Linear Regression Models

Materials from Philippe Rigollet's Lecture Notes of High Dimensional Statistics

May 17, 2026

Outline

Preliminaries

Definition 1 (Sub-Gaussian Variable)

A random variable $X \in \mathbb{R}$ is said to be $\texttt{sub}-\texttt{Gaussian}$ with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}.$$

In this case we write $X \sim \texttt{subG}(\sigma^2)$.

Preliminaries

Definition 2 (Sub-Gaussian Vector)

A random vector $X \in \mathbb{R}^d$ is said to be $\texttt{sub}-\texttt{Guassian}$ with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and $u^T X$ is sub-Gaussian with variance proxy σ^2 for any unit vector $u \in \mathsf{S}^{d-1}$. In this case we write $X \sim \texttt{subG}_d(\sigma^2)$.

Preliminaries

Definition 3 (Sub-Gaussian Matrix)

A random vector $X \in \mathbb{R}^{d\times T}$ is said to be $\texttt{sub}-\texttt{Guassian}$ with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and $u^T X v$ is sub-Gaussian with variance proxy σ^2 for any unit vector $u \in \mathsf{S}^{d-1}$, $v \in \mathsf{S}^{d-1}$. In this case we write $X \sim \texttt{subG}_{d\times T}(\sigma^2)$.

Preliminaries

Theorem 4

Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2 and $\mathcal{B}_2 = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq 1\}$. Then

$$\mathbb{E}[\max_{\theta \in \mathcal{B}_2} \theta^T X] = \mathbb{E}[\max_{\theta \in \mathcal{B}_2} |\theta^T X|] \leq 4\sigma\sqrt{d} \quad \text{and} \quad \mathbb{P}(\max_{\theta \in \mathcal{B}_2} \theta^T X > t) \leq 6^d e^{-\frac{t^2}{8\sigma^2}},$$

for any $t > 0$. Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds that

$$\max_{\theta \in \mathcal{B}_2} \theta^T X = \max_{\theta \in \mathcal{B}_2} |\theta^T X| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

Preliminaries

Let $\mathcal{V}(\mathrm{P})$ denote a set of finite number of vertices, a convex polytope P is defined as $\mathrm{P} = \mathtt{conv}(\mathcal{V}(\mathrm{P}))$

Preliminaries

Let $\mathcal{V}(\mathsf{P})$ denote a set of finite number of vertices, a convex polytope P is defined as $\mathsf{P} = \texttt{conv}(\mathcal{V}(\mathsf{P}))$

Theorem 5

Let P be a polytope with N vertices $v^{(1)}, \ldots, v^{(N)} \in \mathbb{R}^d$ and let $X \in \mathbb{R}^d$ be a random vector such that, $[v^{(i)}]^T X$, $i = 1, \ldots, N$ are sub-Gaussian random variables with variance proxy σ^2. Then

$$\mathbb{E}[\max_{\theta \in \mathsf{P}} \theta^T X] \leq \sigma\sqrt{2\log N}, \quad \text{and} \quad \mathbb{E}[\max_{\theta \in \mathsf{P}} |\theta^T X|] \leq \sigma\sqrt{2\log(2N)}.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\left(\max_{\theta \in \mathsf{P}} \theta^T X > t\right) \leq Ne^{-\frac{t^2}{2\sigma^2}}, \quad \text{and} \quad \mathbb{P}\left(\max_{\theta \in \mathsf{P}} |\theta^T X| > t\right) \leq 2Ne^{-\frac{t^2}{2\sigma^2}}.$$

Fixed Design Linear Regression

Consider the following model:

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ is sub-Gaussian with variance proxy σ^2 and $\mathbb{E}[\epsilon] = 0$. We assume that $x \in \mathbb{R}^d$ and $f(x) = x^T\theta^*$ for some unknown θ^*.

Fixed Design Linear Regression

Consider the following model:

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, 2, \dots, n,$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$ is sub-Gaussian with variance proxy σ^2 and $\mathbb{E}[\epsilon] = 0$. We assume that $x \in \mathbb{R}^d$ and $f(x) = x^T\theta^*$ for some unknown θ^*.

In contrast to the random design where $x_1, \dots, x_n$ are **stochastic**, the fixed design assumes that $x_1, \dots, x_n$ are **deterministic**. In such setting, we observe $\mu^* = (f(x_1), \dots, f(x_n))^T \in \mathbb{R}^n$ and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \in \mathbb{R}^n$ is sub-Gaussian with variance proxy σ^2.

Fixed Design Linear Regression

We focus on the *Mean Squared Error*(MSE) as a measure of performance. It is defined by

$$\texttt{MSE}(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_n(x_i) - f(x_i)\right)^2.$$

Or equivalently,

$$\texttt{MSE}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i^*)^2 = \|\hat{\mu} - \mu^*\|_2^2$$

Fixed Design Linear Regression

Let the design vectors $x_1, \ldots, x_n \in \mathbb{R}^d$ store in a $n \times d$ matrix $\mathbb{X}$, whose jth row is given by x_j^T. The linear regression model can be written in the matrix form:

$$Y = \mathbb{X}\theta^* + \epsilon, \qquad \text{(Linear Model)}$$

where $Y = (Y_1, \ldots, Y_n)^T$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Moreover,

$$\texttt{MSE}(\mathbb{X}\hat{\theta}) = \frac{1}{n}\|\mathbb{X}(\hat{\theta} - \theta^*)\|_2^2 = \left(\hat{\theta} - \theta^*\right)^T \frac{\mathbb{X}^T\mathbb{X}}{n} \left(\hat{\theta} - \theta^*\right)$$

Unconstrained Least Squares Estimator

Definition 6 (Least Squares Estimator)

Let the least squares estimator $\hat{\theta}^{LS}$ be any vector such that

$$\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|Y - \mathbb{X}\theta\|_2^2.$$

Unconstrained Least Squares Estimator

Definition 6 (Least Squares Estimator)

Let the least squares estimator $\hat{\theta}^{LS}$ be any vector such that

$$\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|Y - \mathbb{X}\theta\|_2^2.$$

We also call $\hat{\mu}^{LS} = \mathbb{X}\hat{\theta}^{LS}$ least squares estimator. By definition of the least squares estimator, $\hat{\mu}^{LS}$ can be regarded as a projection of Y onto the space spans by the columns of $\mathbb{X}$.

Unconstrained Least Squares Estimator

Definition 6 (Least Squares Estimator)

Let the least squares estimator $\hat{\theta}^{LS}$ be any vector such that

$$\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|Y - \mathbb{X}\theta\|_2^2.$$

We also call $\hat{\mu}^{LS} = \mathbb{X}\hat{\theta}^{LS}$ least squares estimator. By definition of the least squares estimator, $\hat{\mu}^{LS}$ can be regarded as a projection of Y onto the space spans by the columns of $\mathbb{X}$.

It is known that the least squares estimators of θ^* and $\mu^* = \mathbb{X}\theta^*$ are maximum likelihood estimators when $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Unconstrained Least Squares Estimator

Proposition 1

The least squares estimator $\hat{\mu}^{LS} = \mathbb{X}\hat{\theta}^{LS}$ satisfies

$$\mathbb{X}^T \hat{\mu}^{LS} = \mathbb{X}^T Y.$$

Moreover, $\hat{\theta}^{LS}$ can be chosen to be

$$\hat{\theta}^{LS} = \left(\mathbb{X}^T \mathbb{X}\right)^{\dagger} \mathbb{X}^T Y,$$

where $(\mathbb{X}^T\mathbb{X})^{\dagger}$ denotes the Moore-Penrose pseudoinverse of $\mathbb{X}^T\mathbb{X}$.

Proof of Proposition 1.

Let $h(\theta) = \|Y - \mathbb{X}\theta\|_2^2$, which is convex and differentiable over $\theta \in \mathbb{R}^d$, so any of its minima $\hat{\theta}^{LS}$ satisfies

$$\nabla_\theta h(\hat{\theta}^{LS}) = 2\mathbb{X}^T(Y - \mathbb{X}\hat{\theta}^{LS}) = 0.$$

Proof of Proposition 1.

Let $h(\theta) = \|Y - \mathbb{X}\theta\|_2^2$, which is convex and differentiable over $\theta \in \mathbb{R}^d$, so any of its minima $\hat{\theta}^{LS}$ satisfies

$$\nabla_\theta h(\hat{\theta}^{LS}) = 2\mathbb{X}^T(Y - \mathbb{X}\hat{\theta}^{LS}) = 0.$$

That is

$$\mathbb{X}^T\mathbb{X}\hat{\theta}^{LS} = \mathbb{X}^T Y.$$

It concludes the proof of the first statement. The second statement follows from the definition of the Moore-Penrose pseudoinverse. □

Unconstrained Least Squares Estimator

Theorem 7

Assume that (Linear Model) holds where $\epsilon \sim \texttt{subG}_n(\sigma^2)$. Then the least squares estimator $\widehat{\theta}^{LS}$ satisfies

$$\mathbb{E}\left[\texttt{MSE}(\mathbb{X}\widehat{\theta}^{LS})\right] = \frac{1}{n}\mathbb{E}\|\mathbb{X}\widehat{\theta}^{LS} - \mathbb{X}\theta^*\|_2^2 \lesssim \sigma^2\frac{r}{n}.$$

where $r = \texttt{rank}(\mathbb{X}^T\mathbb{X})$. Moreover, for any $\delta > 0$, with probability $1-\delta$, it holds

$$\texttt{MSE}(\mathbb{X}\widehat{\theta}^{LS}) \lesssim \sigma^2\frac{r + \log(1/\delta)}{n}.$$

Proof of Theorem 7.

By definition of $\widehat{\theta}^{LS}$,

$$\|Y - \mathbb{X}\widehat{\theta}^{LS}\|_2^2 \leq \|Y - \mathbb{X}\theta^*\|_2^2 = \|\epsilon\|_2^2.$$

Proof of Theorem 7.

By definition of $\hat{\theta}^{LS}$,

$$\|Y - \mathbb{X}\hat{\theta}^{LS}\|_2^2 \leq \|Y - \mathbb{X}\theta^*\|_2^2 = \|\epsilon\|_2^2.$$

Moreover,

$$\|Y - \mathbb{X}\hat{\theta}^{LS}\|_2^2 = \|\mathbb{X}\theta^* + \epsilon - \mathbb{X}\hat{\theta}^{LS}\|_2^2 = \|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}^{LS}\|_2^2 - 2\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*) + \|\epsilon\|_2^2$$

Therefore, we get

$$\|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}^{LS}\|_2^2 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*) = 2\|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}^{LS}\|_2 \frac{\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}^{LS}\|_2}$$

□

Proof of Theorem 7.

As $\widehat{\theta}^{LS}$ depends on ϵ, it is difficult to control the term

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2}$$

To remove this dependency, we apply a technique "sup-out" $\widehat{\theta}^{LS}$

Proof of Theorem 7.

As $\widehat{\theta}^{LS}$ depends on ϵ, it is difficult to control the term

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2}$$

To remove this dependency, we apply a technique "sup-out" $\widehat{\theta}^{LS}$

Let $\Phi = [\varphi_1, \dots, \varphi_2] \in \mathbb{R}^{n \times r}$ be an orthonormal basis of the `range` of $\mathbb{X}$. Hence, there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\widehat{\theta}^{LS} - \theta^*) = \Phi\nu$.

Proof of Theorem 7.

As $\widehat{\theta}^{LS}$ depends on ϵ, it is difficult to control the term

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2}$$

To remove this dependency, we apply a technique "sup-out" $\widehat{\theta}^{LS}$

Let $\Phi = [\varphi_1, \dots, \varphi_2] \in \mathbb{R}^{n \times r}$ be an orthonormal basis of the `range` of $\mathbb{X}$. Hence, there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\widehat{\theta}^{LS} - \theta^*) = \Phi\nu$. It yields

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2} = \frac{\epsilon^T \Phi\nu}{\|\Phi\nu\|_2} = \frac{\epsilon^T \Phi\nu}{\|\nu\|_2} = \frac{\widetilde{\epsilon}^T \nu}{\|\nu\|_2} \leq \sup_{u \in \mathcal{B}_2} \widetilde{\epsilon}^T u.$$

where $\mathcal{B}_2$ is the unit ball of $\mathbb{R}^r$ and $\widetilde{\epsilon} = \Phi^T \epsilon$.

Proof of Theorem 7.

As $\widehat{\theta}^{LS}$ depends on ϵ, it is difficult to control the term

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2}$$

To remove this dependency, we apply a technique "sup-out" $\widehat{\theta}^{LS}$

Let $\Phi = [\varphi_1, \dots, \varphi_2] \in \mathbb{R}^{n \times r}$ be an orthonormal basis of the range of $\mathbb{X}$. Hence, there exists $\nu \in \mathbb{R}^r$ such that $\mathbb{X}(\widehat{\theta}^{LS} - \theta^*) = \Phi\nu$. It yields

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2} = \frac{\epsilon^T \Phi\nu}{\|\Phi\nu\|_2} = \frac{\epsilon^T \Phi\nu}{\|\nu\|_2} = \frac{\tilde{\epsilon}^T \nu}{\|\nu\|_2} \leq \sup_{u \in \mathcal{B}_2} \tilde{\epsilon}^T u.$$

where $\mathcal{B}_2$ is the unit ball of $\mathbb{R}^r$ and $\tilde{\epsilon} = \Phi^T \epsilon$. Thus

$$\|\mathbb{X}\theta^* - \mathbb{X}\widehat{\theta}^{LS}\|_2^2 \leq 4 \sup_{u \in \mathcal{B}_2} (\tilde{\epsilon}^T u)^2.$$

Proof of Theorem 7.

For any $u \in \mathsf{S}^{r-1}$, it holds $\|\Phi u\|_2^2 = u^T \Phi^T \Phi u = u^T u = 1$, which implies $\Phi u \in \mathsf{S}^{r-1}$.

As $\epsilon \sim \mathtt{subG}_n(\sigma^2)$, by definition we have

$$\mathbb{E}\left[e^{s\epsilon^T \Phi u}\right] \leq e^{\frac{s^2\sigma^2}{2}}, \quad \forall s \in \mathbb{R}.$$

Proof of Theorem 7.

For any $u \in \mathsf{S}^{r-1}$, it holds $\|\Phi u\|_2^2 = u^T \Phi^T \Phi u = u^T u = 1$, which implies $\Phi u \in \mathsf{S}^{r-1}$.

As $\epsilon \sim \mathtt{subG}_n(\sigma^2)$, by definition we have

$$\mathbb{E}\left[e^{s\tilde{\epsilon}^T u}\right] = \mathbb{E}\left[e^{s\epsilon^T \Phi u}\right] \leq e^{\frac{s^2\sigma^2}{2}}, \quad \forall s \in \mathbb{R}.$$

Therefore, $\tilde{\epsilon} \sim \mathtt{subG}_r(\sigma^2)$ and $\tilde{\epsilon}_i \sim \mathtt{subG}(\sigma^2)$

Proof of Theorem 7.

For any $u \in \mathsf{S}^{r-1}$, it holds $\|\Phi u\|_2^2 = u^T \Phi^T \Phi u = u^T u = 1$, which implies $\Phi u \in \mathsf{S}^{r-1}$. As $\epsilon \sim \mathtt{subG}_n(\sigma^2)$, by definition we have

$$\mathbb{E}\left[e^{s\tilde{\epsilon}^T u}\right] = \mathbb{E}\left[e^{s\epsilon^T \Phi u}\right] \leq e^{\frac{s^2\sigma^2}{2}}, \quad \forall s \in \mathbb{R}.$$

Therefore, $\tilde{\epsilon} \sim \mathtt{subG}_r(\sigma^2)$ and $\tilde{\epsilon}_i \sim \mathtt{subG}(\sigma^2)$

$$\mathbb{E}[\sup_{u \in \mathcal{B}_2} (\tilde{\epsilon}^T u)^2] = \mathbb{E}\|\tilde{\epsilon}\|_2^2 = \sum_{i=1}^{r} \mathbb{E}[\tilde{\epsilon}_i^2]$$

where the last inequality comes from the property of sub Gaussian variable that $\mathbb{E}[X^2] \leq 4\sigma^2$.

Proof of Theorem 7.

For any $u \in \mathsf{S}^{r-1}$, it holds $\|\Phi u\|_2^2 = u^T \Phi^T \Phi u = u^T u = 1$, which implies $\Phi u \in \mathsf{S}^{r-1}$. As $\epsilon \sim \mathtt{subG}_n(\sigma^2)$, by definition we have

$$\mathbb{E}\left[e^{s\tilde{\epsilon}^T u}\right] = \mathbb{E}\left[e^{s\epsilon^T \Phi u}\right] \leq e^{\frac{s^2\sigma^2}{2}}, \quad \forall s \in \mathbb{R}.$$

Therefore, $\tilde{\epsilon} \sim \mathtt{subG}_r(\sigma^2)$ and $\tilde{\epsilon}_i \sim \mathtt{subG}(\sigma^2)$

$$\mathbb{E}[\sup_{u \in \mathcal{B}_2} (\tilde{\epsilon}^T u)^2] = \mathbb{E}\|\tilde{\epsilon}\|_2^2 = \sum_{i=1}^{r} \mathbb{E}[\tilde{\epsilon}_i^2] \leq 4\sigma^2 r$$

where the last inequality comes from the property of sub Gaussian variable that $\mathbb{E}[X^2] \leq 4\sigma^2$.

Putting together, we conclude the proof the first statement. □

Proof of Theorem 7.

By the Theorem 4, with probability at least $1-\delta$, it holds that

$$\sup_{u\in\mathcal{B}_2}(\tilde{\epsilon}^T u)^2 \leq \left(4\sigma\sqrt{r}+2\sigma\sqrt{2\log(1/\delta)}\right)^2 \leq 32\sigma^2 r + 16\sigma^2\log(1/\delta)$$

where the last inequality comes from the fact that $(a+b)^2 \leq 2a^2+2b^2$. □

If $d \leq n$ and $B := \frac{\mathbb{X}^T\mathbb{X}}{n}$ has rank d, then we have

$$\|\hat{\theta}^{LS} - \theta^*\|_2^2 \leq \frac{\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS})}{\lambda_{\min}(B)}, \quad \textbf{(}\lambda_{\min}(B) > 0 \textbf{ by assumption)}.$$

and we can use Theorem 7 to bound $\|\hat{\theta}^{LS} - \theta^*\|_2^2$ directly.

If $d \leq n$ and $B := \frac{\mathbb{X}^T\mathbb{X}}{n}$ has rank d, then we have

$$\|\hat{\theta}^{LS} - \theta^*\|_2^2 \leq \frac{\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS})}{\lambda_{\min}(B)}, \quad \textbf{(}\lambda_{\min}(B) > 0 \textbf{ by assumption)}.$$

and we can use Theorem 7 to bound $\|\hat{\theta}^{LS} - \theta^*\|_2^2$ directly.

For a $d \times d$ symmetric matrix B, it has eigen-decomposition $U\Sigma U^T$ where $U^TU = UU^T = I$. Any vector $u \in S^{d-1}$ can be expressed by $U\alpha$ for some $\alpha \in \mathbb{R}^d$ satisfied $\|\alpha\|_2 = 1$.

If $d \leq n$ and $B := \frac{\mathbb{X}^T\mathbb{X}}{n}$ has rank d, then we have

$$\|\hat{\theta}^{LS} - \theta^*\|_2^2 \leq \frac{\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS})}{\lambda_{\min}(B)}, \quad \textbf{($\lambda_{\min}(B) > 0$ by assumption)}.$$

and we can use Theorem 7 to bound $\|\hat{\theta}^{LS} - \theta^*\|_2^2$ directly.

For a $d \times d$ symmetric matrix B, it has eigen-decomposition $U\Sigma U^T$ where $U^TU = UU^T = I$. Any vector $u \in S^{d-1}$ can be expressed by $U\alpha$ for some $\alpha \in \mathbb{R}^d$ satisfied $\|\alpha\|_2 = 1$.

Hence, for any $u \in \mathrm{S}^{d-1}$,

$$u^TBu = (U\alpha)^TU\Sigma U^T(U\alpha) = \alpha^T\Sigma\alpha = \sum_{i=1}^{d} \lambda_i(B)\alpha_i^2 \geq \lambda_{\min}(B)\sum_{i=1}^{n}\alpha_i^2 = \lambda_{\min}(B).$$

Constrained Least Squares Estimator

Let $K \subset \mathbb{R}^d$ be a symmetric convex set. If we know a `priori` that $\theta^* \in K$, we may prefer a `constrained least squares` estimator $\hat{\theta}_K^{LS}$ defined by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|Y - \mathbb{X}\theta\|_2^2.$$

Constrained Least Squares Estimator

Let $K \subset \mathbb{R}^d$ be a symmetric convex set. If we know a `priori` that $\theta^* \in K$, we may prefer a `constrained least squares` estimator $\hat{\theta}_K^{LS}$ defined by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|Y - \mathbb{X}\theta\|_2^2.$$

And similarly,

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) \leq 2 \sup_{\theta \in K-K} (\epsilon^T \mathbb{X}\theta)$$

where $K - K \triangleq \{x - y : x, y \in K\}$.

Constrained Least Squares Estimator

Let $K \subset \mathbb{R}^d$ be a symmetric convex set. If we know a `priori` that $\theta^* \in K$, we may prefer a `constrained least squares` estimator $\hat{\theta}_K^{LS}$ defined by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|Y - \mathbb{X}\theta\|_2^2.$$

And similarly,

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) \leq 2 \sup_{\theta \in K-K} (\epsilon^T \mathbb{X}\theta)$$

where $K - K \triangleq \{x - y : x, y \in K\}$. If K is symmetric$(x \in K \Rightarrow -x \in K)$ and convex, then $K - K = 2K$ so that

Constrained Least Squares Estimator

Let $K \subset \mathbb{R}^d$ be a symmetric convex set. If we know a `priori` that $\theta^* \in K$, we may prefer a `constrained least squares` estimator $\hat{\theta}_K^{LS}$ defined by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|Y - \mathbb{X}\theta\|_2^2.$$

And similarly,

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) \leq 2 \sup_{\theta \in K-K} (\epsilon^T \mathbb{X}\theta)$$

where $K - K \triangleq \{x - y : x, y \in K\}$. If K is symmetric$(x \in K \Rightarrow -x \in K)$ and convex, then $K - K = 2K$ so that

$$2 \sup_{\theta \in K-K} (\epsilon^T \mathbb{X}\theta) = 4 \underbrace{\sup_{v \in \mathbb{X}K} (\epsilon^T v)}_{\text{a measure of the size of } \mathbb{X}K},$$

where $\mathbb{X}K \triangleq \{X\theta : \theta \in K\} \subset \mathbb{R}^n$.

ℓ_1 Constrained Least Squares

Theorem 8

Let $K = \mathcal{B}_1$ be the unit ℓ_1 ball of $\mathbb{R}^d$, $d \geq 2$ and assume that θ^*. Moreover, assume the conditions of Theorem 6 and that the columns of $\mathbb{X}$ are normalized in such a way that $\max_j \|\mathbb{X}_j\|_2 \leq \sqrt{n}$. Then the constrained least squares estimator $\hat{\theta}^{LS}_{\mathcal{B}_1}$ satisfies

$$\mathbb{E}\left[\texttt{MSE}(\mathbb{X}\theta^{LS}_{\mathcal{B}_1})\right] = \frac{1}{n}\mathbb{E}\|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*\|_2^2 \lesssim \sigma\sqrt{\frac{\log d}{n}}.$$

Moreover, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1}) \lesssim \sigma\sqrt{\frac{\log(d/\delta)}{n}}.$$

- $\mathcal{B}_1 = \left\{x \in \mathbb{R}^d : \sum_{i=1}^d |x_i| \leq 1\right\}$.

Proof of Theorem 8.

From the considerations preceding the theorem, we got that

$$\|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*\|_2^2 \leq 4 \sup_{v \in \mathbb{X}K} \left(\epsilon^T v\right)^2.$$

Proof of Theorem 8.

From the considerations preceding the theorem, we got that

$$\|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*\|_2^2 \leq 4 \sup_{v \in \mathbb{X}K} \left(\epsilon^T v\right)^2.$$

Note that $K = \mathcal{B}_1 = \text{conv}(\{e_1, -e_1, \ldots, e_d, -e_d\})$ and

$$\mathbb{X}K \subset \mathsf{P}_{\mathbb{X}} \triangleq \text{conv}\{\mathbb{X}_1, -\mathbb{X}_1, \ldots, \mathbb{X}_d, -\mathbb{X}_d\}.$$

Therefore,

$$\|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*\|_2^2 \leq 4 \sup_{v \in \mathbb{X}K} \left(\epsilon^T v\right)^2 \leq 4 \sup_{v \in \mathsf{P}_{\mathbb{X}}} \left(\epsilon^T v\right)^2.$$

Proof of Theorem 8.

From the considerations preceding the theorem, we got that

$$\|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*\|_2^2 \leq 4 \sup_{v \in \mathbb{X}K} \left(\epsilon^T v\right)^2.$$

Note that $K = \mathcal{B}_1 = \texttt{conv}(\{e_1, -e_1, \ldots, e_d, -e_d\})$ and

$$\mathbb{X}K \subset \mathsf{P}_{\mathbb{X}} \triangleq \texttt{conv}\{\mathbb{X}_1, -\mathbb{X}_1, \ldots, \mathbb{X}_d, -\mathbb{X}_d\}.$$

Therefore,

$$\|\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_1} - \mathbb{X}\theta^*\|_2^2 \leq 4 \sup_{v \in \mathbb{X}K} \left(\epsilon^T v\right)^2 \leq 4 \sup_{v \in \mathsf{P}_{\mathbb{X}}} \left(\epsilon^T v\right)^2.$$

Since $\epsilon \sim \texttt{subG}_n(\sigma^2)$, then for any column $\mathbb{X}_j$ such that $\|\mathbb{X}_j\|_2 \leq \sqrt{n}$, the random variable $\epsilon^T \mathbb{X}_j \sim \texttt{subG}(n\sigma^2)$ and so does $-\epsilon^T \mathbb{X}_j \sim \texttt{subG}(n\sigma^2)$ (**Show by Yourself**). □

Proof of Theorem 8.

Applying Theorem 5, we get the bound on $\mathbb{E}[\texttt{MSE}(\mathbb{X}\hat{\theta}_K^{LS})]$:

$$\mathbb{E}[\texttt{MSE}(\mathbb{X}\hat{\theta}_K^{LS})] \leq \frac{4}{n}\mathbb{E}[\sup_{v \in \mathsf{P}_{\mathbb{X}}} \epsilon^T v] \leq \frac{4}{n}\sigma\sqrt{n}\sqrt{2\log 2d} = 4\sigma\sqrt{\frac{2\log 2d}{n}}.$$

Proof of Theorem 8.

