COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q4 (a) (i) Projecting the class means to 1-D, we have

$$\mu_k^y = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k} y_n = \frac{1}{N_k}\sum_{n\in\mathcal{C}_k} \mathbf{w}^\mathsf{T}\mathbf{x}_n = \mathbf{w}^\mathsf{T}\frac{1}{N_k}\sum\nolimits_{n\in\mathcal{C}_k}\mathbf{x}_n = \mathbf{w}^\mathsf{T}\boldsymbol{\mu}_k, \qquad k=1,2$$

(2 marks, K)

(ii) The class variances in the FDA-projected space are

$$\begin{aligned}
(\sigma_k^y)^2 &= \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}(\mathbf{w}^\mathsf{T}\mathbf{x}_n - \mu_k^y)^2, \quad k=1,2.\\
&= \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}(\mathbf{w}^\mathsf{T}\mathbf{x}_n - \mathbf{w}^\mathsf{T}\boldsymbol{\mu}_k)^2\\
&= \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\|(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\mathbf{w}\|^2\\
&= \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\mathbf{w}\right]^\mathsf{T}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\mathbf{w}\right]\\
&= \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{w}^\mathsf{T}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\mathbf{w}.
\end{aligned}$$

(5 marks, KA)

(iii) The optimal projection vector $\mathbf{w}^*$ can be obtained by maximizing $J(\mathbf{w})$ with respect to $\mathbf{w}$. Using the results in (a) and (b), we have

$$\begin{aligned}
\mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmax}}\,\frac{(\mu_1^y - \mu_2^y)^2}{(\sigma_1^y)^2 + (\sigma_2^y)^2},\\
&= \underset{\mathbf{w}}{\text{argmax}}\,\frac{\left[\mathbf{w}^\mathsf{T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^\mathsf{T}\left[\mathbf{w}^\mathsf{T}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]}{\sum_{k=1}^2 \frac{1}{N_k}\sum_{n\in\mathcal{C}_k}\mathbf{w}^\mathsf{T}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\mathsf{T}\mathbf{w}}\\
&= \underset{\mathbf{w}}{\text{argmax}}\,\frac{\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\mathbf{w}\right]^\mathsf{T}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\mathsf{T}\mathbf{w}\right]}{\mathbf{w}^\mathsf{T}\mathbf{S}_W\mathbf{w}}\\
&= \underset{\mathbf{w}}{\text{argmax}}\,\frac{\mathbf{w}^\mathsf{T}\mathbf{S}_B\mathbf{w}}{\mathbf{w}^\mathsf{T}\mathbf{S}_W\mathbf{w}}.
\end{aligned}$$

(6 marks, AE)

COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

(b) The covariance of $\mathbf{x}$'s is the expectation of $(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathsf{T}}$:

$$\begin{aligned}\mathbb{E}\{(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\} &= \mathbb{E}\{(\mathbf{V}\mathbf{z}+\boldsymbol{\epsilon})(\mathbf{V}\mathbf{z}+\boldsymbol{\epsilon})^{\mathsf{T}}\}\\ &= \mathbf{V}\mathbb{E}\{\mathbf{z}\mathbf{z}^{\mathsf{T}}\}\mathbf{V}^{\mathsf{T}} + \mathbb{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^{\mathsf{T}}\}\\ &= \mathbf{V}\mathbf{I}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}\\ &= \mathbf{V}\mathbf{V}^{\mathsf{T}} + \boldsymbol{\Sigma}.\end{aligned}$$

(6 marks, AE)

(c) (i) The function $\phi(\mathbf{x})$ is to map $\mathbf{x}$ to a high-dimensional space.

(2 marks, K)

(ii) Using a non-linear kernel avoids evaluating the dot products in very high dimension space defined by ϕ, which could be of infinite dimension.

(2 marks, A)

(iii) $\phi(\mathbf{x}) = \mathbf{x}$, i.e., a linear function.

(2 marks, E)

COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q5 (a) (i) A perceptron with one neuron can only produce one decision plane, which will fail to separate the samples in this problem into two classes.

(3 marks, KA)

(ii) The neural network has a hidden layer with two non-linear neurons. Each neuron will give one decision plane. The samples of the two classes in this question spread to three regions in the input space, which require at least two lines for perfect separation. Therefore, two hidden neurons will be needed. However, the network can solve this problem only if the hidden neurons are nonlinear. This is because if they are linear, y is a linear function of $\mathbf{x}$, i.e., $y = \mathbf{W}\mathbf{x} + b$, which can provide one line only.

(4 marks, AE)

(b) (i) When $a_k = -\infty$, $y_k = 0$.
When $a_k = +\infty$, $y_k = 1$ according to L'Hospitals rule.
For other values of a_k, we have $0 < y_k < 1$ because $\sum_{k=1}^{2} y_k = 1$.

(4 marks)

(ii) Because t_k can be either 0 or 1, we have $t_2 = 1 - t_1$. Also, because $\sum_{k=1}^{2} y_k = 1$, we have $y_2 = 1 - y_1$. Put these two equations into E_{mce}, we have

$$\begin{aligned} E_{\text{mce}} &= -\sum_{\mathbf{x}\in\mathcal{X}} \sum_{k=1}^{2} t_k \log y_k \\ &= \sum_{\mathbf{x}\in\mathcal{X}} [-t_1 \log y_1 - t_2 \log y_2] \\ &= \sum_{\mathbf{x}\in\mathcal{X}} [-t_1 \log y_1 - (1 - t_1) \log(1 - y_1)] \\ &= E_{\text{bce}} \end{aligned}$$

(4 marks, AE)

(iii) This means that we do not need to use two outputs for binary classification problems. Simply use a network with one output and use binary cross-entropy as the loss function will do the job.

