UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2022/23

ADVANCED MACHINE LEARNING

Duration 120 mins (2 hours)

This paper is a WRITE-ON examination paper.

You ***must*** write your Student ID on this Page and must not write your name anywhere on the paper.

All answers should be written within the designated boxes in this examination paper and sufficient space is provided for each question.

If, for some reason, space is required to complete or correct an answer to a question, use the "Additional Space" provided on the facing or adjacent page to the question. Clearly indicate which question the answer corresponds to.

No credit will be given for answers presented elsewhere and without clear indication of to what question they correspond. Blue answer books may be used for scratch; they will be discarded without being looked at.

Answer all parts of the question in section A (40 marks)
and ALL three questions from section B (20 marks each)

Student ID:

Question	Mark	*Arithmetic checked*	*Double Marked*
A1	/40		
B2	/20		
B3	/20		
B4	/20		
Total:	/100		

University approved calculators MAY be used.

A foreign language translation dictionary (paper version) is permitted provided it contains no notes, additions or annotations.

10 page examination paper

 Version 1

Section A

A 1

(a) Explain why in boosting it is important to use as decorrelated trees as possible and explain how this is achieved in random forest. [5 marks]

The idea of boosting is to reduce the variance in the prediction of the machine by averaging over many machines. This only works if the machines are decorrelated. In random forest this is achieved both by using booststapping and by building decision trees that use only a small randomly sampled subset of the features.

(b) Multilayer perceptrons (MLPs) are famously universal approximators (they can approximate any smooth function up to arbitrary precision). This should make them susceptible to massively overfitting a training set. Explain why, despite this, MLPs often have good generalisation performance. [5 marks]

Although the class of MLPs (considering any architecture) is universal approximators, a particular MLP (with a fixed number of hidden units) will not be a universal approximator. In practice we choose the architecture of an MLP to fit the complexity of the function we are trying to learn (depending on the size of the training set). This prevents overfitting.

(c) Show that if $\lambda > 0$ is an eigenvalue of $\mathbf{C} = \mathbf{X}\,\mathbf{X}^{\mathsf{T}}$ then it is also an eigenvalue of $\mathbf{D} = \mathbf{X}^{\mathsf{T}}\mathbf{X}$, where $\mathbf{X}$ is a matrix. [5 marks]

(Algebraic questions requiring a good grasp of linear algebra that underlies are lot of duality in machine learning.)

If $\lambda \neq 0$ is an eigenvalue of C then there exists some eigenvector v ($\neq 0$) such that $\mathbf{C}\,v = \lambda\,v$. Multiplying on the left by $\mathbf{X}^{\mathsf{T}}$ then

$$\mathbf{X}^{\mathsf{T}}\mathbf{C}\,v = \lambda\,\mathbf{X}^{\mathsf{T}}\,v$$

but $\mathbf{X}^{\mathsf{T}}\mathbf{C} = \mathbf{X}^{\mathsf{T}}\,(\mathbf{X}\,\mathbf{X}^{\mathsf{T}}) = (\mathbf{X}^{\mathsf{T}}\,\mathbf{X})\,\mathbf{X}^{\mathsf{T}} = \mathbf{D}\,\mathbf{X}^{\mathsf{T}}$ (where we have used the associativity of matrix multiplication and the definitions of C and D). Thus

$$\mathbf{D}\,\mathbf{X}^{\mathsf{T}}\,v = \lambda\,\mathbf{X}^{\mathsf{T}}\,v$$

or $\mathbf{D}u = \lambda\,u$ where $u = \mathbf{X}^{\mathsf{T}}\,v \neq 0$ is an eigenvector of D with eigenvalue λ.

(d) If $\|\boldsymbol{x}\|$ is a proper norm, use the triangular inequality ($\|\boldsymbol{x}+\boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$), linearity of a norm ($\|a\,\boldsymbol{x}\| = a\,\|\boldsymbol{x}\|$) and the definition of convexity, to show that the norm is convex. [5 marks]

(The students have not seen this. The algebra is simple, but the students need a good understanding to do the proof.)

