COMPSCI 514: Optional Problem Set 5

Due: December 12th by 11:59pm in Gradescope.

This problem set is optional. If you complete it, the points will count towards extra credit on top of your prior problem sets.

Instructions:

- Each group should work together to produce a single solution set. One member should submit a solution pdf to Gradescope, marking the other members as part of their group.
- You may talk to members of other groups at a high level about the problems but not work through the solutions in detail together.
- You must show your work/derive any answers as part of the solutions to receive full credit.

1. Convex Functions and Sets (15 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 $f(\vec{x}) = \|\vec{x}\|_2$.
   (b) (1 point) $f : \mathbb{R}^d \to \mathbb{R}$ with $f(\vec{x}) = \|\vec{x} - \vec{c}\|_2$, where $\vec{c}$ is some fixed vector.
   (c) (1 point) $f : \mathbb{R}^{n \times d} \to \mathbb{R}$ with $f(A) = \text{rank}(A)$.
   (d) (1 point) $f : \mathbb{R}^{n \times n} \to \mathbb{R}$ with $f(A) = \text{tr}(A)$.
2. 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) $\{\vec{x} : f(\vec{x}) \le c\}$ 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) $\{A \in \mathbb{R}^{n \times d} : \text{rank}(A) \le k\}$ where k is some fixed integer.
   (d) (1 point) $\{\vec{x} \in \mathbb{R}^n : \vec{x}(i) \in [0, 1] \text{ for all } i \text{ and } \sum_{i=1}^d \vec{x}(i) = 1\}$.
3. Consider the following optimization problem involving the Laplacian $\mathbf{L} \in \mathbb{R}^{n \times n}$ of a graph.
$$\min_{\vec{x} \in \mathbb{R}^n : \|\vec{x}\|_2 = 1 \text{ and } \vec{x}^T \vec{1} = 0} \vec{x}^T \mathbf{L} \vec{x}.$$
   (a) (1 point) Where have we seen this optimization problem before? What is the solution?
   (b) (2 points) Prove that a sum of two convex functions is always convex.
   (c) (2 points) Prove that the objective function $f(\vec{x}) = \vec{x}^T \mathbf{L} \vec{x}$ is convex. **Hint:** It may be helpful to use part (2) here along with the formula for $\vec{x}^T \mathbf{L} \vec{x}$ in terms of 'smoothness' of $\vec{x}$ over the graph.
   (d) (2 points) Is the above a convex optimization problem over a convex constraint set?

2. Gradient Descent with a Decaying Step Size (8 points)

In class 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 Cauchy-Schwarz that $\frac{1}{t}\sum_{i=1}^t d_i \leq \frac{1}{\sqrt{t}}\sqrt{\sum_{i=1}^t d_i^2}$.

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 $\hat{\theta} = \arg\min_{\vec{\theta}_1,\ldots,\vec{\theta}_t} f(\vec{\theta}_i)$,

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

5. (2 points) In our analysis in class and above, we show that $f(\hat{\theta}) - f(\vec{\theta}_*) \leq \epsilon$ 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 $f(\bar{\theta}) - f(\vec{\theta}_*) \leq \epsilon$. This strategy is often used, e.g., when using stochastic gradient descent for large datasets, since determining the best iterate can be much more expensive than just storing a running average. **Hint:** Use the bound in part (3) along with the assumption that f is convex.