Homework 1

Luis Vitor Zerkowski - 14895730

September 19, 2023

1 Multivariate Calculus

1.a

Since $x \in \mathbb{R}^m$, and we are applying the sigmoid function to each one of the vector's entries, we have:

$$\sigma(x) = \begin{bmatrix} \sigma(x_1) \\ \vdots \\ \sigma(x_m) \end{bmatrix}$$

Thus, getting the gradient of the sigmoid function applied to the vector, $\nabla_x \sigma(x)$, is just a matter of applying the derivative of the sigmoid function to each one of the entries.

We then start by computing the derivative of the sigmoid function:

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

By the quotient rule, the sum rule on $(1+e^{-x})$ and also the chain rule on $(1+e^{-x})$:

$$= \frac{0-(-1)e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} =$$

$$= \frac{1}{1+e^{-x}} \frac{e^{-x}}{1+e^{-x}} = \sigma(x) \frac{e^{-x}+1-1}{1+e^{-x}} =$$

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

Having the derivative of the sigmoid function in hands, we know that:

$$\nabla_x \sigma(x) = \begin{bmatrix} \frac{\partial \sigma(x_1)}{x_1} \\ \vdots \\ \frac{\partial \sigma(x_m)}{x_m} \end{bmatrix} = \begin{bmatrix} \sigma(x_1)(1-\sigma(x_1)) \\ \vdots \\ \sigma(x_m)(1-\sigma(x_m)) \end{bmatrix}$$

1.b

Let's start by doing the multiplication with index notation to make it easier to compute the derivatives. Since $X \in \mathbb{R}^{n \times n}$ and $w \in \mathbb{R}^{n \times 1}$, multiplying both should give us $Xw \in \mathbb{R}^{n \times 1}$. So we do:

$$[Xw]_i = \sum_{p=1}^{n} X_{ip} w_p$$

Applying the derivatives w.r.t w on the i-th row, we get:

$$\frac{\partial}{\partial w_j}[Xw]_i = \frac{\partial}{\partial w_j} \sum_{p=1}^{n} X_{ip} w_p =$$

By the sum rule and also using the fact that X_{ip} is just a constant:

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

We now end up with a Kronecker Delta:

$$\frac{\partial w_p}{\partial w_j} = \delta_{jp} = \begin{cases} 1, & \text{if } j = p \\ 0, & \text{otherwise} \end{cases}$$

And then we can do:

$$\sum_{p=1}^{n} X_{ip} \frac{\partial w_p}{\partial w_j} = \sum_{p=1}^{n} X_{ip} \delta_{jp} = X_{ij}$$

Thus the derivative $\frac{\partial}{\partial w} f$ with $f = Xw$ give us exactly $\frac{\partial}{\partial w} f = X$.

1.c

We start again by doing the multiplication with index notation to make it easier to compute the derivatives. Since $X \in \mathbb{R}^{n \times n}$ and $w \in \mathbb{R}^{n \times 1}$, the multiplication $w^T X w$ should give us a real number. So we do:

$$w^T X w = \sum_{p=1}^{n} \sum_{q=1}^{n} w_p X_{pq} w_q$$

Applying the derivatives w.r.t w, we get:

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

By the sum rule and also using the fact that X_{pq} is just a constant:

$$= \sum_{p=1}^{n} \sum_{q=1}^{n} X_{pq} \frac{\partial}{\partial w_i} w_p w_q =$$

By the product rule:

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

Using the Kronecker Delta once again:

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

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

$$= w^T X^T + w^T X = w^T (X^T + X)$$

Thus the derivative $\frac{\partial}{\partial w} f$ with $f = w^T X w$ give us exactly $\frac{\partial}{\partial w} f = w^T (X^T + X)$.

1.d

We start again by doing the multiplication with index notation to make it easier to compute the derivatives. Let's also call $(x - As) = v$ to simplify the problem for now. Since $\Sigma^{-1} \in \mathbb{R}^{m \times m}$ and $v \in \mathbb{R}^{m \times 1}$, the multiplication $v^T \Sigma^{-1} v$ should give us a real number. So we do:

$$v^T \Sigma^{-1} v = \sum_{p=1}^{m} \sum_{q=1}^{m} v_p \Sigma_{pq}^{-1} v_q$$

Applying the derivatives w.r.t Σ^{-1}, we get:

$$\frac{\partial v^T \Sigma^{-1} v}{\partial \Sigma_{ij}^{-1}} = \frac{\partial}{\partial \Sigma_{ij}^{-1}} \sum_{p=1}^{m} \sum_{q=1}^{m} v_p \Sigma_{pq}^{-1} v_q =$$

By the sum rule and also using the fact that v_p and v_q are just constants:

$$= \sum_{p=1}^{m} \sum_{q=1}^{m} v_p v_q \frac{\partial}{\partial \Sigma_{ij}^{-1}} \Sigma_{pq}^{-1} =$$

And since $\frac{\partial}{\partial \Sigma_{ij}^{-1}} \Sigma_{pq}^{-1} = 1$ if $i = p, j = q$ and $\frac{\partial}{\partial \Sigma_{ij}^{-1}} \Sigma_{pq}^{-1} = 0$ otherwise, it follows:

$$= \sum_{p=1}^{m} \sum_{q=1}^{m} v_p v_q \frac{\partial}{\partial \Sigma_{ij}^{-1}} \Sigma_{pq}^{-1} = v_i v_j = (x - As)(x - As)^T$$

Thus the derivative $\frac{\partial}{\partial \Sigma^{-1}} (x - As)^T \Sigma^{-1} (x - As)$ give us exactly $(x - As)(x - As)^T$.

1.e

We start by writing $\varsigma(x)$ with $x \in \mathbb{R}^n$ in its vector form:

$$\varsigma(x) = \begin{bmatrix} \frac{e^{x_1}}{\sum_{j=1}^{n} e^{x_j}} \\ \vdots \\ \frac{e^{x_n}}{\sum_{j=1}^{n} e^{x_j}} \end{bmatrix}$$

Thus, getting the derivative of ς w.r.t. the vector x is just a matter of applying the partial derivatives to each one of the rows.

$$\frac{\partial}{\partial x} \varsigma(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} \frac{e^{x_1}}{\sum_{j=1}^{n} e^{x_j}} & \cdots & \frac{\partial}{\partial x_n} \frac{e^{x_1}}{\sum_{j=1}^{n} e^{x_j}} \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} \frac{e^{x_n}}{\sum_{j=1}^{n} e^{x_j}} & \cdots & \frac{\partial}{\partial x_n} \frac{e^{x_n}}{\sum_{j=1}^{n} e^{x_j}} \end{bmatrix}$$

We then start by computing the derivative of a cell to fill out the matrix afterwards. For x_i we have:

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

$$= \frac{\partial}{\partial x_i} \frac{e^{x_i}}{e^{x_1} + \ldots + e^{x_n}} =$$

Applying the quotient rule to the fraction and then the sum rule to $(e^{x_1} + \ldots + e^{x_n})$, we get:

$$= \frac{e^{x_i}(e^{x_1} + \ldots + e^{x_n}) - e^{x_i}(0 + \ldots + e^{x_i} + \ldots + 0)}{(e^{x_1} + \ldots + e^{x_n})^2} = \frac{e^{x_i}(e^{x_1} + \ldots + e^{x_n}) - e^{2x_i}}{(e^{x_1} + \ldots + e^{x_n})^2} =$$

$$= \frac{e^{x_i}}{(e^{x_1} + \ldots + e^{x_n})} - \frac{e^{x_i}}{(e^{x_1} + \ldots + e^{x_n})} \frac{e^{x_i}}{(e^{x_1} + \ldots + e^{x_n})} = \varsigma(x)_i - \varsigma(x)_i^2$$

Now for x_k with $k \neq i$, we have:

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

$$= \frac{\partial}{\partial x_k}\frac{e^{x_i}}{e^{x_1}+\ldots+e^{x_n}} =$$

Applying the quotient rule to the fraction and then the sum rule to $(e^{x_1}+\ldots+e^{x_n})$, we get:

$$= \frac{0(e^{x_1}+\ldots+e^{x_n}) - e^{x_i}(0+\ldots+e^{x_k}+\ldots+0)}{(e^{x_1}+\ldots+e^{x_n})^2} = \frac{-e^{x_i+x_k}}{(e^{x_1}+\ldots+e^{x_n})^2} =$$

$$= -\frac{e^{x_i}}{(e^{x_1}+\ldots+e^{x_n})}\frac{e^{x_k}}{(e^{x_1}+\ldots+e^{x_n})} = -\varsigma(x)_i\varsigma(x)_k$$

With all the matrix entries in hand, we can write the derivative down as $\frac{\partial}{\partial x}\varsigma(x) = \operatorname{diag}(\varsigma(x)) - \varsigma(x)\varsigma(x)^T$.

2 Full Analysis of a Distribution: Poisson Distribution

2.a

We first start by showing the distribution is well defined. So for an arbitrary $X = k$, we have:

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

And now if we show that every part of the expression is non-negative, we know that their multiplications or quotients will be non-negative.

Since $\lambda > 0$ by definition, we know $\lambda^k = \prod_{i=1}^{k}\lambda > 0$ because we are simply multiplying a positive number k times.

Now for $e^{-\lambda}$, we have $e^{-\lambda} = \frac{1}{e^{\lambda}}$ which is non-negative, because e^{λ} is non-negative and if we take the limit of the fraction with $\lambda > 0$ going to infinity, we get $\lim_{\lambda\to\infty}\frac{1}{e^{\lambda}} = 0$.

For the last part, we know $k! = k(k-1)\ldots 1 > 0$, since k is the number of occurrences of an event and so it has to be a non-negative number and even it $k = 0$, we know $0! = 1$.

Since all the parts of the expression are non-negative, we know the whole expression is non-negative.

Now for showing that the Poisson distribution is normalized, we sum over all the possibilities of the events. So we have:

$$\sum_{k=0}^{\infty} p(X = k) = \sum_{k=0}^{\infty}\frac{\lambda^k e^{-\lambda}}{k!} =$$

But since $e^{-\lambda}$ does not depend on on k, we can take it out of the summation as a constant multiplying al the terms.

$$= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} =$$

Now using the given Taylor expansion for the exponential function, we have:

$$= e^{-\lambda}e^{\lambda} = 1$$

And we just showed that $\sum_{k=0}^{\infty} p(X = k) = 1$, so the distribution is normalized indeed.

2.b

We start by computing the mean of the distribution. So by definition, we have:

$$E[X] = \sum_{k=0}^{\infty} kp(X = k) = \sum_{k=0}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} =$$

Evaluating the summation on $k = 0$ and summing the rest, we have:

$$= 0 + \sum_{k=1}^{\infty} k\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} =$$

Using again the fact that $e^{-\lambda}$ is a constant to the summation over k and splitting λ^k in a convenient way, we have:

$$= e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} =$$

Now changing variables $j = k - 1$ and using the given Taylor expansion of the exponential function again, we get:

$$= e^{-\lambda}\lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

So the mean of the Poisson distribution is $E[X] = \lambda$.

Now for the variance, we start again by the definition:

$$\sigma^2(X) = E[X^2] - E[X]^2$$

Since we have $E[x]$, we know that $E[x]^2 = \lambda^2$, so we just need to compute $E[X^2]$:

$$E[X^2] = \sum_{k=0}^{\infty} k^2 p(X = k) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} =$$

Separating $e^{-\lambda}$ from the summation again, splitting the first term, $k = 0$, of the summation and also splitting λ^k, we have:

$$= e^{-\lambda}(0 + \sum_{k=1}^{\infty} k\frac{\lambda\lambda^{k-1}}{(k-1)!}) = e^{-\lambda}\lambda(\sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{(k-1)!}) =$$

Manipulating the expression to get a $(k - 1)$ term, we have:

$$= e^{-\lambda}\lambda \sum_{k=1}^{\infty}((k-1)\frac{\lambda^{k-1}}{(k-1)!} + \frac{\lambda^{k-1}}{(k-1)!}) = e^{-\lambda}\lambda(\sum_{k=1}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}) =$$

Now computing the next term, $k = 1$, of the first summation, $\sum_{k=1}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!}$, we get:

$$= e^{-\lambda}\lambda(0 + \sum_{k=2}^{\infty}(k-1)\frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}) = e^{-\lambda}\lambda(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!})) =$$

Changing variables $j = k - 2$ and $z = k - 1$, and finally using the given Taylor expansion of the exponential function again, we have:

$$= e^{-\lambda}\lambda(\lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} + \sum_{z=0}^{\infty} \frac{\lambda^z}{z!})) = e^{-\lambda}\lambda(\lambda e^{\lambda} + e^{\lambda}) = \lambda^2 + \lambda$$

Going back to the definition of variance, we get:

$$\sigma^2(X) = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

So the variance of the Poisson distribution is $\sigma^2(X) = \lambda$.

2.c

Given a fixed time interval Δt, the Poisson distribution changes a bit. Normally λ refers to the average rate of occurrences of an event per unit time, but now we are working with a specific time interval, which leads to:

$$p(X = k) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}$$

For the likelihood of the events, we have to compute $p(k_1, \dots, k_N|\lambda)$, but since they are i.i.d, we have:

$$p(k_1, \dots, k_N|\lambda) = \prod_{i=1}^{N} p(k_i) = \prod_{i=1}^{N} \frac{(\lambda \Delta t)^{k_i} e^{-\lambda \Delta t}}{k_i!}$$

Writing the log-likelihood, thus, gives us:

$$\log \prod_{i=1}^{N} \frac{(\lambda \Delta t)^{k_i} e^{-\lambda \Delta t}}{k_i!} = \sum_{i=1}^{N} \log(\frac{(\lambda \Delta t)^{k_i} e^{-\lambda \Delta t}}{k_i!}) =$$

$$= \sum_{i=1}^{N} \log((\lambda \Delta t)^{k_i} e^{-\lambda \Delta t}) - \sum_{i=1}^{N} \log(k_i!) = \sum_{i=1}^{N} \log(\lambda \Delta t)^{k_i} + \sum_{i=1}^{N} -\lambda \Delta t - \sum_{i=1}^{N} \log(k_i!) =$$

$$= \sum_{i=1}^{N} k_i \log(\lambda \Delta t) - \sum_{i=1}^{N} \log(k_i!) - N\lambda \Delta t$$

So the log-likelihood of the events assuming i.i.d and that all events are Poisson processes measured in a fixed interval time Δt is given by $\sum_{i=1}^{N} k_i \log(\lambda \Delta t) - \sum_{i=1}^{N} \log(k_i!) - N\lambda \Delta t$.

2.d

For this exercise we start by taking log-likelihood derivative w.r.t λ and equal to zero:

$$\lambda_{ML} = \arg\max_{\lambda} \sum_{i=1}^{N} k_i \log(\lambda \Delta t) - \sum_{i=1}^{N} \log(k_i!) - N\lambda \Delta t$$

Which leads to:

$$\frac{\partial}{\partial \lambda}(\sum_{i=1}^{N} k_i \log(\lambda \Delta t) - \sum_{i=1}^{N} \log(k_i!) - N\lambda \Delta t) = 0$$

By the sum rule, we can differentiate each part of the equation separately and then put them all together. So we have:

$$\frac{\partial}{\partial \lambda} \sum_{i=1}^{N} k_i \log(\lambda \Delta t) = \sum_{i=1}^{N} k_i \frac{1}{\lambda \Delta t} \Delta t = \sum_{i=1}^{N} \frac{k_i}{\lambda}$$

$$\frac{\partial}{\partial \lambda} \sum_{i=1}^{N} \log(k_i!) = 0$$

$$\frac{\partial}{\partial \lambda} N\lambda \Delta t = N\Delta t$$

With all parts in hand, we continue:

$$\sum_{i=1}^{N} \frac{k_i}{\lambda} - N\Delta t = 0 \implies \frac{1}{\lambda} \sum_{i=1}^{N} k_i = N\Delta t \implies$$

$$\implies \sum_{i=1}^{N} k_i = N\Delta t \lambda \implies \lambda_{ML} = \frac{\sum_{i=1}^{N} k_i}{N\Delta t}$$

So the maximum likelihood estimator is $\lambda_{ML} = \frac{\sum_{i=1}^{N} k_i}{N\Delta t}$.

2.e

Since λ_{ML} represents an estimate of the average amount of people entering a store every minute, we just need to multiply it by a factor of 60 to get an estimator for an hour. So we have $\lambda_h = 60 * \lambda_{ML}$.

2.f

To find λ_{MAP}, we start by detailing its expression:

$$\lambda_{MAP} = \arg\max_{\lambda} p(\lambda|k) =$$

From Bayes rule, we get:

$$= \arg\max_{\lambda} \frac{p(k|\lambda)p(\lambda)}{p(k)} = \arg\max_{\lambda} \frac{p(k|\lambda)p(\lambda)}{\int p(k|\lambda)p(\lambda)\partial\lambda}$$

As the question states, it can be hard to compute the integral in the denominator of the above expression. But because we are working with maximum a posteriori estimation, we are using the arg max operator on λ. This means we are looking through all the possible values of λ, and all of them will have the same common denominator, so we can simply ignore the integral and work on finding the λ that maximizes $p(k|\lambda)p(\lambda)$ without any effect on the estimator found. We then continue our computations:

$$\lambda_{MAP} = \arg\max_{\lambda} p(k|\lambda)p(\lambda) =$$

Using the log trick, we get:

$$= \arg\max_{\lambda} \log(p(k|\lambda)p(\lambda)) = \arg\max_{\lambda} \log p(k|\lambda) + \log p(\lambda)$$

Since we already have the log-likelihood, we just need to compute the expression for the prior $\log p(\lambda)$ before continuing with our λ_{MAP} computations. Considering the given prior distribution, we have:

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

$$= \alpha_1 \log \alpha_2 - \log \Gamma(\alpha_1) + (\alpha_1 - 1)\log\lambda - \alpha_2\lambda =$$

Expanding the Gamma function $\Gamma(x) = (x-1)!$:

$$= \alpha_1 \log \alpha_2 - (\sum_{i=1}^{\alpha_1 - 1} \log(\alpha_1 - i)) + (\alpha_1 - 1)\log\lambda - \alpha_2\lambda$$

Going back to the original equation $\arg\max_{\lambda} \log p(k|\lambda) + \log p(\lambda)$, we get:

$$\arg\max_{\lambda} \log p(k|\lambda) + \log p(\lambda) = \arg\max_{\lambda} \sum_{i=1}^{N} k_i \log\lambda - \sum_{i=1}^{N} \log(k_i!) - N\lambda + \alpha_1 \log\alpha_2 - (\sum_{i=1}^{\alpha_1 - 1} \log(\alpha_1 - i)) + (\alpha_1 - 1)\log\lambda - \alpha_2\lambda =$$

$$= \arg\max_{\lambda} \log\lambda \sum_{i=1}^{N} k_i - \sum_{i=1}^{N} \log(k_i!) + \alpha_1 \log\alpha_2 - (\sum_{i=1}^{\alpha_1} \log(\alpha_1 - i)) + (\alpha_1 - 1)\log\lambda - (N + \alpha_2)\lambda =$$

$$= \arg\max_{\lambda} ((\sum_{i=1}^{N} k_i) + \alpha_1 - 1)\log\lambda - (N + \alpha_2)\lambda - \sum_{i=1}^{N} \log(k_i!) + \alpha_1 \log\alpha_2 - (\sum_{i=1}^{\alpha_1 - 1} \log(\alpha_1 - 1))$$

But to get to the value that maximizes the above expression w.r.t λ, we would take its derivative and equal it to zero, which would zero out all the terms that do not depend on λ. Therefore, we can rewrite the expression with only the terms that depend on λ without compromising the maximum a posteriori estimation:

$$\lambda_{MAP} = \arg\max_{\lambda} ((\sum_{i=1}^{N} k_i) + \alpha_1 - 1)\log\lambda - (N + \alpha_2)\lambda$$

Exactly what was expected.

2.g

To find the MAP estimator we need to take the derivative of the expression we got on the last question w.r.t λ and equal it to zero. So we have:

$$\frac{\partial}{\partial\lambda}((\sum_{i=1}^{N} k_i) + \alpha_1 - 1)\log\lambda - (N+\alpha_2)\lambda = 0$$

By the sum rule:

$$\frac{(\sum_{i=1}^{N} k_i) + \alpha_1 - 1}{\lambda} - (N+\alpha_2) = 0$$

$$(\sum_{i=1}^{N} k_i) + \alpha_1 - 1 = (N+\alpha_2)\lambda$$

$$\lambda_{MAP} = \frac{(\sum_{i=1}^{N} k_i) + \alpha_1 - 1}{(N+\alpha_2)}$$

Thus, the MAP estimator is $\lambda_{MAP} = \frac{(\sum_{i=1}^{N} k_i)+\alpha_1-1}{(N+\alpha_2)}$.

2.h

Since we are only interested in showing the posterior distribution is also a Gamma distribution, we can ignore the evidence, which works as a normalizing constant, and compute $p(\lambda|k) \propto p(k|\lambda)p(\lambda)$:

$$p(k|\lambda)p(\lambda) = \prod_{i=1}^{N} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2\lambda} =$$

$$= \frac{\lambda^{\sum_{i=1}^{N} k_i} e^{-N\lambda}}{\prod_{i=1}^{N} k_i!} \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2\lambda} \propto \lambda^{(\sum_{i=1}^{N} k_i)+\alpha_1-1} e^{-(N+\alpha_2)\lambda}$$

And from the last expression, we observe that $p(k|\lambda)p(\lambda)$ is indeed proportional to a Gamma distribution with parameters $\alpha_1' = (\sum_{i=1}^{N} k_i) + \alpha_1$ and $\alpha_2' = N + \alpha_2$, so $p(\lambda|k)$ is indeed proportional to a Gamma distribution.

3 General Multiple Outputs Linear Regression

3.a

We can inspect the dimensions of the parameter W by understanding the dimensions of the other parameters around it. We know that $y(x, W) \in \mathbb{R}^{K\times 1}$, since we want to predict an output for each one of the K targets. We also know that $\phi(x)$ is an M-dimensional vector, so $\phi(x) \in \mathbb{R}^{M\times 1}$. By using these facts and the given equation $y(x, W) = W^T\phi(x)$, we conclude that $W^T \in \mathbb{R}^{K\times M}$, which then leads to the answer $W \in \mathbb{R}^{M\times K}$.

