Name: _____

Intermediate Bayesian Modeling
Midterm Exam

Instructions: Attempt as many parts of as many problems as possible. Show enough work to convince me that you know what you are doing. Write your answer to each problem on the blank paper provided. Write on only one side of each page. Do well!

1. (40 pts.) Video sharing websites like YouTube are interested in how many views are received by videos on their platform. Let u be the number of views received by a particular video in a one-hour span. Assume that u is well-modeled by a $\text{Pois}(\theta)$ distribution. Further assume that a Bayesian statistician puts a $\text{Gamma}\left(1, {}^{1}/_{150}\right)$ prior on the rate parameter θ for this Poisson distribution.

   (a) *Find the marginal distribution for u based on this model and prior. (15 pts.)*

$$\begin{aligned} f(u) &= \int_{S(\theta)} f(u \mid \theta) p(\theta) d\theta \\ &= \int_{S(\theta)} \frac{\theta^u \exp(-\theta)}{u!} \times \frac{1}{150} \exp\left(-\frac{\theta}{150}\right) d\theta \\ &= \frac{1}{150 u!} \int_{S(\theta)} \theta^u \exp\left(-\frac{151}{150}\theta\right) d\theta \\ &= \frac{1}{150\,\Gamma(u+1)} \times \frac{150\,\Gamma(u+1)}{151\left(\frac{151}{150}\right)^u} \\ &\quad \times \int_{S(\theta)} \frac{\left(\frac{151}{150}\right)^{u+1}}{\Gamma(u+1)} \theta^u \exp\left(-\frac{151}{150}\theta\right) d\theta \\ &= \frac{150^u}{151^{u+1}} \end{aligned}$$

   (b) *Describe how this marginal distribution can be used to test whether seeing $u = 1280$ views in one hour is consistent with the given model and prior. (10 pts.)*

   Using Box's marginal p-value test, we can measure how unlikely we would be to see data like what was seen given our model and prior. Here, the marginal distribution is strictly decreasing in u, so the values of u which are less likely than $u = 1280$ are all values above 1280. Then we can calculate the probability up to u by summing the individual probabilities and,

$$p_B(u) = 1 - \frac{1}{151} \sum_{i=0}^{u} \left(\frac{150}{151}\right)^u.$$

(c) *Consider a collection of exponential random variables* $v_i \mid \theta \overset{\text{iid}}{\sim} \text{Exp}\left({}^{1}/_{\theta}\right)$*, which can be thought of as the waiting times (in hours) between new viewers watching the previously discussed video. Explain the conditions under which identical Bayesian inferences will be made for* θ *using either the* u *data or the* v_i *data. (15 pts.)*

Identical Bayesian inferences will be made about θ when the likelihood principle holds—that is, when the likelihood of θ under the two models are proportional to one another—and when the same prior is used for θ. (In other words, identical inferences will be made whenever the kernel of the posterior distribution is the same for both approaches.)

Assuming that we use the same prior for both approaches, we only need likelihood proportionality. Under the Poisson approach, we have $L(\theta) \propto \theta^u \exp(-\theta)$. Under the exponential approach, the joint likelihood for waiting times up to the u^{th} viewing would be

$$L(\theta \mid v_1, ..., v_u) \propto \theta^u \exp(-\theta \sum_{i=1}^{u} v_i).$$

Then we can see that we will make identical inferences—the likelihoods are proportional to one another—if the time it takes for us to reach the u^{th} view is exactly 1 hour, $\sum_{i=1}^{u} v_i = 1$, and we don't consider data past those u views.

2. (40 pts.) Recall that in the development of the Deviance Information Criterion, the penalization term p_D is given by

$$p_D = \text{E}_{\theta|y}[-2\ell(\theta \mid y)] + 2\ell(\hat{\theta} \mid y),$$

where $\ell(\theta \mid y)$ is the log-likelihood for some parameter θ given observed data y, and where $\hat{\theta}$ is an estimator for θ.

