CS446: Machine Learning, Fall 2018, Homework 1

Name: Aahan Agrawal (agrawl10)

Collaborated with Udit Samani (usamani2), Ankit Gohel (ankitng2)

Problem (3)

Solution:
(a)

By definition of being L-smooth:

$$f(w_i) \leq f(w_{i-1}) + \nabla f(w_{i-1})^T(w_i - w_{i-1}) + \frac{L}{2}\|w_i - w_{i-1}\|^2$$

Since $w_i - w_{i-1} = -\gamma \nabla f(w_{i-1})$

$$= f(w_{i-1}) - \gamma\|\nabla f(w_{i-1})\|^2 + \frac{L}{2}\|\gamma \nabla f(w_{i-1})\|^2$$

Substituting in $\gamma = \frac{1}{L}$

$$\begin{aligned}
&= f(w_{i-1}) - \frac{1}{L}\|\nabla f(w_{i-1})\|^2 + \frac{L}{2L^2}\|\nabla f(w_{i-1})\|^2 \\
&= f(w_{i-1}) - \frac{1}{L}\|\nabla f(w_{i-1})\|^2 + \frac{1}{2L}\|\nabla f(w_{i-1})\|^2 \\
&= f(w_{i-1}) - \frac{1}{2L}\|\nabla f(w_{i-1})\|^2
\end{aligned}$$

Now rearrange the inequality to obtain:

$$\begin{aligned}
f(w_{i-1}) - f(w_i) &\geq \frac{1}{2L}\|\nabla f(w_{i-1})\|^2 \\
2L\left(f(w_{i-1}) - f(w_i)\right) &\geq \|\nabla f(w_{i-1})\|^2 \\
\sum_{i=1}^{T} \frac{2L\left(f(w_{i-1}) - f(w_i)\right)}{T} &\geq \sum_{i=1}^{T} \frac{\|\nabla f(w_{i-1})\|^2}{T}
\end{aligned}$$

The LHS is telescoping sum, so it reduces to:

$$2\frac{f(w_0) - f(w_T)}{\gamma T} \geq \sum_{i=1}^{T} \frac{\|\nabla f(w_{i-1})\|^2}{T}$$

(b)

By definition of being L-smooth:

$$f(w_{i-1}) + \nabla f(w_{i-1})^T(w_i - w_{i-1}) + \frac{L}{2}\|w_i - w_{i-1}\|^2 \geq f(w_i)$$

Since $w_i - w_{i-1} = -\gamma g(w_{i-1})$

$$f(w_{i-1}) + \nabla f(w_{i-1})^T(-\gamma g(w_{i-1})) + \frac{L\gamma^2}{2}\|g(w_{i-1})\|^2 \geq f(w_i)$$

Rearranging the inequality:

$$\frac{L\gamma^2}{2}\|g(w_{i-1})\|^2 - \nabla f(w_{i-1})^T(\gamma g(w_{i-1})) \geq f(w_i) - f(w_{i-1})$$
$$\gamma\left(\frac{L\gamma}{2}\|g(w_{i-1})\|^2 - \nabla f(w_{i-1})^T g(w_{i-1})\right) \geq f(w_i) - f(w_{i-1})$$

Now we *add* the term $\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2$ to the LHS inside the expression multiplied by γ

Adding it requires that we subtract $\frac{1}{2}\|g(w_{i-1})^2\|$ from the LHS inside the expression multiplied by γ and add $\gamma/2\|\nabla f(w_{i-1})\|^2$ to the RHS

$$\gamma\left(\left(\frac{1}{2}\right)\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + (\frac{L\gamma - 1}{2})\|g(w_{i-1})\|^2)\right) \geq f(w_i) - f(w_{i-1}) + \frac{\gamma}{2}\|\nabla f(w_{i-1})\|^2$$
$$\gamma\left(\left(\frac{1}{2}\right)\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + (\frac{L\gamma - 1}{2})\|g(w_{i-1})\|^2)\right) - f(w_i) + f(w_{i-1}) \geq \frac{\gamma}{2}\|\nabla f(w_{i-1})\|^2$$

Multiply out by 2

$$\gamma\left(\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + (L\gamma - 1)\|g(w_{i-1})\|^2)\right) - 2f(w_i) + 2f(w_{i-1}) \geq \gamma\|\nabla f(w_{i-1})\|^2$$

Let us flip orientations to make this consistent with the proof statement.

$$\gamma\|\nabla f(w_{i-1})\|^2 \leq \gamma\left(\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + (L\gamma - 1)\|g(w_{i-1})\|^2)\right) - 2f(w_i) + 2f(w_{i-1})$$

