STAT 499: Undergraduate Research

Week 6: Recap on main theorems

Facheng Yu

Compiled on: 2024-02-15, 19:28:53

6.1 Sparsity models

Let $\theta^* \in \mathbb{R}^d$ be an unknown regression vector. Suppose that we observe $y \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$ via the linear model:

$$y = \mathbf{X}\theta^* + w$$

Hard sparsity The support set of θ^* is defined as

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

The hard sparsity requires $s := |S(\theta^*)|$ substantially smaller than d. Under the sparsity assumption, we may have a unique linear solution of the least squares estimator.

6.2 Basis pursuit linear program

Basis pursuit linear program. When $w \equiv \mathbf{0} \in \mathbb{R}^n$, consider such a program:

$$\min_{\theta\in\mathbb{R}^d} \|\theta\|_1 \text{ such that } \mathbf{X}\theta = y. \tag{6.1}$$

Assume that there is a vector $\theta^* \in \mathbb{R}^d$ whose support is $S \subset \{1, \ldots, d\}$ such that $y = \mathbf{X}\theta^*$.

Nullspace. $\mathsf{null}(\mathbf{X}) = \{\Delta \in \mathbb{R}^d : \mathbf{X}\Delta = 0\}$. which is the feasible space for (6.1).

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

Proposition 6.1 *If we want the solution of (6.1) to be unique and exactly θ^*, it is equivalent to require that*

$$\mathsf{null}(\mathbf{X}) \cap \mathbb{T}(\theta^*) = \{0\}. \tag{6.2}$$

Proposition 6.2 *Define $\mathbb{C}(S) = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$. Then, it holds that*

$$\mathbb{T}(\theta^*) \subset \mathbb{C}(S).$$

Restricted nullspace property. Based on (6.2), we give the following definition: The matrix $\mathbf{X}$ satisfies the restricted nullspace property with respect to S if

$$\mathsf{null}(\mathbf{X}) \cap \mathbb{C}(S) = \{0\}. \tag{6.3}$$

Theorem 6.3 *If $\mathbf{X}$ satisfies the restricted nullspace property, the following two properties are equivalent:*

(a) For any vector $\theta^* \in \mathbb{R}^d$ with support S, the basis pursuit program (6.1) applied with $y = \mathbf{X}\theta^*$ has unique solution $\widehat{\theta} = \theta^*$.

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

6.3 From basis pursuit program

Suppose that the noise vector $w \in \mathbb{R}^n$ is a non-degenerated random vector.

Extension of the basis pursuit program. The extension relaxes the constraints of the basis pursuit program, i.e., y does not have to $\mathbf{X}\theta$ for some θ. The extended program can be written as

$$\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{such that } \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 \leq b^2 \tag{6.4}$$

for some noise tolerance $b > 0$.

The program above can be shown as equivalent as such a program:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 \right\} \quad \text{such that } \|\theta\|_1 \leq R \tag{6.5}$$

for some radius $R > 0$.

Lasso program. To eliminate the constraint, one can consider the lasso program as well:

$$\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}. \tag{6.6}$$

Here $\lambda_n > 0$ is a regularization parameter to be chosen by the user.

Proposition 6.4 (Equivalent programs) *Suppose that (6.4), (6.5), and (6.6) are convex programs. Then it holds that*

(i) For any $b > 0$, there exists $\lambda \geq 0$ such that program (6.4) and program (6.6) are equivalent;

(ii) For any $R > 0$, there exists $\lambda \geq 0$ such that program (6.5) and program (6.6) are equivalent.

Note: The proof need to use the strong duality of the Lagrangian program and the minimax theorem. And the proof of minimax theorem is provided below.

6.4 Estimation in noisy settings

Extension of restricted nullspace. Define the set

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

RE condition. The matrix $\mathbf{X}$ satisfies the restricted eigenvalue (RE) condition over S with parameters (k, α) if

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq k\|\Delta\|_2^2 \text{ for all } \Delta \in \mathbb{C}_\alpha(S). \tag{6.7}$$

Assumption 6.5 (Lasso assumptions) *Assume that*

($\mathbf{A_1}$) *The vector θ^* is supported on a subset $S \subset \{1, \dots, d\}$ with $|S| = s$.*

($\mathbf{A_2}$) *The design matrix $\mathbf{X}$ satisfies the RE condition with parameter $(k, 3)$:*

$$\frac{1}{n}\|\mathbf{X}\Delta\|_2^2 \geq k\|\Delta\|_2^2 \textit{ for all } \Delta \in \mathbb{C}_3(S).$$

