MATH888: High-dimensional probability and statistics Fall 2021

Lecture 38 — December 6th, 2021

Sebastien Roch, UW-Madison *Scribe: Joseph Shenouda*

1 Overview

In the last lecture we proved both the weak and fast LASSO rates and introduced the *inchorence* property to prove the fast LASSO rate.

In this lecture we discuss when we can achieve this incoherence property and conclude by introducing a theorem on oracle inequalities, using a dictionary of functions as opposed to a simple linear model.

2 Main Section

Recall the Lasso estimator problem, we assume that our observed data $Y \in \mathbb{R}^n$ is generated according to the following model,

$$Y = \mathbb{X}\theta^* + \epsilon \tag{1}$$

Where $||\theta^*||_0 \leq k$ (i.e. θ^* is sparse, having at most k non-zero terms). Under this assumption the Lasso estimator is $\hat{\theta}^{\mathcal{L}}$ such that,

$$\hat{\theta}^{\mathcal{L}} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n}||Y - \mathbb{X}\theta||_2^2 + 2\tau||\theta||_1 \right\} \tag{2}$$

We showed that if $\mathbb{X}$ satisfies the following,

$$\left| \frac{\mathbb{X}^T\mathbb{X}}{n} - I_p \right|_\infty \leq \frac{1}{32k} \tag{3}$$

then it has the incoherence property denoted as INC(k) and using this INC(k) condition for $\mathbb{X}$ we can prove the following theorem,

Theorem 1 (Fast Rate for Lasso)**.** *Fix $n \geq 2$. Assume the linear model $Y = \mathbb{X}\theta^* + \epsilon$ holds where $||\epsilon||_{\Psi_2} \leq \sigma$. Moreover assume that $||\theta^*||_0 \leq k$ and that $\mathbb{X}$ satisfies INC(k). Then the Lasso estimator with regularization parameter defined by*

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

Then with probability $1 - \delta$

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

We proved this theorem in the last lecture.

Remark: In practice the value for σ is an unknown hyperparameter and can be adjusted using cross validation.

2.1 Validity of Incoherence Assumption

The fast rate for Lasso estimators depends on this incoherence assumption. It is natural to ask if this assumption makes sense for high dimensional problems when $p >> n$. For $p < n$ as $k \to \infty$ the incoherence assumption is equivalent to orthogonality of the matrix $\mathbb{X}$. So here we show that that there exists a matrix that satisfies INC(k) even for $p > n$.

We use a probabilistic method to prove this existence by finding a probability measure which assigns positive probability to objects satisfying this property, implying that there must exist object that satisfy said property. The proof is from Proposition 2.16 [1].

Proposition 2 (2.16)**.** *Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be a random matrix with iid entries X_{ij} taking values in $\{+1, -1\}$ uniformly. Then the incoherence of $\mathbb{X}$ is k with probability $1 - \delta$ as long as*

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

Proof. To show INC(k) we need that,

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

To prove this statement we will show that the diagonals of $\frac{1}{n}[\mathbb{X}^T\mathbb{X}]_{jj}$ are close to 1 and that the off-diagonals $\frac{1}{n}[\mathbb{X}^T\mathbb{X}]_{ij}$ are all close to 0.

We start by showing that the diagonals of $\frac{\mathbb{X}^T\mathbb{X}}{n}$ are close to 0 notice that the jth diagonal term is

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2$$

Since each $X_{ij} \in \{+1, -1\}$ we have that $\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2 = 1$. Therefore,

$$\left|\left|\frac{(\mathbb{X}^T\mathbb{X})_{jj}}{n} - 1\right|\right|_\infty = 0 \leq \frac{1}{32k}$$

For the off-diagonals where $i \neq j$,

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

Since $X_{\ell i}$ and $X_{\ell j}$ are independent and uniform we have that their product is also a uniform random variable taking values $\{+1, -1\}$ and thus it is a bounded random variable so it has subgaussian norm $||X_{\ell i}X_{\ell j}||_{\psi_2} \lesssim 1$

