Roll. No:	Name:

Total Marks: 15, Total Time: 35mins

Instructions

1. The quiz is divided into two sections: short answer questions, and problems that require a detailed solution. For the first section, provide the final answer for the first three questions and a short justification for the last question. For the second section, provide detailed answers showing all the necessary steps.
2. Use rough sheets for any calculations *if necessary,* and do not submit the rough sheets. Do not use a pencil for writing the answers.
3. Assume standard data whenever you feel that the given data is insufficient. However, do quote your assumptions explicitly.

I Short answer questions

2 1. Consider a m-strongly convex and L-smooth function f_1 with $m = 10^{-3}$ and $L = 1$. Let f_2 be a convex and L-smooth function with $L = 1$. Suppose we run GD with stepsize 1 for both functions. Let $f(x_n) - f(x^*)$ (with obvious notation) denote the optimization error after n iterations of GD. Which of the following statements is true regarding the number of iterations required to ensure the optimization error is less than 10^{-3}?

I: The number of GD iterations for f_1 is lower than that for f_2.
II: The number of GD iterations for f_2 is lower than that for f_1.
III: The number of GD iterations for f_1 and f_2 are comparable.

2 2. Provide a bound using Oh-notation for the number of iterations n of GD so that $f(x_n) - f(x^*) \leq \epsilon$ (with obvious notation) for an objective function that is (i) convex and smooth; and (ii) strongly-convex and smooth. Ignore the constants and specify the dependence on ϵ alone.

Answer:

Function type	Iteration complexity
convex and smooth	
strongly-convex and smooth	

2 3. Let f be a smooth function that is not necessarily convex. For a given $\epsilon > 0$, a point $\bar{\theta}$ is an ϵ stationary point if it satisfies $\left\|\nabla f(\bar{\theta})\right\|^2 \leq \epsilon$. Specify the number of iterations required to find an ϵ stationary point for a stochastic gradient algorithm in the following two settings:

   1. unbiased gradient information;
   2. biased gradient information with a bias $c_1\delta^2$ and variance c_2/δ^2, where δ is a sensitivity parameterthat for bias-variance tradeoff.

   Use big-oh notation to specify the complexity in terms of ϵ.

4. Consider a L-smooth function f that satisfies the following property: Every stationary point $\bar{x}$ (i.e., $\nabla f(\bar{x}) = 0$) is a global minimum. Suppose we perform a gradient descent (GD) update as follows:

$$x_{k+1} = x_k - \alpha \nabla f(x_k),$$

where $\alpha < 1/L$.

[.5] (a) Does the GD algorithm converge to a global minimum of f?

[.5] (b) What is number of iterations to find an ϵ stationary point (as defined in the previous question)?

[1] (c) Suppose the GD algorithm find an ϵ stationary point, say $\bar{x}$, in N iterations. Does this imply a bound on $f(\bar{x}) - f(x^*)$? Why or why not?

II. Problems that require a detailed solution (Answer any one)

Note: If more than one question is answered, then the first answer will be considered for evaluation.

1. Let $f(x) = x^2 + 3\sin^2(x)$.

[3] (a) Show that f satisfies the PL-condition.

[2] (b) Suppose we form gradient descent for finding the minima of this function, i.e., an update of the following form:

$$x_{k+1} = x_k - \alpha f'(x_k).$$

Provide a bound on $f(x_n) - f(x^*)$ with an appropriate choice of α.

[2] (c) Consider the case where we have noisy function observations, i.e., for any x, the optimization algorithm is given an observation of the form $\hat{f}(x) = f(x) + \epsilon$, where ϵ is a random variable with standard normal distribution. How would you find x^* for the function defined above in this setting? Specify the algorithm precisely.

[1] (d) *BONUS*: Specify the convergence rate in big-Oh notation for the previous part (proof not needed).

2. Let $f(x) = \sum_{i=1}^{m} f_i(x)$, where f is a L-smooth function, and $\|\nabla f_i(x)\|^2 \leq \sigma^2$, for $i = 1, \ldots, m$. Do note that f is *not* necesarily convex.

Answer the following:

[1] (a) For minimizing f, write the update iteration of the SGD algorithm with stepsize denoted by α_k and iterate by x_k.

[3] (b) Show the following bound holds for SGD algorithm from the part above:

$$\mathbb{E}\left[f(x_{k+1}) - f(x_k)\right] \leq -\alpha_k \|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L\sigma^2. \tag{1}$$

[3] (c) Fix n, set $\alpha = \frac{c}{\sqrt{n}}$ for some constant c. Is there a choice for the constant c such that the following bound holds for the SGD algorithm with stepsize α:

$$\min_{0 \leq k \leq n-1} \mathbb{E}\left[\|\nabla f(x_k)\|_2^2\right] \leq \sqrt{\frac{2(f(x_0) - f(x^*))L\sigma^2}{n}},$$

where x^* is a global minimum of f. Show your work in arriving at the bound above for suitable α.

1 (d) *BONUS*: Is the bound in the part above the best achievable using a stochastic gradient algorithm? Or can it be improved?