Statistics 210B Lecture 17 Notes

Daniel Raban

March 17, 2022

1 Introduction to Sparse Linear Regression

1.1 High-dimensional linear regression

Consider the following high-dimensional linear model, with $y = X\theta^* + w \in \mathbb{R}^n$, where $X \in \mathbb{R}^{n \times d}$ is the **design matrix** and $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ is the **response**. We write the design matrix as

$$X = \begin{bmatrix} x_1^\top \\ \vdots \\ y_n^\top \end{bmatrix}, \qquad x_i \in \mathbb{R}^d, i = 1, \dots, n$$

and the parameter as

$$\theta^* = \begin{bmatrix} \theta_1^* \\ \vdots \\ \theta_n^* \end{bmatrix}.$$

We interpret

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

as noise. We can also write the problem in the scalar form

$$y_i = \langle x_i, \theta^* \rangle + w_i \qquad i = 1, \dots, n.$$

Our task is that we observe (X, y), and we want to estimate $\theta^* \in \mathbb{R}^d$.

The classical asymptotic regime is that the dimension d is fixed, and the sample size n is large. We will focus on the high dimensional regime, in which both d and n are large, and $d > n$. In high dimensions, least squares will not give a consistent estimate. We need

some further assumptions on θ^* and X so that consistent estimation is possible in high dimensions.

We will assume a *sparcity assumption*.

Definition 1.1. For $\theta^* \in \mathbb{R}^d$, define the **support** as

$$S(\theta^*) = \{j \in [d] : \theta_j^* \neq 0\}.$$

We will assume that $|S(\theta^*)| \leq s$. If $S(\theta^*)$ is known, then $n \geq s$ is enough for consistent estimation. We can look at the least-squares problem

$$\min_{\theta_S} \|y - X_S\theta_S\|_2^2, \qquad \text{where } \theta_S = (\theta_i)_{i\in S} \in \mathbb{R}^{|S|},$$

$$X_S = \begin{bmatrix} x_{1,S}^\top \\ \vdots \\ x_{n,S}^\top \end{bmatrix} \in \mathbb{R}^{n\times|S|}, \qquad \text{where } x_{i,S} = (x_{i,j})_{j\in S} \in \mathbb{R}^{|S|}.$$

which will have a unique minimizer.

Because of this, we will focus on when $S(\theta^*)$. We will show that $n \geq s\log(d/s)$. The interesting regime for this problem is when $s \ll n \ll d$.

1.2 Recovery in the noiseless setting

In the noiseless setting, we have

$$y = X\theta^* \in \mathbb{R}^n, \qquad \theta^* \in \mathbb{R}^d,$$

where θ^* is s-sparse. Our task is to recover θ^* given (X, y). If $n < d$, there will be infinite solutions θ such that $y = X\theta$. The **null space** of X is

$$\text{Null}(X) := \{\Delta \in \mathbb{R}^d : X\Delta = 0\}.$$

For all $\Delta \in \text{Null}(X)$, $\theta = \theta * +\Delta$ satisfies $y = X\theta$. The **feasible space** of $y = X\theta$ is the affine space $\theta^* + \text{Null}(X) = \{\theta^* + \Delta : \Delta \in \text{Null}(X)\}$. This gives infinitely many solutions.

To find θ^*, we can use **ℓ_0-norm minimization**:

$$\min_{\theta: y=X\theta} \|\theta\|_0, \qquad \|\theta\|_0 = \sum_{i=1}^d \mathbb{1}_{\{\theta_i\neq 0\}}.$$

However, this is computationally hard because this norm is not convex. To solve this problem, we need to search over $S \subseteq [d]$, where $|S|$ is from $1, 2, \ldots, s$, and look at whether there is a solution of $y = X_S\theta_S$. The complexity of this problem is

$$\Theta\left(\sum_{k=1}^{s-1}\binom{d}{k}\right) \approx d^s,$$

which is exponential in the sparsity. We would prefer polynomial complexity.

Instead, it is more efficient to consider the convex relaxation of ℓ_1-norm minimization:

$$\min_{y=X\theta} \|\theta\|_1 = \sum_{i=1}^{d} |\theta|.$$

This problem was called **basis pursuit** in the original 1994 paper by Chen, Donoho, and Saunders.[1] If we consider the convex dual problem, then we get the **LASSO** problem, as introduced by Tibshirani. This ℓ_1-norm minimization problem can be reformulated as a linear program and solved efficiently.

Our question is as follows: What is the condition such that the solution

$$\widehat{\theta} := \arg\min_{\theta} \{\|\theta\|_1 : y = X\theta\}$$

equals the original θ^*?

1.3 A sufficient condition for exact recovery

Fix $\theta^* \in \mathbb{R}^d$ with $S(\theta) = s$. We want some condition of $X \in \mathbb{R}^{n\times d}$ such that

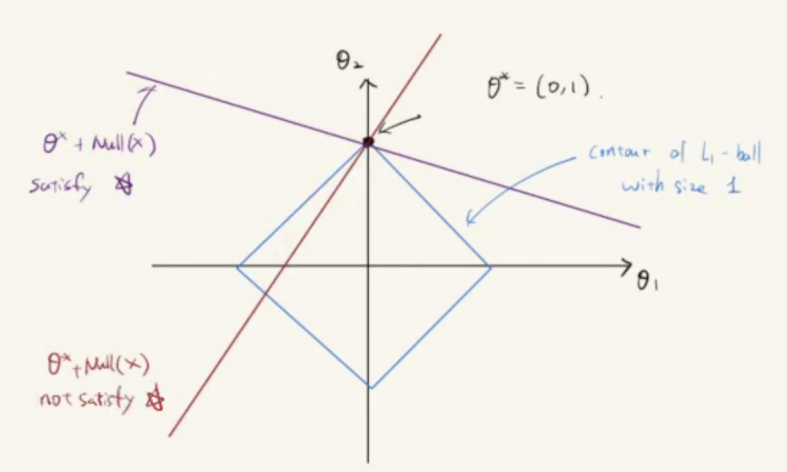
$$\arg\min_{\theta} \{\|\theta\|_1 : X\theta^* = X\theta\} = \theta^*.$$

Notice that $X\theta^* = X\theta$ means that $\theta \in \text{Null}(X) + \theta^*$, so this condition can be reformulated as

$$\forall \theta \in \theta^* + \text{Null}(X) \setminus \{\theta^*\}, \qquad \|\theta\|_1 > \|\theta_*\|_1.$$

When will this property hold?

Example 1.1. To gain some intuition, consider the case with $d = 2$, $n = 1$, and with $\dim \text{Null}(X) = 1$. Then $\theta^* + \text{Null}(X)$ is an affine space passing through θ^*.

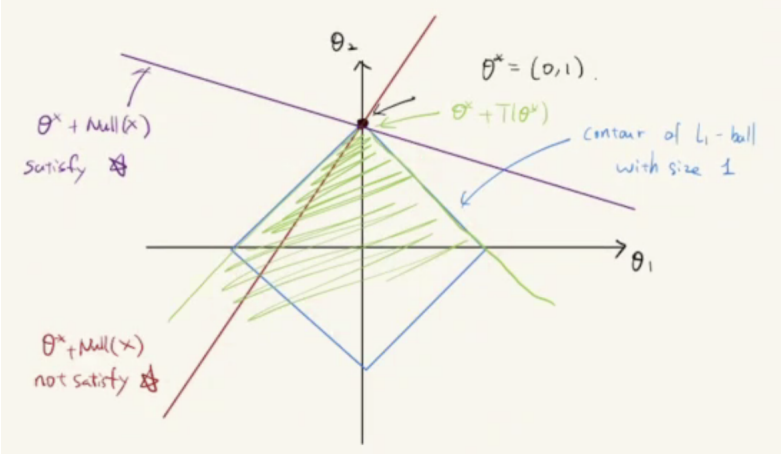


[1] This paper was not published until 1998.

We can define the **tangent cone**

$$T(\theta^*) := \{\Delta \in \mathbb{R}^d : \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1 \text{ for some } t > 0\}.$$

This enters the picture as follows.



We can see from the picture that we will not have exact recovery exactly when $\theta^* + \text{Null}(X)$ intersects the tangent cone at more than one point.

We will get exact recovery when

$$\theta^* + \text{Null}(X) \cap \theta^* + T(\theta^*) = \{\theta^*\},$$

which is equivalent to the condition

$$\text{Null}(X) \cap T(\theta^*) = \{0\}.$$

This is a necessary and sufficient condition for exact recovery.of θ^*. This condition involves the interplay between properties of X and properties of θ^*.

