W14 Problem Sheet Explanation: Maximum Likelihood

Question 1

If the concept of maximum likelihood estimation is bit confusing, you can find a good explanation here.

So, in this question it is stated that we are given information on the current, I and the resistance, R and we are trying to estimate the voltage, V. So, in this situation, our student can change the current or resistance and measure the other, so the student will have pairs of data. Let's say that the student measures these values to be:

$$I = [5.1, 2.5, 1.6]$$

$$R = [1, 2, 3]$$

Now, we have the equation $I = \frac{V}{R} + \epsilon$ and we know that V is a fixed value however we do not know what that value is. If we did not have an error in the equation, then we could easily work out V, however that is not the case.

In this question we have been given $p(D|V)$ - this is the probability that we observe our data (so our I and R data) assuming we know what the *true value* of V is.
Often, you will be given $p(d_i|V)$ which is the probability that you observe a singular data point (so I_i and R_i). To extend this to get the probability that the whole data set has occurred we just take the product of the singular case - this product is the equation in which we have been given.

We start with the equation:

$$p(D|V) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(I_i - \frac{V}{R_i})^2}{\sigma^2}} \tag{1}$$

Now, this is our **likelihood function** - a function showing us the probability that our data points came from the distribution with set values for V.

The aim of maximum likelihood estimation is to find the value of V which is most likely to have given us the data set D, i.e the value of V which gives us the highest probability $p(D|V)$.

So, we have an equation and we need to find the input which gives us the maximum output - how have we achieved this in the past? We take the derivative of the function and solve for 0 - this will give us our maximum point.

Now, we would do this with just our likelihood function, however taking derivatives of lots of products can be really quite painful. One way around this is to take the log of the function, which quite nicely turns any products into summations, meaning we can take the derivatives of individual parts. We normally use the natural logarithm as this helps to cancel out any exponentials.

So, these are the steps for calculating the MLE which we will follow for this example:

1. Get the likelihood function by taking the product of $p(d_i|\theta)$ to get $p(D|\theta)$.
2. Get the log-likelihood function.
3. Calculate the derivative of the log-likelihood function.
4. Set the derivative to be equal to 0 and solve to get an estimate of the parameters, θ.

Step 1: Get the likelihood function
This has already been done for us - it is equation (1).

Step 2: Get the log-likelihood function

$$\begin{aligned}
\mathcal{L}(D|V) &= \ln p(D|V) && (2)\\
&= \ln\left(\prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(I_i-\frac{V}{R_i})^2}{\sigma^2}}\right) && (3)\\
&= \sum_i \left(\ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(I_i-\frac{V}{R_i})^2}{\sigma^2}}\right)\right) && (4)\\
&= \sum_i \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_i \ln\left(e^{-\frac{1}{2}\frac{(I_i-\frac{V}{R_i})^2}{\sigma^2}}\right. && (5)\\
&= \left(n * \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\right) + \sum_i \left(-\frac{1}{2}\frac{(I_i-\frac{V}{R_i})^2}{\sigma^2}\right) && (6)
\end{aligned}$$

In the working above, we went from (3) to (4) by using the multiplicative rule for logarithms, i.e $\ln(a) + \ln(b) = \ln(ab)$. This rule is also used to get from (4) to (5). To get the (6), we simplify both terms: the first term does not contain i and so a summation over values of i is the same as an addition of the term n times. The second term is simplified through the cancellation of $\ln$ and e.

Step 3: Calculate the derivative
Next, let's calculate the derivative of our log-likelihood function. We are trying to

find the value of V which gives us the maximum value of $\mathcal{L}(D|V)$ and hence we calculate the derivative with respect to V.

$$\frac{d}{dV}\mathcal{L}(D|V) = \frac{d}{dV}\left(n\ln\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{d}{dV}\left(\sum_i -\frac{1}{2}\frac{(I_i - \frac{V}{R_i})^2}{\sigma^2}\right) \tag{7}$$

$$= 0 + \sum_i \frac{d}{dV}\left(-\frac{1}{2}\frac{(I_i - \frac{V}{R_i})^2}{\sigma^2}\right) \tag{8}$$

$$= \sum_i -\frac{1}{2}\cdot 2 \cdot -\frac{1}{R_i}\cdot\frac{(I_i - \frac{V}{R_i})}{\sigma^2} \tag{9}$$

$$= \sum_i \frac{I_i - \frac{V}{R_i}}{R_i\sigma^2} \tag{10}$$

Step 4: Set derivative = 0 and solve
Now, we want to find the value V_{ML} which makes the derivative 0:

$$0 = \sum_i \frac{I_i - \frac{V_{ML}}{R_i}}{R_i\sigma^2} \tag{11}$$

$$\iff \quad 0 = \frac{1}{\sigma^2}\sum_i\left(\frac{I_i}{R_i} - \frac{V_{ML}}{R_i^2}\right) \tag{12}$$

$$\iff \quad 0 = \sum_i\left(\frac{I_i}{R_i} - \frac{V_{ML}}{R_i^2}\right) \tag{13}$$

$$\iff \quad \sum_i \frac{V_{ML}}{R_i^2} = \sum_i \frac{I_i}{R_i} \tag{14}$$

$$\iff \quad V_{ML} = \sum_i \frac{I_i}{R_i} \Big/ \sum_i \frac{1}{R_i^2} \tag{15}$$

Question 2

Finding the MAP value (Maximum A Posteriori) is also known as Bayesian inference. We normally use Bayesian inference over maximum likelihood estimation when we don't have many data points to estimate from.

Bayesian inference comes from Bayes' rule, which is given as:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

In terms of when we consider data and parameters, this is:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

Each of these terms is actually given a name, which match up to the following:

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

Maximum Likelihood Estimate uses the likelihood only, $p(D|\theta)$, to calculate an estimate for θ. MAP builds on this by also introducing a prior.

A simple way to think of what a prior is, $p(\theta)$, is to think of it as some extra information about which parameters are more likely to be valid or not. Consider that you found a coin on the floor and you want to estimate θ = the probability of flipping a head. If we flip the coin twice and get 2 tails, then using MLE we might get the estimate that $\theta = 0$. However, if this coin was found on the ground then we would probably assume that it is more likely that $\theta = 0.5$, which contradicts our MLE estimate. This is essentially what a prior is - an assumption about what the parameters are. When multiplied with the likelihood, this will give a higher weighting to those parameters which are more likely.

To get the posterior, we do not need to calculate the evidence, $p(D)$, as this can be considered a normalising constant. *Note: A normalising constant within probability functions is a constant multiplied by the function so that the sum of all probabilities add up to 1.*

Hence, we can write:

$$Posterior \propto Likelihood \times Prior$$

This is why we can use $p(D|\theta)p(\theta)$ to get an estimate for θ.

Let's go through the steps for calculating the MAP:

Step 1: Calculate the likelihood x prior

$$\begin{aligned} p(D|\theta)p(\theta) &= b\, e^{-(3-\theta)^2} \cdot c\, e^{-\theta(\theta-1)} && (16) \\ &= b\, c\, e^{-(3-\theta)^2-\theta(\theta-1)} && (17) \end{aligned}$$

Step 2: Calculate the natural logarithm

$$\begin{aligned} \ln\left(p(D|\theta)p(\theta)\right) &= \ln\left(b\, c\, e^{-(3-\theta)^2-\theta(\theta-1)}\right) && (18) \\ &= \ln b + \ln c + \ln e^{-(3-\theta)^2-\theta(\theta-1)} && (19) \\ &= \ln b + \ln c + -(3-\theta)^2 - \theta(\theta-1) && (20) \end{aligned}$$

Step 3: Calculate the derivative

$$\begin{aligned} \frac{d}{d\theta}\left(\ln p(D|\theta)p(\theta)\right) &= \frac{d}{d\theta}\left(-(3-\theta)^2 - \theta(\theta-1)\right) && (21) \\ &= -1 \cdot -2(3-\theta) + (-1)\cdot(\theta-1) + (-\theta)\cdot 1 && (22) \\ &= 2(3-\theta) - 2\theta + 1 && (23) \\ &= -4\theta + 7 && (24) \end{aligned}$$

Step 4: Set derivative to 0 and solve

$$
\begin{aligned}
0 &= -4\theta_{MAP} + 7 && (25)\\
4\theta_{MAP} &= 7 && (26)\\
\theta_{MAP} &= 7/4 && (27)
\end{aligned}
$$

Question 3

Now, let's use both methods above for the next question.

Maximum Likelihood Estimation
First, we want to calculate $p(X|\theta)$. We have been given a table in which that if we knew the value of θ, we could work out the probability that each value of X would come up. So, we have been given $p(x_i|\theta)$, i.e $p(x_i = 1|\theta) = \frac{\theta}{3}$ and we just need to take the product over each value of X to get $p(X|theta)$.

