

# Notes for Lecture 11

Scribe: Tingkai Jia

## 1 Convergence of SGD for Strongly Convex Problems

**Problem Definition:**

$$\min_{\mathbf{x} \in \mathbf{R}^d} F(\mathbf{x}) \triangleq \mathbb{E}_{\xi}[f(\mathbf{x}; \xi)]$$

**Assumption 1.1.** Given  $\xi_0, \dots, \xi_{t-1}$ ,  $g(\mathbf{x}_t, \xi_t)$  is an unbiased estimator of  $\nabla F(\mathbf{x}_t)$ , i.e.,

$$\mathbb{E}[g(\mathbf{x}_t, \xi_t) | \xi_0, \dots, \xi_{t-1}] = \nabla F(\mathbf{x}_t)$$

**Assumption 1.2.** For all  $\mathbf{x}$ , we have

$$\mathbb{E}[\|g(\mathbf{x}, \xi)\|_2^2] \leq \sigma^2.$$

**Theorem 1.3** (SGD with fixed stepsizes). Suppose  $F(\mathbf{x})$  is  $L$ -smooth and  $\mu$ -strongly convex, with Assumption 1.1 and 1.2, if  $\eta_t = \eta \leq \frac{1}{2L}$ , then SGD achieves

$$\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] \leq (1 - 2\mu\eta)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \frac{\eta\sigma^2}{2\mu}$$

*Proof.* Using the SGD update rule, we have

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_t - \eta g(\mathbf{x}_t; \xi_t) - \mathbf{x}^*\|_2^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta(\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t) + \eta^2 \|g(\mathbf{x}_t; \xi_t)\|_2^2. \end{aligned} \quad (1)$$

Since  $\mathbf{x}_t$  is indep. of  $\xi_t$ , we obtain

$$\begin{aligned} \mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t)] &= \mathbb{E}[\mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t) | \xi_0, \dots, \xi_{t-1}]] \\ &= \mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top \mathbb{E}[g(\mathbf{x}_t; \xi_t) | \xi_0, \dots, \xi_{t-1}]] \\ &= \mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top \nabla F(\mathbf{x}_t)]. \end{aligned} \quad (2)$$

Furthermore, strong convexity gives

$$\begin{aligned} \langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle &= \langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^* \rangle \geq \mu \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \\ &\Rightarrow \mathbb{E}[\langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle] \geq \mu \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] \end{aligned} \quad (3)$$

Combine (1), (2), (3) and Assumption 1.2 to obtain

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] &= \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] - 2\eta \mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t)] + \eta^2 \mathbb{E}[\|g(\mathbf{x}_t; \xi_t)\|_2^2] \\ &\leq \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] - 2\mu\eta \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \eta^2 \sigma^2 \\ &= (1 - 2\mu\eta) \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \eta^2 \sigma^2 \\ \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] - \frac{\eta\sigma^2}{2\mu} &\leq (1 - 2\mu\eta) \left( \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] - \frac{\eta\sigma^2}{2\mu} \right), \end{aligned}$$

thus we obtain

$$\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] - \frac{\eta\sigma^2}{2\mu} \leq (1 - 2\mu\eta)^t \left( \mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2] - \frac{\eta\sigma^2}{2\mu} \right)$$

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] &\leq (1 - 2\mu\eta)^t \mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2] + \frac{\eta\sigma^2}{2\mu} (1 - (1 - 2\mu\eta)^t) \\
&\leq (1 - 2\mu\eta)^t \mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2] + \frac{\eta\sigma^2}{2\mu} (1 - (1 - \frac{\mu}{L})^t) \\
&\leq (1 - 2\mu\eta)^t \mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2] + \frac{\eta\sigma^2}{2\mu}.
\end{aligned}$$

Then we finish the proof.  $\square$

**Theorem 1.4** (SGD with diminishing stepsizes). *Suppose  $F(\mathbf{x})$  is  $L$ -smooth and  $\mu$ -strongly convex, with Assumption 1.1 and 1.2, if  $\eta_t = \frac{\theta}{t+1}$  for some  $\theta > \frac{1}{2\mu}$ , then SGD achieves*

$$\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] \leq \frac{\alpha_\theta}{t+1}$$

where  $\alpha_\theta = \max \left\{ \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2, \frac{2\theta^2\sigma^2}{2\mu\theta-1} \right\}$ .

*Proof.* Like fixed stepsizes situation, we first use the SGD update rule to have

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 &= \|\mathbf{x}_t - \eta_t g(\mathbf{x}_t; \xi_t) - \mathbf{x}^*\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta_t (\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t) + \eta_t^2 \|g(\mathbf{x}_t; \xi_t)\|_2^2.
\end{aligned} \tag{4}$$

We also have

$$\mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t)] = \mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top \nabla F(\mathbf{x}_t)]. \tag{5}$$

Furthermore, strong convexity gives

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle &= \langle \nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^* \rangle \geq \mu \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \\
&\Rightarrow \mathbb{E}[\langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle] \geq \mu \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2]
\end{aligned} \tag{6}$$

Combining (4), (5), (6) and Assumption 1.2, we obtain

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] &= \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] - 2\eta_t \mathbb{E}[(\mathbf{x}_t - \mathbf{x}^*)^\top g(\mathbf{x}_t; \xi_t)] + \eta_t^2 \mathbb{E}[\|g(\mathbf{x}_t; \xi_t)\|_2^2] \\
&\leq \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] - 2\mu\eta_t \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \eta_t^2 \sigma^2 \\
&= (1 - 2\mu\eta_t) \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \eta_t^2 \sigma^2
\end{aligned}$$

Then we use induction to complete the following proof.

