STA510 2022 mandatory assignment 3

Deadline: Friday October 21st 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.

Problems marked with an $^{\text{R}}$ should be solved in R, the others are theory questions.

It is OK to use any code from lectures, examples etc. AS LONG AS THESE ARE PROPERLY REFERENCED TO.

Along with this set of exercises, an additional file `M3data.txt` is needed in order to solve problem 3.

Problem 1 "Monte Carlo integration"

Consider the following integrals:

$$A = \int_0^\infty \exp(-x)\sqrt{x}dx$$

$$B = \int_0^{10} \exp(-x)\sqrt{x}dx$$

$$C = \int_{-1}^1 \sqrt{1+\cos(x)}dx$$

$$D = \int_{-1}^1 \sqrt{1+\sin(x)}dx$$

For checking your code, the values of the integrals are $A \approx 0.8862269255$, $B \approx 0.8860764953$, $C \approx 2.712040395$ and $D \approx 1.917702155$.

a) Write down an expression for a Monte Carlo estimator of A based on n random draws from an $Exponential(1)$ distribution.

b) Write down an expression for a Monte Carlo estimator of B based on n random draws from an $Exponential(1)$ distribution.

c) Write down an expressions for Monte Carlo estimators of C and D based on n random draws from a $Uniform(-1, 1)$ distribution.

d) Write down an expressions for importance sampling estimators of A and B based on n random draws from a $Gamma(3/2, 1)$ distribution. Do you, based on theory, expect these to work well?

e) Write down an expressions for importance sampling estimators of A and B based on n random draws from a $Gamma(4, 1/4)$ distribution (i.e. with scale parameter equal to $1/4$). Do you, based on theory, expect these to work well?

f) Write down an expressions for antithetic Monte Carlo estimators of C and D based on random draws from a $Uniform(-1, 1)$ distribution and n evaluations of the integrands. Do you, based on theory, expect these to work well?

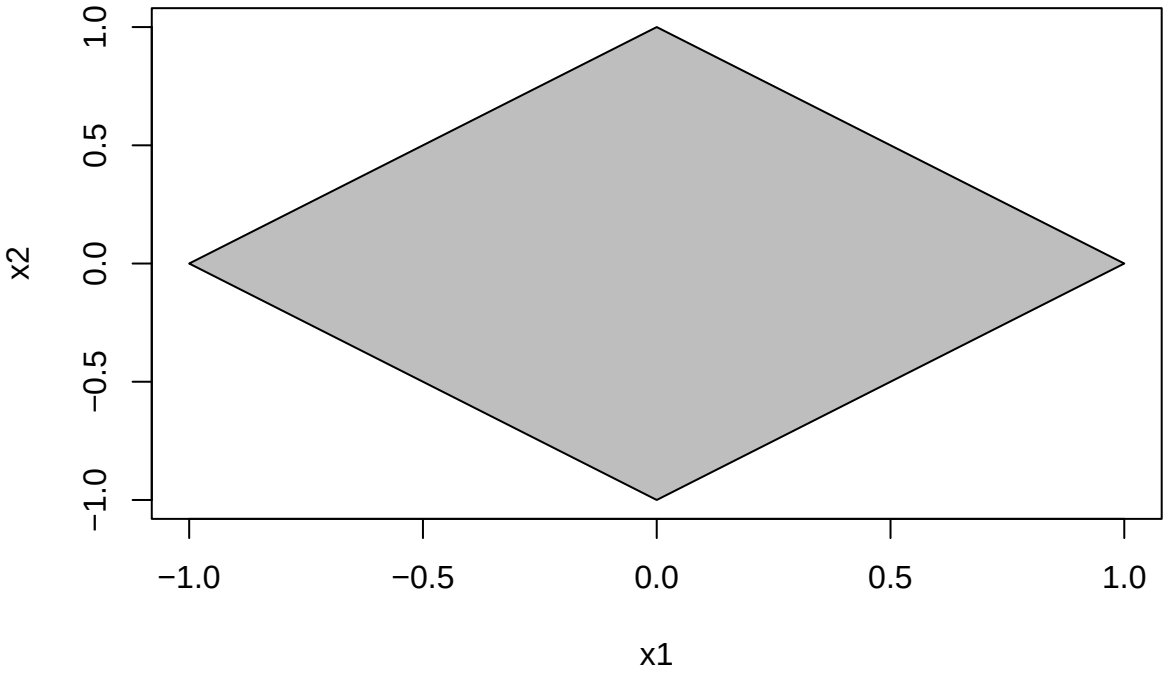
g)^R Implement the estimators you proposed in points d), e) and f), and verify that your theoretical predictions concerning the behavior of the estimators are correct.

Problem 2 "A cross-polytope"

In financial engineering, cross-polytopes play an important role for specifying so-called gross exposure constraints on a portfolio of financial assets.