Applying Theorem 5, we get the bound on $\mathbb{E}[\texttt{MSE}(\mathbb{X}\hat{\theta}_K^{LS})]$:

$$\mathbb{E}[\texttt{MSE}(\mathbb{X}\hat{\theta}_K^{LS})] \leq \frac{4}{n}\mathbb{E}[\sup_{v\in \mathsf{P}_{\mathbb{X}}} \epsilon^T v] \leq \frac{4}{n}\sigma\sqrt{n}\sqrt{2\log 2d} = 4\sigma\sqrt{\frac{2\log 2d}{n}}.$$

And for any $t > 0$,

$$\mathsf{P}\left(\texttt{MSE}(\mathbb{X}\hat{\theta}_K^{LS})) > t\right) \leq \mathsf{P}\left(\sup_{v\in \mathsf{P}_{\mathbb{X}}} (\epsilon^T v) > nt/4\right) \leq 2de^{-\frac{nt^2}{32\sigma^2}}$$

To conclude the proof, we find t such that

$$2de^{-\frac{nt^2}{32\sigma^2}} \leq \delta \Leftrightarrow t^2 \geq \frac{32\sigma^2}{n}\log\frac{2d}{\delta}$$

□

ℓ_0 constrained least squares

Define the ℓ_0 (pseudo) norm of vector $\theta \in \mathbb{R}^d$ as

$$\|\theta\|_0 = \sum_{j=1}^{d} \mathbb{1}(\theta_j \neq 0).$$

A vector with "small" ℓ_0 norm is called a `sparse` vector. More precisely, θ is a k-`sparse` vector if $\|\theta\|_0 \leq k$.

ℓ_0 constrained least squares

Define the ℓ_0 (pseudo) norm of vector $\theta \in \mathbb{R}^d$ as

$$\|\theta\|_0 = \sum_{j=1}^{d} \mathbb{1}(\theta_j \neq 0).$$

A vector with "small" ℓ_0 norm is called a `sparse` vector. More precisely, θ is a k-`sparse` vector if $\|\theta\|_0 \leq k$.

The `support` of θ is defined as

$$\texttt{supp}(\theta) = \{j \in \{1, \ldots, d\} : \theta_j \neq 0\},$$

and $\|\theta\|_0 = \texttt{card}(\texttt{supp}(\theta)) := |\texttt{supp}(\theta)|$.

$$\lim_{q \to 0^+} \sum_{j=1}^{d} |\theta_j|^q = \|\theta\|_0.$$

ℓ_0 constrained least squares

Denote $\mathcal{B}_0(k)$ the ℓ_0 ball of $\mathbb{R}^d$, i.e., the set of k-sparse, defined by

$$\mathcal{B}_0(k) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\}.$$

ℓ_0 constrained least squares

Denote $\mathcal{B}_0(k)$ the ℓ_0 ball of $\mathbb{R}^d$, i.e., the set of k-sparse, defined by

$$\mathcal{B}_0(k) = \{\theta \in \mathbb{R}^d : \|\theta\|_0 \leq k\}.$$

Theorem 9

Fix a positive integer $k \leq d/2$. Let $K = \mathcal{B}_0(k)$ be set of k-sparse vectors of $\mathbb{R}^d$ and assume that $\theta^* \in \mathcal{B}_0(k)$. Moreover, assume the conditions of Theorem 7. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$\mathtt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2}{n} \log \binom{d}{2k} + \frac{\sigma^2 k}{n} + \frac{\sigma^2}{n} \log(1/\delta).$$

Proof of Theorem 9.

Similar to the proof of Theorem 7, we get

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) = 2\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2 \frac{\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2}$$

Proof of Theorem 9.

Similar to the proof of Theorem 7, we get

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) = 2\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2 \frac{\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2}$$

Since both $\hat{\theta}_K^{LS}$ and θ^* are in $\mathcal{B}_0(k)$, we have $\hat{\theta}_K^{LS} - \theta^* \in \mathcal{B}_0(2k)$.

Proof of Theorem 9.

Similar to the proof of Theorem 7, we get

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) = 2\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2 \frac{\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2}$$

Since both $\hat{\theta}_K^{LS}$ and θ^* are in $\mathcal{B}_0(k)$, we have $\hat{\theta}_K^{LS} - \theta^* \in \mathcal{B}_0(2k)$.

For any $S \subset \{1, \ldots, d\}$, let $\mathbb{X}_S$ the $n \times |S|$ submatrix of $\mathbb{X}$ that is obtained from the column of $\mathbb{X}_j$, $j \in S$ of $\mathbb{X}$. Denote by $r_S \leq |S|$ the rank of $\mathbb{X}_S$ and let $\Phi_S = [\phi_1, \ldots, \phi_{r_S}] \in \mathbb{R}^{n \times r_S}$ be an orthonormal basis of the column span of $\mathbb{X}_S$.

Proof of Theorem 9.

Similar to the proof of Theorem 7, we get

$$\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) = 2\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2 \frac{\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}\hat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2}$$

Since both $\hat{\theta}_K^{LS}$ and θ^* are in $\mathcal{B}_0(k)$, we have $\hat{\theta}_K^{LS} - \theta^* \in \mathcal{B}_0(2k)$.

For any $S \subset \{1, \ldots, d\}$, let $\mathbb{X}_S$ the $n \times |S|$ submatrix of $\mathbb{X}$ that is obtained from the column of $\mathbb{X}_j$, $j \in S$ of $\mathbb{X}$. Denote by $r_S \leq |S|$ the rank of $\mathbb{X}_S$ and let $\Phi_S = [\phi_1, \ldots, \phi_{r_S}] \in \mathbb{R}^{n \times r_S}$ be an orthonormal basis of the column span of $\mathbb{X}_S$.

Moreover, for any $\theta \in \mathbb{R}^d$, define $\theta(S) \in \mathbb{R}^{|S|}$ be the vector with coordinates θ_j, $j \in S$. □

Proof of Theorem 9.

Denote by $\widehat{S} = \text{supp}(\widehat{\theta}_K^{LS} - \theta^*)$, we have $|\widehat{S}| \leq 2k$ and there exists $\nu \in \mathbb{R}^{r_{\widehat{S}}}$ such that

$$\mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*) = \mathbb{X}_{\widehat{S}} \left(\widehat{\theta}_K^{LS}(\widehat{S}) - \theta^*(\widehat{S}) \right) = \Phi_{\widehat{S}} \nu.$$

Proof of Theorem 9.

Denote by $\widehat{S} = \text{supp}(\widehat{\theta}_K^{LS} - \theta^*)$, we have $|\widehat{S}| \leq 2k$ and there exists $\nu \in \mathbb{R}^{r_{\hat{s}}}$ such that

$$\mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*) = \mathbb{X}_{\widehat{S}} \left(\widehat{\theta}_K^{LS}(\widehat{S}) - \theta^*(\widehat{S}) \right) = \Phi_{\widehat{S}} \nu.$$

Therefore,

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*)\|_2} = \frac{\epsilon^T \Phi_{\widehat{S}} \nu}{\|\nu\|_2} \leq \max_{|S| \leq 2k} \sup_{u \in \mathcal{B}_2^{r_S}} [\epsilon^T \Phi_S] u = \max_{|S| = 2k} \sup_{u \in \mathcal{B}_2^{r_S}} [\epsilon^T \Phi_S] u,$$

where the last equality comes from the fact the $\mathbb{R}^{|S|}$ is the subspace of $\mathbb{R}^{2k}$ if $|S| \leq 2k$ and $\mathcal{B}_2^{r_S}$ is the unit ball of $\mathbb{R}^{r_S}$.

Proof of Theorem 9.

Denote by $\widehat{S} = \text{supp}(\widehat{\theta}_K^{LS} - \theta^*)$, we have $|\widehat{S}| \leq 2k$ and there exists $\nu \in \mathbb{R}^{r_{\widehat{S}}}$ such that

$$\mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*) = \mathbb{X}_{\widehat{S}} \left(\widehat{\theta}_K^{LS}(\widehat{S}) - \theta^*(\widehat{S}) \right) = \Phi_{\widehat{S}} \nu.$$

Therefore,

$$\frac{\epsilon^T \mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}(\widehat{\theta}_K^{LS} - \theta^*)\|_2} = \frac{\epsilon^T \Phi_{\widehat{S}} \nu}{\|\nu\|_2} \leq \max_{|S| \leq 2k} \sup_{u \in \mathcal{B}_2^{r_S}} [\epsilon^T \Phi_S] u = \max_{|S| = 2k} \sup_{u \in \mathcal{B}_2^{r_S}} [\epsilon^T \Phi_S] u,$$

where the last equality comes from the fact the $\mathbb{R}^{|S|}$ is the subspace of $\mathbb{R}^{2k}$ if $|S| \leq 2k$ and $\mathcal{B}_2^{r_S}$ is the unit ball of $\mathbb{R}^{r_S}$. It yields

$$\|\mathbb{X}\widehat{\theta}_K^{LS} - \mathbb{X}\theta^*\|_2^2 \leq 4 \max_{|S| = 2k} \sup_{u \in \mathcal{B}_2^{r_S}} (\widetilde{\epsilon}^T u)^2,$$

$\widetilde{\epsilon}_S = \Phi_S^T \epsilon \sim \texttt{subG}_{r_S}(\sigma)$. □

Proof of Theorem 9.

Using a union bound, we get for any $t > 0$,

$$\mathbb{P}\left(\max_{|S|=2k} \sup_{u \in \mathcal{B}_2^{r_S}} (\tilde{\epsilon}^T u)^2 > t\right) \leq \sum_{|S|=2k} \mathbb{P}\left(\sup_{u \in \mathcal{B}_2^{r_S}} (\tilde{\epsilon}^T u)^2 > t\right)$$

It follows from the proof of Theorem 4 that for any $|S| = 2k$,

$$\mathbb{P}\left(\sup_{u \in \mathcal{B}_2^{r_S}} (\tilde{\epsilon}^T u)^2 > t\right) \leq 6^{2k} e^{-\frac{t}{8\sigma^2}}, \quad \forall t > 0.$$

□

Proof of Theorem 9.

Putting together, for any $t > 0$,

$$\begin{aligned}\mathbb{P}(\texttt{MSE}(\mathbb{X}\hat{\theta}_K^{LS}) > t) &\leq \mathbb{P}(\frac{4}{n}\max_{|S|=2k}\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde{\epsilon}^T u)^2 > t) \\ &\leq \sum_{|S|=2k}\mathbb{P}\left(\sup_{u\in\mathcal{B}_2^{r_S}}(\tilde{\epsilon}^T u)^2 > nt/4\right) \\ &\leq \binom{d}{2k}6^{2k}e^{-\frac{nt}{32\sigma^2}}\end{aligned}$$

and

$$\binom{d}{2k}6^{2k}e^{-\frac{nt}{32\sigma^2}} \leq \delta \Leftrightarrow t > \frac{32\sigma^2}{n}\left(\log\binom{d}{2k} + 2k\log(6) + \log\frac{1}{\delta}\right)$$

□

Lemma 10

For any integers $1 \leq k \leq n$, it holds

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

Lemma 10

For any integers $1 \leq k \leq n$, it holds

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

Proof.

Show by in induction and note that $(1 + \frac{1}{k})^k \leq e$. □

ℓ_0 constrained least squares

Corollary 11

Under the assumption of Theorem 9, for any $\delta > 0$, with probability $1 - \delta$, it holds

$$\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_0(k)}) \lesssim \frac{\sigma^2 k}{n} \log\left(\frac{ed}{2k}\right) + \frac{\sigma^2 k}{n} \log(6) + \frac{\sigma^2}{n} \log(1/\delta).$$

ℓ_0 constrained least squares

Corollary 12

Under the assumption of Theorem 9,

$$\mathbb{E}\left[\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_0(k)})\right] \lesssim \frac{\sigma^2 k}{n}\log\left(\frac{ed}{k}\right).$$

Proof of Corollary 12.

For any $H \geq 0$

$$\begin{aligned}\mathbb{E}\left[\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_0(k)})\right] &= \int_0^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t \\ &= \int_0^H \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t + \int_H^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t\end{aligned}$$

Proof of Corollary 12.

For any $H \geq 0$

$$\begin{aligned}
\mathbb{E}\left[\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_0(k)})\right] &= \int_0^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t \\
&= \int_0^H \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t + \int_H^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t \\
&\leq H + \int_0^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq n(\xi + H))\mathrm{d}\xi
\end{aligned}$$

Proof of Corollary 12.

For any $H \geq 0$

$$\begin{aligned}
\mathbb{E}\left[\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}_{\mathcal{B}_0(k)})\right] &= \int_0^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t \\
&= \int_0^H \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t + \int_H^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq nt)\mathrm{d}t \\
&\leq H + \int_0^\infty \mathbb{P}(\|\mathbb{X}\hat{\theta}^{LS}_K - \mathbb{X}\theta^*\| \geq n(\xi + H))\mathrm{d}\xi \\
&\leq H + \binom{d}{2k} 6^{2k} \int_0^\infty e^{-\frac{n(\xi+H)}{32\sigma^2}} \mathrm{d}\xi \\
&= H + \binom{d}{2k} 6^{2k} e^{-\frac{nH}{32\sigma^2}} \frac{32\sigma^2}{n},
\end{aligned}$$

where the second inequality comes from the proof of Theorem □

Proof of Corollary 12.

Take H such that

$$\binom{d}{2k} 6^{2k} e^{-\frac{nH}{32\sigma^2}} = 1.$$

It yields

$$H \lesssim \frac{\sigma^2 k}{n} \log\left(\frac{ed}{n}\right),$$

which completes the proof. □

The Gaussian Sequence Model

The Gaussian sequence model is as follows:

$$Y_i = \theta_i^* + \epsilon_i, \quad i = 1, \ldots, d \qquad \text{(Gaussian Sequence Model)}$$

where $\epsilon_1, \ldots, \epsilon_d$ are i.i.d $\mathcal{N}(0, \sigma^2)$ random variables. The goal here is to estimate the unknown vector θ^*.

The Gaussian Sequence Model

The Gaussian sequence model is as follows:

$$Y_i = \theta_i^* + \epsilon_i, \quad i = 1, \dots, d \qquad \text{(Gaussian Sequence Model)}$$

where $\epsilon_1, \dots, \epsilon_d$ are i.i.d $\mathcal{N}(0, \sigma^2)$ random variables. The goal here is to estimate the unknown vector θ^*.

Note that (Gaussian Sequence Model) is a special case of the fixed design (Linear Model) with $n = d$, and $\mathbb{X} = I_d$.

The sub-Gaussian Sequence Model

Assumption ORT

The design matrix satisfies

$$\frac{\mathbb{X}^T\mathbb{X}}{n} = I_d,$$

where I_d denotes the identity matrix of $\mathbb{R}^d$.

The sub-Gaussian Sequence Model

Assumption ORT

The design matrix satisfies

$$\frac{\mathbb{X}^T\mathbb{X}}{n} = I_d,$$

where I_d denotes the identity matrix of $\mathbb{R}^d$.

- Assumption ORT allows for cases where $d \leq n$ but not $d > n$ (high dimensional case) due to rank constraint. $\left(\texttt{rank}(\frac{\mathbb{X}^T\mathbb{X}}{n}) = d \leq \min\{d, n\}\right)$
- The d columns of $\mathbb{X}$ are orthogonal in $\mathbb{R}^n$ and all have norm $\sqrt{n}$.

The sub-Gaussian Sequence Model

Under Assumption ORT, it follows from (Linear Model) that

$$y := \frac{1}{n}\mathbb{X}^T Y = \frac{\mathbb{X}^T\mathbb{X}}{n}\theta^* + \frac{1}{n}\mathbb{X}^T\epsilon$$
$$= \theta^* + \xi,$$

where $\xi = (\xi_1, \ldots, \xi_d) \sim \mathtt{subG}_d(\sigma^2/n)$.

The sub-Gaussian Sequence Model

Under Assumption ORT, it follows from (Linear Model) that

$$y := \frac{1}{n}\mathbb{X}^T Y = \frac{\mathbb{X}^T\mathbb{X}}{n}\theta^* + \frac{1}{n}\mathbb{X}^T\epsilon$$
$$= \theta^* + \xi,$$

where $\xi = (\xi_1, \ldots, \xi_d) \sim \mathtt{subG}_d(\sigma^2/n)$.

Under Assumption ORT,

- The Linear Model is equivalent to the sub-Gaussian Sequence Model Gaussian Sequence Model up a transformation of the data Y and a change of variable for the variance.
- For any $\hat{\theta} \in \mathbb{R}^d$,

$$\mathtt{MSE}(\mathbb{X}\hat{\theta}) = (\hat{\theta} - \theta^*)^T \frac{\mathbb{X}^T\mathbb{X}}{n}(\hat{\theta} - \theta^*) = \|\hat{\theta} - \theta^*\|_2^2$$

The sub-Gaussian Sequence Model

For any $\theta \in \mathbb{R}^d$, Assumption ORT yields,

$$\begin{aligned}\|y-\theta\|_2^2 &= \|\frac{1}{n}\mathbb{X}^T Y - \theta\|_2^2 \\ &= \|\theta\|_2^2 - \frac{2}{n}\theta^T \mathbb{X} Y + \frac{1}{n^2} Y^T \mathbb{X}\mathbb{X}^T Y \\ &= \frac{1}{n}\|\mathbb{X}\theta\|_2^2 - \frac{2}{n}(\mathbb{X}\theta)^T Y + \frac{1}{n}\|Y\|_2^2 + Q \\ &= \frac{1}{n}\|Y - \mathbb{X}\theta\|_2^2 + Q,\end{aligned}$$

(Equivalence to LSE)

where Q is a constant that does not depend on θ and is defined by

$$Q = \frac{1}{n^2} = \frac{1}{n^2} Y^T \mathbb{X}\mathbb{X}^T Y - \frac{1}{n}\|Y\|_2^2.$$

The sub-Gaussian Sequence Model

For any $\theta \in \mathbb{R}^d$, Assumption ORT yields,

$$\begin{aligned}\|y-\theta\|_2^2 &= \|\frac{1}{n}\mathbb{X}^T Y - \theta\|_2^2 \\ &= \|\theta\|_2^2 - \frac{2}{n}\theta^T \mathbb{X} Y + \frac{1}{n^2} Y^T \mathbb{X}\mathbb{X}^T Y \\ &= \frac{1}{n}\|\mathbb{X}\theta\|_2^2 - \frac{2}{n}(\mathbb{X}\theta)^T Y + \frac{1}{n}\|Y\|_2^2 + Q \\ &= \frac{1}{n}\|Y - \mathbb{X}\theta\|_2^2 + Q,\end{aligned}$$

(Equivalence to LSE)

where Q is a constant that does not depend on θ and is defined by

$$Q = \frac{1}{n^2} = \frac{1}{n^2} Y^T \mathbb{X}\mathbb{X}^T Y - \frac{1}{n}\|Y\|_2^2.$$

This implies that the least squares estimator $\widehat{\theta}^{LS}$ is equal to y (By the optimal condition).

The sub-Gaussian Sequence Model

The preceding discussion to be summarized by a sightly more general model called $\texttt{sub} - \texttt{Gaussian sequence model}$:

$$y = \theta^* + \xi \quad \in \mathbb{R}^d \qquad \text{(Sub-Gaussian Sequence Model)}$$

where $\xi \sim \texttt{subG}_d(\sigma^2/n)$.

The sub-Gaussian Sequence Model

The preceding discussion to be summarized by a sightly more general model called $\texttt{sub} - \texttt{Gaussian sequence model}$:

$$y = \theta^* + \xi \quad \in \mathbb{R}^d \qquad \text{(Sub-Gaussian Sequence Model)}$$

where $\xi \sim \texttt{subG}_d(\sigma^2/n)$.

- We define this model independently of Assumption ORT and thus for any values of n and d.

Sparsity Adaptive Thresholding Estimators

If we knew a prior that θ was k sparse, we could employ directly Corollary 11 to obtain that with probability $1-\delta$, we have

$$\texttt{MSE}(\mathbb{X}\widehat{\theta}^{LS}_{\mathcal{B}_0(k)}) \leq C_\delta \frac{\sigma^2 k}{n} \log\left(\frac{ed}{2k}\right). \qquad (\ell_0 \text{ constraint estimator with known sparsity})$$

Sparsity Adaptive Thresholding Estimators

Assume the (Sub-Gaussian Sequence Model). If nothing is known about θ^*, it is natural to estimate it using the least squares estimator $\hat{\theta}^{LS} = y$. In this case,

$$\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}) = \|y - \theta^*\|_2^2 = \|\xi\| \leq C_\delta \frac{\sigma^2 d}{n},$$

where the last inequality holds with probability at least $1 - \delta$. It is consistent with (ℓ_0 constraint estimator with known sparsity) if $k = Cd$ for some positive constant $C \leq 1$.

Sparsity Adaptive Thresholding Estimators

Assume the (Sub-Gaussian Sequence Model). If nothing is known about θ^*, it is natural to estimate it using the least squares estimator $\hat{\theta}^{LS} = y$. In this case,

$$\texttt{MSE}(\mathbb{X}\hat{\theta}^{LS}) = \|y - \theta^*\|_2^2 = \|\xi\| \leq C_\delta \frac{\sigma^2 d}{n},$$

where the last inequality holds with probability at least $1 - \delta$. It is consistent with (ℓ_0 constraint estimator with known sparsity) if $k = Cd$ for some positive constant $C \leq 1$.

However, this approach does not use the fact that k may be much smaller than d, which happens when θ^* has many zero coordinate.

Sparsity Adaptive Thresholding Estimators

If $\theta_j^* = 0$, then, $y_j = \xi_j$, a sub-Gaussian variable with variance proxy σ^2/n. In particular, we know that with probability at least $1 - \delta$,

$$|\xi_j| \leq \sigma \sqrt{\frac{2\log(2/\delta)}{n}} = \tau.$$

Sparsity Adaptive Thresholding Estimators

If $\theta_j^* = 0$, then, $y_j = \xi_j$, a sub-Gaussian variable with variance proxy σ^2/n. In particular, we know that with probability at least $1 - \delta$,

$$|\xi_j| \leq \sigma \sqrt{\frac{2\log(2/\delta)}{n}} = \tau.$$

The consequences of this inequality are interesting:

- **If we observe $|y_i| \gg \tau$, then it must correspond to $\theta_j^* \neq 0$.**

Sparsity Adaptive Thresholding Estimators

If $\theta_j^* = 0$, then, $y_j = \xi_j$, a sub-Gaussian variable with variance proxy σ^2/n. In particular, we know that with probability at least $1 - \delta$,

$$|\xi_j| \leq \sigma \sqrt{\frac{2\log(2/\delta)}{n}} = \tau.$$

The consequences of this inequality are interesting:

- **If we observe $|y_i| \gg \tau$, then it must correspond to $\theta_j^* \neq 0$.**
- **If $|y_j| \leq \tau$ is smaller, then θ_j^* cannot be very large. By the triangle inequality, $|\theta_j^*| \leq |y_j| + |\xi_j| \leq 2\tau$. Therefore, we loose at most 2τ by choosing $\widehat{\theta}_j = 0$.**

Sparsity Adaptive Thresholding Estimators

If $\theta_j^* = 0$, then, $y_j = \xi_j$, a sub-Gaussian variable with variance proxy σ^2/n. In particular, we know that with probability at least $1 - \delta$,

$$|\xi_j| \leq \sigma \sqrt{\frac{2\log(2/\delta)}{n}} = \tau.$$

The consequences of this inequality are interesting:

- **If we observe $|y_i| \gg \tau$, then it must correspond to $\theta_j^* \neq 0$.**
- **If $|y_j| \leq \tau$ is smaller, then θ_j^* cannot be very large. By the triangle inequality, $|\theta_j^*| \leq |y_j| + |\xi_j| \leq 2\tau$. Therefore, we loose at most 2τ by choosing $\widehat{\theta}_j = 0$.**

Such observations lead us to consider the **hard thresholding estimator**.

Sparsity Adaptive Thresholding Estimators

Definition 13 (Hard Thresholding Estimator)

The hard thresholding estimator with threshold $2\tau > 0$ is denoted by $\widehat{\theta}^{\texttt{HRD}}$ and has coordinates

$$\widehat{\theta}_j^{\texttt{HRD}} = \begin{cases} y_j & \text{if } |y_j| > 2\tau, \\ 0 & \text{if } |y_j| \leq 2\tau, \end{cases}$$

for $j = 1, \ldots, d$. In short, we can write $\widehat{\theta}_j^{\texttt{HRD}} = y_j \mathbb{1}(|y_j| > 2\tau)$

Sparsity Adaptive Thresholding Estimators

Definition 13 (Hard Thresholding Estimator)

The hard thresholding estimator with threshold $2\tau > 0$ is denoted by $\widehat{\theta}^{\text{HRD}}$ and has coordinates

$$\widehat{\theta}_j^{\text{HRD}} = \begin{cases} y_j & \text{if } |y_j| > 2\tau, \\ 0 & \text{if } |y_j| \leq 2\tau, \end{cases}$$

for $j = 1, \ldots, d$. In short, we can write $\widehat{\theta}_j^{\text{HRD}} = y_j \mathbb{1}(|y_j| > 2\tau)$

For each individual ξ_j, the inequality

$$|\xi_j| \leq \sigma \sqrt{\frac{2\log(2/\delta)}{n}} \quad \text{holds with prob. at least } 1 - \delta.$$

For the hard thresholding estimator, we should consider τ such that $|\xi_j| \leq \tau$ holds simultaneously for all j, which can be achieved by controlling the maxima.

Sparsity Adaptive Thresholding Estimators

By the inequality of suprema of sub-Gaussian variables, we have

$$\max_{1\le j\le d} |\xi_j| \le \sigma\sqrt{\frac{2\log(2d/\delta)}{n}} \quad \text{holds with prob. at least } 1-\delta.$$

It yields the following theorem.

Sparsity Adaptive Thresholding Estimators

Theorem 14

Consider the (Linear Model) under the Assumption (ORT) or, equivalently, the (Sub-Gaussian Sequence Model). Then the hard thresholding estimator $\widehat{\theta}^{\mathtt{HRD}}$ with threshold

$$2\tau = 2\sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$$

enjoys the following two properties on the same event $\mathcal{A}$ such that $\mathbb{P}(\mathcal{A}) \geq 1-\delta$:

(1) If $\|\theta\|_0 = k$,

$$\mathtt{MSE}(\mathbb{X}\widehat{\theta}^{\mathtt{HRD}}) = \|\widehat{\theta}^{\mathtt{HRD}} - \theta^*\|_2^2 \lesssim \sigma^2 \frac{k\log(2d/\delta)}{n}.$$

(2) If $\min_{j\in \mathtt{supp}(\theta^*)} > 3\tau$, then

$$\mathtt{supp}(\widehat{\theta}^{\mathtt{HRD}}) = \mathtt{supp}(\theta^*).$$

Proof of Theorem 14.

