Convex Optimization: Homework 1

Instructor: Yuanzhi Li
Carnegie Mellon University
Due: February 2th

1 Convex sets

In this problem you want to show (meaning: prove in mathematical language) that some commonly used sets are convex.

1.1 The polytope (10 points)

A d-dimensional polytope $\mathcal{P}$ is a set in $\mathbb{R}^d$, defined as the set of points $x \in \mathbb{R}^d$ satisfying the following constraints: For an integer $m > 0$, for m vectors $a_1, \cdots, a_m \in \mathbb{R}^d$ and m values $b_i \in \mathbb{R}$:

$$\forall i \in [m] : \langle a_i, x \rangle \leq b_i \tag{1.1}$$

Show that $\mathcal{P}$ is either an empty set or a convex set.
(Hint: Use the basic definition for convex set, the proof should be within a few lines)

1.2 The unit ball (10 points)

A d-dimensional unit ball $\mathcal{B}$ is a set in $\mathbb{R}^d$, defined as the set of points $x \in \mathbb{R}^d$ satisfying the following constraints:

$$\sum_{i \in [d]} x_i^2 \leq 1$$

Where x_i is the i-th coordinate of x.
Show that $\mathcal{B}$ is a convex set.
(Hint: You can use the following fact: For any value $a, b \in \mathbb{R}$, $2ab \leq a^2 + b^2$. Now, use the basic definition for convex set and apply this inequality d times on each coordinate)

1.3 The linear transformation (10 points)

Suppose $\mathcal{D}$ is a convex set in $\mathbb{R}^d$ and $A \in \mathbb{R}^{d \times d}$ is a matrix and $b \in \mathbb{R}^d$ is a vector.
Show that the following set is also convex:

$$\{x \in \mathbb{R}^d \mid Ax + b \in \mathcal{D}\}$$

1.4 Ellipsoid (10 points)

Use the previous two subproblems to show that an Ellipsoid is convex: An Ellipsoid $\mathcal{E}$ in $\mathbb{R}^d$ is a set defined as: for some matrix A in $\mathbb{R}^{d \times d}$,

$$\mathcal{E} = \{x \in \mathbb{R}^d \mid x^\top A^\top A x \leq 1\} \tag{1.2}$$

2 Convex functions

In this problem you want to show that some commonly used functions are convex. You can use the basic definition or the alternative definition mentioned in class in the second lecture.

2.1 1-d convex function (10 points)

Show that the following functions: $f(x) = x$, $f(x) = e^x$, $f(x) = x \log x$, $f(x) = \log(1 + e^x)$ are convex functions.

(Hint: You can use the Hessian condition of convex functions in Lecture 2)

2.2 The linear transformation (10 points)

Suppose f is a convex function over $\mathbb{R}^d$, show that for any positive integer d', any matrix $A \in \mathbb{R}^{d' \times d}$ and any vector $b \in \mathbb{R}^{d'}$, $g(x) := f(Ax + b)$ is a convex function over $\mathbb{R}^{d'}$ as well.

2.3 The non-negative summation (10 points)

Suppose $f_1, \cdots, f_m$ are convex functions over $\mathbb{R}^d$, show that for any $\lambda_1, \cdots, \lambda_m \geq 0$,

$$f(x) := \sum_{i \in [m]} \lambda_i f_i(x)$$

is a convex function over $\mathbb{R}^d$.

2.4 The max operation (10 points)

Suppose $f_1, \cdots, f_m$ are convex functions over $\mathbb{R}^d$, show that

$$f(x) = \max \{f_1(x), \cdots, f_m(x)\}$$

is a convex function over $\mathbb{R}^d$.

What about $g(x) = \min \{f_1(x), \cdots, f_m(x)\}$?

2.5 Logistic Regression (10 points)

Show that the logistic regression problem considered in Lecture 1 is convex.

(Hint: Combine the linear transformation, 1-d function $\log(1 + e^x)$ and the non-negative summation problems)

3 Smoothness of the functions

3.1 1-d smooth function (10 points)

Show that the following functions are L-smooth for some $L \leq 4$: $f(x) = \cos(x)$, $f(x) = \tanh(x)$ and $f(x) = \log(1 + e^x)$.

(Hint: You can use the Hessian condition for smoothness in Lecture 2)

3.2 Linear transformation (10 points)

Show that if f is a L-smooth function (not necessarily convex) in $\mathbb{R}^d$, then for every orthogonal matrix $U \in \mathbb{R}^{d\times d}$ (meaning $UU^\top = I$) and every vector $b \in \mathbb{R}^d$,

$$g(x) = f(Ux + b)$$

is a L-smooth function over $\mathbb{R}^d$ as well.

4 Gradient Descent and (Basic) Mirror Descent

4.1 Gradient Descent with a Politician (10 points)

Suppose there is an unknown L-smooth convex function f which is in the hands of a politician. At every iteration t, you can query the politician the gradient of this function at a point x_t. However, the politician will not give you $\nabla f(x_t)$ directly, instead, he will simply give you a direction g_t such that $\langle \nabla f(x_t), g_t\rangle > 0$ when $\|\nabla f(x_t)\|_2 \neq 0$. In the unlikely event that $\|\nabla f(x_t)\|_2 = 0$, the politician will give you 0 and retire.

Show that as long as $g_t \neq 0$, there is an $\eta > 0$ such that the update $x_{t+1} = x_t - \eta g_t$ satisfies:

$$f(x_{t+1}) < f(x_t)$$

Bonus (10 points): How to find this η? Give an algorithm that is as efficient as possible.

4.2 The distance to x^* (20 points)

Show that for convex and L-smooth function f, when $\eta \leq \frac{1}{L}$, the update rule $x_{t+1} = x_t - \eta\nabla f(x_t)$ is indeed decreasing the distance from x_t to x^*, in the sense that

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2$$

(Hint: Replace $x_{t+1} - x^*$ by $x_t - \eta\nabla f(x_t) - x^*$ on the left hand side, then use the fact that $\|x_t - \eta\nabla f(x_t) - x^*\|_2^2 = \|x_t - x^*\|_2^2 - 2\eta\langle\nabla f(x_t), x_t - x^*\rangle + \eta^2\|\nabla f(x_t)\|_2^2$.)

4.3 Bonus problem: Optimization with Error (50 points)

Li Hua is a student in this course. After listening to the professor's lecture, he found a major flaw to apply the proof of the gradient descent algorithm in Lecture 2 in real-world problems: Li Hua points out that proof only applies when the update $x_{t+1} = x_t - \nabla f(x_t)$ is *exact*. However, a real machine only operates on finite precision and the update, in most of the cases, can not be computed exactly. In mathematical language, Li Hua says that the actual update on a real machine is

$$x_{t+1} = x_t - \nabla f(x_t) + \delta_t$$

Where $\|\delta_t\|_2 \leq \delta$ for some $\delta > 0$ depending on the machine precision, is the error term. Li Hua claims that the professor cheated and the proof will not work in this case. Now, you would like to support your professor by showing that when f is convex and L-smooth, when $\eta \leq \frac{1}{2L}$ and $\delta \leq \frac{1}{10L}$, one still have:

$$x_T \leq x^* + \frac{\|x^* - x_0\|_2^2}{\eta T} + O(L\delta\|x_0 - x^*\|_2 + L\delta^2)$$