- When  $k = 0$ , it is surely true that

$$\mathbb{E}[\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2] \leq \alpha_\theta = \max \left\{ \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2, \frac{2\theta^2\sigma^2}{2\mu\theta-1} \right\}.$$

- When  $k = t$ , we assume our theorem is true.
- When  $k = t + 1$ , it follows that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] &\leq (1 - 2\mu\eta_t) \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \eta_t^2 \sigma^2 \\
&\leq \left(1 - \frac{2\mu\theta}{t+1}\right) \frac{\alpha_\theta}{t+1} + \frac{\theta^2\sigma^2}{(t+1)^2} \\
&\leq \left(1 - \frac{2\mu\theta}{t+1}\right) \frac{\alpha_\theta}{t+1} + \frac{2\mu\theta-1}{2(t+1)^2} \alpha_\theta
\end{aligned}$$

$$\begin{aligned}
&\leq \left( \frac{1}{t+1} - \frac{2\mu\theta + 1}{2(t+1)^2} \right) \alpha_\theta \\
&\leq \left( \frac{1}{t+1} - \frac{1}{(t+1)^2} \right) \alpha_\theta \\
&= \frac{t}{(t+1)^2} \alpha_\theta = \frac{t(t+2)}{(t+1)^2} \cdot \frac{\alpha_\theta}{t+2} \\
&\leq \frac{\alpha_\theta}{t+2}.
\end{aligned}$$

Thus we finish the proof.  $\square$

## 2 Convergence of SGD for Convex Problems

**Theorem 2.1.** Suppose  $F(\mathbf{x})$  is  $L$ -smooth and convex, with Assumption 1.1 and 1.2, then SGD achieves

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sum_{k=0}^t \sigma^2 \eta_k^2}{2 \sum_{k=0}^t \eta_k},$$

where  $\tilde{\mathbf{x}}_t = \sum_{k=0}^t \frac{\eta_k}{\sum_{i=0}^t \eta_i} \mathbf{x}_k$ . If we choose  $\eta_t = \mathcal{O}(1/\sqrt{t})$ , then we have

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \mathcal{O}\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sigma^2 \log t}{\sqrt{t}}\right).$$

*Proof.* By convexity of  $F$ , we have

$$\begin{aligned}
F(\mathbf{x}^*) &\geq F(\mathbf{x}_t) + (\mathbf{x} - \mathbf{x}_t)^\top \nabla F(\mathbf{x}_t) \\
\Rightarrow \mathbb{E}[(\mathbf{x} - \mathbf{x}_t)^\top \nabla F(\mathbf{x}_t)] &\geq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)].
\end{aligned}$$

This together with (1) and (2) implies

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] &\leq \mathbb{E}[(\mathbf{x} - \mathbf{x}_t)^\top \nabla F(\mathbf{x}_t)] \\
&= \mathbb{E}[(\mathbf{x} - \mathbf{x}_t)^\top g(\mathbf{x}_t; \xi_t)] \\
&\leq \frac{1}{2\eta_t} (\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2) + \frac{1}{2} \eta_t \sigma^2 \\
2\eta_t \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] &\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \eta_t^2 \sigma^2.
\end{aligned}$$

Sum recursively to obtain

$$\begin{aligned}
\sum_{k=0}^t 2\eta_k \mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)] &\leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 + \sigma^2 \sum_{k=0}^t \eta_k^2 \\
&\leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sigma^2 \sum_{k=0}^t \eta_k^2.
\end{aligned}$$

Setting  $\nu_t = \frac{\eta_t}{\sum_{k=0}^t \eta_k}$ , yields

$$\sum_{k=0}^t 2\nu_k \mathbb{E}[F(\mathbf{x}_k) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sigma^2 \sum_{k=0}^t \eta_k^2}{\sum_{k=0}^t \eta_k}.$$

With Jensen's inequality, we finally obtain

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sum_{k=0}^t \sigma^2 \eta_k^2}{2 \sum_{k=0}^t \eta_k}.$$

Then if setting  $\eta_t = \mathcal{O}(1/\sqrt{t})$ , with the fact that  $2(\sqrt{t+1} - 1) \leq \sum_{k=0}^t \frac{1}{\sqrt{k}} \leq 2\sqrt{t}$  and  $\sum_{k=0}^t \frac{1}{k} \leq \log t + 1$ , we have

$$\mathbb{E}[F(\tilde{\mathbf{x}}_t) - F(\mathbf{x}^*)] \leq \mathcal{O}\left(\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 + \sigma^2 \log t}{\sqrt{t}}\right).$$

□