MATH888: High-dimensional probability and statistics Fall 2021

Lecture 36 — December 1, 2021

Sebastien Roch, UW-Madison *Scribe: Thanasis Pittas*

1 Overview

In Lecture 34, we saw an estimator for sparse linear regression with strong error guaranties. However, that estimator was computationally inefficient when the sparsity level was unknown. In Lecture 35 we introduced the sub-Gaussian sequence model and observed that under the assumption ORT (imposing that $\frac{1}{n}\mathbb{X}\mathbb{X}^T = I_p$), that model is essentially equivalent to that of linear regression, up to a transformation of the data and a scaling of the variance. In this lecture, we give an estimator for the sparse sub-Gaussian sequence model that is computationally efficient even if the sparsity level is unknown. This implies an estimator for sparse linear regression that has the same statistical guarantees as that of Lecture 34 as well as improved computational complexity.

2 Sparse Sub-Gaussian Sequence Model

2.1 Model Definition

Recall the sparse sub-Gaussian Sequence Model: The observation is a vector $y \in \mathbb{R}^p$ which is generated by the process

$$y = \theta^* + \xi \ ,$$

where $\xi \sim \mathrm{subG}_p(\sigma^2/n)$. Following the notation of the previous lectures, we denote by $\|\theta^*\|_0$ the sparsity of θ^*, that is, $\|\theta^*\|_0 := \sum_{i=1}^p \mathbb{1}\{\theta_i^* \neq 0\}$. The goal is to find an estimator $\hat{\theta} = \hat{\theta}(y)$ such that $\|\hat{\theta} - \theta^*\|_2$ is small with high probability.

2.2 Hard Thresholding Estimator

Since every entry of y is a perturbed version of the corresponding entry of θ^* and since the noise ξ is concentrated, we are motivated to threshold the entries of y, hoping that this will distinguish the zero entries of θ^* from the non-zero entries. We thus define the *hard threshold estimator*

$$\hat{\theta}_j^H = \begin{cases} 0, & |y_j| \leq 2\tau \ , \\ y_j, & |y_j| > 2\tau \ , \end{cases} \tag{1}$$

where the threshold τ remains to be specified. Formally, we show the following (also see Theorem 2.11 in [1]).

Theorem 1. *Let C be a large enough constant. If $\tau = \sigma\sqrt{(2C/n)\log(2p/\delta)}$, the estimator $\hat{\theta}^H$ from Equation (1) satisfies*

$$\|\hat{\theta}^H - \theta^*\|_2^2 \lesssim \frac{\sigma^2}{n}\|\theta^*\|_0 \log\left(\frac{2p}{\delta}\right) ,$$

with probability at least $1-\delta$.

In comparison to the estimator from two lectures ago, we note that $\hat{\theta}^H$ is computationally efficient while achieving the same (optimal) error rate.

Proof. We start with the following claim, saying that the selection of τ is such that all additive errors are small.

Claim 2. *Let $\mathcal{A} = \{|\xi_j| \le \tau \; \forall j\}$. Then $\mathbb{P}[A^c] \le \delta$.*

Proof of Claim 2. By writing the complement of $\mathcal{A}$ as a union and using union bound, we have that

$$\mathbb{P}[A^c] = \mathbb{P}\left[\bigcup_{j=1}^{p}\{|\xi_j| > \tau\}\right] \le \sum_{j=1}^{p}\mathbb{P}\left[\{|\xi_j| > \tau\}\right] \le 2p\exp\left(-\frac{\tau^2}{C2(\sigma^2/n)}\right) \le \delta ,$$

where for the second inequality we use that $\xi \sim \mathrm{subG}(\sigma^2/n)$, which means that if e_j denotes the j-th vector of the orthonormal basis of $\mathbb{R}^p$, then $\xi_j = \langle e_j, \xi\rangle \sim \mathrm{subG}(\sigma^2/n)$. The last inequality uses $\tau = \sigma\sqrt{(2C/n)\log(2p/\delta)}$. □

Next, we note that under the event $\mathcal{A}$, the contribution to the error from all coordinates is small.

Claim 3. *Under the event $\mathcal{A} = \{|\xi_j| \le \tau \; \forall j\}$, the following hold for all $j \in \{1, \dots, p\}$:*

1. *If $\theta_j^* = 0$, then $\hat{\theta}_j^H = 0$.*
2. *If $\theta_j^* \neq 0$, then $|\hat{\theta}_j^H - \theta_j^*| \le 3\tau$.*

Proof of Claim 3. For the first case, assume that $\theta_j^* = 0$. Then, $y_j = \xi_j$ and thus

$$|y_j| = |\xi_j| \le \tau \le 2\tau ,$$

which means that $\theta_j^H = 0$. For the second case, $\theta_j^* \neq 0$, we examine two sub-cases:

1. If $|y_j| > 2\tau$, then $\hat{\theta}_j^H = y_j$ and thus $|\hat{\theta}_j^H - \theta_j^*| = |\xi_j| \le \tau \le 3\tau$.
2. If $|y_j| \le 2\tau$, then $\hat{\theta}_j^H = 0$ and thus $|\hat{\theta}_j^H - \theta_j^*| = |\theta_j^*| = |y_j - \xi_j| \le 3\tau$ by the triangle inequality.

□

Having these two claims at hand, we note that $\|\hat{\theta}^H - \theta^*\|_2^2 = \sum_{j=1}^{p}(\hat{\theta}_j^H - \theta_j^*)^2 \le 9\tau^2\|\theta^*\|_0$. Plugging the value of τ completes the proof of Theorem 1. □

2.3 Soft Thresholding Estimator

An alternative way of thresholding the values of y is used in the following estimator, for which we get the same guarantee as in Theorem 1.

$$\hat{\theta}_j^S = \begin{cases} 0, & |y_j| \le 2\tau \ , \\ y_j - 2\tau, & y_j > 2\tau \ , \\ y_j + 2\tau, & y_j < -2\tau \ , \end{cases} \tag{2}$$

or, in a more compact form,

$$\hat{\theta}_j^S = \left(1 - \frac{2\tau}{|y_j|}\right)_+ y_j$$

Theorem 4. *Let C be a sufficiently large constant. If $\tau = \sigma\sqrt{(2C/n)\log(2p/\delta)}$, the estimator $\hat{\theta}^S$ from Equation (2) satisfies*

$$\|\hat{\theta}^S - \theta^*\|_2^2 \lesssim \frac{\sigma^2}{n}\|\theta^*\|_0 \log\left(\frac{2p}{\delta}\right) \ ,$$

with probability at least $1-\delta$.

The proof of this is the same as that of Theorem 1, up to a small modification of the proof of Claim 3. More specifically, the case $\theta_j^* = 0$ remains the same. Regarding the other case, we again have two sub-cases: If $|y_j| > 2\tau$, then $\hat{\theta}_j^S = y_j - 2\tau$ and thus $|\hat{\theta}_j^S - \theta_j^*| = |\xi - 2\tau| \le 3\tau$. The second sub-case does not need any changes.

2.3.1 LASSO estimator

The reason that we consider the soft thresholding estimator is that it is the solution of an optimization problem of the following nice form.

$$\hat{\theta}^S = \arg\min_{\theta\in\mathbb{R}^p} \left\{\|y-\theta\|_2^2 + 4\tau\|\theta\|_1\right\}.$$

It can be easily checked that this is an equivallent form of the soft thresholding estimator by writing $\|y-\theta\|_2^2 + 4\tau\|\theta\|_1 = \sum_{i=1}^p \left((y_i-\theta_i)^2 + 4\tau|\theta_i|\right)$ and solving the optimization problem for every coordinate separately.

To slightly generalize the above, we define the LASSO estimator

$$\hat{\theta}^{LASSO} = \arg\min_{\theta\in\mathbb{R}^p} \left\{\frac{1}{n}\|y-\mathbb{X}\theta\|_2^2 + 4\tau\|\theta\|_1\right\}.$$

Based on the discussion of the last lecture, under the assumption ORT, the above estimator is equivalent to the soft thresholding estimator. However, the LASSO estimator is natural to consider even in the absence of that assumption.

Because of its particular form, there are algorithms for computing $\hat{\theta}^{LASSO}$ numerically. One such method is coordinate descent:

1. Until Convergence:

   (a) For $j = 1, \ldots, p$:

      i. $\theta_j \leftarrow R_j \left(1 - 2\tau/|R_j|\right)_+$, where $R_j = \mathbb{X}_j^T \left(y - \sum_{k \neq j} \theta_k \mathbb{X}_k\right)$.

Using standard optimization results, it can be shown that this method converges to the true solution. It remains to show statistical guaranties for $\hat{\theta}^{LASSO}$, which will leave for the next lecture.

References

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