(BONUS) Exercise List: Proving convergence of the Stochastic Gradient Descent for smooth and convex functions.

Robert M. Gower

February 10, 2019

1 Introduction

Consider the problem

$$w^* \in \arg\min_{w} \left(\frac{1}{n} \sum_{i=1}^{n} f_i(w) \overset{\text{def}}{=} f(w) \right), \tag{1}$$

where we assume that $f(w)$ is μ–strongly quasi-convex

$$f(w^*) \geq f(w) + \langle w^* - w, \nabla f(w) \rangle + \frac{\mu}{2}\|w - w^*\|^2, \tag{2}$$

and each f_i is convex and L_i–smooth

$$f_i(w+h) \leq f_i(w) + \langle \nabla f_i(w), h \rangle + \frac{L_i}{2}\|h\|^2, \quad \text{for } i = 1, \ldots, n. \tag{3}$$

Here we will provide a modern proof of the convergence of the SGD algorithm

$$w^{t+1} \quad = \quad w^t - \gamma^t \nabla f_{i_t}(w^t), \quad \text{where } i_t \sim \frac{1}{n}. \tag{4}$$

The result we will prove is given in the following theorem.

Theorem 1.1. Assume f is μ-quasi-strongly convex and the f_i's are convex and L_i–smooth. Let $L_{\max} = \max_{i=1,\ldots,n} L_i$ and let

$$\sigma^2 \quad \overset{\text{def}}{=} \quad \sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(w^*)\|^2. \tag{5}$$

Choose $\gamma^t = \gamma \in (0, \frac{1}{2L_{\max}}]$ for all t. Then the iterates of SGD given by (4) satisfy:

$$\mathbb{E}\|w^t - w^*\|^2 \leq (1 - \gamma\mu)^t \|w^0 - w^*\|^2 + \tfrac{2\gamma\sigma^2}{\mu}. \tag{6}$$

2 Proof of Theorem 1.1

We will now give a modern proof of the convergance of SGD.

Ex. 1 — Let $\mathbb{E}_t[\cdot] \stackrel{\text{def}}{=} \mathbb{E}\left[\cdot \,|\, w^t\right]$ and consider the tth iteration of the SGD method (4). Show that

$$\mathbb{E}_t\left[\nabla f_{i_t}(w^t)\right] = \nabla f(w^t).$$

Answer (Ex. 1) — Since $i_t \sim 1/n$ we have that

$$\mathbb{E}_t\left[\nabla f_{i_t}(w^t)\right] = \sum_{i=1}^n \frac{1}{n}\nabla f_i(w^t) = \nabla f(w^t).$$

Ex. 2 — Let $\mathbb{E}_t[\cdot] \stackrel{\text{def}}{=} \mathbb{E}\left[\cdot \,|\, w^t\right]$ be the expectation conditioned on w^t. Using a step of SGD (4) show that

$$\mathbb{E}_t\left[\|w^{t+1}-w^*\|^2\right] \quad = \quad \|w^t - w^*\|^2 - 2\gamma\left\langle w^t - w^*, \nabla f(w^t)\right\rangle + \gamma^2 \sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w^t)\|^2. \tag{7}$$

Answer (Ex. 2) — By using (4) we have that

$$\|w^{t+1}-w^*\|^2 \quad = \quad \|w^t - w^*\|^2 - 2\gamma\left\langle w^t - w^*, \nabla f_{i_t}(w^t)\right\rangle + \gamma^2\|\nabla f_i(w^t)\|^2. \tag{8}$$

Since i_t is the only random variable conditioned on w^t we have that

$$\mathbb{E}_t\left[\left\langle w^t - w^*, \nabla f_{i_t}(w^t)\right\rangle\right] = \left\langle w^t - w^*, \mathbb{E}_t\left[\nabla f_{i_t}(w^t)\right]\right\rangle = \left\langle w^t - w^*, \nabla f(w^t)\right\rangle.$$

Consequently applying $\mathbb{E}_t[\cdot]$ to (8) gives the result.

Ex. 3 — Now we need to bound the term $\sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w^t)\|^2$ to continue the proof. We break this into the following steps.

