Optimization for Machine Learning HW 3

Shuyue Jia
BUID: U62343813

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:

Proof. From the hint, we know that $\|\Pi_{\|\mathbf{w}\| \leq D}[x] - \mathbf{w}_\star\|^2 \leq \|x - \mathbf{w}_\star\|^2$ for all t and all vectors x. Thus,

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \eta_t \nabla \ell(\mathbf{w}_t, z_t) - \mathbf{w}_\star\|^2 \\ &\leq \|(\mathbf{w}_t - \mathbf{w}_\star) - \eta_t \nabla \ell(\mathbf{w}_t, z_t)\|^2 \\ &\leq \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\eta_t \langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star \rangle + \eta_t^2 \nabla \|\ell(\mathbf{w}_t, z_t)\|^2. \end{aligned} \tag{1}$$

Thus, we obtain

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

From the convexity and subgradient, we know that $\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \le \langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star \rangle$. As a result, we have

$$\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \le \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}. \tag{3}$$

Thus, we can obtain,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \le \mathbb{E}\left[\sum_{t=1}^{T} \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]. \tag{4}$$

□

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

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

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

Solution:
Proof. Following the results of the past problem, we sum telescopes,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\le \mathbb{E}\left[\sum_{t=1}^{T} \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right] \\
&\le \mathbb{E}[\frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{2\eta_1} + \sum_{t=1}^{T-1} \left(\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 (\frac{1}{2\eta_{t+1}} - \frac{1}{2\eta_t}) \right) \\
&\quad + \sum_{t=1}^{T} \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}] \\
&\le \mathbb{E}\left[\frac{D^2}{2\eta_1} + 4D^2 \sum_{t=2}^{T-1} (\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}) + \sum_{t=1}^{T} \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right] \\
&= \mathbb{E}\left[\frac{D^2}{2\eta_1} + 4D^2 (\frac{1}{2\eta_T} - \frac{1}{2\eta_1}) + \sum_{t=1}^{T} \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right] \\
&\le \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T} \eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right].
\end{aligned} \tag{5}$$

□

(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 $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 1 + \int_1^T \frac{dx}{\sqrt{x}}$).

Solution:

Proof. From the above problem and $\|\nabla \ell(\mathbf{w}_t, z_t)\| \leq G$, we know that

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T} \eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right] \\
&\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T} \eta_t G^2}{2}\right] \\
&\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{\sum_{t=1}^{T} \frac{D}{G\sqrt{t}} G^2}{2}\right] \\
&\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{1}{2} DG \sum_{t=1}^{T} \frac{1}{\sqrt{t}}\right] \\
&\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{1}{2} DG \left(1 + \int_1^T \frac{dx}{\sqrt{x}}\right)\right].
\end{aligned} \tag{6}$$

We know that

$$\int_1^T \frac{dx}{\sqrt{x}} = 2\sqrt{x}\Big|_1^T = 2\sqrt{T} - 2\sqrt{1} = 2\sqrt{T} - 2. \tag{7}$$

Thus, we can get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{1}{2} DG \left(1 + \int_1^T \frac{dx}{\sqrt{x}}\right)\right] \\
&\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + \frac{1}{2} DG (1 + 2\sqrt{2} - 2)\right] \\
&\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + DG\sqrt{T} - \frac{1}{2} DG\right].
\end{aligned} \tag{8}$$

Finally, we can obtain

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\leq \mathbb{E}\left[\frac{2D^2}{\eta_T} + DG\sqrt{T} - \frac{1}{2} DG\right] \\
&\leq O(DG\sqrt{T}).
\end{aligned} \tag{9}$$

$\square$

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:

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

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:

Proof. We have

$$\begin{aligned} \mathcal{L}(\mathbf{w}_{t+1}) &\le \mathcal{L}(\mathbf{w}_t) + \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{H}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\| \\ &\le \mathcal{L}(\mathbf{w}_t) - \eta_t \langle \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t \rangle + \frac{H\eta_t^2}{2}\|\mathbf{g}_t\|^2. \end{aligned} \tag{1}$$

Thus, we have

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta_t \,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta_t^2 G^2}{2}. \tag{2}$$

Then, we will sum over $\tau + 1$ to T and divide $T - \tau$:

$$\frac{1}{T-\tau}\mathbb{E}[\sum_{t=\tau+1}^{T} \mathcal{L}(\mathbf{w}_{t+1})] \le \frac{1}{T-\tau}\mathbb{E}[\sum_{t=\tau+1}^{T} \mathcal{L}(\mathbf{w}_t)] - \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} \eta_t \,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] + \frac{1}{T-\tau}\frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2. \tag{3}$$

