Advanced Course in Optimization and Machine Learning

July 22, 2016

Contents

Course Overview

While the introductory course (67731) on *convex optimization* focuses on formulating and recognizing convex optimization problems, the emphasis of our course is more algorithmic. We revisit some of the most fundamental problems in optimization and introduce *efficient methods* for solving these problems. Starting from core techniques, we proceed to cover some of the recent major achievements in the area of mathematical programming. In particular, we discuss the emerging interplay between optimization and *machine learning*.

0.1 Case Study: Ridge Regression

As an example, consider the optimization problem associated with the commonly used method of *ridge regression*. We are given a sequence of *vector instances* $x_1, \ldots, x_n \in \mathbb{R}^d$ and corresponding *labels* $y_1, \ldots, y_n \in \mathbb{R}$. For a *regularization* parameter $\lambda > 0$, the objective is given by

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2}(w^\top x_i - y_i)^2 + \frac{\lambda}{2}\|w\|^2 \ . \tag{1}$$

We also assume[1] that $\|x\|_i \leqslant 1$ for all i. We next observe that exactly minimizing the objective is equivalent to solving a linear system. Precisely, denoting $C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$ and $b = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$, an equivalent objective is given by

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} w^\top (C + \lambda I) w - w^\top b \ .$$

Since $C + \lambda I$ is positive definite, the objective is convex. By first-order conditions, minimizing the problem is equivalent to solving the system $Cw = b$. This can be done by standard methods in numerical linear algebra like Gaussian Elimination. Forming the matrix C takes $O(nd^2)$ and computing the inverse takes $O(d^\omega)$, where $\omega < 2.373$ is the current value of the matrix multiplication constant.

When n and d are huge, it is often not practical to invert or even to store the matrix C. Instead of computing an exact minimizer, we now seek for an efficient

[1] As we shall see later, this assumption is w.l.o.g.

iterative method that converges to an optimal solution as fast as possible. We begin with the well known Gradient Descent algorithm. Noting that the gradient at any point w is $Cw + \lambda w - b$, we can apply the Gradient Descent (GD) algorithm. Note that computing Cw does not require neither the computation nor the storage of C. Instead, we iteratively process all the x_i's and sum the terms $x_i(x_i^\top w)$. Thus, the runtime per iteration[2] is $O(nd)$.

The next natural question is how many iterations are needed in order to obtain a good approximation. For a desired accuracy $\epsilon > 0$, we will establish the upper bound $O(\lambda^{-1} \log(1/\epsilon))$ on the number of iterations until reaching ϵ-approximation. We refer to the quantity λ^{-1} as the condition number of the problem. We remark that the dependence on this quantity can be improved using either the Conjugate Gradient (CG) method or the Accelerated Gradient Descent (AGD) method of Nesterov.

In the last decade, stochastic methods have become indispensable tools in optimization and in particular in machine learning optimization problems. More concretely, stochastic first-order methods such as Stochastic Gradient Descent (SGD) use only a random subsequence (a.k.a. *mini-batch*) of $(x_1, y_1), \ldots, (x_n, y_n)$ in order to form an unbiased estimate of the gradient on each round. While this modification reduces the complexity per iteration, it also introduces an undesired noise and consequently, the convergence rate deteriorates. For example, the convergence rate of the SGD, which uses only a single random example for each estimate, scales at least linearly with $1/\epsilon$ (rather than logarithmically).

Recently, several fast stochastic methods have been developed to remedy this situation. Roughly speaking, the suggested methods employ sophisticated techniques in order to reduce the variance induced by the estimation process. We will study the SDCA algorithm ([Shalev-Shwartz and Zhang, 2013]) and show that its runtime per iteration is $O(d)$ and its convergence rate is $O((n + \lambda^{-1}) \log(1/\epsilon))$. That is, it is faster than GD by factor $\min\{n, \lambda^{-1}\}$. As with GD, we can improve the dependence on λ^{-1} by using acceleration techniques.

Last, consider the regime where d is moderate but n and λ^{-1} are extremely large. We will (hopefully) study recently developed random *linear sketching* methods. Roughly speaking, given a matrix A, a linear sketch of A is a smaller random matrix AS. Based on Johnson-Lindenstrauss type results, a *sketch-and-solve* approach can be used to significantly reduce the overall computation. For our case we will see an algorithm due to [Clarkson and Woodruff, 2013] whose runtime is[3] $\tilde{O}(d^3 + nd)$. We summarize the comparison in Figure 1.

[2]Note that the complexity per iteration of GD can be bounded by $N + d$, where N is the number of nonzeros in the sequence $x_1, \ldots, x_n$. While a trivial bound on N is nd, many machine learning problems are very *sparse*, resulting in a much better upper bound on N. This is an important advantage of first-order methods over Gaussian elimination which has does not exploit data sparsity.

[3]Throughout these notes, the notation $\tilde{O}(\cdot)$ has nearly the same meaning as $O(\cdot)$; the only difference is that we hide polylogarithmic dependencies.

Method	overall runtime
Gaussian elimination	$nd^2 + d^\omega$
GD	$dn\gamma^{-1}$
SDCA	$d(n + \gamma^{-1})$
Sketch-and-solve	$d^3 + nd$

Figure 1: Comparison between numerical methods for approximately minimizing linear regression with respect the square loss (equivalently, approximately solving a linear system). We compare overall runtimes while ignoring logarithmic terms.

Course organization: In the first part of the course we study convex analysis while closely following the first three chapters of [Borwein and Lewis, 2010]. We then proceed to cover fundamental deterministic methods for unconstrained (convex) optimization such as Gradient Descent and Newton's method. Borrowing ideas from the first two parts of the course, we next study efficient *interior point* and *cutting-plane* methods for constrained optimization. Finally, we will study advanced stochastic methods such as SDCA and linear sketching.

Part I

Convex Analysis

Chapter 1

Background

1.1 Euclidean Spaces

We start by reviewing some basic properties of Euclidean spaces. We denote by $\mathbf{E}$ an arbitrary Euclidean space, that is, $\mathbf{E}$ is a finite-dimensional real vector space, equipped with an inner product[1] $\langle \cdot, \cdot \rangle$. The inner product induces a *norm* by $\|x\| = \sqrt{\langle x, x \rangle}$. The *unit ball* is defined by $B = \{x : \|x\| \leqslant 1\}$. For two sets $C, D \subseteq \mathbf{E}$ and a subset $\Lambda \subseteq \mathbb{R}$, we define

$$C + D = \{x + y : x \in C, y \in D\} \quad , \quad \Lambda C = \{\lambda x : \lambda \in \Lambda, x \in C\} \ .$$

The set of nonnegative reals is denoted by $\mathbb{R}_+$. A set $C \subseteq \mathbf{E}$ is called a *cone* if $\mathbb{R}_+ C = C$. An important example of a cone is the nonnegative orthant $\mathbb{R}^n_+ := \{x \in \mathbb{R}^n : \text{each } x_i \geqslant 0\}$ is a cone.

Once we have the notion of norm, a topology is naturally defined. The interior of a set $D \subseteq \mathbf{E}$, denoted $\operatorname{int} D$, consists of all the points x for which there exists $r > 0$ such that $x + rB \subseteq D$. The set D is said to be *open* if $D = \operatorname{int} D$. We say that $\bar{x} \in E$ is the limit point of a sequence $(x_n) \subseteq \mathbf{E}$ if $\lim_{n\to\infty} \|x_n - \bar{x}\| = 0$. The closure of D, denoted $\operatorname{cl} D$, consists of all the limit points of D. A set $D \subseteq \mathbf{E}$ is *closed* if it contains all of its limit points, i.e., if $D = \operatorname{cl} D$. The *boundary* of a set $D \subseteq \mathbf{E}$ is defined by $\operatorname{bd} D = \operatorname{cl} D \backslash \operatorname{int} D$. As an example, the interior of the nonnegative orthant, $\mathbb{R}^n_+$, is the positive orthant, $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : \text{each } x_i > 0\}$. Conversely, $\operatorname{cl} \mathbb{R}^n_{++} = \mathbb{R}^n_+$. Hence, the nonnegative orthant is closed (but not open) whereas the positive orthant is open (but not closed). An exercise shows that a set D is closed if and only if its complement $\mathbf{E} \backslash D$ is open. A set D is *bounded* if there exists $r > 0$ such that $D \subseteq rB$. Last, a set $D \subseteq \mathbf{E}$ is *compact* if it is closed and bounded.

It can be seen that $x_n \to \bar{x}$ if and only if for every index $i \in \dim(E)$, the i-th coordinate of x_n converges to the i-th coordinate of $\bar{x}$. This fact leads to important generalization of results from univariate calculus.

[1] Unless stated otherwise, it is always assumed that vectors in $\mathbf{E}$ are expressed according to some (arbitrary) orthonormal basis (e.g., in $\mathbb{R}^n$ we simply pick the standard basis).

Theorem 1.1.1 ***(Bolzano-Weierstrass)*** *Bounded sequences in* $\mathbf{E}$ *have convergent subsequences.*

A real-valued function f defined over a subset $D \subseteq \mathbf{E}$ is said to be continuous at $\bar{x} \in D$ if for every sequence x_n that converges to x, $f(x_i) \to f(\bar{x})$. It can be verified that for any $\alpha \in \mathbb{R}$, the level set $\{x \in D : f(x) \leqslant \alpha\}$ is closed providing that D is closed. The following important result provides sufficient conditions for existence of an optimal solution for a minimization problem.

Theorem 1.1.2 ***(Weierstrass)*** *Suppose that the set $D \subseteq E$ is nonempty and closed, and that the level sets of the continuous function $f : D \to \mathbb{R}$ are bounded. Then, f has a global minimizer in D, i.e., there exists $\bar{x} \in D$ such that $f(\bar{x}) \leqslant f(x)$ for all $x \in D$.*

Proof Let $v = \inf\{f(x) : x \in D\}$. For an arbitrary $z \in D$, consider the set $C = \{x \in D : f(x) \leqslant f(z)\}$. By assumption, C is bounded. Since f is continuous and D is closed, C is also closed (see Exercise (1.1.2)). We now restrict our attention to the compact set C and show that f attains its minimum on C. By definition of v, there exists a sequence (x_n) in C which satisfies $f(x_n) \to v$. According to Theorem 1.1.1, we can pick a convergent subsequence (x_{n_k}). Since C is closed, the limit of this subsequence, $\bar{x}$, also belongs to C. The continuity of f implies that $f(\bar{x}) = \lim_{k\to\infty} f(x_{n_k}) = v$ (in particular, we conclude that $v \neq -\infty$). ■

In particular, a continuous function defined over a compact set has a global minimizer (and also a global maximizer).

Convexity

Recall that a set C is *convex* if for every two points $x, y \in C$ and $\alpha \in [0, 1]$, the point $\alpha x + (1 - \alpha)y$ belongs to C. Let $D \subseteq \mathbf{E}$. For any sequence $x_1, \ldots, x_m \in D$ and $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m_+$ such that $\sum_{i=1}^m \alpha_i = 1$, the vector $\sum_{i=1}^m \alpha_i x_i$ is called a convex combination of $x_1, \ldots, x_m$. It follows by induction on m that a convex set consists exactly of all the convex combinations of its points. Arbitrary intersection of convex sets is convex. The convex hull of a subset $D \in \mathbf{E}$, denoted $\operatorname{conv} D$, is the intersection of all the convex sets that contain D, thus, is is the smallest convex set that contains D. An exercise shows that $\operatorname{conv} D$ consists exactly of all the convex combinations of points from D. The following elementary property of convex hulls is proven useful in many applications (see Exercise (1.1.12)).

Lemma 1.1.1 *If a linear function f defined over $\operatorname{conv} D$ has a global minimizer (similarly, maximizer) in $\operatorname{conv} D$, then it also has a global minimizer (maximizer) in D.*

Proof Let $\bar{x} = \sum_{i=1}^{m} \alpha_i x_i$ be a convex combination of points from D that attains the minimum. Let $x_j \in \arg\min\{f(x_i) : i \in [m]\}$. Then, by the linearity of f

$$f(\bar{x}) = \sum_{i=1}^{m} \alpha_i f(x_i) \geqslant \sum_{i=1}^{m} \alpha_i f(x_j) = f(x_j) \ .$$

The proof for the case that $\bar{x}$ attains the maximum of f is analogous. ■

Given a convex set $C \subseteq \mathbf{E}$, we say that a function $f : C \to \mathbb{R}$ is *convex* if for all $x, y \in C$ and $\alpha \in [0,1]$, $f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha)f(y)$. The function f is *strictly convex* if the above inequality is strict whenever x and y are distinct and $\alpha \in (0,1)$. Examples of convex functions include affine functions and norms. For a fixed vector $\bar{x} \in \mathbf{E}$, the function $x \mapsto \frac{1}{2}\|x - \bar{x}\|^2$ defined over $\mathbf{E}$ is strictly convex. The function f is said to be *concave* if $-f$ is convex.

The following basic properties of convex functions are reviewed in the exercises. It is often eaiser to verify the convexity of a function by restricting it to a line. Namely, a function $f : C \to \mathbb{R}$ is (strictly) convex if and only if for every $x, y \in C$ $(x \neq y)$, the function $\phi : [0,1] \to \mathbb{R}$ defined by $\phi(t) = f(x + t(y - x))$ is (strictly) convex. The epigraph of the function $f : C \to \mathbb{R}$ is the set $\{(x, s) \in C \times \mathbb{R} : f(x) \leqslant s\}$. It is easy to verify that f is convex if and only if its epigraph is convex. Finally, It can be seen that if a strictly convex function attains its minimum, then it is unique.

Exercises

Exercise 1.1.1 *Let $(\mathbf{E}, \langle \cdot, \cdot \rangle)$ be an inner product space.*

1. *Prove the Cauchy-Schwarz inequality: for every $x, y \in \mathbf{E}$, $|\langle x, y \rangle| \leqslant \sqrt{\langle x, x \rangle}\sqrt{\langle y, y \rangle}$. An equality holds if and only if x and y are linearly dependent.*

2. *Show that the function $\|x\| = \sqrt{\langle x, x \rangle}$ defined over E is indeed a norm. (Hint: The only nontrivial property is the triangle inequality. Given $x, y \in \mathbf{E}$, use the Cauchy-Schwarz inequality to bound $\langle x + y, x + y \rangle$ from above.)*

3. *For a fixed $y \in \mathbf{E}$, show that the function $x \mapsto \langle x, y \rangle$ is continuous.*

Exercise 1.1.2 *Show that any level set of the form $\{x \in D : f(x) \leqslant \alpha\}$ of a continuous function $f : D \to \mathbb{R}$ is closed providing that $D \subseteq \mathbf{E}$ is closed.*

Exercise 1.1.3 *Compute the interiors and the boundaries of the following subsets of $\mathbf{E}$. Deduce which of the sets are closed or open.*

1. $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : \text{each } x_i > 0\}$

2. $\mathbb{R}^n_{+} = \{x \in \mathbb{R}^n : \text{each } x_i \geqslant 0\}$

3. $\mathbb{S}^n_{++} = \{A \in \mathbb{R}^{d \times d} : A \text{ is symmetric and positive definite}\}$

4. $\mathbb{S}^n_+ = \{A \in \mathbb{R}^{d\times d} : A \text{ is symmetric and positive semidefinite}\}$

5. ($\star$) *A linear subspace $L \subseteq \mathbf{E}$. Distinguish between the cases* $\dim(L) = \dim(\mathbf{E})$ *and* $\dim(L) < \dim(\mathbf{E})$. *(Hint: Consider the latter case and let $v_1, \ldots, v_m$ be an orthogonal basis for L. Consider also a completion of $v_1, \ldots, v_m$ to an orthogonal basis, $v_1, \ldots, v_n$ of $\mathbf{E}$. Note that a vector $\bar{x}$ belongs to L if and only if for every $i > m$, $\langle \bar{x}, x_i \rangle = 0$. Use the continuity of the inner product (Exercise (1.1.1)).)*

6. *For a point w and a scalar b, consider the halfspaces $\{x : \langle w, x\rangle \leqslant b\}, \{x : \langle w, x\rangle < b\}$. Also, consider the hyperplane $\{x : \langle w, x\rangle = b\}$.*

Exercise 1.1.4

1. *Show that a set D is closed if and only if its complement $\mathbf{E}\backslash D$ is open.*

2. ($\star$) *Show that the only two sets in $\mathbf{E}$ which are both closed and open are $\varnothing$ and $\mathbf{E}$.*

Exercise 1.1.5 *Show that an arbitrary union of open sets is open. Similarly, an arbitrary intersection of closed sets is closed. Deduce that* $\operatorname{int}(D)$ *is equal to the union of all the open sets that are contained in D (thus, it is the largest open set contained in D) and* $\operatorname{cl}(D)$ *is equal to the intersection of all the closed sets that contain D (thus, it is the smallest closed set that contains D).*

Exercise 1.1.6 *Show that $x_n \to \bar{x}$ in $\mathbf{E}$ if and only if for every index $i \in \dim(E)$, the i-th coordinate of x_n converges to the i-th coordinate of $\bar{x}$.*

Exercise 1.1.7 *Prove Theorem 1.1.1.*

Exercise 1.1.8 *Suppose that the set $D \subseteq E$ is nonempty and closed and let $f : D \to \mathbb{R}$ be a continuous function. Assume that for every α, the set $\{x : f(x) \geqslant \alpha\}$ is bounded. Show that f has a global maximizer in D. Deduce that a continuous function f defined over a compact set D has both a global minimizer and a global maximizer.*

Exercise 1.1.9 *We recall some basic operations that preserve convexity of sets. Prove the following:*

1. *Arbitrary intersection of convex sets is convex.*

2. *If $C, D \subseteq \mathbf{E}$ are convex, then $C + D$ is convex. In particular, convexity is invariant to translation.*

Exercise 1.1.10 *Show that for any $D \subseteq \mathbf{E}$, $conv(D)$ consists exactly of all the convex combinations of elements of D.*

Exercise 1.1.11 *Prove that the closure of a convex set $C \subseteq \mathbf{E}$ is convex. Conclude that for any set $D \subseteq \mathbf{E}$, $\mathrm{cl}(\mathit{conv}\, D)$ is the smallest closed convex set containing D.*

TODO: affine sets, relative interior, the interior of a convex set is convex

Exercise 1.1.12 ***(Zero-sum games)*** *A two-player* zero-sum game *is defined as follows. Let $A \in \mathbb{R}^{n\times m}$ be a matrix. The game consists of one round in which the row player decides on a row and the column player decides on a column. From reasons that will become clear shortly, we associate the i-th row with the standard basis vector $e_i \in \{e_1, \ldots, e_n\}$, and the j-th column with $e_j \in \{e_1, \ldots, e_m\}$. Given a pair of decisions (a.k.a. strategies), (e_i, e_j), the payoff of the row player and the loss of the column player are equal to $A_{i,j}$. Naturally, while the row player wishes to maximize its payoff, the column player would like to minimize its loss. Therefore, we consider two important quantities:*[2]

$$Mm = \max_{e_i : i\in[n]} \min_{e_j : j\in[m]} A_{i,j} \quad , \quad mM = \min_{e_j : j\in[m]} \max_{e_i : i\in[n]} A_{i,j} \ .$$

The left quantity corresponds to the scenario where the column player can choose its action based on the decision of the row player. The right quantity corresponds to the opposite scenario. A pair of strategies (i, j) is called an equilibrium *if given that the strategy of the column player is e_j, the strategy e_i of the row player is optimal (i.e., $A_{q,j} \leqslant A_{i,j}$ for all $q \in [n]$), and vice versa ($A_{i,s} \geqslant A_{i,j}$ for all $s \in [m]$).*

1. *Prove the minimax inequality: $Mm \leqslant mM$. Describe a game in which the inequality is strict.*

2. *Show that there exists an equilibrium if and only if $Mm = mM$. Furthermore, a pair of strategies (e_i, e_j) forms an equilibrium iff $Mm = A_{i,j} = mM$.*

3. *Suppose that we modify the game by allowing each player to select a probability distribution over its set of* pure *strategies, i.e., the row player is allowed to select a probability distribution $p \in \Delta_n = \{p' \in \mathbb{R}^n : \text{each } p'_i \geqslant 0,\ \sum p'_i = 1\}$ and the column player is allowed to select a vector $q \in \Delta_m$. Given a pair of strategies (p, q), the payoff of the row player (analogously, the negative loss of the column player) is now defined as the expected payoff:*

$$\sum_{i=1}^{n}\sum_{i=1}^{m} p_i q_j A_{i,j} = p^\top A q \ .$$

The quantities Mm and mM are now defined by

$$Mm = \max_{p\in\Delta_n} \min_{q\in\Delta_m} p^\top A q \quad , \quad mM = \min_{q\in\Delta_m} \max_{p\in\Delta_n} p^\top A q \ .$$

[2] For preciseness, the expression $\max_{e_i : i\in[n]} \min_{e_j : j\in[m]} A_{i,j}$ is defined as follows: Consider the function $\phi : [n] \to \mathbb{R}$ defined by $\phi(i) = \max_{j\in[m]} A_{i,j}$. Then $\max_{e_i : i\in[n]} \min_{e_j : j\in[m]} A_{i,j} = \max_{i \in [n]} \phi(i)$.

The definition equilibrium is generalized analogously. The minimax inequality certainly holds. The minimax theorem due to von Neumann says that these quantities are equal, and therefore, there exists an equilibrium (we will hopefully see a non-standard proof of this result).

(a) Show that $\Delta_n = conv\{e_1, \ldots, e_n\}$.

(b) Show that there exists a pure strategy e_i *for the row player that attains the value* mM. *That is, for all* $q \in \Delta_m$, *there exists* $i \in [n]$ *such that*

$$e_i^\top Aq = \max_{p \in \Delta_n} p^\top Aq \ .$$

Similarly, there exists a pure strategy e_j *for the column player that attains the value* Mm.

Exercise 1.1.13 *Let* $C \subseteq \mathbf{E}$ *be a convex set. Show that a function* $f : C \to \mathbb{R}$ *is (strictly) convex if and only if for every* $x, y \in C$, *the function* $\phi_{x,y} : [0,1] \to \mathbb{R}$ *defined by* $\phi_{x,y}(t) = f(x + t(y - x))$ *is (strictly) convex.*

Exercise 1.1.14 *Prove the following:*

1. *For a fixed vector* $\bar{x} \in \mathbf{E}$, *the function* $x \mapsto \frac{1}{2}\|x - \bar{x}\|^2$ *defined over* $\mathbf{E}$ *is strictly convex.*

2. *Let* $C \subseteq \mathbf{E}$ *be a convex set and* $f : C \to \mathbb{R}$ *be a strictly convex function. If* f *attains its minimum over* C, *then it is unique.*

Exercise 1.1.15 *We recall some basic operations that preserve convexity of functions. Prove the following:*

1. *Given a convex set* $C \subseteq \mathbf{E}$ *and convex functions* $(f_i)_{i \in I}$ *defined over* C, *show that* $f = \sup f_i$ *is convex.*

2. *Let* g *be a linear function from* $\mathbf{E}$ *to another Euclidean space,* $\mathbf{Y}$, $f : \mathbf{Y} \to \mathbb{R}$ *be a convex function and let* $b \in \mathbf{Y}$. *The function* $x \mapsto f(g(x) + b)$ *defined over* $\mathbf{E}$ *is convex.*

Exercise 1.1.16 *Let* $C \subseteq \mathbf{E}$ *and* $f : C \to \mathbb{R}$. *Show that* f *is convex if and only if its epigraph is convex.*

1.2 Symmetric Matrices

Linear algebra plays a significant role in our studies. In this short section we consider the vector space $\mathbb{S}^d$ of symmetric $d \times d$ matrices. We make this vector space into a Euclidean space by defining the Frobenius inner product. Note that for any two $n \times n$ matrices X and Y, $\mathrm{tr}(X^\top Y) = \sum_{i=1} \sum_{j=1}^d X_{i,j} Y_{i,j}$. It is easily seen that the bilinear form $\langle \cdot, \cdot \rangle$ defined by $\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$ forms an inner product. It induces the Frobenius norm, $\|X\|_F = \sqrt{\langle X, X \rangle} = \sqrt{\sum_{i=1}^d \sum_{j=1}^d X_{i,j}^2}$. In the sequel, we always equip the space $\mathbb{S}^d$ with the Frobenius inner product and norm.

According to the spectral theorem, any matrix $X \in \mathbb{S}^d$ has d eigenvalues which we denote and order by $\lambda_1(X) \geqslant \ldots \geqslant \lambda_d(X)$. We also define the vector $\lambda(X) = (\lambda_1(X), \ldots, \lambda_d(X))$. Let $\mathbf{O}^d$ be the group of $d \times d$ *orthogonal* matrices. Then, any $X \in \mathbb{S}^d$ can be written in *eigenvalue decomposition* (EVD) form as

$$X = U(\mathrm{diag}\, \lambda(X))U^\top = \sum_{i=1}^d \lambda_i(X) u_i u_i^\top \ ,$$

where the matrix $U \in \mathbf{O}^d$, and the operator diag maps a d-dimensional vector x into a diagonal matrix X with $X_{i,i} = x_i$. The i-th column of U, denoted u_i, is the eigenvector of X corresponding to the eigenvalue $\lambda_i(X)$. An exercise shows that $\|X\|_F = \|\lambda(X)\|$.

An important subset of the space $\mathbb{S}^d$ is the convex cone $\mathbb{S}^d_+$ of symmetric *positive semidefinite* matrices, which consists of the matrices $X \in \mathbb{S}^d$ that satisfy $z^\top X z \geqslant 0$ for all $z \in \mathbb{R}^d$. We also consider the subset $\mathbb{S}^d_{++} \subseteq \mathbb{S}^n_+$ of *positive definite* matrices, for which the above inequality is strict for all $z \in \mathbb{R}^n$. It can be seen that $X \in \mathbb{S}^d_+$ (respectively, $X \in \mathbb{S}^d_{++}$) if and only if all the eigenvalues of X are nonnegative (positive). This two subsets induce an ordering relation over $\mathbb{S}^d$; For two matrices $X, Y \in \mathbb{S}^d$, we write $X \preceq Y$ (equivalently, $X - Y \preceq 0$) if $Y - X \in \mathbb{S}^d_+$. Analogously, we write $X \prec Y$ (equivalently, $X - Y \prec 0$) if $Y - X \in \mathbb{S}^d_{++}$.

Note that if $X, Y \in \mathbb{S}^n$, the Cauchy-Schwartz inequality implies that

$$\langle X, Y \rangle \leqslant \|X\|_F \|Y\|_F = \|\lambda(X)\| \|\lambda(Y)\|$$

We conclude this section by showing an important refinement of this result, while demonstrating an application of Lemma 1.1.1. We rely on the following theorem, whose proof is deferred to a later stage of the course. Recall that a square matrix is called *doubly stochastic* if its entries lie in the range $[0, 1]$ and the entries in every column and row sum up to 1. A square matrix is called a *permutation matrix* if its entries lie in $\{0, 1\}$ and each row and column contain exactly one entry of 1.

Theorem 1.2.1 ***(Birkhoff)*** *The convex hull of the set of $n \times n$ permutation matrices is the set of $n \times n$ doubly stochastic matrices.*

We also need the following simple lemma. For a vector $x \in \mathbb{R}^d$, we denote by $[x]$ the vector with the same components permuted into decreasing order.

Lemma 1.2.1 *For any two vectors $x, y \in \mathbb{R}^d$, $\langle x, y\rangle \leqslant \langle [x], [y]\rangle$.*

Theorem 1.2.2 ***(Fan)*** *For any two symmetric $d \times d$ matrices X and Y,*

$$\langle X, Y\rangle \leqslant \langle \lambda(X), \lambda(Y)\rangle$$

Moreover, equality holds if and only if X and Y have a simultaneous ordered spectral decomposition: there exists $U \in \mathbf{O}^d$ such that $X = U(\operatorname{diag}\lambda(X))U^\top$ and $Y = U(\operatorname{diag}\lambda(Y))U^\top$.

Proof Consider the spectral decompositions $X = \sum_{i=1}^d \lambda_i(X) u_i u_i^\top$ and $Y = \sum_{i=1}^d \lambda_i(Y) v_i v_i^\top$. Then,

$$\langle X, Y\rangle = \sum_{i=1}^d \sum_{j=1}^d \lambda_i(X)\lambda_j(Y)\operatorname{tr}(u_i u_i^\top v_j v_j^\top) = \sum_{i=1}^d \sum_{j=1}^d \lambda_i(X)\lambda_j(Y)(\langle u_i, v_j\rangle)^2 \ .$$

Denote $z_{i,j} = (\langle u_i, v_j\rangle)^2$. It can be verified that the $d \times d$ matrix, $Z = (z_{i,j})$ is doubly stochastic[3]. Hence,

$$\langle X, Y\rangle = \lambda(X)^\top Z \lambda(Y) \ .$$

Now consider the problem of maximizing $A \mapsto \lambda(X)^\top A \lambda(Y)$ over all doubly stochastic matrices. Combining Theorem 1.2.1 and Lemma 1.1.1, we know that it suffices to consider this maximization problem over all permutation matrices. The desired inequality follows from Lemma 1.2.1. The condition for equality is left as an exercise. ■

Additional notions such as *projection* matrices, *adjoint* and *singular value decomposition* (SVD) are reviewed in the exercises.

Exercises

Exercise 1.2.1 *Let $A : \mathbf{E} \to \mathbf{Y}$ be a linear transformation between two Euclidean spaces. Recall that the* null space *(a.k.a.* kernel*) of A, denoted by $N(A)$, consists of all the vectors x for which $Ax = 0$. The column space of A, denoted $C(A)$ is the set $\{y \in \mathbf{Y} : (\exists x \in \mathbf{E}) \ \ s.t. \ \ Ax = y\}$. The adjoint of A, denoted $A^\star$, is the unique linear transformation from $\mathbf{Y}$ to $\mathbf{E}$ that satisfies $\langle A^\star y, x\rangle = \langle y, Ax\rangle$ for all $x \in \mathbf{E}$ and $y \in \mathbf{Y}$. Fixing bases for $\mathbf{E}$ and $\mathbf{Y}$, the associated matrices which we also denote by A and $A^\star$, satisfy $A^\star = A^\top$.*

1. *Show that for any linear transformation A, the column space of A coincides with the orthogonal complement of the null space of $A^\star$.*

[3]For example, $\sum_{j=1}^n Z_{1,j} = \|V^\top u_1\|^2 = \|u_1\|^2 = 1$.

2. ($\star$) *Show that the column space of $AA^\star$ coincides with the column space of A. (Hint: Prove that the null space of $AA^\star$ is included in the null space of $A^\star$. For this purpose, note that if $v \neq 0$ belongs to the null space of $AA^\star$, then $0 = \langle v, AA^\star v\rangle = \langle A^\star v, A^\star v\rangle$.)*

Exercise 1.2.2 *Show that the bilinear form $\langle \cdot, \cdot \rangle$, defined over pairs of $n \times n$ matrices by $\langle X, Y\rangle = \mathrm{tr}(X^\top Y)$, forms an inner product.*

Exercise 1.2.3 *Show that $\mathbb{S}^d_+$ is indeed a convex cone.*

Exercise 1.2.4 *Let $X \in \mathbb{S}^d$. Show that $X \in \mathbb{S}^d_+$ (respectively, $X \in \mathbb{S}^d_{++}$) if and only if all the eigenvalues of X are nonnegative (positive).*

Exercise 1.2.5 *An (orthogonal)[4] linear projection is a linear transformation P from a space $\mathbf{E}$ to itself that satisfies $P^2 = P$ and its column space is orthogonal to its null space.*

1. *Show that P is a linear (orthogonal) projection if and only if the associated matrix can be written as $P = UU^\top$, where $U \in \mathbb{R}^{d\times k}$ has orthonormal columns ($k \leqslant n$ is the rank of P).*

2. *Conclude that a projection matrix is positive semidefinite and k of its eigenvalues are one, while the rest are zero.*

Exercise 1.2.6 *Recall the SVD theorem: Every matrix $X \in \mathbb{R}^{d\times n}$ can be written in* singular value decomposition *(SVD) form as $X = U\Sigma V^\top$, where $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{n\times n}$ have orthonormal columns, and $\Sigma \in \mathbb{R}^{d\times n}$ is a diagonal matrix. The columns of U and V, denoted $u_1, \ldots, u_d$ and $v_1, \ldots, v_n$ (respectively), are named left and right singular vectors, respectively. The $p = \min\{d, n\}$ diagonal entries of Σ, denoted $\sigma_1(X), \ldots, \sigma_p(X)$, are called singular values and we always assume that $\sigma_1(X) \geqslant \ldots \geqslant \sigma_p(X)$. We also define the vector $\sigma(X) = (\sigma_1(X), \ldots, \sigma_p(X))$. Note that X can also be written as $X = \sum_{i=1}^r \sigma_i(X) u_i v_i^\top$, where r is the rank of X (the rest of the singular values are 0).*

1. *Following the above notation, show that $XX^\top \in \mathbb{S}^d_+$ and both the SVD and the EVD of $XX^\top$ are given by $XX^\top = \sum_{i=1}^p \sigma_i(X)^2 u_i u_i^\top$.*

2. *Let $Y \in \mathbb{S}^d$. Show that if $Y = \sum_{i=1}^d \lambda_i(Y) u_i u_i^\top$ is the EVD of X, then its SVD is given by $Y = \sum_{i\in[d]} |\lambda_i(Y)| u_i v_i^\top$, where $v_i = \mathrm{sgn}(\lambda_i(X)) u_i$.*

Exercise 1.2.7

[4] A linear projection is only assumed to satisfy the property $P^2 = P$. We restrict our attention to orthogonal projections.

1. *Show that the trace is invariant under cyclic permutations: for any sequence of matrices $A_1, \ldots, A_m$ for which $\prod_{i=1}^m A_i$ is a square matrix,*

$$\operatorname{tr}(A_1 \cdot \ldots \cdot A_m) = \operatorname{tr}(A_2 \cdot A_3 \cdot \ldots \cdot A_m \cdot A_1) = \ldots = \operatorname{tr}(A_m \cdot A_1 \cdot \ldots \cdot A_{m-1}) \ .$$

2. *Show that the trace is similarity-invariant: for $A, P \in \mathbb{R}^{d\times d}$, $\operatorname{tr}(P^{-1}AP) = \operatorname{tr}(A)$ providing that P is invertible.*

3. *Show that the trace is unitary-invariant: if $A \in \mathbb{R}^{d\times d}$ and $U \in \mathbb{R}^{m\times d}$ has orthogonal columns ($m \geqslant d$), then*

$$\operatorname{tr}(UAU^\top) = \operatorname{tr}(A) \ .$$

4. *Show that the Frobenius norm is unitary-invariant: if $A \in \mathbb{R}^{d\times n}$ and $U \in \mathbb{R}^{m\times d}$ has orthogonal columns ($m \geqslant d$), then $\|UA\|_F = \|A\|_F$.*

5. *Conclude that for any matrix X, $\|X\|_F = \|\sigma(X)\|$, where σ is the vector consisting of the singular values of X (see Exercise (1.2.6)). Also, if $X \in \mathbb{S}^d$, then $\|X\|_F = \|\lambda(X)\|$.*

6. *Show that the Frobenius norm is submutiplicative: for any two matrices A, B for which AB is defined, $\|AB\| \leqslant \|A\|\|B\|$.*

Exercise 1.2.8 *Prove Lemma 1.2.1.*

Exercise 1.2.9

1. *Prove that following analog of Fan's inequality: for two matrices $X, Y \in \mathbb{S}^d$,*

$$\langle X, Y\rangle \geqslant \sum_{i=1}^{d} \lambda_i(X)\lambda_{d+1-i}(Y) \ .$$

2. *($\star$) Complete the proof of Fan's theorem: show that if the matrices $X, Y \in \mathbb{S}^d$ satisfy $\langle X, Y\rangle = \langle \lambda(X), \lambda(Y)\rangle$, then there exists $U \in \mathbf{O}^n$ such that $X = U(\operatorname{diag}\lambda(X))U^\top$ and $Y = U(\operatorname{diag}\lambda(Y))U^\top$. (Hint: Consider the matrix $X+Y$ and denote its spectral decomposition by $U(\lambda(X+Y))U^\top$. Let $\tilde{X} = U(\operatorname{diag}\lambda(X))U^\top$, $\tilde{Y} = U(\operatorname{diag}\lambda(Y))U^\top$. Prove the equality $\langle \tilde{X}, X+Y\rangle = \langle X, X+Y\rangle$ and use the Cauchy-Schwarz inequality (more precisely, use the condition for equality in Cauchy-Schwarz inequality) to conclude that $\tilde{X} = X$. Similarly, show that $\tilde{Y} = Y$.)*

Exercise 1.2.10 ***(PCA via Fan's Theorem)*** *Consider the PCA problem: we are given a sequence of vectors, $x_1, \dots, x_n \in \mathbb{R}^d$, and a parameter $k \in [d]$. We would like to find a projection of the vectors onto a k-dimensional subspace of $\mathbb{R}^d$ such that the sum of squared ℓ_2-distances between points and their projection is minimized. Following the notation from Exercise (1.2.5), we consider the minimization of*

$$P \mapsto \sum_{i=1}^{n} \|x_i - Px_i\|^2$$

over all $d \times d$ projection matrices of rank k. Let $A = \sum_{i=1}^n x_i x_i^\top$.

1. *Show that above minimization problem is equivalent to maximizing*

$$P \mapsto \langle P, A \rangle$$

   over all $d \times d$ projection matrices of rank k.

2. *Use Fan's inequality to deduce that if $u_1, \dots, u_k$ denote the leading eigenvectors of $A = \sum_{i=1}^n x_i x_i^\top$ (i.e. u_i corresponds to $\lambda_i(A)$), then $P = \sum_{i=1}^k u_i u_i^\top$ is an optimal solution for the above optimization problems.*

Exercise 1.2.11 ***(Courant–Fischer–Weyl min-max principle)*** *Let $A \in \mathbb{S}^d$ and denote its spectral decomposition by $A = \sum_{i=1}^d \lambda_i u_i u_i^\top$, where $\lambda_1 \geqslant \dots \geqslant \lambda_d$. Prove the Courant–Fischer–Weyl min-max principle: for every $i \in [d]$, the following identities hold:*

$$\lambda_i(A) = \max_{\substack{x \neq 0: \\ x \perp \{u_1, \dots, u_{i-1}\}}} \frac{x^\top A x}{x^\top x} \,,$$

and u_i is a corresponding $\arg\max$. *Similarly,*

$$\lambda_i(A) = \min_{\substack{x \neq 0: \\ x \perp \{u_d, \dots, u_{d-i+1}\}}} \frac{x^\top A x}{x^\top x} \,,$$

and u_i is a corresponding $\arg\min$.

Chapter 2

Inequality Constraints

2.1 Optimality Conditions and Basic Separation

The significance of derivatives in finding optimal solutions of optimization problems is already known to us from a basic calculus course. In this section we investigate further analytical notions while emphasizing the interplay between derivatives and convexity in characterizing optimal solutions. This will lead us to a central theme in convex analysis — separation theorems.

Consider the problem of minimizing a function $f : C \to \mathbb{R}$, where C is a subset of $\mathbf{E}$. A point $\bar{x}$ is called a local minimizer of f on C if there exists a ball around $\bar{x}$ such that $f(\bar{x}) \leqslant f(x)$ for all $x \in D$ that lies in this ball. The *directional derivative* of f at $\bar{x}$ in a direction $d \in \mathbf{E}$ is

$$f'(\bar{x}; d) = \lim_{t \downarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} \ , \tag{2.1}$$

when this limit exists. When the map $d \mapsto f'(\bar{x}; d)$ is linear, i.e., there exists $a \in \mathbf{E}$ such that $f'(\bar{x}; d) = \langle a, d \rangle$, then we say that f is (Gâteaux) differentiable with (Gâteaux) derivative $\nabla f(\bar{x}) = a$. When f is differentiable[1] at every point in C, we say that f is differentiable on C.

Given a convex set $C \subseteq \mathbf{E}$ and a point $\bar{x} \in C$, the *normal cone* to C at $\bar{x}$, denoted $N_C(\bar{x})$, is the set of all vectors $d \in \mathbf{E}$ that satisfy $\langle d, x - \bar{x} \rangle \leqslant 0$ for all $x \in C$. It is clear that if $\bar{x} \in \operatorname{int}(C)$, then $N_C(\bar{x})$ consists exactly of the vector 0. The normal cone gives us an elegant necessary condition for optimality.

Theorem 2.1.1 ***(First-order necessary condition)*** *Suppose that $C \subseteq \mathbf{E}$ is convex and $\bar{x}$ is a local minimizer of the function $f : C \to \mathbb{R}$. Then for any point $x \in C$, the directional derivative $f'(\bar{x}; x - \bar{x})$, if it exists, satisfies $f'(\bar{x}; x - \bar{x}) \geqslant 0$. In particular, if f is differentiable at $\bar{x}$ then $-\nabla f(\bar{x}) \in N_C(\bar{x})$. If, in addition $\bar{x} \in \operatorname{int}(C)$ then $\nabla f(\bar{x}) = 0$.*

[1] While Gâteaux differentiability is slightly weaker than the usual (a.k.a. Fréchet) differentiability, it suffices for our needs. The difference between these notions is explained in Exercise (2.1.1).

Proof Assume that some point $x \in C$ satisfies $f'(\bar{x}; x - \bar{x}) < 0$. Then, for small enough $t \in (0, 1)$, we have[2] $f(\bar{x} + t(x - \bar{x})) - f(\bar{x})/t < 0$, contradicting the local minimality of $\bar{x}$. If f is differentiable, then for all $x \in C$, $\langle -\nabla f(\bar{x}), x - \bar{x} \rangle = -f'(\bar{x}; x - \bar{x}) \leqslant 0$. Thus, $-\nabla f(x) \in N_C(\bar{x})$. If, in addition $\bar{x}$ belongs to the interior of C, then $N_C(\bar{x}) = \{0\}$ so $\nabla f(x)$ must be zero. ■

When convexity of the function is assumed, we can derive analogous sufficient conditions for *global* optimality. An exercise shows that for any $\bar{x}, x$ in a convex set $C \subseteq \mathbf{E}$, the function $t \in (0, 1] \to \frac{f(\bar{x}+t(x-\bar{x}))-f(\bar{x})}{t}$ is nondecreasing providing that f is convex. In particular, $f(x) - f(\bar{x}) \geqslant f'(\bar{x}; x - \bar{x})$. This implies the following result.

Theorem 2.1.2 ***(First-order sufficient condition)*** *Suppose that f is a convex function defined over the convex set $C \subseteq \mathbf{E}$. Then for any points $\bar{x}, x \in C$, the directional derivative $f'(\bar{x}; x - \bar{x})$ exists in $[-\infty, \infty)$. If the condition $f'(\bar{x}; x - \bar{x}) \geqslant 0$ holds for all $x \in C$, or in particular if $-\nabla f(\bar{x}) \in N_C(\bar{x})$, then $\bar{x}$ is a global minimizer of f on C.*

In particular, critical points of a convex functions are global minimizers. Notably, a milder "local" notion of convexity is often sufficient for optimality. The following result is typical.

Theorem 2.1.3 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice continuously differentiable[3] function. If $\bar{x}$ is a critical point and $\nabla^2 f(\bar{x})$ is positive definite, then $\bar{x}$ is a local minimizer. Conversely, if f is twice continuously differentiable and $\bar{x}$ is a local minimizer, then $\nabla^2 f(\bar{x})$ is positive definite.*

We now establish several important consequences of Theorem 2.1.1 and Theorem 2.1.2.

Corollary 2.1.1 ***(First-order conditions for linear constraints)*** *Let $C \subseteq \mathbf{E}$ be a convex set, $f : C \to \mathbb{R}$, a linear map $A : \mathbf{E} \to \mathbf{Y}$, and a point $b \in Y$. Consider the optimization problem*

$$\inf\{f(x) : Ax = b\}$$

If $\bar{x}$ is a local minimizer f is differentiable at $\bar{x}$, then $\nabla f(\bar{x}) \in A^\star \mathbf{Y}$. Conversely, if f is convex and $\nabla f(x) \in A^\star \mathbf{Y}$, then $\bar{x}$ is a global minimizer.

As we shall see later, Corollary 2.1.1 is a direct consequence of the more general KKT conditions.