For any vectors, x, and y and any scalar $a \in [0,1]$ then

$$\|a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}\| \leq \|a\,\boldsymbol{x}\| + \|(1-a)\,\boldsymbol{y}\| = a\,\|\boldsymbol{x}\| + (1-a)\,\|\boldsymbol{y}\|$$

where the first inequality follows from the triangular inequality and the second equality from the linearity of the norm. However,

$$\|a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}\| \leq a\,\|\boldsymbol{x}\| + (1-a)\,\|\boldsymbol{y}\|$$

is the defining equation of convexity.

(e) Use the fact norms are convex to argue that an elastic net with a loss function

$$L(\boldsymbol{w}) = \sum_{i=1}^{m} \left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_i - y_i\right)^2 + \alpha\|\boldsymbol{w}\|_{L_1} + \beta\,\|\boldsymbol{w}\|_{L_2}^2$$

has a unique minimum. [5 marks]

The quadratic minima and the two norms are all convex (up) functions. The sum or convex functions is convex. We note that the L_2 norm, $\|w\|_{L_2}^2$, is strongly convex which ensures that the minimum is unique.

(f) Explain the main advantage and disadvantage of stochastic gradient descent (SGD) compared to full gradient descent. [5 marks]

The overwhelming advantage of SGD is that it is far faster to compute the gradient with respect to a mini-batch than computing the gradient using the whole training set.

A significant disadvantage is that the gradient estimates can have very high added noise. Thus the algorithm will typically wander around. It will also never converge.

(g) The multinomial distribution $\mathrm{Multi}(\boldsymbol{k}|n,\boldsymbol{p})$ describes the likelihood of observed counts $\boldsymbol{k} = (k_1, k_2, \ldots, k_d)$ for d possible outcomes, where the total number of counts is $\sum_{i=1}^{d} k_i = n$. The vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_d)$ is a vector of probabilities (summing to 1) where p_i is the probability of outcome i occuring. Show that the Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha})$ is a conjugate prior to the multinomial likelihood $\mathrm{Multi}(\boldsymbol{k}|n,\boldsymbol{p})$ where

$$\mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \Gamma(\alpha_0)\prod_{i=1}^{d}\frac{p_i^{\alpha_i-1}}{\Gamma(\alpha_i)} \qquad \mathrm{Multi}(\boldsymbol{k}|n,\boldsymbol{p}) = n!\prod_{i=1}^{d}\frac{p_i^{k_i}}{k_i!}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ is a vector of parameters that controls the Dirichlet distribution with $\alpha_0 = \sum_{i=1}^{n}\alpha_n$. Derive the update equation for the parameters $\boldsymbol{\alpha}$. [5 marks]

We only need to consider the functional form with respect to p_i. Thus the posterior is proportional to

$$f(\boldsymbol{p}|\boldsymbol{k}) \propto \prod_{i=1}^{d}\frac{p_i^{k_i}}{k_i!}\prod_{i=1}^{d}\frac{p_i^{\alpha_i-1}}{\Gamma(\alpha_i)} \propto \prod_{i=1}^{n} p_i^{\alpha_i+k_i-1} \propto \mathrm{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}+\boldsymbol{k}).$$

The update equation is thus $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{k}$.

(h) Describe how the minimum description length formalism is used for model selection.

[5 marks]

In the minimum description length formalism we choose a model that generates a prediction that minimises the length of the model plus the length of the residual errors between the prediction of the model and the true data. The description length of model is a minimum encoding of the parameters of the model, while the length message needed to communicate the residual errors is usually taken to be the Shannon bound.

