Optimization for Machine Learning HW 2

name here

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 $T > 1$,

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

   (Hint: There are a number of different ways to show this. One way starts by showing that $\frac{T-k+1}{T-k} \sum_{t=k+1}^{T} a_t = \sum_{t=k}^{T} a_t + \frac{1}{T-k} \sum_{t=k}^{T} (a_t - a_k)$ 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 $\frac{T}{(T-k)(T-k+1)} = \frac{T}{T-k} - \frac{T}{T-k+1}$).

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:

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

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:

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

   (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.

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

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