Lemma 2.1.1 ***(The projection lemma)*** *Let $C \subseteq \mathbf{E}$ be a closed and convex set. For every point $y \in \mathbf{E}$, the projection $P_C(y) = \arg\min_{x \in C} \|x - y\|$ is uniquely defined. Moreover, for any point $\bar{x} \in C$, $y - \bar{x} \in N_C(\bar{x})$ if and only if $\bar{x} = P_C(y)$.*

[2] Note that $\bar{x} + t(x - \bar{x})$ belongs to C since C is convex.

[3] We rely on the following fact from multivariate analysis: the Hessian of a twice continuously differentiable function is symmetric.

Proof We may assume that $y \notin C$. Define the function $f : E \to \mathbb{R}$ by $f(x) = \|x - y\|^2/2$. It is easily seen that the level sets of f are bounded. According to Weierstrass theorem, f admits a minimizer on C. Since f is strictly convex, the minimizer is unique. Next, we note that for any $\bar{x}, x \in C$, $f'(\bar{x}; x - \bar{x}) = \langle \bar{x} - y, x - \bar{x} \rangle$. If $\bar{x} = P_C(y)$, then for $d = y - \bar{x}$, by first order necessary conditions (2.1.1), we have

$$0 \geqslant \langle d, x - \bar{x} \rangle \Rightarrow d \in N_C(\bar{x}) \ .$$

Conversely, by first-order sufficient conditions (2.1.2), if $d = y - \bar{x} = f \in N_C(\bar{x})$, then $\bar{x}$ is a (unique) global minimizer. ■

As an example, consider projecting onto the unit ball, B. A quick picture reveals that for $y \notin B$, the projection is given by $y \mapsto y/\|y\|$. The projection lemma gives us a simple way to validate our informal proof. Indeed, denoting $\bar{x} = y/\|y\|$, for all $x \in B$ we have

$$\begin{aligned} \langle y - y/\|y\|, x - y/\|y\| \rangle &= \langle y - y/\|y\|, -y/\|y\| \rangle + \langle y - y/\|y\|, x \rangle \\ &= -(\|y\| - 1) + (\|y\| - 1)\langle y/\|y\|, x \rangle) \\ &\leqslant -(\|y\| - 1) + (\|y\| - 1) = 0 \ , \end{aligned}$$

where the inequality follows from Cauchy-Schwarz inequality together with the fact that $\|y\| - 1 \geqslant 0$ and both $y/\|y\|$ and x belong to B.

Finally, the basic separation demonstrates the interplay between analytic and geometric concepts.

Theorem 2.1.4 ***(Basic separation)*** *Suppose that $C \subseteq \mathbf{E}$ is closed and convex and the point $y \in E$ does not lie in C. Then, there exist a nonzero $a \in \mathbf{E}$ and $b \in \mathbb{R}$ such that*

$$b < \langle a, y \rangle \quad , \quad b \geqslant \langle a, x \rangle \quad (\forall x \in C) \ .$$

Proof Define $f : C \to \mathbb{R}$ by $f(x) = \|y - x\|^2/2$. We already know that there exists a unique minimizer $\bar{x} = P_C(y)$ which also satisfies

$$(\forall x \in C) \quad \langle y - \bar{x}, x - \bar{x} \rangle \leqslant 0 \tag{2.2}$$

Choosing $a = y - \bar{x}$, $b = \langle y - \bar{x}, \bar{x} \rangle$ yields:

$$\langle a, y \rangle - b = \|y\|^2 - \langle \bar{x}, y \rangle - b = \|y\|^2 - 2\langle \bar{x}, y \rangle + \|\bar{x}\|^2 = \|y - \bar{x}\|^2 > 0 \Rightarrow \langle a, y \rangle > b$$

$$(\forall x \in C) \quad \langle a, x \rangle = \langle y - \bar{x}, x \rangle \leqslant \langle y - \bar{x}, \bar{x} \rangle = b \ ,$$

where the last inequality follows from (2.2). ■

It follows that a closed convex set coincides with the intersection of all the closed affine halfspaces that contain it.

Exercises

Exercise 2.1.1 *Let $C \subseteq \mathbf{E}$ and let $\bar{x} \in \mathrm{int}(C)$. Recall that a function $f : C \to \mathbb{R}$ is Fréchet differentiable at $\bar{x}$ if*

$$f(\bar{x}+d) - (f(\bar{x}) + \langle \nabla f(\bar{x}), d\rangle) = o(\|d\|)$$

where the function $o : \mathbb{R}_+ \to \mathbb{R}$ satisfies $\lim_{r \downarrow 0} \frac{o(r)}{r} = 0$.

1. *Show that if f is Fréchet differentiable at $\bar{x}$, then for any $d \in \mathbf{E}$, $f'(\bar{x}; d) = \langle \nabla f(x), d\rangle$. Conclude that Fréchet differentiability implies (Gâteaux) differentiability.*

2. *Show that if the function is (Gâteaux) differentiable and the limit $f'(\bar{x}; d)$ is uniform[4] over d, then the function is Fréchet differentiable at $\bar{x}$.*

Exercise 2.1.2 *We extend Taylor's Theorem to multivariate real-valued functions.*

1. *Mean-value theorem: Let $f : \mathbf{E} \to \mathbb{R}$ be a differentiable function. Show that for every $x, y \in \mathbf{E}$, there exists $\alpha \in (0, 1)$ such that*

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

2. *Assume now that f is twice differentiable. Prove the following important consequence of the multivariate Taylor's theorem: for any two points x, y, there exists $\alpha \in [0, 1]$ such that*

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x + \alpha(y - x))(y - x) \ ,$$

(Hint for both parts: consider the function $\phi : [0, 1] \to \mathbb{R}$ defined by $\phi(t) = f(x + t(y - x))$.)

Exercise 2.1.3 ***(First-order characterization of convexity)*** *Let $C \subseteq E$ be a open convex set and let $f : C \to \mathbb{R}$. Prove the following facts:*

1. *For any $\bar{x}, x \in C \subseteq \mathbf{E}$, the function $t \in (0, 1] \to \frac{f(\bar{x}+t(x-\bar{x}))-f(\bar{x})}{t}$ is monotonically nondecreasing providing that f is convex.*

2. *Conclude that if f is convex then for any $\bar{x}, x \in C$, $f'(\bar{x}; x-\bar{x})$ exists in $[-\infty, \infty)$ and $f'(\bar{x}; x - \bar{x}) \leqslant f(x) - f(\bar{x})$. In particular, if f is convex and differentiable, then*

$$\nabla f(\bar{x})^\top (x - \bar{x}) = f'(\bar{x}; x - \bar{x}) \leqslant f(x) - f(\bar{x}) \ , \tag{2.3}$$

Conclude the proof of Theorem 2.1.2.

[4] That is, for any $\epsilon > 0$, there exists $s > 0$ such that for any $0 < t \leqslant s$ and any $d \in \mathbf{E}$, $\left|\frac{f(\bar{x}+td)-f(\bar{x})}{t} - f'(\bar{x}; d)\right| \leqslant \epsilon$ (note that for every ϵ, there exists a choice of $s > 0$ that works for all $d \in \mathbf{E}$).

3. *Conversely, if f is differentiable on C and Equation (2.3) holds for every $\bar{x}, x \in C$, then f is convex. (Hint: for $x, y \in C$, $\alpha \in [0,1]$ and $z = \alpha x + (1-\alpha)y$, consider the expression $\alpha \cdot f'(z; x-z) + (1-\alpha) f'(z; y-z)$.).*

4. *Assuming that f is differentiable, f is strictly convex if and only if Equation (2.3) is strict whenever $x \neq y$.*

5. *Assume that f is differentiable. Show that f is convex if and only if for all $y, x \in C$, we have*
$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geqslant 0$$
Furthermore, f is strictly convex if and only if the inequality is strict whenever $x \neq y$. (Hint: use Exercise (2.1.2)).

6. *Let $I \subseteq \mathbb{R}$ be an open interval and assume that $f : I \to \mathbb{R}$ is differentiable. Show that f is (strictly) convex if and only if f' is monotonically nondecreasing (increasing).*

TODO: Bregman distance

Exercise 2.1.4 *Prove Corollary 2.1.1. (Hint: Recall from Exercise (1.2.1) that the null space of A coincides with the complement of the column space of $A^\star$.)*

Exercise 2.1.5 ***(Second-order characterization of convexity)*** *Let $C \subseteq \mathbb{R}^d$ be a open convex set and let $f : C \to \mathbb{R}$ be a twice differentiable function.*

1. *Show that if $\nabla^2 f(x)$ is positive semidefinite (positive definite) for every x, then f is (strictly) convex. Furthermore, if f is twice continuously differentiable and convex, then $\nabla^2 f(x)$ is positive semidefinite. (Hint: use Exercise (2.1.2)).*

2. *Conclude that if $C = I \subseteq \mathbb{R}$ is an open interval, then f is (strictly) convex if and only if $f'' \geqslant 0$. furthermore, if $f'' > 0$, then f is strictly convex.*

3. *Is it true that if f is strictly convex and twice continuously differentiable, then its Hessian at any point is positive definite?*

4. *Prove Theorem 2.1.3.*

Exercise 2.1.6 *The above discussion of first and the second-order characterizations of convex functions is limited to functions with open doamins. The following result is useful:*

1. *Prove that a continuous function $f : \operatorname{cl} C \to \mathbb{R}$ is convex if and only if its restriction to C is convex.*

2. *Given an example of C and $f : \operatorname{cl} C \to \mathbb{R}$ such that $f|_C$ is strictly convex but f is only convex.*

Exercise 2.1.7 *Let $f = h \circ g$, where $h, g : \mathbb{R} \to \mathbb{R}$ are differentiable.*

1. *Show that $f''(x) = h''(g(x))(g'(x))^2 + g''(x)h'(g(x))$.*
2. *Conclude that f is convex under each of the following conditions:*
   (a) *The function h is convex and nondecreasing and g is convex.*
   (b) *The function h is convex and nonincreasing and g is conave.*
3. *Extend the above results to the case where $g : \mathbb{R}^d \to \mathbb{R}$.*
4. *Derive analogous conditions for the concavity of f.*

Exercise 2.1.8 *Let $C \subseteq \mathbf{E}$ be an open convex set. Show that the twice continuously differentiable function $f : C \to \mathbb{R}$ is convex if and only if for every $x, y \in C$, the function $\phi_{x,y} : \mathbb{R} \to \mathbb{R}$, defined by $\phi_{x,y}(t) = f(x + t(y - x))$, satisfies $\phi''_{x,y}(0) \geqslant 0$. Furthremore, if for any such $\phi_{x,y}$, $\phi''_{x,y}(0) > 0$ if $x \neq y$, then f is strictly convex. (Hint: Use Exercise (1.1.13), Exercise (2.1.5) and Exercise (2.1.6). Show that for any $\phi_{x,y}$ and any $t \in (0, 1)$, $\phi''_{x,y}(t) = \phi''_{x+\alpha(y-x),y}(0)/(1 - t)^2$.)*

Exercise 2.1.9*

1. *Consider the function $f : \mathbb{S}^d_{++} \to \mathbb{R}$ defined by $f(X) = \log \det X$.*
   (a) *Show that $\nabla f(X) = X^{-1}$. (Hint: Note that for any $D \in \mathbb{S}^n$ and small enough $t > 0$, $f(X + tD) = \log \det (X(I + tX^{-1}D)) = \log \det X + \log \det(I + tX^{-1}D)$. Also note that $X^{-1}D$ is similar to the positive semidefinite matrix $X^{-1/2}DX^{-1/2}$. Express the directional derivative using the eigenvalues of $X^{-1/2}DX^{-1/2}$.)*
   (b) *Show that f is convex. (Hint: use Exercise (2.1.3).)*
2. *Consider the function $f : \mathbb{S}^n_{++} \to \mathbb{R}$ defined by $f(X) = \operatorname{tr} X^{-1}$.*
   (a) *Show that $\nabla f(X) = -X^{-2}$. (Hint: Use the relation $X + tD = X(I - (-tX^{-1}D))$. Verify the identity $(I - A)^{-1} = \sum_{i=0}^{\infty} A^i$ which holds for any matrix A whose maximal singular value is at most 1.)*
   (b) *Prove that f is convex. (Hint: use Exercise (2.1.8).)*

Exercise 2.1.10

1. *Let $C \subseteq \mathbf{E}$ be a closed and convex set. For any $y \in \mathbf{E}$ and $x \in C$, $\|x - P_C(y)\| \leqslant \|x - y\|$.*

2. *Show that the projection from* $\mathbf{E} = \mathbb{R}^d$ *onto* $\mathbb{R}^d_+$ *is given* $y \mapsto y^+$, *where* $y_i^+ = \max\{0, y_i\}$. *Similarly, show that the projection from* $\mathbb{S}^d$ *onto* $\mathbb{S}^d_+$ *is given by* $U(\operatorname{diag}\lambda)U^\top \mapsto U\operatorname{diag}(\lambda^+)U^\top$.

3. *Let* $V \subseteq E$ *be a linear subspace and denote by* $\{v_1, \ldots, v_k\}$ *an orthonormal basis for* V. *Show that the projection mapping onto* V *is given by the linear map* $x \mapsto \sum_{i=1}^k v_i v_i^\top x$.

Exercise 2.1.11* ***(Supporting Hyperplane Theorem)*** *Let* $C \subseteq \mathbf{E}$ *be a convex set and let* $\bar{x} \notin C$. *Show that there exists* $a \in \mathbf{E}$ *such that*

$$\sup_{x \in C} \langle x, a \rangle \leqslant \langle \bar{x}, a \rangle \ .$$

(Hint: Consider the case $\bar{x} \in \operatorname{cl}(C)$. *We can pick a sequence* $(y_n) \subsetneq C$ *that converges to* $\bar{x}$ *(why?). Use the separation theorem to find* $a_n \in \mathbf{E}$, $b_n \in \mathbb{R}$ *which separate between* y_n *and* C. *Crucially, we may assume w.l.o.g. that for all* n, $\|a_n\|$ *and* $|b_n|$ *are at most* c *for some constant* $c > 0$.*)*

Exercise 2.1.12 ***(Strong Separation)*** *Suppose that the sets* $C, D \subseteq \mathbf{E}$ *are closed and convex and* D *is also bounded.*

1. *Show that the set* $D - C$ *is closed and convex.*

2. *Deduce that if* C *and* D *are disjoint, there there exists a nonzero* $a \in E$ *such that*
$$\inf_{x \in D} \langle a, x \rangle > \sup_{y \in C} \langle a, y \rangle$$

3. *Show that no such separation exists for the sets* $C = \{(x_1, x_2) : x_1 > 0, x_2 \geqslant 1/x_1\}$ *and* $D = \{(x_1, x_2) : x_2 = 0\}$.

2.2 Theorems of the Alternative

A prominent consequence of the basic separation theorem is a family of results called "theorems of the alternative". We discuss two such results due to Gordan and Farkas.

Theorem 2.2.1 *For any sequence of points $a_1 \ldots, a_m \in \mathbf{E}$, exactly one of the following systems has a solution:*

$$0 = \sum_{i=1}^{m} \lambda_i a_i \ , \ \textit{each } \lambda_i \geqslant 0, \ \sum_{i=1}^{m} \lambda_i = 1 \ .$$

$$(\forall i \in [m]) \ \ \langle a_i, x \rangle < 0 \quad , x \in \mathbf{E} \ .$$

Proof Consider the set $C = \operatorname{conv}\{a_1, \ldots, a_m\}$. Assume that $0 = \sum_{i=1}^{m} \lambda_i a_i \in C$ and assume by contradiction that the second system has a solution $\bar{x} \in \mathbf{E}$. Then, we obtain a contradiction by

$$0 = \langle \sum_{i=1}^{m} \lambda_i a_i, \bar{x} \rangle = \sum_{i=1}^{m} \lambda_i \langle a_i, \bar{x} \rangle < 0$$

(since all the λ_i's are nonnegative and at least one of them is positive). Assume now that the first system has no solution. It is not hard to verify that the set C is closed and convex. The Basic separation theorem implies that there exists $x \in \mathbf{E}$ and $b \in \mathbb{R}$ such that

$$\langle 0, x \rangle > b \ \ , \ \ (\forall i \in [m]) \ \langle a_i, x \rangle \leqslant b \ .$$

Since b must be negative, it follows that the second system has a solution. ■

For the second theorem of alternative, we need the following important result.

Theorem 2.2.2 ***(Carathéodory's Theorem)*** *Let $\{a_i : i \in I\} \subseteq \mathbf{E}$ be a finite set. For each $J \subseteq I$, we consider the* finitely *generated cone*

$$C_J = \{\sum_{j \in J} \mu_j a_j : \textit{each } \mu_j \geqslant 0\}$$

Also, we consider the convex hull $C = \textit{conv}\{a_i : i \in I\}$.

1. *The set C_I is equal to the union of all those cones C_J for which the set $\{a_j : j \in J\}$ is independent.*
2. *The set C_I is closed.*
3. *Every vector in C can be expressed as a convex combination of at most* $\dim \mathbf{E}+1$ *vectors from $\{a_i : i \in I\}$.*

The proof is outlined in the exercises.

Lemma 2.2.1 ***(Farkas)*** *For any set of points $a_1, \ldots, a_m$ and $c \in \mathbf{E}$, exactly one of the following systems has a solution:*

$$c = \sum_{i=1}^{m} \lambda_i a_i, \quad (\forall i \in [m])\ \lambda_i \geqslant 0 \ .$$

$$\langle c, w \rangle > 0, \quad (\forall i \in [m])\ \ \langle a_i, w \rangle \leqslant 0, \quad w \in \mathbf{E} \ .$$

Proof The fact that if the first system has a solution then the second does not have a solution follows as in Gordan theorem. Denote by $C = \{\sum_{i=1}^{m} \mu_i a_i : \text{each } \mu_i \geqslant 0\}$. The set C is convex and according to Theorem 2.2.2, it is also closed. The basic separation theorem implies that if the first system has no solution, then there exists $w \in \mathbf{E}$ and $b \in \mathbb{R}$ such that $\langle x, w \rangle \leqslant b$ for all $x \in C$ and $\langle c, w \rangle > b$. We argue that that b can be chosen to be 0. First, $b \geqslant 0$ since $0 \in C$. Next, we observe that if there exist $\bar{x} \in C$ such that $\langle \bar{x}, w \rangle > 0$, then $\sup_{x \in C} \langle x, w \rangle = \infty$. Thus, if the first system has no solution then the second system has a solution. ■

Application: Duality of Linear Programming

Consider a linear program (LP) of the form

$$\begin{array}{ll} \max & c^\top x \\ \text{s.t.} & Ax \leqslant b \ , \ x \geqslant 0 \end{array} \tag{2.4}$$

where $c \in \mathbb{R}^d$, $A \in \mathbb{R}^{n \times d}, b \in \mathbb{R}^n$ and inequality between vectors is understood component-wise. The dual linear program is given by

$$\begin{array}{ll} \min & b^\top y \\ \text{s.t.} & A^\top y \geqslant b \ , \ y \geqslant 0 \ . \end{array} \tag{2.5}$$

The former problem is called the *primal* problem and the second is called the *dual* problem. The weak-duality theorem follows immediately from the definitions of the problems.

Lemma 2.2.2 ***(Weak Duality for Linear Programming)*** *For any two feasible solutions x and y for the primal and the dual programs, we have $c^\top x \leqslant b^\top y$.*

The strong duality theorem establishes conditions under which an equality is obtained. This result has numerous applications in continuous and discrete optimization.

Theorem 2.2.3 ***(Strong Duality for Linear Programming)*** *The primal and the dual programs described above satisfy exactly one of the following properties:*

1. *Both are infeasible (i.e., there are no vectors that satisfy the constraints).*
2. *The primal problem is unbounded and the dual is infeasible.*
3. *The dual problem is unbounded and the primal is infeasible.*
4. *Both problems have optimal solutions, denoted $x^\star$ and $y^\star$, respectively. They satisfy the equality*
$$c^\top x^\star = b^\top y^\star \ . \tag{2.6}$$

Proof It follows from the weak duality Theorem that these are indeed all the possible scenarios. We prove the strong duality equation (2.6). Denote the rows of A by $a_1, \ldots, a_n$. Let $I = \{i \in [n] : a_i^\top x^\star = b_i\}$ be the constraints that are tight at $x^\star$. We argue that $c = \sum_{i \in I} \lambda_i a_i$ for some nonnegative vector of coefficients, λ. Otherwise, according to Farkas's lemma, there exists $v \in \mathbb{R}^d$ such that $\langle v, c \rangle > 0$ and $\langle v, a_i \rangle \leqslant 0$ for all $i \in I$. It is clear then that for sufficiently small $\epsilon > 0$, the vector $x^\star + \epsilon v$ is a feasible solution with $c^\top (x^\star + \epsilon v) > c^\top x^\star$, contradicting the optimality of $x^\star$.

Define $y \in \mathbb{R}^n$ by
$$y_i = \begin{cases} \lambda_i & i \in I \\ 0 & i \notin I \end{cases}$$
Then,
$$y^\top b = \sum_{i \in I} y_i b_i = \sum_{i \in I} y_i (a_i^\top x^\star) = c^\top x^\star \ .$$
Since y is feasible, the theorem follows using the Weak duality theorem. ■

Note that the vector y constructed during the proof satisfies the following property: for every $i \in [n]$, either the *dual variable* y_i is zero or the corresponding constraint is tight. This is not a coincidence.