A polytope in d-dimensional space is defined to be the points $\mathbf{x} = (x_1, \ldots, x_d)$ so that $\sum_{i=1}^{d} |x_i| \leq c$ for some $c > 0$.

In two dimensions (i.e. $d = 2$), the cross-polytope with $c = 1$ is the shaded region in the plot below:



In this problem, we are concerned with estimating the volume of a d-dimensional cross-polytope (for moderate values of d, say smaller than 10). To this end, we use Monte Carlo integration, as the volume V_c of the polytope may be written as

$$V_c = \int_{-c}^{c} \cdots \int_{-c}^{c} \mathbf{1}(\mathbf{x}) dx_1 \ldots dx_d$$

where

$$\mathbf{1}(x) = \begin{cases} 1 & \text{if } \sum_i |x_i| \leq c \\ 0 & \text{otherwise} \end{cases}$$

It is first proposed to use the Monte Carlo estimator

$$\hat{V}_c = \frac{(2c)^d}{n} X, \text{ where } X = \sum_{j=1}^{n} \mathbf{1}(\mathbf{u}_j)$$

and $\mathbf{u}_j = (u_{1,j}, \ldots, u_{d,j})$ and each $u_{i,j} \sim$ iid $Uniform(-c, c)$.

a) Show that $X \sim binomial(n, V_c(2c)^{-d})$.

It can be shown that $V_1 = 2^d/(d!)$ (i.e. the volume for $c = 1$, see e.g. https://en.wikipedia.org/wiki/Cross-polytope). Hence in this case, $X \sim binomial(n, 1/(d!))$

b) Work out the number of simulations n needed so that the *relative standard error* $SD(\hat{V}_1)/V_1$ is smaller than some constant r. Comment on what you find.

c)^R Write a function for estimating V_c for a given dimension d, c and n. Test your function with $c = 1$ and $d = 2, 4, 8$

Now we consider importance sampling using iid $N(0, \sigma^2)$ proposals. The resulting estimator of V_c

$$\tilde{V}_c = \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbf{1}_c(\mathbf{z}_j)}{\prod_{i=1}^{d} \mathcal{N}(z_{i,j}|0, \sigma^2)}$$

where $\mathbf{z}_j = (z_{1,j}, \ldots, z_{d,j})$ and all $z_{i,j} \sim$ iid $N(0, \sigma^2)$. Further, $\mathcal{N}(x|\mu, \sigma^2)$ is the probability density function of a $N(\mu, \sigma^2)$ random variable evaluated at x.

d)^R Can you improve the relative standard error of $\hat{V}_1$ using the importance sampling estimate $\tilde{V}_1$ for $d = 8$? Explore this numerically, including finding a good value of σ.

In financial engineering, one is sometimes tasked with computing the probability of a multivariate normal random variable $N(\mu, \Sigma)$ being inside a cross-polytope, i.e.

$$\alpha = P\left(\sum_{i=1}^{d} |x_i| < c\right) = \int \cdots \int \mathbf{1}(\mathbf{x}) \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right) d\mathbf{x}$$

Consider $d = 4, c = 1$ and

$$\mu = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 1 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1 \end{pmatrix}.$$

e)^R Using a Monte Carlo method of your choice, estimate the probability α above. Present your best estimate of α along with a measure of the uncertainty in that estimate.

Problem 3

You have a data set consisting of $n = 239$ observations of some random variable X, which has density $f(x)$. The data set is available in the file *M3data.txt*. Your are tasked with estimating $f(1)$, i.e. the value of the true density at $x = 1$.

For this purpose, you use the kernel density estimation to find a point estimate for $f(1)$ as follows:

```
x <- read.table("M3data.txt")$V1 # read the data
# estimator of f(1) based on KDE:
f1.hat <- function(x){return(density(x,from=1,to=1,n=1)$y)}
# estimate of the full data set:
f1.o <- f1.hat(x)
print(paste0("Estimate of f(1) : ",f1.o))
```

```
## [1] "Estimate of f(1) : 0.383732864871009"
```

a)^R Generate $B = 2000$ bootstrap estimates of $f(1)$, and display these using a histgram. Estimate the bias and standard deviation of the KDE-based estimate of $f(1)$ based on the bootstrap estimates.

An alternative method for estimating $f(1)$ is using a histogram. In this case, estimate would be

$$\hat{\theta} = \frac{A}{2nw} \text{ where } A = \text{ number of observations in the interval } (1 - w, 1 + w)$$

The bin half-width may be choosen using Scott's Normal Reference rule to yield

$$\hat{w} = \frac{3.49}{2}\hat{\sigma}n^{-1/3}$$

b)^R Implement the estimator $\hat{\theta}$ (based on Scott's Normal Reference rule) in a function that takes a data set of length n as the only argument. Obtain an estimate of $f(1)$ based on the data set above.

c)^R Obtain bootstrap-based estimates of the bias and standard deviation of $\hat{\theta}$ based on the data set above. Which estimator would you prefer - $\hat{\theta}$ or the KDE-based above?

