Optimization for Machine Learning HW 6

Shuyue Jia
BUID: U62343813

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:
   The intuition is to design a **Bisection method** that meets the above requirements:

Algorithm 1 The Designed Bisection Algorithm

Require: First-order oracle for convex function, T, a, b
Initialize $a \leftarrow -1, b \leftarrow 1$
for $t = 1$ to T **do**
  $c \leftarrow \frac{a+b}{2}$ % Midpoint of the interval
  gradient$_c \leftarrow$ oracle.gradient(c) % Gradient at the midpoint
  if gradient$_c > 0$ **then**
    $b \leftarrow c$
  else
    $a \leftarrow c$
  end if
end for
return $\frac{a+b}{2}$ % Final approximation

   Firstly, the bisection method halves the interval $[a, b]$ in each iteration, so after T iterations, the interval size is reduced to $(b - a)/2^T$.

   Since the function $\mathcal{L}$ is H-smooth, the absolute value of its gradient is bounded by H everywhere. Thus, the Lipschitz continuity of the gradient implies that $\mathcal{L}(c) - \mathcal{L}(w_\star) \leq \frac{H}{2}(b - a)$ for any $c \in [a, b]$.

   By combining steps 1 and 2, we have $\mathcal{L}(c) - \mathcal{L}(w_\star) \leq \frac{H}{2^T}(b - a)$.

   Then, let $\hat{w}$ be the final approximation returned by the bisection method. Since $\hat{w}$ lies in the final interval $[a, b]$, we can write $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \leq \frac{H}{2^T}(b - a)$.

   Next, by substituting $b - a$ with the initial interval size, we get $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \leq \frac{H}{2^T}(b - a) = \frac{H}{2^T}(2) = \frac{H}{2^{T-1}}$.

   Finally, since $2^{T-1} = 2^T/2$, we can rewrite the result as $\mathcal{L}(\hat{w}) - \mathcal{L}(w_\star) \leq O(H2^{-2T})$, 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 $p_\star = \operatorname{argmin} \mathbb{E}[\ell(w, Z)] = \mathcal{L}(w)$ where $\ell(w, z) = (w - z)^2$. After observing T coin flips $z_1, \ldots, z_T$, 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:

(1) ***Proof.*** The loss function is given by $\ell(w, z) = (w - z)^2$, and our goal is to find the probability $p_\star$ that minimizes the expected loss $\mathbb{E}[\ell(w, Z)]$, where Z, a Bernoulli random variable, is the outcome of a coin flip.

$$\mathbb{E}\left[(z_1 - p_\star)^2\right] = \mathrm{Var}(Z) = p_\star(1 - p_\star). \tag{1}$$

The expected loss is given by

$$\mathbb{E}[\ell(w, Z)] = \sum_{z \in \{0,1\}} \ell(w, z) P(Z = z). \tag{2}$$

For our coin flip, $p_\star$ minimizes this expectation. Now, let's calculate the expected loss for the natural prediction $\hat{p} = \frac{z_1 + \cdots + z_T}{T}$ after observing T coin flips $z_1, \ldots, z_T$.

Then, we will have:

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(\hat{p}) - \mathcal{L}(p_\star)] &= \mathbb{E}\left[\left(\frac{z_1 + \cdots + z_T}{T} - p_\star\right)^2 - (p_\star - p_\star)^2\right] \\
&= \frac{1}{T^2}\mathbb{E}\left[(z_1 + \cdots + z_T - T p_\star)^2\right] \\
&= \frac{1}{T^2}\sum_{t=1}^{T} \mathbb{E}\left[(z_t - p_\star)^2\right] \quad \text{(by linearity of expectation)} \\
&= \frac{1}{T^2} T \mathbb{E}\left[(z_1 - p_\star)^2\right] \quad \text{(all terms are identical)} \\
&= \frac{1}{T}\mathbb{E}\left[(z_1 - p_\star)^2\right] \\
&= \frac{1}{T} p_\star(1 - p_\star).
\end{aligned} \tag{3}$$

□

(2) ***Explanation*** for why this does not contradict the $\frac{1}{\sqrt{T}}$ lower bound for stochastic convex optimization:

The lower bound of $\frac{1}{\sqrt{T}}$ is a convergence rate for optimization problems where the goal is to find the optimal parameter $w_\star$ in a convex function. In this case, we are not optimizing a convex function directly; instead, we are estimating a probability parameter $p_\star$ based on observed coin flips.

The rate $\frac{1}{\sqrt{T}}$ arises in optimization problems where the objective function is smooth and has a Lipschitz continuous gradient. In the coin flip scenario, the loss function is not differentiable at $p_\star = 0$ and $p_\star = 1$, so it doesn't satisfy the smoothness conditions typically assumed in optimization settings.