MML — Exercises — Example Answers

December 11, 2021

Contents

1 Least-squares solution

Exercise: The loss function for linear regression is

$$L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left(y_n - \boldsymbol{\phi}(x_n)^T\boldsymbol{\theta}\right)^2 = ||\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}||^2$$

1. Find the gradient of the loss function with respect to $\boldsymbol{\theta}$.
2. Find $\boldsymbol{\theta}$ for which $L(\boldsymbol{\theta})$ is minimized.
3. Demonstrate $\boldsymbol{\theta}$ is in fact a minimum for $L(\boldsymbol{\theta})$.

Notice $\boldsymbol{\Phi}(X)$ contains all the samples evaluated using the different basis functions used. $\boldsymbol{\Phi}(X)$ is usually denoted as the *design matrix* and for N data points and M basis functions, it represents the following.

$$\boldsymbol{\Phi}(X) = \begin{pmatrix} \boldsymbol{\phi}(x_1)^T \\ \vdots \\ \boldsymbol{\phi}(x_N)^T \end{pmatrix} = \begin{pmatrix} \phi_1(x_1) & \dots & \phi_M(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_M) & \dots & \phi_M(x_M) \end{pmatrix}$$

Solution: Let us compute the gradient with respect to $\boldsymbol{\theta}$. For illustrative purposes, we will use the vector form of the loss function.

$$\begin{aligned} \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\left(||\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}||^2 \right) = \frac{\partial}{\partial \boldsymbol{\theta}}\left(\left(\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}\right)^T \left(\mathbf{y} - \boldsymbol{\Phi}(X)\boldsymbol{\theta}\right) \right) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}}\left(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\theta}^T\boldsymbol{\Phi}(X)^T\mathbf{y} + \boldsymbol{\theta}^T\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta} \right) \\ &= -2\boldsymbol{\Phi}(X)^T\mathbf{y} + 2\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta} \end{aligned}$$

Now, we calculate $\boldsymbol{\theta}^*$ for which the gradient is zero, which should minimize the loss function.

$$\left.\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = 0$$

$$0 = -2\boldsymbol{\Phi}(X)^T\mathbf{y} + 2\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta}^*$$

$$\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta}^* = \boldsymbol{\Phi}(X)^T\mathbf{y}$$

$$\boldsymbol{\theta}^* = \left(\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\right)^{-1}\boldsymbol{\Phi}(X)^T\mathbf{y}$$

Notice that the following quantity

$$\boldsymbol{\Phi}(X)^\dagger = \left(\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\right)^{-1}\boldsymbol{\Phi}(X)^T$$

is commonly known as the *pseudo-inverse* of the matrix $\boldsymbol{\Phi}(X)$, which can be regarded as a generalization for the inverse of a non-square matrix.

Therefore, the $\boldsymbol{\theta}^*$ which minimizes the loss function should be

$$\boldsymbol{\theta}^* = \boldsymbol{\Phi}(X)^\dagger\mathbf{y}$$

We still need to confirm that $\boldsymbol{\theta}^*$ is actually a minimum of $L(\boldsymbol{\theta})$. We can do so by computing the Hessian. We know that if the resulting Hessian is positive definite, $\boldsymbol{\theta}^*$ is a minimum for the least squares problem.

$$\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \frac{\partial}{\partial \boldsymbol{\theta}}\left(-2\boldsymbol{\Phi}(X)^T\mathbf{y} + 2\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)\boldsymbol{\theta} \right) = 2\boldsymbol{\Phi}(X)^T\boldsymbol{\Phi}(X)$$

We need to prove that $\mathbf{\Phi}(X)^T\mathbf{\Phi}(X)$ is positive definite. Recall $\mathbf{\Phi}(X) \in \mathbb{R}^{N\times M}$.

$$\mathbf{\Phi}(X)^T\mathbf{\Phi}(X) \text{ is positive definite } \iff \mathbf{z}^T\mathbf{\Phi}(X)^T\mathbf{\Phi}(X)\mathbf{z} > 0, \quad \forall \mathbf{z} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$$

$$\mathbf{z}^T\mathbf{\Phi}(X)^T\mathbf{\Phi}(X)\mathbf{z} = \Big(\mathbf{\Phi}(X)\mathbf{z}\Big)^T\mathbf{\Phi}(X)\mathbf{z} = ||\mathbf{\Phi}(X)\mathbf{z}||^2 \geq 0$$

$\mathbf{\Phi}(X)^T\mathbf{\Phi}(X)$ is positive definite if we can show that $\mathbf{\Phi}(X)\mathbf{z} \neq 0, \quad \forall \mathbf{z} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$. For that, we make two assumptions which usually hold in the least squares problem. The first one is that $N \geq M$, which means that we have more (or equal) data points than basis functions. The second one is that the design matrix $\mathbf{\Phi}(X)$ is full rank. We can re-arrange the matrix product $\mathbf{\Phi}(X)\mathbf{z}$ as a linear combination of each basis function $\boldsymbol{\phi}_i(\mathbf{x})$, where $\boldsymbol{\phi}_i(\mathbf{x})$ is the i-th basis function evaluated on each data point.

$$\mathbf{\Phi}(X)\mathbf{z} = \begin{pmatrix}\boldsymbol{\phi}_1(\mathbf{x}) & \boldsymbol{\phi}_2(\mathbf{x}) & \dots & \boldsymbol{\phi}_M(\mathbf{x})\end{pmatrix}\begin{pmatrix}z_1\\ z_2\\ \vdots\\ z_M\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix}$$

$$\sum_{i=1}^{M}\boldsymbol{\phi}_i(\mathbf{x})z_i = \mathbf{0}$$

Since $\mathbf{\Phi}(X)$ is full rank and $M \leq N$, all the terms $\boldsymbol{\phi}_i(\mathbf{x})$ are linearly independent and thus, there exists no $\mathbf{z}$ for which we can obtain $\mathbf{0}$ in the previous expression. Consequently, we have that $||\mathbf{\Phi}(X)\mathbf{z}||^2 > 0$ and thus, $\mathbf{\Phi}(X)^T\mathbf{\Phi}(X)$ is positive definite, which means that $\boldsymbol{\theta}^*$ is in fact a minimum for the least squares problem.

2 Exercises chapter 5

In some of the solutions, earlier solutions are re-used. In an exam, you need to ensure to state what identity is used, or you may need to prove sub-results if requested. In an exam, you may refer to earlier derivation in your exam transcript, **but you must do so clearly and unambiguously**, e.g. with an equation number.

5.1.

$$f(x) = \log(x^4)\sin(x^3)$$

$$f'(x) = \frac{\partial \log(x^4)}{\partial x}\sin(x^3) + \log(x^4)\frac{\partial \sin(x^3)}{\partial x}$$

$$f'(x) = \frac{1}{x^4}4x^3\sin(x^3) + \log(x^4)\cos(x^3)3x^2$$

$$f'(x) = \frac{4}{x}\sin(x^3) + 3x^2\log(x^4)\cos(x^3)$$

5.2.

$$f(x) = \frac{1}{1+\exp(-x)}$$

$$f'(x) = \frac{-1}{(1+\exp(-x))^2}\exp(-x)(-1)$$

$$f'(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

5.3.

$$f(x) = \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

$$f'(x) = \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)\big(\frac{-1}{2\sigma^2}2(x-\mu)\big)$$

$$f'(x) = \frac{(\mu - x)}{\sigma^2}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

5.5.

- $f_1(\mathbf{x}) = \sin(x_1)\cos(x_2), \quad \mathbf{x} \in \mathbb{R}^2$

$$\frac{\partial f_1}{\partial \mathbf{x}} \in \mathbb{R}^{1\times 2}$$

$$\begin{aligned} \frac{\partial f_1}{\partial \mathbf{x}} &= \left[\frac{\partial f_1}{\partial x_1}, \frac{\partial f_1}{\partial x_2}\right] \\ &= \Big[\cos(x_1)\cos(x_2), -\sin(x_1)\sin(x_2)\Big] \end{aligned}$$

- $f_2(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\frac{\partial f_2}{\partial \mathbf{x}} \in \mathbb{R}^{1\times n}$$

We can solve this directly using basic rules of vector calculus

$$\frac{\partial f_2}{\partial \mathbf{x}} = \frac{\partial(\mathbf{x}^T\mathbf{y})}{\partial \mathbf{x}} = \mathbf{y}^T$$

We can confirm this result holds by confirming the notation used in the lectures. First, let us calculate the value $f_2(\mathbf{x}, \mathbf{y})$

$$f_2(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

$$\frac{\partial f_2}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n}\frac{\partial x_j}{\partial x_i}y_i = \sum_{i=1}^{n}\delta_{ij}y_i = y_j$$

$$\frac{\partial f_2}{\partial \mathbf{x}} = \left[\frac{\partial f_2}{\partial x_1}, \ldots, \frac{\partial f_2}{\partial x_n}\right] = [y_1, \ldots, y_n] = \mathbf{y}^T$$

- $\mathbf{f}_3(x) = \mathbf{x}\mathbf{x}^T, \quad \mathbf{x} \in \mathbb{R}^n$

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{x}} \in \mathbb{R}^{(n\times n)\times n}$$

$$\frac{\partial \mathbf{f}_3}{\partial \mathbf{x}} = \mathbf{C} \quad \text{where } C \text{ is a 3D tensor.}$$

$$C_{ijk} = \frac{\partial f_3(\mathbf{x})_{ij}}{\partial x_k}$$

$$\mathbf{x}\mathbf{x}^T = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\begin{pmatrix} x_1 & \ldots & x_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1x_2 & \ldots & x_1x_n \\ x_2x_1 & x_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ x_nx_1 & \ldots & \ldots & x_n^2 \end{pmatrix}$$

$$C_{ijk} = \frac{\partial(x_ix_j)}{\partial x_k} = \frac{\partial x_i}{\partial x_k}x_j + \frac{\partial x_j}{\partial x_k}x_i = \delta_{ik}x_j + \delta_{jk}x_i = \begin{cases} 0 & \text{if } k \neq i \text{ and } k \neq j \\ x_i & \text{if } k = j \text{ and } i \neq j \\ x_j & \text{if } k = i \text{ and } i \neq j \\ 2x_i & \text{if } k = i = j \end{cases}$$

5.6.

- $f(\mathbf{t}) = \sin\big(\log(\mathbf{t}^T\mathbf{t})\big) \quad \mathbf{t} \in \mathbb{R}^D$

  We directly apply the chain rule

$$\frac{\partial f}{\partial \mathbf{t}} = \frac{\partial \sin\big(\log(\mathbf{t}^T\mathbf{t})\big)}{\partial \log(\mathbf{t}^T\mathbf{t})} \cdot \frac{\partial \log(\mathbf{t}^T\mathbf{t})}{\partial (\mathbf{t}^T\mathbf{t})} \cdot \frac{\partial (\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}}$$

  All of the terms are one dimensional except for $\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} \in \mathbb{R}^{1\times D}$. Let us calculate the value using the notation for vector calculus in the lectures. As in 5.5, we first calculate the value of $\mathbf{t}^T\mathbf{t}$ and its derivative w.r.t. t_i.

$$\mathbf{t}^T\mathbf{t} = \sum_{i=1}^{D} t_i^2, \quad \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_i} = 2t_i$$

$$\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial \mathbf{t}} = \left[\frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_1}, \ldots, \frac{\partial(\mathbf{t}^T\mathbf{t})}{\partial t_D}\right] = [2t_1 \ldots, 2t_D] = 2\mathbf{t}^T$$

  We can now use this result to proceed with the derivative of $f(\mathbf{t})$.

$$\frac{\partial f}{\partial \mathbf{t}} = \cos\big(\log(\mathbf{t}^T\mathbf{t})\big) \cdot \frac{1}{\mathbf{t}^T\mathbf{t}} \cdot 2\mathbf{t}^T$$

$$\frac{\partial f}{\partial \mathbf{t}} = 2\mathbf{t}^T \frac{\cos\big(\log(\mathbf{t}^T\mathbf{t})\big)}{\mathbf{t}^T\mathbf{t}}$$

- $g(\mathbf{X}) = tr(\mathbf{AXB}), \quad \mathbf{A} \in \mathbb{R}^{D\times E}, \mathbf{X} \in \mathbb{R}^{E\times F}, \mathbf{B} \in \mathbb{R}^{F\times D}$

  Eq. 101 from Matrix cookbook gives us the direct result.

$$\frac{\partial g}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{B}^T$$

  Alternative proof: Use index notation introduced in the course.

$$g(\mathbf{X}) = tr(\mathbf{AXB}) = \sum_{i=1}^{D} (\mathbf{AXB})_{ii}$$

  In order to fully compute $g(\mathbf{X})$, we need to calculate $(\mathbf{AXB})_{ii}$

$$(\mathbf{AXB})_{ii} = \sum_{k=1}^{F} (\mathbf{AX})_{ik} b_{ki} = \sum_{k=1}^{F} \left(\sum_{l=1}^{E} a_{il} x_{lk}\right) b_{ki}$$

  Thus

$$g(\mathbf{X}) = \sum_{i=1}^{D}\sum_{k=1}^{F}\sum_{l=1}^{E} a_{il} x_{lk} b_{ki}$$

  Now we can just calculate the derivative using index notation

$$\frac{\partial g}{\partial x_{nm}} = \frac{\partial}{\partial x_{nm}} \sum_{i=1}^{D}\sum_{k=1}^{F}\sum_{l=1}^{E} a_{il} x_{lk} b_{ki} = \sum_{i=1}^{D}\sum_{k=1}^{F}\sum_{l=1}^{E} a_{il} \frac{\partial x_{lk}}{\partial x_{nm}} b_{ki} = \sum_{i=1}^{D}\sum_{k=1}^{F}\sum_{l=1}^{E} a_{il} \delta_{ln} \delta_{km} b_{ki}$$

  Notice that in the last expression, all the terms in the summation cancel except when $k = m$ and $l = n$. Therefore

$$\frac{\partial g}{\partial x_{nm}} = \sum_{i=1}^{D} a_{in} b_{mi} = \sum_{i=1}^{D} b_{mi} a_{in} = (\mathbf{BA})_{mn}$$

Using this last result, we can calculate the derivative w.r.t. **X**.

$$\frac{\partial g}{\partial \mathbf{X}} = (\mathbf{BA})^T = \mathbf{A}^T\mathbf{B}^T$$

Alternative proof 2: Use properties 4.19 and 5.100 from the course book. From 4.19

$$g(\mathbf{X}) = tr(\mathbf{AXB}) = tr(\mathbf{XBA}) = tr(\mathbf{XC}), \quad \mathbf{C} = \mathbf{BA}$$

and from 5.100

$$\frac{\partial g}{\partial \mathbf{X}} = \frac{\partial tr(\mathbf{XC})}{\partial \mathbf{X}} = tr\left(\frac{\partial(\mathbf{XC})}{\partial \mathbf{X}}\right), \quad \text{where } \frac{\partial(\mathbf{XC})}{\partial \mathbf{X}} \in \mathbb{R}^{(E\times E)\times(E\times F)}$$

We need to calculate $\frac{\partial(\mathbf{XC})_{ij}}{\partial X_{kl}}$, and we find convenient to write the pairs i, j of the product $\mathbb{I}\mathbf{XC}$, where $\mathbb{I} \in \mathbb{R}^{E\times E}$ is the identity matrix.

$$(\mathbb{I}\mathbf{XC})_{ij} = \sum_{e=1}^{E}\sum_{f=1}^{F} \delta_{ie}x_{ef}c_{fj}$$

$$\frac{\partial(\mathbf{XC})_{ij}}{\partial X_{kl}} = \frac{\partial(\mathbb{I}\mathbf{XC})_{ij}}{\partial X_{kl}} = \delta_{ik}c_{lj}$$

in the prevous expression, all the terms in the sum vanish except the ones that contain x_{kl} in it.

