COMPSCI 514: Problem Set 5

Due: 12/10 by 11:59pm in Gradescope.

Instructions:

- You are allowed to work on this problem set in a group of up to three members.
- You should choose your group from within your own class (either online or in-person).
- You may talk to members of other groups at a high level about the problems but **not work through the solutions in detail together**.
- Each group should **submit a single solution set**: one member should upload a pdf to Gradescope, marking the other members as part of their group in Gradescope.
- You must show your work/derive any answers as part of the solutions to receive full credit.

Core Competency Problems

1. Convex Functions and Sets (12 points)

1. For each of the functions below, either prove that it is convex, or give a counter example showing that it is not.
   (a) (1 point) $f : \mathbb{R}^d \to \mathbb{R}$ with $f(\vec{x}) = \|\vec{x} - \vec{c}\|_2$, where $\vec{c} \in \mathbb{R}^d$ is some fixed vector.
   (b) (1 point) $f : \mathbb{R}^{n \times d} \to \mathbb{R}$ with $f(A) = \text{rank}(A)$.
   (c) (1 point) $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ with $f(A) = \text{tr}(A)$.
2. A set $\mathcal{S}$ is convex if $\vec{a}, \vec{b} \in \mathcal{S}$ and $\lambda \in [0, 1]$ implies $\lambda\vec{a} + (1 - \lambda)\vec{b} \in \mathcal{S}$. For each of the sets below, either prove that it is convex, or give a counter example showing that it is not.
   (a) (1 point) $\{\vec{x} : f(\vec{x}) \leq c\}$ where c is any scalar constant and f is a convex function.
   (b) (1 point) $\{\vec{y} \in \mathbb{R}^n : \exists \vec{x} \in \mathbb{R}^d \text{ with } \vec{y} = A\vec{x}\}$, where $A \in \mathbb{R}^{n \times d}$ is any fixed matrix.
   (c) (1 point) $\{A \in \mathbb{R}^{n \times d} : \text{rank}(A) \leq k\}$ where k is some fixed integer.
3. (4 points) For a convex set $\mathcal{S}$, the projection function $P_{\mathcal{S}}(\vec{z})$ returns $\vec{y} \in \arg\min_{\vec{y} \in \mathcal{S}} \|\vec{z} - \vec{y}\|_2$. Show that for any $\vec{b} \in \mathbb{R}^d$ with $\vec{b} \neq 0$ and $W \in \mathbb{R}$, the set $\mathcal{S}_{\vec{b},W} = \{\vec{v} \in \mathbb{R}^d : \langle \vec{b}, \vec{v} \rangle \geq W\}$ is convex (2 points). What is the projection function $P_{\mathcal{S}_{\vec{b},W}}(\vec{z})$? Prove that this is the projection function (2 points).
4. (2 points) Show that for any convex set $\mathcal{S}$, the projection function is unique. I.e., for any $\vec{z}$, there is a unique $\vec{y} \in \mathcal{S}$ with $\vec{y} = \arg\min_{\vec{y} \in \mathcal{S}} \|\vec{z} - \vec{y}\|_2$. **Hint:** Suppose there are two distinct $\vec{y}_1, \vec{y}_2$ such that $\|\vec{z} - \vec{y}_1\|_2^2 = \|\vec{z} - \vec{y}_2\|_2^2 = \min_{\vec{y} \in \mathcal{S}} \|\vec{z} - \vec{y}\|_2^2$. Try to show that this implies there exists a vector $\vec{y} \in \mathcal{S}$ that is even closer to $\vec{z}$ than $\vec{y}_1$ and $\vec{y}_2$.

2. Gradient Descent (12 points)

1. (2 points) Let $f : \mathbb{R} \to \mathbb{R}$ be a convex and $f(\theta^*) = \min_\theta f(\theta)$. Prove that $f'(\theta) \leq 0$ if $\theta < \theta^*$ and $f'(\theta) \geq 0$ if $\theta > \theta^*$. **Hint:** You may want to use the alternative definition of convexity that $f(y) \geq f(x) + f'(x)(y - x)$.