Part I

Using that each f_i is L_i–smooth and convex and using Lemma A.1 in the appendix show that

$$\sum_{i=1}^n \frac{1}{2nL_i}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \quad \leq \quad f(w) - f(w^*). \tag{9}$$

Hint: Remember that $\nabla f(w^*) = 0$.
Now let $L_{\max} = \max_{i=1,\ldots,n} L_i$ and conlude that

$$\sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \quad \leq \quad 2L_{\max}(f(w) - f(w^*)). \tag{10}$$

Part II

Using (10) and Definition 5 show that

$$\sum_{i=1}^n \frac{1}{n}\|\nabla f_i(w)\|^2 \quad \leq \quad 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2. \tag{11}$$

Answer (Ex. I) — From Lemma A.1 we have, after re-arranging, that

$$\frac{1}{2L_i}\|\nabla f_i(w) - \nabla f_i(y)\|_2^2 \quad \leq \quad f_i(w) - f_i(y) + \langle \nabla f_i(y), y - w \rangle \,. \tag{12}$$

Plugin $y = w^*$, dividing the above by n and summing over $i = 1, \ldots, n$ gives

$$\sum_{i=1}^{n} \frac{1}{n}\frac{1}{2L_i}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \quad \leq \quad f(w) - f(w^*) + \langle \nabla f(w^*), w^* - w \rangle \,, \tag{13}$$

where we used that $\sum_{i=1}^{n} \frac{1}{n} f_i(w) = f(w)$. The result (9) now follows from that $\nabla f(w^*) = 0$. Finally (10) follows from $L_{\max} \geq L_i$ so that

$$\sum_{i=1}^{n} \frac{1}{2nL_{\max}}\|\nabla f_i(y) - \nabla f_i(w)\|_2^2 \leq \sum_{i=1}^{n} \frac{1}{2nL_i}\|\nabla f_i(w) - \nabla f_i(w^*)\|_2^2 \leq f(w) - f(w^*).$$

Answer (Ex. II) — Using that $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$ we have that

$$\begin{aligned} \sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(w) \pm \nabla f_i(w^*)\|^2 \quad & \leq \quad 2\sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(w) - \nabla f_i(w^*)\|^2 + 2\sum_{i=1}^{n} \frac{1}{n}\|\nabla f_i(w^*)\|^2 \\ & \overset{(10)+(5)}{\leq} \quad 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2. \end{aligned} \tag{14}$$

Ex. 4 — Using (11) together with (7) and the strong quasi-convexity (2) of $f(w)$ show that

$$\mathbb{E}_t\left[\|w^{t+1} - w^*\|^2\right] \quad \leq \quad (1 - \mu\gamma)\|w^t - w^*\|^2 + 2\gamma(2\gamma L_{\max} - 1)(f(w^t) - f(w^*)) + 2\sigma^2\gamma^2. \tag{15}$$

Answer (Ex. 4) — Follow immediatly.

Ex. 5 — Using that $\gamma \in (0, \frac{1}{2L_{\max}}]$ conclude the proof by taking expectation again, and unrolling the recurrence.

Answer (Ex. 5) — Since $\gamma \in (0, \frac{1}{2L_{\max}}]$ we have that $(2\gamma L_{\max} - 1) \leq 0$. Furthermore $f(w^t) - f(w^*) \geq 0$ thus, by taking expectation and using the tower, from (15) we have that

$$\mathbb{E}\left[\|w^{t+1} - w^*\|^2\right] \quad \leq \quad (1 - \mu\gamma)\mathbb{E}\left[\|w^t - w^*\|^2\right] + 2\sigma^2\gamma^2. \tag{16}$$

Let $r_t = \mathbb{E}\left[\|w^{t+1} - w^*\|^2\right]$. The above gives the following recurrence

$$\begin{aligned} r_{t+1} \quad & \leq \quad (1 - \mu\gamma)r_t + 2\sigma^2\gamma \\ & \leq \quad (1 - \mu\gamma)^2 r_{t-1} + (1 - \mu\gamma)2\sigma^2\gamma^2 + 2\sigma^2\gamma^2 \\ & \vdots \\ & \leq \quad (1 - \mu\gamma)^{t+1} r_0 + \sum_{j=0}^{t}(1 - \mu\gamma)^j 2\sigma^2\gamma^2. \end{aligned}$$

