Q-3)

a)

It deals with the notion of Strong Convexity.

Now equation, $f_n(\beta) = \frac{1}{2N} \| y - X\beta \|_2^2$ is always Convex. But not Strongly Convex when $N < P$. (Statement 1)

Defination of Strong Convexity: Given a differentiable function $f: R^P \to R$, we say that it is strongly convex with parameter $\gamma > 0$ at $\theta \in R^P$ if the inequality

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

hold for all $\theta' \in R^P$.

In particular the function f is strongly convex with parameter γ around $\beta^* \in R^n$ if and only if the min eigen value of the Hession matrix $\nabla^2 f(\beta)$ is atleast γ for all vector β in the neighbourhood of β^*.

From Statement 1

We get $\nabla^2 f(\beta) = X^T X / N$ for all $\beta \in R^P$. Thus, the least-square loss is strongly convex if and only if the eigen values of the $p \times p$ positive semidefinite matrix $X^T X$ are uniformly bounded away from zero.

Here $X^T X$ has Rand $\{N, P\}$ & hence is always Rank deficient & not Strongly Convex.

So we Relax notion of strong Convexity.

It is only necessary to impose a type of strong convexity condition for some subset $C \subset \mathbb{R}^p$ of possible perturbation vectors $v \in \mathbb{R}^p$. We say that a function f satisfies <u>Restricted Strong Convexity</u> at β^* with respect to C if there is a constant $\gamma > 0$ such that

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

and for all $\beta \in \mathbb{R}^p$ in a neighbourhood of β^*

For $f_n \rightarrow f_n(\beta) = \frac{1}{2N} \|y - X\beta\|_2^2$ (Linear Regression), this notion is equivalent to lower bounding the Restricted eigenvalues of the model matrix, requiring

$$\frac{\frac{1}{N} v X^T X v}{\|v\|_2^2} \geq \gamma \text{ for all nonzero } v \in C$$

b> To explain $G(\hat{v}) \leq G(0)$

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

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

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

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

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

Now, $G(\hat{v}) \leq G(0)$, since

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

Now Above inequality can be verified from the fact

$$\|y - X\hat{\beta}\|_2^2 \leq \|y - X\beta^*\|_2^2$$

$$\|\hat{\beta}\|_1 \leq \|\beta^*\|_1$$

c)

<u>Given</u> : $G(v) = \frac{1}{2N}\|y - X(\beta^* + v)\|_2^2 + \lambda_N \|\beta^* + v\|_1$ (11·20)

Putting $y = X\beta^* + w$ in $G(\hat{v}) \leq G(0)$

$$\Rightarrow G(\hat{v}) \leq G(0) : \frac{1}{2N}\|X\beta^* + w - X\beta^* - X\hat{v}\|_2^2 + \lambda_N \|\beta^* + \hat{v}\|_1 \leqslant$$

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

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

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

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

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

$$\Rightarrow \frac{X\hat{v}}{2N} \leqslant \frac{w^T X\hat{v}}{N} + \lambda_N \{\|\beta^*\|_1 - \|\beta^* - \hat{v}\|_1\} \quad (11\cdot21)$$

d) Given: $\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{w^T X\hat{v}}{N} + \lambda_N \{ \|\beta^*\|_1 - \|\beta^* + \hat{v}\|_1 \}$ (11.21)

Now, $\|\beta^*\|_1 = \|\beta^*_s + \beta^*_{sc}\|_1$

$\|\beta^*\|_1 = \|\beta^*_s\|_1$ $\because$ $\beta^*_{sc} = 0$, and

$\|\beta^* + \hat{v}\|_1 = \|\beta^*_s + \hat{v}_s + \hat{v}_{sc}\|_1$

$= \|\beta^*_s + \hat{v}_s\|_1 + \|\hat{v}_{sc}\|_1$ {Disjoint Joint}

$= \|\beta^*_s - (-1)\hat{v}_s\|_1 + \|\hat{v}_{sc}\|_1$

$\geq \|\beta^*_s\|_1 - \|\hat{v}_s\|_1 + \|\hat{v}_{sc}\|_1$ $\{\|a-b\|_1 \geq \|a\|_1 - \|b\|_1\}$

