36-755: Advanced Statistical Theory 1 **Fall 2016**

Lecture 10: October 5

Lecturer: Alessandro Rinaldo *Scribes: Arun Sai Suggala*

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

10.1 LASSO

10.1.1 Fast rates for LASSO

In the last class we derived *slow rates* for LASSO under minimal assumptions. Now we will derive fast rates for LASSO under Restricted Eigenvalue(RE) condition.

Theorem 10.1 *Assume that*

- $supp(\theta^*) = S$ *where* $|S| = s > 0$.
- X *satisfies RE*$(3, \kappa)$ *with respect to* S, *where* $\kappa > 0$ *is a constant.*
- $\lambda_n \geq \frac{2}{n}\|X^T\epsilon\|_\infty$.

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

$$\frac{1}{n}\|X\left(\hat{\theta} - \theta^*\right)\|_2^2 \leq 9\lambda_n^2 \frac{s}{\kappa},$$

and

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n.$$

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

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{\epsilon^T X\Delta}{n} + \lambda_n[\|\theta^*\|_1 - ||\hat{\theta}||_1].$$

Since θ^* is S-sparse, we know

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

Plugging this into the Basic inequality yields:

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|^2 \tag{10.1}$$
$$\leq 2\frac{\|X^T\epsilon\|_\infty}{n}||\hat{\Delta}||_1 + 2\lambda_n\left(||\theta^*_S||_1 - ||\theta^*_S + \hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1\right) \tag{10.2}$$
$$\leq 2\frac{\|X^T\epsilon\|_\infty}{n}||\hat{\Delta}||_1 + 2\lambda_n\left(||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1\right) \quad \text{(by triangle inequality)} \tag{10.3}$$
$$\leq \lambda_n||\hat{\Delta}_S||_1 + \lambda_n||\hat{\Delta}_{S^c}||_1 + 2\lambda_n||\hat{\Delta}_S||_1 - 2\lambda_n||\hat{\Delta}_{S^c}||_1 \tag{10.4}$$
$$= \lambda_n\left(3||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1\right) \tag{10.5}$$
$$\Rightarrow \hat{\Delta} \in C(3, S). \tag{10.6}$$

We now prove the first claim. Note that:

$$\|\hat{\Delta}_S\|_1 \leq \sqrt{s}\|\hat{\Delta}_S\|_2 \leq \sqrt{s}\|\hat{\Delta}\|_2.$$

We now use $RE(3,\kappa)$ condition to get the following inequalities:

$$\implies \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq 3\lambda_n\|\hat{\Delta}_S\|_1 \leq 3\sqrt{s}\lambda_n\|\hat{\Delta}\|_2 \leq 3\lambda_n\frac{\sqrt{s}}{\sqrt{\kappa}}\frac{1}{\sqrt{n}}\|X\hat{\Delta}\|_2$$

$$\implies \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq 9\lambda_n^2\frac{s}{\kappa}.$$

To prove the second claim, we again use $RE(3,\kappa)$ condition on the above inequality.

$$\kappa\|\hat{\Delta}\|_2^2 \leq \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq 9\lambda_n^2\frac{s}{\kappa}$$

$$\implies \|\hat{\Delta}\|_2 \leq 3\lambda_n\frac{\sqrt{s}}{\kappa}.$$

■

10.1.2 Model Selection Property of LASSO

Lets now look at the model selection properties of LASSO. The following theorem shows that LASSO can recover the true signed support under appropriate conditions. Please refer to [WM09] for a proof of the theorem.

Theorem 10.2 *Assume that the true parameter θ^* satisifes $supp(\theta^*) = S$ where $|S| = s > 0$, and the design matrix $X \in \mathbb{R}^{n\times p}$ satisfies the following conditions:*

- $\max_{i\in[p]}\|X_i\|_2 \lesssim \sqrt{n}$
- $\lambda_{min}(\frac{1}{n}X_S^TX_S) \geq C_{min} > 0$
- *(Irrepresentability)* $\max_{i\in S^c}\|X_i^TX_S(X_S^TX_S)^{-1}\|_1 \leq \alpha < 1$

Then for $\lambda_n \gtrsim \|\frac{1}{n}X^T\epsilon\|_\infty$, the following statements hold with high probability:

- *LASSO has a unique solution.*
- *(ℓ_∞ convergence rate)* $\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \lambda_n \left[\| \left(\frac{1}{n} X_S^T X_S\right)^{-1} \|_\infty + \frac{4\sigma}{\sqrt{C_{min}}} \right]$.
- *(Sign Consistency) If* $\min_{i \in S} |\theta_i^*| > \lambda_n \left[\| \left(\frac{1}{n} X_S^T X_S\right)^{-1} \|_\infty + \frac{4\sigma}{\sqrt{C_{min}}} \right]$, *then* $\hat{\theta}$ *has the correct signed support i.e., supp*$(\hat{\theta}) =$ *supp*(θ^*) *and sgn*$(\hat{\theta}) =$ *sgn*(θ^*).

10.1.3 Oracle Inequalities

Until now, we assumed that the data $\{X_i, Y_i\}_{i=1}^n$ is generated from a linear model. But what if this assumption is violated? What can we say about the estimator in this case? When can we say our estimator is good? We introduce *oracle inequalities* as a means to measure the performance of an estimator if the model is misspecified.