Summing up the geometric series we have that

$$\sum_{j=0}^{t}(1-\mu\gamma)^j = \frac{1-(1-\mu\gamma)^{t+1}}{1-(1-\mu\gamma)} \leq \frac{1}{\mu\gamma}.$$

Thus

$$r_{t+1} \leq (1-\mu\gamma)^{t+1} r_0 + \frac{2\sigma^2\gamma^2}{\mu\gamma} = (1-\mu\gamma)^{t+1} r_0 + \frac{2\sigma^2\gamma}{\mu}. \quad \square \tag{17}$$

Ex. 6 — BONUS importance sampling: Let $i_t \sim p_i$ in the SGD update (4), where $p_i > 0$ are probabilities with $\sum_{i=1}^n p_i = 1$. What should the p_i's be so that SGD has the fastest convergence?

3 Decreasing step-sizes

Based on Theorem 1.1 we can introduce a decreasing stepsize.

Theorem 3.1 (Decreasing stepsizes)**.** Let f be μ–strongly quasi-convex and each f_i be L_i–smooth and convex. Let $\mathcal{K} \overset{\text{def}}{=} L_{\max}/\mu$ and

$$\gamma^t = \begin{cases} \frac{1}{2L_{\max}} & \text{for} \quad t \leq 4\lceil \mathcal{K} \rceil \\ \frac{2t+1}{(t+1)^2\mu} & \text{for} \quad t > 4\lceil \mathcal{K} \rceil. \end{cases} \tag{18}$$

If $t \geq 4\lceil \mathcal{K} \rceil$, then SGD iterates given by (4) satisfy:

$$\mathbb{E}\|w^t - w^*\|^2 \leq \frac{\sigma^2}{\mu^2}\frac{8}{t} + \frac{16}{e^2}\frac{\lceil \mathcal{K} \rceil^2}{t^2}\|w^0 - w^*\|^2. \tag{19}$$

Proof. Let $\gamma_t \overset{\text{def}}{=} \frac{2t+1}{(t+1)^2\mu}$ and let t^* be an integer that satisfies $\gamma_{t^*} \leq \frac{1}{2L_{\max}}$. In particular this holds for

$$t^* \geq \lceil 4\mathcal{K} - 1 \rceil.$$

Note that γ_t is decreasing in t and consequently $\gamma_t \leq \frac{1}{2L_{\max}}$ for all $t \geq t^*$. This in turn guarantees that (6) holds for all $t \geq t^*$ with γ_t in place of γ, that is

$$\mathbb{E}\|r^{t+1}\|^2 \leq \frac{t^2}{(t+1)^2}\mathbb{E}\|r^t\|^2 + \frac{2\sigma^2}{\mu^2}\frac{(2t+1)^2}{(t+1)^4}. \tag{20}$$

Multiplying both sides by $(t+1)^2$ we obtain

$$\begin{aligned} (t+1)^2\mathbb{E}\|r^{t+1}\|^2 &\leq t^2\mathbb{E}\|r^t\|^2 + \frac{2\sigma^2}{\mu^2}\left(\frac{2t+1}{t+1}\right)^2 \\ &\leq t^2\mathbb{E}\|r^t\|^2 + \frac{8\sigma^2}{\mu^2}, \end{aligned}$$

where the second inequality holds because $\frac{2t+1}{t+1} < 2$. Rearranging and summing from $j = t^* \ldots t$ we obtain:

$$\sum_{j=t^*}^{t} \left[(j+1)^2 \mathbb{E}\|r^{j+1}\|^2 - j^2 \mathbb{E}\|r^j\|^2 \right] \leq \sum_{j=t^*}^{t} \frac{8\sigma^2}{\mu^2}. \tag{21}$$

Using telescopic cancellation gives

$$(t+1)^2 \mathbb{E}\|r^{t+1}\|^2 \leq (t^*)^2 \mathbb{E}\|r^{t^*}\|^2 + \frac{8\sigma^2 (t-t^*)}{\mu^2}.$$

