

Problem Set 1 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

January 29, 2018

Problem 1: Existence of the Subgradients

- (A) Let \mathcal{X} be a convex set. Prove that that given any convex function $f: \mathcal{X} \rightarrow \mathbb{R}$ and any $x \in \mathcal{X}$, there exists at least one vector g , called a *subgradient* of f at x , such that $f(y) \geq f(x) + \langle g, y - x \rangle$ for all $y \in \mathcal{X}$.

To establish this claim, you may follow the steps below. We will only prove the existence under slightly restricted assumptions, but you can assume that the vector g above exists in full generality.

- (A.1) Define the *Epigraph* of f , $\text{Epi}(f) := \{(x, t) \in \mathcal{X} \times \mathbb{R} : f(x) \leq t\}$. Prove the $\text{Epi}(f)$ is convex.
- (A.2) Recall the following definitions from real analysis:

Definition 1 (Boundary and Interior).

Using the separating hyperplane theorem from the notes (the full version, which applies to arbitrary convex sets not just compact ones), prove the supporting hyperplane theorem.

Theorem 1 (Supporting Hyperplane). Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set, and let $x \in \text{Bd}(\mathcal{C})$. Then, there exists a nonzero $w \in \mathbb{R}^n$ such that, for all $y \in \mathcal{C}$, $\langle w, y - x \rangle \geq 0$.

Hint: Find two (not-necessarily compact!) convex sets to apply the separating hyperplane theorem. You might want $\text{Int}(\mathcal{C})$ to be one of them - and you should check that $\text{Int}(\mathcal{C})$ is convex

- (A.3) Using part *i)* and *ii)*, prove the existence of a subgradient at $x \in \mathcal{X}$. You may assume that $x \in \text{Int}(\mathcal{X})$ to avoid annoying edge cases.
- (B) Let $\{f_i\}_{i \in I}$ be a (possibly infinite, uncountable) family of convex functions, and suppose that $f_i(x) < \infty$ for all $x \in \mathcal{X}$. Show that $f(x) := \sup_i f_i(x)$ is convex on \mathcal{X} (you may assume $f(x)$ is finite).
- (C) Using what we've proven about subgradients, prove that a function $f : \mathcal{X} \rightarrow \mathbb{R}$ is convex if and only if it can be written as the supremum of affine functions (e.g. supremum of functions of the form $f_i(x) = \langle a_i, x \rangle + b_i$)

Problem 2: Properties of Subgradients

Let f be a convex function over a domain \mathcal{X} . We will assume $x \in \text{Int}(\mathcal{X})$.

- (A) Show by way of example that the subgradient is not necessarily unique, but that the set of all subgradients is closed and convex. We will denote this set $\partial f(x)$.
- (B) Show that f has a directional derivative in each direction. Use this to conclude that a convex f is differentiable at x only if $\partial f(x) = \{\nabla f(x)\}$.
- (C) Show that if $g_1 \in \partial f_1(x)$ and $g_2 \in \partial f_2(x)$, then $g_1 + g_2 \in \partial(f_1 + f_2)(x)$.
- (D) Let $f(x) = \sup_i g_i(x)$ which g_i convex. Show that $\text{Conv}\{\partial g_i(x) | g_i(x) = f(x)\} \subseteq \partial f$.
- (E) *This problem is likely to be changed in further revisions* Prove that if $f_i(x) = w_i^\top x b_i$ are a compact family of affine functions (i.e. $\{(w_i, b_i)\} \subset \mathbb{R}^{n+1}$ is compact), then the converse is true, namely $\partial f \subset \text{Conv}\{\partial g_i(x) | g_i(x) = f(x)\}$.

Problem 3: Subgradients of Norms

- (A) Subgradient of the ℓ_1 and ℓ_∞ -norms
- (A.1) Prove that, for all $x \in \mathbb{R}^n$, $\|x\|_1 = \sup_{y: \|y\|_\infty \leq 1} \langle x, y \rangle$, $\|x\|_\infty = \sup_{y: \|y\|_1 \leq 1} \langle x, y \rangle$.
- (A.2) Compute $\partial\|x\|_1$ and $\partial\|x\|_\infty$
- (B) Subgradient of the L_1 -norm
- (B.1) Let $A \in \mathbb{R}^{m \times n}$. Let $\sigma_i(\cdot)$ denote the i -th singular value of a matrix. Using the inequality $\sum_{i=1}^{\min(n,m)} \sigma_i(AB) \leq \sum_{i=1}^{\min(n,m)} \sigma_i(A)\sigma_i(B)$ for all $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$ (this is non-trivial, see [this Stack Exchange](#)), prove the following: For all $X \in \mathbb{R}^{m \times n}$,