End of question A1

Section B

B 2

(a) Show that the expected generalisation given by

$$\mathbb{E}_{\mathcal{D}}\big[E(\mathcal{D})\big] = \mathbb{E}_{\mathcal{D}}\left[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right]$$

can be written as the sum of a bias term, B, and variance term V where

$$B = \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2, \quad V = \mathbb{E}_{\mathcal{D}}\left[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)^2\right]$$

where $\hat{f}_m(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\Big[\hat{f}(\boldsymbol{x}|\mathcal{D})\Big]$ is the prediction made by averaging over all machines.

[10 marks]

(Students find this derivation difficult because it requires understanding expectations. Being rather conceptual students often struggle with this. Two marks for: (i) writing the first line, (ii) using the trick of adding and subtracting the prediction of the mean machine, (iii) multiplying out the square, (iv) arguing the cross term vanishes (v) explaining the final answer.)

The first step is to subtract and add the response of the mean machine

$$\begin{aligned}
\bar{E}_G = \mathbb{E}_{\mathcal{D}}\big[E_G(\mathcal{D})\big] &= \mathbb{E}_{\mathcal{D}}\left[\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\ \left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\,\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\,\mathbb{E}_{\mathcal{D}}\left[\left(\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)\ + \left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)\right)^2\right] \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})\Bigg(\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)^2 + \left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right] \\
&\qquad\qquad + \mathbb{E}_{\mathcal{D}}\Big[2\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)\left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)\Big]\Bigg)
\end{aligned}$$

The last term vanishes on taking the expectation. This leaves us with the variance term plus the bias.

(b) Explain in words (1) the *bias* and (2) the *variance* terms and (3) the dilemma. [6 marks]

(i) The *bias* is the generalisation performance of the mean machine (i.e. the prediction made by taking the mean response averaged over machines trained with every possible training set of a given size).

(ii) The *variance* measures how the responses of the individual machines vary from machine to machine.

(iii) To get good generalisation we want to both reduce the bias and variance. To reduce the bias we need a complex machine, but that will typically lead to an increase of the variance. Conversely a simple machine is likely to have a small variance, but a high bias.

(c) Consider a classification problem where $\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})$ is the softmax output of a learning machine trained on a dataset $\mathcal{D}$ for class c. Let $\hat{m}_c(\boldsymbol{x}) = \mathbb{E}_\mathcal{D}\Big[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\Big]$ be the output of the mean machine for class c (i.e. the outputs averaged over machines trained on all possible datasets). Consider the cross-entropy loss

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{(\boldsymbol{x},y)\in\mathcal{D}}\sum_{c\in\mathcal{C}} [\![y = c]\!] \log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\Big)$$

where $[\![y = c]\!]$ is an indicator function equal to 1 if the target y is equal to class c, and 0 otherwise. Show that the expected loss over all training sets can be written as the expected loss for the mean machine (a bias) plus an additional term (a variance). Use Jensen's inequality ($\mathbb{E}\left[\log(X)\right] \leq \log\big(\mathbb{E}\left[X\right]\big)$) to show the second term is non-negative. [4 marks]

We can add and subtract the loss for the mean machine

$$\begin{aligned}\bar{L} &= -\mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_\mathcal{D}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!] \log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\Big)\right]\right] \\ &= -\mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right] - \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_\mathcal{D}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!] \log\left(\frac{\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})}{\hat{m}_c(\boldsymbol{x})}\right)\right]\right]\end{aligned}$$

The first term acts like a bias. The second (variance-like) term is

$$V = -\mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!]\, \mathbb{E}_\mathcal{D}\Big[\log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\Big)\Big]\right] + \mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right]$$

But using Jensen's inequality
$\mathbb{E}_\mathcal{D}\Big[\log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\Big)\Big] \leq \log\Big(\mathbb{E}_\mathcal{D}\Big[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta}_\mathcal{D})\Big]\Big) = \log(\hat{m}_c(\boldsymbol{x}))$**. Thus**

$$V \geq -\mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right] + \mathbb{E}_{(\boldsymbol{x},y)}\left[\sum_{c\in\mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right] = 0$$

that is, the second term is non-negative.