Substibuting these relation into inequality (11.21) yields

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

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{w^T X\hat{v}}{N} + \lambda_N \{ \|\hat{v}_s\|_1 - \|\hat{v}_{sc}\|_1 \} \quad \text{—— } 1$$

from Holder's inequality :→

$\|f^T g\|_1 \leq \|f\|_p \|g\|_q$, where $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$

Here, $\|w^T X\hat{v}\|_1 = \|X^T w\|_\infty \|\hat{v}\|_1$

Applying holder's in 1

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{w^T X\hat{v}}{N} + \lambda_N \{ \|\hat{v}_s\|_1 - \|\hat{v}_{sc}\|_1 \} \leq \quad \text{—— (11.22)}$$

$$\frac{\|X^T w\|_\infty \|\hat{v}\|_1}{N} + \lambda_N \{ \|\hat{v}_s\|_1 - \|\hat{v}_{sc}\|_1 \}$$

e) We know that,

$$\|\hat{v}_S\|_1 \leq \sqrt{k}\,\|\hat{v}_S\|_2 \leq \sqrt{k}\,\|\hat{v}\|_2$$

We will use this inequality 1

e)

Since $\frac{1}{N}\|X^T w\|_\infty \leq \frac{\lambda_N}{2}$ by assumption, eq. 11.22 becomes

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

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

$\because \|v\|_1 = \|\hat{v}_S\|_1 + \|\hat{v}_{S^c}\|_1$ {Disjoint set}

$$\Rightarrow \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{3\lambda_N}{2}\|\hat{v}_S\|_1 - \frac{1}{2}\lambda_N\|\hat{v}_{S^c}\|_1 \leq \sqrt{k}\times\frac{3}{2}\lambda_N\|\hat{v}\|_2 - \frac{1}{2}\lambda_N\|\hat{v}_{S^c}\|_1$$

$\because \|\hat{v}_S\|_1 \leq \sqrt{k}\|\hat{v}_S\|_2 \leq \sqrt{k}\,\|\hat{v}\|_2$

Hence,

$$\Rightarrow \frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{\lambda_N}{2}\{\|\hat{v}_S\|_1 + \hat{v}_{S^c}\|_1\} + \lambda_N\{\|\hat{v}_S\|_1 - \|\hat{v}_{S^c}\|_1\} \leq \frac{3}{2}\sqrt{k}\,\lambda_N\|\hat{v}\|_2 \tag{11.23}$$

f) Lemma 11.1 allows us to apply γ-RE condition (11.10) to $\hat{v}$,

i.e. $\dfrac{\frac{1}{N} v X^T X v}{\|v\|_2^2} \geq \gamma$ OR $\frac{1}{N} \|X\hat{v}\|_2^2 \geq \gamma \|\hat{v}\|_2^2$.

Combining this with with inequality 11.23 gives the lower bound

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

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

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

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

Scanned by CamScanner

9 Given inequality : $\lambda_N \geq \frac{2\|X w\|_\infty}{N}$

Using this in inequality (11.22) yields

$$\frac{\|X\hat{v}\|_2^2}{2N} \leq \frac{\lambda_N}{\alpha}\|\hat{v}\|_1 + \lambda_N\{\|\hat{v}_s\|_1 - \|\hat{v}_{sc}\|_1\} \leq \frac{3}{\alpha}\sqrt{k}\,\lambda_N\|\hat{v}\|_2 \quad (11.23)$$

This also implies,

$$\Rightarrow \quad 0 \leq \frac{\lambda_N}{\alpha}\|\hat{v}\|_1 + \lambda_N\{\|\hat{v}_s\|_1 - \|\hat{v}_{sc}\|\}$$

$$\Rightarrow \quad 0 \leq \lambda_N\{\|\hat{v}_s\|_1 + \|\hat{v}_{sc}\|_1\} + 2\lambda_N\{\|\hat{v}_s\|_1 - \|\hat{v}_{sc}\|\}$$

$$\Rightarrow \quad 0 \leq 3\lambda_N\|\hat{v}_s\|_1 - \lambda_N\|\hat{v}_{sc}\|_1$$

$$\Rightarrow \quad \|\hat{v}_{sc}\|_1 \leq 3\|\hat{v}_s\|_1 \quad \{\text{Proves Lemma 11.1}\}$$

