Optimization for Machine Learning Review HW

SOLUTIONS

All parts of each question are equally weighted. 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.

1. Suppose that you have a dataset of $N = 10^8$ examples $z_1, \ldots, z_N$ and a loss function $\ell(w, z)$ relating model parameters w to examples z. Suppose that ℓ is G-Lipschitz and H-smooth and you are interested in minimizing the training error $\hat{L}(w) = \frac{1}{N}\sum_{i=1}^{N} \ell(w, z_i)$. Assume that it takes 1 millisecond to compute a gradient $\nabla\ell(w, z_i)$. For this problem we will ignore all other computing costs. Let $w_\star = \operatorname{argmin} \hat{\mathcal{L}}(w)$ and let w_1 be some given point. assume that $\|w_1 - w_\star\| \leq R$ and $\mathcal{L}(w_1) - \mathcal{L}(w_\star) \leq \Delta$ for some given R and Δ.

   (a) Suppose ℓ is also convex. If you run stochastic gradient descent (sampling a new z_t uniformly at random from $z_1, \ldots, z_N$ for each iteration), provide an upper bound on how long it will take (just counting gradient computation time) to find a point $\hat{w}$ satisfying $\mathbb{E}[\hat{L}(\hat{w}) - \hat{L}(w_\star)] \leq \epsilon$ as a function of ϵ, G, H, Δ, R and N (you may not need all variables in your expression). Describe how you will will find the point $\hat{w}$, and justify your expression for ϵ. You may provide your answer in any units you like.

   Solution:

   Recall that with a learning rate of $\eta = \frac{R}{G\sqrt{T}}$, stochastic gradient descent starting from initial point $\mathbf{w}_1$ will ensure after T iterations:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \frac{RG}{\sqrt{T}}$$

   Thus, if we set $\hat{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t$, by Jensen inequality we have

$$\mathbb{E}[\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star)] \leq \frac{RG}{\sqrt{T}}$$

   This algorithm takes T milliseconds to run. Now, if we set $T \geq \frac{R^2G^2}{\epsilon^2}$ and run for T milliseconds, we obtain $\mathbb{E}[\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star)] \leq \epsilon$.

   (b) Without supposing that ℓ is convex, if you still run stochastic gradient descent, provide an upper bound on how long it will take (again just counting gradient computation time) to find a point $\hat{w}$ satisfying $\mathbb{E}[\|\nabla\hat{\mathcal{L}}(\hat{w})\|] \leq \epsilon$ as a function of ϵ, G, H, Δ, R and N (you may not need all variables in your expression). Describe you will will find the point $\hat{w}$, and justify your expression for ϵ.

   Solution:

   We have shown in class that with $\eta = \frac{\sqrt{2\Delta}}{G\sqrt{HT}}$, SGD guarantees:

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \leq G\sqrt{2\Delta HT}$$

so that we can take $\hat{\mathbf{w}}$ uniformly at random from $\mathbf{w}_1, \ldots, \mathbf{w}_T$ to obtain:

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \leq \frac{\sqrt{G\sqrt{2\Delta H}}}{T^{1/4}}$$

This will again take T milliseconds. Therefore, with $T = \frac{2G^2\Delta H}{\epsilon^4}$ milliseconds, we obtain $\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|] \leq \epsilon$.

(c) Suppose that ℓ is non-convex and you have one day to train your model to obtain a point with the smallest possible value of $\mathbb{E}[\|\nabla \hat{\mathcal{L}}(\hat{w})\|] \leq \epsilon$. Is it reasonable to use variance reduction? Why or why not?

Solution:
Yes it is reasonable. Non-convex variance reduction is theoretically superior to SGD regardless of how long you have to run.

(d) Suppose that ℓ is convex and you have one day to train your model to obtain a point with the smallest possible value of $\mathbb{E}[\hat{L}(\hat{w}) - \hat{L}(w_\star)]$. Is it reasonable to use variance reduction? Why or why not? (hint: this question can be subtle. For the purposes of this question, you may interpret "use variance reduction" to mean "use the algorithm for convex variance reduction discussed in class exactly as discussed". However, you may also propose alternative ways to use variance reduction, in which case your answer may vary. Regardless, you must justify your answer).

Solution:
In the convex setting, in order to have gains from variance reduction in our theoretical analysis, we need to compute several full passes over the dataset. Since $N = 10^8$, this requires at least 10^8 milliseconds per pass. Unfortunately, there are about $8.7 \cdot 10^7$ milliseconds in a day, so we will not have time to complete even a single pass. As a result, variance reduction may not be so helpful.

