Optimization for Machine Learning HW 4

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

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

   Solution:

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

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

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

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

   Solution:

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

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

   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: