_n),$$

and using that $\nabla h(p) = [\log p_i + 1]_{i=1}^n$, we find

$$h(q) = h(p) + \sum_{i=1}^n (q_i - p_i) \log p_i + \frac{1}{2} \sum_{i=1}^n \frac{(q_i - p_i)^2}{\widetilde{q}_i}.$$

Using the Cauchy-Schwarz inequality, we have

$$\left(\sum_{i=1}^n |q_i - p_i|\right)^2 = \left(\sum_{i=1}^n \sqrt{\widetilde{q}_i} \frac{|q_i - p_i|}{\sqrt{\widetilde{q}_i}}\right)^2 \leqslant \left(\sum_{i=1}^n \widetilde{q}_i\right)\left(\sum_{i=1}^n \frac{(q_i - p_i)^2}{\widetilde{q}_i}\right).$$

Of course, this gives

$$h(q) \geqslant h(p) + \sum_{i=1}^n (q_i - p_i) \log p_i + \frac{1}{2} \|p - q\|_1^2.$$

Rearranging this, we have $h(q) - h(p) - \langle \nabla h(p), q - p \rangle = \sum_{i=1}^n q_i \log \frac{q_i}{p_i}$, or that

$$D_{kl}\left(q|p\right) \geqslant \frac{1}{2} \|p - q\|_1^2 = 2 \|P - Q\|_{TV}^2.$$

This is the result. □

B. Questions and Exercises

Exercises for Lecture 2

Question 1: Let $\pi_C(x) := \mathrm{argmin}_{y \in C} \|x - y\|_2$ denote the Euclidean projection of x onto the set C, where C is closed convex. Show that the projection is a Lipschitz mapping, that is,

$$\|\pi_C(x_0) - \pi_C(x_1)\|_2 \leqslant \|x_0 - x_1\|_2$$

for all vectors x_0, x_1. Show that, even if C is compact, this inequality cannot (in general) be improved.

Question 2: Let $\mathbb{S}_n = \{A \in \mathbb{R}^{n \times n} : A = A^T\}$ be the set of $n \times n$ symmetric matrices. Let $f(X) = \lambda_{\max}(X)$ for $X \in \mathbb{S}_n$. Show that f is convex and compute $\partial f(X)$.

Question 3: A convex function f is called λ strongly convex with respect to the norm $\| \cdot \|$ on the (convex) domain X if for any $x, y \in X$, we have

$$f(y) \geqslant f(x) + \langle g, y - x \rangle + \frac{\lambda}{2}\|x - y\|^2$$

for all $g \in \partial f(x)$. Recall that a function f is L-Lipschitz continuous with respect to the norm $\|\cdot\|$ on the domain X if

$$|f(x) - f(y)| \leqslant L \|x - y\| \text{ for all } x, y \in X.$$

Let f be λ-strongly convex w.r.t. $\|\cdot\|$ and h_1, h_2 be L-Lipschitz continuous convex functions with respect to the norm $\|\cdot\|$. For $i = 1, 2$ define

$$x_i = \arg\min_{x \in X}\{f(x) + h_i(x)\}.$$

Show that

$$\|x_1 - x_2\| \leqslant \frac{2L}{\lambda}.$$

Hint: You may use the fact, demonstrated in the notes, that if h is L-Lipschitz and convex, then $\|g\|_* \leqslant L$ for all $g \in \partial h(x)$, where $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$.

Question 4 (Hölder's inequality)**:** Let x and y be vectors in $\mathbb{R}^n$ and $p, q \in (1, \infty)$ be conjugate, that is, satisfy $1/p + 1/q = 1$. In this question, we will show that $\langle x, y \rangle \leqslant \|x\|_p \|y\|_q$, and moreover, that $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms. (The result is essentially immediate in the case that $p = 1$ and $q = \infty$.)

(a) Show that for any $a, b \geqslant 0$ and any $\eta \geqslant 0$, we have

$$ab \leqslant \frac{\eta^p}{p} a^p + \frac{1}{\eta^q q} b^q.$$

Hint: use the concavity of the logarithm and that $1/p + 1/q = 1$.

