Final Exam for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

May 10, 2018

Rules

- The final is due on Gradescope at 8:00pm Friday, May 11th.
- Collaboration is forbidden.
- Please ask all questions via *private* messages on Piazza. Do not email me and do not post publicly (even if you think "nothing will be given away").
- I will be maintaining a list of errata on Piazza. Please consult this before asking a question about a typo that's already been addressed.
- The final must be LaTeXed. Points will be deducted otherwise. Remember to write $\langle x, y \rangle$ instead of $< x, y >$; the latter is ugly, a pain to follow when grading, and will result in loss of points.
- You may use anything in the course notes or one of the course texts as given. In particular, you can cite theorems and equations from course notes and materials, but please point me to the lecture number and the statement you are citing. You may also use any standard references on calculus, analysis or linear algebra. You may not Google anything or use the web (except perhaps to find a standard textbook to download)
- Optional problems are not for credit, but may give you good Karma.

- Lastly, please do not be daunted by the length of the exam. Most of the space is taken up by background, definitions, and a couple optional problems.

Problem 1: Making Heads and Tails of Matrices

Let $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^d$ be a linear map, and consider the problem of minimizing

$$\min_{X\in\mathbb{R}^{n\times n}} \Psi(X) \quad \text{s.t.} \quad \text{rank}(X) \leqslant r$$
$$\text{where } \Psi(X) := \tfrac{1}{2}\|\mathcal{A}(X) - b\|_2^2 .$$

where $b \in \mathbb{R}^d$. We assume that there exists an $X^* \in \mathbb{R}^{n\times n}$ with $\text{rank}(X^*) \leqslant r$ satisfying $\mathcal{A}(X^*) = b$.

You might the following quick review helpful. Given $X, Y \in \mathbb{R}^{n\times n}$, recall that the matix inner product and Frobenius can be expressed as

$$\langle X, Y\rangle = \sum_{ij} X_{ij}Y_{ij} \text{ and } \|X\|_F^2 = \langle X, X\rangle . \tag{1}$$

(A) We say that $\mathcal{A}$ satisfies the (k, δ_k)-RIP if, for all matrix X with rank at most k, $(1 - \delta_k)\|X\|_{\text{F}}^2 \leqslant \|\mathcal{A}(X)\|_2^2 \leqslant (1 + \delta_k)\|X\|_{\text{F}}^2$. Show that if $\mathcal{A}$ satisfies the $(2r, \delta_{2r})$-RIP for any $\delta_{2r} \in (0, 1)$, then X^* is the unique matrix with $\text{rank}(X) \leqslant r$ satisfying $\Psi(X) = 0$.

(B) Show that if $d < nr$, $\mathcal{A}$ cannot satisfy the (r, δ)-RIP property for any $\delta \in (0, 1)$.

(C) Let $\text{svd}_r(X)$ denote the rank-r singular value truncation of a matrix. That is, if $X = U\Sigma V^\top$, $\text{svd}_r(X) = U\Sigma_r V^\top$, where Σ_r is obtained by setting all but the r-largest entries of Σ to zero.

The operation $\text{svd}_r(X)$ is then the Euclidean projection (in $\|\cdot\|_F$) onto a non-convex set. What is this set (you don't need to prove anything), and why is this set non-convex?

(D) As with any linear operator, we can define the tranpose of $\mathcal{A}$, denoted $\mathcal{A}^\top : \mathbb{R}^d \to \mathbb{R}^{n\times n}$. This is the unique linear operator such that, for all $v \in \mathbb{R}^d$ and $X \in \mathbb{R}^{n\times n}$

$$\langle \mathcal{A}^\top(v), X\rangle = \langle v, \mathcal{A}(X)\rangle . \tag{2}$$

Equivalently, if we view X as an $n \times n$ vector and $\mathcal{A}$ as $d \times (n \times n)$ matrix, then $\mathcal{A}^\top$ is just the matrix transpose of $\mathcal{A}$ (*hint: use the inner product definition of $\mathcal{A}^\top$ to answer all subsequent questions. DO NOT try to write things out as matrices*).

