CS754 HW3 Q1

Q1a) Equation 11.15 states,

$$\|\hat{\beta} - \beta^*\|_2 \le \frac{C\sigma}{\gamma}\sqrt{\frac{\tau k \log p}{N}}$$

Equation 11.16 states,

$$\|\hat{\beta} - \beta^*\|_2 \le c\sigma\sqrt{\frac{k \log(ep/k)}{N}}$$

Theorem 3 in class state,

$$\|\theta^* - \theta\|_2 \le \frac{C_0\|\theta - \theta_S\|_1}{\sqrt{S}} + C_1\varepsilon.$$

θ^* corresponds to $\hat{\beta}$, θ corresponds to θ, S corresponds to k (sparsity)

$\varepsilon \propto \sigma\sqrt{N}$, because the noise vector has N elements, each i.i.d Gaussians with variance σ, and mean zero, and ε is proportional to the l_2 norm of the vector w i.e noise vector.

$$\mathbb{E}[\|w\|_2^2] = \mathbb{E}\left[\sum_{i=1}^{N} w_i^2\right] = N\sigma^2$$

$\therefore$ $\varepsilon \propto \sigma\sqrt{N}$. Note that this considers Gaussian noise.

→ Number of measurements

Eq^n 11.15 & 11.16 are inversely proportional to $\frac{1}{\sqrt{N}}$.

In theorem 3, the error bound second term varies proportional to $\sqrt{N}$

→ Signal Sparsity
~~Theorem~~ 11.15 and 11.16 are proportional to $\sqrt{k}$, in ~~Theorem~~ 11.16, there is also a $\log(p/k)$ term, which gives a tighter bound.
In Theorem 3, the 1st term of the error bound is inversely proportional to $\sqrt{S}$, which is the sparsity.

→ Noise standard deviation.
Eqn 11.15 & 11.16 are directly proportional to σ.
In Theorem 3, the 2nd term with epsilon is directly proportional to σ, since $\varepsilon \propto \sigma$.

→ ~~One~~ Signal dimension
Eqn 11.15 & 11.16 are directly proportional to $\log(p)$, where p is the signal dimension.
In Theorem 3, there is no direct relation with signal dimension, but, it is used to get a lower bound for no. of measurements needed for Compressed sensing.

$$N \geq C \log(m/\delta) \|\theta\|_0 \mu^2(\Psi, \Phi)$$

→ Intuitive Result:
~~Theorem~~ Eqn 11.15 & 11.16 are more intuitive because of the proportionality. Acc to eqn 11.15 & 11.16, as no of measurements increases, error decreases which seems better than Theorem 3, which suggests otherwise. Similarly, lower $k(S)$, intuitively should give lesser error, but it gives higher error according to Theorem 3. Also, the influence of signal dimension is indirect.

Q1 c) Equation 11.20 states

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

~~where~~ Now, Let us consider $J(\beta)$

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

This is minimised by $\beta = \hat{\beta}$.
Note that $G(\nu)$ is nothing but $J(\beta)$, but with $\nu = \beta - \beta^*$.
Thus, if $J(\beta)$ is minimised by $\beta = \hat{\beta}$, $G(\nu)$ is minimised by $\hat{\nu} = \hat{\beta} - \beta^*$.

Thus, $G(\nu)$ is minimised by $\hat{\nu} = \hat{\beta} - \beta^*$

$\therefore$ $G(\hat{\nu})$ is minimum value for $G(\nu)$

$\therefore$ $G(\hat{\nu}) \leq G(\nu) \; \forall \; \nu$

$\therefore$ $G(\hat{\nu}) \leq G(0)$.

d) Equation 11.20 states

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

Using $G(\hat{\nu}) \leq G(0)$, we get,

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

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

Also, $w = y - X\beta^*$.

