Notes for Lecture 5

Scribe: Tingkai Jia

1 Polyak-Lojasiewicz Condition

Without a strong convexity condition, some functions can still achieve linear convergence if they satisfy Polyak-Lojasiewicz Condition shown as

$$f(\mathbf{x}) - f(\mathbf{x}_*) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x})\|^2.$$

With PL condition, we obtain the following theorem.

Theorem 1. *Suppose f satisfies PL condition and is L-smooth. If $\eta_t \equiv \eta = \frac{1}{L}$, then*

$$f(\mathbf{x}_t) - f(\mathbf{x}_*) \leq \left(1 - \frac{\mu}{L}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}_*))$$

Proof. Since f is L-smooth, we firstly have

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2}\|\nabla \mathbf{x}_{t+1} - \mathbf{x}_t\|^2,$$

substitute $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$, we then get

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2} \cdot \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2$$

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2.$$

With the above fact, we can obtain

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_*) \leq f(\mathbf{x}_t) - f(\mathbf{x}_*) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$$

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_*) \leq f(\mathbf{x}_t) - f(\mathbf{x}_*) - \frac{\mu}{L}(f(\mathbf{x}_t) - f(\mathbf{x}_*)) = \left(1 - \frac{\mu}{L}\right)(f(\mathbf{x}_t) - f(\mathbf{x}_*)),$$

where the last inequality comes from PL condition, and apply it recursively, we can complete the proof. □

Here is an example of achieving linear convergence by applying the PL condition.

Example 1 (over-parameterized linear regression)**.** *Suppose $m < n$, with m data samples $\{\mathbf{a}_i \in \mathbb{R}^n, y_i \in \mathbb{R}\}_{1 \leq i \leq m}$, a linear model try to best fits the data*

$$\textbf{\textit{minimize}}_{\mathbf{x} \in \mathbb{R}^n} \quad f(\mathbf{x}) = \frac{1}{2}\sum_{i=1}^{m}\left(\mathbf{a}_i^\top \mathbf{x} - y_i\right)^2$$

It is obvious that the hessian matrix of f is rank-deficient when $m < n$, so f is not strongly convex and we have infinitely many of local minima that make f attain its minimum value of 0. But f satisfies PL condition and can still get linear convergence as the following theorem show.

Theorem 2. *Suppose that* $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m]^\top \in \mathbb{R}^{m\times n}$ *has rank* m*, and* $\eta_t \equiv \eta = \frac{1}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}$*. Then GD obeys*

$$f(\mathbf{x}_t) - f(\mathbf{x}_*) \le \left(1 - \frac{\lambda_{\min}(\mathbf{A}\mathbf{A}^\top)}{\lambda_{\max}(\mathbf{A}\mathbf{A}^\top)}\right)^t (f(\mathbf{x}_0) - f(\mathbf{x}_*))$$

Proof. First we can see $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})$, then we have

$$\|\nabla f(\mathbf{x})\|_2^2 = (\mathbf{A}\mathbf{x} - \mathbf{y})^\top \mathbf{A}\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{y})$$

$$\|\nabla f(\mathbf{x})\|_2^2 \ge \lambda_{\min}(\mathbf{A}\mathbf{A}^\top)\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 = 2\lambda_{\min} f(\mathbf{x}),$$

where the last inequality means f satisfies PL condition if we choose $\mu = \lambda_{\min}(\mathbf{A}\mathbf{A}^\top)$. Then apply 1, we immediately finish the proof. □

2 Convex and Smooth Functions Minimization

Lemma 1. *Suppose* f *is convex and* L*-smooth. If* $\eta_t \equiv \eta = \frac{1}{L}$*, then*

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \le \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|_2^2,$$

where $\mathbf{x}_*$ *is any minimizer of* $f(\cdot)$*.*

Proof. It follows that

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 &= \|\mathbf{x}_t - \mathbf{x}_* - \eta(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_*))\|_2^2 \\ &= \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - 2\eta\langle \mathbf{x}_t - \mathbf{x}_*, \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_*)\rangle + \eta^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_*)\|_2^2 \\ &\le \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - \frac{2\eta}{L}\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_*)\|_2^2 + \eta^2\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_*)\|_2^2 \\ &= \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|_2^2.\end{aligned}$$

□

Theorem 3. *Suppose* f *is convex and* L*-smooth. If* $\eta_t \equiv \eta = \frac{1}{L}$*, then GD obeys*

$$f(\mathbf{x}_t) - f(\mathbf{x}_*) \le \frac{2L\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{t},$$

where $\mathbf{x}_*$ *is any minimizer of* $f(\cdot)$*.*

Proof. First we use convexity and Cauchy-Schwartz to get

$$f(\mathbf{x}_*) - f(\mathbf{x}_t) \ge \langle \nabla f(\mathbf{x}_t), \mathbf{x}_* - \mathbf{x}_t\rangle \ge -\|\nabla f(\mathbf{x}_t)\|_2\|\mathbf{x}_t - \mathbf{x}_*\|_2$$

$$\|\nabla f(\mathbf{x}_t)\|_2 \ge \frac{f(\mathbf{x}_t) - f(\mathbf{x}_*)}{\|\mathbf{x}_t - \mathbf{x}_*\|_2} \ge \frac{f(\mathbf{x}_t) - f(\mathbf{x}_*)}{\|\mathbf{x}_0 - \mathbf{x}_*\|_2}.$$

Setting $\triangle_t := f(\mathbf{x}_t) - f(\mathbf{x}_*)$, consider the fact also mentioned in 1

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \le -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2,$$

combining the above two bounds yield

$$\triangle_{t+1} - \triangle_t \le -\frac{\triangle_t^2}{2L\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2} =: -\frac{1}{w_0}\triangle_t^2.$$

Dividing both sides by $\triangle_t \triangle_{t+1}$ and rearranging the terms give

$$\frac{1}{\triangle_{t+1}} \geq \frac{1}{\triangle_t} + \frac{\triangle_t}{w_0 \triangle_{t+1}} \geq \frac{1}{\triangle_t} + \frac{1}{w_0},$$

consider the last inequality recursively, we easily get

$$\frac{1}{\triangle_t} \geq \frac{1}{\triangle_0} + \frac{t}{w_0} \geq \frac{t}{w_0}$$

$$\triangle_t \leq \frac{w_0}{t} = \frac{2L\|\mathbf{x}_0 - \mathbf{x}_*\|_2^2}{t},$$

thus we complete the proof. □