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

Lecture 11: February 26

Lecturer: Pradeep Ravikumar *Scribes: Jing Mao, Zhaojie Gong, Junyan Jiang*

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

11.1 Sparse linear models in high dimensions

Linear model is largely used in machine learning and statistics. Typically in low-dimensional instantiation, the number of predictors d is substantially less than the sample size n. In contrast, we are going to explore the high-dimensional regime, which allows scaling that $d \asymp n$ or even $d \gg n$.

11.1.1 Problem formulation

Suppose that we observe $y_i \in \mathbb{R}, x_i \in \mathbb{R}^d$ for $i = 1, 2, \ldots, n$. Then the linear model is of the form

$$y_i = {\theta^*}^T x_i + w_i$$

, where $w_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. noise variables and $\theta^* \in \mathbb{R}^d$. In fixed design, $\{x_i\}_{i=1}^n$ are fixed whereas in random design, each $x_i \sim P_x$ i.i.d.

When the number of samples $n < d$, the linear system is under-determined and we need to equip the model with some form of low-dimensional structure.

Definition 11.1 *The hard sparsity assumption states that the support set of* θ^**,*

$$S(\theta^*) := \{j \in \{1, 2, \ldots d\} \mid \theta_j^* \neq 0\}$$

has cardinality $|S(\theta^*)| < n$*.*

Definition 11.2 *The p-norm of vector* θ *is*

$$\|\theta\|_p = \left(\sum_{i=1}^{d} |\theta_i|^p \right)^{1/p}$$

When $p = 0$, $\|\theta\|_0 = \sum_{i=1}^d \mathbb{I}(\theta_i \neq 0)$, which corresponds to hard sparsity. For *weak sparsity*, $\|\theta\|_p \leqslant C$ which gives a set of θ.

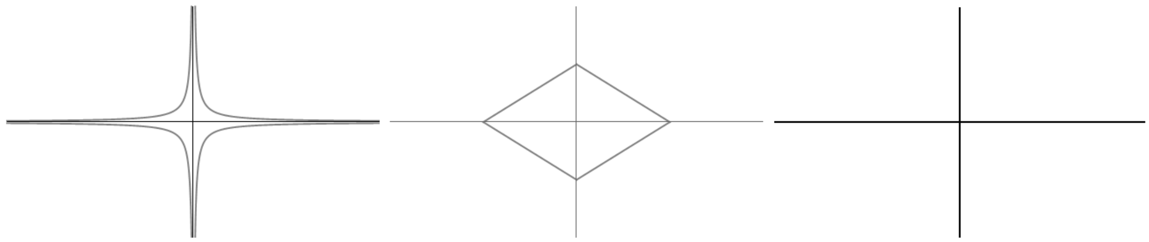


Figure 11.1: Illustration of ℓ_p for parameter $p \in [0, 1]$. (a) with $p < 1$ (b) $p = 1$ (convex) (c) $p = 0$

Example 1 *(Gaussian Sequence Model) In this model, we make observations of the form*

$$y_i = \sqrt{n}\theta_i^* + w_i \qquad i = 1, 2, \ldots, n$$

where $n = d$ *and* $\mathbf{y} = (\sqrt{n}\mathbf{I}_n)\theta^* + \mathbf{w}$.

Example 2 *(Lifting and non-linear functions) Consider polynomial functions of the form*

$$f_\theta(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \ldots + \theta_q t^q$$

Where we observe n samples $\{(t_i, y_i)\}_{i=1}^n$. *We could then define the matrix* $\mathbf{X}$ *as*

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^q \\ 1 & t_2 & t_2^2 & \ldots & t_2^q \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & t_n & t_n^2 & \ldots & t_n^q \end{bmatrix}$$

More generally, we formulate

$$f_\theta(t) = \sum_{j=1}^{d} \theta_j \phi_j(t)$$

where $\{\phi_1, \ldots \phi_d\}$ *are known basis functions. Then we have* $y = \mathbf{X}\theta + w$*, where* $X_{ij} = \phi_j(t_i)$.

11.1.2 Recovery in noiseless setting

Consider $\mathbf{X} \in \mathbb{R}^{n \times d}$ where $n < d$. In noiseless setting, we assume that $\exists \theta^*$ s.t. $y = \mathbf{X}\theta^*$ and $\|\theta^*\|_0 = s^* \ll d$. In this case, we consider the following optimization problem

$$\min_\theta \|\theta\|_0 \qquad \text{s.t. } y = \mathbf{X}\theta$$

The approach to solve the above problem works as following:

for $s = 1, \ldots, d$,
  for all $S \subseteq \{1, \ldots, d\}$ s.t. $|S| = s$
    check if $\exists \theta_s$ s.t. $y = \mathbf{X}_s\theta_s$

The complexity of this approach is then $\sum_{j=1}^{s^*} \binom{d}{j} \asymp d^{s^*}$, which would be computationally expensive if s^* is large.

We could also approximate this non-convex optimization problem with a convex program by changing $\|\theta\|_0$ to $\|\theta\|_1$. This gives the following optimization problem

$$\min_\theta \|\theta\|_1 \qquad \text{s.t. } y = \mathbf{X}\theta$$

which is known as the *basis pursuit linear program.*

11.2 Exact recovery and restricted nullspace

We define the set

$$T(\theta^*) = \{\Delta \mid \|\theta^* + \Delta\|_1 \leqslant \|\theta^*\|_1\}$$

and note that the null space of $\mathbf{X}$ is defined as

$$\text{null}(\mathbf{X}) = \{\Delta \mid \mathbf{X}\Delta = 0\}$$

We have the following theorem.

Theorem 11.3 *θ^* is the unique solution to the above problem iff $T(\theta^*) \cap null(\mathbf{X}) = \{\mathbf{0}\}$.*

Proof: If $T(\theta^*) \cap \text{null}(\mathbf{X}) \neq \{\mathbf{0}\}$ then

$$\exists \bar{\Delta} \in T(\theta^*) \cap \text{null}(\mathbf{X})$$

We have

$$\|\theta^* + \bar{\Delta}\|_1 \leqslant \|\theta^*\|_1$$

and

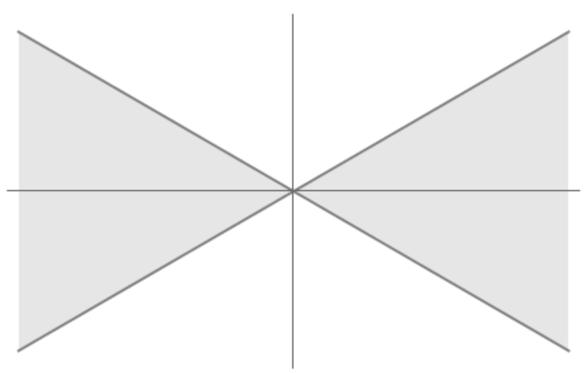
$$\mathbf{X}(\theta^* + \bar{\Delta}) = \mathbf{X}\theta^* + \mathbf{X}\bar{\Delta} = y$$

Then θ^* is not the unique solution.
The other direction is similar. For more details, please refer to theorem 7.1 in the textbook. ■

We define the set

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

corresponding to a cone of vectors.
In the two dimensional case, when S has only one element, the cone can be shown as follows.



The shade area corresponds to $|\Delta_1| \geqslant |\Delta_2|$

Proposition 11.4

$$T(\theta^*) \subset C(S)$$

where S is the support of θ^*.

Proof: In this proof we define $\Delta_S \in \mathbb{R}^d$ as

$$(\Delta_S)_j = \begin{cases} \Delta_j & j \in S \\ 0 & \text{otherwise} \end{cases}$$

$\forall \Delta \in T(\theta^*)$,

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

Then

$$\|\Delta_{S^c}\|_1 \leqslant \|\Delta_S\|_1$$
$$\Delta \in C(S)$$

■

Proposition 11.5 *Given the above definition, we have*

$$C(S) \subset \bigcup_{\theta:\theta_{S^c}=0} T(\theta)$$

Proof: Say $\Delta \in C(S)$.

In this case, $\Delta \in C(S) \Rightarrow \|\Delta_{S^c}\|_1 \leqslant \|\Delta_S\|_1$. We want to show: $\exists \theta^*$ such that $\theta^*_{S^c} = 0$ and $\Delta \in T(\theta^*)$. By setting $\delta^*_s = -2\Delta_s$, we have:

$$\begin{aligned} \left\|\theta^* + \Delta\right\|_1 &= \left\|\theta^*_s + \Delta_s\right\|_1 + \|\Delta_{S^c}\|_1 \\ &= \left\|\theta^*_s\right\|_1 - \|\Delta_s\|_1 + \|\Delta_{S^c}\|_1 \\ &\leqslant \left\|\theta^*_s\right\|_1 \end{aligned}$$

■

Theorem 11.6 *The following two statements are equivalent:*

(a) For any θ^* with support S, θ^* is the unique solution of the basis pursuit.

(b) $\mathbf{X}$ satisfies the restricted nullspace property with respect to S.