2. (2 points) We say a $\vec{\theta}$ is a *local minimum* for a function $f : \mathbb{R}^d \to \mathbb{R}$ if there exists $\epsilon > 0$ such that $f(\vec{\theta}) < f(\vec{\theta'})$ for all θ' such that $\|\vec{\theta} - \vec{\theta'}\|_2 < \epsilon$. Show that if f is convex then there can at most one local minimum $\vec{\theta}^*$.

3. (2 points) In our gradient descent analysis we used the fact $\frac{1}{t}\sum_{i=1}^{t} f(\vec{\theta}^{(i)}) \leq f(\vec{\theta}^*) + \epsilon$ implies that $f(\hat{\theta}) \leq f(\vec{\theta}^*) + \epsilon$ for the best iterate $\hat{\theta} = \arg\min_{\vec{\theta}^{(1)},\ldots,\vec{\theta}^{(t)}} f(\vec{\theta}^{(i)})$. Prove that if we instead set $\bar{\theta} = \frac{1}{t}\sum_{i=1}^{t} \vec{\theta}^{(i)}$ (i.e., $\bar{\theta}$ is the average iterate) then we also have $f(\bar{\theta}) \leq f(\vec{\theta}^*) + \epsilon$. **Hint:** Use that f is convex.

4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz function, i.e., $\|\nabla f(\theta)\|_2 \leq G$ for all θ.

   (a) (2 points) If $\theta^{(i+1)} = \theta^{(i)} - \eta \nabla f(\theta^{(i)})$, give an upper bound on $\|\theta^{(i+1)} - \theta^{(i)}\|_2$ in terms of η and G.

   (b) (2 points) In our fixed step size gradient algorithm we set $t = \frac{R^2G^2}{\epsilon^2}$ and $\eta = \frac{R}{G\sqrt{t}}$. Under these settings, what is the worst case increase in function value from step i to step $i+1$? I.e., give an upper bound on $f(\theta^{(i+1)}) - f(\theta^{(i)})$. Does this make intuitive sense? **Hint:** Use part (a).

   (c) (2 points) Consider the case of projected gradient descent over a convex set $\mathcal{S}$. So $\theta^{(i+1)} = P_{\mathcal{S}}(\theta^{out})$ for $\theta^{out} = \theta^{(i)} - \eta \nabla f(\theta^{(i)})$. Show that the bounds of (a), (b) still hold.

3. Gradient Descent with a Decaying Step Size (6 points)

We showed that gradient descent with step size $\eta = \frac{R}{G\sqrt{t}}$ converges to an ϵ approximate minimizer in $t = \frac{R^2G^2}{\epsilon^2}$ steps, for a convex G-Lipschitz function starting from an initial point $\vec{\theta}_1$ within a radius R of the optimum. This fixed step size analysis requires that we pick ϵ ahead of time and set η based on ϵ. However, in many applications we don't want to fix ϵ, but want to attain higher and higher accuracy as we run for longer. Here, we will analyze a variant of gradient descent with a gradually decreasing step size that allows us to do this.

Consider gradient descent with the update $\vec{\theta}_{i+1} = \vec{\theta}_i - \eta_i \vec{\nabla} f(\vec{\theta}_i)$, where the step size is set as

$$\eta_i = \frac{f(\vec{\theta}_i) - f(\vec{\theta}_*)}{\|\vec{\nabla} f(\vec{\theta}_i)\|_2^2}.$$

Note that using this step size requires knowledge of $f(\vec{\theta}_*)$, but not of $\vec{\theta}_*$, which may be reasonable in some settings. More complex approaches can remove the need to know this value.

1. (2 points) Let $d_i = f(\vec{\theta}_i) - f(\vec{\theta}_*)$ be our error at step i. Prove that with the above step size:

$$d_i^2 \leq G^2 \cdot \left(\|\vec{\theta}_i - \vec{\theta}_*\|_2^2 - \|\vec{\theta}_{i+1} - \vec{\theta}_*\|_2^2 \right).$$

   Hint: Start with the single step analysis shown in class, applied with step size η_i.

