Learning with Sparsity

Zhenwei Lin

日期：2020 年 8 月 14 日

摘　要

本文主要为几种稀疏模型的介绍：

关键词： Lasso; Ridege

1 Linear Regression

dataset is $\{(x_i, y_i)\}_{i=1}^{N}$,and the model is

$$\begin{aligned} y &= x^T\beta + \epsilon \\ E(\epsilon) &= 0 \\ Var(\epsilon) &< \infty \end{aligned} \tag{1}$$

2 Computational perspective

$$\min_{\beta} \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i^T\beta)^2 = \frac{1}{2}\|Y - X\beta\|^2$$

where

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \in R^N, x_{n\times p} = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} \tag{2}$$

为什么这里不把 β_0 单独给出，因为该数据是被标准化过后的数据，可以看到标准化后的数据其 β_0 为 0.

$$\nabla L(\beta) = x^T(Y - X\beta) = 0 \Rightarrow (x^T x)\beta = x^T Y$$

1. if X^TX is invertible ,we have $\hat{\beta} = (X^TX)^{-1}X^TY$
2. if $n < p$ means the size of sample is too small, we have rank(X^TX) = rank$(XX^T) \leq$ n
3. collinearity

To solve these problems, using **Tikhoov Regulation** $\hat{\beta_\lambda} = (X^TX + \lambda I)^{-1}X^TY$ where X^TX is a semidefinite matrix,and I is a positive definitive matrix.

Tikhoov Regulation in model means Ridge Regression.

$$\begin{aligned} \min_{\beta} \quad & \frac{1}{2}\|Y - x\beta\|^2 + \lambda\|\beta\|^2 \\ \Rightarrow \nabla L(\beta) &= -X^T(Y - X\beta) + 2\lambda\beta = 0 \\ \Rightarrow \hat{\beta}_\lambda &= (X^TX + \lambda I)^{-1}X^TY \end{aligned}$$

3 Bayesian Perspective

下面从贝叶斯的角度来对其进行一定的解释。

$$\begin{aligned} y_i &= x_i^T\beta + \epsilon_i \quad \epsilon_i \sim N(0,1) \\ \therefore f(y_i|x_i) &= \frac{1}{\sqrt{2\pi}}exp\left\{-\frac{(y_i - x_i^T\beta)^2}{2}\right\} \\ \Rightarrow \max_\beta \log\prod_{i=1}^{N} f(y_i|x_i) &\iff \min_\beta \frac{1}{2}\|Y - x\beta\|^2 \end{aligned}$$

by $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$

prior(先验信息)(对 β 的理解)

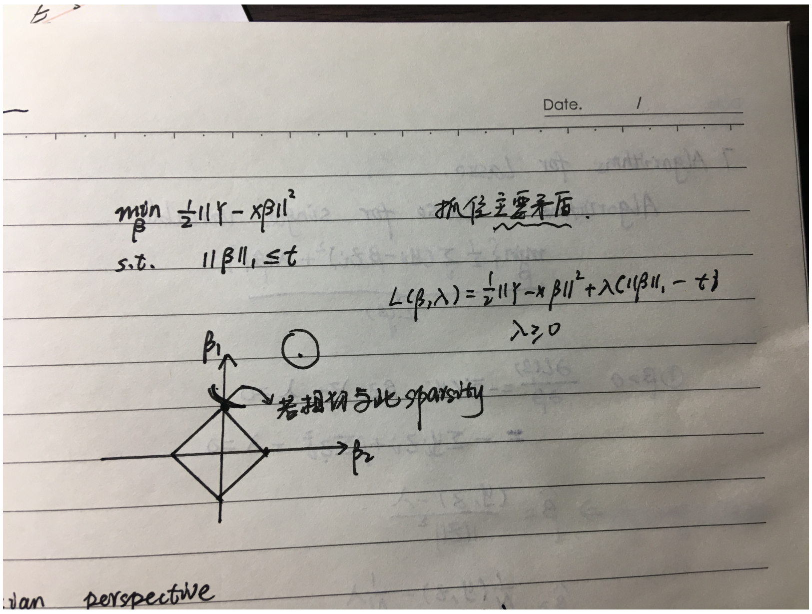
$$\begin{aligned} f(\beta) &= \frac{\lambda^{\frac{p}{2}}}{\sqrt{2\pi}}exp\left\{-\lambda\|\beta\|^2\right\} \\ \because f(y_i|x_i) &= \frac{1}{\sqrt{2\pi}}exp\left\{-\frac{(y_i - x_i^T\beta)^2}{2}\right\} \\ \therefore p(\beta|y,x) &= \frac{P(Y|\beta,x)P(\beta|x)P(x)}{P(Y|x)P(x)} = \frac{P(Y|\beta,x)P(\beta|x)}{P(Y|x)} \propto P(Y|\beta,x)P(\beta|x) = P(Y|\beta,x)P(\beta) \\ &\iff \max_\beta P(Y|x,\beta)P(\beta) \iff \min_\beta \frac{1}{2}\|Y - X\beta\|^2 + \tilde{\lambda}\|\beta\|^2 \quad where \tilde{\lambda} = \frac{\lambda}{2} \end{aligned}$$

