Statistics 210B Lecture 19 Notes

Daniel Raban

March 31, 2022

1 Efficient Error Estimation for Noisy, Sparse Linear Regression

1.1 Recap: introduction to noisy, sparse linear regression

We are investigating sparse linear regression, with the model $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^n, \qquad X \in \mathbb{R}^{n \times d}, \qquad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix}, \qquad \theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{bmatrix}.$$

We assume the sparsity condition $|S(\theta^*)| \le s$. Given (y, X), our task is to estimate θ^*. We had three formulations of the LASSO problem:

1. The λ **formulation**:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\},$$

2. 1**-norm constrained formulation**:

$$\arg\min_{\theta} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\} \qquad \text{s.t. } \|\theta\|_1 \le R$$

3. The **error constrained formulation**:

$$\arg\min_{\theta} \{\|\theta\|_1\} \qquad \text{s.t. } \frac{1}{2n} \|y - X\theta\|_2^2 \le b^2.$$

Given these three formulations, how can we give a tight upper bound of the estimation error $\|\widehat{\theta} - \theta_*\|_2$? Last time, in the noiseless setting, we had the restricted nullspace condition $\text{Null}(X) \cap \mathbb{C}(S) = \{0\}$, which was sufficient for exact recovery of θ_*. In this noisy setting, we will have the **restricted eigenvalue condition**, which will be sufficient for efficient estimation.

1.2 The restricted eigenvalue condition

Recall the $\mathbb{C}$ cone

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

We can modify this by adding a parameter:

$$\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\}.$$

In this extended definition, $\mathbb{C}(S) = \mathbb{C}_1(S)$. If we let $\alpha \to 0$, we get

$$\mathbb{C}_0(S) = \Delta \in \mathbb{R}^d : S(\Delta) = S\}.$$

Later we will focus on the $\mathbb{C}_\alpha$ cone for $\alpha = 3$.

Definition 1.1. $X \in \mathbb{R}^{n\times d}$ satisfies the **restricted eigenvalue condition** over $S \subseteq [d]$ with parameter (κ, α) (denoted $\mathrm{RE}(S, (\kappa, \alpha))$) if

$$\langle \Delta, (\tfrac{1}{n}X^\top X)\Delta\rangle = \frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \qquad \forall \Delta \in \mathbb{C}_\alpha(S).$$

This is called the restricted eigenvalue condition because the condition

$$\langle \Delta, (\tfrac{1}{n}X^\top X)\Delta\rangle \geq \kappa\|\Delta\|_2^2 \qquad \forall \Delta \in \mathbb{R}^d$$

is equivalent to $\lambda_{\min}(\frac{1}{n}X^\top X) \geq \kappa$.

Here is some intuition. We can think of the RE condition as a sort of strong convexity for the objective function. Suppose we define the objective function

$$L_n(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2,$$

which we want to minimize to get a minimizer $\widehat{\theta}$. The Hessian is

$$\nabla^2 L_n(\theta) = \frac{1}{n}X^\top X \in \mathbb{R}^d.$$

When the sample size is large, we know that there is concentration:

$$\sup_{\theta\in\mathbb{R}^d} |L_n(\theta) - \mathbb{E}[L_n(\theta)]| \leq \text{small},$$

but we want to bound $\|\widehat{\theta} - \theta^*\|_2$. If the Hessian is lower bounded by a large number, then the objective function will grow very fast around the minimizer. On the other hand, a

weak bound may mean that the objective function grows too slowly around the minimizer.

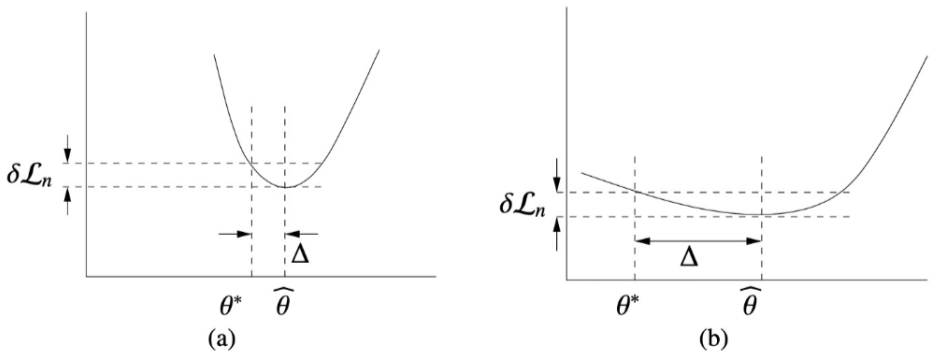

Figure 7.5 Illustration of the connection between curvature (strong convexity) of the cost function, and estimation error. (a) In a favorable setting, the cost function is sharply curved around its minimizer $\widehat{\theta}$, so that a small change $\delta\mathcal{L}_n := \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$ in the cost implies that the error vector $\Delta = \widehat{\theta} - \theta^*$ is not too large. (b) In an unfavorable setting, the cost is very flat, so that a small cost difference $\delta\mathcal{L}_n$ need not imply small error.

