36-755: Advanced Statistics **Fall 2017**

Lecture 13: Wednesday, October 11

Lecturer: Alessandro Rinaldo *Scribe: Benjamin LeRoy*

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

13.1 Visualizing Shrinkage

Recall that for Ridge regression:

$$\begin{aligned}\hat{\beta}_{ridge} &= \text{argmin}_{\beta\in\mathbb{R}^d}||Y - X\beta||^2 + \lambda||\beta||^2, \quad \lambda \geq 0 \\ &= (X^TX + \lambda I_d)^{-1}X^TY\end{aligned}$$

To motivate approaching ridge regression in terms of spectral decomposion observe the following about Ordinary Least Squares when $r = rank(X) \leq \min\{n, d\}$.

We can decompose $X = U\Lambda V^T$ where Λ diagonal, with r non-zero values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$. We'll express $U = [u_1, ..., u_d]$. Then:

$$\begin{aligned}X\hat{\beta}_{ols} &= X(X^TX)^+X^T \\ &= \sum_{i=1}^{R} u_i u_i^T Y\end{aligned}$$

Back to ridge regression we have:

$$X\hat{\beta}_{ridge} = X(X^TX + \lambda\,\mathbb{I}_d)X^TY$$

Plugg in in SVD of X and noticing that

$$X^TX + \lambda\,\mathbb{I}_d \quad = \quad V\Lambda^2V^T + \lambda\,\mathbb{I}_d \quad = \quad V(\Lambda^2 + \mathbb{I}_d)V^T$$

we have that

$$\begin{aligned}X\hat{\beta}_{ridge} &= U\Lambda V^TV(\Lambda^2 + \lambda\,\mathbb{I})^{-1}V^TV\Lambda U^TY \\ &= U\Lambda(\Lambda^2 + \lambda\,\mathbb{I})^{-1}\Lambda U^TY && \text{def of V (orthonormal structure gets } V^TV = \mathbb{I}_d) \\ &= UHU^TY\end{aligned}$$

where $H = \begin{bmatrix} \frac{\sigma_1^2}{\sigma_1^2+\lambda} & & & & 0 \\ & \ddots & & & \\ & & \frac{\sigma_r^2}{\sigma_r^2+\lambda} & & \\ & & & 0 & \\ & & & & \ddots \\ 0 & & & & 0 \end{bmatrix}$.

Which means we can express

$$X\hat{\beta}_{ridge} = \sum_{i=1}^{r} u_i \frac{\sigma_i^2}{\sigma_i^2 + \lambda} u_i^T Y$$

This can be thought of as a weighted projection onto the PC directions of X with a shrinkage by λ, especially comparing to $X\hat{\beta}_{ols} = \sum_{i=1}^{r} u_i u_i^T Y$.

To really think about the shrinkage seen in ridge regression (and Lasso and best subset selection) we focus on the basic case where $Y \sim (\mu, \sigma^2 \mathbb{I}) \in \mathbb{R}^d$ with the goal of estimating μ. Under these assumptions we observe:

$\hat{\mu}$	argmin representation	reduction	comment
$\hat{\mu}_{mle}$	$= Y$		
$\hat{\mu}_{ridge}$	$= \text{argmin}_{\mu \in \mathbb{R}^d} \|\|Y - \mu\|\|^2 + \lambda \|\|\mu\|\|^2$	$= \frac{Y}{1+\lambda}$	shrinks $\to 0$
$\hat{\mu}_{lasso}$	$= \text{argmin}_{\mu \in \mathbb{R}^d} \|\|Y - \mu\|\|^2 + \lambda \|\|\mu\|\|_1$	$= \text{soft}_{\lambda/2}(Y)$	where $\text{soft}_{\lambda/2}(Y) = \begin{cases} x - \lambda/2 & x > \lambda/2 \\ 0 & \lvert x \rvert \leq \lambda/2 \\ x + \lambda/2 & x < -\lambda/2 \end{cases}$
$\hat{\mu}_{\text{best subset}}$	$= \text{argmin}_{\mu \in \mathbb{R}^d} \|\|Y - \mu\|\|^2 + \lambda \|\|\mu\|\|_0$	$= \text{soft}_{\lambda/2}(Y)$	where $\text{hard}_{\sqrt{\lambda}}(Y) = \begin{cases} x & \lvert x \rvert > \sqrt{\lambda} \\ 0 & \lvert x \rvert < \sqrt{\lambda} \end{cases}$

Figure 13.1 provides a visual of each of these shrinkage functions compared to the OLS function ($y = x$).

