# Quiz 5

**Quiz Rules:** closed-book, closed-notes; no computers, no calculators, no phones; if you are working on papers, have only a pen/pencil on your desk; if you are working with digital devices, keep them in airplane mode; work independently for at most **90 minutes**

*On my honor, I pledge that I followed the Quiz Rules and I have neither given nor received unpermitted aid on this quiz.*

**signature:** Chanwoo **print name:** Chanwoo Lee

**start time:** 9:40 **end time:** 10:40

Suppose that we observe

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}^* + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ has orthonormal columns and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Consider the regularized optimization for estimating $\boldsymbol{w}^*$

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \frac{\lambda}{p}\|\boldsymbol{w}\|_p^p$$

for $p = 1$ or 2 and $\lambda > 0$.

1. Show/explain how the optimization problem above can be transformed into an equivalent optimization of the form

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\widetilde{\boldsymbol{y}} - \boldsymbol{w}\|_2^2 + \frac{\lambda}{p}\|\boldsymbol{w}\|_p^p$$

and explain how $\widetilde{\boldsymbol{y}}$ is related to $\boldsymbol{y}$.

$$\arg\min_{w} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{p}\|w\|_p^p$$

$$= \arg\min_{w} \frac{1}{2}\left(\|y\|^2 - 2 y^t X w + \|w\|^2\right) + \frac{\lambda}{p}\|w\|_p^p$$

$X^tX = I$ & $y$ is constant, $X$ is const

$$= \arg\min_{w} \frac{1}{2}\|X^t y - w\|^2 + \frac{\lambda}{p}\|w\|_p^p$$

$$\therefore \tilde{y} = X^t y$$

(continued on next page)

2. Let $\lambda = 0$.

 (a) Give an expression for $\widehat{\boldsymbol{w}}$, the solution to optimization above, and

 (b) Consider the prediction error for a new observation of the form $y = \boldsymbol{x}^T \boldsymbol{w}^* + \epsilon$, for arbitrary, fixed $\boldsymbol{x} \in \mathbb{R}^n$ and independent noise $\epsilon \sim \mathcal{N}(0, 1)$. Show that the expected squared prediction error

$$\mathbb{E}[(y - \boldsymbol{x}^T \widehat{\boldsymbol{w}})^2] \;=\; \|\boldsymbol{x}\|_2^2 + 1\ .$$

(a) $\min_w \underbrace{\frac{1}{2}\|y - Xw\|_2^2}_{=L(w)}$

$\frac{\partial L(w)}{\partial w} = (X^t X) w - X^t y = 0$ $\qquad$ $\frac{\partial^2 L(w)}{\partial w^2} = I$ "p.d"

$\therefore\ \hat{w} = X^t y \quad (\because X^t X = I)$

(b) $E(\hat{w}) = X^t X w^* = w^*$ $\qquad$ $y \sim N(Xw^*, I)$

$Var(\hat{w}) = I$

$$\begin{aligned} E(y - x^t \hat{w})^2 &= E(x^t(w^* - \hat{w}) + \varepsilon)^2 \\ &= E(\varepsilon^2) + E(x^t w^* - x^t \hat{w})^2 \\ &= 1 + E((x^t\hat{w} - E(x^t\hat{w}))^2) \\ &= 1 + Var(x^t \hat{w}) \\ &= 1 + x^t I x = 1 + \|x\|_2^2 \end{aligned}$$

(continued on next page)

3. Let $\lambda > 0$ and $p = 2$.

(a) Give an expression for $\widehat{\boldsymbol{w}}$, the solution to optimization above in this case, and

(b) Consider the prediction error for a new observation of the form $y = \boldsymbol{x}^T \boldsymbol{w}^* + \epsilon$, as above. Derive an expression for the expected squared prediction error $\mathbb{E}[(y - \boldsymbol{x}^T \widehat{\boldsymbol{w}})^2]$, and show that it reduces to the expression in the MLE case when $\lambda = 0$.

