Mandatory assignment 1

Deadline: Friday September 23th at 12:00 (noon, Norwegian time).

Hand in on Canvas. Submissions should be of **either** of the following types

- Preferred: Submit two files: One R markdown (Rmd) file containing both theory answers and R-code, and a pdf-file with the output you obtain when running (knitting) you R markdown file. See tutorial to get started.
- Submit two files: one pdf-file with a report containing the answers to the theory questions, and one file including the R-code.

The first line of R-code should be: `rm(list=ls())`. Check that the Rmd/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 try to add some comments to explain important parts of the code. The file ending of the R-code file should be .Rmd, .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.

The following files are associated with this assignment:

- `lawn.R` (see Problem 3)

and should be stored in your working directory when solving the exercise (but should not be submitted). Problems marked with an $^{\mathrm{R}}$ should be solved in R, the others are theory questions.

Finally, note that some of the problems are similar (but not equal) to those in the first mandatory assignment of 2020 (which has published solution on Canvas).

Problem 1:

In this problem, we consider a random variable X with probability density function given by

$$f(x) = \begin{cases} \frac{x}{2}, & \text{if } 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

a) Verify that this is a proper probability density function.
Find the mean $E(X)$ and the variance $Var(X)$.
Also find $E\left(\frac{1}{2}(X+1)\right)$ and $E(X^3)$.

b) Find the cumulative distribution function $F(x)$ associated with X, and use this to calculate $P(X < 1)$ and $P(0.5 < X < 1.5)$.
Find also the inverse cumulative distribution function $F^{-1}(u)$ (Hint: You should get a quadratic equation which generally have two solutions, but only one of these solutions (which?) gives sensical results here.)

c)$^{\mathrm{R}}$ Write a function (with single argument n) which produces n random numbers with density (1) using the inverse transform method. Ideally, the function should not contain any for-loops.
Check that your function produces correct results by comparing a histogram (use argument `probability=TRUE`) of many random variables to the density (1).

d) Using a uniform(0,2) proposal distribution g, do the calculations required to obtain an accept-reject method for drawing random numbers with density (1).

e)$^{\mathrm{R}}$ Write a function (with single argument n) which produces n random numbers with density (1) using the accept-reject method.
Check the results in the same way as in point c.
Which method would you prefer in this case - accept-reject or inverse transform? The function `system.time` is useful in this regard.

f)$^{\mathrm{R}}$ Check the answers you got for expectations, variance and probabilities in points a) and b) above by simulation. For $E(X)$, also find a simulation-based 95% confidence interval. Use either the inverse transform or accept-reject algorithm for generating random numbers with density (1).

g)$^{\mathrm{R}}$ Plot a kernel density estimate of $f(x)$ based on $n = 100000$ simulated random variables and compare it to the true $f(x)$. Comment on what you find.

Problem 2:

In this problem, we let $X_1, X_2, \ldots, X_n$ be iid $Exp(1)$ random variables (i.e. exponential random variables with expectation equal to 1, to simulate n of these, use `rexp(n)`).

a) Find the mean and variance of $A = \sum_{i=1}^{n} X_i$.
Find the mean and variance of $B = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Which distribution does $C = \sqrt{n}(B-1)$ have when n is "large"?

b)$^{\mathrm{R}}$ For $n = 1000$, check the results you got in a) by simulating many independent replications of each of A, B, C (note, to simulate a single A, B or C, you need to simulate n Xs).

Suppose now that you did not know $E(X)$, but in stead wished to estimate it using B as a point estimate and $[B \pm 1.96S/\sqrt{n}]$ as an approximate 95% confidence interval, where S is the estimate of standard deviation calculated from the random sample $X_1, \ldots, X_n$.

c)$^{\mathrm{R}}$ For each of $n = 10, 100, 1000$, estimate the probability that the above confidence interval contains the true $E(X)$. Comment on what you find.

Problem 3:

Consider the probability distribution with density

$$f(x) = \frac{1}{3}(\exp(-x) + \exp(-x/2)),\ x \geq 0$$

a) Derive a method (of your choice) for simulating random variables with density $f(x)$.

b)$^{\mathrm{R}}$ Implement the method you choose in a), and check your code by comparing a histogram against the true density.

