36-710: Advanced Statistical Theory **Fall 2018**

Lecture 15: October 22

Lecturer: Alessandro Rinaldo *Scribe: Theresa Gebert*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Recall that last time we proved the slow rate for the LASSO in mean estimation. Today we will derive conditions under which the LASSO *does* achieve best subset selection. As we will see, this is only true under very strong assumptions.

One key (but strong) assumption is a condition on the eigenvalues of X^TX. This is known as the RE (restricted eigenvalue) condition. In the first part of this lecture, we will use this condition to prove the so-called *fast rate* of the LASSO. In the second part of this lecture, we will look into "sparsistency," which shows LASSO's ability to recover the true support of β^*. Next week, we will cover so-called Oracle Inequalities.

15.1 Fast Rate of the LASSO

As usual, let $S \subseteq \{1, \dots, d\}$. This is the notation we will use for a particular class of vectors in $\mathbb{R}^d$:

$$C_\alpha(S) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}.$$

In English, $C_\alpha(S)$ is the set of vectors where the values of coefficients in S are sufficiently larger than the coefficients not in S. You could imagine that if the coefficients in vs. not in S are not "distinguishable" enough from each other, it might be too difficult for us to recover the coefficients in S.

Definition 15.1 *$X \in \mathbb{R}^{d \times n}$ satisfies the $\mathsf{RE}(\alpha, k)$ condition for $k > 0$ when*

$$\frac{1}{n}\|X\Delta\|^2 \leq k\|\Delta\|^2$$

for all $\Delta \in C_\alpha(S)$.

Note that this requirement is quite artificial. Now we will see a theorem that uses this definition to prove a fast convergence rate for the LASSO. The first three requirements are very familiar to us: they are all we needed to assume for the slow rate of the LASSO in the last lecture. The fifth assumption is what is so strong, and necessary, for a faster rate on the errors $\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|^2$ and $\|\hat{\beta} - \beta^*\|$. (It is worth noting that this theorem implicitly focuses on $d > n$.)

Theorem 15.2 *Assume that:*

1. *$Y = X\beta^* + \epsilon$;*
2. *X is fixed;*
3. *$\epsilon \in \mathsf{SG}(\sigma^2)$;*

4. $S = support(\beta^*)$, $|S| = s$;
5. X *satisfies the* $\mathsf{RE}(3, k)$ *condition for some* $k > 0$.

Then, if

$$\lambda_n \geq 2\frac{\|X^T\epsilon\|_\infty}{n}$$

any LASSO solution $\hat{\beta}$ *satisfies*

1. $\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|^2 \leq 9\lambda_n^2 \frac{s}{k}$;
2. $\|\hat{\beta} - \beta^*\| \leq \frac{3}{k}\sqrt{s}\lambda_n$.

Note two things: (1) as Heejong pointed out, the last condition is always true if X^TX is positive semidefinite; (2) that there is no way to check that condition without checking over all possible S, at which point you may as well just do brute-force subset selection.

Remark 15.3 *If we plug in* $\lambda \sim \sigma\sqrt{\frac{\log d}{n} + \log(1/\delta)}$ *then we get essentially the same guarantees as best subset selection* ($s = \|\beta^*\|_0$).

Now we will prove the theorem. Our strategy is to (1) show that for the choice of λ_n given in the theorem, $\hat{\Delta} = \hat{\beta} - \beta^* \in C_3(S)$ (by deriving a basic inequality and using Hölder's Inequality and the Triangle Inequality); (2) use the RE condition and simply rearrange. Note that this proof technique is the same for other norms (not just for LASSO).

Proof: We first show that, for this choice of λ_n, $\hat{\Delta} = \hat{\beta} - \beta^* \in C_3(S)$. We get the basic inequality from the fact that $\hat{\beta}$ minimizes the LASSO equation:

$$\frac{1}{2n}\|Y - X\hat{\beta}\|^2 + \lambda_n\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^*\|^2 + \lambda_n\|\beta^*\|_1.$$

Plugging in the fact that $Y = X\beta^* + \epsilon$ and rearranging, we get:

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \frac{\epsilon^T X\hat{\Delta}}{n} + \lambda_n[\|\beta^*\|_1 - \|\hat{\beta}\|_1]. \tag{15.1}$$

Note that by the way we have defined $\hat{\Delta}$ and the definition of S, we have that:

$$\|\beta^*\|_1 - \|\hat{\beta}\|_1 = \|\beta^*_S\|_1 - \|\beta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\beta}_{S^c}\|_1 \tag{15.2}$$

since $\|\beta^*_S\|_1 + \|\hat{\beta}_{S^c}\|_1 = \|\beta^*_S\|_1 - \|\beta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1$, but $\|\hat{\beta}_{S^c} - \beta^*_{S^c}\|_1 = \|\hat{\beta}_{S^c}\|_1$ because $\beta^*_{S^c} = 0$.

