SI251 - Convex Optimization, Spring 2022
Homework 2

Due on Apr 10, 2022, before class

Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- You are required to write down all the major steps towards making your conclusions; otherwise you may obtain limited points ($\leq 20\%$) of the problem.
- Write your homework in English; otherwise you will get no points of this homework.
- Do your homework by yourself. Any form of plagiarism will lead to 0 point of this homework. If more than one plagiarisms during the semester are identified, we will prosecute all violations to the fullest extent of the university regulations, including but not limited to failing this course, academic probation, or expulsion from the university.
- No late submission will be accepted.
- If you have any doubts regarding the grading, you need to contact the instructor or the TAs within two days since the grade is announced.
- Handwritten assignment is acceptable, but we prefer a LaTex version.

I. Gradient Descent for Unconstrained Problems

1.(*Unconstrained optimization*) Let f be differentiable, m-strongly convex, M-smooth and with minimizer x^*. In class we proved geometric convergence of the error $\|x^t - x^*\|_2$. In this exercise, we explore how to prove convergence in the function value difference $f(x^t) - f(x^*)$ for gradient descent with step size $\alpha = 1/M$.

(1) Prove that:

$$f(x^{t+1}) - f(x^*) \leq f(x^t) - f(x^*) - \frac{1}{2M}\|\nabla f(x^t)\|_2^2 \tag{1}$$

This shows that we have a descent method. (5 points)

(2) Prove that:

$$\frac{m}{M}(f(x^t) - f(x^*)) \leq \frac{1}{M}\|\nabla f(x^t)\|_2^2 \tag{2}$$

(10 points)

Solution:(1) Prove that:

$$f\left(x^{l+1}\right) - f\left(x^*\right) \leq f\left(x^l\right) - f\left(x^*\right) - \frac{1}{2M}\left\|\nabla f\left(x^l\right)\right\|_2^2$$

This shows that we have a descent method. (5 points) Solution: We have by M-smoothness

$$f\left(x^{l+1}\right) - f\left(x^l\right) - \left\langle \nabla f\left(x^l\right), x^{l+1} - x^l \right\rangle \leq \frac{M}{2}\left\|x^{l+1} - x^l\right\|_2^2.$$

Furthermore, note that by our gradient descent method, $x^{l+1} - x^l = -\frac{1}{M}\nabla f\left(x^l\right)$. Substituting this in, we see that

$$f\left(x^{l+1}\right) - f\left(x^l\right) + \frac{1}{M}\left\|\nabla f\left(x^l\right)\right\|_2^2 \leq \frac{1}{2M}\left\|\nabla f\left(x^l\right)\right\|_2^2.$$

Rearranging, and adding $-f\left(x^*\right)$ to both sides, we obtain

$$f\left(x^{l+1}\right) - f\left(x^*\right) \leq f\left(x^l\right) - f\left(x^*\right) - \frac{1}{2M}\left\|\nabla f\left(x^l\right)\right\|_2^2$$

(2) From the characteristic of m-strong convexity, we have

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle \leq \frac{1}{2m}\|\nabla f(x) - \nabla f(y)\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

We substitute $y = x^l$, the output at step l, and $x = x^*$, the minimum. Since x^* is a minimum, $\nabla f\left(x^*\right) = 0$. Therefore

$$f\left(x^l\right) - f\left(x^*\right) \leq \frac{1}{2m}\left\|\nabla f\left(x^l\right)\right\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

Multiplying both sides by m/M we obtain the desired result

$$\frac{m}{M}\left(f\left(x^l\right) - f\left(x^*\right)\right) \leq \frac{1}{2M}\left\|\nabla f\left(x^l\right)\right\|_2^2$$

2.(*Backtracking line search.*) Suppose f is strongly convex with $mI \preceq \nabla^2 f(x) \preceq MI$. Let Δx be a decent direction at x. Show that the backtracking stopping condition holds for

$$0 < t \leq -\frac{\nabla f(x)^T \Delta x}{M\|\Delta x\|_2^2} \tag{3}$$

Use this to give an upper bound on the number of backtracking iterations. (10 points)
The upper bound $\nabla^2 f(x) \preceq MI$ implies

$$f(x + t\Delta x) \leq f(x) + t\nabla f(x)^T \Delta x + (M/2)t^2 \Delta x^T \Delta x$$

hence $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$ if

$$t(1 - \alpha)\nabla f(x)^T \Delta x + (M/2)t^2 \Delta x^T \Delta x \leq 0$$

i.e., the exit condition certainly holds if $0 \leq t \leq t_0$ with

$$t_0 = -2(1 - \alpha)\frac{\nabla f(x)^T \Delta x}{M \Delta x^T \Delta x} \geq -\frac{\nabla f(x)^T \Delta x}{M \Delta x^T \Delta x}$$

Assume $t_0 \leq 1$. Then $\beta^k t \leq t_0$ for $k \geq \log\left(1/t_0\right) / \log(1/\beta)$.

II. Gradient Descent for Constrained Problems

1. Consider a constrained optimization problem $min_{x\in\mathcal{C}} f(x)$ where $\mathcal{C}$ is a compact convex set, and f is convex and has a continuous derivative. The conditional gradient method with stepsizes $\{\alpha^t\}_{t=0}^{\infty}$ generates a sequence of the form

$$x^{t+1} = (1-\alpha^t)x^t + \alpha^t z^t \tag{4}$$

where $z^t \in \arg\, min_{z\in\mathcal{C}}\langle \nabla f(x^t), z\rangle$. Compute the form of these updates for the following cases:

(1) $\mathcal{C} = \{x \in \mathbb{R}^d \mid \|x\|_1 \leq 1\}$ (10 points)

(2) $\mathcal{C} = \{X \in \mathbb{R}^{d\times d} \mid \sum_{j=1}^{d} \sigma_j(X) \leq 1\}$ where $\sigma_j(X)$ is the j^{th} singular value. (10 points)

(1) Let z be a vector with $\|z\|_1 = 1$. As mentioned in class, the problem is equivalent to maximizing $\left\langle -\nabla f\left(x^l\right), z\right\rangle$. Then we have

$$\left\langle -\nabla f\left(x^l\right), z\right\rangle \leq \left\|\nabla f\left(x^l\right)\right\|_\infty \|z\|_1$$

by Cauchy-Schwartz, where inequality is obtained when

$$z = -\operatorname{sign}\left(\left[\nabla f\left(x^l\right)\right]_{i^*}\right) \epsilon_{i^*}$$

with $i^* = \arg\min_{i=1}\ldots,d - \left|\left[\nabla f\left(x^l\right)\right]_i\right|$ and e_i the standard basis vectors. The update is therefore given by

$$x^l = \left(1-\alpha^l\right)x^l + \alpha^l z^l = \left(1-\alpha^l\right)x^l + \alpha^l\left(-\operatorname{sign}\left(\left[\nabla f\left(x^l\right)\right]_{i^*}\right) e_{i^*}\right)$$

(2) We want to solve the following optimization problem

$$\arg\min_{Z\in C}\left\langle \nabla f\left(X^l\right), Z\right\rangle = \arg\min_{Z\in C} \operatorname{tr}\left(Z^T \nabla f\left(X^l\right)\right)$$

Using SVD, we can write $\nabla f\left(X^l\right) = U\Lambda V^T$, where U and V are unitary, and Λ is a diagonal matrix; also define $\hat{Z} = U^T Z V$. We have

$$\begin{aligned}
\arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(Z^T\nabla f\left(X^l\right)\right) &= \arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(Z^T\left(U\Lambda V^T\right)\right) \\
&= \arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(V^T Z^T U\Lambda\right) \\
&= \arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(\left(U^T Z V\right)^T \Lambda\right) \\
&= U\left[\arg\min_{Z\in C} \operatorname{tr}\left(\hat{Z}^T\Lambda\right)\right] V^T
\end{aligned}$$

where we used invariance of the trace operator under cyclic permutations. Now we seek to find a $\dot{Z}^*$ such that the above is minimized. Note that the off diagonal entries do not eect the trace since Λ is diagonal. Hence, w.l.o.g. we take $\tilde{Z}^+$to be diagonal, and the problem reduces to part (a), i.e. the problem of finding the diagonal vector $\hat{z}$ of Z given the diagonal vector of Λ which is equivalent to the singular vector v of $\nabla f\left(X^l\right)$, i.e.

$$\arg\min_{Z\in C} \operatorname{tr}\left(\hat{Z}^T\Lambda\right)\Big) = \operatorname{diag}\left(\arg\min_{\|\hat{A}_1\leq 1\|}\langle v, \hat{z}\rangle\right)$$

As a consequence, defining $i^* = \arg\max_i |\Lambda_{ii}|$, the minimum is attained by defining Z^* as:

$$\left[\hat{Z}^*\right]_{ii} = \begin{cases} 0 & \text{if } i \neq i^* \\ -\operatorname{sign}\left([\Lambda_{ii}]\right) & \text{if } i = i^* \end{cases}$$

Therefore, in our update, Z^t is given by $Z^l = UZ^*V^T$.

