Lecture 9: Lasso Estimator

Isabella Zhu

6 March 2025

§1 Summary

What we have so far:

- $\text{MSE}(\mathbb{X}\hat{\theta}_{B_0(s)}) \lesssim \frac{\sigma^2 s}{n} \log\left(\frac{ed}{s}\right)$ if $\theta^* \in B_0(s)$.
- $\text{MSE}(\mathbb{X}\hat{\theta}_{B_1(s)}) \lesssim \sigma\sqrt{\log d/n}$ if $\theta^* \in B_1$.
- Adapt to s using hard threshold, $\mathbb{X}^T\mathbb{X}/n = I_d$, $\text{MSE}(\mathbb{X}\hat{\theta}^{HARD}) \lesssim \frac{\sigma^2 s}{n} \log d$.

§2 The Lasso Estimator

We are continuing under the linear regression context. We are now assuming

$$\theta^* \in B_0(s) \text{ or } |\theta^*|_1 \leq R$$

for some R.

Definition 2.1. Fix $\tau > 0$. The Lasso estimator $\hat{\theta}^L$ is defined as

$$\hat{\theta}^L \in \operatorname*{argmin}_{\theta \in \mathbb{R}} \frac{1}{n}|Y - \mathbb{X}\theta|_2^2 + 2\tau|\theta|_1$$

Claim 2.2 — If $\frac{\mathbb{X}^T\mathbb{X}}{n} = I_D$, then $\hat{\theta}^L = \hat{\theta}^{SOFT}$.

§2.1 Slow Rate for Lasso

Definition 2.3. **Slow rate** refers to $\Theta\left(\frac{1}{\sqrt{n}}\right)$. **Fast rate** refers to $\Theta\left(\frac{1}{n}\right)$.

Theorem 2.4

Assume $|X_j|_2 \leq n$ and $2\tau = 2\sigma\sqrt{\frac{2\log(2d)}{n}} + 2\sigma\sqrt{2\log(1/\delta)/n}$. Then, MSE of Lasso estimator is

$$\text{MSE}(\mathbb{X}\hat{\theta}^L) \leq 4\tau|\theta^*|_1 \lesssim \sigma|\theta^*|_1\sqrt{\frac{\log d}{n}}$$

with probability at least $1 - \delta$.

> **Remark 2.5.** We use L1 norm instead of L2 norm because L1 encourages more sparsity.

Proof. By definition of $\hat{\theta}$, we have

$$\frac{1}{n}|Y - \mathbb{X}\hat{\theta}|_2^2 + 2\tau|\hat{\theta}|_1 \leq \frac{1}{n}|Y - \mathbb{X}\theta^*|_2^2 + 2\tau|\theta^*|_1$$

Like we've done five hundred other times already, we can rearrange

$$\frac{1}{n}|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq \frac{2}{n}\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^* \rangle + 2\tau|\theta^*|_1 - 2\tau|\hat{\theta}|_1 = \frac{2}{n}\langle \mathbb{X}^T\epsilon, \hat{\theta} - \theta^* \rangle + 2\tau|\theta^*|_1 - 2\tau|\hat{\theta}|_1$$

By Holder, we can bound

$$2\langle \mathbb{X}^T\epsilon, \hat{\theta} - \theta^* \rangle \leq 2|\mathbb{X}^T\epsilon|_\infty|\hat{\theta} - \theta^*|_1$$

Note that $|\mathbb{X}^T\epsilon|_\infty = \max|\mathbb{X}_j^T\epsilon|$ so $\mathbb{X}_j^T\epsilon$ is subgaussian with variance proxy $\sigma^2 n$. Therefore, with high probability,

$$2|\mathbb{X}^T\epsilon|_\infty|\hat{\theta} - \theta^*|_1 \leq 2 \cdot n\tau|\hat{\theta} - \theta^*|_1$$

Therefore, we get that

$$\frac{1}{n}|\mathbb{X}\hat{\theta} - \mathbb{X}\star|_2^2 \leq 2\tau|\hat{\theta} - \theta^*|_1 + 2\tau|\theta^*|_1 - 2\tau|\hat{\theta}|_1 \leq 2\tau|\theta^*|_1 + 2\tau|\theta^*|_1 = 4\tau|\theta^*|_1$$

as desired.

> **Remark 2.6.** This only requires $|X_j|_2 \leq \sqrt{n}$ as opposed to $\frac{\mathbb{X}^T\mathbb{X}}{n} = I_d$, so this works for n much less than d. $\hat{\theta}^L$ is similar to $\hat{\theta}^{HARD}$, adaptive to s but still requires σ and δ.

§2.2 Fast Rate for Lasso

We need $\frac{\mathbb{X}^T\mathbb{X}}{n} \approx I_d$.

Definition 2.7. Incoherence. The design matrix $\mathbb{X}$ satisfies incoherence (INC(k)) with parameter k if

$$\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right|_\infty \leq \frac{1}{32k}.$$

This translates to the following conditions on our $\mathbb{X}_j$s:

$$\left|\frac{|\mathbb{X}_j|_2^2}{n} - 1\right| \leq \frac{1}{32k}, \quad \left|\frac{|\mathbb{X}_i^T\mathbb{X}_j|_2^2}{n} - 0\right| \leq \frac{1}{32k}$$

Lemma 2.8

Fix $k \leq d$ and assume INC(k). Then for $S \subset \{1, \dots d\}$, $|S| \leq k$, and $\theta \in \mathbb{R}^d$, s.t.

$$|(\theta_{S^c})|_1 \leq 3|\theta_S|_1, \text{ (cone condition)}$$

then it holds

$$|\theta|_2^2 \leq \frac{2|\mathbb{X}\theta|_2^2}{n} \text{ (restricted eigenvalue condition)}$$

We have

$$(\theta_S)_j = \begin{cases} \theta_j & \text{if } j \in S \\ 0 & \text{otherwise} \end{cases}$$

S^c denotes the complement of S.

