Optimization for Machine Learning HW 2

Shuyue Jia
BUID: U62343813

Due: 9/20/2023

All parts of each question are equally weighted. When solving one question/part, you may assume the results of all previous questions/parts. This HW provides an alternative analysis of SGD in the convex setting that provides a convergence bound for the *last iterate*: $\mathbb{E}[\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star)] = \tilde{O}(1/\sqrt{T})$.

1. Prove the following technical identity: for any sequence of numbers $a_1, \dots, a_T$ with $T > 1$,

$$T \cdot a_T = \sum_{t=1}^{T} a_t + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} (a_t - a_k)$$

(Hint: There are a number of different ways to show this. One way starts by showing that $\frac{T-k+1}{T-k} \sum_{t=k+1}^{T} a_t = \sum_{t=k}^{T} a_t + \frac{1}{T-k} \sum_{t=k}^{T} (a_t - a_k)$ and uses induction on k. Another is to rearrange the terms in the sums to directly show equality. For this, you might want to show the useful identity $\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t = \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k$, valid for all a and b. You might also want to observe that $\frac{T}{(T-k)(T-k+1)} = \frac{T}{T-k} - \frac{T}{T-k+1}$).

Proof. We will start with the second hint: "Another is to rearrange the terms in the sums to directly show equality. For this, you might want to show the useful identity $\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t = \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k$, valid for all a and b. You might also want to observe that $\frac{T}{(T-k)(T-k+1)} = \frac{T}{T-k} - \frac{T}{T-k+1}$)."

Firstly, we need to prove the hint identity, *i.e.*, $\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t = \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k$, because it will be useful to solve this problem. We first expand the summation of a_t from $t = k$ to T:

$$\begin{aligned}
&\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t \\
&= \sum_{k=1}^{T-1} b_k \left(a_k + a_{k+1} + \ldots + a_T\right) \\
&= \sum_{k=1}^{T-1} \left(b_k a_k + b_k a_{k+1} + \ldots + b_k a_T\right) \\
&= \sum_{k=1}^{T-1} b_k a_k + \sum_{k=1}^{T-1} b_k a_{k+1} + \ldots + \sum_{k=1}^{T-1} b_k a_T \\
&= a_1 \sum_{k=1}^{T-1} b_k + a_2 \sum_{k=1}^{T-1} b_k + \ldots + a_{T-1} \sum_{k=1}^{T-1} b_k + a_T \sum_{k=1}^{T-1} b_k \\
&= \left(a_1 + a_2 + \ldots + a_{T-1} + a_T\right) \sum_{k=1}^{T-1} b_k
\end{aligned} \tag{1}$$

Then, for the right-hand sum, we can expand the summation of a_t and b_k:

$$
\begin{aligned}
&\sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k \\
=&(a_1 + a_2 + \ldots + a_{T-1}) \sum_{k=1}^{1} b_k + (a_1 + a_2 + \ldots + a_{T-1}) \sum_{k=1}^{2} b_k \\
&+ \ldots + (a_1 + a_2 + \ldots + a_{T-1}) \sum_{k=1}^{T-1} b_k + a_T \sum_{k=1}^{T-1} b_k \\
=&(a_1 + a_2 + \ldots + a_{T-1}) \left(\sum_{k=1}^{1} b_k + \sum_{k=1}^{2} b_k + \ldots + \sum_{k=1}^{T-1} b_k \right) + a_T \sum_{k=1}^{T-1} b_k \\
=&(a_1 + a_2 + \ldots + a_{T-1}) \sum_{k=1}^{T-1} b_k + a_T \sum_{k=1}^{T-1} b_k \\
=&(a_1 + a_2 + \ldots + a_T) \sum_{k=1}^{T-1} b_k
\end{aligned}
\tag{2}
$$

As a result, we can prove this identity $\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t = \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k$ holds true for all a and b.

Secondly, let $b_k = \frac{T}{(T-k)(T-k+1)}$ to simplify computation. Then, we obtain:

$$
\begin{aligned}
&\sum_{t=1}^{T} a_t + \sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} (a_t - a_k) \\
&= \sum_{t=1}^{T} a_t + \sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t - \sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_k
\end{aligned}
\tag{3}
$$

