**36–755 Advanced Statistical Theory** | **Carnegie Mellon University**
**Fall 2016** | **Department of Statistics**

# Lecture 9: September 28th, 2016

*Lecturer: Alessandro Rinaldo* <arinaldo@cmu.edu> *Scribe: Neil Spencer*



## 9.1 Lasso

Consider a regression framework where $Y$ is an $n \times 1$ vector, $X$ is a $n \times d$ matrix, $\theta^*$ is a $d \times 1$ vector, and $\epsilon$ is a $n \times 1$ vector. Further assume that

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

with $\epsilon \in \mathrm{SG}_n(\sigma^2)$. In the LASSO, we use estimate $\hat{\theta}$ to estimate $\theta$, where $\hat{\theta}$ us the solution to

$$\hat{\theta} \in \text{argmin}_{\theta \in \mathbb{R}^d} \left( \frac{1}{2n} ||Y - X\theta||^2 + \lambda_n ||\theta||_1 \right). \tag{9.1}$$

Equation (1) defines a convex optimization problem that produces sparse solutions depending on $\lambda_n$. The parameter $\lambda_n$ is chosen by the user. It can be thought of as $\lambda(n, d, \sigma)$ because the choice will depend on those values.

Equation (1) has solutions for both $d \leq n$ and $d > n$. There can be multiple optimal solution $\hat{\theta}$, but the maximizing value $X\hat{\theta}$ is unique. For a discussion of the uniqueness of solutions to the Lasso problem, see [1].

The basic inequality [2] is a useful inequality for proving results pertaining to the Lasso. It is given below as Lemma 1.1. It is used to prove Theorem 1.2.

**Lemma 9.1.** *In the Lasso set-up, if $\theta^*$ is the true parameter value and $\hat{\theta}$ is the lasso solution, then*

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

*Proof.*

$$\begin{aligned}
\frac{1}{2n} \left( ||\epsilon||^2 + \left|\left| X(\hat{\theta} - \theta^*) \right|\right|^2 - 2\epsilon^T X(\hat{\theta} - \theta^*) \right) + \lambda_n ||\hat{\theta}||_1 &= \frac{1}{2n} \left|\left| \epsilon - X(\hat{\theta} - \theta^*) \right|\right|^2 + \lambda_n ||\hat{\theta}||_1 \\
&= \frac{1}{2n} \left|\left| Y - X\hat{\theta} \right|\right|^2 + \lambda_n ||\hat{\theta}||_1 \\
&\leq \frac{1}{2n} ||Y - X\theta^*||^2 + \lambda_n ||\theta^*||_1 \\
&= \frac{1}{2n} ||\epsilon||^2 + \lambda_n ||\theta^*||_1
\end{aligned}$$

□

**Theorem 9.2.** *If*

$$\lambda_n \geq \left|\left|\frac{X^T\epsilon}{n}\right|\right|_\infty = max_{J=1,\ldots,d}\left|\frac{X_J^T\epsilon}{n}\right|,$$

*then any Lasso solution satisfies*

$$\left|\left|\frac{X(\hat{\theta}-\theta^*)}{n}\right|\right|^2 \leq 4\,||\theta^*||_1\,\lambda_n.$$

*Proof.* Lemma 1.1 provides

$$\begin{aligned}
\frac{1}{2n}\left|\left|X(\hat{\theta}-\theta^*)\right|\right|^2 &\leq \epsilon^T\frac{X(\hat{\theta}-\theta^*)}{n} + \lambda_n(||\theta^*||_1 - ||\hat{\theta}||_1) \text{ by Basic inequality (Lemma 1.1)}\\
&\leq \frac{1}{n}\left|\left|X^T\epsilon\right|\right|_\infty \left|\left|\hat{\theta}-\theta^*\right|\right|_1 + \lambda_n(||\theta^*||_1 - ||\hat{\theta}||_1) \text{ by Holder Inequality}\\
&\leq \frac{1}{n}\left|\left|X^T\epsilon\right|\right|_\infty \left(||\theta^*||_1 + ||\hat{\theta}||_1\right) + \lambda_n(||\theta^*||_1 - ||\hat{\theta}||_1) \text{ by Triangle Inequality}\\
&\leq \left(\frac{1}{n}\left|\left|X^T\epsilon\right|\right|_\infty - \lambda_n\right)||\hat{\theta}||_1 + \left(\frac{1}{n}\left|\left|X^T\epsilon\right|\right|_\infty + \lambda_n\right)||\hat{\theta}||_1\\
&\leq 2\lambda_n||\theta^*||_1.
\end{aligned}$$

□

**What is a good choice for $\lambda_n$?**
Recall that $\epsilon \in \text{SG}(\sigma^2)$ and assume $\max_{J=1\ldots D}||X_J|| \leq C\sqrt{n}$ for some $C > 0$. Then for $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|\left|\frac{\epsilon^T X}{n}\right|\right|_\infty \geq t\right) &\leq \mathbb{P}\left(\max_J|X_J^T\epsilon| \geq tn\right)\\
&\leq \sum_J \mathbb{P}\left(|X_J^T\epsilon| \geq tn\right)\\
&\leq \sum_J \mathbb{P}\left(\frac{|X_J^T\epsilon|}{||X_J||} \geq \frac{tn}{||X_J||}\right)\\
&\leq 2d\exp\left(\frac{-t^2 n}{2\sigma^2\max||X_J||^2}\right) \text{ by Subgaussianity}\\
&\leq 2d\exp\left(\frac{-t^2 n}{2\sigma^2 C^2}\right) \text{ because } ||X_J||^2 < C^2 n\\
&\leq \delta
\end{aligned}$$

if we choose

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

Consider $\delta = 1/n$. Then with probability $1 - 1/n$,

$$\frac{\left|\left|X(\hat{\theta}-\theta^*)\right|\right|^2}{n} \leq ||\theta^*||_1\sqrt{\frac{2\sigma^2 C\left(\log(n) + \log(2d)\right)}{n}}$$

