UNIVERSITY OF AMSTERDAM

University of Amsterdam

Homework Assignment 1

Machine Learning I

2024

Pedro M. P. Curvo
15713725

Contents

1 Multivariate Calculus

a)

$\boldsymbol{\sigma}$ is applied to the vector $\boldsymbol{x}$ element-wise, meaning $\sigma(\boldsymbol{x})_i = \sigma(x_i)$. Which implies that

$$\boldsymbol{\sigma}(\boldsymbol{x}) = \begin{bmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_n) \end{bmatrix} = \begin{bmatrix} \frac{1}{1+e^{-x_1}} \\ \vdots \\ \frac{1}{1+e^{-x_n}} \end{bmatrix}.$$

Then, the derivative of $\sigma(x)_i$ with respect to x_k is

$$\begin{aligned} \frac{\partial}{\partial x_k}\sigma(x)_i &= \frac{\partial}{\partial x_k}\frac{1}{1+e^{-x_i}} \\ &= \frac{e^{-x_i}}{(1+e^{-x_i})^2}\delta_{ik} \\ &= \frac{e^{-x_i}+1-1}{(1+e^{-x_i})^2}\delta_{ik} \\ &= \left(\frac{1}{1+e^{-x_i}} - \frac{1}{(1+e^{-x_i})^2}\right)\delta_{ik} \\ &= \left(\sigma(x)_i - \sigma(x)_i^2\right)\delta_{ik} \\ &= \sigma(x)_i\left(1-\sigma(x)_i\right)\delta_{ik} \end{aligned}$$

Thus, we can say that

$$\nabla_{\boldsymbol{x}}\boldsymbol{\sigma} = \begin{bmatrix} \sigma(x)_1\left(1-\sigma(x)_1\right) & 0 & \cdots & 0 \\ 0 & \sigma(x)_2\left(1-\sigma(x)_2\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma(x)_n\left(1-\sigma(x)_n\right) \end{bmatrix} = \operatorname{diag}(\boldsymbol{\sigma}(\boldsymbol{x})) - \operatorname{diag}(\boldsymbol{\sigma}^2(\boldsymbol{x})).$$

b)

The function $\boldsymbol{f}$ is defined as $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w}$, where $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$.
We can then say that $f_i = \sum_{j=1}^{n} X_{ij} w_j$. Hence, if we derivate f_i with respect to w_k, we get:

$$\frac{\partial f_i}{\partial w_k} = \frac{\partial}{\partial x_k} \sum_{j=1}^{n} X_{ij} w_j = X_{ik}.$$

Therefore,

$$\nabla_{\boldsymbol{w}} \boldsymbol{f} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nn} \end{bmatrix} = \boldsymbol{X}.$$

c)

The function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$ is given as $\boldsymbol{f} = \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{w}$. We can also express this as:

$$\boldsymbol{f} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i X_{ij} w_j.$$

If we differentiate $\boldsymbol{f}$ with respect to w_k, we obtain:

$$\begin{aligned}
\frac{\partial f}{\partial w_k} &= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial}{\partial w_k} w_i X_{ij} w_j \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} (\delta_{ik} X_{ij} w_j + w_i X_{ij} \delta_{jk}) \\
&= \sum_{j=1}^{n} X_{kj} w_j + \sum_{i=1}^{n} X_{ik} w_i
\end{aligned}$$

Therefore, the Jacobian of $\boldsymbol{f}$ with respect to $\boldsymbol{w}$ is:

$$\nabla_{\boldsymbol{w}} f = \boldsymbol{w}^T (\boldsymbol{X} + \boldsymbol{X}^T).$$

d)

We know that

$$\boldsymbol{\varsigma}(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}.$$

And that

$$\frac{\partial \boldsymbol{\varsigma}(\boldsymbol{x})}{\partial \boldsymbol{x}} \in \mathbb{R}^{n \times n}.$$

Taking the derivative of $\boldsymbol{\varsigma}(\boldsymbol{x})_i$ with respect to x_k, we obtain:

$$\begin{aligned}
\frac{\partial \varsigma(x)_i}{\partial x_k} &= \frac{\partial}{\partial x_k} \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \\
&= \frac{\delta_{ik} e^{x_i} \sum_{j=1}^{n} e^{x_j} - e^{x_i} e^{x_k}}{\left(\sum_{j=1}^{n} e^{x_j}\right)^2} \\
&= \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \left(\delta_{ik} - \frac{e^{x_k}}{\sum_{j=1}^{n} e^{x_j}} \right) \\
&= \varsigma(x)_i \left(\delta_{ik} - \varsigma(x)_k \right).
\end{aligned}$$

Therefore, the gradient of $\boldsymbol{\varsigma}(\boldsymbol{x})$ with respect to x is:

$$\begin{aligned}
\nabla_{\boldsymbol{x}} \boldsymbol{\varsigma}(\boldsymbol{x}) &= \begin{bmatrix} \varsigma(x)_1 \left(1 - \varsigma(x)_1\right) & -\varsigma(x)_1 \varsigma(x)_2 & \cdots & -\varsigma(x)_1 \varsigma(x)_n \\ -\varsigma(x)_2 \varsigma(x)_1 & \varsigma(x)_2 \left(1 - \varsigma(x)_2\right) & \cdots & -\varsigma(x)_2 \varsigma(x)_n \\ \vdots & \vdots & \ddots & \vdots \\ -\varsigma(x)_n \varsigma(x)_1 & -\varsigma(x)_n \varsigma(x)_2 & \cdots & \varsigma(x)_n \left(1 - \varsigma(x)_n\right) \end{bmatrix} \\
&= \begin{bmatrix} \varsigma(x)_1 - \varsigma(x)_1^2 & -\varsigma(x)_1 \varsigma(x)_2 & \cdots & -\varsigma(x)_1 \varsigma(x)_n \\ -\varsigma(x)_2 \varsigma(x)_1 & \varsigma(x)_2 - \varsigma(x)_2^2 & \cdots & -\varsigma(x)_2 \varsigma(x)_n \\ \vdots & \vdots & \ddots & \vdots \\ -\varsigma(x)_n \varsigma(x)_1 & -\varsigma(x)_n \varsigma(x)_2 & \cdots & \varsigma(x)_n - \varsigma(x)_n^2 \end{bmatrix} \\
&= \begin{bmatrix} \varsigma(x)_1 & 0 & \cdots 0 & \\ 0 & \varsigma(x)_2 & \cdots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \varsigma(x)_n \end{bmatrix} - \begin{bmatrix} \varsigma(x)_1^2 & \varsigma(x)_1 \varsigma(x)_2 & \cdots & \varsigma(x)_1 \varsigma(x)_n \\ \varsigma(x)_2 \varsigma(x)_1 & \varsigma(x)_2^2 & \cdots & \varsigma(x)_2 \varsigma(x)_n \\ \vdots & \vdots & \ddots & \vdots \\ \varsigma(x)_n \varsigma(x)_1 & \varsigma(x)_n \varsigma(x)_2 & \cdots & \varsigma(x)_n^2 \end{bmatrix} \\
&= \operatorname{diag}(\boldsymbol{\varsigma}(\boldsymbol{x})) - \boldsymbol{\varsigma}(\boldsymbol{x}) \boldsymbol{\varsigma}(\boldsymbol{x})^T.
\end{aligned}$$

e)

Defining $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}$ as $\boldsymbol{f} = \|\boldsymbol{X\theta} - \boldsymbol{y}\|_2^2$, with $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{\theta} \in \mathbb{R}^d, \boldsymbol{y} \in \mathbb{R}^n$, we can expand this as:

$$f = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} X_{ij}\theta_j - y_i \right)^2 .$$

If we differentiate f with respect to θ_k, we obtain:

$$\begin{aligned}
\frac{\partial f}{\partial \theta_k} &= \frac{\partial}{\partial \theta_k} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} X_{ij}\theta_j - y_i \right)^2 \\
&= 2 \sum_{i=1}^{n} \left(\sum_{j=1}^{n} (X_{ij}\theta_j) - y_i \right) X_{ik} \\
&= 2 \sum_{i=1}^{n} X_{ik} \left([X\theta]_i - y_i \right). \\
&= 2[[X\theta - y]^T X]_k. \\
&= 2[(\theta^T X^T - y^T)X]_k. \\
&= 2[\theta^T X^T X - y^T X]_k.
\end{aligned}$$

We can then rewrite this as:

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{f} = 2(\boldsymbol{\theta}^T \boldsymbol{X}^T \boldsymbol{X} - \boldsymbol{y}^T \boldsymbol{X}).$$

If we then set the gradient to zero, we obtain:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \boldsymbol{f} &= 0 \\
2(\boldsymbol{\theta}^T \boldsymbol{X}^T \boldsymbol{X} - \boldsymbol{y}^T \boldsymbol{X}) &= 0 \\
\boldsymbol{\theta}^T \boldsymbol{X}^T \boldsymbol{X} &= \boldsymbol{y}^T \boldsymbol{X} \\
\boldsymbol{\theta}^T &= \boldsymbol{y}^T \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \quad \text{(if } \boldsymbol{X}^T \boldsymbol{X} \text{ is invertible, i.e, if all features are independent)} \\
\boldsymbol{\theta} &= (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T y.
\end{aligned}$$

Note: $\left((\boldsymbol{X}^T \boldsymbol{X}\right)^{-1})^T = (\boldsymbol{X}^T \boldsymbol{X})^{-1}$ because $\boldsymbol{X}^T \boldsymbol{X}$ is a symmetric matrix, hence its inverse is also symmetric.

f)

The formula we obtained in the previous question fails if $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible. This can happen if the columns of $\boldsymbol{X}$ are linearly dependent. In this case, the matrix $\boldsymbol{X}^T\boldsymbol{X}$ will have a zero determinant and, therefore, will not be invertible. Another problem that can arise when computing the inverse of this matrix is numerical instability, i.e., if the determinant of the matrix is close to zero, the inverse can have large values [1] that can lead to overflow errors.

If we add the term $\lambda\|\boldsymbol{\theta}\|_2^2$ to the previous function, we obtain:

$$\begin{aligned} f &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij}\theta_j - y_i\right)^2 + \lambda\|\theta\|_2^2 \\ &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij}\theta_j - y_i\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2. \end{aligned}$$

If we differentiate f with respect to θ_k, we obtain:

$$\begin{aligned} \frac{\partial f}{\partial \theta_k} &= \frac{\partial}{\partial w_k}(\sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij}\theta_j - y_i\right)^2 + \lambda\sum_{j=1}^{n}\theta_j^2) \\ &= 2\sum_{i=1}^{n}\left(\sum_{j=1}^{n} X_{ij}\theta_j - y_i\right)X_{ik} + 2\lambda\theta_k \\ &= 2\sum_{i=1}^{n} X_{ik}\left([Xw]_i - y_i\right) + 2\lambda\theta_k \\ &= 2[[Xw - y]^T X]_k + 2\lambda\theta_k \\ &= 2[(w^T X^T - y^T)X]_k + 2\lambda\theta_k \\ &= 2[w^T X^T X - y^T X]_k + 2\lambda\theta_k \end{aligned}$$

We can then rewrite this as:

$$\nabla_{\boldsymbol{\theta}} f = 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta} - y) + 2\lambda\boldsymbol{\theta}.$$

If we then set the gradient to zero, we obtain:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} f &= 0 \\ 2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) + 2\lambda\boldsymbol{\theta} &= 0 \\ \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}^T\boldsymbol{y} + \lambda\boldsymbol{\theta} &= 0 \\ \boldsymbol{X}^T\boldsymbol{y} &= \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} + \lambda\boldsymbol{\theta} \\ \boldsymbol{\theta} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \end{aligned}$$

[1] $A^{-1} = \frac{1}{|A|}Adj(A)$

This term solves the problem of the matrix $\boldsymbol{X}^T\boldsymbol{X}$ not being invertible. In simple terms, $\boldsymbol{X}^T\boldsymbol{X}$ is either Positive Definite or Positive Semi-Definite. If it's Positive Definite, it's invertible. If it's Positive Semi-Definite, then the smallest eigenvalue might be zero, which is also the one that causes the matrix to be singular. By adding the term $\lambda\boldsymbol{I}$, we ensure that the smallest eigenvalue is greater than zero, making the matrix invertible, because we shift all the eigenvalues by λ (see Eq. (1)). It also helps with numerical stability, as it ensures that the matrix is not close to being singular, depending on the value of λ.

$$\begin{aligned} \det(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I} - \lambda_{\text{eigen}}\boldsymbol{I}) = 0, \quad &\text{(definition of eigenvalues)} \\ \det(\boldsymbol{X}^T\boldsymbol{X} - \lambda'\boldsymbol{I}) = 0 \quad &\text{(taking } \lambda' = \lambda_{\text{eigen}} - \lambda) \end{aligned} \tag{1}$$

which gives the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$. Thus, we see that the eigenvalues are shifted by λ. Since $\lambda' \geq 0$ and $\lambda > 0$, we can say that the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}$ are greater than or equal to λ, which makes the matrix invertible.

2 Full analysis of a distribution: Exponential distribution

a)

To compute the expectancy of the exponential distribution, we can use the formula:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx.$$

Hence, for the exponential distribution, we have:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^{\infty} x\lambda e^{-\lambda x} dx \\
&= -\left(\int_0^{\infty} x(-\lambda e^{-\lambda x}) dx \right) \\
&= -\left(\left[x e^{-\lambda x} \right]_0^{\infty} - \int_0^{\infty} e^{-\lambda x} dx \right) \\
&= -\left(0 - 0 - \left[-\frac{1}{\lambda} e^{-\lambda x} \right]_0^{\infty} \right) \\
&= \frac{1}{\lambda} (0 + 1) \\
&= \frac{1}{\lambda}.
\end{aligned}$$

b)

As demonstrated in the previous question, the expectancy of the exponential distribution is $\frac{1}{\lambda}$, which implies that $\mu = \frac{1}{\lambda}$.

In this question, a student arrives in average every 2 minutes, which implies that $\lambda = \frac{1}{2}$.

Therefore,

$$P(X \leq x) = \begin{cases} 1 - e^{-\frac{x}{2}} & \text{, if } x \geq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{2}$$

Hence, the probability that the time between two arrivals is less than 1 minute is given by:

$$P(X \leq 1) = 1 - e^{-\frac{1}{2}} = 1 - e^{-0.5} \approx 0.3935.$$

c)

The PDF of the exponential distribution is given, for $\lambda \geq 0$, as:

$$p(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{, if } x =\geq 0 \\ 0 & \text{, otherwise} \end{cases} \tag{3}$$

Since the dataset comprises independent and identically distributed (i.i.d.) samples, the likelihood is the product of the PDFs of each sample. Therefore, the likelihood of the dataset is given by:

$$p(\boldsymbol{x}|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

The log-likelihood is then:

$$\log p(\boldsymbol{x}|\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

d)

To find the maximum likelihood estimate (MLE) of the parameter λ, we need to find the value of λ that maximizes the log-likelihood function. To do this, we can take the derivative of the log-likelihood function with respect to λ and set it to zero:

$$\begin{aligned}
\frac{\partial}{\partial \lambda} \log p(\boldsymbol{x}|\lambda) &= 0 \\
\frac{n}{\lambda} - \sum_{i=1}^{n} x_i &= 0 \\
\frac{n}{\lambda} &= \sum_{i=1}^{n} x_i \\
\lambda &= \frac{n}{\sum_{i=1}^{n} x_i} \\
\lambda &= \frac{1}{\mu},
\end{aligned}$$

where μ is the mean of the dataset ($\mu = \bar{x}$).

Thus,

$$\lambda_{ML} = \frac{1}{\mu}$$

e)

By Bayes' theorem, the posterior distribution of λ given the data is:

$$p(\lambda|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\lambda)p(\lambda)}{p(\boldsymbol{x})},$$

where $p(\lambda)$ is the prior distribution of λ and $p(\boldsymbol{x})$ is the marginal likelihood of the data.

To find the MAP estimate of λ, we need to find the value of λ that maximizes the posterior distribution. To do this, we can take the derivative of the posterior distribution with respect to λ and set it to zero, which is equivalent to:

$$\lambda_{\text{MAP}} = argmax_{\lambda} p(\lambda|\boldsymbol{x})$$

Since we are differentiating with respect to λ, we can ignore the denominator $p(\boldsymbol{x})$ as it does not depend on λ, i.e, when we differentiate and set to zero, this value will vanish. Therefore, the MAP estimate of λ is given by:

$$\lambda_{\text{MAP}} = argmax_{\lambda} p(\mathbf{x}|\lambda)p(\lambda)$$

Then we can take the logarithm of the posterior distribution and differentiate it with respect to λ. Because the logarithm is a monotonically increasing function, the maximum of the posterior distribution will be the same as the maximum of the logarithm of the posterior distribution.

Hence, the MAP estimate of λ is given by:

$$\begin{aligned}\lambda_{\text{MAP}} &= argmax_{\lambda} \log p(\boldsymbol{x}|\lambda)p(\lambda) \\ &= argmax_{\lambda} \log p(\mathbf{x}|\lambda) + \log p(\lambda)\end{aligned}$$

Now, knowing that the prior for the parameter λ is given by the Gamma distribution with hyper-parameters α_1 and α_2:

$$p(\lambda|\alpha_1, \alpha_2) = \text{Gamma}(\lambda|\alpha_1, \alpha_2) = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}\lambda^{\alpha_1 - 1}e^{-\alpha_2\lambda},$$

we can then write the MAP estimate of λ as:

$$\begin{aligned}\lambda_{\text{MAP}} &= argmax_{\lambda} \log p(\mathbf{x}|\lambda) + \log p(\lambda) \\ &= argmax_{\lambda} \log \left(\lambda^n e^{-\lambda \sum_{i=1}^n x_i}\right) + \log \left(\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}\lambda^{\alpha_1 - 1}e^{-\alpha_2\lambda}\right) \\ &= argmax_{\lambda} n \log \lambda - \lambda \sum_{i=1}^{n} x_i + (\alpha_1 - 1)\log \lambda - \alpha_2\lambda \\ &= argmax_{\lambda} (n + \alpha_1 - 1)\log \lambda - \lambda \left(\sum_{i=1}^{n} x_i + \alpha_2\right).\end{aligned}$$

f)

Using the expression from the previous question, we can differentiate the expression with respect to λ and set it to zero to find the MAP estimate of λ:

$$\begin{aligned}
\frac{\partial}{\partial \lambda}\left((n+\alpha_1-1)\log\lambda - \lambda\left(\sum_{i=1}^{n} x_i + \alpha_2\right)\right) &= 0 \\
\frac{n+\alpha_1-1}{\lambda} - \sum_{i=1}^{n} x_i - \alpha_2 &= 0 \\
\frac{n+\alpha_1-1}{\lambda} &= \sum_{i=1}^{n} x_i + \alpha_2 \\
\lambda &= \frac{n+\alpha_1-1}{\sum_{i=1}^{n} x_i + \alpha_2}.
\end{aligned}$$

Thus,

$$\lambda_{\text{MAP}} = \frac{n+\alpha_1-1}{\sum_{i=1}^{n} x_i + \alpha_2}.$$

g)

To show that the posterior distribution of an exponential distribution with a Gamma prior is also a Gamma distribution, we can write the posterior distribution as:

$$p(\lambda|\mathbf{x}, \alpha_1, \alpha_2) \propto p(\mathbf{x}|\lambda)p(\lambda|\alpha_1, \alpha_2).$$

The denominator of the posterior distribution is the marginal likelihood of the data, which is independent of λ. Therefore, we can ignore it when finding the form of the posterior distribution.

This leads to:

$$\begin{aligned} p(\lambda|\mathbf{x}, \alpha_1, \alpha_2) &\propto p(\mathbf{x}|\lambda)p(\lambda|\alpha_1, \alpha_2) \\ &\propto \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda} \\ &\propto \lambda^{n+\alpha_1-1} e^{-\lambda\left(\sum_{i=1}^n x_i + \alpha_2\right)}. \end{aligned}$$

This is the form of a Gamma distribution with parameters $n + \alpha_1$ and $\sum_{i=1}^n x_i + \alpha_2$. Moreover, if we think on the denominator of the posterior distribution that we ignored before, since its a constant and its only purpose is to normalize the posterior distribution, if the dependent part of the posterior distribution is a Gamma distribution, then the denominator must be the normalization constant of the Gamma distribution.

With this result, we can say that the Gamma distribution is a conjugate prior for the exponential distribution.

3 General Multiple Outputs Linear Regression

a)

We know that $\boldsymbol{y} \in \mathbb{R}^K$ and $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^M$. Therefore, we can conclude that $\mathbf{W}^T \in \mathbb{R}^{K \times M}$, so $\mathbf{W} \in \mathbb{R}^{M \times K}$.

b)

The likehood is given by:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X}, \boldsymbol{\Sigma}) &= \mathcal{N}\left(\mathbf{t}|\mathbf{y}(\mathbf{X}, \mathbf{W}), \boldsymbol{\Sigma}\right) \\ &= \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{t} - \mathbf{y}(\mathbf{X}, \mathbf{W}))^T \boldsymbol{\Sigma}^{-1}(\mathbf{t} - \mathbf{y}(\mathbf{X}, \mathbf{W}))\right) \end{aligned}$$

Considering we have a dataset with N samples, the likelihood is given by:

$$\begin{aligned} p(\mathbf{T}|\mathbf{X}, \boldsymbol{\Sigma}) &= \prod_{n=1}^{N} p(\boldsymbol{t}_n|\mathbf{X}, \boldsymbol{\Sigma}) \\ &= \prod_{n=1}^{N} \frac{1}{(2\pi)^{K/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))\right) \\ &= \frac{1}{(2\pi)^{NK/2}|\boldsymbol{\Sigma}|^{N/2}} \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T \boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))\right) \end{aligned}$$

Now, taking the logarithm of the likelihood, we obtain:

$$\begin{aligned} \log p(\mathbf{T}|\mathbf{X}, \boldsymbol{\Sigma}) &= -\frac{NK}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W}))^T \boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{W})) \\ &= -\frac{NK}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(x_n))^T \boldsymbol{\Sigma}^{-1}(\mathbf{t}_n - \mathbf{W}^T\boldsymbol{\phi}(x_n)) \end{aligned}$$

$\boldsymbol{\Sigma}$ is a positive semi-definite matrix, so its determinant is non-negative. Therefore, the log-likelihood is well-defined, assuming that $\boldsymbol{\Sigma}$ is invertible.

c)

To find the maximum likelihood estimate (MLE) of the parameter $\mathbf{W}$, we need to find the value of $\mathbf{W}$ that maximizes the log-likelihood function. To do this, we can take the derivative of the log-likelihood function with respect to $\mathbf{W}$ and set it to zero:

$$\frac{\partial}{\partial \mathbf{W}} \log p(\mathbf{T}|\mathbf{X}, \mathbf{\Sigma}) = 0$$

$$\frac{\partial}{\partial \mathbf{W}} \left(-\frac{NK}{2} \log(2\pi) - \frac{N}{2} \log |\mathbf{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{t}_n - \mathbf{W}^T \phi(\boldsymbol{x}_n))^T \mathbf{\Sigma}^{-1} (\mathbf{t}_n - \mathbf{W}^T \phi(\boldsymbol{x}_n)) \right) = 0$$

$$\frac{\partial}{\partial \mathbf{W}} \left(-\frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{t}_n - \mathbf{W}^T \phi(x_n))^T \mathbf{\Sigma}^{-1} (\boldsymbol{t}_n - \mathbf{W}^T \phi(\boldsymbol{x}_n)) \right) = 0$$

$$-\frac{1}{2} \frac{\partial}{\partial \mathbf{W}} \sum_{n=1}^{N} (\boldsymbol{t}_n - \mathbf{W}^T \phi(\boldsymbol{x}_n))^T \mathbf{\Sigma}^{-1} (\boldsymbol{t}_n - \mathbf{W}^T \phi(\boldsymbol{x}_n)) = 0$$

$$\sum_{n=1}^{N} \mathbf{\Sigma}^{-1} (\boldsymbol{t}_n - \mathbf{W}^T \phi(\boldsymbol{x}_n)) \phi(\boldsymbol{x}_n)^T = 0, \quad \text{using the Hint}$$

$$\sum_{n=1}^{N} \boldsymbol{t}_n \phi(\boldsymbol{x})^T - \mathbf{W}^T \phi(\boldsymbol{x}_n) \phi(\boldsymbol{x}_n)^T = 0$$

$$\sum_{n=1}^{N} \boldsymbol{t}_n \phi(\boldsymbol{x}_n)^T = \mathbf{W}^T \phi(\boldsymbol{x}_n) \phi(\boldsymbol{x}_n)^T$$

$$\sum_{n=1}^{N} \boldsymbol{t}_n \phi(\boldsymbol{x}_n)^T (\phi(\boldsymbol{x}_n) \phi(\boldsymbol{x}_n)^T)^{-1} = \mathbf{W}^T$$

$$\sum_{n=1}^{N} (\phi(\boldsymbol{x}_n) \boldsymbol{\phi}(\boldsymbol{x}_n)^T)^{-1} \phi(\boldsymbol{x}_n) \boldsymbol{t}_n^T = \mathbf{W}$$

$$(\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi} \boldsymbol{T} = \mathbf{W}$$

Thus,

$$\mathbf{W_{ML}} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi} \boldsymbol{T},$$

where $\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \\ \boldsymbol{\phi}(\boldsymbol{x}_2)^T \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_N)^T \end{bmatrix}$ and $\boldsymbol{T} = \begin{bmatrix} \boldsymbol{t}_1^T \\ \boldsymbol{t}_2^T \\ \vdots \\ \boldsymbol{t}_N^T \end{bmatrix}$.

Note: $\phi(\boldsymbol{x})\phi(\boldsymbol{x})^T$ is symmetric, so its inverse is also symmetric.

d)

First, start by substituing $\mathbf{\Sigma^{-1}}$ by $\mathbf{\Omega}$:

$$\begin{aligned}
\log p(\mathbf{T}|\mathbf{X}, \mathbf{\Sigma}) &= -\frac{NK}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T\mathbf{\Sigma}^{-1}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n)) \\
&= -\frac{NK}{2}\log(2\pi) - \frac{N}{2}\log\frac{1}{|\mathbf{\Omega}|} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T\mathbf{\Omega}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n)) \\
&= -\frac{NK}{2}\log(2\pi) + \frac{N}{2}\log|\mathbf{\Omega}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T\mathbf{\Omega}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))
\end{aligned}$$

Now, we can differentiate the log-likelihood with respect to $\mathbf{\Omega}$ and set it to zero:

$$\begin{aligned}
&\frac{\partial}{\partial\mathbf{\Omega}}\log p(\mathbf{T}|\mathbf{X}, \mathbf{\Omega}) = 0 \\
0 &= \frac{\partial}{\partial\mathbf{\Omega}}\left(-\frac{NK}{2}\log(2\pi) + \frac{N}{2}\log|\mathbf{\Omega}| - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T\mathbf{\Omega}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))\right) \\
0 &= \frac{N}{2}\mathbf{\Omega}^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T \\
\mathbf{\Omega}^{-1} &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T \\
\mathbf{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))(\boldsymbol{t}_n - \mathbf{W}^T\phi(\boldsymbol{x}_n))^T
\end{aligned}$$

Note: Some steps used the fact that since $\mathbf{\Sigma}$ is symmetric, thus its inverse is also symmetric, hence $(\mathbf{\Omega}^{-1})^T = \mathbf{\Omega}^{-1}$.

4 Counting Fish

a)

After observing these four samples, I would say that the total number of fishes N could be the maximum number I observed, which is 9.

b)

We know that the fishes are numbered from 1 to N. We also know that the probability of observing a fish is uniform, i.e., $p(x_i) = \frac{1}{N}$. If we compute the likelihood we then know that:

$$\begin{aligned} p(x_1, \cdots, x_k|N) &= \prod_{i=1}^{k} p(x_i|N) \\ &= \frac{1}{N^k} \quad \text{, with the contraint being that } N \geq \max(x_1, \cdots, x_k) \end{aligned}$$

We use this constraint because it is not possible to have observed a fish with a number larger than N.

Now, we would take the derivative of the log-likelihood and set the derivative to zero, but looking at the likelihood, we can see that the maximum is infinity when $N = 0$, since its a monotonically decreasing function. However, considering our constraint, we get that $N = \max(x_1, \cdots, x_k)$.

We can then conclude, based on the maximum likelihood estimate, that the $N_{ML} = \max(x_1, \cdots, x_k)$. In our case, $N_{ML} = 9$.

c)

To determine the bias we first need to look on what could be a good estimator for the total number of fishes, which could be the expectancy for the maximum value. I did this in the previous question so $N_{MLE} = \max(x_1, \cdots, x_k) \rightarrow \mathbb{E}[N_{MLE}] = \mathbb{E}[\max(x_1, \cdots, x_k)]$.

I'm going to demonstrate two ways for solving this, being the first the most correct one, however can be mathematically complex.

First way

Before we showed that $\hat{N}_{MLE} = \max(x_1, \cdots, x_k)$. Meaning that to find the expectancy of this estimator, we can start by findind the expectancy of the maximum value of the sample.

From the perspective of a frequentist approach, we can say that the probability of the observed maximum $M = \max(x_1, \cdots, x_k)$ being less than or equal to m is given by:

$$P(M \leq m) = \frac{m^k}{N^k}$$

We can think of this as the following: the observations are independent and identically distributed, and the probability of observing a fish with number m is $\frac{1}{N}$. Hence,

$$P(M \leq m) = P(x_1 \leq m, \cdots, x_k \leq m) = \prod_{i=1}^{k} P(x_i \leq m) = \left(\frac{m}{N}\right)^k$$

We can also think that we are filling k slots for N fishes and they can be repeated.

Given this, we have that:

$$\mathbb{E}[M] = \sum_{m=k}^{N} mP(M = m) = \sum_{m=k}^{N} m\left(P(M \leq m) - P(M \leq m-1)\right) = \sum_{m=k}^{N} m\left(\left(\frac{m}{N}\right)^k - \left(\frac{m-1}{N}\right)^k\right)$$

Note that we did not not differentiate the cumulative distribution function, because we are dealing with discrete values.

Now, here is when the maths gets a little complex, we have a sum of powers of integers, which can be solved by Faulhaber's formula, but it is quite complex.

However, if we assume that N is large, we can do the following:

$$
\begin{aligned}
\mathbb{E}[M] &= \sum_{m=k}^{N} m \left(\left(\frac{m}{N}\right)^k - \left(\frac{m-1}{N}\right)^k \right) \\
&= \sum_{m=k}^{N} m \left(\left(\frac{m}{N}\right)^k - \left(\frac{m}{N} - \frac{1}{N}\right)^k \right) \\
&= \sum_{m=k}^{N} m \left(\left(\frac{m}{N}\right)^k - \left(\frac{m}{N}\right)^k \left(1 - \frac{1}{m}\right)^k \right) \\
&= \sum_{m=k}^{N} m \left(\left(\frac{m}{N}\right)^k - \left(\frac{m}{N}\right)^k \left(1 - \frac{k}{m}\right) \right) \quad \text{, using a Taylor Series} \\
&= \sum_{m=k}^{N} m \left(\frac{m}{N}\right)^k \left(1 - \left(1 - \frac{k}{m}\right)\right) \\
&= \sum_{m=k}^{N} m \left(\frac{m}{N}\right)^k \frac{k}{m} \quad \text{, approximating using an Integral} \\
&= k \left(N \int_0^1 x^k dx \right) \quad \text{, because } \frac{m}{N} \leq 1 \text{ and the N accounts for the fact that is discrete} \\
&= N \frac{k}{k+1}
\end{aligned}
$$

