1st Homework Assignment

Platon Karageorgis
Machine Learning I

November 21, 2024

1 Multivariate Calculus

a) We have the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ where $x \in R^m$ and we want to calculate its derivative. This is calculated in the following way:

$$\frac{\partial}{\partial x_i}\sigma(x) = \frac{\partial}{\partial x_i}\left(\frac{1}{1+e^{-x}}\right)$$

We set $u = 1 + e^{-x}$ and apply the chain rule:

$$\frac{\partial}{\partial u}\left(\frac{1}{u}\right) \cdot \frac{\partial}{\partial x}\left(1+e^{x}\right) = \frac{-1}{u^2} \cdot -e^{-x}$$

We substitute x back in the place of u and we get:

$$\frac{-1}{\left(1+e^{-x}\right)^2} \cdot -e^{-x} = \frac{e^{x}}{\left(1+e^{-x}\right)^2}$$

Therefore we conclude that the derivative of the sigmoid function is:

$$\frac{\partial}{\partial x_i}\sigma(x) = \frac{e^{x}}{\left(1+e^{-x}\right)^2}$$

b) We have $X \in R^{n,n}, w \in R^n$ and we need to calculate the derivative of $f = Xw$ for a w_i. This is computed in the following manner:

$$\left[\frac{df}{dw}\right]_{ij} = \frac{\partial f_i}{\partial w_j} = \frac{\partial}{\partial w_j}\left(Xw\right)_i = \frac{\partial}{\partial w_j}\sum_{p=1}^{n}\left(X_{ip}w_p\right)$$

We move the derivation inside the summation:

$$\sum_{p=1}^{n}\frac{\partial}{\partial w_j}\left(X_{ip}w_p\right)$$

$$\sum_{p=1}^{n}X_{ip}\delta_{jp} = X_{ij} = X$$

In the final steps we first replaced the Kronecker delta as it was inside the summation, and then replaced the index p with j as the delta was applied. Finally, since for a derivation of the jth element of f by an ith X we get the X_{ij} element of X, then we can generalize this and conclude that the total derivative of $\frac{df}{dx}$ will be X. Hence, we have found the answer to this question.

c) We derive f in terms of w in the following manner:

$$\frac{\partial f}{\partial w_i} = \frac{\partial \left(w^T X w\right)}{\partial w_i} = \frac{\partial}{\partial w_i}\left(\sum_{p=1}^{n}\sum_{q=1}^{n} w_p X_{pq} w_q\right) = \sum_{p=1}^{n}\sum_{q=1}^{n} \frac{\partial}{\partial w_i}\left(w_p X_{pq} w_q\right)$$

We then apply the product rule:

$$\sum_{p=1}^{n}\left(\frac{\partial}{\partial w_i}\left(w_p\right)\sum_{q=1}^{n} X_{pq} w_q\right) + \sum_{p=1}^{n}\left(w_p \sum_{q=1}^{n}\frac{\partial}{\partial w_i}\left(X_{pq} w_q\right)\right)$$

$$= \sum_{p=1}^{n}\left(\delta_{ip}\sum_{q=1}^{n} X_{pq} w_q\right) + \sum_{p=1}^{n}\left(wp\sum_{q=1}^{n} X_{pq}\delta_{iq}\right)$$

We remove the Kronecker deltas as they are inside the summations, and then we replace the corresponding indices p, q with i, j respectively.

$$\sum_{q=1}^{n} X_{iq} w_q + \sum_{p=1}^{n} w_p X_{pi}$$

Subsequently, we write this in matrix form to visualize the total derivative.

$$\begin{pmatrix} \frac{df}{dw_1} \\ \frac{df}{dw_2} \\ \vdots \\ \frac{df}{dw_n} \end{pmatrix} = \begin{pmatrix} \sum_{q=1}^{n} X_{1q} w_q + \sum_{p=1}^{n} w_p X_{p1} \\ \sum_{q=1}^{n} X_{2q} w_q + \sum_{p=1}^{n} w_p X_{p2} \\ \vdots \\ \sum_{q=1}^{n} X_{nq} w_q + \sum_{p=1}^{n} w_p X_{pn} \end{pmatrix}$$

So, the final result is:

$$\left[\frac{\partial f}{\partial w}\right]_i = Xw + w^T X$$

d) The goal of this exercise is to compute the Jacobian matrix of the softmax function ς(x)i. The derivative of the softmax function is defined as:

$$\frac{d\varsigma(x)_i}{dx} = \frac{\partial \varsigma(x)_i}{\partial x_k} = \frac{\partial}{\partial x_k}\left(\frac{\exp(x_i)}{\sum_{j=1}^{n}\exp(x_j)}\right)$$

After applying the quotient rule the relation transforms to:

$$\frac{\partial}{\partial x_k}\left(\exp(x_i)\right)\cdot\sum_{j=1}^{n}\exp(x_j) - \exp(x_i)\cdot\frac{\partial}{\partial x_k}\left(\sum_{j=1}^{n}\exp(x_j)\right)$$

$$= \frac{\delta_{ik} \exp(x_i) \cdot \sum_{j=1}^{n} \exp(x_j) - \exp(x_i) \cdot \sum_{j=1}^{n} \frac{\partial}{\partial x_k} (\exp(x_j))}{\left(\sum_{j=1}^{n} \exp(x_j)\right)^2}$$

$$= \frac{\delta_{ik} \exp(x_i) \sum_{j=1}^{n} \exp(x_j) - \exp(x_i) \exp(x_k)}{\left(\sum_{j=1}^{n} \exp(x_j)\right)^2} \tag{2}$$

In order to continue, we need to separate the solution into 2 cases, where Case 1 will assume that $i = k$ and Case 2 that $i \neq k$.

