Q1. (a) Comparison of bounds in equns. 11.15 and 11.16 to Theorem 03 done in class.

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

$y \in \mathbb{R}^N$; $X \in \mathbb{R}^{N \times P}$; $\beta \in \mathbb{R}^P$

Here, $y = X\beta + \omega$ where ω is a vector of zero-mean i.i.d gaussian noise ($\omega \in \mathbb{R}^N$)

Theorem 03 (class):

$\min \|\beta\|_1$ such that $\|y - X\beta\|_2^2 \leq \varepsilon$ has soln β^* (where $y = A\beta + \omega$), then

$$\|\beta^* - \beta\|_2 \leq \frac{C}{\sqrt{S}} \|\beta - \beta_S\|_1 + C_1 \varepsilon$$

where β_S is created by taking top 'S' largest magnitude elements of β and setting the rest to zero (location of elements not changed). and $N \geq C S \log(P/S)$

Equ^n 11.15 $\quad \|\hat{\beta} - \beta^*\|_2 \leq C\, \sigma/\gamma \sqrt{\frac{\tau k \log P}{N}}$

Eqn 11.16 $\quad \|\hat{\beta} - \beta^*\|_2 \leq C\sigma \sqrt{\frac{k \log (eP/k)}{N}}$

k is similar to S in theorem 3

σ is std dev of noise ω vector

I. For Theorem 3:-

1. Signal sparsity:- Error decreases with increasing S (signal sparsity) due to $1/\sqrt{S}$ factor
2. No. of measurements:- Error decreases as it increase
3. Signal dimension:- As P increases while N is constant, error increases
4. noise std. deviation:- error $\propto$ ε (noise)

II. Equn 11.15

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{c\sigma}{\gamma}\sqrt{\frac{k\log p^{c}}{N}}$$

$k \to$ sparsity level

$\sigma \to$ noise std. dev

$N \to$ no. of measurements

$p \to$ signal dimension

a) N :- $\hat{v}$ (error) decreases by $\frac{1}{\sqrt{N}}$ factor

b) Signal sparsity (k) :- $\hat{v} \propto \sqrt{k}$

c) σ :- $\hat{v} \propto \sigma$

d) p :- $\hat{v} \propto \sqrt{\log p}$

III. Equn 11.16

$$\|\hat{\beta} - \beta^*\| \leq c\sigma\sqrt{\frac{k\log(ep/k)}{N}}$$

This is same as equn 11.15 except that the error $\hat{v}$ has refined logarithmic dependence on signal sparsity (K) i.e., $\sqrt{k\log\left(\frac{ep}{k}\right)}$

Which bound is more intuitive :-

Theorem 03 provides an explicit trade-off between sparsity and reconstruction error via the L1-norm. While equations 11.15 and 11.16 provide probabilistic guarantees and explicitly depend on measurements N and signal dimension p.

Hence, if we want statistical guarantee in higher dimensions, equations 11.15 and 11.16 are more useful and intuitive to understand.

Q1.
(b) Define Restricted Eigen Value Condition.

First, let us define strong convexity property for a function $f: \mathbb{R}^p \to \mathbb{R}$ [an extension of the general convex property]

$f: \mathbb{R}^p \to \mathbb{R}$ (a differentiable fn) is said to be "strongly convex" with parameter $\gamma > 0$ at $\theta \in \mathbb{R}^p$ if

$$f(\theta') - f(\theta) \geq \nabla f(\theta)^T (\theta' - \theta) + \frac{\gamma}{2} \|\theta' - \theta\|_2^2$$

$\forall \; \theta' \in \mathbb{R}^p$.

[$\nabla f(\theta)^T$ is the gradient of f at θ]

If f is twice continuously differentiable, strong convexity can also be formulated as:-

if $\text{Min}\left(\text{Eig}\left[\nabla^2 f(\beta)\right]\right) \geq \gamma > 0 \;\; \forall \; \beta$ in the neighbourhood of β^*

then, function f is strongly convex with parameter γ around $\beta^* \in \mathbb{R}^p$.