Therefore we just need to show that the probability that the off-diagonals are close to 0 is high,

$$
\begin{aligned}
\mathbb{P}\left(\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I\right\|_\infty > t\right) \\
&= \mathbb{P}\left(\max_{i\neq j}\left|\frac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}X_{\ell j}\right| > t\right) \\
&\leq \sum_{i\neq j}\mathbb{P}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}X_{\ell j}\right| > t\right) \quad \text{(Union Bound)}
\end{aligned}
$$

Now $\frac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}X_{\ell j}$ is a sum of independent subgaussain random variables with mean 0 so we can apply Hoeffding's inequality to get,

$$
\begin{aligned}
\sum_{i\neq j}\mathbb{P}\left(\left|\frac{1}{n}\sum_{\ell=1}^{n} X_{\ell i}X_{\ell j}\right| > t\right) &\leq \sum_{i\neq j} 2\exp\left(-\frac{(nt)^2}{2n}\right) \\
&\leq 2p^2\exp\left(-\frac{(nt)^2}{2n}\right)
\end{aligned}
$$

If we let $t = \frac{1}{32k}$ then,

$$
\mathbb{P}\left(\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I\right\|_\infty > \frac{1}{32k}\right) \leq 2p^2\exp\left(-\frac{(n/32k)^2}{2n}\right)
$$

And,

$$
2p^2\exp\left(-\frac{(n/32k)^2}{2n}\right) \leq \delta
$$

For $n \gtrsim k^2\log(p/\delta)$ □

3 Oracle Inequalities

Sometimes the assumption that your data comes from the linear model $y = \mathbb{X}\theta^* + \epsilon$ is to strong of an assumption.

Instead let's assume the data is generated by some arbitrary function $f(x)$ so that the model becomes,

$$
Y_i = f(X_i) + \epsilon_i \tag{4}
$$

Now consider a dictionary of functions $\mathcal{H} = \{\varphi_1, ..., \varphi_M\}$ where $\varphi_j : \mathbb{R}^p \to \mathbb{R}$. And we want to estimate the unknown f using a linear combination of the functions in the dictionary.

$$f \approx \varphi_\theta := \sum_{j=1}^{M} \theta_j \varphi_j \tag{5}$$

The vector $\varphi_\theta(X_i) \in \mathbb{R}^n$ is defined as,

$$\varphi_\theta(X_i) := \sum_{j=1}^{M} \theta_j \varphi_j(X_i) \tag{6}$$

So again we can use the same MSE to find our LASSO estimator,

$$\hat{\theta^{\mathcal{L}}} \in \arg\min_{\theta \in \mathbb{R}^m} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \varphi_\theta(X_i))^2 + 2\tau ||\theta||_1 \right\} \tag{7}$$

We will present an oracle inequality for the Lasso estimator similar to what we showed earlier. Recall that to prove the fast Lasso estimator rate we needed the incoherence property. For this setting define the matrix $\Phi \in \mathbb{R}^{n \times M}$ with elements $\Phi_{ij} = \varphi_j(X_i)$. The incoherence property is then,

$$INC(k) : \left|\left| \frac{\Phi^T \Phi}{n} - I_m \right|\right|_\infty \leq \frac{1}{32k} \tag{8}$$

Under this condition we have the following theorem from [1]

Theorem 3. *Assume that $||\epsilon||_{\psi_2} \leq \sigma$ and let Φ have INC(k) then if,*

$$\tau = C\sigma \left\{ \sqrt{\frac{\log M}{n}} + \sqrt{\frac{\log 1/\delta}{n}} \right\} \tag{9}$$

and $||\theta^*||_0 \leq k$ then with probability $1 - \delta$.

$$\frac{1}{n} ||\varphi_{\hat{\theta}^{\mathcal{L}}} - f||_2^2 \lesssim \inf_{\theta \in \mathbb{R}^M, ||\theta||_0 \leq k} \left\{ \frac{1}{n} ||\varphi_\theta - f||_2^2 + \frac{\sigma^2}{n} ||\theta||_0 \log(M/\delta) \right\} \tag{10}$$

References

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