

# MSiA 400 Lab Assignment 1

Due Oct 24 at 11:59pm

Instructions: Please submit a report file that includes: a short answer, related code, printouts, etc. for each problem (where necessary). 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  $\{S_i \mid i = 1, \dots, 9\}$  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  $S_1 \rightarrow S_3 \rightarrow S_5 \rightarrow S_9$ , 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  $t_{11}, t_{12}, \dots, t_{19}, t_{21}, t_{22}, \dots, t_{99}$ , where  $t_{ij}$  represents a transition from State  $i$  to State  $j$ , for  $i, j \in \{1, \dots, 9\}$ . 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  $i, j \in \{1, \dots, 9\}$ . 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`. Hint: `colSums()` adds all rows for each column.

### 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.

### Problem 1c

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

### Problem 1d

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

### Problem 1e

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

### 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)? Hint: Modify the mean first passage time equations, with time spent at each state.

### 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  $\text{var}[p(x)] = \frac{1}{\lambda^2}$ . 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.

### 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  $\lambda = 1, 2, 4$ .

### 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  $\lambda$ .

### Problem 3a

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

### 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).

### Problem 3c

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