36-709: Advanced Statistical Theory I **Spring 2020**

Lecture 13: March 3

Lecturer: Siva Balakrishnan *Scribe: Mike Stanley*

In the previous lecture, we continued studying properties of noisy sparse linear models. Namely, we considered models of the form $y = X\theta^* + \epsilon$, where $\epsilon \in \mathbb{R}^d$ is a noise vector.

In order to show that the solution to the Lasso program is well controlled, we defined the *Restricted Eigenvalue Condition*, which is a generalization of the *Restricted Nullspace Property* that we used when describing characteristics of the solution to basis pursuit.

Definition 13.1 *The matrix X satisfies the* restricted eigenvalue *(RE) condition over S with parameters (κ, α) if*

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \tag{13.1}$$

for all $\Delta \in \mathbb{C}_\alpha(S)$, where $\mathbb{C}_\alpha(S) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\}$.

In particular, we showed in the previous lecture that if X satisfies RE, has normalized columns, and θ^* is s-sparse, we can obtain the following bound on the estimation error, $\Delta := \widehat{\theta} - \theta^*$.

$$\|\Delta\|_2^2 \leq \frac{s \log p}{n} \tag{13.2}$$

This particular bound comes from assuming gaussian noise.

13.1 Prediction Error Bounds

Above, we provide a bound on the parameter estimation error, here provide a bound on $\|X\Delta\|_2^2$. Namely, we show the following:

Theorem 13.2 *Suppose we choose $\lambda \geq 2\|\frac{X^T X}{n}\|_\infty$. Then $\frac{\|X\Delta\|_2^2}{n} \leq C\lambda\|\theta^*\|_1$ for some C.*

We know that for various statistical models, the choice of $\lambda = C\sigma\sqrt{\frac{\log p}{n}}$ is valid with high probability. With this in mind, we can derive the so called "slow rate" for the LASSO, namely

$$\frac{\|X\Delta\|_2^2}{n} \leq C\sigma\|\theta^*\|_1\sqrt{\frac{\log p}{n}} \tag{13.3}$$

Note, there is also a fast rate for the LASSO, which makes extra sparsity assumptions about θ^*. See the derivation at the end of the previous lecture notes.

Proof:

As a quick aside suppose we first consider the constrained Lasso formulation, namely, minimize $\frac{1}{2n}\|y - X\theta\|_2^2$ subject to $\|\widehat{\theta}\|_1 \leq \|\theta^*\|_1$. By this constraint, we see that $\|\Delta\|_1 \leq \|\widehat{\theta}\|_1 + \|\theta^*\|_1 \leq 2\|\theta^*\|_1$. Using the basic inequality, we obtain that,

$$\frac{\|X\Delta\|_2^2}{2n} \leq \frac{(X^T\epsilon)^T\Delta}{n} \leq \|\frac{X^T\epsilon}{n}\|_\infty\|\Delta\|_1 \leq 2\|\frac{X^T\epsilon}{n}\|_\infty\|\theta^*\|_1.$$

Under the column normalization, and σ-sub-Gaussian noise assumptions we have seen previously that,

$$\|\frac{X^T\epsilon}{n}\|_\infty \leq C\sqrt{\frac{\log p}{n}}, \tag{13.4}$$

yielding our desired slow rate prediction error bound.

Now returning to the Lagrangian LASSO. Using the basic inequality we obtain that,

$$0 \leq \frac{\|X\Delta\|_2^2}{2n} \leq \frac{(X^T\epsilon)^T\Delta}{n} + \lambda\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\}. \tag{13.5}$$

From this we obtain two conclusions. First using our condition on λ we see that,

$$0 \leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda\{\|\theta^*\|_1 - \|\widehat{\theta}\|_1\}.$$

This (after writing $\|\Delta\|_1 \leq \|\widehat{\theta}\|_1 + \|\theta^*\|_1$) implies that $\|\widehat{\theta}\|_1 \leq 3\|\theta^*\|_1$.

Now, one again from (13.5) we see that,

$$\frac{\|X\Delta\|_2^2}{2n} \leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda\{\|\theta^*\|_1\},$$

and so, $\frac{\|X\Delta\|_2^2}{n} \leq C\lambda\|\theta^*\|_1$. Once again using our usual scaling for λ (under the assumptions of column normalization, and σ-sub-Gaussian noise), we once again obtain the slow rate prediction error bound of (13.4).

■

13.2 Intuition for Restricted Eigenvalue Condition

Under the constrained form of the Lasso, we are minimizing the cost function $\mathcal{L}_n(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$ subject to an ℓ_1 constraint with some radius. As our n grows with data, we

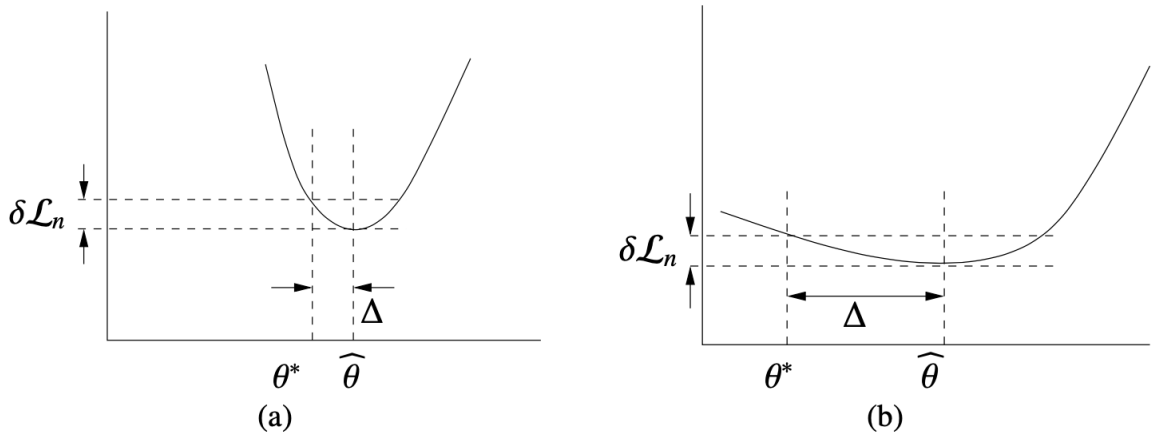

Figure 13.1: Simple example of how curvature allows up to related the closeness of the loss function to closeness in the parameter estimate (Wainwright pg. 208)

would expect that $\mathcal{L}_n(\theta^*) \approx \mathcal{L}_n(\widehat{\theta})$. We would ultimately like to know what closeness in cost function implies about the parameter estimation error. Define $\delta\mathcal{L}_n := \mathcal{L}_n(\theta^*) - \mathcal{L}_n(\widehat{\theta})$. Then figure 13.2 can give us some intuition in one dimension as to why curvature is important when reasoning about the parameter error, Δ, from the cost function error, $\delta\mathcal{L}_n$.

Analogously, when taking the problem into d dimensions, the curvature of the cost function is captured by the structure of its Hessian matrix. For the LASSO, we can compute the Hessian:

$$\nabla^2\mathcal{L}_n = \frac{1}{n}X^TX \tag{13.6}$$

Guaranteeing that the eigenvalues of the above matrix are uniformly bounded away from zero, i.e.

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa\|\Delta\|_2^2 > 0 \tag{13.7}$$

for all $\Delta \in \mathbb{R}^d - \{0\}$, gives us that we have an analogous curvature to the loss function shown in example 13.2.

While it is intuitively clear that curvature should help us obtain rates for parameter estimation (say in the ℓ_2 sense) it is perhaps less clear why under the RE condition (and sparsity of θ^*) we were able to obtain faster rates for the in-sample *prediction error* (compare the bounds at the end of last lecture and the beginning of this lecture).

Intuitively, this is related to the phenomenon of localization. Let us focus on the constrained

LASSO. Using the basic inequality we have that the in-sample prediction error:

$$\frac{\|X\Delta\|_2^2}{2n} \leq \Big\langle \frac{X^T\epsilon}{n}, \Delta \Big\rangle.$$

Now in order to bound the right hand side, there are two possible strategies – we simply upper bound it directly using Holder's inequality, and this leads to the slow rate. We are upper bounding the RHS quite naively.

Alternatively, if we know that $\|\Delta\|_2$ is small (which is true under RE) then we might instead upper bound the RHS more tightly. Concretely, if we know that $\|\Delta\|_2^2 \leq \frac{\|X\Delta\|_2^2}{n\kappa}$ (RE) and $\|\Delta\|_1 \leq 4\|\Delta_S\|_1$ (cone condition) then we could obtain tighter bounds. Roughly (we did this more precisely at the end of the last lecture),

$$\frac{\|X\Delta\|_2^2}{2n} \leq \Big\langle \frac{X^T\epsilon}{n}, \Delta \Big\rangle \leq \sup_{\|\Delta\|_2^2 \leq \frac{\|X\Delta\|_2^2}{n\kappa}, \|\Delta\|_1 \leq 4\|\Delta_S\|_1} \Big\langle \frac{X^T\epsilon}{n}, \Delta \Big\rangle,$$

and this in turn leads to the fast rate. In effect, we are using the RE/curvature condition to argue that $\|\Delta\|_2$ must be small (localizing the empirical process on the RHS), and that this in turn must mean that the prediction error must be very small (leading to fast rates).

13.3 Support Recovery

As above, we are interested in the linear model:

$$y = X\theta^* + \epsilon \tag{13.8}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\|\theta^*\|_0 = s$. Furthermore, we obtain our parameter estimate:

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{2n}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1 \tag{13.9}$$

We are interested to determine the conditions that allow $\text{supp}(\widehat{\theta}) = \text{supp}(\theta^*)$.

We know that $\|\widehat{\theta} - \theta^*\|_2 \lesssim \sqrt{\frac{s\log p}{n}}$ and hence $\|\widehat{\theta} - \theta^*\|_\infty \lesssim \sqrt{\frac{s\log p}{n}}$.

One strategy is to threshold the elements of $\widehat{\theta}$, i.e. $\widehat{S} = \text{supp}(T(\widehat{\theta}))$ where no false inclusions would mean $\widehat{S} \subseteq S$. Using the above bound, we can reason that if

$$\theta_j^* > \sqrt{\frac{s\log p}{n}} \tag{13.10}$$

then $\widehat{S}$ will contain element j. In general, we refer to conditions like this one, which bound from below the magnitude of any non-zero coefficient, in order to subsequently ensure support recovery a $\beta_{\min}$ condition.

Before we can prove any results regarding how well the above thresholding idea works, we must define *mutual incoherence*.

We start with a deterministic design matrix X, and let $S = \{i \in [p] | \theta_i^* \neq 0\}$, $|S| = k$. *Mutual incoherence* is a condition on X_S, i.e. the columns of X restricted to the sparse non-zero subset, such that

$$\max_{j \in S^c} \|(X_S^T X_S)^{-1} X_S^T x_j\|_1 \leq 1 - \alpha \tag{13.11}$$

Intuitively, we are attempting to restrict how well covariates in the complement set S^c align with the actual support set.

We make the following additional assumption about X:

$$\lambda_{\min}\Big(\frac{X_S^T X_S}{n}\Big) > 0 \tag{13.12}$$

We won't go into much detail about this result (see the Wainwright book) but the essence is that the mutual incoherence condition, minimal signal strength condition, and minimum eigenvalue condition will imply that the LASSO solution selects the correct support with high probability.

13.3.1 Primal-Dual Witness Construction as Proof Technique

To prove this result, we introduce the idea of a subgradient and the *primal-dual witness method*. Naively, to find the minimum of a convex cost function, we take a derivative and find the vector at which the derivative equals zero. And indeed, the $\widehat{\theta}$ that is recovered is optimal. Namely, we find the $\widehat{\theta}$ such that

$$\frac{\partial J}{\partial \theta} = \frac{-X^T(y - X\theta^*)}{n} + \lambda \frac{\partial \|\theta\|_1}{\partial \theta} = 0 \tag{13.13}$$

which will be true at $\theta = \widehat{\theta}$. Note, J refers to the constrained LASSO equation. However, the cost function is not differentiable because of the ℓ_1 norm.

We can side-step this complication using subgradients. As the name suggests, we want to find a function that lower bounds the gradient of our actual ℓ_1 constraint. Our subgradient, $\widehat{z}$ can then be used write down the zero-subgradient condition, or "KKT" condition.

$$\frac{-X^T(y - X\theta^*)}{n} + \lambda_n \widehat{z} = 0 \tag{13.14}$$

For LASSO, we can intuitively define the subgradient function as a modified sign function. Namely,

$$\partial\|\theta\|_1 = \text{sign}(\theta_j) = \begin{cases} 1 & \theta_j > 1 \\ -1 & \theta_j < 1 \\ [-1, 1] & \theta_j = 0 \end{cases}$$

For the last condition, notice that if $\theta_j = 0$ our subgradient value can be any real number between -1 and 1. As such, we say that $\widehat{z} \in \partial\|\widehat{\theta}\|_1$. Furthermore, we say that a pair $(\widehat{\theta}, \widehat{z}) \in \mathbb{R}^p \times \mathbb{R}^p$ is primal-dual optimal if $\widehat{\theta}$ is a minimizer and $\widehat{z} \in \partial\|\widehat{\theta}\|_1$. The primal-dual witness method constructs such a pair.

The following shows the Primal-Dual Witness Construction:

1. Set $\widehat{\theta}_{S^c} = 0$
2. Find $(\widehat{\theta}_S, \widehat{z}_S)$ such that

$$\widehat{\theta}_S = \arg\min_{\theta} \frac{1}{2n}\|y - X_S\theta_S\|_2^2 + \lambda_n\|\theta_S\|_1 \tag{13.15}$$

   and $\widehat{z}_S = \text{sign}(\widehat{\theta}_S)$.

3. Find $\widehat{z}_{S^c}$ such that KKT conditions are true.

Once again the remaining details (of how to complete Step 3) can be found in the Wainwright book.