(a) $\underset{w}{\operatorname{argmin}} \ \frac{1}{2} \| \underset{\sim}{X^t y} - w \|^2 + \frac{\lambda}{2} \|w\|_2^2$

$$\frac{\partial L(w)}{\partial w} = w - X^t y + \lambda w = 0 \qquad \frac{\partial^2 L(w)}{\partial w^2} = (1+\lambda) I \quad \text{"p.d"}$$

$$\therefore \quad \hat{w} = \frac{1}{1+\lambda} X^t y$$

(b) $E(\hat{\omega}) = \frac{1}{1+\lambda} X^t X w^* = \frac{1}{1+\lambda} w^*$

$Var(\hat{\omega}) = \left(\frac{1}{1+\lambda}\right)^2 I$

$E(x^t \hat{w}) = \frac{1}{1+\lambda} x^t w^*$

$var(x^t \hat{\omega}) = \left(\frac{1}{1+\lambda}\right)^2 x^t x = \left(\frac{1}{1+\lambda}\right)^2 \|x\|^2$

$$E((y - x^t \hat{\omega})^2) = E((x^t(w^* - \hat{w}) + \varepsilon)^2)$$

$$= E(\varepsilon^2) + E(x^t(w^* - \hat{\omega}))^2$$

$$= 1 + MSE_{x^t w^*}(x^t \hat{w})$$

$$= 1 + \left(\frac{1}{1+\lambda} x^t w^* - x^t w^*\right)^2 + \left(\frac{1}{1+\lambda}\right)^2 \|x\|_2^2$$

$$= 1 + \left(\frac{\lambda}{1+\lambda}\right)^2 (x^t w^*)^2 + \left(\frac{1}{1+\lambda}\right)^2 \|x\|_2^2$$

Put $\lambda = 0$ then

$\Rightarrow 1 + \|x\|_2^2$ it reduces to expression in MLE case

(continued on next page)

4. Let $\lambda > 0$ and $p = 1$.

(a) What value of $\lambda$ would you suggest for this case and why?

(b) Suppose that $\boldsymbol{w}^*$ has only $k < n$ nonzero elements. Consider the prediction error for a new observation of the form $y = \boldsymbol{x}^T\boldsymbol{w}^* + \epsilon$, as above. Show that the expected squared prediction error can be bounded as follows

$$\mathbb{E}[(y - \boldsymbol{x}^T\widehat{\boldsymbol{w}})^2] \;\leq\; (2\log n + 1)(k+1)\|\boldsymbol{x}\|^2 + 1\;.$$

Hint: You may use the soft-thresholding result from class which states that the solution to the optimization in (2), for appropriately chosen $\lambda$, produces an estimator $\widehat{\boldsymbol{w}}$ satisfying the bound

$$\mathbb{E}[\|\boldsymbol{w}^* - \widehat{\boldsymbol{w}}\|_2^2] \;\leq\; (2\log n + 1)\Big(1 + \sum_{i=1}^{n}\min(|w_i^*|^2, 1)\Big)\;.$$

$$\frac{1}{2}\|\tilde{y} - w\|_2^2 + \lambda\|w\|_1 \Rightarrow \hat{w}_i = \operatorname{sign}(\tilde{y}_i)\max(|\tilde{y}_i| - \lambda, 0)$$

(a) I suggest $\lambda = \sqrt{2\log(n)}$ because $\max|\varepsilon_i| \approx \sqrt{2\log(n)}$

then we have $P(|\tilde{y}_i| > \lambda \mid w_i = 0) \approx 0$

$P(|\tilde{y}_i| > \lambda \mid |w_i| \gg \lambda) \approx 1$

(with the above fact)

$\Rightarrow$ we can bound $MSE(\hat{w})$ well

(b) $E((y - x^t\hat{w})^2)$

$= E([x^t(w^* - \hat{w}) + \varepsilon]^2) = E[(x^t(w^* - \hat{w}))^2] + E(\varepsilon^2)$

$= \underbrace{E[(x^t(w^* - \hat{w}))^2]}_{(*)} + 1$

$(*)$ $E([x^t(w^* - \hat{w})]^2) \underset{\text{C-S Inequality}}{\leq} E[\|x\|^2\|w^* - \hat{w}\|^2] = \|x\|^2 E\|w^* - \hat{w}\|^2$

$\leq \|x\|^2(2\log n + 1)\Big(1 + \underbrace{\sum_{i=1}^{n}\min(|w_i^*|^2, 1)}_{\leq K\ (\because \text{ at most } k \text{ non zero elts})}\Big)$

(end of the quiz)

$$(*) \leq \|x\|^2 (2\log n + 1)(1+K)$$

$$\therefore \; E\left((y - x^t \hat{w})^2\right) = E\left[\left[x^t (w^* - \hat{w})\right]^2\right] + 1$$

$$\leq \|x\|^2 (2\log n + 1)(1+K) + 1$$