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

Lecture 13: October 15

Lecturer: Alessandro Rinaldo *Scribes: Leqi Liu*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

13.1 Continue on Least Squares Regression

We start by finishing the proof of the prediction risk of least squares regression.

Theorem 13.1 *Let $\epsilon \in SG_n(\sigma^2)$ be a vector of independent sub-gaussian random variables. Then, for some universal constant $c > 0$,*

$$\mathbb{P}\left(\frac{1}{n}||X(\hat{\beta}-\beta^*)||^2 \geq c\sigma^2\left(\frac{r+\log(1/\delta)}{n}\right)\right) \leq \delta$$

for all $\delta \in (0,1)$ and $r = rank(X^TX)$.

Proof: In the last lecture, using the basic inequality and *sup-out* trick, we get that

$$||X(\hat{\beta}-\beta^*)||^2 \leq 4 \sup_{x\in\mathbb{S}^{r-1}} (x^T\tilde{\epsilon})^2,$$

which implies that

$$\mathbb{E}[||X(\hat{\beta}-\beta^*)||^2] \leq 4\sigma^2 r.$$

To get a high probability bound, we first note that $\forall t > 0$,

$$\mathbb{P}\left(4\sup_{x\in\mathbb{S}^{r-1}}(x^T\tilde{\epsilon})^2 \geq t\right) = \mathbb{P}\left(2\max_{z\in\mathcal{N}_{\frac{1}{2}}} z^T\tilde{\epsilon} \geq \sqrt{t}\right) \leq |\mathcal{N}_{\frac{1}{2}}|e^{\frac{-t}{8\sigma^2}}$$

where $\mathcal{N}_{\frac{1}{2}}$ is a $\frac{1}{2}$−cover of $\mathbb{S}^{r-1}$ and we know that $|\mathcal{N}_{\frac{1}{2}}| \leq 5^r$ from the previous lecture. The last inequality is done by union bound and the fact that $\tilde{\epsilon}$ is sub-gaussian. To get the desired bound, we just need to set the RHS to be δ. ■

Remark: If $\frac{X^TX}{n}$ has rank d ($d < n$), then $||\hat{\beta}-\beta^*||^2 \leq c\sigma^2\frac{1}{\lambda_{\min}(\frac{X^TX}{n})}\frac{r+\log(\frac{1}{\delta})}{n}$ where $\frac{1}{\lambda_{\min}(\frac{X^TX}{n})}$ depends on d and is the extra price we need pay for estimation. We can see that inference on the mean, i.e. $X\beta^*$ is much easier than inference on the parameter, i.e. β^*.

13.2 Penalized Least Squares Regression

We will now move to penalized least squares regression. Our assumption stays the same as before: $Y = X\beta^* + \epsilon$. The penalized least squares regression is:

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^d} ||Y - X\beta||^2 + \lambda f(\beta),$$

where $\lambda \geq 0$ is a tuning parameter, $f : \mathbb{R}^d \to \mathbb{R}_+$ is the penalty/regularization term.

13.2.1 Ridge Regression

Ridge regression is defined for penalized least squares regression with $f(\beta) = ||\beta||^2$. By taking the derivative of the objective function and set it to zero, we get a unique minimizer:

$$\widehat{\beta}_{\text{Ridge}} = (XX^T + \lambda I_d)^{-1} X^T Y.$$

Note that as we have proved in Assignment 3, since $XX^T + \lambda I_d$ is positive-definite, $(XX^T + \lambda I_d)^{-1}$ always exists.

Intuition of $\widehat{\beta}_{\text{Ridge}}$ Recall that $X\widehat{\beta}_{\text{OLS}} = \sum_{j=1}^{r} u_j u_j^T Y$ where $u_1, \cdots u_r$ are the orthonormal vectors in $\mathbb{R}^n$ forming the columns of U in the singular vector decomposition of X, i.e. $X = U\Lambda V^T$ and

$$\Lambda = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \dots & \sigma_r \end{bmatrix}$$

with $\sigma_j > 0, \forall j = 1, \cdots, r$. For ridge regression, $X\widehat{\beta}_{\text{Ridge}} = \sum_{j=1}^{r} u_j u_j^T Y \frac{\sigma_j^2}{\lambda + \sigma_j^2}$. We see that the predictions $X\widehat{\beta}_{\text{Ridge}}$ is still a projection on the column space of X yet puts more weight on direction where $X^T X$ has more "variance" and downweigh directions with smaller singular values.

"Drawbacks" of $\widehat{\beta}_{\text{Ridge}}$ $\widehat{\beta}_{\text{Ridge}}$ is dense, i.e. $\forall j = 1, \cdots, d$, $\widehat{\beta}_{\text{Ridge}}(j) \neq 0$. In high-dimensional settings $(d > n)$, ridge regression doesn't make a lot of sense since we assume that $\text{supp}(\beta^*) = \{j : \beta^*(j) \neq 0\}$ is small compared to d. Under the sparsity regime, we would like to estimate β^* with an estimator $\hat{\beta}$ that is sparse: $\text{supp}(\hat{\beta})$ is small. In fact, what we would really like is $\text{supp}(\hat{\beta}) \approx \text{supp}(\beta^*)$, which is similar to performing model selection.

13.2.2 Thresholding

Then, what would be an ideal $f(\beta)$ to get a sparse solution?

13.2.2.1 Best Subset Selection

When $f(\beta) = ||\beta||_0$, the estimator $\widehat{\beta}_{l_0}$ is the best subset selection estimator. Computing $\widehat{\beta}_{l_0}$ is really hard: it requires computing $\sum_{j=1}^{d} \binom{d}{j}$ OLS, which requires inverting a matrix at each time. However, it has good theoretical properties: with probability at least $1 - \delta$, $\delta \in (0, 1)$, when choosing $\lambda \asymp \sigma^2 \frac{\log d}{n}$,

$$\frac{1}{n} ||X(\widehat{\beta}_{l_0} - \beta^*)||^2 \leq C ||\beta^*||_0 \sigma^2 \frac{\log(\frac{ed}{\delta})}{n}.$$

13.2.2.2 LASSO

A natural relaxation to the above is setting $f(\beta) = ||\beta||_1$ instead.

Advantages of LASSO LASSO is a convex problem, so we can solve it efficiently. At the same time, it has some "model selection properties," since it produces sparse solutions. Note that LASSO is equivalent to the problem $\min_{\beta \in \mathbb{R}^d} ||\beta||_1$ such that $||Y - X\beta||^2 \leq b$, $b > 0$.

Theorem 13.2 *If $\lambda_n > \frac{|X^T \epsilon|_\infty}{n}$, then all LASSO solutions satisfy that $\frac{1}{n}||X(\widehat{\beta}_{LASSO} - \beta^*)||^2 \leq 4\lambda_n ||\beta^*||_1$.*

This is the so-called slow rate of LASSO since the RHS depends on n instead of $\sqrt{n}$. Next lecture, we will see how, by making additional assumptions, we can recover the fast rate.

13.2.2.3 Comparison among the three penalized least squares regression

Assume that X is an orthogonal matrix, i.e. $X^T X = I$. We have the following results: $X\widehat{\beta}_{\text{OLS}} = Y$, $X\widehat{\beta}_{\text{Ridge}} = Y/(1+\lambda)$, $X\widehat{\beta}_{\text{LASSO}} = S_{\frac{\lambda}{2}}(Y)$ and $X\widehat{\beta}_{l_0} = H_{\sqrt{\lambda}}(Y)$ where soft-thresholding is

$$S_{\frac{\lambda}{2}}(x) = \begin{cases} x - \frac{\lambda}{2} & \text{if } x \geq \frac{\lambda}{2} \\ 0 & \text{if } |x| < \frac{\lambda}{2} \\ x + \frac{\lambda}{2} & \text{if } x \leq -\frac{\lambda}{2} \end{cases}$$

and hard-thresholding is

$$H_{\sqrt{\lambda}}(x) = \begin{cases} x & \text{if } x \geq |\sqrt{\lambda}| \\ 0 & \text{if } x < |\sqrt{\lambda}| \end{cases}$$

Below (Figure 13.1) is a visualization of the mean estimation/prediction of the three shrinkage estimators (ridge, lasso, l_0) compared with the mean estimation of OLS (dotted line y=x):

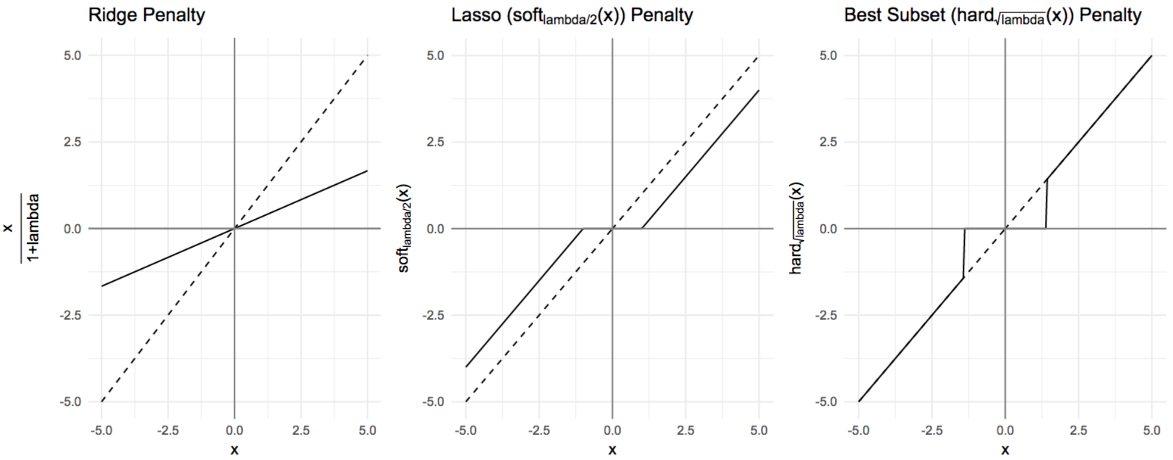

Figure 13.1: This is borrowed from the scribe of Lecture 13 done by Benjamin LeRoy in Fall 2017.