Consider the following model for (X, Y), $X \in \mathbb{R}^d$, $Y \in \mathbb{R}$:

$$Y = f(X) + \epsilon,$$

where $f : \mathbb{R}^d \to \mathbb{R}$, $\epsilon \sim SG(\sigma^2)$. Assume a dictionary of functions $\{f_1, f_2, \dots f_M\}$ from $\mathbb{R}^d$ to $\mathbb{R}$. Given $n-$observations $\{X_i, Y_i\}_{i=1}^n$ we wish to estimate f using a linear combination of the functions in dictionary:

$$f_\theta(X) = \sum_{i=1}^M \theta_i f_i(X), \qquad \text{where } (\theta_1, \theta_2, \dots \theta_M) \in \mathbb{R}^M.$$

Note that f need not be in span$\{f_1, f_2, \dots f_M\}$.

Remark: We can recover linear regression if $M = d$ and $f_i(X) = X^{(i)}$ where $X^{(i)}$ is the i^{th} coordinate of X.

Definition 10.3 *For any estimator* $\hat{f}(X)$ *of* $f(X)$ *based on* $\{X_i, Y_i\}_{i=1}^n$, *define its risk as:*

$$R(\hat{f}) = \mathbb{E}[MSE(\hat{f})] = \frac{1}{n}\mathbb{E}[\sum_{i=1}^n (\hat{f}(X_i) - f(X_i))^2].$$

Definition 10.4 *Let* $K \subseteq \mathbb{R}^M$. *The oracle on* K *is the function* f_{θ^*} *such that:*

$$R(f_{\theta^*}) \leq R(f_\theta) \quad \forall \theta \in K.$$

We want to do as well as the oracle risk $R(f_{\theta^*})$. An estimator satisfies *oracle inequality* in expectation if

$$R(\hat{f}) \leq cR(f_\theta^*) + \Phi(n, M, f).$$

where $c \geq 1$. We ideally want c to be close to 1 and $\Phi \to 0$.

Theorem 10.5 ***(Oracle inequality for least squares)*** *Let* $\epsilon_1, \dots \epsilon_n \in SG_n(\sigma^2)$. *Then the following holds with probability at least* $1 - \delta$*:*

$$MSE(f_{\hat{\theta}_{LS}}) \leq \inf_{\theta \in \mathbb{R}^M} MSE(f_\theta) + c\sigma^2 \frac{M}{n} \log \frac{1}{\delta}.$$

where $\hat{\theta}_{LS}$ *is the least squares estimator with design matrix* $\Phi_{n \times M}$ *where* $\Phi_{ij} = f_j(X_i)$.

Proof: Let $\theta^* = \arg\min_{\theta \in \mathbb{R}^M} MSE(f_\theta)$. With a slight abuse of notation, let $f_\theta = [f_\theta(X_1),\ f_\theta(X_2),\ \dots f_\theta(X_n)]^T$. Since $\hat{\theta}_{LS}$ is the least squares estimator, we have:

$$\|Y - f_{\hat{\theta}_{LS}}\|_2^2 \leq \|Y - f_{\theta^*}\|_2^2.$$

Subsituting $Y = f + \epsilon$ in the above equation we get:

$$\|f - f_{\hat{\theta}_{LS}}\|_2^2 - \|f - f_{\theta^*}\|_2^2 \leq 2\epsilon^T (f_{\hat{\theta}_{LS}} - f_{\theta^*}).$$

Also note that f_{θ^*} is the orthogonal projection of f on span$\{f_1, f_2, \dots f_M\}$. So

$$f - f_{\theta^*} \perp f_{\theta^*}, f_{\hat{\theta}_{LS}}$$

Using this in the above inequality we get:

$$\|f_{\hat{\theta}_{LS}} - f_{\theta^*}\|_2^2 \leq 2\epsilon^T (f_{\hat{\theta}_{LS}} - f_{\theta^*}).$$

$$\implies \|f_{\hat{\theta}_{LS}} - f_{\theta^*}\|_2 \leq 2 \left\langle \epsilon, \frac{(f_{\hat{\theta}_{LS}} - f_{\theta^*})}{\|f_{\hat{\theta}_{LS}} - f_{\theta^*}\|_2} \right\rangle$$

Note that $(f_{\hat{\theta}_{LS}} - f_{\theta^*})$ lies in the column span of Φ which is an $M-$dimensional space. Let $\tilde{\Phi}_{n \times M}$ be an orthonormal basis of column span of Φ. There exists a v such that $(f_{\hat{\theta}_{LS}} - f_{\theta^*}) = \tilde{\Phi} v$. Substituting this in the RHS of the above inequality we get:

$$\|f_{\hat{\theta}_{LS}} - f_{\theta^*}\|_2 \leq 2\|\tilde{\Phi}^T \epsilon\|_2 \lesssim \sigma \sqrt{M \log \frac{1}{\delta}},$$

where the last inequality holds with probability at least $1 - \delta$. ■

References

[WM09] Wainwright, Martin J, "Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso)," *IEEE transactions on information theory*, 2009, pp. 2183–2202.