General Notes

- You will submit a minimum of two files, the core files must conform to the following naming conventions (including capitalization and underscores). 123456789 is a placeholder, please replace these nine digits with your nine-digit Bruin ID. The files you must submit are:

  1. *123456789_stats102c_hw5.Rmd*: Your markdown file which generates the output file of your submission.
  2. *123456789_stats102c_hw5.html/pdf*: Your output file, either a PDF or an HTML file depending on the output you choose to generate.

  If you fail to submit any of the required core files you will receive ZERO points for the assignment. If you submit any files which do not conform to the specified naming convention, you will receive (at most) half credit for the assignment.

- **Your .Rmd file must knit**. If your .Rmd file does not knit you will receive (at most) half credit for the assignment.
  The two most common reason files fail to knit are because of workspace/directory structure issues and missing include files. To remedy the first, ensure all of the file paths in your document are relative paths pointing at the current working directory. To remedy the second, simply make sure you upload any and all files you source or include in your .Rmd file.

- Your coding should adhere to the tidyverse style guide: https://style.tidyverse.org/.

- Any functions/classes you write should have the corresponding comments as the following format.

```
my_function <- function(x, y, ...){
#A short description of the function
#Args:
#x: Variable type and dimension
#y: Variable type and dimension
#Return:
#Variable type and dimension
Your codes begin here
}
```

NOTE: *Everything* you need to do this assignment is here, in your class notes, or was covered in discussion or lecture.

- Please **DO NOT** look for solutions online.
- Please **DO NOT** collaborate with anyone inside (or outside) of this class.
- Please work **INDEPENDENTLY** on this assignment.
- **EVERYTHING** you submit **MUST** be 100% your, original, work. Any student suspected of plagiarizing, in whole or in part, any portion of this assignment, will be **immediately** referred to the Dean of Student's office without warning.

Problem 1: The weather on any given day on a tropical island could be rainy, sunny or cloudy, and the probability of tomorrow's weather only depends on today's weather and not any other previous days. Suppose that we obtained the transition probabilities as below.

P(Rainy tomorrow | Rainy today) = 0.5

P(Rainy tomorrow | Cloudy today) = 0.5

P(Sunny tomorrow | Cloudy today) = 0.5

P(Rainy tomorrow | Sunny today) = 0.25

P(Cloudy tomorrow | Rain today) = 0.25

P(Cloudy tomorrow | Sunny today) = 0.25

(a) Please find the transition matrix and draw a transition state diagram, with state space {1: Rainy, 2: Cloudy, 3: Sunny}.

(b) Suppose today is sunny. What is the expected weather two days from now?

(c) Is this Markov chain irreducible? Explain. Can you find a stationary distribution?

Problem 2: Consider a Markov chain with the state space S = 0, 1, 2, 3, 4, 5 and transition matrix

$$\begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

, where $\alpha_i > 0$ and $\sum \alpha_i = 1$. Determine, in the long run, the probability of being in stat 0. Does it depend on the initial stat X_0?

Problem 3: Suppose the Markov chain is defined by $X^{(t+1)} = \alpha X^{(t)} + \epsilon_t$, where $\epsilon_t \sim N(0, 1)$. Write R code to simulate this Markov chain with $X^{(0)} \sim N(0, 1)$ for $t \leq 10^4$ and $\alpha = 0.9$. Check if your sample fits the distribution $N(0, 1/(1 - \alpha^2))$.

Problem 4: We want to model the number of siblings people in a certain population have. We can model the number of siblings a person has as a Poisson random variable Y for some unknown mean parameter λ. The probability mass function of $Y \sim Pois(\lambda)$ is

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad \text{for } y = 0, 1, 2, \ldots.$$

Suppose, before observing any data, we model our prior beliefs about λ by a gamma distribution Gamma(α, β) with hyperparameters $\alpha, \beta > 0$, i.e,

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \quad \text{for } \lambda > 0,$$

where $\Gamma(z) = \displaystyle\int_0^\infty x^{z-1}e^{-x}dx$. Note that

- The prior mean is $E(\lambda) = \dfrac{\alpha}{\beta}$.
- The prior variance is $Var(\lambda) = \dfrac{\alpha}{\beta^2}$.
- The prior mode is $\text{mode}(\lambda) = \begin{cases} \dfrac{\alpha - 1}{\beta} & \text{if } \alpha > 1, \\ 0 & \text{if } \alpha \leq 1. \end{cases}$

Suppose we observe data $y_1, y_2, \ldots, y_n \overset{iid}{\sim} Pois(\lambda)$. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$.

(a) Show that the gamma distribution is a conjugate prior for the Poisson likelihood. More specifically, show that the posterior distribution $\pi(\lambda|y)$ is of the form

$$\pi(\lambda|y) \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i, \beta + n\right).$$

(b) Show that the posterior mean $E(\lambda|y)$ is a weighted average of the prior mean and the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. That is, find the weight w so that

$$E(\lambda|y) = w\frac{\alpha}{\beta} + (1 - w)\bar{y}.$$

(c) What is $\lim_{n\to\infty} E(\lambda|y)$? What does this limit represent?

(d) Use the `dgamma()` function in R to visualize the prior and posterior densities for hyperparameters $\alpha = 8, \beta = 4$ when the data is summarized by:

 (i) $\sum_{i=1}^{n} y_i = 20$ and $n = 5$

 (ii) $\sum_{i=1}^{n} y_i = 80$ and $n = 20$

 Represent the sample mean on the plot to show how the posterior distribution is a compromise between the prior distribution and the data. Clearly indicate the different components of your plot.

 Hint: For a Gamma(α, β) distribution, α is the `shape` parameter and β is the `rate` parameter.

(e) For scenarios (i) and (ii) in part (d), find and interpret 95% quantile-based credible intervals.

Problem 5: The Cauchy distribution is a continuous probability distribution. It is often used in statistics as the canonical example of a "pathological" distribution, and the density is defined as:

$$f(x|\gamma, \eta) = \frac{1}{\gamma\pi[1 + (\frac{x-\eta}{\gamma})]}, \ -\infty < x < \infty, \ \gamma > 0,$$

where η is the location parameter, specifying the location of the peak of the distribution.

(a) Please write an algorithm using the Metropolis-Hastings sampler to generate samples from the Cauchy distribution with $\gamma = 1$ and $\eta = 0$. You may use $N(\mu, \sigma)$ as the proposal distribution.

(b) Implement your algorithm in R to generate 10,000 samples and show the histogram.

(c) Compare the generated samples with `qcauchy` in R. Please conduct a exploratory analysis to determine if your samples agree with the output of `qcauchy`.

Problem 6: For the example of the Bayesian prediction application (8.7) discussed in class, suppose data from a test on a second batch of parts is collected. Here 25 parts were tested for 24 hours. Five failures were observed at times 6.1, 7.3, 7.4, 14.8, 18.1, and the other 20 parts were not observed to fail during the test.

(a) Please derive the posterior distribution with the new data, and design a Metropolis algorithm using only the gamma proposal distribution.

(b) Implement the algorithm, and find the shape parameter such that the chain is mixing well.