If $i = k$:
In this case, from equation (2):

$$(2) \rightarrow \frac{\exp(x_k) \sum_{j=1}^{n} \exp(x_j) - \exp(x_k)^2}{\left(\sum_{j=1}^{n} \exp(x_j)\right)^2} = \exp(x_k)(\frac{\left(\sum_{j=1}^{n} \exp(x_j) - \exp(x_k)\right)}{\sum_{j=1}^{n} \exp(x_j)) \cdot \sum_{j=1}^{n} \exp(x_j)}$$

$$= \varsigma(x)_k \cdot \frac{\sum_{j=1}^{n} \exp(x_j) - \exp(x_k)}{\sum_{j=1}^{n} \exp(x_j)}$$

$$= \varsigma(x)_k \left(1 - \varsigma(x)_k\right)$$

If $i \neq k$:
From equation (2), we have:

$$(2) \xrightarrow{\delta_{ik} = 0} \frac{-\exp(x_i) \exp(x_k)}{\left(\sum_{j=1}^{n} \exp(x_j)\right)^2}$$

$$= \frac{-\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)} \cdot \frac{-\exp(x_k)}{\sum_{j=1}^{n} \exp(x_j)}$$

$$= -\varsigma(x)_i \varsigma(x)_k$$

Therefore:

$$\frac{\partial \varsigma(x)_i}{\partial x_k} = \begin{cases} \varsigma(x)_k \left(1 - \varsigma(x)_k\right) & \text{if } i = k \\ -\varsigma(x)_i \varsigma(x)_k & \text{if } i \neq k \end{cases}$$

To write it in matrix form, we need to perceive the result that we just concluded in. When $i = k$, we will only be in the diagonal of the Jacobian matrix, whereas in the other case we need to be off the diagonal. Therefore, knowing that AA^T for matrix A is the outer product which gives us the off-diagonal elements, the final matrix form is:

$$\left(\frac{\partial \varsigma(x)}{\partial x}\right)_{ik} = J(\varsigma(x)) = \text{diag}(\varsigma(x)) - \varsigma(x)\varsigma(x)^T$$

To delve into further, the first matrix will contain the diagonal values of the Jacobian matrix and the rest of the elements will be 0. Afterwards, the subtraction of the outer product keeps intact the diagonal values but causes each non-diagonal element to have the corresponding values of the 2nd case where i is not equal to k. The matrix will look like this:

$$\begin{pmatrix} \frac{\partial \varsigma(x)_1}{\partial x_1} & \frac{\partial \varsigma(x)_1}{\partial x_2} & \cdots & \frac{\partial \varsigma(x)_1}{\partial x_n} \\ \frac{\partial \varsigma(x)_2}{\partial x_1} & \frac{\partial \varsigma(x)_2}{\partial x_2} & \cdots & \frac{\partial \varsigma(x)_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \varsigma(x)_n}{\partial x_1} & \frac{\partial \varsigma(x)_n}{\partial x_2} & \cdots & \frac{\partial \varsigma(x)_n}{\partial x_n} \end{pmatrix}$$

e) We have the following:

$$\|X\theta - y\|_2^2, \quad X \in \mathbb{R}^{n \times d}, \quad \theta \in \mathbb{R}^d, \quad y \in \mathbb{R}^n$$

We set:

$$f(\theta) = \|X\theta - y\|_2^2$$

And we proceed to calculate the derivative of f.

$$\left[\frac{df}{d\theta}\right]_i = \frac{\partial}{\partial \theta_i} \cdot (X\theta - y)^T \cdot ((X\theta - y))$$

The last step stems from the definition of the L_2 norm, as the squared Euclidean norm of a vector v is equal to the dot product of v with itself. In our case $v = X\theta - y$.
We continue to expand this relation:

$$\frac{\partial}{\partial \theta_i} \cdot \left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y\right) =$$

We remove the last term as it is unrelated to θ:

$$\frac{\partial}{\partial \theta_i} \cdot \left(\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta\right) =$$

Moreover, we observe that the third term is equivalent to the second term using the transpose of matrix product rule $(AB)^T = B^T A^T$:

$$\frac{\partial}{\partial \theta_i} \cdot \left(\theta^T X^T X\theta - 2\theta^T X^T y\right) =$$

In order to simplify the calculations we will calculate the derivative of each of the two terms separately. We begin with the first one:

$$\frac{\partial}{\partial \theta_i}\left(\theta^T X^T X\theta\right) = \frac{\partial}{\partial \theta_i}\left(\sum_{p=1}^{d}\sum_{q=1}^{d} \theta_p X_{p,q}^T X_{p,q} \theta_q\right)$$