(b) Show that $\langle x, y \rangle \leqslant \frac{\eta^p}{p} \|x\|_p^p + \frac{1}{\eta^q q} \|y\|_q^q$ for all $\eta > 0$.

(c) Using the result of part (b), show that $\langle x, y \rangle \leqslant \|x\|_p \|y\|_q$.

(d) Show that $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms.

Exercises for Lecture 3

Question 5: In this question and the next, we perform experiments with (stochastic) subgradient methods to train a handwritten digit recognition classifier (one to recognize the digits $\{0, 1, \ldots, 9\}$). A warning: we use optimization notation here, consistent with Example 3.4, which is non-standard for typical machine learning or statistical learning applications.

We represent a multiclass classifier using a matrix

$$X = [x_1 \; x_2 \; \cdots \; x_k] \in \mathbb{R}^{d \times k},$$

where there are k classes, and the predicted class for a data vector $a \in \mathbb{R}^d$ is

$$\operatorname*{argmax}_{l \in [k]} \langle a, x_l \rangle = \operatorname*{argmax}_{l \in [k]} \{[X^T a]_l\}.$$

We represent data as pairs $(a, b) \in \mathbb{R}^d \times \{1, \ldots, k\}$, where a is the data point (features) and b the label of the data point. We use the *multiclass hinge* loss function

$$F(X; (a, b)) = \max_{l \neq b} \left[1 + \langle a, x_l - x_b \rangle\right]_+$$

where $[t]_+ = \max\{t, 0\}$ denotes the positive part. We will use stochastic gradient descent to attempt to minimize

$$f(X) := \mathbb{E}_P[F(X; (A, B))] = \int F(X; (a, b)) dP(a, b),$$

where the expectation is taken over pairs (A, B).

(a) Show that F is convex.

(b) Show that $F(X; (a, b)) = 0$ if and only if the classifer represented by X has a *large margin*, meaning that

$$\langle a, x_b \rangle \geqslant \langle a, x_l \rangle + 1 \text{ for all } l \neq b.$$

(c) For a pair (a, b), give a way to calculate a vector $G \in \partial F(X; (a, b))$ (note that $G \in \mathbb{R}^{d \times k}$).

Question 6: In this problem, you will perform experiments to explore the performance of stochastic subgradient methods for classification problems, specifically, a handwritten digit recognition problem using zip code data from the United States Postal Service (this data is taken from the book [24], originally due to Yann Le Cunn). The data—training data `zip.train`, test data `zip.test`, and information file `zip.inf`—are available for download from the zipped tar file `http://web.stanford.edu/~jduchi/PCMIConvex/ZIPCodes.tgz`. Starter code is available for `julia` and `Matlab` at the following urls.

i. For Julia: `http://web.stanford.edu/~jduchi/PCMIConvex/sgd.jl`

ii. For Matlab: `http://web.stanford.edu/~jduchi/PCMIConvex/matlab.tgz`

There are two methods left un-implemented in the starter code: the `sgd` method and the `MulticlassSVMSubgradient` method. Implement these methods (you may find the code for unit-testing the multiclass SVM subgradient useful to double check your implementation). For the SGD method, your stepsizes should be proportional to $\alpha_i \propto 1/\sqrt{i}$, and you should project X to the Frobenius norm ball

$$B_r := \{X \in \mathbb{R}^{d \times k} : \|X\|_{\mathrm{Fr}} \leqslant r\}, \text{ where } \|X\|_{\mathrm{Fr}}^2 = \sum_{ij} X_{ij}^2.$$

We have implemented a pre-processing step that also *kernelizes* the data representation. Let the function $K(a, a') = \exp(-\frac{1}{2\tau}\|a - a'\|_2^2)$. Then the kernelized data representation transforms each datapoint $a \in \mathbb{R}^d$ into a vector

$$\phi(a) = \left[K(a, a_{i_1})\ K(a, a_{i_2})\ \cdots\ K(a, a_{i_m})\right]^\top$$

where $i_1, \ldots, i_m$ is a random subset of $\{1, \ldots, N\}$ (see `GetKernelRepresentation`.)

Once you have implemented the `sgd` and `MulticlassSVMSubgradient` methods, use the method `RunExperiment` (Julia/Matlab). What performance do you get in classification? Which digits is your classifier most likely to confuse?

Question 7: In this problem, we give a simple bound on the rate of convergence for stochastic optimization for minimization of strongly convex functions. Let C denote a compact convex set and f denote a λ-strongly convex function with respect to the ℓ_2-norm on C, meaning that

$$f(y) \geqslant f(x) + \langle g, y - x\rangle + \frac{\lambda}{2}\|x - y\|_2^2 \text{ for all } g \in \partial f(x),\ x, y \in C.$$

Consider the following stochastic gradient method: at iteration k, we

i. receive a noisy subgradient g_k with $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$

ii. perform the projected subgradient step

$$x_{k+1} = \pi_C(x_k - \alpha_k g_k).$$

Show that if $\mathbb{E}[\|g_k\|_2^2] \leqslant M^2$ for all k, then with the stepsize choice $\alpha_k = \frac{1}{\lambda k}$, we have the convergence guarantee

$$\mathbb{E}\left[\sum_{k=1}^K (f(x_k) - f(x^*))\right] \leqslant \frac{M^2}{2\lambda}(\log K + 1).$$

Exercises for Lecture 4

Question 8: We saw in the lecture that if we use mirror descent,

$$x_{k+1} = \operatorname*{argmin}_{x \in C}\left\{\langle g_k, x\rangle + \frac{1}{\alpha_k} D_h(x, x_k)\right\},$$

in the stochastic setting with $\mathbb{E}[g_k \mid x_k] \in \partial f(x_k)$ then we have the *regret* bound

$$\mathbb{E}\left[\sum_{k=1}^K (f(x_k) - f(x^\star))\right] \leqslant \mathbb{E}\left[\frac{1}{\alpha_K} R^2 + \frac{1}{2}\sum_{k=1}^K \alpha_k \|g_k\|_*^2\right].$$

Here we have assumed that $D_h(x^\star, x_k) \leqslant R^2$ for all k. We now use this inequality to prove Corollary 4.3.3.

In particular, choose the stepsize α_k adaptively at the kth step by optimizing the convergence bound up to the current iterate, that is, set

$$\alpha_k = R\left(\sum_{i=1}^{k} \|g_i\|_*^2\right)^{-\frac{1}{2}},$$

based on the previous subgradients. Prove that in this case one has

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^*))\right] \leqslant 3R\mathbb{E}\left[\left(\sum_{k=1}^{K}\|g_k\|_*^2\right)^{\frac{1}{2}}\right]$$

Conclude Corollary 4.3.3.

Hint: An intermediate step, which may be useful, is to prove the following inequality: for any non-negative sequence $a_1, a_2, \ldots, a_k$, one has

$$\sum_{i=1}^{k} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leqslant 2\sqrt{\sum_{i=1}^{k} a_i}.$$

Induction is one natural strategy.

Question 9 (Strong convexity of ℓ_p-norms)**:** Prove the claim of Example 4.3. That is, for some fixed $p \in (1, 2]$, if $h(x) = \frac{1}{2(p-1)}\|x\|_p^2$, show that h is strongly convex with respect to the ℓ_p-norm.

