Exercise 3.1 MLE for the Bernoulli/ binomial model

$$
\begin{aligned}
\frac{d}{d\theta}p(D|\theta) &= \frac{d}{d\theta}(\theta^{N_1}(1-\theta)^{N_0}) &(1)\\
&= N_1\theta^{N_1-1}(1-\theta)^{N_0} - N_0\theta^{N_1}(1-\theta)^{N_0 1} &(2)\\
&= \theta^{N_1-1}(1-\theta)^{N_0-1}(N_1(1-\theta) - N_0\theta) &(3)\\
&= \theta^{N_1-1}(1-\theta)^{N_0-1}(N_1 - N\theta) &(4)\\
\therefore \theta_{\text{MLE}} &= \frac{N_1}{N} &(5)
\end{aligned}
$$

Exercise 3.2 Marginal likelihood for the Beta-Bernoulli model

$$
\begin{aligned}
p(D) &= \frac{[(\alpha_1)\cdots(\alpha_1+N_1-1)][(\alpha_0)\cdots(\alpha_0+N_0-1)]}{(\alpha)\cdots(\alpha+N-1)} &(6)\\
&= \frac{[(\alpha_1)\cdots(\alpha_1+N_1-1)][(\alpha_0)\cdots(\alpha_0+N_0-1)]}{(\alpha_1+\alpha_0)\cdots(\alpha_1+\alpha_0+N-1)} &(7)\\
&= \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1+\alpha_0+N)}[(\alpha_1)\cdots(\alpha_1+N_1-1)][(\alpha_0)\cdots(\alpha_0+N_0-1)] &(8)\\
&= \frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1+\alpha_0+N)}\frac{\Gamma(\alpha_1+N_1)}{\Gamma(\alpha_1)}\frac{\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_0)} &(9)\\
&= \frac{\Gamma(\alpha_1+N_1)\Gamma(\alpha_0+N_0)}{\Gamma(\alpha_1+\alpha_0+N)}\frac{\Gamma(\alpha_1+\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} &(10)
\end{aligned}
$$

Exercise 3.3 Posterior prdictive for Beta-Binomial model

$$
\begin{aligned}
Bb(1|\alpha_1', \alpha_0', 1) &= \frac{B(1+\alpha_1', 1-1+\alpha_0')}{B(\alpha_1', \alpha_0')}\binom{1}{1} &(11)\\
&= \frac{B(1+\alpha_1', \alpha_0')}{B(\alpha_1', \alpha_0')} &(12)\\
&= \frac{\Gamma(\alpha_1'+\alpha_0')}{\Gamma(\alpha_1')\Gamma(\alpha_0')}\frac{\Gamma(1+\alpha_1')\Gamma(\alpha_0')}{\Gamma(1+\alpha_1'+\alpha_0')} &(13)\\
&= \frac{\Gamma(\alpha_1'+\alpha_0')}{\Gamma(\alpha_1')\Gamma(\alpha_0')}\frac{\alpha_1'\Gamma(\alpha_1')\Gamma(\alpha_0')}{(\alpha_1'+\alpha_0')\Gamma(\alpha_1'+\alpha_0')} &(14)\\
&= \frac{\alpha_1'}{\alpha_1'+\alpha_0'} &(15)
\end{aligned}
$$

Exercise 3.4 Beta updating from censored likelihood

$$
\begin{aligned}
p(\theta, X < 3) &= p(\theta)p(X < 3|\theta) && (16)\\
&= p(\theta)(\sum_{k=0}^{2} p(X = k|\theta)) && (17)\\
&= p(\theta)(\sum_{k=0}^{2} \theta^k (1-\theta)^{(5-k)}) && (18)\\
&= Beta(\theta|1, 1)(\sum_{k=0}^{2} \theta^k (1-\theta)^{(5-k)}) && (19)\\
&= \sum_{k=0}^{2} \theta^k (1-\theta)^{(5-k)} && (20)
\end{aligned}
$$

Exercise 3.5 Uninformative prior for log-odds ratio

$$
\begin{aligned}
p(\theta) &= p(\phi)|\frac{d\phi}{d\theta}| && (21)\\
&= p(\phi)\theta^{-1}(1-\theta)^{-1} && (22)\\
&\propto Beta(\theta|0, 0)\ (\because p(\phi) \propto 1) && (23)
\end{aligned}
$$

Exercise 3.6 MLE for the Poisson distribution

$$
\begin{aligned}
D &= \{x_1, x_2, \ldots, x_N\} && (24)\\
p(D|\lambda) &= \Pi_{i=1}^{N} Poi(x_i|\lambda) && (25)\\
&= e^{-N\lambda} \frac{\lambda^{\sum_{i=1}^{N} x_i}}{\Pi_{i=1}^{N}(x_i!)} && (26)\\
\log p(D|\lambda) &= -N\lambda + \sum_{i=1}^{N} x_i \log\lambda - \sum_{i=1}^{N} \log x_i! && (27)\\
\frac{\partial}{\partial\lambda} \log p(D|\lambda) &= -N + \frac{1}{\lambda}\sum_{i=1}^{N} x_i && (28)\\
\therefore \lambda_{MLE} &= \frac{1}{N}\sum_{i=1}^{N} x_i && (29)
\end{aligned}
$$

Exercise 3.7 Bayesian analysis of the Poisson distribution

(a)

$$
\begin{aligned}
p(\lambda|D) &\propto p(\lambda)p(D|\lambda) && (30)\\
&\propto \lambda^{a-1}e^{-\lambda b}e^{-N\lambda}\frac{\lambda^{\sum_{i=1}^N x_i}}{\Pi_{i=1}^N(x_i!)} && (31)\\
&= e^{-\lambda(N+b)}\frac{\lambda^{\sum_{i=1}^N x_i}}{\Pi_{i=1}^N(x_i!)} && (32)\\
&\propto Ga(\lambda|a+\sum_{i=1}^N x_i, b+N) && (33)
\end{aligned}
$$

(b)

$$
\begin{aligned}
\frac{a+\sum_{i=1}^N x_i}{b+N} &\to \frac{1}{N}\sum_{i=1}^N x_i && (34)\\
&= \lambda_{MLE} && (35)
\end{aligned}
$$

Exercise 3.8 MLE for the uniform distribution

(a)

$$
\begin{aligned}
D &= \{x_1, \ldots, x_N\} && (36)\\
p(D|a) &= \Pi_{i=1}^N \frac{1}{2a} I(x_i \in [-a, a]) && (37)
\end{aligned}
$$

If $\forall i \ -a \leq x_i \leq a$, then $p(D|a) = \frac{1}{(2a)^n}$. More smaller a, more larger $p(D|a)$.
$\hat{a} = \max\{|x_1|, \ldots, |x_N|\}$

(b)

$$
\begin{aligned}
p(x_{n+1}|\hat{a}) &= \frac{1}{2\hat{a}} I(x_{n+1} \in [-\hat{a}, \hat{a}]) && (38)\\
&= \begin{cases} 0 & (x_{n+1} \notin [-\hat{a}, \hat{a}]) \\ \frac{1}{2\hat{a}} & (x_{n+1} \in [-\hat{a}, \hat{a}]) \end{cases} && (39)
\end{aligned}
$$

(c)

If we use MLE approach, the probability between $-\hat{a}$ and $\hat{a}$ is 0. Bayesian approach with introducing a wide range prior is better.

Exercise 3.9 Bayesian analysis of the uniform distribution

$$
\begin{aligned}
p(\theta|D) &= \frac{p(D,\theta)}{p(D)} && (40)\\
&= \begin{cases} \frac{(N+K)b^N}{K} p(D,\theta) & (m \le b) \\ \frac{(N+K)m^{N+K}}{Kb^K} p(D,\theta) & (m > b) \end{cases} && (41)\\
&= \begin{cases} \frac{(N+K)b^{N+K}}{\theta^{N+K+1}} I(\theta \ge \max(D,b)) & (m \le b) \\ \frac{(N+K)m^{N+K}}{\theta^{N+K+1}} I(\theta \ge \max(D,b)) & (m > b) \end{cases} && (42)\\
&= (N+K)\{\max(D,b)\}^{N+K}\theta^{-(N+K+1)} I(\theta \ge \max(D,b)) && (43)\\
&= \mathrm{Pareto}(\theta|N+K, \max(D,b)) && (44)
\end{aligned}
$$

Exercise 3.10 Taxicab (tramcar) problem

(a)

$$
\begin{aligned}
p(\theta|\{100\}) &= \mathrm{Pareto}(\theta|1,100) && (45)\\
&= 100\theta^{-2} I(\theta \ge 100) && (46)
\end{aligned}
$$

(b)

mean

not exist

mode

100

median

$$\int_{100}^{x} 100\theta^{-2}I(\theta \geq 100)d\theta = \frac{1}{2} \tag{47}$$

$$100\int_{100}^{x} \theta^{-2}d\theta = \frac{1}{2} \tag{48}$$

$$-100[\frac{1}{\theta}]_{100}^{x} = \frac{1}{2} \tag{49}$$

$$\frac{1}{x} - \frac{1}{100} = -\frac{1}{200} \tag{50}$$

$$x = 200 \tag{51}$$

(c)

$$p(D'|D,\alpha) = \int_{\theta} p(D'|\theta)p(\theta|D,\alpha)d\theta \tag{52}$$

$$= \int_{\theta} U(x|0,\theta)\mathrm{Pareto}(\theta|1,m)d\theta \tag{53}$$

$$= \int_{\theta} \frac{1}{\theta}I[0 \leq x \leq \theta]m\theta^{-2}I(\theta \geq m)d\theta \tag{54}$$

$$= m\int_{\theta} \theta^{-3}I[0 \leq x \leq \theta]I[\theta \geq m]d\theta \tag{55}$$

$$= m[-\frac{1}{2}\theta^{-2}]_{\max(x,m)}^{\infty} \tag{56}$$

$$= \frac{m}{2\{\max(x,m)\}^2} \tag{57}$$

(d)

$$p(x = 100|D,\alpha) = \frac{1}{200} \tag{58}$$

$$p(x = 50|D,\alpha) = \frac{1}{200} \tag{59}$$

$$p(x = 150|D,\alpha) = \frac{100}{2 \cdot 150^2} = \frac{1}{450} \tag{60}$$

(e)

- More observations, more accurate.
- Thy hyper-parameter of prior should be set by other information.

Exercise 3.11 Bayesian analysis of the exponential distribution

(a)

$$\log p(D|\theta) = N \log \theta - \theta \sum_{i=1}^{N} x_i \tag{61}$$

$$\frac{d}{d\theta} p(D|\theta) = \frac{N}{\theta} - \sum_{i=1}^{N} x_i \tag{62}$$

$$\hat{\theta}_{\text{MLE}} = \frac{1}{\bar{x}} \tag{63}$$

(b)

$$\hat{\theta}_{\text{MLE}} = \frac{1}{\frac{5+6+4}{3}} \tag{64}$$

$$= 1/5 \tag{65}$$

(c)

$$E[\theta] = \int_0^{\infty} \theta p(\theta) d\theta \tag{66}$$

$$= \lambda \theta e^{-\lambda\theta} d\theta \tag{67}$$

$$= \frac{1}{\lambda} \tag{68}$$

$$\hat{\lambda} = 3 \tag{69}$$

If we use hint,

$$p(\theta) \propto \text{Ga}(\theta|1, \lambda) \tag{70}$$

$$E[\theta] = \frac{1}{\lambda} \tag{71}$$

(d)

$$p(\theta|D,\hat{\lambda}) \propto p(D|\theta)p(\theta|\hat{\lambda}) \tag{72}$$
$$= \theta^N \exp(-\theta\sum_{i=1}^{N} x_i)\hat{\lambda}\exp(-\hat{\lambda}\theta) \tag{73}$$
$$= \hat{\lambda}\theta^N \exp(-\theta(\hat{\lambda}+\sum_{i=1}^{N} x_i)) \tag{74}$$
$$\propto \mathrm{Ga}(\theta|N+1,\hat{\lambda}+\sum_{i=1}^{N} x_i) \tag{75}$$

(e)

The exponential prior is NOT conjugate to the exponential likelihood.
(The Gamma prior is conjugate to the exponential likelihood.)

(f)

$$E[\theta|D,\hat{\lambda}] = \frac{N+1}{\hat{\lambda}+\sum_{i=1}^{N} x_i} \tag{76}$$

(g)

$\hat{\lambda}$ can be thought as pseudo sample of machine time. If the expert is collect, this example is reasonable.

Exercise 3.12 MAP estimation for the Bernoulli with non-conjugate priors

(a)

$$p(D|\theta) = \theta^{N_1}(1-\theta)^{N-N_1} \tag{77}$$
$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{78}$$
$$\propto \begin{cases} \theta^{N_1}(1-\theta)^{N-N_1} & (\theta = 0.4 \text{ or } 0.5) \\ 0 & (\text{otherwise}) \end{cases} \tag{79}$$
$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta\in\{0.4,0.5\}} \theta^{N_1}(1-\theta)^{N-N_1} \tag{80}$$

(b)

When N is small, the effect of α and β of the conjugate prior is large and $\theta = 0.41$ may not be estimated. On the other hand, the non-conjugate prior will emit $\theta = 0.4$ or 0.5, so the non-conjugate prior is better.

When N is large, $\hat{\theta}_{\mathrm{MAP}}$ of the conjugate prior converges to 0.41, but $\hat{\theta}_{\mathrm{MAP}}$ of the non-conjugate prior converges to 0.4. Since, the conjugate prior is better.