(a) *Prove that if* $\hat{\theta}$ *is the posterior mean for* θ*,* p_D *must be non-negative whenever the likelihood is log-concave (*$\frac{d^2}{d\theta^2}\ell(\theta \mid y) < 0$ *for all* θ*). (25 pts.)*

According to Jensen's inequality, if a function $f(u)$ is concave, then $\text{E}_u[f(u)] \leq f\left(\text{E}_u[u]\right)$. Because the likelihood function is log-concave, this inequality holds for ℓ—that is, $\text{E}_{\theta|y}[\ell(\theta \mid y)] \leq \ell(\text{E}_{\theta|y}[\theta] \mid y)$. If we multiply by a constant -2, this will reverse the inequality, giving us

$$\begin{aligned} \text{E}_{\theta|y}[-2\ell(\theta \mid y)] &\geq -2\ell(\text{E}_{\theta|y}[\theta] \mid y) \\ \text{E}_{\theta|y}[-2\ell(\theta \mid y)] + 2\ell(\text{E}_{\theta|y}[\theta] \mid y) &\geq 0. \end{aligned}$$

This completes the proof.

(b) *In certain circumstances and for certain choices of* $\hat{\theta}$*, the quantity* p_D *may be negative. This would be inconvenient, since* p_D *is a penalty term. If* $\hat{\theta}$ *is chosen to be the posterior median, what additional constraint or constraints—beyond log-concavity—could be imposed to guarantee that* p_D *is positive? (You can explain this in words; you don't need*

to show it mathematically.) (15 pts.)

The simplest additional constraint is to require that the the posterior be symmetric. If the posterior were symmetric, the mean and median of the posterior would be the same value, and the argument above would hold for the median as well. Other constraints may also be possible, but are more difficult to show if the Jensen's inequality argument isn't preserved.

3. *(40 pts.) Let X_i, $i \in \mathcal{N}$ be an infinite exchangeable sequence of random quantities and define Y_h and Y_k to be the averages of h and of k random quantities from among the X_i's. Assume De Finetti's law of large numbers applies here—that is, for any ε and θ, there exist H and K such that*

$$h \geq H, k \geq K \longrightarrow \Pr\left[|Y_h - Y_k| > \varepsilon\right] < \theta.$$

If $\Phi_n(\xi) = \Pr\left[Y_n \leq \xi\right]$, prove that $\lim_{n\to\infty} \Phi_n(\xi)$ exists.

By the law of total probability,

$$\Pr\left[Y_n \leq \xi\right] = \Pr\left[Y_n \leq \xi \cap Y_h > \xi + \varepsilon\right] + \Pr\left[Y_n \leq \xi \cap Y_h \leq \xi + \varepsilon\right].$$

Observe that

$$\begin{aligned}
\Pr\left[Y_n \leq \xi \cap Y_h > \xi + \varepsilon\right] &= \Pr\left[Y_n \leq \xi < Y_h - \varepsilon\right] \\
&\leq \Pr\left[Y_n < Y_h - \varepsilon\right] \\
&= \Pr\left[\varepsilon < Y_h - Y_n\right] \\
&\leq \Pr\left[\varepsilon < |Y_h - Y_n|\right],
\end{aligned}$$

with the final inequality holding because $Y_h - Y_n \leq |Y_h - Y_n|$ irrespective of Y_h and Y_n.

We note that this is the form of De Finetti's law of large numbers, which we've assumed to hold here. Therefore, for any $\varepsilon, \theta > 0$, $\exists\, N_{\varepsilon,\theta}, H_{\varepsilon,\theta} \in \mathcal{N}$ such that $\forall\, n \geq N_{\varepsilon,\theta}, h \geq H_{\varepsilon,\theta}$, $\Pr\left[\varepsilon < |Y_h - Y_n|\right] < \theta$. Combining this result with the above development, we obtain the following result:

For any choice of $\varepsilon, \theta > 0$, there exist $N_{\varepsilon,\theta}, H_{\varepsilon,\theta} \in \mathcal{N}$ such that $n \geq N_{\varepsilon,\theta}, h \geq H_{\varepsilon,\theta}$ implies $\Pr\left[Y_n \leq \xi \cap Y_h > \xi + \varepsilon\right] < \theta$.

Now considering the second term of the total probability sum, by the definition of joint probability we have

$$\begin{aligned}
\Pr\left[Y_n \leq \xi \cap Y_h \leq \xi + \varepsilon\right] &= \Pr\left[Y_n \leq \xi \mid Y_h \leq \xi + \varepsilon\right] \times \Pr\left[Y_h \leq \xi + \varepsilon\right] \\
&\leq \Pr\left[Y_h \leq \xi + \varepsilon\right].
\end{aligned}$$

Then using the same $N_{\varepsilon,\theta}, H_{\varepsilon,\theta}$ above, for any $\varepsilon, \theta > 0$ we have

$$\Pr\left[Y_n \leq \xi\right] < \theta + \Pr\left[Y_h \leq \xi + \varepsilon\right].$$