Hint: Let $\Psi(t) = \frac{1}{2(p-1)}t^{2/p}$ and $\phi(t) = |t|^p$, noting that $h(x) = \Psi(\sum_{j=1}^{n}\phi(x_j))$. Then by a Taylor expansion, this question is equivalent to showing that for any $w, x \in \mathbb{R}^n$, we have

$$x^\top \nabla^2 h(w) x \geqslant \|x\|_p^2$$

where, defining the shorthand vector $\nabla\phi(w) = [\phi'(w_1) \ \cdots \ \phi'(w_n)]^\top$, we have

$$\begin{aligned}\nabla^2 h(w) = \Psi''\left(\sum_{j=1}^{n}\phi(w_j)\right)\nabla\phi(w)\nabla\phi(w)^\top \\ + \Psi'\left(\sum_{j=1}^{n}\phi(w_j)\right)\operatorname{diag}\left(\phi''(w_1), \ldots, \phi''(w_n)\right).\end{aligned}$$

Now apply an argument similar to that used in Example 4.2 to show the strong convexity of $h(x) = \sum_j x_j \log x_j$, but applying Hölder's inequality instead of Cauchy-Schwarz.

Question 10 (Variable metric methods and AdaGrad)**:** Consider the following variable-metric method for minimizing a convex function f on a convex set $C \subset \mathbb{R}^n$:

$$x_{k+1} = \operatorname*{argmin}_{x \in C}\left\{\langle g_k, x\rangle + \frac{1}{2}(x - x_k)^\top H_k (x - x_k)\right\},$$

where $\mathbb{E}[g_k] \in \partial f(x_k)$. In the lecture, we showed that

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^\star))\right] \leqslant \frac{1}{2}\mathbb{E}\left[\sum_{k=2}^{K}\left(\|x_k - x^\star\|_{H_k}^2 - \|x_k - x^\star\|_{H_{k-1}}^2\right) + \|x_1 - x^\star\|_{H_1}^2\right]$$
$$+ \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K}\|g_k\|_{H_k^{-1}}^2\right].$$

(a) Let

$$H_k = \operatorname{diag}\left(\sum_{i=1}^{k} g_i g_i^\top\right)^{\frac{1}{2}}$$

be the diagonal matrix whose entries are the square roots of the sum of the squares of the gradient coordinates. (This is the AdaGrad method.) Show that

$$\|x_k - x^\star\|_{H_k}^2 - \|x_k - x^\star\|_{H_{k-1}}^2 \leqslant \|x_k - x^\star\|_\infty \operatorname{tr}(H_k - H_{k-1}),$$

where $\operatorname{tr}(A) = \sum_{i=1}^{n} A_{ii}$ is the trace of the matrix

(b) Assume that $R_\infty = \sup_{x \in C}\|x - x^\star\|_\infty$ is finite. Show that with any choice of diagonal matrix H_k, we obtain

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^\star))\right] \leqslant \frac{1}{2}R_\infty \mathbb{E}[\operatorname{tr}(H_K)] + \frac{1}{2}\mathbb{E}\left[\sum_{k=1}^{K}\|g_k\|_{H_k^{-1}}^2\right].$$

(c) Let $g_{k,j}$ denote the jth coordinate of the kth subgradient. Let H_k be chosen as above. Show that

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^\star))\right] \leqslant \frac{3}{2}R_\infty \sum_{j=1}^{n}\mathbb{E}\left[\left(\sum_{k=1}^{K} g_{k,j}^2\right)^{\frac{1}{2}}\right].$$

(d) Suppose that the domain $C = \{x : \|x\|_\infty \leqslant 1\}$. What is the expected regret of AdaGrad? Show that (to a numerical constant factor we ignore) this expected regret is *always* smaller than the expected regret bound for standard projected gradient descent, which is

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^\star))\right] \leqslant O(1)\sup_{x \in C}\|x - x^\star\|_2\,\mathbb{E}\left[\sum_{k=1}^{K}\|g_k\|_2^2\right]^{\frac{1}{2}}.$$

Hint: Use Cauchy-Schwarz

(e) As in the previous sub-question, assume that $C = \{x : \|x\|_\infty \leqslant 1\}$. Suppose that the subgradients are such that $g_k \in \{-1, 0, 1\}^n$ for all k, and that for each coordinate j we have $\mathbb{P}(g_{k,j} \neq 0) = p_j$. Show that AdaGrad has convergence guarantee

$$\mathbb{E}\left[\sum_{k=1}^{K}(f(x_k) - f(x^\star))\right] \leqslant \frac{3\sqrt{K}}{2}\sum_{j=1}^{n}\sqrt{p_j}.$$

What is the corresponding bound for standard projected gradient descent? How much better can AdaGrad be?