Now, we take into account the definition of the trace for any 4D tensor $\mathbf{A} \in \mathbb{R}^{(N\times N)\times(P\times Q)}$ given in the course book:

$$tr(\mathbf{A})_{ij} = \sum_{k=1}^{N} a_{kkij}, \quad \text{where } tr(\mathbf{A}) \in \mathbb{R}^{P\times Q}$$

We use this definition to calculate our result.

$$tr\left(\frac{\partial(\mathbf{XC})}{\partial \mathbf{X}}\right)_{ij} = \sum_{k=1}^{E} \frac{\partial(\mathbf{XC})_{kk}}{\partial X_{ij}} = \sum_{k=1}^{E} \delta_{ki}c_{jk} = c_{ji}$$

all the terms will be 0 except when $k = i$.

$$tr\left(\frac{\partial(\mathbf{XC})}{\partial \mathbf{X}}\right) = \mathbf{C}^T = (\mathbf{BA})^T = \mathbf{A}^T\mathbf{B}^T$$

5.7.

a. $f(z) = \log(1+z), \quad z = \mathbf{x}^T\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^D$

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial \mathbf{x}} = \frac{\partial \log(1+z)}{\partial z}\frac{\partial(\mathbf{x}^T\mathbf{x})}{\partial \mathbf{x}} = \frac{2\mathbf{x}^T}{1+z} = \frac{2\mathbf{x}^T}{1+\mathbf{x}^T\mathbf{x}}$$

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^D, \quad \frac{\partial f}{\partial z} \in \mathbb{R}, \quad \frac{\partial z}{\partial \mathbf{x}} \in \mathbb{R}^D$$

b. $f(\mathbf{z}) = \sin(\mathbf{z}), \quad \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}, \quad \mathbf{A} \in \mathbb{R}^{E \times D}, \mathbf{x} \in \mathbb{R}^D, \mathbf{b} \in \mathbb{R}^E$

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \sin(\mathbf{z})}{\partial \mathbf{z}} \frac{\partial (\mathbf{A}\mathbf{x} + \mathbf{b})}{\partial \mathbf{x}}$$

Notice that $\frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{E \times E}$. We already know that $\sin(\cdot)$ is applied to each element independently, thus

$$\frac{\partial f_i}{\partial z_j} = \begin{cases} 0 & \text{if } i \neq j \\ \cos(z_i) & \text{if } i = j \end{cases}$$

We also have $\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{E \times D}$. Although this has already shown in the lectures, let us review the result $\frac{\partial \mathbf{z}}{\partial \mathbf{x}}$ using the notation of the course.

$$z_i = \sum_{j=1} A_{ij} x_j + b_i$$

We can now easily compute $\frac{\partial z_i}{\partial x_j}$

$$\frac{\partial z_i}{\partial x_j} = A_{ij}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \mathbf{A}$$

Let us use all the previous results to compute the derivative of $f(\mathbf{z})$ w.r.t. $\mathbf{x}$.

$$\frac{\partial f}{\partial \mathbf{x}} = diag(\cos(\mathbf{z}))\mathbf{A}, \quad \text{where } diag(\mathbf{a}) = \begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & a_N \end{pmatrix}, \quad \mathbf{a} \in \mathbb{R}^N$$

$$\frac{\partial f}{\partial \mathbf{x}} = diag(\cos(\mathbf{A}\mathbf{x} + \mathbf{b}))\mathbf{A}$$

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{E \times D}, \quad \frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{E \times E}, \quad \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \in \mathbb{R}^{E \times D}$$

5.8.

a. $f(z) = \exp(-\frac{1}{2}z), \quad z = g(\mathbf{y}) = \mathbf{y}^T \mathbf{S}^{-1} \mathbf{y}, \quad \mathbf{y} = h(\mathbf{x}) = \mathbf{x} - \boldsymbol{\mu}, \quad \mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^D, \mathbf{S} \in \mathbb{R}^{D \times D}$

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial z} \frac{\partial z}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \frac{\partial \exp(-\frac{1}{2}z)}{\partial z} \frac{\partial (\mathbf{y}^T \mathbf{S}^{-1} \mathbf{y})}{\partial \mathbf{y}} \frac{\partial (\mathbf{x} - \boldsymbol{\mu})}{\partial \mathbf{x}} = \exp\left(-\frac{1}{2}z\right)\left(-\frac{1}{2}\right)\mathbf{y}^T(\mathbf{S}^T + \mathbf{S}^{-T})\mathbb{I} \\ &= -\frac{1}{2}\exp\left(-\frac{1}{2}\Big((\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})\Big)\right)(\mathbf{x} - \boldsymbol{\mu})^T(\mathbf{S}^T + \mathbf{S}^{-T}) \end{aligned}$$

where $\mathbf{S}^{-T} = \left(\mathbf{S}^{-1}\right)^T$, and we use (5.107) to calculate $\frac{\partial (\mathbf{y}^T \mathbf{S}^{-1} \mathbf{y})}{\partial \mathbf{y}}$.

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^D, \quad \frac{\partial f}{\partial z} \in \mathbb{R}, \quad \frac{\partial z}{\partial \mathbf{y}} \in \mathbb{R}^D, \quad \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \in \mathbb{R}^{D \times D}$$

b. $f(\mathbf{x}) = tr(\mathbf{x}\mathbf{x}^T + \sigma^2\mathbb{I}), \quad \mathbf{x} \in \mathbb{R}^D$

Let us expand $f(x)$.

$$\begin{aligned} f(x) &= \sum_{i=1}^{D} \left((\mathbf{x}\mathbf{x}^T)_{ii} + \sigma^2 \right) \\ &= \sum_{i=1}^{D} (\mathbf{x}\mathbf{x}^T)_{ii} + D\sigma^2 = \sum_{i=1}^{D} x_i^2 + D\sigma^2 \end{aligned}$$

From previous exercices, we know that $(\mathbf{x}\mathbf{x}^T)_{ij} = x_i x_j$. Therefore

$$\frac{\partial f}{\partial \mathbf{x}} = \frac{\partial \left(\sum_{i=1}^{D} x_i^2 + D\sigma^2 \right)}{\partial \mathbf{x}} = 2\mathbf{x}^T$$

c. $f(\mathbf{z}) = \tanh(\mathbf{z}) \in \mathbb{R}^M, \quad \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}, \quad \mathbf{x} \in \mathbb{R}^N, \mathbf{A} \in \mathbb{R}^{M \times N}, \mathbf{b} \in \mathbb{R}^M$

$$\begin{aligned} \frac{\partial f}{\partial \mathbf{x}} &= \frac{\partial f}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \tanh(\mathbf{z})}{\partial \mathbf{z}} \frac{\partial (\mathbf{A}\mathbf{x} + \mathbf{b})}{\partial \mathbf{x}} = diag\left(1 - \tanh^2(\mathbf{z})\right) \mathbf{A} \\ &= diag\left(1 - \tanh^2(\mathbf{A}\mathbf{x} + \mathbf{b})\right) \mathbf{A} \end{aligned}$$

where we used $\frac{d \tanh(a)}{da} = 1 - \tanh^2(a)$.

Dimensions are

$$\frac{\partial f}{\partial \mathbf{x}} \in \mathbb{R}^{M \times N}, \quad \frac{\partial f}{\partial \mathbf{z}} \in \mathbb{R}^{M \times M}, \quad \frac{\partial z}{\partial \mathbf{x}} \in \mathbb{R}^{M \times N}$$

5.9.

$$g(\mathbf{z}, \boldsymbol{\nu}) := \log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}, \boldsymbol{\nu}), \quad \mathbf{z} = t(\boldsymbol{\epsilon}, \boldsymbol{\nu})$$

We apply the chain rule straightforwardly.

$$\begin{aligned} \frac{d}{d\boldsymbol{\nu}} g(\mathbf{z}, \boldsymbol{\nu}) &= \frac{\partial g(\mathbf{z}, \boldsymbol{\nu})}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \boldsymbol{\nu}} + \frac{\partial g(\mathbf{z}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \\ &= \frac{\partial \left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}, \boldsymbol{\nu}) \right)}{\partial \mathbf{z}} \frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} + \frac{\partial \left(\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z}, \boldsymbol{\nu}) \right)}{\partial \boldsymbol{\nu}} \\ &= \left(\frac{1}{p(\mathbf{x}, \mathbf{z})} \frac{\partial p(\mathbf{x}, \mathbf{z})}{\partial \mathbf{z}} - \frac{1}{q(\mathbf{z}, \boldsymbol{\nu})} \frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \mathbf{z}} \right) \frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} - \frac{1}{q(\mathbf{z}, \boldsymbol{\nu})} \frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \\ &= \frac{1}{p(\mathbf{x}, \mathbf{z})} \frac{\partial p(\mathbf{x}, \mathbf{z})}{\partial \mathbf{z}} \frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} - \frac{1}{q(\mathbf{z}, \boldsymbol{\nu})} \left(\frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \mathbf{z}} \frac{\partial t(\boldsymbol{\epsilon}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} + \frac{\partial q(\mathbf{z}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} \right) \end{aligned}$$

3 Sets and probabilities

Let us denote the three axioms of probability theory:

1. The probability of an event E is a non-negative number.

   $$E \in \mathbb{R}, \quad P(E) \geq 0$$

2. The probability that at least one of the events will occur is 1.

   $$P(\Omega) = 1$$

3. Any countable sequence of disjoint sets $E_1, E_2, \dots$ satisfies

   $$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) \quad \text{when } E_i \cap E_j = \varnothing, \quad \forall i, j, \quad i \neq j$$

Demonstrate the following properties.

a. $P(\neg A) = 1 - P(A)$

b. $P(\varnothing) = 0$, where $\varnothing$ is the empty set

c. $0 \leq P(A) \leq 1$

d. $A \subseteq B \implies P(A) \leq P(B)$
   Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

f. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_i \leq A_{i+1} \forall i$ then:

   $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \longrightarrow \infty} P(A_i)$$

   Hint: Use axiom 3.

   *: This question is beyond the course content.

Solution:

a. $P(\neg A) = 1 - P(A)$

   Let us consider a collection of events contained in the sample space $\{A_1, \dots, A_N\} \subseteq \Omega$. Let us select the first i events (where $i \leq N$) and denote them as A. The rest of them will be the complementary set, denoted as $\neg A$.

   $$A = \{A_1, \dots, A_i\} \qquad \neg A = \{A_{i+1}, \dots, A_N\} \qquad A \cap \neg A = \varnothing$$

   Using axiom (2), we have

   $$P(A_1 \cup A_2 \cup \cdots \cup A_N) = P(\Omega) = 1$$

   And using axiom (3), we have

   $$P(A_1, \dots, A_i) + P(A_{i+1}, \dots, A_N) = P(A_1 \cup A_2 \cup \cdots \cup A_N)$$

   $$P(A) + P(\neg A) = 1$$

   Thus

   $$P(\neg A) = 1 - P(A)$$

b. $P(\varnothing) = 0$, where $\varnothing$ is the empty set

We can just consider the sample space, Ω, where its complementary is the empty set $\varnothing$. Using the previous property and axiom 2, we have.

$$P(\Omega) = 1$$

$$P(\varnothing) = P(\neg\Omega) = 1 - P(\Omega) = 1 - 1 = 0$$

$$P(\varnothing) = 0$$

c. $0 \leq P(A) \leq 1$

Here we can also use property (a) and the first axiom. Consider any event A.

$$P(A) \geq 0$$

$$P(\neg A) = 1 - P(A) \geq 0$$

$$1 \geq P(A)$$

We can join the previous inequalities and obtain the following.

$$0 \leq P(A) \leq 1$$

d. $A \subseteq B \implies P(A) \leq P(B)$

Hint: Consider the following definition. $B \backslash A = \{x \in B : x \notin A\}$

We can construct B as the union of two disjoint sets.

$$B = B \backslash A \cup A$$

where $B \backslash A \cap A = \varnothing$ by definition of $B \backslash A$. Let us use axiom 3 and 1.

$$P(B) = P(B \backslash A) + P(A) \geq P(A)$$

where by axiom 1, we have $P(B \backslash A) \geq 0$. Thus

$$P(A) \leq P(B)$$

e. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Let us define the union $(A \cup B)$ in terms of two disjoint sets.

$$(A \cup B) = A \cup B \backslash A$$

where $A \cap B \backslash A = \varnothing$ by definition. Using axiom 3 we have

$$P(A \cup B) = P(A) + P(B \backslash A)$$

To calculate $P(B \backslash A)$, let us define B in terms of A, and the union of two disjoint sets.

$$B = (B \cap A) \cup (B \backslash A)$$

where $(B \cap A) \cap (B \backslash A) = \varnothing$ by definition. Using also axiom 3, we have.

$$P(B) = P(B \cap A) + P(B \backslash A)$$

$$P(B \backslash A) = P(B) - P(B \cap A)$$

Therefore, the probability of $(A \cup B)$ is the following

$$P(A \cup B) = P(A) + P(B \backslash A) = P(A) + P(B) - P(B \cap A)$$

f. (*) if $\{A_i\}_{i=1}^{\infty} \subseteq \Omega$ and $A_{i-1} \subseteq A_i \quad \forall i > 0$ then:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \longrightarrow \infty} P(A_i)$$

Hint: Use axiom 3.

Let us define the following

$$A := \bigcup_{i=1}^{\infty} A_i$$

We would like to write A in terms of disjoint sets so as to use axiom 3.

$$A_{i-1} \subseteq A_i \quad \forall i > 0 \implies A = \bigcup_{i=1}^{\infty} A_i \backslash A_{i-1}$$

The previous expression holds if we have $A_0 = \varnothing$. Notice this new expression can be regarded as starting with A_1 and adding the information form $A_2, A_3, \dots$ which is not previously considered (e.g $A_2 \backslash A_1, A_3 \backslash A_2, \dots$). Since this construction is a union of disjoint sets, we now can use axiom 3.

