HW5
Jiahong Hu
Jh3561
STAT 4400

Problem 1

Problem 1.
part a

$H(x) = -P(1)\log(1) - P(2)\log(2) - P(4)\log(4) - P(8)\log(8)$

by Lagrange with $P(1) + P(2) + P(4) + P(8) = 1$

$\mathcal{L}(x) = H(x) - \lambda[P(1) + P(2) + P(4) + P(8) - 1]$

$$\left.\begin{aligned} \frac{\partial L}{\partial P(1)} &= -\log P(1) - 1 - \lambda = 0 \\ \frac{\partial L}{\partial P(2)} &= -\log P(2) - 1 - \lambda = 0 \\ \frac{\partial L}{\partial P(4)} &= -\log P(4) - 1 - \lambda = 0 \\ \frac{\partial L}{\partial P(8)} &= -\log P(8) - 1 - \lambda = 0 \end{aligned}\right\} \Rightarrow P(1) = P(2) = P(4) = P(8) = \frac{1}{4}$$

part b.

$H(x) = -\sum_{i=1}^{4} P_i \log_2 P_i \quad 0 \le P_i \le 1 \quad P_i \log_2 P_i \le 0 \quad H(x) \ge 0$

$\Rightarrow \because H(x) \ge 0$

$\therefore$ The smallest possible value for $H(x)$ is 0

Consider set $\{x_1, \cdots, x_n\}$

if $P(X_1 = 1)$ while $P(x_j) = 0$, for $j > 1$, $H(P) = -1\log(1) = 0$.

In this case, the $\Sigma P_j = 1$, H is the smallest part c

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$\frac{\partial L}{\partial P_i} = -\log P_i - 1 + \lambda = 0$$

$$\frac{\partial L}{\partial \gamma} = \sum P_i - 1 = 0$$

$$\Rightarrow \begin{cases} \log P_i = -1 + \alpha \\ \sum P_i = 1 \end{cases} \Rightarrow P_i = \frac{1}{2^n}$$

$$\Rightarrow H_{max} = -\sum P_i \log_2 P_i = -\frac{1}{2^n} \log_2 \frac{1}{2^n} = n \log_2 (2)$$

part d.

$$P_{int} = \{P(X_0 = 0), P(X_0 = 1)\} = (\tfrac{1}{2}, \tfrac{1}{2})$$

$$P = \begin{bmatrix} P_{0\to 1} & P_{1\to 1} \\ P_{0\to 0} & P_{1\to 0} \end{bmatrix} = \begin{bmatrix} \tfrac{1}{2} & \tfrac{1}{2} \\ \tfrac{1}{2} & \tfrac{1}{2} \end{bmatrix}$$

part e.

It has lower entropy compared to the chain above, according to the conclusion we got above.

part f.

$$P_t = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}$$

$$P_t \cdot P_{eq} = P_{eq}$$

$$\begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \end{bmatrix} = \begin{bmatrix} p_0 \\ p_1 \end{bmatrix} \Rightarrow \begin{cases} p_0 = \frac{3}{4} p_1 \\ 1 = p_0 + p_1 \end{cases}$$

$$\Rightarrow \begin{cases} p_0 = \frac{3}{7} \\ p_1 = \frac{4}{7} \end{cases} \qquad \Rightarrow P_{eq} = \begin{bmatrix} p_0 \\ p_1 \end{bmatrix} = \begin{bmatrix} \frac{3}{7} \\ \frac{4}{7} \end{bmatrix}$$

part g.

by part (f)

$$\lim_{n \to \infty} P(X_n = 1) = \frac{4}{7}$$

$$\lim_{n \to \infty} P(X_n = 0) = \frac{3}{7}.$$

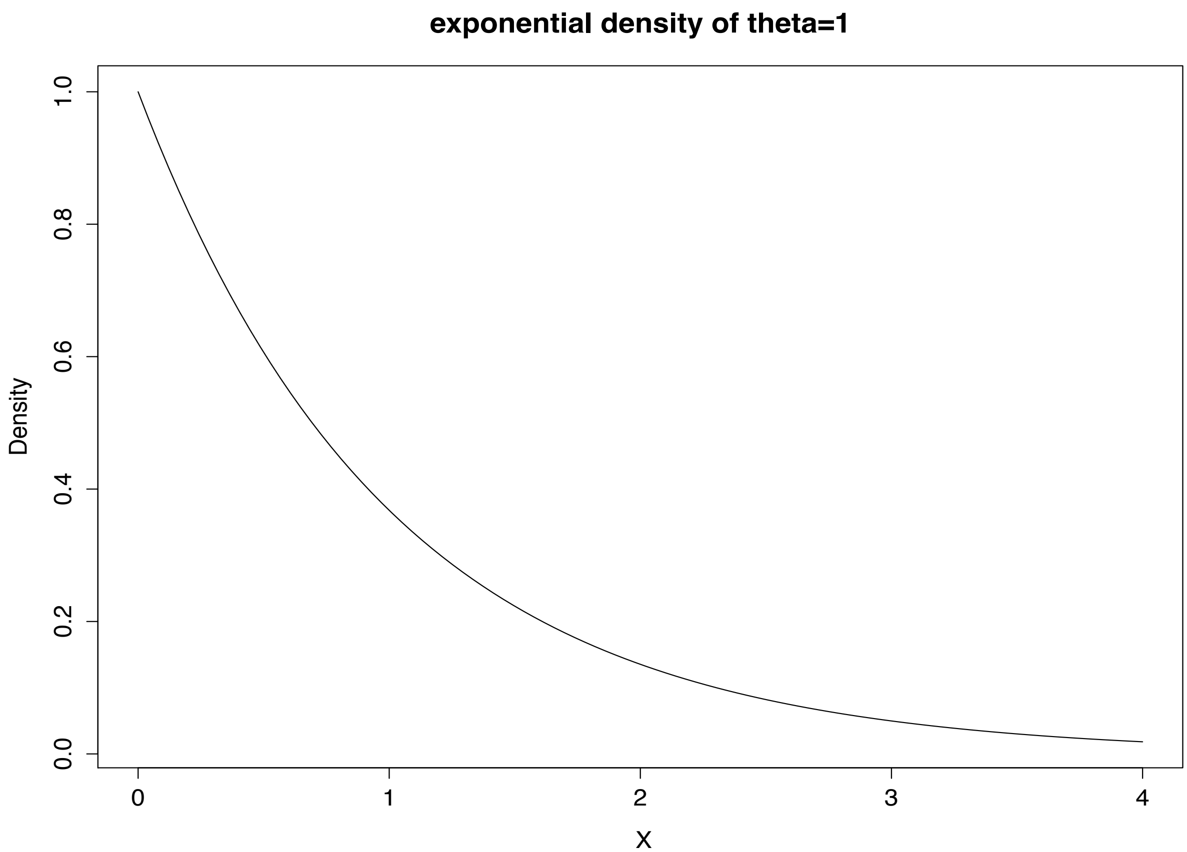
part h. $\quad X_1 \sim Ber(\frac{4}{7}) \quad P(X_i = 0) = \frac{3}{7} \quad P(X_i = 1) = \frac{4}{7}$

$$H(X_1) = -\sum_{i=1} \log P_i \, P_i = -\frac{3}{7} \log_2 (\frac{3}{7}) - \frac{4}{7} \log(\frac{4}{7}) = 0.985$$

$$H(X_1 \cdots X_n) = H(X_1) + (n-1) H(X_2 | X_1)$$
$$= 0.985 + (n-1) \left(\frac{3}{7} \left(\frac{3}{5} \log \frac{3}{5} + \frac{2}{5} \log \frac{2}{5} \right) + \frac{4}{7} \left(\frac{3}{10} \log \frac{3}{10} + \frac{7}{10} \log \frac{7}{10} \right) \right)$$
$$= 0.985 + (n-1) 0.9197$$