Now consider the quantity $\Pr[Y_h \leq \xi - \varepsilon] - \theta$. Using the same law of total probability trick, we can see that

$$\Pr[Y_h \leq \xi - \varepsilon] - \theta = \Pr[Y_h \leq \xi - \varepsilon \cap Y_n > \xi] + \Pr[Y_h \leq \xi - \varepsilon \cap Y_n \leq \xi] - \theta.$$

Taking the first term, we can see that

$$\begin{aligned}\Pr[Y_h \leq \xi - \varepsilon \cap Y_n > \xi] &= \Pr[Y_h + \varepsilon \leq \xi < Y_n] \\ &\leq \Pr[Y_h + \varepsilon < Y_n] \\ &= \Pr[\varepsilon < Y_n - Y_h] \\ &\leq \Pr[\varepsilon < |Y_n - Y_h|],\end{aligned}$$

so once again we have $\Pr[Y_h \leq \xi - \varepsilon \cap Y_n > \xi] < \theta$ under the same $N_{\varepsilon,\theta}, H_{\varepsilon,\theta}$ as before.

The second term here undergoes a similar development:

$$\begin{aligned}\Pr[Y_h \leq \xi - \varepsilon \cap Y_n \leq \xi] &= \Pr[Y_h \leq \xi - \varepsilon \mid Y_n \leq \xi] \times \Pr[Y_n \leq \xi] \\ &\leq \Pr[Y_n \leq \xi].\end{aligned}$$

So we have

$$\begin{aligned}\Pr[Y_h \leq \xi - \varepsilon] - \theta &< \theta + \Pr[Y_n \leq \xi] - \theta \\ &= \Pr[Y_n \leq \xi].\end{aligned}$$

Taken together, these two developments give us the following set of inequalities:

$$\Pr[Y_h \leq \xi - \varepsilon] - \theta < \Pr[Y_n \leq \xi] < \Pr[Y_h \leq \xi + \varepsilon] + \theta,$$

which we can express using simpler notation as

$$\Phi_h(\xi - \varepsilon) - \theta < \Phi_n(\xi) < \Phi_h(\xi + \varepsilon) + \theta.$$

At this point, it is good to revisit what we need to prove. We want to show that $\lim_{n\to\infty} \Phi_n(\xi)$ exists. Since $\Phi_n(\xi) \in \mathcal{R}$, it is sufficient to show that the sequence $\Phi_n(\xi)$ is Cauchy—that is, there exist $H_\eta, K_\eta \in \mathcal{N}$ such that for all $h \geq H_\eta, k \geq K_\eta$, $|\Phi_h(\xi) - \Phi_k(\xi)| < \eta$. This is very nearly what we have above, except that we have θ instead of η and our $\Phi_h(\cdot)$ statements include ε which, we know from De Finetti's law of large numbers, can be arbitrarily small.

Observe that the Φ functions are nondecreasing. Then

$$\Phi_h(\xi - \varepsilon) \leq \Phi_h(\xi) \leq \Phi_h(\xi + \varepsilon),$$

and it follows that

$$\Phi_h(\xi - \varepsilon) - \theta \leq \Phi_h(\xi) - \theta < \Phi_h(\xi) + \theta \leq \Phi_h(\xi + \varepsilon) + \theta.$$

Then if $\Phi_n(\xi) - \Phi_h(\xi - \varepsilon) < \theta$ and $\Phi_h(\xi + \varepsilon) - \Phi_n(\xi) < \theta$, it necessarily follows that $\Phi_n(\xi) - \Phi_h(\xi) < \theta$ and $\Phi_h(\xi) - \Phi_n(\xi) < \theta$. This is identical to saying $|\Phi_n(\xi) - \Phi_h(\xi)| < \theta$, so for our above choice of $N_{\varepsilon,\theta}, H_{\varepsilon,\theta}$, we have established that this sequence is Cauchy. Since it is a Cauchy sequence in $\mathcal{R}$, it converges, completing the proof.□

4. (40 pts.) A common tool for thinking about discrete probabilities is the Pólya urn scheme. We imagine an urn containing α red balls and β blue balls. Balls are drawn from the urn and observed. This is analogous to sampling from a population with two options—successes or failures on a Bernoulli trial, say. Different distributions can be modeled in this way based on whether or how balls are replaced in the urn after being drawn.

   Imagine you are at the decennial Statistics Carnival, playing a game. In this game, a carnival worker has you draw a ball from an urn. Then you return the ball to the carnival worker. The carnival worker then takes one of three actions (always the same after each draw):

   (1) The carnival worker does not return the ball to the urn.
   (2) The carnival worker returns the ball to the urn.
   (3) The carnival worker returns the ball to the urn, adding another ball of the same color.

   You will draw 10 balls, and then you will guess which action the carnival worker has been taking. The game begins with one red ball and one blue ball in the urn.

   (a) *The carnival worker indicates that the three replacement actions are equally probable. Explain why the carnival worker is obviously lying. (5 pts.)*

   It is impossible to draw 10 balls if the carnival worker chooses Action 1. Either the game will fail to be playable after two draws, or the carnival worker self-evidently did not choose Action 1. (Technically, it is possible for the carnival worker to choose each of the three actions equiprobably, so it may be unfair to say that they are lying. It is fair to say that Action 1 has zero probability in the following discussion of the problem, however, since it would be deterministically either verified or rejected by the successful action of picking a third ball.)

   (b) *What probability distribution is followed by Y, the total number of red balls seen in 10 draws, if the carnival worker takes Action 2. (10 pts.)*

   If the carnival worker takes Action 2, then the same number of balls of each type will be in the urn after each draw. This means each draw can be thought of as an independent trial with probability $\frac{\alpha}{\alpha+\beta}$ of a success. Here, since $\alpha = \beta = 1$, Y follows a $\text{Bin}\left(10, {}^1\!/_2\right)$ distribution.

   (c) *Action 3 gives rise to a beta-binomial distribution with parameters $n = 10$, $\alpha = 1$, and $\beta = 1$. This is the same as the marginal distribution for Y if Y were distributed* $\text{Bin}(10, p)$ *with a* $\text{Beta}(1, 1)$ *prior on the probability of a red ball, p. Find the probability mass function for this beta-binomial distribution. (10 pts.)*

This is just another marginal distribution calculation:

$$\begin{aligned}
f(y) &= \int_{S(p)} \binom{10}{y} p^y (1-p)^{10-y} \times \frac{\Gamma(1+1)}{\Gamma(1)\,\Gamma(1)} p^{1-1}(1-p)^{1-1} dp \\
&= \binom{10}{y} \int_{S(p)} p^y (1-p)^{10-y} dp \\
&= \binom{10}{y} \frac{\Gamma(y+1)\,\Gamma(11-y)}{\Gamma(12)} \int_{S(p)} \frac{\Gamma(12)}{\Gamma(y+1)\,\Gamma(11-y)} p^y (1-p)^{10-y} dp \\
&= \frac{10!}{11!} \times \frac{y!}{y!} \times \frac{(10-y)!}{(10-y)!} \\
&= \frac{1}{11}\, I_{\{0,\ldots,10\}}(y)
\end{aligned}$$

I've left the support for y off until the last line for notational simplicity, but it is good to reintroduce it there. Note that this means y is distributed Discrete Uniform over the values 0 to 10.

(d) *Assume that you've seen $Y = 8$ red balls in 10 draws. Calculate the Bayes factor comparing Model A—the carnival worker replaces balls according to Action 2, and Model B—the carnival worker replaces balls according to Action 3. (15 pts.)*

Recall that the Bayes factor is given by

$$BF_{A,B} = \frac{\Pr[y \mid A]}{\Pr[y \mid B]} = \frac{\int_{S(p)} f(y \mid p, A) p(p \mid A) dp}{\int_{S(p)} f(y \mid p, B) p(p \mid B) dp}.$$

Under Action 2, we know that the data follow a $\text{Bin}\left(10, {}^1\!/_2\right)$ distribution—in other words, we have a prior giving point mass 1 to $p = 0.5$. The probability of seeing $y = 8$ under Action 2, then, is $f(8 \mid A) = \binom{10}{8}/2^{10} = 0.044$. Under Action 3, we know that the marginal distribution for y is discrete uniform and $f(8 \mid B) = \frac{1}{11} = 0.091$. Then the Bayes factor comparing Model A to Model B is $BF_{A,B} = \frac{0.044}{0.091} = 0.483$, meaning that these data are about half as likely to occur if the carnival worker picks Action 2 than if the carnival worker picks Action 3.

(e) *EXTRA CREDIT: Pretend $\alpha = \beta = 10$. What probability distribution would be followed by Y if the carnival worker took Action 1?*

This would be a sampling-without-replacement scheme from a fixed population, which is how we define the hypergeometric distribution.