Define the event

$$\mathcal{A} = \left\{ \max_j |\xi_j| \leq \tau \right\}$$

and by the inequality of maxima of sub-Gaussian variables $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$.

Proof of Theorem 14.

Define the event

$$\mathcal{A} = \left\{ \max_j |\xi_j| \leq \tau \right\}$$

and by the inequality of maxima of sub-Gaussian variables $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$. On the event $\mathcal{A}$, the following holds for any $j = 1, \ldots, d$. First, observe that

$$|y_j| > 2\tau \Rightarrow |\theta_j^*| \geq |y_j| - |\xi_j| > \tau \qquad (\clubsuit)$$

and

$$|y_j| \leq 2\tau \Rightarrow |\theta_j^*| \leq |y_j| + |\xi_j| \leq 3\tau. \qquad (\spadesuit)$$

□

Proof of Theorem 14.

It yields

$$\begin{aligned}|\widehat{\theta}_j^{\texttt{HRD}} - \theta_j^*| &= |y_j - \theta_j^*|\mathbb{1}(|y_j| > 2\tau) + |\theta_j^*|\mathbb{1}(|y_j| \le 2\tau) \\ &\le \tau\mathbb{1}(|y_j| > 2\tau) + |\theta_j^*|\mathbb{1}(|y_j| \le 2\tau)\end{aligned}$$

Proof of Theorem 14.

It yields

$$
\begin{aligned}
|\widehat{\theta}_j^{\texttt{HRD}} - \theta_j^*| &= |y_j - \theta_j^*| \mathbb{1}(|y_j| > 2\tau) + |\theta_j^*| \mathbb{1}(|y_j| \le 2\tau) \\
&\le \tau \mathbb{1}(|y_j| > 2\tau) + |\theta_j^*| \mathbb{1}(|y_j| \le 2\tau) \\
&\le \tau \mathbb{1}(|\theta_j^*| > \tau) + |\theta_j^*| \mathbb{1}(|\theta_j^*| \le 3\tau) && \text{by } (\clubsuit) \text{ and } (\spadesuit) \\
&\le 4 \min(|\theta_j^*|, \tau). && \textbf{(Check by Yourself)}
\end{aligned}
$$

Proof of Theorem 14.

It yields

$$\begin{aligned}
|\widehat{\theta}_j^{\texttt{HRD}} - \theta_j^*| &= |y_j - \theta_j^*|\mathbb{1}(|y_j| > 2\tau) + |\theta_j^*|\mathbb{1}(|y_j| \le 2\tau) \\
&\le \tau\mathbb{1}(|y_j| > 2\tau) + |\theta_j^*|\mathbb{1}(|y_j| \le 2\tau) \\
&\le \tau\mathbb{1}(|\theta_j^*| > \tau) + |\theta_j^*|\mathbb{1}(|\theta_j^*| \le 3\tau) && \text{by } (\clubsuit) \text{ and } (\spadesuit) \\
&\le 4\min(|\theta_j^*|, \tau). && \textbf{(Check by Yourself)}
\end{aligned}$$

It yields

$$\begin{aligned}
\|\widehat{\theta}^{\texttt{HRD}} - \theta^*\|_2^2 &= \sum_{j=1} 6d|\widehat{\theta}^{\texttt{HRD}} - \theta_j^*|^2 \\
&\le 16\sum_{j=1}^{d} \min(|\theta_j|^2, \tau^2) \le 16\|\theta\|_0\tau^2.
\end{aligned}$$

This complete the proof of (1). □

Proof of Theorem 14.

To prove (2), note that if $\theta_j^* \neq 0$, then $|\theta_j^*| > 3\tau$. It yields

$$|y_j| = |\theta_j^* + \xi| \geq |\theta_j^*| - |\xi| > 3\tau - \tau = 2\tau.$$

Therefore $\widehat{\theta}_j^{\texttt{HRD}} \neq 0$ so that $\texttt{supp}(\theta^*) \subset \texttt{supp}(\widehat{\theta}^{\texttt{HRD}})$.

Proof of Theorem 14.

To prove (2), note that if $\theta_j^* \neq 0$, then $|\theta_j^*| > 3\tau$. It yields

$$|y_j| = |\theta_j^* + \xi| \geq |\theta_j^*| - |\xi| > 3\tau - \tau = 2\tau.$$

Therefore $\widehat{\theta}_j^{\mathtt{HRD}} \neq 0$ so that $\mathtt{supp}(\theta^*) \subset \mathtt{supp}(\widehat{\theta}^{\mathtt{HRD}})$.

Next, if $\widehat{\theta}_j^{\mathtt{HRD}} \neq 0$, then $|\widehat{\theta}^{\mathtt{HRD}}| = |y_j| > 2\tau$. It yields

$$|\theta_j^*| = |y_j - \xi_j| > 2\tau - \tau > \tau.$$

Therefore, $|\theta_j^*| \neq 0$ and $\mathtt{supp}(\widehat{\theta}^{\mathtt{HRD}}) \subset \mathtt{supp}(\theta^*)$.

Proof of Theorem 14.

To prove (2), note that if $\theta_j^* \neq 0$, then $|\theta_j^*| > 3\tau$. It yields

$$|y_j| = |\theta_j^* + \xi| \geq |\theta_j^*| - |\xi| > 3\tau - \tau = 2\tau.$$

Therefore $\widehat{\theta}_j^{\mathtt{HRD}} \neq 0$ so that $\mathtt{supp}(\theta^*) \subset \mathtt{supp}(\widehat{\theta}^{\mathtt{HRD}})$.

Next, if $\widehat{\theta}_j^{\mathtt{HRD}} \neq 0$, then $|\widehat{\theta}^{\mathtt{HRD}}| = |y_j| > 2\tau$. It yields

$$|\theta_j^*| = |y_j - \xi_j| > 2\tau - \tau > \tau.$$

Therefore, $|\theta_j^*| \neq 0$ and $\mathtt{supp}(\widehat{\theta}^{\mathtt{HRD}}) \subset \mathtt{supp}(\theta^*)$.

Hence, it concludes that $\mathtt{supp}(\widehat{\theta}^{\mathtt{HRD}}) = \mathtt{supp}(\theta^*)$ □

Sparsity Adaptive Thresholding Estimators

Similar results can be obtained for the **soft thresholding** estimator $\hat{\theta}^{\mathtt{SFT}}$ defined by

$$\hat{\theta}_j^{\mathtt{SFT}} = \begin{cases} y_j - 2\tau & \text{if } y_j > 2\tau, \\ y_j + 2\tau & \text{if } y_j < -2\tau, \\ 0 & \text{if } |y_j| \leq 2\tau, \end{cases}$$

In short, we can write

$$\hat{\theta}_j^{\mathtt{SFT}} = \left(1 - \frac{2\tau}{|y_j|}\right)_+ y_j.$$

The BIC and Lasso Estimators

It can be shown that the hard and soft thresholding estimators are solutions of the following penalized empirical risk minimization problems:

$$\widehat{\theta}^{\texttt{HRD}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \|y - \theta\|_2^2 + 4\tau^2 \|\theta\|_0 \right\}$$

$$\widehat{\theta}^{\texttt{SFT}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \|y - \theta\|_2^2 + 4\tau \|\theta\|_1 \right\}$$

The BIC and Lasso Estimators

In view of (Equivalence to LSE), under the Assumption ORT, the variational definitions can be written as

$$\widehat{\theta}^{\texttt{HRD}} = \arg\min \left\{ \frac{1}{n} \|Y - X\theta\|_2^2 + 4\tau^2 \|\theta\|_0 \right\}$$

$$\widehat{\theta}^{\texttt{SFT}} = \arg\min \left\{ \frac{1}{n} \|Y - X\theta\|_2^2 + 4\tau \|\theta\|_1 \right\}$$

The BIC and Lasso Estimators

Definition 15

Fix $\tau > 0$ and assume (Linear Model). The BIC(Bayes Information Criterion) estimator of θ^* in is defined by an $\widehat{\theta}^{\mathtt{BIC}}$ such that

$$\widehat{\theta}^{\mathtt{BIC}} \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \|Y - \mathbb{X}\theta\|_2^2 + \tau^2 \|\theta\|_0 \right\}.$$

Moreover the Lasso estimator of θ^* in is defined by any $\widehat{\theta}^{\mathcal{L}}$ such that

$$\widehat{\theta}^{\mathcal{L}} \in \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{n} \|Y - \mathbb{X}\theta\|_2^2 + 2\tau \|\theta\|_1 \right\}.$$

The BIC and Lasso Estimators

Computing the BIC estimator can be proved to be NP-hard in the worst case. In particular, no computational method is known to be significantly faster than the brute force search among all 2^d sparsity patterns.

$$\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}=\min_{0\le k\le d}\left\{\min_{\theta:\|\theta\|_0=k}\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}$$

The BIC and Lasso Estimators

Computing the BIC estimator can be proved to be NP-hard in the worst case. In particular, no computational method is known to be significantly faster than the brute force search among all 2^d sparsity patterns.

$$\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}=\min_{0\le k\le d}\left\{\min_{\theta:\|\theta\|_0=k}\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}$$

To compute $\min_{\theta:\|\theta\|_0=k}\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2$, we need to compute $\binom{d}{k}$ least squares estimators on a space of size k. Each costs $O(k^3)$(matrix inverstion).

The BIC and Lasso Estimators

Computing the BIC estimator can be proved to be NP-hard in the worst case. In particular, no computational method is known to be significantly faster than the brute force search among all 2^d sparsity patterns.

$$\min_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}=\min_{0\le k\le d}\left\{\min_{\theta:\|\theta\|_0=k}\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2+\tau^2\|\theta\|_0\right\}$$

To compute $\min_{\theta:\|\theta\|_0=k}\frac{1}{n}\|Y-\mathbb{X}\theta\|_2^2$, we need to compute $\binom{d}{k}$ least squares estimators on a space of size k. Each costs $O(k^3)$(matrix inverstion).Therefore, the total cost of the brute force search is

$$C\sum_{k=0}^{d}\binom{d}{k}k^3=\Theta(d^3 2^d).$$

The BIC and Lasso Estimators

Solvers for the Lasso estimator:

- Coordinate gradient descent implemented in the **glmnet** in **R**.
- LARS that computes the entire regularization path, i.e., the solution of the convex problem for all value of τ. It relies on the fact that, as a function of τ, the solution $\widehat{\theta}^{\mathcal{L}}$ is a piecewise linear function (with values in $\mathbb{R}^d$.) LARS is slow for very large problem.
- FISTA.
- Stochastic gradient descent for very large d and very large n when computing $\|Y - \mathbb{X}\theta\|_2^2$ may be computationally expensive.

Analysis of the BIC estimator

Theorem 16

Assume that the (Linear Model) holds such that $\epsilon \sim \mathtt{subG}_n(\sigma^2)$. Then, then BIC estimator $\widehat{\theta}^{\mathtt{BIC}}$ with regularization parameter

$$\tau^2 = 16\log(6)\frac{\sigma^2}{n} + 32\frac{\sigma^2\log(ed)}{n} \qquad (\spadesuit)$$

satisfies

$$\mathtt{MSE}(\mathbb{X}\widehat{\theta}^{\mathtt{BIC}}) = \frac{1}{n}\|\mathbb{X}\widehat{\theta}^{\mathtt{BIC}} - \mathbb{X}\theta^*\|_2^2 \lesssim \|\theta^*\|_0\sigma^2\frac{\log(ed/\delta)}{n}$$

with probability at least $1-\delta$.

Analysis of the BIC estimator

Theorem 16

Assume that the (Linear Model) holds such that $\epsilon \sim \mathtt{subG}_n(\sigma^2)$. Then, then BIC estimator $\widehat{\theta}^{\mathtt{BIC}}$ with regularization parameter

$$\tau^2 = 16\log(6)\frac{\sigma^2}{n} + 32\frac{\sigma^2\log(ed)}{n} \qquad (\spadesuit)$$

satisfies

$$\mathtt{MSE}(\mathbb{X}\widehat{\theta}^{\mathtt{BIC}}) = \frac{1}{n}\|\mathbb{X}\widehat{\theta}^{\mathtt{BIC}} - \mathbb{X}\theta^*\|_2^2 \lesssim \|\theta^*\|_0\sigma^2\frac{\log(ed/\delta)}{n}$$

with probability at least $1-\delta$.

- **The theorem shows that $\widehat{\theta}^{\mathtt{BIC}}$ adapts to the unknown sparsity of θ^*, just like $\widehat{\theta}^{\mathtt{HRD}}$. And it holds under no assumption on the design matrix $\mathbb{X}$.**

Proof of Theorem 16.

We begin as usual by noting that

$$\frac{1}{n}\|Y - \mathbb{X}\widehat{\theta}^{\texttt{BIC}}\|_2^2 + \tau^2\|\widehat{\theta}^{\texttt{BIC}}\|_0 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + \tau^2\|\theta^*\|_0 \leq \frac{1}{n}\|\epsilon\|_2^2 + \tau^2\|\theta^*\|_0$$

Proof of Theorem 16.

