

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 22, 2018

Problem 1: Existence of the Subgradients

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 $y \in \mathcal{X}$. You will do so following the steps below.

i) 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.

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 $\|x - y\| \leq \epsilon$. 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 $\text{Bd}(\mathcal{X}) := \overline{\mathcal{X}} - \text{Int}(\mathcal{X}) = \{x \in \overline{\mathcal{X}} : x \notin \text{Int}(\mathcal{X})\}$.

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 $x \in \text{Bd}(\mathcal{C})$. Then, there exists a $w \in \mathbb{R}^d$ such that, for all $y \in \mathcal{C}$, $\langle w, y - x \rangle \geq 0$.

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

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 $g : \mathcal{X} \rightarrow \mathbb{R}$ is convex if and only if it can be written as the supremum of affine functions (e.g. supremum of functions of the form $g_\alpha(x) = \langle a_\alpha, x \rangle + b_\alpha$)

Problem 2: Properties of Subgradients

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

ii) Show that a convex f is differentiable at x if and only if $\partial f(x) = \{\nabla f(x)\}$.

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) Let $f(x) = \sup_\alpha g_\alpha(x)$ which g_α convex. Show that $\text{Conv}\{\partial g_\alpha(x) | g_\alpha(x) = f(x)\} \subseteq \partial f$.

(v) Prove that if $g_\alpha(x) = w_\alpha^\top x + b_\alpha$ are a compact family of affine functions (i.e. $\{(w_\alpha, b_\alpha)\} \subset \mathbb{R}^{d+1}$ is compact), then the converse is true, namely $\partial f \subset \text{Conv}\{\partial g_\alpha(x) | g_\alpha(x) = f(x)\}$.

Problem 3: Subgradients of Norms

A) Subgradient of the L_1 and L_∞ -norms

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

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 $\sum_{i=1}^{\min(n,m)} \sigma_i(AB) \leq \sum_{i=1}^{\min(n,m)} \sigma_i(A)\sigma_i(B)$ for all $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$, prove the follow: 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.

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

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 and only if, for all $x \in \mathcal{X}$ and all $g \in \partial f(x)$, $\|g\|_* \leq L$.

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 $\text{Exp}[g_t | g_1, \dots, g_{t-1}] \in \partial f(x_t)$, and $\sup_t \text{Exp}[\|g_t\|^2] \leq L$ and $\eta = \frac{LR}{\sqrt{T}}$. Prove that

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

You fill in the

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

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

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.

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

ii) Show that there exists a universal constant 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 receives responses g_t such that $\text{Exp}[g_t] = \nabla f(x_t)$ and $|g_t| \leq 1$ a.s. Then, there is a 0-smooth, 1-Lipschitz function f and a mechanism for generating noisy subgradients g_t such, for any algorithm using g_t as gradient queries, the iterate x_{T+1} satisfies

$$\text{Exp}[f(x_{T+1})] - \inf_{x \in [-1, 1]} f(x) \geq c/\sqrt{T} \quad (5)$$

Problem 5: Generalized Projections

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 $\Phi : \mathcal{D} \rightarrow \mathbb{R}$ be a strictly convex, continuously differentiable map such that $\|\nabla \Phi(x)\|$ diverges on $\text{Bd}(\mathcal{D})$, and $\nabla \Phi(\mathcal{D}) = \mathbb{R}^d$. We call Φ a *mirror map*.

A) Define the *Bregman Divergence*

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

and the associated Φ projection

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

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

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

$$(\nabla \Phi(\Pi_{\mathcal{X}}^{\Phi}(y)) - \nabla \Phi(y))^{\top} (\Pi_{\mathcal{X}}^{\Phi}(y) - x) \leq 0 \quad (8)$$

and conclude that

$$D_{\Phi}(x, \Phi_x(y)) + D_{\Phi}(\Phi_x(y), y) \leq D_{\Phi}(x, y) \quad (9)$$

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 $f : \mathcal{X} \rightarrow \mathbb{R}$ 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

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

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

D) A common setup for mirror descent is on the simplex, where $\mathcal{D} = \{x : x_i \geq 0 \forall i \in [d]\}$, and $\mathcal{X} := \{x \in \mathcal{D} : \|x\|_1 = 1\}$. Given an iterate x_t , compute the updates y_{t+1} and x_{t+1} .