Notes for Lecture 4

Scribe: Tingkai Jia

1 Quadratic Minimization

To learn about the convergence rate of GD, we begin with quadratic objective functions

$$\textbf{minimize}_{\mathbf{x}} \quad f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{b}^\top \mathbf{x}$$

for some $n \times n$ matrix $\mathbf{Q} \succ 0$, where $\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{b}$. Now consider GD process shown as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t).$$

Convergence rate: if $\eta_t \equiv \eta = \frac{2}{\lambda_1(\mathbf{Q})+\lambda_n(\mathbf{Q})}$, then

$$\|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \left(\frac{\lambda_1(\mathbf{Q}) - \lambda_n(\mathbf{Q})}{\lambda_1(\mathbf{Q}) + \lambda_n(\mathbf{Q})}\right)^t \|\mathbf{x}_0 - \mathbf{x}_*\|_2,$$

where $\lambda_1(\mathbf{Q})$ (resp. $\lambda_n(\mathbf{Q})$) is the largest (resp. smallest) eigenvalue of $\mathbf{Q}$.

Proof. First we can easily get the gradient of objective function $\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x} - \mathbf{b}$, and the unique optimal solution is $\mathbf{x}_* = \mathbf{Q}^{-1}\mathbf{b}$. Then according to the GD update rule, by subtracting $\mathbf{x}_*$ on both sides, we get

$$\mathbf{x}_{t+1} - \mathbf{x}_* = \mathbf{x}_t - \mathbf{x}_* - \eta_t \nabla f(\mathbf{x}_t) = \mathbf{x}_t - \mathbf{x}_* - \eta_t(\mathbf{Q}\mathbf{x}_t - \mathbf{b})$$
$$= \mathbf{x}_t - \mathbf{x}_* - \eta_t \mathbf{Q}(\mathbf{x}_t - \mathbf{x}_*) = (\mathbf{I} - \eta_t \mathbf{Q})(\mathbf{x}_t - \mathbf{x}_*),$$

taking l_2 norm on both sides of the last equation, we get

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 = \|(\mathbf{I} - \eta_t \mathbf{Q})(\mathbf{x}_t - \mathbf{x}_*)\|_2 \leq \|\mathbf{I} - \eta_t \mathbf{Q}\|_2 \|\mathbf{x}_t - \mathbf{x}_*\|_2.$$

Now we want to get the optimal convergence rate by setting η_t, which means minimizing the max eigenvalue of matrix $\mathbf{I} - \eta_t \mathbf{Q}$ (For a symmetric positive definite matrix, the singular values are equal to the eigenvalues). Then we observe that

$$\|\mathbf{I} - \eta_t \mathbf{Q}\|_2 = \max\{|1 - \eta_t \lambda_1(\mathbf{Q})|, |1 - \eta_t \lambda_n(\mathbf{Q})|\} = \frac{\lambda_1(\mathbf{Q}) - \lambda_n(\mathbf{Q})}{\lambda_1(\mathbf{Q}) + \lambda_n(\mathbf{Q})},$$

the last equation means we set $\eta_t \equiv \eta = \frac{2}{\lambda_1(\mathbf{Q})+\lambda_n(\mathbf{Q})}$ to make these two terms equal. $\square$

2 Equivalent characterizations of L-smoothness

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function. Then the following properties are equivalent characterizations of L-smoothness of f:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{A}$$

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq L\|\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{B}$$

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{C}$$

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{D}$$

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{E}$$

Proof $A \Rightarrow B$: By Cauchy-Schwartz, we have

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \|\mathbf{x} - \mathbf{y}\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2^2.$$

Proof $B \Rightarrow C$: Define the function $G : [0, 1] \to \mathbb{R}$ as

$$G(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), t(\mathbf{y} - \mathbf{x}) \rangle,$$

so that $G(0) = 0$ and $G(1) = f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$. By the fundamental theorem of calculus, we have

$$\begin{aligned} G(1) - G(0) &= \int_0^1 G'(t)\, \mathrm{d}t = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle\, \mathrm{d}t. \\ &= \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), t(\mathbf{y} - \mathbf{x}) \rangle \, \frac{1}{t}\, \mathrm{d}t. \\ &\leq L\|\mathbf{y} - \mathbf{x}\|_2^2 \int_0^1 t\, \mathrm{d}t = \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2. \end{aligned}$$

Proof $C \Rightarrow D$: We begin with a useful auxiliary lemma:

Lemma 1. *Consider a differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ satisfying condition (C) and with its global minimum achieved at some $\mathbf{v}^*$. Then*

$$g(\mathbf{v}) - g(\mathbf{v}^*) \geq \frac{1}{2L}\|\nabla g(\mathbf{v})\|_2^2 \quad \textit{for all } \mathbf{v} \in \mathbb{R}^d.$$

Proof. We have

$$\begin{aligned} g(\mathbf{v}^*) = \inf_{\mathbf{u} \in \mathbb{R}^d} g(\mathbf{u}) &\leq \inf_{\mathbf{u} \in \mathbb{R}^d} \left\{ g(\mathbf{v}) + \langle \nabla g(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle + \frac{L}{2}\|\mathbf{v} - \mathbf{u}\|_2^2 \right\} \\ &= g(\mathbf{v}) - \frac{1}{2L}\|\nabla g(\mathbf{v})\|_2^2, \end{aligned}$$

where the last step follows by showing that the minimum of the quadratic program over $\mathbf{u}$ is achieved at $\mathbf{u}^* = \mathbf{v} - \frac{1}{L}\nabla g(\mathbf{v})$, and then performing some algebra.

