TMA4300 - Exercise 1

Stochastic simulation

Marcus A. Engebretsen and Gina Magnussen

Problem A

Probability integral transform, rejection sampling and bivariate techniques

1. Sampling from g(x) - Probability integral transform

The probability density function

$$g(x) = \begin{cases} cx^{\alpha-1}, & 0 < x < 1 \\ ce^{-x} & 1 \leq x \\ 0 & \text{otherwise,} \end{cases}$$

is given with c as a normalising constant and $\alpha \in (0, 1)$.

a)

The cumulative distribution function $G_X(x)$ is then found by integrating over the different domains,

$$G_X(x) = \begin{cases} \frac{c}{\alpha} & 0 < x < 1, \\ \frac{c}{\alpha} + ce^{-1} - ce^{-x} & 1 \leq x \\ 0 & \text{else} \end{cases}$$

By using the property that the area under a pdf should integrate to 1, we find that $c = \frac{c\alpha}{e+\alpha}$
The probability integral transform, setting $u = G_X(x)$ and solving for x, gives

$$x = G^{-1}(u) = \left(u\frac{e+\alpha}{e}\right)^{\frac{1}{\alpha}}, \qquad 0 < x < 1$$
$$x = G^{-1}(u) = -log\left[\left(\frac{e+\alpha}{e\alpha}\right) - \frac{u}{c}\right] = -\log\left[\frac{1}{c}(1-u)\right] \qquad 1 \leq x$$

the inverse expressions of the cumulative distribution function.

b)

R function

R plot: gsample

2. Gamma distribution with $\alpha \in (0, 1)$, $\beta = 1$ generated by rejection sampling

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}, & 0 < x \\ 0 & \text{else} \end{cases}$$

with $\alpha \in (0, 1)$ and $\beta = 1$.

a)

The acceptance probability R:

$$\frac{1}{\gamma}\frac{f(x)}{g(x)}$$

$$\gamma_1 = \frac{f(x)}{g(x)} = \begin{cases} \frac{1}{\Gamma(\alpha)}\frac{e^{-x}}{c} & 0 < x < 1 \\ \frac{1}{\Gamma(\alpha)}\frac{x^{\alpha-1}}{c} & x \geq 1 \end{cases}$$

$$\frac{\partial \gamma_1}{\partial x} = \begin{cases} -\frac{1}{\Gamma(\alpha)}\frac{e^{-x}}{c} & 0 < x < 1 \\ \frac{\alpha-1}{\Gamma(\alpha)}\frac{x^{\alpha-2}}{c} & x \geq 1 \end{cases}$$

Solving $\frac{\partial \gamma_1}{\partial x} = 0$ for x and inserting into the expression for R will now maximize the acceptance probability given the constraint that $f(x) < \frac{1}{\gamma}g(x)$. CHECK CONSTANTS!!

$$\gamma_1 = \begin{cases} \frac{1}{\Gamma(\alpha)}\frac{1}{c} & 0 < x < 1 \\ \frac{1}{\Gamma(\alpha)}\frac{1}{c} & x \geq 1 \end{cases}$$

which in turn gives an acceptance probability

$$R = \begin{cases} e^{-x} & 0 < x < 1 \\ x^{\alpha-1} & x \geq 1 \end{cases}$$

b) R function

3. Ratio of uniforms - gamma, $\alpha > 1, \beta = 1$.

Consider now the same distribution, but with parameters $\alpha > 1$ and $\beta = 1$. The ratio of uniforms method is used to simulate samples from this distribution. With C_f, f^*, a, b_+ and b_- as given in formula (3) and (4) in the exercise description, we find that