End of question B2

B 3

(a) Show that for the mapping

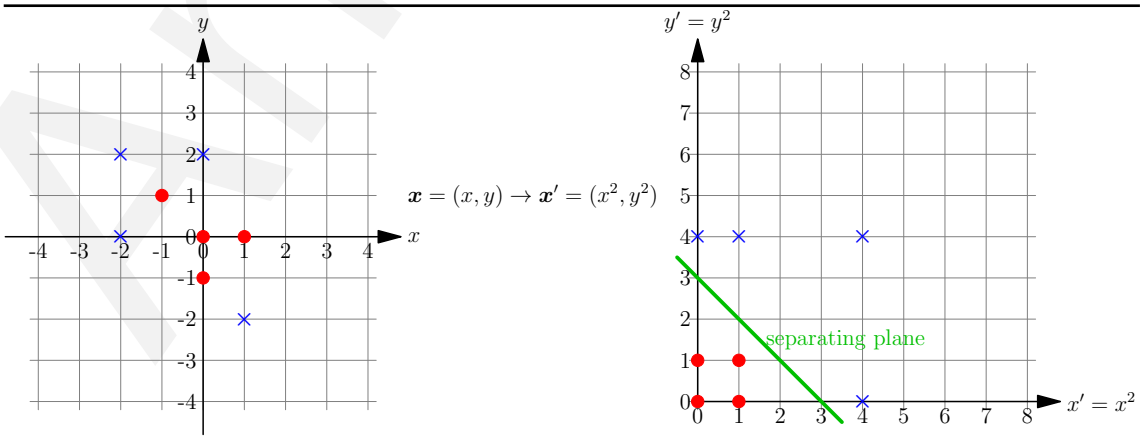
$$\boldsymbol{x} = (x_1, x_2, x_3)^\mathsf{T} \rightarrow \vec{\phi}(\boldsymbol{x}) = (x_1^2, x_2^2, x_3^2, \sqrt{2}\,x_1\,x_2,\ \sqrt{2}\,x_1\,x_3,\ \sqrt{2}\,x_2\,x_3)^\mathsf{T}$$

the kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \left\langle \vec{\phi}(\boldsymbol{x}), \vec{\phi}(\boldsymbol{y}) \right\rangle$ is equal to $(\boldsymbol{x}^\mathsf{T}\boldsymbol{y})^2$. [5 marks]

(This is an extension of the 2-D example shown in the lecture.)

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{y}) &= \left\langle \vec{\phi}(\boldsymbol{x}), \vec{\phi}(\boldsymbol{y}) \right\rangle \\ &= (x_1^2, x_2^2, x_3^2, \sqrt{2}\,x_1\,x_2,\ \sqrt{2}\,x_1\,x_3,\ \sqrt{2}\,x_2\,x_3) \\ &\quad (y_1^2, y_2^2, y_3^2, \sqrt{2}\,y_1\,y_2,\ \sqrt{2}\,y_1\,y_3,\ \sqrt{2}\,y_2\,y_3)^\mathsf{T} \\ &= x_1^2y_1^2 + x_2^2y_2^2 + x_3^2y_2^2 + 2\,x_1\,y_1\,x_2\,y_2 + 2\,x_1\,y_1\,x_3\,y_3 + 2\,x_2\,y_2\,x_3\,y_3 \\ &= (x_1\,y_1 + x_2\,y_2 + x_3\,y_3)^2 = (\boldsymbol{x}^\mathsf{T}\boldsymbol{y})^2 \end{aligned}$$

(b) Show how the data points $\{\boldsymbol{x}_i = (x_i, y_i) | i = 1, 2, \ldots\}$ shown below transform under the mapping $\boldsymbol{x} = (x_i, y_i) \rightarrow \boldsymbol{x}' = (x_i^2, y_i^2)$ and sketch the position of the maximal margin dividing plane in the new feature space. [5 marks]



(c) Give three conditions that any positive semi-definite kernel, $K(\boldsymbol{x}, \boldsymbol{y})$, should satisfy.

[6 marks]

- **For any function** $f(\boldsymbol{x})$

$$\iint f(\boldsymbol{x})\, K(\boldsymbol{x}, \boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} \geq 0.$$

- **The eigenvalues are non-negative.**
- **Can be written as an inner-product**

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{y}) \rangle = \sum_i \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{y})$$

for some (possibly infinite) vector function $\boldsymbol{\phi}(\boldsymbol{x})$**.**

(d) Using properties of positive semi-definite kernels to show that

$$K^{(3)}(\boldsymbol{x}, \boldsymbol{y}) = K^{(2)}(\boldsymbol{x}, \boldsymbol{y})\, K^{(1)}(\boldsymbol{x}, \boldsymbol{y})$$

is positive semi-definite if $K^{(1)}(\boldsymbol{x}, \boldsymbol{y})$ and $K^{(2)}(\boldsymbol{x}, \boldsymbol{y})$ are positive semi-definite.

[4 marks]

As $K^{(1)}(\boldsymbol{x}, \boldsymbol{y})$ **and** $K^{(2)}(\boldsymbol{x}, \boldsymbol{y})$ **are positive-definite we can write**

$$K^{(1)}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \phi_i^{(1)}(\boldsymbol{x})\, \phi_i^{(1)}(\boldsymbol{y}) \qquad K^{(2)}(\boldsymbol{x}, \boldsymbol{y}) = \sum_j \phi_j^{(2)}(\boldsymbol{x})\, \phi_j^{(2)}(\boldsymbol{y})$$

and thus we can write

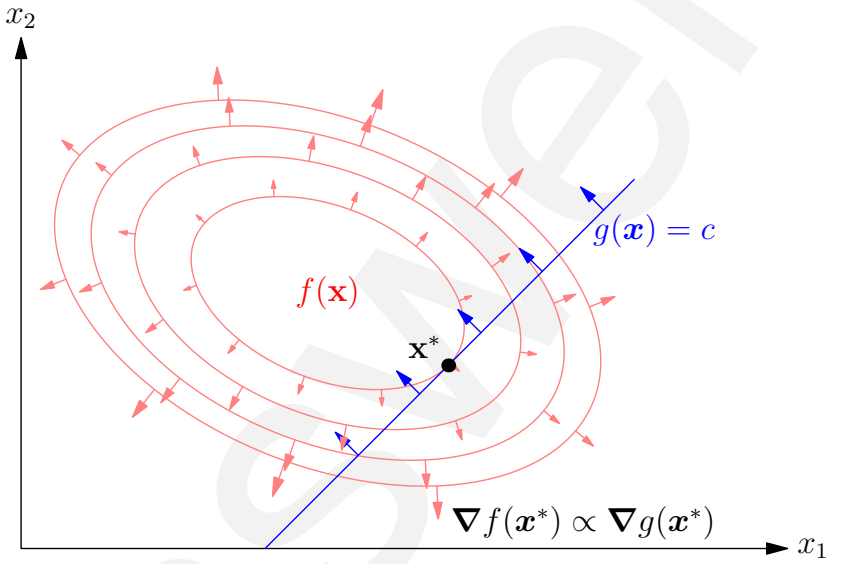
$$\begin{aligned} K^{(3)}(\boldsymbol{x}, \boldsymbol{y}) &= \sum_{i,j} \phi_i^{(1)}(\boldsymbol{x})\, \phi_j^{(1)}(\boldsymbol{x})\, \phi_i^{(2)}(\boldsymbol{y})\, \phi_j^{(2)}(\boldsymbol{y}) \\ &= \sum_{i,j} \phi_{ij}^{(3)}(\boldsymbol{x})\, \phi_{ij}^{(3)}(\boldsymbol{y}) = \left\langle \boldsymbol{\phi}^{(3)}(\boldsymbol{x}), \boldsymbol{\phi}^{(3)}(\boldsymbol{y}) \right\rangle \end{aligned}$$

where $\boldsymbol{\phi}^{(3)}(\boldsymbol{x})$ **is a vector with components** $\phi_{ij}^{(3)}(\boldsymbol{x}) = \phi_i^{(1)}(\boldsymbol{x})\, \phi_j^{(2)}(\boldsymbol{x})$**.**
Therefore $K^{(3)}(\boldsymbol{x}, \boldsymbol{y})$ **is positive semi-definite.**

End of question B3

B 4

(a) Below we show contour lines for a quadratic minimum $f(\boldsymbol{x})$ and a constraint $g(\boldsymbol{x}) = x_2 - x_1 = c$. Plot the gradient $\nabla f(\boldsymbol{x})$ at various points along the contour lines and $\nabla g(\boldsymbol{x})$ at various points along the constraint. Mark the point that minimises $f(\boldsymbol{x})$, subject to the constraint $g(\boldsymbol{x}) = c$. Write down the condition for the minimum points. [10 marks]



(2 marks showing gradient for $f(\boldsymbol{x})$, 2 marks showing gradient for constraint $g(\boldsymbol{x}) - c = 0$, 3 marks for identifying $\boldsymbol{x}^*$, 3 marks for noting $\nabla f(\boldsymbol{x}^*) \propto \nabla g(\boldsymbol{x}^*)$.)

(b) Consider for a dataset $\{\boldsymbol{x}_i | i = 1, 2, \ldots, n\}$. Subtracting the mean and projecting onto a vector $\boldsymbol{v}$ gives us a number $z_i = \boldsymbol{v}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})$. Show that the direction $\boldsymbol{v}$, with $\|\boldsymbol{v}\|^2 = 1$, that maximises the variance of the set of numbers $\{z_i | i = 1, 2, \ldots, n\}$, is given by the eigenvector of the covariance matrix with the largest eigenvalue. [10 marks]

We consider the variance

$$\sigma^2(\boldsymbol{v}) = \frac{1}{n-1}\sum_{i=1}^{n} z_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{v}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})\right)^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{v}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu})\,(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\,\boldsymbol{v} = \boldsymbol{v}^{\mathsf{T}}\left(\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})\,(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}\right)\boldsymbol{v} = \boldsymbol{v}^{\mathsf{T}}\mathbf{C}\,\boldsymbol{v}$$

where

$$\mathbf{C} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})\,(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathsf{T}}$$

is the empirical covariance matrix. To compute the direction that maximises this variance we write a Lagrangian

$$\mathcal{L} = \boldsymbol{v}^{\mathsf{T}}\mathbf{C}\,\boldsymbol{v} - \lambda\,\left(\|\boldsymbol{v}\|^2 - 1\right)$$

which is extremised when

$$\boldsymbol{\nabla}\mathcal{L} = 2\mathbf{C}\,\boldsymbol{v} - 2\,\lambda\,\boldsymbol{v}$$

or $\mathbf{C}\,\boldsymbol{v} = \lambda\,\boldsymbol{v}$. That is, the direction is an eigenvector of $\mathbf{C}$. But, for an eigenvector, $\boldsymbol{v}$ of $\mathbf{C}$

$$\sigma^2(\boldsymbol{v}) = \boldsymbol{v}^{\mathsf{T}}\,\mathbf{C}\,\boldsymbol{v} = \lambda\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v} = \lambda.$$

Thus the direction, v, that maximises the variance of the data projected onto v is given by the eigenvector with the largest eigenvalue.

(4 marks for showing $\sigma^2(v) = \boldsymbol{v}^{\mathsf{T}}\mathbf{C}\boldsymbol{v}$, 2 marks for writing down Lagrangian, 2 marks for solving Lagrangian and two marks for arguing that we need to select the eigenvector with the largest eigenvalue.)

End of question B4

END OF PAPER