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) Denote the Lagrangian as

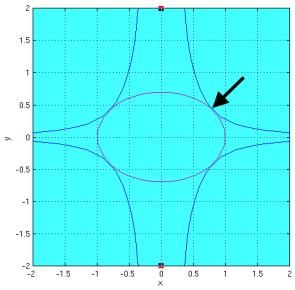
$$L(x, y, \lambda) = x^2 y + \lambda(x^2 + 2y^2 - 1),$$

where λ is a Lagrange multiplier. Then, we take the derivative of L with respect to x, y and λ, respectively, and set the resulting derivatives to 0.

$$\begin{aligned}
\frac{\partial L}{\partial x} &= 2xy + 2x\lambda = 0 \quad \Longrightarrow \quad \lambda = -y \\
\frac{\partial L}{\partial y} &= x^2 + 4y\lambda = 0 \quad \Longrightarrow \quad x = 2y \\
\frac{\partial L}{\partial \lambda} &= x^2 + 2y^2 - 1 = 0 \\
&\Longrightarrow \quad 4y^2 + 2y^2 - 1 = 0 \\
&\Longrightarrow \quad 6y^2 = 1 \Longrightarrow \quad y^* = \sqrt{\frac{1}{6}} \\
&\Longrightarrow \quad x^* = 2\sqrt{\frac{1}{6}} = \sqrt{\frac{2}{3}}
\end{aligned}$$

(10 marks, KA)

(b) Contour plot



(8 marks, KA)

(c) Small σ leads to sharp Gaussian functions, which means that the kernel function will only have non-zero value when $\mathbf{x}_i$ is very close to $\mathbf{x}_j$. This causes a large number of islands (or spots) in the input space, where each island corresponds to one support vectors. The decision boundaries have lots of sharp bends. On the other hand, large σ leads to flat Gaussian functions, which means that $K(\mathbf{x}_i, \mathbf{y}_j)$ is close to 1.0 even though $\mathbf{x}_i$ and $\mathbf{x}_j$ could be far apart. This results in very smooth decision boundaries.

(7 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			

Q5 (a) Assume that the data in $\mathcal{X}$ are i.i.d. Then, the conditional likelihood of $\mathcal{Y}$ is given by

$$\mathcal{L}(\mathcal{Y}|\mathcal{X},\boldsymbol{\beta}) = \prod_{i=1}^{N} \mathcal{N}(y_i|\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}, \sigma_y^2).$$

The conditional likelihood of y_i is a Gaussian distribution with mean $\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ and variance σ_y^2 because

$$\mu_{y_i} = \mathbb{E}\{y_i\} = \boldsymbol{\beta}^{\mathsf{T}}\mathbb{E}\{\mathbf{x}_i\} + \mathbb{E}\{\boldsymbol{\epsilon}_i\} = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$$

and

$$\begin{aligned}\sigma_{y_i}^2 &= \mathbb{E}\{y_i^2\} - (\mathbb{E}\{y_i\})^2 \\ &= \mathbb{E}\{(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i)^2 + 2\epsilon_i\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i + \epsilon_i^2\} - (\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})^2 \\ &= \sigma^2.\end{aligned}$$

The log-likelihood of $\mathcal{Y}$ is

$$\log \mathcal{L}(\mathcal{Y}|\mathcal{X},\boldsymbol{\beta}) = \sum_{i=1}^{N}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(y_i - \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i)^2}{2\sigma^2}\right].$$

Then, the maximum-likelihood estimate is

$$\boldsymbol{\beta}_{\mathrm{ML}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{N} -\frac{(y_i - \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i)^2}{2\sigma^2} \tag{1}$$

$$= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_i)^2 \tag{2}$$

$$= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{3}$$

which is the same as $\boldsymbol{\beta}_{\mathrm{LS}}$.

(13 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) (i)

$$\begin{aligned}
\hat{x}_{t|t} &= \frac{1}{t}\left[\sum_{i=1}^{t-1} z_i + z_t\right] \\
&= \frac{1}{t}\sum_{i=1}^{t-1} z_i + \frac{1}{t} z_t \\
&= \frac{t-1}{t} \cdot \frac{1}{t-1}\sum_{i=1}^{t-1} z_i + \frac{1}{t} z_t \\
&= \frac{t-1}{t}\hat{x}_{t|t-1} + \frac{1}{t} z_t \\
&= \hat{x}_{t|t-1} + \frac{1}{t}\left(z_t - \hat{x}_{t|t-1}\right)
\end{aligned}$$

(7 marks, AE)

(ii) The recursive formula in (b)(i) is the state update equation of the Kalman filter

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K}_t(\mathbf{z}_t - \mathbf{H}_t\hat{\mathbf{x}}_{t|t-1})$$

in which $\mathbf{H}_t = 1$ and $\mathbf{K}_t = \frac{1}{t}$. Note that the covariance update formula of the Kalman filter is

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_{t|t-1}.$$

In the context of this problem, we have $\mathbf{P}_{t|t}$ being the variance of the estimate $\hat{\mathbf{x}}_{t|t}$. As $\mathbf{K}_t = \frac{1}{t}$, the variance of $\hat{\mathbf{x}}_{t|t}$ decreases when t increases.

(5 marks, AE)

DEPARTMENT / INSTITUTE / SCHOOL / CENTRE OF EIE

SOLUTION & MARKING SCHEME (2018/ 2019)

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) Essay type questions. Marks will be given according to (1) validity of arguments, (2) evidences supporting the arguments and (3) clarity of writing.

(15 marks, AE)

(b) Situations include (1) feature dimension very high or the number of training vectors is smaller than the feature dimension, (2) the data do not follow a Gaussian mixture distribution, e.g., rolling a die, and (3) the data is categorical, e.g., gender, month, voting preference, etc. The reasons are that under these situations, the inverse of the covariance matrix does not exists or the covariance matrix is almost singular. Also, it does not make sense to use a Gaussian density function to fit categorical data.

(5 marks, AE)

(c) The maximum rank of the within-class covariance matrix is 1000 (or 999 if the data is not zero mean), which is the same as the dimension of the within-class covariance matrix. Numerical error is likely to occur during the computation of its inverse. As the rank of the between-class covariance matrix is 9 (no. of classes - 1), the maximum dimension of the LDA projected vectors is 9. A better approach is to use PCA followed by LDA. If PCA is used, we may project the data to M-dimensional space first, where $9 < M << 1000$. The 1,000 M-dimensional vectors will allow us to compute the inverse of the within-class covariance matrix without numerical difficulty.

(5 marks, AE)