Optimization for Machine Learning HW 4

SOLUTIONS

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 $\mathbb{E}[\|\mathbf{g}_t - \nabla \mathcal{L}(\mathbf{w}_t)\|^2] \le \sigma_t^2$ for some sequence of numbers $\sigma_1, \sigma_2, \ldots, \sigma_T$. Suppose $\eta \le \frac{1}{H}$, 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

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

From smoothness, we have:

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

taking expectations:

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

by bias-variance decomposition, we have $\mathbb{E}[\|\mathbf{g}_t\|^2] \le \mathbb{E}[\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \sigma_t^2]$:

$$\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \eta \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{g}_t \rangle + \frac{H}{2}\eta^2 (\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \sigma_t^2)\right]$$

using $\mathbb{E}[\mathbf{g}_t] = \nabla \mathcal{L}(\mathbf{w}_t)$:

$$\begin{aligned} &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \eta \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \frac{H}{2}\eta^2 (\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \sigma_t^2)\right] \\ &= \mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - (\eta - \eta^2 H/2) \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \frac{H}{2}\eta^2 \sigma_t^2\right] \end{aligned}$$

using $\eta \le 1/H$:

$$\le \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \frac{\eta}{2} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 + \frac{H}{2}\eta^2 \sigma_t^2]$$

now rearrange terms:

$$\mathbb{E}\left[\frac{\eta}{2} \|\nabla \mathcal{L}(\mathbf{w}_t)\|^2\right] \le \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1}) + \frac{H}{2}\eta^2 \sigma_t^2]$$

now, sum over all t and telescope the RHS:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\frac{\eta}{2}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] &\le \mathbb{E}\left[\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star) + \frac{H}{2}\eta^2\sigma_t^2\right] \\
&\le \Delta + \frac{H}{2}\eta^2\sigma_t^2 \\
\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2\right] &\le \frac{2\Delta}{\eta} + H\eta\sum_{t=1}^{T}\sigma_t^2
\end{aligned}$$

2. Suppose that $\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, z)]$ and $\mathcal{L}$ is H-smooth and $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z) - \nabla\mathcal{L}(\mathbf{w}_t)\|^2] \le \sigma^2$ 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 $z_{t,1}, \dots, z_{t,t}$ and set:

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

Show that

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

Observe that this procedure fits into the framework analyzed in the previous question: we just need to calculate σ_t.

Since $\mathbf{g}_t$ is an average of t independent quantities with mean $\nabla\mathcal{L}(\mathbf{w}_t)$, we have:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{g}_t - \nabla\mathcal{L}(\mathbf{w}_t)\|^2] &= \frac{1}{t^2}\,\mathbb{E}\left[\|\sum_{i=1}^{t}(\nabla\ell(\mathbf{w}_t, z_{t,i}) - \nabla\mathcal{L}(\mathbf{w}_t))\|^2\right] \\
&= \frac{1}{t^2}\,\mathbb{E}\left[\sum_{i=1}^{t}\|\nabla\ell(\mathbf{w}_t, z_{t,i}) - \nabla\mathcal{L}(\mathbf{w}_t))\|^2 + \sum_{i\ne j}\langle\nabla\ell(\mathbf{w}_t, z_{t,i}) - \nabla\mathcal{L}(\mathbf{w}_t)), \nabla\ell(\mathbf{w}_t, z_{t,j}) - \nabla\mathcal{L}(\mathbf{w}_t))\rangle\right]
\end{aligned}$$

using $\mathbb{E}[\nabla\ell(\mathbf{w}_t, z_{t,i}) - \nabla\mathcal{L}(\mathbf{w}_t))] = 0$ and $z_{t,i}$ independent from $z_{t,j}$:

$$\begin{aligned}
&= \frac{1}{t^2}\,\mathbb{E}\left[\sum_{i=1}^{t}\|\nabla\ell(\mathbf{w}_t, z_{t,i}) - \nabla\mathcal{L}(\mathbf{w}_t))\|^2\right] \\
&\le \frac{\sigma^2}{t}
\end{aligned}$$

Thus, by the previous question's result:

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2] &\le \frac{2\Delta}{\eta} + H\eta\sum_{t=1}^{T}\sigma_t^2 \\
&\le \frac{2\Delta}{\eta} + H\eta\sum_{t=1}^{T}\frac{\sigma^2}{t} \\
&= 2\Delta H + \sigma^2\sum_{t=1}^{T}\frac{1}{t} \\
&\le 2\Delta H + \sigma^2(1 + \log(T))
\end{aligned}$$

where in the last line we have used the identity $\sum_{t=1}^{T} \leq 1 + \int_1^T \frac{dx}{x} \leq 1 + \log(T)$ that has been proven in previous homeworks.

Now, note that $1 + \log(T)$ is $O(\log(T))$ to finish the result.

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 here is sufficient).

At the tth iteration, we evaluate t gradients. Thus the total number of gradient evaluations is:

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

From this we can conclude:

$$N \leq T^2 \implies T \leq \sqrt{N} \tag{1}$$

$$N \geq \frac{T^2}{4} \implies \frac{1}{T} \leq \frac{2}{\sqrt{N}} \tag{2}$$

Now, by Cauchy-Schwarz:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\mathbf{w}_t)\|^2}$$

so that by Jensen:

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

NOTE: at this point it would be fine to just skip some steps and insert your bounds on N to obtain the big-O statement. These solutions will do it in a bit more detail just to be instructional.

applying (1) and (2):

$$\leq \sqrt{\frac{4\Delta H + 2\sigma^2(1+\log(\sqrt{N}))}{\sqrt{N}}}$$

using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ twice:

$$\begin{aligned}
&\leq \frac{2\sqrt{\Delta H} + \sigma\sqrt{2} + \sigma\sqrt{\log(N)}}{N^{1/4}} \\
&= O\left(\frac{\sqrt{\log(N)}}{N^{1/4}}\right)
\end{aligned}$$

When comparing to what we get with batch size 1, notice that in the past we have used $\eta_t = \frac{1}{t}$ to get the rate $O\left(\frac{\sqrt{\log(T)}}{T^{1/4}}\right)$ (e.g. see Theorem 2 in Notes 4) . Since batch-size 1 implies $N = T$, this is actually *the same* rate as we just obtained. However, in the previous homework we obtained the rate $O\left(\frac{\log(T))^{1/4}\sqrt{\log\log(T)}}{T^{1/4}}\right)$, which is a little better. This is being unfair to the batch-size argument however: if we were to instead make the batch size at time t something like $t\log(t)$, we could also match this result from the previous homework. In general for large N there is a perfect correspondence between asymptotic rates with constant learning rate $1/H$ and increasing batch sizes $B_1, B_2, \ldots$ and rates with batch size 1 by decreasing learning rates $\eta_t = \frac{1}{\sqrt{B_t}}$.