3.b

For each target vector, we have:

$$p(t_i|W,\Sigma) = N(t_i|y(x_i,W),\Sigma) = \frac{1}{\sqrt{(2\pi)^K \det\Sigma}} e^{\frac{-1}{2}(t_i - y(x_i,W))^T\Sigma^{-1}(t_i - y(x_i,W))}$$

So the log-likelihood is given by:

$$\log p(t_i|W,\Sigma) = \log \frac{1}{\sqrt{(2\pi)^K \det\Sigma}} e^{-\frac{1}{2}(t_i - y(x_i,W))^T\Sigma^{-1}(t_i - y(x_i,W))} =$$

$$= (-\frac{1}{2}(t_i - y(x_i,W))^T\Sigma^{-1}(t_i - y(x_i,W))) - \log\sqrt{(2\pi)^K \det\Sigma} =$$

$$= (-\frac{1}{2}(t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W))) - \frac{1}{2}\log((2\pi)^K \det \Sigma) =$$

$$= (-\frac{1}{2}(t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W))) - \frac{K}{2}\log(2\pi) - \frac{1}{2}\log\det \Sigma) =$$

$$= -\frac{1}{2}((t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W)) + K\log 2\pi + \log\det \Sigma)$$

Now that we have the log-likelihood for each target, we can compute the log-likelihood for all N independent observations. It will turn out to be a sum of log terms, since we can multiply the likelihoods for all the independent observations. So we have:

$$p(T|W, \Sigma) = \prod_{i=1}^{N} N(t_i|y(x_i, W), \Sigma)$$

$$\log p(T|W, \Sigma) = \sum_{i=1}^{N} \log N(t_i|y(x_i, W), \Sigma) =$$

And now using the previous results:

$$= \sum_{i=1}^{N} -\frac{1}{2}((t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W)) + K\log 2\pi + \log\det \Sigma) =$$

$$= -\frac{1}{2}(NK\log 2\pi + N\log\det \Sigma + \sum_{i=1}^{N}(t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W)))$$

So the log-likelihood of T is given by $\log p(T|W, \Sigma) = -\frac{1}{2}(NK\log 2\pi + N\log\det \Sigma + \sum_{i=1}^{N}(t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W)))$.

3.c

To find the maximum likelihood solution W_{ML}, we need to take the derivative w.r.t W of the log-likelihood function computed above and equal it to zero. We start by computing the derivative w.r.t W^T and proceed to prove that its equal to the transpose of the derivative w.r.t W. So we have:

$$\frac{\partial}{\partial W^T}(-\frac{1}{2}(NK\log 2\pi + N\log\det \Sigma + \sum_{i=1}^{N}(t_i - y(x_i, W))^T \Sigma^{-1}(t_i - y(x_i, W)))) =$$

We remove the first two terms $\frac{NK}{2}\log 2\pi$ and $\frac{N}{2}\log\det \Sigma$ since their derivative w.r.t W^T is zero. So we get:

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

And now by the sum rule:

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

Making use of the hint given to us $\frac{\partial}{\partial A}(x - As)^T W(x - As) = -2W(x - As)s^T$ and the fact that the derivative w.r.t the transpose is equal to the transpose of the derivative w.r.t the orignal matrix , we have:

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

Transforming it to the matrix form, we get:

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

Now equaling it to zero, we have:

$$\Phi^T T\Sigma^{-1} - \Phi^T\Phi W\Sigma^{-1} = 0$$
$$\Phi^T\Phi W\Sigma^{-1} = \Phi^T T\Sigma^{-1}$$

But we know that the covariance matrix is symmetric and positive definite, thus it has an inverse - that we've been using - and we can do:

$$\Phi^T\Phi W\Sigma^{-1}\Sigma = \Phi^T T\Sigma^{-1}\Sigma$$
$$\Phi^T\Phi W = \Phi^T T$$

We also know that $\Phi^T\Phi$ should be invertible as long as we get linear independent x input vectors. Assuming we do, we then have:

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

So now we have shown that the maximum likelihood solution is $W_{ML} = (\Phi^T\Phi)^{-1}\Phi^T T$ and thus is independent of the covariance matrix Σ. We just need to show that this result is valid by proving the derivative of the likelihood w.r.t W^T is equal to the transpose of the derivative of the likelihood w.r.t W:

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

Focusing on the derivative term and expanding to then use index notation, we have:

$$\frac{\partial}{\partial W}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i)) = \frac{\partial}{\partial W}(t_i^T - \phi(x_i)^T W)\Sigma^{-1}(t_i - W^T\phi(x_i)) =$$
$$= \frac{\partial}{\partial W}(t_i^T\Sigma^{-1} - \phi(x_i)^T W\Sigma^{-1})(t_i - W^T\phi(x_i)) =$$
$$= \frac{\partial}{\partial W}(t_i^T\Sigma^{-1}t_i - \phi(x_i)^T W\Sigma^{-1}t_i - t_i^T\Sigma^{-1}W^T\phi(x_i) + \phi(x_i)^T W\Sigma^{-1}W^T\phi(x_i))$$