Problem 2

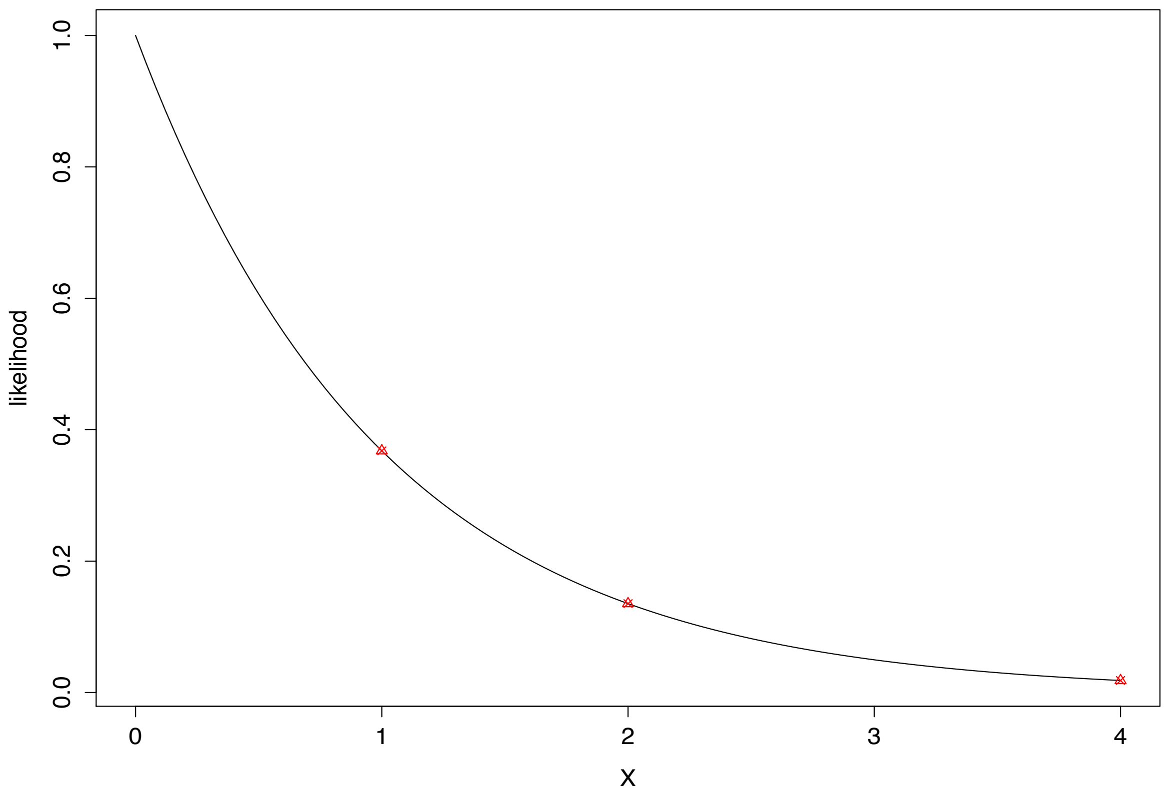
Part a



Part b

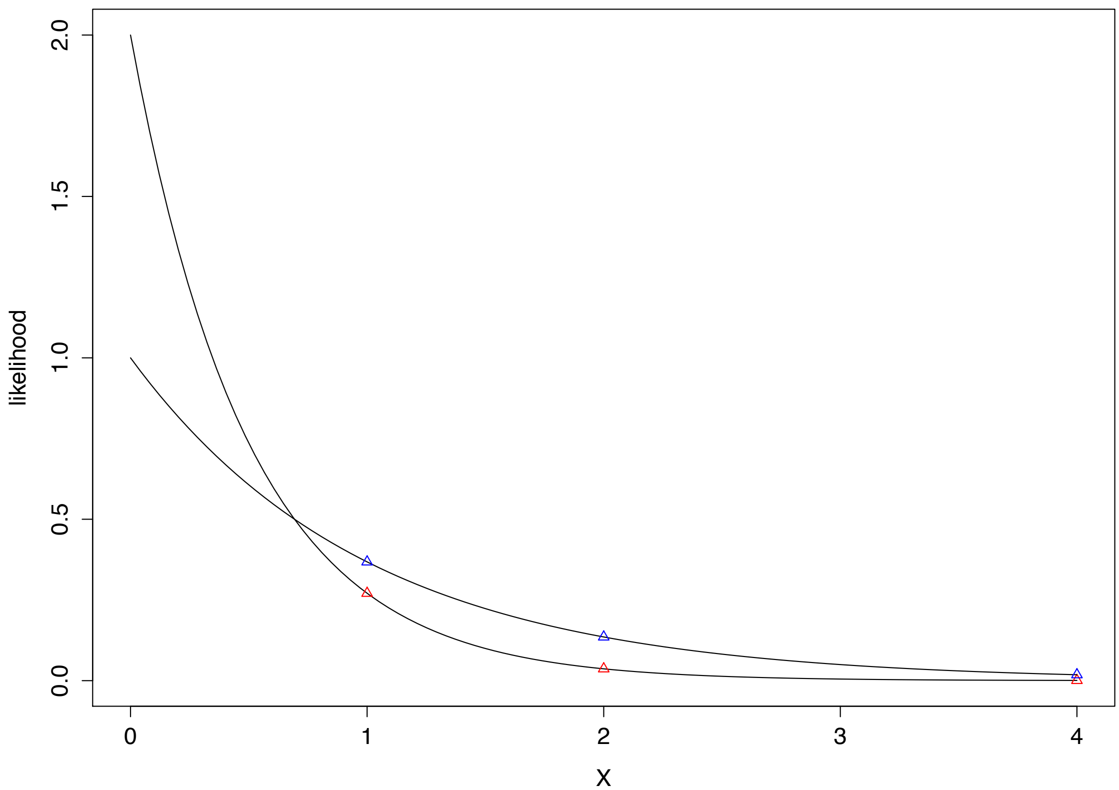
The likelihood for each point under the exponential distribution with theta =1 is calculated as exp(-x). The likelihood of the dataset X={1,2,4} is calculated as below.
L=p(x1,x2,x3|that=1)=p(x1|theta=1)*p(x2|theta=1)*p(x3|theta=1)=exp(-7).

The plot of likelihood for the three points is shown as below.



Part c

As shown in graph below, the likelihood for each point in the set X={1,2,4} when theta equals to 2 is lower than that while theta is 1. In the mathematical proof, p(x1,x2,x3|theta=2)=2exp(-2x1)*2exp(-2x2)*2exp(-2x3)=8exp(-14)<exp(-7). Hence, a higher theta results in a smaller likelihood value.



2 (d) Show that if the data is exponentially distributed with a gamma prior the posterior is again a gamma. $P(x_i|\theta) = \theta e^{-\theta x_i}$

$$P(\theta|x_1, \cdots x_n) \propto P(\theta|\alpha_0, \beta_0) \cdot L(x_1, \cdots x_n|\theta)$$

$$\propto \text{Gamma}(\theta|\alpha_0, \beta_0) \prod_{i=1}^{n} P(x_i|\theta)$$

$$\propto \prod_{i=1}^{n} \theta e^{-\theta x_i} \cdot \frac{\theta^{\alpha_0 - 1} \beta_0^{\alpha_0} e^{-\beta_0 \theta}}{\Gamma(\alpha_0)}$$

$$\propto \theta^n e^{-\theta \sum_{i=1}^{n} x_i} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0 - 1} e^{-\beta_0 \theta}$$

$$\propto \theta^{n+\alpha_0-1} \exp(-\theta(\beta_0 + \sum_{i=1}^{n} x_i)) \cdot \cancel{\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}}$$

$$\propto \theta^{n+\alpha_0-1} \exp(-\theta(\beta_0 + \sum_{i=1}^{n} x_i))$$

$$P(\theta|x_1, \cdots x_n) \sim \text{Gamma}(\alpha^*, \beta^*)$$

$$\alpha^* = n + \alpha_0$$

$$\beta^* = \beta_0 + \sum_{i=1}^{n} x_i$$

2(e) Show that if $p(x|\theta)$ is an exponential family model and $q(\theta)$ its nature conjugate prior, the posterior $\pi(\theta|x_{1:n})$ under n observations can be computed as the posterior given a x_n, using prior $\tilde{q}(\theta)$: $\pi(\theta|x_{1:n-1})$

$$\pi(\theta|x_1, \cdots x_n) = \frac{P(x_1, \cdots x_n|\theta) q(\theta)}{P(x_1, \cdots x_n)} = \frac{\prod_{i=1}^{n} P(x_i|\theta) q(\theta)}{P(x_1, \cdots, x_n)} = \frac{\prod_{i=1}^{n} P(x_i|\theta) q(\theta)}{\prod_{i=1}^{n} P(x_i)}$$

$$= \frac{P(x_n|\theta)}{P(x_n)} \left(\frac{\prod_{i=1}^{n-1} P(x_i|\theta)}{\prod_{i=1}^{n-1} P(x_i)} q(\theta) \right)$$

$$\tilde{q}(\theta) = \pi(\theta|x_1, \cdots x_{n-1}) = \frac{\prod_{i=1}^{n-1} P(x_i|\theta) q(\theta)}{\prod_{i=1}^{n-1} P(x_i)}$$

$$= \frac{P(x_n|\theta)}{P(x_n)} \pi(\theta|x_1 \cdots x_{n-1}) = \frac{P(x_n|\theta)}{P(x_n)} \tilde{q}(\theta).$$

Note: $x_1, \cdots x_n$ are iid.

Part 2f

$$\pi(\theta \mid x_1, \cdots x_n, \alpha_0, \beta_0) = \frac{P(x_n \mid \theta)\, \pi(\theta \mid x_1, x_2, \ldots x_{n-1})}{P(x_n)}$$

$$= \frac{P(x_n \mid \theta)\, \tilde{q}(\theta)}{P(x_n)}$$

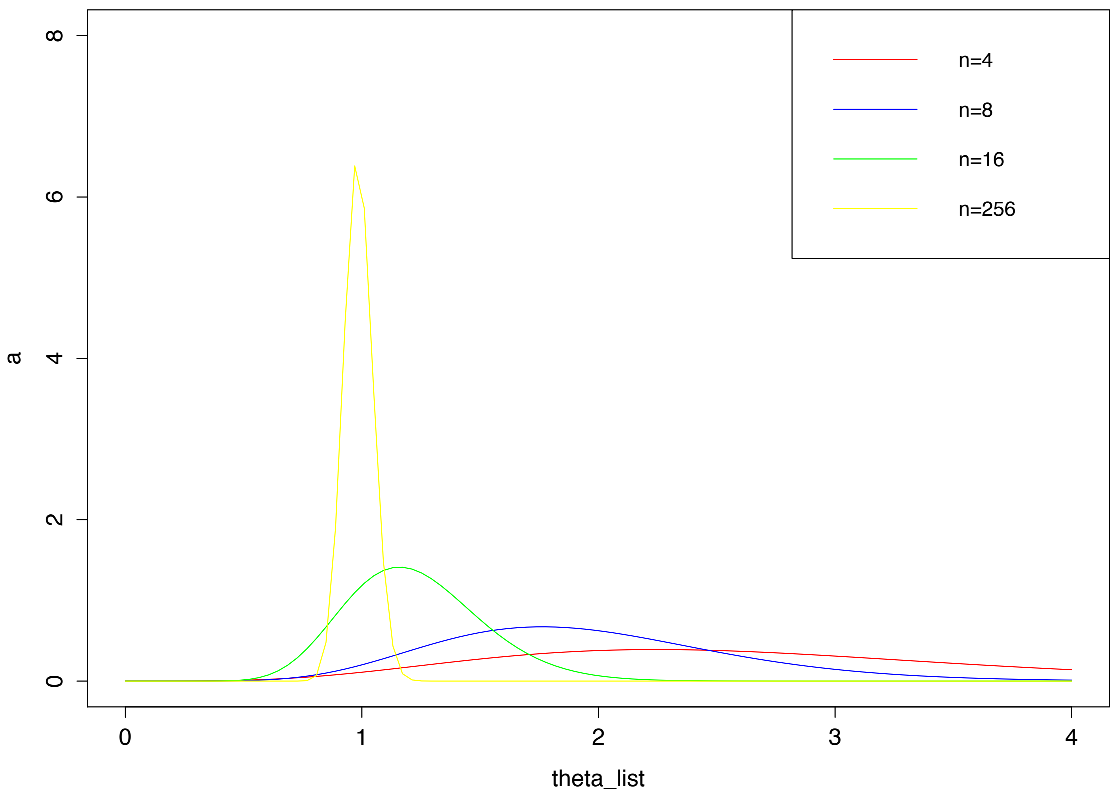
$$= P(x_n \mid \theta) \cdot \beta_{n-1}^{\alpha_{n-1}}\, \theta^{\alpha_{n-1}-1} \exp(-\theta \beta_{n-1})$$

$$= \theta e^{-\theta x_n} \cdot \beta_{n-1}^{\alpha_{n-1}}\, \theta^{\alpha_{n-1}} \exp(-\theta \beta_{n-1})$$

$$\propto \theta^{(\alpha_{n+1})-1} \exp(-\theta(\beta_{n-1} + x_n)$$

$$\propto \text{Gamma}(\alpha_n + 1, \beta_{n-1} + x_n).$$

Part g



As n increasing, the variance of the posterior distribution decrease and the mean of the distribution become closer to the true value, which is 1 in this case.