Computational Statistics (732A90) Lab6

Anubhav Dikshit(anudi287) and Thijs Quast(thiqu264)

2 March 2019

Contents

Question 1: Genetic algorithm

In this assignment, you will try to perform one-dimensional maximization with the help of a genetic algorithm.

1. Define the function

$$f(x) := \frac{x^2}{e^x} - 2e^{-}\frac{(9sinx)}{(x^2+x+1)}$$

```
f <- function(x){
((x^2)/exp(x))-2*exp(-9*sin(x)/(x^2+x+1))
}
```

2. Define the function crossover(): for two scalars x and y it returns their "kid" as $\frac{(x+y)}{2}$.

```
crossover <- function(x,y){
(x+y)/2
}
```

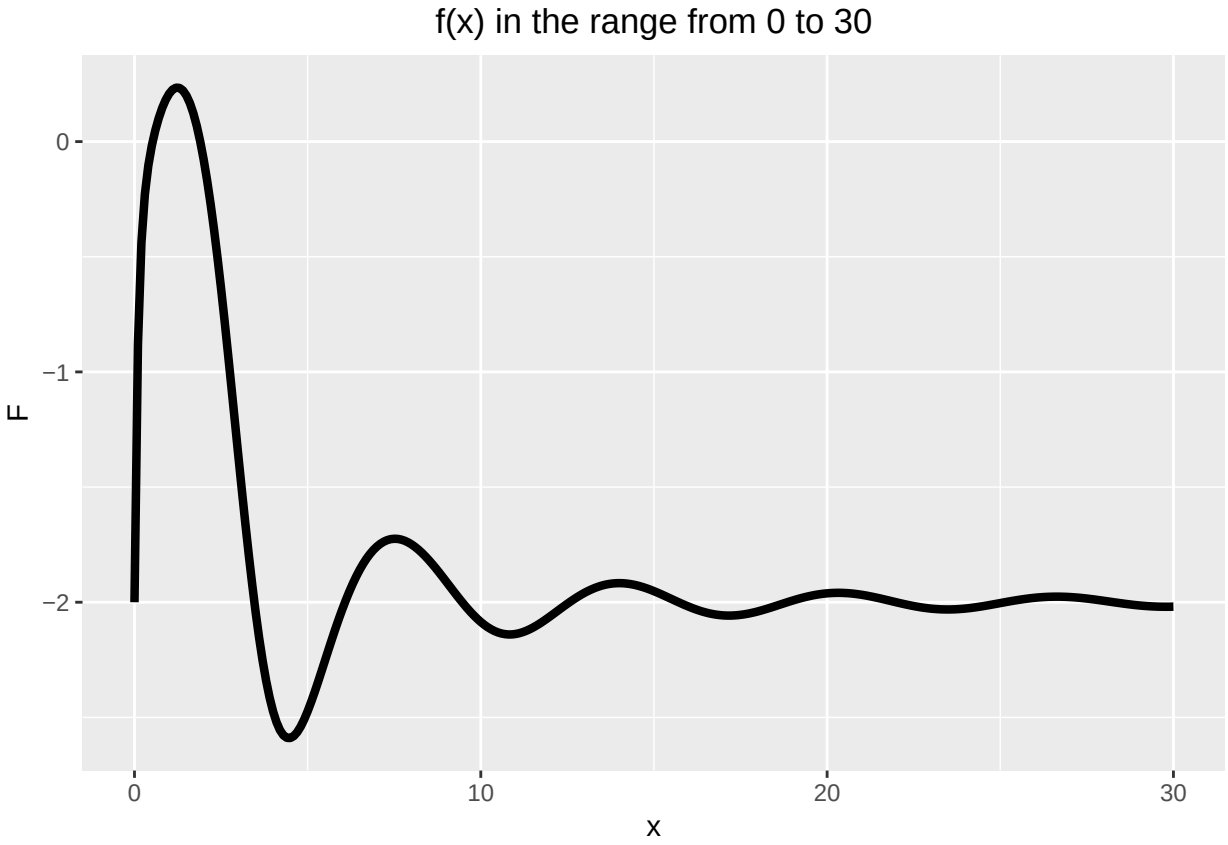
3. Define the function mutate() that for a scalar x returns the result of the integer division x2 mod 30. (Operation mod is denoted in R as %%).

```
mutate <- function(x){
x^2%%30
}
```

4. Write a function that depends on the parameters maxiter and mutprob and:

(a) Plots function f in the range from 0 to 30. Do you see any maximum value?

```
df = data.frame(x = seq(0,30,0.1),F = sapply(X = seq(0,30,0.1),FUN = f))
ggplot(data = df)+
geom_line(mapping = aes(x = x,y = F),
color = 'black',
size = 1.5)+
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```




```
cat("The maximum values of the function is at x = 2 and the value is =",df[which.max(df$F),2])
```

```
## The maximum values of the function is at x = 2 and the value is = 0.2341007
```

(b) Defines an initial population for the genetic algorithm as X = (0,5,10,15. . . ..30).

(c) Computes vector Values that contains the function values for each population point.

(d) Performs maxiter iterations where at each iteration

- Two indexes are randomly sampled from the current population, they are further used as parents (use sample()).
- One index with the smallest objective function is selected from the current population, the point is referred to as victim (use order()).
- Parents are used to produce a new kid by crossover. Mutate this kid with probability mutprob (use crossover(), mutate()).
- The victim is replaced by the kid in the population and the vector Values is updated.
- The current maximal value of the objective function is saved.

(e) Add the final observations to the current plot in another colour.