Corollary 2.2.1 ***(Complementary Slackness for Linear Programming)*** *The following conditions are equivalent:*

1. *The vectors $x^\star$ and $y^\star$ are optimal solutions for the primal and the dual problems, respectively.*
2. *For every $i \in [n]$, either $(Ax^\star)_i = b_i$ or $y_i^\star = 0$. Similarly, for every $j \in [d]$, either $(A^\top y^\star)_j = c_j$ or $x_j^\star = 0$.*

Exercises

Exercise 2.2.1 *Let $a_1, \ldots, a_m \in \mathbf{E}$. Show that $\mathrm{conv}\{a_1, \ldots, a_m\}$ is compact.*

Exercise 2.2.2* ***(Carathéodory's Theorem)***

1. *Prove Theorem 2.2.2. (Hint: verify the fact that a finite union of closed sets is closed. Hence, given the first part of the lemma, it suffices to prove the second part for any C_J which is generated by an independent set of vectors. For the last part of the lemma, note that if $x = \sum_{i=1}^m \mu_i a_i \in C$, then $(x,1) = \sum_{i=1}^m \mu_i (a_i,1)$.)*

2. *Conclude that for every set $P \subseteq \mathbb{R}^d$, every point $x \in \operatorname{conv} P$ belongs to the convex hull of a finite set $P' \subseteq P$ of size at most $d+1$.*

Exercise 2.2.3

1. *Deduce Gordan's theorem from the Farkas lemma. (Hint: consider the augmented vectors $(a_i, 1)$ for all i.)*

2. *($\star$) Deduce the Farkas lemma from Gordan's theorem.*

Exercise 2.2.4 *Prove Lemma 2.2.2.*

Exercise 2.2.5 *Prove Corollary 2.2.1.*

Exercise 2.2.6* ***(The minimax theorem)*** *Consider a zero-sum game as formulated in Exercise (1.1.12), where the players are allowed to play mixed strategies. Following the same notations, we already know that $Mm \leqslant mM$. In this question we prove an equality. This result is called von Neumann theorem.*

1. *Assume by contradiction that $Mm < mM =: v$. We consider a modified game in which all the values are decreased by v, i.e., the game matrix, $\tilde{A} \in \mathbb{R}^{n \times m}$, satisfies $\tilde{A}_{i,j} = A_{i,j} - v$. Denoting the rows of $\tilde{A}$ by $\tilde{a}_1, \ldots, \tilde{a}_n$, use Exercise (1.1.12) to conclude that for every $p \in \Delta_n$ there exists $q \in \{e_1, \ldots, e_m\}$ such that $p^\top \tilde{A} q < 0$.*

2. *Deduce that the convex hull of $\{\tilde{a}_1, \ldots, \tilde{a}_n\}$ and the set $\mathbb{R}^m_+$ are disjoint.*

3. *Use Exercise (2.1.12) to obtain the contradiction $mM < v$.*

2.3 Max-functions and Lagrangian Duality

We now establish optimality conditions for one of the most common forms of an optimization problem:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g_i(x) \leqslant 0 \quad \text{for } i = 1, \ldots, m \\ & x \in C \ . \end{array} \tag{2.7}$$

where C is a convex subset of $\mathbf{E}$ and where f and $g_1, \ldots, g_m$ are differentiable functions defined over C. We say that $x \in C$ is *feasible* if $g_i(x) \leqslant 0$ for all $i \in [m]$. The *feasible region* is the set of feasible points. The problem is said to be feasible if the feasible region is not empty. A point $x \in C$ is a local minimizer if for all close feasible points x, $f(x) \geqslant f(\bar{x})$.

A key idea in optimization is the association of unconstrained optimization problems with constrained minimization problems. In this section we present two such methodologies: *max-functions* and *Lagrangian*. We next explain each of these concepts.

Let $\bar{x} \in \text{int}(C)$ (in our context, $\bar{x}$ can be viewed as a candidate local minimizer). Define $h_0 : C \to \mathbb{R}$ by $h_0(x) = f(x) - f(\bar{x})$, and for $i = 1, \ldots, m$, let $h_i = g_i$. We define the max-function $h(x) = \max_{i \in \{0,\ldots,m\}} h_i(x)$. Note that while the h_i's are smooth, the function h is certainly not smooth. Still, the directional derivative exists and has the following useful property.

Lemma 2.3.1 *Let $x \in \text{int}(C)$. Define the index set $I = \{i \in \{0, \ldots, m\} : h_i(x) = h(x)\}$. Then, for all $d \in \mathbf{E}$,*

$$h'(x; d) = \max_{i \in \{0,\ldots,m\}} \langle \nabla h_i(x), d \rangle \ .$$

Proof Let $d \in \mathbf{E}$. By continuity, we may assume w.l.o.g.[5] that $I = \{0, \ldots, m\}$. For all $i \in I$, we have

$$\liminf_{t \downarrow 0} \frac{h(x + td) - h(x)}{t} \geqslant \lim_{t \downarrow 0} \frac{h_i(x + td) - h_i(x)}{t} = \langle \nabla h_i(x), d \rangle \ .$$

Suppose

$$\limsup_{t \downarrow 0} \frac{h(x + td) - h(x)}{t} > \max_i \langle \nabla h_i(x), d \rangle \ .$$

Then, we can take a sequence $(t_n)_{n \in \mathbb{N}}$ of positive scalars such that $t_n \downarrow 0$ and

$$(\forall n \in \mathbb{N}) \quad \frac{h(x + t_n d) - h(x)}{t_n} > \max_i \langle \nabla h_i(x), d \rangle + \epsilon \ .$$

[5] All other functions do not affect $h'(x; d)$.

for some $\epsilon > 0$. For each $n \in \mathbb{N}$, there exists an index $i = i(n) \in I$ such that $h(x + td) = h_i(x + td)$. At least one index $j \in I$ appears in the sequence $(i(n))$ infinitely many times. In the limit we obtain the contradiction

$$\langle \nabla h_j(x), d\rangle > \max_i \langle \nabla h_i(x), d\rangle + \epsilon \ .$$

■

The *Lagrangian* associated with the optimization problem (2.7) is the function $L : C \times \mathbb{R}^m_+ \to \mathbb{R}$ defined by

$$L(x; \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \ . \tag{2.8}$$

Let $\bar{x}$ be a feasible point. We define the *active set* $I(\bar{x}) = \{i \in [m] : g_i(x) = 0\}$. We say that $\lambda \in \mathbb{R}^m_+$ is a *Lagrange multiplier vector* for $\bar{x}$ if $\bar{x}$ is a critical point of $L(; \lambda)$, i.e.,

$$\nabla f(\bar{x}) + \sum_{i=1}^{m} \lambda_i g_i(\bar{x}) = 0,$$

and *complementary slackness* holds: $\lambda_i = 0$ for any $i \notin I(\bar{x})$.

The Lagrangian provides a very simple sufficient condition for optimality. Note that if λ is a Lagrange multiplier for $\bar{x}$, then $\sum_{i=1}^m \lambda_i g_i(\bar{x}) = 0$. Therefore, $L(\bar{x}, \lambda) = f(\bar{x})$. For any other feasible point $x \in C$, it is easy to check that $L(x, \lambda) \leqslant f(x)$. The following conclusion is immediate.

Theorem 2.3.1 *Consider the optimization problem (2.7). Assume that $\bar{x} \in C$ is a feasible solution and there exists $\lambda \in \mathbb{R}^m_+$ such that $\bar{x}$ minimizes the function $L(\cdot; \lambda)$ and $\lambda_i = 0$ for all $i \notin I(\bar{x})$. Then $\bar{x}$ is a global minimizer. In particular, if $x \in \operatorname{int}(C)$, $f, g_1, \ldots, g_m$ are convex and there exists a Lagrange multiplier vector for $\bar{x}$, then $\bar{x}$ is a global minimizer.*

Proof The first part was proved above. When $f, g_1, \ldots, g_m$ are convex, $\bar{x}$ minimizes the function $L(\cdot, \lambda)$ if and only if $\bar{x}$ is a critical point of $L(\cdot, \lambda)$. The "in particular" part is now clear. ■

We turn to discuss necessary conditions. In particular, we aim at establishing necessary conditions that ensure the existence of a Lagrange multiplier vector. We begin with the following theorem.

Theorem 2.3.2 ***(Fritz John Conditions)*** *Assume that $\bar{x} \in \operatorname{int}(C)$ is a local minimizer of the problem (2.7). Then, there exist $\lambda_0, \lambda_i \in \mathbb{R}_+$ $(i \in I(\bar{x}))$, not all zero, such that*

$$\lambda_0 \nabla f(x) + \sum_{i \in I(x)} \lambda_i \nabla g_i(x) = 0 \ . \tag{2.9}$$

Proof Consider the function $h(x)$ defined above. It can be seen that for $i = 0, \ldots, m$, $h_i(\bar{x}) = 0$. Therefore, $h(\bar{x}) = 0$. Furthermore, if x is close enough to $\bar{x}$, then $h(x) \geqslant h_0(x) \geqslant 0$. Hence, $\bar{x}$ is a local minimizer of h. According to Lemma 2.3.1 and Theorem 2.1.1, for all $d \in \mathbf{E}$,

$$0 \leqslant h'(x; d) = \max\{\langle \nabla f(x), d\rangle, \langle \nabla g_i(x), d\rangle : i \in I(x)\} \ .$$

Therefore, for all $d \in \mathbf{E}$, both $\langle \nabla f(\bar{x}), d\rangle \geqslant 0$ and

$$(\forall i \in I(\bar{x})) \quad \langle \nabla g_i(\bar{x}), d\rangle \geqslant 0 \ .$$

Putting it differently, there is no $d \in \mathbf{E}$ such that $\langle \nabla g_i(\bar{x}), d\rangle < 0$ for all $i \in I(\bar{x})$. The feasibility of Equation (2.9) follows now from Gordan's theorem. ■

Note that the theorem does not yield a Lagrange multiplier. We need to impose additional assumptions in order to rule out the possibility that $\lambda_0 = 0$. For example, one such condition is linear independence of the set $\{\nabla g_i(\bar{x}) : i \in I(\bar{x})\}$. We now discuss a weaker condition. We say that the *Mangasarian-Fromovitz constraint qualification* holds at $\bar{x}$ if there is a direction $d \in \mathbf{E}$ such that $\langle \nabla g_i(\bar{x}), d\rangle < 0$ for all $i \in I(\bar{x})$.

Theorem 2.3.3 ***(Karush-Kuhn-Tucker (KKT) conditions)*** *Suppose that the conditions in Theorem 2.3.2 hold and that Mangasarian-Fromovitz constraint qualification holds at $\bar{x}$. Then, there exists a Lagrange multiplier vector for $\bar{x}$.*

Proof Theorem 2.3.2 implies the feasibility of Equation (2.9). By the opposite direction of Gordan's theorem, the constraint qualification implies that $\lambda_0 \neq 0$. Dividing Equation (2.9) by λ_0, we conclude the existence of a Lagrange multiplier vector. ■

Exercises

Exercise 2.3.1 *Compute the directional derivatives of the absolute value function $|\cdot|$ at the point 0.*

Exercise 2.3.2 ***(The entire regularization path)*** *Let $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be differentiable functions. We consider two minimization problems:*

$$\textit{Regularized minimization:} \quad \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) + \lambda \|w\|^2 \tag{2.10}$$

$$\textit{Constrained minimization:} \quad \min_{w : \|w\| \leqslant B} \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{2.11}$$

where B and λ are two positive scalars. Show that if $w^\star$ is a global minimizer of the regularized problem, then there exists $B > 0$ such that $w^\star$ is a minimizer of the constrained problem. Conversely, if $f_1, \ldots, f_n$ are convex and $w^\star \neq 0$ is a global minimizer of the constrained problem (2.11), then there exists $\lambda > 0$ such that $w^\star$ is also a global minimizer of the corresponding regularized problem (2.10).

Exercise 2.3.3 *Show that the Mangasarian-Fromovitz constraint qualification is indeed weaker than linear independence. That is, if* $a_1, \ldots, a_m \in \mathbf{E}$ *are linearly independent, then there exits* $d \in \mathbf{E}$ *such that* $\langle d, a_i \rangle < 0$ *for all* $i \in [m]$.

Chapter 3

Nonsmooth Optimization and Lagrangian Duality

3.1 Subgradients and Convex Functions

We are already familiar with the following useful characterization of differentiable convex functions: for any point x in the domain, the first-order approximation around x lies below the function. This is another example for how a local information yields a global information on f. In the absence of differentiability, this property is captured by *subgradients*.

Definition 3.1.1 *Let $C \subseteq \mathbf{E}$, $f : C \to \mathbb{R}$ and $\bar{x} \in C$. A vector $a \in E$ is a subgradient of f at $\bar{x}$ if for all $x \in C$,*

$$\langle a, x - \bar{x} \rangle \leqslant f(x) - f(\bar{x})$$

The subdifferential of f at $\bar{x}$, denoted $\partial f(\bar{x})$, is the set of all subgradients of f at $\bar{x}$.

The subgradient provides a trivial yet important optimality condition: a point $\bar{x}$ is a global minimizer of a function f is and only if $0 \in \partial f(\bar{x})$.

Our geometric intuition tells us that the subdifferentials of a convex function are never empty. Furthermore, if f is differentiable, the subdifferential consists exactly of one element, the gradient. The simplest illustrating example is the absolute value function. At any point $x \neq 0$, the function is differentiable and the only subgradient is $\operatorname{sgn}(x)$. For $x = 0$, any slope $\alpha \in [-1, 1]$ satisfies the definition. It turns out that subgradients characterize convex function in a very strong way.

Theorem 3.1.1 *Let $C \subseteq \mathbf{E}$ be a convex set and let $f : C \to \mathbb{R}$. If $\partial f(x) \neq \varnothing$ for all $x \in C$, then f is convex. Conversely, if f is convex, then the subdifferential of any point $\bar{x} \in \operatorname{int} C$ is not empty. Moreover, we have the following relationship between subgradients and directional derivatives: For any $d \in \mathbf{E}$,*

$$f'(\bar{x}; d) = \max\{\langle a, d \rangle : a \in \partial f(\bar{x})\} \ .$$

In particular, if f is differentiable, $\partial f(\bar{x})$ consists exactly of $\nabla f(\bar{x})$.

The proof of the first part of the theorem (sufficient condition for convexity) is analogous to the third part of Exercise (2.1.3). Throughout the rest of this section, we prove the second part by relating the subgradient to the directional derivative. Before proceeding, by taking a quick picture, we see that existence of subgradients can be proved by means of separation theorems. Namely, the point $(x, f(x))$ can be separated from the epigraph of f, $\{(x, y) : f(x) < y\}$. Indeed, this approach is outlined in the exercises.

Throughout the rest of this section, let us fix a convex set $C \subseteq \mathbf{E}$, a convex function $f : C \to \mathbb{R}$ and $\bar{x} \in \operatorname{int} C$. We begin with the following simple result.

Lemma 3.1.1 *An element $a \in E$ is a subgradient of f at $\bar{x}$ if and only if $\langle \phi, \cdot \rangle \leqslant f'(\bar{x}; \cdot)$.*

Proof If a is a subgradient, then for any $d \in E$ and $t > 0$,

$$\frac{f(\bar{x} + td) - f(\bar{x})}{t} \geqslant \frac{\langle a, td \rangle}{t} = \langle a, d \rangle \ .$$

In particular, the inequality is satisfied when $t \downarrow 0$. Therefore, $\langle a, \cdot \rangle \leqslant f'(x; \cdot)$. Assume now that $\langle a, d \rangle \leqslant f'(\bar{x}; d)$ for all $d \in \mathbf{E}$. Then, since f is convex, we know that $f'(\bar{x}; d) \leqslant f(\bar{x} + d) - f(\bar{x})$ for all d, so we obtain that a is a subgradient. ■

The following lemma provides some (additional) properties of the directional derivative. We use the following definitions: a function $g : \mathbf{E} \to \mathbb{R}$ is said to be *sublinear* if for all $d, e \in \mathbf{E}$ and $\lambda, \mu \in \mathbb{R}_+$, $g(\lambda d + \mu e) \leqslant \lambda g(d) + \mu g(e)$. A function $g : \mathbf{E} \to \mathbb{R}$ is said to be *subaditive* if for all $d, e \in \mathbf{E}$, $g(d + e) \leqslant g(d) + g(e)$. The function g is *positively homogeneous* if for all $d \in \mathbf{E}$ and $\lambda \in \mathbb{R}_+$, $g(\lambda d) = \lambda g(d)$. An exercise shows that g is sublinear if and only if g is subaditive and positively homogeneous. Also, if g is sublinear, then $g(-x) \geqslant -g(x)$ for all x (Exercise (3.1.6*)).

Lemma 3.1.2
The directional derivative satisfies the following properties:

1. *For any direction $d \in E$, the directional derivative $f'(\bar{x}; d)$ exists in $\mathbb{R}$.*
2. *Sublinearity in the second argument: the function $f'(\bar{x}; \cdot)$ is sublinear.*

Now let $d \in \mathbf{E}$ and denote $\mu = f'(\bar{x}; d)$. We know that $\mu \in \mathbb{R}$. Define the linear subspace $S = \{\alpha d : \alpha \in \mathbb{R}\}$ and the linear function $\Lambda : S \to \mathbb{R}$ by $\Lambda(\alpha d) = \alpha\mu$. Since $f'(\bar{x}; \alpha d) = \alpha\mu$ when $\alpha \geqslant 0$ and for $\alpha < 0$,

$$f'(\bar{x}; \alpha d) \geqslant -f'(\bar{x}; |\alpha| d) = \alpha\mu \ ,$$

we have that $\Lambda(\cdot) \leqslant f'(\bar{x}; \cdot)|_S$ and $\lambda(\alpha d) = f'(\bar{x}; \alpha d)$ for $\alpha \geqslant 0$. In particular, $\Lambda(d) = f'(\bar{x}; d)$. It would be great if we could extend Λ to the entire space while preserving the relation $\Lambda(\cdot) \leqslant f'(\bar{x}; \cdot)$. This gives us an excuse to prove the following fundamental theorem.

Theorem 3.1.2 ***(Hahn-Banach extension)*** *Suppose $p : \mathbf{E} \to \mathbb{R}$ is a sublinear function and $\Lambda : S \to \mathbb{R}$ is linear, where S is a subspace of $\mathbf{E}$. If $\Lambda \leqslant P|_S$, then there exists an extension of Λ, $\bar{\Lambda} : \mathbf{E} \to \mathbb{R}$, which satisfies $\bar{\Lambda}|_S = \Lambda$ and $\bar{\Lambda} \leqslant p$.*

Proof Let $x_1 \in \mathbf{E} \backslash S$. We show how to extend Λ to $S_1 = \operatorname{span}(S \cup \{x_1, \})$. Since the dimension is finite,[1] by repeating this argument we conclude the proof.

Note that every vector $v \in S_1$ can be uniquely written as $x + tx_1$, where $x \in S$ and $t \in \mathbb{R}$. We define $\bar{\Lambda} : S_1 \to \mathbb{R}$ by $\bar{\Lambda}(x + tx_1) = f(x) + \mu t$, where μ will be chosen later. It is clear that $\bar{\Lambda}$ is linear and that $\bar{\Lambda}|_S = \Lambda$. In order to satisfy the relation $\bar{\Lambda} \leqslant p$, it suffices to ensure that for every $x, y \in S$,

$$\bar{\Lambda}(y + x_1) \leqslant p(y + x_1) \ , \ \bar{\Lambda}(x - x_1) \leqslant p(x - x_1) \ . \tag{3.1}$$

The Sublinearity of p and our assumptions on Λ imply that

$$p(y + x_1) + p(x - x_1) \geqslant p(y + x) \geqslant \lambda(y + x) = \Lambda(y) + \Lambda(x) \ .$$

Hence,

$$p(y + x_1) - \Lambda(y) \geqslant \Lambda(x) - p(x - x_1) \ .$$

We now choose $\mu = \sup_{x \in S} \Lambda(x) - p(x - x_1)$. It is straightforward to check that μ satisfies Equation (3.1). ∎

We leave it as an easy exercise to complete the proof of Theorem 3.1.1.

Exercises

Exercise 3.1.1 *Let $C \subseteq \mathbf{E}$, $f : C \to \mathbb{R}$. Show that a point $\bar{x} \in C$ is a global minimizer of a function f is and only if $0 \in \partial f(\bar{x})$.*

Exercise 3.1.2 *Use the supporting hyperplane theorem to deduce that the subdifferential of a convex function $f : C \to \mathbb{R}$ at any point $x \in \operatorname{int}(C)$ is not empty.*

Exercise 3.1.3 *Prove Lemma 3.1.2 (Hint: you may rely on Exercise (3.1.6*).)*

Exercise 3.1.4 *Complete the proof of Theorem 3.1.1.*

Exercise 3.1.5 *Use Lemma 3.1.1 in order to provide an alternative proof for the fact that if f is differentiable at $x \in \operatorname{int} C$, then $\partial f(x) = \{\nabla f(\bar{x})\}$.*

Exercise 3.1.6* *Let $g : \mathbf{E} \to \mathbb{R}$. Prove the following facts:*

[1] The theorem is also true in infinite-dimensional spaces. However, a standard proof of this result in the infinite-dimensional case usually involves the axiom of choice (although this axiom can be replaced by the weaker ultrafilter theorem).

1. *The function g is sublinear if and only if g is subaditive and positively homogeneous.*

2. *Show that if g is sublinear then for all x, $g(-x) \geqslant -g(x)$.*

Exercise 3.1.7 *Justify the argument that we made during the proof of Theorem 3.1.2: in order to satisfy the relation $\bar{\Lambda} \leqslant p$, it suffices to ensure that for every $x, y \in S$,*

$$\bar{\Lambda}(y + x_1) \leqslant p(y + x_1) \ , \ \bar{\Lambda}(x - x_1) \leqslant p(x - x_1) \ .$$

Exercise 3.1.8

1. *Let $\|\cdot\|$ be an arbitrary norm defined over $\mathbf{E}$. Compute $\partial f(x)$ at any point $x \in \mathbf{E}$.*

2. *($\star$) Let $f : \mathbf{E} \to \mathbb{R}$ be a sublinear function. Show that*

$$\partial f(0) = \{a \in E : (\forall x \in E) \ \ \langle \phi, x \rangle \leqslant f(x)\} \ ,$$

and for all nonzero $\bar{x} \in E$,

$$\partial f(\bar{x}) = \{a \in \partial f(0) : \langle a, \bar{x} \rangle = f(\bar{x})\} \ .$$

3.2 The Value Function

In this section we describe an alternative approach to the KKT conditions (Theorem 2.3.3) while considering convex but not necessarily differentiable objective and inequality constraints. We begin by extending the type of functions the definition of a convex function. We consider extended real-valued functions $f : \mathbf{E} \to [-\infty, +\infty]$. The domain of f is the set $\operatorname{dom} f = \{x \in \mathbf{E} : f(x) \leqslant +\infty\}$. In order to avoid the appearance of the expression $+\infty - \infty$, we need to generalize the definition of a convex function. We say that f is convex if its epigraph $\{(x, s) \in \mathbf{E} \times \mathbb{R} : f(x) \leqslant s\}$ is convex. It can be verified that f is convex if and only if for all $x, y \in \operatorname{dom} f$ and every $\alpha \in [0, 1]$, $f(\alpha x + (1 - \alpha)y) \leqslant \alpha f(x) + (1 - \alpha)f(y)$. Furthermore, the domain of a convex function is convex.