2. (1 point) Argue via the Cauchy-Schwarz Inequality that $\frac{1}{t}\sum_{i=1}^{t} d_i \leq \frac{1}{\sqrt{t}}\sqrt{\sum_{i=1}^{t} d_i^2}$. **Hint:** Cauchy-Schwarz Inequality states that $(\vec{x} \cdot \vec{y})^2 \leq \|\vec{x}\|_2^2 \cdot \|\vec{y}\|_2^2$ for all vectors $\vec{x}, \vec{y}$.

3. (2 points) Use parts (1) and (2) to show that after t steps:

$$\frac{1}{t}\sum_{i=1}^{t}\left[f(\vec{\theta}_i) - f(\vec{\theta}_*)\right] \leq \frac{GR}{\sqrt{t}}.$$

4. (1 point) Conclude that for any $\epsilon > 0$, after $t = \frac{G^2R^2}{\epsilon^2}$ steps, letting $\hat{\theta} = \arg\min_{\vec{\theta}_1,\ldots,\vec{\theta}_t} f(\vec{\theta}_i)$,

$$f(\hat{\theta}) - f(\vec{\theta}_*) \leq \epsilon.$$

Extra Credit Challenge Problem

C1. The Power of Message Passing (10 points)

Let G be an undirected, unweighted, d-regular graph on n nodes. I.e., a graph where every node has degree d. Assume that G has no self-loops. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be its adjacency matrix. Let $\mathbf{B} = \mathbf{A} + \mathbf{I}$. Think of each node in the graph as a user, and the edges as representing communication links between the users.

1. (2 points) Prove that $d+1$ is one of the eigenvalues of $\mathbf{B}$.

2. (2 points) Prove that if G is connected then the absolute value of each eigenvalues of $\mathbf{B}$ is at most $d+1$ and that there is only one eigenvector with an eigenvalue that has absolute value $d+1$. **Hint:** First prove that if $\vec{v}$ is an eigenvector with eigenvalue λ then for all i,

$$|\lambda| \cdot |\vec{v}(i)| \leq \sum_{j\in\Gamma(i)\cup\{i\}} |\vec{v}(j)|$$

and use this is argue that $|\lambda| \leq d+1$ and that $\lambda = d+1$ implies all entries of $\vec{v}$ are equal. Here $\Gamma(i)$ denotes the set of neighbors of the i^{th} node.

Consider now that setting where the i user has some initial value z_i and wants to estimate the average value $\mu = \frac{1}{n}\sum_{i=1}^{n} z_i$. Consider the following simple distributed averaging process: each user sets their initial estimate of the average to $\mu_i^{(0)} = z_i$. Then, at each step, each user sends its current estimate of the average $\mu_i^{(t)}$ to all of its neighbors in the network. Each user then updates their estimate to be the average of their neighbors' estimates and their own. I.e., they set $\mu_i^{(t+1)} = \frac{1}{d+1}\sum_{j\in\Gamma(i)\cup\{i\}} \mu_j^{(t)}$. Here $\Gamma(i)$ denotes the set of neighbors of the i^{th} node.

3. (2 points) Write this averaging process as a linear algebraic equation involving $\mathbf{B}$, the vector of estimates at time t, $\vec{\mu}^{(t)} \in \mathbb{R}^n$, and the vector of estimates at time $t+1$, $\vec{\mu}^{(t+1)} \in \mathbb{R}^n$.

4. (2 points) Show that we can write $\vec{\mu}^{(0)} = \mu \cdot \vec{1} + c_2\vec{v}_2 + \ldots + c_n\vec{v}_n$, where $\vec{1} \in \mathbb{R}^n$ is the all ones vector, $\vec{v}_2, \ldots, \vec{v}_n \in \mathbb{R}^n$ are orthonormal eigenvectors of $\mathbf{A}$, which are all orthonormal to $\vec{1}$, and $c_2, \ldots, c_n$ are some coefficients.

5. (2 points) Show that similarly, we can write $\vec{\mu}^{(t)} = \mu \cdot \vec{1} + \left(\frac{\lambda_2}{d+1}\right)^t c_2\vec{v}_2 + \ldots + \left(\frac{\lambda_n}{d+1}\right)^t c_n\vec{v}_n$, where $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$ are the eigenvalues of $\mathbf{B}$. Argue that $\lim_{t\to\infty} \mu_i^{(t)} = \mu$.