Chapter 7: Sparse Linear Models in High-dimensions:

7.1a) $$\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y-\theta\|_2^2 + \frac{1}{2}\lambda^2\|\theta\|_0 \right\} = \min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y-\theta\|_2^2 + \frac{1}{2}\lambda^2 \right\}$$

$$= 0$$

$$\theta^* = y$$

An evaluation at θ equals zero:

$$\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y-\theta\|_2^2 + \frac{1}{2}\lambda^2 \right\} = \min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y\|^2 + \frac{1}{2}\lambda^2 \right\}$$

$$> 0$$

$$|y| > \lambda$$

Hard-threshold:

$$H_\lambda(y) = \begin{cases} y_i & \text{if } |y| \geq \lambda \\ 0 & \text{otherwise.} \end{cases}$$

b) $$\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2}\|y-\theta\|_2^2 + \lambda\|\theta\|_1 \right\} = -\|y-\theta\| + \lambda$$

$$= 0$$

$$\theta^* = |y| - \lambda$$

An evaluation at θ^* equals zero:

$$\theta^* = |y| - \lambda \quad ; \quad |y| = \lambda$$

$$= 0$$

Hard-threshold:

$$T_\lambda(y) = \begin{cases} \text{sign}(y)(|y|-\lambda) & |y| \geq \lambda \\ 0 & \text{otherwise} \end{cases}$$

7.2 $$B_q(R_q) = \left\{ \theta \in \mathbb{R}^d \;\middle|\; \sum_{j=1}^{d} |\theta_j|^q \leq R_q \right\}$$

$$B_{w(q)}(R_q) = \left\{ \theta \in \mathbb{R}^d \;\middle|\; |\theta|_{(j)} \leq C_j^{-\alpha} \text{ for } i = 1, \dots, d \right\}$$

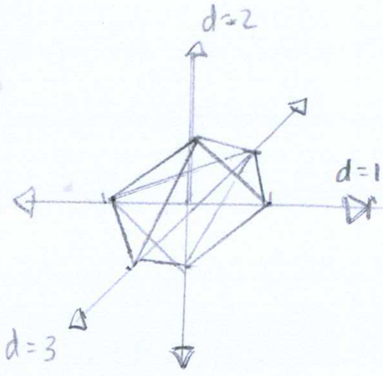
$|\theta|_j$ = order statistic of θ^* where $|\theta|_1 = \max|\theta_j|$ and $|\theta|_d = \min|\theta_j|$

a) Star Shape Set $= \{ C \subseteq \mathbb{R}^d \mid \theta \in C \to t\theta \in C \text{ for all } t \in [0,1] \}$

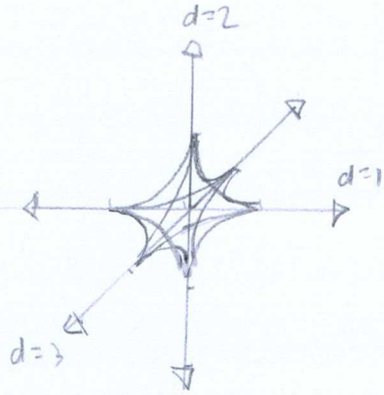
Original Strong l_q-ball : $B_q(R_q) = \{ \theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q \}$

Scaled Strong l_q-ball : $tB_q(R_q) = \{ t\theta \in \mathbb{R}^d \mid t^q \sum_{j=1}^{d} |\theta_j|^q \leq R_q \}$

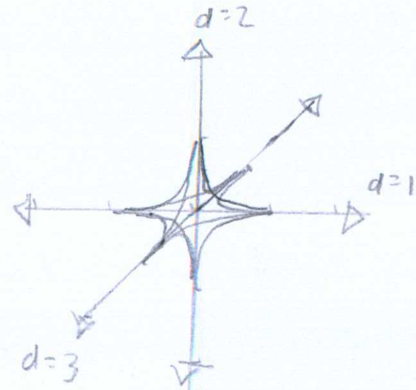
A plot of $B_q(1) = \{ \theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq 1 \}$



$\boxed{q=1}$



$\boxed{q=2}$



$\boxed{q=3}$

b) $\alpha > 1/q \quad B_{w(\alpha)}(C) \subseteq B_q(R_q)$

$$B_{w(\alpha)}(C)^q = \{ \theta^q \in \mathbb{R}^d \mid |\theta|^q \leq \left(\frac{c}{j^\alpha}\right)^q \text{ for } j = 1, \dots, d \}$$

$$= \{ \theta^q \in \mathbb{R}^d \mid |\theta|^q \leq \frac{c^q}{j} \text{ for } j = 1, \dots, d \}$$

$$\subseteq \{ \theta \in \mathbb{R}^d \mid \sum |\theta|^q \leq R_q \}$$

$$\subseteq B_q(R_q)$$

c) $$\Pi_s(\theta^*) = \underset{\|\theta\|_0 \leq s}{\operatorname{argmin}} \|\theta - \theta^*\|_2^2$$

$$= \{ \theta \in \mathbb{R} \mid \|\theta\|_0 \leq s \}$$

d) $\theta^* \in B_q(R_q)$:

$$\|\Pi_s(\theta - \theta^*)\|_2^2 = \sum_{j=s+1}^{d} |\theta_j|^2$$

$$\leq |\theta_s^*|^{2-q} \sum_{j=s+1}^{d} |\theta_j^*|^2$$

$$\leq \left(\frac{1}{s} \sum_{j=s+1}^{s} |\theta_i|^2\right)^{\frac{2-q}{q}} \sum_{j=s+1}^{d} |\theta_i^*|^q$$

$$\le \left(\frac{1}{s}\right)^{\frac{2-q}{2}} \left(\sum_{i=1}^{s} |\theta_i|^q\right)^{\frac{2-q}{q}} \sum_{j=s+1}^{d} |\theta_j|^q$$

$$\le \left(\frac{1}{s}\right)^{\frac{2-q}{q}} \left(\sum_{i=1}^{s} |\theta_i|^q\right)^{\frac{2}{q}}$$

$$\le (R_q)^{2/q} s^{1-2/q}$$

7.3.

a) $S \subset [1, 2, \ldots, d]$

$$\frac{X_S^T X_S}{n} = \begin{bmatrix} \frac{(X_S)_1^T (X_S)_1}{n} & \frac{(X_S)_2^T (X_S)_1}{n} & \cdots & \frac{(X_S)_s^T (X_S)_1}{n} \\ \frac{(X_S)_1^T (X_S)_2}{n} & \frac{(X_S)_2^T (X_S)_2}{n} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(X_S)_1^T (X_S)_s}{n} & \cdots & \cdots & \frac{(X_S)_s^T (X_S)_s}{n} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \frac{(X_S)_2^T (X_S)_1}{n} & \cdots & \frac{(X_S)_s^T (X_S)_1}{n} \\ \frac{(X_S)_1^T (X_S)_2}{n} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{(X_S)_1^T (X_S)_s}{n} & \cdots & \cdots & 1 \end{bmatrix}$$

$$\gamma_{min}\left(\frac{X_S^T X_S}{n}\right) = v^T \frac{X_S^T X_S}{n} v$$

$$= \sum_{i=1}^{s} v_i^2 + 2 \sum_{i<j}^{s} \frac{(X_S)_i^T (X_S)_j}{n} v_i v_j \quad , \qquad |v_i|^2 = 1$$

$$= 1 + 2 \sum_{i<j}^{s} \frac{(X_S)_i^T (X_S)_j}{n} v_i v_j$$

$$\ge 1 - 2 \frac{\gamma}{s} \sum_{i<j}^{s} |v_i v_j|$$

$$\ge 1 - \frac{\gamma}{s} \left((|v_1| + \cdots + |v_s|)^2 - \sum_{i=1}^{s} v_i^2 \right)$$

$$\geq 1 - \frac{\gamma}{s}\left(s\|v\|_1^2 - \|v\|_2^2\right)$$

$$\geq 1 - \frac{\gamma}{s}(s-1) \quad , \quad \|v\|_2^2 = 1$$

$$\geq c(\gamma)$$

$$\gamma_{min}\left(\frac{X_S^T X_S}{n}\right) = \inf v^T \frac{X_S^T X_S}{n} v$$

$$\geq c(\gamma)$$

(Definition 7.7)

The matrix X satisfies a restricted nullspace with S if

$C(S) \cap null(X) = \{0\}$ in an l_1-norm.

b) If $\gamma < 1/3$, $\gamma_{min}\left(\frac{X_S^T X_S}{n}\right) = 1 - \gamma$

$$\geq 1 - \gamma \frac{s-1}{s}$$

$$\geq 1 - \frac{1}{3}\frac{s-1}{s}$$

$$\geq \frac{2}{3} - \frac{1}{3s}$$

$$\boxed{\text{The derivation intersects zero at } s = \tfrac{1}{2} \text{, as with the definition of 7.7.}}$$

$$\gamma_{min}\left(\frac{X_S^T X_S}{n}\right) \geq 2/3$$

$$\geq \frac{2}{3} - \frac{1}{3s}$$

(Proposition 7.9)

If a pairwise coherence satisfies the bound $\delta_{PW}(X) \leq \frac{1}{3s}$ then additionally the nullspace property for all subsets S of cardinality at most s.

(Definition 7.10) (Restricted Isometry Property)

For a given integer $s \in \{1, \ldots, d\}$, $X \in \mathbb{R}^{n \times d}$ satisfies the restricted isometry property of order s with constant $\delta(X) > 0$ if $\left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right| \leq \delta_s(X)$ for subsets S of size s.

7.4

a) $\delta_{PW}(X) \le \delta_S(X) \longleftrightarrow \delta_S(X) \le s\,\delta_{PW}(X)$

$$\min_{j=k}\left|\frac{\langle X_j, X_k\rangle}{n}\right| \le \left|\frac{\langle X_j, X_k\rangle}{n}\right| \longleftrightarrow \left|\frac{\langle X_j, X_k\rangle}{n}\right| \le s \circ \max_{j\neq k}\left|\frac{\langle X_j, X_k\rangle}{n}\right|$$

$$(1-\gamma)\left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right| \le \left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right| \leftrightarrow \left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right| \le s(1+\gamma)\left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right|$$

$$(1-\gamma)\left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right| \le \left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right| \le s(1+\gamma)\left|\left|\left|\frac{X_S^T X_S}{n} - I\right|\right|\right|$$

$$\delta_{PW}(X) \le \delta_S(X) \le s\,\delta_{PW}(X)$$

b) $\delta_S(W) = \left|\left|\left|\frac{X^T X}{n} - I\right|\right|\right|_2 =$

$$= \sqrt{s} \circ \max_{j \neq k}\left|\frac{\langle X_j, X_k\rangle}{n}\right|_2$$

