Mandatory assignment 2

Deadline: Friday October 25 at 23:59.

Read carefully through the information about the mandatory assignments in the file *mandatorySTA510.pdf* found in the file folder *Course information* on Canvas. Notice in particular that the assignment should be solved individually.

Hand in on Canvas. If you like, you can make the solution as a R Markdown document. If so, submit both the .rmd file and the complete report as a .html file. A template that you may use is available on Canvas.
Alternatively, you can submit two files, one file including the R-code solving the problems marked with an R and one pdf-file with a report containing the answers to the theory questions. The first line of the R-code file should be: `rm(list=ls())`. Check that the R-code file runs before you submit it. Use comments in the R-code to clearly identify which question each part of the R-code belong to. Also add brief comments to explain important parts of the code. The file ending of the R-code file should be .R or .r. The report can be handwritten and scanned to pdf-file, or written in your choice of text editor and converted to pdf. Cite the sources you use.

Problem 1

In the introductory lecture, we encountered the *birthday problem*: in a group og N persons, what is the probability that at least k have birthday on the same day?

We will now consider a continuous-time version of this problem where people enter a room according to a Poisson process $(N_t)_{t\geq 0}$ with rate $\lambda = 1$. Each person is independently marked with one of 365 birthdays, where all birthdays are equally likely. Let $X_1, X_2, \ldots,$ be the interarrival times for the process of people entering the room. Let T be the first time when two people in the room share the same birthday.

a) Argue that $T = \sum_{i=1}^{K} X_i$, where the X_i's are independent of K and where K is the number of people in the room the first time that two people share the same birthday.
Show that $E(T) = E(K)$.

By analytical methods, using properties of Poisson processes, it can be shown that

$$E(T) = \int_0^\infty (1 + \frac{t}{365})^{365} e^{-t} dt \tag{1}$$

b)R Estimate the integral in (1) using the general basic Monte Carlo integration method with 1000 simulations. Next, compute the sample variance and use this to compute an estimate of the number of simulations needed in order to estimate the integral in (1) with a precision of $e = 0.01$ and confidence of $\approx 90\%$. At last, run the basic (crude) Monte Carlo integration method once more in order to estimate the integral in (1), this time with the number of simulations just computed.

We will now consider a generalization of the above continuous-time birthday problem. As in a)-b), the task is to estimate the expected time it takes until the first time when two people in the room share the same birthday. However, we now assume that the persons arrive according to a non-homogeneous Poisson process (NHPP) with rate

$$\lambda(t) = t \tag{2}$$

c) Two ways of simulating data from NHPP models have been discussed in the lectures. Why is the thinning method less suitable than the transformation method in this case?
Describe a simulation algorithm for the continuous-time birthday problem in the NHPP case based on the transformation method.
Additionally, explain how to use the basic (crude) Monte Carlo method in order to estimate $\Lambda(t) = \int_0^t \lambda(s)ds$ for any fixed $t > 0$, and explain why the method of antithetic variables can be applied in this case.

d)$^{\mathrm{R}}$ Implement in R the simulation algorithm for the continuous-time birthday problem with NHPP described in c) and estimate both $E(T)$ and $E(K)$ for this case. Do these values seem to agree also in this particular non-homogeneous case?

Problem 2

In the first part of this problem we will consider two random walk models: the random walk (X_n) on a complete graph and the random walk (Y_n) on a discrete circle. We assume that both graphs have $n = 11$ vertices. The transition probabilities of the two processes are thus given as follows. For all $i, j \in \{1, \dots, 11\}$, we have

$$P(X_{n+1} = i \mid X_n = j) = \frac{1}{11} \tag{3}$$

$$P(Y_{n+1} = i \mid Y_n = j) = \begin{cases} \frac{1}{2} & \text{if } i \equiv j \pm 1 \mod 11; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

(Thus, Y_n jumps right or left with equal probability. If $Y_n = 1$ it either jumps to 2 or 11, and if $Y_n = 11$ it either jumps to 10 or 1).

a) Show that both models are examples of regular Markov chains and compute their limiting probability distributions, $\pi_{complete}$ and π_{circle}, respectively.

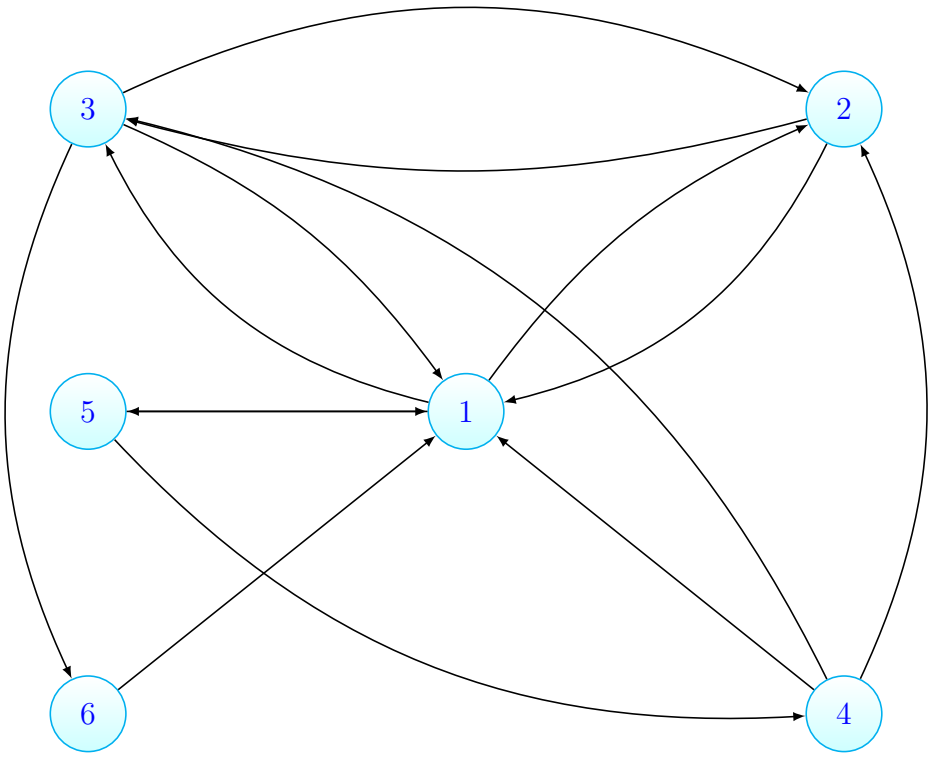
Recall that Zipfs law (which we also encountered in Problem 4 of Mandatory Assignment 1) with parameters $s > 0$ and $n \in \{1, 2, \dots\}$ is given by

$$P(X = k) = \frac{1/k^s}{\sum_{i=1}^{n}(1/i^s)}, \quad \text{for} \quad k = 1, \dots, n. \tag{5}$$

b)[R] Use the random walk on the complete graph to initiate the basic Metropolis-Hasting algorithm and to simulate Zipfs law with $n = 11$ and $s = 3$.
Run an additional simulation where you instead use the random walk on the discrete circle in order to initiate the basic Metropolis-Hasting algorithm.
In both cases, start the random walk at vertex 1 and run 10.000 simulation steps, and apply the law of large numbers (LLN) for Markov chains in order to estimate Zipfs law. To compare how well the two simulation methods approximate the exact distribution of Zipfs law, compute (for both cases) their *total variation distance* to Zipfs law. Here, for two measures μ and ν on $S = \{1, \dots, n\}$, the total variation distance is given by

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{i=1}^{n} |\mu_i - \nu_i| \tag{6}$$

We will now consider a random walk model on the less regular graph with directed edges depicted below. Imagine that the vertices of this graph represents webpages and that there is a directed link between vertex v and vertex w (thus, with arrow pointing at w) if v contains a link to page w. Now, consider a *random surfer* $(Z_n)_{n\geq 0}$ that jumps from his current webpage v by randomly choosing an outgoing link (e.g. if the random surfer is at webpage 3, it can jump to websites $1, 2$ and 6).



c)[R] Implement a simulation algorithm in R that simulates the just described process of the random surfer, starting the Markov chain by sampling from the uniform measure on $\{1, \dots, 6\}$. Estimate the limiting distribution of (Z_n) by running the process for $n = 10.000$ steps and by applying the law of large numbers for Markov chains. Based on your estimate, make a ranking of the webpages according to how often they were visited by the random surfer.

Consider now the slight modification of the Markov chain in c) where, in each step, the random surfer tosses a (unfair) coin: if heads, it proceeds as described on the previous page; if tails, its next visited website is chosen uniformly at random from all the 6 vertices.

d)$^{\mathrm{R}}$ Implement a simulation algorithm in R that simulates the modified version of the random surfer. Perform the simulation for the two cases where $P(\text{heads}) = 0.2$ and $P(\text{heads}) = 0.85$. As in c), start from the uniform measure on $\{1, \ldots, 6\}$. For each case, estimate the limiting distribution by running the process for $n = 10.000$ steps and by applying the law of large numbers for Markov chains. Does the coin tossing have any notable effect on the limiting distribution when compared to the distribution obtained in c)?

Problem 3

You want to sell an item within a time window $[0, 1]$. Once you receive a bid, you must accept/reject it immediately. In order to maximise your profit, you are considering the following three selling strategies:

Strategy A: Fix a price $\theta \in [0, 1]$ and accept first bid over θ.

Strategy B: Fix a price $\theta \in [0, 1]$ and accept first bid which (at time t) is larger than $\theta \cdot (1 - \frac{t^2}{2})$.

Strategy C: Accept first bid which (at time t) is larger than $1 - t$.

Based on previous experiences, you assume that bids arrive at times of a Poisson process with rate $\lambda = 1$. Moreover, the bids are modelled as i.i.d. unif$[0, 1]$ random variables.

a) Show that, under strategy A, $E(\text{profit}) = \frac{1+\theta}{2}(1 - e^{(\theta-1)})$.

b)$^{\mathrm{R}}$ Implement an algorithm in R that simulates the selling process as described above under strategy B. Use this approach to estimate the expected profit for each $\theta_i = \frac{i}{10}$, $i = 1, \ldots, 10$.

c) Using that $E(\text{profit}) = \int_0^1 P(\text{profit} > t)dt$, show that under strategy C,

$$E(\text{profit}) = 1 - e^{-1/2} \int_0^1 e^{\frac{t^2}{2}} dt = 1 - e^{-1/2} \int_0^1 e^{\frac{(t-1)^2}{2}} dt \tag{7}$$

d) Describe a simulation algorithm based on the hit and miss method that can be applied to estimate (7). Calculate how many simulations one would need to run in order to have a approximate 95% probability that the estimate is at most a margin of $e = 0.02$ from the true value. Additionally, by choosing a suitable proposal function, describe a simulation algorithm based on the method of importance sampling that approximates (7).

e)$^{\mathrm{R}}$ Implement in R the simulation algorithm based on the hit and miss method described in d) to estimate the integral in (7).
Based on your calculations in a), b) and e), present the estimates for all the three strategies in a table. Which strategy gives the highest expected profit?