1.3 Bounds on ℓ_2 error

Our setting is

$$Y = X\theta^* + w, \qquad X \in \mathbb{R}^{n\times d}, \theta^* \in \mathbb{R}^d, w \in \mathbb{R}^n,$$

where $s \ll n \ll d$. We make two assumptions:

(A1): $S(\theta^*) = S \subseteq [d]$, where $|S| = s$.

(A2): X satisfies $\mathrm{RE}(S, (\kappa, \alpha = 3))$.

(A2) is a bit of an abstract condition. Later, we will show that Gaussian random matrices satisfy (A2) when the sample size n is larger enough than the sparsity level s.

Theorem 1.1. *Under assumptions (A1) and (A2),*

(a) λ formulation: Take the Lagrangian parameter $\lambda_n \geq 2\|\frac{X^\top w}{n}\|_\infty$. Then

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{\lambda_n}$$

(b) 1-norm constraint formulation: Take $R = \|\theta^*\|_1$. Then

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{4}{\kappa}\sqrt{s}\left\|\frac{X^\top w}{n}\right\|_\infty.$$

(c) Error constraint formulation: Let $b^2 \geq \frac{\|w\|_2^2}{2n}$. *Then*

$$\|\widehat{\theta} - \theta^*\| \leq \frac{4}{\kappa}\sqrt{s}\left\|\frac{X^\top w}{n}\right\| + \frac{2}{\sqrt{\kappa}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}}.$$

In all these cases, we have the 1-norm bound

$$\|\widehat{\theta} - \theta^*\|_1 \leq 4\sqrt{s}\|\widehat{\theta} - \theta^*\|_2.$$

Remark 1.1. This theorem is fully deterministic. There is no probability happening, and this theorem is entirely due to algebra.

Remark 1.2. The bound $\frac{1}{\kappa}$ is independent of n.

Remark 1.3. People generally think that the λ formulation is best because the bound is not so sensitive to the choice of the hyperparameter λ_n. In the second formulation, it is also difficult to pick R because we do not know what $\|\widehat{\theta^*}\|_1$ is.

In all cases, the error bound is $\sqrt{s}\|\frac{X^\top w}{n}\|_\infty$, and it is difficult to know what the typical size of this is. We make a further assumption: Assume X is deterministic with $\mathrm{RE}(S, (\kappa, 3))$ with $\max_{j\in[d]} \frac{\|x_j\|_2}{\sqrt{n}} \leq C$, where $x_j \in \mathbb{R}^n$ is the j-th column of X. Let $w \sim \mathrm{sG}(\sigma)$ with $\mathbb{E}[w] = 0$.

If these assumptions hold, then we claim that

$$\left\|\frac{X^\top w}{n}\right\|_\infty = \max_{i\in[d]} |\langle X_j, w\rangle/n| \lesssim \sigma\sqrt{\frac{\log d}{n}}.$$

Here, $\langle x_j, w\rangle/n \sim \mathrm{sG}(\sigma\sqrt{1/n})$. This tells us that

$$\|\widehat{\theta} - \theta^*\|_2 \lesssim \sqrt{s}\left\|\frac{X^\top w}{n}\right\|_\infty \lesssim \sqrt{\frac{s\log d}{n}}.$$

So we will have efficient estimation as long as $n \gg (\sigma^2 \vee 1)s\log d$.

1.4 Proof of RE condition bounds

The overall strategy is two steps:

1. Derive a basic inequality (the zero order optimality condition)
2. Algebraic manipulation.

Proof. (b): let's prove the 1-norm constraint formulation,

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2n}\|y - X|theta\|_2^2 \qquad \text{s.t. } \|\theta\|_1 \leq \|\theta^*\|_1 = R.$$

By the optimality of $\widehat{\theta}$, we know

$$\frac{1}{2n}\|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{2n}\|y - X\theta^*\|_2^2.$$

This is the zero order optimality condition. (The first order optimality condition for optimizing $f(x)$ subject to $g(x) \leq 0$ is $\nabla f(\widehat{x}) = \lambda \nabla g(\widehat{x})$, where λ is a scalar.) Here the right hand side is $\frac{1}{2n}\|w\|_2^2$, and the left hand side is $\frac{1}{2n}\|w + X(\theta^* - \widehat{\theta})\|_2^2$. So we have

$$\begin{aligned}\|w\|_2^2 &\geq \|w + X(\theta^* - \widehat{\theta})\|_2^2 \\ &= \|w\|_2^2 + 2\langle w, X(\theta^* - \widehat{\theta})\rangle + \|X(\theta^* - \widehat{\theta})\|_2^2\end{aligned}$$

Denote $\widehat{\Delta} = \widehat{\theta} - \theta^*$, which is what we want to bound. We can solve this to get

$$\|X\widehat{\Delta}\|_2^2 \leq 2\langle w, X\widehat{\Delta}\rangle.$$

Thus, our basic inequality is:

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \leq \frac{2}{n} w^\top X\widehat{\Delta}.$$

If $\widehat{\Delta} \in \mathbb{C}_\alpha(S)$, the left hand side can be lower bounded by

$$\frac{1}{n}\|X\widehat{\Delta}\|_2^2 \geq \kappa\|\widehat{\Delta}\|_2^2,$$

using the restricted eigenvalue condition. To check why $\widehat{\Delta} \in \mathbb{C}_\alpha(S)$, note that the condition $\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1$ tells us that $\widehat{\Delta} \in \mathbb{C}(S) \subseteq \mathbb{C}_3(S)$.