Dividing the above by $(t+1)^2$ gives

$$\mathbb{E}\|r^{t+1}\|^2 \leq \frac{(t^*)^2}{(t+1)^2} \mathbb{E}\|r^{t^*}\|^2 + \frac{8\sigma^2(t-t^*)}{\mu^2(t+1)^2}. \tag{22}$$

For $t \leq t^*$ we have that (6) holds, which combined with (22), gives

$$\begin{aligned} \mathbb{E}\|r^{t+1}\|^2 &\leq \frac{(t^*)^2}{(t+1)^2} \left(1 - \frac{\mu}{2L_{\max}}\right)^{t^*} \|r^0\|^2 \\ &+ \frac{\sigma^2}{\mu^2 (t+1)^2} \left(8(t-t^*) + \frac{(t^*)^2}{\mathcal{K}} \right). \end{aligned} \tag{23}$$

Choosing t^* that minimizes the second line of the above gives $t^* = 4\lceil \mathcal{K} \rceil$, which when inserted into (23) becomes

$$\begin{aligned} \mathbb{E}\|r^{t+1}\|^2 &\leq \frac{16 \lceil \mathcal{K} \rceil^2}{(t+1)^2} \left(1 - \frac{1}{2\mathcal{K}}\right)^{4\lceil \mathcal{K} \rceil} \|r^0\|^2 \\ &\quad + \frac{\sigma^2}{\mu^2} \frac{8(t - 2\lceil \mathcal{K} \rceil)}{(t+1)^2} \\ &\leq \frac{16 \lceil \mathcal{K} \rceil^2}{e^2 (t+1)^2} \|r^0\|^2 + \frac{\sigma^2}{\mu^2} \frac{8}{t+1}, \end{aligned} \tag{24}$$

where we have used that $\left(1 - \frac{1}{2x}\right)^{4x} \leq e^{-2}$ for all $x \geq 1$. □

A Appendix: Auxiliary smooth and convex lemma

As a consequence of the f_i's being smooth and convex we have that f is also smooth and convex. In particular f is convex since it is a convex combination of the f_i's. This gives us the following useful lemma.

Lemma A.1. If f is both L–smooth

$$f(z) \leq f(w) + \langle \nabla f(w), z - w \rangle + \frac{L}{2} \|z - w\|_2^2 \tag{25}$$

and convex

$$f(z) \geq f(y) + \langle \nabla f(y), z - y \rangle \,, \tag{26}$$

then we have that

$$f(y) - f(w) \quad \leq \quad \langle \nabla f(y), y - w \rangle - \frac{1}{2L} \|\nabla f(y) - \nabla f(w)\|_2^2. \tag{27}$$

Proof. To prove (27), it follows that

$$\begin{aligned} f(y) - f(w) \quad &= \quad f(y) - f(z) + f(z) - f(w) \\ &\overset{(26)+(25)}{\leq} \quad \langle \nabla f(y), y - z \rangle + \langle \nabla f(w), z - w \rangle + \frac{L}{2} \|z - w\|_2^2. \end{aligned}$$

To get the tightest upper bound on the right hand side, we can minimize the right hand side in z, which gives

$$z = w - \frac{1}{L}(\nabla f(w) - \nabla f(y)). \tag{28}$$

Substituting this in gives

$$\begin{aligned} f(y) - f(w) \quad &= \quad \left\langle \nabla f(y), y - w + \frac{1}{L}(\nabla f(w) - \nabla f(y)) \right\rangle \\ &\quad\; -\frac{1}{L} \langle \nabla f(w), \nabla f(w) - \nabla f(y) \rangle + \frac{1}{2L} \|\nabla f(w) - \nabla f(y)\|_2^2 \\ &= \quad \langle \nabla f(y), y - w \rangle - \frac{1}{L} \|\nabla f(w) - \nabla f(y)\|_2^2 + \frac{1}{2L} \|\nabla f(w) - \nabla f(y)\|_2^2 \\ &= \quad \langle \nabla f(y), y - w \rangle - \frac{1}{2L} \|\nabla f(w) - \nabla f(y)\|_2^2. \quad \square \end{aligned}$$