Statistical Computing for Scientists and Engineers

Homework 2

Jiale Shi

September/21/2018

1 Problem 1

(a) Obtain analytic forms of: The posterior distribution of Eq. (3) and the marginal posterior distribution over α and β:$p(\alpha, \beta|y)$ by using Eq. (4), Eq.(5) and the hint provided.

Answer:

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta) \\ &\propto (\alpha+\beta)^{-5/2} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha+y_j-1} (a-\theta_j)^{\beta+n_j-y_j-1} \end{aligned} \tag{1}$$

$$\begin{aligned} p(\alpha, \beta|y) &= \frac{p(\theta, \alpha, \beta|y)}{p(\theta|\alpha, \beta, y)} \\ &\propto \frac{(\alpha+\beta)^{-5/2} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha+y_j-1} (a-\theta_j)^{\beta+n_j-y_j-1}}{\prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1} (a-\theta_j)^{\beta+n_j-y_j-1}} \\ &\propto \frac{(\alpha+\beta)^{-5/2} \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}{\prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}} \end{aligned} \tag{2}$$

(b) Plot the marginal posterior density $p(\alpha, \beta|y)$ as a function of the transformed variables $\log\left(\frac{\alpha}{\beta}\right)$ and $\log(\alpha+\beta) \in [(-1.3, -2.3); (1, 5)]$. Obtain the corresponding value of (α, β).

Answer: Let $X = \log\frac{\alpha}{\beta}, Y = \log(\alpha+\beta)$. Then $\beta = \frac{\exp(Y)}{1+\exp(X)}, \alpha = \exp(X)\beta$. Using the python code that provided to us.

$X = -1.79, Y = 2.99$,then the corresponding value of $(\alpha, \beta) = (2.85, 17.04)$

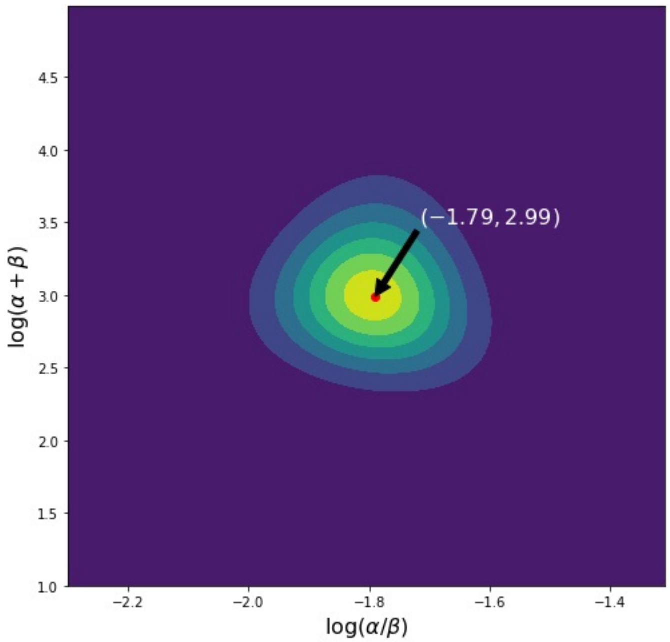

Figure 1: the marginal posterior density $p(\alpha, \beta|y)$ as a function of the transformed variables $\log\left(\frac{\alpha}{\beta}\right)$ and $\log(\alpha+\beta)$

2 Problem 2

Jeffrey's prior and maximum entropy prior: Consider a random variable x described by a Possion distribution:

$$x \sim p(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} \tag{3}$$

(a) Determine the Jeffrey prior π^J for θ. Is the scale invariant prior $\pi_0(\theta) = \frac{1}{\theta}$ preferable to π^J? Why?

Answer:

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \tag{4}$$

Therefore,

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln p(x|\theta)\right] = \frac{\theta}{\theta^2} = \frac{1}{\theta} \tag{5}$$

Therefore the Jeffreys' prior is given by:

$$\pi^J = [I(\theta)]^{1/2} = \theta^{-1/2} \tag{6}$$

The scale invariant prior $\pi_0(\theta) = \frac{1}{\theta}$ is not preferable to π^J because they are not the same function.

(b) Find the maximum entropy prior for θ for the reference measure π^J subject to the constraints $E^\pi[\theta] = 1$, $Var^\pi[\theta] = 1$.

Answer: considering the reference measure as $\pi_{ref} = \pi^J \propto \theta^{-1/2}$.

The maximum entropy prior under the constraints that the prior mean and variance of θ are both 1:

Two constrains, therefore, $K = 2$.

$E^\pi[\theta] = 1$, $g_1(\theta) = \theta$.

$Var^\pi[\theta] = 1 = E[(\theta - 1)^2]$, $g_2(\theta) = (\theta - 1)^2$.

$$\hat{\pi} = \frac{\pi_{ref}(\theta) \exp\Big(\sum_{k=1}^K \lambda_k g_k(\theta)\Big)}{\int \pi_{ref}(\theta) \exp\Big(\sum_{k=1}^K \lambda_k g_k(\theta)\Big)} \tag{7}$$

In this problem,

$$\hat{\pi} \propto \theta^{-1/2} \exp\big(\lambda_1 \theta + \lambda_2 (\theta - 1)^2\big) \tag{8}$$

(c) Find the maximum entropy prior for θ for the reference measure π_0 subject to the constraints $E^\pi[\theta] = 1$, $Var^\pi[\theta] = 1$.

Answer: Considering the reference measure as $\pi_{ref} = \pi_0 \propto \theta^{-1}$.

The maximum entropy prior under the constraints that the prior mean and variance of θ are both 1:

Two constrains, therefore, $K = 2$.

$E^\pi[\theta] = 1$, $g_1(\theta) = \theta$.

$Var^\pi[\theta] = 1 = E[(\theta - 1)^2]$, $g_2(\theta) = (\theta - 1)^2$.

$$\hat{\pi} = \frac{\pi_{ref}(\theta) \exp\Big(\sum_{k=1}^K \lambda_k g_k(\theta)\Big)}{\int \pi_{ref}(\theta) \exp\Big(\sum_{k=1}^K \lambda_k g_k(\theta)\Big)} \tag{9}$$

In this problem,

$$\hat{\pi} \propto \theta^{-1} \exp\big(\lambda_1 \theta + \lambda_2 (\theta - 1)^2\big) \tag{10}$$

3 Problem 3

Laplace approximation: the data set $X = (X_1, ..., X_n)$ presents the number of the wins of a football team in the past n home games. We can model this using

$$X_i \sim g(x_i|\theta) = \theta(\theta+1)x_i^{\theta-1}(1-x_i), x_i = (0,1) \tag{11}$$

with parameter $\theta > 0$. Unfortunately, this model does not have any corresponding, useful, conjugate prior. But it is acceptable to impose a prior model on θ with Gamma distribution.

(a) Derive the posterior PDF of θ.

Answer:

$$\begin{aligned} p(\theta|x) &= Gamma(\theta; a, b)\prod_{i=1}^{n} p(x_i|\theta) \\ &= \frac{b^a\theta^{a-1}\exp\{-b\theta\}}{\Gamma(a)}\theta^n(\theta+1)^n\prod_{i=1}^{n} x_i^{\theta-1}(1-x_i) \end{aligned} \tag{12}$$

(b) Using Laplace approximation, find a normal distribution but approximates the posterior distribution using n = 20.

$$\sum_{x=i} \ln X_i = -4.59 \tag{13}$$

and a = b = 1 where a and b are the hyperparameters of the gamma distribution $Gamma(a, b)$.

Answer:

$n = 20; a = b = 1$

$$p(\theta|x) = \frac{\exp\{-\theta\}}{\Gamma(1)}\theta^{20}(\theta+1)^{20}\prod_{i=1}^{20} x_i^{\theta-1}(1-x_i) \tag{14}$$

$$\log p(\theta|x) = -\theta + 20\log(\theta(\theta+1)) + (\theta-1)\sum_{i}^{20}\log(x_i) + C \tag{15}$$

the first derivative

$$\frac{d\log p(\theta|x)}{d\theta} = -1 + \frac{20}{\theta} + \frac{20}{\theta+1} + \sum_{i}^{20}\log(x_i) = 0 \tag{16}$$

$$\theta^{MAP} \approx 6.69 \tag{17}$$

the second derivative

$$A = -\frac{d^2\log p(\theta|x)}{d\theta^2}\theta = \theta^{MAP} = 6.69 = -\frac{20}{\theta^2} - \frac{20}{(\theta+1)^2} \approx -(-0.785) = 0.785 = \frac{1}{\sigma^2} \tag{18}$$

Therefore,

$$
\begin{aligned}
p(\theta|x) &\approx (2\pi)^{-10}|A|^{1/2}\exp\left\{-\frac{1}{2}(\theta-\theta^{MAP})^2A\right\} \\
&\approx (2\pi)^{-10}(0.785)^{1/2}\exp\left\{-\frac{1}{2}(\theta-6.69)^2 0.785\right\} \\
&\approx (2\pi)^{-10}(\frac{1}{\sigma^2})^{1/2}\exp\left\{-\frac{1}{2\sigma^2}(\theta-\theta^{MAP})^2\right\}
\end{aligned}
\tag{19}
$$

where $\frac{1}{\sigma^2} = 0.785, \theta^{MAP} = 6.69$

4 Problem 4

Monte Carlo integration: Consider the following function,

$$f(x) = x^3 + 5x\cos x \tag{20}$$

(a) Calculate the integral $I = \int_a^b f(x)dx$ with $a = 3$ and $b = 4$ using Monte Carlo integration with $N = 10000$ samples. Compare this value with the exact solution.

Answer:

$$\begin{aligned} I_{exact} &= \int_a^b f(x)dx \\ &= \int_a^b (x^3 + 5x\cos x)dx \\ &= \left[\frac{x^4}{4} + 5x\sin x + 5\cos x\right]|_a^b \\ &= \left[\frac{x^4}{4} + 5x\sin x + 5\cos x\right]|_3^4 \\ &\approx 28.178894351627594 \end{aligned} \tag{21}$$

Using Monte Carlo integration with $N = 10,000$ samples. using the python's function numpy.random.uniform(3,4,10000).

The integral I from Monte Carlo integration

$$I_{MC} = 28.192521342751455 \approx I_{exact} \tag{22}$$

the error is about 0.05%.

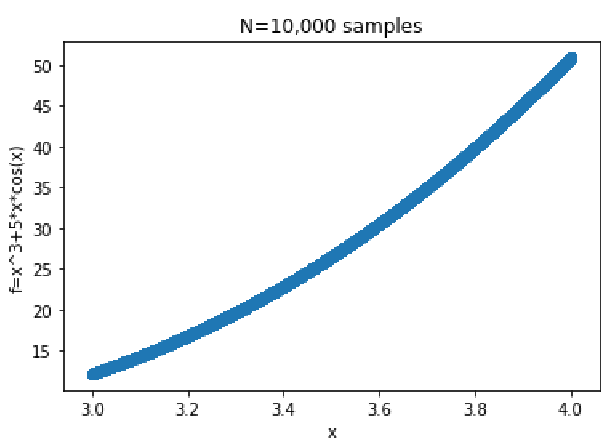

Figure 2: Monte Carlo integration for P4a with N= 10,000

(b) Check the relation between the number of samples N and solution accuracy by plotting the error for $N = [10, 1000]$.

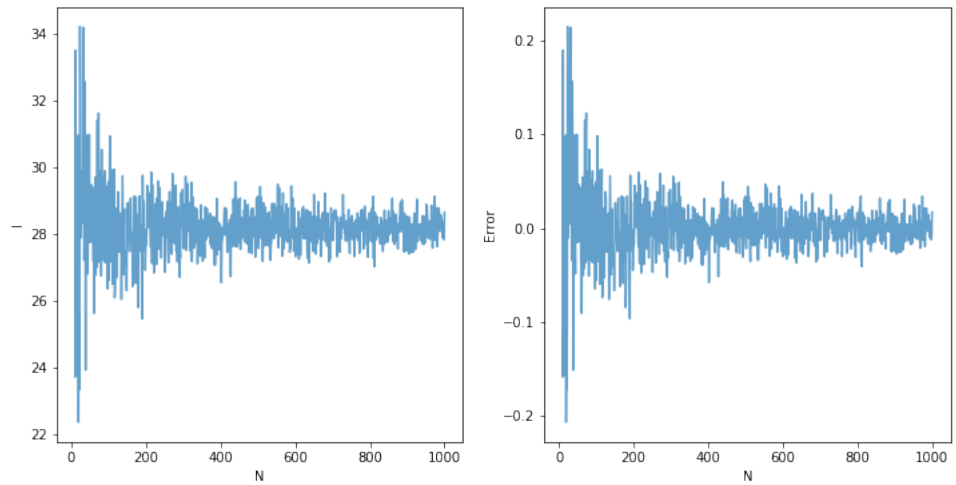

Figure 3: a. Integral I for N = [10,1000] b. Error for N = [10,1000]

Answer:

From the plot, we find that when N is small, the error is very large, but when N increases, the error becomes smaller.

(c) For $N = 100, 1000, 10000$ and 100000 repeat the MC integration for m = 10000 times. Plot the histogram of the results of MC integration for each N. Use the law of large numbers to justify the trend in the histograms.

Answer:

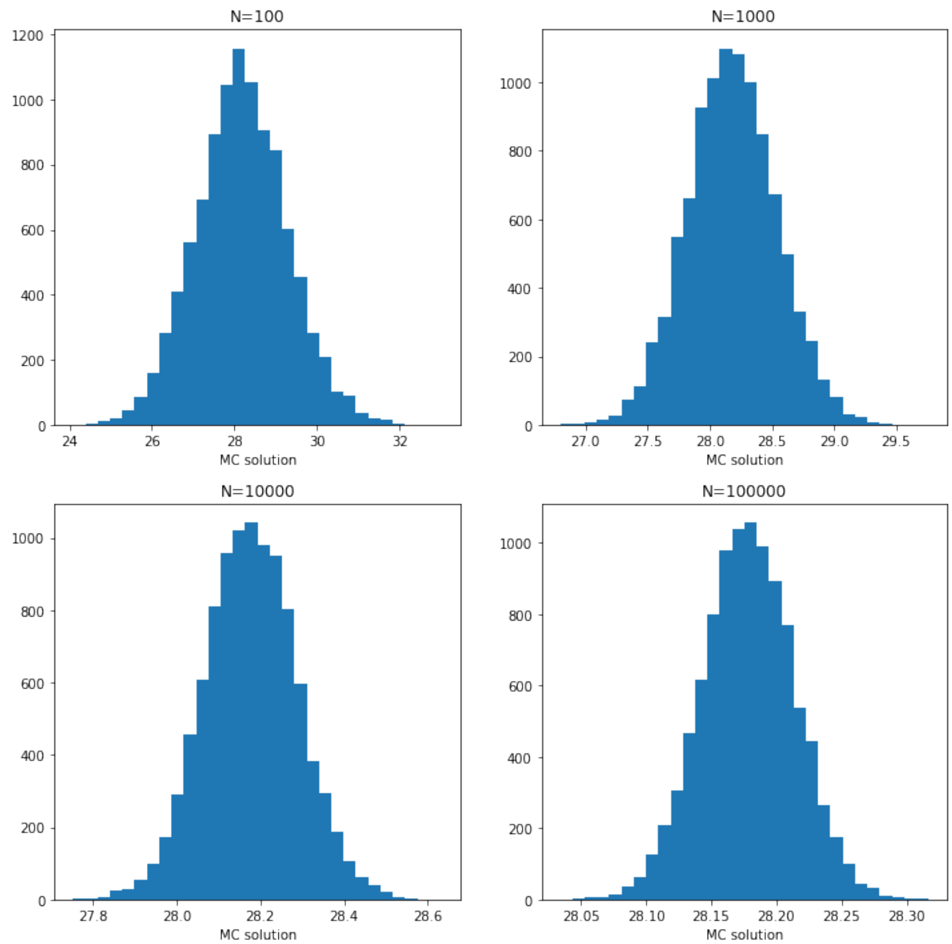

Figure 4: Histogram of the results of MC for each N

Using the law of large numbers to justify the trend in the histogram. $Var[\bar{X_N}] = \frac{\sigma^2}{N}$, therefore when N increases, $Var[\bar{X_N}] = \frac{\sigma^2}{N}$ decreases, the histogram becomes more like a Gaussian distribution.

5 Problem 5

Bayesian Information Criterion (BIC): suppose we toss a biased coin where probability of heads ($x = 1$) is θ_1. However, we only know about the outcome through an unreliable friend of ours, Joey, who can be trusted with a probability θ_2. Let us call this report y. This means that we can write down $p(y|x, \theta_2)$ as

(a) What is the joint probability distribution $p(x, y|\theta_1, \theta_2$? Write your name in a table.

Answer:

	y=0	y=1
x=0	$(1-\theta_1)\theta_2$	$(1-\theta_1)(1-\theta_2)$
x=1	$\theta_1(1-\theta_2)$	$\theta_1\theta_2$

(b) Consider we have the outcomes

$$\begin{aligned} x &= (1, 1, 0, 1, 1, 0, 0) \\ x &= (1, 0, 0, 0, 1, 0, 1) \end{aligned} \tag{23}$$

Find the maximum likelihood estimate for θ_1 and θ_2.

$$p(\theta_1, \theta_2|X, Y) = \theta_1^4(1-\theta_1)^3\theta_2^4(1-\theta_2)^3 \tag{24}$$

$$\log p(\theta_1, \theta_2|X, Y) = 4\log\theta_1 + 3\log(1-\theta_1) + 4\log\theta_2 + 3\log(1-\theta_2) \tag{25}$$

$$\frac{\log p}{d\theta_2} = \frac{4}{\theta_2} - \frac{3}{1-\theta_2} = 0; \theta_2 = \frac{4}{7} \tag{26}$$

$$\frac{\log p}{d\theta_1} = \frac{4}{\theta_1} - \frac{3}{1-\theta_1} = 0; \theta_1 = \frac{4}{7} \tag{27}$$

(c)We denote this model with M_2, where index 2 stands for the number of parameters in the model. Find $p(D|\hat{\theta_1}, \hat{\theta_2}, M_2)$ where $\hat{\theta}$ denotes the MLE solution for parameter θ.

Answer:

From b, $\theta_1 = \frac{4}{7}$ and $\theta_2 = \frac{4}{7}$

	y=0	y=1
x=0	$\frac{12}{49}$	$\frac{9}{49}$
x=1	$\frac{12}{49}$	$\frac{16}{49}$

$$p(X, Y|\theta_1, \theta_2, M_2) = (\frac{4}{7})^8(\frac{3}{7})^6 \approx 7.044 \times 10^{-5} \tag{28}$$

(d) If we also denote a model with 4 parameters $\bar{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$ that represents $p(x, y|\bar{\theta}) = \theta_{x,y}$. Find the MLE of $\bar{\theta}$.

Answer:

	y=0	y=1
x=0	$\theta_{0,0}$	$\theta_{0,1}$
x=1	$\theta_{1,0}$	$\theta_{1,1}$

$$\begin{aligned} p(\bar{\theta}|X,Y) &= \theta_{0,0}^2\theta_{0,1}\theta_{1,0}^2\theta_{1,1}^2 \\ \theta_{0,1} &= 1-\theta_{0,0}-\theta_{1,0}-\theta_{1,1} \end{aligned} \tag{29}$$

$$p(\bar{\theta}|X,Y) = \theta_{0,0}^2(1-\theta_{0,0}-\theta_{1,0}-\theta_{1,1})\theta_{1,0}^2\theta_{1,1}^2 \tag{30}$$

$$\log p = 2\log\theta_{0,0} + \log(1-\theta_{0,0}-\theta_{1,0}-\theta_{1,1}) + 2\log\theta_{1,0} + 2\log\theta_{1,1} \tag{31}$$

$$\frac{d\log p}{d\theta_{0,0}} = \frac{2}{\theta_{0,0}} - \frac{1}{1-\theta_{0,0}-\theta_{1,0}-\theta_{1,1}} = 0 \tag{32}$$

$$\frac{d\log p}{d\theta_{1,0}} = \frac{2}{\theta_{1,0}} - \frac{1}{1-\theta_{0,0}-\theta_{1,0}-\theta_{1,1}} = 0 \tag{33}$$

$$\frac{d\log p}{d\theta_{1,1}} = \frac{2}{\theta_{1,1}} - \frac{1}{1-\theta_{0,0}-\theta_{1,0}-\theta_{1,1}} = 0 \tag{34}$$

$$\begin{aligned} \theta_{0,0} &= \frac{2}{7} \\ \theta_{0,1} &= \frac{1}{7} \\ \theta_{1,0} &= \frac{2}{7} \\ \theta_{1,1} &= \frac{2}{7} \end{aligned} \tag{35}$$

(e) Find $p = (D|\hat{\theta}, M_4)$ where $\hat{\theta}$ denotes the MLE solution for parameters $\bar{\theta}$.
Answer:

	y=0	y=1
x=0	$\frac{2}{7}$	$\frac{1}{7}$
x=1	$\frac{2}{7}$	$\frac{2}{7}$

$$p(X,Y|\theta_{0,0},\theta_{0,1},\theta_{1,0},\theta_{1,1},M_2) = (\frac{2}{7})^6(\frac{1}{7}) \approx 7.771\times 10^{-5} \tag{36}$$

(f) Find the Bayesian Information Criterion for M_2 and M_4. Which model is preferred by this criterion?

Answer: From the Bayesian Information Criterion for M_2 and M_4,
For M_2,$k=2, N=7, L = p(X,Y|\theta_1,\theta_2,M_2) = (\frac{4}{7})^8(\frac{3}{7})^6 \approx 7.044\times 10^{-5}$

$$BIC = k\log(N) - 2\log(L) \approx 23.01 \tag{37}$$

For M_4,$k = 3$(there is a constrain for the four parameters, therefore only three parameters are free parameters, and k should be 3), $N = 7$, $L = p(X, Y|\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1}, M_2) = (\frac{2}{7})^6(\frac{1}{7}) \approx 7.771 \times 10^{-5}$

$$BIC = k\log(N) - 2\log(L) \approx 24.76 \tag{38}$$

Since $BIC_{M_4} > BIC_{M_2}$, M_2 is more preferred by this criterion.

6 Problem 6

Maximum Likelihood Estimation (MLE) and Maximum A Posterior (MAP): Consider a random variable x described by

(a) Derive the maximum likelihood estimate (MLE) (λ_{MLE})

Answer: The likelihood is given by

$$p(x|\lambda) = \prod_{i=1}^{n} \lambda \exp\{-\lambda x_i\} \tag{39}$$

The log likelihood:

$$\ln p = n \ln \lambda - \lambda \sum x_i \tag{40}$$

Now set set derivative w.r.t λ to 0:

$$\frac{d \ln p}{d\lambda} = \frac{n}{\lambda} + \sum x_i = 0 \tag{41}$$

Therefore,

$$\lambda_{MLE} = \frac{1}{\bar{x}}. \left(\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \right) \tag{42}$$

(b)Obtain an analytic form of the posterior distribution of Eq.(11) and Derive the maximum a posterior estimator (MAP) λ_{MAP} as a function of α, β.

Answer: Let us consider the data $X = x_1, x_2, ..., x_n$. The posterior distribution $p(\lambda|X)$ is given by:

$$\begin{aligned} p(\lambda|X) &= \frac{p(\lambda|X)p(\lambda)}{\int p(\lambda|X)p(\lambda)} \\ &\propto p(\lambda|X)p(\lambda) \\ &\lambda^n \exp\left\{ -\lambda \sum_{i=1}^{N} x_i \right\} Gamma(\alpha, \beta) \\ &\lambda^n \exp\left\{ -\lambda \sum_{i=1}^{N} x_i \right\} \lambda^{\alpha-1} \exp\{-\beta\lambda\} \\ &e^{-\lambda(\sum_{i=1}^{N} x_i + \beta)} \lambda^{n+\alpha-1} \end{aligned} \tag{43}$$

$$p(\lambda|X) \propto Gamma(\alpha + n, \sum_{i=1}^{N} x_i + \beta) \tag{44}$$

The log posterior:

$$\log p(\lambda|X) \propto -\lambda(\sum_{i=1}^{N} x_i + \beta) + (n + \alpha - 1) \log \lambda \tag{45}$$

$$0 = \frac{d\log p(\lambda|X)}{d\lambda} = -(\sum_{i=1}^{N} x_i + \beta) + \frac{n+\alpha-1}{\lambda} \tag{46}$$

$$\lambda_{MAP} = \frac{n+\alpha-1}{\sum_{i=1}^{N} x_i + \beta} \tag{47}$$

(c) Generate $N = 20$ samples drawn from an exponential distribution with parameter $\lambda = 0.2$. Fix $\beta = 100$ and vary α over the range(1,40) using a step-size of 1.

Compute the corresponding MLE and MAP estimates for λ

For each α, compute the mean squared $error^2$ of both estimates compared against the true value and then plot the mean squared error as a function of α.

Now, fix $\alpha = 30$, $\theta = 100$ and vary N over the range(1,500) using a step-size of 1. Plot the mean squared error for each N of the corresponding estimates and explain under what condition is the MAP estimator better.

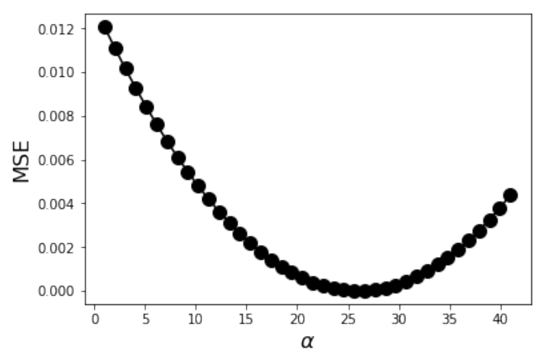

Figure 5: MSE as a function of α

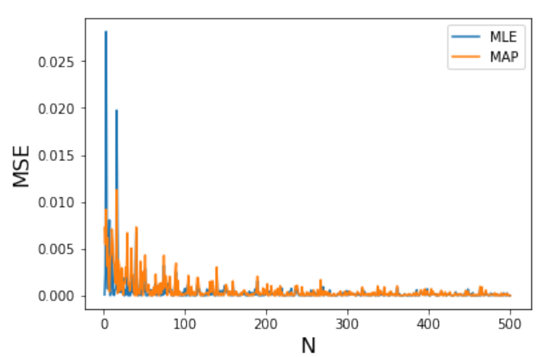

Figure 6: MSE as a function of N

From Figure 6, when N is small, then the MAP estimator would be better than the MLE estimator.