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

Lecture 12: February 27

Lecturer: Siva Balakrishnan *Scribe: Lantian Xu*

12.1 Recap

1. Basis pursuit: $\widehat{\theta} = \arg\min \|\theta\|_1$, want to find $y = X\theta^*$, $X \in R^{n\times d}$, θ^* s-sparse (Thm 7.8 MJW book, etc)
2. Restricted Nullspace Property(RN): $\{\text{null}(X) \cap \{\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}\} = \{0\}$
3. Pairwise Incoherence(PI): $\|\frac{X^TX}{n} - I\|_\infty \leq \delta$
4. Restricted Isometry Property(RIP): For any $|S| \leq s$, $\|\frac{X_S^TX_S}{n} - I_S\|_{op} \leq \delta$
5. If PI holds with $\delta \leq \frac{1}{2s} \Rightarrow$ RN holds
6. If RIP holds with $\delta \leq \frac{1}{3} \Rightarrow$ RN holds
7. Notice that for a matrix $X \in R^{n\times d}$ with $N(0,1)$ entries, PI holds with high probability if $n \gtrsim s^2 \log d$, RIP holds with high probability if $n \gtrsim s\log(\frac{ed}{s})$

12.2 Today

12.2.1 Intuition for RIP

In general we want to solve

$$\begin{aligned} \widehat{\theta} = &\arg\min \|\theta\|_0 \\ &\text{subject to } \; y = X\theta \end{aligned} \tag{12.1}$$

but it is hard to compute. We replace the troublesome L_0 norm by L_1 norm to get $\widehat{\theta}$.
For $\|\widehat{\theta}\|_0 \leq \|\theta^*\|_0 \leq s$, $\|\Delta\|_0 \leq 2s$. In order to argue that θ^* is unique, it suffices to show that $\|X\Delta\|_2^2 > 0$ for any (at most) 2s-sparse vector Δ.

Let us now see that RIP in fact implies this condition. Observe that,

$$\left\|\frac{X\Delta}{\sqrt{n}}\right\|_2^2 = \frac{\Delta^TX^TX\Delta}{n} = \Delta^T(\frac{X^TX}{n} - I)\Delta + \|\Delta\|_2^2 \geq -\delta\|\Delta\|_2^2 + \|\Delta\|_2^2 > 0$$

holds when $\|\frac{X_S^TX_S}{n} - I_S\|_{op} < 1$, $\forall\ |S| \leq 2s$. This shows that an RIP condition is sufficient for the success of the L_0 minimization program.

12.2.2 LASSO

Suppose now we observe a vector $y \in R^n$ and a matrix $X \in R^{n \times d}$ that are linked via the standard linear model $y = X\theta^* + \epsilon$, where θ^* is s-sparse and ϵ is a vector of noise variables whose entries have distribution $N(0, \sigma^2)$. Our goal is to find a parameter estimation error bound $\|\Delta\|_2 := \|\widehat{\theta} - \theta^*\|_2$ between the LASSO solution $\widehat{\theta}$ and the unknown regression vector θ^*.

We have different constrained forms of the LASSO, for example

$$\widehat{\theta} = \arg\min_{\|\beta\|_1 \leq t} \frac{1}{2n}\|y - X\beta\|_2^2 \tag{12.2}$$

or

$$\widehat{\theta} = \arg\min_{\beta} \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{12.3}$$

These convex programs are equivalent due to Lagrangian duality theory.

In low dimension setting, we can obtain a bound without s-sparse condition:

$$\lambda_{\min}(\frac{X^TX}{n})\|\Delta\|_2^2 \leq \frac{\|X\Delta\|_2^2}{2n} \leq \left|\frac{(X^T\epsilon)^T\Delta}{n}\right| \leq \frac{\|X^T\epsilon\|_2\|\Delta\|_2}{n}$$

if $\lambda_{\min}(\frac{X^TX}{n}) > 0$. While in high dimension setting the eigenvalue condition not always holds (for example, when $d > n$ the rank of $\frac{X^TX}{n}$ would be at most n). For this reason, we have to come up with some useful conditions in high dimension.

Definition 12.1 (Restricted Eigenvalue condition) We say a matrix X satisfies the restricted eigenvalue (RE) condition over S with parameters (κ, α) if for any $\|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1$,

$$\left\|\frac{X\Delta}{\sqrt{n}}\right\|_2^2 \geq \kappa\|\Delta\|_2^2$$

In class we discussed $\alpha = 3$ case (as 7.3.2 (A2), MJW book).

