Linear Regression (cont'd)

Tuesday, January 18, 2022 9:23 AM

Recall the standard linear model:

$$y = X\theta^* + w, \quad z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$$

and the least-squares estimator (assuming $n >> d$):

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta, \{z_i\}_{i=1}^n) \triangleq \frac{1}{2}\|y - X\theta\|_2^2$$

<u>Last time</u>: $\|\hat{\theta} - \theta^*\|_2 \le \frac{2}{\lambda_{min}(X^TX)} \cdot \|X\| \cdot \|w\|_2$ (estimation error)

Assuming $X_{ij} \sim \mathcal{N}(0,1)$ iid, for any $\eta \geq 0$, whp

$$\sqrt{n} - c(\sqrt{d} + \eta) \le \sigma_i(X) \le \sqrt{n} + c(\sqrt{d} + \eta)$$

for some constant $c > 0$. Hence, whp

$$\frac{\|X\|}{\lambda_{min}(X^TX)} = \frac{\sigma_1(X)}{\sigma_d^2(X)} \le \frac{\sqrt{n} + c(\sqrt{d} + \eta)}{(\sqrt{n} - c(\sqrt{d} + \eta))^2}$$

Further, assuming $w_i \sim \mathcal{N}(0, \sigma^2)$ iid, then we want to bound $\|w\|_2$. Observe that

$$\|w\|_2^2 = \sum_{i=1}^n w_i^2 = \sigma^2 \sum_{i=1}^n g_i^2 \quad \text{where } g_i \sim \mathcal{N}(0,1)$$

$$\Rightarrow \mathbb{E}[\|w\|_2^2] = \sigma^2 n \Rightarrow \mathbb{E}[\|w\|_2] \le \sigma\sqrt{n}$$

So we expect that $\|w\|_2 \sim c' \cdot \sigma \cdot \sqrt{n}$ for some constant $c' > 0$ whp. (Exercise)

<u>Q</u>: How to find $\hat{\theta}$ algorithmically?

Recall that we assume X has full column rank in order to get the estimation error bound. This implies $\lambda_{min}(X^TX) > 0$; $X^TX \succ 0$. Note that

$$\nabla^2 \mathcal{L}(\theta; \{z_i\}_{i=1}^n) = X^TX$$

Hence, $\mathcal{L}(\cdot\,; \{z_i\}_{i=1}^n)$ is strongly convex.

<u>Definition / Claim</u>: We say that $f: \mathbb{R}^d \to \mathbb{R}$ is strongly convex with modulus $c > 0$ if any of the following equivalent conditions hold:

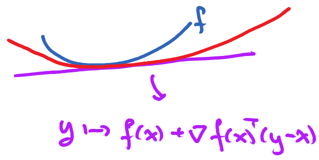
(1) For any $x, y \in \mathbb{R}^d$ and $\alpha \in [0,1]$,

$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y) - \frac{1}{2} c\, \alpha (1-\alpha) \|x-y\|_2^2$$

(2) The function $x \mapsto f(x) - \frac{1}{2} c \|x\|_2^2$ is convex.

(3) (In the presence of differentiability) For $x, y \in \mathbb{R}^d$,

$$f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} c \|y-x\|_2^2$$



(4) (In the presence of twice differentiability)

For any $x \in \mathbb{R}^d$,

$$v^T \nabla^2 f(x) v \ge c \cdot \|v\|_2^2 \quad \forall v \in \mathbb{R}^d.$$

<u>Definition</u>: Let $f: \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. We say that f has L-Lipschitz continuous gradient for some $L > 0$ if for all $x, y \in \mathbb{R}^d$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \le L \cdot \|x-y\|_2.$$

<u>Proposition 1</u>: Let $f: \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, c-strongly convex, and have L-Lipschitz continuous gradient. Then, for all $x, y \in \mathbb{R}^d$,

$$(\nabla f(x) - \nabla f(y))^T (x-y) \ge \frac{cL}{c+L} \|x-y\|_2^2 + \frac{1}{c+L} \|\nabla f(x) - \nabla f(y)\|_2^2$$

(Exercise)

With the above results, let us analyze the gradient descent method for solving

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} f(\theta)$$

where $f: \mathbb{R}^d \to \mathbb{R}$ satisfies the assumptions in Proposition 1.

The update formula is given by

$$\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla f(\theta^k), \quad \alpha_k > 0 \text{ step size.}$$

<u>Theorem:</u> Let $f: \mathbb{R}^d \to \mathbb{R}$ be as above. Suppose that

$\alpha_k = \alpha \in \left(0, \frac{2}{C+L}\right]$. Then, the sequence $\{\theta^k\}_{k \geq 0}$ satisfies

↳ constant step size

$$\| \theta^k - \hat{\theta} \|_2^2 \leq \left(1 - \frac{2\alpha c L}{C+L}\right)^k \| \theta^0 - \hat{\theta} \|_2^2$$

(linear convergence)

<u>Proof:</u> We compute

$$\| \theta^{k+1} - \hat{\theta} \|_2^2 = \| \theta^k - \alpha \nabla f(\theta^k) - \hat{\theta} \|_2^2$$

$$= \underline{\| \theta^k - \hat{\theta} \|_2^2} - 2\alpha \nabla f(\theta^k)^T (\theta^k - \hat{\theta}) + \underline{\alpha^2 \| \nabla f(\theta^k) \|_2^2}.$$

Observe that $\nabla f(\hat{\theta}) = 0$. Thus, by Proposition 1,

$$(\nabla f(\theta^k) - \underbrace{\nabla f(\hat{\theta})}_{=0})^T (\theta^k - \hat{\theta}) \geq \underline{\frac{cL}{C+L} \| \theta^k - \hat{\theta} \|_2} + \underline{\frac{1}{C+L} \| \nabla f(\theta^k) \|_2^2}$$

Hence,

$$\| \theta^{k+1} - \hat{\theta} \|_2^2 \leq \left(1 - \frac{2\alpha c L}{C+L}\right) \| \theta^k - \hat{\theta} \|_2^2 + \alpha \left(\alpha - \frac{2}{C+L}\right) \| \nabla f(\theta^k) \|_2^2$$

$$\leq \left(1 - \frac{2\alpha c L}{C+L}\right) \| \theta^k - \hat{\theta} \|_2^2.$$