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A_i \backslash A_{i-1}\right) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1})$$

The infinite summation is in fact defined as a limit.

$$P(A) = \sum_{i=1}^{\infty} P(A_i \backslash A_{i-1}) = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} P(A_i \backslash A_{i-1})$$

Notice the result in exercise (d), where we obtained $P(B) = P(B \backslash A) + P(A)$ for $A \subseteq B$. Therefore

$$P(A_i) = P(A_i \backslash A_{i-1}) + P(A_{i-1})$$

$$P(A_i \backslash A_{i-1}) = P(A_i) - P(A_{i-1})$$

$$P(A) = \lim_{n \longrightarrow \infty} \sum_{i=1}^{n} P(A_i) - P(A_{i-1}) = \lim_{n \longrightarrow \infty} \left(\sum_{i=1}^{n} P(A_i) - \sum_{i=1}^{n-1} P(A_i)\right) = \lim_{n \longrightarrow \infty} P(A_n)$$

where we used $P(A_0) = P(\varnothing) = 0$. In conclusion,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P(A) = \lim_{i \longrightarrow \infty} P(A_i)$$

4 Lecture 'Moments' exercise solutions

- $\mathbb{E}_{X,Y}[X+Y] = \mathbb{E}_X[X] + \mathbb{E}_Y[Y]$

$$\begin{aligned}
\mathbb{E}_{X,Y}[X+Y] &= \iint_{x,y\in\mathbb{R}} (x+y)p(x,y)dxdy \\
&= \iint_{x,y\in\mathbb{R}} xp(x,y)dxdy + \iint_{x,y\in\mathbb{R}} yp(x,y)dxdy \\
&= \int_{x\in\mathbb{R}} xp(x)dx + \int_{y\in\mathbb{R}} yp(y)dy \\
&= \mathbb{E}_X[X] + \mathbb{E}_Y[Y]
\end{aligned}$$

- $\mathbb{V}_X[X] = \mathbb{E}_X[X^2] - (\mathbb{E}_X[X])^2$

$$\begin{aligned}
\mathbb{V}_X[X] &= \mathbb{E}_X\Big[\big(X - \mathbb{E}_X[X]\big)^2\Big] \\
&= \mathbb{E}_X\Big[X^2 - 2\mathbb{E}_X[X]X + \big(\mathbb{E}_X[X]\big)^2\Big] \\
&= \mathbb{E}_X[X^2] - 2\mathbb{E}_X[X]\mathbb{E}_X[X] + \big(\mathbb{E}_X[X]\big)^2 = \\
&= \mathbb{E}_X[X^2] - \big(\mathbb{E}_X[X]\big)^2
\end{aligned}$$

- $\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y], \quad$ if Y and X are independent.

$$\begin{aligned}
\mathbb{V}_{X,Y}[X+Y] &= \mathbb{E}_{X,Y}\Big[\big(X+Y-\mathbb{E}_{X,Y}[X+Y]\big)^2\Big] = \mathbb{E}_{X,Y}\Big[\Big((X+Y) - \big(\mathbb{E}_X[X] + \mathbb{E}_Y[Y]\big)\Big)^2\Big] \\
&= \mathbb{E}_{X,Y}\Big[\Big(\big(X - \mathbb{E}_X[X]\big) + \big(Y - \mathbb{E}_Y[Y]\big)\Big)^2\Big] \\
&= \mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)^2 + \big(Y - \mathbb{E}_Y[Y]\big)^2\Big] + 2\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big] \\
&= \mathbb{E}_X\Big[\big(X - \mathbb{E}_X[X]\big)^2\Big] + \mathbb{E}_Y\Big[\big(Y - \mathbb{E}_Y[Y]\big)^2\Big] + 2\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big] \\
&= \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big]
\end{aligned}$$

$\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big]$ is the covariance between x and y. Let us show that it is 0 for independent variables, that is, for $p(x,y) = p(x)p(y)$.

$$\begin{aligned}
Cov[x,y] &= \mathbb{E}_{X,Y}\Big[XY + \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_X[X]Y - \mathbb{E}_Y[Y]X\Big] \\
&= \mathbb{E}_{X,Y}[XY] + \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_Y[Y]\mathbb{E}_X[X] \\
&= \mathbb{E}_{X,Y}[XY] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \iint_{x,y\mathbb{R}} xyp(x,y)dxdy - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \int_{x\in\mathbb{R}} xp(x)dx \int_{y\in\mathbb{R}} yp(y)dy - \mathbb{E}_X[X]\mathbb{E}_Y[Y] \\
&= \mathbb{E}_X[X]\mathbb{E}_Y[Y] - \mathbb{E}_X[X]\mathbb{E}_Y[Y] = 0
\end{aligned}$$

Notice this is true since we used $p(x,y) = p(x)p(y)$. Therefore, the previous identity holds.

$$\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y]$$

- $\mathbb{V}_{X,Y}[X-Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y], \quad$ if Y and X are independent.

$$\begin{aligned}\mathbb{V}_{X,Y}[X-Y] &= \mathbb{E}_{X,Y}\Big[\big(X - Y - \mathbb{E}_{X,Y}[X-Y]\big)^2\Big] = \mathbb{E}_{X,Y}\bigg[\Big((X-Y) - \big(\mathbb{E}_X[X] - \mathbb{E}_Y[Y]\big)\Big)^2\bigg] \\ &= \mathbb{E}_{X,Y}\bigg[\Big(\big(X - \mathbb{E}_X[X]\big) - \big(Y - \mathbb{E}_Y[Y]\big)\Big)^2\bigg] \\ &= \mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)^2 + \big(Y - \mathbb{E}_Y[Y]\big)^2\Big] - 2\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big] \\ &= \mathbb{E}_X\Big[\big(X - \mathbb{E}_X[X]\big)^2\Big] + \mathbb{E}_Y\Big[\big(Y - \mathbb{E}_Y[Y]\big)^2\Big] \\ &= \mathbb{V}_X[X] + \mathbb{V}_Y[Y]\end{aligned}$$

We already showed that $\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big] = 0$ if x and y are independent.

- $\mathbb{V}_X[cX] = c^2\mathbb{V}_X[X]$

$$\begin{aligned}\mathbb{V}_X[cX] &= \mathbb{E}_X\Big[\big(cX - \mathbb{E}_X[cX]\big)^2\Big] \\ &= \mathbb{E}_X\Big[c^2X^2 - 2cX\mathbb{E}_X[cX] + \big(\mathbb{E}_X[cX]\big)^2\Big] \\ &= \mathbb{E}_X\Big[c^2X^2 - 2c^2X\mathbb{E}_X[X] + c^2\big(\mathbb{E}_X[X]\big)^2\Big] \\ &= c^2\mathbb{E}_X\Big[X^2 - 2X\mathbb{E}_X[X] + \big(\mathbb{E}_X[X]\big)^2\Big] \\ &= c^2\mathbb{E}_X\Big[\big(X - \mathbb{E}_X[X]\big)^2\Big] \\ &= c^2\mathbb{V}_X[X]\end{aligned}$$

- $\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2Cov[X,Y]$

  We have already proven that this expression holds.

$$\mathbb{V}_{X,Y}[X+Y] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2\mathbb{E}_{X,Y}\Big[\big(X - \mathbb{E}_X[X]\big)\big(Y - \mathbb{E}_Y[Y]\big)\Big] = \mathbb{V}_X[X] + \mathbb{V}_Y[Y] + 2Cov[X,Y]$$

- Show that sampling with replacement gives an unbiased gradient estimator.

  Let us consider $f(x)$, where x is a random variable, where $x \sim X$. We would like to calculate the gradient of the following quantity.

$$\nabla\mathbb{E}_x[f(x)]$$

  Since this might yield complex analytical calculation, we define an estimator which calculates the previous value using sampling with replacement. In other words, we consider $\{x_1, \ldots, x_N\}$ which are i.i.d. and sampled according to x, $x_i \sim X$. The estimator is defined as follows.

$$\nabla\mathbb{E}_x[f(x)] \approx \nabla\frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

Let us prove that this estimator is unbiased.

$$\begin{aligned}
&\mathbb{E}_{x_1,\dots,x_N}\left[\nabla\frac{1}{N}\sum_{i=1}^{N}f(x_i)\right]-\nabla\mathbb{E}_x[f(x)], && \mathbb{E}_X[\nabla X]=\nabla\mathbb{E}_X[X]\\
=&\nabla\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{x_i}[f(x_i)]-\nabla\mathbb{E}_x[f(x)], && \mathbb{E}_{X,Y}[X+Y]=\mathbb{E}_X[X]+\mathbb{E}_Y[Y]\\
=&\nabla\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_x[f(x)]-\nabla\mathbb{E}_x[f(x)], && x_i\sim X, x\sim X \implies \mathbb{E}_{x_i}[f(x_i)]=\mathbb{E}_x[f(x)]\\
=&\nabla\frac{N}{N}\mathbb{E}_x[f(x)]-\nabla\mathbb{E}_x[f(x)]=0
\end{aligned}$$

Thus, the estimator is unbiased.

- In terms of v, the variance of a gradient estimator using a single element in the minibatch, compute the variance for a minibatch of size M.

  Let us consider a single element in the minibatch $\{x_1\}$.

  $$v=\mathbb{V}_{x_1}\big[\nabla f(x_1)\big]$$

  Let us now compute the variance for a minibatch of size M, with $\{x_1,\dots,x_M\}$ being i.i.d. and sampled from X, as before $x_i\sim X$.

  $$\begin{aligned}
\mathbb{V}_{x_1,\dots,x_M}\left[\nabla\frac{1}{M}\sum_{i=1}^{M}f(x_i)\right]&=\frac{1}{M^2}\mathbb{V}_{x_1,\dots,x_M}\left[\nabla\sum_{i=1}^{M}f(x_i)\right], \qquad \mathbb{V}_X[cX]=c^2\mathbb{V}_X[X]\\
&=\frac{1}{M^2}\mathbb{V}_{x_1,\dots,x_M}\left[\sum_{i=1}^{M}\nabla f(x_i)\right]\\
&=\frac{1}{M^2}\sum_{i=1}^{M}\mathbb{V}_{x_i}\big[\nabla f(x_i)\big], \quad X,Y \text{ ind.} \implies \mathbb{V}_{X,Y}[X+Y]=\mathbb{V}_X[X]+\mathbb{V}_Y[Y]\\
&=\frac{1}{M^2}\sum_{i=1}^{M}v=\frac{M}{M^2}v=\frac{v}{M}
\end{aligned}$$

  Thus

  $$\mathbb{V}_{x_1,\dots,x_M}\left[\nabla\frac{1}{M}\sum_{i=1}^{M}f(x_i)\right]=\frac{v}{M}$$

  The variance of the estimator decreases with a factor of M for increasing sizes of minibatch.

5 Exercises chapter 6

6.1

$$p(x) = \sum_{y \in Y} p(x,y) = \begin{bmatrix} 0.01 + 0.05 + 0.1 \\ 0.02 + 0.1 + 0.05 \\ 0.03 + 0.05 + 0.03 \\ 0.1 + 0.07 + 0.05 \\ 0.1 + 0.2 + 0.04 \end{bmatrix} = \begin{bmatrix} 0.16 \\ 0.17 \\ 0.11 \\ 0.22 \\ 0.34 \end{bmatrix}$$

$$p(y) = \sum_{x \in X} p(x,y) = \begin{bmatrix} 0.01 + 0.02 + 0.03 + 0.1 + 0.1 \\ 0.05 + 0.1 + 0.05 + 0.07 + 0.2 \\ 0.1 + 0.05 + 0.03 + 0.05 + 0.04 \end{bmatrix} = \begin{bmatrix} 0.26 \\ 0.47 \\ 0.27 \end{bmatrix}$$

$$p(x|Y = y_1) = \frac{p(x, Y = y_1)}{p(Y = y_1)} = \frac{1}{0.26}\begin{bmatrix}0.01, & 0.02, & 0.03, & 0.1, & 0.1\end{bmatrix}$$

$$p(x|Y = y_1) \approx \begin{bmatrix}0.038, & 0.077, & 0.115, & 0.385, & 0.385\end{bmatrix}$$

$$p(y|X = x_3) = \frac{p(y, X = x_3)}{p(X = x_3)} = \frac{1}{0.11}\begin{bmatrix}0.03, & 0.05, & 0.03\end{bmatrix} \approx \begin{bmatrix}0.273, & 0.273, & 0.454\end{bmatrix}$$

6.4
Bag 1: 4 mangos, 2 apples.

$$p(mango|heads) = \frac{2}{3} \quad p(apple|heads) = \frac{1}{3}$$

Bag 2: 4 mangos, 4 apples.

$$p(mango|tails) = \frac{1}{2} \quad p(apple|tails) = \frac{1}{2}$$

Heads and tails distrib. is:

$$p(heads) = 0.6 \quad p(tails) = 0.4$$

Mango is presented as evidence. Therefore, we apply Bayes' rule to infer the probability that we picked the mango from bag 2. This is equivalent to computing the posterior distribution of obtaining *tails* given that a mango is taken.

$$p(tails|mango) = \frac{p(tails, mango)}{p(mango)}$$

$$p(tails, mango) = p(mango|tails)p(tails) = 0.5 \cdot 0.4 = 0.2$$

$$p(heads, mango) = p(mango|heads)p(heads) = \frac{2}{3} \cdot 0.6 = 0.4$$

$$p(mango) = p(tails, mango) + p(heads, mango) = 0.2 + 0.4$$

Finally

$$p(tails|mango) = \frac{p(tails, mango)}{p(mango)} = \frac{0.2}{0.2 + 0.4} = \frac{1}{3}$$

The probability of taking a mango from bag 2 is 0.333.

Note on why we need $p(tails|mango)$ and not $p(tails, mango)$: The answer is simple. Since we know that the outcome of the experiment is a *mango*, we consider it as evidence. Therefore, the posterior is the answer to this question.

6.6

Let us prove (6.44): $\mathbb{V}_X[X] = \mathbb{E}_X[X^2] - \big(\mathbb{E}_X[X]\big)^2$. This is already solved above in this document.

$$\begin{aligned}
\mathbb{V}_X[X] &= \mathbb{E}_X\left[\big(X - \mathbb{E}_X[X]\big)^2\right] \\
&= \mathbb{E}_X\left[X^2 - 2\mathbb{E}_X[X]X + \big(\mathbb{E}_X[X]\big)^2\right] \\
&= \mathbb{E}_X[X^2] - 2\mathbb{E}_X[X]\mathbb{E}_X[X] + \big(\mathbb{E}_X[X]\big)^2 = \\
&= \mathbb{E}_X[X^2] - \big(\mathbb{E}_X[X]\big)^2
\end{aligned}$$

6.11

Let us consider the random variables x, y with joint distribution $p(x, y)$.

$$\begin{aligned}
\mathbb{E}_X[x] &= \mathbb{E}_Y\Big[\mathbb{E}_X[x|y]\Big] \\
&= \int_{y\in\mathbb{R}} \mathbb{E}_X[x|y]p(y)dy \\
&= \int_{y\in\mathbb{R}}\int_{x\in\mathbb{R}} xp(x|y)p(y)dxdy \\
&= \int_{x\in\mathbb{R}} x\left(\int_{y\in\mathbb{R}} p(x,y)dy\right)dx \\
&= \int_{x\in\mathbb{R}} xp(x)dx = \mathbb{E}_X[x]
\end{aligned}$$

Notice that we just swapped the integrals and marginalized y over the joint distribution to obtain $p(x)$.

6 Week 3 exercise

Show that solving linear regression using gradient descent with momentum, if converges, converges to $\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$.

