Assignment 3

Harshvardhan Agarwal
200050050

Parth Dwivedi
200050100

October 2021

1 Question 1

1.1 Maximum Likelihood Estimate

We shall aim to find an expression for the Maximum Likelihood Estimate:-

$$\begin{aligned}\hat{\mu}^{\mathrm{ML}} &= arg\ max_{\mu} P(x_1, x_2, \ldots, x_N|\mu) \\ &= arg\ max_{\mu} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\end{aligned}$$

Maximizing the expression above is equivalent to maximizing the log of the function, since log is a monotonically increasing function.
To do this, we take the derivative of the log of the function and set it to 0.

$$\begin{aligned}\frac{d}{d\mu} log \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} &= \frac{d}{d\mu} \sum_{i=1}^{N} \left(-\frac{(x_i-\mu)^2}{2\sigma^2} - log(\sigma\sqrt{2\pi}) \right) \\ &= \sum_{i=1}^{N} -\frac{\mu - x_i}{\sigma^2}\end{aligned}$$

Setting it to 0, we get:-

$$\begin{aligned}\mu &= \frac{\sum_{i=1}^{N} x_i}{N} \\ \therefore\ \hat{\mu}^{\mathrm{ML}} &= \frac{\sum_{i=1}^{N} x_i}{N}\end{aligned}$$

1.2 Maximum A-Posteriori Estimate

We shall aim to find an expression for the Maximum A-Posteriori Estimate:-

$$\begin{aligned}\hat{\mu}^{\mathrm{MAP}_1} &= arg\ max_{\mu} P(\mu|x_1, x_2, \ldots, x_N) \\ &= arg\ max_{\mu} \frac{P(x_1, x_2, \ldots, x_N|\mu)P(\mu)}{\int_{\mu} P(x_1, x_2, \ldots, x_N, \mu) d\mu}\end{aligned}$$

Since the denominator is not a function of μ, it will have no effect on $\hat{\mu}^{\mathrm{MAP}_1}$, hence it can be neglected.
So our expression now becomes:-

$$
\begin{aligned}
\hat{\mu}^{\text{MAP}_1} &= arg\ max_\mu \left[P(x_1, x_2, \ldots, x_N|\mu)P(\mu)\right] \\
&= arg\ max_\mu \left[P(\mu)\prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]
\end{aligned}
$$

To maximize this, as we did before, we take the derivative of the log of this expression with respect to μ and set it to 0.

1.2.1 Case 1 : Gaussian Prior

$$
P(\mu) = \frac{1}{\sigma_{prior}\sqrt{2\pi}}e^{-\frac{(\mu-\mu_{prior})^2}{2\sigma_{prior}^2}}
$$

So,

$$
\begin{aligned}
&\frac{d}{d\mu}log\left[\frac{1}{\sigma_{prior}\sqrt{2\pi}}e^{-\frac{(\mu-\mu_{prior})^2}{2\sigma_{prior}^2}}\prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right] \\
=&\frac{d}{d\mu}\left[-log(\sigma_{prior}\sqrt{2\pi}) - \frac{(\mu-\mu_{prior})^2}{2\sigma_{prior}^2} + \sum_{i=1}^{N}\left(-\frac{(x_i-\mu)^2}{2\sigma^2} - log(\sigma\sqrt{2\pi})\right)\right] \\
=&-\frac{d}{d\mu}\left[\frac{(\mu-\mu_{prior})^2}{2\sigma_{prior}^2} + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\
=&-\left[\frac{\mu-\mu_{prior}}{\sigma_{prior}^2} + \sum_{i=1}^{N}\frac{\mu-x_i}{\sigma^2}\right]
\end{aligned}
$$

Setting this to zero, we get:-

$$
\begin{aligned}
\mu\left(\frac{1}{\sigma_{prior}^2} + \frac{N}{\sigma^2}\right) &= \left(\frac{\mu_{prior}}{\sigma_{prior}^2} + \frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right) \\
\implies \mu &= \frac{\left(\frac{\mu_{prior}}{\sigma_{prior}^2} + \frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_{prior}^2} + \frac{N}{\sigma^2}\right)} \\
&= \frac{\sigma^2\mu_{prior} + \sigma_{prior}^2\sum_{i=1}^{N}x_i}{\sigma^2 + N\sigma_{prior}^2} \\
\therefore\ \hat{\mu}^{\text{MAP}_1} &= \frac{\sigma^2\mu_{prior} + \sigma_{prior}^2\sum_{i=1}^{N}x_i}{\sigma^2 + N\sigma_{prior}^2}
\end{aligned}
$$

1.2.2 Case 2 : Uniform Prior

$$
P(\mu) = \begin{cases} 0 & x < 9.5 \\ \frac{1}{2} & 9.5 \leq x \leq 11.5 \\ 0 & 11.5 < x \end{cases}
$$

The aim is to maximize the log of the expression, but here the function is 0 $\forall\ x \in (-\infty, 9.5) \cup (11.5, \infty)$

So, we only consider the part of the function in the range $[9.5, 11.5]$. So,

$$\begin{aligned}
\frac{d}{d\mu}log\left(\frac{1}{2}\prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) &= \frac{d}{d\mu}\left[-log(2) - \sum_{i=1}^{N}\left(\frac{(x_i-\mu)^2}{2\sigma^2} + log(\sigma\sqrt{2\pi})\right)\right] \\
&= -\sum_{i=1}^{N}\frac{\mu - x_i}{\sigma^2}
\end{aligned}$$

Setting this to 0, we get:-

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

But we only considered the function in the range$[9.5, 11.5]$.

If the value of μ thus obtained is greater than 11.5, then this means that the function achieves its largest value at 11.5, since it is $0 \; \forall \; \mu > 11.5$

Likewise, if the value of μ is less than 9.5, then the function achieves its largest value at 9.5, since it is $0 \; \forall \; \mu < 9.5$

Therefore:-

$$\hat{\mu}^{\mathrm{MAP}_2} = \begin{cases} 9.5 & \frac{\sum_{i=1}^{N} x_i}{N} < 9.5 \\ \frac{\sum_{i=1}^{N} x_i}{N} & 9.5 \leq \frac{\sum_{i=1}^{N} x_i}{N} \leq 11.5 \\ 11.5 & 11.5 < \frac{\sum_{i=1}^{N} x_i}{N} \end{cases}$$

1.3 Box Plot

Plot for relative error of the three estimates from μ_{true}.