$$= \sqrt{s}\,\delta_{PW}(X)$$

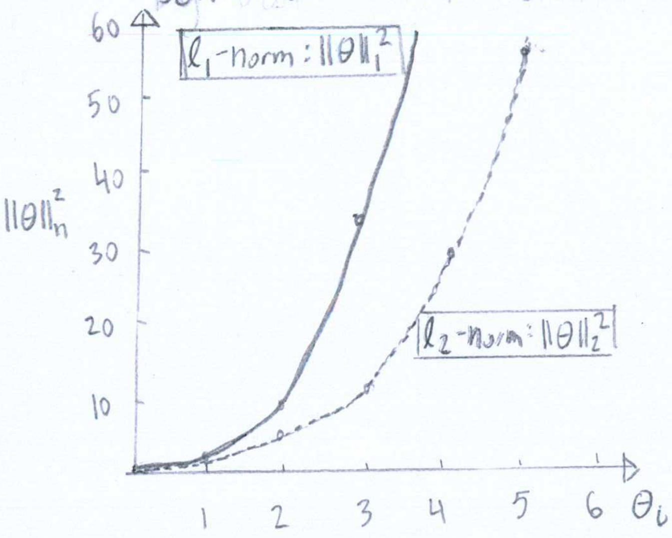
(Definition 7.7) The matrix X satisfies the "restricted eigenvalue condition" over S with parameter $\frac{1}{n}\|X\Delta\|_2^2 \ge \kappa\|\Delta\|_2^2$ for all $\Delta \in C_\infty(S)$.

7.5 $C_1(S) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \le \alpha_1 \|\theta\|_1\}$

$C_2(S) = \{\theta \in \mathbb{R}^d \mid \|\theta_2\|_2 \le \alpha_2 \|\theta\|_2\}$

Definition 7.7 defines $\frac{1}{n}\|X\theta\|_2^2 \ge \kappa\|\theta\|_2^2$

$$\ge \frac{\gamma_2}{s}\|\theta\|_2^2$$



Since $\frac{\gamma_1}{s}\|\theta\|_1^2 \ge \frac{\gamma_2}{s}\|\theta\|_2^2$

$$\frac{1}{n}\|X\theta\|_2^2 \ge \frac{\gamma_1}{s}\|\theta\|_1^2$$

Where $\|\theta\|_1^2 = \left(\sum|\theta_i|\right)^2$

$\|\theta\|_2^2 = \sum|\theta_i|^2$

7.6 $||\theta||_{v(1)} = \sum_{i=1}^{d} w_i |\theta_i|$ where $w \in \mathbb{R}^d$

Function: $\min_{\theta \in \mathbb{R}^d} ||\theta||_{v(1)}$

Constraint: $X\theta = y$

a) X requirements as a random matrix:

① X has real elements and size $n \times m$

Proof: No complex values

② Data in X is random with independent and identically distributed Gaussian data.

Proof: Independent data has no conditional probability:

$$P(X_i | X_j) != P(X_i)$$

Identically distributed data cumulates in an exact fashion:

$$F(X_i) = F(X_j) \ \forall X \in I$$ where $F(X)$ = cumulative distribution function.

③ The matrix is orthonormal with one's on the diagonals when expectedly squared.

Proof:

$$\mathbb{E}[X^T X] = \delta_{ij}$$

b) Restricted Isometry Property:

Books definition: $\delta_{pw}(X) \leq \delta_s(X) \leq s\,\delta_{pw}(X)$

$$(1-\gamma)\left\|\frac{X_s^T X_s}{n} - 1\right\|_2^2 \leq \left\|\frac{X_s^T X_s}{n} - 1\right\|_2^2 \leq (1+\gamma)\left\|\frac{X_s^T X_s}{n} - 1\right\|_2^2$$

Other book definitions: $(1-\gamma)||C||_2^2 \leq ||AC||_2^2 \leq (1+\gamma)||C||_2^2$

where $C = \left\|\frac{X_s^T X_s}{n} - 1\right\|$

"Every set of columns are an orthonormal basis set -or- eigenvalues have rational spacing."

Pairwise Coherence:

Books definition: $\sigma_{PW}(X) = (1-\gamma) \left\| \frac{X_S^T X_S}{n} - I \right\|_2^2$

$$\leq \frac{1}{3s}$$

Other book definition: With a sparse matrix $\in \mathbb{R}^{m \times n}$, a sample size of $m \geq k \log n$ represents data with high probability.

With k measurements.

"An average data sample needs $\sim\frac{1}{3}$ of the original matrix for exact representation. Carnegie Melon argues $\sim\frac{1}{2}$ of the original matrix bests representation."

As $t \to \infty$,

$$\|\theta\|_{V(t)} = \min_j \sum_{i=1}^{d} \theta_i$$

$$= \min_j \sum_{i=1}^{d} X^T \theta_i X^{-1}$$

$$= \min_j \sum_{i=1}^{d} \frac{w_i}{w_j} \theta_j$$

$$= \min_j \frac{1}{w_j} \sum_{i=1}^{d} w_i \theta_j \qquad \text{where } w_j = \begin{cases} 1 & j \in S \\ t & \text{otherwise} \end{cases}$$

$$\lim_{t \to \infty} \|\theta\|_{V(t)} = \lim_{t \to \infty} \min_j \frac{1}{w_j} \sum_{i=1}^{d} w_i \theta_i$$

$$= \lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{d} w_i \theta_i$$

$$= 0$$

7.7. $X \in \mathbb{R}^{n\times d}$: i.i.d : $N(0,1)$

a.) $S \in \{1, 2, \ldots, d\}$: $(1-\gamma)\delta_{pw}(X) = (1-\gamma)\left\| \frac{X_S^T X_S}{n} - 1 \right\|$

$$\leq (1-\gamma)\sqrt{\frac{\log d}{n}}$$

Chapter 6 Equation #40
"logarithm of a Gaussian bound"

$$\leq \frac{1}{3S}$$

Pairwise coherence
Proposition 7.9

$$\leq \frac{1}{3}\sqrt{\frac{\log d}{n}}$$

Problem's advice of
$n = s^2 \log d$

$$(1-\gamma)\sqrt{\frac{\log d}{n}} \leq \frac{1}{3}\sqrt{\frac{\log d}{n}}$$

$$(1-\gamma) \leq \frac{1}{3}$$

$$\frac{2}{3} \leq \gamma$$

High-probability
a random eigenvalue
is above $2/3$

b) $(1-\delta_{2S})\delta_{pw}(X) \leq \left\| \sum_{i=1}^{N} \frac{\langle X\theta, X\theta \rangle}{n} \right\|_2^2$

$$\delta_{pw}(X) \leq \frac{1}{(1-\delta_{2S})}\left\| \sum_{i=1}^{N} \frac{\langle X\theta, X\theta \rangle}{n} \right\|_2^2$$

$$\leq \frac{1}{(1-\delta_{2S})}\left\| \sum_{i=1}^{N} \theta\left[\frac{X_S^T X_S}{n} - 1\right]\theta \right\|_2^2$$

$$\leq \frac{\delta_{2S}}{(1-\delta_{2S})}\sum_{i=1}^{N}\|\theta\|_2^2$$

where $\|\theta\|_2^2 = \|\delta_S(X)\|_2^2$

$$\leq \frac{\delta_{2S}}{(1-\delta_{2S})}\sum_{i=1}^{N}\|\delta_S(X)\|_2^2$$

$$\leq \frac{\delta_{2S}}{(1-\delta_{2S})}\left\{\|\delta_{pw}(X)\|_1 + \|\delta_S^c\|_1\right\}$$

$\delta_{2S} \leq 1/3$ when $\frac{\|\delta_S^c\|}{\|\delta_{pw}\|} = 1$

(Theorem 7.16) A random matrix $X \in \mathbb{R}^{n \times d}$, in which each row $X_i \in \mathbb{R}^d$ is independent and identically distributed from a $N(0, \Sigma)$ distribution. Then there are universal positive constants $c_1 < 1 < c_2$ such that:

$$\frac{\|X\theta\|_2^2}{n} \geq c_1 \|\sqrt{\Sigma}\,\theta\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\theta\|_1^2 \quad \text{for all } \theta \in \mathbb{R}^d$$

with probability $\frac{1 - e^{-n/32}}{1 - e^{-n/32}}$.

7.8

(Convex Hull)

$$\text{conv}(C) = \{X \in C \mid \sum_{i=1}^{K} \alpha_k X_k,\ \alpha_i \geq 0,\ \sum_{i=1}^{K} \alpha_i = 1\}$$

"The smallest set containing C"

(Support Function -or- Support Set)

$$\text{Supp}(f) = \{X \in X \mid f(x) \neq 0\}$$

"A set of points in X where f is non-zero."

7.9 $L_0(k) = B_2(1) \cap B_0(k) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1 \text{ and } \|\theta\|_0 \leq k\}$

$L_1(k) = B_2(1) \cap B_1(\sqrt{k}) = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2 \leq 1 \text{ and } \|\theta\|_1 \leq \sqrt{k}\}$

a) $L_0(k) \subset L_1(k)$

$$\overline{\text{conv}}(L_0(k)) \subset \overline{\text{conv}}(L_1(k))$$

since $\overline{\text{conv}}(L_1(k)) \subset L_1(k)$

$$\overline{\text{conv}}\, L_0(k) \subset L_1(k)$$

b) $B_2(1) \cap B_1(\sqrt{k}) \leq 2(B_2(1) \cap B_0(k))$

$$\|\theta\| \leq \sqrt{k} \leq 2\|\theta\|_1 \leq k$$

$$L_1(k) \subseteq 2 \circ \overline{\text{conv}}(L_0(k))$$

7.10

a From Definition 7.12: $\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2$ for all $\Delta \in C_\alpha(S)$ with (v, k)-parameters.

$$|\theta^T \Gamma \theta| \geq k\|\theta\|_\alpha^2$$

$$\geq \begin{cases} 12\sigma\|\theta\|_2^2 \\ 12\sigma \dfrac{\|\theta\|_1^2}{s} \end{cases}$$

Where (12σ, ℓ-norm $= 1, 2$) are the parameters.

b) l_2-norm: $\gamma_{min}(|\theta^T T \theta|) \leq |\theta^T T \theta|_2 \leq \gamma_{max}(|\theta^T T \theta|_2)$

$$(1-\gamma)\,12\cdot\sigma\|\theta\|_2^2 \leq 12\cdot\sigma\|\theta\|_2^2 \leq (1+\gamma)\,12\sigma\,\|\theta\|_2^2$$

l_1-norm: $\gamma_{min}(|\theta^T T \theta|_1) \leq |\theta^T T \theta|_1 \leq \gamma_{max}(|\theta^T T \theta|_1)$

$$(1-\gamma)\frac{12\sigma\|\theta\|_1^2}{s} \leq \frac{12\cdot\sigma\|\theta\|_1}{s} \leq (1+\gamma)\frac{12\cdot\sigma\|\theta\|_1^2}{s}$$

c) R.I.P. conditions for a high-dimensional matrix are:

1) A matrix value inside equals zero. This is a "noiseless" setting and restricted nullspace with subsequent properties.

2) The variance decays at a $\frac{k\log P}{n}$ rate where k = sparsity, p = data dimension, and n = number of values.

Matrices regularly violate this chapter's properties, such as:

A) Autoregressive processes where the current time depends on past values because a regular variance.

B) A dataset where the minimum and maximum variance depend on each other because rank's degeneracy.

C) Highly degenerate matrices with isolated phenomena.

7.11.

a) $\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2$ is from Restricted Eigenvalue theorem.

When k is $\frac{c_1}{2}\gamma_{min}(\Sigma)$, then

$\frac{1}{n}\|X\theta\|_2^2 \geq \frac{c_1}{2}\gamma_{min}(\Sigma)$ fits, but not for Equation 7.31.

The relationship, $|S| \geq \frac{c_1}{2c_2} \frac{\gamma_{min}(\Sigma)}{\rho^2(\Sigma)} (1+\alpha)^{-2} \frac{n}{\log d}$ models.

Equation 7.31 as a violation. Average variance has issues with degenerate processes, for example sinusoidal waves. In these cases, min variance minus maximum variance is more than the average.

b. $\{\Sigma^{(0)}\} = \{-4,-3,-2,-1,0,1,2,3,4\}$

$$\Sigma^2 = \frac{\sum_{i=1}^{a} (X_i - \bar{X})^2}{n}$$

$$= \frac{(-4-0)^2 + (-3-0)^2 + (-2-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2 + (2-0)^2 + (3-0)^2 + (4-0)^2}{9}$$

$$= 6.\overline{66}$$

Expected variance: $\frac{1}{n}\sum_{i=1}^{9} X_i \|\Sigma\|_2^2 = \frac{(-4-3-2-1+0+1+2+3+4)\sqrt{6.66}^2}{9}$

$$= 0$$

Through equation 7.31, a violation of an example:

$$\frac{1}{n}\sum_{i=1}^{9} X_i \|\Sigma\|_2^2 \geq c_1 \|\sqrt{\Sigma}\,\Sigma\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\Sigma\|_1^2$$

Where $c_1 = c_2 = 1$, d = data dimensions $= 1$, n = # measurements $= 9$.

$$0 \overset{!}{\geq} 1 \cdot \|\sqrt{6.66} \circ 6.66\|_2^2 - 6.66 \frac{\log(1)}{9} \|6.66\|_1^2$$

$$\overset{!}{\geq} 17.81$$

The model violates the property in Equation 7.31 because the regular values from -4 to 4.

(Theorem 7.19) - Lasso Oracle Inequality

A Lagrangian Lasso with regularization parameter $\lambda \geq 2\|\frac{X^T w}{n}\|$

$$\|\hat{\theta}-\theta^*\|_2^2 \leq \frac{144}{c_1^2} \frac{\lambda_n^2}{\bar{K}^2}|S| + \frac{16}{c_1}\frac{\lambda_n}{\bar{K}}\|\theta^*_{S^c}\|_1 + \frac{32c_2}{c_1}\frac{\rho^2(\Sigma_i)}{K}\frac{\log d}{n}\|\theta^*_S\|_1^2$$

for cardinality $|S| \leq \frac{c_1}{64c_2}\frac{\bar{K}}{\rho^2(\Sigma_i)}\frac{n}{\log d}$.

7.12
$$B_q(R_q) = \{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\theta_j|^q \leq R_q\}$$

- λ represents noise in $Y = Xt + \lambda$ and also with Lagrangian Lasso's.
- When the parameters enter Theorem 7.19:

$$\lambda = 0 \; ; \; c_1 = \frac{32\bar{K}}{R_q\left(\frac{\sigma^2 \log d}{n}\right)^{q/2}} \; ; \; c_2 = c_0 R_q \; ; \; \rho^2(\Sigma) = \sigma^2$$

$$\|\hat{\theta}-\theta^*\|_2^2 \leq \frac{144}{c_1^2}\frac{\lambda_n^2}{\bar{K}^2}|S| + \frac{16}{c_1}\frac{\lambda_n}{\bar{K}}\|\theta_{S^c}^*\|_1 + \frac{32c_2}{c_1}\frac{\rho^2(\Sigma)}{K}\frac{\log d}{n}\|\theta\|^2$$

$$\leq c_0 R_q \left(\frac{\sigma^2 \log d}{n}\right)^{1-q/2}$$

7.13: $y = X\theta^* + w$; $w \sim N(0, \sigma^2 I_{n\times n})$; $\theta \in \mathbb{R}^d$;

The book suggests a lambda, $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$.

This **lambda** value relates a Gaussian function with d-measurements and n-dimensions.

From equation 7.25a, $\|\hat{\theta}-\theta^*\| \leq \frac{3}{K}\sqrt{S}\,\lambda_n$

$$\leq \frac{3}{K}\sqrt{S}\left(4\sigma\sqrt{\frac{\log d}{n}}\right)$$

$$\leq \frac{6\sigma}{\gamma}\sqrt{\frac{\log d}{n}}$$

Where $K=2$ and $\gamma = \frac{1}{\sqrt{S}}$

7.14 $X \in \mathbb{R}^{n \times d}$ with rows as $N(0, \Sigma)$

The restricted eigenvalue equation describes a regression limit near a function's minimum.

$$\|\Sigma' \Delta\|_\infty \geq 2 \left\|\frac{X^T W}{n}\right\|_\infty \|\Delta\|_2$$

$$\geq \|\Delta\|_2$$

$$\geq \frac{\|\Delta\|_1}{2\sqrt{s}} \quad \text{where} \quad \frac{\|\Delta\|_1}{2\sqrt{s}} \leq \|\Delta\|_2$$

$$\geq \frac{\gamma}{2} \|\Delta\|_1 \quad \text{and} \quad \gamma = \frac{1}{\sqrt{s}}$$

7.15. Equation 7.19: $\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\}$ such that $\|\theta\|_1 \leq R$ where $R = \|\theta^*\|$

a) $\mathbb{P}\left[\frac{Z}{C\sigma} \geq c_1 \sqrt{\frac{s \log(ed/s)}{n}} + \delta \right]$

$$\mathbb{P}\left[\frac{Z}{C\sigma} - c_1 \sqrt{\frac{s \log(ed/s)}{n}} \geq \delta \right] = \mathbb{P}\left[\frac{\lambda Z}{C\sigma} - \lambda c_1 \sqrt{\frac{s \log(ed/s)}{n}} \geq \lambda \delta \right]$$

$$\approx \mathbb{P}\left[e^{\frac{1 + \lambda \mathbb{E}[z]}{C\sigma} + \frac{\lambda^2 \mathbb{E}[z^2]^2}{2C\sigma} + \dots - \frac{\lambda \mathbb{E}[z]}{C\sigma}} \geq e^{\lambda \delta} \right]$$

$$\leq \mathbb{P}\left[e^{\frac{\lambda^2 \mathbb{E}[z^2]^2}{2C\sigma}} \geq e^{\lambda \delta} \right]$$

$$\operatorname*{argmin}_{\lambda} \left\{ \frac{\lambda^2 \mathbb{E}[z]^2}{2C\sigma} - \lambda \delta \right\} = 0$$

$$\lambda^* = \frac{C\sigma\delta}{\mathbb{E}[z^2]^2}$$

$$\mathbb{P}\left[e^{\frac{\lambda^2 \mathbb{E}[z]^2}{2C\sigma}} \geq e^{\lambda \delta} \right] = \mathbb{P}\left[e^{\left(\frac{C\sigma\delta}{\mathbb{E}[z]^2}\right)^2 \frac{\mathbb{E}[z^2]^2}{2C\sigma}} \geq e^{\left(\frac{C\sigma\delta}{\mathbb{E}[z^2]}\right)\delta} \right]$$

$$\approx e^{\frac{-C\sigma\delta^2}{2\mathbb{E}[z^2]^2}}$$

$$\approx e^{\frac{-C\delta^2}{2\sigma}}$$

$$\approx C_2 e^{-C_3 n \delta^2}$$

where $c_1 = \sqrt{\frac{n}{s \log(ed/s)}} \mathbb{E}[z]$; $C_2 = 1$; $C_3 = \frac{2n\sigma}{C}$

$$\mathbb{P}\left[\frac{Z}{C\sigma} \geq C_1\sqrt{\frac{s\log(ed/s)}{n}} + \delta\right] \overset{?}{\leq} C_2 e^{-C_3 \cdot n \cdot \delta^2}$$

b) $\|\hat{\theta} - \theta^*\| \asymp \frac{\sigma}{K}\sqrt{\frac{s\log(ed/s)}{n}}$

$K\|\hat{\theta} - \theta^*\| = \sigma\sqrt{\frac{s\log(ed/s)}{n}}$

The restricted eigenvalue condition: $\frac{\|X\Delta\|_2^2}{n} \geq K\|\Delta\|_2^2$

$$\geq \sigma\sqrt{\frac{s\log(ed/s)}{n}}$$

Left-handed Probability:

Method #1: $\mathbb{P}\left[\frac{\|\Delta\|}{C\sigma} \geq C_1\sqrt{\frac{s\log(ed/s)}{n}} + \delta\right] \leq \underbrace{C_2 e^{-C_3 n \delta^2}}_{\text{"Right-side of a Gaussian"}}$

$$\underbrace{1 - C_2 e^{-C_3 n \delta^2}}_{\text{"Left-side of a Gaussian}} \leq \mathbb{P}\left[\frac{\|\Delta\|}{C\sigma} \geq C_1\sqrt{\frac{s\log(ed/s)}{n}} + \delta\right] \leq C_2 e^{-C_3 n \delta^2}$$

Method #2: $\mathbb{P}[X > Y] = P[Z > 0]$ where $Z = X - Y$

$= 1 - \text{erf}\left(\frac{-\mu}{\sqrt{2}\sigma}\right)$ for a Gaussian $\sim N(x|\mu,\sigma^2)$

$= 1 - \phi\left(\frac{-\mu}{\sigma}\right)$

$$\mathbb{P}\left[\frac{\|\Delta\|}{C\sigma} \geq C_1\sqrt{\frac{s\log(ed/s)}{n}} + \delta\right] = \mathbb{P}\left[\frac{\|\Delta\|}{C\sigma} - \mu > \delta\right]$$

$$= 1 - \frac{1}{\sqrt{2\pi}}\int_0^{\frac{\|\Delta\|}{C\sigma} - \mu} e^{-\delta^2/2}\, d\delta$$

$$= 1 - \frac{1}{\sqrt{\pi}}\int_0^{\frac{\|\Delta\|}{C\sigma} - \mu} e^{-z^2}\, dz$$ where $z = \delta/\sqrt{2}$

$dz = d\delta/\sqrt{2}$

$$= 1 - \frac{1}{\sqrt{\pi}}\sqrt{\pi}\cdot\text{erf}(z)$$

$$= 1 - \text{erf}\left(\frac{\delta}{\sqrt{2}}\right)$$

$$= 1 - e^{-\delta^2/2}\left[-\frac{\sqrt{\frac{2}{\pi}}}{x} + \frac{\sqrt{\frac{2}{\pi}}}{x^3} + \cdots\right]$$

$= 1 - c_2' e^{-c_2 n \delta^2}$

where $c_1' = \sqrt{\frac{n}{s\log(ed/s)}}\,\mu\,;\ c_2' = -\frac{\sqrt{\frac{2}{\pi}}}{x} + \frac{\sqrt{\frac{2}{\pi}}}{x^3} + \cdots$

$c_3' = \frac{1}{2n}$

(Types of Lassos)

a) Lagrangian Lasso with regularization parameters lower bounded at $\lambda_n \geq 2\|\frac{X^T W}{n}\|_\infty$: $\|\hat{\theta} - \theta\|_2 \leq \frac{3}{k}\sqrt{s}\,\lambda_n$

b) Any solution with a constrained lasso, $\min\{\frac{1}{2n}\|y - X\theta\|_2^2\}$, satisfies the bound : $\|\hat{\theta} - \theta\|_2 \leq \frac{4}{k}\sqrt{s}\,\|\frac{X^T W}{n}\|$

b) Any solution of the relaxed basis pursuit program with $b^2 \geq \frac{\|W\|_2^2}{2n}$ satisfies : $\|\hat{\theta} - \theta\|_2 \leq \frac{4}{k}\sqrt{s}\,\|\frac{X^T W}{n}\| + \frac{2}{\sqrt{k}}\sqrt{b^2 - \frac{\|W\|^2}{2n}}$

In addition, all three satisfy the 1-norm :

$$\|\hat{\theta} - \theta\|_1 \leq 4\sqrt{s}\,\|\hat{\theta} - \theta^*\|_2$$

7.16

a) $\hat{\theta} \in \arg\min\{\frac{1}{2n}\|y - X\theta\|_2^2 + \lambda_n \|\theta\|_{v(1)}\}$

$\theta^* = \frac{y}{x} + 2\max_{j=1\ldots d} W$

$\|\hat{\theta} - \theta^*\| \leq 3\|\hat{\theta} - \theta^*\|_{s, v(1)}$ where the "cone's set" belongs to $C_3(S:V) = \{\Delta \in \mathbb{R}^d \mid \|\Delta_{S^c}\|_{v(1)} \leq 3\|\Delta_S\|_{v(1)}\}$

and $\Delta = \|\hat{\theta} - \theta^*\|$

b) A Lagrangian Lasso : $\|\hat{\theta} - \theta\|_2 \leq \frac{1}{2n}\|X\hat{\Delta}\|_2^2$

$\leq \frac{W^T X \hat{\Delta}}{n} + \lambda\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}$

$$\|\theta^*\|_1 - \|\hat\theta\|_1 = \|\theta_S^*\|_1 - \|\theta_S^* + \hat\Delta_S\|_1 - \|\hat\Delta_{S^c}\|_1$$

$$\|\hat\theta - \theta^*\| \le \frac{1}{n}\|X\hat\Delta\|_2^2$$

$$\le 2\frac{W^T X\hat\Delta}{n} + 2\cdot\lambda_n\{\|\theta_S^*\|_1 - \|\theta_S^* + \hat\Delta_S\|_1 - \|\hat\Delta_{S^c}\|_1\}$$

$$\le 2\left\|\frac{X^T W}{n}\right\|_\infty \|\hat\Delta_1\|_1 + 2\lambda_n\{\|\hat\Delta_S\|_1 - \|\hat\Delta_S\|_1\}$$

$$\le 2\lambda\{3\|\hat\Delta_S\|_1 - \|\hat\Delta_{S^c}\|_1\}$$

$$k\|\hat\theta - \theta^*\| \le 6\lambda_n\sqrt{s}\,\|\hat\Delta\|_2$$

$$\|\hat\theta - \theta^*\| \le \frac{6}{k}\lambda_n\sqrt{\sum_{j\in S} v_j^2}$$

c) The "weighted" lasso is helpful with a non-convergent regression. This case is from a zero coefficient in the least square's parameter equation, a division by zero, or false estimator. The weight, and second term in $\min\{\frac{1}{2}\|y - X\theta\|^2 + \lambda\sum w_i|\theta|\}$ avoids large magnitude coeffigeants through penalization. A real-world case for the weight is geometric regression of a square, circle, triangle or ellipse. The ellipse equation, $Ax^2 + By^2 + Cxy + Dx + Ey + F = 0$ needs this additional constraint from personal experience.

7.17

a) $\hat{\Theta} \in \underset{\Theta \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{\sqrt{n}} \|y - X\Theta\|_2 + \gamma \|\Theta\|_1 \right\} = 0$

$$\gamma^* = \frac{2X}{\sqrt{n}} \|y - X\Theta\| \quad \text{where } y = X\Theta + w$$

$$= 2 \left\| \frac{X^T w}{n} \right\| \quad \text{which is also a solution for Lagrangian Lasso's.}$$

b) $\hat{\Theta} \in \operatorname{argmin} \left\{ \frac{1}{\sqrt{n}} \sqrt{(y - X\Theta)^2} + \gamma \|\Theta\|_1 \right\} = 0$

$$= \frac{2}{2} \frac{1}{\sqrt{n}} \frac{(-X)(y - X\Theta)}{\sqrt{(y - X\Theta)^2}} + \gamma \frac{d\|\Theta\|_1}{d\Theta}$$

$$= \frac{\frac{1}{n} X^T (X\hat{\Theta} - y)}{\frac{1}{\sqrt{n}} \|y - X\hat{\Theta}\|_2} + \gamma \hat{z}$$

c) $\hat{\Delta} = \hat{\Theta} - \Theta^*$

$$\frac{1}{n} \|X\hat{\Delta}\|_2^2 \leq \frac{1}{n} \|X(\hat{\Theta} - \Theta^*)\|_2^2$$

$$\leq \left\langle \Delta, \frac{X^T w}{n} \right\rangle + \gamma \frac{\|y - X\Theta\|_2}{\sqrt{n}} \|\Delta\|_1$$