Let's see how to reformulate this tangent cone. In our example, $d = 2$, $S = \{2\}$, and $\theta^* = (0, 1)^\top$. Then

$$\begin{aligned} T(\theta^*) &= \{(\Delta_1, \Delta_2) : \exists t > 0, \|(0, 1) + (t\Delta_1 + \Delta_2)\|_1 \leq \|(0, 1)\|_1\} \\ &= \{(\Delta_1, \Delta_2) : \exists t > 0, t|\Delta_1| + |1 + t\Delta_2| \leq 1\} \qquad = \{(\Delta_1, \Delta_2) : |\Delta_1| \leq |\Delta_2|, \Delta_2 \leq 0\}. \end{aligned}$$

In general, suppose $S(\theta^*) = S \subseteq [d]$. Then we can express the tangent cone as

$$T(\theta^*) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1, |\Delta_i||\theta_i^*| \leq 0 \forall i \in S\}, \qquad \Delta_S = (\Delta_i)_{i \in S}, \Delta_{S^c} = (\Delta_i)_{i \in S^c}.$$

Define the cone

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

Then $T(\theta^*) \subseteq \mathbb{C}(S)$ for any $S(\theta^*) = s$. A sufficient condition for exact recovery is that

$$\mathbb{C}(S) \cap \text{Null}(X) = \{0\}.$$

Definition 1.2. Let $X \in \mathbb{R}^{n\times d}$ with $S \subseteq [d]$. We say taht X satisfies the **restricted nullspace property with respect to** S ($\mathrm{RN}(S)$) if

$$\mathbb{C}(S) \cap \mathrm{Null}(X) = \{0\}.$$

Theorem 1.1. *The following are equivalent:*

(a) For all $\theta^* \in \mathbb{R}^d$ with $S(\theta^*) = S$,

$$\arg\min_{\theta}\{\|\theta\|_1 : X\theta_* = X\theta\} = \theta^*\}.$$

(b) X satisfies the $\mathrm{RN}(S)$, i.e.

$$\mathbb{C}(S) \cap \mathrm{Null}(X) = \{0\}.$$

Earlier, we said that $\mathrm{RN}(S)$ was only a *sufficient* condition for exact recovery. But this theorem says that it is necessary to have exact recovery for *any* θ^* with $S(\theta^*) = S$.

Proof. (b) $\implies$ (a): Let $\widehat{\theta} \in \arg\min_\theta\{\|\theta\|_1 : X\theta^* = X\theta\}$. Then $\|\widehat{\theta}\|_1 \le \|\widehat{\theta}^*\|_1$. Now suppose we define $\widehat{\Delta} = \widehat{\theta} - \theta^* \in \mathrm{Null}(X)$; we want to show that $\widehat{\Delta} \in \mathbb{C}(S)$. Then we have

$$\begin{aligned}
\|\theta^*_S\|_1 &= \|\theta^*_1\| \\
&\ge \|\theta^* + \widehat{\Delta}\|_1 \\
&= \|\theta^*_S + \widehat{\Delta}_S\|_1 + \|\underbrace{\theta^*_{S^c}}_{=0} + \widehat{\Delta}_{S^c}\|_1
\end{aligned}$$

Using the triangle inequality,

$$\ge \|\theta^*_S\|_1 - \|\widehat{\Delta}_S\|_1 + \|\widehat{\Delta}_{S^c}\|_1.$$

Cancelling $\|\theta^*_S\|_1$ on both sides, we get $\|\widehat{\Delta}_{S^c}\|_1 \le \|\widehat{\Delta}_S\|$. That is, $\widehat{\Delta} \in \mathbb{C}(S) \cap \mathrm{Null}(X)$. By our assumption, this means $\widehat{\Delta} = 0$, so $\widehat{\theta} = \theta^*$.

(a) $\implies$ (b): Let $\widetilde{\theta} \in \mathrm{Null}(X) \setminus \{0\}$. We want to construct a θ^* so that to recover θ^*, we need $\mathrm{RN}(S)$. We will not prove this direction because it is mostly more algebra. □

What are examples of matrices satisfying $\mathrm{RN}(S)$? For a random matrix $X \in \mathbb{R}^{n\times d}$ with $X_{i,j} \overset{\text{iid}}{\sim} N(0,1)$, $\mathrm{RN}(S)$ is satisfied with high probability as long as $n \gtrsim s\log(d/s)$. This is one of the main components of **compressed sensing**. If you want to estimate a sparse signal, you can apply a random matrix and solve this ℓ_1 minimization problem.