We now consider the following algorithm:

Input: Parameters $\eta \in (0, 1/2)$, $X_0 \leftarrow 0$. **for** $t = 0, 1, 2, \ldots$ **do**
  $Y_{t+1} \leftarrow X_t - \eta\mathcal{A}^\top(\mathcal{A}(X_t) - b)$;
  $X_{t+1} \leftarrow \text{svd}_r(Y_{t+1})$;
end

Algorithm 1: PGD

Suppose that $\mathcal{A}$ satisfies the $(2r, \delta_{2r})$-RIP, and set $\eta = \frac{1}{1+\delta_{2r}}$. Moreover, define the function:

$$F(A, B) := \langle \mathcal{A}^T(\mathcal{A}(B) - b), A - B \rangle + \frac{1 + \delta_{2r}}{2} \|A - B\|_{\mathrm{F}}^2 \tag{3}$$

(D.1) Prove that $\Psi(X_{t+1}) - \Psi(X_t) \leqslant F(X_{t+1}, X_t)$

(D.2) Prove that $F(X_{t+1}, X_t) \leqslant F(X_*, X_t)$ (*Hint: try writing* $F(A, X_t)$ *as another function* $G(A, Y_{t+1})$)

(D.3) Conclude that $\Psi(X_{t+1}) \leqslant \frac{\delta_{2r}}{1-\delta_{2r}} \|\mathcal{A}(X_t - X_*)\|_2^2$ (note this is the Euclidean norm on $\mathbb{R}^d$). (There is no typo, there is no extra factor of 2 in this inequality)

(E) Prove the following theorem:

Theorem 1. Let $\mathcal{A}$ be $(2r, \delta_{2r})$-RIP for $\delta_{2r} \leqslant 1/3$, and let X^* denote the unique rank r minimizer of $\Psi(X)$. Then, the iterates of Algorithm 1 satisfy:

$$\|X_t - X_*\|_F^2 \leqslant \frac{1}{1 - \delta_{2r}} \cdot \left(\frac{2\delta_{2r}}{1 - \delta_{2r}} \right)^t \|b\|^2 \tag{4}$$

Problem 2: Non-Convex Black Box Optimization

A function $f : [0,1] \to \mathbb{R}$ is said to be semi-strictly unimodal if there exists an interval $[a,b] \in [0,1]$ such that $f(a) = f(b) = f(x)$ for all $x \in [a,b]$, and f is strictly decreasing on $[0,a]$ and strictly increasing on $[b,1]$. We will consider the *zeroth-order* black box model, where the the algorithm is allowed to query the function value $f(x_1), \ldots, f(x_T)$ at iterates $x_1, \ldots, x_T$. Here, x_{t+1} is allowed to depend on $(x_s, f(x_s))_{s=1}^t$.

(A) Show every non-constant convex function is semi-strictly unimodal.

(B) Given an example of a non-convex function which is semi-strictly unimodal.

(C) Consider your favorite black-box convex optimization algorithm for convex functions $f : [0,1] \to \mathbb{R}$ (e.g. Golden Section Search from class, or ternary search). Show if f is only semi-strictly unimodal (but not necessarily convex), the algorithm still finds an x such that there exists an $x^* \in \arg\min_{x \in [0,1]} f(x)$ with $|x - x^*| \leqslant \epsilon$ in $\mathcal{O}(\log(1/\epsilon))$ function queries. *No need to go through the analysis of the entire algorithm, just pinpoint the step in the analysis that uses convexity and generalize it to semi-strictly unimodal functions.*

(D) A function $f : [0,1] \to \mathbb{R}$ is said to be non-strictly unimodal if there exists an interval $[a,b] \in [0,1]$ such that $f(a) = f(b) = f(x)$ for all $x \in [a,b]$, and f is non-increasing on $[0,a]$ and non-decreasing on $[b,1]$. Show that *no zeroth orther black box algorithm* can be guaranteed to find find a point x within ϵ of an optimum x^* of a non-strictly unimodal algorithm, even if f is guaranteed to be continuous.