So plugging in 15.1 to 15.2 we get:

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \frac{\epsilon^T X\hat{\Delta}}{n} + \lambda_n[\|\beta^*_S\|_1 - \|\beta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1]$$

and using Hölder's Inequality on the first term of the right-hand side (and moving the 2 to the other side):

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|^2 \leq 2\frac{\|X^T\epsilon\|_\infty\|\hat{\Delta}\|_1}{n} + 2\lambda_n[\|\beta^*_S\|_1 - \|\beta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1]$$

and using the Triangle Inequality on $\|\beta_S^* - \hat{\Delta}_S\|_1 \geq \|\beta_S^*\|_1 - \|\hat{\Delta}_S\|_1$:

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|^2 \leq 2\frac{\|X^T\epsilon\|_\infty\|\hat{\Delta}\|_1}{n} + 2\lambda_n[\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1]$$

and finally using the fact that $\lambda_n \leq 2\|X^T\epsilon\|_\infty\|\Delta\|_1/n$:

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|^2 \leq \lambda_n\|\hat{\Delta}\|_1 + 2\lambda_n[\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1].$$

Split $\|\hat{\Delta}\|_1$ into $\|\hat{\Delta}_S\|_1$ and $\|\hat{\Delta}_{S^c}\|_1$:

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|^2 \leq \lambda_n\|\hat{\Delta}_S\|_1 + \lambda_n\|\hat{\Delta}_{S^c}\|_1 + 2\lambda_n[\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1]$$

and combine the right-hand side to yield

$$0 \leq \lambda_n[3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1]$$

which implies that $\|\hat{\Delta}_{S^c}\|_1 \leq 3\|\hat{\Delta}_S\|_1$, which means that $\hat{\Delta} = \hat{\beta} - \beta^* \in C_3(S)$. Now, notice that

$$\lambda_n[3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1] \leq 3\lambda_n\|\hat{\Delta}_S\|_1 \leq 3\lambda_n\hat{\Delta}\|_1 \leq 3\lambda_n\sqrt{s}\|\hat{\Delta}_S\|_2 \leq 3\lambda_n\sqrt{s}\|\hat{\Delta}\|_2$$

because $\forall\ x \in \mathbb{R}^d, \|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\|x\|_2$. Now we use the $\mathsf{RE}(3,k)$ condition to show:

$$\frac{1}{n}\|X\hat{\Delta}\|^2 \leq 3\lambda_n\sqrt{s}\frac{\|X\hat{\Delta}\|}{\sqrt{n}}\frac{1}{\sqrt{k}}$$

dividing out the right-hand side and squaring yields:

$$\frac{1}{n}\|X\hat{\Delta}\|^2 \leq (3\lambda_n\sqrt{s/k})^2$$

which implies that

$$\frac{\|X\hat{\Delta}\|^2}{n} \leq 9\lambda_n^2\frac{s}{k}$$

as claimed.

Similarly for estimating β^* by the $\mathsf{RE}(3,k)$ condition:

$$\|\hat{\Delta}\| = \|\hat{\beta} - \beta^*\| \leq \frac{1}{\sqrt{n}}\|X\hat{\Delta}\|\frac{1}{\sqrt{k}} \leq 3\lambda_n\sqrt{s/k}\frac{1}{\sqrt{k}} = 3\lambda_n\sqrt{s}/k.$$

■

15.2 Best Subset Selection

We can do model selection in high dimensions as well; again, with very strong assumptions. Notice that nothing we have done so far shows that the LASSO recovers the *true* support. The result we will discuss next is due to [1] (which was at first rejected from the Annals of Statistics, then accepted three years later in IEEE).

Let $S = \text{support}(\beta^*)$. Ideally we want to recover S, the true support. See Theorem 7.21 in [2].

Theorem 15.4 *Assume that*

1. $\lambda_{min}\frac{X_S^T X_S}{n} \geq C_{min} > 0$;
2. $\max_J \|X_J\| \leq \sqrt{n}$;
3. $\max_{J \in S^c} |X_J^T X_S (X_S^T X_S)^{-1} \mathsf{sgn}(\beta^*)| < 1$;
4. $\|X_{S^c}^T X_S (X_S^T X_S)^{-1}\|_\infty < 1 - \gamma$ *for* $\gamma \in (0, 1)$,

then

1. *support*$(\hat{\beta}) \subseteq S$;
2. $\|\hat{\beta}_S - \beta_S^*\| \leq \lambda_n g_n(\lambda_n)$, *where*

$$g_n(\lambda_n) = \left\| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right\|_\infty + \frac{4\sigma}{\sqrt{c_{min}}}.$$

The first two assumptions in this theorem are not strong, and the third one is known as the β_{min} condition. It is just a technicality. The fourth condition is called the *incoherence condition*, and this is very strong. It cannot be verified in practice.

If we want to achieve *sparsistency*, then

$$\min_{j \in S} |\beta_j^*| > g_n(\lambda_n) \implies \text{support}(\hat{\beta}) = S.$$

This is not really useful in practice and requires very strong assumptions. The incoherence condition is essentially necessary to gain this result. In addition, many papers you read on this assume X is fixed, when in practice we obviously want a random design.

Next week, we will be covering Oracle Inequalities.

References

[1] M. J. Wainwright. "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso)." *IEEE Transactions on Information Theory*. pp. 2183–2202. May 2009.

[2] M. J. Wainwright. "High-Dimensional Statistics: A Non-Asymptotic Viewpoint." *Cambridge University Press*. Feb 2019.