Exercises for Lecture 5

Question 11: In this problem, we prove a lower bound for strongly convex optimization problems. Suppose at each iteration of the optimization procedure, we receive a noisy subgradient g_k satisfying

$$g_k = \nabla f(x_k) + \xi_k, \ \xi_k \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

To prove a lower bound for optimization procedures, we use the functions

$$f_\nu(x) = \frac{\lambda}{2}(x - \nu\delta)^2, \ \nu \in \{\pm 1\}.$$

Let $f_\nu^\star = 0$ denote the minimum function values for f_ν on $\mathbb{R}$ for $\nu = \pm 1$.

(a) Recall the separation between two functions f_1 and f_{-1} as defined previously (5.1.4),

$$d_{\text{opt}}(f_{-1}, f_1; C) :=$$
$$\sup\left\{\delta \geqslant 0 : \begin{array}{l} f_1(x) \leqslant f_1^\star + \delta \text{ implies } f_{-1}(x) \geqslant f_{-1}^\star + \delta \\ f_{-1}(x) \leqslant f_{-1}^\star + \delta \text{ implies } f_1(x) \geqslant f_1^\star + \delta \end{array} \text{ for any } x \in C.\right\}.$$

When $C = \mathbb{R}$ (or, more generally, as long as $C \supset [-\delta, \delta]$), show that

$$d_{\text{opt}}(f_{-1}, f_1; C) \geqslant \frac{\lambda}{2}\delta^2.$$

(b) Show that the Kullback-Leibler divergence between two normal distributions $P_1 = N(\mu_1, \sigma^2)$ and $P_2 = N(\mu_2, \sigma^2)$ is

$$D_{\text{kl}}\left(P_1 | P_{-1}\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}.$$

(c) Use Le Cam's method to show the following lower bound for stochastic optimization: for any optimization procedure $\widehat{x}_K$ using K noisy gradient evaluations,

$$\max_{\nu \in \{-1,1\}} \mathbb{E}_{P_\nu}[f_\nu(\widehat{x}_K) - f_\nu^\star] \geqslant \frac{\sigma^2}{32\lambda K}.$$

Compare the result with the regret upper bound in problem 7. *Hint:* If P_ν^K denotes the distribution of the K noisy gradients for function f_ν, show that

$$D_{\text{kl}}\left(P_1^K | P_{-1}^K\right) \leqslant \frac{2K\lambda^2\delta^2}{\sigma^2}.$$

Question 12: Let $C = \{x \in \mathbb{R}^n : \|x\|_\infty \leqslant 1\}$, and consider the collection of functions $\mathcal{F}$ where the stochastic gradient oracle $g : \mathbb{R}^n \times \mathcal{S} \times \mathcal{F} \to \{-1, 0, 1\}^n$ satisfies

$$\mathbb{P}(g_j(x, S, f) \neq 0) \leqslant p_j$$

for each coordinate $j = 1, 2, \ldots, n$. Show that, for large enough $K \in \mathbb{N}$, a minimax lower bound for this class of functions and the given stochastic oracle is

$$\mathfrak{M}_K(C, \mathcal{F}) \geqslant c\frac{1}{\sqrt{K}}\sum_{j=1}^{n}\sqrt{p_j},$$

where $c > 0$ is a numerical constant. How does this compare to the convergence guarantee that AdaGrad gives?