We begin as usual by noting that

$$\frac{1}{n}\|Y - \mathbb{X}\hat{\theta}^{\texttt{BIC}}\|_2^2 + \tau^2\|\hat{\theta}^{\texttt{BIC}}\|_0 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + \tau^2\|\theta^*\|_0 \leq \frac{1}{n}\|\epsilon\|_2^2 + \tau^2\|\theta^*\|_0$$

It implies

$$\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2 \leq n\tau^2\|\theta^*\|_0 + 2\epsilon^T\mathbb{X}(\hat{\theta}^{\texttt{BIC}} - \theta^*) - n\tau^2\|\hat{\theta}^{\texttt{BIC}}\|_0$$

□

Proof of Theorem 16.

First, note that

$$2\epsilon^T(\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*) = 2\epsilon^T \left(\frac{\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*}{\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2} \right) \|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2$$

$$\leq 2 \left[\epsilon^T \left(\frac{\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*}{\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2} \right) \right]^2 + \frac{1}{2}\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2$$

where we used the inequality $2ab = 2(\sqrt{2a} \cdot \frac{1}{\sqrt{2}}b) \leq 2a^2 + \frac{1}{2}b^2$.

Proof of Theorem 16.

First, note that

$$2\epsilon^T(\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*) = 2\epsilon^T \left(\frac{\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*}{\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2} \right) \|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2$$

$$\leq 2 \left[\epsilon^T \left(\frac{\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*}{\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2} \right) \right]^2 + \frac{1}{2}\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2$$

where we used the inequality $2ab = 2(\sqrt{2a} \cdot \frac{1}{\sqrt{2}}b) \leq 2a^2 + \frac{1}{2}b^2$.

Putting together, it yields

$$\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2 \leq 2n\tau^2\|\theta^*\|_0 + 4\left[\epsilon^T \mathcal{U}(\hat{\theta}^{\texttt{BIC}} - \theta^*)\right]^2 - 2n\tau^2\|\hat{\theta}^{\texttt{BIC}}\|_0 \qquad (\clubsuit)$$

where $\mathcal{U}(\hat{\theta}^{\texttt{BIC}} - \theta^*) = \frac{\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*}{\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2}$. □

Proof of Theorem 16.

Next, we need to "sup out" $\widehat{\theta}^{\texttt{BIC}}$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta \in \mathbb{R}^d} = \max_{1 \le k \le d} \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S}.$$

Proof of Theorem 16.

Next, we need to "sup out" $\widehat{\theta}^{\texttt{BIC}}$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta \in \mathbb{R}^d} = \max_{1 \le k \le d} \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S}.$$

Applied to the above inequality, it yields

$$\begin{aligned} & 4\left[\epsilon^T \mathcal{U}(\widehat{\theta}^{\texttt{BIC}} - \theta^*)\right]^2 - 2n\tau^2 \|\widehat{\theta}^{\texttt{BIC}}\|_0 \\ & \le \max_{1 \le k \le d} \left\{ \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S} 4\left[\epsilon^T \mathcal{U}(\theta - \theta^*)\right]^2 - 2n\tau^2 k \right\} \end{aligned}$$

Proof of Theorem 16.

Next, we need to "sup out" $\widehat{\theta}^{\texttt{BIC}}$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta \in \mathbb{R}^d} = \max_{1 \le k \le d} \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S}.$$

Applied to the above inequality, it yields

$$\begin{aligned}
& 4\left[\epsilon^T \mathcal{U}(\widehat{\theta}^{\texttt{BIC}} - \theta^*)\right]^2 - 2n\tau^2 \|\widehat{\theta}^{\texttt{BIC}}\|_0 \\
& \le \max_{1 \le k \le d} \left\{ \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S} 4\left[\epsilon^T \mathcal{U}(\theta - \theta^*)\right]^2 - 2n\tau^2 k \right\} \\
& \le \max_{1 \le k \le d} \left\{ \max_{|S|=k} \sup_{u \in \mathcal{B}_2^{r_{S,*}}} 4\left[\epsilon^T \Phi_{S,*} u\right]^2 - 2n\tau^2 k \right\},
\end{aligned}$$

Proof of Theorem 16.

Next, we need to "sup out" $\widehat{\theta}^{\texttt{BIC}}$. To that end, we decompose the sup into a max over cardinalities as follows:

$$\sup_{\theta \in \mathbb{R}^d} = \max_{1 \le k \le d} \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S}.$$

Applied to the above inequality, it yields

$$\begin{aligned}
& 4\left[\epsilon^T \mathcal{U}(\widehat{\theta}^{\texttt{BIC}} - \theta^*)\right]^2 - 2n\tau^2 \|\widehat{\theta}^{\texttt{BIC}}\|_0 \\
& \le \max_{1 \le k \le d} \left\{ \max_{|S|=k} \sup_{\mathbf{supp}(\theta)=S} 4\left[\epsilon^T \mathcal{U}(\theta - \theta^*)\right]^2 - 2n\tau^2 k \right\} \\
& \le \max_{1 \le k \le d} \left\{ \max_{|S|=k} \sup_{u \in \mathcal{B}_2^{r_{S,*}}} 4\left[\epsilon^T \Phi_{S,*} u\right]^2 - 2n\tau^2 k \right\},
\end{aligned}$$

where $\Phi_{S,*} = [\phi_1, \ldots, \phi_{r_{S,*}}]$ is an orthonormal basis of the set $\{\mathbb{X}_j, j \in S \cup \mathbf{supp}(\theta^*)\}$ of columns of $\mathbb{X}$ and $r_{S,*} \le |S| + \|\theta^*\|_0$ is the dimension of this column span. □

Proof of Theorem 16.

Using the union bounds, we get for any $t > 0$,

$$\mathbb{P}\left(\max_{1\le k\le d}\left\{\max_{|S|=k}\sup_{u\in\mathcal{B}_2^{r_{S,*}}} 4\left[\epsilon^T\Phi_{S,*}u\right]^2 - 2n\tau^2k\right\} \ge t\right)$$
$$\le \sum_{k=1}^{d}\sum_{|S|=k}\mathbb{P}\left(\sup_{u\in\mathcal{B}_2^{r_{S,*}}}[\epsilon^T\Phi_{S,*}u]^2 \ge \frac{t}{4} + \frac{1}{2}n\tau^2k\right)$$

□

Proof of Theorem 16.

Moreover, from Theorem 4, we get for $|S| = k$,

$$\mathbb{P}\left(\sup_{u\in\mathcal{B}_2^{r_{S,*}}}\left[\epsilon^T\Phi_{S,*}u\right]^2 \geq \frac{t}{4}+\frac{1}{2}n\tau^2k\right) \leq 2\cdot 6^{r_{S,*}}\exp\left(-\frac{t/4+n\tau^2k/2}{8\sigma^2}\right)$$

Proof of Theorem 16.

Moreover, from Theorem 4, we get for $|S| = k$,

$$\mathbb{P}\left(\sup_{u\in\mathcal{B}_2^{r_{S,*}}}\left[\epsilon^T\Phi_{S,*}u\right]^2 \geq \frac{t}{4}+\frac{1}{2}n\tau^2k\right) \leq 2\cdot 6^{r_{S,*}}\exp\left(-\frac{t/4+n\tau^2k/2}{8\sigma^2}\right)$$

$$\leq 2\exp\left(-\frac{t}{32\sigma^2}-\frac{n\tau^2k}{16\sigma^2}+(k+\|\theta^*\|_0)\log(6)\right) \qquad (r_{S,*}\leq |S|+\|\theta^*\|_0)$$

$$\leq \exp\left(-\frac{t}{32\sigma^2}-2k\log(ed)+\|\theta^*\|_0\log(12)\right)$$

where, in the last inequality, we used the definition (♠) of τ. □

Proof of Theorem 16.

Putting together, we get

$$\mathbb{P}\left(\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2 \geq 2n\tau^2\|\theta^*\|_0 + t\right) \leq \sum_{k=1}^{d}\sum_{|S|=k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right)$$

$$= \sum_{k=1}^{d}\binom{d}{k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right)$$

Proof of Theorem 16.

Putting together, we get

$$\begin{aligned}
\mathbb{P}\left(\|\mathbb{X}\widehat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2 \geq 2n\tau^2\|\theta^*\|_0 + t\right) &\leq \sum_{k=1}^{d}\sum_{|S|=k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&= \sum_{k=1}^{d}\binom{d}{k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&\leq \sum_{k=1}^{d}\exp\left(-\frac{t}{32\sigma^2} - k\log(ed) + \|\theta^*\|_0\log(12)\right)
\end{aligned}$$

Proof of Theorem 16.

Putting together, we get

$$\begin{aligned}
\mathbb{P}\left(\|\mathbb{X}\hat{\theta}^{\texttt{BIC}} - \mathbb{X}\theta^*\|_2^2 \geq 2n\tau^2\|\theta^*\|_0 + t\right) &\leq \sum_{k=1}^{d}\sum_{|S|=k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&= \sum_{k=1}^{d}\binom{d}{k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&\leq \sum_{k=1}^{d}\exp\left(-\frac{t}{32\sigma^2} - k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&= \sum_{k=1}^{d}(ed)^{-k}\exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(12)\right) \\
&\leq \exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(12)\right)
\end{aligned}$$

Proof of Theorem 16.

Putting together, we get

$$\begin{aligned}
\mathbb{P}\left(\|\mathbb{X}\hat{\theta}^{\mathtt{BIC}} - \mathbb{X}\theta^*\|_2^2 \geq 2n\tau^2\|\theta^*\|_0 + t\right) &\leq \sum_{k=1}^{d}\sum_{|S|=k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&= \sum_{k=1}^{d}\binom{d}{k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&\leq \sum_{k=1}^{d}\exp\left(-\frac{t}{32\sigma^2} - k\log(ed) + \|\theta^*\|_0\log(12)\right) \\
&= \sum_{k=1}^{d}(ed)^{-k}\exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(12)\right) \\
&\leq \exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(12)\right)
\end{aligned}$$

where, in the third inequality, we apply $\binom{d}{k} \leq \left(\frac{ed}{k}\right)^k$. □

Proof of Theorem 16.

To conclude the proof, choose $t = 32\sigma^2\|\theta^*\|_0 \log(12) + 32\sigma^2 \log(1/\delta)$ and observe that combined with (♣), it yields with probability $1 - \delta$

$$\begin{aligned}\|\mathbb{X}\widehat{\theta}^{\mathtt{BIC}} - \mathbb{X}\theta^*\|_2^2 &\leq 2n\tau^2\|\theta\|_0^* + t \\ &= 64\sigma^2 \log(ed)\|\theta^*\|_0 + 64\log(12)\sigma^2\|\theta^*\|_0 + 32\sigma^2\log(1/\delta) \\ &\leq 224\|\theta^*\|_0\sigma^2\log(ed) + 32\sigma^2\log(1/\delta)\end{aligned}$$

where we have used the definition (♠) of τ. □

Slow rate for the Lasso estimator

Theorem 17

Assume that the (Linear Model) holds where $\epsilon \sim \texttt{subG}(\sigma^2)$. Moreover, assume that the columns of $\mathbb{X}$ are normalized in such a way that $\max_j \|\mathbb{X}_j\|_2 \leq \sqrt{n}$. Then, the Lasso estimator $\widehat{\theta}^{\mathcal{L}}$ with regularization parameter

$$2\tau = 2\sigma\sqrt{\frac{2\log(ed)}{n}} + 2\sigma\sqrt{\frac{2\log(1/\delta)}{n}} \qquad (\diamondsuit)$$

satisfies

$$\texttt{MSE}(\mathbb{X}\widehat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\widehat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq 4\|\theta^*\|_1\sigma\sqrt{\frac{2\log(2d)}{n}} + 4\|\theta^*\|_1\sigma\sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability $1-\delta$. Moreover, there exists a numerical constant $C > 0$ such that

$$\mathbb{E}[\texttt{MSE}(\mathbb{X}\widehat{\theta}^{\mathcal{L}})] \leq C\|\theta^*\|_1\sigma\sqrt{\frac{\log(2d)}{n}}.$$

Proof of Theorem 17.

From the definition of $\hat{\theta}^{\mathcal{L}}$, it holds

$$\frac{1}{n}\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 + 2\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1.$$

Proof of Theorem 17.

From the definition of $\hat{\theta}^{\mathcal{L}}$, it holds

$$\frac{1}{n}\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 + 2\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1.$$

It implies

$$\begin{aligned}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 &\leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + 2n\tau(\|\theta^*\|_1 - \|\hat{\theta}^{\mathcal{L}}\|) \\ &\leq 2\|\mathbb{X}^T\epsilon\|_\infty\|\hat{\theta}^{\mathcal{L}}\|_1 - 2n\tau\|\hat{\theta}^L\|_1 + 2\|\mathbb{X}^T\epsilon\|_\infty\|\theta^*\|_1 + 2n\tau\|\theta^*\|_1 \\ &= 2(\|\mathbb{X}^T\epsilon\|_\infty - n\tau)\|\hat{\theta}^{\mathcal{L}}\|_1 + 2(\|\mathbb{X}^T\epsilon\|_\infty + n\tau)\|\theta^*\|_1,\end{aligned}$$

where the Hölder's inequality is used. □

Proof of Theorem 17.

Since $\max_j \|\mathbb{X}_j\|_2 \leq \sqrt{n}$, we have $\mathbb{X}_j^T \epsilon \sim \texttt{subG}(n\sigma^2)$.

Proof of Theorem 17.

Since $\max_j \|\mathbb{X}_j\|_2 \leq \sqrt{n}$, we have $\mathbb{X}_j^T \epsilon \sim \texttt{subG}(n\sigma^2)$. Observe now that for any $t > 0$,

$$\mathbb{P}(\|\mathbb{X}^T \epsilon\|_\infty > t) = \mathbb{P}(\max_{1 \leq j \leq d} |\mathbb{X}_j^T \epsilon| > t)$$

Proof of Theorem 17.

