Sparse linear models in high dimensions

Chen Liu

March 30, 2022

Problem formulation

Least-squares prediction for situations that number of features d is substantially less than the number of samples n has a long history. However, when $d > n$ we have to impose additional structure on the parameter θ. Give the data $X \in R^{n \times d}$, here each row represents the the data sample $x_i \in R^d$. The target $y \in R^n$ and parameter $\theta \in R^d$. Sometimes we also assume the existence of noise $w \in R^n$

$$y = X\theta^* + w$$

Example

Formulation of LASSO

$$min_\theta \frac{1}{2}|y - X\theta\|_2^2 + \lambda\|\theta\|_1$$

It can be solved with proximal gradient descent. Here we want to know under what conditions we can get a accurate estimation of the oracle parameter

Recovery in the noiseless setting

Firstly, we consider the setting without noise, so we can suppose that the data is generated by the following equation, here θ^* is the oracle parameter.

$$y = X\theta^*$$

Here $X \in R^{n\times d}$ and $d > n$. We also assume that the oracle parameter $\theta*$ is sparse, denote the non-zero index set as S and the complement S^c.

Recovery in the noiseless setting

Here we define the problem,

$$min\|\theta\|_1 \qquad s.t. \qquad X\theta = y$$

And we consider the possible solution.

- $null(X) = \{\Delta \in R^d | X\Delta = 0\}$
- tangent cone $T(\theta^*) = \{\Delta \in R^d | \|\theta^* + t\Delta\|_1 \leq \|\theta^*\|_1\}$
- $C(S) = \{\Delta \in R^d | \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$

Recovery in the noiseless setting

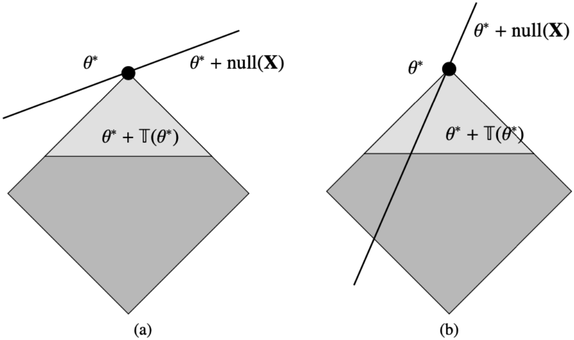

Figure 7.2 Geometry of the tangent cone and restricted nullspace property in $d = 2$ dimensions. (a) The favorable case in which the set $\theta^* + \text{null}(\mathbf{X})$ intersects the tangent cone only at θ^*. (b) The unfavorable setting in which the set $\theta^* + \text{null}(\mathbf{X})$ passes directly through the tangent cone.

Recovery in the noiseless setting

Definition (restricted nullspace property)

The matrix X satisfies the restricted nullspace property with respect to S if $C(S) \cap null(X) = \{0\}$.

Theorem

The following two properties are equivalent
(a) For any vector $\theta^* \in R^d$ *with support* S*, the problem applied with* $y = X\theta$ *has unique solution* $\hat{\theta} = \theta^*$*.*
(b) The matrix X satisfies the restricted nullspace property with respect to S*.*

Recovery in the noiseless setting

Proof.

for (b) to (a) $\hat{\theta}$ is optimal, we have $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$. Define $\hat{\Delta} = \hat{\theta} - \theta^*$,

$$\begin{aligned} \|\theta^*_S\|_1 = \|\theta^*\|_1 &\geq \|\theta^* + \hat{\Delta}\|_1 \\ &= \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \\ &\geq |\theta^*_S\|_1 - \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \end{aligned}$$

Here $\hat{\Delta} \in C(S)$ and $X\hat{\Delta} = 0$. So $\hat{\Delta} = 0$ □

Recovery in the noiseless setting

Proof.

for (a) to (b) take $\theta \in null(X) \setminus \{0\}$. Take $y = X\theta_S$
for the problem

$$min\|\beta\|_1 s.t. \quad X\beta = y$$

We have solutions θ_S and $-\theta_{S^c}$. Due to the uniqueness,
$\|\theta_S\|_1 < \|\theta_{S^c}\|_1$ □

Recovery in the noiseless setting

Incoherence parameter of the design matrix

$$\delta_{pw}(X) = max_{j,k=1,\cdots,d} \| \frac{< X_j, X_k >}{n} \mathbf{1}[j = k] \|$$

Proposition

If the pairwise incoherence satisfies the bound $\delta_{pw}(X) < \frac{1}{3s}$, then the restricted nullspace property holds for all subsets S of cardinality at most S.

