CS203A: Assignment 2

Priya Gole (200727)

---

1. (15+15 marks) **Endsem allocation**

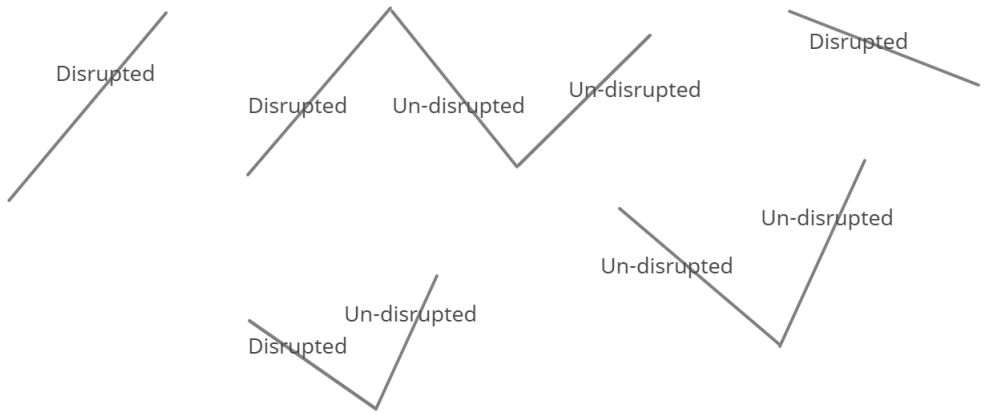
   You are allocated as the Tutor of CS203, with n students. Rajat has created 2 sets of Endsem papers to decrease cheating. He has asked you to help decide which paper should be given to whom. You scraped through the data on Hello, and found out who have been project partners in previous courses, as they will be friends now. Thus, you have found out m friendship connections among the students. You reported this to Rajat, and he said he is fine with any allocation that disrupts atleast half of the friendship connections. A friendship connection is disrupted if the students get different sets of papers.

   (a) You are really busy, and just randomly allocated each student to set 1 or set 2. Show that the expected value of disrupted friendship connections is $\frac{m}{2}$.

   (b) Getting expected value is not enough, you need to find a proper allocation. But you cannot go over all the 2^n allocations as $n \approx 150$. Using the construction for pairwise independence given in class, show that you can find an allocation with at least half of the friendship connections disrupted in poly(n)-time.

   Solution:

   (a) We have n students and m friendship connections.
   Representing m friendship connections in the graph, let there be t disconnected graphs as shown (the remaining nodes are not represented in graph):



   Suppose, k of the connections are disrupted (and $m - k$ are still intact).
   Now, label all the connections in the graph as shown in the figure above.

   For a disconnected graph (let the number of edges be e), the probability of that particular labels to take is $\left(\frac{1}{2}\right)^e$.
   Take the left most connection (one disrupted edge), the 2 students must get different sets, i.e., first student can get any set 1 or 2, say (s)he gets set 1 then the second student must get set 2 and the probability of that happening is $\left(\frac{1}{2}\right)$.
   For any disconnected graph, the first student can get any set 1 or 2, and to maintain the connectivity (connection disrupted or not) the set of all other students in that graph is fixed. The probability of getting a particular set is $\left(\frac{1}{2}\right)$ so,

the probability of getting the given connectivity $= \left(\frac{1}{2}\right)^{no.\ of\ students-1} = \left(\frac{1}{2}\right)^{e}$

For the entire graph (containing t disconnected graphs), the probability of a particular connectivity is,

$$\begin{aligned} P &= \left(\frac{1}{2}\right)^{e_1} \left(\frac{1}{2}\right)^{e_2} \cdots \left(\frac{1}{2}\right)^{e_t} \\ &= \left(\frac{1}{2}\right)^{e_1+e_2+\cdots+e_t} \\ &= \left(\frac{1}{2}\right)^{m} \end{aligned}$$

Hence, probability of each combination is $\left(\frac{1}{2}\right)^{m}$ (constant).

Let X be a random variable denoting the number of disrupted connections. So, X takes up values $0 \leq X \leq m$.

Suppose, k of the connections are disrupted (and $m-k$ are still intact), number of ways of choosing these k connections from m connections is $\binom{m}{k}$.
So, the probability of $X = k$ is $\binom{m}{k}\left(\frac{1}{2}\right)^{m}$

Hence, expected value of disrupted friendship is

$$\begin{aligned} E[X] &= \sum_{k=0}^{m} k\binom{m}{k}\left(\frac{1}{2}\right)^{m} \\ &= \left(\frac{1}{2}\right)^{m} \sum_{k=0}^{m} k\binom{m}{k} \qquad \textit{equation 1} \end{aligned}$$

We know,

$$\begin{aligned} (1+a)^m &= \sum_{k=0}^{m} \binom{m}{k} a^k \qquad \textit{differentiating this wrt a we get} \\ m(1+a)^{m-1} &= \sum_{k=0}^{m} k\binom{m}{k} a^{k-1} \qquad \textit{put a = 1} \\ m \cdot (2)^{m-1} &= \sum_{k=0}^{m} k\binom{m}{k} \end{aligned}$$

Hence, putting this in equation 1 we get,

$$\begin{aligned} E[X] &= \left(\frac{1}{2}\right)^{m} \sum_{k=0}^{m} k\binom{m}{k} \\ &= \left(\frac{1}{2}\right)^{m} \cdot m \cdot (2)^{m-1} \\ E[X] &= \frac{m}{2} \end{aligned}$$

(b)

□

2. (5+10+10+15 marks) **Estimating the number of tickets**

You are given a bag full of N tickets numbered $1, \ldots, N$(N is unknown to you). You can take out tickets one at a time, note their label, and put them back in the bag. Your task is to estimate N. We will do this in the same way as we estimated π in lecture:

(a) Assume you drew out k tickets. What will be the expected value of the mean of these tickets ? Calculate N in terms of this mean, call this $\tilde{N}$.

(b) Chernoff bound can be extended to work on the case when the Random Variables take values other than $\{0, 1\}$. This is known as Hoeffding's inequality. Use it to find a lower bound on the probability that the error in N, using the above calculation, will be less than $\delta N(\delta < 1/2)$. (in terms of N, δ, k)

(c) Assume k, N are odd. In calculation of part (a), instead of using the value of mean, we use the median of the labels of tickets drawn. Prove a lower bound of $1 - 2e^{-\frac{k(1+2\delta)^2}{2(3-2\delta)}}$ on the probability that the error in N using the median will be less than $\delta N(\delta < 1/2)$. (in terms of N, δ, k)

(d) Start with a random hidden value of N in range $10^4 - 10^6$. Write a function that gives k values from $[N]$ when queried with equal probability. Use these values to calculate $\tilde{N}$ as in part (a) and (c), and plot them with respect to increasing $k \leq 1000$. Repeat this estimation for a total of 3 different N, and put the plots in the main answer file. Submit the code you used to generate these plots, along with a readme on how to execute the code, zipped together with the main answer file into a single .zip file.

Solution:

(a) We draw out k tickets.
Each ticket can take up N values,, i.e., $\{1, 2, 3, \cdots, N\}$, so number of combinations we can get on drawing k tickets is N^k.
The probability of one combination $= p = \left(\frac{1}{N}\right)^k$

Let X be a random variable denoting the mean of tickets. Hence, the expected mean of the tickets is,

$$E[X] = \sum_{X=x} x \cdot p(x)$$

The probability of occurrence of all combinations is same $= p$ (constant). Hence, we can write,

$$E[X] = \sum_{X=x} x \cdot p(x)$$
$$E[X] = \sum_{X=x} x \cdot p$$
$$E[X] = p \cdot \sum_{X=x} x$$

Now, let S be equal to sum of all mean and let A denote the set of all combinations of tickets we can draw.

$$S = \sum_{X=x} x$$
$$= \sum_{\{x_1, x_2, \cdots, x_k\} \in A} \frac{x_1 + x_2 + x_3 + \cdots + x_k}{k}$$

Hence, S is sum of elements of all possible combinations divided by k.

To find sum of elements of all combinations:
We have total of N^k combinations.
For first drawn ticket, each of 1 to N occur equal number of times. So,
1 occurs N^{k-1} times,
2 occurs N^{k-1} times,
3 occurs N^{k-1} times,
$\cdots$
N occurs N^{k-1} times.

Hence, sum for the first ticket is (S_1),

$$\begin{aligned} S_1 &= 1 \cdot N^{k-1} + 2 \cdot N^{k-1} + \cdots + N \cdot N^{k-1} \\ &= (1 + 2 + \cdots + N)N^{k-1} \\ &= \frac{(N)(N+1)}{2} \cdot N^{k-1} \\ &= \frac{N^k(N+1)}{2} \end{aligned}$$

Similarly,with the same symmetric argument, we can find sum of all tickets in the combination. i.e., $S_1 = S_2 = \cdots = S_k$ Hence,

$$\begin{aligned} S &= \frac{S_1 + S_2 + \cdots + S_k}{k} \\ &= \frac{N^k(N+1)k}{2k} \\ &= \frac{N^k(N+1)}{2} \end{aligned}$$

Hence,

$$\begin{aligned} E[X] &= p \cdot S \\ &= \left(\frac{1}{N}\right)^k \cdot \frac{N^k(N+1)}{2} \\ E[X] &= \frac{N+1}{2} \end{aligned}$$

Hence, $\boxed{\tilde{N} = 2E[X] - 1}$

(b) Let $X_1, X_2, \cdots, X_n$ be independent random variable such that $a_i \leq X_i \leq b_i$. Let S_n be the sum of these random variables $S_n = X_1 + X_2 + \cdots + X_n$.

Then, Hoeffding's inequality states that, for all $t > 0$:

$$P(|S_n - E[S_n]| \geq t) \leq 2e^{\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)}$$

Here, we have k random variables (k tickets) which take up values from 1 to N, i.e., for $0 \leq k \leq N$, $1 \leq X_i \leq N$.
Let $S_k = \sum_{i=1}^{k} X_i$.

Let Y be a random variable denoting average of random variable X, i.e., $Y = \frac{S_k}{k}$.
Then by Hoeffding's inequality,

$$P(|S_k - E[S_k]| \geq t) \leq 2e^{\left(\frac{-2t^2}{\sum_{i=1}^{k}(N-1)^2}\right)}$$

$$P(|S_k - E[S_k]| \geq t) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$P\left(\left|\frac{S_k}{k} - \frac{E[S_k]}{k}\right| \geq \frac{t}{k}\right) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$P\left(|Y - E[Y]| \geq \frac{t}{k}\right) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$P\left(\left|Y - \frac{N_{exp}+1}{2}\right| \geq \frac{t}{k}\right) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)} \qquad \textit{From (a)}$$

$$P\left(\left|\frac{2Y - N_{exp} - 1}{2}\right| \geq \frac{t}{k}\right) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$P\left(\left|\frac{N - N_{exp}}{2}\right| \geq \frac{t}{k}\right) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$P\left(|N - N_{exp}| \geq \frac{2t}{k}\right) \leq 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

$$P\left(|N - N_{exp}| < \frac{2t}{k}\right) \geq 1 - 2e^{\left(\frac{-2t^2}{k(N-1)^2}\right)}$$

We have $\frac{2t}{k} = \delta N$, hence substituting t in the above inequality we get,

$$P\left(|N - N_{exp}| < \frac{2t}{k}\right) \geq 1 - 2e^{\left(\frac{-2\delta^2 N^2 k^2}{4k(N-1)^2}\right)}$$

$$P\left(|N - N_{exp}| < \delta N\right) \geq 1 - 2e^{\left(\frac{-\delta^2 N^2 k}{2(N-1)^2}\right)}$$

(c) We are given N, k as odd.
On rearranging all the combinations in an ascending order, by symmetry we can say that, the number of times 1 occur as median is same as the number of times N occurs as median.
Let the number of times 1 occurs as median be n which is also equal to number of times N occurs as median. Hence for these combinations, average of medians is

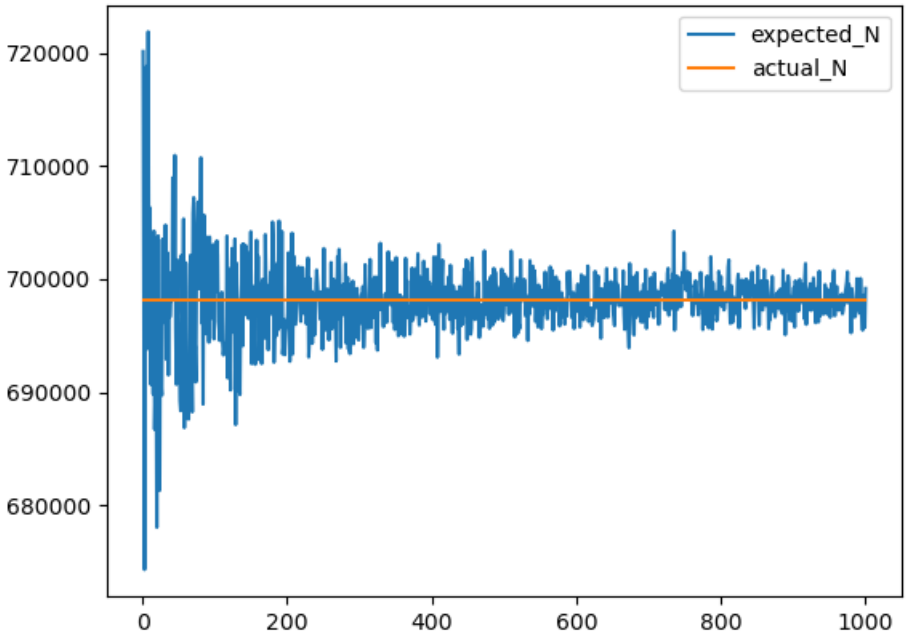
$$M_1 = \frac{1 \cdot n + N \cdot n}{2n}$$
$$= \frac{N+1}{2}$$

Similarly,
the number of times 2 occur as median is same as the number of times $N-1$ occurs as median and for these combination too, average of medians is $\frac{N+1}{2}$.

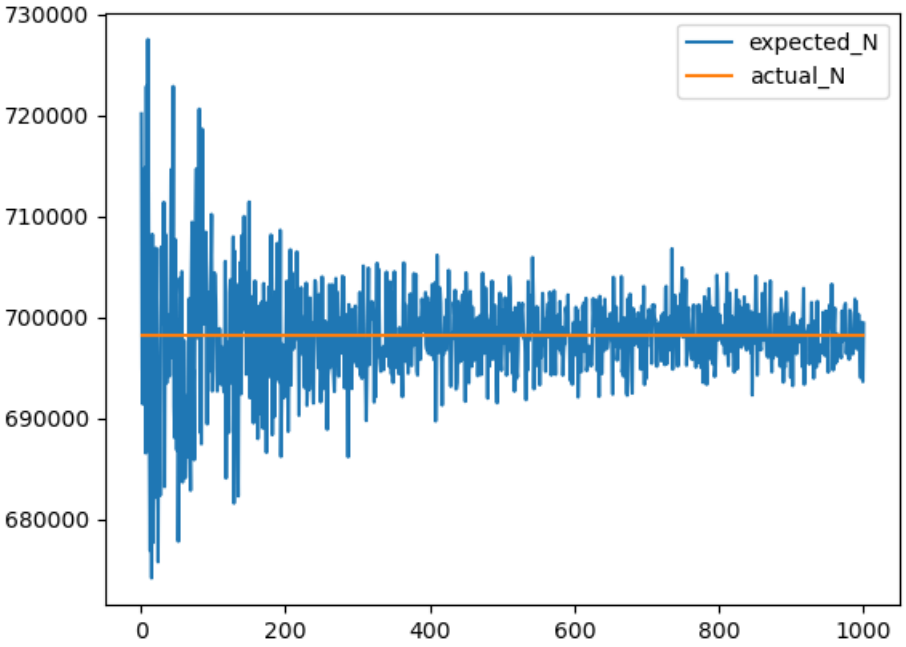
Hence, extending the same argument, we can say that the average of medians is $\frac{N+1}{2}$.

(d) The random experiment is performed 100 times and the estimated N for k going from 1 to 1000 is plotted below:

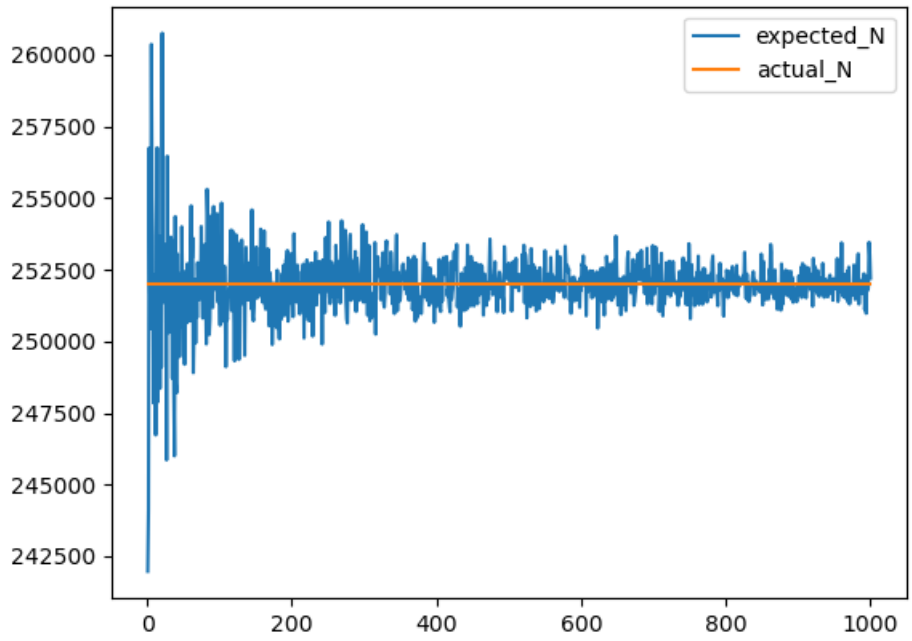
Prediction 1: When N takes up the value = 698209
1(a) Estimated N using mean:



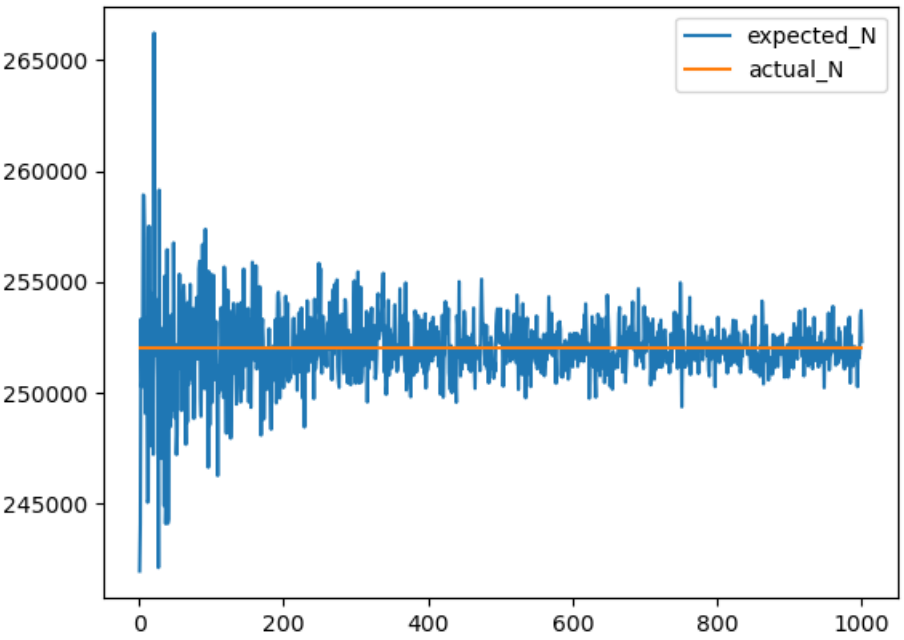
1(b) Estimated N using median:



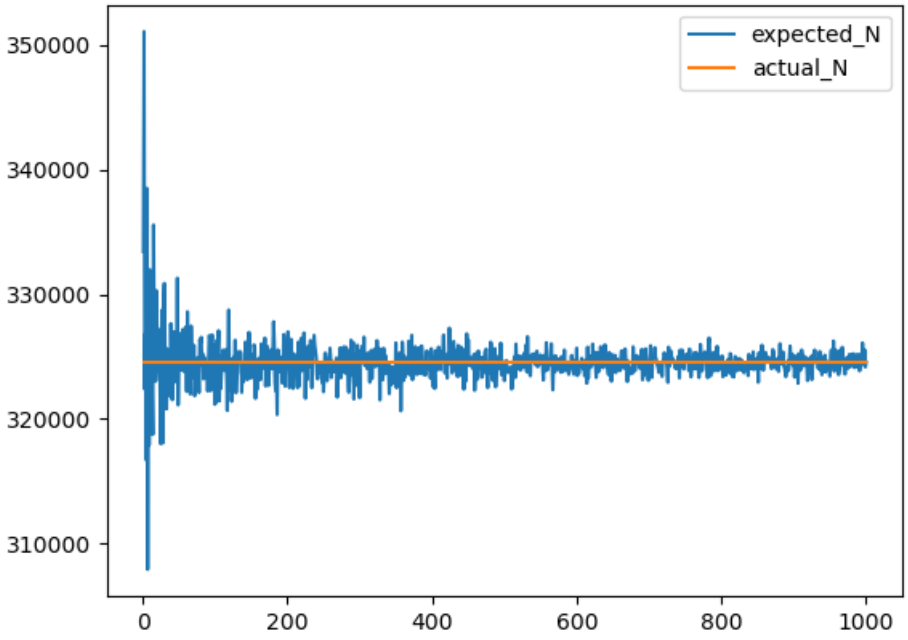
Prediction 2: When N takes up the value = 252009
2(a) Estimated N using mean:



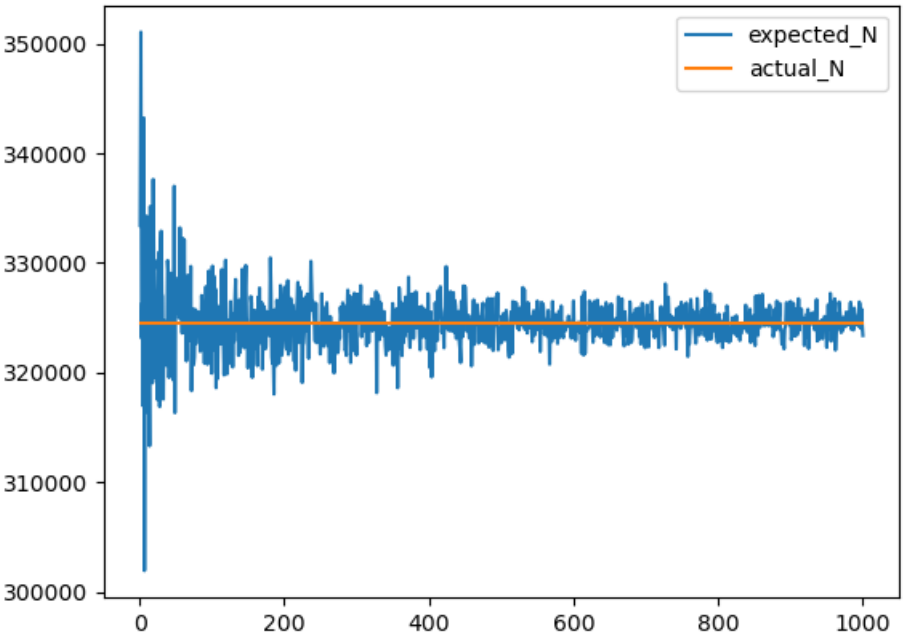
2(b) Estimated N using median:



Prediction 3: When N takes up the value = 324488
1(a) Estimated N using mean:



1(b) Estimated N using median:



□

3. (15+15 marks) **Markov Chain**

Consider a homogeneous regular Markov chain with state space S of size $|S|$, and transition matrix M. Suppose that M is symmetric and entry-wise positive.

(a) Show that all the eigenvalues of M are bounded by 1 and that the uniform distribution is the unique stationary probability distribution for M.

(b) Starting from the stationary distribution, express the probability of returning to the same state as the state at $t = 0$ after $n \in \mathbb{N}$ steps in terms of the eigenvalues of M. Compute the limit of the above probability as $n \to \infty$.

You might find the second part to be easier than the first. Feel free to assume the first part and finish the second part (even when you can't prove the first part).

Solution:

(a) For the eigenvalues λ of matrix M, we have $M\vec{X} = \lambda\vec{X}$ where $\vec{X} \neq 0$.

Let, $\vec{X} = [x_1\ x_2\ x_3 \cdots x_s]^T$
Let $x_{max} = max\{|x_1|, |x_2|, |x_3|, \cdots |x_s|\}$
x_{max} can not be equal to zero, as, if $x_{max} = 0$ then all x_i's will be zero, which can not happen as $\vec{X} \neq 0$.

We have M, a symmetric transition matrix. Let M be denoted as follows,

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} & \cdots & m_{1s} \\ m_{21} & m_{22} & \cdot\cdot & \cdot\cdot & m_{2s} \\ m_{31} & \cdot\cdot & \cdot\cdot & \cdot\cdot & m_{3s} \\ \cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot \\ \cdot & \cdot\cdot & \cdot\cdot & \cdot\cdot & \cdot \\ m_{s1} & \cdot\cdot & \cdot\cdot & \cdot\cdot & m_{ss} \end{bmatrix}$$

Let x_{max} occur at index i. Therefore,

$$M\vec{X} = \lambda\vec{X} \qquad \textit{expanding the } k^{th} \textit{ index, we get}$$

$$\sum_{j=1}^{s} m_{kj}x_j = \lambda x_k$$

$$\Big|\sum_{j=1}^{s} m_{kj}x_j\Big| = |\lambda x_k|$$

$$\sum_{j=1}^{s} m_{kj}|x_j| \geq |\lambda x_k|$$

$$\sum_{j=1}^{s} m_{kj}|x_k| \geq |\lambda x_k|$$

$$|x_k|\sum_{j=1}^{s} m_{kj} \geq |\lambda||x_k|$$

$$|x_k| \geq |\lambda||x_k|$$

$$1 \geq |\lambda|$$

Now, for $\vec{X}$ to be a probability distribution vector, x_i's must be non-negative and sum upto 1.

Expanding $M\vec{X} = \lambda\vec{X}$ we get

$$\begin{aligned} m_{11}x_1 + m_{12}x_2 + \cdots + m_{1s}x_s &= \lambda x_1 \\ m_{21}x_1 + m_{22}x_2 + \cdots + m_{2s}x_s &= \lambda x_2 \\ &\cdots \\ m_{s1}x_1 + m_{s2}x_2 + \cdots + m_{ss}x_s &= \lambda x_s \end{aligned}$$

Sum of all rows is one as it is a transition matrix. It is also given that the matrix is symmetric hence, sum of all columns must be one.

Summing all the above equations we get,

$$\begin{aligned} x_1 + x_2 + x_3 + \cdots + x_s &= \lambda(x_1 + x_2 + x_3 + \cdots + x_s) \\ \lambda &= 1 \end{aligned}$$

Hence, $M\vec{X} = \vec{X}$
We know, the stationary probability distribution is unique.
Take $\vec{X} = \begin{bmatrix} \frac{1}{s} & \frac{1}{s} & \cdots & \frac{1}{s} \end{bmatrix}$
This satisfies all equation $M\vec{X} = \vec{X}$.
Hence, the uniform distribution is the unique stationary probability distribution for M.

(b) We have to find the probability of returning to the same state as the initial state after n steps. Let X be a random variable denoting the state.

$$\begin{aligned} P &= \sum_{i\in[s]} P\left(X_n = i \;\middle|\; X_0 = i\right) \cdot P(X_0 = i) \\ &= \sum_{i\in[s]} P\left(X_n = i \;\middle|\; X_0 = i\right) \cdot \left(\frac{1}{s}\right) \\ &= \left(\frac{1}{s}\right) \sum_{i\in[s]} P\left(X_n = i \;\middle|\; X_0 = i\right) \\ &= \left(\frac{1}{s}\right) \sum_{i\in[s]} (M^n)_{ii} \\ &= \left(\frac{1}{s}\right) \cdot (\textit{trace of } M^n) \end{aligned}$$

Let λ_i's be the eigenvalues of M. Then the eigenvalues of M^n will be $(\lambda_i)^n$.We know trace of matrix is the sum of all eigenvalues. Hence,

$$\begin{aligned} P &= \left(\frac{1}{s}\right) \cdot (\textit{trace of } M^n) \\ &= \left(\frac{1}{s}\right) \cdot \sum_i (\lambda_i)^n \end{aligned}$$

We know, $|\lambda| \leq 1$ and one of the eigenvalue is $\lambda = 1$, hence

$$lim_{n\to\infty} \sum_i (\lambda_i)^n = 1$$

Therefore, as $n \to \infty$,

$$lim_{n\to\infty} P = \frac{1}{s}$$

□