Note: This lemma and its proof are of independent interest, as they show how gradient descent with step size $1/L$ can be thought of as minimizing a linear approximation along with a quadratic regularization term scaled by $L/2$.

Let us now show that $C \Rightarrow D$. For a fixed $\mathbf{x} \in \mathbb{R}^d$, define the function

$$g_x(\mathbf{z}) = f(\mathbf{z}) - \langle \nabla f(\mathbf{x}), \mathbf{z} \rangle.$$

Note that g_x is convex, differentiable and minimized when $\mathbf{z} = \mathbf{x}$, and it satisfies our smoothness condition. Hence, the preceding lemma with $\mathbf{v}^* = \mathbf{x}$ and $\mathbf{v} = \mathbf{y}$ implies that

$$g_x(\mathbf{y}) - g_x(\mathbf{x}) \geq \frac{1}{2L}\|\nabla g_x(\mathbf{y})\|_2^2 = \frac{1}{2L}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2^2.$$

A little bit of calculation shows that

$$g_x(\mathbf{y}) - g_x(\mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle,$$

which completes the proof. □

Proof $D \Rightarrow E$**:** We have

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,$$

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{1}{2L} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \quad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Adding these inequalities yields E.
Proof $E \Rightarrow A$**:** By Cauchy-Schwartz, we have

$$\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \|\mathbf{x} - \mathbf{y}\|_2.$$

3 Equivalent characterizations of μ-strong convexity

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function. Then the following properties are equivalent characterizations of μ-strong convexity of f:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \geq \mu \|\mathbf{x} - \mathbf{y}\|_2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{A}$$

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{B}$$

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{C}$$

$$f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{1}{2\mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{D}$$

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \leq \frac{1}{\mu} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2 \qquad \text{for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{E}$$

Note that all of these conditions can be obtained from the L-smoothness conditions by:

- flipping all the inequality signs, and
- replacing L by μ everywhere

4 Strongly Convex and Smooth Functions Minimization

We can generalize quadratic minimization to a broader class of problems

$$\textbf{minimize}_{\mathbf{x}} \quad f(\mathbf{x})$$

where $f(\cdot)$ is L-strongly convex and μsmooth, which means $0 \preceq \mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$ for $\forall \mathbf{x}$. Now consider GD process shown as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t).$$

Convergence rate: if $\eta_t \equiv \eta = \frac{2}{\mu + L}$, then

$$\|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^t \|\mathbf{x}_0 - \mathbf{x}_*\|_2,$$

where $\kappa = L/\mu$ is condition number and $\mathbf{x}_*$ is optimal solution.

Proof. It is seen from the fundamental theorem of calculus that

$$\nabla f(\mathbf{x}_t) = \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_*) = \left(\int_0^1 \nabla^2 f(\mathbf{x}_\tau) \, \mathrm{d}\tau \right) (\mathbf{x}_t - \mathbf{x}_*)$$

where $\mathbf{x}_\tau = \mathbf{x}_t + \tau(\mathbf{x}_* - \mathbf{x}_t)$. Then according to the GD update rule, by subtracting $\mathbf{x}_*$ on both sides, we get

$$\mathbf{x}_{t+1} - \mathbf{x}_* = \mathbf{x}_t - \mathbf{x}_* - \eta_t \nabla f(\mathbf{x}_t) = \mathbf{x}_t - \mathbf{x}_* - \eta_t \left(\int_0^1 \nabla^2 f(\mathbf{x}_\tau)\, \mathrm{d}\tau \right) (\mathbf{x}_t - \mathbf{x}_*)$$

$$= \left(\mathbf{I} - \eta_t \int_0^1 \nabla^2 f(\mathbf{x}_\tau)\, \mathrm{d}\tau \right) (\mathbf{x}_t - \mathbf{x}_*),$$

taking l_2 norm on both sides of the last equation, we get

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2 = \| \left(\mathbf{I} - \eta_t \int_0^1 \nabla^2 f(\mathbf{x}_\tau)\, \mathrm{d}\tau \right) (\mathbf{x}_t - \mathbf{x}_*)\|_2.$$

$$\leq \sup_{0 \leq \tau \leq 1} \|\mathbf{I} - \eta_t \nabla^2 f(\mathbf{x}_\tau)\|_2 \|\mathbf{x}_t - \mathbf{x}_*\|_2 \leq \frac{L - \mu}{L + \mu} \|\mathbf{x}_t - \mathbf{x}_*\|_2.$$

The last inequality refers to the quadratic minimization, but it is impossible to get the maximum and minimum eigenvalue of matrix $\nabla^2 f(\mathbf{x}_\tau)$ on unknown $\mathbf{x}_\tau$, so we have replaced them with L and μ respectively, which also means we set $\eta_t \equiv \eta = \frac{2}{\mu + L}$. □