Thus, we can obtain

$$\begin{aligned} \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} \eta_t \,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] &\le \frac{1}{T-\tau}\mathbb{E}[\sum_{t=\tau+1}^{T} \mathcal{L}(\mathbf{w}_t)] - \frac{1}{T-\tau}\mathbb{E}[\sum_{t=\tau+1}^{T} \mathcal{L}(\mathbf{w}_{t+1})] + \frac{1}{T-\tau}\frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2 \\ &\le \frac{1}{T-\tau}\mathbb{E}[\sum_{t=\tau+1}^{T} \mathcal{L}(\mathbf{w}_t)] + \frac{1}{T-\tau}\frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2. \end{aligned} \tag{4}$$

Because $\mathcal{L}(\mathbf{w}) \in [0, M]$ for all $\mathbf{w}$, we will have

$$\begin{aligned} \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} \eta_t \,\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] &\le \frac{1}{T-\tau}\mathbb{E}[\sum_{t=\tau+1}^{T} \mathcal{L}(\mathbf{w}_t)] + \frac{1}{T-\tau}\frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2 \\ &\le \frac{1}{T-\tau}M + \frac{1}{T-\tau}\frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2. \end{aligned} \tag{5}$$

Use the fact that $\eta_T \le \eta_t$ for all t:

$$\frac{\eta_T}{T-\tau}\sum_{t=\tau+1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \frac{1}{T-\tau}M + \frac{1}{T-\tau}\frac{HG^2}{2}\sum_{t=\tau+1}^{T} \eta_t^2. \tag{6}$$

Finally, we have

$$\begin{aligned}\frac{1}{T-\tau}\sum_{t=\tau+1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\le \frac{1}{\eta_T(T-\tau)}M + \frac{1}{\eta_T(T-\tau)}\frac{HG^2}{2}\sum_{t=\tau+1}^{T}\eta_t^2 \\ &\le \frac{1}{\eta_T(T-\tau)}\left(M + \frac{HG^2}{2}\sum_{t=\tau+1}^{T}\eta_t^2\right).\end{aligned} \tag{7}$$

□

(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:

Proof. From Eqn. (2), we know that

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta_t\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] + \frac{H\eta_t^2G^2}{2}. \tag{8}$$

Thus, we will have

$$\eta_t\,\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] + \frac{H\eta_t^2G^2}{2}. \tag{9}$$

By summing up from $t = \tau + 1$ to T, we will have

$$\begin{aligned}\sum_{t=\tau+1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\le \sum_{t=\tau+1}^{T}\frac{\mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})]}{\eta_t} + \sum_{t=\tau+1}^{T}\frac{HG^2}{2}\eta_t \\ &= \frac{\mathbb{E}[\mathcal{L}(\mathbf{w}_{\tau+1})]}{\eta_{\tau+1}} - \frac{\mathbb{E}[\mathcal{L}(\mathbf{w}_T)]}{\eta_{\tau_T}} + \sum_{t=\tau+1}^{T-1}(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t})\,\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \\ &+ \frac{HG^2}{2}\sum_{t=\tau+1}^{T}\eta_t \\ &\le M\sqrt{\tau+1} + M(\sqrt{T} - \sqrt{\tau+1}) + \frac{HG^2}{2}\sum_{t=\tau+1}^{T}\eta_t \\ &\le M\sqrt{T} + \frac{HG^2}{2}\sum_{t=\tau+1}^{T}\eta_t \\ &\le M\sqrt{T} + \frac{HG^2}{2}\sum_{t=\tau+1}^{T}\frac{1}{\sqrt{t}}.\end{aligned} \tag{10}$$

Thus, we finally have

$$
\begin{aligned}
\mathbb{E}[\|\nabla \mathcal{L}(\hat{\mathbf{w}})\|^2] &= \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \\
&\leq \frac{1}{T} \sum_{t=\tau+1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \\
&\leq M \frac{\sqrt{T}}{T} + \frac{1}{T} \frac{HG^2}{2} \sum_{t=\tau+1}^{T} \frac{1}{\sqrt{t}} \\
&\leq M \frac{1}{\sqrt{T}} + \frac{1}{T} \frac{HG^2}{2} \sum_{t=\tau+1}^{T} \frac{1}{\sqrt{t}} \\
&\leq O\left(\frac{1}{\sqrt{T}}\right).
\end{aligned} \tag{11}
$$

$\square$