MSiA 400 Lab Assignment 1

Due Oct 13 at 5pm

Please submit a report file that includes: a short answer, related code, printouts, etc. for each problem (where necessary). Push your answers to Github. All programming must be in R (or R Markdown).

Problem 1

You will analyze data from a website with 8 pages (plus a 9th state, indicating that the user has left the website). Formulate a Markov chain for this website where each state $\{S_i \mid i = 1, \cdots, 9\}$ corresponds to a page. Each visitor starts at the home page (Page 1), then browses from page-to-page until he/she leaves the website. So, a sample path may be $S_1 \to S_3 \to S_5 \to S_9$, corresponding to a visitor starting on the home page, moving to Page 3, then Page 5, then leaving the website.

Attached is the dataset `webtraffic.txt`, which records the paths of 1000 visitors (rows). The data has 81 columns labeled $t11, t12, \cdots, t19, t21, t22, \cdots, t99$, where t_{ij} represents a transition from State i to State j, for $i, j \in \{1, \cdots, 9\}$. Each visitor has a 1 in column t_{ij} if the visitor clicked from Page i to Page j, and 0 elsewhere. For example, the aforementioned sample path would have 1's in columns t_{13}, t_{35}, and t_{59} and 0's elsewhere.

Problem 1a

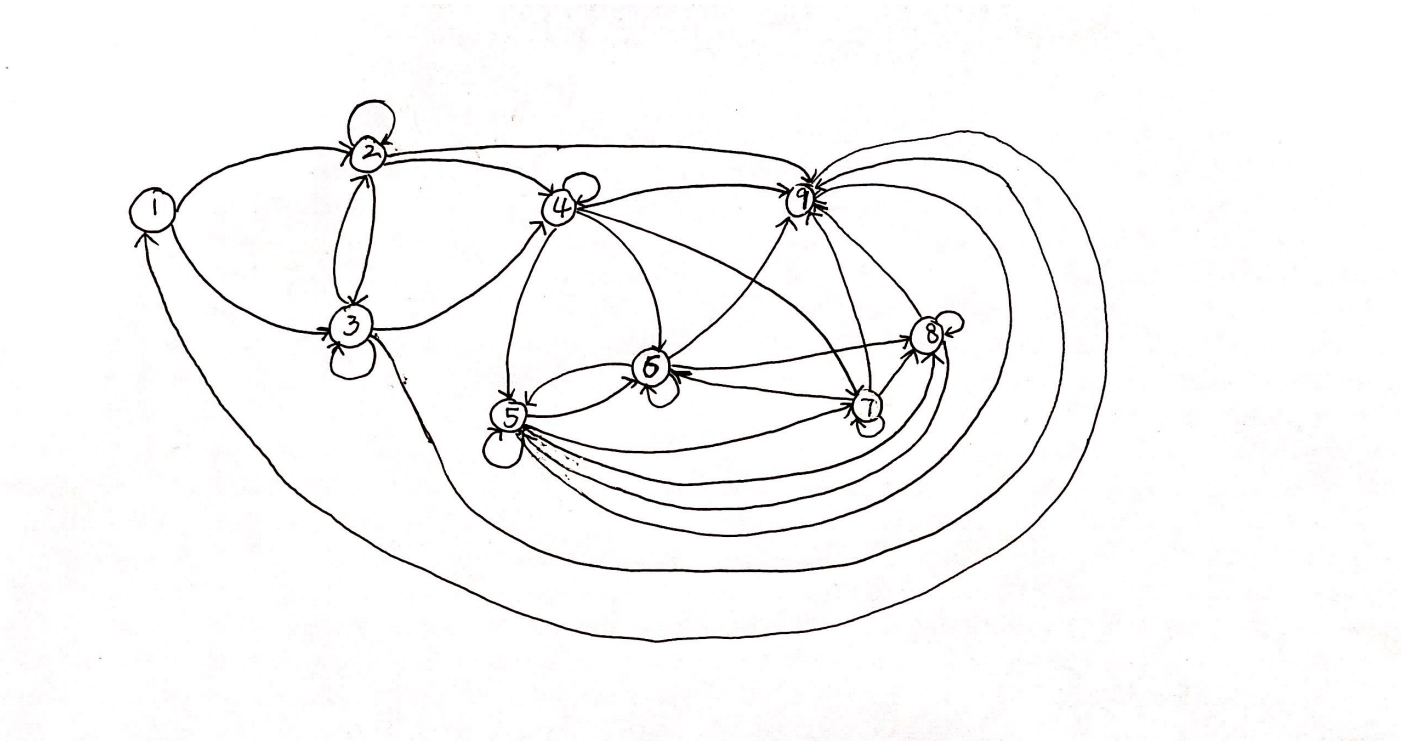
Construct a 9 by 9 matrix `Traffic` that counts total traffic from State i to State j, for $i, j \in \{1, \cdots, 9\}$. Note that `Traffic` has 0's in row 9 and column 1. Set `Traffic[9,1]=1000`. (This is equivalent to making each user return to the home page after they leave the website.) Display `Traffic`. `colSums()` adds all rows for each column.

```
data <- read.delim("webtraffic.txt")
col_total <- colSums(data)
Traffic <- matrix(col_total, nrow = 9, ncol = 9, byrow = TRUE)
Traffic[9, 1] = 1000
Traffic
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
##  [1,]    0  447  553    0    0    0    0    0    0
##  [2,]    0   23  230  321    0    0    0    0   63
##  [3,]    0  167   43  520    0    0    0    0   96
##  [4,]    0    0    0   44  158  312  247    0  124
##  [5,]    0    0    0    0   22   52   90  127  218
##  [6,]    0    0    0    0   67   21    0  294   97
##  [7,]    0    0    0    0    0   94    7  185   58
##  [8,]    0    0    0    0  262    0    0   30  344
##  [9,] 1000    0    0    0    0    0    0    0    0
```

Problem 1b

Draw a directed graph where each node represents a state, and each arrow from State i to State j has positive (non-zero) traffic (i.e., `Traffic[i,j]>0`). This may be submitted as a TikZ graph (or using your graphing program of choice) or a picture of a hand-drawn graph (provided it is legible). Is the Markov chain irreducible? Is the Markov chain ergodic? Explain.



This Markov chain is irreducible because all states communicate with each other. This Marcov chain is ergodic because it is recurrent and aperiodic.

Problem 1c

Construct and display the one-step transition probability matrix `P` (using the Maximum Likelihood estimate, i.e., $p_{ij} = \frac{\text{Traffic}[i,j]}{\sum_{j=1}^{9} \text{Traffic}[i,j]}$).

```
row_total <- rowSums(Traffic)
P <- Traffic / row_total
P
```

```
##       [,1]       [,2]       [,3]       [,4]      [,5]       [,6]       [,7]
##  [1,]    0 0.44700000 0.55300000 0.00000000 0.0000000 0.00000000 0.00000000
##  [2,]    0 0.03610675 0.36106750 0.50392465 0.0000000 0.00000000 0.00000000
##  [3,]    0 0.20217918 0.05205811 0.62953995 0.0000000 0.00000000 0.00000000
##  [4,]    0 0.00000000 0.00000000 0.04971751 0.1785311 0.35254237 0.27909605
##  [5,]    0 0.00000000 0.00000000 0.00000000 0.0432220 0.10216110 0.17681729
##  [6,]    0 0.00000000 0.00000000 0.00000000 0.1398747 0.04384134 0.00000000
##  [7,]    0 0.00000000 0.00000000 0.00000000 0.0000000 0.27325581 0.02034884
##  [8,]    0 0.00000000 0.00000000 0.00000000 0.4119497 0.00000000 0.00000000
##  [9,]    1 0.00000000 0.00000000 0.00000000 0.0000000 0.00000000 0.00000000
##             [,8]      [,9]
##  [1,] 0.00000000 0.0000000
##  [2,] 0.00000000 0.0989011
##  [3,] 0.00000000 0.1162228
##  [4,] 0.00000000 0.1401130
##  [5,] 0.24950884 0.4282908
##  [6,] 0.61377871 0.2025052
##  [7,] 0.53779070 0.1686047
##  [8,] 0.04716981 0.5408805
##  [9,] 0.00000000 0.0000000
```

Problem 1d

What is the probability of a visitor being on Page 5 after 5 clicks?

```
a <- c(1, rep(0, 8))
prob5 <- a %*% P %*% P %*% P %*% P %*% P
prob5[5]
```

```
## [1] 0.1315178
```

The probability of a vistor being on Page 5 after 5 clicks is 0.1315178.

Problem 1e

Compute and display the steady-state probability vector `Pi`, solving the system of equations (as demonstrated in lab).

```
Q <- t(P) - diag(9)
Q[9, ] <- rep(1, 9)
rhs <- c(rep(0, 8), 1)
Pi <- solve(Q, rhs)
Pi
```

```
## [1] 0.15832806 0.10085497 0.13077897 0.14012033 0.08058898 0.07583914 0.05446485
## [8] 0.10069664 0.15832806
```

Problem 1f

The following table represents the average time (in minutes) that a visitor spends on each page:

Page	1	2	3	4	5	6	7	8
Min	0.1	2	3	5	5	3	3	2

What is the average time a visitor spends on the website (until he/she first leaves)? Modify the mean first passage time equations, with time spent at each state.

```
B <- P[1:8, 1:8]
Q <- diag(8) - B
rhs <- c(0.1, 2, 3, 5, 5, 3, 3, 2)
m <- solve(Q, rhs)
m[1]
```

```
## [1] 14.563
```

The average time a visotor spends on the website is 14.563

Problem 2

Use Monte Carlo integration to estimate the integral $\int_0^\infty e^{-\lambda x} \sin x dx$ for $\lambda > 0$. Use the exponential distribution $p(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, which has variance $\text{var}\left[p(x)\right] = \frac{1}{\lambda^2}$. Note, here $g(x) = \frac{\sin x}{\lambda}$. To generate random variables from the exponential distribution, you may first draw $X \sim \text{unif}(0, 1)$, then let $Y = -\frac{\ln X}{\lambda}$.

Problem 2a

Determine the number of samples required to achieve an error tolerance of 10^{-3} with 99% confidence.

$$n \geq \frac{\frac{1}{\lambda^2}}{(10^{-3})^2 0.01}$$

$$n \geq \frac{10^7}{\lambda^2}$$

Problem 2b

Compute the approximation (using the number of samples obtained in Problem 2a) and verify that it is within tolerance by comparing to the exact solution: $\int_0^\infty e^{-\lambda x} \sin x dx = \frac{1}{1+\lambda^2}$. Numerically evaluate for each of $\lambda = 1, 2, 4$.

```
set.seed(400)
n1 <- 10000000 / (1^2)
n2 <- 10000000 / (2^2)
n3 <- 10000000 / (4^2)
X1 <- runif(n1, 0, 1)
X2 <- runif(n2, 0, 1)
X3 <- runif(n3, 0, 1)
Y1 <- -log(X1) / 1
Y2 <- -log(X2) / 2
Y3 <- -log(X3) / 4
I1 <- sum(sin(Y1)) / n1
I2 <- sum(sin(Y2) / 2) / n2
I3 <- sum(sin(Y3) / 4) / n3
I1
```

```
## [1] 0.499993
```

```
I2
```

```
## [1] 0.199819
```

```
I3
```

```
## [1] 0.05892617
```

Problem 3

Obtain draws from the gamma distribution $p(x) = \frac{x^{k-1}}{\Gamma(k)\theta^k}\exp\left(-\frac{x}{\theta}\right)$ using MCMC. Use the exponential distribution $q(x|\lambda) = \lambda e^{-\lambda x}$ as q, with your previous iterate as λ.

Problem 3a

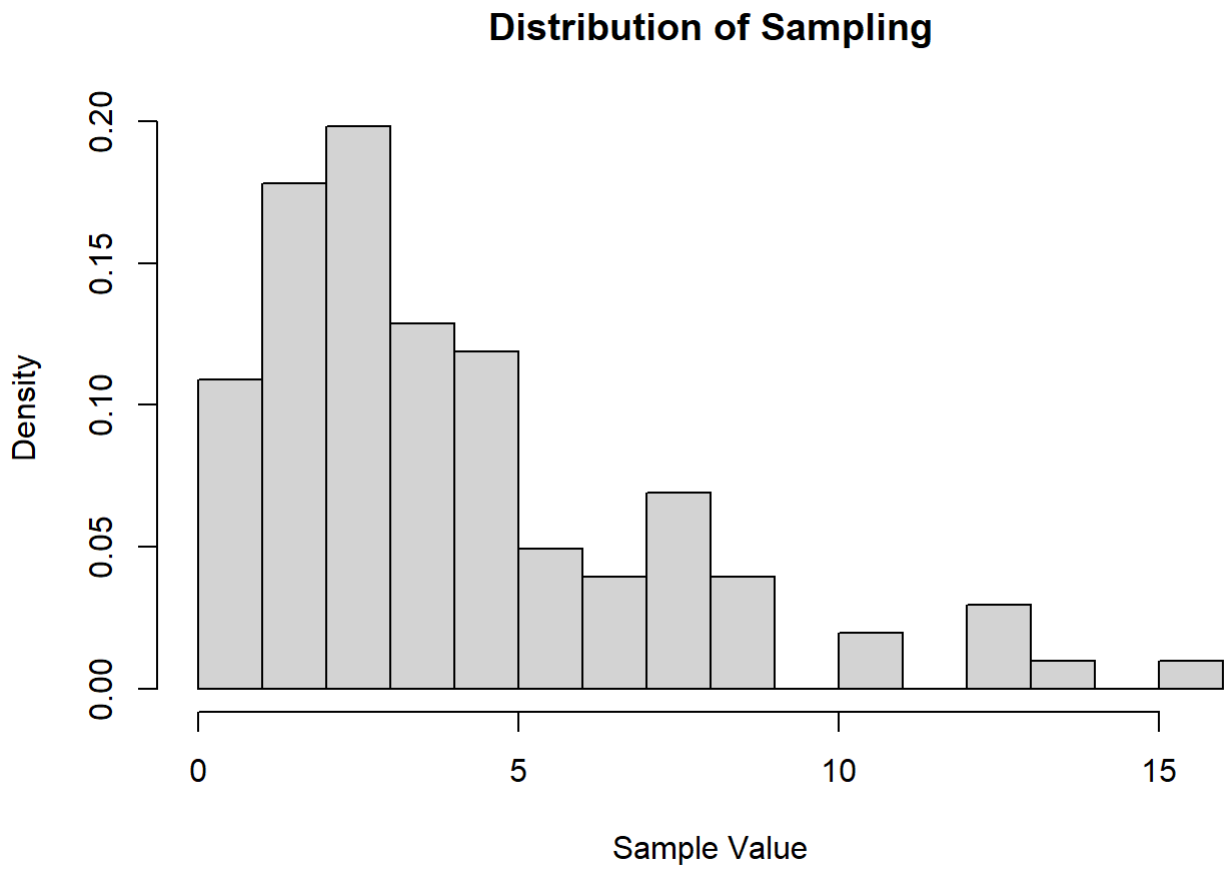
Which MCMC algorithm (Metropolis, Metropolis-Hastings, or Gibbs) is better suited for this problem?

Metropolis-Hastings

Problem 3b

Using a burn-in period of 5000 samples and keeping every 100 samples, generate 100 samples from the gamma distribution with shape $k = 2$ and scale $\theta = 2$. Use the algorithm you chose in Problem 3a and write your own sampler (as opposed to using a function from a package).

```
set.seed(99999)
q <- function(x, lambda){
  return (lambda * exp(-lambda * x))
}
f <- function(x, k = 2, theta = 2){
  return (x ^ (k - 1) * exp(-x / theta))
}
x_list = c(1, rep(0, 14999))
for(t in 0 : 14999){
  if(t == 0){
    x = 1
  }else{
    x = -log(runif(1, 0, 1)) / x_list[t+1]
  }
  a = f(x) * q(x_list[t+1], x) / (f(x_list[t+1]) * q(x, x_list[t+1]))
  u = runif(1, 0, 1)
  if(u <= a){
    x_list[t+2] = x
  }else{
    x_list[t+2] = x_list[t+1]
  }
}
hist(x_list[seq(5000, 15000, 100)], freq = FALSE, breaks = 20, main = "Distribution of Sampling"
, xlab = "Sample Value")
```



Problem 3c

Are the samples generated in Problem 3b sufficiently random? How can you tell?

The samples generated are not sufficiently random, because its distribution is not perfectly following Gamma distribution.