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 21, 2018

Problem 1: Existence of the Subgradients

Let $\mathcal{X}$ be a convex set. Prove that that given any convex function $f : \mathcal{X} \to \mathbb{R}$ and any $x \in \mathcal{X}$, there 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 $y \in \mathcal{X}$. You will do so following the steps below.

i) 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.

ii) Recall the following definitions from real analysis

Definition 1 (Boundary and Interior)**.** For a set $\mathcal{X} \subset \mathbb{R}^d$, its closure $\overline{\mathcal{X}}$ is defined as the set of all $x \in \mathbb{R}^d$ (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 $\mathrm{Bd}(\mathcal{X}) := \overline{\mathcal{X}} - \mathrm{Int}(\mathcal{X}) = \{x \in \overline{\mathcal{X}} : x \notin \mathrm{Int}(\mathcal{X})\}$.

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 $\mathcal{C} \subset \mathbb{R}^d$ be a convex set, and let $x \in \mathrm{Bd}(\mathcal{C})$. Then, there exists a $w \in \mathbb{R}^d$ 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*

iii) Using part i) and ii), prove the existence of a subgradient.

iv) Let $\{g_\alpha\}$ be a (possibly infinite, uncountable) family of convex functions, and suppose that $g_\alpha(x) < \infty$ for all $x \in \mathcal{X}$. Show that $g(x) := \sup_\alpha g_\alpha(x)$ is convex on $\mathcal{X}$ (you may assume $g(x)$ is finite).

v) Using what we've proven about subgradients, prove that a function $g : \mathcal{X} \to \mathbb{R}$ is convex if and only if it can be written as the supremum of affine functions (e.g. supremum of functions of the form $g_\alpha(x) = \langle a_\alpha, x \rangle + b_\alpha$)

Problem 2: Properties of Subgradients

i) Show by way of example that the subgradient is not necssarily unique, but that the set of all subgradients is convex. We will denote this *set* $\partial f(x)$.

ii) Show that if f is differentiable at x, $\partial f(x) = \{\nabla f(x)\}$.

iii) Show that if $g_1 \in \partial f_1(x)$ and $g_2 \in \partial f_2(x)$, then $g_1 + g_2 \in \partial(f_1 + f_2)(x)$.

vi) Consider the following theorem

Theorem 2. Let Ω be a convex domain, and let $f(x) = \sup_\alpha g_\alpha(x)$, where g_α are convex functions and finite on Ω. Suppose f is also finite on Ω. Then for all $x \in \Omega$, $\partial f = \mathrm{Conv}\{\partial g_\alpha(x) | \sup_x g_\alpha(x) = f(x)\}$

Prove one direction of the theorem, $\partial f \subseteq \mathrm{Conv}\{\partial g_\alpha(x) | \sup_x g_\alpha(x) = f(x)\}$

v) (Hard) Prove the other direction of Theorem 2. Here you may want to work in epigraph space.

Problem 3: Subgradients of Norms

A) Subgradient of the L_1 and L_∞-norms

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

ii)) Compute $\partial \|x\|_1$ and $\partial \|x\|_\infty$

A) Subgradient of the L_1-norm

i)) 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 $B \in \mathbb{R}^{n \times m}$, prove the follow: 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.

ii)) 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 $\mathbb{R}^d$. Define the dual norm $\|x\|_* := \sup_{x:\|x\| \leqslant 1} \langle x, y \rangle$.

i) Show that the dual norm is a norm, and describe its subgradient.

ii) 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 $x \in \mathcal{X}$ and all $g \in \partial f(x)$, $\|g\|_* \leqslant L$.

Problem 4: Extensions for Gradient Descent

A) Generalizations of SGD

i)) 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 $\mathrm{Exp}[g_t | g_1, \ldots, g_{t-1}] \in \partial f(x_t)$, and $\sup_t \mathrm{Exp}[\|g_t\|^2] \leqslant L$. Prove that

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

You fill in the $\ldots$.

ii) 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 $g_t \in \partial f_t(x_t)$, and $\eta = \frac{LR}{\sqrt{T}}$. Prove that

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

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 resturns an estimate $\widehat{\sigma}$ of σ from observing $z_1, z_2, \ldots, z_T$ satisfies $\mathbb{P}\mathrm{r}[\widehat{\sigma} \neq \sigma] \geqslant \frac{1}{4}$.

i) Construct a function on f_σ such that $\mathrm{Exp}[z_i|\sigma] = \nabla f_\sigma(x)$ for all $x \in \Omega$. What is the optimum $x^*_\sigma$of f_σ? What is the “smoothness” of f_σ.

ii) Show that there universal constants c such that the following hold: Fix a $T \in \mathbb{N}$, and let $\mathcal{A}$ be an algorithm which is allowed to make T queries $x_t \in [-1, 1]$ and g_t, where $\mathcal{A}$ decides x_t, and recieves responses g_t such that $\mathrm{Exp}[g_t] = \nabla f(x_t)$ and $|g_t| \leqslant 1$ as.s. Then, there is a 0-smooth, 1-Lipschitz function f and a mechanism generating responses g_t such that the iterate x_T satisfies

$$\mathrm{Exp}[f(x_1)] - \inf_{x\in[-1,1]} f(x) \geqslant c/\sqrt{T} \tag{5}$$

Mirror Descent

In this problem, we introduce a useful generalization of gradient descent. Let $\mathcal{X} \subseteq \mathcal{D} \subseteq \mathbb{R}^d$ 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 $\nabla\Phi(\mathcal{D}) = \mathbb{R}^d$. We call Φ a *mirror map*.

A) Define the *Bregman Divergence*

$$D_\Phi(x,y) = f(x) - f(y) - \nabla f(y)^\top (x-y) \tag{6}$$

and the associated Φ projection

$$\Pi^\Phi_{\mathcal{X}}(y) := \arg\min_{x\in\mathcal{X}} D_{\Phi(x,y)} \tag{7}$$

Show that $\Phi(x) = \frac{1}{2}\|x\|_2^2$ is a mirror map, 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 $y \in \mathcal{D}$,

$$(\nabla\Phi(\Pi^\Phi_{\mathcal{X}}(y)) - \nabla\Phi(y))^\top (\Pi^\Phi_{\mathcal{X}}(y) - x) \leqslant 0 \tag{8}$$

and conclude that

$$D_\phi(x, \Phi_x(y)) + D_\phi(\Phi_x(y), y) \leqslant D_\Phi(x,y) \tag{9}$$

What does this reduce to when $\Phi(x) = \frac{1}{2}\|x\|_2^2$?

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

$$f(\sum_{s=1}^{T} x_s) - \min_{x\in\mathcal{X}} f(x) \leqslant \frac{D(x,\pi)}{\eta} + \eta\frac{L^2 T}{\kappa} \tag{10}$$

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 $\mathcal{D} : \{x : x_i > 0 \forall i \in [d]\}$, and $\mathcal{X} := \{x \in \mathcal{D} : \|x\| = 1\}$. Given an iterate x_t, compute the updates y_{t+1} and x_{t+1}.