36-709: Advanced Statistical Theory I	**Spring 2020**

Lecture 11: February 25

Lecturer: Siva Balakrishnan *Scribe: Motolani Olarinre*

In the previous class, we looked at basis pursuit, where in we aim to obtain a sparse solution to the high dimensional problem

$$\widehat{\theta} = \arg\min \|\theta\|_1 \quad \text{such that} \quad X\theta = y$$

i.e in the noiseless setting, and where X has a non-trivial null space. Suppose the solution θ^* has support S, then θ^* can be recovered using basis pursuit if the Restricted Nullspace (RN) property is satisfies, i.e.

$$\text{null}(X) \cap \{\Delta \in \mathbb{R}^d | \, \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\} = \{0\}$$

Today, we will look at sufficient conditions for the RN property to hold.

11.1 Pairwise Incoherence

We define the pairwise incoherence parameter as the smallest δ_{pw} such that

$$\left\| \frac{X^T X}{n} - \mathrm{I} \right\|_\infty \leq \delta_{pw}(X) \tag{11.1}$$

We claim that if $\delta_{pw}(X) \leq \frac{1}{2s}$, then for all subsets $\{S \subset \{1, 2, ..., d\} \, \text{s.t.} \, |S| \leq s\}$, RN(S) (restricted nullspace property with respect to S) holds.

11.1.1 Example

An example of a matrix that satisfies the pairwise incoherence property (and for which RN holds) is the matrix X s.t. $X_{ij} \sim N(0, 1)$. Then we have the following high probability bound on the error:

$$\left\| \frac{X^T X}{n} - \mathrm{I} \right\|_\infty \leq \sqrt{\frac{\log d}{n}}$$

In this case, if we have number of samples $n \geq s^2 \log d$, then we get the bound $\sqrt{\frac{\log d}{n}} \leq \frac{1}{2s}$, and the matrix X has pairwise incoherence.

11.1.2 Proof: pairwise incoherence implies RN

We can prove that the bound on pairwise incoherence implies that RN(S) holds in the following way: Pick some vector θ such that $X\theta = 0$. For some set S s.t. $|S| \leq s$, we have $\theta = \theta_S + \theta_{S^c}$, and $X\theta_S = -X\theta_{S^c}$. We can lower bound the l2 norm of the right side as:

$$\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^T \frac{X^T X}{n} \theta_S = \theta_S^T \left(\frac{X^T X}{n} - \mathrm{I} \right) \theta_S + \|\theta_S\|_2^2 \\
&\overset{(i)}{\geq} - \left\| \frac{X^T X}{n} - \mathrm{I} \right\|_\infty \cdot \|\theta\|_1^2 + \|\theta_S\|_2^2 \\
&\overset{(ii)}{\geq} -\delta \|\theta\|_1^2 + \|\theta_S\|_2^2 \\
&\overset{(iii)}{\geq} \|\theta_S\|_2^2 (1 - s\delta)
\end{aligned} \tag{11.2}$$

Inequality (i) comes from the inequality: $u^T M v \leq \|M\|_\infty \|u\|_1 \|v\|_1$
Inequality (ii) comes from 11.1 above.
Inequality (iii) comes from the inequality: $\|\theta_S\|_1 \leq \sqrt{s} \|\theta_S\|_2$

Since we know that $X\theta_S = -X\theta_{S^c}$, we can also upper bound the l2 norm of the right side as:

$$\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right| = \left| \theta_S^T \left(\frac{X^T X}{n} - \mathrm{I} \right) \theta_{S^c} + \underbrace{\theta_S^T \theta_{S^c}}_{=0} \right| \\
&\leq \delta \|\theta_S\|_1 \|\theta_{S^c}\|_1 \leq \delta \sqrt{s} \|\theta_S\|_2 \|\theta_{S^c}\|_1
\end{aligned} \tag{11.3}$$

Relating (11.2) to (11.3), we have

$$\|\theta_S\|_2^2 (1 - s\delta) \leq \delta \sqrt{s} \|\theta_S\|_2 \|\theta_{S^c}\|_1$$

$$\|\theta_S\|_2 \leq \frac{\delta \sqrt{s}}{(1 - s\delta)} \|\theta_{S^c}\|_1$$

$$\|\theta_S\|_1 \leq \sqrt{s} \|\theta_S\|_2 \leq \frac{s\delta}{(1 - s\delta)} \|\theta_{S^c}\|_1$$

Assuming $\delta \leq \frac{1}{2s}$, then $\|\theta_S\|_1 \leq \|\theta_{S^c}\|_1$ therefore the restricted nullspace property holds. To summarize: In order to obtain a solution to the problem $X\theta = y$, where n ¡ d, if $n \geq s^2 \log d$, then the pairwise incoherence parameter $\delta_{pw} \leq \frac{1}{2s}$. This implies that the restricted nullspace property holds, and basis pursuit will result in a unique solution.

