Theorem 11.1b

Derives error bounds on Minimum of the objective function:

$$J(\beta) = \frac{1}{2N} \| \vec{y} - X\vec{\beta} \|^2 + \lambda_N \|\beta\|_1$$

λ_N: Regularisation Parameter

$\beta \in \mathbb{R}^p$: unknown sparse signal.

$\underset{N \times 1}{\vec{y}} = X\vec{\beta} + \vec{w}$ Measurement Vector

$X_{N \times p}$ (p ↓ covariates)

$\underset{N \times 1}{\vec{w}}$: $\mathcal{N}(0, \sigma^2)$

$X \in \mathbb{R}^{N \times p}$: Sensing Matrix with Normalised Columns

minimizer of $J(\vec{x})$: known by LASSO in statistics.

Side-stuff

if $N > p$ high N Error ↓

if $N < p$ assume more structure (here sparseness)

a) Define Restricted Eigen Value Condition

Least Squares Objective Function $f_N(\beta) = \frac{1}{2N} \| y - X\beta \|^2$

→ $\frac{1}{N} \frac{v X^T X v}{\|v\|_2^2} \geq \gamma$ $\forall$ non-zero $v \in C$ (Constrain Set)

γ ↓ some const

b) Starting from 11.20 on page 309 explain why $G(\hat{v}) \leq G(0)$

$$G(v) := \frac{1}{2N} \| y - X(\beta^* + v) \|_2^2 + \lambda_N \| \beta^* + v \|_1$$

β^* original

$$G(0) = \frac{\| y - X\beta^* \|_2^2}{2N} + \lambda \|\beta^*\|_1$$

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

$$G(\hat{v}) = \frac{1}{2N} \| y - X\hat{\beta} \|_2^2 + \lambda_N \|\hat{\beta}\|_1$$

Since $\hat{\beta}$ is the minimiser of LASSO

we have that

$G(\hat{v}) \leq G(v)$

$G(\hat{v}) \leq G(0)$

c) Do the algebra to obtain 11.21

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{w^T X \hat{v}}{N} + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \}$$

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

$G(\hat{v}) \leq G(0)$

$$\therefore \frac{1}{2N} \|\vec{y} - X\beta^* - X\hat{v}\|_2^2 + \lambda_N \|\beta^* + \hat{v}\|_1 \leq \frac{1}{2N} \|\vec{y} - X\beta^*\|_2^2 + \lambda_N \|\beta^*\|_1$$

$$\frac{1}{2N} \|\vec{y} - X\beta^* - X\hat{v}\|_2^2 \leq \frac{1}{2N} \|\vec{y} - X\beta^*\|_2^2 + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \}$$

→ Last Eq^n on pg 308 of PDF states $y = X\beta^* + w$

$$\therefore \frac{1}{2N} \|w - X\hat{v}\|_2^2 \leq \frac{1}{2N} \|w\|_2^2 + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \}$$

$$\therefore \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{(w^T)(X\hat{v})}{N} + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \}$$

c) Holder's Inequality as taught in EE734 $\quad \|X\|_p = E[|X|^p]^{\frac{1}{p}}$

For $1 \leq p \leq \infty$ and $\frac{1}{p} + \frac{1}{q} = 1$

if Random Variables X and Y satisfy $0 < \|X\|_p, \|Y\|_q < \infty$

then $|E[XY]| \leq \|X\|_p \|Y\|_q$

$(w^T X)(\hat{v})$ where $\hat{v}_i$: i^{th} entry.

Let $\{B_1, B_2, \ldots, B_p\}$ partition Ω

let $Y = \sum \frac{\hat{v}_i \, I(w \in B_i)}{N}$ and $Z = \sum \frac{(w^T X_i) \, I(w \in B_i)}{N}$

$$\therefore E[YZ] = \frac{(w^T X)(\hat{v})}{N}$$

$$\|Z\|_\infty = \underbrace{\|w^T X\|_\infty}_{1}$$

$$\|Y\|_1 = \underbrace{\|\hat{v}\|_1}_{N}$$

$$\therefore E[YZ] \leq \|Y\|_1 \|Z\|_\infty$$

hence 11.22

I was earlier going to state "Trivially follows from Holder's" but then decided to elaborate ;)

C) Derive 11.23

$$\underbrace{\frac{\|X\hat{v}\|_2^2}{2N}}_{} \leq \frac{\lambda_N}{2}\left\{\|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1\right\} + \lambda_N\left\{\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1\right\} \leq \frac{3}{2}\sqrt{k}\,\lambda_N\|\hat{v}\|_2$$

Theorem 11.1 b) assumes $\lambda_N \geq \underbrace{\frac{2\|X^T w\|_\infty}{N}}_{}$ (A)