Recovery in the noisy setting

Here for the noisy setting

$$y = X\theta^* + w$$

$$min \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$$

$$min \frac{1}{2n} \|y - X\theta\|_2^2 \qquad \|\theta\|_1 \leq \|\theta^*\|$$

$$min \|\theta\|_1 \qquad s.t. \qquad \frac{1}{2n} \|y - X\theta\|_2^2 \leq b^2$$

b^2 measures the tolerance of noise.
And we define the set $C_\alpha(S) = \{\Delta \in R^d | \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}$

Recovery in the noisy setting

Definition

The matrix X satisfies the restricted eigen (RE) condition over S with parameters (k, α) if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2 \forall \Delta \in C_\alpha(S)$$

Assumptions

- The vector θ^* is supported on a subset S
- The design matrix satisfies RE with $(k, 3)$

Recovery in the noisy setting

Theorem

Under the two assumptions.

(a) Any solution with $\lambda \leq 2\|\frac{X^T w}{n}\|_\infty$, $\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{k}\sqrt{s}\lambda$

(b) Any solution satisfies the bound $\|\hat{\theta} - \theta^*\|_2 \leq \frac{4}{k}\sqrt{s}\|\frac{X^T w}{n}\|_\infty$

(c) With $b^2 \geq 2\|\frac{X^T w}{n}\|_\infty$, $\|\hat{\theta} - \theta^*\|_2 \leq \frac{4}{k}\sqrt{s}\|\frac{X^T w}{n}\|_\infty + \frac{2}{\sqrt{k}}\sqrt{b^2 - \frac{\|w\|_2^2}{2n}}$

Recovery in the noisy setting

Proof.

(b)For optimal solution $\hat{\theta}$, we have $\frac{1}{2}\|y - X\hat{\theta}\|_2^2 \leq \frac{1}{2}\|y - X\theta^*\|_2^2$. Let $\hat{\Delta} = \hat{\theta} - \theta^*$. We can get the following inequality

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq \frac{2w^T X\hat{\Delta}}{n}$$

Using Holder $\frac{2w^T X\hat{\Delta}}{n} \leq \|\frac{w^T X}{n}\|_\infty \|\hat{\Delta}\|_1$. In addition,

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \leq 2\|\hat{\Delta}_S\|_1 \leq 2\sqrt{s}\|\hat{\Delta}\|_2$$

Using the RE condition, $\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2$. Then we can get the final result. □

Recovery in the noisy setting

Proof.

(c) We have $\frac{1}{2n}\|y - X\theta^*\|_2^2 = \frac{1}{2n}\|w\|_2^2 \leq b^2$.
let $\hat{\Delta} = \hat{\theta} - \theta^*$,we have $\frac{1}{2n}\|y - X\hat{\theta}\|_2^2 \leq \frac{1}{2n}\|y - X\theta^*\|_2^2 + b^2 - \frac{1}{2n}\|w\|_2^2$,
rearranging the term, we can get

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq 2\frac{w^T X\Delta}{n} + 2(b^2 - \frac{1}{2n}\|w\|_2^2)$$

$$k\|\hat{\Delta}\|_2^2 \leq 4\sqrt{s}\|\hat{\Delta}\|_2\|\frac{w^T X}{n}\|_\infty + 2(b^2 - \frac{1}{2n}\|w\|_2^2)$$

□

Recovery in the noisy setting

Proof.

(a) we have $L(\theta) = \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1$. For optimal solution $\hat{\theta}$, we have $L(\hat{\theta}) \leq L(\theta^*)$ and $L(\theta^*) = \frac{1}{2n}\|w\|_2^2 + \lambda\|\theta^*\|_1$. let $\hat{\Delta} = \hat{\theta} - \theta^*$.So we can get

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{w^T X\hat{\Delta}}{n} + \lambda(\lambda\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$