Allowing the value $+\infty$ helps us in formulating optimization programs. For example, instead of explicitly restricting the feasible region to a set $C \subseteq \mathbf{E}$, we may add the *indicator function* δ_C, where $\delta_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise, to the objective. We usually prefer to exclude the possibility that f takes the value $-\infty$ (e.g., the definition of value function below requires us to allow this value. However, we then provide conditions under which it does not takes this value). We say that f is *proper* if $\operatorname{dom} f \neq \varnothing$ and f never takes the value $-\infty$. The definition of directional derivative and subdifferential are easily extended to proper functions (see Exercise (3.2.3)).

We establish optimality conditions for convex optimization problems of the following form:

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g_i(x) \leqslant 0 \quad \text{for } i = 1, \ldots, m \end{array} \tag{3.2}$$

where $f, g_1, \ldots, g_m : \mathbf{E} \to [-\infty, \infty]$ are convex but not necessarily differentiable. Denoting by $g(x) = (g_1(x), \ldots, g_m(x))$, the Lagrangian $L : \mathbf{E} \times \mathbb{R}^+_m \to \mathbb{R}$ is defined by $L(x, \lambda) = f(x) + \lambda^\top g(x)$. In our context, a vector $\lambda \in \mathbb{R}^m_+$ is a Lagrange multiplier vector for a feasible solution $\bar{x}$ if $\bar{x}$ minimizes $L(\cdot, \lambda)$ and complementary slackness holds: $\lambda_i = 0$ whenever $g_i(\bar{x}) = 0$. Certainly, the sufficient conditions stated in Theorem 2.3.1 apply also here.

Theorem 3.2.1 *If the point $\bar{x}$ is feasible and λ is a Lagrange multiplier vector for $\bar{x}$, then $\bar{x}$ is optimal.*

Note that the proof does not rely on convexity.

The main object in our approach is the value function $v : \mathbb{R}^m \to \mathbb{R}$, defined by

$$v(b) = \inf\{f(x) : g(x) \leqslant b\} .$$

Clearly, v is monotonic in the following sense: $v(b) \leqslant v(a)$ if $b - a \in \mathbb{R}^m_+$. Note that if $\bar{x} \in \operatorname{dom} f$ is optimal for the program (3.2), then $v(0) = f(\bar{x})$. Also, by definition,

for all $x \in \mathbf{E}$, $f(x) \geqslant v(g(x))$. In the absence of differentiability, we will resort to the subdifferential of the value function. For this purpose, we need to ensure that v is proper. The following constraint qualification suffices:

$$\text{Slater condition:} \quad \exists \hat{x} \in \operatorname{dom} f \in \mathbf{E} \text{ s.t. } (\forall i \in [m])\ g_i(\hat{x}) < 0\ .$$

Lemma 3.2.1 *Suppose that $\bar{x} \in \operatorname{dom} f$ is optimal for the convex program (3.2) and that the Slater condition holds. Then v is proper.*

Proof By assumption, $v(0) = f(\bar{x}) \in (-\infty, \infty)$. The slater condition implies that $0 \in \operatorname{int}(\operatorname{dom} v)$. Assume by contradiction that there exits $\mu \in \mathbb{R}^m$ with $v(\mu) = -\infty$. By considering a line of the form $[-s\mu, s\mu] \subseteq \operatorname{dom} v$ and using the convexity of v, we obtain a contradiction. Filling the missing details is left as an exercise. ■

Theorem 3.2.2 ***(Lagrangian necessary conditions)*** *Suppose that $\bar{x} \in \operatorname{dom} f$ is optimal for the convex program (3.2) and that the Slater condition holds. Then there exists a Lagrange multiplier vector for $\bar{x}$*

Proof According to Lemma 3.2.1, the value function v is proper and $0 \in \operatorname{int}(\operatorname{dom} v)$. According to Theorem 3.1.1, the subdifferential $\partial v(0)$ is not empty. Let a be any subgradient of v at 0. By definition of the subgradient and the above elementary properties of the value function, for every $b \in \mathbb{R}^m_+$,

$$f(\bar{x}) = v(0) \leqslant v(b) - \langle a, b\rangle \leqslant v(0) - \langle a, b\rangle = f(\bar{x}) - \langle a, b\rangle\ . \tag{3.3}$$

It follows that $\lambda := -a \in \mathbb{R}^m_+$. In particular, substituting $b = g(\bar{x})$, we obtain that $f(\bar{x}) \leqslant L(\bar{x}, \lambda)$. Since the coordinates of both $g(\bar{x})$ and λ are nonnegative, we conclude that $\lambda_i = 0$ whenever $g_i(\bar{x}) < 0$. We complete the proof by showing that for every $x \in \mathbf{E}$, $L(x, \lambda) \geqslant f(\bar{x})$. Indeed, for every $x \in \mathbf{E}$,

$$f(x) \geqslant v(g(x)) \geqslant v(0) + \langle a, g(x)\rangle = f(\bar{x}) - \lambda^\top g(x)\ .$$

By rearraning we obtain $L(x, \lambda) \geqslant f(\bar{x})$. Hence,

$$L(\bar{x}, \lambda) \leqslant f(\bar{x}) \leqslant L(x, \lambda)\ .$$

We deduce that λ is a Lagrange multiplier vector for $\bar{x}$. ■

Note that if $f, g_1, \ldots, g_m$ are differentiable, the KKT conditions are recaptured. Indeed, it is not hard to see that in this case, the Slater condition is equivalent to the Mangasarian-Fromovitz constraint qualification.

Exercises

Exercise 3.2.1 *Let* $f : \mathbf{E} \to [-\infty, \infty]$.

1. *Prove that* f *is convex if and only if for all* x, y *in* $\operatorname{dom} f$ *(or* $\mathbf{E}$ *if* f *is proper) and every* $\alpha \in [0, 1]$,

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

2. *Show that if* f *is convex, then its domain is convex.*

Exercise 3.2.2 *Complete the proof of Lemma 3.2.1.*

Exercise 3.2.3 *For a proper function* $f : \mathbf{E} \to [-\infty, \infty]$, *we define the subdifferential of* f *at every point* $x \in \operatorname{dom} f$ *as in Definition 3.1.1. For* $x \notin \operatorname{dom} f$, *we define* $\partial f(x) = \varnothing$. *Convince yourself that all the results from Section 3.1 (in particular, Theorem 3.1.1) extend to this setting, where the conditions* $x \in C, x \in \operatorname{int} C$ *are respectively replaced by* $x \in \operatorname{dom} f, x \in \operatorname{int}(\operatorname{dom} f)$.

Exercise 3.2.4 *Establish the equivalence between the Slater condition and the Mangasarian-Fromovitz constraint qualification asserted at' the end of the section.*

Exercise 3.2.5 **TODO: John's Ellipsoid**

3.3 Lagrangian Duality

Consider the convex program (3.2). We now state without a proof[2] a fundamental duality result for convex programs. We consider two progarms:

$$\text{Primal: } p = \inf_{x \in \mathbf{E}} \sup_{\lambda \in \mathbb{R}^m_+} L(x; \lambda) \quad ,$$

$$\text{Dual: } d = \sup_{\lambda \in \mathbb{R}^m_+} \inf_{x \in \mathbf{E}} L(x; \lambda) \quad .$$

It is straighforward to verify that $d \leqslant p$ and that p is equal to the optimal value of the convex program (3.2). The term $p - d$ is named the *duality gap*. The following theorem establishes conditions under which the duality gap is zero.

Theorem 3.3.1 *Suppose that the Slater condition holds for the primal problem. Then the duality gap is zero.*

Exercises

Exercise 3.3.1 *Prove that $d \leqslant p$.*

Exercise 3.3.2 *Show that p is equal to the optimal value of the convex program (3.2)*

[2] Due to the lack of time, we will not see the proof of this important result. Sections 3.3, 4.1, 4.2, 4.3 in [Borwein and Lewis, 2010] provide all the details.

Part II

Optimization

Chapter 4

Condition Number

4.1 Solving Linear Systems using Gradient Descent

We start with one of the most fundamental problems in numerical computation: solving a system of linear equations. Precisely, given a matrix $A \in \mathbb{R}^{d\times d}$ and a vector $b \in \mathbb{R}^d$, our goal is to find a vector x such that $Ax = b$. From a basic course in linear algebra, we know that when A is invertible, a unique solution is given by $x^\star = A^{-1}b$. We also know that $x^\star$ can be computed by applying Gaussian elimination, whose (computational) complexity is dominated by the runtime of matrix multiplication (currently $O(d^{2.37})$). While we are not aware of any exact solver with better worst-case complexity, it turns out that if we are satisfied with approximate solutions, then in many applications we can do much better by applying iterative optimization methods. Concretely, when $A \in \mathbb{S}^d_{++}$, solving the system $Ax = b$ is equivalent to minimizing of the following convex quadratic form:

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) = \frac{1}{2}x^\top Ax - b^\top x \right\} \ . \tag{4.1}$$

Based on this equivalence, we show that a Gradient Descent-based solver, which we simply call GD, essentially reduces the problem of approximating $A^{-1}b$ to the computation of a relatively small number of matrix-vector products with the matrix A. The complexity of this method depends on the *condition number* of A, which is defined by

$$\kappa(A) = \frac{\lambda_1(A)}{\lambda_d(A)} \ , \tag{4.2}$$

where $\lambda_1(A) \geqslant \ldots \lambda_d(A) > 0$ are the eigenvalues of A.

Most of this section is devoted to the analysis of the convergence rate of GD. As will be apparent soon, in our context, it is natural to evaluate the quality of an approximate solution $\bar{x}$ according its distance from $x^\star$ as measured by A; denoting $\|z\|_A = \sqrt{z^\top Az}$, we call $\bar{x}$ an ϵ-approximate solution if $\|\bar{x} - x^\star\|_A^2 \leqslant \epsilon$. Note that since A is pd, the function $\|\cdot\|$ is indeed a norm.

Theorem 4.1.1 *Let $A \succ 0$ be a $d \times d$ matrix and let $b \in \mathbb{R}^d$ be a vector. Given an accuracy parameter $\epsilon > 0$, GD finds an ϵ-approximate solution after $O(\kappa(A)\log(\|x\|^\star/\epsilon))$ iterations with overall complexity $O((t_A + d)\kappa(A)\log(\|x^\star\|_A/\epsilon))$, where t_A is the runtime of multiplying A with a vector.*

An extremely important observation is that (unlike Gaussian elimination), the GD-based solver exploits sparsity of the matrix A. Namely, we note if A has only $m = o(d^2)$ nonzero entries, then t_A scales linearly with m. Therefore, assuming that $\log(1/\epsilon)$ is negligible, GD is superior to Gaussian elimination if $\kappa_A(t_A + d) = o(d^{2.737})$. In words, GD is superior when A is well-conditioned (and possibly sparse).

Gradient Descent-based solver for linear systems

Starting from $x_0 = 0$, the algorithm GD maintains an approximation, x_t, of the minima of f (4.1), $x^\star = A^{-1}b$, according to the following update rule:

$$x_{t+1} = x_t - \eta_t \nabla f(x_t) = x_t - \eta_t(Ax_t - b) = x_t - \eta_t(A(x_t - x^\star)) \ ,$$

where $\eta_t \in \mathbb{R}_+$, the *step size* at time t, is a parameter that we tune later. Of course, it would be better to move in the direction $I(x_t - x^\star)$ (rather than the direction of the gradient, $A(x_t - x^\star)$), but computing this direction is hard as solving the problem. This point of view suggests that the "closer" is A to the identity matrix, the faster is the convergence of GD to $x^\star$. Indeed, this intuition is affirmed, where the notion of distance from the identity matrix is captured by the condition number.

Theorem 4.1.1 may be proved in various ways. Our technique will lead to a unified treatment of GD's convergence rate in a more general setting. We start by observing that for any two vectors, $x, y \in \mathbb{R}^d$,

$$f(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}(y - x)^\top A(y - x) \ . \tag{4.3}$$

To see this, simply note that right-hand side is the second degree Taylor approximation of f near x. Since f is quadratic, the approximation is accurate, i.e., an equality holds. In particular, since $\nabla f(x^\star) = 0$, we have $f(x) - f(x^\star) = \frac{1}{2}(y - x)^\top A(y - x) = \frac{1}{2}\|y - x\|_A^2$.

We would like to understand what is the impact of relying only on first-order information. Denote $\lambda_i = \lambda_i(A)$ for all i. Since $A \succ 0$, we know from the Courant minmax principle that $\lambda_d\|y - x\|^2 \leqslant \|y - x\|_A^2 \leqslant \lambda_1\|y - x\|^2$. Therefore, we have the following upper and lower bounds on the first order approximation of f near x:

$$D_f(y, x) := f(y) - \big(f(x) + \nabla f(x)^\top (y - x)\big) \in \left[\frac{\lambda_d}{2}\|y - x\|^2, \frac{\lambda_1}{2}\|y - x\|^2\right] \ . \tag{4.4}$$

The function $D_f(y, x)$ is called the Bregman distance with f for the points y, x (note that D is not symmetric, does it does not form a metric). Based on this inequality, we can analyze the progress of GD.

Lemma 4.1.1 *Let $x \in \mathbb{R}^d$, $x_+ = x - \frac{1}{\lambda_1}\nabla f(x)$. Then, $f(x_+) - f(x) \leqslant -\frac{1}{2\lambda_1}\|\nabla f(x)\|^2$.*

Proof By substituting $y = x - \eta\nabla f(x)$ in (4.4), we obtain

$$f(y) - f(x) \leqslant -\eta\nabla f(x)^\top \nabla f(x) + \frac{\lambda_1\eta^2}{2}\|\nabla f(x)\|^2 \ .$$

Optimizing over η yields $\eta^\star = 1/\lambda_1$. Substituting this value in the bound implies the lemma. ■

Next, we relate the magnitude of the gradient to the suboptimality.

Lemma 4.1.2 *For any $x \in \mathbb{R}^d$, $f(x) - f(x^\star) \leqslant \frac{1}{2\lambda_d}\|\nabla f(x)\|^2$.*

Proof Using (4.4), we have

$$f(x^\star) \geqslant f(x) + \nabla f(x)^\top (x^\star - x) + \frac{\lambda_d}{2}\|x^\star - x\|^2 \geqslant \min_{z\in\mathbb{R}^d} f(x) + \nabla f(x)^\top z + \frac{\lambda_d}{2}\|z\|^2$$
$$\underbrace{=}_{z^\star = -\lambda_d^{-1}\nabla f(x)} f(x) - \frac{1}{2\lambda_d}\|\nabla f(x)\|^2 \ .$$

By rearranging, we conclude the bound. ■

The proof of Theorem 4.1.1 is now almost straightforward. Denoting $\Delta_t = f(x_t) - f(x^\star)$, we have

$$\Delta_{t+1} \leqslant \Delta_t - \frac{1}{2\lambda_1}\|\nabla f(x_t)\|^2 = \Delta_t - \frac{2\lambda_d}{2\lambda_1}\Delta_t \leqslant \Delta_t \exp(-\lambda_d/\lambda_1) \ .$$

Therefore, $\Delta_T \leqslant \exp(-T\lambda_d/\lambda_1)\Delta_0$. By rearranging and recalling that $\|x_t - x^\star\|_A^2 = 2\Delta_T$, we conclude the claimed convergence rate. The runtime per iteration is $t_A + d$. We thus conclude Theorem 4.1.1.

One disadvantage of the suggested implementation is that the step size depends on $\lambda_1(A)$, which is usually unknown. In Exercise (4.1.3) we suggest a different strategy that leads to an identical bound.

Degenerate linear systems

As we shall see soon, in applications it is often the case that $A \in \mathbb{S}_+^d$ (i.e., A is a $d \times d$ positive semidefinite matrix. In this case, A might be singular) and b belongs to the column space of A (hence the system $Ax = b$ is still solvable). Does Theorem 4.1.1 hold in this case? Certainly, in our situation the statement in Theorem 4.1.1 is erroneous. First, since $\lambda_d(A)$ might be zero, the condition number $\lambda_1(A)/\lambda_d(A)$ is not defined. Furthermore, since there might be no unique solution, $x^\star$ is not well defined. Intuitively, it seems that the we can remedy the situation by restricting ourselves to the column space of A. Indeed, it can be seen that for any solution $\bar{x}$ to

$Ax = b$, the projection of $\bar{x}$ onto the column space of A is also a solution. Moreover, all of these projections coincide with the vector $A^\dagger b$, which we denote by $x^\star$ (see Exercise (4.1.2)). Furthermore, it is easily seen that the iterates x_t maintained by GD lie in the column space of A. Finally, we redefine the condition number of A as

$$\kappa(A) = \lambda_1(A)/\lambda_{\text{rank}}(A) \ . \tag{4.5}$$

We are now able to show the following generalization of Theorem 4.1.1.

Theorem 4.1.2 *Let $A \succeq 0$ be a $d \times d$ matrix and let $b \in \mathbb{R}^d$ be a vector in the column space A. Denote $x^\star = A^\dagger b$. Given an accuracy parameter $\epsilon > 0$, GD finds an ϵ-approximate solution[1] after $O(\kappa(A)\log(\|x\|_A^\star/\epsilon))$ iterations with overall complexity $O((t_A + d)\kappa(A)\log(\|x^\star\|_A/\epsilon))$, where $\kappa(A)$ is defined in Equation (4.5) and t_A is the runtime of multiplying A with a vector.*

The proof is outlined in Exercise (4.1.4*).

Application: Least Squares

The method of Least Squares is arguably the simplest and the most popualar approach for regression analysis in statistics and sachine searning. The associated optimization problem is defined as following. We are given as an input a sequence of vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ together with a corresponding sequence of labels $y_1, \ldots, y_n \in \mathbb{R}$. The objective is given by

$$\min \left\{ L(w) = \frac{1}{2n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 : w \in \mathbb{R}^d \right\} \ . \tag{4.6}$$

Note that an equivalent problem is given by

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} w^\top A w - w^\top b \ .$$

where $A = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$ and $b = \frac{1}{n}\sum_{i=1}^n y_i x_i$. As we know, this problem is equivalent to the system $Aw = b$. Clearly, $A \succeq 0$, and since the column space of A is equal to $\text{span}\{x_1, \ldots, x_n\}$, b belongs to the column space of A. We conclude that Theorem 4.1.2 holds. In particular, since $L(w) - L(w^\star) = \frac{1}{2}\|w - w^\star\|^2$, we conclude that GD finds $\bar{w}$ with $L(\bar{w}) \leqslant \min_{w \in \mathbb{R}^d} L(w) + \epsilon$ in time $O((t_A + d)\kappa(A)\log(\|w^\star\|/\epsilon))$, where $w^\star = A^\dagger b$.

[1] The definition of ϵ-approximate minima refers to the function $\|\cdot\|_A$. Note that when A is only assumed to be positive semidefinite, $\|\cdot\|_A$ is a semi-norm rather than a norm. However, when restricted to the column space of A, $\|\cdot\|_A$ forms a norm. See Exercise (4.1.1).

Exercises

Exercise 4.1.1 *(**Mahalanobis norm**) Let $A \succ 0$ be a $d \times d$ matrix.*

1. *Show that the bilinear form, $\langle \cdot, \cdot \rangle_A : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, defined by $\langle x, y \rangle_A := x^\top A y$ forms an inner product. Conclude that the function $z \mapsto \sqrt{z^\top A z}$ is a norm.*[2]

2. *Recall that the quadratic function defined in (4.1) satisfies $f(\bar{x}) - f(x^\star) = \|\bar{x} - x^\star\|_A$. Conclude that $x^\star$ is (indeed) unique.*

3. *Show that when $A \succeq 0$, $\|\cdot\|_A$ forms a semi-norm, i.e., it satisfies the triangle inequality and $\|\alpha x\|_A = \alpha \|x\|$ for every scalar $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^d$ (however, there might be $x \neq 0$ with $\|x\|_A = 0$). Furthermore, show that when restricted to the column space of A, $\|\cdot\|_A$ forms a norm.*

Exercise 4.1.2 *(**The Moore-Penrose pseudoinverse**) Let $B \in \mathbb{R}^{d \times n}$ be any matrix of rank r and denote its SVD by $B = \sum_{i=1}^r \sigma_i u_i v_i^\top$. The pseudoinverse of B is defined by $B^\dagger = \sum_{i=1}^r \sigma_i^{-1} v_i u_i^\top$.*

1. *Show that $BB^\dagger$ forms a projection matrix onto the column space of B (i.e., $(BB^\dagger)^2 = BB^\dagger$ and the range of $BB^\dagger$ coincides with the range of B). Similarly, show that $B^\dagger B$ is a projection matrix onto the row space of B (i.e., $(B^\dagger B)^2 = B^\dagger B$ and the range of $B^\dagger B$ coincides with the range of $B^\top$). .*

2. *Let A be an $n \times d$ matrix and let $b \in \mathbb{R}^n$ be a vector in the column space of A. Show that $A^\dagger b$ solves the system $Ax = b$. Moreover, if $\bar{x}$ is any solution to $Ax = b$, then its projection to the column space of A coincides with $A^\dagger b$.*

3. *Let $A \succ 0$ be a $d \times d$ matrix of rank r and assume that let $A = \sum_{i=1}^r \lambda_i u_i u_i^\top$ is the eigendecomposition of A. Show that $A^\dagger = \sum_{i=1}^r \lambda_i^{-1} u_i u_i^\top$.*

Exercise 4.1.3 *Consider a variant of GD whose step size is defined by $\eta_t = \frac{\|\nabla f(x_t)\|^2}{\|\nabla f(x_t)\|_A^2}$. Show that Theorem 4.1.1 holds with respect to this variant. (Hint: What is the relation between η_t and $1/\lambda_1$?)*

Exercise 4.1.4* *Prove Theorem 4.1.2. (Hint: Denote by $U \in \mathbb{R}^{d \times r}$ the matrix whose i-th column is the i-th leading eigenvector of A. Consider the quadratic problem*

$$\min_{x \in \mathbb{R}^r} \left\{ \tilde{f}(x) = \frac{1}{2} x^\top \tilde{A} x - \tilde{b}^\top x \right\} , \tag{4.7}$$

where $\tilde{A} = U^\top A U$ and $\tilde{b} = U^\top b$.)

[2] You may prove this fact directly. Instead, you can show that for every inner product $\langle \cdot, \cdot \rangle$, the function $z \mapsto \sqrt{\langle z, z \rangle}$.