$$\leq \left\langle \Delta, \frac{X^T w}{n} \right\rangle + \gamma \frac{\|y - X\Theta\|_2}{\sqrt{n}} \left\{ \|\Delta_{s^c}\|_1 + \|\hat{\Delta}_{s^c}\| \right\}$$

d) $\gamma_n \geq 2 \cdot \frac{\|X^T w\|_\infty}{\sqrt{n} \|w\|_2}$

$$\|\hat{\Theta} - \Theta^*\| = \|\Delta_{s^c}\|_{v(1)}$$

$$\leq 3\|\Delta_{s^c}\|_{v(1)} \quad \text{where } C_3(s, v) = \{ \Delta \in \mathbb{R}^d \mid \|\Delta_{s^c}\|_{v(1)} \leq 3\|\Delta_{s^c}\|_{v(1)} \}$$

e) From the R.E. condition, $k\|\hat{\Theta} - \Theta^*\|_2 \leq \frac{1}{n} \|X^T \Delta\|_2^2$

$$\|\hat{\Theta} - \Theta^*\|_2 \leq \frac{1}{kn} \|X^T \Delta\|_2^2$$

$$\leq \gamma \sqrt{s} \frac{\|w\|_2}{\sqrt{n}} \quad \text{where } w = X^T \Delta$$

7.18 $X \in \mathbb{R}^{n \times d}$ satisfies $\delta_{pw} \leq \frac{1}{2s}$ and "restricted" eigenvalues.

Mutual coherence: $$\max_{j \in S^c} \|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \leq \alpha$$

The proof is weak because false negatives across real datasets. A false negative example has variance values not in the variance lasso, such as a "big" and ever-expansive dataset. The lassos require parameters.

From the "restricted" eigenvalue condition,

$$\frac{1}{n}\|X X_j\|_2^2 + \alpha \|X X_j\|_1 \leq \alpha \|X_S^T X_S\|_1$$

$$\frac{1}{n}\|X^T X^* - X^T X_S^0\|_2^2 \leq \alpha \|X_S^T X_S\|_1 - \alpha \|X_S^T X_j^*\|_1$$

$$\leq \alpha \|X_S^T X_S\| - \alpha \|X_S^T X_S\|$$

$$\frac{1}{n}\|X^T X^* - X^T X_S^0\|_2^2 \leq \alpha (\|X_S^T X_S - X_S^T X_S\|)$$

$$\frac{1}{n} \frac{\|X^T X^* - X^T X_S^0\|_2^2}{\|X_S^T X_S - X_S^T X_j\|} \leq \alpha$$

$$\max_{j \in S} \|(X_S^T X_S)^{-1} X_S^T X_S\|_1 \leq \alpha$$

7.19 $$\max_{j \in S} \|\Sigma_i (\Sigma_i)^{-1}\|_1 \leq \alpha < 1$$

a) The "restricted" eigenvalue condition: $\frac{1}{n}\|X\Delta\| \geq \alpha \|\Delta\|_1^2$

$$\frac{1}{n}\|X_j X_S - W_j Z\| \leq \alpha \|X_S^T X_S\|_1$$

$$\frac{1}{n}\|X_j X_S\| \leq \alpha + \|W_j^T X_S (X_S^T X_S)^{-1} Z\|$$

b) $\mathbb{P}\left[\frac{1}{n}\max_j |X_j^T X_S (X_S^T X_S)^{-1} z| \geq \delta\right]$

$= \mathbb{P}\left[\max_j |X_j^T X_S (X_S^T X_S)^{-1} z| \geq n\delta\right]$

Chapter two descibes a standard sub-Gaussian tail:

$\mathbb{P}[\max |z_j| \geq t] \leq 2(d-s)\, e^{\frac{-nt^2}{2c^2\sigma^2}}$

where $Z_j = X_j [I_n - X_S(X_S^T X_S)^{-1} X_S^T]\left(\frac{w}{n}\right)$

$\mathbb{P}\left[\max_j |z_j| \geq \frac{\sigma}{\sqrt{C_{min}}}\left\{\sqrt{\frac{2\log s}{n}} + \delta\right\}\right] \leq 2 \cdot e^{\frac{-n\delta^2}{2}}$

The above chapter represents a simplification for the bounds in a sub-Gaussian.

Wainright suggests $n \geq \frac{16 s \log(d-s)}{(1-\alpha)\sqrt{C_{min}}}$, a similar setup.

$\mathbb{P}\left[\max_j |X_j^T X_S (X_S^T X_S)^{-1} z| \geq n\delta\right] \leq 2 \cdot 2 e^{\frac{-n\delta^2}{2}}$

$\leq \alpha'$

and the $2 \cdot e^{\frac{-n\delta^2}{2}} = \frac{1 + e^{\frac{-n\delta^2}{2}}}{2}$

$\leq \frac{1}{2}(1+\alpha)$

17.20

a) $\hat{\Theta} = \min_\Theta \left\{\frac{1}{2n}\|y - X\Theta\|_2^2\right\}$

$\Theta^* = \frac{y}{X}$

b) $\gamma_\ell \|\Delta\|_2^2 \leq \frac{\|X\Delta\|_2^2}{n} \leq \gamma_n \|\Delta\|_2^2$ where $\|X\|^2 = N^2(0, \sigma^2)$

$= \sigma^2 s \log(ed/s)$

From Problem 5.7

$$\|\Delta\|_2^2 = \|\hat{\theta} - \theta^*\|_2^2 \lesssim \frac{\gamma_\mu}{\gamma_\ell} \sigma^2 s \log(ed/s)$$

The relationship from problem 5.7 is extensive because parameters, approximations, and zero-mean.