Step 1
In X we have two 0s, three 1s, three 2s and two 3s (note that the ordering does not matter). So:

$$
\begin{aligned}
p(X|\theta) &= p(x_i = 3|\theta) \cdot p(x_i = 0|\theta) \cdot ... \cdot p(x_i = 1|\theta) && (28)\\
&= \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{\theta}{3}\right)^3 \cdot \left(\frac{2(1-\theta)}{3}\right)^3 \cdot \left(\frac{(1-\theta)}{3}\right)^2 && (29)
\end{aligned}
$$

Step 2
Now, we want log-likelihood, so:

$$
\begin{aligned}
\ln p(X|\theta) &= \ln\left(\frac{2\theta}{3}\right)^2 + \ln\left(\frac{\theta}{3}\right)^3 + \ln\left(\frac{2(1-\theta)}{3}\right)^3 + \ln\left(\frac{(1-\theta)}{3}\right)^2 && (30)\\
&= 2\ln\left(\frac{2\theta}{3}\right) + 3\ln\left(\frac{\theta}{3}\right) + 3\ln\left(\frac{2(1-\theta)}{3}\right) + 2\ln\left(\frac{(1-\theta)}{3}\right) && (31)\\
&= 2\ln\left(\frac{2}{3}\right) + 2\ln\theta + 3\ln\left(\frac{1}{3}\right) + 3\ln\theta + 3\ln\left(\frac{2}{3}\right) && (32)\\
&\quad + 3\ln(1-\theta) + 2\ln\left(\frac{1}{3}\right) + 2\ln(1-\theta) && (33)\\
&= 2\ln\theta + 3\ln\theta + 3\ln(1-\theta) + 2\ln(1-\theta) + C && (34)\\
&= 5\ln\theta + 5\ln(1-\theta) + C && (35)
\end{aligned}
$$

In step (31) the powers were brought in front of the logarithms. Step (32) split each term up into logarithms of the constants and the θ terms. Steps (34) and (35) take all the terms which do not include θ and group it into a single constant term C.

Step 3
Now, let's get the derivative:

$$\frac{d}{d\theta}\left(\ln p(X|\theta)\right) = \frac{d}{d\theta}5\ln\theta + \frac{d}{d\theta}5\ln(1-\theta) + \frac{d}{d\theta}C \tag{36}$$
$$= \frac{5}{\theta} + (-1)\cdot\frac{5}{(1-\theta)} + 0 \tag{37}$$
$$= \frac{5}{\theta} - \frac{5}{1-\theta} \tag{38}$$

Step 4
Now, let's set the derivative equal to 0 and solve to get our estimate.

$$0 = \frac{5}{\theta_{ML}} - \frac{5}{1-\theta_{ML}} \tag{39}$$
$$\iff \quad \frac{5}{1-\theta_{ML}} = \frac{5}{\theta_{ML}} \tag{40}$$
$$\iff \quad 5\theta_{ML} = 5(1-\theta_{ML}) \tag{41}$$
$$\iff \quad 10\theta_{ML} = 5 \tag{42}$$
$$\implies \quad \theta_{ML} = \frac{1}{2} \tag{43}$$

Maximum A Posteriori Estimate
Now, let's use your other method - we have been given the extra information that $p(\theta) = b\,\theta\,(1-\theta)$ so let's go through our steps.

Step 1
We will use what we calculated before during our working for the MLE and multiply it by $p(\theta)$.

$$p(X|\theta)p(\theta) = \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{\theta}{3}\right)^3 \cdot \left(\frac{2(1-\theta)}{3}\right)^3 \cdot \left(\frac{(1-\theta)}{3}\right)^2 \cdot b\cdot\theta\cdot(1-\theta) \tag{44}$$

Step 2
Instead of recalculating how to simplify down the logarithm, we can reuse our steps from earlier.

$$\ln\left(p(X|\theta)p(\theta)\right) = 5\ln\theta + 5\ln(1-\theta) + C + \ln b + \ln\theta + \ln(1-\theta) \tag{45}$$
$$= 6\ln\theta + 6\ln(1-\theta) + C + \ln b \tag{46}$$
$$= 6\ln\theta + 6\ln(1-\theta) + C \tag{47}$$

Step 3

$$\frac{d}{d\theta}\ln\left(p(X|\theta)p(\theta)\right) = \frac{6}{\theta} - \frac{6}{1-\theta} \tag{48}$$

Step 4

$$
\begin{aligned}
&0 = \frac{6}{\theta_{MAP}} - \frac{6}{1-\theta_{MAP}} && (49)\\
\iff \quad & \frac{6}{1-\theta_{MAP}} = \frac{6}{\theta_{MAP}} && (50)\\
\iff \quad & 6\theta_{MAP} = 6(1-\theta_{MAP}) && (51)\\
\iff \quad & 12\theta_{MAP} = 6 && (52)\\
\implies \quad & \theta_{MAP} = \frac{1}{2} && (53)
\end{aligned}
$$