Proof: We first prove $(a) \Longrightarrow (b)$. For a given $\theta^* \in \text{null}(\mathbf{X})\backslash\{\mathbf{0}\}$, consider the basis pursuit problem

$$\min_{\beta \in \mathbb{R}^d} \|\beta\|_1 \;\; s.t. \; \mathbf{X}\beta = \mathbf{X}[\theta^*_S \; 0]^T$$

By assumption, the unique optimal solution will be $\beta' = [\theta_S^* \ 0]^T$. Since $\mathbf{X}\theta^* = 0$, the vector $[0 \ -\theta_{S^c}^*]^T$ is also a solution. By uniqueness, we have $\left\|\beta'\right\|_1 > \|\beta\|_1$. This gives us $\left\|\theta_S^*\right\|_1 < \left\|\theta_{S^c}^*\right\|_1$ and therefore $\theta^* \notin C(S)$.

Then we prove $(b) \implies (a)$. If θ^* is not a unique solution of the basis pursuit, we have $T(\theta^*) \cap \operatorname{null}(\mathbf{X}) \neq \{\mathbf{0}\}$. Since $T(\theta^*) \subset C(S)$, $C(S) \cap \operatorname{null}(\mathbf{X}) \neq \{\mathbf{0}\}$. Thus, $\mathbf{X}$ does not satisfies the restricted nullspace property. ■

11.3 Sufficient conditions for restricted nullspace

In this section, we discuss about the ways to check $C(s) \cap \operatorname{null}(\mathbf{X}) = \{\mathbf{0}\}$. Remember that $\mathbf{X} \in \mathbb{R}^{n \times d}$.

Definition 11.7 *The pairwise incoherence $\delta_{PW}(\mathbf{X})$ is defined as*

$$\delta_{PW}(\mathbf{X}) := \max_{j \neq k} \left| \frac{\langle X_j, X_k \rangle}{n} \right|$$

We hope that $\delta_{PW}(\mathbf{X})$ is small. For an orthogonal $\mathbf{X}$, $\delta_{PW}(\mathbf{X})$ achieve its smallest value 0 for $j \neq k$. On the other hand, if there are two columns X_j and X_k that are really close to each other, it is difficult to say which one is more important. For example, if $X_j = X_k$, we will have $\theta_j X_j + \theta_k X_k = (\theta_j + \theta_k) X_j$, and $\delta_{PW}(\mathbf{X})$ will be large in this case.

Theorem 11.8 *If the pairwise incoherence satisfies the bound*

$$\delta_{PW}(\mathbf{X}) \leqslant \frac{1}{3s}$$

then $\mathbf{X}$ satisfies RNP for all S such that $|S| \leqslant s$.

The definition of pairwise incoherence property can be further extended to the restricted isometric property.

Definition 11.9 $\mathbf{X}$ *satisfies the restricted isometric property (RIP) of order s with constant $\delta_s(\mathbf{X})$ if*

$$|||\mathbf{X}_S^T \mathbf{X}_S / n - \mathbf{I}_s|||_2 \leqslant \delta_s(\mathbf{X})$$

for all S such that $|S| \leqslant s$.

Here, $\mathbf{X}_S$ is defined as the sub-matrix formed by a set of columns in $\mathbf{X}$, where the indices of the columns are defined by S.

The l_2-operation norm of a matrix is defined as its maximum singular value:

$$|||\mathbf{A}|||_2 := \sup_{u \neq 0} \frac{\|\mathbf{A}u\|}{\|u\|}$$

When $s = 1$, the restricted isometric property can be rewritten as:

$$\left| \frac{\left\|X_j\right\|_2^2}{n} - 1 \right| \leqslant \delta_1(\mathbf{X})$$

When $s = 2$, the left hand side can be rewritten as:

$$\frac{\mathbf{X}_S^T \mathbf{X}_S}{n} - \mathbf{I}_s = \begin{bmatrix} \left\|X_j\right\|_2^2 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & \left\|X_k\right\|_2^2 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{11.1}$$

If we assume that all columns of $\mathbf{X}$ are normalized to $\left\|X_j\right\|_2^2 = n$, we have

$$\frac{\mathbf{X}_S^T \mathbf{X}_S}{n} - \mathbf{I}_s = \begin{bmatrix} 0 & \frac{\langle X_j, X_k \rangle}{n} \\ \frac{\langle X_j, X_k \rangle}{n} & 0 \end{bmatrix} \tag{11.2}$$

whose l_2-norm is exactly $\max_{j \neq k} \left| \frac{\langle X_j, X_k \rangle}{n} \right|$, the same as the form of pairwise incoherence $\delta_{PW}(\mathbf{X})$.