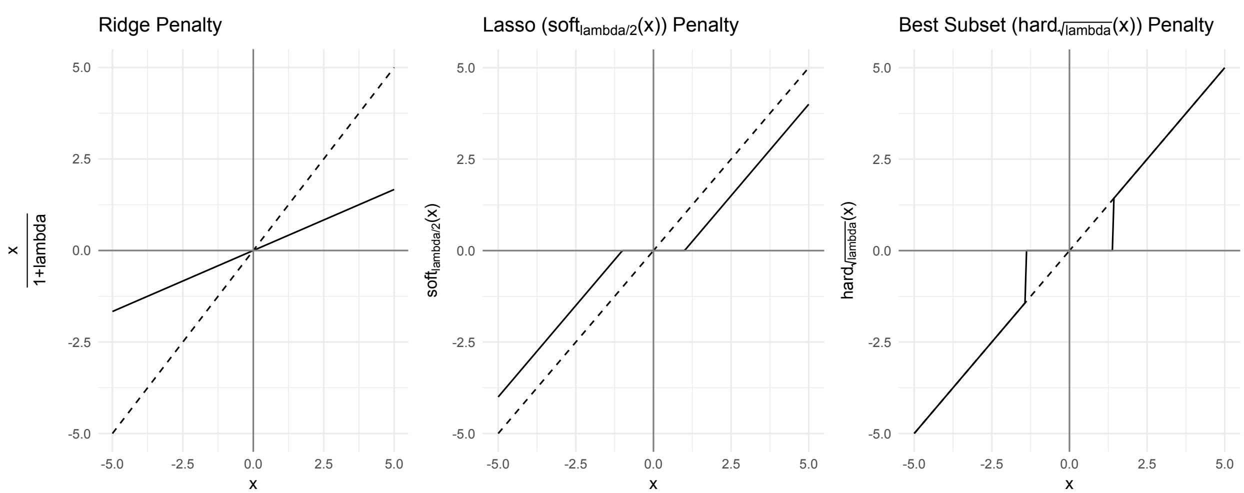

Figure 13.1: Different shrinkage lines for the basic case

When X is orthogonal in the standard regression case and $\mu = X\beta$ then we have:

$$\hat{\mu}_{ridge} = \frac{X^T Y}{1 + \lambda} = \frac{\hat{\beta}_{ols}}{1 + \lambda} \qquad \hat{\mu}_{lasso} = \text{soft}_{\lambda/2}(\hat{\beta}_{ols}) \qquad \hat{\mu}_{\text{best subset}} = \text{hard}_{\sqrt{\lambda}}(\hat{\beta}_{ols})$$

13.2 Fast Rates for Lasso

13.2.1 Reminders

In the last lecture we saw that Lasso could give us slow rates (compared to the best subset selection) if $\lambda_n \geq \frac{||X^T \epsilon||_\infty}{n}$. Specifically that if the constraint on λ_n held then for $c > 0$:

$$\frac{1}{2n}||X(\hat{\beta}_{lass} - \beta^*||^2 \leq 4||\beta^*||_1 \lambda \qquad \text{with prob } \geq 1 - \frac{1}{n^c}$$

Where, with assumptions of sub-Gaussian noise and bounded covariates, we saw this had order $o\left(\sigma\sqrt{\frac{\log n + \log d}{n}}\right)$.

This was slower than with the best subset selection where, as a reminder:

$$\hat{\beta}_{\text{best subset}} = \text{argmin}_{\beta \in \mathbb{R}^d} \frac{1}{2n}||Y - X\beta||^2 + \lambda||\beta||_0$$

and which yields preformance of order $||\beta^*||_0 \frac{\sigma^2}{n}(\log d + \log n)$.

Additionally, recall that if $\lambda_{min}\left(\frac{X^T X}{n}\right) > c$ then $\frac{1}{2}||\hat{\beta}_{lasso} - \beta^* * ||^2 \leq \frac{4}{c}||\beta^*||_1 \lambda_n$.

13.2.2 Fast Rates for Lasso

In order to get fast rates for the Lasso, we need the restrcted eigenvalue (RE) condition defined as below.

Definition 13.1 *X satifies the $RE(\alpha, \kappa)$ condition with $\alpha > 1$, $\kappa > 0$ for some $S \subseteq \{1, ..., d\}$ if*

$$\frac{1}{n}||X\Delta||^2 \geq \kappa||\Delta||^2 \qquad \textit{for all} \quad \Delta \in C_\alpha(S) = \{x \in \mathbb{R}^d : ||x_{S^C}||_1 \leq \alpha||x_S||_1\}$$

Aside: The constraint $\frac{1}{n}||X\Delta||^2 \geq \kappa||\Delta||^2$ can be though of as forcing $||X\Delta||^2$ to have at least some amount of curvature on the S dimensions. One can think about this like bounding the derivative in those dimensions from below as Δ is the difference between $\hat{\beta}$ and β.

Theorem 13.2 *Assuming the following conditions:*

1) $Y = X\beta^* + \epsilon \qquad \epsilon \in SG(\sigma^2)$ *independent*
2) *supp*$(\beta^*) = \{i : \beta_i \neq 0\} = S$
3) *X satifies the $RE(3, \kappa)$ conditional with respect to S*

then if $\lambda_n \geq \frac{2||X^T \epsilon||_\infty}{n}$ we have that

$$\frac{1}{n}||X\hat{\Delta}||^2 \leq 9\lambda_n^2 \frac{|S|}{\kappa} \quad \textit{and} \quad ||\hat{\Delta}|| \leq 3\sqrt{|S|}\frac{\lambda_n}{\kappa}$$

Aside: This is the same preformance as best subset selection noting that $|S| = ||\beta^*||_0$.

