Optimization for Machine Learning HW 6

Due: 11/17/2023

All parts of each question are equally weighted. When solving one question/part, you may assume the results of all previous questions/parts. You may also assume all previous homework results and results from class or lecture notes, but please explain which result you are using when you use it.

1. We have seen that for H-smooth convex objectives, no first-order algorithm can faster than $O(H/T^2)$ in a *dimension-free* manner. That is, the "hard" function we studied is a very high dimensional function. However, if we restrict to considering $\mathcal{L} : \mathbb{R} \to \mathbb{R}$, the situation is quite different. Provide a first-order algorithm that, given a first-order oracle for a convex function $\mathcal{L} : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{L}$ is H-smooth and achieves its minimum at some $|w_\star| \leq 1$, then after T iterations the algorithm outputs $\hat{w}$ that satisfies $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \leq O(H2^{-2T})$. Be careful: you need to have the 2 in the exponent since 2^{-T} is NOT $O(2^{-2T})$.

   Solution:

2. Suppose you are trying to identify the bias of a coin. We model a coin flip as a "1" if it comes up heads, and "0" otherwise, and let $p_\star$ be the probability that it comes up heads. If $Z \in \{0, 1\}$ is the outcome of a coin flip, it holds that $p_\star = \operatorname{argmin} \mathbb{E}[\ell(w, Z)] = \mathcal{L}(w)$ where $\ell(w, z) = (w - z)^2$. After observing T coin flips $z_1, \dots, z_T$, you make the natural prediction $\hat{p} = \frac{z_1 + \cdots + z_T}{T}$. Show that $\mathbb{E}[\mathcal{L}(\hat{p}) - \mathcal{L}(p_\star)] = \frac{p_\star(1-p_\star)}{T}$. Explain why this does *not* contradict our $\frac{1}{\sqrt{T}}$ lower bound for stochastic convex optimization?

   Solution: