Optimization for Machine Learning Final Exam

name

Due: 12/18/2023

You MAY NOT collaborate with other students on this final.

When solving one question/part, you may assume the results of all previous questions/parts. You may also assume all previous homework results and results from class or lecture notes, but please explain which result you are using when you use it.

Please justify all answers unless explicitly instructed not to in the question statement.

1. (5pts) Suppose that $\mathcal{L}(\mathbf{w})$ is a convex function, and $\mathbf{w}_1, \dots, \mathbf{w}_T$ are such that:

$$\sum_{t=1}^{T} t(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)) \le T^{3/2}$$

   You know the identities of the points $\mathbf{w}_1, \dots, \mathbf{w}_T$, but you *do not* have any other information about $\mathcal{L}$ (e.g. you cannot compute its values or its gradients). Provide a *deterministic* point $\hat{\mathbf{w}}$ as a function of $\mathbf{w}_1, \dots, \mathbf{w}_T$ such that

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \le O\left(\frac{1}{\sqrt{T}}\right)$$

   Solution:

2. (5pts) Specify a function $f : \mathbb{R} \to \mathbb{R}$ that is both 1-strongly convex and 10-Lipschitz, or prove that no such function exists.

   Solution:

3. (5pts) Specify a function $f : \mathbb{R} \to \mathbb{R}$ that is both 1-smooth and 2 strongly convex, or prove that no such function exists.

   Solution:

4. (a) (5pts) *Newton's method* employs the following update: $\mathbf{w}_{t+1} = \mathbf{w}_t - \nabla^2 \mathcal{L}(\mathbf{w}_t)^{-1} \nabla \mathcal{L}(\mathbf{w}_t)$. Suppose that $\mathcal{L}$ is a convex quadratic function (that is, $\mathcal{L}$ has the form $\mathcal{L}(\mathbf{w}) = \mathbf{w}^\top A \mathbf{w} + \langle \mathbf{w}, \mathbf{v} \rangle + c$ for some positive semi-definite matrix A, vector $\mathbf{w}$ and scalar c). Show that in this case, no matter what $\mathbf{w}_1$ is, $\mathbf{w}_2 = \operatorname{argmin} \mathcal{L}$ whenever A is strictly positive-definite.

   Solution:

(b) (5pts) A friend asks if the reason the result in part (a) is able to avoid the $\Omega(1/T^2)$ lower bound we learned in class is that the analysis was restricted to quadratic loss functions rather than general smooth convex loss functions. You tell them no: the lower bound applies even to algorithms that consider only quadratic losses. Why?

Solution:

(c) (5pts) What is the true reason that the result in part (a) can avoid the $\Omega(1/T^2)$ lower bound?

Solution:

5. (10pts) Let us call a differentiable function q-star-convex if for all $\mathbf{w}$, $\mathcal{L}(\mathbf{w}_\star) \ge \mathcal{L}(\mathbf{w}) + q\langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{w}_\star - \mathbf{w}\rangle$. Convex functions satisfy this condition with $q = 1$ (although $q = 1$ does not imply convexity). $q > 1$ indicates some non-convexity. Suppose that $\mathcal{L} = \mathbb{E}[\ell(\mathbf{w}, z)]$ is q-star-convex, and satisfies $\|\mathbf{w}_\star\| \le D$ for some known D. Further, suppose that $\ell(\mathbf{w}, z)$ is differentiable and G-Lipschitz in $\mathbf{w}$ for all z. Show that stochastic gradient descent with an appropriate learning rate guarantees:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \le O\left(\frac{qDG}{\sqrt{T}}\right)$$

Solution:

6. Consider the following "communication constrained" situation: A "server" holds a training dataset, and when provided with a point $\mathbf{w}$ it will be able to randomly sample some example z and compute a gradient of a loss $\ell(\mathbf{w}, z)$ such that $\mathbb{E}[\ell(\mathbf{w}, z)] = \mathcal{L}(\mathbf{w})$ for some H-smooth function $\mathcal{L}$. It is guaranteed that each coordinate of $\nabla \ell(\mathbf{w}, z) \in \mathbb{R}^d$ lies in $[-1, 1]$ for all $\mathbf{w}$ and z. Unfortunately, in order to preserve outgoing bandwidth, the server will not tell you the actual gradients $\nabla \ell(\mathbf{w}, z)$ it computes. Instead it will give you a d-bit string $C(\nabla \ell(\mathbf{w}, z)) \in \{\pm 1\}^d$ (the C stands for "compressed"). The ith coordinate of $C(\nabla \ell(\mathbf{w}, z))$ is set randomly by the formula:

$$C(\nabla \ell(\mathbf{w}, z))[i] = \begin{cases} 1 & \text{with probability } \frac{1+\nabla \ell(\mathbf{w}, z)[i]}{2} \\ -1 & \text{with probability } \frac{1-\nabla \ell(\mathbf{w}, z)[i]}{2} \end{cases}$$

(a) (5pts) Show that $\mathbb{E}[C(\nabla \ell(\mathbf{w}, z))] = \nabla \mathcal{L}(\mathbf{w})$, where the expectation is over both the randomness in the choice of z as well as the randomness in the function C.

Solution:

(b) (10 pts) Suppose you perform SGD with these compressed gradients:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta C(\nabla \ell(\mathbf{w}_t, z))$$

Show that after T iterations, with an appropriate learning rate,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le O\left(\frac{\sqrt{dH(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}}{\sqrt{T}}\right)$$

Solution: