3M1 Mathematical Methods, 2023

Linear Algebra

1. (a) i. $H_{ij} = \partial f/\partial x_i \partial x_j = \partial f/\partial x_j \partial x_i$.

ii. $\boldsymbol{H}$ is positive-definite, then $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} > 0$ for all non-zero $\boldsymbol{v}$. Therefore $f(\boldsymbol{x})$ must be a local minimum.

$\boldsymbol{H}$ is negative-definite, then $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} < 0$ for all non-zero $\boldsymbol{v}$. Therefore $f(\boldsymbol{x})$ must be a local maximum.

$\boldsymbol{H}$ is indefinite, then there exist vectors $\boldsymbol{u}^T\boldsymbol{H}\boldsymbol{u} < 0$ and $\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{w} > 0$ therefore $f(\boldsymbol{x})$ is a saddle point.

$\boldsymbol{H}$ is semi-definite, then we cannot say if $f(\boldsymbol{x})$ is a minimum, maximum or saddle point.

iii. To be a norm, $\|\boldsymbol{x}\|_H$ must be real and zero only when $\boldsymbol{x} = 0$. This requires that $\boldsymbol{H}$ be positive-definite.

(b) i.

$$\begin{aligned}\boldsymbol{r}_{k+1} &= \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_{k+1} \\ &= \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_k + w(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)) \\ &= (\boldsymbol{I} - w\boldsymbol{A})\boldsymbol{r}_k\end{aligned}$$

Therefore $\boldsymbol{r}_k = (\boldsymbol{I} - w\boldsymbol{A})^k \boldsymbol{r}_0$.

ii.

$$\begin{aligned}\boldsymbol{x}_{k+1} - \boldsymbol{x} &= \boldsymbol{x}_k + w(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k) - \boldsymbol{x} \\ &= \boldsymbol{x}_k + w(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x}_k) - \boldsymbol{x} \\ &= (\boldsymbol{I} - w\boldsymbol{A})(\boldsymbol{x}_k - \boldsymbol{x})\end{aligned}$$

Therefore $\boldsymbol{e}_k = (\boldsymbol{I} - w\boldsymbol{A})^k \boldsymbol{e}_0$.

iii. For $\boldsymbol{e}_{k+1} = (\boldsymbol{I} - w\boldsymbol{A})\boldsymbol{e}_k$, note that

$$\|\boldsymbol{e}_{k+1}\| \le \|\boldsymbol{I} - w\boldsymbol{A}\|\|\boldsymbol{e}_k\|.$$

Clearly convergence requires $\|\boldsymbol{I} - w\boldsymbol{A}\| < 1$. In the 2-norm this implies that $\lambda_{\max}(\boldsymbol{I} - w\boldsymbol{A}) < 1$. Noting that eigenvalues of $(\boldsymbol{I} - w\boldsymbol{A})$ are $1 - w\lambda_i$, where λ_i are the eigenvalues of $\boldsymbol{A}$, convergence requires that $|1 - w\lambda_i| < 1$. Therefore convergence requires that $0 < w < 2/\lambda_{\max}(\boldsymbol{A})$.

Optimality:

$$\begin{aligned}\|\boldsymbol{I} - w\boldsymbol{A}\|_2 &= \max_i |1 - w\lambda_i(\boldsymbol{A})| \\ &= \max_i (|1 - w\lambda_{\min}(\boldsymbol{A})|, |1 - w\lambda_{\max}(\boldsymbol{A})|)\end{aligned}$$

Aim is to minimise $\|\boldsymbol{I} - w\boldsymbol{A}\|$, which happens when $1 - w\lambda_{\min}(\boldsymbol{A})$ and $|1 - w\lambda_{\max}(\boldsymbol{A})|$ are equal in magnitude but opposite in sign, i.e.

$$1 - w\lambda_{\max}(\boldsymbol{A}) = -1 + w\lambda_{\min}(\boldsymbol{A}).$$

which gives

$$w_{\text{opt}} = \frac{2}{\lambda_{\max}(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{A})}.$$

Optimisation

For D-dimensional column vector $\mathbf{x} = [x_1, x_2, \cdots, x_D]^\mathsf{T} \in \mathbb{R}^D$ consider the product of logistic functions

$$s(\mathbf{x}) = \prod_{d=1}^{D} \frac{1}{1 + \exp(-x_d)}$$

In addition consider the Gaussian distribution over $\mathbf{x} \in \mathbb{R}^D$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

with mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and $D \times D$ covariance matrix $\boldsymbol{\Sigma}$. For the probability

$$P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\mathcal{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} s(\mathbf{x}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\mathcal{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int_{\mathbb{R}^D} s(\mathbf{x}) \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{x}$

1. Show that the point $\mathbf{x}^* \in \mathbb{R}^D$ yielding the maximum value of $P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be obtained by minimising the function

$$f(\mathbf{x}) = \mathbf{1}^\mathsf{T} \mathbf{g}(\mathbf{x}) + \frac{1}{2} \mathbf{x}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \mathbf{x}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$$

   where the D-dimensional vector of ones is denoted as $\mathbf{1} = [1, 1, \cdots, 1]^\mathsf{T}$ and $\mathbf{g}(\mathbf{x})$ is an element-wise application of the function $g(\cdot)$ which acts on each component of the vector $\mathbf{x}$ such that $\mathbf{g}(\mathbf{x}) = [g(x_1), g(x_2), \cdots, g(x_D)]^\mathsf{T}$ with each $g(\cdot)$ defined appropriately. [20%]

   $f(\mathbf{x}) = -\log P(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and noting that many of the constants are independent of $\mathbf{x}$ yields the quadratic term $\frac{1}{2}\mathbf{x}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \mathbf{x} - \mathbf{x}^\mathsf{T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ the negative logarithm of the product of logistic functions gives $\sum_d \log(1 + \exp(-x_d)) = \sum_d g(x_d)$ where each $g(x_d) = \log(1 + \exp(-x_d))$ and the sum can be written in vector form as given.

2. Derive and justify the necessary condition for a point $\mathbf{x}^* \in \mathbb{R}^D$ to be a strong local minimum of the function $f(\mathbf{x})$. [15%]

   For a point $\mathbf{x}^*$ to be a local minimum and as there are no constraints so that all directions in $\mathbb{R}^D$ are feasible then the necessary condition is that $\nabla f(\mathbf{x}^*) = 0$ which for the function should be $\mathbf{g}'(\mathbf{x}^*) + \boldsymbol{\Sigma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}) = 0$ where $\mathbf{g}'(\mathbf{x}^*)$ is a component wise defined vector with elements $g'(.) = -(1 + \exp(\cdot))^{-1}$.

3. Derive and justify the sufficient condition for a point $\mathbf{x}^* \in \mathbb{R}^D$ to be a strong local minimum of the function $f(\mathbf{x})$. [15%]

For an arbitrary vector $\mathbf{d}$ then $\mathbf{d}^{\mathsf{T}}(g''(\mathbf{x}^*) + \boldsymbol{\Sigma}^{-1})\mathbf{d} > 0$ must be satisfied for $\mathbf{x}^*$ to be a strong local minimum. Where the non-zero elements of the diagonal matrix $g''(\mathbf{x}^*)$ are the second derivatives $g''(x_i^*) = \frac{\exp(x_d)}{(1+\exp(x_d))^2}$.

4. Derive a Newton optimisation method for the function $f(\mathbf{x})$ acting on $\mathbf{x} \in \mathbb{R}^D$. [20%]

   The Newton Scheme takes the gradient and Hessian terms such that $\mathbf{x} \leftarrow \mathbf{x} - (g''(\mathbf{x}^*) + \boldsymbol{\Sigma}^{-1})^{-1}(\mathbf{g}'(\mathbf{x}^*) + \boldsymbol{\Sigma}^{-1}(\mathbf{x}^* - \boldsymbol{\mu}))$.

5. For the specific case where $D = 2$, and $\boldsymbol{\Sigma}$ is a diagonal covariance matrix, assess whether $f(\mathbf{x})$ is convex in $\mathbb{R}^2$ and state the implication on the nature of the point $\mathbf{x}^*$. [30%]

   The Hessian is $\mathbf{D} + g^{''}(\mathbf{x}_k)$ where $\mathbf{D}$ is diagonal with terms D_{11} and D_{22} which are the reciprocal values of the diaognal terms in the covariance matrix. This yields a determinant $(D_{11} + g^{''}(x_1))(D_{22} + g^{''}(x_2))$ that has to be strictly greater than zero for all $\mathbf{x}$ for it to be convex. It is clear that when the negative logarithm of the logistic function is used for $g(\cdot)$ then this will be the case, ie. the strict positivity will hold due to the positive value of the function for all values of argument. This therefore implies that the minimum is in fact a unique global minimum.

3M1 Mathematical Methods, 2022

1. *Discrete State Space Models*

(a) Every element of $\mathbf{P}$ must be positive and the sum of the elements in each row must be 1. [10%]

(b) The general expression to be satisfied is

$$\pi = \pi \mathbf{P}$$

However as the structure of the matrix means that the first m states are *transient* then the expression can be simplified to

$$\pi_1 = \mathbf{0}; \quad \pi_2 = \pi_2 \mathbf{P}_{22}$$

The distribution does not depend on the initial state if the states associated with $\mathbf{P}_{22}$ are recurrent. [20%]

(c) If $\mathbf{P}_{12} = \mathbf{0}$ then there may be two distinct stationary distributions, one based on $\mathbf{P}_{11}$ the other on $\mathbf{P}_{22}$. The selection of the stationary distribution depends on the initial state in this case. [10%]

(d) (i) The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0.5 & b & 0.0 & 0.0 & a \\ 0.0 & 0.0 & 0.5 & 0.5 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 1.0 \\ 0.0 & 1.0 & 0.0 & 0.0 & 1.0 \\ 1.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

[15%]

(d) (ii) When $a = 0.5$ this means that $b = 0.0$. From Part (c) this means that the network can be analysed seperatey and the stationary distribiution wil depend on the initial state.

$$\begin{bmatrix} x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix} = \begin{bmatrix} x_2 & x_3 & x_4 \end{bmatrix}$$

This yields

$$\begin{aligned} x_3 + x_4 &= x_2 \\ 0.5x_2 &= x_3 \\ 0.5x_2 &= x_4 \end{aligned}$$

This yields the following solution

$$\begin{bmatrix} x_2 & x_3 & x_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

For states 1 and 5

$$\begin{bmatrix} x_1 & x_5 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 1.0 & 0.0 \end{bmatrix} = \begin{bmatrix} x_1 & x_5 \end{bmatrix}$$

This yields

$$\begin{aligned} 0.5x_1 + x_5 &= x_1 \\ 0.5x_1 &= x_5 \end{aligned}$$

This yields the following solution

$$\begin{bmatrix} x_1 & x_5 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \end{bmatrix}$$

The stationary distribution depends on the starting state (Part b). The final distribution is thus of the form

$$\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{bmatrix} = \begin{bmatrix} \frac{2\alpha}{3} & \frac{1-\alpha}{2} & \frac{1-\alpha}{4} & \frac{1-\alpha}{4} & \frac{\alpha}{3} \end{bmatrix}$$

where $0 \leq \alpha \leq 1$. [25%]

(d)(iii) When $b = 0.1$ states 1 and 5 are transient states. By inspection states 2, 3 and 4 have period 2, the paths are:

$$\begin{aligned} 2 \rightarrow 3 \rightarrow 2 \\ 2 \rightarrow 4 \rightarrow 2 \end{aligned}$$

As the matrix has got a perod of two two of the eigenvalues are 1 and -1. The complete list is 9.3007e-01, -4.3007e-01, -1.0, 0.0, 1.0 but only two are obtainable with no calculations. [20%]