Finally, from Eqn. (1) and Eqn. (2), we know that $\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t = \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k$. Then, we will have

$$
\begin{aligned}
&\sum_{t=1}^{T} a_t + \sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} (a_t - a_k) \\
&= \sum_{t=1}^{T} a_t + \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k - \sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_k \\
&= \sum_{t=1}^{T} a_t + \sum_{t=1}^{T-1} T(\frac{1}{T-t} - \frac{1}{T}) a_t + T(1 - \frac{1}{T}) a_T - \sum_{k=1}^{T-1} b_k (T - K + 1) a_k \\
&= \sum_{t=1}^{T} a_t + \sum_{t=1}^{T-1} \frac{t}{T-t} a_t + (1 - \frac{1}{T}) a_T - \sum_{k=1}^{T-1} \frac{T}{T-k} a_k \\
&= \sum_{t=1}^{T} a_t + \sum_{t=1}^{T-1} (\frac{t}{T-t} - \frac{T}{T-t}) + T a_T - a_T \\
&= a_T + \sum_{t=1}^{T-1} a_t - \sum_{t=1}^{T-1} a_t + T a_T - a_T \\
&= T a_T
\end{aligned}
\tag{4}
$$

$\square$

2. Consider stochastic gradient descent with a constant learning rate η: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla \ell(\mathbf{w}_t, z_t)$. Suppose that ℓ is convex and G-Lipschitz. Show that for all k:

$$\sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] \leq \frac{\eta(T-k+1)G^2}{2}$$

Proof. The proof is actually very similar to the proof for the **Theorem 3.2** of Stochastic Gradient Descent. All expectations presented here are not over the randomness of the algorithm, *i.e.*, over the choices $z_1, ..., z_T$. Besides, we denote $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t)$ for simplicity.

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{w}_t - \eta \mathbf{g}_t - \mathbf{w}_k\|^2\right] \\ &= \mathbb{E}\left[\|\left(\mathbf{w}_t - \mathbf{w}_k\right) - \eta \mathbf{g}_t\|^2\right] \\ &= \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}_k\|^2 - 2\eta \left\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_k \right\rangle + \eta^2 \|\mathbf{g}_t\|^2\right] \end{aligned} \tag{1}$$

By rearranging the above equation, we have:

$$\begin{aligned} &\mathbb{E}\left[\left\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_k \right\rangle\right] \\ &= \frac{\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2 - \|\mathbf{w}_t - \mathbf{w}_k\|^2\right]}{2\eta} + \frac{\eta\, \mathbb{E}\left[\|\mathbf{g}_t\|^2\right]}{2} \end{aligned} \tag{2}$$

Then, we start from the perspective of unbiasedness and convexity, and we can also observe that $\mathbf{w}_t$ is a deterministic function of $z_1, ..., z_{t-1}$, and that $\mathbf{g}_t$ is independent of $z_1, ..., z_{t-1}$ given $\mathbf{w}_t$,

$$\begin{aligned} &\mathbb{E}\left[\left\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_k \right\rangle\right] \\ &= \mathop{\mathbb{E}}_{z_1,...,z_{t-1}}\left[\mathbb{E}[\left\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{w}_k \right\rangle | z_1, ..., z_{t-1}]\right] \\ &= \mathop{\mathbb{E}}_{z_1,...,z_{t-1}}\left[\left\langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_k \right\rangle\right] \\ &= \mathbb{E}\left[\left\langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_k \right\rangle\right] \\ &\geq \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)\right] \end{aligned} \tag{3}$$

Thus, according to Eqn. (2) and Eqn. (3), we have:

$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)\right] \leq \frac{\mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2 - \|\mathbf{w}_t - \mathbf{w}_k\|^2\right]}{2\eta} + \frac{\eta\, \mathbb{E}\left[\|\mathbf{g}_t\|^2\right]}{2} \tag{4}$$

Next, we sum $\mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)\right]$ from $t = k$ to T,

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=k}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)\right] \\ &\leq \sum_{t=k}^{T} \frac{\eta\, \mathbb{E}\left[\|\mathbf{g}_t\|^2\right]}{2} - \frac{\mathbb{E}\left[\|\mathbf{w}_{T+1} - \mathbf{w}_k\|^2\right]}{2\eta} \\ &\leq \eta \frac{\sum_{t=k}^{T} \mathbb{E}\left[\|\mathbf{g}_t\|^2\right]}{2} \end{aligned} \tag{5}$$

Finally, since ℓ is convex and G-Lipschitz, we have $\mathbf{g}_t = \nabla \ell(\mathbf{w}_t, z_t) \leq G$. As a result, we get:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=k}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)\right] &\leq \frac{\eta \sum_{t=k}^{T} G^2}{2} \\ &\leq \frac{\eta\left(T - k + 1\right) G^2}{2} \end{aligned} \tag{6}$$

$\square$

3. Show that for G-Lipschitz convex losses, SGD with constant learning rate $\eta = \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|}{G\sqrt{T}}$ guarantees:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star)] \leq O\left(\frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G \log(T)}{\sqrt{T}}\right)$$

(Hint: you will need to show $\sum_{t=1}^{T} \frac{1}{t} \leq 1+\log(T)$. As an intermediate step, try showing $\sum_{t=2}^{T} \frac{1}{t} \leq \int_1^T \frac{dt}{t}$ - note the sum starts at 2. Drawing a picture might help).

