Optimization for Machine Learning HW 3

Due: 9/27/2023

All parts of each question are equally weighted. When solving one question/part, you may assume the results of all previous questions/parts.

1. This question explores the use of *time-varying* learning rates. Suppose $\mathcal{L}(\mathbf{w}) = \mathbb{E}_z[\ell(\mathbf{w}, z)]$ is a convex function, and suppose $D \geq \|\mathbf{w}_1 - \mathbf{w}_\star\|$ for some $\mathbf{w}_1$ and $\mathbf{w}_\star = \operatorname{argmin} \mathcal{L}(\mathbf{w})$. In class, we showed that if $\|\nabla \ell(\mathbf{w}, z)\| \leq G$ for all z and $\mathbf{w}$, then stochastic gradient descent with learning rate $\eta = \frac{D}{G\sqrt{T}}$ satisfies

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

However, in order to set this learning rate, we needed to use knowledge of D, G and T. This question helps show a way to avoid needing to know T.

(a) To do this, we will consider *projected* stochastic gradient descent with *varying learning rate*. Suppose we start at $\mathbf{w}_1 = 0$. Then the update is:

$$\mathbf{w}_{t+1} = \Pi_{\|\mathbf{w}\| \leq D}\left[\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t)\right]$$

where $\Pi_{\|\mathbf{w}\| \leq D}[x] = \operatorname{argmin}_{\|\mathbf{w}\| \leq D} \|x - \mathbf{w}\|$. Notice that $\Pi_{\|\mathbf{w}\| \leq D}[\mathbf{w}_\star] = \mathbf{w}_\star$ by definition of D. Show that

$$\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star \rangle \leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}$$

And conclude:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

(You may use without proof the identity $\|\Pi_{\|\mathbf{w}\| \leq D}[x] - \mathbf{w}_\star\|^2 \leq \|x - \mathbf{w}_\star\|^2$ for all t and all vectors x. This follows because $\|\mathbf{w}_\star\| \leq D$.)

Solution:

(b) Next, show that so long as η_t satisfies $\eta_t \leq \eta_{t-1}$ for all t, we have:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T} \eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$

(hint: at some point you will probably need to show $\|\mathbf{w}_t - \mathbf{w}_\star\|^2 (\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}) \leq 2D^2 (\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}})$).

Solution:

(c) Next, consider the update

$$\mathbf{w}_{t+1} = \Pi_{\|\mathbf{w}\|\le D}\left[\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t)\right]$$

where we set $\eta_t = \frac{D}{G\sqrt{t}}$. Recalling our assumption that $\|\nabla \ell(\mathbf{w}_t, z_t)\| \le G$ with probability 1, Show that

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

This allows you to handle any T value without having the algorithm know T ahead of time. (Hint: you may want to show that $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \le 1 + \int_1^T \frac{dx}{\sqrt{x}}$).

Solution:

2. This question is an exercise in understanding the non-convex SGD analysis. In the notes, Theorem 5.3 discusses how to use varying learning rate η_t proportional to $\frac{1}{\sqrt{t}}$ to obtain a non-convex convergence rate of:

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|^2] \le O\left(\frac{\log(T)}{\sqrt{T}}\right)$$

In this question, we will remove the logarithmic factor by adding an extra assumption.

(a) Suppose that $\mathcal{L}$ is H-smooth, $\|\nabla \ell(\mathbf{w}, z)\| \le G$ for all $\mathbf{w}$ and z, and further that $\mathcal{L}(\mathbf{w}) \in [0, M]$ for all $\mathbf{w}$ (this last assumption is slightly stronger than we have assumed in class). Consider the SGD update:

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

Suppose η_t is an arbitrary deterministic learning rate schedule satisfying $\eta_{t+1} \le \eta_t$ for all t (i.e. the learning rate never increases). Show that for all $\tau < T$:

$$\frac{1}{T-\tau}\mathbb{E}\left[\sum_{t=\tau+1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] \le \frac{1}{\eta_T(T-\tau)}\left(M + \frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2\right)$$

Solution:

(b) Next, consider $\eta_t = \frac{1}{\sqrt{t}}$. In class, we considered choosing $\hat{\mathbf{w}}$ *uniformly* at random from $\mathbf{w}_1, \dots, \mathbf{w}_T$. Instead, produce a *non-uniform* distribution over $\mathbf{w}_1, \dots, \mathbf{w}_T$ such that choosing $\mathbf{w}_T$ from this distribution satisfies:

$$\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|^2] \le O\left(\frac{1}{\sqrt{T}}\right)$$

where the $O(\cdot)$ notation hides constants that do not depend on T. That is, you should find some $p_1, \dots, p_T$ such that you set $\hat{\mathbf{w}} = \mathbf{w}_t$ with probability p_t. The uniform case is $p_t = 1/T$ for all t. If it helps, you may assume that T is divisible by any natural number (e.g. you can assume T is even if you want). Note that such an assumption is not required.

Solution:

BONUS (c) Assume that $\mathcal{L}$ is H-smooth, $\|\nabla \ell(\mathbf{w}, z)\| \le G$ for all $\mathbf{w}$ and z, and $\mathbf{w}_1$ is such that $\mathcal{L}(\mathbf{w}_1) - \inf_{\mathbf{w}} \mathcal{L} \le \Delta$ (note that this is *the same* as our usual assumptions in class). Devise a sequence of learning rates such that:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] \le O\left(\frac{(HG^2 \log\log(T) + \Delta)\sqrt{\log(T)}}{\sqrt{T}}\right)$$

where the $O(\cdot)$ notation hides constants that may depend on G, Δ and H but *not* T.

Solution: