Lecture 11: October 9 - Slow rate for the LASSO. The RE condition.

Lecturer: Alessandro Rinaldo *Scribes: Kwangho Kim*

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

This lecture's notes illustrate some uses of various $\LaTeX$ macros. Take a look at this and imitate.

1.1 Penalized Least Squares

1.1.1 Penelization

For linear regression problem we defined last time

$$Y = X\beta^* + \epsilon$$

- If $Rank(X^TX) = d$, how can we estimate $X\beta^*$ and $\beta*$ when d grows with n?
- If $Rank(X) = n$, then you can fit the data perfectly.

More generally people have modeled that

$$\underset{\beta}{argmin}\frac{1}{2n}||Y - X\beta||^2 + \lambda_n f(\beta)$$

, where λ_n is a tuning parameter and $f(\beta)$ is a complexity panelty.

For example, we have Ridge regression when $f(\beta) = ||\beta||_2^2$. Ridge regression yields "dense solution" where all the coefficients of $\hat{\beta}$ are nonzero.

1.1.2 Model Selection Property

Assume cardinality of $S = \{i : \beta_i^* \neq 0\}$ is small compared to d. We want to estimate S when $|S| \ll d$.

In this caes, least square solutions are dense. So what if your solution is also very sparse and $\hat{S} = \{i : \hat{\beta}_i^* \neq 0\}$ is close to S?

One way to do this is with best subset selection:

$$\underset{\beta}{argmin}\frac{1}{2n}||Y - X\beta||^2 + \lambda_n||\beta||_0$$

However, this requires fitting $\sum_{j=1}^d \binom{d}{j}$ least squares, each of which requires matrix inversion. So this will be computationally infeasible.

Thus, people have come up with a compromise : LASSO as below

$$\underset{\beta}{argmin} \frac{1}{2n}||Y - X\beta||^2 + \lambda_n f||\beta||_1$$

LASSO has the following properties:

- Unlike ridge regression, solution can be sparse depending on λ_n
- This is still convex problem
- Fast algorithm exists - e.g. lars, glmnet, etc.
- Solution is unique if columns of X are in general position.
- * Lasso does not have a model selection property *unless* we put a strong assumption. Hence we usually cannot make inference on β while still we use it as a good tool for prediction.

1.1.3 Slow Rate for Lasso

Theorem 1.1 *If $\lambda_n \geq \frac{1}{n}||X^T\epsilon||_\infty$, then for the Lasso solution β we have*

$$\frac{1}{n}||X(\hat{\beta} - \beta^*)||^2 \leq 4||\beta^*||_1\lambda_n$$

Proof: Start with basic inequality,

$$\frac{1}{n}||X(\hat{\beta} - \beta^*)||^2 \leq \frac{\epsilon^T X(\hat{\beta} - \beta^*)}{n} + \lambda_n(||\beta^*||_1 - ||\hat{\beta}||_1)$$

which comes from the last lecture. From this inequality, we can proceed as

$$\begin{aligned}
\frac{1}{n}||X(\hat{\beta} - \beta^*)||^2 &\leq \frac{\epsilon^T X(\hat{\beta} - \beta^*)}{n} + \lambda_n(||\beta^*||_1 - ||\hat{\beta}||_1) \\
&\leq \frac{||\epsilon^T X||_\infty}{n}||\hat{\beta} - \beta^*||_1 + \lambda_n(||\beta^*||_1 - ||\hat{\beta}||_1) \quad \text{(Holder)} \\
&\leq \Big(\frac{||X^T\epsilon||_\infty}{n} - \lambda_n\Big)||\hat{\beta}||_1 + \Big(\frac{||X^T\epsilon||_\infty}{n} + \lambda_n\Big)||\beta^*||_1 \quad \text{(Triangle inequality)} \\
&\leq \Big(\frac{||X^T\epsilon||_\infty}{n} + \lambda_n\Big)||\beta^*||_1 \quad \text{(By assumption)} \\
&\leq 2\lambda_n||\hat{\beta}||_1 \quad \text{(By assumption)}
\end{aligned}$$

■

Now, when is $\lambda_n \geq \frac{1}{n}||X^T\epsilon||_\infty$ true? If $\epsilon \in SG_n(\sigma^2)$ and $max_j||X_j|| \leq \sqrt{Cn}$ so that all the covariance have roughly the same order, then we have

$$\begin{aligned}
P\Big(\max_j \frac{1}{n}||X_j^T\epsilon||_\infty \geq t\Big) &\leq \sum_j P\Big(|X_j^T\epsilon| \geq tn\Big) \\
&= \sum_j P\Big(\frac{|X_j^T\epsilon|}{||X_j||} \geq \frac{tn}{||X_j||}\Big) \\
&\leq 2d\exp -\frac{t^n}{2\sigma^2 C} \quad \text{(Hoffding)}
\end{aligned}$$

with probability $1-\delta$. Hence the choice of λ_n should be the one satisfying $\lambda_n \leq \sqrt{\frac{2\sigma^2 C}{n}(\log 1/\delta + \log d)}$. With this chice of λ_n, we finally have the following lamma.

Lemma 1.2 *With above choice of λ_n, with high probability of $1-\frac{1}{n^c}$ for some $c>0$, we have*

$$\frac{1}{n}||X(\hat{\beta}-\beta^*)||^2 \lesssim ||\beta^*||_1 \sigma \sqrt{\frac{\log n + \log d}{n}}$$

Compare the result of lemma 1.2 to what we can obtain from the best subset selection method:

$$\frac{1}{n}||X(\hat{\beta}-\beta^*)||^2 \lesssim ||\beta^*||_0 \sigma^2 \frac{\log n + \log d}{n}.$$

From above, we can deduce that why lemma 1.2 is called "slow rate".

1.1.4 Fast Rate for Lasso

To obtain faster rate, we need additional stronger assumption on $\frac{X^T X}{n}$.

Definition ($Re(\alpha,\kappa)$ condition) Design matrix X satisfies the restricted eigenvalue condition with parameters $\alpha > 1$ and $\kappa > 0$ and $S \in \{1,...,d\}$, if

$$\frac{1}{n}||X\Delta||^2 \geq \kappa ||\Delta||^2 \forall \Delta \in C_\alpha(s) := \{\Delta \in \mathbb{R}^d : ||\Delta_{S^c}||_1 \leq \alpha ||\Delta_S||_1\}$$

where Δ_{S^c} indicates a coordinate of Δ outside S.

intuition. Let $\Delta = \hat{\beta}-\beta^*$. Then we know that $\frac{1}{n}||X(\hat{\beta}-\beta^*)||^2 = \frac{1}{n}||X\Delta||^2$ can be small even if $||\Delta||^2$ is large. Because the function $\Delta \to \frac{1}{n}||X\Delta||^2$ may be flat at $\hat{\Delta}$.

To prevent this, we would need that

$$\frac{1}{n}||X\Delta||^2 \geq ||\hat{\Delta}||_k^2.$$

This holds if $k = \lambda_{min}(\frac{X^X}{n})$. But this does not happen in general. If $d > n$, $Re(\alpha,\kappa)$ requires this behavior along all possible directions Δ, but only along directions in $C_\alpha(s)$.