Need to find maximum of $f^*(x)$ to find the bounds for the area C_f. Since $f^*(x)$ is a concave function, we solve

$$\frac{df^*(x)}{dx} = (\alpha-1)x^{\alpha-2} - x^{\alpha-1}e^{-x} = 0 \implies x = \alpha - 1 \frac{d(x^2 f^*(x))}{dx} = 0 \implies x = \alpha + 1$$

$$\begin{aligned} f^*(\alpha-1) &= (\alpha-1)^{\alpha-1}e^{1-\alpha} \\ a &= \sqrt{\sup_x f^*(x)} = \sqrt{f^*(\alpha-1)} = \sqrt{(\alpha-1)^{\alpha-1}e^{1-\alpha}} \\ b_+ &= \sqrt{\sup_{x\geq 0} x^2 f^*(x)} = \sqrt{(\alpha+1)^2 f^*(\alpha+1)} = (\alpha+1)^{\frac{\alpha+1}{2}}e^{-\frac{\alpha+1}{2}} \\ b_- &= (\sqrt{\sup_{x\leq 0} x^2 f^*(x)?}) = 0 \end{aligned}$$

$$C_f = [0,a]\times[b_-,b_+] = \left[0, \sqrt{(\alpha-1)^{\alpha-1}e^{1-\alpha}}\right] \times \left[0, (\alpha+1)^{\frac{\alpha+1}{2}}e^{-\frac{\alpha+1}{2}}\right]$$

Ratio of uniforms:

$$x_1 = \frac{u_2}{u_1} \qquad\qquad x_2 = u_1$$

Problem B

The Dirichlet distribution: Simulating using known relations

$$
\begin{aligned}
f_z(z,\alpha) = dz_1...dz_k &\propto (z_1^{\alpha_1-1})e^{-z_1}...(z_k^{\alpha_k-1})e^{-z_k}dz_1...dz_k \\
&= z_1^{\alpha_1-1}...z_k^{\alpha_k-1}e^{-(z_1+...+z_k)}dz_1...dz_k \\
&= z_1^{\alpha_1-1}...z_k^{\alpha_k-1}e^{-v}dz_1...dz_k
\end{aligned}
$$

where $v = -(z_1 + ... + z_k)$

Change of variables

$$z_i = x_i \cdot v \implies dz_i = dx_i \cdot v + x_i dv$$

Using $\sum_{i=1}^{k} x_i = 1$ and $dx_1 + ... + dx_k = 0$ Define $w = dz_1 + ... + dz_{k-1} = [dx_1 + ... + dx_{k-1}]v + [x_1 + ... + x_k]dv$

Then

$$
\begin{aligned}
dz_k &= dx_k v + x_k v \\
&= -[dx_1 + ... + dx_{k-1}]v + [1 - [x_1 + ...x_{k-1}]]dv \\
&= dv - ([dx_1 + ... + dx_{k-1}]v + [x_1 + ...x_{k-1}]dv) \\
&= dv - w
\end{aligned}
$$

Using exterior algebra:

$$
\begin{aligned}
dz_1 \wedge ... \wedge dz_{k-1} \wedge dz_k &= (dz_1 \wedge ... \wedge dz_{k-1}) \wedge (dv - w) \\
&= ... \\
&= v^{k-1} dx_1 \wedge ... \wedge dx_{k-1} \wedge dv
\end{aligned}
$$

Filling into the expression gives:

$$
\begin{aligned}
f_z(z,\alpha)dz_1...dz_k &\propto (x_1 v)^{\alpha_1-1}...(x_{k-1}v)^{\alpha_{k-1}-1}(v(1-[x_1+...+x_{k-1}]))^{\alpha_k-1}e^{-v^{k-1}}dx_1 \wedge ... \wedge dx_{k-1} \wedge dv \\
&= v^{\alpha_1+...+\alpha_{k-1}}e^{-v}dv \left(x_1^{\alpha_1-1}...x_{k-1}^{\alpha_{k-1}-1}\right)\left(1-\sum_{i=1}^{k-1}x_i\right)^{\alpha_{k-1}} dx_1...dx_{k-1}dv
\end{aligned}
$$

Problem C

A toy Bayesian model: Birthdays

The probability that two or more students has their birthday on the same day can be simulated as follows: Randomly assign a birth date to the 35 in a given NTNU class, and check for duplicate dates. If so, count this event as a success. If not, assign dates randomly again and to the same check. Do this many times. The estimate is then the number of successes divided by the number of trials.

1: Independent birthdays, equally likely

```
# Birthdays
sim <- 1000000
stud <- 35
count <- 0
```

```r
for (i in 1:sim){
  bdays <- round(runif(stud)*365)
  if (sum(duplicated(bdays))>=1){
    count <- count + 1
  }
}

prob <- count/sim
print(prob)
```

```
## [1] 0.813821
```

Exact calculation: The probability that two or more students has their birthday on the same day, is the complement of the event that no students has their birthday on the same day.

$$
\begin{aligned}
P(\text{no. of students with same birthday} \geq 2) &= 1 - P(\text{no students with same birthday}) \\
&= 1 - \frac{365!}{330!\, 365^{35}} = 0.8144
\end{aligned}
$$

The difference between the simulated and exact answer is of order 10^{-3} or better with enough simulations. Thus, the simulated probability is quite good.

2: Bayesian model

a)

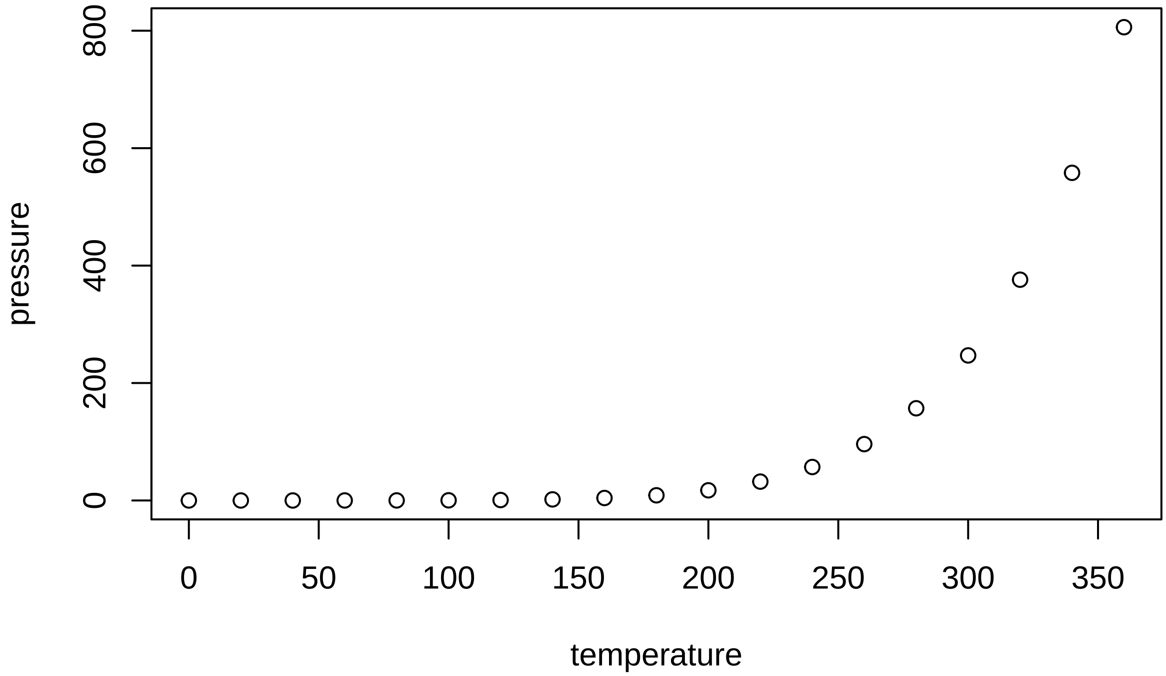
The posterior distribution of (q_1, q_2, q_3, q_4) is

$$
\begin{aligned}
P(q_{1:4}) &\propto P(x_{1:4}|q_{1:4})P(x_{1:4}) \\
&\propto \prod_{i=1}^{k} q_i^{x_i} \cdot \prod_{i=1}^{k} q_i^{\alpha_i - 1} \\
&\propto \prod_{i=1}^{k} q_i^{x_i + \alpha_i - 1} \\
&\propto \prod_{i=1}^{k} q_i^{\bar{\alpha}_i - 1}
\end{aligned}
$$

where $(x_{1:4}) = (x_1, x_2, x_3, x_4)$ and similary for q. Thus, the posterior distribution is Dirichlet?? with parameters $\bar{\alpha}i = x_i + \alpha_i$.

b)

The posterior distribution of (q_1, q_2, q_3, q_4) is Dirichlet with parameters $\bar{\alpha_i}$, $i = 1, ..., 4$.



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.