And we have $\|\theta^*\|_1 - \|\hat{\theta}\|_1 = \|\theta_s^*\|_1 - \|\theta_s^* + \Delta_s\|_1 - \|\Delta_{S^C}\|_1$

$$\begin{aligned}
0 \leq \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq & \frac{2w^T X\hat{\Delta}}{n} + 2\lambda(\|\theta_s^*\|_1 - \|\theta_s^* + \Delta_s\|_1 - \|\Delta_{S^C}\|_1) \\
\leq & 2\|\frac{w^T X}{n}\|_\infty \|\hat{\Delta}\|_1 + 2\lambda(\|\Delta_s\|_1 - \|\Delta_{S^C}\|_1) \\
\leq & \lambda(3\|\Delta_S\|_1 - \|\Delta_{S^C}\|_1)) \leq 3\sqrt{s}\lambda\|\hat{\Delta}\|_2^2
\end{aligned}$$

Finally using $\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2$ □

Recovery in the random design setting

Here $\rho(\Sigma)$ is maximum diagonal entry of Σ

Theorem

Consider a random matrix X^n, each row $x_i \in R^d$ is drawn i.i.d from $N(0, \Sigma)$, then there are universal positive constants $c_1 < 1 < c_2$ that for all $\theta \in R^d$

$$\frac{\|X\theta\|_2^2}{n} \geq c1\|\sqrt{\Sigma}\theta\|_2^2 - c2\rho^2(\Sigma)\frac{logd}{n}\|\theta\|_1^2$$

with probability at least $1 - \frac{e^{\frac{-n}{32}}}{1-e^{\frac{-n}{32}}}$

Proof.

By rescaling, we can focus on the ellipse
$S^{d-1}(\Sigma) = \{\in R^d | \|\sqrt{\Sigma\theta}\| = 1\}$. Define the function as
$g(t) = 2\rho\sqrt{\frac{logd}{n}}t$
here the bad event is

$$\epsilon = \{X^{n\times d} | inf_{\theta\in S^{d-1}} \frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2g(\|\theta\|_1)\}$$

So we have to bound $P[\epsilon]$
For a pair set $K(r_l, r_u) = \{\theta \in S^{d-1}(\Sigma) | g(\|\theta\|_1) \in [r_l, r_u]\}$, And the event A is defined as

$$A(r_l, r_u) = \{inf_{\theta\in K} \frac{\|X\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 2r_u\}$$

□

Recovery in the random design setting

Lemma

For any pair of radii $0 \leq r_l < r_u$ *wehave*

$$P[A(r_l, r_u)] \leq e^{\frac{-n}{32}} e^{-\frac{n}{2} r_u^2}$$

Moreover, for $\mu = \frac{1}{4}$*, we have*

$$\epsilon \subset A(0, \mu) \cup (\cup_{l=0}^{\infty} A(2^{l-1}\mu, 2^l \mu))$$

With this lemma , we can derive

$$P[\epsilon] \leq P[A(0, \mu)] + \sum_{l=0}^{\infty} P[A(2^{l-1}\mu, 2^l \mu)] \leq e^{\frac{-n}{32}} \{\sum_{l=0}^{\infty} e^{\frac{n}{2} 2^{2l} \mu^2}\}$$

with $\mu = \frac{1}{4} and 2^{2l} \geq 2l$ we have

$$P[\epsilon] \leq e^{\frac{-n}{32}} \{\sum_{l=0}^{\infty} e^{\frac{n}{2} 2^{2l} \mu^2}\} \leq \frac{e^{\frac{-n}{32}}}{1 - e^{\frac{-n}{32}}}$$

Recovery in the random design setting

Theorem ((Lasso oracle inequality)

Under the previous condition, with $\lambda \geq 2\|\frac{2X^T w}{n}\|_\infty$*. For any* $\theta^* \in R^d$ *any optimal solution* $\hat{\theta}$ *satifies the following bound,*

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{144}{c_1^2}\frac{\lambda^2}{k^2}|S| + \frac{16}{c_1}\frac{\lambda}{k}\|\theta^*_{S^C}\|_1 + \frac{32c_2}{c_1}\frac{\rho^2(\Sigma)}{k}\frac{logd}{n} + \|\theta^*_{S^C}\|_1^2$$

with $|S| \leq \frac{c_1}{64c_2}\frac{k}{\rho^2(\Sigma)}\frac{n}{logd}$

Proof.

For any $\theta^* \in R^d$, we have

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{\lambda}{2}\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 + 2\|\theta^*_{S^c}\|_1\}$$

it implies, $\|\hat{\Delta}\|_1^2 \leq (4\|\hat{\Delta}_S\|_1 + 2\|\theta^*_{S^c}\|_1)^2 \leq 32|S|\|\hat{\Delta}\|_2^2 + 8\|\theta^*_{S^c}\|_1^2$
Combined with the previous theorem, we have,

$$\begin{aligned}\frac{\|X\hat{\Delta}\|_2^2}{n} \geq & \{c_1 k - 32c_2\rho^2\|S\|\frac{logd}{n}\} * \|\hat{\Delta}\|_2^2 - 8c_2\rho^2\frac{logd}{n}\|\theta^*_{S^c}\|_1^2 \\ & \geq c_1\frac{k}{2}c_1\|\hat{\Delta}\|_2^2 - 8c_2\rho^2\frac{logd}{n}\|\theta^*_{S^c}\|_1^2\end{aligned}$$

Proof.

if $c_1 \frac{k}{4}\|\hat{\Delta}\|_2^2 \geq 8c_2\rho^2\frac{logd}{n}$, we have

$$c_1\frac{k}{4}\|\hat{\Delta}\|_2^2 \leq \frac{\lambda}{2}3\sqrt{|S|}\|\hat{\Delta}\|_2 + 2\lambda\|\theta^*_{S^C}\|_1$$

from this inequality, we can obtain that,

$$\|\hat{\Delta}\|_2^2 \leq \frac{144}{c_1^2}\frac{\lambda^2}{k^2}|S| + \frac{16}{c_1}\frac{\lambda}{k}\|\theta^*_{S^C}\|_1$$

else we have$c_1\frac{k}{4}\|\hat{\Delta}\|_2^2 < 8c_2\rho^2\frac{logd}{n}$, it implies that,

$$\|\hat{\Delta}\|_2^2 < \frac{32c_2\rho^2 logd}{c_1 kn}$$

□

Bounds on prediction error

For the recovery with noise $y = X\theta + w$, suppose that $w \sim N(0, \sigma^2)$.
We have $\frac{1}{n}E[\|y - X\hat{\theta}\|_2^2] = \frac{1}{n}E[\|X(\hat{\theta} - \theta^*)\|_2^2] + \sigma^2$.
So we use the term $\frac{1}{n}\|X(\hat{\theta} - \theta^*)\|_2^2$ to estimate trhe prediction error.

Theorem (Prediction error bounds)

Consider $\lambda \leq 2\|\frac{X^T w}{n}\|_\infty$ (a) Any optimal solution $\hat{\theta}$ satisfies the following bound,

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq 12\|\theta^*\|_1\lambda$$

(b) ifθ^* is supported on S and design matrix X satisfies $(k, 3)$-RE condition over S, then we have,

$$\frac{\|X(\hat{\theta} - \theta^*)\|_2^2}{n} \leq \frac{9}{k}s\lambda^2$$

Bounds on prediction error

Proof.

(a) Denote $\hat{\Delta} = \hat{\theta} - \theta$.

$$0 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{w^T X\hat{\Delta}}{n} + \lambda(\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$

$$\|\frac{w^T X\hat{\Delta}}{n}\|_1 \leq \|\frac{w^T X}{n}\|_\infty \|\hat{\Delta}\|_1 \leq \frac{\lambda}{2}(\|\theta^*\|_1 + \|\hat{\theta}\|_1)$$

$$0 \leq \frac{\lambda}{2}(\|\theta^*\|_1 + \|\hat{\theta}\|_1) + \lambda(\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$

So we can get $\|\hat{\Delta}\|_1 \leq 4\|\theta^*\|_1$

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \frac{\lambda}{2}\|\Delta\|_1 + \lambda(\|\theta^*\|_1 - \|\theta^* + \hat{\Delta}\|_1)\|\Delta\|_1 \leq \frac{3\lambda}{2}\|\hat{\Delta}\|_1$$

$\square$

Bounds on prediction error

Proof.

