CS215 Assignment 3
Bayesian Estimation

Josyula Venkata Aditya
210050075

Kartik Sreekumar Nair
210050083

November 1, 2022

Contents

Preface

This assignment is done in a group. Group members:

1. **Josyula Venkata Aditya - 210050075**
2. **Kartik Sreekumar Nair - 210050083**

The submission folder contains four folders(in addition to this file and the problem statement file):

code/problem 1 - contains 1 *MATLAB*(`.m`) file, corresponding to problem 1.

code/problem 2 - contains 1 *MATLAB*(`.m`) file, corresponding to problem 2.

results/fig - contains 2 folders corresponding to the problems; each folder contains the plot corresponding to that problem. This is the main folder of plots, please perform the evaluations of plots using this folder.

results/eps - contains all the plots, but in `.eps` format, used in the report. This format may not be readable from your device, hence for evaluation purposes, please check the plots in `results/fig` folder.

report - contains the reports of the 3 problems.

We have used f to denote PDFs throughout this report.

Problem 1

Use the Matlab function randn() to generate a data sample of N points drawn from a Gaussian distribution with mean $\mu_{\text{true}} = 10$ and standard deviation $\sigma_{\text{true}} = 4$. Consider the problem of using the data σ_{prior} to get an estimate $\hat{\mu}$ of this Gaussian mean, assuming it is unknown, when the standard deviation σ_{true} is known. Consider using one of the two prior prior distributions on the mean:

(i) a Gaussian prior with mean $\mu_{\text{prior}} = 10.5$ and standard deviation $\sigma_{\text{prior}} = 1$

$$f_1(\mu) = \frac{1}{\sqrt{2\pi}\sigma_{\text{prior}}} e^{-\frac{(\mu-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}} \tag{1.1}$$

(ii) a uniform prior over $[9.5, 11.5]$.

$$f_2(\mu) = \begin{cases} \frac{1}{2} & 9.5 \leq \mu \leq 11.5 \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

Consider various sample sizes $N = 5, 10, 20, 40, 60, 80, 100, 500, 10^3, 10^4$. For each sample size N, repeat the following experiment $M \geq 100$ times: generate the data, get the maximum likelihood estimate $\hat{\mu}^{\text{ML}}$, get the maximum-a-posteriori estimates $\hat{\mu}_1^{\text{MAP}}$ and $\hat{\mu}_2^{\text{MAP}}$, and measure the relative errors $\frac{|\hat{\mu}-\mu_{\text{true}}|}{\mu_{\text{true}}}$ for all three estimates.

Ans: We are given the pdf of the data given its mean μ,

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma_{\text{true}}} e^{-\frac{(x-\mu)^2}{2\sigma^2_{\text{true}}}} \tag{1.3}$$

For the entire data of N sample points, we get the Likelihood L:

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{N} f_X(x_i|\mu) && (1.4) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_{\text{true}}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2_{\text{true}}}} && (1.5) \end{aligned}$$

$$\begin{aligned} \log L(\mu) &= \sum_{i=1}^{N} \left(\log\left(\frac{1}{\sqrt{2\pi}\sigma_{\text{true}}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2_{\text{true}}} \right) && (1.6) \\ &= c_0 - \frac{1}{2\sigma^2_{\text{true}}} \sum_{i=1}^{N} (x_i - \mu)^2 && (1.7) \\ &= c_0 - \frac{N}{2\sigma^2_{\text{true}}} \left(\mu^2 - 2\frac{\sum x_i}{N}\mu + \frac{\sum x_i^2}{N} \right) && (1.8) \\ &= c_1 - \frac{N}{2\sigma^2_{\text{true}}} (\mu - \frac{\sum x_i}{N})^2 && (1.9) \end{aligned}$$

where c_0 and c_1 are expressions not containing μ.
The ML estimate $\hat{\mu}^{\text{ML}}$ is

$$\hat{\mu}^{\text{ML}} = \arg\max_{\mu} L(\mu) = \arg\max_{\mu} \log L(\mu) \tag{1.10}$$

$$= \arg\min_{\mu} \frac{N}{2\sigma^2_{\text{true}}}(\mu - \frac{\sum x_i}{N})^2 \tag{1.11}$$

$$= \frac{\sum x_i}{N} = \overline{x}, \tag{1.12}$$

the sample mean.