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

Now, $\|w - X\hat{\nu}\|_2^2 = \|w\|_2^2 + \|X\hat{\nu}\|_2^2 - 2w^T X\hat{\nu}$

Substituting,

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

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

This is equation 11.21. Hence, proved.

e) $\frac{1}{N} w^T X\hat{\nu} = \frac{1}{N}\langle X^T w, \hat{\nu}\rangle$

This is dot product of $X^T w$ and $\hat{\nu}$.
If we take the largest term of $X^T w$ & multiply that with all terms of $\hat{\nu}$, that will be larger than the dot product.
Also, sum of all terms of $\hat{\nu}$ is less than or equal to the L1 norm of $\hat{\nu}$.

$$\therefore \langle X^T w, \hat{\nu}\rangle \leq \|X^T w\|_\infty \sum_{i=1}^{p} \hat{\nu}_i \leq \|X^T w\|_\infty \|\hat{\nu}\|_1$$

$$\therefore \frac{1}{N} w^T X\hat{\nu} \leq \frac{1}{N}\|X^T w\|_\infty \|\hat{\nu}\|_1$$

Using 11.21, LHS ≥ 0

$$\therefore 0 \leq \frac{\|X\hat{\nu}\|_2^2}{2N} \leq \frac{w^T X\hat{\nu}}{N} + \lambda_N(\|\beta^*\|_1 - \|\beta^*+\hat{\nu}\|_1)$$

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

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

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

$$\therefore -\|\beta^*+\hat{\nu}\|_1 \leq \|\beta^*\|_1 - \|\hat{\nu}\|_1 \quad \text{— (1)}$$

$$\leq \left(\frac{\|X^T w\|_\infty}{N} - \lambda_N\right)\|\hat{\nu}\|_1 + 2\lambda_N\|\beta^*\|_1 \quad (\text{Using (1)})$$

Now, $\frac{1}{N}\|X^T w\|_\infty \le \frac{\lambda_N}{2}$.

$$\therefore \le -\frac{\lambda_N}{2}\|\hat{v}\|_1 + 2\lambda_N\|\beta^*\|_1$$

$$= \frac{\lambda_N}{2}\left(-\|\hat{v}\|_1 + 4\|\beta^*\|_1\right)$$

$$\therefore 0 \le \frac{\lambda_N}{2}\left(-\|\hat{v}\|_1 + 4\|\beta^*\|_1\right)$$

$$\therefore \|\hat{v}\|_1 \le 4\|\beta^*\|_1 \le 4R_1$$

Eq 11.22 states

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

Since $\|\hat{v}\|_1 = \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1$,

$\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1 \le \|\hat{v}\|_1$.

$$\therefore \frac{\|X\hat{v}\|_2^2}{2N} \le \frac{\|X^T w\|_\infty\|\hat{v}\|_1}{N} + \lambda_N\|\hat{v}\|_1$$

$$= \left(\frac{\|X^T w\|_\infty}{N} + \lambda_N\right)\|\hat{v}\|_1$$

$$\le \left(\frac{\lambda_N}{2} + \lambda_N\right) 4R_1$$

$$= 6\lambda_N R_1$$

Hence, 11.25 a is proved.

From eqn 11.23,

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

Lemma 11.1 states

$$\frac{\gamma}{2}\|\hat{\nu}\|_2^2 \leq \frac{3}{2}\lambda_N \sqrt{k}\,\|\hat{\nu}\|_2$$

$$\therefore \|\hat{\nu}\|_2 \leq \frac{3\lambda_N\sqrt{k}}{\gamma}$$

$$\therefore \frac{\|X\hat{\nu}\|_2^2}{N} \leq \frac{9k\lambda_N^2}{\gamma} = \frac{9|S|\lambda_N^2}{\gamma}$$

$$\hat{\nu} = (\hat{\beta} - \beta^*)$$

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

b) Let $X^{N\times p}$ be the model (design) matrix, and let S be a set of indices representing the support of the parameter vector β^*.

The restricted eigenvalue condition is defined over a set of the form

$$C(S;\alpha) := \{v \in \mathbb{R}^p : \|v_{S^c}\|_1 \leq \alpha \|v_S\|_1\}$$

for some $\alpha \geq 1$.

v_S denotes the vector v with elements from subset S.

The restricted eigenvalue condition for the model matrix requires that

$$\frac{1}{N}\frac{v^T X^T X v}{\|v\|_2^2} \geq \gamma \quad \forall \text{ non zero } v \in C$$

γ here, is a parameter.