2. Given a vector $\mathbf{y} \in \mathbb{R}^d$, a matrix $\mathbf{Y} \in \mathbb{R}^{d\times m}$ and the parameter $\gamma \in \mathbb{R}_{++}$.

(1) Solve the following projection onto the probability simplex problem:

$$\begin{aligned} \min_x \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{s.t.} \quad & \mathbf{x} \in \Delta, \end{aligned} \tag{5}$$

where $\Delta := \{\mathbf{x} \in \mathbb{R}_+^d \mid \mathbf{1}^T\mathbf{x} = 1\}$. (10 points)

Solution: (1) We can rewrite the original problem as follows

$$\begin{aligned} \underset{x}{\text{minimize}} & \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \\ \text{subject to} & \sum_{i=1}^{d} x_i = 1 \\ & x_i \geq 0. \end{aligned}$$

The Lagrangian of the above problem is

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\alpha};\theta) = \|\boldsymbol{x}-\boldsymbol{y}\|_2^2 + \theta\left(\sum_{i=1}^{d}\boldsymbol{x}_i - 1\right) - \langle\boldsymbol{\alpha},\boldsymbol{x}\rangle$$

where $\theta \in \mathbb{R}$ and $\boldsymbol{\alpha} \in \mathbb{R}_+^d$ are both Lagrange multipliers. Differentiating with respect to x_i and comparing to zero gives the optimality condition

$$\frac{d\mathcal{L}}{dx_i} = 2\left(\boldsymbol{x}_i - \boldsymbol{y}_i\right) + \theta - \boldsymbol{\alpha}_i = 0$$

The complementary slackness KKT condition implies that whenever $x_i > 0$ we must have that $\boldsymbol{\alpha}_i = 0$. Thus, if $\boldsymbol{x}_i > 0$, we obtain

$$\boldsymbol{x}_i = \boldsymbol{y}_i - \frac{1}{2}\theta.$$

By the above equation, we have

$$\sum_{i=1}^{d}\boldsymbol{x}_i = \sum_{i=1}^{m}\boldsymbol{x}_i = \sum_{i=1}^{m}\boldsymbol{y}_i - \frac{1}{2}\theta = 1$$

where m is the cardinality of $\mathcal{I} = \{i \mid \boldsymbol{x}_i > 0\}$. Therefore, $\theta = \frac{2}{m}\left(\sum_{i=1}^{m}\boldsymbol{x}_i - 1\right)$. Given θ, we can characterize the optimal solution for $\boldsymbol{x}$ as

$$\boldsymbol{x}_i = \max\left\{\boldsymbol{y}_i - \frac{1}{2}\theta, 0\right\}.$$

(2) Solve the following projection onto the l_1 ball problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & \|\mathbf{x}-\mathbf{y}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_1 \leq \gamma. \end{aligned} \tag{6}$$

(10 points)

Solution: We do so by presenting a reduction to the problem of projecting onto the simplex given in (1).
- Note that if $\|\boldsymbol{x}\|_1 \leq \gamma$ then the solution to original problem is $\boldsymbol{x} = \boldsymbol{y}$.
- Assume that $\|\boldsymbol{x}\|_1 > \gamma$. Thus, the optimal solution must be on the boundary of the constraint set and we can replace the inequality constraint $\|\boldsymbol{x}\|_1 \leq \gamma$ with an equality constraint $\|\boldsymbol{x}\|_1 = \gamma$.
- Notice that each non-zero component of the optimal solution $\boldsymbol{x}^*$ shares the sign of its counterpart in $\boldsymbol{y}$, i.e., $x_i^* y_i \geq 0$ for all i (This is easily to be verified).

Based on the above observations and the symmetry of the objective. Let $\boldsymbol{u}$ be a vector obtained by taking the absolute value of each component of $\boldsymbol{y}$, $u_i = |y_i|$. We replace original problem with

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^d}{\text{minimize}} \quad & \|\boldsymbol{\beta}-\boldsymbol{u}\|_2^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_1 \leq \gamma, \\ & \beta \geq 0. \end{aligned}$$

Once we obtain the solution for the above problem we construct the optimal of original problem by setting $x_i = \text{sign}\,(y_i)\,\beta_i$.