The right hand side can be upper bounded by viewing the scalar $w^\top X\widehat{\Delta}$ as the product of the vectors $w^\top X$ and $\widehat{\Delta}$:

$$\frac{2}{n} w^\top X\widehat{\Delta} \leq \frac{2}{n}\|X^\top w\|_\infty \cdot \|\widehat{\|}_1.$$

Since $\widehat{\Delta} \in \mathbb{C}(S)$, we can efficiently bound the 1-norm in terms of the 2-norm:

$$\|\widehat{\Delta}\|_1 = \|\widehat{\Delta}_{S^c}\|_1 + \|\widehat{\Delta}_S\|_1 \leq 2\|\widehat{\Delta}_S\|_1 \leq 2\sqrt{s}\|\widehat{\Delta}_S\|_2 \leq 2\sqrt{s}\|\widehat{\Delta}\|_2.$$

Using this in our inequality and dividing by κ on both sides gives

$$\|\widehat{\Delta}\|_2\| \leq \frac{4\sqrt{s}}{\kappa}\left\|\frac{X^\top w}{n}\right\|_\infty. \qquad \square$$

Remark 1.4. If instead of bounding by $\|X^\top w\|_\infty \cdot \|\widehat{\|}_1$, we try to bound by $\|X^\top w\|_2 \cdot \|\widehat{\|}_2$, then we get $\|\widehat{\Delta}\|_2 \le \frac{2}{\kappa}\|X^\top w/n\|_2 \sim \sqrt{\frac{d}{n}}$. This is worse than the rate $\sqrt{\frac{\log d}{n}}$.

The proof of (c) follows the same lines:

Proof. The error-constraint formulation

$$\widehat{\theta} = \arg\min_\theta\{\|\theta\|_1\} \qquad \text{s.t. } \frac{1}{2n}\|y - X\theta\|_2^2 \le b^2.$$

gives (using $y - X\widehat{\theta} = w - X\widehat{\Delta}$).

$$\begin{cases} \|\widehat{\theta}\|_1 \le \|\theta^*\|_1, \\ \frac{1}{2n}\|w + X\widehat{\Delta}\|_2 \le \frac{1}{2n}\|w\|_{@}^2 + \left(b^2 - \frac{1}{2n}\|w\|_2^2\right) \end{cases}$$

The algebra proceeds the same as for (b), but we have to keep track of the additive term $\frac{2}{\sqrt{\kappa}}\sqrt{b^2 - \frac{\|w\|_2^2}{n}}$. □

The proof of (a) has slightly different reasoning:

Proof. We first show that when $\lambda_n \ge 2\|\frac{X^\top w}{n}\|_\infty$, we have $\widehat{\Delta} \in \mathbb{C}_3(S)$. By optimality, we have

$$\frac{1}{2n}\|w + X\widehat{\Delta}\|_2^2 + \lambda_n\|\theta^* + \widehat{\Delta}\|_1 \le \frac{1}{2n}\|w\|_2^2 + \lambda_n\|\theta^*\|_1.$$

This gives us the Lagrangian basic inequality

$$\frac{1}{2n}\|X\widehat{\Delta}\|_2^2 \le \frac{w^\top X^\top \widehat{\Delta}}{n} + \lambda_n(\|\theta^*\|_1 - \|\theta^* + \widehat{\Delta}\|_1)$$

We can upper bound the right hand side by

$$\begin{aligned} &\le \left\|\frac{X^\top w}{n}\right\|_\infty \|\widehat{\Delta}\|_1 + \lambda_n(\|\theta_S^*\|_1 - \|\theta_S^* + \widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1) \\ &\le \left\|\frac{X^\top w}{n}\right\|_\infty \|\widehat{\Delta}\|_1 + \lambda_n(\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1) \\ &\le \frac{\lambda_n}{2}(3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1. \end{aligned}$$

This upper bound must be nonnegative, so

$$\|\widehat{\Delta}_{S^c}\|_1 \le 3\|\widehat{\Delta}_S\|_1,$$

which means that $\widehat{\Delta} \in \mathbb{C}_3(S)$. Now, by the RE condition and this bound we have shown,

$$\frac{\kappa}{2}\|\widehat{\Delta}\|_2^2 \le \frac{\lambda_n}{2}(3\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1)$$

$$\leq \frac{\lambda_n}{2} 3\sqrt{s} \|\widehat{\Delta}\|_2.$$

Canceling a factor of $\|\widehat{\Delta}\|_2$ on both sides, we get $\|\widehat{\Delta}\|_2 \leq \frac{3\lambda_n}{\kappa}\sqrt{s}$. □

Next time, we will show that the RE condition is satisfied with high probability for Gaussian random matrices.