Q 2)

(a) Def^n: $X \in R^{N \times P}$, $\beta \in R^P$, $y \in R^N$

We say the model matrix X satisfies the restricted eigenvalue condition with respect to subset $C \in R^P$ if there is a constant $\gamma > 0$ such that

$$\frac{v^T X^T X v}{N \|v\|_2^2} \geq \gamma \text{ for all non zero } v \in C.$$

Explanation according to text:
We want our loss function $f_N(\beta)$ to be strictly convex where $f_N(\beta) = \frac{\|y - X\beta\|_2^2}{2N}$

Then the hessian $\nabla^2 f_N(\beta) = \left(\frac{X^T X}{N}\right)$ should have its eigenvalues uniformly bounded away from 0. But rank of $X^T X$ is $\min(N, p)$ which makes it rank deficient and hence not strongly convex, whenever $N < p$.

So, we relax the condition by allowing the function to be strictly convex in subset $C \in R^P$. By restrictly strong convexity at β^* w.r.t. C if there is constant $\gamma > 0$ s.t.

$$\frac{v^T \nabla^2 f(\beta) v}{\|v\|_2^2} \geq \gamma \text{ for all non zero } v \in C$$

& for all $\beta \in R^P$ in neighbourhood of β^*.
In specific case of linear regression, where $\nabla^2 f(\beta) = X^T X / N$, we get the restricted eigenvalue condition.

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

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

Now, $\hat{v} = \hat{\beta} - \beta^*$.

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

$= \frac{1}{2N}\|y - X\hat{\beta}\|_2^2 + \lambda_N \|\hat{\beta}\|_1$ —— ②.

We know $\hat{\beta}$ is minimizer of $J(\beta)$

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

$\therefore J(\hat{\beta}) \leq J(\beta^*)$

From ① & ②,

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

(c) From previous part $G(\hat{v}) \leq G(0)$

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

Since $y = X\beta^* + w$.,

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

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

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

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

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

(since $w^T X\hat{v}$ is scalar & scalar^T = scalar)

Rearranging we get,

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

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

(d) β^* is a s-sparse vector.

$S = S(\beta^*)$ is the set of indices which represent the non-zero support of β^*. $|S(\beta^*)| = s$

$x_S \in R^{|S|}$ is the subvector indexed by the elements of S and x_{S^c} is defined in analogous manner.

Now, since $\beta^*_{S^c} = 0$, we have $\|\beta^*\|_1 = \|\beta^*_S\|_1$

Using this, we have

$$\|\beta^* + \hat{v}\|_1 = \|\beta^*_S + \hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1$$

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

← {by triangle inequalities}

Thus,

$$\|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \leq \|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1$$

Using above inequality on the previous part we get,

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{w^T X \hat{v}}{N} + \lambda_N (\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1)$$

← (1)

Holder's inequality: (from wikipedia)

Let (S, Σ, μ) be measure space and $p, q \in [1, \infty)$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$. Then,

for all measurable real or complex valued functions f and g,

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

Using holder's inequality for $p = \infty$, $q = 1$, & $f = w^T X$, $g = \hat{v}$,

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

Since $w^T X \hat{v}$ is a scalar,

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

Also $\|w^T X\|_\infty = \|X w^T\|_\infty$

Using this in eqn ①,

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{\|X^T w\|_\infty \|\hat{v}\|_1}{N} + \lambda_n (\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1)$$

(e) By assumption, $\frac{\|X^T w\|_\infty}{N} \leq \frac{\lambda_N}{2}$

$$\therefore \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{\lambda_N}{2}\|\hat{v}\|_1 + \lambda_N\left(\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1\right)$$

$$\leq \frac{\lambda_N}{2}\left(\|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1\right) + \lambda_N\left(\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1\right)$$

$$\leq \frac{3\lambda_N}{2}\|\hat{v}_S\|_1 - \frac{\lambda_N}{2}\|\hat{v}_{S^c}\|_1$$

$$\leq \frac{3}{2}\lambda_N\|\hat{v}_S\|_1 \quad \text{o} \left\{\frac{\lambda_N}{2}\|\hat{v}_{S^c}\|_1 \text{ is a positive term}\right\}$$

$$\therefore \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{3}{2}\lambda_N\|\hat{v}_S\|_1$$

~~Now, since as we know $\|\hat{v}_S\|_1 = \|\hat{v}\|$~~

Now, we use the fact that if v is a k sparse vector then by a Cauchy Schwartz inequality $\|v\|_1 = \sum_{i=1}^{k}|v_i| \cdot 1 \leq (\sum|v_i|^2)^{1/2}(\sum_i 1^2)^{1/2}$

$$\leq \sqrt{k}\,\|v\|_2$$

Therefore,

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{3}{2}\lambda_N\|\hat{v}_S\|_1 \leq \frac{3}{2}\lambda_N\sqrt{k}\,\|\hat{v}_S\|_2$$

$$\leq \frac{3}{2}\lambda_N\sqrt{k}\,\|\hat{v}\|_2 \quad \{\|\hat{v}_S\|_2 \leq \|\hat{v}\|_2\}$$

$$\therefore \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{3}{2}\lambda_N\sqrt{k}\,\|\hat{v}\|_2$$

(f) From previous part,

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{3}{2}\lambda_N\sqrt{K}\|\hat{v}\|_2 \quad \text{—— (1)}$$

Lemma 11.1 states that if $\lambda_N \geq 2\left\|\frac{X^T w}{N}\right\|_\infty > 0$,

then the error $\hat{v} = \hat{\beta} - \beta^*$ associated with any lasso solution $\hat{\beta}$ belongs to cone set $C(S;3)$

Now, we apply the restricted eigenvalue condition to $\hat{v}$ belonging to the cone set

$$\therefore \frac{1}{N}\|X\hat{v}\|_2^2 \geq \gamma\|\hat{v}\|_2^2$$

$$\frac{1}{2N}\|X\hat{v}\|_2^2 \geq \frac{\gamma}{2}\|\hat{v}\|_2^2 \quad \text{—— (2)}$$

From (1) & (2),

$$\frac{\gamma}{2}\|\hat{v}\|_2^2 \leq \frac{3}{2}\lambda_N\sqrt{K}\|\hat{v}\|_2$$

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

$$\frac{\gamma}{2}\|\hat{v}\|_2 \leq \frac{3}{2}\lambda_N\sqrt{K}$$

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

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

Hence proved eqn 11.14b (i.e. Theorem 11.1 (b))

(g) The inequality $\lambda_N \geq 2\frac{\|X^T w\|_\infty}{N}$

shows up in part (e); which gives us the cone set $C(S;3)$ for $\hat{v}$

It is explained as follows,

from part (e), we have,

$$0 \leq \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{3\lambda_N}{2}\|\hat{v}_S\|_1 - \frac{\lambda}{2}\|\hat{v}_{S^c}\|_1$$

$$\therefore \frac{\lambda}{2}\|\hat{v}_{S^c}\|_1 \leq \frac{3\lambda_N}{2}\|\hat{v}_S\|_1$$

$$\therefore \|\hat{v}_{S^c}\|_1 \leq 3\|\hat{v}_S\|_1$$

This is the cone constraint for $C(S;3)$

(h) As explained in part (a), whenever $N<p$, the matrix X^TX is rank deficient and hence not strongly convex.

We relax the condition, and impose the strong convexity condition on a subset $C \in \mathbb{R}^p$ for all possible vectors $\hat{v}$.

For the function $y = X\beta^* + w$, we get restricted strong convexity with respect to the subset C if there is a constant $\gamma > 0$ s.t.

$$\frac{\hat{v}^T X^T X v}{N\|\hat{v}\|_2^2} \geq \gamma \quad \text{for all } v \in C, v \neq 0$$

Now, it turns out that if we solve the regularised version of lasso, with a "suitable" choice of λ_N, the lasso error satisfies the constraint $\|\hat{v}_{S^c}\|_1 \leq 3\|\hat{v}_S\|_1$, which is the cone constraint.

(i) Example 11.1 for classical linear Gaussian model says,

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{c\sigma}{\gamma}\sqrt{\frac{\tau k \log n}{m}}$$

with probability $1 - 2e^{-\frac{1}{2}(\tau-2)\log p}$ and $w \in R^m$ is Gaussian with iid $N(0,\sigma^2)$ distribution.

Theorem 3 $A = \phi\psi$ of size $m \times n$ has RIP of order $2s$ where $\delta_{2s} < 0.41$. Let solution of following be θ^* (measurement vector $y = \phi\psi\theta + \eta$):

$$\min \|\theta\|_1 \text{ s.t. } \|y - \phi\psi\theta\|_2^2 \leq \epsilon$$

Then we have,

$$\|\theta^* - \theta\|_2 \leq \frac{C_0}{\sqrt{s}}\|\theta - \theta_s\|_1 + C_1\epsilon$$

Advantages of theorem 3 over ex 11.1 bound are:

(i) The theorem 3 works for any sparsity level. ~~given it follows RIP with $\delta_{2s} \leq 0.41$~~, But 11.1 requires the sparsity level to be known.

ii) Theorem 3 also give bounds for compressible signals (which have small values in some orthonormal basis) 11.1 considers only sparse signals.

iii) Theorem 3 is more general than 11.1. 11.1 requires proper choice of λ_N according to theorem 11.1 b & restricted eigenvalue condition.

(iv) 11.1 is only for gaussian noise which theorem 3 is for any general noise.

Advantages of 11.1 over theorem 3:

(i) 11.1 requires restricted eigenvalue condition required by theorem 11.1(b), instead of RIP condition. RIP can be costly.

(ii) The ε value needs to be fixed properly according to noise level for theorem 3. This can be avoided by 11.1.

(iii) Even with knowledge of support set, since model has k free parameters no method can achieve squared l_2-error that decays more quickly than $\frac{k}{N}$.

Apart from logarithmic factor, the lasso rate matches the best possible one could achieve even if subset $S(\beta^*)$ were known a-priori. The rate given by 11.1 ea. is known to be minimax optimal meaning it cannot be improved upon by any estimator.

So the bound by 11.1 gives good gaurantees.

(j) Lasso: Constraint: $\|X^T w\|_\infty \leq \frac{N\lambda_N}{2}$

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

Dantzig: Constraint: $\|A^T e\|_\infty \leq \lambda$

Bound: ~~$\|\hat{\beta} - \beta^*\|_2$~~

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

Both λ in Dantzig & λ_N in Lasso are similar. Both give upper bound on the largest value of $\|X^T w\|_\infty$ i.e. interaction between measurement matrix & noise vector. The term $\sqrt{k}\lambda$ appear in both.. In fact for case when $\sigma_k(x)_1 = 0$ for Dantzig, we get $\|\hat{x} - x\|_2 \leq C_3\sqrt{k}\lambda$ similar to lasso $\|\hat{\beta} - \beta^*\|_2 \leq \frac{3}{\gamma}\sqrt{k}\lambda_N$.