Proof: Under these assumptions we first show that $\hat{\Delta} \in C_3(S)$: Recall the basic inequality (via Δ inequality and expansion) we obtained in the last lecture:

$$0 \leq \frac{1}{2n}||X\hat{\Delta}||^2 \leq \frac{\epsilon^T X\hat{\Delta}}{n} + \lambda_n\Big(||\beta^*||_1 - ||\hat{\beta}||_1\Big) \tag{13.1}$$

Since β^* is S-sparse and recalling that $\hat{\Delta} = \hat{\beta} - \beta^*$, then

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

From this we can observe that

$$\begin{aligned} \frac{1}{n}||X\hat{\Delta}||^2 \leq \frac{2}{n}&||X^T\epsilon||_\infty||\hat{\Delta}||_1 && \textbf{(i)} \\ &+ 2\lambda\Big(||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^C}||_1\Big) && \textbf{(ii)} \end{aligned} \tag{13.2}$$

Where **(i)** comes from Holder's inequality and **(ii)** comes a subsitution into equation 13.1 from the triangle inequality giving

$$\begin{aligned} ||\beta^*_S||_1 &\leq ||\hat{\Delta}_S||_1 + ||\beta^*_S + \hat{\Delta}_S|| \\ \Leftrightarrow ||\hat{\Delta}_S||_1 &\geq ||\beta^*_S||_1 - ||\beta^*_S + \hat{\Delta}_S|| \end{aligned}$$

Using the fact that $\frac{2||X^T\epsilon||_\infty}{n} \leq \lambda_n$ from the theorem assumptions, we have that we can constrain $\frac{1}{n}||X\hat{\Delta}||^2$ in equation 13.2 by

$$\begin{aligned} &\leq \lambda||\hat{\Delta}_S||_1 + \lambda_n||\hat{\Delta}_{S^C}||_1 + 2\lambda_n(||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^C}||_1) \\ &\leq \lambda_n(3||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^C}||_1) \end{aligned}$$

This implies that $\hat{\Delta} \in C_3(S)$, so we can use the fact that $\frac{||X\hat{\Delta}||^2}{n} \geq ||\hat{\Delta}||^2\kappa$.

So, we can constraint $\frac{1}{n}||X\hat{\Delta}||^2$ in the following way,

$$\begin{aligned} \frac{1}{n}||X\hat{\Delta}||^2 &\leq \lambda_n(3||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^C}||_1) \\ &\leq 3\lambda_n(||\hat{\Delta}||_1) \\ &\leq 3\lambda_n\sqrt{|S|}||\hat{\Delta}_S||_2 && \text{because } x \in \mathbb{R}^d\text{: } ||x||_2 \leq ||x||_1 \leq \sqrt{d}||x||_2 \\ &\leq 3\lambda_n\sqrt{|S|}||\hat{\Delta}||_2 \\ &\leq 3\lambda_n\sqrt{|S|}\frac{||X\hat{\Delta}||}{\sqrt{n}} && \text{from the RE condition we showed first.} \end{aligned}$$

Observing that both sides of the equation has a multiple of $\frac{||X\hat{\Delta}||}{\sqrt{n}}$ we can obtain:

$$\begin{aligned} \frac{1}{\sqrt{n}}||X\hat{\Delta}|| &\leq 3\lambda_n\sqrt{\frac{|S|}{\kappa}} \\ \frac{1}{n}||X\hat{\Delta}||^2 &\leq 9\lambda_n^2\frac{|S|}{\kappa} \end{aligned}$$

which gives us the first part of the conclusion. Additionally, from the RE condition we have that $\frac{||X\hat{\Delta}||^2}{n} \geq ||\hat{\Delta}||^2\kappa$ which leads to

$$||\hat{\Delta}||\sqrt{\kappa} \leq \frac{||X\hat{\Delta}||}{\sqrt{n}} \leq 3\lambda_n\sqrt{\frac{|S|}{\kappa}}$$

Which provides us with the fact that $||\hat{\Delta}|| = ||\hat{\beta}_{lasso} - \beta^*|| \leq 3\lambda_n\sqrt{\frac{|S|}{\kappa}}$.

In order to obtain the fast rate we need with high probability we have a λ_n such that $\lambda_n \geq \frac{2||X^T\epsilon||_\infty}{n}$. If the columns of X are normalized so that they have norm $O(\sqrt{n})$ then you can take $\lambda_n \asymp \sigma\sqrt{\frac{\log n + \log d}{n}}$ and the assumption will hold with probability $\geq 1 - \frac{1}{n^c}$.

■