Now, isolating N, we have that:

$$N = \frac{k+1}{k}\mathbb{E}[M]$$

Which gives a bias of:

$$\text{Bias} = \mathbb{E}[N] - N = \mathbb{E}[M] - \frac{k+1}{k}\mathbb{E}[M] = -\frac{\mathbb{E}[M]}{k} + 1$$

In our case, we have that:

$$\text{Bias} = 9 - 11,25 = -2,25$$

This Bias is negative, meaning the $\hat{N}_{ML}$ underestimates the total number of fishes, as we expected.

Second way

First let's keep in mind that we observed a sample that had a repeated value, and I will deal with this in the end.

First, let's consider we observe k samples, each one different. And in this sample, imagine that we observed the maximum. So, our sample with size k has the maximum m and $k-1$ numbers that are smaller than m. So by thinking on all the possibilities we know that all the combinations possible for "withdraing" such fishes are $\binom{m-1}{k-1}$ (m - 1 numbers for k - 1 positions). And we know that the total number of possibilities for generating a sample of size k is $\binom{N}{k}$. Henceforth,

$$P(M = m) = \frac{\binom{m-1}{k-1}}{\binom{N}{k}}$$

Now, we want to calculate the expectancy of the maximum value, which is given by:

$$E[M] = \sum_{m=k}^{N} mP(M = m)$$

Hence,

$$\begin{aligned}
E[M] &= \sum_{m=k}^{N} m\frac{\binom{m-1}{k-1}}{\binom{N}{k}} \\
&= \sum_{m=k}^{N} m\frac{\binom{m-1}{k-1}}{\binom{N}{k}}\frac{k}{k} \\
&= \sum_{m=k}^{N} \frac{k}{\binom{N}{k}}\frac{m}{k}\binom{m-1}{k-1} \\
&= \sum_{m=k}^{N} \frac{k}{\binom{N}{k}}\frac{m}{k}\frac{(m-1)!}{(k-1)!(m-k)!} \\
&= \sum_{m=k}^{N} \frac{k}{\binom{N}{k}}\frac{m!}{k!(m-k)!} \\
&= \frac{k}{\binom{N}{k}}\sum_{m=k}^{N}\binom{m}{k} \\
&= \frac{k}{\binom{N}{k}}\binom{N+1}{k+1} \\
&= \frac{k(N+1)!}{(k+1)!(N-k)!}\frac{k!(N-k)!}{N!} \\
&= \frac{k(N+1)}{k+1}.
\end{aligned}$$

If we know assume that $E[M] \approx M$,

$$\begin{aligned}
M &= \frac{k(N+1)}{k+1} \\
M(k+1) &= k(N+1) \\
M(k+1) - k &= kN \\
N &= \frac{M(k+1)}{k} - 1.
\end{aligned}$$

Calculating the bias, we have:

$$\begin{aligned}
\text{Bias} &= E[N] - N \\
&= M - (\frac{M(k+1)}{k} - 1) \\
&= M - M - \frac{M}{k} + 1 \\
&= -\frac{M}{k} + 1.
\end{aligned}$$

Since k is at most equal to M, the bias is negative as we expected. The maximum observed in the sample underestimates the total number. Also note that the 1 is there to ensure, that if we observe the entire poplation, $N = k = M$, the bias is zero.

Now, using our values. We have a sample with size 4, however one fish was repeated and the demonstration above was for a sample with all different values. So, we need to consider this by saying that the k above is the number of different values observed. Hence, in our case, $k = 3$ and $M = 9$. This gives us that:

$$N = \frac{9(3+1)}{3} - 1 = 11.$$

And the bias is:

$$\text{Bias} = -\frac{9}{3} + 1 = -2.$$

Which gives the same result as the previous one, considering we only care about the integer part of the number.

We can then conclude that our estimator $\hat{N}_{ML}$ underestimates the total number of fishes, as expected, since we might not have observed the maximum of the whole sample, and we might have observed a sample that is not representative of the whole population. In our case, the bias is -2, stating that the estimator underestimates the total number of fishes by 2.