III. Subgradient Methods

1. On the slide of Subgradient Methods, we have the Lemma 4.1.

Lemma (4.1). Projected subgradient update rule $\mathbf{x}^{t+1} = \mathcal{P}c(\mathbf{x}^t - \eta_t\mathbf{g}^t)$ obeys

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - 2\eta_t(f(x^t) - f^{opt}) + \eta_t^2\|\mathbf{g}^t\|_2^2, \tag{7}$$

where $\mathbf{g}^t$ is any subgradient of f at $\mathbf{x}^t$.
When f is μ-strongly convex, show that Lemma 4.1 can be improved to

$$\|\mathbf{x}^{t+1} - \mathbf{x}^*\|_2^2 \leq (1 - \mu\eta_t)\|\mathbf{x}^t - \mathbf{x}^*\|_2^2 - 2\eta_t(f(x^t) - f^{opt}) + \eta_t^2\|\mathbf{g}^t\|_2^2. \tag{8}$$

(15 points) Solution: Follow the proof of Lemma 4.1 on the slide,

$$\left\|x^{t+1} - x^*\right\|_2^2 \leq \left\|x^t - x^*\right\|_2^2 - 2\eta_t \left\langle x^t - x^*, g^t \right\rangle + \eta_t^2 \left\|g^t\right\|_2^2.$$

Since f is μ-strongly convex, then we have

$$f\left(x^*\right) - f\left(x^t\right) \geq \left\langle x^* - x^t, g^t \right\rangle + \frac{\mu}{2}\left\|x^t - x^*\right\|_2^2.$$

Combine these two inequalities, we complete our proof.

2. For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, a subgradient at $\mathbf{x} \in \mathbb{R}^d$ is a vector $\mathbf{g} \in \mathbb{R}^d$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^T(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{y} \in \mathbb{R}^d. \tag{9}$$

In this case, we write $\mathbf{g} \in \partial f(\mathbf{x})$. We say that f is sub-differentiable when $\partial f(\mathbf{x})$ is non-empty for each $\mathbf{x} \in \mathbb{R}^d$.

(1) Show that $\mathbf{x}^*$ is a minimizer of f if $\mathbf{0} \in \partial f(\mathbf{x})$. (10 points)

(2) Show that f is Lipschitz with parameter L (i.e., $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}$,$\mathbf{y} \in \mathbb{R}^d$) if and only if $\|\mathbf{g}\|_2 \leq L$ for all subgradients vectors $\mathbf{g}$.(10 points)

(1) Solution: Suppose $0 \in \partial f\left(x^*\right)$. Then for all $y \in \mathbb{R}^d$,

$$f(y) \geq f\left(x^*\right) + \langle 0, y - x^* \rangle = f\left(x^*\right)$$

so $f\left(x^*\right)$ is a minimizer.
(2) Solution: ($\Rightarrow$). Suppose f is L-Lipschitz; choose $x, y \in \mathbb{R}^d$, and let g be a subgradient vector at x. Then,

$$\langle g, y - x \rangle \leq f(y) - f(x) \leq L\|y - x\|$$

Rearranging,

$$\frac{\langle g, y - x \rangle}{\|y - x\|} \leq L$$

Since inequality in fact holds for all $y \in \mathbb{R}^d$, we choose $y = g + x$ to obtain the desired result.
($\Leftarrow$). Suppose $\|g\|_2 \leq L$ for all subgradient vectors g. Choose $x, y \in \mathbb{R}^d$. The subgradient condition tells us

$$f(y) - f(x) \geq \langle g, y - x \rangle$$

where g is a subgradient at x. Multiplying both sides by (-1),

$$\begin{aligned} f(x) - f(y) &\leq \langle g, x - y \rangle \\ &\leq \|g\|\|x - y\| \quad \text{(Cauchy-Schwarz)} \\ &\leq L\|x - y\| \end{aligned}$$

In a similar fashion, by switching the roles of x and y, and letting h be a subgradient vector at y, we have

$$f(y) - f(x) \leq \|h\|\|x - y\| \leq L\|x - y\|$$

Hence, we conclude that $|f(y) - f(x)| \leq L\|x - y\|$ and f is L-Lipschitz.