(3 marks, E)

(c) Because finding extrema means finding both the maximum and the minimum subject to the required constraint, we define the Lagrangian function as

$$L(x, y, \lambda) = x + y - \lambda(x^2 + y^2 - 2)$$

COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Take the derivatives of L and set the results to 0, we have

$$\frac{\partial L}{\partial x} = 1 - 2\lambda x = 0 \Longrightarrow \lambda = \frac{x}{2}$$
$$\frac{\partial L}{\partial y} = 1 - 2\lambda x = 0 \Longrightarrow \lambda = \frac{y}{2}$$
$$\frac{\partial L}{\partial \lambda} = x^2 + y^2 - 2 = 0 \Longrightarrow x^2 + y^2 = 2$$

(4 marks, A)

The first two equations suggest that $x = y$. Substituting $x = y$ into the third equation, we obtain

$$x^2 + x^2 = 2 \Longrightarrow x^2 = 1 \Longrightarrow x = \pm 1$$

Because $\frac{\partial f(x,y)}{\partial x} = \frac{\partial f(x,y)}{\partial y} = 1$, the slope is positive. Therefore, the maximum is 2 and it occurs at $(1, 1)$. The minimum is -2 and it occurs at $(-1, -1)$.

(3 marks, AE)

COURSE: EIE6207 YEAR: 6
SUBJECT: Theoretical Fundamental and Engineering Approaches for Intelligent Signal and Information Processing

	SUBJECT EXAMINER	INTERNAL MODERATOR / ASSESSOR	EXTERNAL EXAMINER	
	M.W. Mak			

Q6 (a)

$$\begin{aligned}
\text{mse}(\hat{\theta}) &= \mathbb{E}\left\{(\hat{\theta}-\theta)^2\right\} \\
&= \mathbb{E}\left\{\left[(\hat{\theta}-\mathbb{E}(\hat{\theta}))+(\mathbb{E}(\hat{\theta})-\theta)\right]^2\right\} \\
&= \mathbb{E}\left\{\left[\hat{\theta}-\mathbb{E}\{\hat{\theta}\right]^2\right\}+\left[\mathbb{E}(\hat{\theta})-\theta\right]^2+2\mathbb{E}\left\{(\hat{\theta}-\mathbb{E}\{\hat{\theta}\})(\mathbb{E}\{\hat{\theta}\}-\theta)\right\} \\
&= \text{var}(\hat{\theta})+\left[\mathbb{E}(\hat{\theta})-\theta\right]^2 \\
&= \text{var}(\hat{\theta})+b^2(\theta)
\end{aligned}$$

Note that the 3rd term of the 3rd equation is 0 because

$$\begin{aligned}
(\hat{\theta}-\mathbb{E}\{\hat{\theta}\})(\mathbb{E}\{\hat{\theta}\}-\theta) &= (\hat{\theta}-\theta-b(\theta))(\theta+b(\theta)-\theta) \\
&= (\hat{\theta}-\theta-b(\theta))b(\theta).
\end{aligned}$$

Taking expectation of the above term, we have

$$b(\theta)\left[\mathbb{E}\{\hat{\theta}\}-\theta-b(\theta)\right]=0.$$

(7 marks, KA)

(b) (i) The likelihood function is the product of N densities:

$$\begin{aligned}
p(\mathbf{x};A) &= \prod_{n=0}^{N-1}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right] \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right]
\end{aligned}$$

The log-likelihood function is

$$\log p(\mathbf{x};A) = -\log\left[(2\pi\sigma^2)^{\frac{N}{2}}\right]-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2$$

(4 marks, K)

(ii) The first derivative of the log-likelihood function is

$$\frac{\partial \log p(\mathbf{x}, A)}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) = \frac{N}{\sigma^2}(\bar{x} - A),$$

where $\bar{x}$ is the sample mean. The 2nd derivative is

$$\frac{\partial^2 \log p(\mathbf{x}, A)}{\partial A^2} = -\frac{N}{\sigma^2}.$$

Therefore, the CRLB of the best estimator of A is

$$\begin{aligned} \text{CRLB}(\hat{A}) &= \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \log p(\mathbf{x}, A)}{\partial A^2}\right]} \\ &= \frac{\sigma^2}{N}. \end{aligned}$$

(5 marks, AE)

(c) (i) If the time-of-flight measures are perfect, $\tau^2 = 0$. Then, using the 3rd equation, we have $K_t = c$. Substituting $K_t = c$ into the 2nd equation of the update formulae, we obtain $\sigma^2_{t|t} = 0$. As a result, the estimate $\hat{x}_{t|t}$ will be perfect.

(3 marks, KA)

(ii) The 3rd equation suggests that $K_t > 0$. As $\sigma^2_{t|t-1}$ and c are larger than 0, we have $\sigma^2_{t|t} < \sigma^2_{t|t-1}$. This means that the variance of the position estimate becomes smaller after taking z_t into account.

(3 marks, AE)

(iii) If τ^2 becomes very large, $K_t \to 0$. The first equation suggests that $\hat{x}_{t|t} = \hat{x}_{t|t-1}$, which automatically ignores z_t when estimating the position.

(3 marks, AE)