10-716: Advanced Machine Learning **Spring 2019**

Lecture 12: February 28

Lecturer: Pradeep Ravikumar *Scribes: Jacob Tyo, Rishub Jain, Ojash Neopane*

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

In the previous lecture we talked about the Restricted Null space Property (RNP). We begin this lecture by recounting the relevant theorems, and then improving on these results with Restricted Eigenvalues (RE).

12.1 Preliminaries

We first recall some definitions:

Definition 12.1 (Restricted Nullspace Property(RNSP)) *A matrix $\boldsymbol{X}$ satisfies the restricted nullspace property (RNSP) with respect to $S \subset \{1, \ldots, d\}$ if*

$$\mathbb{C}(S) \cap \text{null}(\boldsymbol{X}) = \{0\} \tag{12.1}$$

where $\mathbb{C}(S) \coloneqq \{\Delta \in \mathbb{R}^d \mid ||\Delta_{S^c}||_1 \leq ||\Delta_S||_1\}$ is the cone of vectors whose ℓ_1-norm off the support is dominated by the ℓ_1-norm on the support.

We are interested in the RNSP because when it is satisfied, we know that solving the Basis Pursuit Linear Program (BPLP)

$$\min_{\theta \in \mathbb{R}^d} ||\theta||_1 \quad \text{such that } \boldsymbol{X}\theta = y \tag{12.2}$$

is equivalent to solving the ℓ_0 regularized problem

$$\min_{\theta \in \mathbb{R}^d} ||\theta||_1 \quad \text{such that } \boldsymbol{X}\theta = y \tag{12.3}$$

which is typically computationally infeasible to solve (since it requires doing a search over a space which grows exponentially with the sparsity of θ). This is formalized by the following theorem:

Theorem 12.2 *The following two properties are equivalent:*

1. *For any vector $\theta^* \in \mathbb{R}^d$ with support S, the BPLP equation 12.2 applied with $y = \boldsymbol{X}\theta^*$ has a unique solution $\hat{\theta} = \theta^*$*
2. *The matrix $\boldsymbol{X}$ satisfies the RNSP with respect to S*

For a given $\boldsymbol{X}$, identifying the subsets S which satisfy the RNSP results in the same difficulties as solving the ℓ_0 regularized problem. To circumvent this, we introduce 2 definitions which allow us to more easily certify when the RNSP holds:

Definition 12.3 (Pairwise Incoherence) *The pairwise incoherence of a design matrix $\boldsymbol{X}$, denoted $\delta_{PW}(\boldsymbol{X})$ is defined as*

$$\delta_{PW}(\boldsymbol{X}) \coloneqq \max_{j \neq k} \frac{|\langle x_j, x_k \rangle|}{n} \tag{12.4}$$

Definition 12.4 (Restricted Isometry Property) *For a given integer $s \in \{1, \ldots, d\}$, we say that $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ satisfies a restricted isometry property (RIP) of order s with constant $\delta_s(\boldsymbol{X}) > 0$ if*

$$\left| \left| \left| \frac{x_s^T x_s}{n} - I_s \right| \right| \right|_2 \leq \delta_s^{(RIP)}(\boldsymbol{X}) \tag{12.5}$$

for all subsets S of size at most s. Here $|||\cdot|||_2$ denotes the ℓ_2-operator norm of a matrix, corresponding to its maximum singular value.

From this, we have the theorems:

Theorem 12.5

$$\delta_{PW}(\boldsymbol{X}) \leq \delta_s^{(RIP)}(x) \leq s\delta_{pw}(x) \tag{12.6}$$

Theorem 12.6 *If the pairwise incoherence of $\boldsymbol{X}$ satisfies the bound*

$$\delta_{PW}(\boldsymbol{X}) \leq \frac{1}{3s} \tag{12.7}$$

then the RNSP holds for all subsets S of cardinality at most s.

Theorem 12.7 *If the RIP constant of order $2s$ is bounded as*

$$\delta_{2s}^{(RIP)}(\boldsymbol{X}) \leq \frac{1}{3} \tag{12.8}$$

then the RNSP holds for any subset S of cardinality $|S| \leq s$

Notice that Theorem 12.6 is a much stronger statement than 12.7, because we can substitute 12.6 into 12.5 to get 12.7.

This result is hard to use, because RNP is hard to check. Attempting to improve upon this result, and further our understanding of sufficient statistics, we will come up with a better condition that corresponds to RNP holding. This leads us to restricted eigenvalues.

12.2 Estimation in Noisy Settings

Everything we have discussed so far has been under the assumption that we don't have any noise in our observations so that we observe the pair $(y, \boldsymbol{X}) \in \mathbb{R}^n \times \mathbb{R}^{n \times d}$ which are related by the linear model

$$y = \boldsymbol{X}\theta^* \tag{12.9}$$

However, in most settings we are interested in solving the more realistic problem

$$y = \boldsymbol{X}\theta^* + w \tag{12.10}$$

which is similar to Problem (12.9) except that there is a noise vector $w \in \mathbb{R}^n$. Analogous to the BPLP, we can define the following equivalent ℓ_1 regularized problems:

$$\min_{\theta \in \mathbb{R}^d} \{\frac{1}{2n}||y - \boldsymbol{X}\theta||_2^2 + \lambda_n||\theta||_1\} \quad \text{(Lagrangian Lasso)} \tag{12.11}$$

$$\min_{\theta \in \mathbb{R}^d} \{\frac{1}{2n}||y - \boldsymbol{X}\theta||_2^2\} \quad \text{such that } ||\theta||_1 \leq R \quad \text{(Constrained Lasso)} \tag{12.12}$$

$$\min_{\theta \in \mathbb{R}^d} ||\theta||_1 \quad \text{such that } \frac{1}{2n}||y - \boldsymbol{X}\theta||_2^2 \leq b^2 \quad \text{(Relaxed Basis Pursuit Program)} \tag{12.13}$$

Remark: Note that all of the above formulations are equivalent in the sense that for specific parameters of λ_n, R, b, the above formulations all return the same solution.

Having defined the analogs of the BPLP, we might be tempted to ask under what conditions can we recover θ^*. However, because of the noise, this seems like an unreasonable question to ask as it will not always be possible to exactly recover θ^*. Instead we ask the following more appropriate questions:

(1) Under what assumptions can we bound the error $||\hat{\theta} - \theta^*||_2$ between the Lasso solution $\hat{\theta}$ and the unknown regression vector θ^*?
(2) Under these assumptions what are bounds we can obtain?

12.2.1 The Restricted Eigenvalue Condition

In this section we will answer the first question: under what assumptions can we bound $||\hat{\theta} - \theta^*||_2$?

To frame the problem, remember that the point of studying this is to determine when we can optimally solve a sparse linear model. Previously, we showed that we can solve such models, but pragmatically it was too computationally complex, and was NP Hard. However, with some simple conditions on the data, we can make this problem tractable.

For intuition with respect to the cone, think of the situation where:

$$Y = X\theta$$
$$S = \text{supp}(\theta)$$

Then, imagine that there exists a $\widetilde{\theta}$ such that $S = \text{supp}(\theta) = \text{supp}(\widetilde{\theta})$ and $X\widetilde{\theta} = 0$. Then:

$$Y = X(\theta + \widetilde{\theta}) = X\theta$$

And thus this is an under-determined and unidentifiable linear system. In summary, if X has an intersection with S, then the problem is unidentifiable. However, if it has a small intersection with S, it is solvable:

$$\text{null}(X) \cap s = \emptyset$$

where $s = \{\Delta | \Delta_{s^c} = 0\}$ (for the L0 problem). To make this problem solvable with L1 loss, more vectors are required. For example, instead of requiring the intersection with just the y-axis, it is now the y-axis plus all of the vectors close to the y-axis, shown in Figure 12.1 . Thus, if there is a null intersection with all vectors that form a cone around the y-axis, then this problem with L1 loss is solvable.

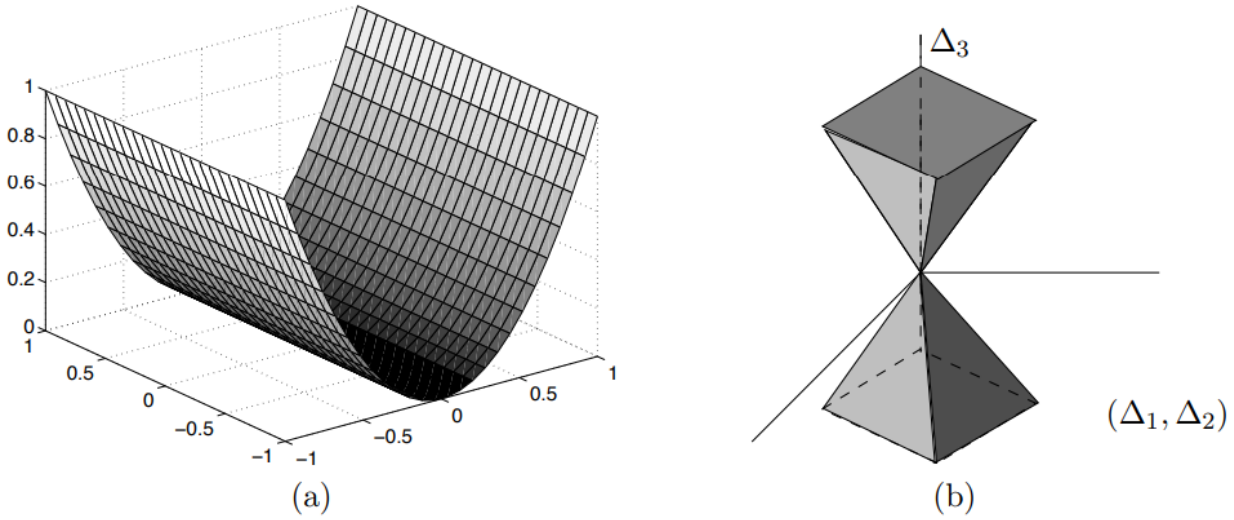

Figure 12.1: (a) shows how in high dimensions, a convex function is often curved in some directions but flat in others. (b) visualizes the cone of vectors close to the y-axis. $C_\alpha(s)$. Figure from [wainwright]

Let:

$$C_\alpha(s) := \left\{ \Delta \in \mathbb{R} \middle| ||\Delta_{s^c}||_1 \leq \alpha ||\Delta_s||_1 \right\} \tag{12.14}$$

Now we want a null space property with respect to α. Think about:

$$\min_{\Delta \neq 0} \frac{||X\Delta||_2}{\sqrt{n}||\Delta||_2}$$

Which represents the minimum eigen value of X. Instead, restrict Δ to lie in the cone of $C_\alpha(s)$. This leads us to a form of the Restricted Eigenvalue condition:

$$\min_{\Delta \in C_\alpha(s), \Delta \neq 0} \frac{||X\Delta||_2}{\sqrt{n}||\Delta||_2} \geq \kappa > 0 \tag{12.15}$$

If Equation 12.15 is satisfied, then we know that RNP is also satisfied.

If RNP were not satisfied, then we know that Equation 12.15 would be equal to 0. Because, in this setting we can enforce that Equation 12.15 is greater than zero, we know that the RNP is satisfied.

For some intuition on why RE is required, think of this as the curvature of the squared loss. If the curvature is not significant enough, then having a small error may not be representative of being close to the true parameter. To better understand why we need this, consider the squared loss:

$$\mathcal{L}_n(\theta) = \frac{1}{2n}||Y - X\theta||_2^2$$

The hessian represents the curvature:

$$\nabla^2 \mathcal{L}_n(\theta) = \left(\frac{X^T X}{n}\right)$$

The above is a $d \times d$ matrix, with rank at most n which is less than d. As previously noted, in machine learning problems we are primarily concerned with minimizing loss. However, getting ϵ-close in loss does not gaurentee that we are ϵ-close in terms of the actual parameter (see Figure 12.2). This information is captured by the hessian, and therefore we want to bound this to ensure that being close in loss means that we are also close in the parameter.

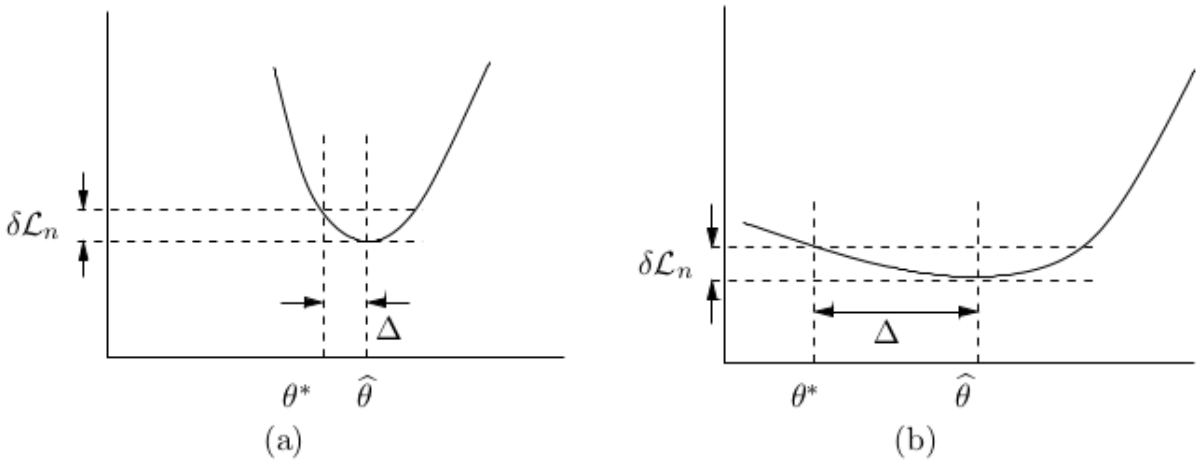

Figure 12.2: Relationship between the curvature of the cost function and estimation error. Figure from [wainwright]

The importance of curvature is problematic for the sparse setting, because some eigenvalues will be equal to zero in some directions, because the loss landscape can be flat in may directions (shown in part (a) of Figure 12.1). However, Equation 12.15 means that we don't care about curvature in all directions. We will see that imposing Equation 12.15 is enough.

12.2.2 Bounds on ℓ_2-error for Hard Sparse Models

Now that we have answered the first question we turn to the second one: under the assumption that the design matrix $\boldsymbol{X}$ satisfies the restricted eigenvalue condition, what bounds can we obtain on $||\hat{\theta} - \theta^*||_2$?

For the remainder of this subsection, we will be operating under the following assumptions:

(A1) The vector θ^* is supported on a subset $S \subseteq \{1, \ldots, d\}$ with $|S| = s$

(A2) The design matrix satisfies the restricted eigenvalue condition over S with parameters $(\kappa, 3)$

Theorem 12.8 *Suppose that assumptions A1 and A1 are satisfied and that*

$$\lambda_n \geq 2\frac{||x^T w||_\infty}{n} \tag{12.16}$$

then

$$||\hat{\theta} - \theta^*||_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n \asymp \sqrt{\frac{s}{n}} \tag{12.17}$$

where $\hat{\theta}$ is a solution to the Largrangian Lasso (Eq. 12.11)

Proof: Assume $\hat{\theta}$ is a solution to the Lagrangian Lasso. This implies

$$L_n(\hat{\theta}) + \lambda_n||\hat{\theta}||_1 \leq L_n(\theta^*) + \lambda_n||\theta^*||_1 \tag{12.18}$$

Now, expanding $L_n(\hat{\theta})$ we see that

$$\begin{aligned}
L_n(\hat{\theta}) &= \frac{1}{2n}||y - \boldsymbol{X}\hat{\theta}||_2^2 && (12.19)\\
&= \frac{1}{2n}||\boldsymbol{X}(\hat{\theta} - \theta^*) + w||_2^2 && \text{(setting } y = \boldsymbol{X}\theta^* + w) \quad (12.20)\\
&= \frac{1}{2n}||\boldsymbol{X}(\hat{\theta} - \theta^*)||_2^2 + \frac{1}{2n}||w||_2^2 - \frac{1}{n}w^t x(\hat{\theta} - \theta^*) && (12.21)
\end{aligned}$$

Similarly we can see that $L_n(\theta^*) = \frac{1}{2n}||w||_2^2$ by setting $\hat{\theta} = \theta^*$ in Eq. 12.21. Plugging these back into Eq. 12.18 and setting $\hat{\Delta} = \hat{\theta} - \theta^*$ we have

$$\begin{aligned}
o &\leq \frac{1}{2n}||\boldsymbol{X}\hat{\Delta}||_2^2 && (12.22)\\
&\leq \underbrace{\frac{1}{n}w^T\boldsymbol{X}\hat{\Delta}}_{\text{Term 1}} + \underbrace{\lambda_n(||\theta^*||_1 - ||\theta^* + \hat{\Delta}||_1}_{\text{Term 2}} && (12.23)
\end{aligned}$$