We have

$$\frac{|\mathbb{X}\theta|_2^2}{n} = \frac{|\mathbb{X}\theta_S|_2^2}{n} + \frac{|\mathbb{X}\theta_{S^c}|_2^2}{n} + 2\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}$$

We deal with each term separately.

$$\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_S = |\theta_S|_2^2 + \theta_S^T\left(\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right)\theta_S$$

By Holder, we know

$$\theta_S^T\left(\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right)\theta_S \geq -|\theta_S|_1^2\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right|_\infty \geq -\frac{|\theta_S|_1^2}{32k}$$

Therefore, we get

$$\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_S \geq |\theta_S|_2^2 - \frac{|\theta_S|_1^2}{32k}$$

We can do a similar thing for $\frac{|\mathbb{X}\theta_{S^c}|_2^2}{n}$, combined with the cone condition, which gives

$$\frac{|\mathbb{X}\theta_{S^c}|_2^2}{n} \geq |\theta_{S^c}|_2^2 - \frac{9|\theta_S|_1^2}{32k}$$

Finally, for the third term,

$$2\left|\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{SC}\right| \leq 2|\theta_S^T\theta_{SC}| + \frac{2}{32k}|\theta_S|_1 \cdot |\theta_{SC}|_1 \leq \frac{6}{32k}|\theta_S|_1^2$$

as θ_S and θ_{SC} are orthogonal.

We need to convert L1 to L2 using Cauchy Schwarz. For first term:

$$\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_S \geq |\theta_S|_2^2 - \frac{|\theta_S|_1^2}{32k} \geq |\theta_S|_2^2 - \frac{|S| \cdot |\theta_S|_2^2}{32k}$$

and similarly for the rest. Using the fact that $|S| \leq k$, we collect terms to get

$$\frac{|\mathbb{X}\theta|_2^2}{2} \geq |\theta_S|_2^2 + |\theta_{SC}|_2^2 - (\frac{1}{32} + \frac{9}{32} + \frac{6}{32})|\theta_S|_2^2 = |\theta|_2^2 - \frac{1}{2}|\theta_S|_2^2 \geq \frac{1}{2}|\theta|_2^2.$$

Theorem 2.9

Assume INC(k) with k equal to the sparsity of θ^* (i.e. $k = |\theta^*|_0$). Fix

$$2\tau = 8\sigma\sqrt{\log(2d)/n} + 8\sigma\sqrt{\log(1/\delta)/n}.$$

Then, the MSE of the lasso estimator is at most

$$\text{MSE}(\mathbb{X}\hat{\theta}^L) \leq 32k\tau^2 \lesssim \frac{\sigma^2|\theta^*|_0}{n}\log(d/\delta)$$

Moreover,

$$|\hat{\theta} - \theta^*|_2^2 \leq 2\text{MSE}(\mathbb{X}\hat{\theta}^L)$$

all happening with probability at least $1 - \delta$.

Proof. For the five hundred millionth time, we start with the good ole basic inequality

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*\rangle + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}|_1$$

We bound

$$2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*\rangle \leq 2|\mathbb{X}^T\epsilon|_\infty \cdot |\hat{\theta} - \theta^*|_1$$

We bound the highest column norm of $\mathbb{X}$. We have

$$|\mathbb{X}_j|_2^2 = (\mathbb{X}^T\mathbb{X})_{jj} \leq n + \frac{n}{32k} \leq 2n$$

by the incoherence property. Therefore, we get

$$2\langle \epsilon, \mathbb{X}\hat{\theta} - \mathbb{X}\theta^*\rangle \leq 2|\mathbb{X}^T\epsilon|_\infty \cdot |\hat{\theta} - \theta^*|_1 \leq 2 \cdot 2n \cdot \frac{\tau}{4} \cdot |\hat{\theta} - \theta^*|_1 = n\tau|\hat{\theta} - \theta^*|_1$$

To summarize, we've proved so far that

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq n\tau|\hat{\theta} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}|_1$$

We add $n\tau|\hat{\theta} - \theta^*|_1$ on both sides.

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 + n\tau|\hat{\theta} - \theta^*|_1 \leq 2n\tau|\hat{\theta} - \theta^*|_1 + 2n\tau|\theta^*|_1 - 2n\tau|\hat{\theta}|_1$$

Now we take the support S into account. We have

$$|\hat{\theta}|_1 = |\hat{\theta}_S|_1 + |\hat{\theta}_{S^c}|_1 \implies |\hat{\theta} - \theta^*|_1 - |\hat{\theta}|_1 = |\hat{\theta}_S - \theta^*|_1 - |\hat{\theta}_S|_1.$$

Putting it together,

$$|\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 + |\mathbb{X}\hat{\theta} - \mathbb{X}\theta^*|_2^2 \leq 2n\tau\left[|\hat{\theta}_S - \theta^*|_1 + |\theta^*|_1 - |\hat{\theta}|_S\right] \leq 4n\tau|\hat{\theta}_S - \theta^*|_1$$

We have that

$$|\hat{\theta} - \theta^*|_1 \leq 4|\hat{\theta}_S - \theta^*|_1 \leftrightarrow |\hat{\theta}_{S^c} - \theta^*_{S^c}| \leq 3|\hat{\theta}_S - \theta^*_S|$$

which is exactly the cone condition! We'll wrap this up next lecture.