$$\|X\|_{\text{op}} := \max_{Y \in \mathbb{R}^{m \times n}: \|Y\|_{\text{nuc}} \leq 1} \langle X, Y \rangle \text{ and } \|X\|_{\text{nuc}} = \max_{Y \in \mathbb{R}^{m \times n}: \|Y\|_{\text{op}} \leq 1} \langle X, Y \rangle, \quad (1)$$

where $\|X\|_{\text{op}} := \sigma_{\max}(X)$, $\|Y\|_{\text{nuc}} := \sum_{i=1}^{\min(n,m)} \sigma_i(Y)$, and $\langle X, Y \rangle := \text{tr}(X^\top Y)$. You may want to refresh yourself on the relationship between traces, eigenvalues and singular values, and some trace tricks. Feel free to use the bound $\sum_i \lambda_i(A) \leq \sum_i \sigma_i(A)$ for any squared matrix A .

(B.2) Compute $\partial\|X\|_{\text{op}}$ and $\partial\|X\|_{\text{nuc}}$. Under what conditions is each subgradient unique?

(C) Let $\|\cdot\|$ be an arbitrary norm (not necessarily Euclidean!) on \mathbb{R}^n . Define the dual norm $\|y\|_* := \sup_{x: \|x\| \leq 1} \langle x, y \rangle$.

(C.1) Show that the dual norm is a norm, and describe its subgradient.

(C.2) Show that for all $g, w \in \mathbb{R}^n$, $|\langle g, w \rangle| \leq \|g\|_* \|w\|$

(C.3) Let f be a convex function on a convex domain \mathcal{X} . Show that f is L -Lipschitz on \mathcal{X} if and only if, for all $x \in \mathcal{X}$, all $g \in \partial f(x)$, and all $y \in \mathcal{X}$, $\langle g, y - x \rangle \leq L\|x - y\|$. Conclude that, if $x \in \text{Int}(\mathcal{X})$, f is L -Lipschitz, and $g \in \partial f(x)$ then $\|g\|_* \leq L$.

Problem 4: Extensions for Gradient Descent

(A) In this exercise, you will show some generalizations of the basic gradient descent analysis we saw in class.

(A.1) Prove the following statement:

Proposition 1. Let Ω be a convex domain of radius R , and let f be a convex function on Ω . Let $x_0 \in \Omega$, and let $x_t = \Pi_\Omega(x_{t-1} - \eta g_t)$, where $\mathbb{E}[g_t | g_1, \dots, g_{t-1}] \in \partial f(x_{t-1})$, and $\sup_t \mathbb{E}[\|g_t\|^2] \leq L^2$ and $\eta = \frac{LR}{\sqrt{T}}$. Prove that

$$\mathbb{E}[f(\frac{1}{T} \sum_{t=1}^T x_t)] \leq \inf_{x \in \Omega} f(x) + \dots \quad (2)$$

You fill in the

(A.2) Prove the following statement:

Proposition 2. Let Ω be a convex domain of radius R , Let f_1, f_2, \dots, f_T be L -Lipschitz, convex functions on Ω . Given any $x_0 \in \Omega$, let $x_t = \Pi_\Omega(x_{t-1} - \eta g_t)$, where $g_t \in \partial f_{t-1}(x_{t-1})$, and $\eta = \frac{LR}{\sqrt{T}}$. Prove that

$$\frac{1}{T} \sum_{t=1}^T f_t(x_t) \leq \inf_{x \in \Omega} \frac{1}{T} \sum_{t=1}^T f_t(x) + \dots \quad (3)$$

You fill in the

- (B) In this problem we show that in the stochastic setting, smoothness of the function f does not help. Let $\Omega = [-1, 1]$, let σ be a random variable with $\mathbb{Pr}[\sigma = 1] = \mathbb{Pr}[\sigma = -1] = 1/2$, fix an $\epsilon \in (0, 1/4)$. Let z_1, z_2, \dots, z_T be T i.i.d random variables, such that $z_i | \sigma$ are mutually independent, and

$$\mathbb{Pr}[z_i = 1 | \sigma] = 1/2 + \sigma\epsilon \text{ and } \mathbb{Pr}[z_i = -1 | \sigma] = 1/2 - \sigma\epsilon \quad (4)$$

You will need the following information

Lemma 1. Let σ and z_1, z_2, \dots, z_T be as above. Then there exists a universal constant C such that, if $T \leq C\epsilon^{-2}$, any algorithm which returns an estimate $\hat{\sigma}$ of σ from observing z_1, z_2, \dots, z_T satisfies $\mathbb{Pr}[\hat{\sigma} \neq \sigma] \geq \frac{1}{4}$, where \mathbb{Pr} is taking over the randomness in σ, z_1, \dots, z_T , and any randomness in the algorithm.

- (B.1) Construct a function on f_σ such that $\mathbb{E}[z_i | \sigma] = \nabla f_\sigma(x)$ for all $x \in \Omega$. What is the optimum x_σ^* of f_σ ? What is the “smoothness” of f_σ ?
- (B.2) Show that there is a universal constant C' such that, for $T \leq C'\epsilon^{-2}$, $\mathbb{E}[f_\sigma(x_{T+1}) - \min_{x \in [-1, 1]} f_\sigma(x)] \geq \epsilon$, where \mathbb{E} is taken over the randomness in σ, z_1, \dots, z_T , and any randomness in the algorithm.

Problem 5: Generalized Projections

In this problem, we introduce a useful generalization of gradient descent. Let $\mathcal{X} \subseteq \mathcal{D} \subseteq \mathbb{R}^n$ be convex sets, and let $\Phi : \mathcal{D} \rightarrow \mathbb{R}^n$ be a strictly convex, continuously differentiable map such that $\|\nabla \Phi(x)\|$ diverges on $\text{Bd}(\mathcal{D})$, and for any sequence $x_n \in \mathcal{D}$ such that $\lim \|x_n\| = \infty$, and $\nabla \Phi(\mathcal{D}) = \mathbb{R}^n$. We call Φ a *mirror map*.

- (A) Define the *Bregman Divergence*

$$D_\Phi(x, y) = f(x) - f(y) - \nabla f(y)^\top (x - y) \quad (5)$$

and the associated Φ projection

$$\Pi_{\mathcal{X}}^\Phi(y) := \arg \min_{x \in \mathcal{X}} D_\Phi(x, y) \quad (6)$$

Show that $\Phi(x) = \frac{1}{2}\|x\|_2^2$ is a mirror map for $\mathcal{D} = \mathbb{R}^n$, and compute $D_\Phi(x, y)$ and explain what $\Pi_{\mathcal{X}}^\Phi(y)$ corresponds to

- (B) Prove that, for all $x \in \mathcal{X}$ and $y \in \mathcal{D}$,

$$(\nabla \Phi(\Pi_{\mathcal{X}}^\Phi(y)) - \nabla \Phi(y))^\top (\Pi_{\mathcal{X}}^\Phi(y) - x) \leq 0 \quad (7)$$

and conclude that

$$D_\Phi(x, \Pi_{\mathcal{X}}^\Phi(y)) + D_\Phi(\Pi_{\mathcal{X}}^\Phi(y), y) \leq D_\Phi(x, y) \quad (8)$$

What does this reduce to when $\Phi(x) = \frac{1}{2}\|x\|_2^2$? For the above, you may use the following lemma:

Lemma 2. Let f be convex, and let \mathcal{X} be a closed convex set on which f is differentiable. Then $x^* \in \arg \min_{x \in \mathcal{X}} f(x)$, if and only if, for all $x \in \mathcal{X}$, $\nabla f(x^*)^\top (x^* - y) \leq 0$ for all $y \in \mathcal{X}$.

- (C) Consider the following algorithm, known as mirror descent. Let $\mathcal{X} \subset \mathcal{D}$ and Φ be as above, let $f : \mathcal{X} \rightarrow \mathbb{R}$ be convex, let $x_1 \in \mathcal{X}$. Fix an $\eta > 0$. For $t \geq 1$, define y_{t+1} such that $\nabla \Phi(y_{t+1}) - \nabla \Phi(x_t) = -\eta g_t$, where $g_t \in \partial f(x_t)$. Prove the following:

Theorem 2. Let $\|\cdot\|$ be an *arbitrary* norm on \mathcal{X} , and suppose that Φ is a κ strongly-convex mirror map with respect to $\|\cdot\|$ on \mathcal{X} . Suppose that f is L -Lipschitz with respect to $\|\cdot\|$. Prove that

$$f\left(\frac{1}{T} \sum_{s=1}^T x_s\right) - \min_{x \in \mathcal{X}} f(x) \leq \frac{D(x^*, x_1)}{\eta} + \eta \frac{L^2 T}{\kappa} \quad (9)$$

Recall that Φ is κ -strongly convex with respect to $\|\cdot\|$ if and only $\Phi(x) - \Phi(y) \leq \nabla \Phi(x)^\top (x - y) + \frac{\kappa}{2} \|x - y\|^2$.

- (D) A common setup for mirror descent is on the simplex, where $\mathcal{D} = \{x : x[i] > 0 \forall i \in [n]\}$, $\mathcal{X} := \{x \in \mathcal{D} : \|x\|_1 = 1\}$, and $\Phi(x) = \sum_i x[i] \log x[i]$. Given an iterate x_t , compute the updates y_{t+1} and x_{t+1} . Here, $x[i]$ is the i -th coordinate of x .

Background

- (A) A ball of radius ϵ about x is the set $\{y : \|y - x\|_2 \leq \epsilon\}$. One can also consider balls with other norms, but they are all qualitatively equivalent to the Euclidean norm.
- (B) For a set $\mathcal{X} \subset \mathbb{R}^n$, its closure $\overline{\mathcal{X}}$ is defined as the set of all $x \in \mathbb{R}^n$ (not necessarily in \mathcal{X}) such that, for all $\epsilon > 0$, there exists a $y \in \mathcal{X}$ such that $\|x - y\| \leq \epsilon$. In other words, for every $\epsilon > 0$, the ball of radius ϵ around x intersects \mathcal{X} . $\text{Int}(\mathcal{X})$ is defined as the set of all points $x \in \mathcal{X}$ such that there exists an $\epsilon > 0$ for which, for all $y : \|x - y\| \leq \epsilon, y \in \mathcal{X}$; in other words, for some $\epsilon > 0$, the ball of radius $\epsilon > 0$ around x lies entirely in \mathcal{X} . Lastly, we define the boundary $\text{Bd}(\mathcal{X}) := \overline{\mathcal{X}} - \text{Int}(\mathcal{X}) = \{x \in \overline{\mathcal{X}} : x \notin \text{Int}(\mathcal{X})\}$.
- (C) A set is said to be *open* if $\mathcal{X} = \text{Int}\mathcal{X}$, and *closed* if $\mathcal{X} \supseteq \text{Bd}(\mathcal{X})$. A set $\mathcal{X} \subset \mathbb{R}^n$ is called compact if and only if it is closed and bounded.
- (D) Given a set of real numbers $\{a_i\}_{i \in I}$ (here I is an index set), $\sup_{i \in I} \{a_i\}$ is the smallest $a \in \mathbb{R}$ such that $a \geq a_i$ for all $i \in I$. If there is no such smallest a , $\sup_{i \in I} \{a_i\} = \infty$. Otherwise, $\sup_{i \in I} \{a_i\} = a_* \in \mathbb{R}$, and for every $\epsilon > 0$, there exists some $i = i(\epsilon) \in I$ such that $a_i \geq a_* - \epsilon$.
- (E) When there exists an i_* such that $a_{i_*} = \sup_{i \in I} \{a_i\}$, we say that the supremum is attained, and may replace \sup with \max for maximum. A finite set always has a maximum. When a maximum exists, we write $\arg \max_{i \in I} \{a_i\} := \{a_i : i \in I, a_i = \{\sup_{i' \in I} a_{i'}\}\}$ to denote the *set* of maximizers.
- (F) $\inf_{i \in I} \{a_i\}$ is defined as the least $a \in \mathbb{R}$ such that $a_i \geq a$ for all $i \in I$, and analogous properties hold.
- (G) Defining $f(x) = \sup_{i \in I} f_i(x)$, means that for every x , compute $\sup_{i \in I} \{f_i(x)\}$.
- (H) A norm is $\|\cdot\|$ is a function from $\mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ such that $\|\alpha x\| = |\alpha| \|x\|$ for any $\alpha \in \mathbb{R}$, $\|x + y\| \leq \|x\| + \|y\|$, and $\|x\| \geq 0$, and $\|x\| = 0 \iff x = 0$.
- (I) A sequence x_n is said to converge to a limit x_* if, for every $\epsilon \geq 0$, there is an $N = N(\epsilon)$ sufficiently large that $\|x_n - x_*\| \leq \epsilon$ for all $n \geq N$. We then write $\lim_{n \rightarrow \infty} x_n = x_*$.
- (J) If f is continuous and $\lim_{n \rightarrow \infty} x_n = x_*$, then $\lim_{x_n \rightarrow \infty} f(x_n) = f(x_*)$. If f is continuous and \mathcal{X} is compact, then $-\infty < \inf_{x \in \mathcal{X}} f(x) \leq \sup_{x \in \mathcal{X}} f(x) < \infty$. Moreover, there exist x_- and $x_+ \in \mathcal{X}$ such that $f(x_-) = \inf_{x \in \mathcal{X}} f(x)$ and $x_+ = \sup_{x \in \mathcal{X}} f(x)$; hence, $\arg \min_{x \in \mathcal{X}} f(x)$ and $\arg \max_{x \in \mathcal{X}} f(x)$ are well-defined, and we can replace \sup and \max with \inf and \min .