References

[1] Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright, *Information-theoretic lower bounds on the oracle complexity of convex optimization*, IEEE Transactions on Information Theory **58** (2012), no. 5, 3235–3249. ←74

[2] P. Assouad, *Deux remarques sur l'estimation*, Comptes Rendus des Séances de l'Académie des Sciences, Série I **296** (1983), no. 23, 1021–1024. ←70

[3] P. Auer, N. Cesa-Bianchi, and C. Gentile, *Adaptive and self-confident on-line learning algorithms*, Journal of Computer and System Sciences **64** (2002), no. 1, 48–75. ←60

[4] K. Azuma, *Weighted sums of certain dependent random variables*, Tohoku Mathematical Journal **68** (1967), 357–367. ←77

[5] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters **31** (2003), 167–175. ←59

[6] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski, *Robust optimization*, Princeton University Press, 2009. ←4

[7] D. P. Bertsekas, *Stochastic optimization problems with nondifferentiable cost functionals*, Journal of Optimization Theory and Applications **12** (1973), no. 2, 218–231. ←22

[8] Dimitri P. Bertsekas, *Convex optimization theory*, Athena Scientific, 2009. ←3, 24

[9] D.P. Bertsekas, *Nonlinear programming*, Athena Scientific, 1999. ←3

[10] Stephen Boyd, John Duchi, and Lieven Vandenberghe, *Subgradients*, 2015. Course notes for Stanford Course EE364b. ←43

[11] Stephen Boyd and Almir Mutapcic, *Stochastic subgradient methods*, 2007. Course notes for EE364b at Stanford, available at http://www.stanford.edu/class/ee364b/notes/stoch_subgrad_notes.pdf. ←43

[12] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004. ←3, 20, 24, 25, 60

[13] Gábor Braun, Cristóbal Guzmán, and Sebastian Pokutta, *Lower bounds on the oracle complexity of nonsmooth convex optimization via information theory*, IEEE Transactions on Information Theory **63** (2017), no. 7. ←74

[14] P. Brucker, *An* $O(n)$ *algorithm for quadratic knapsack problems*, Operations Research Letters **3** (1984), no. 3, 163–166. ←33, 46, 54

[15] Sébastien Bubeck and Nicoló Cesa-Bianchi, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Foundations and Trends in Machine Learning **5** (2012), no. 1, 1–122. ←60

[16] N. Cesa-Bianchi, A. Conconi, and C. Gentile, *On the generalization ability of on-line learning algorithms*, IEEE Transactions on Information Theory **50** (2004September), no. 9, 2050–2057. ←43

[17] Thomas M. Cover and Joy A. Thomas, *Elements of information theory, second edition*, Wiley, 2006. ←74, 78

[18] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, Advances in neural information processing systems 27, 2014. ←43

[19] David L. Donoho, Richard C. Liu, and Brenda MacGibbon, *Minimax risk over hyperrectangles, and implications*, Annals of Statistics **18** (1990), no. 3, 1416–1437. ←70

[20] D.L. Donoho, *Compressed sensing*, Technical report, stanford university, 2006. ←31

[21] John C. Duchi, *Stats311/EE377: Information theory and statistics*, 2015. ←74

[22] John C. Duchi, Elad Hazan, and Yoram Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, Journal of Machine Learning Research **12** (2011), 2121–2159. ←56, 60

[23] John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra, *Efficient projections onto the ℓ_1-ball for learning in high dimensions*, Proceedings of the 25th international conference on machine learning, 2008. ←33, 46, 54

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning*, Second, Springer, 2009. ←81

[25] Elad Hazan, *The convex optimization approach to regret minimization*, Optimization for machine learning, 2012. ←43

[26] Elad Hazan, *Introduction to online convex optimization*, Foundations and Trends in Optimization **2** (2016), no. 3–4, 157–325. ←4

[27] J. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I*, Springer, New York, 1993. ←3, 21, 24, 74

[28] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II*, Springer, New York, 1993. ←3, 24

[29] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal, *Fundamentals of convex analysis*, Springer, 2001. ←24

[30] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of the American Statistical Association **58** (March 1963), no. 301, 13–30. ←76

[31] I. A. Ibragimov and R. Z. Has'minskii, *Statistical estimation: Asymptotic theory*, Springer-Verlag, 1981. ←4, 62, 74

[32] Rie Johnson and Tong Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in neural information processing systems 26, 2013. ←43

[33] Lucien Le Cam, *Asymptotic methods in statistical decision theory*, Springer-Verlag, 1986. ←4, 62, 65