Hence the inequality $\lambda_N \geq \frac{2\|X w\|_\infty}{N}$ used to prove lemma 11.1

h>

From Defination of Restricted eigenvalues

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

Now what constraint set C are relevant

for appropriate choices of the l_1-ball radius - or equivalently, of the regularization parameter λ_N - it turns out that the lasso error satisfies a cone cone constraint of the form

$$\|\hat{v}_{S^c}\| \leq \alpha \|\hat{v}_S\|,$$

Now using this constrain we can successfully prove lemma 11.1 hence

Given a regularization parameter $\lambda_N \geq 2\|X^T w\|_\infty / N > 0$, any estimate $\hat{\beta}$ from the regularized lasso (11.3) satisfies the bound

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

Gurukul
Page No.:
Date: \ \

(i) * Adv. of Theorem 11.1(b) over Theorem 3

By Theorem 11.1(b) we know if $\lambda_N \geq 2\|X^T w\|_\infty / N > 0$, then $\hat{B}$ an estimate of lasso regularized lasso statisfies bound.

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

but with theorem 3, if A is RIP obeying matrix of $\delta_{2S} \leq 0.414$ then

$$\|\hat{B} - B^*\| \leq \frac{C_0}{\sqrt{\ }} \frac{4\sigma \sqrt{1+\delta_{2S}}}{1-\delta_{2S}(\sqrt{2}+1)}$$

where B^* is $|S| = K$ sparse. $\hat{B}$ is the solution for $\min \|B\|_1$ s.t. $\|Y - AB\|_2 \leq$

So, by Example 11.1 we can calculate the propable choice of λ_N i.e

$\lambda_N = 2\sigma \sqrt{\tau \frac{\log p}{N}}$ where $\tau > 2$ is a valid choice with high propability

$\therefore$ theorem 11.1(b) becomes,

$$\|\hat{B} - B^*\|_2 \leq \frac{c\sigma}{\gamma} \sqrt{\frac{\tau K \log p}{N}}$$

Page No.:
Date: \ \

(1) → which tells us that error is bounded $\frac{K}{N}$ i.e l_2-error decays more quickly with $\frac{K}{N}$, This error bound is much better than error bound given by Theorem 3.

(2) → Even if we knows the support of β^* then also the lasso rate (i.e 11.16 bound) is best possible w.r.t to the Theorm3

→ ~~As we decays by increase the N i.e length of the vector~~

(3) → As we increase the no. of measurment the error bounds will decrease drastically But this is not the case withe Theorem 3 because it doesnot depends on the number of measurment you take nor the sparsity (when β^* is K sparse)

→ ~~Therefore Also size of log p~~

(4) → Also the effect of log p is very less with increase in the size input vector on error bound.

Gurukul
Page No.:
Date: \ \

* Adv of Theorem 3 over Theorem 11.1b

→ Theorem 3 depends on δ_{2s} value matrix X, & Theorem 11.1b depends on γ (restricted eigenvalue)

restricted
i.e for strong convexity

$$\frac{1}{N} \frac{v X^T X v}{\|v\|_2} \geq \gamma$$ of all non zero $v \in C$

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

→ Error bounds of Theorem 3 & 11.1b depends δ_{2s} & γ respectively. Therefore finding γ w.r.t δ_{2s} is much more difficult than ~~tat~~ that of δ_{2s}.
Because if we know X satisfy RIP of $\delta_{2s} \leq 0.414$. the range of δ_{2s} is much more limited say 0 to 0.414 or 0 to 1.

→ But in case of γ it is much more complicated, creating C set is a difficult job, also γ can be from 0 to ∞
Therefore, γ finding is more difficult than δ_{2s}

→ Also Error bound of ~~theorem~~ theorem 3 when B^* is K sparse is depend of δ_{2s}.
so if we are able design X with least δ_{2s} then we will have least error bound.
But in case Theorem 11.1b it depends on too many ~~pa~~ ~~parameters~~ parameters.