COMPSCI 514: Problem Set 4

Due: 5/8 by 11:59pm in Gradescope.

Instructions:

- You are allowed to work on this problem set in a group of up to three members.
- 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.
- The problem set is meant for your own practice. We strongly discourage the use of LLMs to directly solve problems.
- Each problem will be graded on the following scale:
  - ✓+: (2 points) Submitted work demonstrates a full understanding of the problem. There may be some errors, omissions, or unclear steps, but overall, a reader would be able to understand how to solve the problem by looking at the submitted work.
  - ✓-: (1 point) Submitted work demonstrates partial understanding of the concepts, but contains significant omissions or errors.
  - X: (0 points) Not completed or submitted work doesn't not provide enough information to determine whether there is understanding of the problem.

1. Convex Functions and Sets

1. Let $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be functions such that f_i is G_i-Lipschitz. Prove that $F(\vec{x}) = \sum_{i=1}^n f_i(\vec{x})$ is G-Lipchitz for $G = \sum_{i=1}^n G_i$.
2. Let $f_1, \ldots, f_n : \mathbb{R}^d \to \mathbb{R}$ be convex functions. Prove that $F(\vec{x}) = \sum_{i=1}^n f_i(\vec{x})$ is convex.
3. Let f and g be convex functions. Is their product $f \cdot g$ a convex function? Either prove that it is or give a counterexample.
4. Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. Is its derivative f' convex? Either prove that it is or give a counterexample.
5. Consider the minimum cut problem on a graph G with n nodes and Laplacian $\mathbf{L}$. We have seen that this problem is equivalent to solving:

$$\min_{\substack{\mathbf{x}\in\{-1,1\}^n \\ \mathbf{x}\neq[1,1,\ldots,1],\ \mathbf{x}\neq[-1,-1,\ldots,-1]}} c(\mathbf{x}) \text{ where } c(\mathbf{x}) = \mathbf{x}^T\mathbf{L}\mathbf{x}.$$

   (a) Prove that the objective function $c(\mathbf{x}) = \mathbf{x}^T\mathbf{L}\mathbf{x}$ is convex. **Hint:** It may be helpful to write $\mathbf{x}^T\mathbf{L}\mathbf{x}$ as a sum of terms and use part (2) here.

   (b) Is min-cut a convex optimization problem over a convex constraint set?

6. Consider the low-rank approximation problem for a given matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$:

$$\min_{\mathbf{B} \in \mathbb{R}^{n \times d}, \text{rank}(\mathbf{B}) \leq k} \|\mathbf{X} - \mathbf{B}\|_F^2.$$

Is the objective function convex? Is the constraint set convex?

2. Gradient Descent with an Adaptive Step Size

In class we prove that gradient descent with step size $\eta = \frac{R}{G\sqrt{t}}$ converges to an ϵ approximate minimizer in $t = \frac{R^2 G^2}{\epsilon^2}$ 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. 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. Prove that $\frac{1}{t} \sum_{i=1}^t d_i \leq \frac{1}{\sqrt{t}} \sqrt{\sum_{i=1}^t d_i^2}$. **Hint:** Apply the Cauchy-Schwarz inequality.

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

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

3. Gradient Descent for Finding Stationary Points

In many applications in machine learning, gradient descent is applied to non-convex functions (e.g., neural network loss functions). Here, we cannot guarantee that the method will converge to an approximate global minimizer. However, we can show that it finds to an approximate *stationary point*, where the gradient is near zero. Such a stationary point is either a near local minima, or a saddle point.

Formally, for a function $f : \mathbb{R}^d \to \mathbb{R}$, we would like to show that the method finds θ such that $\|\nabla f(\theta)\|_2 \leq \epsilon$. To prove this, we must make an additional assumption about our function – that the gradient itself is Lipschitz. In particular, we will assume that for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla f(\theta_1) - \nabla f(\theta_2)\|_2 \leq L \cdot \|\theta_1 - \theta_2\|_2$. A function f satisfying this condition is said to be *L-smooth*.

1. A corollary of L-smoothness is that for nearby points, a linear approximation to the function is fairly accurate. Formally, we will need that, for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$f(\theta_2) - f(\theta_1) \leq \nabla f(\theta_1)^T(\theta_2 - \theta_1) + \frac{L}{2}\|\theta_1 - \theta_2\|_2^2.$$

   Consider gradient descent with the update $\theta_{i+1} = \vec{\theta_i} - \eta \nabla f(\theta_i)$, where f is L-smooth and $\eta = 1/L$. Use the above corollary to prove that $f(\theta_{i+1}) - f(\theta_i) \leq -\frac{1}{2L}\|\nabla f(\theta_i)\|_2^2$.

2. Prove that $\frac{1}{2L}\sum_{i=0}^{t-1}\|\nabla f(\theta_i)\|_2^2 \leq f(\theta_0) - f(\theta_t)$. **Hint:** Use a telescoping sum.

3. Prove that if our starting point satisfies $f(\theta_0) - \min_\theta f(\theta) \leq R$ then after $t = \frac{2RL}{\epsilon^2}$ steps, at least some θ_i for $i \in 0, \ldots, t-1$ must satisfy $\|\nabla f(\theta_i)\|_2 \leq \epsilon$.

4. Assume further that f is convex. Does the upper bound on $\|\nabla f(\theta_i)\|_2$ from part (3) itself imply that $f(\theta_i) - \min_\theta f(\theta)$ is small? I.e., that θ_i is a near global minimizer? Why or why not?