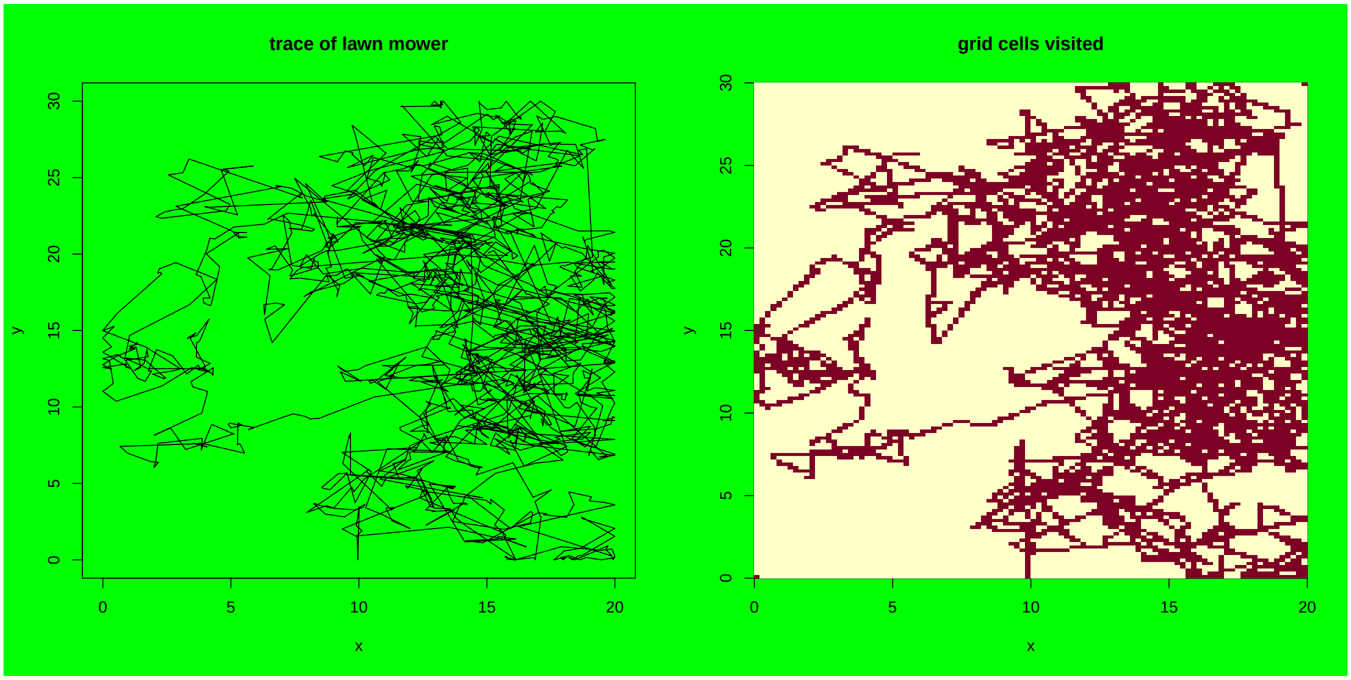

Figure 1: Realization of the lawn mower process with $Etime = 1.0$. Around 37% of the grid cells were visited.

Problem 4:

In this problem, we consider the (simplified) simulation-based design optimization of the control mechanism for a prototype of a robotic lawn mower.

You consider a (virtual) rectangular lawn 20 meters in the x-direction, and 30 meters in the y-direction. The lawn mower generally moves straight ahead with a speed of 1 meter per second (just to keep things simple), and only changes direction if either of the following happens:

- When hitting the boundaries of the lawn, the mower does an inelastic bounce. I.e. if the mower hits the left- or right hand side boundaries, the sign of the velocity in the x-direction is switched, and if the mower hits the lower- or upper boundaries, the sign of the velocity in the y-direction is switched.
- At exponentially distributed times (since the last change in direction), the mower randomly chooses a new direction. These exponentially distributed times have expectation $Etime$.

The effect of choosing different values of $Etime$ is illustrated in Figures 1 and 2. The left panel in both figures show (the centerpoint of the mower) 1500 seconds worth of mowing. In Figure 1, $Etime = 1$, i.e. random direction changes occur on average every second. In Figure 2, $Etime = 1000$ which effectively turns of the random direction changes.

