Optimization for Machine Learning HW 2

SOLUTIONS

October 8, 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 $T > 1$,

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

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

$$\sum_{k=1}^{T-1} b_k \sum_{t=k}^{T} a_t = \sum_{t=1}^{T-1} a_t \sum_{k=1}^{t} b_k + a_T \sum_{k=1}^{T-1} b_k$$

For any k, we have that the sum $\sum_{t=k+1}^{T} (a_t - a_k)$ has $T - k$ terms. Therefore,

$$\begin{aligned}
\sum_{t=k}^{T} a_t + \frac{1}{T-k} \sum_{t=k}^{T} (a_t - a_k) &= \sum_{t=k}^{T} a_t + \frac{1}{T-k} \sum_{t=k+1}^{T} (a_t - a_k) \\
&= \sum_{t=k}^{T} a_t + \left(\frac{1}{T-k} \sum_{t=k+1}^{T} a_t \right) - a_k \\
&= \left(1 + \frac{1}{T-k} \right) \sum_{t=k+1}^{T} a_t \\
&= \frac{T-k+1}{T-k} \sum_{t=k+1}^{T} a_t
\end{aligned}$$

rearranging, we have:

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

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

Now, sum over k from 1 to $T-1$ and see that the LHS telescopes:

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

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:

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

We will proceed by rewriting the usual convergence argument for SGD, only with $\mathbf{w}_k$ in place of $\mathbf{w}_\star$:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2 &= \|\mathbf{w}_t - \eta\nabla\ell(\mathbf{w}_t, z_t) - \mathbf{w}_k\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}_k\|^2 - 2\eta\langle \nabla\ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_k\rangle + \eta^2\|\nabla\ell(\mathbf{w}_t, z_t)\|^2 \\
\langle \nabla\ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_k\rangle &\le \frac{\|\mathbf{w}_t - \mathbf{w}_k\|^2}{2\eta} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2}{2\eta} + \frac{\eta G^2}{2} \\
\sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] &\le \sum_{t=k}^{T} \frac{\|\mathbf{w}_t - \mathbf{w}_k\|^2}{2\eta} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2}{2\eta} + \frac{\eta G^2}{2} \\
&= -\frac{\|\mathbf{w}_{T+1} - \mathbf{w}_k\|^2}{2\eta} + \frac{\eta G^2(T-k+1)}{2}
\end{aligned}$$

which establishes the desired claim.

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

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

First, let's prove a small lemma: $\sum_{t=1}^{T} \frac{1}{t} \le 1 + \log(T)$. To see this, notice that for $t-1 \le x \le t$, $\frac{1}{x} \ge \frac{1}{t}$. Therefore $\int_{t-1}^{t} \frac{dx}{x} \le \int_{t-1}^{t} \frac{dx}{t} = \frac{1}{t}$. Thus, $\sum_{t=1}^{T} \frac{1}{t} = 1 + \sum_{t=2}^{T} \frac{1}{t} = 1 + \sum_{t=2}^{t} \int_{t-1}^{t} \frac{dx}{x} = 1 + \int_{1}^{T} \frac{dx}{x} = 1 + \log(T)$. Now, we can proceed with the rest of the problem.

Set $a_t = \mathbb{E}[\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star)]$ in part 1. Then we have:

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

now, by SGD analysis in class/lecture notes:

$$\le \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta T G^2}{2} + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)]$$

next, by part 2:

$$\begin{aligned}
&\le \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta T G^2}{2} + \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \frac{\eta G^2(T-k+1)}{2} \\
&= \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta T G^2}{2} + \sum_{k=1}^{T-1} \frac{T\eta G^2}{T-k} \\
&= \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta T G^2}{2}\left(1 + \sum_{s=1}^{T-1} \frac{1}{s}\right) \\
&\le \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2\eta} + \frac{\eta T G^2}{2}(2 + \log(T))
\end{aligned}$$

now, set $\eta = \frac{\|\mathbf{w}_\star - \mathbf{w}_1\|}{G\sqrt{T}}$:

$$\leq O\left(\|\mathbf{w}_\star - \mathbf{w}_1\| G\sqrt{T}\log(T)\right)$$

Dividing both sides by T yields the result.