By the sum rule, we can compute each derivative separately:

1. $$\frac{\partial}{\partial W}t_i^T\Sigma^{-1}t_i = 0$$

2. $$\frac{\partial}{\partial W}\phi(x_i)^T W\Sigma^{-1}t_i \implies \sum_{p=1}^{M}\sum_{q=1}^{K}\sum_{r=1}^{K}\frac{\partial}{\partial W_{m,n}}\phi(x_i)_p W_{p,q}\Sigma^{-1}_{q,r}(t_i)_r = \sum_{r=1}^{K}\phi(x_i)_m\Sigma^{-1}_{n,r}(t_i)_r \implies$$
$$\implies \frac{\partial}{\partial W}\phi(x_i)^T W\Sigma^{-1}t_i = \phi(x_i)(\Sigma^{-1}t_i)^T$$

3. $$\frac{\partial}{\partial W}t_i^T\Sigma^{-1}W^T\phi(x_i) \implies \sum_{p=1}^{K}\sum_{q=1}^{K}\sum_{r=1}^{M}\frac{\partial}{\partial W_{m,n}}(t_i)_p\Sigma^{-1}_{p,q}W_{q,r}\phi(x_i)_r = \sum_{p=1}^{K}(t_i)_p\Sigma^{-1}_{p,m}\phi(x_i)_n \implies$$
$$\implies \frac{\partial}{\partial W}t_i^T\Sigma^{-1}W^T\phi(x_i) = (t_i^T\Sigma^{-1})^T\phi(x_i)^T = \phi(x_i)(\Sigma^{-1}t_i)^T$$

4. $$\frac{\partial}{\partial W}\phi(x_i)^T W\Sigma^{-1}W^T\phi(x_i) \implies \sum_{p=i}^{M}\sum_{q=1}^{K}\sum_{r=1}^{K}\sum_{s=1}^{M}\frac{\partial}{\partial W_{m,n}}\phi(x_i)_p W_{p,q}\Sigma^{-1}_{q,r}W_{r,s}\phi(x_i)_s =$$

By the product rule:

$$= \sum_{p=i}^{M}\sum_{q=1}^{K}\sum_{r=1}^{K}\sum_{s=1}^{M}(\phi(x_i)_m \Sigma^{-1}_{n,r} W_{r,s}\phi(x_i)_s + \phi(x_i)_p W_{p,q}\Sigma^{-1}_{q,m}\phi(x_s)_n) =$$

$$= \sum_{r=1}^{K}\sum_{s=1}^{M}\phi(x_i)_m \Sigma^{-1}_{n,r} W_{r,s}\phi(x_i)_s + \sum_{p=i}^{M}\sum_{q=1}^{K}\phi(x_i)_p W_{p,q}\Sigma^{-1}_{q,m}\phi(x_s)_n \implies$$

$$\implies \frac{\partial}{\partial W}\phi(x_i)^T W \Sigma^{-1} W^T \phi(x_i) = \phi(x_i)(\Sigma^{-1}W^T\phi(x_i))^T + \phi(x_i)\phi(x_i)^T W\Sigma^{-1} =$$

Using the symmetry of Σ and thus Σ^{-1} :

$$= \phi(x_i)\phi(x_i)^T W\Sigma^{-1} + \phi(x_i)\phi(x_i)^T W\Sigma^{-1} = 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1}$$

Putting all the terms together, we have:

$$\frac{\partial}{\partial W}(t_i - W^T\phi(x_i))^T\Sigma^{-1}(t_i - W^T\phi(x_i)) = -\phi(x_i)(\Sigma^{-1}t_i)^T - \phi(x_i)(\Sigma^{-1}t_i)^T + 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1} =$$

$$= -2\phi(x_i)(\Sigma^{-1}t_i)^T + 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1} = -2\phi(x_i)t_i^T\Sigma^{-1} + 2\phi(x_i)\phi(x_i)^T W\Sigma^{-1} =$$
$$= -2\phi(x_i)(t_i^T - \phi(x_i)^T W)\Sigma^{-1}$$

And if we put the obtained equation back to the original expression, we get:

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

Exactly the expression we got by using the hint and thus proving that indeed the derivative of the likelihood w.r.t W^T is equal to the transpose of the derivative of the likelihood w.r.t W.