Exercise 4.1.5 *Let $x_1, \dots, x_n \in \mathbb{R}^d$. Show that the column space of $A = \sum_{i=1}^n x_i x_i^\top$ is equal to* $\operatorname{span}\{x_1, \dots, x_n\}$.

Exercise 4.1.6 *Consider the objective associated with Ridge regression:*

$$\min \left\{ L(w) = \frac{1}{2n} \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \frac{\lambda}{2} \|w\|^2 : \ w \in \mathbb{R}^d \right\} ,$$

where $x_1, \dots, x_n \in \mathbb{R}^d$, $y_1, \dots, y_n \in \mathbb{R}$ and $\lambda > 0$ is a regularization parameter.

1. *Show that Ridge regression can be written as a standard least squares problem (4.6).*
2. *Derive an upper bound on the convergence rate and the runtome of GD when applied to the Ridge regression objetive.*

4.2 Acceleration using Conjugate Gradient

In view of Theorem 4.1.1 and Theorem 4.1.2, a natural question which arises is whether we can have a better complexity? In particular, while the dependence on $1/\epsilon$ seems satisfactory, does the linear dependence on the condition number can be improved? In this section we present the Conjugate Gradient (CG) method that leads to a quadratic saving in terms of the dependence on the condition number while having the same computational complexity per iteration. We also discuss the optimality of this method. Our exposition highlights the connection to the area of approximation theory. Namely, we associate every first order algorithm with a matrix polynomial. By these means, the study of the convergence rate is essentially reduced to the study of extremal properties of such polynomials.

A simple induction that when applying GD to a convex quadratic form, its t-th iterate, x_t, belongs to the set $\mathcal{K} := \mathrm{span}\{b, Ab, \ldots, A^{t-1}b\}$, named the *Krylov subspace* of order t. However, the vector x_t may not attain the minimal value of the quadratic form f over $\mathcal{K}$. On the contrary, the Conjugate Gradient algorithm does ensure this property. In the next two parts we address the two following question. (1) How does CG find x_t efficiently? (2) How does this property imply the claimed quadratic saving?

Computing conjugate directions efficiently

Let $\{v_0, \ldots, v_{t-1}\}$ be any basis for the Krylov subspace of order t. We would like to find the minimizer of the quadratic form f over $\mathcal{K}$. Since $x^\star = A^{-1}b$, we consider the following minimization problem:

$$\min_{\alpha_0,\ldots,\alpha_{t-1}} \|A^{-1}b - \sum_{i=1}^{t} \alpha_i v_i\|_A^2 = \|x^\star\|_A^2 - 2\sum_{i=0}^{t-1} \alpha_i v_i^\top b + \sum_{i=0}^{t-1}\sum_{j=0}^{t-1} \alpha_i \alpha_j \langle v_i, v_j\rangle_A$$
$$\|x^\star\|_A^2 - 2\alpha V^\top b + \alpha^\top Z \alpha \ . \tag{4.8}$$

where $\langle v_i, v_j\rangle = v^\top A v_j$, $V \in \mathbb{R}^{d\times t}$ is the matrix whose columns are $v_0, \ldots, v_{t-1}$ and $Z = V^\top A V$. The optimal solution is

$$\alpha^\star = Z^{-1} V^\top b \ , \tag{4.9}$$

Therefore, computing $\alpha^\star$ requires the inversion of A, which is clearly undesired. This should not surprise as since we can not expect obtaining a simpler problem by using an arbitrary change of basis. However, if the $v_0, \ldots, v_{t-1}$ are A-orthogonal (a.k.a. A-conjugate), i.e. $\langle v_i, v_j\rangle_A = 0$ for all $i \neq j$, then the problem (4.9) becomes much easier. Namely, a simple exercise shows that the optimal solution is given by

$$(\forall i) \quad \alpha_i^\star = \frac{v_i^\top b}{\|v_i\|_A^2} \ . \tag{4.10}$$

Computing each of the coefficients α_i can be carried out in time $t_A + d$. Thus, if we compute α_{i-1} at the i-th iteration, we would have the same complexity as GD, but now we also ensure that x_t is the optimal vector in $\mathcal{K}$. In particular, an exact solution is found after at most d iterations.

Naturally, the next question is how can we compute an A-orthogonal set efficiently. A naive method is to start with some arbitrary independent set of vectors (e.g., the standard basis) and apply the Gram-Schmidt process with respect to $\langle \cdot, \cdot \rangle_A$. In order to generate t orthonormal vectors, we need to compute t^2 inner products w.r.t. A, so this preprocessing scheme would run in time $O(t^2(t_A + d))$. For large t, the resulted scheme is not efficient. We next describe a rather clever iterative method that computes A-orthogonal vectors on-the-fly (rather than as a preprocessing step).

Starting from $v_0 = b$, we proceed inductively, where at time t, we compute $\tilde{v}_t = Av_{t-1}$ and then A-orthogonalize $\tilde{v}_t$ with respect to $v_0, \ldots, v_{t-1}$

$$v_t = \tilde{v}_t - \sum_{s=0}^{t-1} \frac{\langle \tilde{v}_t, v_s \rangle_A}{\langle v_s, v_s \rangle_A} v_s \ . \tag{4.11}$$

It follows that $\mathrm{span}\{v_0, \ldots, v_t\} = \mathrm{span}\{b, \ldots, A^t b\}$ and $\{v_0, \ldots, v_t\}$ is A-orthonormal[3]. Crucially, it also follows that for every s, $Av_s \in \mathrm{span}\{v_0, \ldots, v_{s+1}\}$. Hence, since A is symmetric, we have

$$\langle \tilde{v}_t, v_s \rangle_A = v_{t-1}^\top A A v_s = \langle Av_s, v_{t-1} \rangle_A = 0 \quad \text{for all } s < t-2 \ .$$

Hence, all but the two last terms in the RHS of (4.11) are canceled. As a result, the A-orthogonalization of v_t takes only $O(t_A + d)$. We leave it as an exercise to write a pseudo-code of CG.

Analysis via Chebyshev polynomials

We next show how the optimality over Krylov subspaces implies the quadratic saving in the dependence on the condition number. The analysis relies on a beautiful connection to approximation theory.

Every vector in the Krylov subspace $\mathcal{K} = \{b, \ldots, A^{t-1}b\}$ can be (uniquely) written as $p(A)b$, where $p(x)$ is a polynomial of degree t. Clearly, this correspondence is also surjective, i.e., every polynomial of degree $t-1$ induces a unique vector in $\mathcal{K}$. Denote the set of polynomials of degree $t-1$ by Σ_{t-1}. By the optimality of x_t over $\mathcal{K}$, we conclude that

$$\|x^\star - x_t\|_A^2 = \min_{p \in \Sigma_{t-1}} \|x^\star - p(A)b\|_A^2 = \|(I - p(A)A)x^\star\|_A^2 \leqslant \|x^\star\|_A^2 \|I - p(A)A\|_2^2 \ , \tag{4.12}$$

where $\|I - p(A)A\|_2$ denotes the spectral norm of $I - p(A)A$ (the inequality in (4.12) is left as an exercise). Given an accuracy parameter $\epsilon' > 0$, we define $\epsilon = \epsilon' / \|x^\star\|_A^2$

[3]More precisely, the set $\{v_0, \ldots, v_t\}$ is independent unless the vectors $b, \ldots, A^t b$ are linearly dependent but in this case we could stop earlier (see Exercise (4.2.1)).

and aim at finding a small as possible $t \in \mathbb{N}$ for which there exists a degree-$(t-1)$ polynomial p which satisfies $\|I - p(A)A\|^2 \leqslant \epsilon$. Equivalently,[4] we can look for a degree-t polynomial q which evaluates 0 to 1 for which $\|q(A)\|^2 \leqslant \epsilon$. Since A is symmetric and, thus, diagonalizable, for every polynomial q,

$$\|q(A)\|^2 \leqslant \max\{q(x)^2 : x \in \{\lambda_1, \ldots, \lambda_d\}\} \ .$$

where $\lambda_i = \lambda_i(A)$. Before proceeding, note that the polynomial $q(x) = \prod_{i=1}^{d}(1 - x/\lambda_i)$ evaluates 0 to 1 and satisfies $\|q(A)\| = 0$. This gives another proof for the fact that CG converges after at most d iterations.

As a first attempt, consider the polynomial

$$q_0(x) = \left(1 - \frac{2x}{\lambda_1 + \lambda_d}\right)^t \ . \tag{4.13}$$

It can be seen that the maximum value attained by q_0 over the interval $[\lambda_d, \lambda_1]$ is $\left(\frac{\kappa(A)-1}{\kappa(A)+1}\right)^t \leqslant \exp^{-2t/(\kappa(A)+1)}$. For $\epsilon > 0$, by letting

$$t_0 = \lceil \kappa \log(1/\epsilon) \rceil \ ,$$

we obtain that $\|q(A)\|_A^2 \leqslant \epsilon$. We thus recover the upper bound of GD. The reason for the improved rate of CG follows from the following powerful result.

Theorem 4.2.1 *For any $t \in \mathbb{N}$, there exists a polynomial $p_{t,d}$ of degree $d = \lceil \sqrt{2t \log(2/\epsilon)} \rceil$ which satisfies*

$$\sup_{x \in [-1,1]} |x^t - p_{t,d}(x)| \leqslant \epsilon \ .$$

The proof of the theorem relies on Chebyshev polynomials, which are ubiquitous in numerical optimization. We sketch the proof in the next part. We now exploit this result in order to deduce a better bound for CG. Note that as x ranges over $[0, \lambda_1 + \lambda_d]$, $z(x) := 1 - 2x/(\lambda_1 + \lambda_d)$ ranges over the interval $[-1, 1]$. Based on Theorem 4.2.1, we conclude the existence of a polynomial $p_{t_0,d}$ of degree at most $d = \lceil \sqrt{2t_0 \log(2/\epsilon)} \rceil = O(\sqrt{\kappa} \log(1/\epsilon))$ which approximates $q_0(z) = z^{t_0}$ up to an error ϵ over the interval $[-1, 1]$. Therefore, the polynomial $q_1(x) = p_{s,d}(z(x))$ approximates $q_0(x)$ up to an error ϵ over $[0, \lambda_1 + \lambda_d]$. Since $q_0(0) = 1$, it follows that the polynomial $q(x) = q_1(x)/q_1(0)$ evaluates 0 to 1 and satisfies $\sup_{z \in [-1,1]} |q(z) - q_0(z)| \leqslant \epsilon/(1-\epsilon)$ which is smaller than 2ϵ providing that $\epsilon < 1/2$. Finally, since $\sup_{z \in [-1,1]} |q_0(z)| \leqslant \epsilon$ for all $z \in [-1, 1]$, it follows that $\sup_{z \in [-1,1]} |q(z)| \leqslant 3\epsilon$. This leads to the promised speedup.

Theorem 4.2.2 *For any $A \in \mathbb{S}^d_{++}$ and $\epsilon > 0$, CG finds an ϵ-approximate solution to the system $Ax = b$ after $O(\min\{d, \sqrt{\kappa(A)} \ln(1/\epsilon)\})$ iterations and its overall complexity is $O((d + t_A) \cdot \min\{d, \sqrt{\kappa(A)} \ln(1/\epsilon)\})$.*

[4] Note that $I - p(A)A$ is a polynomial of degree t. Also, every polynomial of degree t that evaluates 0 to 1 can be written in such a way.

Optimality of the Bounds

The convergence rate of CG is optimal in the following sense. First, we define a first-order algorithm (for minimization of a quadratic positive definite objective) to be any method that iteratively maintains a solution $x_t \in \mathbb{R}^d$ while having an access to a *first-order* oracle that given a point $x \in \mathbb{R}^d$, returns the pair $(f(x), \nabla f(x))$. We also assume for simplicity that the x_t lies in the span of the observed gradients (and $x_0 = 0$).[5] This is the only information given to the method regarding the function f. In particular, the identity of A and b is not known in advance. The lower bound is as follows: for any $0 < \lambda_d \leqslant \lambda_1$, there exists a matrix $A \in \mathbb{S}^d_{++}$ whose eigenvalues lie in the range $[\lambda_d, \lambda_1]$, and for which any first-order algorithm needs at least $\Omega(\min\{d, \sqrt{\lambda_1/\lambda_d}\log(1/\epsilon)\})$ iterations in order to converge to an ϵ-approximate minimizer. We outline the proof in the exercises.

Appendix

Chebyshev polynomials: proof sketch of Theorem 4.2.1

For a nonnegative integer d, we define the Chebyshev polynomial of degree d in a recursive manner:

$$T_0(x) = 1, \; T_1(x) = x \; .$$

$$T_d(x) = 2xT_{d-1}(x) - T_{d-2}(x) \quad \text{for } d \geqslant 2 \; .$$

By rearranging, we obtain the relation

$$x\, T_d(x) = (T_{d+1}(x) + T_{d-1}(x))/2 \; . \tag{4.14}$$

For convenience, we define $T_d(x) = T_{|d|}(x)$ for all $d \in \mathbb{N}\backslash\mathbb{Z}$ and it is easy to verify that the above recursive definition holds for all $d \in \mathbb{Z}$. We now explore some important properties of these polynomials. First, an exercise reveals that for any $\theta \in \mathbb{R}$, $T_d(\cos(\theta)) = \cos(d\theta)$. This property has several important consequences, but for now we only exploit this fact to deduce that $|T_d(x)| \leqslant 1$ for all $x \in [-1, 1]$.

We next sketch the proof of Theorem 4.2.1 by relating the monomial x^t to a weighted sum of Chebyshev polynomials, where the weights are induced by a random walk. Consider a random walk which starts at 0 and at each time step i make the moves ± 1 with equal probability. We associate with the moves a sequence of (i.i.d.) random variables $Y_1, Y_2, \ldots$. The position after t steps is denoted by $Y_{1:t} = \sum_{i=1}^t Y_i$. We use this process to derive an approximation for z^t. Let us begin by observing the following consequence of Equation (4.14).

Lemma 4.2.1 *For any $t \in \mathbb{N}$, $\mathbb{E}_{Y_1,\ldots,Y_t}[T_{Y_{1:t}}(x)] = x^t$.*

[5] These assumptions can be avoided, see bibliographic remarks.

In words, x^t is a weighted sum of monomials of a lower degree, where the weight of each monomial is proportional to the probability that a random walk with t steps would end at the corresponding position. Clearly, the expected position is 0. Moreover, we can use concentration of measure in order to approximate x^t using a polynomial of degree roughly $\sqrt{t}$. Namely, based on Chernoff bound, we know that the probability that $|Y_{1:t}| \geqslant d = \sqrt{2s\log(2/\epsilon)}$ is at most ϵ. By using the fact that $|T_d(x)| \leqslant 1$ for all $x \in [-1,1]$, it can be shown that the polynomial

$$p_{t,d}(x) = \mathbb{E}_{Y_1,\ldots,Y_t}[T_{Y_{1:t}} \cdot \mathbf{1}_{[Y_{1:t} \leqslant d]}] \ ,$$

which is simply a degree-d truncation of x^t, approximates x^t up to an error ϵ. Filling the missing parts of the proof is left as an exercise.

Exercises

Exercise 4.2.1 *Let $A \in \mathbb{S}^d_{++}$ and let $b \in \mathbb{R}^d$. Assume that for some t, the set $\{b, Ab, \ldots, A^t b\}$ is linearly dependent. Conclude that the solution $x^\star = A^{-1}b$ lies in $\operatorname{span}\{b, Ab, \ldots, A^{t-1}b\}$. In particular, conclude that CG finds an exact solution to the system $Ax = b$ after at most d iterations.*

Exercise 4.2.2 *Write a pseudo-code of the Conjugate Gradient method. It should be made explicit that the only information available to the method on the function $f_{A,b}$ is via an oracle which given a vector x, returns $\nabla f_{a,b}(x)$ (in particular, A and b are not known in advance).*

Exercise 4.2.3 *Prove the inequality in (4.12).*

Exercise 4.2.4 *Extend the analysis of CG to the case where A is positive semidefinite and b lies in the column space of A.*

Exercise 4.2.5* ***(Better bound for clustered eigenvalues)*** *Prove the following bound of CG: Let $A \in \mathbb{S}^d_{++}$, b that lies in column space of A and $\epsilon > 0$. Suppose that all but m of the eigenvalues of A line in the range $[a,b]$. Then CG finds an ϵ-approximate minimizer after at most $m + O(\sqrt{b/a}\log(\|x\|_A/\epsilon))$ iterations.*

Exercise 4.2.6

1. *For any $\theta \in \mathbb{R}$, $T_d(\cos(\theta)) = \cos(d\theta)$.*
2. *Conclude that $\max_{x\in[-1,1]} |T_d(x)| = 1$ and the extreme value is obtained by $x_i = \cos(j\theta/d)$ for $j \in \{0,\ldots,d\}$, where the sign of $T_d(x)$ alternates at these points.*
3. *Draw $T_d|_{[-1,1]}$ for $d \in \{0,\ldots,4\}$.*

Exercise 4.2.7* *Prove Lemma 4.2.1 and complete the proof of Theorem 4.2.1. (Hint: Note that* $\sup_{x\in[-1,1]} |x^t - p_{t,\hat{d}}(x)| = \sup_{x\in[-1,1]} |\mathbb{E}_{Y_1,\ldots,Y_t}[T_{Y_{1:t}}(x)\cdot \mathbf{1}_{[Y_{1:t}>\hat{d}]}]|$*. What can you say about* $\mathbb{E}_{Y_1,\ldots,Y_t}[\mathbf{1}_{[Y_{1:t}>\hat{d}]}]$*?)*

Exercise 4.2.8* ***(Lower bound for quadratic optimization in the black-box model)*** *We prove the lower bound stated at the end of the lecture. To simplify the proof we consider an infinite-dimensional space (and establish the dependence on the condition number). We also assume that the first-order method starts from the point* $x_0 = 0$ *and always maintain a solution in the span of the previous gradients.*

Let $0 < \alpha \leqslant \beta$ *be two scalars and denote* $\kappa = \frac{\beta}{\alpha}$*. Let* $A : \ell_2 \to \ell_2$ *be*[6] *an infinite tridiagonal matrix with* 2 *on the diagonal and* -1 *on the upper and lower diagonals (i.e.,* $A_{i,j} = -1$ *for all* i, j *such that* $|i-j| = 1$*). Consider the function*

$$f(x) = \frac{\alpha(\kappa-1)}{8}(x^\top A x - 2e_1^\top x) + \frac{\alpha}{2}\|x\|^2 \ .$$

Let $\mathcal{A}$ *be an algorithm whose access to the function* f *is via the black-box model and denote the iterates of* $\mathcal{A}$ *by* $x_0, x_1, \ldots$*.*

1. *Show that* $0 \preceq A \preceq 4I$*.*
2. *Show that finding* $\bar{w}$ *that satisfies* $f(\bar{w}) \leqslant \min_{w\in\mathbb{R}^d} f(w) + \epsilon$ *is equivalent to solving a linear system with a matrix* A *that satisfies* $\kappa(A) = \kappa$*.*
3. *Show that for for every* t*,* $x_t \in \mathrm{span}(\{e_1, \ldots, e_t\})$*.*
4. *Show that the optimal solution, denoted* $x^\star$*, satisfies the following infinite set of equations*

$$1 - \frac{2(\kappa+1)}{\kappa-1}x_1^\star + x_{i+2}^\star = 0$$
$$x_{i-1}^\star - \frac{2(\kappa+1)}{\kappa-1}x_i^\star + x_{i+1}^\star = 0, \quad i \geqslant 2$$

5. *Conclude that* $x_i^\star = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i$*.*
6. *Show that* $f(x_t) - f(x^\star) \geqslant \frac{\alpha}{2}\|x_t - x^\star\|^2 \geqslant \frac{\alpha}{2}\sum_{i=t+1}^{\infty} x_t^{\star 2}$*.*
7. *Conclude that unless* $t = \Omega(\sqrt{\kappa}\ln(\alpha/\epsilon))$*,* $f(x_t) - f(x^\star) > \epsilon$*.*

Exercise 4.2.9 **TODO: Lanczos for a general symmetric matrix**

[6] The space ℓ_2 is the inner-product space consisting of all sequences $(x_1, x_2, \ldots) \in \mathbb{R}^\infty$ such that $\sum_{i=1}^{\infty} x_i^2 < \infty$.

4.3 Unconstrained Smooth Convex Optimization

In the previous section we proved an upper bound on the convergence rate of GD for quadratic functions. A careful reader would observe that the analysis holds for a wider class of objective functions; Namely, during the analysis we have not relied on the fact that the second-order Taylor approximation is accurate (4.3). Instead, we derived lower and upper bounds on the second-order term (4.4) and relied solely on these bounds during the analysis. We now exploit this fact in order to significantly extend the scope of our results. For simplicity, we focus now on unconstrained optimization. We always assume that minimizers exist.

Definition 4.3.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. We say that f is β-smooth if the gradient of f is β-Lipschitz, i.e., for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f(x) - f(y)\| \leqslant \beta \|x - y\| \ .$$

Lemma 4.3.1 *If $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth, then*

$$D_f(y, x) = f(y) - \big(f(x) + \nabla f(x)^\top (y - x)\big) \leqslant \frac{\beta}{2}\|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^d \ . \tag{4.15}$$

Furthermore, if f is twice continuously differentiable, then both the β-smoothness of f and Equation (4.15) are equivalent to the condition $\lambda_1(\nabla^2 f(x)) \leqslant \beta$ for all $x \in \mathbb{R}^d$.

Definition 4.3.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. We say that f is α-strongly convex if for all $x, y \in \mathbb{R}^d$ and $v \in \partial f(x)$,*

$$f(y) - \big(f(x) + v^\top (y - x)\big) \geqslant \frac{\alpha}{2}\|y - x\|^2 \ .$$

(In particular, if f is differentiable at x, then $D_f(y, x) \geqslant \frac{\alpha}{2}\|y - x\|^2$ for all $y \in \mathbb{R}^d$.)

Note that a strongly convex function is strictly convex and therefore, minimizers of strongly convex functions are unique. An important 1-strongly convex function is the Tikhonov regularization $\frac{1}{2}\|w\|^2$ (it is also 1-smooth). If f is α-strongly convex function and g is μ-strongly convex, then $f + g$ is $\alpha + \mu$-strongly convex function. Also, for any positive scalar $\lambda > 0$, λf is $\lambda\alpha$-strongly convex. If g is only assumed to be convex, then the sum $f + g$ is α-strongly convex.