We indicate by $\mathbf{x}$ the sequence of sample points $(x_1, x_2, \cdots x_N)$. The posterior is

$$\mathcal{P} = f(\mu|\mathbf{x}) = \frac{f(\mathbf{x}|\mu)f(\mu)}{f(\mathbf{x})} \tag{1.13}$$

$$= \frac{1}{f(\mathbf{x})} L(\mu) f(\mu) \tag{1.14}$$

$$\log \mathcal{P} = c + \log L(\mu) + \log f(\mu) \tag{1.15}$$

where c is an expression that doesn't depend on μ.

We find the MAP estimates for both the priors. f_1 corresponds to the first prior(gaussian) and f_2 corresponds to the second prior(uniform).

(i)

$$\hat{\mu}_1^{\text{MAP}} = \arg\max_{\mu} \mathcal{P} = \arg\max_{\mu} \log \mathcal{P} \tag{1.16}$$

$$= \arg\max_{\mu} (c + \log L(\mu) + \log f_1(\mu)) \tag{1.17}$$

$$= \arg\max_{\mu} \left(-\frac{N}{2\sigma^2_{\text{true}}}(\mu - \overline{x})^2 - \frac{(\mu - \mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}} \right) \tag{1.18}$$

$$= \arg\min_{\mu} \left(\frac{N}{2\sigma^2_{\text{true}}}(\mu - \overline{x})^2 + \frac{(\mu - \mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}} \right) \tag{1.19}$$

We now differentiate the objective function to find an extreme point.

$$\frac{\partial}{\partial \mu} \left[\frac{N}{2\sigma^2_{\text{true}}}(\mu - \overline{x})^2 + \frac{(\mu - \mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}} \right]_{\mu=\mu_0} = \left. \frac{N}{\sigma^2_{\text{true}}}(\mu - \overline{x}) + \frac{(\mu - \mu_{\text{prior}})}{\sigma^2_{\text{prior}}} \right|_{\mu=\mu_0} = 0 \tag{1.20}$$

Observe that the objective function is a quadratic expression in μ with a positive leading coefficient, hence, the extreme point will be a minima and the value of μ at that point will be the $\arg\min$

$$\mu_0 = \hat{\mu}_1^{\text{MAP}} = \frac{\sigma^2_{\text{true}}\mu_{\text{prior}} + N\sigma^2_{\text{prior}}\overline{x}}{\sigma^2_{\text{true}} + N\sigma^2_{\text{prior}}} \tag{1.21}$$

(ii)

$$\hat{\mu}_2^{\text{MAP}} = \arg\max_{\mu} \mathcal{P} \tag{1.22}$$

$$= \arg\max_{\mu \in [9.5, 11.5]} \mathcal{P}, \tag{1.23}$$

as $\mathcal{P}(\mu_1) > 0 = \mathcal{P}(\mu_2)$ when $\mu_1 \in [9.5, 11.5]$, $\mu_2 \notin [9.5, 11.5]$.

$$\therefore \hat{\mu}_2^{\text{MAP}} = \underset{\mu \in [9.5,11.5]}{\arg\max}\ c + \log L(\mu) + \log f_2(\mu) \tag{1.24}$$

$$= \underset{\mu \in [9.5,11.5]}{\arg\max}\ \log L(\mu) \tag{1.25}$$

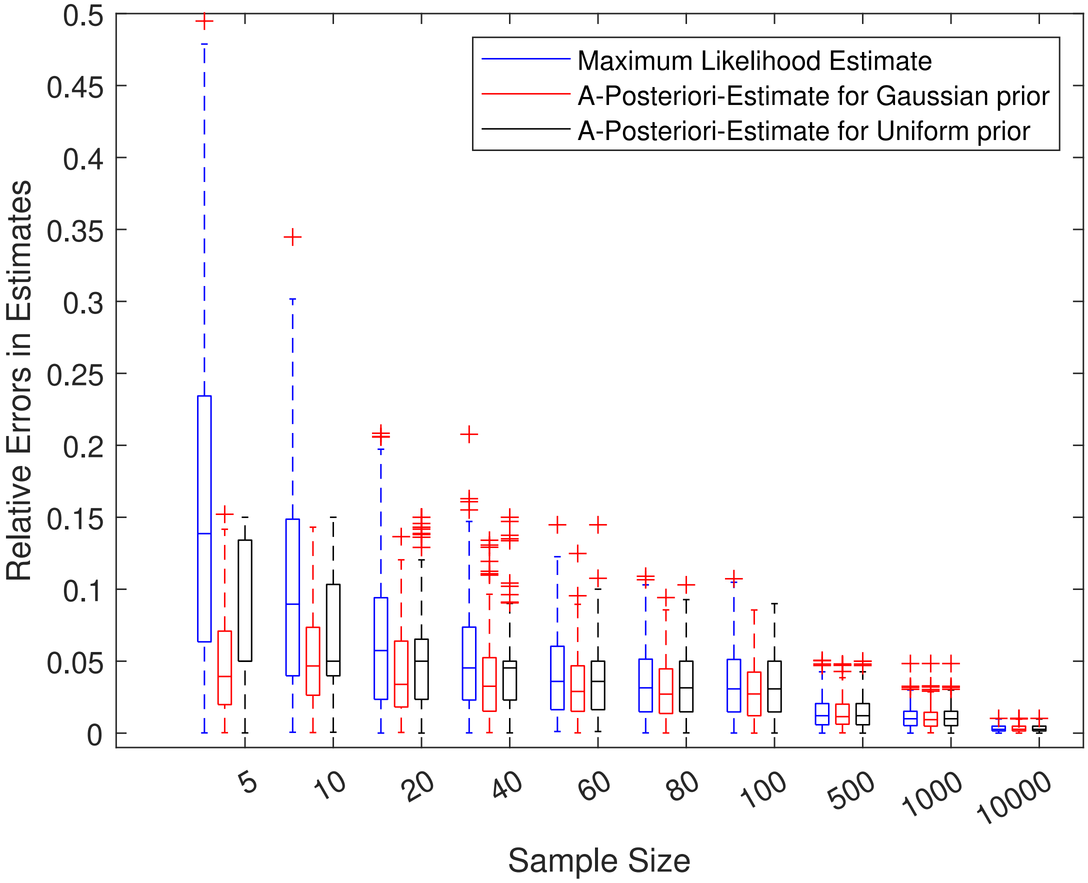
$$= \underset{\mu \in [9.5,11.5]}{\arg\min}\ (\mu - \overline{x})^2 \tag{1.26}$$

$$\hat{\mu}_2^{\text{MAP}} = \begin{cases} 9.5 & \overline{x} < 9.5 \\ \overline{x} & 9.5 \leq \overline{x} \leq 11.5 \\ 11.5 & 11.5 < \overline{x} \end{cases} \tag{1.27}$$

This is equivalent to our ML estimate *clamped* between the bounds of our uniform prior.

(a) Plot a single graph that shows the relative errors for each value of N as a box plot (use the Matlab boxplot() function), for each of the three estimates.

Ans:



(b) Interpret what you see in the graph.

(i) What happens to the error as N increases ?
Ans: We can clearly see from the boxplot that the distribution of relative error decreases to 0 as N increases for all the 3 estimators.

(ii) Which of the three estimates will you prefer and why ?

Ans: We would be picking the MAP estimator with the Gaussian prior. We know that the MLE might not be a good enough estimate to the true value when the sample size is small, it is also evident from the above boxplot. Bayesian estimates can help in such cases if we have a good prior and the error converges to 0 as $N \to \infty$.

We can observe from the boxplot that both the size of the errors and the distribution of errors is narrower for the MAP estimator with Gaussian prior than the other 2 estimators. Hence, we would be choosing this estimator over the other 2.

We make an interesting observation here about the uniform prior. Let the true value of μ be μ_0 and the lower and upper bounds of the uniform prior be l and u respectively. We observe that the sample mean, $\overline{x}$, tends to μ_0 as $N \to \infty$. If the uniform prior was chosen such that $\mu_0 < l$ or $\mu_0 > u$, then, $\hat{\mu_2}^{\text{MAP}}$ would tend to 9.5 or 11.5(according as $\mu_0 < l$ or $\mu_0 > u$). Hence, in this case, our prior was "good" because $\mu_0 = 10$ lies in between $l = 9.5$ and $u = 11.5$ such that $l < \mu_0 < u$ and hence, $\hat{\mu_2}^{\text{MAP}}$ gives the same result as the sample mean asymptotically and hence, the relative error tends to 0.

Problem 2

Use the Matlab function rand() to generate a data sample of N points from the uniform distribution on $[0, 1]$. Transform the resulting data x to generate a transformed data sample where each datum $y := -\frac{1}{\lambda}\log(x)$ with $\lambda = 5$. The transformed data y will have some distribution with parameter λ; what is its analytical form ? Use a Gamma prior on the parameter λ, where the Gamma distribution has parameters $\alpha = 5.5$ and $\beta = 1$.

$$f(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \tag{2.1}$$

Consider various sample sizes $N = 5, 10, 20, 40, 60, 80, 100, 500, 10^3, 10^4$. For each sample size N, repeat the following experiment $M \geq 100$ times: generate the data, get the maximum likelihood estimate $\hat{\lambda}^{\mathrm{ML}}$, get the Bayesian estimate as the posterior mean $\hat{\lambda}^{\mathrm{PosteriorMean}}$, and measure the relative errors $\frac{|\hat{\lambda}-\lambda_{\mathrm{true}}|}{\lambda_{\mathrm{true}}}$ for both the estimates.

Ans:

$$X \sim \text{Uniform}[0, 1] \tag{2.2}$$

We take

$$Y = g(X) \tag{2.3}$$

where

$$g(x) = -\frac{1}{\lambda}\log x, \tag{2.4}$$

and its inverse

$$g^{-1}(y) = e^{-\lambda y}. \tag{2.5}$$

Using transformation of random variables,

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{dg^{-1}(y)}{dy}\right| \tag{2.6}$$

Given $\lambda > 0$

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y} & 0 \leq e^{-\lambda y} \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

$$= \begin{cases} \lambda e^{-\lambda y} & 0 \leq y \leq \infty \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

This is an exponential distribution with parameter λ.

We look at the likelihood given a prior on λ, given N sample points. We denote the sequence of sample points by $\mathbf{y} = (y_1, y_2, \cdots, y_N)$

$$L(\lambda) = f(\mathbf{y}|\lambda) = \prod_{i=1}^{N} f_Y(y_i|\lambda) \tag{2.9}$$

$$\log L(\lambda) = \sum_{i=1}^{N} \log f_Y(y_i|\lambda) \tag{2.10}$$

$$= \sum_{i=1}^{N} (\log(\lambda) - \lambda y_i) \tag{2.11}$$

$$= N\log(\lambda) - \lambda \sum_{i=1}^{N} y_i, \tag{2.12}$$

given $y_i > 0 \ \forall 1 \leq i \leq N$.

$$\hat{\lambda}^{\text{ML}} = \arg\max_{\lambda} \log L \tag{2.13}$$

We would now like to find the ML estimate which is obtained by maximizng the likelihood function. This can be done by differentiating the likelihood function and setting its value to 0.

$$\frac{\partial}{\partial\lambda}\left[\log L\right]_{\lambda=\hat{\lambda}^{\text{ML}}} = \left.\frac{N}{\lambda} - \sum_{i=1}^{N} y_i\right|_{\lambda=\hat{\lambda}^{\text{ML}}} = 0. \tag{2.14}$$

$$\hat{\lambda}^{\text{ML}} = \frac{N}{\sum_{i=1}^{N} y_i} = \frac{1}{\overline{y}} \tag{2.15}$$

(a) Derive a formula for the posterior mean

Ans: The posterior mean is

$$\hat{\lambda}^{\text{PosteriorMean}} = E_{f(\lambda|\mathbf{y})}[\lambda] \tag{2.16}$$

$$\hat{\lambda}^{\text{PosteriorMean}} = \int f(\lambda|\mathbf{y})\lambda d\lambda \tag{2.17}$$

$$= \int_0^\infty \frac{f(\mathbf{y}|\lambda)f(\lambda)}{f(\mathbf{y})}\lambda d\lambda \tag{2.18}$$

$$= \frac{\int_0^\infty f(\mathbf{y}|\lambda)f(\lambda)\lambda d\lambda}{\int_0^\infty f(\mathbf{y}|\lambda)f(\lambda)d\lambda} \tag{2.19}$$

$$f(\mathbf{y}|\lambda)f(\lambda) = \lambda^N e^{-\lambda N\overline{y}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \tag{2.20}$$

$$= k\,\lambda^{\alpha+N-1}e^{-\lambda(N\overline{y}+\beta)} \tag{2.21}$$

We evaluate, using integration by parts,

$$\int_0^\infty f(\mathbf{y}|\lambda)f(\lambda)\lambda d\lambda = \int_0^\infty k\,\lambda^{\alpha+N}e^{-\lambda(N\overline{y}+\beta)} \tag{2.22}$$

$$= \left[-k\,\frac{\lambda^{\alpha+N}e^{-(N\overline{y}+\beta)\lambda}}{(N\overline{y}+\beta)}\right]_0^\infty + \frac{\alpha+1}{(N\overline{y}+\beta)}\int_0^\infty k\,\lambda^{\alpha+N-1}e^{-\lambda(N\overline{y}+\beta)}d\lambda \tag{2.23}$$

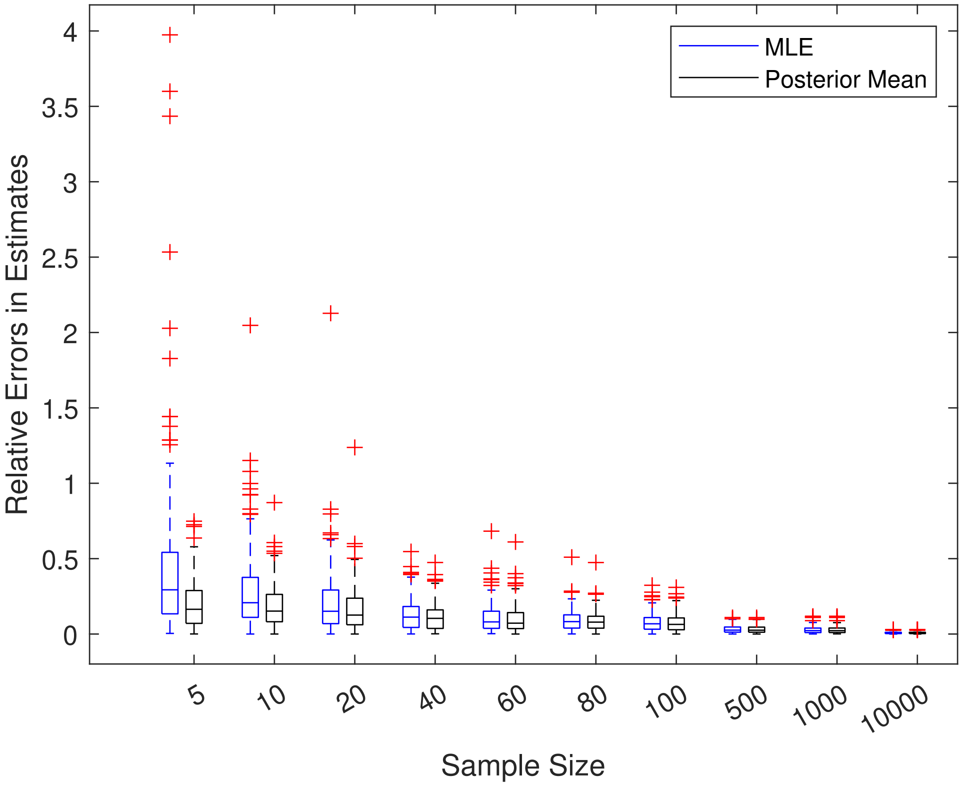
$$= \frac{\alpha+N}{\beta+N\overline{y}}\int_0^\infty f(\mathbf{y}|\lambda)f(\lambda)d\lambda \tag{2.24}$$

Therefore

$$\hat{\lambda}^{\text{PosteriorMean}} = \frac{\alpha+N}{\beta+N\overline{y}} \tag{2.25}$$

(b) Plot a single graph that shows the relative errors for each value of N as a box plot (use the Matlab boxplot() function), for both the estimates.

Ans:



(c) Interpret what you see in the graph.

(i) What happens to the error as N increases ?
Ans: We can clearly see from the boxplot that the distribution of relative error decreases as N increases until it converges to 0 for both the estimators.

(ii) Which of the two estimates will you prefer and why ?
Ans: We can observe that both the estimators are *consistent*. ML estimators might not be good estimators when the sample sizes are small, as is evident from the graph - they have a larger distribution of errors. In turn, Bayesian estimates can be quite helpful in such situations if our prior and loss functions are well chosen.

It is apparent from the graph that for the *Posterior Mean* estimator, the distributions of errors are narrower and the sizes of the errors are lesser compared to the ML estimator. Hence, we would prefer the *Posterior Mean* estimator.

Problem 3

Suppose random variable X has a uniform distribution over $[0, \theta]$, where the parameter θ is unknown.

$$f_X(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{o.w.} \end{cases} \tag{3.1}$$

Consider a Pareto distribution prior on θ, with a scale parameter $\theta_m > 0$ and a shape parameter $\alpha > 1$, as

$$f(\theta) = \begin{cases} c(\theta_m/\theta)^\alpha & \theta \geq \theta_m \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

(a) Find the maximum-likelihood estimate $\hat{\theta}^{\text{ML}}$ and the maximum-a-posteriori estimate $\hat{\theta}^{\text{MAP}}$.

Ans: The likelihood, given N i.i.d sample points denoted by $\mathbf{x} = (x_1, x_2, \cdots, x_N)$

$$L(\theta) = f(\mathbf{x}|\theta) = \prod_{i=1}^{N} f_X(x_i|\theta) \tag{3.3}$$

$$= \begin{cases} \theta^{-N} & 0 \leq x_i \leq \theta \ \ \forall 1 \leq i \leq N \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

$$\hat{\theta}^{\text{ML}} = \arg\max_{\theta} L \tag{3.5}$$

We would like to maximize L, Hence, we would require that $0 \leq x_i \leq \theta \ \forall 1 \leq i \leq N$, L would be 0 otherwise

$$\hat{\theta}^{\text{ML}} = \arg\max_{\theta:\theta \geq \max_i x_i} L \tag{3.6}$$

$$= \arg\max_{\theta:\theta \geq \max_i x_i} \theta^{-N} \tag{3.7}$$

as $\theta > 0$ and θ^{-N} is strictly decreasing.

$$\therefore \hat{\theta}^{\text{ML}} = \max_i x_i \tag{3.8}$$

For MAP

The Posterior,

$$\mathcal{P} = f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})} \tag{3.9}$$

$$= k' L(\theta) f(\theta) \tag{3.10}$$

$$= \begin{cases} k\theta^{-\alpha-N} & \theta \geq x_i \ \ \forall 1 \leq i \leq N, \theta_m \\ 0 & \text{otherwise} \end{cases} \tag{3.11}$$

The Maximum a-Posteriori estimate is given by,

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \mathcal{P} = \underset{\theta:\theta\geq x_i,\theta_m}{\arg\max}\, \theta^{-\alpha-N} \tag{3.12}$$

$$= \max(\theta_m, \max_i x_i) = \max(\theta_m, \hat{\theta}^{\text{ML}}), \tag{3.13}$$

using similar arguments as before.

(b) Does $\hat{\theta}^{\text{MAP}}$ tend to $\hat{\theta}^{\text{ML}}$ as the sample size tends to infinity ? Is this desirable or not ?

Ans: We have $\hat{\theta}^{\text{ML}} = \max_i x_i$ and $\hat{\theta}^{\text{MAP}} = \max(\theta_m, \max_i x_i)$. Let the original distribution from which data was sampled have the true parameter θ_0. We know that ML estimators are consistent. Hence,

$$\hat{\theta}^{\text{ML}} \to \theta_0 \text{ as } N \to \infty \tag{3.14}$$

Now, we shall see that this need not be the case with $\hat{\theta}^{\text{MAP}}$.

First we observe that $\max_i x_i \leq \theta_0$. We consider the following possibilities for θ_m

a. $\theta_m > \theta_0 \geq \max_i x_i \implies \max(\theta_m, \max_i x_i) = \theta_m$.
   So, for such choices of the prior, we have $\hat{\theta}^{\text{ML}} \to \theta_0$ and $\hat{\theta}^{\text{MAP}} = \theta_m$ as $N \to \infty$

b. $\theta_0 \geq \theta_m \geq \max_i x_i$. $\max_i x_i$ is the same as the ML estimate, hence, we can say that as $N \to \infty$, $\max_i x_i \to \theta_0$. Now, $\max(\theta_m, \max_i x_i)$ as $N \to \infty = \max(\theta_m, \theta_0) = \theta_0$.
   For such choices of the prior, $\hat{\theta}^{\text{ML}} = \hat{\theta}^{\text{MAP}}$ as $N \to \infty$.

So overall, if we have a *good* informative prior(i.e. $\theta_m \leq \theta_0$), then, the MAP estimate would be a better choice than the ML estimate as it would be better than ML estimate for smaller datasets and tends to the true value for larger datasets.

The answer to the question *Is this desirable* depends really on the quality of the prior chosen. If we are sure that our prior is *good*, then we can go ahead with the MAP estimator. However if we are only dealing with large datasets, then we can safely use the ML estimator.

(c) Find an estimator of the mean of the posterior distribution $\hat{\theta}^{\text{PosteriorMean}}$.

Ans:

$$\hat{\theta}^{\text{PosteriorMean}} = E_{f(\theta|\mathbf{x})}[\theta] \tag{3.15}$$

$$= \int \mathcal{P}\theta d\theta \tag{3.16}$$

$$= \frac{\int f(\mathbf{x}|\theta)f(\theta)\theta d\theta}{\int f(\mathbf{x}|\theta)f(\theta)d\theta} \tag{3.17}$$

$$= \frac{\int_{\max(\theta_m,\max_i x_i)}^{\infty} k\theta^{-\alpha-N}\theta d\theta}{\int_{\max(\theta_m,\max_i x_i)}^{\infty} k\theta^{-\alpha-N} d\theta} \tag{3.18}$$

$$= \frac{\alpha+N-1}{\alpha+N-2}\max(\theta_m, \max_i x_i) \tag{3.19}$$

(d) Does $\hat{\theta}^{\text{PosteriorMean}}$ tend to $\hat{\theta}^{\text{ML}}$ as the sample size tends to infinity ? Is this desirable or not ?

Ans: We see that

$$\lim_{N\to\infty} \hat{\theta}^{\text{PosteriorMean}} = \lim_{N\to\infty}\left(\frac{\alpha+N-1}{\alpha+N-2}\right)\lim_{N\to\infty}\left(\max\left(\theta_m, \max_i x_i\right)\right) \tag{3.20}$$

$$= 1 \cdot \lim_{N\to\infty}\left(\max\left(\theta_m, \max_i x_i\right)\right) \tag{3.21}$$

$$= \lim_{N\to\infty} \hat{\theta}^{\text{MAP}} \tag{3.22}$$

The asymptotic behaviour of $\hat{\theta}^{\text{PosteriorMean}}$ is same as the asymptotic behaviour of $\hat{\theta}^{\text{MAP}}$. Hence, the rest of the reasoning is same as (b) part and is being omitted here.