We move outside of the inside summations the matrices that are not dependent on θ and remove the transpose from X.

$$\frac{\partial}{\partial\theta_i}\left(X_{q,p}X_{p,q}\sum_{p=1}^{d}\sum_{q=1}^{d}(\theta_p\theta_q)\right)$$

We apply the product rule:

$$\frac{\partial}{\partial\theta_i}\left(X_{q,p}X_{p,q}\sum_{p=1}^{d}\sum_{q=1}^{d}(\delta_{i,p}\theta_q+\delta_{i,q}\theta_p)\right)$$

$$=\frac{\partial}{\partial\theta_i}\left(X_{q,p}X_{p,q}\cdot\left(\sum_{p=1}^{d}\theta_q+\sum_{q=1}^{d}\theta_p\right)\right)$$

$$=\frac{\partial}{\partial\theta_i}\left(\sum_{p=1}^{d}X_{q,i}X_{i,q}\theta_q+\sum_{q=1}^{d}X_{i,p}X_{p,i}\theta_p\right)$$

$$\overset{(1)}{=}\left[X^TX\theta\right]_i+\left[\theta^TX^TX\right]_i=2\left[\theta^TX^TX\right]_i=2\theta^TX^TX$$

Note that in (1) we were able to complete this step using the rule $[v^TA]_i=\sum_p A_{p,i}v_p$ as well as the rule $[AV]_i=\sum_p A_{i,p}v_p$. Also, in the subsequent step we just used this identity $(AB)=B^TA^T$. Finally, in the last step we essentially say that if the derivative of a random i is $2\left[\theta^TX^TX\right]_i$ then the generalized version that holds for all the i values in θ is written as θ^TX^TX.

Having computed the first derivative, we continue with the second one:

$$\frac{\partial}{\partial\theta_i}\left(\theta^TX^Ty\right)=\frac{\partial}{\partial\theta_i}\left(\sum_{p=1}^{d}\sum_{q=1}^{d}\theta_pX_{p,q}^Ty_q\right)$$

$$=\frac{\partial}{\partial\theta_i}\left(\sum_{p=1}^{d}\sum_{q=1}^{d}\theta_pX_{q,p}y_q\right)$$

$$=\sum_{p=1}^{d}\sum_{q=1}^{d}\left(\frac{\partial}{\partial\theta_i}\theta_pX_{q,p}y_q\right)=\sum_{p=1}^{d}\sum_{q=1}^{d}(\delta_{i,p}X_{q,p}y_q)$$

$$=X_{q,i}y_q\overset{(2)}{=}\left[y^TX\right]_i=y^TX$$

In (2) we are allowed to do this step due to the rule $[v^TA]_i=\sum_p A_{p,i}v_p$ that can be applied in the reverse manner as well. As far as the last step is concerned, if the derivative of a random i is $\left[y^TX\right]_i$, then we can conclude that the total derivative for all the elements will be y^TX.
Now that we found both derivatives we can plug them into the original relation, set it as 0, and solve it:

$$\frac{\partial}{\partial\theta}\|X\theta-y\|_2^2=2\theta^TX^TX-2y^TX=0$$

$$\theta^T X^T X = y^T X \Leftrightarrow$$

Transpose the relation like we did multiple times, in order to remove the transpose from theta.

$$X^T X\theta = y^T X \Leftrightarrow$$

We do a left-side multiplication by the inverse of $X^T X$. Note
that by theory we know that $X^T X$ is always a square matrix and hence we know that it has an inverse.

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

Therefore, we have calculated theta!
BONUS This formula assumed in the final step that the matrix $X^\top X$ is always invertible, but in the case that the matrix is singular, it is not. For $X^\top X$ to be singular, all of its columns must be linearly dependent. This would mean that from the d columns, only 1 actually gives us information, and the rest have no contribution. In case
it is singular, by matrix theory, we know that it has at least 1 eigenvalue equal to 0. Thus, $X^\top X$ can't be invertible as invertible matrices must have only positive eigenvalues.

We add $\lambda||\theta||_2^2$ to our function and calculate the derivative again. Due to the sum rule we get:

$$f(\theta) = (X\theta - y)^\top (X\theta - y) + \lambda||\theta||_2^2$$

$$\frac{\partial f}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left((X\theta - y)^\top (X\theta - y)\right) + \frac{\partial}{\partial \theta_i}\left(\lambda||\theta||_2^2\right)$$

Since we already calculated the first part of this relation, we compute the second:

$$\frac{\partial}{\partial \theta_i}\left(\lambda||\theta||_2^2\right) = \lambda\frac{\partial}{\partial \theta_i}\left(\sum_{j=1}^{d}\theta_j^2\right) = \lambda\sum_{j=1}^{d}\frac{\partial}{\partial \theta_i}\theta_j^2$$

$$= 2\lambda\sum_{j=1}^{d}\delta_{i,j}\theta_j = 2\lambda\theta_i = 2\lambda\theta$$

So we replace this result into the original relation along with the derivative of $(X\theta - y)^\top (X\theta - y)$ that we derived in the previous question:

$$(X^\top X + \lambda I)\theta = X^\top y$$

By setting this to zero e get:

$$2\theta X^T X - 2y^T X + 2\lambda\theta = 0 \Leftrightarrow$$

$$\left(X^T X + \lambda I\right)\theta = y^T X$$

$$\theta = \left(X^T X + \lambda I\right)^{-1} y^T X$$

And now we can tell this formula will never fail. That is because $(X^\top X + \lambda I)$ is always invertible.

Proof
To begin with, $(X^\top X + \lambda I)$ is symmetric because $X^\top X$ is symmetric and λI is symmetric ($(X^\top X)^\top = X^\top X$ and the proof for $A^\top = A$ is trivial.)
Also, for any non-zero vector $v \in \mathbb{R}^d$:

$$v^\top (X^\top X + \lambda I)v = v^\top X^\top X v + \lambda v^\top I v = ||Xv||_2^2 + \lambda ||v||_2^2$$

Both $||Xv||_2^2$ and $\lambda||v||2^2$ are non-negative as norms, but $\lambda||v||_2^2$ is also strictly positive since $\lambda > 0$ and $||v||_2^2 > 0$ as v is non-zero by definition and $\lambda > 0$ given by the exercise. Hence, $v^\top (X^\top X + \lambda I)v > 0$ for all non-zero v, and this proves that $X^\top X + \lambda I$ is positive definite. But, a positive definite matrix has only positive eigenvalues, it is full rank and most importantly, is invertible. Therefore, we proved that adding the term λI where $\lambda > 0$ solves the issue the original formula had.

2 Full analysis of a distribution: Exponential distribution

a) Since X $\sim$ Exp(λ) we know that X follows the distribution:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Therefore, according to the theory (and since the distribution has 0 probability for the interval $(-\infty, 0)$), we know that:

$$\mathbb{E}_{X \sim \text{Exp}(\lambda)}[X] = \int_0^\infty x \lambda e^{-\lambda x}\, dx \tag{1}$$

The product rule for differentiation states:

$$\frac{d}{dx}\left[u(x) \cdot v(x)\right] = u'(x) \cdot v(x) + u(x) \cdot v'(x)$$

$$u(x) \cdot v'(x) = \frac{d}{dx}\left[u(x) \cdot v(x)\right] - u'(x) \cdot v(x)$$

$$\int u(x) \cdot v'(x)\, dx = u(x) \cdot v(x) - \int u'(x) * v(x)\, dx$$

So, we will apply this product rule to relation (1) in order to be able to transform the integral into a form where we can solve it. We set:

$$\begin{cases} u = x \\ du = dx \\ \int dv = \int \lambda e^{-\lambda x} \\ dv = \lambda e^{-\lambda x} \\ v = -e^{-\lambda x} \end{cases}$$

And we substitute in (1):

$$\mathbb{E}_{X\sim \text{Exp}(\lambda)}[X] = -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty -e^{-\lambda x}\, dx$$

The first term is 0 because as $x \to \infty$, $e^{-\lambda x}$ decays to 0 faster than x grows to infinity. Therefore, $-xe^{-\lambda x}$ approaches 0. At $x = 0$, the term $-xe^{-\lambda x}$ is also 0. Hence,

$$-xe^{-\lambda x}\Big|_0^\infty = -xe^{-\lambda x}\big|_0^\infty = 0 - 0 = 0$$

Then we are left with the second term and by calculating the integral we get:

$$\int_0^\infty e^{-\lambda x}\, dx = \left[\frac{-e^{-\lambda x}}{\lambda}\right]_0^\infty = -(0 - (-\frac{1}{\lambda})) = \frac{1}{\lambda}$$

Therefore, the expected value is:

$$\mathbb{E}_{X\sim \text{Exp}(\lambda)}[X] = \frac{1}{\lambda}$$

b) To find the probability of two consecutive students arriving in less than 1 minutes, we will use the expected value that we derived in the previous question. The exponential distribution models the time between successive events in a process where events happen continuously and independently at a constant rate. Thus, by using the expected value of X that follows the exponential distribution we should get the average time between the two events, meaning the arrivals of the two consecutive students.
Since the average time is two minutes in this case we get:

$$\mathbb{E}_{X\sim \text{Exp}(\lambda)}[X] = 2 \Leftrightarrow \frac{1}{\lambda} = 2 \Leftrightarrow \lambda = \frac{1}{2}$$

Then, we need to calculate the probability of X being less than 1, as the students arrive in less than 1 minute. To do this, we use the cumulative distribution. The CDF provides the probability that X is less than or equal to a given value x. In this case, we have $x = 1$.

$$P(X < 1) = 1 - e^{-\lambda \cdot 1}$$

We substitute $\lambda = \frac{1}{2}$:

$$P(X < 1) = 1 - e^{-\frac{1}{2}}$$

Approximating $e^{-\frac{1}{2}}$ gives:

$$P(X < 1) \approx 1 - 0.606 = 0.394$$

Therefore the probability is **39.4%**

c) X_i follows an exponential distribution with a parameter λ so the PDF should be:

$$p(x \mid \text{Exp}(\lambda)) = \lambda e^{-\lambda x}$$

Then, according to our theory the formula for calculating the maximum likelihood for N i.i.d. samples is:

$$\mathbf{w}_{\text{ML}} = \arg\max_{\mathbf{w}} p(D \mid \mathbf{w}) = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(x_i \mid \mathbf{w})$$

Plugging in our X dataset along with the exponential distribution we get:

$$p(X \mid \lambda) = \prod_{i=1}^{N} p(x_i \mid \lambda) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i} = \lambda^N e^{-\lambda \sum_{i=1}^{N} x_i}$$

We take the logarithm of this as we are trying to calculate the log-likelihood. By theory, maximizing the log-likelihood is equivalent to maximizing the likelihood so the previous relation transforms to:

$$\log p(X \mid \lambda) = \log \left(\lambda^N e^{-\lambda \sum_{i=1}^{N} x_i} \right)$$

$$\log p(D \mid \lambda) = \log \left(\lambda^N \right) + \log \left(e^{-\lambda \sum_{i=1}^{N} x_i} \right)$$

$$\log p(D \mid \lambda) = N \log \lambda - \lambda \sum_{i=1}^{N} x_i$$

The relation above is the log-likelihood for the parameter λ given the dataset X.

d) To find the parameter that maximizes the log-likelihood, we need to find the derivative of the log-likelihood function in terms of λ and subsequently set the derivative to be 0. By solving for lambda the equation that will be created, we compute the exact x_axis value for which the function has a maximum.

Given the log-likelihood function we calculated in the previous question, we now calculate its derivative:

$$\frac{d}{d\lambda} \left(\log p(D \mid \lambda) \right) = \frac{d}{d\lambda} \left(N \log \lambda - \lambda \sum_{i=1}^{N} x_i \right)$$

$$\frac{d}{d\lambda} \left(N \log \lambda \right) = \frac{N}{\lambda}$$

$$\frac{d}{d\lambda}\left(-\lambda\sum_{i=1}^{N}x_i\right) = -\sum_{i=1}^{N}x_i$$

$$\frac{d}{d\lambda}\left(\log p(D \mid \lambda)\right) = \frac{N}{\lambda} - \sum_{i=1}^{N}x_i$$

Subsequently, we set it as 0 and solve for λ

$$\frac{N}{\lambda} - \sum_{i=1}^{N}x_i = 0$$

$$\frac{N}{\lambda} = \sum_{i=1}^{N}x_i$$

$$\lambda = \frac{N}{\sum_{i=1}^{N}x_i}$$

Therefore, we have computed the maximum likelihood estimator λ_{ML}.

e) Following the directions of the exercise we will attempt to first prove the relation of the *Hint*.

From Bayes rule we know that:

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

We also know from the exercise that:

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} p(\lambda \mid x)$$

So the above relation transforms to:

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} \frac{p(x \mid \lambda)\, p(\lambda)}{p(x)}$$

According to probability theory (also stated at Bishop Chapter 1.2), since we know that the evidence $p(x)$ is a constant we can derive that:

$$\lambda_{\text{MAP}} \propto p(x \mid \lambda) \cdot p(\lambda)$$

In addition, we have learned from the lectures that adding the log when optimizing the distribution does not affect the maximum, therefore we can say:

$$\lambda_{\text{MAP}} \propto \log\left(p(x \mid \lambda) \cdot p(\lambda)\right)$$

Having proved the *Hint*, we now add to this relation the likelihood function we derived in question *c*, as well as the prior distribution of λ which is the Gamma function:

$$\lambda_{\text{MAP}} = N \log \lambda - \lambda \sum_{i=1}^{N} x_i + \log \left(\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{\alpha_1 - 1} \cdot e^{-\alpha_2 \lambda} \right)$$

According to Bishop (Chapter 1.2) it is possible to scale the log-likelihood by a positive constant as it does not interfere with the location of the maximum, therefore we can set the term $\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}$ equal to 1. By rearranging the rest of the terms we get:

$$\lambda_{\text{MAP}} = (\alpha_1 - 1) \log \lambda - \alpha_2 \lambda + N \log \lambda - \lambda \sum_{i=1}^{N} x_i \tag{1}$$

By optimizing this relation we will get the value of λ_{MAP} that gives us the maximum likelihood.

f) To find λ_{MAP} all we have to do is to derive (1) and set the result to 0 like we did in the previous question.

We start by deriving (1):

$$\frac{d}{d\lambda} \left[(\alpha_1 - 1) \log \lambda - \alpha_2 \lambda + N \log \lambda - \lambda \sum_{i=1}^{N} x_i \right]$$

The derivative is:

$$\frac{\alpha_1 - 1 + N}{\lambda} - \alpha_2 - \sum_{i=1}^{N} x_i$$

Set the derivative equal to zero:

$$\frac{\alpha_1 - 1 + N}{\lambda_{\text{MAP}}} - \alpha_2 - \sum_{i=1}^{N} x_i = 0$$

Solve for λ_{MAP}:

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

$$\lambda_{\text{MAP}} = \frac{\alpha_1 - 1 + N}{\alpha_2 + \sum_{i=1}^{N} x_i} \tag{2}$$

So, the MAP Estimator λ_{MAP} is described in relation (2).

g) To find an analytical result for the posterior distribution we need to compute $p(\lambda|x)$ but we can't use any of the results we have calculated so far. The reason is that until now we used estimation techniques but in this case we want the exact result. Hence, for the prior distribution we need to use the Gamma distribution formula for $p(\lambda)$ where $\alpha 1$ and $\alpha 2$ are its corresponding parameters, and for the likelihood function $p(x|\lambda)$ we use the formula of the exponential distribution.

So we have:

$$p(\lambda \mid x) = \frac{p(x \mid \lambda) \cdot p(\lambda)}{\int p(x \mid \lambda) \cdot p(\lambda)\, d\lambda}$$

And after plugging the correct values:

$$p(\lambda \mid x) = \frac{\lambda^N \cdot e^{-\lambda \sum_{i=1}^N X_i} \cdot \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{\alpha_1 - 1} \cdot e^{-\alpha_2 \lambda}}{\int \lambda^N \cdot e^{-\lambda \sum_{i=1}^N X_i} \cdot \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{\alpha_1 - 1} \cdot e^{-\alpha_2 \lambda}\, d\lambda} \tag{3}$$

We hold this relation and continue to calculate the integral separately.

We have:

$$\int \lambda^N \cdot e^{-\lambda \sum_{i=1}^N X_i} \cdot \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{\alpha_1 - 1} \cdot e^{-\alpha_2 \lambda}\, d\lambda$$

We simplify the relation to:

$$\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \int \lambda^{N + \alpha_1 - 1} \cdot e^{-\lambda \cdot (\alpha_2 + \sum_{i=1}^N X_i)}\, d\lambda$$

Subsequently, we set: $\alpha_1' = N + \alpha_1$ and $\alpha_2' = \sum_{i=1}^N X_i + \alpha_2$

And the relation transforms to:

$$\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \int \lambda^{\alpha_1' - 1} \cdot e^{-\lambda \cdot (\alpha_2')}\, d\lambda$$

Moreover, we know from theory that every distribution integrates to 1. If we had the term $\frac{\alpha_2'^{\alpha_1'}}{\Gamma(\alpha_1')}$ then we would have exactly the Gamma distribution for α_1' and α_2' as our parameters and this would integrate to 1. Therefore, since we are only missing this constant, the integral integrates to it reciprocal.

Hence we now have:

$$\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \int \lambda^{\alpha_1' - 1} \cdot e^{-\lambda \cdot (\alpha_2')}\, d\lambda = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_1')}{\alpha_2'^{\alpha_1'}}$$

Now that we have calculated the integral we plug it in back to relation (3):

$$p(\lambda \mid x) = \frac{\lambda^N \cdot e^{-\lambda \sum_{i=1}^N X_i} \cdot \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{\alpha_1 - 1} \cdot e^{-\alpha_2 \lambda}}{\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_1')}{\alpha_2'^{\alpha_1'}}} \tag{3}$$

Moreover, we do the same simplifications we did before, for the numerator and we substitute for the new alpha parameters.

$$p(\lambda \mid x) = \frac{\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{N + \alpha_1 - 1} \cdot e^{-\lambda \cdot (\alpha_2 + \sum_{i=1}^N X_i)}}{\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_1')}{\alpha_2'^{\alpha_1'}}} \tag{3}$$

$$p(\lambda \mid x) = \frac{\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \lambda^{\alpha_1'-1} \cdot e^{-\lambda \cdot (\alpha_2')}}{\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_1')}{\alpha_2'^{\alpha_1'}}} \tag{3}$$

Finally, we cancel the two identical constants $\frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)}$ and we move the final constant on the denominator to the numerator.

$$p(\lambda \mid x) = \frac{\alpha_2'^{\alpha_1'}}{\Gamma(\alpha_1')} \cdot \lambda^N \cdot e^{-\lambda \sum_{i=1}^N X_i} \cdot \lambda^{\alpha_1'-1} \cdot e^{-\alpha_2' \lambda}$$

We can easily see that this is exactly Gamma$(\alpha_1', \alpha 2')$ and therefore we have proved that the posterior distribution is indeed a Gamma distribution.

3 General Multiple Outputs Linear Regression

a) To find the dimensions of the parameter matrix W all we need to do is find the dimensions of $\phi(x)$ and $y(x, W)$, as these are connected with the model given by the exercise:

$$y(x, W) = W^T \phi(x)$$

According to the exercise, $\phi(x)$ is an $Mx1$ vector, and $y(x, W)$ outputs a k-vector, as the observation matrix T has an NxK size, where n is the number of the total observations and k must be the size of the output vector of the function y. Therefore, matrix W has to be a KxM matrix in order to multiply itself with $\phi(x)$ and output a k-vector.

b) The log-likelihood function is described by the following relation:

$$\log p\left(t|W, \Sigma\right) = arg_W max \log \prod_{i=1}^{N} \left(2\pi^{\frac{-k}{2}} |\Sigma|^{\frac{-1}{2}} \exp^{\frac{-1}{2}(t_i - W^T\phi(x_i))^T \Sigma^{-1} (t_i - W^T\phi(x_i))} \right)$$

We simplify the relation by moving the log inside the product:

$$\sum_{i=1}^{N} \left(\log(2\pi^{\frac{-k}{2}}) + log(|\Sigma|^{\frac{-1}{2}}) + log\left(\exp^{\frac{-1}{2}(t_i - W^T\phi(x_i))^T \Sigma^{-1} (t_i - W^T\phi(x_i))} \right) \right)$$

We discard the first two terms as they do not depend on w.

$$\sum_{i=1}^{N} \left(log\left(\exp^{\frac{-1}{2}(t_i - W^T\phi(x_i))^T \Sigma^{-1} (t_i - W^T\phi(x_i))} \right) \right)$$

And the final formula for the log-likelihood is:

$$-\frac{1}{2} \sum_{i=1}^{N} (t_i - W^T\phi(x_i))^T \Sigma^{-1} (t_i - W^T\phi(x_i))$$

c) We start by using the formula that we derived on the previous question and we will calculate its derivative:

$$\frac{\partial}{\partial W_j}\left(-\frac{1}{2}\sum_{i=1}^{N}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i))\right)_j$$

Using the formula given by the exercise we get:

$$\sum_{i=1}^{N}\left(-\frac{1}{2}\right)\Sigma^{-1}(t_i - W^\top\phi(x_i))\phi(x_i)^\top$$

We discard $-\frac{1}{2}$.

$$\sum_{i=1}^{N}\left(\Sigma^{-1}(t_i - W^\top\phi(x_i))\phi(x_i)^\top\right)$$

Set it to 0 and solve in terms of Φ, T:

$$\sum_{i=1}^{N}\left(\Sigma^{-1}(t_i - W^\top\phi(x_i))\phi(x_i)^\top\right) = 0$$

We remove Σ^{-1} by multiplying by Σ (obviously Σ^{-1} is invertible).

$$\sum_{i=1}^{N}\left((t_i - W^\top\phi(x_i))\phi(x_i)^\top\right) = 0$$

$$\sum_{i=1}^{N}(t_i\phi(x_i)^\top) - \sum_{i=1}^{N}W^\top\phi(x_i)\phi(x_i)^\top = 0$$

$$\sum_{i=1}^{N}(t_i\phi(x_i)^\top) = \sum_{i=1}^{N}W^\top\phi(x_i)\phi(x_i)^\top$$

Finally, we convert these to matrices. By theory we know that the feature vector Φ is a column.

$$T^\top\Phi = W^\top\Phi^\top\Phi \quad \Longleftrightarrow \quad \Phi^\top\Phi W = T\Phi$$

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

Regarding the dependence of w on the covariance matrix Σ, it is independent because ultimately in the final relation Σ does not appear anywhere. Intuitively, we understand that this is normal as w is trying to fit the data whilst Σ describes the overall noise in the data.

d) We will calculate the log-likelihood again but this time it will be in terms of Σ. The initial formula we derived for the log-likelihood is:

$$\log p\left(t|W,\Sigma\right) = arg_{\Sigma}max \log \prod_{i=1}^{N} \left(2\pi^{\frac{-k}{2}} |\Sigma|^{\frac{-1}{2}} \exp^{\frac{-1}{2}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i))}\right)$$

$$= \sum_{i=1}^{N} \left(\log(2\pi^{\frac{-k}{2}}) + log(|\Sigma|^{\frac{-1}{2}}) + log\left(\exp^{\frac{-1}{2}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i))}\right)\right)$$

We discard the constant value $2\pi^{\frac{-k}{2}}$ and move $\frac{-N}{2}log(|\Sigma|)$ outside of the summation.

$$-\frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}\left(t_i - W^{\top}\phi(x_i)\right)^{\top}\Sigma^{-1}\left(t_i - W^{\top}\phi(x_i)\right)$$

Let:

$$r_i = t_i - W^{\top}\phi(x_i)$$

Also, we make the substitution suggested in the exercise $\Omega = \Sigma^{-1}$. Note that $\log|\Omega^{-1}|$ can be transformed to $-\log|\Omega|$:

$$\log p\left(t|W,\Sigma\right) = \frac{N}{2}\log|\Omega| - \frac{1}{2}\sum_{i=1}^{N} r_i^{\top}\Omega r_i$$

Having arrived in the final relation of the log-likelihood, we will now compute the derivative of each term separately: For the first term we have:

$$\frac{\partial}{\partial\Omega}\left(\frac{N}{2}\log(|\Omega|)\right)$$

Using the formula given in the exercise the derivative of this relation is:
(Note that Ω^{-1} is symmetric as Ω is symmetric so $(\Omega^{-1})^T = \Omega^{-1}$)

$$\frac{N}{2}\Omega^{-1}$$

Finally, for the second term:

$$\frac{\partial}{\partial\Omega}\left(\frac{-1}{2}\sum_{i=1}^{N} r_i^T\Omega r_i\right)$$