Lemma 4.3.2 *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Then f is α-strongly convex if and only if $\lambda_d(\nabla^2 f(x)) \geqslant \alpha$ for all x.*

The quadratic objective we considered in the previous section was λ_1-smooth and λ_d-strongly convex. As we mentioned above, the analysis in the previous section merely relied on these two facts. We thus conclude the following important generalization of

Theorem 4.1.1. The condition number of a β-smooth and α-strongly convex function f is defined by

$$\kappa(f) = \beta/\alpha \ .$$

In our context, $\bar{x}$ is an ϵ-approximate minimizer for a function $f : \mathbb{R}^d \to \mathbb{R}$ if $f(\bar{x}) \leqslant f(x) + \epsilon$ for all x.

Theorem 4.3.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a β-smooth and α-strongly convex function that admits a (unique) minimizer. Given an accuracy parameter $\epsilon > 0$, GD finds an ϵ-approximate solution after $O(\kappa(A)\log(\Delta_0/\epsilon))$ iterations, where $\Delta_0 = f(x_0) - f(x^\star)$.*

Application: Regularized Loss Minimization for Linear Prediction

We consider a regularized risk minimization (RLM) objective of the form:

$$\min \left\{ L(w) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(w^\top x_i) + \frac{\lambda}{2}\|w\|^2 : w \in \mathbb{R}^d \right\} \ , \tag{4.16}$$

where $x_1, \ldots, x_n \in \mathbb{R}^d$ are vector instances, $\phi_1, \ldots, \phi_n$ are univariate convex functions which are usually associated with a sequence of labels, $y_1, \ldots, y_n \in \mathbb{R}$, and $\lambda \geqslant 0$ is a (Tikhonov) regularization parameter. For example, the standard Ridge regression objective is obtained by defining $\phi_i(z) = (y_i - z)^2$. Another important example is logistic Ridge regression, corresponding to $\phi_i(z) = \log(1 + \exp(-y_i z))$. We assume now that

$$\text{each } \phi_i \text{ is } \beta\text{-smooth and } \alpha\text{-strongly convex} \ . \tag{4.17}$$

(where α might be zero)[7]. For example, it can be verified that the square loss is 1-smooth and 1-strongly convex and the logistic loss is 1/4-smooth and 0-strongly convex. We would like to analyze the smoothness and the strong convexity parameters of L. For simplicity, we consider the case where each ϕ_i is twice differentiable. Using the chain rule, we see that the Gradient and the Hessian of L are given by

$$\nabla L(w) = \frac{1}{n}\sum_{i=1}^{n} \phi'(w^\top x_i)x_i + \lambda w, \quad \nabla^2 L(w) = \frac{1}{n}\sum_{i=1}^{n} \phi''(w^\top x_i)x_i x_i^\top + \lambda I \ .$$

Denoting $C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$, we see that the largest eigenvalue of the Hessian at any point w is at most $\beta\lambda_1(C) + \lambda$ and the smallest eigenvalue is at least $\alpha\lambda_d(C) + \lambda$ (where both α and $\lambda_d(C)$ can be zero). This immediately implies that

$$f \text{ is } (\beta\lambda_1(C) + \lambda)\text{-smooth and } (\alpha\lambda_d(C) + \lambda)\text{-strongly convex} \ . \tag{4.18}$$

[7] We use the following convention: a convex but not necessarily strongly convex function is said to be 0-strongly convex function.

In fact, the above is true even when the ϕ_i's are not twice differentiable. Here is one concrete implication: note that $\lambda_1(C) \leqslant \max_{i\in[n]} \|x_i\|^2 =: R^2$. Hence, the convergence rate of GD when applied to logistic ridge regression is $O(R^2\lambda^{-1}\log(1/\epsilon))$. Since every gradient computation costs $O(nd)$, the overall runtime is $O(R^2\lambda^{-1}nd\log(1/\epsilon))$.

Nesterov's Accelerated Gradient Descent

While we were able to extend the analysis of GD from the quadratic case to the smooth and strongly convex case, no such extension is known for CG. Fortunately, in 1983, Nesterov came up with the Accelerated Gradient Descent (AGD) method which similarly to CG, enjoys a quadratic saving in terms of the dependence on the condition number. The modification is simple: the update rule of AGD forms a linear combination of the current and the previous updates of GD. While the modification is simple, the intuition behind it is not transparent. Indeed, several recent works attempt to derive alternative explanations for AGD (or closer versions of AGD). We know from the previous lecture that the bound of AGD is optimal w.r.t. first-order methods (indeed, quadratic functions form a subclass of smooth and strongly convex functions).

Reducing smooth convex optimization to smooth and strongly convex optimization

We now consider the case where $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth and convex but not necessarily strongly convex (or alternatively, the strong convexity parameter is tiny). In this case GD has the following convergence rate.

Theorem 4.3.2 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and β-smooth function and let $x^\star$ be a minimizer of f. Suppose that we run GD with step size $\eta = \beta^{-1}$. Then, for any $\epsilon > 0$, after at most $\frac{2\beta\|x^\star\|^2}{\epsilon} + 1$ iterations, GD arrives at an ϵ-approximate minimizer.*

Instead of directly proving that GD enjoys the claimed complexity, we propose a similar algorithm with slightly worse convergence rate. Our strategy is to reduce the smooth case to the smooth and strongly convex case. Specifically, instead of directly applying GD to the function f, we will apply it to the function $\tilde{f}(x) = f(x) + \frac{\lambda}{2}\|x\|^2$. Note that the function $\tilde{f}$ is $(\beta+\lambda)$-smooth and λ-strongly convex. As the convergence rate of GD scales with the ratio $(\beta+\lambda)/\lambda$, the larger is λ, the faster is the convergence. However, we should not perturb the function f too much. Namely, λ should be small enough such that an approximate minimizer to $\tilde{f}$ would also form an approximate minimizer of f. Let us formalize this requirement. Suppose that $\hat{x}$ is

an $\epsilon/2$-approximate minimizer to $\tilde{f}$. Letting $x^\star$ be any minimizer of f, we have

$$f(\hat{x}) = \tilde{f}(\hat{x}) - \frac{\lambda}{2}\|\hat{x}\|^2 \leqslant \tilde{f}(x^\star) + \epsilon/2 - \frac{\lambda}{2}\|\hat{x}\|^2 = f(x^\star) + \frac{\lambda}{2}\|x^\star\|^2 + \epsilon/2 - \frac{\lambda}{2}\|\hat{x}\|^2$$
$$\leqslant f(x^\star) + \epsilon/2 + \frac{\lambda}{2}\|x^\star\|^2 \ .$$

Therefore, by choosing $\lambda = \epsilon\|x^\star\|^{-2}$, we obtain ϵ-approximation for the original objective. Note that runtime of this strategy yields the bound $O\left(\frac{\beta\|x^\star\|^2}{\epsilon}\log(\Delta_0/\epsilon)\right)$ which is worse than the stated bound by factor $\log(\Delta/\epsilon)$. We can avoid this extra dependence by applying the above strategy more carefully (see Exercise (4.3.5*)).

From algorithmic point of view, the reduction approach has several disadvantages. First, $x^\star$ is not known so we need to estimate a bound on its norm. In machine learning applications, this often requires us to tune λ (e.g., via cross validation). Another disadvantage is that the accuracy should be fixed before running the algorithm (however, this is not the case in Exercise (4.3.5*)). However, the reduction strategy is quite elegant and general (see next subsection). It allows us to avoid spending several lectures on analyzing the rate of first-order methods for each class of of functions. Furthermore, it sheds a light on the relation between different classes of functions.

Acceleration for The Smooth Case

A slight modification to AGD yields the bound $O\left(\frac{\beta\|x^\star\|}{\sqrt{\epsilon}}\right)$ for the β-smooth case. Thus, we obtain an improvement of factor $1/\sqrt{\epsilon}$. This bound is optimal w.r.t. first-order methods. It is an easy exercise to obtain a slightly weaker bound by reducing to the smooth and strongly convex case and applying AGD. One can show that this bound is optimal w.r.t. first-order methods.

Exercises

Exercise 4.3.1 *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is twice continuously differentiable. Prove the equivalence stated in Lemma 4.3.1.*

Exercise 4.3.2* *Prove the first part of Lemma 4.3.1. (Hint: Consider the function $\phi(t) = f(x + t(y - x))$. Use the fundamental theorem of calculus together with the Cauchy-Schwarz inequality.)*

Exercise 4.3.3 *Prove Lemma 4.3.2.*

Exercise 4.3.4
Consider an RLM objective where each ϕ_i is β-smooth and α-strongly convex.

1. *Prove (4.18).*

2. *Deduce the bound on the convergence rate of GD for the following cases:*
   (a) *Regularized logistic regression.*
   (b) *Regularized squared hinge-loss:* $\phi_i(z) = \max\{0, (1 - y_i z)^2\}$.
3. *Consider the case where* $\alpha > 0$ *but* $\lambda_d(C) = 0$. *Prove that the restriction of the function* L *to the subspace spanned by* $\{x_1, \ldots, x_n\}$ *is* $\alpha\lambda_r(C) + \lambda$, *where* r *is the rank of* C. *Deduce a bound on the convergence of GD for this case.*

Exercise 4.3.5* *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a convex and* β*-smooth function and let* $x^\star$ *be a minimizer of* f. *In this question we use the reduction technique in order to suggest an algorithm that attains the stated in Theorem 4.3.2 . The basic idea is to apply the reduction in an iterative manner, where each time we decrease the suboptimality by a constant multiplicative factor.*

Let $\hat{x}_0 = x_0 = 0$ *and* σ_0 *be a positive scalar. At each time* $t = 0, 1, \ldots$, *we run GD on the function*

$$f_t(x) = f(x) + \frac{\sigma_t}{2}\|x\|^2$$

with the starting point $\hat{x}_t$ *for* $q_t = O(\beta/\sigma_t)$ *iterations. We call the resulted point* $\hat{x}_{t+1}$ *and define* $\sigma_{t+1} = \sigma_t/2$. *For the sake of the analysis, let* $x_t^\star$ *be the minimizer of* f_t. *Denote by* $\Delta_t = f(\hat{x}_t) - f(x^\star)$ *and* $\tilde{\Delta}_t = f_t(\hat{x}_t) - f_t(x_t^\star)$.

1. *Show that for all* t, $\|x_t^\star\| \leqslant \|x^\star\|$.
2. *Show that* $\tilde{\Delta}_0 \leqslant \Delta_0$.
3. *Show that for all* t, *the condition number of* f_t *is* $(\beta + \sigma_t)/\sigma_t$. *Conclude that by appropriately setting the constants in the definition of* q_t, *we ensure that for all* $t = 1, 2, \ldots$,
$$f_{t-1}(\hat{x}_t) - f_{t-1}(x_{t-1}^\star) \leqslant \frac{\tilde{\Delta}_{t-1}}{4} \ .$$
4. *Show that for all* $t = 1, 2, \ldots$,
$$\tilde{\Delta}_t = f_{t-1}(\hat{x}_t) - f_{t-1}(x_t^\star) + \frac{\sigma_t}{2}(\|x_t^\star\|^2 - \|\hat{x}_t\|^2) \leqslant \frac{\tilde{\Delta}_{t-1}}{4} + \frac{\sigma_t}{2}\|x^\star\|^2 \ .$$
5. *Suppose we choose* σ *with* $\sigma \leqslant \Delta_0/\|x^\star\|^2$. *Use induction to deduce that* $\tilde{\Delta}_t \leqslant 2^{-t}\Delta_0$.
6. *Use induction to deduce that* $\Delta_t \leqslant 2^{-(t-1)}\Delta_0$
7. *Conclude that the suggested algorithm attains the bound stated in Theorem 4.3.2.*

Exercise 4.3.6

1. *Use the reduction from the smooth case to the smooth and strongly convex in order to prove a bound of* $O(\|x^\star\|\beta \log(1/\epsilon)/\sqrt{\epsilon})$ *for the* β*-smooth case.*

2. *Use Exercise (4.3.5*) to avoid the extra dependence on* $\log(1/\epsilon)$.

TODO: multiclass

4.4 Computing Eigenvalues: Power Method vs. Lanczos Method

In this lecture we consider the task of computing approximating the k leading eigenvalues of a symmetric matrix $A \in \mathbb{R}^{d\times d}$. In many senses, this challenge is similar to the problem of solving a linear system. For example, exact computation of the eigenvalue decomposition (EVD) can be carried out in time $O(d^\omega)$. Hence, the starting points are identical: since this computational cost is usually prohibitively expensive (and since the corresponding methods do not exploit sparsity conditions), we seek for more efficient iterative methods which provably converge to the solution of the problem.

We present two algorithms, namely the Power method and the Lanczos method. We will see that the relation between the Lanczos method and the Power method resembles the relation between the Gradient Descent and the Conjugate Gradient. In particular, our presentation of the Lanczos method demonstrates another important implication of Theorem 4.2.1.

While the problems share a lot in common, we will shortly observe a significant dissimilarity between the problems: while approximately solving linear systems is equivalent to quadratic convex problem, the problem of approximating the leading eigenvalues is not convex!

For simplicity, we focus on the case where $A \in \mathbb{R}^{d\times d}$ psd and $k = 1$. We denote the spectral (a.k.a. EVD) decomposition of A by $A = \sum_{i=1}^{d} \lambda_i u_i u_i^\top$. Recall that

$$\lambda_1(A) = \max_{x\in\mathbb{R}^d : x\neq 0} \left\{ f(x) = \frac{x^\top A x}{x^\top x} \right\}$$

In our context, an ϵ-approximate maximizer for f is a vector $v \neq 0$ which satisfies $f(v) \geqslant (1-\epsilon)\lambda_1$. Remarkably, the objective f is substantially not convex. Nevertheless, we have quite fast methods for this task.

Power method

Perhaps, the most popular method is the Power method (a.k.a. Power iteration)[8]. Let $x_1 = v$ be a vector chosen uniformly at random from the unit sphere. The power method repeatedly multiplies v by A (and possibly normalize the outcome vector on every round). Denoting the presentation of v according to the eigenbasis by $\sum_{i=1}^{n} \alpha_i u_i$, we have that

$$x_t = A^{t-1} v = \sum_{i=1}^{d} \alpha_i \lambda_i^{t-1} u_i \ . \tag{4.19}$$

[8]The standard method for drawing such a vector is to draw each coordinate i.i.d. according to $\mathcal{N}(0,1)$ and normalize the outcome

We use the following fact:

$$P(\alpha_1^2 \geqslant 1/(9d)) \geqslant 1/2 \ . \tag{4.20}$$

Our analysis is conditioned on the event corresponding to (4.20), i.e., in the sequel we simply assume that (4.20) holds. The main idea behind the Power method is as follows: as we proceed, we expect that for any λ_i which is strictly smaller than λ_1, the gap between λ_1^t to λ_i^t would be large. As a consequence, we expect that x_t would converge to a vector that lies in the eigenspace corresponding to λ_1. Indeed, we will shortly prove the following theorem.

Theorem 4.4.1 *Let $\epsilon > 0$. With probability at least $1/2$, after $t = O\left(\frac{1}{2\epsilon}\log(9n/\epsilon)\right)$, the vector x_t maintained by the Power method satisfies $\frac{x^\top Ax}{x^\top x} \geqslant (1-\epsilon)\lambda_1$. The overall computational complexity is $O\left(t_A \frac{1}{\epsilon}\log(n/\epsilon)\right)$.*

Lanczos method

Note that the vector x_t maintained by the Power method belongs to the Krylov subspace $K_t = \{v, Av, \ldots, A^{t-1}v\}$. Analogously to CG, the Lanczos method maintains the relation

$$x_t \in \arg\max_{x \in K} f(x) \ .$$

The implementation is quite similar to CG. We construct an orthonormal basis, $v_0, \ldots, v_{t-1}$, for K_t (in this case, the v_i's are orthonormal w.r.t. the standard inner product rather than w.r.t. to the inner product induced by A). Similarly to CG, by choosing $v_0 = v$, we ensure that the Gram-Schmidt process can be implemented efficiently. We leave it as an exercise to show that for any $q < s - 1 \leqslant t$, $v_q^\top A v_s = 0$, and consequently, the set $\{v_0, \ldots, v_{t-1}\}$ can be constructed in time $O(t(t_A + d))$. Let $V \in \mathbb{R}^{d \times t}$ be the matrix whose columns are $v_0, \ldots, v_{t-1}$. Since $VV^\top$ is a projection matrix onto K_t, for any nonzero $x \in K_t$ we have

$$\frac{x^\top Ax}{x^\top x} = \frac{x^\top VV^\top AVV^\top x}{x^\top V \underbrace{V^\top V}_{=I} V^\top x} = \frac{(V^\top x)^\top B (V^\top x)}{(V^\top x)^\top (V^\top x)} \ ,$$

where $B = V^\top AV \in \mathbb{R}^{t \times t}$. Note that the map $x \in K_t \mapsto V^\top x$ is a bijection (it maps every vector in K_t to its coefficients according to the orthonormal basis V). Also, the matrix B is tridiagonal. It follows that we can compute x_t as follows:

1. Compute the matrix B in time $O(t(t_A + d))$.
2. Find an exact leading eigenvector z of B in time $O(t^3)$ (in fact, since B is tridiagonal, one can show that this step actually costs $O(t^2)$).
3. Compute $x_t = Vz$ in time td.

All in all, the computation of x_t costs $O(t(t_A + d) + t^2)$. We obtain the following improved bound.

Theorem 4.4.2 *Let $\epsilon > 0$. With probability at least $1/2$, after $t = O\left(\frac{1}{\sqrt{\epsilon}} \log(9n/\epsilon)\right)$, the vector x_t maintained by the Lanczos method satisfies $\frac{x^\top A x}{x^\top x} \geqslant (1-\epsilon)\lambda_1$. The overall computational complexity is $O\left(t_A \frac{1}{\sqrt{\epsilon}} \log(n/\epsilon)\right)$.*

Analysis of the Methods

We next analyze the the convergence rate of the Lanczos method. Along the way, we will also conclude the desired bound on the convergence of the Power method. Therefore, we will deduce Theorem 4.4.1 and Theorem 4.4.2. Given an accuracy parameter $\epsilon > 0$, we would like to obtain an upper bound on the minimal t for which we have that $\frac{\lambda_1 - f(x_t)}{\lambda_1} \leqslant \epsilon$. Note that for both methods, every vector $x \in K_t$ can be written as $p(A)v$ for some polynomial $p \in \Sigma_{t-1}$, we have that

$$x_t = \arg\max_{p \in \Sigma_{t-1}} \frac{v^\top p(A) A p(A) v}{v^\top v} \ .$$

Since $v = \sum_{i=1}^d \alpha_i u_i$, $A = \sum_{i=1}^d \lambda_i u_i u_i^\top$ and $p(A) = \sum_{i=1}^d p(\lambda_i) u_i u_i^\top$, we have that

$$f(x_t) = \max_{p \in \Sigma_{t-1}} \frac{\sum_{i=1}^d \alpha_i^2 p(\lambda_i)^2 \lambda_i}{\sum_{i=1}^d \alpha_i^2 p(\lambda_i)^2} \ .$$

Therefore, for any $p \in \Sigma_{t-1}$,

$$\frac{\lambda_1 - f(x_t)}{\lambda_1} = 1 - \frac{f(x_t)}{\lambda_1} \leqslant \frac{\sum_{i=1}^n \alpha_i^2 p(\lambda_i)^2 (1 - \lambda_i/\lambda_1)}{\sum_{i=1}^n \alpha_i^2 p(\lambda_i)^2} \ .$$

We now split the sum in the enumerator into two parts, depending on whether $1 - \lambda_i/\lambda_1 \leqslant \epsilon$. Let $I \subseteq [n]$ be the set of indices for which the corresponding inequality holds. It follows that

$$\begin{aligned}
\frac{\lambda_1 - f(x_t)}{\lambda_1} &\leqslant \epsilon + \frac{\sum_{i \notin I}^n \alpha_i^2 p(\lambda_i)^2 (1 - \lambda_i/\lambda_1)}{\sum_{i=1}^n \alpha_i^2 p(\lambda_i)^2} \\
&\leqslant \epsilon + \frac{\sum_{i \notin I} \alpha_i^2 p(\lambda_i)^2}{\sum_{i=1}^n \alpha_i^2 p(\lambda_i)^2} \\
&\leqslant \epsilon + \frac{\sum_{i \notin I} \alpha_i^2 p(\lambda_i)^2}{\alpha_1^2 p(\lambda_1)^2} \\
&\leqslant 9n \max_{i \notin I} (p(\lambda_i)^2 / p(\lambda_1)^2) \ .
\end{aligned}$$

It is seen that for $t_0 = 1 + \left\lceil \frac{1}{2\epsilon} \log(9n/\epsilon) \right\rceil$, for every $x \in [0, (1-\epsilon)\lambda_1]$, the polynomial $p_0(x) = x^{t_0 - 1}$

$$p(x)^2 / p(\lambda_1)^2 \leqslant (1-\epsilon)^{2(t_0-1)} \leqslant \epsilon/n \ .$$

Consequently,

$$\frac{\lambda_1 - f(x_{t_0})}{\lambda_1} \leqslant 2\epsilon \ .$$

As a byproduct, we just obtained a proof for Theorem 4.4.1. As in the analysis of CG, the improved rate of the Lanczos method can be proved by means of approximation theory. Namely, from Theorem 4.2.1 we know that there exits a polynomial $p_{t_0,d}$ of degree $d = \left\lceil \sqrt{2t_0 \log(2n/\epsilon)} \right\rceil = O\left(\frac{1}{\sqrt{\epsilon}} \log(n/\epsilon) \right)$ which approximates the polynomial p_0 over $[-1, 1]$ up to an error ϵ/n. Since $\lambda_i/\lambda_1 \in [0, 1]$ for all i, we deduce a speedup of factor $1/\sqrt{\epsilon}$ for the Lanczos method.

The condition number of eigenvalue problems

Lastly, we discuss the rates of the Lanczos method and the Power method under the assumption that λ_2 is strictly smaller than λ_1. Intuitively, the larger is the gap, we expect the convergence to be faster. This intuition is affirmed by the following theorem whose proof is left as an exercise.

