Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

1 (**Murphy 2.16**) Suppose $\theta \sim \text{Beta}(a, b)$ such that

$$\mathbb{P}(\theta; a, b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the Beta function and $\Gamma(x)$ is the Gamma function. Derive the mean, mode, and variance of θ.

Mean *The mean value is the expected value;*

$$\begin{aligned}
\mathbb{E}(\theta) &= \int \theta \mathbb{P}(\theta; a, b) d\theta = \int \theta \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1} d\theta \\
&= \frac{1}{B(a,b)} \int \theta^{a}(1-\theta)^{b-1} d\theta \\
&= \frac{B(a+1,b)}{B(a,b)} \\
&= \left[\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right] \\
&= \left[\frac{a\Gamma(a)\Gamma(b)}{a+b\Gamma(a+b)}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right] \\
&= \boxed{\frac{a}{a+b}}
\end{aligned}$$

Variance *We know that this is* $\mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2$*;*

$$\begin{aligned}
\mathbb{E}[\theta^2] - \mathbb{E}[\theta]^2 &= \int \theta^2 \mathbb{P}(\theta; a, b) d\theta - \frac{a^2}{(a+b)^2} \\
&= \frac{1}{B(a,b)} \int \theta^{a+1}(1-\theta)^{b-1} d\theta - \frac{a^2}{(a+b)^2} \\
&= \frac{B(a+2,b)}{B(a,b)} - \frac{a^2}{(a+b)^2} \\
&= \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} - \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} - \frac{a^2}{(a+b)^2} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\
&= \frac{(a^2+a)(a+b) - a^2(a+b+1)}{(a+b+1)(a+b)^2} \\
&= \boxed{\frac{ab}{(a+b+1)(a+b)^2}}
\end{aligned}$$

Mode *If we visualize a distribution, the point at which there is the greatest number of values is the mode, it is also the only global extrema, as such, it is the only point where*$\nabla_\theta = 0$*;*

$$\begin{aligned}
0 &= \nabla_\theta \left(\frac{1}{B(a,b)} \theta^{a-1}(1-\theta)^{b-1} \right) \\
&= \nabla_\theta (\theta^{a-1}(1-\theta)^{b-1} \\
&= (a-1)\theta^{a-2}(1-\theta)^{b-1} - (b-1)\theta^{a-1}(1-\theta)^{b-2}
\end{aligned}$$

Therefore we can now solve for the specific value of θ *that makes this true;*

$$\begin{aligned}
(a-1)\theta^{a-2}(1-b)^{b-1} &= (b-1)\theta^{a-1}(1-\theta)^{b-2} \\
(a-1)(1-\theta) &= (b-1)\theta \\
\theta &= \frac{a-1}{a+b-2}
\end{aligned}$$

■

2 (**Murphy 9**) Show that the multinoulli distribution

$$\mathrm{Cat}(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{K} \mu_i^{x_i}$$

is in the exponential family and show that the generalized linear model corresponding to this distribution is the same as multinoulli logistic regression (softmax regression).

We are searching for something of the form $\mathbf{p}(\mathbf{y};\boldsymbol{\eta} = b(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^T T(\mathbf{y}) + a(\boldsymbol{\eta})\right)$, as we did in class I am going to force the multinoulli distribution $\prod_{i=1}^{K} \mu_i^{x_i}$ into its exponential form:

$$\begin{aligned}
\prod_{i=1}^{K} \mu_i^{x_i} &= \exp\left[\log\left(\prod_{i=1}^{K} \mu_i^{x_i}\right)\right] && (1) \\
&= \exp\left[\sum_{i=1}^{K} \log(\mu_i^{x_i})\right] && (2) \\
&= \exp\left[\sum_{i=1}^{K} x_i \log(\mu_i)\right] && (3)
\end{aligned}$$

Now lets note that since this is a multinoulli distribution;

$$\mu_k = 1 - \sum_{i=1}^{K-1} \mu_i \text{ and } x_k = 1 - \sum_{i=1}^{K-1} x_i$$

$$\begin{aligned}
&= \exp\left[\sum_{i=1}^{K-1} x_i \log(\mu_i) + \left(1 - \sum_{i=1}^{K-1} x_i\right) \log(\mu_k)\right] && (4) \\
&= \exp\left[\sum_{i=1}^{K-1} x_i \log\left(\frac{\mu_i}{\mu_k}\right) + \log(\mu_k)\right] && (5)
\end{aligned}$$

Lets continue by defining $\boldsymbol{\eta}$ as follows;

$$\boldsymbol{\eta} = \begin{pmatrix} \log\left(\frac{\mu_1}{\mu_k}\right) \\ \vdots \\ \log\left(\frac{\mu_{k-1}}{\mu_k}\right) \end{pmatrix}$$

It follows that we can define μ_k as $1 - \sum_{i=1}^{K-1} \mu_k \exp(\eta_i)$ or equivalently as $\frac{1}{1+\sum_{i=1}^{K-1} exp(\eta_i)}$ Thus picking up where we left off;

$$= \exp\left[\boldsymbol{\eta}^T \mathbf{x} + \log\left(\frac{1}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)}\right)\right] \qquad (6)$$

This is in exponential form with $b(y) = 1$, $T(y) = \mathbf{x}$ and $a(\boldsymbol{\eta}) = -\log(1 + \sum_{i=1}^{K-1} \exp(\eta_i))$

■