Problem 4

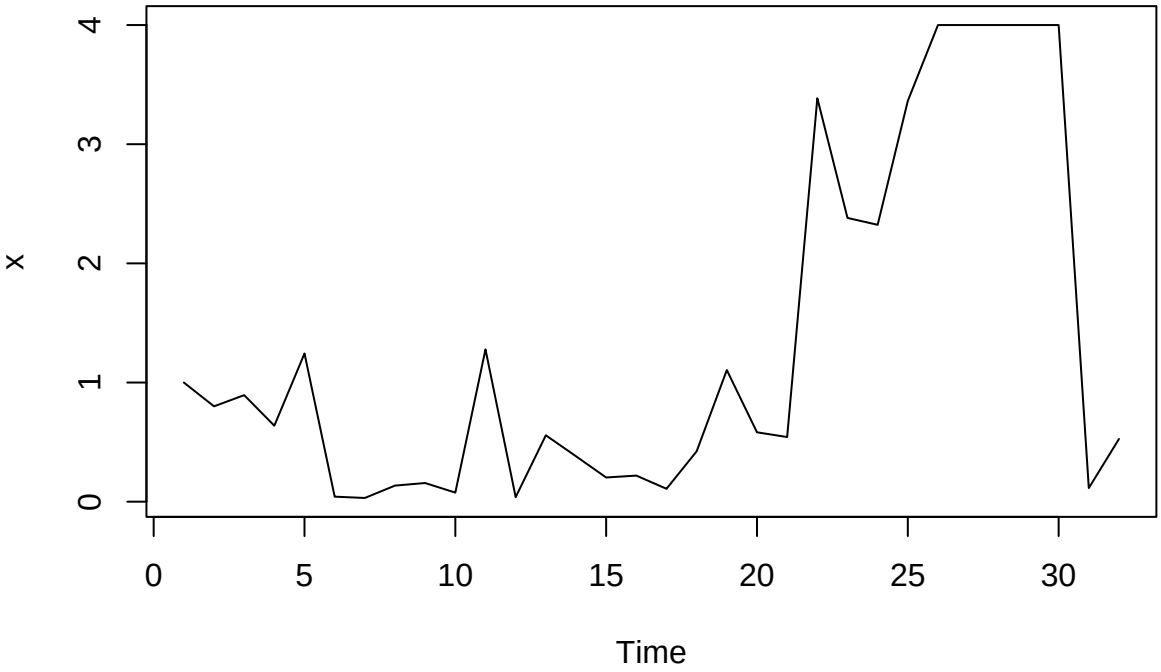
A wind power producer assumes that the daily production of power x_t (in day t, in suitable units) follows the process

$$x_t = \min(4.0, [0.4 + 0.7x_{t-1} + 0.1x_{t-2}]\varepsilon_t),\ \varepsilon_t \sim \text{ iid } Exponetial(1).$$

As can be seen 4.0 is the maximum capacity of the production facility.

Suppose today is day $t = 2$, and $x_2 = 0.8$, and further, that yesterday's production was $x_1 = 1.0$. Below is a piece of code that simulates production for the next 30 days (i.e. days $t = 3, \ldots, 32$) and displays this realization

```
n <- 32
x <- numeric(n)
x[1] <- 1.0
x[2] <- 0.8
for( i in 3:n){
  x[i] <- min(4.0,rexp(1)*(0.4+0.7*x[i-1]+0.1*x[i-2]))
}
ts.plot(x)
```



To answer the following questions, you would need to modify the code so that if produces many realizations of future power production.

a)^R Make graphical representations of the distribution of

- The total production in the next 30 days, i.e. $T = \sum_{t=3}^{32} x_t$.
- The number of days at full capacity, i.e. for how many days are $x_t = 4.0$?
- The smallest daily output, i.e. $\min_t x_t$.

Also find the mean and $(25\%, 50\%, 75\%)$-percentiles of the distributions.

It is of particular interest to estimate $E(T)$ and $P(T > 30)$ with high precision (the company has already sold 30 units of power in the market, and need to know the probability of fulfilling that contract...). For this purpose, you try out using antithetic simulations. Here it is convenient to recall that an $Exponential(1)$ random variable may be simulated as $-\log(U)$ where $U \sim Uniform(0, 1)$. Also recall that you must compare the same number of simulations in regular and antithetic simulations.

b)^R Compare regular Monte Carlo simulations of $E(T)$ and $P(T > 30)$ and antithetic simulations of the same quantities. Comment on what you find.

If the company fails to produce the 30 units of power the company has sold, it needs to by the shortfall, i.e. $S = \max(0, 30 - T)$ from the power market at the end of the month. The price of one unit of power is assumed to have a gamma distribution with shape parameter $\alpha = 10$ and scale parameter $\beta = 1$.

c)^R How much money must the company aside in order to be 95% certain to have sufficient funds are available to by any shortfall from the power market?