MATH888: High-dimensional probability and statistics Fall 2021

Lecture 37 — December 3, 2021

Sebastien Roch, UW-Madison *Scribe: Kwangmoon Park, Sebastien Roch*

1 Overview

At the end of the Lecture 36, we introduced LASSO estimator for sparse linear regression set up, which also employs thresholding method. In this lecture, we examine the Mean Squared Error(MSE) upper bound of LASSO estimator. Specifically, we will first prove a weak MSE bound of the LASSO estimator, and will show that LASSO achieves the rate $\frac{\sigma^2}{n}\log(d)$ with additional assumption: Incoherence.

2 LASSO estimator

2.1 Model Definition

The LASSO estimator of θ^* is defined by any $\hat{\theta}^{\mathcal{L}}$ such that

$$\hat{\theta^{\mathcal{L}}} \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+2\tau\|\theta\|_1\right\}, \tag{1}$$

where $Y = \mathbb{X}\theta^* + \epsilon$, for $Y \in \mathbb{R}^n$, $\mathbb{X} \in \mathbb{R}^{n\times d}$ and some random ϵ. Note that $\|\cdot\|_2$ and $\|\cdot\|_1$ are L_1 and L_2 norm respectively, and τ is a thresholding parameter. The LASSO estimator penalizes the size of the choice of θ in addition to the MSE, and thus it is sometimes referred to as L_1 penalized linear regression model.

2.2 Weak rate of the LASSO

Here, we suggest a MSE bound for (1) with relatively weak assumptions: Sub-Gaussainity and normalized $\mathbb{X}$. Observe that the bound obtained from the theorem below is weaker than the one we saw in the last lecture.

Theorem 1 (Weak rate of the LASSO)**.** *Assume that the linear model $Y = \mathbb{X}\theta^* + \epsilon$ holds, where $\epsilon \sim subG_n(\sigma^2)$. Additionally, assume that each j-th column of $\mathbb{X}$ $\mathbb{X}_j$ satisfies $\max_j \|\mathbb{X}_j\|_2 \le \sqrt{n}$. Then, with probability at least $1-\delta$, the LASSO estimator in (1) with regularization parameter*

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

satisfies the MSE bound

$$\mathit{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\hat{\theta}^*\|_2^2 \le 4\|\theta^*\|_1\sigma\sqrt{\frac{2\log(2d)}{n}} + 4\|\theta^*\|_1\sigma\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Note that we have $\sigma\sqrt{\frac{\log(2d)}{n}}$ in the bound, which is slower than the bound $\sigma^2\frac{\log(2d)}{n}$ for Hard/Soft Thresholding estimator from the last lecture. Let us first prove this Theorem 1 and show that we can obtain same rate for the LASSO estimator with additional assumptions.

Proof. By the definition of the LASSO estimator $\hat{\theta^{\mathcal{L}}}$, we have

$$\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 + 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \le \|Y - \mathbb{X}\hat{\theta}^*\|_2^2 + 2n\tau\|\hat{\theta}^*\|_1, \tag{3}$$

since the LASSO estimator is the minimizer of the objective function in (1).

By using Holder's inequality after plugging in $Y = \mathbb{X}\theta^* + \epsilon$ on the left hand side of (3), we have

$$\begin{aligned}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\hat{\theta}^*\|_2^2 &\le 2\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*)^2 + 2n\tau(\|\hat{\theta}^{\mathbb{L}}\|_1 - \|\theta^*\|_1)\\ &\le 2\|\mathbb{X}^T\epsilon\|_\infty\|\hat{\theta}^{\mathcal{L}}\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 + 2\|\mathbb{X}^T\epsilon\|_\infty\|\theta^*\|_1 - 2n\tau\|\theta^*\|_1\\ &= 2(\|\mathbb{X}^T\epsilon\|_\infty - n\tau)\|\hat{\theta}^{\mathcal{L}}\|_1 + 2(\|\mathbb{X}^T\epsilon\|_\infty + n\tau)\|\theta^*\|_1. \end{aligned} \tag{4}$$

We will show that the first part of (4) is at most 0 and the second part of (4) is at most $4n\tau\|\theta * \|_1$ with probability at least $1-\delta$ with appropriate choice of t below.

By Sub-Gaussianity of ϵ, we have the tail probability bound

$$\begin{aligned}\mathbb{P}(\|\mathbb{X}^T\epsilon\|_\infty \ge t) &= \mathbb{P}(\max_j |\mathbb{X}_j^T\epsilon| \ge t)\\ &\le 2d\exp\left(-\frac{t^2}{2n\sigma^2}\right).\end{aligned}$$

Therefore, by taking $t = \sqrt{2n\log(2d)} + \sigma\sqrt{2n\log(1/\delta)} = n\tau$, we get with probability at least $1-\delta$, (4) $\le 4n\tau\|\theta * \|_1$, which implies

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\hat{\theta}^*\|_2^2 \le 4n\tau\|\theta * \|_1,$$

and it is the same as the bound in the theorem above with $\tau = \sigma\sqrt{\frac{2\log(2d)}{n}} + \sigma\sqrt{\frac{2\log(1/\delta)}{n}}$.

□

2.3 Fast rate for the LASSO

In order to obtain faster MSE bound for the LASSO, we need an additional assumption called Incoherence defined as below.

Assumption 2 (Incoherecne)**.** *We say that the matrix $\mathbb{X}$ has incoherence k or denote it as INC(k) for some integer $k > 0$ if*

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

where the $\|\cdot\|_\infty$ denotes the largest element of A in absolute sense.

Note that this INC(k) is an approximate version of assumption ORT introduced in the last lecture, and in case $k \to \infty$, we can obtain ORT from INC(k). Recall that we needed $d \leq n$ in ORT. However, INC allows $d \gg n$, which is more reasonable in high-dimensional situation.

We use this INC assumption to employ a Lemma that will be introduced below, and we will use the Lemma to prove faster bound theorem for the LASSO. Let us take a look at the Lemma.

For any $\Delta \in \mathbb{R}^d$, $S \subset \{1, \dots, d\}$, define Δ_S to be the vector with coordinates

$$\Delta_{S,j} = \begin{cases} \Delta_j \text{ if } j \in S \\ 0 \text{ otherwise .} \end{cases}$$

With this, we have $\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1$, and the following Lemma holds.

Lemma 3. *Fix a positive integer $k \leq d$ and assume that $\mathbb{X}$ satisfies assumption INC(k). Then, for any $S \in \{1, \dots, d\}$ such that $|S| \leq k$ and any $\Delta \in \mathbb{R}^d$ that satisfies the cone condition*

$$\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1, \tag{5}$$

it holds that

$$\|\Delta\|_1 \leq 2\frac{\|\mathbb{X}\Delta\|_2^2}{n}.$$

Note that in our proof of the fast rate of convergence, we will regard the difference between θ^*, and $\hat{\theta}^{\mathcal{L}}$ in our model as Δ.

For some intuition behind the conditions in this lemma, see Figures 7.2 and 7.6 in [2]. Let us now prove the lemma.

Proof. By the definition of Δ_S, we have

$$\frac{\|\mathbb{X}\Delta\|_2^2}{n} = \frac{\|\mathbb{X}\Delta_S\|_2^2}{n} + \frac{\|\mathbb{X}\Delta_{S^c}\|_2^2}{n} + 2\Delta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\Delta_S. \tag{6}$$

We bound each of the three terms in the RHS of (6).

I. Firstly, it follows from the incoherence condition that

$$\frac{\|\mathbb{X}\Delta_S\|_2^2}{n} = \Delta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\Delta_S = \|\Delta_S\|_2^2 + \Delta_S^T\left(\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right)\Delta_S \geq \|\Delta_S\|_2^2 - \frac{\|\Delta_S\|_1^2}{32k}. \tag{7}$$

II. By the same arguent in **I** and the cone condition (5), we have

$$\frac{\|\mathbb{X}\Delta_{S^c}\|_2^2}{n} \geq \|\Delta_{S^c}\|_2^2 - \frac{\|\Delta_{S^c}\|_1^2}{32k} \geq \|\Delta_{S^c}\|_2^2 - \frac{9\|\Delta_S\|_1^2}{32k}. \tag{8}$$

III. Finally, for the cross product term, by the incoherence condition and Cauchy-Swartz inequality, we have

$$\begin{align} 2\Delta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\Delta_S &\leq \frac{2}{32k}\|\Delta_S\|_1\|\Delta_{S^c}\|_2 \tag{9} \\ &\leq \frac{6}{32k}\|\Delta_S\|_1^2, \tag{10} \end{align}$$

where the last inequality holds by the cone condition (5).

Therefore, combining **I**, **II**, **III** and given the condition that $|S| \leq k$, we have

$$\begin{aligned}\frac{\|\mathbb{X}\Delta\|_2^2}{n} &\geq \|\Delta_S\|_2^2 + \|\Delta_{S^c}\|_2^2 - \frac{16|S|}{32k}\|\Delta_S\|_2^2 \\ &\geq \frac{1}{2}\|\Delta\|_2^2\end{aligned}$$

□

One remark from this Lemma is that given the incoherence assumption, the cone condition implies the strong positive curvature, and it is related to how much the LASSO solution is close to the true parameter θ^*.

Now that we have Lemma 3, we are ready to prove the fast rate for the LASSO stated as below.

Theorem 4 (Fast rate of the LASSO)**.** *Fix $n \geq 2$, and assume the linear model $Y = \mathbb{X}\theta^* + \epsilon$, where $\epsilon \sim subG_n(\sigma^2)$. In addition, assume that $\|\theta^*\|_0 \leq k$ and that $\mathbb{X}$ satisfies assumption INC(k). Then the LASSO estimator $\hat{\theta}^{\mathcal{L}}$ with τ such that*

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

satisfies

$$MSE(\mathbb{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n} \tag{11}$$

and

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n} \tag{12}$$

with probability at least $1 - \delta$.

Proof. Similar to the proof of Theorem1, by the definition of the LASSO estimate, we have

$$\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 \leq \|Y - \mathbb{X}\theta^*\|_2^2 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1. \tag{13}$$

By adding $n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1$ on both sides of (13) and with some arranging of the equations after plugging in $Y = \mathbb{X}\theta^* + \epsilon$, we get

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1. \tag{14}$$

Let us remove the cross product term on the right hand side of (14). Once we apply Holder's inequality and use similar steps as in the proof of Theorem 1, with probability at least $1 - \delta$, we get

$$\begin{aligned}\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) &\leq \|\epsilon^T\mathbb{X}\|_\infty\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \\ &\leq \frac{n\tau}{2}\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1,\end{aligned}$$

where we used the fact that $\|\mathbb{X}_j\|_2^2 \leq n + 1/(32k) \leq 2n$. Thus, the cross product term cancels out with the second term of the LHS of (14). Taking $S = \text{supp}(\theta^*)$, we get

$$\begin{aligned}
(14) &\leq 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \\
&= 2n\tau\|\hat{\theta}_S^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}_S^{\mathcal{L}}\|_1 \\
&\leq 4n\tau\|\hat{\theta}_S^{\mathcal{L}} - \theta^*\|_1,
\end{aligned}$$

where the first equality is from the fact that $2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 = 2n\tau\|\hat{\theta}_{S^c}^{\mathcal{L}}\|_1$ on S^c and the fact that $2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 = 2n\tau\|\hat{\theta}_S^{\mathcal{L}} - \theta^*\|_1$ on S and that $2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 = 2n\tau\|\hat{\theta}_S^{\mathcal{L}}\|_1 + 2n\tau\|\hat{\theta}_{S^c}^{\mathcal{L}}\|_1$. The last inequality is from the triangle inequality.

In particular, we have $\|\hat{\theta}_{S^c}^{\mathcal{L}} - \theta_{S^c}^*\|_1 \leq 3\|\hat{\theta}_S^{\mathcal{L}} - \theta_S^*\|_1$, because we have

$$\begin{aligned}
n\tau\{\|\hat{\theta}_S^{\mathcal{L}} - \theta_S^*\|_1 + \|\hat{\theta}_{S^c}^{\mathcal{L}} - \theta_{S^c}^*\|_1\} &= n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \\
&\leq \|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1. \\
&\leq 4n\tau\|\mathbb{X}\hat{\theta}_S^{\mathcal{L}} - \mathbb{X}\theta^*\|_1 \\
&= 4n\tau\|\mathbb{X}\hat{\theta}_S^{\mathcal{L}} - \mathbb{X}\theta_S^*\|_1.
\end{aligned}$$

Then, with $\Delta = \hat{\theta}^{\mathcal{L}} - \theta^*$, it satisfies the cone condition (5). Thus, using the Lemma 3 and Cauchy-Swartz inequality, since $|S| \leq k$, we have

$$\begin{aligned}
\|\hat{\theta}_S^{\mathcal{L}} - \theta^*\|_1 &\leq \sqrt{|S|}\|\hat{\theta}_S^{\mathcal{L}} - \theta^*\|_2 \\
&\leq \sqrt{|S|}\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_2 \\
&\leq \sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2. \qquad (15)
\end{aligned}$$

Combining the result (15) with the fact that

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq \|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 4n\tau\|\hat{\theta}_S^{\mathcal{L}} - \theta^*\|_1 \leq 4n\tau\sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2,$$

we eventually get

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 &\leq \left(4n\tau\sqrt{\frac{2k}{n}}\right)^2 \\
&\leq 32n\tau^2 k,
\end{aligned}$$

which concludes the bound on MSE.

For the coefficient's bound $\|\hat{\theta}^L - \theta^*\|_2^2$, we again Lemma 3 and get

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_2^2 \leq 2\text{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) \leq 64k\tau^2. \qquad (16)$$

□

In fact, we do not even need incoherence condition INC(k) for the bound in Theorem 4, but the conclusion of Lemma 3,

$$\inf_{|S|\leq k} \inf_{\theta \in C_s} \left\{ \frac{\|\mathbb{X}\theta\|_2^2}{n\|\theta\|_2^2} \geq \kappa \right\},$$

where $\kappa = {}^{1}\!/_{2}$ and C_s is defined as $C_S = \{\|\theta_{S^c}\|_2 \leq 3\|\theta_S\|_1\}$.

Eventually, the LASSO estimator (1) gives equivalent bound up to a constant C as the Hard/Soft Thresholding estimators that we introduced in the last lecture.

References

[1] Philippe Rigollet and Jan-Christian Hütter, *18.657: High Dimensional Statistics Lecture Notes*, MIT, 2017.

[2] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint.* CUP.