Let us divide by γ

$$\|\nabla f(w_{i-1})\|^2 \leq \|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + (L\gamma - 1)\|g(w_{i-1})\|^2 + \frac{-2f(w_i) + 2f(w_{i-1})}{\gamma}$$

Setting $\gamma = \frac{1}{2L}$ and observing that $\|g(w_{i-1})^2\| \geq 0$

$$\|\nabla f(w_{i-1})\|^2 \leq \|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + (L\frac{1}{2L} - 1)\|g(w_{i-1})\|^2 + \frac{-2f(w_i) + 2f(w_{i-1})}{\gamma}$$
$$\|\nabla f(w_{i-1})\|^2 \leq \|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + \frac{-2f(w_i) + 2f(w_{i-1})}{\gamma}$$

If we take the expectation with respect to $f(w_i)$ and conditional on w_{i-1}, this gives

$$\mathbb{E}\|\nabla f(w_{i-1})\|^2 \leq \mathbb{E}\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 + \frac{2\mathbb{E}\left(f(w_{i-1}) - f(w_i)\right)}{\gamma}$$

If we take the expectation, this time with an underlying probability measure over values that w_{i-1} can take on, then the same foregoing inequality holds. This time, however, the expectation is a total one.

By hypothesis, we know that $\mathbb{E}\|\nabla f(w_{i-1}) - g(w_{i-1})\|^2 \leq \sigma^2$, giving

$$\mathbb{E}\|\nabla f(w_{i-1})\|^2 \leq \sigma^2 + \frac{2\mathbb{E}\left(f(w_{i-1}) - f(w_i)\right)}{\gamma}$$

Taking the average, we then get:

$$\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}\|\nabla f(w_{i-1})\|^2 \leq \sigma^2 + \frac{1}{T}\sum_{i=1}^{T} \frac{2\mathbb{E}\left(f(w_{i-1}) - f(w_i)\right)}{\gamma}$$

After distributing $\mathbb{E}$ on the RHS, and cancelling terms in the telescoping series:

$$\frac{1}{T}\sum_{i=1}^{T} \mathbb{E}\|\nabla f(w_{i-1})\|^2 \leq \sigma^2 + \frac{2\mathbb{E}\left(f(w_0) - f(w_T)\right)}{\gamma T}$$

(c)

$$\begin{aligned}\|w_{t+1} - w^\star\|^2 &= \|(w_{t+1} - w_t) - (w^\star - w_t)\|^2 \\ &= \|w_t - w^\star\|^2 + \|w_{t+1} - w_t\|^2 - 2(w_{t+1} - w_t)^T(w^\star - w_t)\end{aligned}$$

Now substitute that $w_{t+1} - w_t = -\gamma g(w_t)$

$$= \|w_t - w^\star\|^2 + \gamma^2 \|g(w_t)\|^2 + 2\gamma(g(w_t)^T)(w^\star - w_t)$$

Apply expectation, conditional on w_t, we can drop the expectation in the first term

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 = \mathbb{E}\left(\|w_t - w^\star\|^2\right) + \gamma^2 \mathbb{E}\left(\|g(w_t)\|^2\right) + 2\mathbb{E}\gamma(g(w_t)^T)(w^\star - w_t)$$

Since $\mathbb{E}g(w_t) = \nabla f(w_t)$ and $\mathbb{E}g(w^\star) = \nabla f(w^\star) = 0$, we have

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 = \mathbb{E}\left(\|w_t - w^\star\|^2\right) + \gamma^2 \mathbb{E}\left(\|g(w_t)\|^2\right) + 2\gamma(\nabla f(w_t) - \nabla f(w^\star))^T(w^\star - w_t)$$

Since expectation is conditional on w_t

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 = \left(\|w_t - w^\star\|^2\right) + \gamma^2 \mathbb{E}\left(\|g(w_t)\|^2\right) - 2\gamma(\nabla f(w_t) - \nabla f(w^\star))^T(w_t - w^\star)$$

(d)

Applying the lemma to our case gives us:

$$\left[\nabla f(w_t) - \nabla f(w^\star)\right]^T (w_t - w^\star) \geq \frac{\mu L}{\mu + L}\|w_t - w^\star\|^2 + \frac{1}{\mu + L}\|\nabla f(w_t) - \nabla f(w^\star)\|^2$$