Solution First note that $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y})$ The update equations for both the parameter and the momentum are

$$\begin{aligned} \boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \gamma\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) + \alpha\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}) + \alpha\Delta\boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t = \alpha\Delta\boldsymbol{\theta}_t - \frac{\gamma}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}). \end{aligned} \tag{1}$$

Now collecting both equations together into a “joint” linear equation:

$$\begin{bmatrix} \boldsymbol{\theta}_{t+1} \\ \Delta\boldsymbol{\theta}_{t+1} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \\ -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_t \end{bmatrix} + \begin{bmatrix} \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \\ \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y} \end{bmatrix}. \tag{2}$$

Then we can apply the derivation of arithmetico–geometric sequences above, and show that

$$\begin{bmatrix} \boldsymbol{\theta}_t \\ \Delta\boldsymbol{\theta}_t \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \\ -\frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X} & \alpha\mathbf{I} \end{bmatrix}^t \begin{bmatrix} \boldsymbol{\theta}_0 - \boldsymbol{\theta}^* \\ \Delta\boldsymbol{\theta}_0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\theta}^* \\ \mathbf{0} \end{bmatrix}, \tag{3}$$

with $\boldsymbol{\theta}^* = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. This equation also says if momentum GD converges, the momentum $\Delta\boldsymbol{\theta}_t$ will vanish to $\mathbf{0}$, which is as expected as $\Delta\boldsymbol{\theta}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} \to \mathbf{0}$.

7 Week 4 exercise

Q1: Convergence analysis of constant step-size gradient descent (GD) for ridge regression:

1. Show that if GD converges, it would converge to $\boldsymbol{\theta}_R^*$.
2. Derive the "safe threshold" for the constant step size γ.

Solution of Q1:

The iterative update of GD for ridge regression is:

$$\boldsymbol{\theta}_{t+1} = ((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y}. \tag{4}$$

Solving the corresponding geometric sequence returns

$$\boldsymbol{\theta}_t = ((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_R^*) + \boldsymbol{\theta}_R^*, \tag{5}$$

where $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ is the minimiser of the loss function. Therefore it means GD, if converges, converges to the right solution. And GD converges if $((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^t(\boldsymbol{\theta}_0 - \boldsymbol{\theta}_R^*) \to \mathbf{0}$.

Applying the analysis techniques of GD for linear regression, we see that it reduces to investigate the eigenvalues of matrix $((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^2$. Therefore we would like to make sure that

$$\lambda_{max} := \lambda_{max}(((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})^2) = \max_{\lambda_x}(1-\gamma\lambda - \frac{\gamma}{\sigma^2}\lambda_x)^2 < 1, \tag{6}$$

where λ_x denotes possible eigenvalue of $\mathbf{X}^\top\mathbf{X}$. Therefore the "safe threshold" for step size selection is

$$\gamma < 2(\lambda + \lambda_{max}(\mathbf{X}^\top\mathbf{X})/\sigma^2)^{-1}. \tag{7}$$

8 Exercises chapter 4

4.1.

$$\mathbf{A} = \begin{pmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \\ 0 & 2 & 4 \end{pmatrix}$$

Sarrus rule. We apply the formula from the book:

$$\begin{aligned} \det(\mathbf{A}) &= 1 \cdot 4 \cdot 4 + 3 \cdot 6 \cdot 0 + 5 \cdot 2 \cdot 2 - 5 \cdot 4 \cdot 0 - 3 \cdot 2 \cdot 4 - 1 \cdot 6 \cdot 2 \\ &= 16 + 0 + 20 - 0 - 24 - 12 = 36 - 36 = 0 \\ &= 0 \end{aligned}$$

Laplace expansion. We also apply the formula from the book:

$$\begin{aligned} \det(\mathbf{A}) &= (-1)^{1+1} \cdot 1 \cdot \begin{vmatrix} 4 & 6 \\ 2 & 4 \end{vmatrix} + (-1)^{1+2} \cdot 3 \cdot \begin{vmatrix} 2 & 6 \\ 0 & 4 \end{vmatrix} + (-1)^{1+3} \cdot 5 \cdot \begin{vmatrix} 2 & 4 \\ 0 & 2 \end{vmatrix} \\ &= 1 \cdot (16 - 12) - 3 \cdot (8 - 0) + 5 \cdot (4 - 0) \\ &= 16 - 12 - 24 + 20 = 36 - 36 = 0 \\ &= 0 \end{aligned}$$

4.2.
Compute the following determinant efficiently.

$$\begin{vmatrix} 2 & 0 & 1 & 2 & 0 \\ 2 & -1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 \\ -2 & 0 & 2 & -1 & 2 \\ -2 & 0 & 0 & 1 & 1 \end{vmatrix}$$

We could compute the determinant using the Laplace expansions. However, this would be inefficient since we would require to compute 5 determinants of 4 × 4 matrices. The most efficient way to do this is to perform the Laplace expansion of rows when all but one numbers are zero. We proceed as follows.

Take out a factor of 2 from the first column, then swap the first and the third row.

$$\begin{vmatrix} 2 & 0 & 1 & 2 & 0 \\ 2 & -1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 \\ -2 & 0 & 2 & -1 & 2 \\ -2 & 0 & 0 & 1 & 1 \end{vmatrix} = 2 \cdot \begin{vmatrix} 1 & 0 & 1 & 2 & 0 \\ 1 & -1 & 0 & 1 & 1 \\ 0 & 1 & 2 & 1 & 2 \\ -1 & 0 & 2 & -1 & 2 \\ -1 & 0 & 0 & 1 & 1 \end{vmatrix} = -2 \cdot \begin{vmatrix} 0 & 1 & 2 & 1 & 2 \\ 1 & -1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 2 & 0 \\ -1 & 0 & 2 & -1 & 2 \\ -1 & 0 & 0 & 1 & 1 \end{vmatrix}$$

Add the first row to the second one swap the first and second columns, and perform Laplace expansion over the first column.

$$= -2 \cdot \begin{vmatrix} 0 & 1 & 2 & 1 & 2 \\ 1 & 0 & 2 & 2 & 3 \\ 1 & 0 & 1 & 2 & 0 \\ -1 & 0 & 2 & -1 & 2 \\ -1 & 0 & 0 & 1 & 1 \end{vmatrix} = 2 \cdot \begin{vmatrix} 1 & 0 & 2 & 1 & 2 \\ 0 & 1 & 2 & 2 & 3 \\ 0 & 1 & 1 & 2 & 0 \\ 0 & -1 & 2 & -1 & 2 \\ 0 & -1 & 0 & 1 & 1 \end{vmatrix} = 2 \cdot 1 \cdot \begin{vmatrix} 1 & 2 & 2 & 3 \\ 1 & 1 & 2 & 0 \\ -1 & 2 & -1 & 2 \\ -1 & 0 & 1 & 1 \end{vmatrix}$$

Substract the first row to the second one, and add the first row to the second and third ones. Then, perform Laplace expansion over the first column again.

$$= 2 \cdot \begin{vmatrix} 1 & 2 & 2 & 3 \\ 1 & 1 & 2 & 0 \\ -1 & 2 & -1 & 2 \\ -1 & 0 & 1 & 1 \end{vmatrix} = 2 \cdot \begin{vmatrix} 1 & 2 & 2 & 3 \\ 0 & -1 & 0 & -3 \\ 0 & 4 & 1 & 5 \\ 0 & 2 & 3 & 4 \end{vmatrix} = 2 \cdot \begin{vmatrix} -1 & 0 & -3 \\ 4 & 1 & 5 \\ 2 & 3 & 4 \end{vmatrix}$$

Transpose the determinant, then substract the first row 3 times to the third one. Perform Laplace expansion over the first column again.

$$= 2 \cdot \begin{vmatrix} -1 & 4 & 2 \\ 0 & 1 & 3 \\ -3 & 5 & 4 \end{vmatrix} = 2 \cdot \begin{vmatrix} -1 & 4 & 2 \\ 0 & 1 & 3 \\ 0 & -7 & -2 \end{vmatrix} = (-1) \cdot 2 \cdot \begin{vmatrix} 1 & 3 \\ -7 & -2 \end{vmatrix}$$

Calculate the resulting value:

$$= -2 \cdot \begin{vmatrix} 1 & 3 \\ -7 & -2 \end{vmatrix} = -2 \cdot (-2 + 21) = -38$$

4.3.

a.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} 1-\lambda & 0 \\ 1 & 1-\lambda \end{vmatrix} = (1-\lambda)^2 = (\lambda - 1)^2$$

we have $\lambda = 1$ with multiplicity 2

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$x = 0, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

b.

$$\mathbf{B} = \begin{pmatrix} -2 & 2 \\ 2 & 1 \end{pmatrix}$$

$$\begin{aligned} \det(\mathbf{B} - \lambda \mathbf{I}) = \begin{vmatrix} -2-\lambda & 2 \\ 2 & 1-\lambda \end{vmatrix} &= (\lambda + 2)(\lambda - 1) - 4 = \lambda^2 + \lambda - 2 - 4 \\ &= \lambda^2 + \lambda - 6 = (\lambda + 3)(\lambda - 2) \end{aligned}$$

we have $\lambda = -3$ with multiplicity 1

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} x + 2y = 0 \\ \cancel{2x + 4y = 0} \end{cases}, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\}$$

we have $\lambda = 2$ with multiplicity 1

$$\begin{pmatrix} -4 & 2 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} \cancel{-4x + 2y = 0} \\ 2x - y = 0 \end{cases}, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$$

4.4.

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 1 & 1 \\ -1 & 1 & -2 & 3 \\ 2 & -1 & 0 & 0 \\ 1 & -1 & 1 & 0 \end{pmatrix}$$

Let us denote the following notation. We refer to a column as C_x and to a row as R_x. E.g. C_2 refers to the second column and $C_1 = C_1 + 2 \cdot C_3$ refers to setting the first column to the addition of the first one and two times the third one.

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & -1 & 1 & 1 \\ -1 & 1-\lambda & -2 & 3 \\ 2 & -1 & -\lambda & 0 \\ 1 & -1 & 1 & -\lambda \end{vmatrix}$$

First, $C_1 = C_1 + 2 \cdot C_2$, and then $C_3 = C_3 - \lambda \cdot C_2$.

$$= \begin{vmatrix} -\lambda-2 & -1 & 1 & 1 \\ 1-2\lambda & 1-\lambda & -2 & 3 \\ 0 & -1 & -\lambda & 0 \\ -1 & -1 & 1 & -\lambda \end{vmatrix} = \begin{vmatrix} -\lambda-2 & -1 & \lambda+1 & 1 \\ 1-2\lambda & 1-\lambda & -2-\lambda-\lambda^2 & 3 \\ 0 & -1 & 0 & 0 \\ -1 & -1 & \lambda+1 & -\lambda \end{vmatrix} = \begin{vmatrix} -\lambda-2 & -1 & \lambda+1 & 1 \\ 1-2\lambda & 1-\lambda & (\lambda-2)(\lambda+1) & 3 \\ 0 & -1 & 0 & 0 \\ -1 & -1 & \lambda+1 & -\lambda \end{vmatrix}$$

We perform Laplace expansion over the third row.

$$= \cancelto{1}{(-1)(-1)} \cdot \begin{vmatrix} -\lambda-2 & \lambda+1 & 1 \\ 1-2\lambda & (\lambda-2)(\lambda+1) & 3 \\ -1 & \lambda+1 & -\lambda \end{vmatrix}$$

We take out a factor of $\lambda+1$ from the second column and perform Laplace expansion over the first row.

$$= (\lambda+1) \cdot \begin{vmatrix} -\lambda-2 & 1 & 1 \\ 1-2\lambda & \lambda-2 & 3 \\ -1 & 1 & -\lambda \end{vmatrix} = (\lambda+1) \cdot \left(-(\lambda+2) \begin{vmatrix} \lambda-2 & 3 \\ 1 & -\lambda \end{vmatrix} - \begin{vmatrix} 1-2\lambda & 3 \\ -1 & -\lambda \end{vmatrix} + \begin{vmatrix} 1-2\lambda & \lambda-2 \\ -1 & 1 \end{vmatrix} \right)$$

$$= (\lambda+1)\Big((\lambda+2)(\lambda-2)\lambda + 3(\lambda+2) + \lambda(1-2\lambda) - 3 + 1 - 2\lambda + \lambda - 2 \Big)$$

$$= (\lambda+1)\Big((\lambda+2)(\lambda-2)\lambda + 3\lambda + 6 + \lambda - 2\lambda^2 - 3 + 1 - 2\lambda + \lambda - 2 \Big)$$

$$= (\lambda+1)\Big((\lambda+2)(\lambda-2)\lambda + -2\lambda^2 + 3\lambda + 2 \Big) = (\lambda+1)\Big((\lambda+2)(\lambda-2)\lambda + (\lambda-2)(-2\lambda-1) \Big)$$

$$= (\lambda+1)(\lambda-2)\Big(\lambda^2 + \cancel{2\lambda} - \cancel{2\lambda} - 1 \Big)$$

$$= (\lambda+1)^2(\lambda-2)(\lambda-1)$$

we have $\lambda = 2$ with multiplicity 1

$$\begin{pmatrix} -2 & -1 & 1 & 1 \\ -1 & -1 & -2 & 3 \\ 2 & -1 & -2 & 0 \\ 1 & -1 & 1 & -2 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 - R_3$ and $R_2 = R_2 + R_4$.

$$\begin{pmatrix} 0 & -2 & -1 & 1 \\ 0 & -2 & -1 & 1 \\ 2 & -1 & -2 & 0 \\ 1 & -1 & 1 & -2 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} 2x = -y + z \\ x = 2w - 2y \\ x = w + y - 2z \end{cases}$$

$$2w - 2y = w + y - 2z$$

$$w = 3y - 2y - 4x$$

$$w = y - 8w + 8y$$

$$w = y \implies x = 0 \implies y = z$$

the equations span the following eigenspace $\left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \right\}$

we have $\lambda = 1$ with multiplicity 1

$$\begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 0 & -2 & 3 \\ 2 & -1 & -1 & 0 \\ 1 & -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 - R_4$.

$$\begin{pmatrix} 0 & -2 & -2 & 0 \\ -1 & -0 & -2 & 3 \\ 2 & -1 & -1 & 0 \\ 1 & -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} x = y \\ w = -2y + 3z \\ x = 2w - y \\ x = w + y - z \end{cases} \implies x = w \implies z = w \implies w = x = y = z$$

the equations span the following eigenspace $\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \right\}$

we have $\lambda = -1$ with multiplicity 2

$$\begin{pmatrix} 1 & -1 & 1 & 1 \\ -1 & 2 & -2 & 3 \\ 2 & -1 & 1 & 0 \\ 1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_2 = R_2 + R_1$ and $R_3 = R_3 - R_1$.

$$\begin{pmatrix} 1 & -1 & 1 & 1 \\ 0 & 1 & -1 & 4 \\ 1 & 0 & 0 & -1 \\ 1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} w = z \\ y = x + 4z \\ x = w + y + z \end{cases} \implies \begin{cases} w = z \\ y = x + 4z \\ y = x + 2z \end{cases} \implies z = 0 \implies x = y$$

the equations span the following eigenspace $\left\{ \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} \right\}$

4.5.

a.

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} = 1 \neq 0 \implies \text{invertible}$$

the matrix is also diagonalizable according to the canonical base $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$

b.

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} = 0 \implies \text{not invertible}$$

however, the matrix is diagonalizable according to the canonical base $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$

c.

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

$$\begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1 \neq 0 \implies \text{invertible}$$

$$\begin{vmatrix} 1-\lambda & 1 \\ 0 & 1-\lambda \end{vmatrix} = (\lambda - 1)^2$$

$\lambda = 1$ with multiplicity 2.

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \text{eigenspace is } \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

The matrix is not diagonalizable because the eigenvectors do not span over $\mathbb{R}^2$.

d.

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

$$\begin{vmatrix} 0 & 1 \\ 0 & 0 \end{vmatrix} = 0 \implies \text{not invertible}$$

$$\begin{vmatrix} -\lambda & 1 \\ 0 & -\lambda \end{vmatrix} = \lambda^2$$

$\lambda = 0$ with multiplicity 2.

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \implies \text{eigenspace is } \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

The matrix is not diagonalizable because the eigenvectors do not span over $\mathbb{R}^2$.

4.6.

a.

$$\mathbf{A} = \begin{pmatrix} 2 & 3 & 0 \\ 1 & 4 & 3 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 2-\lambda & 3 & 0 \\ 1 & 4-\lambda & 3 \\ 0 & 0 & 1-\lambda \end{vmatrix} &= (1-\lambda)\Big((2-\lambda)(4-\lambda)-3\Big) \\ &= (1-\lambda)(\lambda^2 - 6\lambda + 5) \\ &= (1-\lambda)(\lambda-5)(\lambda-1) \\ &= -(\lambda-1)^2(\lambda-5) \end{aligned}$$

we have $\lambda = 5$ with multiplicity 1

$$\begin{pmatrix} -3 & 3 & 0 \\ 1 & -1 & 3 \\ 0 & 0 & -4 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$\begin{cases} x = y \\ z = 0 \end{cases}$, the equations span the following eigenspace $\left\{ \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \right\}$

we have $\lambda = 1$ with multiplicity 2

$$\begin{pmatrix} 1 & 3 & 0 \\ 1 & 3 & 3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$\begin{cases} x = -3y \\ z = 0 \end{cases}$, the equations span the following eigenspace $\left\{ \begin{bmatrix} -3 \\ 1 \\ 0 \end{bmatrix} \right\}$

The matrix is not diagonalizable because the basis formed by the eigenvectors does not span over $\mathbb{R}^3$, but over $\mathbb{R}^2$.

b.

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 1 & 0 & 0 \\ 0 & -\lambda & 0 & 0 \\ 0 & 0 & -\lambda & 0 \\ 0 & 0 & 0 & -\lambda \end{vmatrix} = (1-\lambda) \begin{vmatrix} -\lambda & 0 & 0 \\ 0 & -\lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix} = -\lambda^3(\lambda - 1)$$

we have $\lambda = 1$ with multiplicity 1

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} x = 0 \\ y = 0 \\ z = 0 \end{cases}, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

we have $\lambda = 0$ with multiplicity 3

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\left\{ w = x \right., \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

The matrix is diagonalizable because the basis formed by the eigenvectors spans over $\mathbb{R}^4$.

4.7.

a.

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -8 & 4 \end{pmatrix}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -8 & 4 - \lambda \end{vmatrix} = (\lambda - 4)\lambda + 8 = \lambda^2 + -4\lambda + 8$$

$$\lambda = \frac{4 \pm \sqrt{16 - 32}}{2} = \frac{4 \pm 4i}{2} = 2 \pm 2i$$

we have $\lambda = 2 - 2i$ with multiplicity 1

$$\begin{pmatrix} -2 + 2i & 1 \\ -8 & 2 + 2i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$8x = (2 + 2i)y \implies 4x = (1 + i)y, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 + i \\ 4 \end{bmatrix} \right\}$$

we have $\lambda = 2 + 2i$ with multiplicity 1

$$\begin{pmatrix} -2 - 2i & 1 \\ -8 & 2 - 2i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$8x = (2 - 2i)y \implies 4x = (1 - i)y, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 - i \\ 4 \end{bmatrix} \right\}$$

The matrix is diagonalizable as follows.

$$\mathbf{A} = \mathbf{PDP}^{-1} \quad \text{with } \mathbf{D} = \begin{pmatrix} 2 + 2i & 0 \\ 0 & 2 - 2i \end{pmatrix}, \text{ and } \mathbf{P} = \begin{pmatrix} 1 - i & 1 + i \\ 4 & 4 \end{pmatrix}$$

b.

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 1 & 1 \\ 1 & 1-\lambda & 1 \\ 1 & 1 & 1-\lambda \end{vmatrix}$$

First $R_2 = R_2 - R_3$, then $C_2 = C_2 + C_3$, and finally Laplace expansion over the second row.

$$= \begin{vmatrix} 1-\lambda & 1 & 1 \\ 0 & -\lambda & \lambda \\ 1 & 1 & 1-\lambda \end{vmatrix} = \begin{vmatrix} 1-\lambda & 2 & 1 \\ 0 & 0 & \lambda \\ 1 & 2-\lambda & 1-\lambda \end{vmatrix} = -\lambda \begin{vmatrix} 1-\lambda & 2 \\ 1 & 2-\lambda \end{vmatrix}$$

$$= -\lambda\Big((\lambda-1)(\lambda-2) - 2\Big) = -\lambda\Big(\lambda^2 - 3\lambda + \cancel{2} - \cancel{2}\Big)$$

$$= -\lambda^2(\lambda - 3)$$

we have $\lambda = 3$ with multiplicity 1

$$\begin{pmatrix} -2 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 + R_2$

$$\begin{pmatrix} -1 & -1 & 2 \\ 1 & -2 & 1 \\ 1 & 1 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

We notice the first and last row are the same except for a factor of -1.

$$\begin{cases} 2z = x + y \\ z = 2y - x \end{cases} \implies 4y - 2x = x + y \implies x = y \implies z = y \implies x = y = z$$

the equations span the following eigenspace $\left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$

we have $\lambda = 0$ with multiplicity 2

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$\Big\{ x = -y - z$, the equations span the following eigenspace $\left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} \right\}$

The matrix is diagonalizable as follows.

$$\mathbf{A} = \mathbf{PDP}^{-1} \quad \text{with } \mathbf{D} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ and } \mathbf{P} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

c.

$$\mathbf{A} = \begin{pmatrix} 5 & 4 & 2 & 1 \\ 0 & 1 & -1 & -1 \\ -1 & -1 & 3 & 0 \\ 1 & 1 & -1 & 2 \end{pmatrix}$$

$$\det(\mathbf{A}-\lambda\mathbf{I}) = \begin{vmatrix} 5-\lambda & 4 & 2 & 1 \\ 0 & 1-\lambda & -1 & -1 \\ -1 & -1 & 3-\lambda & 0 \\ 1 & 1 & -1 & 2-\lambda \end{vmatrix}$$

First $R_3 = R_3 + R_4$, then we take $2-\lambda$ as a factor and then $R_2 = R_2 - R_3$.

$$= \begin{vmatrix} 5-\lambda & 4 & 2 & 1 \\ 0 & 1-\lambda & -1 & -1 \\ 0 & 0 & 2-\lambda & 2-\lambda \\ 1 & 1 & -1 & 2-\lambda \end{vmatrix} = (2-\lambda) \begin{vmatrix} 5-\lambda & 4 & 2 & 1 \\ 0 & 1-\lambda & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & -1 & 2-\lambda \end{vmatrix} = (2-\lambda) \begin{vmatrix} 5-\lambda & 4 & 2 & 1 \\ 0 & 1-\lambda & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & -1 & 2-\lambda \end{vmatrix}$$

We perform the Laplace expansion over the second row. Then, $R_1 = R_1 - R_2$

$$= (2-\lambda)(1-\lambda) \begin{vmatrix} 5-\lambda & 2 & 1 \\ 0 & 1 & 1 \\ 1 & -1 & 2-\lambda \end{vmatrix} = (2-\lambda)(1-\lambda) \begin{vmatrix} 5-\lambda & 1 & 0 \\ 0 & 1 & 1 \\ 1 & -1 & 2-\lambda \end{vmatrix}$$

We perform Laplace expansion over the first row.

$$= (2-\lambda)(1-\lambda)\left((5-\lambda) \begin{vmatrix} 1 & 1 \\ -1 & 2-\lambda \end{vmatrix} - \begin{vmatrix} 0 & 1 \\ 1 & 2-\lambda \end{vmatrix} \right)$$

$$= (2-\lambda)(1-\lambda)\Big((5-\lambda)(2-\lambda+1)+1\Big) = (2-\lambda)(1-\lambda)\Big((\lambda-5)(\lambda-3)+1\Big)$$

$$= (2-\lambda)(1-\lambda)(\lambda^2-8\lambda+16) = (\lambda-2)(\lambda-1)(\lambda-4)^2$$

we have $\lambda = 4$ with multiplicity 2

$$\begin{pmatrix} 1 & 4 & 2 & 1 \\ 0 & -3 & -1 & -1 \\ -1 & -1 & -1 & 0 \\ 1 & 1 & -1 & -2 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 + R_2$

$$\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -3 & -1 & -1 \\ -1 & -1 & -1 & 0 \\ 1 & 1 & -1 & -2 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} y = -z \\ w = -y - z \\ 3x = -y - z \end{cases} \implies x = 0 \implies w = -y, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix} \right\}$$

There is no need to compute the rest of the eigenvectors since we with this result we can conclude that they will not form a basis in $\mathbb{R}^4$ and thus, the matrix $\mathbf{A}$ is not diagonalizable. However, let us compute the rest of the eigenvectors for illustrative purposes.

we have $\lambda = 2$ with multiplicity 1

$$\begin{pmatrix} 3 & 4 & 2 & 1 \\ 0 & -1 & -1 & -1 \\ -1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

We ignore fourth row, and $R_1 = R_1 + 3R_3$.

$$\begin{pmatrix} 0 & 1 & 5 & 1 \\ 0 & -1 & -1 & -1 \\ -1 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 + R_2$.

$$\begin{pmatrix} 0 & 0 & 4 & 0 \\ 0 & -1 & -1 & -1 \\ -1 & -1 & 1 & 0 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} y = 0 \\ x = -z \\ x = -w \end{cases} \implies \text{the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix} \right\}$$

we have $\lambda = 1$ with multiplicity 1

$$\begin{pmatrix} 4 & 4 & 2 & 1 \\ 0 & 0 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 + 2R_4$ and $R_3 = R_3 + R_4$.

$$\begin{pmatrix} 6 & 6 & 0 & 3 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_4 = R_4 + R_3$.

$$\begin{pmatrix} 6 & 6 & 0 & 3 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} z = 0 \\ y = -z \\ w = -x \end{cases} \implies y = z = 0, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \right\}$$

As mentioned before, this matrix is diagonalizable because the basis of eigenvectors span over $\mathbb{R}^3$, and not $\mathbb{R}^4$.

d.

$$\mathbf{A} = \begin{pmatrix} 5 & -6 & -6 \\ -1 & 4 & 2 \\ 3 & -6 & -4 \end{pmatrix}$$

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 5-\lambda & -6 & -6 \\ -1 & 4-\lambda & 2 \\ 3 & -6 & -4-\lambda \end{vmatrix}$$

First $R_1 = R_1 - R_3$, then we factor out $\lambda - 2$ from the resulting first row and after that, $C_3 = C_3 + C_1$.

$$= \begin{vmatrix} 2-\lambda & 0 & \lambda-2 \\ -1 & 4-\lambda & 2 \\ 3 & -6 & -4-\lambda \end{vmatrix} = (\lambda-2)\begin{vmatrix} -1 & 0 & 1 \\ -1 & 4-\lambda & 2 \\ 3 & -6 & -4-\lambda \end{vmatrix} = (\lambda-2)\begin{vmatrix} -1 & 0 & 0 \\ -1 & 4-\lambda & 1 \\ 3 & -6 & -1-\lambda \end{vmatrix}$$

We perform the Laplace expansion over the first row.

$$= -(\lambda-2)\begin{vmatrix} 4-\lambda & 1 \\ -6 & -1-\lambda \end{vmatrix} = -(\lambda-2)\Big((\lambda-4)(\lambda+1)+6\Big) = -(\lambda-2)(\lambda^2-3\lambda-4+6)$$

$$= -(\lambda-2)(\lambda^2-3\lambda+2) = -(\lambda-2)^2(\lambda-1)$$

we have $\lambda = 2$ with multiplicity 2.

$$\begin{pmatrix} 3 & -6 & -6 \\ -1 & 2 & 2 \\ -3 & -6 & -6 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

We notice the three rows represent the same equation.

$$x = 2y + 2z, \quad \text{the equations span the following eigenspace } \left\{ \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix} \right\}$$

we have $\lambda = 1$ with multiplicity 1.

$$\begin{pmatrix} 4 & -6 & -6 \\ -1 & 3 & 2 \\ 3 & -6 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = \frac{1}{2}R_1$, and $R_3 = R_3 + 3R_2$

$$\begin{pmatrix} 2 & -3 & -3 \\ -1 & 3 & 2 \\ 0 & 3 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$R_1 = R_1 + R_3$, and then $R_2 = R_2 - R_3$

$$\begin{pmatrix} 2 & 0 & -2 \\ -1 & 0 & 1 \\ 0 & 3 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} x = z \\ z = -3y \end{cases}, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ -\frac{1}{3} \\ 1 \end{bmatrix} \right\}$$

The matrix is diagonalizable as follows.

$$\mathbf{A} = \mathbf{PDP}^{-1} \quad \text{with } \mathbf{D} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \text{ and } \mathbf{P} = \begin{pmatrix} 2 & 2 & 1 \\ 1 & 0 & -\frac{1}{3} \\ 0 & 1 & 1 \end{pmatrix}$$

4.8.

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix}3 & 2\\ 2 & 3\\ 2 & -2\end{pmatrix} \cdot \begin{pmatrix}3 & 2 & 2\\ 2 & 3 & -2\end{pmatrix} = \begin{pmatrix}13 & 12 & 2\\ 12 & 13 & -2\\ 2 & -2 & 8\end{pmatrix}$$

$$\det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix}13-\lambda & 12 & 2\\ 12 & 13-\lambda & -2\\ 2 & -2 & 8-\lambda\end{vmatrix} = \begin{vmatrix}13-\lambda & 12 & 2\\ 25-\lambda & 25-\lambda & 0\\ 2 & -2 & 8-\lambda\end{vmatrix} = \begin{vmatrix}1-\lambda & 12 & 2\\ 0 & 25-\lambda & 0\\ 4 & -2 & 8-\lambda\end{vmatrix}$$

$$= (25-\lambda)\begin{vmatrix}1-\lambda & 2\\ 4 & 8-\lambda\end{vmatrix} = (25-\lambda)\Big((1-\lambda)(8-\lambda)-8\Big) = (25-\lambda)(\lambda^2-9\lambda+8-8)$$

$$(25-\lambda)(\lambda^2-9\lambda) = -(\lambda-25)(\lambda-9)\lambda$$

We have $\lambda = 25$ with multiplicity 1.

$$\begin{pmatrix}-12 & 12 & 2\\ 12 & -12 & -2\\ 2 & -2 & 17\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{cases}6x = 6y + z\\ 2x = 2y + 17z\end{cases} \implies z = 0, x = y, \text{ the equations span the following eigenspace } \left\{\begin{bmatrix}1\\ 1\\ 0\end{bmatrix}\right\},$$

$$\text{if normalized } \left\{\begin{bmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\\ 0\end{bmatrix}\right\}$$

We have $\lambda = 9$ with multiplicity 1.

$$\begin{pmatrix}4 & 12 & 2\\ 12 & 4 & -2\\ 2 & -2 & -1\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{pmatrix}16 & 16 & 0\\ 12 & 4 & -2\\ 2 & -2 & -1\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{pmatrix}16 & 16 & 0\\ 16 & 0 & -4\\ 2 & -2 & -1\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{cases}x = -y\\ 4x = z\end{cases} \quad \text{the equations span the following eigenspace } \left\{\begin{bmatrix}1\\ -1\\ 4\end{bmatrix}\right\}, \text{ if normalized } \left\{\begin{bmatrix}\frac{1}{3\sqrt{2}}\\ -\frac{1}{3\sqrt{2}}\\ \frac{4}{3\sqrt{2}}\end{bmatrix}\right\}$$

We have $\lambda = 0$ with multiplicity 1.

$$\begin{pmatrix}13 & 12 & 2\\ 12 & 13 & -2\\ 2 & -2 & 8\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{pmatrix}25 & 25 & 0\\ 12 & 13 & -2\\ 1 & -1 & 4\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{pmatrix}1 & 1 & 0\\ 2 & 0 & 4\end{pmatrix}\begin{pmatrix}x\\ y\\ z\end{pmatrix} = \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}$$

$$\begin{cases} x = -y \\ x = -2z \end{cases} \quad \text{the equations span the following eigenspace } \left\{ \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix} \right\}, \text{ if normalized } \left\{ \begin{bmatrix} -\frac{2}{3} \\ \frac{2}{3} \\ \frac{1}{3} \end{bmatrix} \right\}$$

The matrix $\mathbf{A}^T\mathbf{A}$ is diagonalizable as follows.

$$\mathbf{A} = \mathbf{PDP}^T \quad \text{with } \mathbf{D} = \begin{pmatrix} 25 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \text{ and } \mathbf{P} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & -\frac{2}{3} \\ \frac{1}{\sqrt{2}} & -\frac{1}{3\sqrt{2}} & \frac{2}{3} \\ 0 & \frac{4}{3\sqrt{2}} & \frac{1}{3} \end{pmatrix}$$

Notice $\mathbf{P}^{-1} = \mathbf{P}^T$ because $\mathbf{P}$ is an orthonormal matrix. Using the information of the previous eigendecomposition, we can obtain the matrix with singular values, taking the root of the non-zero eigenvalues.

$$\mathbf{A} = \mathbf{U\Sigma V}^T \quad \text{with } \mathbf{\Sigma} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}, \text{ and } \mathbf{V} = \mathbf{P}$$

We only need to compute the value of $\mathbf{U}$. This is done as follows.

$$\mathbf{u}_1 = \frac{1}{\sigma_1}\mathbf{A}\mathbf{v}_1 = \frac{1}{5}\begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} = \frac{1}{5}\begin{pmatrix} \frac{5}{\sqrt{2}} \\ \frac{5}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\mathbf{u}_2 = \frac{1}{\sigma_2}\mathbf{A}\mathbf{v}_2 = \frac{1}{3}\begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}\begin{pmatrix} \frac{1}{3\sqrt{2}} \\ -\frac{1}{3\sqrt{2}} \\ \frac{4}{3\sqrt{2}} \end{pmatrix} = \frac{1}{9\sqrt{2}}\begin{pmatrix} 9 \\ -9 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

$$\mathbf{U} = \Big(\mathbf{u}_1, \mathbf{u}_2\Big) = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

Therefore, the SVD of $\mathbf{A}$ is the following.

$$\mathbf{A} = \mathbf{U\Sigma V}^T \quad \text{with } \mathbf{\Sigma} = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{pmatrix}, \text{ and } \mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{3\sqrt{2}} & -\frac{2}{3} \\ \frac{1}{\sqrt{2}} & -\frac{1}{3\sqrt{2}} & \frac{2}{3} \\ 0 & \frac{4}{3\sqrt{2}} & \frac{1}{3} \end{pmatrix}, \text{ and } \mathbf{U} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

4.9.

$$\mathbf{A} = \begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix}$$

$$\mathbf{A}^T\mathbf{A} = \begin{pmatrix} 2 & -1 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$$

$$\det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = 25 - 10\lambda + \lambda^2 - 9 = \lambda^2 - 10\lambda + 16 = (\lambda - 8)(\lambda - 2)$$

We have $\lambda = 8$ with multiplicity 1.

$$\begin{pmatrix} -3 & 3 \\ 3 & -3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$x = y, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\}, \text{ if normalized } \left\{ \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \right\}$$

We have $\lambda = 2$ with multiplicity 1.

$$\begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$x = -y, \text{ the equations span the following eigenspace } \left\{ \begin{bmatrix} -1 \\ 1 \end{bmatrix} \right\}, \text{ if normalized } \left\{ \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \right\}$$

The matrix $\mathbf{A}^T\mathbf{A}$ is diagonalizable as follows.

$$\mathbf{A} = \mathbf{PDP}^T \quad \text{with } \mathbf{D} = \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix}, \text{ and } \mathbf{P} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

Notice $\mathbf{P}^{-1} = \mathbf{P}^T$ because $\mathbf{P}$ is an orthonormal matrix. Using the information of the previous eigendecomposition, we can obtain the matrix with singular values, taking the root of the non-zero eigenvalues.

$$\mathbf{A} = \mathbf{U\Sigma V}^T \quad \text{with } \mathbf{\Sigma} = \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}, \text{ and } \mathbf{V} = \mathbf{P}$$

We only need to compute the value of $\mathbf{U}$. This is done as follows.

$$\mathbf{u}_1 = \frac{1}{\sigma_1}\mathbf{A}\mathbf{v}_1 = \frac{1}{2\sqrt{2}} \begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 4 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$\mathbf{u}_2 = \frac{1}{\sigma_2}\mathbf{A}\mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1, \mathbf{u}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Therefore, the SVD of $\mathbf{A}$ is the following.

$$\mathbf{A} = \mathbf{U\Sigma V}^T \quad \text{with } \mathbf{\Sigma} = \begin{pmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}, \text{ and } \mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}, \text{ and } \mathbf{U} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

4.10.

$$\mathbf{A} = \begin{pmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{pmatrix}$$

We already computed the SVD of this particular matrix in previous exercises. Using results from 4.8, the rank-1 approximation of $\mathbf{A}$ is the following.

$$\hat{\mathbf{A}}(1) = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T = 5 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{pmatrix} = 5 \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} = \begin{pmatrix} \frac{5}{2} & \frac{5}{2} & 0 \\ \frac{5}{2} & \frac{5}{2} & 0 \end{pmatrix}$$

4.11.

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, show that the matrices $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ possess the same non-zero eigenvalues. First, consider the SVD of matrix $\mathbf{A}$

$$\mathbf{A} = \mathbf{U\Sigma V}^T$$

and the products $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$

$$\mathbf{A}^T\mathbf{A} = (\mathbf{U\Sigma V}^T)^T \mathbf{U\Sigma V}^T = \mathbf{V\Sigma}^T\mathbf{U}^T\mathbf{U\Sigma V}^T = \mathbf{V\Sigma}^T\mathbf{\Sigma V}^T$$

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T$$

where $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ since both are orthonormal matrices. The previous two expressions can be regarded as the eigendecompositions of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ respectively.

Let us consider the case $m = n$. Since $\mathbf{\Sigma}$ would be a diagonal square matrix (and thus $\mathbf{\Sigma} = \mathbf{\Sigma}^T$), the products $\mathbf{\Sigma}\mathbf{\Sigma}^T$ and $\mathbf{\Sigma}^T\mathbf{\Sigma}$ would result in the same diagonal matrix, which would imply that both matrices have exactly the same eigenvalues.

Now consider $m < n$. The contrary case ($n < m$) would be equivalent, since it would suffice to transpose $\mathbf{A}$ to obtain the same condition as here. Let us denote the products $\mathbf{\Sigma}\mathbf{\Sigma}^T$ and $\mathbf{\Sigma}^T\mathbf{\Sigma}$ for $m < n$.

$$\mathbf{\Sigma}\mathbf{\Sigma}^T = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_m & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & \sigma_m \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & \sigma_m^2 \end{pmatrix}, \quad \mathbf{\Sigma}\mathbf{\Sigma}^T \in \mathbb{R}^{m \times m}$$

$$\mathbf{\Sigma}^T\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & \sigma_m \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_m & 0 & \dots & 0 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 & \dots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_m^2 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{\Sigma}^T\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$$

The first product $\mathbf{\Sigma}\mathbf{\Sigma}^T$ denotes the eigenvalues of $\mathbf{A}\mathbf{A}^T$, which we observe that has m eigenvalues. The second one, $\mathbf{\Sigma}^T\mathbf{\Sigma}$ denotes the eigenvalues of $\mathbf{A}^T\mathbf{A}$ (it has n eigenvalues in this case). If we observe the previous calculations, we notice that the first m eigenvalues of $\mathbf{A}^T\mathbf{A}$ are exactly the eigenvalues of $\mathbf{A}\mathbf{A}^T$, and the rest $n-m$ eigenvalues are 0. Thus, we have just proven that $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ have the same non-zero eigenvalues. Recall that if $n < m$, it suffices to transpose the original matrix, and perform the same reasoning.

4.12.

The spectral norm, along with Theorem 4.24, is defined as follows

$$||\mathbf{A}||_2 = \max_{\mathbf{x}} \frac{||\mathbf{A}\mathbf{x}||_2}{||\mathbf{x}||_2} = \sigma_1$$

where σ_1 is the highest singular value of $\mathbf{A} \in \mathbb{R}^{m \times n}$. We want to prove that Theorem 4.24 holds.

Let us begin by expanding the euclidean norm $||\mathbf{A}\mathbf{x}||_2$

$$||\mathbf{A}\mathbf{x}||_2 = \sqrt{(\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{x})} = \sqrt{\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x}}$$

We have $\mathbf{A}^T\mathbf{A}$, which is diagonalizable according to Theorem 4.21, which states that a symmetric matrix can always be diagonalized. Let us prove that $\mathbf{A}^T\mathbf{A}$ is symmetric

$$(\mathbf{A}^T\mathbf{A})^T = \mathbf{A}^T(\mathbf{A}^T)^T = \mathbf{A}^T\mathbf{A}$$

Since $\mathbf{A}^T\mathbf{A}$ is symmetric, one can eigendecompose this matrix into the following expression.

$$\mathbf{A}^T\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^T$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix formed by the eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{P}$ is the matrix formed by an orthonormal basis of eigenvectors. It is crucial to use an orthonormal basis, since we will have

$\mathbf{P}^{-1} = \mathbf{P}^T$. We note that the diagonal matrix $\mathbf{D}$ is constructed with the eigenvalues in descending order, being λ_1 the highest eigenvalue of $\mathbf{A}^T\mathbf{A}$. Let us continue developing the first expression.

$$||\mathbf{Ax}||_2 = \sqrt{\mathbf{x}^T\mathbf{A}^T\mathbf{Ax}} = \sqrt{\mathbf{x}^T\mathbf{PDP}^T\mathbf{x}}$$

Now we denote the value $\mathbf{y} = \mathbf{P}^T\mathbf{x}$. Since $\mathbf{P}$ is an orthonormal matrix, the quantity $\mathbf{y}$ preserves the norm.

$$||\mathbf{y}||_2 = \mathbf{y}^T\mathbf{y} = \mathbf{x}^T\mathbf{PP}^T\mathbf{x} = \mathbf{x}^T\mathbf{Ix} = ||\mathbf{x}||_2, \quad \text{where } \mathbf{PP}^T = \mathbf{I}.$$

Let us now formulate the spectral norm in terms of $\mathbf{y}$. Note that maximizing w.r.t. $\mathbf{x} \neq \mathbf{0}$ is equivalent to maximizing w.r.t. $\mathbf{y} \neq \mathbf{0}$

$$||\mathbf{Ax}||_2 = \sqrt{\mathbf{x}^T\mathbf{A}^T\mathbf{Ax}} = \sqrt{\mathbf{x}^T\mathbf{PDP}^T\mathbf{x}} = \sqrt{\mathbf{y}^T\mathbf{D}\mathbf{y}^T} \implies ||\mathbf{A}||_2 = \max_{\mathbf{y}} \frac{\sqrt{\mathbf{y}^T\mathbf{D}\mathbf{y}^T}}{||\mathbf{y}||_2}$$

And if we take a normalized vector $\mathbf{y}$, we have $||\mathbf{y}||_2 = 1$. Thus, we are looking for the following

$$\max_{\mathbf{y}} \sqrt{\sum_{i=1}^{n} y_i^2\lambda_i}, \quad \text{for } ||\mathbf{y}||_2 = 1$$

Notice that if the eigenvalues are sorted in descending order, the $\mathbf{y}$ which maximizes this quantity will be the one which takes the maximum eigenvalue. Therefore, we set $y_1 = 1$ and $y_i = 0, \forall i \neq 1$. Any other configuration of $\mathbf{y}$ results in obtaining a lower value. Consequently

$$||\mathbf{A}||_2 = \max_{\mathbf{y}} \sqrt{\sum_{i=1}^{n} y_i^2\lambda_i} = \sqrt{\lambda_1} = \sigma_1$$

This maximum value is σ_1, which is the highest singular value.

9 Week 6 exercises

Exercise 6.13 (Theorem 6.15) Given a continuous random variable X, with *cdf* $F_X(x)$, show that the random variable $Y := F_X(X)$ is uniformly distribute

$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(F_X(X) \leq y) \\
&= P(X \leq F_X^{-1}(y)) \\
&= F_X(F_X^{-1}(y)) \\
&= y
\end{aligned}$$

Thus, we have that $Y \sim Uniform(0,1)$ because $F_Y(y)$ is the *cdf* of a uniform distribution.

Is the variance estimator unbiased? Let $\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(h_n - \hat{\mu})^2$ be the variance estimator. We begin by calculating the expected value $\mathbb{E}[\hat{\sigma}^2]$.

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(h_n - \hat{\mu})^2\right] = \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}(h_n^2 - 2\hat{\mu}h_n + \hat{\mu}^2)\right] \\
&= \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}h_n^2\right] - \frac{2}{N}\mathbb{E}\left[\sum_{n=1}^{N}\hat{\mu}h_n\right] + \frac{1}{N}\mathbb{E}\left[\sum_{n=1}^{N}\hat{\mu}^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[h_n^2\Big] - \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}\Big[\hat{\mu}h_n\Big] + \mathbb{E}\Big[\hat{\mu}^2\Big]
\end{aligned}$$

Recall that we have $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}h_n$. Let us use the mean estimator to calculate the previous expression.

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[h_n^2\Big] - \frac{2}{N}\sum_{n=1}^{N}\mathbb{E}\left[\frac{1}{N}\sum_{m=1}^{N}h_m h_n\right] + \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}h_n\right)^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[h_n^2\Big] - \frac{2}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\Big[h_n h_m\Big] + \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\Big[h_n h_m\Big] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[h_n^2\Big] - \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}\mathbb{E}\Big[h_n h_m\Big]
\end{aligned}$$

The first term of this expression is calculated by recalling a well-known variance identity. Let us consider h_n to be sampled from the Gaussian distribution we aim to compute the estimator for, i.e. $h_n \sim \mathcal{N}(\mu, \sigma^2)$.

$$\sigma^2 = \mathbb{E}\Big[h_n^2\Big] - \mathbb{E}\Big[h_n\Big]^2 \implies \mathbb{E}\Big[h_n^2\Big] = \sigma^2 + \mathbb{E}\Big[h_n\Big]^2 = \sigma^2 + \mu^2$$

Now we compute the second term of the expression. Notice the expectation of the product between two samples, $\mathbb{E}\Big[h_n h_m\Big]$. We find two cases:

- If $n = m$, the two samples will be correlated since they are equal, and we will have

  $$\mathbb{E}\Big[h_n h_n\Big] = \mathbb{E}\Big[h_n^2\Big] = \sigma^2 + \mu^2$$

  This happens exactly N times along the double summation.

- If $n \neq m$ the two samples are independent, thus

  $$\mathbb{E}\Big[h_n h_m\Big] = \mathbb{E}\Big[h_n\Big]\mathbb{E}\Big[h_m\Big] = \mu^2$$

  This happens exactly $N \times (N-1)$ times along the double summation.

With all this information we are ready to calculate the variance estimator.

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \cancel{\frac{1}{N}} \cancel{\sum_{n=1}^{N}} \left(\sigma^2 + \mu^2\right) - \frac{1}{N^2}\Big(N\left(\sigma^2 + \mu^2\right) + N(N-1)\mu^2\Big) \\
&= \sigma^2 + \mu^2 - \frac{1}{N}\sigma^2 - \frac{1}{N}\mu^2 - \frac{N-1}{N}\mu^2 \\
&= \frac{N-1}{N}\sigma^2 + \cancel{\frac{N-1}{N}\mu^2} - \cancel{\frac{N-1}{N}\mu^2}
\end{aligned}$$

The estimator is not unbiased because $\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$

$$\mathbb{E}[\hat{\sigma}^2] = \frac{N-1}{N}\sigma^2$$

Show that if we know $\mu = 0$, the variance estimator is unbiased If we know that $\mu = 0$, then we can simply use this information to calculate the value of $\hat{\mu}$, i.e. $\hat{\mu} = 0$.

$$\begin{aligned}
\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(h_n - \hat{\mu})^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N} h_n^2\right] \\
&= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\Big[h_n^2\Big] = \frac{1}{N}\sum_{n=1}^{N}\left(\sigma^2 + \mu^2\right) = \sigma^2 + \mu^2 = \sigma^2
\end{aligned}$$

Since we have $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$, we conclude that the variance estimator is unbiased provided that $\mu = 0$.

Maximum Likelihood variance estimator in linear regression Let us consider maximum likelihood estimation in linear regression. A typical probabilistic approach is to assume that the variable which one aims to predict is the result of a linear transformation plus some Gaussian noise of variance σ^2.

$$y_n = \mathbf{w}^T\mathbf{x}_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2), \quad n = \{1, \dots, N\}$$

The likelihood of a single sample n can be expressed as follows

$$p(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n, \sigma^2)$$

where $\mathbf{w}^T\mathbf{x}$ is the mean of the distribution and σ^2 is the variance which take into account the error addition. With this information we can calulate the log-likelihood of our data composed of N samples, $\{\mathbf{X}, \mathbf{y}\} = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$. Recall that we assume that the samples are i.i.d.

$$\begin{aligned}
\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2) = \log \prod_{n=1}^{N} p(y_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) &= \sum_{n=1}^{N} \log \mathcal{N}(y_n|\mathbf{w}^T\mathbf{x}_n, \sigma^2) \\
&= \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2}\right)\right) \\
&= \sum_{n=1}^{N}\left(-\frac{1}{2}\log\sigma^2 - \frac{1}{2}\log 2\pi - \frac{(y_n - \mathbf{w}^T\mathbf{x}_n)^2}{2\sigma^2}\right) \\
&= -\frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2
\end{aligned}$$

The maximum likelihood variance estimator can be computed by taking the derivative of the log-likelihood w.r.t the variance and solving for zero.

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\Big(\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2)\Big) &= \frac{\partial}{\partial\sigma^2}\Bigg(-\frac{N}{2}\log\sigma^2-\frac{N}{2}\log 2\pi-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2\Bigg)\\ &= -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2\end{aligned}$$

When solving for zero, we will have the maximum likelihood variance estimator $\hat{\sigma}_{ML}$.

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\Big(\log p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2)\Big)\Big|_{\sigma=\hat{\sigma}_{ML}} &= 0\\ -\frac{N}{2\hat{\sigma}_{ML}^2}+\frac{1}{2\hat{\sigma}_{ML}^4}\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2 &= 0\\ \frac{N}{\hat{\sigma}_{ML}^2} &= \frac{1}{\hat{\sigma}_{ML}^4}\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2\\ \hat{\sigma}_{ML}^2 &= \frac{1}{N}\sum_{n=1}^{N}(y_n-\mathbf{w}^T\mathbf{x}_n)^2\end{aligned}$$

This is the value of the maximum likelihood estimator. However, we do not know what is the value of $\mathbf{w}$. The best approach to solve this problem is to set $\mathbf{w}$ to the maximum likelihood estimator for this parameter, $\hat{\mathbf{w}}_{ML}$. Consequently we have

$$\hat{\sigma}_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}\big(y_n-\hat{\mathbf{w}}_{ML}^T\mathbf{x}_n\big)^2$$

10 Chapter 6 exercises (part 2)

Lecture Exercise If we reorder the vector $[\mathbf{y}, \mathbf{x}]$ what would the mean and covariance be to represent the same Gaussian.

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \mid \begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{xx} & \mathbf{\Sigma}_{xy}\\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_{yy}\end{bmatrix}\right)$$

We can begin by exploring the exponential terms of the joint distribution

$$-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{\Sigma}_{xx}^{-1}(\mathbf{x}-\mathbf{a}) - \frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{\Sigma}_{yy}^{-1}(\mathbf{y}-\mathbf{b}) - \frac{1}{2}(\mathbf{x}-\mathbf{a})^T\mathbf{\Sigma}_{xy}^{-1}(\mathbf{y}-\mathbf{b}) - \frac{1}{2}(\mathbf{y}-\mathbf{b})^T\mathbf{\Sigma}_{yx}^{-1}(\mathbf{x}-\mathbf{a})$$

Notice that we can reorder the terms such that we can write the probability distribution as $[\mathbf{y}, \mathbf{x}]$. This would result as follows.

$$p\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{y}\\ \mathbf{x}\end{bmatrix} \mid \begin{bmatrix}\mathbf{b}\\ \mathbf{a}\end{bmatrix}, \begin{bmatrix}\mathbf{\Sigma}_{yy} & \mathbf{\Sigma}_{yx}\\ \mathbf{\Sigma}_{xy} & \mathbf{\Sigma}_{xx}\end{bmatrix}\right)$$

Either expressions are equivalent since they result in the same distribution.

Exercise 6.5 Consider the time-series model:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + +\mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + +\mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

a. Form of $p(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T)$?

We can write down the joint probability over $\{\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T\}$ using conditional probabilities and the conditional independencies in the generative process of the time-series.

$$\begin{aligned} p(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T) &= p(\mathbf{x}_0)p(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T|\mathbf{x}_0) = p(\mathbf{x}_0)p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_2, ..., \mathbf{x}_T|\mathbf{x}_0, \mathbf{x}_1)\\ &= p(\mathbf{x}_0)p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}_0)p(\mathbf{x}_3, ..., \mathbf{x}_T|\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2)\\ &= p(\mathbf{x}_0)p(\mathbf{x}_1|\mathbf{x}_0)p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{x}_0)\dots p(\mathbf{x}_T|\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_{T-1}) \end{aligned}$$

If we observe the data generation process of the time-series, we find the following conditional independencies.

$$p(\mathbf{x}_t|\mathbf{x}_0, \dots, \mathbf{x}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1}), \quad 1 \le t \le T$$

Thus, the joint probability factorizes as follows

$$p(\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T) = p(\mathbf{x}_0)\prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1})$$

b. Assume $p(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$.

(a) Compute $p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)$.

$$\begin{aligned} p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t) &= \frac{1}{p(\mathbf{y}_1, ..., \mathbf{y}_t)} p(\mathbf{x}_{t+1}, \mathbf{y}_1, ..., \mathbf{y}_t)\\ &= \frac{1}{p(\mathbf{y}_1, ..., \mathbf{y}_t)} \int_{\mathbf{x}_t \in \mathbb{R}} p(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{y}_1, ..., \mathbf{y}_t) d\mathbf{x}_t\\ &= \frac{1}{p(\mathbf{y}_1, ..., \mathbf{y}_t)} \int_{\mathbf{x}_t \in \mathbb{R}} p(\mathbf{y}_1, ..., \mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t) d\mathbf{x}_t\\ &= \frac{1}{p(\mathbf{y}_1, ..., \mathbf{y}_t)} \int_{\mathbf{x}_t \in \mathbb{R}} p(\mathbf{y}_1, ..., \mathbf{y}_t, \mathbf{x}_t)p(\mathbf{x}_{t+1}|\mathbf{x}_t) d\mathbf{x}_t\\ &= \int_{\mathbf{x}_t \in \mathbb{R}} p(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_t)p(\mathbf{x}_{t+1}|\mathbf{x}_t) d\mathbf{x}_t \end{aligned}$$

To finish the computation, we need to marginalize $\mathbf{x}_t$.

$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t) = \int_{\mathbf{x}_t \in \mathbb{R}} \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\mathcal{N}(\mathbf{x}_{t+1}|\mathbf{A}\mathbf{x}_t, \mathbf{Q})d\mathbf{x}_t$$

Recall that marginalizing a variable from a Gaussian distribution results to another Gaussian distribution. First, let us denote the exponential terms of the product of these two Gaussians.

$$-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_t)^T\boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_t) - \frac{1}{2}(\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)^T\mathbf{Q}^{-1}(\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}_t)$$

First, we will expand the products and keep the quadratic and linear terms with respect to $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$.

$$\begin{aligned}
&= -\frac{1}{2}\mathbf{x}_t^T\boldsymbol{\Sigma}_t^{-1}\mathbf{x}_t - \frac{1}{2}\mathbf{x}_t^T\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\mathbf{x}_t - \frac{1}{2}\mathbf{x}_{t+1}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1} + \mathbf{x}_t^T\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1} + \mathbf{x}_t^T\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t + const \\
&= -\frac{1}{2}\mathbf{x}_t^T\Big(\boldsymbol{\Sigma}_t^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)\mathbf{x}_t + \mathbf{x}_t^T\Big(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1}\Big) - \frac{1}{2}\mathbf{x}_{t+1}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1} + const
\end{aligned}$$

We denote $\boldsymbol{\Lambda} = \Big(\boldsymbol{\Sigma}_t^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)$ and $\mathbf{m} = \Big(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1}\Big)$. Let us now rearrange the terms of $\mathbf{x}_t$ so that we can compute the integral. More specifically, let us consider just the terms that depend on $\mathbf{x}_t$ using the two definitions, $\boldsymbol{\Lambda}$ and $\mathbf{m}$.

$$-\frac{1}{2}\mathbf{x}_t^T\boldsymbol{\Lambda}\mathbf{x}_t + \mathbf{x}_t^T\mathbf{m} = -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\Lambda}^{-1}\mathbf{m})^T\boldsymbol{\Lambda}(\mathbf{x}_t - \boldsymbol{\Lambda}^{-1}\mathbf{m}) + \frac{1}{2}\mathbf{m}^T\boldsymbol{\Lambda}^{-1}\mathbf{m}$$

Notice that we can now integrate all the terms. The ones which depend on $\mathbf{x}_t$ will disappear as they are part of the integral of a normalized distribution, which translates as a zero in the exponential terms. The rest of the terms will remain unchanged. The last step of this computation is to complete the square with respect to x_{t+1} using the rest of the terms. We have

$$\begin{aligned}
&= \frac{1}{2}\mathbf{m}^T\boldsymbol{\Lambda}^{-1}\mathbf{m} - -\frac{1}{2}\mathbf{x}_{t+1}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1} + const \\
&= \frac{1}{2}\Big(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1}\Big)^T\boldsymbol{\Lambda}^{-1}\Big(\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1}\Big) - \frac{1}{2}\mathbf{x}_{t+1}^T\mathbf{Q}^{-1}\mathbf{x}_{t+1} + const \\
&= -\frac{1}{2}\mathbf{x}_{t+1}^T\Big(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\mathbf{Q}^{-1}\Big)\mathbf{x}_{t+1} + \mathbf{x}_{t+1}\mathbf{Q}^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t + const
\end{aligned}$$

We can extract both the covariance and mean from the quadratic and linear terms respectively. The covariance matrix is the following

$$\boldsymbol{\Gamma}_{t+1} = \Big(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\mathbf{Q}^{-1}\Big)^{-1}$$

We can simplify the previous expression by making use of the Woodbury matrix identity (you can google search it if you are curious). We will use the following identity.

$$(\mathbf{A} - \mathbf{U}\mathbf{C}\mathbf{V})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} - \mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}\mathbf{A}^{-1}$$

with $\mathbf{A} = \mathbf{Q}^{-1}$, $\mathbf{U} = \mathbf{Q}^{-1}\mathbf{A}$, $\mathbf{C} = \boldsymbol{\Lambda}^{-1}$, and $\mathbf{V} = \mathbf{A}^T\mathbf{Q}^{-1}$ we can simplify the expression of the covariance matrix.

$$\begin{aligned}
\boldsymbol{\Gamma}_{t+1} &= \Big(\mathbf{Q}^{-1} - \mathbf{Q}^{-1}\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\mathbf{Q}^{-1}\Big) \\
&= \mathbf{Q} + \mathbf{Q}\mathbf{Q}^{-1}\mathbf{A}\Big(\boldsymbol{\Sigma}_t^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A} - \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\mathbf{A}^T \\
&= \mathbf{Q} + \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T
\end{aligned}$$

The mean is computed using the previous result and the linear terms w.r.t. $\mathbf{x}_{t+1}$.

$$\begin{aligned}\mathbf{m}_{t+1} &= \Big(\mathbf{Q} + \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T\Big)\mathbf{Q}^{-1}\mathbf{A}\Big(\boldsymbol{\Sigma}_t^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t \\ &= \Big(\mathbf{A} + \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)\Big(\boldsymbol{\Sigma}_t^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t \\ &= \mathbf{A}\Big(\mathbf{I} + \boldsymbol{\Sigma}_t\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)\Big(\boldsymbol{\Sigma}_t^{-1}(\mathbf{I} + \boldsymbol{\Sigma}_t\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A})\Big)^{-1}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t \\ &= \mathbf{A}\Big(\mathbf{I} + \boldsymbol{\Sigma}_t\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)\Big(\mathbf{I} + \boldsymbol{\Sigma}_t\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\mu}_t \\ &= \mathbf{A}\boldsymbol{\mu}_t\end{aligned}$$

Therefore, the marginal distribution $p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)$ is distributed as follows.

$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t) = \mathcal{N}\Big(\mathbf{x}_{t+1}|\mathbf{A}\boldsymbol{\mu}_t, \mathbf{Q} + \mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T\Big)$$

(b) Compute $p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)$.

$$\begin{aligned}p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t) &= p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{y}_1, ..., \mathbf{y}_t) \\ &= p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})\end{aligned}$$

where we use $p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}) = p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathbf{y}_1, ..., \mathbf{y}_t)$ given the conditional independencies of the data generative process. We will not write down more derivations to this value, since this is just the joint distribution of two gaussians. We can simply denote it as the following product.

$$p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t) = \mathcal{N}\Big(\mathbf{x}_{t+1}|\mathbf{A}\boldsymbol{\mu}_t, \boldsymbol{\Gamma}_{t+1}\Big)\mathcal{N}\Big(\mathbf{y}_{t+1}|\mathbf{C}\mathbf{x}_{t+1}, R\Big), \quad \boldsymbol{\Gamma}_{t+1} = \mathbf{Q}+\mathbf{A}\boldsymbol{\Sigma}_t\mathbf{A}^T$$

(c) Compute $p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_{t+1})$.

$$\begin{aligned}p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_{t+1}) &= \frac{1}{p(\mathbf{y}_1, ..., \mathbf{y}_{t+1})}p(\mathbf{x}_{t+1}, \mathbf{y}_1, ..., \mathbf{y}_{t+1}) \\ &= \frac{1}{p(\mathbf{y}_1, ..., \mathbf{y}_{t+1})}p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)p(\mathbf{y}_1, ..., \mathbf{y}_t) \\ &= \frac{p(\mathbf{y}_1, ..., \mathbf{y}_t)}{p(\mathbf{y}_1, ..., \mathbf{y}_{t+1})}p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)\end{aligned}$$

As hard as this expression might seem, we can solve this by recognizing the expression as a posterior distribution, which is in fact proportional to the joint distribution as follows

$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_{t+1}) \propto p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)$$

We only need to complete the square with respect to $\mathbf{x}_{t+1}$ and we are done. Let us denote the exponential terms of the joint probability $p(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_t)$.

$$-\frac{1}{2}(\mathbf{x}_{t+1} - \mathbf{A}\boldsymbol{\mu}_t)^T\boldsymbol{\Gamma}_{t+1}^{-1}(\mathbf{x}_{t+1} - \mathbf{A}\boldsymbol{\mu}_t) - \frac{1}{2}(\mathbf{y}_{t+1} - \mathbf{C}\mathbf{x}_{t+1})^T\mathbf{R}^{-1}(\mathbf{y}_{t+1} - \mathbf{C}\mathbf{x}_{t+1})$$

We just need to keep the quadratic and linear terms with respect to $\mathbf{x}_{t+1}$.

$$\begin{aligned}&= -\frac{1}{2}\mathbf{x}_{t+1}^T\boldsymbol{\Gamma}_{t+1}^{-1}\mathbf{x}_{t+1} - \frac{1}{2}\mathbf{x}_{t+1}^T\mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\mathbf{x}_{t+1} + \mathbf{x}_{t+1}^T\boldsymbol{\Gamma}_{t+1}^{-1}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{x}_{t+1}^T\mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1} + const \\ &= -\frac{1}{2}\mathbf{x}_{t+1}^T\Big(\boldsymbol{\Gamma}_{t+1}^{-1} + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\Big)\mathbf{x}_{t+1} + \mathbf{x}_{t+1}^T\Big(\boldsymbol{\Gamma}_{t+1}^{-1}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1}\Big) + const\end{aligned}$$

We can extract the covariance matrix from the quadratic term

$$\mathbf{\Sigma}_{t+1} = \Big(\mathbf{\Gamma}_{t+1}^{-1} + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\Big)^{-1}$$

Since these results are familiar with the Kalman filter equations, we will continue developing them until we obtain the original results. We can start by using the Woodbury matrix identity

$$\Big(\mathbf{\Gamma}_{t+1}^{-1} + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\Big)^{-1} = \mathbf{\Gamma}_{t+1} - \mathbf{\Gamma}_{t+1}\mathbf{C}^T\Big(\mathbf{R} + \mathbf{C}\mathbf{\Gamma}_{t+1}\mathbf{C}^T\Big)^{-1}\mathbf{C}\mathbf{\Gamma}_{t+1}$$

We identify the *Kalman gain matrix* $\mathbf{K}_{t+1}$, which is expressed as follows

$$\mathbf{K}_{t+1} = \mathbf{\Gamma}_{t+1}\mathbf{C}^T\Big(\mathbf{R} + \mathbf{C}\mathbf{\Gamma}_{t+1}\mathbf{C}^T\Big)^{-1}$$

Using this definition, we can find the original expression of the covariance matrix.

$$\begin{aligned}\mathbf{\Sigma}_{t+1} &= \mathbf{\Gamma}_{t+1} - \mathbf{K}_{t+1}\mathbf{C}\mathbf{\Gamma}_{t+1}\\ &= \Big(\mathbf{I} - \mathbf{K}_{t+1}\mathbf{C}\Big)\mathbf{\Gamma}_{t+1}\end{aligned}$$

Once we find the covariance, we can calculate the mean using the previous result and the linear terms w.r.t $\mathbf{x}_{t+1}$

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &= \mathbf{\Sigma}_{t+1}\Big(\mathbf{\Gamma}_{t+1}^{-1}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1}\Big)\\ &= \mathbf{\Sigma}_{t+1}\mathbf{\Gamma}_{t+1}^{-1}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{\Sigma}_{t+1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1}\\ &= \Big(\mathbf{I} - \mathbf{K}_{t+1}\mathbf{C}\Big)\mathbf{\Gamma}_{t+1}\mathbf{\Gamma}_{t+1}^{-1}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{\Sigma}_{t+1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1}\\ &= \mathbf{A}\boldsymbol{\mu}_t - \mathbf{K}_{t+1}\mathbf{C}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{\Sigma}_{t+1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1}\\ &= \mathbf{A}\boldsymbol{\mu}_t - \mathbf{K}_{t+1}\mathbf{C}\mathbf{A}\boldsymbol{\mu}_t + \Big(\mathbf{\Gamma}_{t+1}^{-1} + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\Big)^{-1}\mathbf{C}^T\mathbf{R}^{-1}\mathbf{y}_{t+1}\\ &= \mathbf{A}\boldsymbol{\mu}_t - \mathbf{K}_{t+1}\mathbf{C}\mathbf{A}\boldsymbol{\mu}_t + \mathbf{K}_{t+1}\mathbf{y}_{t+1}\\ &= \mathbf{A}\boldsymbol{\mu}_t + \mathbf{K}_{t+1}\Big(\mathbf{y}_{t+1} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_t\Big)\end{aligned}$$

where we have a massive simplification on the terms over $\mathbf{y}_{t+1}$ if we use the following well-known identity: $(\mathbf{A}^{-1}+\mathbf{B}^T\mathbf{D}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{D}^{-1} = \mathbf{A}\mathbf{B}^T(\mathbf{B}\mathbf{A}\mathbf{B}^T+\mathbf{D})^{-1}$. Consequently,

$$\Big(\mathbf{\Gamma}_{t+1}^{-1} + \mathbf{C}^T\mathbf{R}^{-1}\mathbf{C}\Big)^{-1}\mathbf{C}^T\mathbf{R}^{-1} = \mathbf{\Gamma}_{t+1}\mathbf{C}^T\Big(\mathbf{R} + \mathbf{C}\mathbf{\Gamma}_{t+1}\mathbf{C}^T\Big)^{-1} = \mathbf{K}_{t+1}$$

After computing the mean and the covariance, we conclude that the posterior distribution $p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_{t+1})$ has the following form

$$p(\mathbf{x}_{t+1}|\mathbf{y}_1, ..., \mathbf{y}_{t+1}) = \mathcal{N}(\mathbf{x}_{t+1}|\boldsymbol{\mu}_{t+1}, \mathbf{\Sigma}_{t+1})$$

$$\boldsymbol{\mu}_{t+1} = \mathbf{A}\boldsymbol{\mu}_t + \mathbf{K}_{t+1}\Big(\mathbf{y}_{t+1} - \mathbf{C}\mathbf{A}\boldsymbol{\mu}_t\Big)$$

$$\mathbf{\Sigma}_{t+1} = \Big(\mathbf{I} - \mathbf{K}_{t+1}\mathbf{C}\Big)\mathbf{\Gamma}_{t+1}$$

Exercise 6.12 Consider the random variable $\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \mathbf{\Sigma}_x)$ and the following linear model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{w}$$

where $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{Q})$.

a. Compute the likelihood $p(y|x)$. To calculate the likelihood, we need to recall that the addition of Gaussian distributions is also a Gaussian distribution. Thus, it is sufficient for us to compute the expectation and covariance of $\mathbf{y}$ given $\mathbf{x}$.

$$\mathbb{E}[\mathbf{y}|\mathbf{x}] = \mathbb{E}[\mathbf{Ax} + \mathbf{b} + \mathbf{w}|\mathbf{x}] = \mathbf{Ax} + \mathbf{b} + \mathbf{0} = \mathbf{Ax} + \mathbf{b}$$

we note that $\mathbb{E}[\mathbf{w}|\mathbf{x}] = \mathbb{E}[\mathbf{w}] = \mathbf{0}$ because of the independence of $\mathbf{x}$ and $\mathbf{w}$.

$$\begin{aligned} cov[\mathbf{y}|\mathbf{x}] &= \mathbb{E}\Big[\big(\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{x}]\big)\big(\mathbf{y} - \mathbb{E}[\mathbf{y}|\mathbf{x}]\big)^T\Big] = \mathbb{E}\Big[\big(\mathbf{Ax} + \mathbf{b} + \mathbf{w} - \mathbf{Ax} - \mathbf{b}\big)\big(\mathbf{Ax} + \mathbf{b} + \mathbf{w} - \mathbf{Ax} - \mathbf{b}\big)^T\Big] \\ &= \mathbb{E}\Big[\mathbf{w}\mathbf{w}^T\Big] = \mathbf{Q} \end{aligned}$$

$p(\mathbf{y}|\mathbf{x})$ is a Gaussian distribution, $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{Q})$.

b. Compute the marginal distribution $p(\mathbf{y})$.

This result is analogous to **Ex. 6.5.b.b.**. If you integrate by first taking into account the terms w.r.t $\mathbf{x}$ and then complete the square w.r.t $\mathbf{y}$ you will get to the following marginal distribution.

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}, \mathbf{Q} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$

c. We perform the following transformation to $\mathbf{y}$

$$\mathbf{z} = \mathbf{Cy} + \mathbf{v}, \quad \mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$$

Let us compute $p(\mathbf{z}|\mathbf{y})$. As before, we notice that the addition of Gaussians is also a Gaussian.

$$\mathbb{E}[\mathbf{z}|\mathbf{y}] = \mathbb{E}[\mathbf{Cy} + \mathbf{v}|\mathbf{y}] = \mathbf{Cy} + \mathbf{0} = \mathbf{Cy}$$

$$\begin{aligned} cov[\mathbf{z}|\mathbf{y}] &= \mathbb{E}\Big[\big(\mathbf{z} - \mathbb{E}[\mathbf{z}|\mathbf{y}]\big)\big(\mathbf{z} - \mathbb{E}[\mathbf{z}|\mathbf{y}]\big)^T\Big] = \mathbb{E}\Big[\big(\mathbf{Cy} + \mathbf{v} - \mathbf{Cy}\big)\big(\mathbf{Cy} + \mathbf{v} - \mathbf{Cy}\big)^T\Big] \\ &= \mathbb{E}\Big[\mathbf{v}\mathbf{v}^T\Big] = \mathbf{R} \end{aligned}$$

With this information we can deduce that $p(\mathbf{z}|\mathbf{y})$ is distributed according to $\mathbf{z}|\mathbf{y} \sim \mathcal{N}(\mathbf{z}|\mathbf{Cy}, \mathbf{R})$.

The marginal distribution $p(\mathbf{z})$ can be obtained in a similar way as previously done, analogous to **Ex. 6.5.b.b.** as well. Just by using the information about the marginal $p(\mathbf{y})$ we can marginalize $\mathbf{z}$ from the posterior distribution $p(\mathbf{y}, \mathbf{z}) = p(\mathbf{z}|\mathbf{y})p(\mathbf{y})$. If you perform the usual steps: first integrating out $\mathbf{y}$ and then completing the square w.r.t $\mathbf{z}$, you will reach to the following result.

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{C}\boldsymbol{\mu}_y, \mathbf{R} + \mathbf{C}\boldsymbol{\Sigma}_y\mathbf{C}^T) = \mathcal{N}(\mathbf{z}|\mathbf{CA}\boldsymbol{\mu}_x + \mathbf{Cb}, \mathbf{R} + \mathbf{CQC}^T + \mathbf{CA}\boldsymbol{\Sigma}_x\mathbf{A}^T\mathbf{C}^T)$$

d. Compute the posterior distribution $p(\mathbf{x}|\mathbf{y})$.

Although it might seem complicated to compute, the posterior distribution for this case is fairly simple. We start by noticing the usual relationship derived from the Bayes' rule.

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{x})p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y})} \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$$

As we can see, the joint distribution is the unnormalized distribution of the posterior. Just by completing the square with respect to $\mathbf{x}$, one can find the normalized distribution. Let us denote the exponential term in the joint distribution.

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) - \frac{1}{2}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})^T\mathbf{Q}^{-1}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})$$

We will take the quadratic and linear terms with respect to $\mathbf{x}$.

$$-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}_x^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\mathbf{x} + \mathbf{x}^T\mathbf{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \mathbf{x}^T\mathbf{A}^T\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{b}) + const$$
$$-\frac{1}{2}\mathbf{x}^T\Big(\mathbf{\Sigma}_x^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)\mathbf{x} + \mathbf{x}^T\Big(\mathbf{A}^T\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{b}) + \mathbf{\Sigma}_x^{-1}\boldsymbol{\mu}_x\Big) + const$$

When completing the square, we can easily extract the mean and covariance terms from the linear and quadratic forms of the exponential respectively. Thus

$$cov[\mathbf{x}|\mathbf{y}] = \Big(\mathbf{\Sigma}_x^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}$$

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = cov[\mathbf{x}|\mathbf{y}]\Big(\mathbf{A}^T\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{b})+\mathbf{\Sigma}_x^{-1}\boldsymbol{\mu}_x\Big) = \Big(\mathbf{\Sigma}_x^{-1}+\mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\Big(\mathbf{A}^T\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{b})+\mathbf{\Sigma}_x^{-1}\boldsymbol{\mu}_x\Big)$$

As we can see, the posterior distribution is a Gaussian expressed as

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\Big(\mathbf{x}|\Big(\mathbf{\Sigma}_x^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\Big(\mathbf{A}^T\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{b}) + \mathbf{\Sigma}_x^{-1}\boldsymbol{\mu}_x\Big), \Big(\mathbf{\Sigma}_x^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A}\Big)^{-1}\Big)$$