2. A friend tells you they have developed a new first-order deterministic optimization algorithm. So long as the function is convex, Lipschitz, smooth and second-order smooth, their algorithm finds a point $\hat{w}$ such that $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \leq O(1/T^3)$ after T gradient evaluations. They tell you it avoids the lower-bound of $1/T^2$ because their algorithm requires $\mathcal{L}$ to be second-order smooth. Is their result plausible? Why or why not?

Solution:
Their result is not plausible. Even though the hypotheses in the standard lower bound may only specify a smooth function, the "hard" function used to establish the lower bound is a quadratic function, which is actually $J = 0$-second order smooth. Thus even assuming second-order smoothness is not enough if you are still using a first-order algorithm.

3. The *support vector machine* (SVM) is a classical model in machine learning. They can be used for binary classification: given a input x, wish to predict $y \in \{-1, 1\}$. The SVM solves this problem by choosing a fixed *feature map* $\phi(x)$ that produces outputs in some vector space, and then predicting with $\hat{y} = \text{sign}(\langle \phi(x), \mathbf{w} \rangle)$ for some parameter $\mathbf{w}$. Technically, the vector space in which $\phi(x)$ and $\mathbf{w}$ live may be infinite dimensional, but for the purposes of this problem, you may assume $\phi(x) \in \mathbb{R}^d$ for some finite d. $\mathbf{w}$ is trained by minimizing the loss:

$$\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, (x, y))] = \mathbb{E}[\max(0, 1 - y\langle \phi(x), \mathbf{w} \rangle)]$$

Suppose that $\|\mathbf{w}_\star\| \leq R$ and $\|\phi(x)\| \leq B$ for all x. Show that with appropriate learning rate, given an

i.i.d. dataset $(x_1, y_1), \ldots, (x_T, y_T)$, you can fine a point $\hat{\mathbf{w}}$ that ensures:

$$\mathbb{E}[\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq O\left(\frac{RB}{\sqrt{T}}\right)$$

Solution:

We will show that the loss ℓ is convex and B-Lipschitz. Then, using SGD with learning rate $\frac{R}{B\sqrt{T}}$ will yield the desired convergence rate.

First, we show that ℓ is B-Lipschitz. Let $\mathbf{w}$ and $\mathbf{w}'$ be arbitary points and let $z = (x, y)$ be an arbitrary example. Since $\|\phi(x)\| \leq B$, we have $|\langle \phi(x), \mathbf{w}\rangle - \langle \phi(x), \mathbf{w}'\rangle| \leq B\|\mathbf{w} - \mathbf{w}'\|$. Let $m = \langle \phi(x), \mathbf{w}\rangle$ and $m' = \langle \phi(x), \mathbf{w}'\rangle$. Then, it suffices to show:

$$|\max(0, 1 - ym) - \max(0, 1 - ym')| \leq |m - m'| \leq B\|\mathbf{w} - \mathbf{w}'\|$$

To see this, first suppose that both max operations are zero. Then it is clear. Next, if both max operations are non-zero, the statement is also immediate. The interesting case is when one max is zero and the other is not. By symmetry, it does not matter which one is non-zero so let us assume $\max(0, 1 - ym) = 1 - ym$ and $\max(0, 1 - ym') = 0$. Then we have $ym < 1$ and $ym' > 1$. Therefore $0 < 1 - ym < ym' - ym$ so that $|1 - ym| \leq |m - m'|$ since $|y| = 1$ and so we are done.

Now we need to show convexity. Let $t \in [0, 1]$. Then we have:

$$\max(0, 1 - y\langle \phi(x), t\mathbf{w} + (1-t)\mathbf{w}'\rangle) = \max(0, t - ty\langle \phi(x), \mathbf{w}\rangle + (1-t) - (1-t)y\langle \phi(x), \mathbf{w}'\rangle)$$

Now, observe that $a+b \leq \max(0, a)+\max(0, b)$ so that $\max(0, a+b) \leq \max(0, a)+\max(0, b)$. Therefore

$$\begin{aligned}\max(0, 1 - y\langle \phi(x), t\mathbf{w} + (1-t)\mathbf{w}'\rangle) &= \max(0, t - ty\langle \phi(x), \mathbf{w}\rangle) + \max(0, (1-t) - (1-t)y\langle \phi(x), \mathbf{w}'\rangle) \\ &= t\max(0, 1 - y\langle \phi(x), \mathbf{w}\rangle) + (1-t)\max(0, 1 - y\langle \phi(x), \mathbf{w}'\rangle)\end{aligned}$$

and so ℓ is convex.