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

$E[f(X)] \geq f[E(X)]$

$f(x) = x^2$

$$\underbrace{\frac{\|\hat{v}_S\|_2^2}{k}}_{} \geq \left(\underbrace{\frac{\|\hat{v}_S\|_1}{k}}_{}\right)^2$$

$$\|\hat{v}\|_2 \geq \|\hat{v}_S\|_2 \geq \underbrace{\frac{\|\hat{v}_S\|_1}{\sqrt{k}}}_{} \quad (B)$$

hence 11.23 follows trivially from 11.22, A and B

(f) Lemma 11.1

Suppose that $\lambda_N \geq 2 \left\| \frac{X^T w}{N} \right\|_\infty > 0$ then error $\hat{v} := \hat{\beta} - \beta^*$ associated with any LASSO solution $\hat{\beta}$ belongs to the cone set $C(S;3)$

and 11.23 is derived completes proof for error bound of 11.14b

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{3}{\gamma} \sqrt{\frac{k}{N}} \sqrt{N} \lambda_N$$

$$\lambda_N \geq 2 \left\| \frac{X^T w}{N} \right\|_\infty$$

Using Lemma 11.1,

Using $v = \hat{v}$ in (11.10) we get

$$\frac{1}{N} \frac{\|X^T \hat{v}\|_2^2}{\|\hat{v}\|_2^2} \geq \gamma \overset{(11.23)}{\Rightarrow} \frac{\gamma \|\hat{v}\|_2^2}{2} \leq \frac{3}{2} \lambda_N \sqrt{k} \|\hat{v}\|_2$$

$$\Rightarrow \|\hat{v}\|_2 \leq \frac{3 \lambda_N \sqrt{k}}{\gamma}$$

g) (i) In Lemma 11.1 $\lambda_N \geq \frac{2 \|X^T w\|_\infty}{N}$

is used to show that the error $\hat{v} := \hat{\beta} - \beta^*$ associated with Lasso sol^n $\hat{\beta}$ belongs to Cone set $C(S;3)$

(ii) In RHS of 11.22

$\frac{\|X^T w\|_\infty}{N} \|\hat{v}\|_1 + \lambda_N \{ \|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1 \}$

↓

In upper bounding $\lambda_N \geq \frac{2 \|X^T w\|_\infty}{N}$ is used.

h) Importance of cone constraint

Consider the un-numbered equation written between 11.19 and 11.20

$$\|\hat{v}\|_1 = \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1 \leq 2\|\hat{v}_S\|_1 \leq 2\sqrt{k}\,\|\hat{v}\|_2$$

Cone constraint is used here $[\hat{v} \in C(S,1)]$
And this bound is used in final upper bounding of 11.23.

S: support set for original sparse β^*

Cone constraint also appears in eqn 11.12 $\|\hat{v}_{S^c}\| \leq 3\|\hat{v}_S\|_1$

Cone constraint can be though of as proxy for strong convexity for vectors in some constraint set

(i) The particular theorem for special case of Gaussian Noise vector gives bound

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{c\sigma}{\gamma}\sqrt{\frac{k \log p}{N}} \qquad \text{w.p} \geq 1 - 2e^{-\frac{1}{2}(c-2)\log p} \quad \text{(with probability)} \qquad (11.15)$$

So we get tighter bounds here (albeit with high probability) as compared to theorem 3.

Advantage of Theorem 3
Bounds for theorem 3 $\|\theta^* - \theta\|_2 \leq \frac{c}{\sqrt{s}}\|\theta - \theta_S\|_1 + c\varepsilon$

Theorem 3 does not require hard sparsity (even if non-zero entries are more than some 's' entries we would get nice bounds as long as those entries are small (relative term), whereas theorem in book requires hard sparsity.

j) Common Thread between bounds on 'Dantzig Selector' and LASSO

LASSO $\|\hat{x}-x\|_2 \le C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_2\varepsilon$

A: RIP δ_{2k} with $\delta_{2k} < \sqrt{2}-1$

$y = Ax + e$ $\quad \|e\|_2 \le \varepsilon$

$B(y) = \{z : \|Az - y\|_2 \le \varepsilon\}$

here k: sparsity.

'Dantzig Selector'

$\|\hat{x}-x\|_2 \le C_0 \frac{\sigma_k(x)_1}{\sqrt{k}} + C_3\sqrt{k}\lambda$

A RIP δ_{2k} $(< \sqrt{2}-1)$

$y = Ax + e$

$\|A^T e\|_\infty \le \lambda$

$B(y) = \{z : \|A^T(Az - y)\|_\infty \le \lambda\}$

A different recovery algorithm based on Dantzig selector in the case when $\|A^T e\|_\infty$ is small.

Both bounds are minimax estimators

Both bounds have roughly the same flow of derivation and nearly same looking results. If the case of Gaussian Noise is considered and if m and n are fixed, and we consider the effect of varying k. then the bound based on Dantzig selector (adaptive to k) improves whereas bound based on LASSO does not.