Optimization for Machine Learning HW 7

Shuyue Jia
BUID: U62343813

Due: 12/4/2022

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):

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

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:

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

For simplicity throughout this problem, you may assume that $\sqrt{8C\frac{H}{\mu}}$ is an integer. You will not need to use any of the internal arguments for how accelerated gradient descent works, or anything in particular about the constant C.

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

Solution:
From **Lemma 24.4** in the Lecture Notes, we know that

$$\mathcal{L}(\mathbf{w}) \geq \mathcal{L}\left(\mathbf{w}_\star\right) + \left\langle \nabla\mathcal{L}\left(\mathbf{w}_\star\right), \mathbf{w} - \mathbf{w}_\star \right\rangle + \frac{\mu}{2}\left\|\mathbf{w} - \mathbf{w}_\star\right\|^2. \tag{1}$$

Note that $\nabla\mathcal{L}\left(\mathbf{w}_\star\right) = 0$, we will have:

$$\mathcal{L}(\mathbf{w}) \geq \mathcal{L}\left(\mathbf{w}_\star\right) + \frac{\mu}{2}\left\|\mathbf{w} - \mathbf{w}_\star\right\|^2. \tag{2}$$

Thus,

$$\frac{\mu\|\mathbf{w} - \mathbf{w}_\star\|^2}{2} \leq \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star). \tag{3}$$

Besides, we know that:

$$\mathcal{L}\left(\mathbf{w}\right) \leq \mathcal{L}\left(\mathbf{w}_\star\right) + \left\langle \nabla\mathcal{L}\left(\mathbf{w}_\star\right), \mathbf{w} - \mathbf{w}_\star \right\rangle + \frac{H}{2}\left\|\mathbf{w} - \mathbf{w}_\star\right\|^2. \tag{4}$$

Note that $\nabla \mathcal{L}(\mathbf{w}_\star) = 0$, we will have:

$$\mathcal{L}(\mathbf{w}) \leq \mathcal{L}(\mathbf{w}_\star) + \frac{H}{2} \|\mathbf{w} - \mathbf{w}_\star\|^2 . \tag{5}$$

Thus, we will have:

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{H\|\mathbf{w} - \mathbf{w}_\star\|^2}{2}. \tag{6}$$

Finally, we have proven:

$$\frac{\mu\|\mathbf{w} - \mathbf{w}_\star\|^2}{2} \leq \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{H\|\mathbf{w} - \mathbf{w}_\star\|^2}{2}. \tag{7}$$

□

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

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

Solution:
From the question, we know that

$$\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{CH\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{T^2}. \tag{1}$$

Since $T = \sqrt{8C\frac{H}{\mu}}$, we will have:

$$\mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{\mu\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{8}. \tag{2}$$

Because the accelerated gradient descent algorithm starting from initial point $\mathbf{w}_1$ outputs a point $\mathbf{w}_T$, we will have:

$$\frac{\mu\|\mathbf{w}_T - \mathbf{w}_\star\|^2}{2} \leq \mathcal{L}(\mathbf{w}_T) - \mathcal{L}(\mathbf{w}_\star). \tag{3}$$

Thus, according to Eqn. (2) and Eqn. (3), we have:

$$\frac{\mu\|\mathbf{w}_T - \mathbf{w}_\star\|^2}{2} \leq \frac{\mu\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{8}. \tag{4}$$

By arranging:

$$\|\mathbf{w}_T - \mathbf{w}_\star\|^2 \leq \frac{\|\mathbf{w}_1 - \mathbf{w}_\star\|^2}{4}. \tag{5}$$

Finally, we will have:

$$\|\mathbf{w}_T - \mathbf{w}_\star\| \leq \frac{1}{2}\|\mathbf{w}_1 - \mathbf{w}_\star\|. \tag{6}$$

□

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 $\mathbf{x}_{R+1}$.

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 (i.e. Algorithm 1).

   Show that this algorithm satisfies for all R:

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

Solution:

From 1, we know that $\mathbf{w}_1 = \mathbf{x}_r$ and $\mathbf{x}_{r+1} = \mathbf{w}_T$. According to the result of **Problem 2**, we will have:

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

Thus, we have:

$$\begin{aligned}
&\|\mathbf{x}_2 - \mathbf{w}_\star\| \le \frac{1}{2}\|\mathbf{x}_1 - \mathbf{w}_\star\|,\\
&\|\mathbf{x}_3 - \mathbf{w}_\star\| \le \frac{1}{2}\|\mathbf{x}_2 - \mathbf{w}_\star\| \le \frac{1}{2^2}\|\mathbf{x}_1 - \mathbf{w}_\star\|,\\
&\dots\\
&\|\mathbf{x}_{r+1} - \mathbf{w}_\star\| \le \frac{1}{2}\|\mathbf{x}_r - \mathbf{w}_\star\| \le \frac{1}{2^r}\|\mathbf{x}_1 - \mathbf{w}_\star\|.
\end{aligned} \tag{2}$$

Since $r = 1 \dots R$, after R iterations, we will have:

$$\|\mathbf{x}_{R+1} - \mathbf{w}_\star\| \le \frac{1}{2}\|\mathbf{x}_r - \mathbf{w}_\star\| \le \frac{1}{2^R}\|\mathbf{x}_1 - \mathbf{w}_\star\|. \tag{3}$$

Since $x_1 = 0$, we finally have:

$$\|\mathbf{x}_{R+1} - \mathbf{w}_\star\| \le \frac{1}{2^R}\|\mathbf{w}_\star\|. \tag{4}$$

□

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:

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

Solution:
From **Problem 3**, we know that:

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

Since $N = R\sqrt{8C\frac{H}{\mu}}$, we will have:

$$\|\mathbf{x}_{R+1} - \mathbf{w}_\star\| \leq \frac{1}{2^{\frac{N}{\sqrt{8C\frac{H}{\mu}}}}}\|\mathbf{w}_\star\|. \tag{2}$$

From Problem 1, we know that:

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{H\|\mathbf{w} - \mathbf{w}_\star\|^2}{2}. \tag{3}$$

Since $\hat{\mathbf{w}} = \mathbf{x}_{R+1}$, we will then have:

$$\mathcal{L}(\mathbf{x}_{R+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{H\|\mathbf{x}_{R+1} - \mathbf{w}_\star\|^2}{2}. \tag{4}$$

According to Eqn. (2), we will have:

$$\mathcal{L}(\mathbf{x}_{R+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{H}{2}\left(\frac{1}{2^{\frac{N}{\sqrt{8C\frac{H}{\mu}}}}}\|\mathbf{w}_\star\|\right)^2. \tag{5}$$

Then, we will have:

$$\mathcal{L}(\mathbf{x}_{R+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{HN}{2}\|\mathbf{w}_\star\|^2 2^{-\sqrt{\frac{\mu}{2HC}}}. \tag{6}$$

By arranging, we will have:

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq 2^{-\sqrt{\frac{\mu}{2CH}}N}\frac{H\|\mathbf{w}_\star\|^2}{2}. \tag{7}$$

Since $2^{-\sqrt{\frac{\mu}{2CH}}N} = \exp\left(-\frac{\log(2)}{\sqrt{2C}}\sqrt{\frac{\mu}{H}}N\right)$, we will finally have:

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq 2^{-\sqrt{\frac{\mu}{2CH}}N}\frac{H\|\mathbf{w}_\star\|^2}{2} = \exp\left(-\frac{\log(2)}{\sqrt{2C}}\sqrt{\frac{\mu}{H}}N\right)\frac{H\|\mathbf{w}_\star\|^2}{2}. \tag{8}$$

□