Since $\max_j \|\mathbb{X}_j\|_2 \le \sqrt{n}$, we have $\mathbb{X}_j^T \epsilon \sim \texttt{subG}(n\sigma^2)$. Observe now that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(\|\mathbb{X}^T \epsilon\|_\infty > t) &= \mathbb{P}(\max_{1 \le j \le d} |\mathbb{X}_j^T \epsilon| > t) \\
&\le \sum_{j=1}^{d} \mathbb{P}(|\mathbb{X}_j^T \epsilon| > t) \quad \text{(Union Bound)} \\
&\le 2d \exp\left(-\frac{t^2}{2n\sigma^2}\right)
\end{aligned}$$

Proof of Theorem 17.

Since $\max_j \|\mathbb{X}_j\|_2 \leq \sqrt{n}$, we have $\mathbb{X}_j^T \epsilon \sim \texttt{subG}(n\sigma^2)$. Observe now that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(\|\mathbb{X}^T\epsilon\|_\infty > t) &= \mathbb{P}(\max_{1\leq j\leq d} |\mathbb{X}_j^T\epsilon| > t) \\
&\leq \sum_{j=1}^{d} \mathbb{P}(|\mathbb{X}_j^T\epsilon| > t) \quad \text{(Union Bound)} \\
&\leq 2d\exp\left(-\frac{t^2}{2n\sigma^2}\right)
\end{aligned}$$

Therefore, taking $t = \sigma\sqrt{2n\log(2d)} + \sigma\sqrt{2n\log(1/\delta)} = n\tau$, we get that with probability $1-\delta$,

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq 4n\tau\|\theta^*\|_1.$$

□

Proof of Theorem 17.

Let

$$Z = \texttt{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2.$$

Proof of Theorem 17.

Let

$$Z = \texttt{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2.$$

From the high-probability bound, for any $\xi > 0$,

$$\mathbb{P}\left(Z \geq \|\theta^*\|_1 \sigma \sqrt{\frac{2\log(2d)}{n}} + 4\|\theta^*\|_1 \sigma \sqrt{\frac{2\xi}{n}}\right) \leq e^{-\xi}.$$

Proof of Theorem 17.

Let

$$Z = \texttt{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2.$$

From the high-probability bound, for any $\xi > 0$,

$$\mathbb{P}\left(Z \geq \|\theta^*\|_1 \sigma \sqrt{\frac{2\log(2d)}{n}} + 4\|\theta^*\|_1 \sigma \sqrt{\frac{2\xi}{n}}\right) \leq e^{-\xi}.$$

Define

$$A = \|\theta^*\|_1 \sigma \sqrt{\frac{2\log(2d)}{n}}, \qquad B = 4\|\theta^*\|_1 \sigma \sqrt{\frac{2}{n}}.$$

Then

$$\mathbb{P}(Z \geq A + B\sqrt{\xi}) \leq e^{-\xi}.$$

□

Proof of Theorem 17.

Equivalently, for any $t \geq A$,

$$\mathbb{P}(Z \geq t) \leq \exp\left\{-\left(\frac{t-A}{B}\right)^2\right\}.$$

Proof of Theorem 17.

Equivalently, for any $t \geq A$,

$$\mathbb{P}(Z \geq t) \leq \exp\left\{-\left(\frac{t-A}{B}\right)^2\right\}.$$

Hence, we obtain

$$\begin{aligned}
\mathbb{E}[Z] &= \int_0^A \mathbb{P}(Z \geq t)\, dt + \int_A^\infty \mathbb{P}(Z \geq t)\, dt \\
&\leq A + \int_A^\infty \exp\left\{-\left(\frac{t-A}{B}\right)^2\right\} dt. \\
&= A + \frac{\sqrt{\pi}B}{2}
\end{aligned}$$

□

Proof of Theorem 17.

Substituting the definitions of A and B, we get

$$\mathbb{E}[\texttt{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}})] \leq \|\theta^*\|_1 \sigma \sqrt{\frac{2\log(2d)}{n}} + 2\sqrt{2\pi}\,\|\theta^*\|_1 \sigma \frac{1}{\sqrt{n}}.$$

Since $\log(2d) \geq \log 2$, there exists a numerical constant $C > 0$ such that

$$\mathbb{E}[\texttt{MSE}(\mathbb{X}\hat{\theta}^{\mathcal{L}})] \leq C\|\theta^*\|_1 \sigma \sqrt{\frac{\log(2d)}{n}}.$$

□

Remark of Theorem 17

- The regularization parameter ($\diamondsuit$) depends on the confidence level δ, which is not the case for the BIC estimator ($\spadesuit$).
- The rate in Theorem 17 is of order $\sqrt{(\log d)/n}$ (**slow rate**), which is much slow than the rate of order $(\log d)/n$ (**fast rate**) for the BIC estimator.
- Fast rates can be achieved by the computationally efficient Lasso estimator but at the cost of a much stronger condition on the design matrix.

Remark of Theorem 17

- The regularization parameter ($\diamondsuit$) depends on the confidence level δ, which is not the case for the BIC estimator ($\spadesuit$).
- The rate in Theorem 17 is of order $\sqrt{(\log d)/n}$ (**slow rate**), which is much slow than the rate of order $(\log d)/n$ (**fast rate**) for the BIC estimator.
- Fast rates can be achieved by the computationally efficient Lasso estimator but at the cost of a much stronger condition on the design matrix.

The BIC estimator uses an ℓ_0-type penalty, which directly penalizes the number of selected variables and therefore exploits the underlying sparsity structure more explicitly than the ℓ_1 penalty. Under suitable conditions, this sharper structural adaptation leads to the fast rate $(\log d)/n$, whereas the basic Lasso slow-rate analysis only controls the maximal noise-covariate correlation and yields the slower rate $\sqrt{(\log d)/n}$.

Incoherence

Assumption INC(k)

We say that the design matrix $\mathbb{X}$ has incoherence k for some integer $k > 0$ if

$$\left| \frac{\mathbb{X}^T \mathbb{X}}{n} - I_d \right|_\infty \leq \frac{1}{14k}$$

where the $|A|_\infty$ denotes the largest element of A in absolute value. Equivalently,

1. For all $j = 1, \ldots, d$,
$$\left| \frac{\|\mathbb{X}_j\|_2^2}{n} - 1 \right| \leq \frac{1}{14k}.$$
2. For all $1 \leq i, j \leq d$, $i \neq j$, we have
$$|\mathbb{X}_i^T \mathbb{X}_j| \leq \frac{1}{14k}.$$

Remark of Assumption INC(k)

- Assumption ORT arises as the limiting case of INC(k) as $k \to \infty$.
- While Assumption ORT requires $d \leq n$, INC(k) may have $d \gg n$.

Incoherence

Proposition 2

Let $\mathbb{X} \in \mathbb{R}^{n \times d}$ be a random matrix with entries $X_{i,j}, i = 1, \dots, n, j = 1, \dots, d$ that are i.i.d. Rademacher (± 1) random variables. Then $\mathbb{X}$ has incoherence k with probability $1 - \delta$ as soon as

$$n \geq 392k^2 \log(1/\delta) + 784k^2 \log(d).$$

It implies that there exits matrices that satisfy Assumption INC(k) for

$$n \gtrsim k^2 \log(d),$$

for some numerical constant C.

- **There exists a matrix that satisfies INC(k) even for $d > n$.**

Proof of Proposition 2.

Let $\epsilon_{i,j} \in \{-1, 1\}$ denote the Rademacher random variable that is on the ith row and jth column of $\mathbb{X}$.

Note first that the jth diagonal entries of $\mathbb{X}^T\mathbb{X}/n$ is given by

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,j}^2 = 1.$$

Proof of Proposition 2.

Let $\epsilon_{i,j} \in \{-1, 1\}$ denote the Rademacher random variable that is on the ith row and jth column of $\mathbb{X}$.

Note first that the jth diagonal entries of $\mathbb{X}^T\mathbb{X}/n$ is given by

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,j}^2 = 1.$$

Moreover, for $j \neq k$, the (j,k)th entry of the $d \times d$ matrix $\frac{\mathbb{X}^T\mathbb{X}}{n}$ is given by

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_{i,j}\epsilon_{i,k} = \frac{1}{n}\sum_{i=1}^{n}\xi_i^{(j,k)},$$

where for each pair, (j,k), $\xi_i^{(j,k)} = \epsilon_{i,j}\epsilon_{i,k}$ so that the random variables $\xi_1^{(j,k)}, \ldots, \xi_n^{(j,k)}$ are iid Rademacher variables. □

Proof of Proposition 2.

Therefore, we get that for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right|_\infty > 0\right) = \mathbb{P}\left(\max_{j\neq k}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i^{(j,k)}\right| > t\right)$$

Proof of Proposition 2.

Therefore, we get that for any $t > 0$,

$$\mathbb{P}\left(\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right|_\infty > 0\right) = \mathbb{P}\left(\max_{j\neq k}\left|\frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)}\right| > t\right)$$

$$\leq \sum_{j\neq k}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)}\right| > t\right) \qquad \text{(Union bound)}$$

Proof of Proposition 2.

Therefore, we get that for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right|_\infty > 0\right) &= \mathbb{P}\left(\max_{j\neq k}\left|\frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)}\right| > t\right) \\
&\leq \sum_{j\neq k}\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \xi_i^{(j,k)}\right| > t\right) && \text{(Union bound)} \\
&\leq \sum_{j\neq k} 2e^{-\frac{nt^2}{2}} && \text{(Hoeffding's inequality)} \\
&\leq d^2 e^{-\frac{nt^2}{2}},
\end{aligned}$$

where we use the fact that $\xi_i^{(j,k)} \sim \mathtt{subG}(1^2)$. □

Proof of Proposition 2.

Taking now $t = 1/(14k)$ yields

$$\mathbb{P}\left(\left|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right| > \frac{1}{14k}\right) \leq d^2 e^{-\frac{n}{392k^2}} \leq \delta$$

for

$$n \geq 392k^2 \log(1/\delta) + 784k^2 \log(d).$$

□

Incoherence

For any $\theta \in \mathbb{R}^d$ and $S \subset \{1, \dots, d\}$, define θ_S to be the vector with coordinates

$$\theta_{S,j} = \begin{cases} \theta_j & \text{if } j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

In particular $\|\theta\|_1 = \|\theta_S\|_1 + \|\theta_{S^c}\|_1$.

Incoherence

Lemma 18

Fix a positive integer $k \leq d$ and assume that $\mathbb{X}$ satisfies assumption INC(k). Then for any $S \in \{1, \dots, d\}$ such that $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ that satisfies the cone condition

$$\|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1, \qquad (\heartsuit)$$

it holds

$$\|\theta_S\|_2^2 \leq 2\frac{\|\mathbb{X}\theta\|_2^2}{n}$$

Proof of Lemma 18.

We have

$$\frac{\|\mathbb{X}\theta\|_2^2}{n} = \frac{1}{n}\|\mathbb{X}\theta_S + \mathbb{X}\theta_{S^c}\|_2^2 = \frac{\|\mathbb{X}\theta_S\|_2^2}{n} + 2\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c} + \frac{\|\mathbb{X}\theta_{S^c}\|_2^2}{n}.$$
$$\geq \frac{\|\mathbb{X}\theta_S\|_2^2}{n} + 2\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}.$$

Proof of Lemma 18.

We have

$$\frac{\|\mathbb{X}\theta\|_2^2}{n} = \frac{1}{n}\|\mathbb{X}\theta_S + \mathbb{X}\theta_{S^c}\|_2^2 = \frac{\|\mathbb{X}\theta_S\|_2^2}{n} + 2\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c} + \frac{\|\mathbb{X}\theta_{S^c}\|_2^2}{n}.$$
$$\geq \frac{\|\mathbb{X}\theta_S\|_2^2}{n} + 2\theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}.$$

Moreover, from the incoherence condition, we have

$$\frac{\|\mathbb{X}\theta_S\|_2^2}{n} = \theta_S^T\frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c} = \|\theta_S\|_2^2 + \theta_S^T\left(\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right)\theta_S$$
$$\geq \|\theta_S\|_2^2 - \frac{\|\theta_S\|_1^2}{14k} \qquad \textbf{(Check By Yourself)}$$

□

Proof of Lemma 18.

By the definition of θ_S and θ_{S^c}, we have $\theta_{S,i}\theta_{S^c,i} = 0$. And then

Proof of Lemma 18.

By the definition of θ_S and θ_{S^c}, we have $\theta_{S,i}\theta_{S^c,i} = 0$. And then

$$\left|\theta_S^T \frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}\right| = \left|\sum_{i=1}^{d}\sum_{j=1}^{d}\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right|$$

Proof of Lemma 18.

By the definition of θ_S and θ_{S^c}, we have $\theta_{S,i}\theta_{S^c,i} = 0$. And then

$$\begin{aligned}
\left|\theta_S^T \frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}\right| &= \left|\sum_{i=1}^{d}\sum_{j=1}^{d}\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right| \\
&= \sum_{i\neq j}\left|\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right| \\
&\leq \frac{1}{14k}\sum_{i\neq j}|\theta_{S,i}|\cdot|\theta_{S^c,j}| \qquad \textbf{(INC}(k)\textbf{)}
\end{aligned}$$

Proof of Lemma 18.