If $d \le n$ and $\lambda_{\min}(X^T X/n) \ge C_{\min} > 0$. Then you can also get a bound for

$$\left|\left|\hat{\theta} - \theta^*\right|\right|^2 \le \frac{||\theta^*||}{c_{\min}\lambda_n}$$

## 9.2 Getting Fast Rates for Lasso

In order to get "fast" rates for the lasso, there needs to be additional assumptions on $X$. These assumptions also provide consistency of estimation of $\theta$.

A very useful condition is the *restricted eigenvalue condition*. In order to define the condition, we need to establish some notation. For $S \subset \{1, 2, \ldots, d\}$ and $\alpha > 0$, define

$$C(\alpha, S) = \left\{\Delta \in \mathbb{R}^d : ||\Delta_{S^c}||_1 \le \alpha \, ||\Delta_S||_1\right\}.$$

**Definition 9.3.** *$X$ satisfies the restricted eigenvalue (RE($\alpha, \kappa$)) condition over $S = \{1, \ldots, d\} \neq \emptyset$ if*

$$\frac{1}{n} ||X\Delta||^2 \ge \kappa \, ||\Delta||^2 \text{ for all } \Delta \in C(\alpha, S).$$

For intuition, think of $\Delta$ as $\hat{\theta} - \theta^*$. We want $||X\Delta||^2/n$ to be small. Note that if it is, this does necessarily mean that $||\Delta||^2$ is small. Especially if

$$\Delta \to \frac{||X\Delta||^2}{n} \tag{9.2}$$

is flat around $\hat{\theta} - \theta^*$. To prevent this, we need the function (9.2) to be very curved. This is true if

$$\frac{||X\Delta||^2}{n} \ge \kappa \, ||\Delta||^2 \text{ for all } \Delta \in \mathbb{R}^d.$$

Unfortunately, this implies that $\lambda_{\min}(X^T X) \ge C_{\min} > 0$ if $d > n$, which is not possible. Instead, we consider the case where function (9.2) is curved only along certain directions.

These directions are $C(\alpha, S)$ where $S$ is defined by the support of $\theta^*$. That is, $s = \{J : \theta^*_J \neq 0\}$.

**Theorem 9.4.** *Assume that*

- *the support of $\theta^*$ is S where $|S| = s > 0$.*
- *$X$ satisfies RE($3, \kappa$) where $\kappa > 0$ with respect to $S$.*
- *$\lambda_n \ge 2\left|\left|\epsilon^T X\right|\right|/n$.*

*Then any Lasso solution $\hat{\theta}$ satisfies*

$$\frac{1}{n}\left|\left|X\left(\hat{\theta} - \theta^*\right)\right|\right|^2 \le 9\lambda_n^2 s\kappa,$$

*and*

$$\left|\left|\hat{\theta} - \theta^*\right|\right| \le \frac{3}{\kappa}\sqrt{s}\lambda_n.$$

*Proof.* First, we show that given our choice $\lambda_n$, $\hat{\Delta} = (\hat{\theta} - \theta^*) \in C(3, S)$. By the Basic inequality,

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

Since $\theta^*$ is $S$-sparse, we know

$$\begin{aligned} ||\theta^*||_1 - ||\hat{\theta}||_1 &= ||\theta^*_S||_1 - \left|\left| \theta^*_S + \hat{\Delta}_S \right|\right|_1 - \left|\left| \hat{\theta}_{S^c} \right|\right|_1 \\ &= ||\theta^*_S||_1 - \left|\left| \theta^*_S + \hat{\Delta}_S \right|\right|_1 - \left|\left| \hat{\Delta}_{S^c} \right|\right|_1. \end{aligned}$$

Plugging this into the Basic inequality yields:

$$\begin{aligned} 0 &\leq \frac{1}{2n} \left|\left| X\hat{\Delta} \right|\right|^2 \\ &\leq 2\frac{\left|\left| X^T \epsilon \right|\right|_\infty}{n} ||\hat{\Delta}||_1 + 2\lambda_n \left( ||\theta^*_S||_1 - ||\theta^*_S + \hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1 \right) \\ &\leq 2\frac{\left|\left| X^T \epsilon \right|\right|_\infty}{n} ||\hat{\Delta}||_1 + 2\lambda_n \left( ||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1 \right) \text{ By triangle inequality} \\ &\leq \lambda_n ||\hat{\Delta}_S||_1 + \lambda)n ||\hat{\Delta}_{S^c}||_1 + 2\lambda_n ||\hat{\Delta}_S||_1 - 2\lambda_n ||\hat{\Delta}_{S^c}||_1 \\ &= \lambda_n \left( 3||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1 \right) \\ &\Rightarrow \hat{\Delta} \in C(3, S). \end{aligned}$$

Note that the fourth line used the fact

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

□

# References

[1] Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[2] Sara A Van de Geer. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.