Now let us substitute this inequality into the expression from part (c)

$$\begin{aligned}
\mathbb{E}\|w_{t+1} - w^\star\|^2 &\leq \left(\|w_t - w^\star\|^2\right) + \gamma^2 \mathbb{E}\left(\|g(w_t)\|^2\right) - \frac{2\mu L\gamma}{\mu + L}\|w_t - w^\star\|^2 - \frac{2\gamma}{\mu + L}\|\nabla f(w_t) - \nabla f(w^\star)\|^2 \\
\mathbb{E}\|w_{t+1} - w^\star\|^2 &\leq (1 - \frac{2\mu L\gamma}{\mu + L})\left(\|w_t - w^\star\|^2\right) + \gamma^2 \mathbb{E}\left(\|g(w_t)\|^2\right) - \frac{2\gamma}{\mu + L}\|\nabla f(w_t) - \nabla f(w^\star)\|^2
\end{aligned}$$

Now let $\gamma = \frac{2}{\mu + L}$

$$\begin{aligned}
\mathbb{E}\|w_{t+1} - w^\star\|^2 &\leq (1 - \frac{4\mu L}{(\mu + L)^2})\left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\mathbb{E}\left(\|g(w_t)\|^2\right) - \frac{4}{(\mu + L)^2}\|\nabla f(w_t) - \nabla f(w^\star)\|^2 \\
&= (1 - \frac{4\mu L}{(\mu + L)^2})\left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\left(\mathbb{E}\left(\|g(w_t)\|^2\right) - \|\nabla f(w_t) - \nabla f(w^\star)\|^2\right) \\
&= (1 - \frac{4\mu L}{(\mu + L)^2})\left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\left(\mathbb{E}\left(\|g(w_t)\|^2\right) - \|\nabla f(w_t)\|^2\right) \\
&= (1 - \frac{4\mu L}{(\mu + L)^2})\left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\underbrace{\left(\mathbb{E}\left(\|g(w_t)\|^2\right) - \|\nabla f(w_t)\|^2\right)}_{\alpha}
\end{aligned}$$

We observe that α is the same as the expression

$$\mathbb{E}\|g(w_t) - \nabla f(w_t)\|^2$$

Since the foregoing expression expands to

$$\begin{aligned}
&\mathbb{E}\left(\|g(w_t)\|^2 - 2g(w_t)^T \nabla f(w_t) + \|\nabla f(w_t)\|^2\right) \\
&= \mathbb{E}\left(\|g(w_t)\|^2\right) - 2\mathbb{E}(g(w_t))^T \nabla f(w_t) + \mathbb{E}\|\nabla f(w_t)\|^2 \\
&= \mathbb{E}\left(\|g(w_t)\|^2\right) - 2\nabla f(w_t)^T \nabla f(w_t) + \|\nabla f(w_t)\|^2 \\
&= \mathbb{E}\left(\|g(w_t)\|^2\right) - 2\nabla f(w_t)^T \nabla f(w_t) + \|\nabla f(w_t)\|^2 \\
&= \mathbb{E}\left(\|g(w_t)\|^2\right) - \|\nabla f(w_t)\|^2
\end{aligned}$$

Thus, we have:

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 \leq (1 - \frac{4\mu L}{(\mu + L)^2}) \left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\mathbb{E}\left(\|g(w_t) - \nabla f(w_t)\|^2\right)$$

Now apply the hypothesis to get:

$$\begin{aligned}
&\leq (1 - \frac{4\mu L}{(\mu + L)^2}) \left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\sigma^2 \\
&\leq \underbrace{(1 - \frac{4\mu L}{(\mu + L)^2})}_{\beta} \left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\sigma^2
\end{aligned}$$

Observe that β reduces to

$$\kappa = \frac{(\mu + L)^2 - 4\mu L}{(\mu + L)^2} = \frac{(\mu - L)^2}{(\mu + L)^2} \in (0, 1)$$

Subsituting κ in:

$$\leq \kappa \left(\|w_t - w^\star\|^2\right) + \frac{4}{(\mu + L)^2}\sigma^2$$

Thus, we have:

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 \leq \kappa \underbrace{\left(\|w_t - w^\star\|^2\right)}_{*} + \frac{4}{(\mu + L)^2}\sigma^2$$

Note that an implicit expectation, conditional on w_t, acts on $*$

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 \leq \kappa \mathbb{E}\underbrace{\left(\|w_t - w^\star\|^2 | w_t\right)}_{*} + \frac{4}{(\mu + L)^2}\sigma^2$$

If we recurse on $\mathbb{E}\left(\|w_t - w^\star\|^2\right)$ T times, each time taking the expectation with respect to w_{t-1}, then we have

$$\mathbb{E}\|w_{t+1} - w^\star\|^2 \leq \kappa^T \|w_0 - w^\star\|^2 + \mathcal{O}(\sigma^2)$$