Theorem 4.4.3 *Assume that $\lambda_1 - \lambda_2 > 0$ and define the condition number $\kappa = \frac{\lambda_1}{\lambda_1 - \lambda_2}$ be the condition number. In this context, we say that x is an ϵ-approximate maximizer if both $f(x) \geqslant (1-\epsilon)\lambda$ and $\langle x/\|x\|, u_1 \rangle \geqslant 1 - \epsilon$. Then, for any ϵ, the following holds with probability at least $1/2$:*

1. *The Power method achieves an ϵ-approximate maximizer after $t = O(\kappa \log(n/\epsilon))$ iterations.*
2. *The Lanczos method achieves an ϵ-approximate maximizer after $t = O(\sqrt{\kappa} \log(n/\epsilon))$ iterations.*

As we see, different optimization problems give rise to different definitions of the condition number.

Exercises

Exercise 4.4.1 *Prove Theorem 4.4.3.*

4.5 Conditioning and Newton's Method

From here on out, the notes will be more sketchy.

Basic framework for Conditioning

1. Preconditioning: instead of minimizing $f(x)$, we minimize $f(g(x))$, where g is an invertible funciton.

2. For now we consider the case where $f : \mathbb{R}^d \to \mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}^d$ is induced by a positive definite matrix of the form $P^{-1/2}$ (denoting the EVD of P by $P = \sum \lambda_i u_i u_i^\top$, then $P^{-1/2} = \sum \lambda_i^{-1/2} u_i u_i^\top$).

3. Consider a stongly convex quadratic objective $f(x) = x^\top Ax + b^\top x$, where $A \succ 0$. We know that the condition number is λ_1/λ_d, where λ_i is the i-th eigenvalue of A. What would happen if we choose P to be the Hessian, i.e., $P = A$? The new objective becomes

$$f(g(x)) = f(P^{-1/2}x) = x^\top \underbrace{P^{-1/2}AP^{-1/2}}_{I} x + (P^{-1/2}b)^\top x \ .$$

   We see that the new objective is perfectly conditioned! We call this P the *Newton's conditioner*.

4. The catch is that computing $P^{-1/2}$ is hard as solving the problem.

5. Still, more sophisticated conditioners are useful for minimizing quadratic objectives (equivalently, solving linear systems). For example, conditioning is widely used in the context of Laplacian solvers (see [Vishnoi, 2012]).

Newton's Method

1. Conditioning can be also incorporated more adaptively (in this case it is called conditioning rather than preconditioning).

2. Suppose that we aim at minimizing a smooth and strongly convex function (which is not necessarily quadratic) over $\mathbb{R}^d$. We already know how to apply Gradient Descent to this problem. A natural extension of the scheme proposed above is to compute the Hessian at each round and use it as a conditioner. Precisely, we maintain two iterates, x_t and $\tilde{x}_t$, where x_t lies in the original space and $\tilde{x}_t$ lies in the so-called conditioned space. At each time t, the relation between x_t and $\tilde{x}_t$ is given by $\tilde{x}_t = P_t^{1/2} x_t$, where $P_t = \nabla^2 f(x_t)$. While the function being minimized in the original space is fixed (namely, this is the function f), we define the conditioned function at time t by $\tilde{f}_t(y) = f(P_t^{-1/2} y)$.

As in the quadratic case, we next see that at the point $\tilde{x}_t$, the function $\tilde{f}_f$ is perfectly conditioned.

3. The gradient of $\tilde{f}_t$ at any point y is $P^{-1/2}\nabla f(P^{-1/2}y)$ and the Hessian is $P^{-1/2}\nabla^2 f(P^{-1/2}y)P^{-1/2}$. In particular,

$$\nabla \tilde{f}_t(\tilde{x}_t) = P^{-1/2}\nabla f(x), \quad \nabla^2 \tilde{f}_t(\tilde{x}_t) = I \ .$$

4. Thus, our strategy is as follows. Start with $x_1 = 0$. At each time t, perform the following update:

   (a) Compute $\tilde{x}_t = P^{1/2}x_t$ and make a step in the direction of the gradient in the condition space, i.e., let

$$\tilde{x}_{t+1} = \tilde{x}_t - \eta\nabla \tilde{f}_t(\tilde{x}_t) = \tilde{x}_t - \eta P^{-1/2}\nabla f(x_t) \ .$$

   (b) Map $\tilde{x}_{t+1}$ back to the original space. Simple calculation gives the folliwing rule:

$$x_{t+1} = P^{-1/2}\tilde{x}_{t+1} = x_t - \eta P^{-1}\nabla f(x_t) = x_t - \eta\nabla^2 f(x_t)^{-1}\nabla f(x_t) \ .$$

   (c) The obtained algorithm is called **Newton's method**. As we expect, under some suitable assumptions (self-concordance), the convergence of Newton's method is independent of the condition number. We refer to [Nesterov and Nesterov, 2004] for more detials.

A Brief Overview of the Interior Point Method

We briefly introduced the main idea behaind *interior point methods*. We refer to Chapter 3-5 in This notes for further reading.

Chapter 5

Online and Stochastic (Convex) Optimization

5.1 Online Convex Optimization

Main refernce: We follow the survey [Shalev-Shwartz, 2011].

1. The online convex optimization model, Regularized Follow the Leader (RFTL), Online Gradient Descent (OGD): sections 2.1-2.4.

5.2 Strongly Convex Regularizers: from Online Gradient Descent to Exponentiated Gradient

1. We covered Section 2.5.

2. We simplified the proof of lemma 2.8 as we considered only the case where w_t belongs to the interior so its gradient/subgradient vanishes. See the full proof in the survey.

3. Recall that S is the decision set and U is the set of competitors. Note that the FTL lemma (lemma 2.1), which is the basis for the analysis of RFTL, provides a bound w.r.t. a competitor $u \in S$. Hence, to ensure that the bounds are valid w.r.t. any competitor in U, we must form some restriction on the relationship between S and U. While the restriction $U \subseteq S$ obviously works, it can be seen that it suffices to require that the closure of S contains the set U. Note that the later restriction holds in all the setups we consider. For example, in the expert setting, we consider the set $S = \{w \in \mathbb{R}^d_{++} : \sum w_i = 1\}$. While this set does not contain the set of competitors $U = \{e_1, \ldots, e_d\}$, the closure of S does contain this set.

4. In the beginning of the next lecture we will prove that the negative entropy is 1-strongly convex w.r.t. $\|\cdot\|_1$.

Exercises

Exercise 5.2.1 ***(Projected Online Gradient Descent)*** *In this question we extend OGD to the constrained setting. Consider the following setup of RFTL. Let $U = S = \{w \in \mathbb{R}^d : \|w\| \leqslant B\}$ for some $B > 0$. Consider the Regularizer $R : \mathbb{R}^d \to \mathbb{R}$ defined by $R(w) = \frac{1}{2}\|w\|_2^2 + I_S(w)$, where $I_S(w) = 0$ if $w \in S$ and $+\infty$ if $w \notin S$. The class of loss functions under consideration consists of all the convex functions defined over some open set $D \supseteq S$ which are L-Lipschitz w.r.t. $\|\cdot\|_2$. As we saw in class, we may assume w.l.o.g. that each f_t is linear, i.e., $f_t(w) = z_t^\top w$ for some z_t.[1] We also derived the regret bound $R_T \leqslant \sqrt{2T}BL$ for the corresponding instance of RFTL.*

We now show that this instance coincides with a lazy[2] version of projected OGD. Namely, the algorithm is described by:

$$w_0 = 0, \;\; w_t = \Pi_S(-\eta \sum_{i<t} z_i) \;,$$

where Π_S is the projection onto the set S, i.e.,

$$\Pi_S(w) = \begin{cases} w & \|w\| \leqslant B \\ \frac{B}{\|w\|} w & \|w\| > B \end{cases} \;.$$

(Hint: denote by $\theta = -\sum_{i<t} z_i$ and show that w_t is the minimizer of $\|w - \eta\theta\|$ over S.)

Exercise 5.2.2 *Consider the model of prediction with expert advice. Each vector in $U = \{e_1, \ldots, e_d\}$ corresponds to an expert. The loss functions are of the form $f_t(w) = z_t^\top w$. The loss of expert i at time t, denoted $z_{t,i}$, lies in the range $[-1, 1]$. The learner is allowed to choose its decision from the set $S = \{w \in \mathbb{R}^d_{++} : \sum w_i = 1\}$.[3] Let us apply the RFTL algorithm with the negative entropy regularizer $R : \mathbb{R}^d_{++} \to \mathbb{R}$ defined by $R(w) = \sum w_i \log w_i$. We call the resulted algorithm Exponentiated Gradient (EG). Show that EG corresponds to the following update rule:*

$$w_1 = (1/d, \ldots, 1/d), \;\; w_{t+1,i} = \frac{w_{t,i} \exp(-\eta z_{t,i})}{\sum_{i=1}^d w_{t,i} \exp(-\eta z_{t,i})} \;.$$

(Hint: use Corollary 2.1.1.)

[1] We simply set $z_t \in \partial f_t(w_t)$ and observed that the regret with respect to the z_t's upper bounds the regret with respect the f_t's.

[2] The algorithm is lazy in the sense that if are only interested in computing the last iterate w_T we do not need to compute all the w_t's and in particular, we do not need to perform all the projections.

[3] Technically, we do not allow the learner to put zero probability on any expert. The analysis shows that this restriction is not harmful.

Exercise 5.2.3 *Recall the formulation of a zero-sum game from Exercise (1.1.12). In Exercise (2.2.6*) we proved the minimax theorem by using the separation theorem. We now provide a constructive proof. Namely, we show that the EG algorithm can be used to approximately find an optimal strategy both for the row player and the column player.*

We assume for simplicity that the entries of the matrix A belong to $[0,1]$. Recall that the column player wishes to minimize its loss. Consider a repeated game where at each round, the column player chooses a probability vector q_t and thereafter the row player responds with its best (pure) response (the fact that the best response is pure follows from Exercise (1.1.12)). We assume that the column player uses the EG algorithm to maintain the probability vector q_t. Let $\bar{p} = \frac{1}{T}\sum_{t=1}^{T} p_t$ and $\bar{q} = \frac{1}{T}\sum_{t=1}^{T} q_t$. Show that for any $\epsilon > 0$, there exists T such that after T rounds, both $\bar{p}^\top A q \geqslant mM - \epsilon$ for all q and $p^\top A \bar{q} \leqslant Mm + \epsilon$ for all p. Conclude that $Mm = mM$.

5.3 Multi-armed Bandit

We follow section 4.2 in [Shalev-Shwartz, 2011]. The bandit setting is very similar to the standard experts setting. In both cases, at time t the learner chooses, possibly at random an expert $i_t \in [d]$. Denoting the loss of expert i at time t by $z_{t,i}$ and letting w_t be the probability vector maintained by the learner, the loss incurred by the learner is the expected loss of the chosen expert, i.e.,

$$\mathbb{E}[f_t(w_t)] = \mathbb{E}_{i_t \sim w_t} = z_t^\top w_t \ .$$

The only (important) difference is the received feedback. While in the experts setting the learner gets to see the losses of all the experts (i.e., it observes the vector z_t), in the bandit setting the only available information to the learner is the loss of the chosen learner i_t (i.e., the learner only observes z_{t,i_t}). We call this problem *multi-armed Bandit* due to its similarity to gambling, where each expert is viewed as a single *arm* and the loss associated with each arm is the gain resulted from picking it.

This setting is very natural. Indeed, in many (learning) tasks we can only observe the outcome of our choice. For instance, this is often the case in source routing and web advertising tasks.[4]

Due to the limited feedback, the learner faces the fundamental exploration-exploitation tradeoff. On one hand, the learner wishes to exploit the information received throughout the learning process and follow the experts that seem most successful. On the other hand, due the adversarial characteristic of our setting, the learner must sample every expert from time to time.

5.3.1 Reducing the bandit setting to the experts setting by devoting fixed amount of time to exploration

We next describe an algorithm named `MAB` that essentially reduces the bandit setting to the standard setting.[5] The main idea is as follows. Let us divide the time interval T into k time subintervals of equal size. Roughly speaking, we will associate every subinterval with a single round of the EG algorithm. As we describe below, at the beginning of every subinterval I_s ($s \in [k]$), `MAB` generates a random vector $\ell_{s-1} \in [-1, 1]^d$ which forms an unbiased estimate to the average loss of the experts during the previous subinterval. Then it passes this vector to EG which in turn, uses this feedback in order to update its weights over the arms. Throughout the following subinterval, `MAB` chooses its actions according to this distribution except for randomly

[4] In source routing we need to find a path between a source and a target. The cost associated with each path is the congestion along this path. Ad placement is the problem of deciding which advertisement to display on a web page. The gain (or negative loss) is associated with the visitors' actions (e.g., whether the user downloaded the software or not).

[5] This part is not covered in [Shalev-Shwartz, 2011].

chosen d time steps[6] in which it sample each arm once. The vector ℓ_s consists of exactly these estimates. Note that indeed, for every $j \in [d]$, $\ell_{s,j}$ is an unbiased estimate of the average loss of the j-th expert during the s-th interval. Note also that the regret of EG w.r.t. to the sequence $\ell_1, \ldots, \ell_k$ is $O(\sqrt{k \log(d)})$. By relating the expected loss of EG to the expected loss of `MAB` , we are able to prove the following result.

Theorem 5.3.1 *Let the number of subintervals be $k = (T/d)^{2/3}$ and assume that $k > d$. The expected regret of the suggested method is $O(T^{2/3}(d \log(d))^{1/3})$.*

Note that to ensure that the average regret is at most ϵ, we need T to be of order $\Omega(d/\epsilon^3)$.

5.3.2 Exp3: Simultaneous exploration-exploitation

Instead of separting between exploration and exploitation, the Exp3 algrotithm peforms these tasks simultaneously. Similarly to `MAB`, the Exp3 feeds the EG algorithms with unbiased estimates of the losses. The main difference is that the size of each interval is 1. Obviously, we can only sample one arm in each block. However, as we next explain, we can still produce unbiased estimates of the loss. Recall that at time t, the learner chooses an action i_t according to the (positive) probability vector w_t. Let

$$\hat{z}_{t,i} = \begin{cases} \frac{z_{t,i}}{w_{t,i}} & i = i_t \\ 0 & i \neq i_t \end{cases}$$

We next observe that $\hat{z}_{t,i}$ is an unbiased estimate of z_t. Indeed,

$$\mathbb{E}[\hat{z}_{t,i}|w_t] = P(i_t = i) \cdot \hat{z}_{t,i} + P(i_t \neq i) \cdot 0 = w_{t,i} \frac{z_{t,i}}{w_{t,i}} = z_{t,i} \ .$$

Next, we argue that the expected loss of the learner w.r.t. the sequence $(\hat{z}_t)_{t=1}^T$ is equal to the expected loss w.r.t. the sequence $(z_t)_{t=1}^T$. Indeed, by using the law of total expectation as follows:

$$\mathbb{E}[\hat{z}_t^\top w_t] = \mathbb{E}[\mathbb{E}[\hat{z}_t^\top w_t | w_t]] = \mathbb{E}[z_t^\top w_t] \ .$$

Similarly, the expected loss of any fixed competitor $u \in \{e_1, \ldots, e_d\}$ w.r.t. $(\hat{z}_t)$ is equal to the expected loss w.r.t. the sequence $(z_t)_{t=1}^T$. Finally, the same argument holds for the regret. Therefore, it suffices to bound the expected regret of EG w.r.t. $(\hat{z}_t)$. To this end, we need the following refined analysis of EG's regret. To simplify matters, we assume that the losses are nonnegative.

[6] Namely, the d chosen time steps are chosen uniformly at random. We do not require that the d choices would be independent so it is easy to make sure that there are no collisions.

Theorem 5.3.2 *For any sequence of nonnegative loss vectors $(z_t)_{t=1}^T$, the regret of the EG algorithm is bounded above by*

$$R_T \leqslant \frac{\log(d)}{\eta} + \eta \sum_{t=1}^{T} \sum_{i=1}^{d} w_{t,i} z_{t,i}^2 \ .$$

The proof is outlined in the exercises. Recall that the bound obtained using the general bound of RFTL is

$$R_T \leqslant \frac{\log(d)}{\eta} + \eta \sum_{t=1}^{T} \|z_t\|_\infty^2 \ .$$

The refined bound replaces the ℓ_∞ norm of each $\hat{z}_t$ (which corresponds to the Lipschitzness of the loss w.t.t. the ℓ_1 norm) with the expression $\sum_{i=1}^d w_{t,i}\hat{z}_{t,i}^2$. Please note that since w_t has positive coordinates, the map $z \mapsto \sqrt{\sum_{i=1}^d w_i z_i^2}$ is a norm. It is called the *local norm* induced by w_t and we denote it by $\|\cdot\|_{w_t}$

In order to bound the expected regret of the Exp3 algorithm, it remains to bound $\mathbb{E}[\|\hat{z}_t\|_{w_t}^2]$ for each t. Using again the law of total expectation, we obtain

$$\mathbb{E}\left[\sum_{i=1}^{d} w_{t,i}\hat{z}_{t,i}^2\right] = \mathbb{E}\left[\sum_{i=1}^{d} w_{t,i} \cdot \mathbb{E}[\hat{z}_{t,i}^2 | w_t]\right] = \mathbb{E}\left[\sum_{i=1}^{d} w_{t,i}(w_{t,i} \cdot \frac{z_{t,i}^2}{w_{t,i}^2} + 0)\right] = \sum_{i=1}^{d} z_{t,i}^2 \leqslant d \ ,$$

(recall that $|z_{t,i}| \leqslant 1$ for all i). We summarize the above in the next theorem.

Theorem 5.3.3 *The regret of the Exp3 algorithm is bounded above by*

$$R_T \leqslant \frac{\log(d)}{\eta} + \eta T d \ .$$

By choosing $\eta = \sqrt{\frac{\log(d)}{Td}}$, we obtain that

$$R_T \leqslant 2\sqrt{Td\log(d)} \ .$$

Note that to ensure that the average regret is at most ϵ, we need T to be of order $\Omega(d/\epsilon^2)$. Therefore, we improve over `MAB` in factor $1/\epsilon$.

Exercises

Exercise 5.3.1 *Prove Theorem 5.3.1. (Hint: a) The overall regret during the exploration rounds is trivially bounded by $2kd$. b) Show that the expected regret of `MAB` at the exploitation rounds is at most T/k times the expected regret of EG on the sequence $\ell_1, \ldots, \ell_k$.)*

Exercise 5.3.2 *Explain why the bound in Theorem 5.3.2 is indeed better than the bound obtained using the RFTL bound.*

Exercise 5.3.3 *In the question we prove Theorem 5.3.2. We first introduce some useful notation. Define the unnormalized weight vectors $\tilde{w}_t$ by*

$$\tilde{w}_t = (1, \dots, 1), \ \tilde{w}_{t+1,i} = \tilde{w}_{t,i} \exp(-\eta z_{t,i})$$

Also, let $\tilde{W}_t = \sum_{i=1}^d \tilde{w}_{t,i}$. Note that $w_{t,i} = \tilde{w}_{t,i}/\tilde{W}_t$. Denote the expected loss of the learner at time t by $\hat{\ell}_t = w_t^\top z_t$ and let $\hat{\ell}_t^2 = \sum_{i=1}^d w_{t,i} z_{t,i}^2$. The proof is divided into two parts. The first part relate $\tilde{W}_t$ to the loss of the learner. The second part related $\tilde{W}_t$ to the loss of any fixed expert.

1. *Show that for any t, $\tilde{W}_{t+1} = \tilde{W}_t \sum_{i=1}^d w_{t,i} \exp(-\eta z_{t,i})$.*
2. *Use the inequalities $\exp(-x) \leqslant 1 - x + x^2$ and $1 - x \leqslant \exp(-x)$ which holds for all x in order to deduce the inequality*

$$\tilde{W}_{T+1} \leqslant d \exp(-\eta \sum_{t=1}^T \hat{\ell}_t + \eta^2 \sum_{t=1}^T \hat{\ell}_t^2) \ .$$

3. *Show that for any expert $j \in [d]$, $\tilde{W}_{T+1} \geqslant \exp(-\eta \sum_{t=1}^T z_{t,j})$*
4. *Deduce the theorem.*

5.4 Stochastic Dual Coordinate Ascent

The main resource for this lecture is [Shalev-Shwartz, 2016].

Chapter 6

Bibliographic Remarks

1. **The Condition number:** In the optimization literatue, a rate that scales logarithmically with $1/\epsilon$ (as in Theorem 4.1.1) is often called a *linear rate*. This stems from the fact that every "new right digit of the answer" takes constant number of iterations.

2. **Accelearation using Conhugate Gradient:** We say that a set of vectors $\{u_1, \ldots, u_i\}$ are A-conjugate if the set $\{u_1, \ldots, u_i\}$ is orthogonal according to $\langle \cdot, \cdot \rangle_A$. This explains the name of the CG method. The presentation in this section is inspired by [Vishnoi, 2012] and [Sachdeva and Vishnoi, 2013]. **TODO: reference to the lower bound (Nemirovski & Yudin, Yossi, Agrawal).**

3. **Smooth Convex Optimization:** **TODO: reference to Hazan's survey. AGD and its variants** The improved reduction detailed in Exercise (4.3.5*) is from [Allen-Zhu and Hazan, 2016].

Bibliography

[Allen-Zhu and Hazan, 2016] Allen-Zhu, Z. and Hazan, E. (2016). Optimal black-box reductions between optimization objectives. *arXiv preprint arXiv:1603.05642*.

[Borwein and Lewis, 2010] Borwein, J. M. and Lewis, A. S. (2010). *Convex analysis and nonlinear optimization: theory and examples*, volume 3. Springer.

[Clarkson and Woodruff, 2013] Clarkson, K. L. and Woodruff, D. P. (2013). Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 81–90. ACM.

[Nesterov and Nesterov, 2004] Nesterov, Y. and Nesterov, I. E. (2004). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer.

[Sachdeva and Vishnoi, 2013] Sachdeva, S. and Vishnoi, N. K. (2013). Faster algorithms via approximation theory. *Theoretical Computer Science*, 9(2):125–210.

[Shalev-Shwartz, 2011] Shalev-Shwartz, S. (2011). Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194.

[Shalev-Shwartz, 2016] Shalev-Shwartz, S. (2016). Sdca without duality, regularization, and individual convexity. *arXiv preprint arXiv:1602.01582*.

[Shalev-Shwartz and Zhang, 2013] Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss. *The Journal of Machine Learning Research*, 14(1):567–599.

[Vishnoi, 2012] Vishnoi, N. K. (2012). Laplacian solvers and their algorithmic applications. *Theoretical Computer Science*, 8(1-2):1–141.