MATH888: High-dimensional probability and statistics Fall 2021

Lecture 38 — December 6, 2021

Sebastien Roch, UW-Madison *Scribe: Lisheng Ren, Sebastien Roch*

1 Overview

In the last lecture we gived the analysis of Lasso estimator for sparse linear regression. In this lecture we will wrap up the analysis and then generalize the setting to sparse oracle inequality.

2 Main Section

2.1 Lasso estimator on sparse linear regression

First recall our definition for Lasso estimator

Definition 1. *Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be n samples on $\mathbb{R}^p$ and $y \in \mathbb{R}^n$ be the labels. We define the Lasso estimator as*

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

Then recall the definition for Incoherence

Definition 2. *We say the matrix $\mathbb{X}$ has incoherence k and denote it as $\text{INC}(k)$ for integer $k > 0$ if*

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

Note we need the incoherence assumption for Fast rate of Lasso estimator.

Theorem 3. *Under $\text{INC}(k)$ assumption, with*

$$\tau = C\sigma\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

then with probability $1-\delta$

$$\frac{1}{n}\|\mathbb{X}\hat{\theta}^L - \mathbb{X}\theta^*\|_2^2 \lesssim \frac{\sigma^2\|\theta^*\|_0\log(p/\delta)}{n}$$

The proof has already been discussed in note for Lecture 37.

Proposition 4 (Proposition 2.16 in [1])**.** *Let $\mathbb{X} \in \{\pm 1\}^{n\times p}$ be a random matrix such that entries are i.i.d Rademacher (± 1) random variables. The incoherence of $\mathbb{X}$ is k w.p. $1-\delta$ as long as*

$$n \gtrsim k^2\log(p/\delta)$$

Proof. For diagonal entries,

$$\left(\frac{\mathbb{X}^T\mathbb{X}}{n}\right)_{j,j} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{X}_{i,j}^2 = 1$$

Thus $\frac{\mathbb{X}^T\mathbb{X}}{n} - I_p$ is 0 on the diagonal.
For non-diagonal entries,

$$\left(\frac{\mathbb{X}^T\mathbb{X}}{n}\right)_{i,j} = \frac{1}{n}\sum_{l=1}^{n}\mathbb{X}_{l,i}\mathbb{X}_{l,j}$$

Note $\mathbb{X}_{l,i}\mathbb{X}_{l,j}$ is also Rademacher random variable which is subgaussian, thus $\|\sum_{l=1}^{n}\mathbb{X}_{l,i}\mathbb{X}_{l,j}\|_{\psi_2} \lesssim n$. Therefore concentration and an union bond on non-diagonal entries, we have

$$\mathbb{P}(\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_p\|_\infty > \frac{1}{32k}) \leq \sum_{i\neq j\in[p]}\mathbb{P}(\frac{1}{n}\sum_{l=1}^{n}\mathbb{X}_{l,i}\mathbb{X}_{l,j} > \frac{1}{32k}) \leq p^2\exp(-\frac{Cn}{k^2})$$

Take the value equal to δ gives the proposition. □

2.1.1 Lasso estimator on sparse oracle equality

The setting of the problem is the following. Let $\phi_1, \cdots, \phi_m : \mathcal{X} \mapsto \mathbb{R}$ be a set of dictionary functions and $\theta \in \mathbb{R}^m$ be a k-sparse vector. We will use ϕ_θ to denote $\phi_\theta = \sum_{i=1}^{m}\theta_i\phi_i$. We assume the true concept is $f = \phi_{\theta^*}$ and assume we observe $\mathbb{X}$ and Y such that

$$Y_i = f(X_i) + \epsilon_i,\ i = 1, \cdots, n$$

where and ϵ is a subgaussian with variance proxy σ^2.

For any estimator $\hat{f}$, we define the mean squared error as

$$\text{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \hat{f}(X_i))^2$$

In this setting, we can reduce the problem to sparse linear regression by considering the matrix Φ with i-th row equal to $[\phi_1(X_i), \cdots, \phi_m(X_i)]$. We define the Lasso estimator $\hat{\theta}^L$ for θ^* as

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

Therefore the same analysis for Lasso estimator can also be applied here.

Theorem 5 (Theorem 3.5 in [1]). *Under* INC(k) *assumption* ($\|\frac{\mathbb{X}'^T\mathbb{X}'}{n} - I_m\|_\infty \leq \frac{1}{32k}$), *with*

$$\tau = C\sigma\left(\sqrt{\frac{\log m}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Let $\hat{f} = \phi_{\hat{\theta}^L}$ *then with probability* $1 - \delta$

$$\frac{1}{n}\sum_{i=1}^{n}(f(\mathbb{X}_i) - \hat{f}(\mathbb{X}_i))^2 \lesssim \frac{\sigma^2\|\theta^*\|_0\log(m/\delta)}{n}$$

and

$$\text{MSE}(\phi_{\hat{\theta}}) \lesssim \inf_{\theta\in\mathbb{R}^m, \|\theta\|_0\leq k}(\text{MSE}(\phi_\theta) + \frac{\sigma^2\|\theta\|_0\log(m/\delta)}{n})$$

The proof of the above is similar to that of Theorem 3 and is omitted.

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.