Now, take f to be the least-squares objective function

$$f_N(\beta) = \frac{1}{2N} \|y - X\beta\|_2^2 .$$

$$\Rightarrow \nabla^2 f(\beta) = \frac{X^T X}{N} \rightarrow$$

$$\left[\begin{array}{l} f(\beta) = \frac{1}{2N}(y - X\beta)^T (y - X\beta) \\ \nabla f(\beta) = \frac{1}{2N}(-2X^T y + 2X^T X\beta) \\ \nabla f(\beta) = \frac{X^T}{N}(X\beta - y) \\ \text{Differentiate again:} \\ \nabla^2 f(\beta) = \left(\frac{XX^T}{N}\right) \end{array}\right]$$

Hence for $f_N(\beta)$ to be strong convex, all its eigenvalues must be strongly bounded away from 0 ($\because \gamma > 0$)

But $X \in \mathbb{R}^{N \times p} \Rightarrow X^T X \in \mathbb{R}^{p \times p}$

$\text{rank}(X^T X) \leq \min(\text{rank}(X), \text{rank}(X^T)) = \min(p, N)$

if $p > N$, $\text{rank}(X^T X) \leq N \Rightarrow X^T X$ is rank-deficient

Hence, at least one eigen value is 0

$\Rightarrow$ $f_N(\beta)$ is not strongly convex

Therefore, define a weaker "restricted strong convexity". condition:

f satisfies restricted strong convexity at β^* wrt $C \in \mathbb{R}^p$ if $\exists\ \gamma > 0$ s.t.:

$$\frac{v^T \nabla^2 f(\beta) v}{\|v\|_2^2} \geq \gamma$$

for all non-zero $v \in C$. and $\forall\ \beta \in \mathbb{R}^p$ in the neighbourhood of β^*.

For the least squares objective function,

where $\nabla^2 f_N(\beta) = \frac{X^T X}{N}$, the condition becomes

$$\boxed{\frac{1}{N} \frac{v X^T X v}{\|v\|_2^2} \geq \gamma > 0} \quad \forall\ v \in C \text{ and } v \neq 0$$

Q1. (c) Equation 11.20

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

(Lagrangian Lasso)

<u>To show</u>:- $G(\hat{v}) \leq G(0)$

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

$\hookrightarrow$ unknown regression vector (ground reality)

optimal ~~value~~ ~~predicted value~~
minimizing equ^n 11.20 ($\hat{v}$)

$$G(0) = \frac{1}{2N} \| y - X\beta^* \|_2^2 + \lambda_N \| \beta^* \|_1$$

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

we know that $\hat{\beta}$ is the minimizer of our lasso for objective function $J(\beta) = \frac{1}{2N} \| y - X\beta \|_2^2 + \lambda_N \| \beta \|_1$

$\Rightarrow$ $G(\hat{v}) \leq G(v)$ ($\hat{v}$ is minimizer of $G(v)$)

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

Q1. (d)

Eqn :- (11.21)

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

we have our starting point

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

From Q1 (c), we have $G(\hat{\nu}) \leq G(0)$

$$\Rightarrow \frac{1}{2N}\|y - X(\beta^* + \hat{\nu})\|_2^2 + \lambda_N \|\beta^* + \hat{\nu}\|_1$$
$$\leq \frac{1}{2N}\|y - X\beta^*\|_2^2 + \lambda_N \|\beta^*\|_1$$

substitute $w = y - X\beta^*$ ($\because y = X\beta^* + w$)

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

$$\Rightarrow \frac{1}{2N}(w - X\hat{\nu})^T(w - X\hat{\nu}) - \frac{1}{2N} w^T w \leq \lambda_N (\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1)$$

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

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

(proved)

Q1. (e) Theorem 11.2

Consider the Lagrangian lasso with a regularization parameter $\lambda_N \geq \frac{2}{N} \|X^T w\|_\infty$