By the definition of θ_S and θ_{S^c}, we have $\theta_{S,i}\theta_{S^c,i} = 0$. And then

$$\begin{aligned}
\left|\theta_S^T \frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}\right| &= \left|\sum_{i=1}^{d}\sum_{j=1}^{d}\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right| \\
&= \sum_{i\neq j}\left|\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right| \\
&\leq \frac{1}{14k}\sum_{i\neq j}|\theta_{S,i}|\cdot|\theta_{S^c,j}| \qquad \textbf{(INC}(k)\textbf{)} \\
&\leq \frac{1}{14k}\|\theta_S\|_1\|\theta_{S^c}\|_1
\end{aligned}$$

Proof of Lemma 18.

By the definition of θ_S and θ_{S^c}, we have $\theta_{S,i}\theta_{S^c,i} = 0$. And then

$$\begin{aligned}
\left|\theta_S^T \frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c}\right| &= \left|\sum_{i=1}^{d}\sum_{j=1}^{d}\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right| \\
&= \sum_{i\neq j}\left|\theta_{S,i}\left(\frac{\mathbb{X}_i^T\mathbb{X}_j}{n}\right)\theta_{S^c,j}\right| \\
&\leq \frac{1}{14k}\sum_{i\neq j}|\theta_{S,i}|\cdot|\theta_{S^c,j}| \qquad \textbf{(INC}(k)\textbf{)} \\
&\leq \frac{1}{14k}\|\theta_S\|_1\|\theta_{S^c}\|_1 \\
&\leq \frac{3}{14k}\|\theta_S\|_1^2 \qquad \textbf{(Lemma's condition)}
\end{aligned}$$

□

Proof of Lemma 18.

Putting together,

$$\begin{aligned}\frac{\|\mathbb{X}\theta\|_2^2}{n} &\geq \frac{\|\mathbb{X}\theta_S\|_2^2}{n} + 2\theta_S^T \frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c} \\ &\geq \|\theta_S\|_2^2 - \frac{\|\theta_S\|_1^2}{14k} - \frac{6}{14k}\|\theta_S\|_1^2 \\ &= \|\theta_S\|_2^2 - \frac{1}{2k}\|\theta_S\|_1^2.\end{aligned}$$

Proof of Lemma 18.

Putting together,

$$\begin{aligned}\frac{\|\mathbb{X}\theta\|_2^2}{n} &\geq \frac{\|\mathbb{X}\theta_S\|_2^2}{n} + 2\theta_S^T \frac{\mathbb{X}^T\mathbb{X}}{n}\theta_{S^c} \\ &\geq \|\theta_S\|_2^2 - \frac{\|\theta_S\|_1^2}{14k} - \frac{6}{14k}\|\theta_S\|_1^2 \\ &= \|\theta_S\|_2^2 - \frac{1}{2k}\|\theta_S\|_1^2.\end{aligned}$$

Since $\|\theta_S\|_1^2 \leq |S|\|\theta_S\|_2^2$ and $|S| \leq k$, we complete the proof. □

Theorem 19

Fix $n \geq 2$. Assume that the (Linear Model) holds where $\epsilon \sim \mathtt{subG}_n(\sigma^2)$. Moreover, assume that $\|\theta^*\|_0 \leq k$ and that $\mathbb{X}$ satisfies assumption INC(k). Then the Lasso estimator $\widehat{\theta}^{\mathcal{L}}$ with regularization parameter defined by

$$2\tau = 8\sigma\sqrt{\frac{\log(2d)}{n}} + 8\sigma\sqrt{\frac{\log(1/\delta)}{n}}$$

satisfies

$$\mathtt{MSE}(\mathbb{X}\widehat{\theta}^{\mathcal{L}}) = \frac{1}{n}\|\mathbb{X}\widehat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \lesssim k\sigma^2\frac{\log(2d/\delta)}{n}$$

and

$$\|\widehat{\theta}^{\mathcal{L}} - \theta^*\|_1 \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}[\mathtt{MSE}(\mathbb{X}\widehat{\theta}^{\mathcal{L}})] \lesssim k\sigma^2\frac{\log(2d)}{n}, \quad \text{and} \quad \mathbb{E}\left[\|\widehat{\theta}^{\mathcal{L}} - \theta^*\|_1\right] \lesssim k\sigma\sqrt{\frac{\log(2d/\delta)}{n}}.$$

Proof of Theorem 19.

From the definition of $\hat{\theta}^{\mathcal{L}}$,

$$\frac{1}{n}\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 + 2\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1$$

Proof of Theorem 19.

From the definition of $\hat{\theta}^{\mathcal{L}}$,

$$\frac{1}{n}\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 + 2\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1$$

Multiplying both sides by n and similar to the proof of Theorem 17, we have

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + 2\tau n\|\theta^*\|_1 - 2\tau n\|\hat{\theta}^{\mathcal{L}}\|_1$$

Proof of Theorem 19.

From the definition of $\hat{\theta}^{\mathcal{L}}$,

$$\frac{1}{n}\|Y - \mathbb{X}\hat{\theta}^{\mathcal{L}}\|_2^2 + 2\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \leq \frac{1}{n}\|Y - \mathbb{X}\theta^*\|_2^2 + 2\tau\|\theta^*\|_1$$

Multiplying both sides by n and similar to the proof of Theorem 17, we have

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + 2\tau n\|\theta^*\|_1 - 2\tau n\|\hat{\theta}^{\mathcal{L}}\|_1$$

And adding both sides $n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1$ yields

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2\tau n\|\theta^*\|_1 - 2\tau n\|\hat{\theta}^{\mathcal{L}}\|_1$$

□

Proof of Theorem 19.

Applying Hölder's inequality, we get

$$\epsilon^T \mathbb{X}(\widehat{\theta}^{\mathcal{L}} - \theta^*) \leq \|\epsilon^T \mathbb{X}\|_\infty \|\widehat{\theta}^{\mathcal{L}} - \theta^*\|_1.$$

Proof of Theorem 19.

Applying Hölder's inequality, we get

$$\epsilon^T \mathbb{X}(\widehat{\theta}^{\mathcal{L}} - \theta^*) \leq \|\epsilon^T \mathbb{X}\|_\infty \|\widehat{\theta}^{\mathcal{L}} - \theta^*\|_1.$$

By the assumption INC(k), we have $\|\mathbb{X}_j\|_2^2 \leq (\frac{1}{14k} + 1)n \leq 2n$. And hence, $\epsilon^T \mathbb{X}_j \sim \texttt{subG}(2n\sigma^2)$.

Proof of Theorem 19.

Applying Hölder's inequality, we get

$$\epsilon^T \mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) \leq \|\epsilon^T \mathbb{X}\|_\infty \|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1.$$

By the assumption INC(k), we have $\|\mathbb{X}_j\|_2^2 \leq (\frac{1}{14k} + 1)n \leq 2n$. And hence, $\epsilon^T \mathbb{X}_j \sim \texttt{subG}(2n\sigma^2)$.

Following the same steps in the proof of Theorem 19, we get

$$\epsilon^T \mathbb{X}(\hat{\theta}^{\mathcal{L}} - \theta^*) \leq \frac{n\tau}{2} \|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1.$$

with probability at least $1 - \delta$. □

Proof of Theorem 19.

Therefore, taking $S = \text{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \\ &= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c}\|_1\end{aligned}$$

Proof of Theorem 19.

Therefore, taking $S = \text{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S\|_1
\end{aligned}$$

Proof of Theorem 19.

Therefore, taking $S = \text{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S\|_1 \\
&\leq 4n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 \qquad \text{(Triangle Inequality)}
\end{aligned} \tag{$\dagger$}$$

Proof of Theorem 19.

Therefore, taking $S = \text{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S\|_1 \\
&\leq 4n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 \qquad \text{(Triangle Inequality)}
\end{aligned} \tag{$\dagger$}$$

Hence

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 4\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 = 4\|\hat{\theta}^{\mathcal{L}}_S - \theta^*_S\|_1.$$

Proof of Theorem 19.

Therefore, taking $S = \text{supp}(\theta^*)$ to be the support of θ^*, we get

$$\begin{aligned}
\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 2n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S + \hat{\theta}^{\mathcal{L}}_{S^c}\|_1 \\
&= 2n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 + 2n\tau\|\theta^*\|_1 - 2n\tau\|\hat{\theta}^{\mathcal{L}}_S\|_1 \\
&\leq 4n\tau\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 \qquad \text{(Triangle Inequality)}
\end{aligned} \tag{$\dagger$}$$

Hence

$$\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 4\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 = 4\|\hat{\theta}^{\mathcal{L}}_S - \theta^*_S\|_1.$$

Subtracting both sides by $\|\hat{\theta}^{\mathcal{L}}_S - \theta^*_S\|_1$ yields

$$\|\hat{\theta}^{\mathcal{L}}_{S^c} - \theta^*_{S^c}\|_1 \leq 3\|\hat{\theta}^{\mathcal{L}}_S - \theta^*_S\|_1.$$

□

Proof of Theorem 19.

So that $\theta = \widehat{\theta}^{\mathcal{L}} - \theta^*$ satisfies the cone condition ($\heartsuit$). Using the Cauchy-Schwartz inequality and Lemma 18 respectively, we get since $|S| \leq k$,

$$\|\widehat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 \leq \sqrt{S}\|\widehat{\theta}^{\mathcal{L}}_S - \theta^*\|_2 \leq \sqrt{\frac{2k}{n}}\|\mathbb{X}\widehat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2$$

Proof of Theorem 19.

So that $\theta = \hat{\theta}^{\mathcal{L}} - \theta^*$ satisfies the cone condition ($\heartsuit$). Using the Cauchy-Schwartz inequality and Lemma 18 respectively, we get since $|S| \leq k$,

$$\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_1 \leq \sqrt{S}\|\hat{\theta}^{\mathcal{L}}_S - \theta^*\|_2 \leq \sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2$$

Combing this result with ($\dagger$), we find

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq \|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 + n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 4n\tau\sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2,$$

and therefore

$$\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2^2 \leq 32nk\tau^2.$$

□

Proof of Theorem 19.

Moreover, it yields

$$n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 4n\tau\sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2$$

and

$$\begin{aligned}\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 4\sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2 \\ &\leq 4\sqrt{\frac{2k}{n}}\sqrt{32nk\tau^2} = 32k\tau.\end{aligned}$$

Proof of Theorem 19.

Moreover, it yields

$$n\tau\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 \leq 4n\tau\sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2$$

and

$$\begin{aligned}\|\hat{\theta}^{\mathcal{L}} - \theta^*\|_1 &\leq 4\sqrt{\frac{2k}{n}}\|\mathbb{X}\hat{\theta}^{\mathcal{L}} - \mathbb{X}\theta^*\|_2 \\ &\leq 4\sqrt{\frac{2k}{n}}\sqrt{32nk\tau^2} = 32k\tau.\end{aligned}$$

The bound in expectation follows using the same argument as in the proof of Corollary 12. □

Remark on the Proof of Theorem 19

Note that all required for the proof was not really incoherence but the conclusion of Lemma 18

$$\inf_{|S|\le k}\inf_{\theta\in\mathcal{C}_S}\frac{\|\mathbb{X}\theta\|_2^2}{n\|\theta_S\|_2^2}\ge\kappa, \qquad \text{(Restrict Eigenvalue Condition)}$$

where $\kappa=\frac{1}{2}$ and $\mathcal{C}_S$ is the cone defined by

$$\mathcal{C}_S=\{\|\theta_{S^c}\|_1\le 3\|\theta_S\|_1\}.$$

Remark on the Proof of Theorem 19

Note that all required for the proof was not really incoherence but the conclusion of Lemma 18

$$\inf_{|S|\le k}\inf_{\theta\in\mathcal{C}_S}\frac{\|\mathbb{X}\theta\|_2^2}{n\|\theta_S\|_2^2}\ge\kappa, \qquad \text{(Restrict Eigenvalue Condition)}$$

where $\kappa=\frac{1}{2}$ and $\mathcal{C}_S$ is the cone defined by

$$\mathcal{C}_S=\{\|\theta_{S^c}\|_1\le 3\|\theta_S\|_1\}.$$

Note that all k-sparse vectors θ are in a cone $\mathcal{C}_S$ with $|S|\le k$ so that the (Restrict Eigenvalue Condition) implies that the smallest eigenvalue of $\mathbb{X}_S$ satisfies $\lambda_{\min}(\mathbb{X}_S)\ge n\kappa$ for all S such that $|S|\le k$.

Remark on the Proof of Theorem 19

Note that all required for the proof was not really incoherence but the conclusion of Lemma 18

$$\inf_{|S|\leq k}\inf_{\theta\in\mathcal{C}_S}\frac{\|\mathbb{X}\theta\|_2^2}{n\|\theta_S\|_2^2}\geq\kappa, \qquad \text{(Restrict Eigenvalue Condition)}$$

where $\kappa=\frac{1}{2}$ and $\mathcal{C}_S$ is the cone defined by

$$\mathcal{C}_S=\{\|\theta_{S^c}\|_1\leq 3\|\theta_S\|_1\}.$$

Note that all k-sparse vectors θ are in a cone $\mathcal{C}_S$ with $|S|\leq k$ so that the (Restrict Eigenvalue Condition) implies that the smallest eigenvalue of $\mathbb{X}_S$ satisfies $\lambda_{\min}(\mathbb{X}_S)\geq n\kappa$ for all S such that $|S|\leq k$.

The (Restrict Eigenvalue Condition) is weaker than the coherence and it can be shown that a design matrix $\mathbb{X}$ of i.i.d Rademacher random variables satisfies the (Restrict Eigenvalue Condition) as soon as $n\geq Ck\log(d)$ with positive probability.