Problem 2.1

Part A

By the definition of the Bernoulli distribution we have,

\[ p(x; \mu) = \mu^x(1 - \mu)^{1-x} \]

\[
\sum_{x=0}^{1} p(x; \mu) = p(x = 0; \mu) + p(x = 1; \mu) \\
= \mu^0(1 - \mu)^1 + \mu^1(1 - \mu)^0 \\
= 1 - \mu + \mu \\
= 1 \quad \text{(proved)}
\]

Part B

\[
E[x] = \sum_{x=0}^{1} x \cdot p(x; \mu) \\
= 0 \cdot p(x = 0; \mu) + 1 \cdot p(x = 1; \mu) \\
= \mu(1 - \mu)^0 \\
= \mu \quad \text{(proved)}
\]

Part C
\begin{align*}
\text{var}[x] &= \sum_{x=0}^{1} (x - \mu)^2 p(x; \mu) \\
&= \mu^2 \cdot p(x = 0; \mu) + (1 - \mu)^2 \cdot p(x = 1; \mu) \\
&= \mu^2 (1 - \mu) + (1 - \mu)^2 \mu \\
&= (1 - \mu)(\mu^2 + \mu - \mu^2) \\
&= \mu(1 - \mu) \quad \text{(proved)}
\end{align*}

**Part D**

By the definition of entropy we have

\begin{align*}
H[x] &= -\sum_{x=0}^{1} p(x; \mu) \ln p(x; \mu) \\
&= -\sum_{x=0}^{1} \mu^x (1 - \mu)^{1-x} \{\ln(\mu^x(1 - \mu)^{1-x})\} \\
&= -(1 - \mu) \ln(1 - \mu) + \mu \ln(\mu) \\
&= -\mu \ln(\mu) - (1 - \mu) \ln(1 - \mu) \quad \text{(proved)}
\end{align*}

**Problem 2.6**

Using the fact that

\[ \int_0^1 \mu^{a-1}(1 - \mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)} \] (1)

show the following for the beta distribution:

\begin{align*}
\mathbb{E}[\mu] &= \frac{a}{a + b} \quad \text{(2)} \\
\text{var}[\mu] &= \frac{ab}{(a + b)^2(a + b + 1)} \quad \text{(3)} \\
\text{mode}[\mu] &= \frac{a - 1}{a + b - 2} \quad \text{(4)}
\end{align*}
Expectation

By definition, $\mathbb{E}[\mu] = \int f(\mu)p(\mu)d\mu$ where $p$ is the beta distribution.

\begin{align*}
\text{Beta}(\mu; a, b) &= \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1 - \mu)^{b-1} \\
\mathbb{E}[\mu] &= \int_0^1 \mu \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1 - \mu)^{b-1} d\mu \\
&= \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^a(1 - \mu)^{b-1} d\mu \\
&= \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a + 1)}{\Gamma((a + 1) + b)} \\
&= \frac{\Gamma(a + b)\Gamma(a + 1)}{\Gamma(a)\Gamma(a + b + (a + 1))} \\
&= \frac{a}{a + b}
\end{align*}

Variance

We can use a previous homework question which showed that variance is equal to the expectation of square minus square of expectation to prove this. Namely,

\begin{equation}
\text{var}[\mu] = \mathbb{E}[\mu^2] - (\mathbb{E}[\mu])^2
\end{equation}

We just need to repeat the process from part a on the first term, and subtract part a.

\begin{align*}
\mathbb{E}[\mu^2] &= \int_0^1 \mu^2 \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1 - \mu)^{b-1} d\mu \\
&= \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1}(1 - \mu)^{b-1} d\mu \\
&= \frac{\Gamma(a + b)\Gamma(a + 2)}{\Gamma(a)\Gamma((a + 2) + b)} \\
&= \frac{\Gamma(a + b)\Gamma(a + 1)}{(a + b)(a + b + 1)} \\
&= \frac{a(a + 1)}{(a + b)(a + b + 1)}
\end{align*}
Then plugging in to equation 6 we get,

$$\text{var}[\mu] = \frac{a(a + 1)}{(a + b)(a + b + 1)} - \frac{a}{a + b}^2$$

$$= \frac{a(a + 1)(a + b)}{(a + b)^2(a + b + 1)} - \frac{a^2(a + b + 1)}{(a + b)^2(a + b + 1)}$$

multiply terms to get same denom

$$= \frac{ab}{(a + b)^2(a + b + 1)}$$

denom in wanted form, just simplify

**Mode**

The mode is the point where the pdf is greatest, so take derivative of beta distribution wrt $\mu$ and set it equal to zero. This will require using the product rule for differentiation.

$$0 = \frac{\partial}{\partial \mu} \text{Beta}(\mu; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\partial}{\partial \mu} \mu^{a-1}(1 - \mu)^{b-1}$$

$$= (a - 1)\mu^{a-2}(1 - \mu)^{b-1} + \mu^{a-1}(b - 1)(1 - \mu)^{b-2}(-1)$$

product rule

$$= \mu^{a-2}(1 - \mu)^{b-2}[(a - 1)(1 - \mu) + \mu(b - 1)(-1)]$$

dist property

$$0 = (a - 1)(1 - \mu) + \mu(b - 1)(-1)$$

second term

$$0 = -\mu[(a - 1) + (b - 1)] + (a - 1)$$

dist property

$$\mu = \frac{a - 1}{a + b - 2}$$

The mode is undefined when $a + b = 2$. Note that there are three values for $\mu$ where the partial derivative is equal to 0. $\mu = 0, 1, \frac{a - 1}{a + b - 2}$. We can analyze these three values by looking at the shape of the Beta distribution. For values of $a, b > 1$, there is a single maximum that corresponds to the mode; while for values of $a, b < 1$, there are two maximums at $\mu = 0, 1$ (see figure 2.2 in section 2.1.1 of Bishop).

**Problem 2.7**

1. 

$$\text{MLE}[\mu] = \frac{m}{m + l}$$

2. 

$$E[\mu] = \frac{a}{a + b}$$
3. Posteriori distribution

\[ p(\mu|m, l, a, b) = \text{Beta}(\mu \mid a + m, b + l) \quad (9) \]

Hence,

\[ E[\mu|m, l, a, b] = \frac{a + m}{(a + m) + (b + l)} \quad (10) \]

From Eq (1), (2), and (4),

\[ E[\mu \mid m, l, a, b] = \frac{a + m}{a + b + m + l} = \frac{a}{a + b + m + l} + \frac{m}{a + b + m + l} \]

\[ = \frac{a + b}{a + b + m + l} \cdot \frac{a}{a + b + m + l} + \frac{m}{a + b + m + l} \cdot \frac{m}{m + l} \]

where \( \lambda = \frac{a + b}{a + b + m + l} \)

\[ E[\mu \mid m, l, a, b] = \lambda E[\mu] + (1 - \lambda) \text{MLE}[\mu] \quad (11) \]

**Problem 2.8**

**Part 1**

As given, \( E_x[x \mid y] \) denotes the expectation of \( x \) under the conditional distribution \( p(x \mid y) \). Hence, we have,

\[ E_x[x \mid y] = \sum_x x \cdot p(x \mid y) \quad (12) \]
Hence, the right hand side of the equation can be simplified as follows,

\[ E_y[E_x[x|y]] = E_y\left[ \sum_x x \cdot p(x|y) \right] \]

\[ = \sum_y p(y) \sum_x x \cdot p(x|y) \]

\[ = \sum_x \sum_y x \cdot p(x|y) \cdot p(y) \]

\[ = \sum_x \sum_y p(x|y) \cdot p(y) \]

\[ = \sum_x p(x, y) \quad \text{Product Rule} \]

\[ = \sum_x x \cdot p(x) \quad \text{Marginalization} \]

\[ = E[x] \]

Thus, we have,

\[ E[x] = E_y[E_x[x|y]] \quad (13) \]

**Part 2**

To prove that,

\[ \text{var}[x] = E_y[\text{var}_x[x|y]] + \text{var}_y[E_x[x|y]] \]

From the previous HW, we know that

\[ \text{var}[f] = E[f(x)^2] - E[f(x)]^2 \quad (14) \]

Therefore, for the conditional distribution we have,

\[ \text{var}_x[x|y] = E_x[x^2|y] - E_x[x|y]^2 \quad (15) \]

Taking the expectation of equation 15 w.r.t. to the distribution of \( y \) we get,

\[ E_y[\text{var}_x[x|y]] = E_y[E_x[x^2|y]] - E_y[E_x[x|y]^2] \quad (16) \]

Also consider the variance of the function \( E_x[x|y] \) w.r.t. to \( y \). Using the property in equation 14 we have,

\[ \text{var}_y[E_x[x|y]] = E_y[E_x[x|y]^2] - E_y[E_x[x|y]]^2 \quad (17) \]
Adding equations 16 and 17 gives us the RHS of the result to be proven. Therefore we get

$$E_y[\text{var}_x[x|y]] + \text{var}_y[E_x[x|y]] = E_y[\text{var}_x[x^2|y]] - E_y[E_x[x^2|y]]$$

$$= \sum_y p(y) \sum_x x^2 \cdot p(x|y) - (\sum_y p(y) \sum_x x \cdot p(x|y))^2$$

$$= \sum_x x^2 \sum_y p(x|y) \cdot p(y) - (\sum_x \sum_y p(x|y) \cdot p(y))^2$$

$$= \sum_x x^2 \sum_y p(x, y) - (\sum_x \sum_y p(x, y))^2$$ \hspace{1cm} \text{Product Rule}$$

$$= \sum_x x^2 \cdot p(x) - (\sum_x x \cdot p(x))^2$$ \hspace{1cm} \text{Marginalization}$$

$$= E[x^2] - E[x]^2$$

$$= \text{var}[x]$$ \hspace{1cm} = \text{LHS (proved)}$$