Setup: Standard Linear Model

\[ Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \quad \text{covariate-response pair} \]

\[ Y = X\theta^* + \epsilon \quad \text{ (t)} \]

\[ \theta^* \in \mathbb{R}^d : \text{ground-truth} \quad \epsilon \in \mathbb{R}^n : \text{noise} \]

\[ X \in \mathbb{R}^{n \times d} : \text{design matrix with } x_i \text{ as } i^{th} \text{ row} \]

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

\[ \hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \| y - X\theta \|_2^2 \quad \text{(x)} \]

Since \( \hat{\theta} \) is optimal for (x) and \( \theta^* \) 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

\[ \| X\hat{\Delta} \|_2^2 \leq 2 \hat{\Delta}^T X^T \epsilon \quad \text{with} \quad \hat{\Delta} = \hat{\theta} - \theta^* \quad \text{(**)} \]

(check using the generative model (t)).

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

\[ \| X\hat{\Delta} \|_2^2 \geq \lambda_{\min}(X^T X) \cdot \| \hat{\Delta} \|_2^2 \]

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

\[ \lambda_{\min}(X^T X) \cdot \| \hat{\Delta} \|_2^2 \leq \| X\hat{\Delta} \|_2^2 \leq 2 \hat{\Delta}^T X^T \epsilon = 2 \| \hat{\Delta} \|_2 \cdot \| X^T \epsilon \|_2 \]

and hence

\[ \| \hat{\Delta} \|_2 \leq \frac{2}{\lambda_{\min}(X^T X)} \| X^T \epsilon \|_2 \quad \text{(statistical error)} \]
Note that the bound just obtained is deterministic. One can obtain high probability bounds for various statistical models on $X$ or $W$. We will return to this later.

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

$$L(\theta; \{z_i, y_i\}_1^n) = \frac{1}{2} \| y - X\theta \|^2_2.$$  \hfill (L)

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

**Definition/Claim:** 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 $\alpha \in [0, 1]$, 
   $$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{1}{2} c \alpha (1-\alpha) \|x-y\|^2_2.$$

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

3. (in the presence of differentiability) $\forall x, y \in \mathbb{R}^d$,
   $$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2} c \|y-x\|^2_2.$$

4. (in the presence of second-order differentiability) $\forall x \in \mathbb{R}^d$,
   $$\nabla^2 f(x) \succeq \frac{1}{2} c \|x\|^2_2 \quad \forall \nu \in \mathbb{R}^d.$$
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 \( \forall x, y \in \mathbb{R}^d, \)

\[
\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2
\]

A consequence of the above notion is the following:

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

\[
(\nabla f(x) - \nabla f(y))^T (x - y) > \frac{c L}{C + L} \| x - y \|_2^2 + \frac{1}{C + L} \| \nabla f(x) - \nabla f(y) \|_2^2.
\]

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

\[
\min_{\theta \in \mathbb{R}^d} f(\theta),
\]

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

\[
\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla f(\theta^k),
\]

where \( \alpha_k > 0 \) is the step size in the \( k \)-th iteration.

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

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

where \( \hat{\theta} \) is the optimal solution to \((P)\).
Theorem 1: We compute

\[ \| \theta^k - \hat{\theta} \|_2^2 = \| \theta^k - \alpha \nabla f(\theta^k) - \hat{\theta} \|_2^2 \]

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

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

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

\[ \geq \frac{c_L}{C+L} \| \theta^k - \hat{\theta} \|_2^2 + \frac{1}{C+L} \| \nabla f(\theta^k) \|_2^2. \]

Hence, by our choice of \( \alpha \), we have

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

\[ \leq (1 - \frac{2\alpha c_L}{C+L}) \| \theta^k - \hat{\theta} \|_2^2. \]

\[ \square \]

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