[34] Erich L. Lehmann and George Casella, *Theory of point estimation, second edition*, Springer, 1998. ←62

[35] Friedrich Liese and Igor Vajda, *On divergences and informations in statistics and information theory*, IEEE Transactions on Information Theory **52** (2006), no. 10, 4394–4412. ←78

[36] David Luenberger, *Optimization by vector space methods*, Wiley, 1969. ←24

[37] Jerrold Marsden and Michael Hoffman, *Elementary classical analysis, second edition*, W.H. Freeman, 1993. ←3

[38] Brendan McMahan and Matthew Streeter, *Adaptive bound optimization for online convex optimization*, Proceedings of the twenty third annual conference on computational learning theory, 2010. ←60

[39] Angelia Nedić, *Subgradient methods for convex minimization*, Ph.D. Thesis, 2002. ←60

[40] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization **19** (2009), no. 4, 1574–1609. ←43, 60

[41] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, Wiley, 1983. ←4, 25, 59, 74

[42] Arkadi Nemirovski, *Efficient methods in convex programming*, 1994. Technion: The Israel Institute of Technology. ←74

[43] Arkadi Nemirovski, *Lectures on modern convex optimization*, 2005. Georgia Institute of Technology. ←43

[44] Y. Nesterov, *Introductory lectures on convex optimization*, Kluwer Academic Publishers, 2004. ←26, 43, 74

[45] Y. Nesterov and A. Nemirovski, *Interior-point polynomial algorithms in convex programming*, SIAM Studies in Applied Mathematics, 1994. ←25

[46] Jorge Nocedal and Stephen J. Wright, *Numerical optimization*, Springer, 2006. ←3, 60

[47] B. T. Polyak, *Introduction to optimization*, Optimization Software, Inc., 1987. ←3, 43

[48] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization **30** (1992), no. 4, 838–855. ←43

[49] R. Tyrell Rockafellar, *Convex analysis*, Princeton University Press, 1970. ←3, 6, 24

[50] Walter Rudin, *Principles of mathematical analysis, third edition*, McGraw-Hill, 1976. ←3

[51] S. Shalev-Shwartz, *Online learning: Theory, algorithms, and applications*, Ph.D. Thesis, 2007. ←47

[52] Shai Shalev-Shwartz, *Online learning and online convex optimization*, Foundations and Trends in Machine Learning **4** (2012), no. 2, 107–194. ←4

[53] Shai Shalev-Shwartz and Tong Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, Journal of Machine Learning Research **14** (2013), 567–599. ←43

[54] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński, *Lectures on stochastic programming: Modeling and theory*, SIAM and Mathematical Programming Society, 2009. ←4

[55] Naum Zuselevich Shor, *Minimization methods for nondifferentiable functions*, translated by Krzystof Kiwiel and Andrzej Ruszczyński, Springer-Verlag, 1985. ←60

[56] Naum Zuselevich Shor, *Nondifferentiable optimization and polynomial problems*, Springer, 1998. ←56, 60

[57] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc B. **58** (1996), no. 1, 267–288. ←31

[58] Alexandre B. Tsybakov, *Introduction to nonparametric estimation*, Springer, 2009. ←63, 74

[59] Abraham Wald, *Contributions to the theory of statistical estimation and testing hypotheses*, Annals of Mathematical Statistics **10** (1939), no. 4, 299–326. ←4, 74

[60] Abraham Wald, *Statistical decision functions which minimize the maximum risk*, Annals of Mathematics **46** (1945), no. 2, 265–280. ←4

[61] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Annals of Statistics **27** (1999), no. 5, 1564–1599. ←4, 63

[62] Bin Yu, *Assouad, Fano, and Le Cam*, Festschrift for lucien le cam, 1997, pp. 423–435. ←4, 63, 74

[63] Martin Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, Proceedings of the twentieth international conference on machine learning, 2003. ←43

Stanford University, Stanford CA 94305
E-mail address: jduchi@stanford.edu