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 30, 2018

Problem 1: Existence of the Subgradients

(A) Let $\mathcal{X}$ be a convex set. Prove that that given any convex function $f\colon \mathcal{X} \to \mathbb{R}$ and any $x \in \mathcal{X}$, there exists at least one vector g, called a *subgradient* of f at x, such that $f(y) \geqslant f(x) + \langle g, y - x\rangle$ 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, $\mathrm{Epi}(f) := \{(x,t) \in \mathcal{X} \times \mathbb{R} : f(x) \leqslant t\}$. Prove the $\mathrm{Epi}(f)$ 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 $x \in \mathrm{Bd}(\mathcal{C})$. Then, there exists a nonzero $w \in \mathbb{R}^n$ such that, for all $y \in \mathcal{C}$, $\langle w, y - x\rangle \geqslant 0$.

Hint: Find two (not-necessarily compact!) convex sets to apply the separating hyperplane theorem. You might want $\mathrm{Int}(\mathcal{C})$ *to be one of them - and you should check that* $\mathrm{Int}(\mathcal{C})$ *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 $f : \mathcal{X} \to \mathbb{R}$ 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 necssarily unique, but *prove* 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) Here, you will be asked to show a partial converse to the above statement. Suppose that $\mathcal{X}$ is a compact set, with non-empty interior, and let $f(x) = \max_{w \in \mathcal{X}} \langle w, x \rangle$. Prove that $\partial f(x) \subset \text{Conv}\{w : \langle w, x \rangle = f(x)\}$. Hint: A key step is to show that if v satisfies $\max_{w \in \mathcal{X} \cup \{v\}} \langle v, x \rangle = \max_{w \in \mathcal{X}} \langle v, x \rangle$ for all $x \in \mathbb{R}^n$, then the separating hyperplane theorem implies $v \in \text{Conv}(\mathcal{X})$.

(F) Using the previous two subproblems, derive a formula for $\partial \| \cdot \|$, where $\| \cdot \|$ is an arbitrary norm. (Hint: Use 1.C)

Problem 3: Subgradients of Norms

(A) Subgradient of the ℓ_1 and ℓ_∞-norms

(A.1) Prove that, for all $x \in \mathbb{R}^n$, $\|x\|_1 = \sup_{y : \|y\|_\infty \leqslant 1} \langle x, y \rangle$, $\|x\|_\infty = \sup_{y : \|y\|_1 \leqslant 1} \langle x, y \rangle$.

(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 $\sum_{i=1}^{\min(n,m)} \sigma_i(AB) \leqslant \sum_{i=1}^{\min(n,m)} \sigma_i(A)\sigma_i(B)$ 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 $X \in \mathbb{R}^{m\times n}$,

$$\|X\|_{\mathrm{op}} := \max_{Y\in\mathbb{R}^{m\times n}\|Y\|_{\mathrm{nuc}}\leqslant 1} \langle X, Y\rangle \text{ and } \|X\|_{\mathrm{nuc}} = \max_{Y\in\mathbb{R}^{m\times n}:\|Y\|_{\mathrm{op}}\leqslant 1} \langle X, Y\rangle\ , \tag{1}$$

where $\|X\|_{\mathrm{op}} := \sigma_{\max}(X)$, $\|Y\|_{\mathrm{nuc}} := \sum_{i=1}^{\min(n,m)} \sigma_i(Y)$, and $\langle X, Y\rangle := \mathrm{tr}(X^\top Y)$. 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 $\sum_i \lambda_i(A) \leqslant \sum_i \sigma_i(A)$ for any squared matrix A.

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

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

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

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

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

Problem 4: Extensions for Gradient Descent

(A) In this exercise, you will show some generalizations of the basic grdient 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 $\mathbb{E}[g_t|g_1, \ldots, g_{t-1}] \in \partial f(x_{t-1})$, and $\sup_t \mathbb{E}[\|g_t\|^2] \leqslant L^2$ and $\eta = \frac{LR}{\sqrt{T}}$. Prove that

$$\mathrm{Exp}[f(\frac{1}{T}\sum_{t=1}^{T} x_t)] \leqslant \inf_{x\in\Omega} f(x) + \ldots \tag{2}$$

You fill in the ....

(A.2) Prove the following statement:

Proposition 2. Let Ω be a convex domain of radius R, Let $f_1, f_2, \ldots, f_T$ 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

$$\frac{1}{T}\sum_{t=1}^{T} f_t(x_t) \leqslant \inf_{x\in\Omega} \frac{1}{T}\sum_{t=1}^{T} f_t(x) + \ldots \tag{3}$$

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 $\mathbb{P}\mathrm{r}[\sigma = 1] = \mathbb{P}\mathrm{r}[\sigma = -1] = 1/2$, fix an $\epsilon \in (0, 1/4)$. Let $z_1, z_2, \ldots, z_T$ be T i.i.d random variables, such that $z_i|\sigma$ are mutually independent, and

$$\mathbb{P}\mathrm{r}[z_i = 1|\sigma] = 1/2 + \sigma\epsilon \text{ and } \mathbb{P}\mathrm{r}[z_i = -1|\sigma] = 1/2 - \sigma\epsilon \tag{4}$$