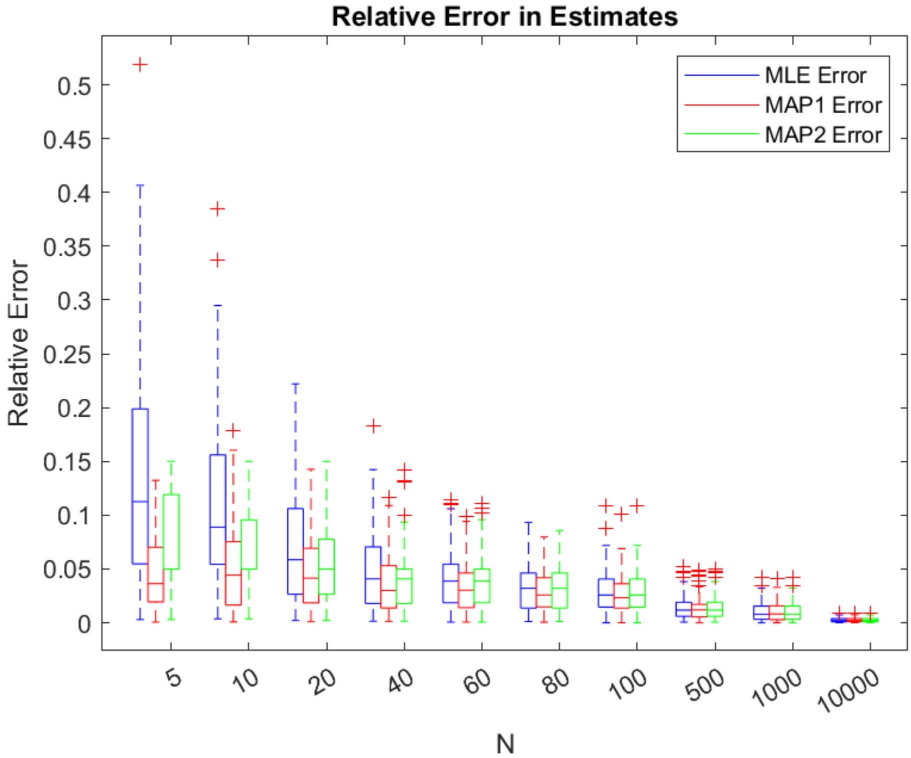

Figure 1: Relative Error from μ_{true}

1.4 Inferences

We can see that for all three estimates, as N increases, the relative error goes closer and closer to 0.

Along with this, we also see that as the value of N increases, the variance of the estimates also becomes closer and closer to 0, i.e. the values lie in a much smaller spread.

This is because all the estimates, as N goes to ∞ tend to $\frac{\sum_{i=1}^{N} x_i}{N}$, which in turn tends to μ_{true}.

$$\begin{aligned}
\lim_{N\to\infty} \hat{\mu}^{\mathrm{MAP}_1} &= \lim_{N\to\infty} \frac{\sigma^2 \mu_{prior} + \sigma_{prior}^2 \sum_{i=1}^{N} x_i}{\sigma^2 + N\sigma_{prior}^2} \\
&= \lim_{N\to\infty} \frac{\frac{\sigma^2 \mu_{prior}}{N} + \sigma_{prior}^2 \frac{\sum_{i=1}^{N} x_i}{N}}{\frac{\sigma^2}{N} + \sigma_{prior}^2} \\
&= \frac{\sum_{i=1}^{N} x_i}{N}
\end{aligned}$$

And $\hat{\mu}^{\mathrm{ML}}$ and $\hat{\mu}^{\mathrm{MAP}_2}$ are already equal to $\frac{\sum_{i=1}^{N} x_i}{N}$ ($\hat{\mu}^{\mathrm{MAP}_2}$ has some boundary conditions but that won't matter in the limit $N \to \infty$ as we shall see).

$$E\left[\frac{\sum_{i=1}^{N} x_i}{N}\right] = \frac{1}{N}\sum_{i=1}^{N} E[x_i]$$

Since x_i is drawn from a Gaussian Distribution with $\mu_{true} = 10$ and $\sigma = 4$, so $E[x_i] = \mu_{true}$,

$$E\left[\frac{\sum_{i=1}^{N} x_i}{N}\right] = \frac{1}{N} N\mu_{true}$$
$$= \mu_{true}$$

Hence, the mean of all estimates in the limit $N \rightarrow \infty$ becomes μ_{true}, and since $\mu_{true} \in [9.5, 11.5]$, the boundary cases in $\hat{\mu}^{\mathrm{MAP}_2}$ don't matter.

And by law of large numbers, as $N \rightarrow \infty$, $Var\left(\frac{\sum_{i=1}^{N} x_i}{N}\right) \rightarrow 0$

Hence, the variance or "spread" of all estimates goes to 0, with value $= \mu_{true}$ in the limit $N \rightarrow \infty$.
So, as N increases, the error should decrease and become closer and closer to 0, which is seen in the graph for all three estimates.

Of the given estimates, the Maximum A-Posteriori Estimate with the Gaussian Prior ($\hat{\mu}^{\mathrm{MAP}_1}$) gives the least average error along with quite a small spread for the initial values, compared to the other two estimates.

The next best estimator would be the Maximum A-Posteriori Estimate with the Uniform Prior($\hat{\mu}^{\mathrm{MAP}_2}$), which is not too different from the Maximum A-Posteriori Estimate with the Gaussian Prior($\hat{\mu}^{\mathrm{MAP}_1}$).

In last place comes the Maximum Likelihood Estimate($\hat{\mu}^{\mathrm{ML}}$), which has a much greater error than the other two for small values of N.

Hence, I would prefer the Maximum A-Posteriori Estimate with the Gaussian Prior($\hat{\mu}^{\mathrm{MAP}_1}$) to obtain an estimate of μ_{true}.

