Bound on Statistical Error for Linear Regression

Setup: Standard Linear Model

- $Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ covariate-response pair
- $y = X\theta^* + w$ — (†)

$\theta^* \in \mathbb{R}^d$: ground-truth ; $w \in \mathbb{R}^n$: noise ;

$X \in \mathbb{R}^{n \times d}$: design matrix with X_i^T as i^{th} row

To appreciate the geometry at play when studying the statistical error, consider the classic ($n >> d$) setting and the least squares estimator:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \ \frac{1}{2} \|y - X\theta\|_2^2 \qquad \text{— (*)}$$

Since $\hat{\theta}$ is optimal for (*) and θ^* is feasible, we have

$$\frac{1}{2}\|y - X\hat{\theta}\|_2^2 \leq \frac{1}{2}\|y - X\theta^*\|_2^2,$$

which implies that

$$\|X\hat{\Delta}\|_2^2 \leq 2\hat{\Delta}^T X^T w \qquad \text{with } \hat{\Delta} = \hat{\theta} - \theta^* \qquad \text{— (**)}$$

(check using the generative model (†))

Assuming that X has full column rank (so that X^TX is invertible), we have, by the Courant-Fischer theorem

$$\|X\hat{\Delta}\|_2^2 \geq \lambda_{\min}(X^TX) \cdot \|\hat{\Delta}\|_2^2$$

Note that $\lambda_{\min}(X^TX) > 0$ here. It follows from (**) that

$$\lambda_{\min}(X^TX) \cdot \|\hat{\Delta}\|_2^2 \leq \|X\hat{\Delta}\|_2^2 \leq 2\hat{\Delta}^T X^T w \leq 2\|\hat{\Delta}\|_2 \cdot \|X^T w\|_2$$

and hence

$$\|\hat{\Delta}\|_2 \leq \frac{2}{\lambda_{\min}(X^TX)} \|X^T w\|_2 \qquad \text{(statistical error)}$$

Note that the bound just obtained is deterministic. One can obtain high probability bounds for various statistical models on X or w. *For details, see Handout 1.*

The key in the above derivation is the invertibility of X^TX, which guarantees that $\lambda_{min}(X^TX) > 0$. Incidentally, X^TX is also the Hessian of the loss function

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

It follows that $\mathcal{L}$ is <u>strongly convex</u>. Such a property plays a crucial role in establishing sharp bounds on the optimization error for various iterative methods as well. Let us elaborate.

<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 holds:

(1) $\forall x, y \in \mathbb{R}^d$ 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) $\forall x, y \in \mathbb{R}^d$,

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

(4) (in the presence of second-order differentiability) $\forall x \in \mathbb{R}^d$,

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

Another notion that is of interest is the following:

Definition: 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 $\forall x, y \in \mathbb{R}^d$,

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

A consequence of the above notions is the following:

Proposition 1 (Exercise) Let $f: \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable, c-strongly convex, and have L-Lipschitz continuous gradient. Then, $\forall x, y \in \mathbb{R}^d$,

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

Using the above proposition, let us study the convergence behavior of the gradient method for solving

$$\min_{\theta \in \mathbb{R}^d} f(\theta), \qquad \text{— (P)}$$

where $f: \mathbb{R}^d \to \mathbb{R}$ is as in Proposition 1. The update formula of the gradient method is given by

$$\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla f(\theta^k), \qquad \text{— (***)}$$

where $\alpha_k > 0$ is the step size in the k^{th} iteration.

Theorem 1: Let $f: \mathbb{R}^d \to \mathbb{R}$ be as in Proposition 1. Suppose that $\alpha_k \equiv \alpha \in \left(0, \frac{2}{c+L}\right]$ in (***). Then, the sequence $\{\theta^k\}$ generated by (***) satisfies

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

where $\hat{\theta}$ is the optimal solution to (P),

Proof: We compute

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

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

Now, observe that by Proposition 1 and the fact that $\nabla f(\hat{\theta}) = 0$,

$$\nabla f(\theta^k)^T(\theta^k - \hat{\theta}) = (\nabla f(\theta^k) - \nabla f(\hat{\theta}))^T(\theta^k - \hat{\theta})$$

$$\geq \frac{cL}{c+L}\|\theta^k - \hat{\theta}\|_2^2 + \frac{1}{c+L}\|\nabla f(\theta^k)\|_2^2.$$

Hence, by our choice of α, we have

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

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

□

Returning to the least squares loss function (L), it is clear that $\mathcal{L}$ satisfies the assumptions of Proposition 1 (check). Thus, when we apply the gradient method to solve (*), we have the optimization error bound given in Theorem 1.