By having a learning rate that changes appropriately over time (called a "schedule") it is possible to eliminate the logarithmic factor, but it is quite difficult to do so - finding such a schedule was open until as recently as 2019! See `https://arxiv.org/abs/1904.12443` for the first such result via a very complicated schedule and analysis. Just this summer, `https://arxiv.org/abs/2307.11134` provided a much tighter analysis with a simpler learning rate.

Proof. In **Problem 1**, we obtain

$$Ta_T = \sum_{t=1}^{T} a_t + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} (a_t - a_k) \tag{1}$$

Here, let $a_t = \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)]$ and $a_t - a_k = \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)]$. Then, we will have:

$$T \cdot \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] = \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] \tag{2}$$

Then, from the results of **Theorem 3.2** in the Lecture Notes, we have:

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] &\leq \sum_{t=1}^{T} \frac{\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_\star\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta\, \mathbb{E}[\mathbf{g}_t]^2}{2} \\ &\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} \end{aligned} \tag{3}$$

By applying Eqn. (3) to Eqn. (2), we get:

$$\begin{aligned} T \cdot \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &= \sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] \\ &\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] \end{aligned} \tag{4}$$

Next, in **Problem 2**, we have:

$$\mathbb{E}\left[\sum_{t=k}^{T} \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)\right] \leq \frac{\eta\,(T-k+1)\,G^2}{2} \tag{5}$$

By applying Eqn. (5) to Eqn. (4), we get:

$$
\begin{aligned}
T \cdot \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] \\
&\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \frac{\eta\,(T-k+1)\,G^2}{2} \\
&\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} + \frac{\eta T G^2}{2} \sum_{k=1}^{T-1} \frac{1}{(T-k)}
\end{aligned}
\tag{6}
$$

Here, since $\sum_{k=1}^{T-1} \frac{1}{(T-k)} = \sum_{k=1}^{T-1} \frac{1}{k} = 1 + \frac{1}{1} + \cdots + \frac{1}{T-1}$, the above inequality can be written as follows,

$$
\begin{aligned}
T \cdot \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} + \frac{\eta T G^2}{2} \sum_{k=1}^{T-1} \frac{1}{(T-k)} \\
&\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta T G^2}{2} + \frac{\eta T G^2}{2} \sum_{k=1}^{T-1} \frac{1}{k}
\end{aligned}
\tag{7}
$$

Further, we know that the learning rate η is constant, *i.e.*, $\eta = \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|}{G\sqrt{T}}$. We use this η in the above inequality.

$$
\begin{aligned}
T \cdot \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]}{2\eta} + \frac{\eta G^2 T}{2} + \frac{\eta G^2 T}{2} \sum_{k=1}^{T-1} \frac{1}{k} \\
&\leq \frac{\mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}_\star\|^2]]G\sqrt{T}}{2\|\mathbf{w}_1 - \mathbf{w}_\star\|} + \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\| G^2 T}{2G\sqrt{T}} + \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G T}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k} \\
&\leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2 G\sqrt{T}}{2\|\mathbf{w}_1 - \mathbf{w}_\star\|} + \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\| G^2 T}{2G\sqrt{T}} + \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G T}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k} \\
&\leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\| G\sqrt{T}}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\| G T}{2\sqrt{T}} + \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G T}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k} \\
&\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G T}{\sqrt{T}} + \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G T}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k}
\end{aligned}
\tag{8}
$$

Here, we learn that $\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log(T)$ from the hint. The proof is as follows,

$$
\begin{aligned}
\sum_{t=1}^{T} \frac{1}{t} &= 1 + \sum_{t=2}^{T} \frac{1}{t} \\
&= 1 + \int_{t=2}^{T} \frac{1}{t} dt \\
&\leq 1 + \int_{t=1}^{T} \frac{1}{t} dt \\
&\leq 1 + (\log(T) - \log(1)) \\
&\leq 1 + \log(T)
\end{aligned}
\tag{9}
$$

Hereby, we get:

$$\begin{aligned}
T \cdot \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| GT}{\sqrt{T}} + \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| GT}{2\sqrt{T}} \sum_{k=1}^{T-1} \frac{1}{k} \\
&\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| GT}{\sqrt{T}} + \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| GT}{2\sqrt{T}} (1 + \log(T-1)) \\
&\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| GT(3 + \log(T-1))}{2\sqrt{T}}
\end{aligned} \tag{10}$$

Finally,

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] &\leq \frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G(3 + \log(T-1))}{2\sqrt{T}} \\
&\leq O\left(\frac{\|\mathbf{w}_\star - \mathbf{w}_1\| G \log(T)}{\sqrt{T}}\right)
\end{aligned} \tag{11}$$

$\square$