In order to assess the relative merit of different values of $Etime$, the complete lawn is divided into 20 cm by 20 cm grid cells[1]. Subsequently, it is counted how many of these cells were visited at least once by the mower. The right panels in Figures 1 and 2 illustrates this, where red-colored cells are visited. Finally the proportion of visited cells (i.e. visited cells/total number of cells) is

[1]To reflect that the actual cutting device of the mower is around 20 cm wide

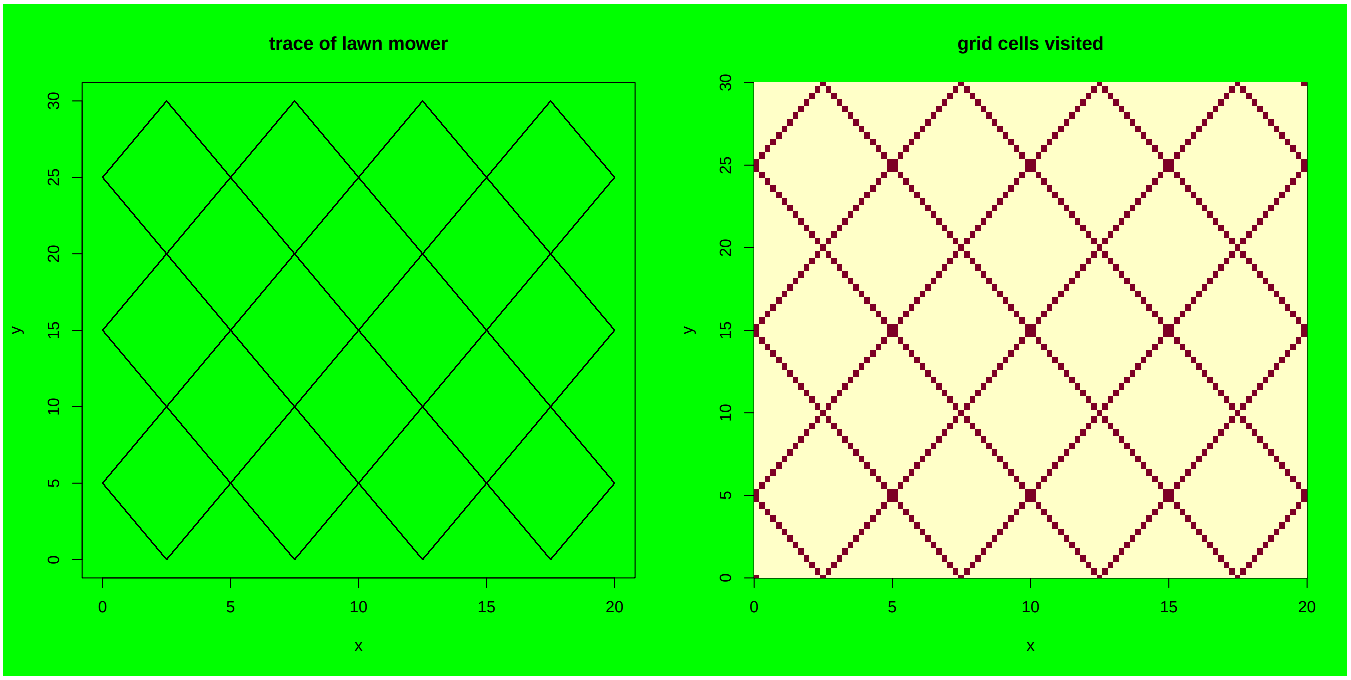

Figure 2: Realization of the lawn mower process with $Etime = 1000.0$. Only around 8% of the grid cells were visited.

taken as a performance measure[2]. In Figure 1, 37% of cells were visited, whereas only 8% of the cells were visited in Figure 2.

R-code for simulation the lawn mower process is provided in the file `lawn.R`. The file consist of a single function with signature

```
sim.lawn<-function(Etime=10.0,plot.trace=FALSE,plot.image=FALSE)
```

You may have to install the package `gplots` in order to run the code[3].

The function takes a fixed value of *Etime* as argument and returns a (random) visit proportion. Setting either `plot.trace` or `plot.image` to `TRUE` produces plots of the types in Figures 1 and 2. Be sure to keep these latter arguments at `FALSE` when running simulations to save computing time.

a)[R] Generate at least 100 visit proportions for $Etime = 5.0$. Make a histogram of the generated data and calculate the (approximate) mean, median and the standard deviation of the distribution of visit proportions for $Etime = 5.0$.

---

[2] This is a reasonable approximation to the actual proportion of the lawn covered by some region of the cutting device

[3] i.e. by running `install.packages("gplots")`

b)[R] Also generate 100 visit proportions for each of $Etime = \{1.0, 10.0, 100.0\}$. Make a box-plot of the different visit proportion distributions. Compare the

- mean visit proportion.
- median visit proportion.
- 0.05−quantile of the visit proportion

for each of the settings ($Etime = \{1.0, 5.0, 10.0, 100.0\}$). Which one of the settings would you prefer for the finished product?