Optimization for Machine Learning HW 6

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 $|w_\star| \le 1$, then after T iterations the algorithm outputs $\hat{w}$ that satisfies $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \le O(H2^{-2T})$ and justify your answer (Hint: you may eventually need to argue that $\mathcal{L}(w) - \mathcal{L}(w_\star) \le \frac{H}{2}\|w - w_\star\|^2$ for all w).

   Solution:

   The algorithm is binary search. We recursively define a sequence of upper bounds b_t and lower bounds a_t such that $a_t \le w_\star \le b_t$ for all t. To start, set $a_1 = -1$ and $b_1 = 1$. Next let $w_t = \frac{a_t+b_t}{2}$ and query the gradient oracle to obtain g_t. If $g_t \ge 0$, then set $b_{t+1} = w_t$ and $a_{t+1} = a_t$. Otherwise, set $b_{t+1} = b_t$ and $a_{t+1} = w_t$ and repeat the process. Put another way, $w_{t+1} = \sum_{i=1}^t 2^{-i} \cdot (-1)^{\text{sign}(g_i)}$. After T iterations, we output $\hat{w} = w_{T+1}$.

   Now, this algorithm maintains the invariant that $w_\star \in [a_t, b_t]$ for all t. To see this, notice that clearly $w_\star \in [a_1, b_1] = [-1, 1]$ since $|w_\star| \le 1$. Next, if $w_\star \in [a_t, b_t]$, then if $g_t \ge 0$, we have by convexity that $\mathcal{L}(w_t + \delta) \ge \mathcal{L}(w_t) + \delta g_t \ge \mathcal{L}(w_t)$ for all positive δ. Therefore, $w_\star \le w_t$ so that $w_t \in [a_t, w_t] = [a_{t+1}, b_{t+1}]$. Conversely, if $g_t < 0$, again by convexity for all positive δ, $\mathcal{L}(w_t - \delta) \ge \mathcal{L}(w_t) - \delta g_t \ge \mathcal{L}(w_t)$ so that $w_\star \in [w_t, b_t] = [a_{t+1}, b_{t+1}]$.

   Finally, since $w_t = \frac{a_t+b_t}{2}$, we have $b_{t+1} - a_{t+1} = \frac{b_t - a_t}{2}$ so that $b_t - a_t \le 2^{2-t}$. Thus, since $w_\star \in [a_T, b_T]$, we have $|w_\star - \hat{w}| \le 2^{2-T}$. Since $\mathcal{L}$ is H-smooth, $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \le \langle \nabla \mathcal{L}(w_\star), \hat{w} - w_\star \rangle + \frac{H}{2}\|\hat{w} - w_\star\|^2 = \frac{H}{2}\|\hat{w} - w_\star\|^2 \le \frac{H}{2} 2^{4-2T}$ which is $O(2^{-2T})$ as desired.

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 $p_\star = \text{argmin}\, \mathbb{E}[\ell(w, Z)] = \mathcal{L}(w)$ where $\ell(w, z) = (w - z)^2$. After observing T coin flips $z_1, \dots, z_T$, you make the natural prediction $\hat{p} = \frac{z_1 + \dots + z_T}{T}$. Show that $\mathbb{E}[\mathcal{L}(\hat{p}) - \mathcal{L}(p_\star)] = \frac{p_\star(1-p_\star)}{T}$. Why does this *not* contradict our $\frac{1}{\sqrt{T}}$ lower bound for stochastic convex optimization?

   Solution:

   We have for any p:

$$\mathcal{L}(p) = \mathop{\mathbb{E}}_{Z}[(p - Z)^2] = \mathbb{E}[p^2 - 2Zp + Z^2] = p^2 - 2p_\star p + p_\star$$

Now, substituting in $\hat{p} = \frac{z_1 + \cdots + z_T}{T}$, we observe:

$$\begin{aligned}
\mathbb{E}[\hat{p}] &= \frac{\mathbb{E}[z_1] + \cdots + \mathbb{E}[z_T]}{T} = p_\star \\
\mathbb{E}[\hat{p}^2] &= \frac{1}{T^2} \sum_{i,j \leq T} \mathbb{E}[z_i z_j] = \frac{1}{T}^2 \left(\sum_{i \neq j} p_\star^2 + \sum_i p_\star \right) \\
&= p_\star^2 + \frac{1}{T}^2 \sum_i (p_\star^2 - p_\star) = p_\star^2 + \frac{p_\star(1 - p_\star)}{T}
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{p})] = p_\star^2 + \frac{p_\star(1 - p_\star)}{T} - 2p_\star^2 + p_\star \\
&= \frac{p_\star(1 - p_\star)}{T} + p_\star(1 - p_\star) \\
&= \frac{p_\star(1 - p_\star)}{T} + \mathcal{L}(p_\star)
\end{aligned}$$

This does not contradict our lower bound because the bound says that there is *some* function for which we cannot converge faster than $O(1/\sqrt{T})$. It does not mean that for *every* function we cannot converge faster than $O(1/\sqrt{T})$.