You will need the following information

Lemma 1. Let σ and $z_1, z_2, \ldots, z_T$ be as above. Then there exists a universal constant C such that, if $T \leqslant C\epsilon^{-2}$, any algorithm which returns an estimate $\widehat{\sigma}$ of σ from observing $z_1, z_2, \ldots, z_T$ satisfies $\mathbb{P}\mathrm{r}[\widehat{\sigma} \neq \sigma] \geqslant \frac{1}{4}$, where $\mathbb{P}\mathrm{r}$ is taking over the randomness in σ, $z_1, \ldots, Z_T$, 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 $x^*_\sigma$of f_σ? What is the "smoothness" of f_σ?

(B.2) Show that there is a universal constant C' such that, for $T \leqslant C'\epsilon^{-2}$, $\mathbb{E}[f_\sigma(x_{T+1}) - \min_{x\in[-1,1]} f_\sigma(x)] \geqslant \epsilon$, where Exp_σ is taken over the randomness in σ, $z_1, \ldots, Z_T$, 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 $\Phi : \mathcal{D} \to \mathbb{R}$ be a strictly convex, continuously differentiable map such that $\|\nabla\Phi(x)\|$ diverges on $\mathrm{Bd}(\mathcal{D})$, 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*

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

and the associated Φ projection

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

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 $\Pi^\Phi_{\mathcal{X}}(y)$ corresponds to

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

$$(\nabla\Phi(\Pi^\Phi_{\mathcal{X}}(y)) - \nabla\Phi(y))^\top(\Pi^\Phi_{\mathcal{X}}(y) - x) \leqslant 0 \tag{7}$$

and conclude that

$$D_\phi(x, \Phi_x(y)) + D_\phi(\Phi_x(y), y) \leqslant D_\Phi(x, y) \tag{8}$$

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 $x \in \mathcal{X}$, $\nabla f(x^*)^\top (x^* - y) \leqslant 0$ 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 $f : \mathcal{X} \to \mathbb{R}$ be convex, let $x_1 \in \mathcal{X}$. Fix an $\eta > 0$. For $t \geqslant 1$, 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

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

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

(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 $\{y : \|y - x\|_2 \leqslant \epsilon\}$. 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 $\|x - y\| \leqslant \epsilon$. In other words, for every $\epsilon > 0$, the ball of radius ϵ around x intersects $\mathcal{X}$. $\mathrm{Int}(\mathcal{X})$ is defined as the set of all points $x \in \mathcal{X}$ such that there exists an $\epsilon > 0$ for which, for all $y : \|x - y\| \leqslant \epsilon$, $y \in \mathcal{X}$; it other words, for some $\epsilon > 0$, the ball of radius $\epsilon > 0$ around x lies entirely in $\mathcal{X}$. Lastly, we define the boundary $\mathrm{Bd}(\mathcal{X}) := \overline{\mathcal{X}} - \mathrm{Int}(\mathcal{X}) = \{x \in \overline{\mathcal{X}} : x \notin \mathrm{Int}(\mathcal{X})\}$.

(C) A set is said to be *open* if $\mathcal{X} = \mathrm{Int}\mathcal{X}$, and closed if $\mathcal{X} \supseteq \mathrm{Bd}(\mathcal{X})$. 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 $a \geqslant a_i$ for all $i \in I$. If there is no such smallest a, $\sup\{a_i\}_{i\in I} = \infty$. Otherwise, $\sup\{a_i\}_{i\in I} = a_* \in \mathbb{R}$, and for every $\epsilon > 0$, there exists some $i = i(\epsilon) \in I$ such that $a_i \geqslant a_* - \epsilon$.

(E) When there exists an i_* such that $a_{i_*} = \sup\{a_i\}_{i\in I}$, 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) $\inf\{a_i\}_{i\in I}$ is defined as the least $a \in \mathbb{R}$ such that $a_i \geqslant a$ 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 $\mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$ such that $\|\alpha x\| = |\alpha|\|x\|$ for any $\alpha \in \mathbb{R}$, $\|x + y\| \leqslant \|x\| + \|y\|$, and $\|x\| \geqslant 0$, and $\|x\| = 0 \iff x = 0$.

(I) A sequence x_n is said to converge to a limit x_* if, for every $\epsilon \geqslant 0$, there is an $N = N(\epsilon)$ sufficiently large that $\|x_n - x_*\| \leqslant \epsilon$ for all $n \geqslant N$. We then write $\lim_{n\to\infty} x_n = x_*$.

(J) If f is continuous and $\lim_{n\to\infty} x_n = x_*$, then $\lim_{x_n\to\infty} f(x_n) = f(x_*)$. If f is continuous and $\mathcal{X}$ is compact, then $-\infty < \inf_{x\in\mathcal{X}} f(x) \leqslant \sup_{x\in\mathcal{X}} < \infty$. Moreover, there exist x_- and $x_+ \in \mathcal{X}$ such that $f(x_i) = \inf_{x\in\mathcal{X}} f(x)$ 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.