Optimization for Machine Learning HW 4

Shuyue Jia
BUID: U62343813

Due: 10/20/2023

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 a little theoretical motivation for some ideas encountered in practice (e.g. [Smith et al., 2018, https://openreview.net/pdf?id=B1Yy1BxCZ]).

1. Suppose that you run the SGD update with a constant learning rate and a gradient estimate $\mathbf{g}_t$: $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\mathbf{g}_t$ where $\mathbb{E}[\mathbf{g}_t] = \nabla\mathcal{L}(\mathbf{w}_t)$. So far, we have considered only the case $\mathbf{g}_t = \nabla\ell(\mathbf{w}_t, z_t)$, but it might be any other random quantity, so long as $\mathbb{E}[\mathbf{g}_t] = \nabla\mathcal{L}(\mathbf{w}_t)$. Suppose that $\mathcal{L}$ is an H-smooth function, and suppose $\mathbb{E}[\|\mathbf{g}_t - \nabla\mathcal{L}(\mathbf{w}_t)\|^2] \leq \sigma_t^2$ for some sequence of numbers $\sigma_1, \sigma_2, \ldots, \sigma_T$. Suppose $\eta \leq \frac{1}{H}$, and let $\Delta = \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)$ where $\mathbf{w}_\star = \operatorname{argmin} \mathcal{L}(\mathbf{w})$. Show that

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] \leq \frac{2\Delta}{\eta} + H\eta\sum_{t=1}^{T}\sigma_t^2$$

Solution:
Proof. Since that $\mathcal{L}$ is an H-smooth function, we will have

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

Now, in deference to the randomness, we take the expected value of both sides:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\,\mathbb{E}\left[\langle\nabla\ell(\mathbf{w}_t), \mathbf{g}_t\rangle\right] + \frac{H\eta^2}{2}\,\mathbb{E}[\|\mathbf{g}_t\|^2]. \tag{2}$$

Since $\mathbb{E}[\mathbf{g}_t] = \nabla\mathcal{L}(\mathbf{w}_t)$, we will have:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\,\mathbb{E}\left[\|\nabla\ell(\mathbf{w}_t)\|^2\right] + \frac{H\eta^2}{2}\,\mathbb{E}[\|\mathbf{g}_t\|^2]. \tag{3}$$

From bias variance decomposition:

$$\begin{aligned} \mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] &\leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta\,\mathbb{E}\left[\|\nabla\ell(\mathbf{w}_t)\|^2\right] + \frac{H\eta^2}{2}\,\mathbb{E}[\|\nabla\ell(\mathbf{w}_t)\|^2 + \sigma_t^2] \\ &= \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \eta(1 - \frac{\eta H}{2})\,\mathbb{E}[\|\nabla\ell(\mathbf{w}_t)\|^2] + \frac{H\eta^2\sigma_t^2}{2}. \end{aligned} \tag{4}$$

Since $\eta \leq \frac{1}{H}$,

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{\eta}{2}\,\mathbb{E}[\|\nabla\ell(\mathbf{w}_t)\|^2] + \frac{H\eta^2\sigma_t^2}{2}. \tag{5}$$

Summing over t and telescoping,

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_1)] \leq -\sum_{t=1}^{T} \frac{\eta}{2}\,\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2] + \frac{H\eta^2}{2}\sum_{t=1}^{T} \sigma_t^2. \tag{6}$$

Thus, we will have:

$$\sum_{t=1}^{T} \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2] \leq \frac{2\Delta}{\eta} + H\eta \sum_{t=1}^{T} \sigma_t^2. \tag{7}$$

$\square$

2. Suppose that $\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, z)]$ and $\mathcal{L}$ is H-smooth and $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z) - \nabla \mathcal{L}(\mathbf{w})\|^2] \leq \sigma^2$ for all $\mathbf{w}$. Consider SGD with constant learning rate $\eta = \frac{1}{H}$, but where the tth iterate uses a minibatch of size t. That is, at each iteration t, we sample t independent random values $z_{t,1}, \ldots, z_{t,t}$ and set:

$$\mathbf{g}_t = \frac{1}{t}\sum_{i=1}^{t} \nabla \ell(\mathbf{w}_t, z_{t,i})$$
$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\mathbf{g}_t}{H}$$

Define $\Delta = \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)$ where $\mathbf{w}_\star = \operatorname{argmin} \mathcal{L}(\mathbf{w})$. Show that

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

Solution:
Proof. Since that $\mathcal{L}$ is an H-smooth function and $\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\mathbf{g}_t}{H}$, we will have

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

Since we know $\eta = \frac{1}{H}$, then we will have:

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}_t) - \frac{1}{H}\langle \nabla \ell(\mathbf{w}_t), \mathbf{g}_t\rangle + \frac{1}{2H}\|\mathbf{g}_t\|^2. \tag{2}$$

Now, in deference to the randomness, we take the expected value of both sides:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] \leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{1}{H}\,\mathbb{E}\left[\langle \nabla \ell(\mathbf{w}_t), \mathbf{g}_t\rangle\right] + \frac{1}{2H}\,\mathbb{E}[\|\mathbf{g}_t\|^2]. \tag{3}$$

From bias variance decomposition and $\mathbb{E}[\|\mathbf{g}_t - \nabla \ell(\mathbf{w}_t)\|^2] \leq \frac{\sigma^2}{t}$:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\mathbf{w}_{t+1})] &\leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{1}{H}\,\mathbb{E}\left[\|\nabla \ell(\mathbf{w}_t)\|^2\right] + \frac{1}{2H}\,\mathbb{E}\left[\|\nabla \ell(\mathbf{w}_t)\|^2 + \frac{\sigma^2}{t}\right] \\
&\leq \mathbb{E}[\mathcal{L}(\mathbf{w}_t)] - \frac{2}{H}\,\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2 + \frac{\sigma^2}{2tH}.
\end{aligned} \tag{4}$$

Summing over t and telescoping,

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_1)] \leq -\frac{2}{H}\sum_{t=1}^{T}\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2] + \frac{\sigma^2}{2H}\sum_{t=1}^{T}\frac{1}{t}. \tag{5}$$

Since we know: $\sum_{t=1}^{T}\frac{1}{t} = 1 + \log(T)$

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_1)] \leq -\frac{2}{H}\sum_{t=1}^{T}\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2] + \frac{\sigma^2}{2H}(1+\log(T)). \tag{6}$$

Thus, we will have:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2] &\leq \frac{\Delta H}{2} + \frac{\sigma^2}{4}(1+\log(T)) \\ &\leq O\left(\Delta H + \sigma^2 \log(T)\right).\end{aligned} \tag{7}$$

$\square$

3. Let N be the total number of gradient evaluations in question 2. Show that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|] \leq O\left(\frac{\sqrt{\log(N)}}{N^{1/4}}\right)$$

where here we consider Δ, H, σ all constant for purposes of big-O. Note that this is the average of $\|\nabla \mathcal{L}(\mathbf{w}_t)\|$ rather than $\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2$. Compare this result to what you might obtain with using a varying learning rate but a fixed batch size (one sentence of comparison here is sufficient).

Solution:
Proof. In **Problem 2**, we proof:

$$\sum_{t=1}^{T}\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2] \leq O\left(\Delta H + \sigma^2 \log(T)\right). \tag{1}$$

Besides, since N is the total number of gradient evaluations,

$$N = \sum_{t=1}^{T} t = \frac{T(T+1)}{2} = \frac{T^2+T}{2} = O(T^2). \tag{2}$$

In other words, we will have

$$T = O(\sqrt{N}). \tag{3}$$

Through the Jensen Inequality, *i.e.*, $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$ we will have:

$$\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2] \geq \mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|]^2. \tag{4}$$

Thus,

$$\sqrt{\mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|^2]} \geq \mathbb{E}[\|\nabla \ell(\mathbf{w}_t)\|]. \tag{5}$$

From Eqn. (2), we can know that $\log(N) = \log(\frac{T^2+T}{2}) \approx 2\log(T)$. Thus, we will have:

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\ell(\mathbf{w}_t)\|] &\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\ell(\mathbf{w}_t)\|^2]} \\
&\leq \sqrt{\frac{1}{T}O(\Delta H + \sigma^2\log(T))} \\
&\leq O\left(\sqrt{\frac{\Delta H + \sigma^2\log(T)}{T}}\right) \\
&\leq O\left(\sqrt{\frac{\Delta H + \sigma^2\frac{\log(N)}{2}}{\sqrt{N}}}\right) \\
&\leq O\left(\sqrt{\frac{\Delta H}{\sqrt{N}} + \frac{\sigma^2\log(N)}{2\sqrt{N}}}\right) \\
&\leq O\left(\sqrt{\frac{2\Delta H + \sigma^2\log(N)}{2\sqrt{N}}}\right) \\
&\leq O\left(\frac{\sqrt{2\Delta H + \sigma^2\log(N)}}{\sqrt{2}N^{\frac{1}{4}}}\right) \\
&\leq O\left(\frac{\sqrt{\log(N)}}{N^{\frac{1}{4}}}\right).
\end{aligned}
\tag{6}
$$

□