Theorem 12.2 (ℓ_2 error between $\widehat{\theta}$(12.2) and θ^*) If θ^* is s-sparse, X satisfies RE condition with parameter (κ,3) and we pick $t = \|\theta^*\|_1$ in (12.2), then

$$\|\widehat{\theta} - \theta^*\|_2 \leq 4\|\frac{X^T\epsilon}{n}\|_\infty \frac{\sqrt{s}}{\kappa} \tag{12.4}$$

For $\|\frac{X^T\epsilon}{n}\|_\infty \lesssim \sqrt{\frac{\log d}{n}}$, $\|\widehat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{s \log d}{n}}$.

Proof: First given $t = \|\theta^*\|_1$ the target vector θ^* is feasible; from RE condition we have

$$\kappa\|\Delta\|_2^2 \leq \left\|\frac{X\Delta}{\sqrt{n}}\right\|_2^2.$$

From $\frac{1}{2n}\|y - X\widehat{\theta}\|_2^2 \leq \frac{1}{2n}\|y - X\theta^*\|_2^2$ and Holder inequality one can derive

$$\frac{\|X\Delta\|_2^2}{n} \leq \left|\frac{2(X^T\epsilon)^T\Delta}{n}\right| \leq 2\|\frac{X^T\epsilon}{n}\|_\infty\|\Delta\|_1$$

On the other hand, following the proof of the analysis of basis pursuit we obtain that $\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1$, under our constraint on θ. This in turn gives $\|\Delta\|_1 \leq 2\|\Delta_S\|_1 \leq 2\sqrt{s}\|\Delta\|_2$; Putting together these pieces yields $\|\widehat{\theta} - \theta^*\|_2 \leq 4\|\frac{X^T\epsilon}{n}\|_\infty\frac{\sqrt{s}}{\kappa}$.

■

Theorem 12.3 (ℓ_2 error between $\widehat{\theta}$(12.3) and θ^*) If θ^* is s-sparse, X satisfies RE condition with parameter (κ,3) and we have a regularization parameter lower bounded as $\lambda \geq 2\|\frac{X^T\epsilon}{n}\|_\infty$, then

$$\|\widehat{\theta} - \theta^*\|_2 \leq \frac{3\lambda\sqrt{s}}{\kappa} \tag{12.5}$$

Proof: Condition (12.3) gives

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

Rearranging yields

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \left|\frac{2(X^T\epsilon)^T\Delta}{n}\right| + \lambda\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\}$$

Rewriting θ^* under s-sparse condition and applying triangle inequality, we have

$$\|\theta^*\|_1 - \|\widehat{\theta}\|_1 = \|\theta_S^*\|_1 - \|\theta_S^* + \Delta_S\|_1 - \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1 - \|\Delta_{S^c}\|_1$$

Plugging in above inequality yields

$$\frac{1}{2n}\|X\Delta\|_2^2 \leq \frac{\lambda}{2}\{3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1\} \tag{12.6}$$

since $\lambda \geq 2\|\frac{X^T\epsilon}{n}\|_\infty$. Here we are in a situation to apply the RE condition: $\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2$. Thus from the inequality in Theorem 12.2,

$$\kappa\|\Delta\|_2^2 \leq \frac{3\lambda}{2}\|\Delta_S\|_1 \leq \frac{3\lambda\sqrt{s}}{2}\|\Delta\|_2 \tag{12.7}$$

as desired.

■

12.2.3 In-sample Prediction Error

We want to come up with some methods which do not need strong assumptions to predict well. In general, we want to get the bound

$$\frac{\|X\Delta\|_2^2}{n} \leq C\lambda\|\theta^*\|_1$$

without assumptions on X or θ^*(remains for later discussion).

On the other hand, a simple inspection of the above proof for the ℓ_2 error shows that under RE and if θ^* is s-sparse we obtain the faster rates for the in-sample prediction error of:

$$\frac{\|X\Delta\|_2^2}{n} \leq \frac{Cs\log p}{n},$$

when λ in the Lagrangian LASSO is chosen as $\lambda \geq 2\|\frac{X^T\epsilon}{n}\|_\infty$.

In more detail, from (12.6) we obtain the bound:

$$\frac{1}{n}\|X\Delta\|_2^2 \leq 3\lambda\|\Delta_S\|_1 \leq 3\lambda\sqrt{s}\|\Delta\|_2.$$

Combining this with the RE condition, to obtain that,

$$\frac{1}{n}\|X\Delta\|_2^2 \leq \frac{3\lambda\sqrt{s}\|X\Delta\|_2}{\sqrt{n\kappa}}.$$

Re-arranging this we obtain that,

$$\frac{\|X\Delta\|_2^2}{n} \leq \frac{Cs\lambda^2}{\kappa},$$

as desired.

□