因为 x 与 β 无关, 所以 $P(\beta|x) = P(\beta)$ 成立。

4 Lasso(least absolute shrinkage and selection operator)

$$\begin{aligned} \min_\beta \quad & \frac{1}{2}\|Y - x\beta\|^2 \\ & s.t.\ \|\beta\|_1 \le t \end{aligned}$$

and the intution could be shown in the following graph.



In Bayesian Perspective: the prior is $P(B) = constant \times exp\left\{-\lambda \|\beta\|_1\right\}$ (called laplace distribution)

$$
\begin{aligned}
p(B) &\propto p(Y|x,\beta)p(\beta) \\
\max_{\beta} log \prod_{i=1}^{N} \exp\left\{-\frac{(y_i - x_i^T\beta)^2}{2}\right\} &\exp\left\{-\lambda \|\beta\|_1\right\} \\
\max_{\beta} -\frac{1}{2}\|Y - x\beta\|^2 &- \lambda \|\beta\|_1 \\
\min_{\beta} \frac{1}{2}\|Y - x\beta\|^2 &+ \lambda \|\beta\|_1 \, (\lambda \geq 0)
\end{aligned}
$$

5 Algorithms for lasso

5.1 Algorithm 1(lasso for single variable)

$$
\underbrace{\min_{\beta}\left\{\frac{1}{2}\sum_i (y_i - \beta z_i)^2 + \lambda \|\beta\|_1\right\}}_{L(\beta)} \tag{3}
$$

1. when $\beta > 0$,

$$
\begin{aligned}
\frac{\partial L(\beta)}{\partial \beta} &= -\sum_i (y_i - \beta z_i) z_i + \lambda = 0 \\
&= -\sum_i y_i z_i + \beta \sum_i z_i^2 + \lambda = 0 \\
\Rightarrow \hat{\beta} &= \frac{(y,z) - \lambda}{\|z\|^2} \Rightarrow \hat{\beta} = \frac{\frac{1}{N}(y,z) - \frac{1}{N}\lambda}{\frac{1}{N}\|z\|^2} \\
\because \frac{1}{N}\|z\|^2 &= 1 \\
\therefore \hat{\beta} &= \frac{1}{N}(y,z) - \frac{1}{N}\lambda = \frac{1}{N}(y,z) - \tilde{\lambda}
\end{aligned} \tag{4}
$$

2. when $\beta < 0$,similarly we have

$$\hat{\beta} = \frac{1}{N}(y,z) + \tilde{\lambda}$$

3. when $\beta = 0$

$$\begin{aligned}
&0 \in \partial L(0) \\
&0 \in -\frac{1}{N}\sum(y_i - \beta z_i)z_i + \lambda\partial|\beta| \\
\overset{\text{by}\beta=0}{\iff}\ &0 \in -\frac{1}{N}(y,z) + \lambda\partial|\beta| \\
&\Rightarrow \frac{1}{N}(y,z) = \lambda\partial|0| = [-\lambda, \lambda]
\end{aligned}$$

summary: we have

$$\hat{\beta} = \begin{cases} \frac{1}{N}(y,z) - \tilde{\lambda} & if\, \frac{1}{N}(y,z) > \tilde{\lambda} \\ 0 & if\, \left|\frac{1}{N}(y,z)\right| \le \tilde{\lambda} \\ \frac{1}{N}(y,z) + \tilde{\lambda} & if\, \frac{1}{N}(y,z) < -\tilde{\lambda} \end{cases}$$

从计算结果来说，其内积即相关性越小，则其数值更偏向于 0。

so $\hat{\beta} = Soft_{\lambda}(\frac{1}{N}(y,z))$

5.2 Algorithm 2(Multivariable of orthogonal Design)

$$X_{n\times p}^T X = I$$

$$\min_{\beta} \quad \frac{1}{2N}\|Y - X\beta\|^2 + \lambda\|\beta\|_1$$

$$\begin{aligned}
\|Y - x\beta\|^2 &= \|Y\|^2 - 2(Y, X\beta) + \beta^T X^T X\beta \\
&= \|\beta\|^2 - 2(X^T Y, \beta) + \|Y\|^2 \\
&= \left\|\beta - X^T Y\right\|^2 + \|Y\|^2 - \left\|X^T Y\right\| \\
\iff min_{\beta}&\left\{\frac{1}{2N}\left\|\beta - x^T Y\right\|^2 + \lambda\|\beta\|_1\right\}
\end{aligned}$$

can be decomposition.

5.3 Algorithm 3 (Cyclic Coordinate Descent)

$$\beta \in \mathbf{R^p} = (\beta_1, \cdots, \beta_p)^T$$

and $\beta_2, \cdots, \beta_p$ is known.

$$\begin{aligned}
&\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(y_i - \sum_{i=2}^{p}\beta_j x_j - \beta_1 x_{i1})^2 + \lambda\sum_{i=2}^{p}|\beta_i| + \lambda|\beta_1|\right\} \\
&\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}(r_{i1} - \beta_1 x_{i1})^2 + \lambda|\beta_1|\right\} \\
&\Rightarrow \hat{\beta}_1 = Soft_{\lambda}(\frac{1}{N}(r_i, x_1))
\end{aligned} \tag{4}$$

按照一个一个坐标轴分开进行优化。

5.4 Algorithm 4 Proximal gradient descent

$$\min\left\{\frac{1}{2N}\|Y-x\beta\|^2+\lambda\|\beta\|_1\right\}$$

and we know $\frac{1}{2N}\|Y-x\beta\|^2$ is a β-smooth function.

so

$$\begin{aligned}
&\|\nabla f(\beta_1)-\nabla f(\beta_2)\| \\
=&\left\|\frac{1}{N}X^T(Y-X\beta_1)-\frac{1}{N}X^T(Y-X\beta_2)\right\| \\
=&\frac{1}{N}\left\|X^TX(\beta_1-\beta_2)\right\| \le \frac{\lambda_{max}(X^TX)}{N}\|\beta_1-\beta_2\| \\
\Rightarrow b =&\frac{\lambda_{max}(X^TX)}{N}
\end{aligned}$$

$$\therefore \beta_{t+1}=prox_{g/b}(\beta_t-\frac{1}{b}\nabla f(\beta))$$

recall in definition, we have

$$prox_{g/b}(z)=argmin_x\left\{\frac{b}{2}\|x-z\|^2+g(z)\right\}$$

$$min_x\left\{\frac{1}{2}\|x-z\|^2+\frac{\lambda}{b}\|x\|_1\right\}\Rightarrow prox_{g/b}=Soft_{\lambda/b}(z)$$

$$\begin{aligned}
\therefore \hat{\beta}_{t+1} =&Soft_{\lambda/b}(\beta_t+\frac{N}{\lambda_{max}(X^TX)}\frac{1}{N}X^T(Y-X\beta_t)) \\
=&Soft_{\lambda/b}(\beta_t-\frac{X^TX}{\lambda_{\max}(X^TX)}\beta_t+\frac{X^TY}{\lambda_{max}(X^TX)}) \\
=&Soft_{\lambda/b}\left\{(I-\frac{X^TX}{\lambda_{max}(X^TX)})\beta_t+\frac{X^TY}{\lambda_{max}(X^TX)}\right\}
\end{aligned}$$

5.5 Algorithm 5 : AGD → FISTA

$$\begin{aligned}
\beta_0 =&0, \alpha_1=\beta_0, t=0 \\
\beta_t =&Soft_{\lambda/b}\left\{(I-\frac{X^TX}{\lambda_{max}(X^TX)})\alpha_t+\frac{X^TY}{\lambda_{max}(X^TX)}\right\} \\
a_{t+1} =&\frac{1+\sqrt{1+4a_t^2}}{2} \\
\alpha_{t+1} =&\beta_t+\frac{a_t-1}{a_{t+1}}(\beta_t-\beta_{t+1})
\end{aligned}$$

and the convergence speed is :$O(\frac{1}{T^2})$

5.6 Algorithm6:ADMM

$$\begin{aligned}
&\min_{\beta} \quad \left\{\frac{1}{2N}\|Y - X\beta\|^2 + \lambda\|\beta\|_1\right\} \\
\Longleftrightarrow &\min_{\beta} \quad \left\{\frac{1}{2N}\|Y - X\beta\|^2\right\} \\
&s.t. \quad \beta - \alpha = 0
\end{aligned}$$

$$\begin{aligned}
L(\beta, \alpha, v) =& \frac{1}{2N}\|Y - x\beta\|^2 + \lambda\|\alpha\|_1 + v^T(\beta - \alpha) + \frac{\rho}{2}\|\beta - \alpha\|^2 \\
\frac{\partial L}{\partial \beta} =& -\frac{1}{N}X^T(Y - X\beta) + v + \rho(\beta - \alpha) \\
=& (\frac{1}{N}X^T X + \rho I)\beta + v - \frac{1}{N}X^T Y - \rho\alpha = 0 \\
\Rightarrow \hat{\beta} =& (\frac{1}{N}X^T X + \rho I)^{-1}(\frac{1}{N}X^T Y - v + \rho\alpha)
\end{aligned}$$

for α :

$$\begin{aligned}
&\min_{\alpha}\left\{\frac{\rho}{2}\|\alpha - \beta\|^2 - v^T(\alpha - \beta) + \lambda\|\alpha\|_1\right\} \\
\Longleftrightarrow &\min_{\alpha}\left\{\frac{\rho}{2}\left\|\alpha - \beta - \frac{v}{\rho}\right\|^2 + \lambda\|\alpha\|_1\right\} \\
\Longleftrightarrow &\min_{\alpha}\left\{\frac{1}{2}\left\|\alpha - \beta - \frac{v}{\rho}\right\|^2 + \frac{\lambda}{\rho}\|\alpha\|_1\right\} \\
&\Rightarrow \hat{\alpha} = Soft_{\lambda/\rho}(\beta + \frac{v}{\rho}) \\
\therefore\ &\beta_{t+1} = (\frac{1}{N}X^T X + \rho I)^{-1}(\frac{1}{N}X^T Y - v + \rho\alpha_t) \\
&\alpha_{t+1} = Soft_{\lambda/\rho}(\beta_{t+1} - \frac{v_t}{\rho}) \\
&v_{t+1} = v_t + \rho(\beta_{t+1} - \alpha_{t+1})
\end{aligned}$$

5.7 Alogrithm 7: Dual Method

$$\begin{aligned}
&\min_{\beta} \quad \left\{\frac{1}{2N}\|Y - X\beta\|^2 + \lambda\|\beta\|_1\right\} \\
&\min_{\beta} \quad \left\{\sup_{\alpha}\left\{(\alpha, Y - X\beta) - \frac{1}{2}\|\alpha\|^2\right\} + \lambda\sup_{\lambda}\left\{(\gamma, \beta) - \delta_{B_\infty}(\gamma)\right\}\right\}
\end{aligned}$$

6 Analysis of LASSO

当我们使用数值方法求解出了 $\hat{\beta} = argmin\left\{\frac{1}{2N}\|Y - X\beta\|^2 + \lambda\|\beta\|_1\right\}$

但是我们考虑真实的模型为 $Y = X^T\beta^* + \epsilon$, where β^* is real coefficient.

这时我们对这个问题有三个方法可以来阐述其误差大小

(a) l_2 error $L_2(\hat{\beta}, \beta^*) = \left\|\hat{\beta} - \beta^*\right\|_2$

(b) Prediction error : $L_p(\hat{\beta}, \beta^*) = \frac{1}{N}\left\|X\beta^* - X\hat{\beta}\right\|^2$

(c) Variable Selection Error : $\left\{j | sign(\hat{\beta}) \neq sign(\beta^*)\right\}$

下面考虑第一种从 l_2-error 的角度来看问题。

denote $\hat{v} = \hat{\beta} - \beta^*$

$$\Rightarrow L_2(\hat{\beta}, \beta^*) = \|\hat{v}\|$$

$$\Rightarrow L_p(\hat{\beta}, \beta^*) = \frac{1}{N}\|x\hat{v}\|_2^2$$

recall: α-strong convex function

$$f(y) - f(x) - (\nabla f(x), y - x) \geq \frac{\alpha}{2}\|y - x\|$$

这里假设 $f(\beta)$ 是一个 α -strong convex function

再对 $f(\hat{\beta})$ 在 β^* 处进行 taylor 展开，则有

$$f(\hat{\beta}) - f(\beta^*) - (\nabla f(\beta^*), \hat{\beta} - \beta^*) = \frac{1}{2N}\hat{v}^T x^T x\hat{v} \geq \frac{\gamma}{2}\|\hat{v}\|^2 \tag{5}$$

由于无法认为 $X^T X$ 满秩，故在某些区域上满足该条件进行思考，鉴于此，我们有 restricted strong convex 成立。

定义 6.1. Restricted Strong convex

$$f(y) - f(x) - (\nabla f(x), y - x) \geq \frac{\gamma}{2}\|y - x\|^2 \quad \forall x, y, y - x \in C$$

定义 6.2. Restricted Eigenvalue Condition(REC)

$$\frac{\hat{v}^T x^T x\hat{v}}{N} \geq \gamma\|\hat{v}\|^2 \quad \hat{v} \in C$$

现在我们先假设 REC 条件成立，考虑以下两个 lasso 形式

1.
$$\begin{aligned} \min \quad & \frac{1}{2}\|Y - X\beta\|^2 \\ s.t. \quad & \|\beta\|_1 \leq R \end{aligned} \tag{6}$$

2. Lagrangian Lasso
$$\min_{\beta}\left\{\frac{1}{2N}\|Y - X\beta\|^2 + \lambda_N\|\beta\|_1\right\} \tag{7}$$

那么 C 是哪个集合呢？我们先有如下符号定义：

$$\|\beta^*\|_0 = k$$

说明其有 k 个位置不为 0

$$\begin{aligned} S &\subseteq \{1, 2, \cdots, p\} \\ v_s &= \left\{v | v_j = 0, j \notin s\right\} \\ v_{s^c} &= \left\{v | v_j = 0, j \in s\right\} \end{aligned}$$

例 6.1.

$$\begin{aligned} v &= (1, 2, 3) \\ s &= \{2\} \\ \Rightarrow v_s &= \{(0, 2, 0)\} \\ \Rightarrow v_{s^c} &= \{(1, 0, 3)\} \end{aligned} \tag{8}$$

因此我们特定的取

$$s = \left\{i|\beta_i^* \neq 0\right\} \Rightarrow \|\beta^*\|_1 = \left\|\beta_s^*\right\|_1$$

引理 6.1. Assume

$$\hat{\beta} \in \underset{\|\beta\|_1 \leq R}{argmin} \left\{\frac{1}{2N} \|Y - X\beta\|^2\right\}$$

$$\hat{v} = \hat{\beta} - \beta^*$$

then $\|\hat{v}_{s^c}\|_1 \leq \|\hat{v}_s\|_1$ where s $= \{j|\beta^* \neq 0\}$

证明. 假设其在边界上达到最优

$$R = \|\beta^*\|_1 \geq \left\|\hat{\beta}\right\|_1 = \|\hat{v} + \beta^*\|_1 = \|\hat{v}_s + \hat{v}_{s^c} + \beta^*\|_1$$

$$= \underbrace{\|\hat{v}_{s^c}\|_1}_{\text{取出 } (\hat{\beta}-\beta^*) \text{ 中 } \beta^*=0 \text{ 的部分}} + \underbrace{\|\beta^* + \hat{v}_s\|_1}_{\text{取出 } (\hat{\beta}-\beta^*) \text{ 中 } \beta^* \neq 0 \text{ 的部分}} \overset{\text{triangle inequality}}{\geq} \|\beta^*\|_1 - \|\hat{v}_s\|_1 + \|\hat{v}_{s^c}\|_1$$

$$\Rightarrow \|\beta^*\|_1 \geq \|\beta^*\|_1 - \|\hat{v}_s\|_1 + \|\hat{v}_{s^c}\|_1$$

□

Definition 6.1.

$$C(S, 1) = \{v| \|v_{s^c}\|_1 \leq 1 \|v_s\|_1\} \tag{9}$$

引理 6.2.

$$\begin{gathered} p \geq q \geq 1, \quad x \in R^d \\ \|x\|_p \leq \|x\|_q \leq d^{\frac{1}{q}-\frac{1}{p}} \|x\|_p \end{gathered} \tag{10}$$

等价范数，范数可以被同样上下限控制住。

定理 6.3. Consider Constrainted LASSO.Assume REC over C(S,1) then

(a) $\left\|\hat{\beta} - \beta^*\right\| \leq \frac{4\sqrt{k}}{\gamma} \left\|\frac{x^T\epsilon}{N}\right\|_\infty$

(b) if we futher assume that $\epsilon \overset{i.i.d}{\sim} N(0, \sigma^2)$ then for any $0 < \delta < 1$ we have

$$\left\|\hat{\beta} - \beta^*\right\|_2 \leq C(\delta)\frac{4\sigma}{\gamma}\sqrt{\frac{klogP}{N}}$$

依概率 $1 - \delta$ 成立，其中 P 为维数 P.

(a)proof: Basci inequality

$$\hat{v} = \hat{\beta} - \beta^*$$

$$\frac{1}{2N} \|Y - X\beta\|^2 \leq \frac{1}{2N} \|Y - X\beta^*\|^2 \qquad Y = X\beta^* + \epsilon$$

basic inequality 成立的原因在于 $\hat{\beta}$ 是其最小化后的结果，而 β^* 却只是其中符合模型的最优解

$$\Rightarrow \left\|Y - X\beta^* + X\beta^* - X\hat{\beta}\right\|^2 \le \|\epsilon\|^2$$

$$\left\|\epsilon + X(\beta^* - \hat{\beta})\right\|^2 \le \|\epsilon\|^2$$

$$\Rightarrow \|X\hat{v}\|^2 \le 2(X^T\epsilon, \hat{v})$$

$$\Rightarrow \frac{1}{N}\|X\hat{v}\|^2 \overset{\text{Cauchy-Schwarz inequality 对任意对偶范数成立}}{\le} 2\left\|\frac{X^T\epsilon}{N}\right\|_\infty \|\hat{v}\|_1$$

$$\overset{\text{by REC}}{\rightarrow} \frac{1}{N}\|X\hat{v}\|_2^2 \ge \gamma\|\hat{v}\|_2^2$$

$$\|\hat{v}\|_1 = \|\hat{v}_s + \hat{v}_{s^c}\|_1 \le 2\|\hat{v}_s\|_1 \le 2\sqrt{k}\|\hat{v}_s\|_2 \le 2\sqrt{k}\|\hat{v}\|_2$$

$$\Rightarrow \gamma\|\hat{v}\|_2^2 \le 2\left\|\frac{X^T\epsilon}{N}\right\|_\infty \|\hat{v}\|_1 \le 4\sqrt{k}\left\|\frac{X^T\epsilon}{N}\right\|_\infty \|\hat{v}\|_2$$

$$\Rightarrow \|\hat{v}\|_2 \le \frac{4\sqrt{k}}{\gamma}\left\|\frac{X^T\epsilon}{N}\right\|_\infty$$

□

(b)proof:

$$\epsilon_i \sim N(0,\sigma^2) \quad X = (x_1, \cdots, x_p) \quad X^T\epsilon = \begin{pmatrix} x_1^T\epsilon \\ x_2^T\epsilon \\ \vdots \\ x_p^T\epsilon \end{pmatrix}$$

$$\frac{x_1^T\epsilon}{N} \sim N(0, \frac{\sigma^2\|x_i\|^2}{N^2}) \overset{\frac{\|x_i\|^2}{N}=1}{=} N(0, \frac{\sigma^2}{N})$$

$$p(|z| \ge t) \overset{z\sim N(0,1)}{\le} 2exp(-\frac{t^2}{2\sigma^2})$$

$$\Rightarrow p(\left|\frac{x_i^T\epsilon}{N}\right| \ge t) \le 2exp(-\frac{Nt^2}{2\sigma^2})$$

$$p(\left\|\frac{x_i^T\epsilon}{N}\right\|_\infty \ge t) \le \sum_{j=1}^{P} p(\frac{|x_i^T\epsilon|}{N} \ge t) = \underbrace{2Pexp(-\frac{Nt^2}{2\sigma^2})}_{\delta}$$

无穷范数表示其中的最大值

$$\Rightarrow log(\frac{\delta}{2p}) = -\frac{Nt^2}{2\sigma^2}$$

$$t = \sqrt{\frac{2\sigma^2}{N}log(\frac{p}{2\delta})} = c(\delta)\cdot\sigma\sqrt{\frac{log(p)}{N}}$$

代入 a 的结论即可得证。 □

这个定理想表达的主要有：

1. σ 越大，$\left\|\hat{\beta} - \beta^*\right\|$ 越大不好估计
2. k 越大，信息量越大，则不好找最优解
3. want $\frac{logP}{N} \rightarrow 0$,we have $O(logP) < O(N)$ 同时要求 k 小点，eg:N=100, 则有 $\Rightarrow P = exp(10)$

Consider Lagrangian LASSO

引理 6.4. If $\lambda_N \geq \frac{2\|x^T\epsilon\|_\infty}{N}$ then

$$\hat{v} \in C(S,3) = \{v|\ \|v_{S^c}\|_1 \leq 3\|v_S\|_1\}$$

证明. By Basic inequality,we have

$$\frac{1}{2N}\left\|Y - X\hat{\beta}\right\|^2 + \lambda_N\left\|\hat{\beta}\right\|_1 \leq \frac{1}{2N}\|Y - X\beta^*\|^2 + \lambda_N\|\beta^*\|_1$$

$$\Rightarrow 0 \leq \frac{1}{N}\|X\hat{v}\|^2 \leq \frac{\left\|X^T\epsilon\right\|_\infty}{N}\|\hat{v}\|_1 + \lambda_N(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1)$$

$$\Rightarrow 0 \leq \frac{\lambda_N}{2}\|\hat{v}\|_1 + \lambda_N(\|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1)$$

$$\left\|\hat{\beta}\right\|_1 = \|\hat{v} + \beta^*\|_1 = \|\hat{v_S} + \beta^* + \hat{v}_{S^c}\|_1 = \|\hat{v}_S + \beta^*\|_1 + \|\hat{v}_{S^c}\|_1 \geq \|\beta^*\|_1 - \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1$$

$$\Rightarrow \|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1 \geq \|\beta^*\|_1 - \left\|\hat{\beta}\right\|_1$$

$$\Rightarrow \frac{1}{2}(\|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1) + (\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|) \geq 0$$

$$\Rightarrow \|\hat{v}_{S^c}\|_1 \leq 2\|\hat{v}_S\|_1$$

□

定理 6.5. let us consider the **Lagrangian LASSO** problem, and assume that x satifies REC over C(S,3), then for any $\lambda_N \geq 2\left\|\frac{X^T\epsilon}{N}\right\|_\infty$,we have

(a) $\left\|\hat{\beta} - \beta^*\right\|_1 \leq \frac{3\sqrt{k}}{\gamma}\lambda_N$

(b) If we futher assume $\epsilon \overset{i.i.d}{\sim} N(0,\sigma^2)$,then we have $\left\|\hat{\beta} - \beta^*\right\| \leq \frac{C(s)}{\gamma}\sigma\sqrt{\frac{klogP}{N}}$