(a) If $\|\beta^*\|_1 \leq R_1$, then any optimal solution $\hat{\beta}$ satisfies

$$\frac{\|X(\hat{\beta} - \beta^*)\|_2^2}{N} \leq 12 R_1 \lambda_N$$

(b) If β^* is supported on a subset S, and the design matrix X satisfies the γ-RE condition over $C(S;3)$, then any optimal solution $\hat{\beta}$ satisfies

$$\frac{\|X(\hat{\beta} - \beta^*)\|_2^2}{N} \leq \frac{9}{\gamma} |S| \lambda_N^2$$

taking $\lambda_N = c\sigma \sqrt{\frac{\log P}{N}}$, the two bounds take the form

$$\frac{\|X(\hat{\beta} - \beta^*)\|_2^2}{N} \leq c_1 \sigma R_1 \sqrt{\frac{\log P}{N}} \text{ and,}$$

$$\frac{\|X(\hat{\beta} - \beta^*)\|_2^2}{N} \leq c_2 \frac{\sigma^2}{\gamma} \frac{|S| \log P}{N}$$

Equation 11.21 (modified) states

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

$$w^T X \hat{\upsilon} \leq \|X^T w\|_\infty \|\hat{\upsilon}\|_1$$

(using Holder's inequality for l_1 and l_∞ norm)

$\Rightarrow$

$$\frac{\|X^T w\|_\infty}{N} \|\hat{\upsilon}\|_1 + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{\upsilon}\|_1 \} \geq \frac{\|X\hat{\upsilon}\|_2^2}{2N} \geq 0$$

$\Rightarrow$

$$0 \leq \frac{\|X^T w\|_\infty}{N} \|\hat{\upsilon}\|_1 + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{\upsilon}\|_1 \} \quad \text{— (1)}$$

From triangle inequality

$$\|\beta^* + \hat{\upsilon}\|_1 \geq \|\hat{\upsilon}\|_1 - \|\beta^*\|_1$$

$$\Rightarrow -\|\beta^* + \hat{\upsilon}\|_1 \leq \|\beta^*\|_1 - \|\hat{\upsilon}\|_1$$

Add $\|\beta^*\|_1$ on both sides

$$\|\beta^*\|_1 - \|\beta^* + \hat{\upsilon}\|_1 \leq 2\|\beta^*\|_1 - \|\hat{\upsilon}\|_1 \quad \text{— (2)}$$

Substitute (2) into (1) to get

$$0 \leq \frac{\|X^T w\|_\infty}{N} \|\hat{\upsilon}\|_1 + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{\upsilon}\|_1 \}$$

$$\leq \left\{ \frac{\|X^T w\|_\infty}{N} - \lambda_N \right\} \|\hat{\upsilon}\|_1 + 2\lambda_N \|\beta^*\|_1 \quad \text{— (3)}$$

Using the assumption of Theorem (2)

$$\frac{\|X^T w\|_\infty}{N} \leq \frac{\lambda_N}{2}$$

putting this in (2), we get

$$\left\{\frac{\|X^T \omega\|_\infty}{N} - \lambda_N\right\}\|\hat{\nu}\|_1 + 2\lambda_N \|\beta^*\|_1$$

$$\leq \frac{\lambda_N}{2}\left\{-\|\hat{\nu}\|_1 + 4\|\beta^*\|_1\right\} \quad \text{--- (4)}$$

$\therefore$ from (1), (3) and (4), we get

$$0 \leq \frac{\|X^T\omega\|_\infty}{N}\|\hat{\nu}\|_1 + \lambda_N\left\{\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1\right\}$$

$$\leq \frac{\|X^T \omega\|_\infty}{N} - \lambda_N \|\hat{\nu}\|_1 + 2\lambda_N \|\beta^*\|_1$$

$$\leq \frac{1}{2}\lambda_N\left\{-\|\hat{\nu}\|_1 + 4\|\beta^*\|_1\right\}$$

$$\therefore \quad \boxed{\|\hat{\nu}\|_1 \leq 4\|\beta^*\|_1 \leq 4R_1} \quad \text{--- (5)}$$

(first first and last inequality in prev exp.)

Again, look at modified inequality 11.21 and eqn (1)

$$\frac{\|X\hat{\nu}\|_2^2}{2N} \leq \frac{\|X^T\omega\|_\infty}{N}\|\hat{\nu}\|_1 + \lambda_N\left\{\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1\right\}$$

From triangle inequality,

$$\|\beta^*\|_1 - \|\beta^* + \hat{\nu}\|_1 \leq \|\hat{\nu}\|_1$$

$$\Rightarrow \frac{\|X\hat{\nu}\|_2^2}{2N} \leq \left\{\frac{\|X^T\omega\|_\infty}{N} + \lambda_N\right\}\|\hat{\nu}\|_1 \leq \frac{3\lambda_N}{2} \times 4R_1 = \underline{\underline{6R_1\lambda_N}}$$

(using (5))

$$\boxed{\frac{\|X(\hat{\beta}-\beta^*)\|}{N} \leq 12 R_1 \lambda_N}$$ (Th 11.2, claim a proved)

$\therefore$

Proof of bound (11.25(b)) $\rightarrow$ Th 11.2 claim b

Inequality 11.23 states

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

(11.23)

It is applicable here since $\frac{\lambda_N}{2} \geq \frac{\|X^T w\|_\infty}{N}$ is given

$$\Rightarrow \frac{\|X\hat{\upsilon}\|_2^2}{2N} \leq \frac{3}{2}\sqrt{k}\lambda_N\|\hat{\upsilon}\|_2 \quad \text{— (b)}$$

Lemma 11.1 states that error vector $\hat{\upsilon}$ lies in cone $C(S;3)$

$\Rightarrow$ Applying γ-RE condition gives

$$\|\hat{\upsilon}\|_2^2 \leq \frac{1}{N\gamma}\|X\hat{\upsilon}\|_2^2 \Rightarrow \|\hat{\upsilon}\|_2 \leq \frac{1}{\sqrt{N\gamma}}\|X\hat{\upsilon}\|_2$$

~~From proof of 11.25 (a),~~

$$\cancel{\frac{\|X\hat{\upsilon}\|_2^2}{2N} \leq \left\{\frac{\|X^T w\|_\infty}{N} + \lambda_N\right\}\|\hat{\upsilon}_1\|}$$

$$\cancel{\leq \frac{3\lambda_N}{2}\|\hat{\upsilon}\|_1} \quad \text{(given bound on } \lambda_N)$$

$$\cancel{\Rightarrow \frac{\|X\hat{\upsilon}\|_2^2}{N} \leq 3\lambda_N\|\hat{\upsilon}\|_1}$$

~~we know that for l_1 and l_0 norm~~

$$\cancel{\|\hat{\upsilon}_0\|_1 = \|\hat{\upsilon}_S\|_1 + \|\hat{\upsilon}_{S^c}\|_1 \leq \|\hat{\upsilon}_S\|_1 \leq 2\sqrt{k}\|\hat{\upsilon}\|_2}$$

$$\Rightarrow \quad \cancel{\frac{\|X\hat{v}\|_2^2}{N}} \leq \cancel{6\lambda_N\sqrt{K}\,\|\hat{v}\|_2}$$

plugging this in above (6) inequality

$$\frac{\|X\hat{v}\|_2^{\cancel{2}}}{\cancel{2}N} \leq \frac{3}{\cancel{2}} \sqrt{K}\,\lambda_N \cdot \frac{1}{\sqrt{N\gamma}} \cancel{\|X\hat{v}\|_2}$$

$$\Rightarrow \quad \frac{\|X\hat{v}\|}{\sqrt{N}} \leq 3\sqrt{K}\,\lambda_N \frac{1}{\sqrt{\gamma}}$$

Squaring both sides, we get

$$\frac{\|X\hat{v}\|^2}{N} \leq \frac{9}{\gamma} K \lambda_N^2$$

But $\quad K = |S|$

$$\therefore \quad \boxed{\frac{\|X\hat{v}\|^2}{N} \leq 9/\gamma \; |S|\,\lambda_N^2} \quad \text{(proved)}$$