(E) Suppose $f : [0,1]$ is non-strictly unimodal and L Lipschitz. Show that any black box algorithm which find an x such that $|f(x) - f(x^*)| \leqslant \epsilon$ requires $\mathcal{O}(L/\epsilon)$ function evaluations. Give a simple algorithm which attains this lower bound.

(F) Generalize the previous question to d dimensions.

Problem 3: Optimization without Gradients

Note: Parts A.1 and A.2 are optional, and Part C does not ask you to show anything. Given a set $\mathcal{S} \subset \mathbb{R}^d$ and $\alpha \in \mathbb{R}$, we let $\alpha\mathcal{S} := \{\alpha x : x \in \mathcal{S}\}$. For $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}^d$, we let $\mathcal{S}_1 + \mathcal{S}_2 := \{x_1 + x_2 : x_1 \in \mathcal{S}_1, x_2 \in \mathcal{S}_2\}$.

(A) Let f be a continuously differentiable convex function. Let $\mathcal{B}^n := \{x \in \mathbb{R}^n : \|x\|_2 \leqslant 1\}$ and $\mathcal{S}^{n-1} := \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. We let $u \sim \mathcal{B}^n$ and $u \sim \mathcal{S}^{n-1}$ denote the uniform measures on $\mathcal{B}^n$ and $\mathcal{S}^{n-1}$, respectively. Define for $\delta > 0$, the δ-smoothings

$$\widehat{f}_\delta(x) := \mathbb{E}_{u \sim \mathcal{B}^n}[f(x + \delta u)] \tag{5}$$

We will prove that

$$\nabla \widehat{f}_\delta(x) = \frac{n}{\delta} \cdot \mathbb{E}_{u \sim \mathcal{S}^{n-1}} f(x + \delta u) u \tag{6}$$

following these three steps

(A.1) (Optional) Prove, using Stoke's theorem from calculus, that

$$\frac{\delta}{\mathrm{vol}(\mathcal{B}^n)} \cdot \nabla \widehat{f}_\delta(x) = \int_{u \in \mathcal{S}^{n-1}} f(x + \delta u) \cdot d\boldsymbol{u} \tag{7}$$

Feel free to use your favorite calculus resources.

(A.2) (Optional) that there exists a dimension-dependent constant $C(n)$ (which you do not need to specify yet) such that, for any continuously differentiable f and $\delta > 0$,

$$\delta \cdot C(n) \cdot \nabla \widehat{f}_\delta(x) = \mathbb{E}_{u \sim \mathcal{S}^{n-1}}[f(x + \delta u) u] \tag{8}$$

(A.3) Using (8), prove that $C(n) = 1/n$ by choosing a particular function f, and computing the values of $\nabla \widehat{f}_\delta(x)$ and $\mathbb{E}_{u \sim \mathcal{S}^{n-1}}[f(x + \delta u) u]$ for that f (note that any f will do).

(B) Let $\mathcal{K} \subset \mathbb{R}^n$ be a convex set, and suppose that $\mathcal{B}^n \subset \mathcal{K}$.

(B.1) Prove that for $\delta \in (0, 1)$, the following set $\mathcal{K}_\delta := (1 - \delta)\mathcal{K}$ is convex and, for all $x \in \mathcal{K}_\delta$, $x + \delta\mathcal{B}^n \subset \mathcal{K}$.

(B.2) Let $F(x) = \Pi_{\mathcal{K}}(x)$ denote the Euclidean projection on $\mathcal{K}$. Show how to use $F(x)$ as a black box to compute projections onto $\mathcal{K}_\delta$.