(b) From previous analysis we have

$$0 \leq \frac{1}{n}\|X\hat{\Delta}\|_2^2 \leq \frac{2w^T X\hat{\Delta}}{n} + 2\lambda(\|\theta_s^*\|_1 - \|\theta_s^* + \Delta_s\|_1 - \|\Delta_{S^c}\|_1)$$
$$\leq 2\|\frac{w^T X}{n}\|_\infty\|\hat{\Delta}\|_1 + 2\lambda(\|\Delta_s\|_1 - \|\Delta_{S^c}\|_1)$$
$$\leq \lambda(3\|\Delta_S\|_1 - \|\Delta_{S^c}\|_1)) \leq 3\sqrt{s}\lambda\|\hat{\Delta}\|_2^2$$

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq 3\lambda\|\hat{\Delta_S}\|_1 \leq 3\sqrt{s}\lambda\|\hat{\Delta_S}\|_2$$

Then we use the RE-condition. □

Varaible Selection

We need two assumptions

- Lower eigenvalue: The smallest eigenvalue of the sample covariance submatrix in- dexed by S is bounded below:

$$\gamma_{min}(\frac{X_S^T X_S}{n}) > c_{min} > 0$$

- Mutual incoherence: There exists some $\alpha \in [0, 1)$ such that

$$max_{j^c} \|(X_s^T X_S)^{-1} X_S^T X_j\|_1 \leq \alpha$$

Varaible Selection

We define $\Pi = I_n - (X_s^T X_S)^{-1} X_S^T X_S$.

Theorem

Consider an S-sparse linear regression model for which the design matrix satisfies the two assumptions. Then for any $\lambda \geq \frac{2}{1-\alpha} \| X_{S^c}^T \Pi_S(X) \frac{w}{n} \|_\infty$. *For* $min \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1$
(a)Uniqueness: There is a unique optimal solution$\hat{\theta}$.
(b)No false inclusion: This solution has its support set $\hat{S}$ *contained within the true support set S . (c) We have the follow bound*

$$\|\hat{\theta}_S - \theta_S^*\|_\infty \leq \|(\frac{X_s^T X_s}{n})^{-1} X_S^T \frac{w}{n}\|_\infty + \lambda \|(\frac{X_s^T X_s}{n})^{-1}\|_\infty$$

The term is denoted as $B(\lambda, X)$
(d) No false exclusion: The Lasso includes all indices, if $min_{i \in S} |\theta_i| > B(\lambda, X)$ *then the selection is consistent.*

Varaible Selection

subgraient, given a convex function $f : R^D \to R$ we say $z \in R^d$ a subgradient and denote it by $z \in \partial f(\ \theta)$ if we have for all $\Delta \in R^d$

$$f(\theta + \Delta) \geq f(\theta) + < z, \Delta >$$

For Lasso we have

$$\frac{1}{n} X^T (X\hat{\theta} - y) + \lambda z = 0$$

Varaible Selection

Primal-dual witness

- Set $\hat{\theta_{S^C}} = 0$
- Determine $(\hat{\theta_S}, \hat{z}_S) \in R^s \times R^s$ by solving

$$\hat{\theta}_S \in argmin_{\theta \in R^s}\{\frac{1}{2n}\|y - X_S\theta_S\|_2^2 + \lambda\|\theta_S\|_1\}$$

- Solve for $\hat{z}_{S^C}$ via zero-subgradient and check whether the strict dual $\|\hat{z}_{S^C}\|_\infty < 1$ holds.

$$\frac{1}{n}\begin{bmatrix} \mathbf{X}_S^T\mathbf{X}_S & \mathbf{X}_S^T\mathbf{X}_{S^c} \\ \mathbf{X}_{S^c}^T\mathbf{X}_S & \mathbf{X}_{S^c}^T\mathbf{X}_{S^c} \end{bmatrix}\begin{bmatrix} \widehat{\theta}_S - \theta_S^* \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \mathbf{X}_S^T w \\ \mathbf{X}_{S^c}^T w \end{bmatrix} + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Varaible Selection

Lemma

If the lower eigen value condition holds, the success of primal-dual witness construction implies that the vector $(\hat{\theta}_S, 0) \in R^d$ is the unique optimal solution of Lasso.

Proof.

When the construction succeeds, then $\hat{\theta} = (\hat{\theta_S}, 0)$ is optimal solution with $\hat{z}$ satisfies $\|\hat{z}_{S^c}\|_\infty$. Now let $\tilde{\theta}$ be any other optimal solution. If we introduce $F(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$, then

$$F(\hat{\theta}) - \lambda < \hat{z}, \tilde{\theta} - \hat{\theta} >= F(\tilde{\theta}) + \lambda(\|\tilde{\theta}\|_1 - < \hat{z}, \tilde{\theta} >)$$

$$F(\hat{\theta}) + < \nabla F(\hat{\theta}), \tilde{\theta} - \hat{\theta} > - F(\tilde{\theta}) = \lambda(\|\tilde{\theta}\|_1 - < \hat{z}, \tilde{\theta} >)$$

$\square$

Varaible Selection

$$\frac{1}{n}\begin{bmatrix} \mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S & \mathbf{X}_S^{\mathrm{T}}\mathbf{X}_{S^c} \\ \mathbf{X}_{S^c}^{\mathrm{T}}\mathbf{X}_S & \mathbf{X}_{S^c}^{\mathrm{T}}\mathbf{X}_{S^c} \end{bmatrix}\begin{bmatrix} \widehat{\theta}_S - \theta_S^* \\ 0 \end{bmatrix} - \frac{1}{n}\begin{bmatrix} \mathbf{X}_S^{\mathrm{T}} w \\ \mathbf{X}_{S^c}^{\mathrm{T}} w \end{bmatrix} + \lambda_n \begin{bmatrix} \widehat{z}_S \\ \widehat{z}_{S^c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Proof.

For construction let $\hat{z}_{S^C} = \frac{1}{n\lambda} X_{S^C}^T X_S(\hat{\theta}_S - \theta_S^*) + X_{S^C}^T(\frac{w}{\lambda n})$

$$\hat{\theta}_S - \theta_S^* = (X_S^T X_S)^{-1} X_S^T w - \lambda n (X_S^T X_S)^{-1} \hat{z}_S$$

$$\hat{z}_{S^C} = X_{S^C}^T X_S (X_S^T X_S)^{-1} \hat{z}_S + X_{S^C}^T \Pi(\frac{w}{\lambda n})$$

denote the first term as μ the second one as V we have $\|z_{S^C}\| + \infty \leq \|\mu\|_\infty + \|V\|_\infty$. The first term is less than α due to Mutual Incoherence assumption. The second one is less than $\frac{1-\alpha}{2}$. so we have $\hat{z}_{S^C} < 1$ $\square$

Corollary

Consider the S -sparse linear model based on a noise vector w with zero-mean i.i.d. σ-sub-Gaussian entries, and a deterministic design matrix X that satisfies assumptions , as well as the C-column normalization condition ($max_j \|X_j\|_2/\sqrt{n} \leq C$). Suppose that we solve the Lagrangian Lasso with regularization parameter

$$\lambda = \frac{2C\sigma}{1-\alpha}\{\sqrt{\frac{2log(d-s)}{n}} + \delta\}$$

for some $\delta > 0$, the optimal solution $\hat{\theta}$ is unique with its support contained within S and satisifies the following bound with probability $1 - 4e^{-\frac{n^2}{2}}$

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \{\sqrt{\frac{2log(d-s)}{n}} + \delta\}\frac{\sigma}{\sqrt{c_{min}}} + \lambda\|(\frac{X_s^T X_s}{n})^{-1}\|_\infty$$

Varaible Selection

Proof.

Firstly we check the λ, consider the following variable $Z_j = X_j^T \Pi_S(X(\frac{w}{n}))$ for $j \in S_C$. We have $\|\Pi_s(X)X_j\|_2 \leq \|X_j\|_2 \leq C\sqrt{n}$ So each Z_j is sub-Gaussian with parameter at most $\frac{C^2\sigma^2}{n}$.

$$P[max|Z_j| \geq t] \leq 2(d-s)e^{-\frac{nt^2}{2\sigma^2 C^2}}$$

For variable $\tilde{Z}_i = e_i^T(\frac{1}{n}X_s^T X_S)^{-1} X_S^T \frac{w}{n}$, and we have

$$\frac{\sigma^2}{n}\|(\frac{1}{n}X_S^T X_S)^{-1}\|_2 \leq \frac{\sigma^2}{c_{min}n}$$

$$P[max_{i\in S}|\tilde{Z}_i| > \frac{\sigma}{\sqrt{c_{min}}}(\sqrt{\frac{2logs}{n}} + \delta)] \leq 2e^{-\frac{n\delta^2}{2}}$$

□