BONUS: Consider SGD with a *varying* learning rate $\eta_t = \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|}{G\sqrt{t}}$. Show that

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

Define $a_t = 2\eta_t\,\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_*)]$, then from part (a),

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

Let's bound two sums separately.

Second sum: Consider the one-step update first:

$$\|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2 = \|\mathbf{w}_t - \mathbf{w}_k\|^2 - 2\eta_t\langle g_t, \mathbf{w}_t - \mathbf{w}_k\rangle + \eta_t^2\|g_t\|^2.$$

Summing over t from k to T and applying convexity gives:

$$\begin{aligned}\sum_{t=k}^{T} 2\eta_t\,\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] &\leq \sum_{t=k}^{T} \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_k\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_k\|^2 + G^2\eta_t^2] \\ &\leq G^2 \sum_{t=k}^{T} \eta_t^2.\end{aligned}$$

Since $\eta_t \leq \eta_{t-1}$ for all t (i.e., η_t is non-increasing) and $\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \geq 0$, we have:

$$\begin{aligned}\sum_{t=k}^{T} a_t - a_k &= \sum_{t=k}^{T} 2\eta_t\,\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] - 2\eta_k\,\mathbb{E}[\mathcal{L}(\mathbf{w}_k) - \mathcal{L}(\mathbf{w}_\star)] \\ &\leq \sum_{t=k}^{T} 2\eta_t\,\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] - 2\eta_t\,\mathbb{E}[\mathcal{L}(\mathbf{w}_k) - \mathcal{L}(\mathbf{w}_\star)] \\ &= \sum_{t=k}^{T} 2\eta_t\,\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_k)] \leq G^2 \sum_{t=k}^{T} \eta_t^2.\end{aligned}$$

Now we plug this and the definition of η_t in the second sum, which gives:

$$\sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} a_t - a_k = \sum_{k=1}^{T-1} \frac{T}{(T-k)(T-k+1)} \sum_{t=k}^{T} \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{t}$$

Note that $\frac{1}{t} \leq \frac{1}{k}$ for all $t \geq k$, and thus $\sum_{t=k}^{T} \frac{1}{t} \leq \frac{T-k+1}{k}$, so:

$$\begin{aligned}&\leq \sum_{k=1}^{T-1} \frac{T\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{(T-k)k} \\ &\leq \sum_{k=1}^{T/2} \frac{T\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{(T-k)k} + \sum_{k=T/2}^{T-1} \frac{T\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{(T-k)k}\end{aligned}$$

Again note that $\frac{T}{T-k} \leq 2$ for $k \leq \frac{T}{2}$ and similarly $\frac{T}{k} \leq 2$ for $k \geq \frac{T}{2}$, so:

$$\begin{aligned}
&\leq \sum_{k=1}^{T/2} \frac{2\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{k} + \sum_{k=T/2}^{T-1} \frac{2\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{T-k} \\
&\leq 4\|\mathbf{w}_1 - \mathbf{w}_*\|^2(1 + \log \frac{T}{2}).
\end{aligned}$$

First sum: For the first sum, we apply the standard derivation of SGD:

$$\sum_{t=1}^{T} a_t = \sum_{t=1}^{T} 2\eta_t \, \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_*)]$$

Apply convexity and one-step update:

$$\begin{aligned}
&\leq \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}_*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}_*\|^2 + \eta_t^2 \|g_t\|^2] \\
&\leq \|\mathbf{w}_1 - \mathbf{w}_*\|^2 + G^2 \sum_{t=1}^{T} \eta_t^2
\end{aligned}$$

Now plug in the definition of η_t and apply the integral bound:

$$\begin{aligned}
&\leq \|\mathbf{w}_1 - \mathbf{w}_*\|^2 + G^2 \sum_{t=1}^{T} \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|^2}{G^2 t} \\
&\leq \|\mathbf{w}_1 - \mathbf{w}_*\|^2(2 + \log T).
\end{aligned}$$

Put two sums together, we get

$$Ta_T = \frac{\|\mathbf{w}_1 - \mathbf{w}_*\|\sqrt{T}}{G} \, \mathbb{E}[\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_*)] \leq \|\mathbf{w}_1 - \mathbf{w}_*\|^2(6 + 5\log T),$$

so rearranging terms proves the bound.