(C) You may use the following theorem (nothing to prove)

Theorem 0.1. *Let $\phi_1, \ldots, \phi_T$ denote a sequence of convex functions on a domain $\mathcal{X}$ and consider the iterates $x_0 \in \mathcal{X}$ and $x_{t+1} = \Pi_{\mathcal{X}}(x_t - \eta g_t)$, where*

- *$\Pi_{\mathcal{X}}$ denotes the Euclidean projection onto $\mathcal{X}$.*

- $\mathbb{E}[g_t | x_1, \ldots, x_t, g_1, \ldots, g_{t-1}] = \nabla \phi_t(x_t)$
- $\mathbb{E}[\|g_t\|^2 | x_1, \ldots, x_t, g_1, \ldots, g_{t-1}] \leqslant G^2$.
- $\max_{x_1, x_2 \in \mathcal{X}} \|x_1 - x_2\|_2 \leqslant D$.

Then there exists a universal constant C such that

$$\mathbb{E}[\sum_{t=1}^{T} \phi_t(x_t) - \min_{x \in \mathcal{X}} \sum_{t=1}^{T} \phi_t(x)] \leqslant C(\eta G^2 T + \frac{D^2}{\eta}) \tag{9}$$

(D) Let $f_1, \ldots, f_T$ be a sequence of convex, G-Lipschitz, continuously differentiable functions, with δ-smoothings $\widehat{(f_t)}_\delta(x_t)$. Let

$$D = \max_{x, x' \in \mathcal{K}} \|x - x'\|_2 \tag{10}$$

You may assume that for each t, there exists some $z_t \in \mathcal{K}$ such that $f_t(z_t) = 0$ (the algorithm does not know this z_t, this is for the analysis).

(D.1) Show that, for every $\eta > 0, \delta \in (0, 1)$, there exists an algorithm which satisfies the following:

- Computes one function evaluation $f_t(\widetilde{x}_t)$ at each round t, and one call to a projection oracle $F(x) = \Pi_{\mathcal{K}}(x)$
- For each t, $\widetilde{x}_t \in \mathcal{K}$ and $\widetilde{x}_t$ depends only on $(f_1, \ldots, f_{t-1})$, $(x_1, \ldots, x_{t-1})$.
- For each t, there exists an x_t such that $\|x_t - \widetilde{x}_t\|_2 \leqslant \delta$ and $x_1, \ldots, x_T$ satisfy

$$\mathbb{E}[\sum_{t=1}^{T} \widehat{(f_t)}_\delta(x_t) - \min_{x \in K_\delta} \sum_{t=1}^{T} \widehat{(f_t)}_\delta(x)] \leqslant C \left(\eta T \cdot \frac{n^2}{\delta^2} \cdot D^2 G^2 + \frac{D^2}{\eta} \right) \tag{11}$$

(D.2) Noting that $D \geqslant 1$, prove the bounds

$$\mathbb{E}[\sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)] \leqslant C \left(\eta T \cdot \frac{n^2}{\delta^2} \cdot D^2 G^2 + \frac{D^2}{\eta} \right) + 3\delta DGT \tag{12}$$

and

$$\mathbb{E}[\sum_{t=1}^{T} f_t(\widetilde{x}_t) - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} f_t(x)] \leqslant C \left(\eta T \cdot \frac{n^2}{\delta^2} \cdot D^2 G^2 + \frac{D^2}{\eta} \right) + 4\delta DGT \tag{13}$$

(D.3) Optimize both bounds above in terms of δ and η (you may be loose up to constant factors that do not depend on n, D, G, T). At what rate do the bounds grow in T? If $f_1 = f_2 = \cdots = f_T$, what rate of convergence do you get for the averages $\frac{1}{T}\sum_{t=1}^{T} x_t$ and $\frac{1}{T}\sum_{t=1}^{T} \widetilde{x}_t$.

(D.4) Compare and contrast the methods to the randomized direction descent algorithm we saw on homework 2. Account for the advantages of each algorithm in its appropriate setting.