Optimization for Machine Learning HW 7

name

Due: 10/27/2023

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.

This homework examines the connection between accelerated algorithms for smooth and strongly-convex functions. In particular, you will devise an algorithm for H-smooth and μ-strongly convex objectives such that after computing N gradient evaluations, the algorithm outputs a $\hat{\mathbf{w}}$ such that (dropping some constants):

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \le \exp\left(-\sqrt{\frac{\mu}{H}}N\right)$$

This is contrast to ordinary gradient descent, for which the guarantee is only $\exp(-\frac{\mu}{H}N)$. Recall that if $\mathcal{L}$ is an H-smooth, convex function, then there is an absolute constant C such that after T gradient evaluations, the accelerated gradient descent algorithm starting from initial point $\mathbf{w}_1$ outputs a point $\mathbf{w}_T$ such that:

$$\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star) \le \frac{CH\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{T^2}$$

For simplicity throughout this problem, you may assume that $\sqrt{8C\frac{H}{\mu}}$ is an integer.

1. Suppose that $\mathcal{L}$ is and H smooth and μ-strongly convex function. Show that $\frac{\mu\|\mathbf{w}-\mathbf{w}_\star\|^2}{2} \le \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \le \frac{H\|\mathbf{w}-\mathbf{w}_\star\|^2}{2}$.

   Solution:
   For the upper bound, from the smoothness Lemma, $\mathcal{L}(\mathbf{w}) \le \mathcal{L}(\mathbf{w}_\star) + \langle \nabla\mathcal{L}(\mathbf{w}_\star), \mathbf{w} - \mathbf{w}_\star\rangle + \frac{H}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2$. Use $\nabla\mathcal{L}(\mathbf{w}_\star) = 0$ and rearrange to see that $\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \le \frac{H\|\mathbf{w}-\mathbf{w}_\star\|^2}{2}$.
   Similarly, for the lower bound we have $\mathcal{L}(\mathbf{w}) \ge \mathcal{L}(\mathbf{w}_\star) + \langle \nabla\mathcal{L}(\mathbf{w}_\star), \mathbf{w} - \mathbf{w}_\star\rangle + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2$. Again, since $\nabla\mathcal{L}(\mathbf{w}_\star) = 0$, we rearrange to see $\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \ge \frac{\mu\|\mathbf{w}-\mathbf{w}_\star\|^2}{2}$.

2. Show that after $T = \sqrt{8C\frac{H}{\mu}}$ iterations of accelerated gradient descent, we have:

   $$\|\mathbf{w}_T - \mathbf{w}_\star\| \le \frac{1}{2}\|\mathbf{w}_1 - \mathbf{w}_\star\|$$

   (hint: can you relate $\|\mathbf{w}_T - \mathbf{w}_\star\|^2$ to $\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star)$?)

   Solution:
   Recall that strongly convex functions satisfy $\frac{\mu}{2}\|\mathbf{w}-\mathbf{w}_\star\|^2 \le \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star)$. Therefore after T iterations,

we have:

$$\begin{aligned}\frac{\mu}{2}\|\mathbf{w}_T - \mathbf{w}_\star\|^2 &\le \mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star)\\ &\le \frac{CH\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{T^2}\\ &\le \frac{CH\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{8CH/\mu}\end{aligned}$$

rearrange and take square roots:

$$\|\mathbf{w}_T - \mathbf{w}_\star\| \le \frac{1}{2}\|\mathbf{w}_1 - \mathbf{w}_\star\|$$

3. Consider an algorithm that runs accelerated gradient descent for $\sqrt{8C\frac{H}{\mu}}$ iterations, then stops, resets $\mathbf{w}_1 = \mathbf{w}_T$, and then restarts and runs accelerated gradient descent for $\sqrt{8C\frac{H}{\mu}}$ iterations and repeats (e.g. Algorithm 1).

Algorithm 1 Restarted Accelerated Gradient Descent

```
Set $x_1 = 0$
for $r = 1 \dots R$ do
    Set $\mathbf{w}_1 = \mathbf{x}_r$
    Initialize and run accelerated gradient descent for $T = \sqrt{8C\frac{H}{\mu}}$ iterations starting from initial iterate $\mathbf{w}_1$, let $\mathbf{w}_T$ be the output.
    Set $\mathbf{x}_{r+1} = \mathbf{w}_T$.
end for
return $\mathbf{x}_{R+1}$.
```

Show that this algorithm statisfies:

$$\|\mathbf{x}_{R+1} - \mathbf{w}_\star\| \le \frac{1}{2^R}\|\mathbf{w}_\star\|$$

Solution:
By the question 1, we have that:

$$\|\mathbf{x}_{r+1} - \mathbf{x}_r\| \le \frac{1}{2}\|\mathbf{x}_r - \mathbf{w}_\star\|$$

Therefore, iterating this recursion r times shows the desired result.

4. Suppose $N = R\sqrt{8C\frac{H}{\mu}}$ for some integer R. Show that after N gradient evaluations, Algorithm 1 outputs a point $\hat{\mathbf{w}} = \mathbf{x}_{R+1}$ that satisfies:

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \le 2^{-\sqrt{\frac{\mu}{2CH}}N}\frac{H\|\mathbf{w}_\star\|^2}{2}$$

Solution:

By Question 3 and Question 2,

$$\begin{aligned}
\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) &\leq \frac{H\|\hat{\mathbf{w}} - \mathbf{w}_\star\|^2}{2} \\
&\leq \frac{H\|\mathbf{w}_\star\|^2}{2 \cdot 2^{2R}} \\
&= \frac{H\|\mathbf{w}_\star\|^2}{2^{\frac{2N}{\sqrt{8C\frac{H}{\mu}}}} H} \\
&= 2^{-\sqrt{\frac{\mu}{2CH}}N} \frac{H\|\mathbf{w}_\star\|^2}{2}
\end{aligned}$$