Again using the formula given we get:

$$-\frac{1}{2}\sum_{i=1}^{N} r_i r_i^{\top}$$

The final derivative after combining the two terms is:

$$\frac{\partial}{\partial \Omega} \log L(\Omega) = \frac{N}{2}\Omega^{-1} - \frac{1}{2}\sum_{i=1}^{N} r_i r_i^\top$$

Now, we set the derivative to zero and solve for Ω:

$$\frac{N}{2}\Omega^{-1} - \frac{1}{2}\sum_{i=1}^{N} r_i r_i^\top = 0 \Leftrightarrow$$

$$N\Omega^{-1} - \sum_{i=1}^{N} r_i r_i^\top = 0 \Leftrightarrow$$

$$\Omega^{-1} = \frac{1}{N}\sum_{i=1}^{N} r_i r_i^\top$$

Ultimately, we substitute back Σ and $x_i - W^\top \phi(x_i)$:

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} \left(x_i - W^\top \phi(x_i)\right)\left(x_i - W^\top \phi(x_i)\right)^\top$$

Which proves that the maximum likelihood solution for Σ is the one that the exercise illustrated.

4 Counting Fish

a) Intuitively, there should be 9 fishes in the aquarium. That's it because if it had $9 + n$ where n is any positive integer, then the probability of observing the same fish more than once, would be decreasing as n grew bigger (and we observed fish #3 twice).

b) We need to compute for which N the ML_e is maximized, but first we will define the distribution that describes the data. Specifically, we have the following uniform distribution:

$$P(X = x|N) = \begin{cases} \frac{1}{N}, N \geq 9 \\ 0, N < 9 \end{cases}$$

The distribution illustrates that for any N above 9, where N depicts the greatest tag on the aquarium, the probability will be $\frac{1}{N}$. In addition, the probability of $N < 9$ is 0 as we have already seen a fish with this tag. We continue now to define the log-likelihood:

$$P(x|N) = arg_N max \prod_{i=1}^{M} \left(\frac{1}{N_m}\right) = \left(\frac{1}{N}\right)^i$$

Normally, we would have to find the derivative and set it to 0, but it is clear in this case that $\left(\frac{1}{N}\right)^i$ will have the greatest value when N is the smallest possible integer (up to 9),

which in our case is 9.

c) The formula of the bias is described by:

$$Bias_\theta = E_{x\sim p(x)}[x] - \theta$$

, where θ is the true value and in our case is $N = 9$, and $E_{x\sim p(x)}[x]$ is the estimated value. If these two values are equal then we have no bias. If the expected value is larger than the true value we have an overestimation of the value, and finally if the expected value is smaller than the true value then we have an underestimation. The last case is not possible in our case because N can't be less than 9.

We already obtained the true value from the previous question, we proceed to calculate the expected value:

$$E\left[X = k\right] = \sum_{i=1}^{M} i \cdot p(i = M)$$

Since we have a discrete uniform distribution, we will use the following formula:

$$P(i = M) = P(i \leq M) - P(i \leq M - 1)$$

We substitute this in our relation:

$$\begin{aligned} E\left[X = k\right] &= \sum_{i=1}^{M} i \left[\left(\frac{i}{N}\right)^k - \left(\frac{i-1}{N}\right)^k \right] \\ &= \sum_{i=1}^{M} i \left(\frac{i}{N}\right)^k - \sum_{i=1}^{M} i \left(\frac{i-1}{N}\right)^k \end{aligned}$$

We make an index change to the second sum in order to match the other term:

$$= \sum_{i=1}^{M} i \left(\frac{i}{N}\right)^k - \sum_{i=0}^{M-1} (i+1) \left(\frac{i}{N}\right)^k$$

For the calculations to continue we need to visualize what the above relation is saying:

The first sum:

$$\sum_{i=1}^{M} i \left(\frac{i}{N}\right)^k = 1 \left(\frac{1}{N}\right)^k + 2 \left(\frac{2}{N}\right)^k + \cdots + M \left(\frac{M}{N}\right)^k.$$

The second sum:

$$\sum_{i=0}^{M-1} (i+1) \left(\frac{i}{N}\right)^k = 1 \left(\frac{0}{N}\right)^k + 2 \left(\frac{1}{N}\right)^k + \cdots + M \left(\frac{M-1}{N}\right)^k.$$

Therefore, after subtracting these two terms we will be left by the first sub-term of the

second sum $1\left(\frac{0}{N}\right)^k$, the largest sub-term of the first term $M\left(\frac{M}{N}\right)^k$, and most importantly, $-\sum_{i=1}^{M-1}\left(\frac{i}{N}\right)^k$. This is because every time in the second term, we cancel the previous sub-term of the first term but we subtract 1 additional time. By adding all these additional subtractions for each term, we get this result. Finally this is described:

$$E[X=k] = 1\left(\frac{0}{N}\right)^k + M\left(\frac{M}{N}\right)^k - \sum_{i=1}^{M-1}\left(\frac{i}{N}\right)^k = M\left(\frac{M}{N}\right)^k - \sum_{i=1}^{M-1}\left(\frac{i}{N}\right)^k$$