11.2 Restricted Isometry Property (RIP)

For a given s, if

$$\left\|\frac{X_S^T X_S}{n} - \mathrm{I}\right\|_{op} \leq \delta_{RIP}(X) \tag{11.4}$$

Where X_S is the matrix X with just the columns indexed by set S selected. For all subsets $\{S \subset \{1, 2, ..., d\}$ s.t. $|S| \leq s\}$, then X is said to have restricted isometry.
We claim that if X satisfies the restricted isometry property with parameters $(2s, \frac{1}{3})$, then RN(S) holds $\forall\ |S| \leq s$.

11.2.1 Example

For matrix X s.t. $X_{ij} \sim N(0, 1)$, we fix some subset $\{S \subset \{1, 2, ..., d\}$ s.t. $|S| \leq s\}$. In order to show that this has RIP$(2s, \frac{1}{3})$, we first give a high probability bound on the right hand side of (11.4) for a the fixed S, then union bound to obtain a high probability bound over all subsets S. Using the bounds on max singular value of covariance matrices, we have:

$$\mathbb{P}\left(\left\|\frac{X_S^T X_S}{n} - \mathrm{I}\right\|_{op} \geq \sqrt{\frac{2s}{n}} + \delta\right) \leq 2\exp\left(\frac{-n\delta^2}{2}\right)$$

Note that if $n \geq Cs$, then the bound holds for fixed S w.p. $\approx 1 - \exp(-n)$.
If we pick $\delta \gg \sqrt{\frac{s\log(\frac{ed}{s})}{n}}$, and then apply the union bound to the above probability over all subsets S s.t $|S| \leq s$, then RIP$(2s, \frac{1}{3})$ will hold for $n \gg s\log(\frac{ed}{s})$
The disadvantage of using RIP is that checking that it holds is computationally intractable.

11.3 LASSO

Assuming the linear model

$$y = X\theta^* + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma^2)$$

is correct, and θ^* is s-sparse. We aim to answer the following questions:
1. Can we ensure $\left\|\widehat{\theta} - \theta^*\right\|_2^2$ is small?
2. What conditions on X do we need to ensure 1?
3. What choice of (n, d, s) do we need to ensure 1?

We can formulate the LASSO problem in the following ways:
1. Regularized version

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1$$

2. Constrained version

$$\widehat{\theta} = \arg\min_{\|\theta\|_1 \leq \mathrm{R}} \frac{1}{2n} \|y - X\theta\|_2^2$$

3.

$$\widehat{\theta} = \arg\min_{\frac{1}{2\mathrm{n}}\|\mathrm{y}-\mathrm{X}\theta\|_2^2 \leq \mathrm{t}} \|\theta\|_1$$

Looking at the regularized version of the LASSO, by the basic inequality:

$$\frac{1}{2n} \left\|y - X\widehat{\theta}\right\|_2^2 + \lambda_n \left\|\widehat{\theta}\right\|_1 \leq \frac{1}{2n} \|\epsilon\|_2^2 + \lambda_n \|\theta^*\|_1$$

Setting $\Delta = \widehat{\theta} - \theta^*$, and noting that $y - X\widehat{\theta} = \epsilon - X\Delta$, we can expand the first term in the equality above so we get:

$$\frac{1}{2n} \|X\Delta\|_2^2 - \frac{1}{n}\langle \epsilon, X\Delta \rangle + \lambda_n \left\|\widehat{\theta}\right\|_1 \leq \lambda_n \|\theta^*\|_1$$

The goal is then to bound $\|\Delta\|_2^2$. In low dimension, a bound on $\|X\Delta\|_2^2$ would provide guarantees for $\|\Delta\|_2^2$. This is however not true in high dimensions, because X has a non-trivial null space.