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

Lecture 14: October 17

Lecturer: Alessandro Rinaldo *Scribes: Maria Jahja*

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

14.1 Lasso

For the model $Y = X\beta^\star + \varepsilon$, where $\varepsilon \in SG_n(\sigma^2)$, the Lasso solution is

$$\hat{\beta} := \hat{\beta}_{\text{Lasso}} \in \arg\min_{\beta \in \mathbb{R}^d} \frac{1}{2n}\|Y - X\beta\|^2 + \lambda_n \|\beta\|_1$$

where $\lambda_n > 0$ is the regularization parameter chosen by the user. The penalty forces the solution to the least-squares problem to have relatively small ℓ_1-norm to promote sparse solutions.

Theorem 14.1 *Let A be the event $\{\lambda_n \geq n^{-1}\|X^T\varepsilon\|_\infty\}$. If A holds, then*

$$\frac{\|X(\hat{\beta} - \beta^\star)\|^2}{n} \leq 4\|\beta^\star\|_1 \lambda_n \tag{14.1}$$

Proof: [Theorem 14.1].
First, we establish the basic inequality

$$\frac{1}{2n}\|X(\hat{\beta} - \beta^\star)\|^2 \leq \frac{\varepsilon^T X(\hat{\beta} - \beta^\star)}{n} + \lambda_n(\|\beta^\star\|_1 - \|\hat{\beta}\|_1). \tag{14.2}$$

This follows from the simple fact that

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

From the linearity of the true model, we can plug in $X\beta^\star + \epsilon$ and simplify to get (14.2).

Now we bound $\frac{1}{2n}\|X(\hat{\beta} - \beta^\star)\|^2$.

$$\begin{aligned}
\frac{1}{2n}\|X(\hat{\beta} - \beta^\star)\|^2 &\leq \frac{\varepsilon^T X(\hat{\beta} - \beta^\star)}{n} + \lambda_n(\|\beta^\star\|_1 - \|\hat{\beta}\|_1) \\
\text{(via Hölder's inequality)} \qquad &\leq \frac{\|X^T\varepsilon\|_\infty \|\hat{\beta} - \beta^\star\|_1}{n} + \lambda_n(\|\beta^\star\|_1 - \|\hat{\beta}\|_1) \\
\text{(Triangle inequality)} \qquad &\leq \frac{\|X^T\varepsilon\|_\infty}{n}(\|\hat{\beta}\|_1 + \|\beta^\star\|_1) + \lambda_n(\|\beta^\star\|_1 - \|\hat{\beta}\|_1) \\
&= \|\hat{\beta}\|_1 \left(\frac{\|X^T\varepsilon\|_\infty}{n} - \lambda_n\right) + \|\beta^\star\|_1 \left(\frac{\|X^T\varepsilon\|_\infty}{n} + \lambda_n\right) \\
&\leq 2\lambda_n\|\beta^\star\|_1
\end{aligned}$$

where the last inequality follows since $\frac{\|X^T\varepsilon\|_\infty}{n} - \lambda_n < 0$ by assumption. ■

There are several things to note in Theorem 14.1. First, the bound depends on the user-chosen parameter λ_n (which itself shows dependency on the sample size n). Second, the framing of the problem does not explicitly write probabilities. Implicitly, we see that A is a random quantity, and we can only learn something if $\mathbb{P}(A)$ tends to one.

14.1.0.1 Choices of λ_n

It is clear that we would like λ_n to be as small as possible, but we need the assumption to still hold. Because we do not observe the true ε, we cannot solve for λ_n. But when can we get A to occur? We will need to add an extra assumption.

Assume $\max_j \|X_n\| \leq \sqrt{Cn}$ where X_j is the jth column of the design matrix X, and we have some $C > 0$. Recall that $\varepsilon \in SG(\sigma^2)$. Then for all $t \geq 0$

$$\begin{aligned}
\mathbb{P}\left(\frac{\|X^T\varepsilon\|_\infty}{n} \geq t\right) &= \mathbb{P}\left(\frac{\max_j |X_j^T\varepsilon|}{n} \geq t\right) \\
\text{(Union bound)} \qquad &\leq \sum_j \mathbb{P}\left(\frac{|H_j^T\varepsilon|}{n} \geq t\right) \\
&= \sum_j \mathbb{P}\left(\frac{|H_j^T\varepsilon|}{n||X_j||_2} \geq \frac{t}{||X_j||_2}\right) \\
\text{(Subgaussianity, } \|X_j\| \leq \sqrt{Cn}) \qquad &\leq 2d\exp\left(-\frac{t^2 n}{2\sigma^2 C}\right) \\
&= \exp\left(-\frac{t^2 n}{2\sigma^2 C} + \log 2d\right) \\
&\leq \delta < 1
\end{aligned}$$

where the last inequality chooses

$$t = \sqrt{\frac{2\sigma^2 C}{n}(\log 1/\delta + \log 2d)}.$$

.

Note that in practice, we do not have σ^2 (otherwise we might solve for t). This result gives us the following corollary.

Corollary 14.2 (Slow rate for the Lasso) *With probability* $1 - \delta$,

$$\frac{\|X(\hat{\beta} - \beta^\star)\|^2}{n} \leq 4\|\beta^\star\|_1 \sigma \sqrt{\frac{c}{n}(\log 1/\delta + \log 2d)}$$

Why do we call this a "slow" rate? Recall best subset selection with the ℓ_0 penalty. We can derive a bound of the order

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

From Corollary 14.2, if we pick $\delta = 1/n$, then

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

with probability $1 - 1/n$ by choosing t accordingly. The square root here is not ideal. However, this is still a good result—the Lasso solution is sparse and works well in high dimensions ($d > n$).

14.1.1 Fast rate for the Lasso

Can we be more computationally efficient, and still do as well as the optimal? Notice if $\lambda_{\min}(X^T X) \geq \gamma_n > 0$, then

$$\|\hat{\beta} - \beta^\star\| \leq \frac{1}{\gamma_n}\|X(\hat{\beta} - \beta^\star)\|^2.$$

If $d > n$, then $\lambda_{\min}(X^T X) = 0$. We see that when $X^T X$ is badly conditioned, the solution is unstable and small perturbations turn into large changes.

To get away from the slow rate, we need to add additional constraints on X. This can be formalized in several ways, but the basic underlying idea is the same: get γ_n away from zero.

One of the more milder ways of controlling γ_n is the *restricted eigenvalue condition* on $X^T X/n$.

Definition 14.3 (Restricted Eigenvalue (RE) condition)

For some $\alpha \geq 1$, and subset $\mathcal{S} \subseteq \{1, \ldots, d\}$, $\mathcal{S} \neq \emptyset$, let

$$\mathcal{C}_\alpha(\mathcal{S}) = \{\Delta \in \mathbb{R}^d : \|\Delta_{\mathcal{S}^\complement}\|_1 \leq \alpha\|\Delta_{\mathcal{S}}\|_1\},$$

where $\mathcal{S}^\complement = \{1, \ldots, d\}\backslash\mathcal{S}$ and $\Delta_{\mathcal{S}} = \{\Delta_j,\ j \in \mathcal{S}\}$.

We say that an $n \times d$ matrix X satisfies the RE(α, κ) condition w.r.t. $\mathcal{S}$ if

$$\frac{1}{n}\|X\Delta\|^2 \geq \kappa\|\Delta\|^2 \qquad \forall \Delta \in \mathcal{C}_\alpha(\mathcal{S})$$

where $\kappa > 0$.

Intuition If we take the vector $\Delta = \hat{\beta} - \beta^\star$, we have shown that $\|X\Delta\|^2/n$ can be small. But from this we cannot conclude that $\|\Delta\|^2$ is small since the function

$$\Delta \to \frac{1}{n}\|X\Delta\|^2 \tag{14.3}$$

may be very flat around $\hat{\beta} - \beta^\star$, i.e. in the unfavorable setting where the loss function is flat around its minimizer $\beta^\star$, it is not necessarily true that a small loss difference implies a small error (for a visual example, see Wainwright Figure 7-5).

If we have that $\lambda_{\min}(X^T X)/n$ is bounded away from zero, say in fact that we have

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

where $0 < \kappa \leq \lambda_{\min}(X^T X/n)$, then we know we have enough curvature. We only need this to be true for certain Δ, specifically we require the function (14.3) to be curved along $\Delta \in \mathcal{C}_\alpha(\mathcal{S})$, where $\mathcal{S}$ is the support of $\beta^\star$.

Context of the RE condition The RE condition was first developed in the field of compressed sensing, and was later adapted to Lasso. We noted the RE condition is one of the milder restrictions; there exist stronger criteria (also historically-rooted in compressed sensing) that we could use to derive fast rates.

One such example is the pairwise incoherence condition

$$\max_{j,k\in\{1,\ldots,d\}} \left| \frac{x_j x_k}{n} - \mathbf{1}(j \neq k) \right| \leq \frac{1}{C_s} \qquad \text{for } C > 0, n \in \mathbb{N}_+$$

which implies the RE condition. This condition is more interpretable: it says that we cannot have overly correlated columns of X.

A natural generalization of the pairwise incoherence condition to larger subsets of columns is the restricted isometry property, where for some $\delta \in (0, 1)$

$$\left\| \frac{X_S^T X_S}{n} - I_k \right\|_{\text{op}} \leq \delta_k$$

for all S of size k, where X_S is a sub-matrix of X with columns in S. This is equivalent to stating that all the eigenvalue of $X_S^T X_S$ fall into the interval $[1 - \delta_k, 1 + \delta_k]$, and has the same idea of restricting correlation between columns.

For futher details, see Wainwright 7.2.3, which describes the relationship between resticted nullspace, restricted eigenvalue, pairwise incoherence, and restricted isometry properties.

Next lecture, we will continue discussion on fast rates of convergence for Lasso.