The Lagrangian Lasso is defined as:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \mathbf{X}\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}. \tag{6.8}$$

Theorem 6.6 (Theorem 7.13 in Wainwright's book) *Under assumptions ($\mathbf{A_1}$) and ($\mathbf{A_2}$), for any solution $\widehat{\theta}$ of the Lagrangian Lasso with $\lambda_n \geq 2\|\frac{\mathbf{X}^T w}{n}\|_\infty$, we have*

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{3}{k}\sqrt{s}\lambda_n.$$

6.5 Concentration

Lemma 6.7 (Markov inequality) *For a non-negative random variable X with $\mathbb{E}[X] < \infty$, it holds that, for any $t > 0$,*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

Lemma 6.8 (Concentration for the Gaussian variable) *Suppose that $X \sim \mathcal{N}(0, \sigma^2)$. it holds that, for any $t > 0$,*

$$\mathbb{P}(X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}.$$

Note: To obtain the concentration for the Gaussian variable, we need to use Markov inequality and the moment generating function of the Gaussian variable.

Proposition 6.9 (Concentration for the maxima) *Suppose that $X_1, \ldots, X_d \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$. Then we have*

$$\mathbb{P}(\max\{X_1, \ldots, X_d\} \geq t) \leq d e^{-\frac{t^2}{2\sigma^2}}.$$

Note: The key step is to use the union bound.

6.6 proof of Minimax theorem

Theorem 6.10 (Minimax theorem) *Let $\phi(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Define*

$$p^* = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y)$$

and

$$d^* = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y)$$

It holds that the gap $p^* - d^*$ is zero if:

- $\mathcal{X}, \mathcal{Y}$ are both convex, and one of them is compact.

- The function ϕ is convex-concave: $\phi(\cdot, y)$ is convex for every $y \in \mathcal{Y}$, and $\phi(x, \cdot)$ is concave for every $x \in \mathcal{X}$.

- The function ϕ is continuous.

To show this, we need following lemmas:

Lemma 6.11 *It holds that*

$$\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\phi(x,y)\le\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi(x,y).$$

Proof: For any $x\in\mathcal{X}$, $y\in\mathcal{Y}$

$$\min_{x'\in\mathcal{X}}\phi(x',y)\le\phi(x,y)\le\max_{y'\in\mathcal{Y}}\phi(x,y').$$

Since the inequality holds for any $x\in\mathcal{X}$ and any $y\in\mathcal{Y}$, we can take max on the left and take min on the right:

$$\max_{y\in\mathcal{Y}}\min_{x'\in\mathcal{X}}\phi(x',y)\le\min_{x\in\mathcal{X}}\max_{y'\in\mathcal{Y}}\phi(x,y').$$

■

Lemma 6.12 *The following statements are equivalent:*

(1) There exists $(x^*,y^*)\in\mathcal{X}\times\mathcal{Y}$ such that for any $x\in\mathcal{X}$, $y\in\mathcal{Y}$,

$$\max_{y\in\mathcal{Y}}\phi(x^*,y)\le\phi(x^*,y^*)\le\min_{x\in\mathcal{X}}\phi(x,y^*).$$

(2) The minimax equation holds:

$$\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\phi(x,y)=\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi(x,y)$$

and

$$x^*=\arg\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi(x,y),$$

$$y^*=\arg\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\phi(x,y).$$

Proof: (1) $\implies$ (2): Take min and max on the left and right respectively:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi(x^*,y)\le\phi(x^*,y^*)\le\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\phi(x,y^*).$$

Together with lemma 3.4, we have

$$\phi(x^*,y^*)=\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\phi(x,y)=\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi(x,y).$$

(2) $\implies$ (1): by the definition of x^* and y^*, we have

$$\max_{y\in\mathcal{Y}}\phi(x^*,y)=\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\phi(x,y)=\max_{y\in\mathcal{Y}}\min_{x\in\mathcal{X}}\phi(x,y)=\min_{x\in\mathcal{X}}\phi(x,y^*).$$

Thus,

$$\phi(x^*,y^*)\le\max_{y\in\mathcal{Y}}\phi(x^*,y)=\min_{x\in\mathcal{X}}\phi(x,y^*)\le\phi(x^*,y^*),$$

which implies that $\phi(x^*,y^*)=\max_{y\in\mathcal{Y}}\phi(x^*,y)=\min_{x\in\mathcal{X}}\phi(x,y^*)$. ■