Now, we will analyze Term 1 and Term 2 separately. For Term 1, we have

$$\begin{aligned}
\frac{1}{n}x^T\boldsymbol{X}\hat{\Delta} &= \langle\frac{1}{n}\boldsymbol{X}^T w, \hat{\Delta}\rangle && (12.24)\\
&\leq ||\frac{1}{n}\boldsymbol{X}^t w||_\infty||\hat{\Delta}||_1 && \text{(Holder)} \quad (12.25)\\
&\leq \frac{\lambda_n}{2}||\hat{\Delta}||_1 && \text{(Assumption on Lambda (Eq. 12.16))} \quad (12.26)\\
&= \frac{\lambda_n}{2}(||\hat{\Delta}_S||_1 + ||\hat{\Delta}_{S^c}||_1) && (12.27)
\end{aligned}$$

For Term 2 we have

$$||\theta^*||_1 - ||\theta^* + \hat{\Delta}||_1 = ||\theta^*_S||_1 - ||\theta^*_S + \hat{\Delta}_s||_1 - ||\hat{\Delta}_{S^c}||_1 \tag{12.28}$$

$$\leq ||\theta^*_S||_1 - ||\theta_s||_1 + ||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1 \tag{12.29}$$

$$= ||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1 \tag{12.30}$$

So that

$$0 \leq \frac{||\boldsymbol{X}\hat{\Delta}||_2}{n} \leq \lambda_n(3||\hat{\Delta}_S||_1 - ||\hat{\Delta}_{S^c}||_1) \tag{12.31}$$

$$\Rightarrow ||\hat{\Delta}_{S^c}||_1 \leq 3||\hat{\Delta}_S||_1 \tag{12.32}$$

so that $\hat{\Delta} \in \mathbb{C}_3(S)$ which implies $\frac{||\boldsymbol{X}\hat{\Delta}||_2}{n} \geq \kappa||\hat{\Delta}||_2^2$. This implies that $\kappa||\hat{\Delta}_s||_2^2 \leq 3\lambda_n||\hat{\Delta}_S||_1 \leq 3\lambda_n\sqrt{s}||\hat{\Delta}_s||_2$. Solving for $||\hat{\Delta}_s||_2$ gives the desired result. ■

Corollary 12.9 *Let* $w_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ $i = \{1, \ldots, n\}$. *Then,*

$$\max_{j=\{1,\ldots,n\}} \frac{||x_j||_2}{\sqrt{n}} \leq c$$

Now analyze $\frac{X_j^T w}{n}$. *Looking at the variance shows:*

$$Var(\frac{X_j^T w}{n}) = \frac{1}{n^2}\mathbb{E}[(X_j^T w)(X_j^T w)] = \frac{\sigma^2}{n^2}X_j^T I X_j = \frac{||X_j||^2\sigma^2}{n^2} \leq \frac{c^2\sigma^2}{n}$$

And therefore X_j *is sub-Gaussian. Then bounding the sum of sub-Gaussian RVs:*

$$\mathbb{P}\Big(\max_{j=1,\ldots,d}|Z_j| > t\Big) \leq 2d\exp\Big\{\frac{-nt^2}{2c^2\sigma^2}\Big\} = 2\exp\Big\{\frac{-nt^2}{2c^2\sigma^2} + logd\Big\}$$

Now let $t_\delta = \frac{\sqrt{2\log(d)}c\sigma}{\sqrt{n}} + c\sigma\delta$, *and therefore:*

$$\mathbb{P}\Big(||\frac{X^T w}{n}||_\infty > t_\delta\Big) \leq 2e^{\frac{-n\delta^2}{2}}$$

This allows us to set λ_n*:*

$$\lambda_n = 2c\sigma\Big(\sqrt{\frac{2\log d}{n}} + \delta\Big)$$

Which implies:

$$||\hat{\theta} - \theta^*||_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n = \frac{6c\sigma}{\kappa}\sqrt{\frac{s\log d}{n}} + \frac{6c\sigma}{\kappa}\sqrt{s}\delta \tag{12.33}$$

The above holds with probability $1 - 2\exp\{\frac{-n\delta^2}{2}\}$. *One interesting observation from this loss is that* $\log d$ *is the only extra loss suffered, and can be though of as the cost of searching.*

References

[wainwright] M. Wainwright, "High Dimensional Statistics," *Prerelease*, 2019