So we can finally state that $W_{ML} = (\Phi^T\Phi)^{-1}\Phi^T T$.

3.d

We repeat the same process but now taking the derivative w.r.t $\Omega = \Sigma^{-1}$:

$$\frac{\partial}{\partial\Omega}(-\frac{1}{2}(NK\log 2\pi + N\log\det\Omega^{-1} + \sum_{i=1}^{N}(t_i - y(x_i, W))^T\Omega(t_i - y(x_i, W)))) =$$

We cancel out $-\frac{NK}{2}\log 2\pi$ since it does not depend on Ω:

$$= \frac{\partial}{\partial\Omega}(-\frac{1}{2}(N\log\det\Omega^{-1} + \sum_{i=1}^{N}(t_i - y(x_i, W))^T\Omega(t_i - y(x_i, W))))$$

Separating both parts and computing them, we have:

$$\frac{\partial}{\partial\Omega}\log\det\Omega^{-1} =$$

By the relationship $\det X^{-1} = \frac{1}{\det X}$, we get:

$$= \frac{\partial}{\partial\Omega}\log\frac{1}{\det\Omega} = \frac{\partial}{\partial\Omega}(-\log\det\Omega)$$

Now using the given identity, we have:

$$\frac{\partial}{\partial\Omega}\log\det\Omega^{-1} = -(\Omega^{-1})^T$$

And for the second part:

$$\frac{\partial}{\partial\Omega}\sum_{i=1}^{N}(t_i - y(x_i, W))^T\Omega(t_i - y(x_i, W)) = \frac{\partial}{\partial\Omega}\sum_{i=1}^{N}(t_i - W^T\phi(x_i))^T\Omega(t_i - W^T\phi(x_i)) =$$

Now to use the second identity that was given to us, we need to transform Ω into Ω^T. We can do that simply by taking the transpose, since we know that covariance matrix is symmetric and the inverse of a symmetric matrix is also symmetric. Therefore, changing the equation to get to use the identity and also by the sum rule, we get:

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

Now putting everything together, we have:

$$\frac{\partial}{\partial\Omega}(-\frac{1}{2}(N\log\det\Omega^{-1} + \sum_{i=1}^{N}(t_i - y(x_i, W))^T\Omega(t_i - y(x_i, W)))) =$$

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

Finally equaling it zero:

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

Going back to Σ, we have:

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

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

And using the symmetry of covariance matrix, we get:

$$\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(t_i - W_{ML}^T\phi(x_i))(t_i - W_{ML}^T\phi(x_i))^T$$

So it's shown that the maximum likelihood solution for Σ is given by $\Sigma_{ML} = \frac{1}{N}\sum_{i=1}^{N}(t_i - W_{ML}^T\phi(x_i))(t_i - W_{ML}^T\phi(x_i))^T$.

4 Bayesian Linear Regression

4.a

To obtain the reported posterior, a Gaussian distribution $N(w|\mu_0, \Sigma_0)$ has been used as the prior. This can be understood exactly by answering the hint question. If no data is observed, the posterior should look like $p(w|t) = N(w|\mu_N, \Sigma_N)$ with $N = 0$, so $\mu_N = \Sigma_0\Sigma_0^{-1}\mu_0 = \mu_0$ and $\Sigma_N^{-1} = \Sigma_0^{-1}$.

4.b

When the precision β approaches zero it means we lost confidence on our model, so adding new information does not update the model parameters. This can also be seen mathematically by $\lim_{\beta\to 0}\Sigma_{N+1}^{-1} = \lim_{\beta\to 0}\Sigma_N^{-1} + \beta\phi_{N+1}\phi_{N+1}^T = \Sigma_N^{-1}$, which also leads to $\lim_{\beta\to 0}\mu_{N+1} = \lim_{\beta\to 0}\Sigma_N(\Sigma_N^{-1}\mu_N + \beta\phi_{N+1}t_{N+1}) = \mu_N$, so $\Sigma_{N+1} = \Sigma_N$ and $\mu_{N+1} = \mu_N$.

4.c

Using the fact that the posterior becomes the prior for a new posterior after a new observation, we can write the updated posterior as:

$$p(w|\Phi_{N+1}, t_{N+1}, \Sigma_0, \mu_0, \beta) = \frac{p(t_{N+1}|w, \Phi_{N+1}, \Sigma_0, \mu_0, \beta)p(w|\Phi_N, t_N, \Sigma_0, \mu_0, \beta)}{\int p(t_{N+1}|w, \Phi_{N+1}, \Sigma_0, \mu_0, \beta)p(w|\Phi_N, t_N, \Sigma_0, \mu_0, \beta)\partial w}$$