2 Question 2

2.1 Maximum Likelihood Estimate

Let us find the distribution of transformed data. Transformation is done as

$$y = f(x) = \frac{-1}{\lambda} log(x)$$
$$x = f^{-1}(y) = e^{-\lambda y}$$

Clearly, f is a monotonic function in the given domain of x.

$$\begin{aligned} P_Y(y) &= P_X(f^{-1}(y)) \cdot \left| \frac{df^{-1}(y)}{dy} \right| \\ &= \lambda e^{-\lambda y} \end{aligned}$$

for $y \geq 0$. So,

$$P_Y(y) = \begin{cases} 0 & y < 0 \\ \lambda e^{-\lambda y} & 0 \leq y \end{cases}$$

Let L represent log of likelihood function. Then

$$P(Y_1, Y_2...Y_N|\lambda) = \lambda^N e^{-\sum_{i=1}^{N} \lambda y_i}$$
$$L = \log(P(Y_1, Y_2...Y_N|\lambda)) = N \log(\lambda) - \lambda \sum_{i=1}^{N} y_i$$

To find the Maximum Likelihood Estimator, we equate derivative of L to zero

$$\frac{\partial L}{\partial \lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} y_i = 0$$
$$\implies \hat{\lambda}^{ML} = \frac{N}{\sum_{i=1}^{N} y_i}$$

2.2 Posterior Mean Estimate

Posterior is given by

$$\begin{aligned} P(\lambda|Y_1, Y_2...Y_N) &= P(Y_1, Y_2...Y_N|\lambda) \cdot \frac{P(\lambda)}{P(Y_1, Y_2...Y_N)} \\ &= \frac{P(Y_1, Y_2...Y_N|\lambda) \cdot P(\lambda)}{\int_\lambda P(Y_1, Y_2...Y_N|\lambda) \cdot P(\lambda) \cdot d\lambda} \end{aligned}$$

Since Prior for λ follows a Gamma distribution with shape parameter α and an inverse scale parameter β.

$$P(\lambda) = \frac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}$$

Posterior Mean is be given by

$$\begin{aligned} E_{P(\lambda|Y_1,Y_2...Y_N)}[\lambda] &= \int_\lambda \lambda \cdot P(\lambda|Y_1, Y_2...Y_N) \cdot d\lambda \\ &= \frac{\lambda \cdot P(Y_1, Y_2...Y_N|\lambda) \cdot P(\lambda)}{\int_\lambda P(Y_1, Y_2...Y_N|\lambda) \cdot P(\lambda) \cdot d\lambda} \\ &= \frac{\int_0^\infty \lambda^{N+\alpha} e^{-\beta\lambda} e^{-\sum_{i=1}^{N} \lambda y_i} d\lambda}{\int_0^\infty \lambda^{N+\alpha-1} e^{-\beta\lambda} e^{-\sum_{i=1}^{N} \lambda y_i} d\lambda} \end{aligned}$$

We know that Gamma function is given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \cdot dx$$
$$\implies \frac{\Gamma(\alpha)}{\beta^\alpha} = \int_0^\infty x^{\alpha-1} e^{-\beta x} \cdot dx$$

Let $y = \sum_{i=1}^N y_i$. Now

$$\begin{aligned} E_{P(\lambda|Y_1,Y_2\ldots Y_N)}[\lambda] &= \frac{\int_0^\infty \lambda^{N+\alpha} e^{-\lambda(\beta+y)} d\lambda}{\int_0^\infty \lambda^{N+\alpha-1} e^{-\lambda(\beta+y)} d\lambda} \\ &= \frac{\Gamma(N+\alpha+1)/(\beta+y)^{N+\alpha+1}}{\Gamma(N+\alpha)/(\beta+y)^{N+\alpha}} \\ &= \frac{N+\alpha}{\beta+y} \\ \implies \hat{\lambda}^{\text{PosteriorMean}} &= \frac{N+\alpha}{\beta+y} \end{aligned}$$

where $y = \sum_{i=1}^N y_i$.

2.3 Box Plot

Plot for relative error of both estimates from λ_{true}.

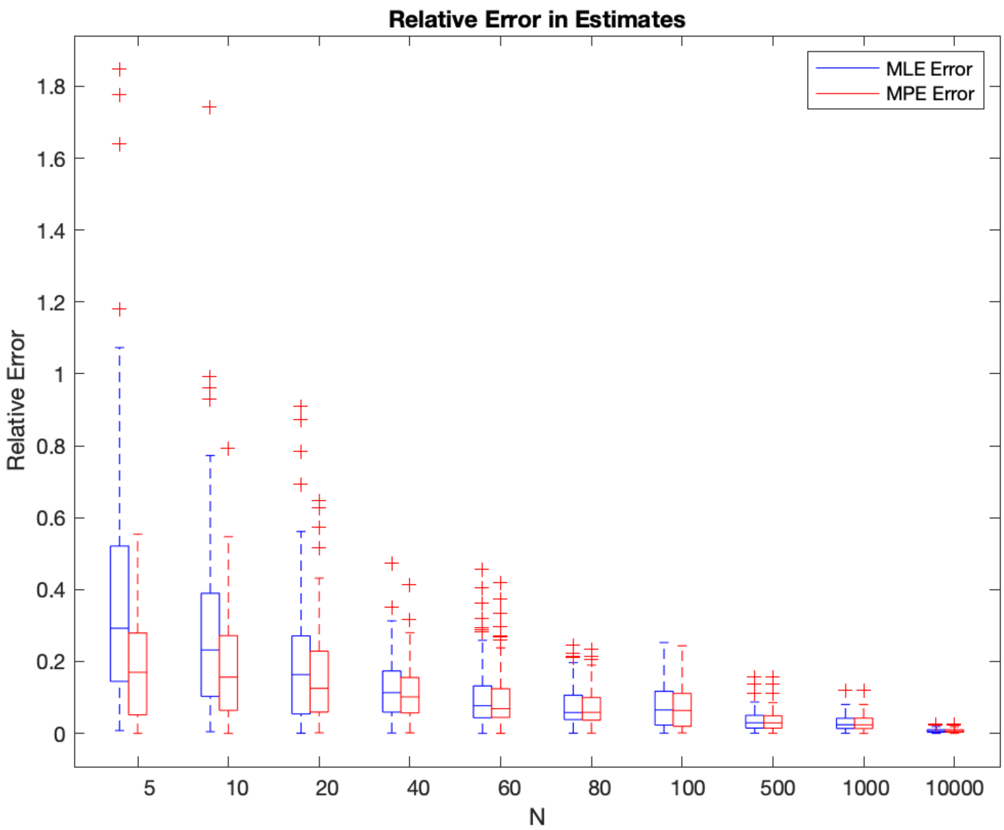

Figure 2: Relative Error from λ_{true}

2.4 Inferences

The relative error gradually decreases as N increases. For large value of N, both the estimators converge to true parameter.
It is clear from the box plots that Posterior Mean Estimator is much better than Maximum Likelihood Estimator for smaller values of N. As the value of N increases, the difference between estimators diminishes.

Here's our Posterior Mean Estimator

$$\hat{\lambda}^{\text{PosteriorMean}} = \frac{N+\alpha}{\beta+y} = \frac{1+\frac{\alpha}{N}}{\frac{y}{N}+\frac{\beta}{N}}$$

As N tends to ∞

$$\lim_{N\to\infty} \hat{\lambda}^{\text{PosteriorMean}} = \frac{N}{y} = \hat{\lambda}^{ML}$$

where $y = \sum_{i=1}^{N} y_i$

Thus we can see that for large value of N, both the estimators converge to same value.

3 Question 3

3.1 Maximum Likelihood Estimate

$$\hat{\theta}^{\text{ML}} = arg\ max_{\theta} P(x_1, x_2, \ldots, x_N|\theta)$$

Now,

$$P(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \theta < x \end{cases}$$

Clearly, $P(x_1, x_2, \ldots, x_N|\theta) = 0 \ \forall \ \theta < max(x_1, x_2, \ldots, x_N)$, so we shall only consider θ such that $\theta >= max(x_1, x_2, \ldots, x_N)$.
In this case:-

$$\begin{aligned} \hat{\theta}^{\text{ML}} &= arg\ max\ _{\theta} \prod_{i=1}^{N} \frac{1}{\theta} \\ &= arg\ max\ _{\theta}\ \frac{1}{\theta^N} \end{aligned}$$

To maximize $\frac{1}{\theta^N}$, we must minimize θ under the initial constraints of $\theta >= max(x_1, x_2, \ldots, x_N)$.

Clearly, this is when $\theta = max(x_1, x_2, \ldots, x_N)$. Therefore:-

$$\hat{\theta}^{\text{ML}} = max(x_1, x_2, \ldots, x_N)$$

3.2 Maximum A-Posteriori Estimate

$$\begin{aligned} \hat{\theta}^{\text{MAP}} &= arg\ max\ _{\theta} P(\theta|x_1, x_2, \ldots, x_N) \\ &= arg\ max\ _{\theta} \frac{P(x_1, x_2, \ldots, x_N|\theta)P(\theta)}{\int_{\theta} P(x_1, x_2, \ldots, x_N, \theta)d\theta} \end{aligned}$$

Since the denominator is not a function of μ, it will have no effect on $\hat{\mu}^{\text{MAP}_1}$, hence it can be neglected. So our expression now becomes:-

$$\hat{\theta}^{\text{MAP}} = arg\ max\ _{\theta} \left[P(x_1, x_2, \ldots, x_N|\theta)P(\theta)\right]$$

As $P(\theta) = 0 \ \forall \ \theta < \theta_m$ and (as done in Maximum Likelihood Estimate) $P(x_1, x_2, \ldots, x_N|\theta) = 0 \ \forall \ \theta < max(x_1, x_2, \ldots, x_N)$, we only consider θ such that $\theta \geq max(x_1, x_2, \ldots, x_N, \theta_m)$.

$$\begin{aligned} \hat{\theta}^{\text{MAP}} &= arg\ max\ _{\theta} \left[k \left(\frac{\theta_m}{\theta}\right)^{\alpha} \prod_{i=1}^{N} \frac{1}{\theta}\right] \\ &= arg\ max\ _{\theta}\ \frac{k\theta_m^{\alpha}}{\theta^{(N+\alpha)}} \end{aligned}$$

where k is the proportionality constant for $P(\theta)$.

Since k and θ_m are constants, to maximize the expression we must minimize θ under the given constraints. Clearly this is when $\theta = max(x_1, x_2, \ldots, x_N, \theta_m)$.

Therefore:-

$$\hat{\theta}^{\text{MAP}} = max(x_1, x_2, \ldots, x_N, \theta_m)$$

3.3 Posterior Distribution

The Posterior Probability Distribution is defined as $P(\Theta|x_1, x_2, \ldots, x_n)$.

$$
\begin{aligned}
P(\theta|x_1, x_2, \ldots, x_n) &= \frac{P(x_1, x_2, \ldots, x_N|\theta)P(\theta)}{\int_\theta P(x_1, x_2, \ldots, x_N, \theta)d\theta} \\
&= \begin{cases} \frac{k\theta_m^\alpha}{\theta^{N+\alpha}} & max(x_1, x_2, \ldots, x_n, \theta_m) \leq \theta \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

where k is a constant, since the denominator and the proportionality constant of $P(\theta)$ do not depend on θ. Now, we know that:-

$$\int_\theta P(\theta|x_1, x_2, \ldots, x_n)d\theta = 1$$

Let $max(x_1, x_2, \ldots, x_n, \theta_m) = \phi$, so:-

$$
\begin{aligned}
\int_{\theta=\phi}^{\infty} \frac{k\theta_m^\alpha}{\theta^{N+\alpha}} &= 1 \\
\implies \frac{k\theta_m^\alpha}{(N+\alpha-1)\phi^{N+\alpha-1}} d\theta &= 1
\end{aligned}
$$

This means that:-

$$k = \frac{(N+\alpha-1)\phi^{N+\alpha-1}}{\theta_m^\alpha}$$

Substituting this value in the expression for $P(\theta|x_1, x_2, \ldots, x_n)$, we get:-

$$P(\theta|x_1, x_2, \ldots, x_n) = \begin{cases} \frac{(N+\alpha-1)\phi^{N+\alpha-1}}{\theta^{N+\alpha}} & \phi \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\phi = max(x_1, x_2, \ldots, x_n, \theta_m)$

Now we have to find posterior mean estimate, $\hat{\theta}^{\text{PosteriorMean}} = E_{P(\Theta|x_1,x_2,\ldots,x_n)}[\Theta]$

$$
\begin{aligned}
E_{P(\Theta|x_1,x_2,\ldots,x_n)}[\Theta] &= \int_{\theta=\phi}^{\infty} \theta \frac{(N+\alpha-1)\phi^{N+\alpha-1}}{\theta^{N+\alpha}} d\theta \\
&= \int_{\theta=\phi}^{\infty} \frac{(N+\alpha-1)\phi^{N+\alpha-1}}{\theta^{N+\alpha-1}} d\theta \\
&= \frac{(N+\alpha-1)\phi}{N+\alpha-2}
\end{aligned}
$$

Hence, we have:-

$$\hat{\theta}^{\text{PosteriorMean}} = \frac{(N+\alpha-1)\phi}{N+\alpha-2}$$

where $\phi = max(x_1, x_2, \ldots, x_n, \theta_m)$

3.4 Inferences as $N \to \infty$

Since $x_1, x_2, \ldots, x_N$ are realised from uniform distribution spread over 0 to θ_{true}, we can say as $N \to \infty$

$$\lim_{N\to\infty} max(x_1, x_2, \ldots, x_N) = \theta_{true}$$

Thus for large N, the three estimates converge to following values

$$\begin{aligned}
\lim_{N\to\infty} \hat{\theta}^{\mathrm{ML}} &= \theta_{true} \\
\lim_{N\to\infty} \hat{\theta}^{\mathrm{MAP}} &= max(\theta_{true}, \theta_m) \\
\lim_{N\to\infty} \hat{\theta}^{\mathrm{PosteriorMean}} &= \lim_{N\to\infty} \frac{(N+\alpha-1)\phi}{N+\alpha-2} \\
&= \lim_{N\to\infty} \phi = max(\theta_{true}, \theta_m)
\end{aligned}$$

Now consider the two cases

3.4.1 Case 1 : $\theta_m \leq \theta_{true}$

$$\begin{aligned}
\lim_{N\to\infty} \hat{\theta}^{\mathrm{ML}} &= \theta_{true} \\
\lim_{N\to\infty} \hat{\theta}^{\mathrm{MAP}} &= \theta_{true} \\
\lim_{N\to\infty} \hat{\theta}^{\mathrm{PosteriorMean}} &= \theta_{true}
\end{aligned}$$

That is both $\hat{\theta}^{\mathrm{MAP}}$ and $\hat{\theta}^{\mathrm{PosteriorMean}}$ converge to $\hat{\theta}^{\mathrm{ML}}$ for large values of N.

3.4.2 Case 2 : $\theta_m > \theta_{true}$

$$\begin{aligned}
\lim_{N\to\infty} \hat{\theta}^{\mathrm{ML}} &= \theta_{true} \\
\lim_{N\to\infty} \hat{\theta}^{\mathrm{MAP}} &= \theta_m \\
\lim_{N\to\infty} \hat{\theta}^{\mathrm{PosteriorMean}} &= \theta_m
\end{aligned}$$

That is $\hat{\theta}^{\mathrm{MAP}}$ and $\hat{\theta}^{\mathrm{PosteriorMean}}$ do not converge to $\hat{\theta}^{\mathrm{ML}}$ for large values of N.

Thus we infer that $\hat{\theta}^{\mathrm{MAP}}$ and $\hat{\theta}^{\mathrm{PosteriorMean}}$ may and may not converge to $\hat{\theta}^{\mathrm{ML}}$ depending upon the prior. Convergence to $\hat{\theta}^{\mathrm{ML}}$ is desirable for both of them, since $\hat{\theta}^{\mathrm{ML}}$ itself converges to θ_{true}.