```
GA <- function(maxiter,mutprob){
population <- data.frame(X = seq(0,30,5),Values = sapply(X = seq(0,30,5),FUN = f))
maximums <- data.frame(X = 0,Values = 0)
for(i in 1:maxiter){
parents <- sample(x = population$X,size = 2)
victim <- population[which.min(population$Values),]
cross_kid <- crossover(parents[1],parents[2])
new_member <- ifelse(runif(1)<= mutprob,mutate(cross_kid),cross_kid)
population[which(population$X == victim$X)[1],] <- c(new_member,f(new_member))
maximums[i,] <- population[which.max(population$Values),]
}
unique(maximums)
}
```

5. Run your code with different combinations of maxiter= 10,100 and mutprob= 0.1,0.5,0.9. Observe the initial population and final population. Conclusions?

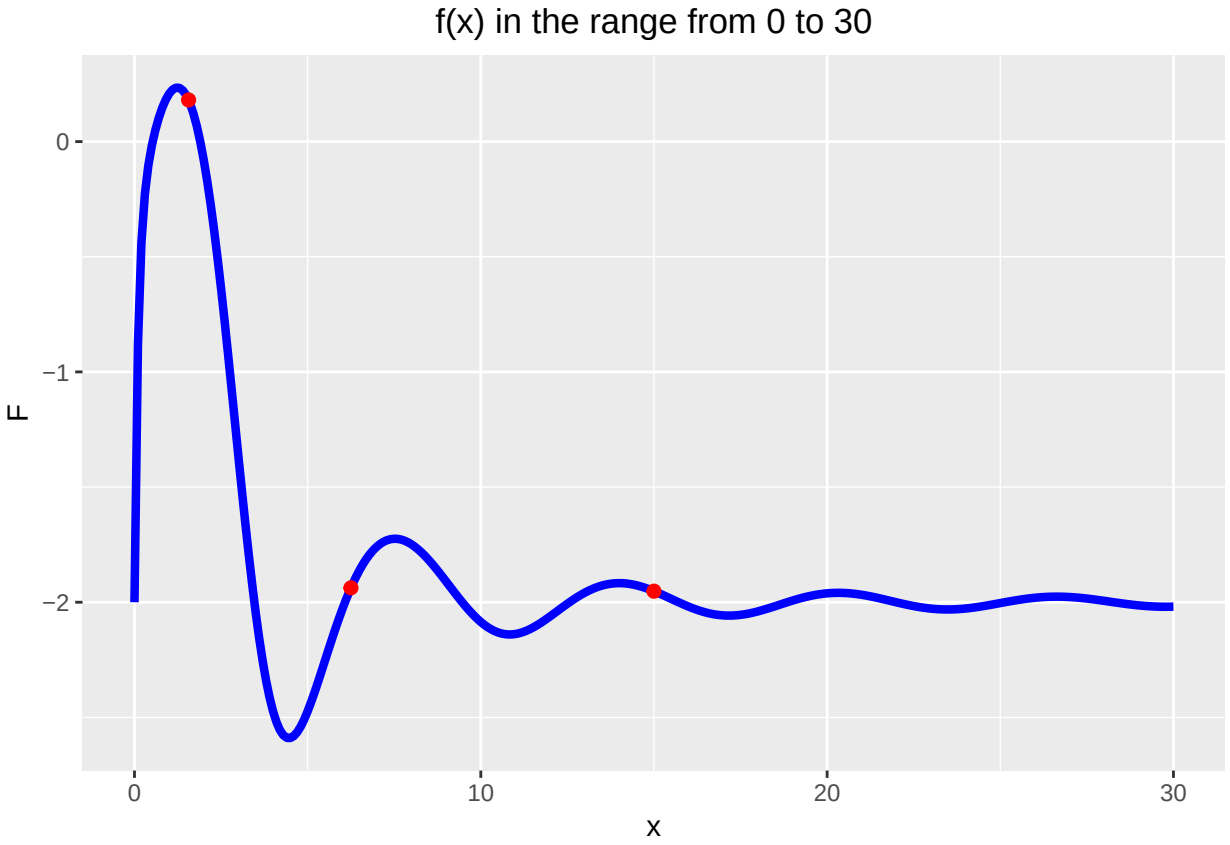
```
df1 <- GA(10,0.1)
df1

ggplot()+
geom_line(data = df ,mapping = aes(x = x,y = F),color = 'blue',size = 1.5)+
geom_point(data = df1 ,mapping = aes(x = X,y = Values),color = 'red', size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```

```
df2 <- GA(10,0.5)
df2
```

```
##          X     Values
## 1  15.0000 -1.9519470
## 5   6.2500 -1.9375289
## 10  1.5625  0.1806634
```

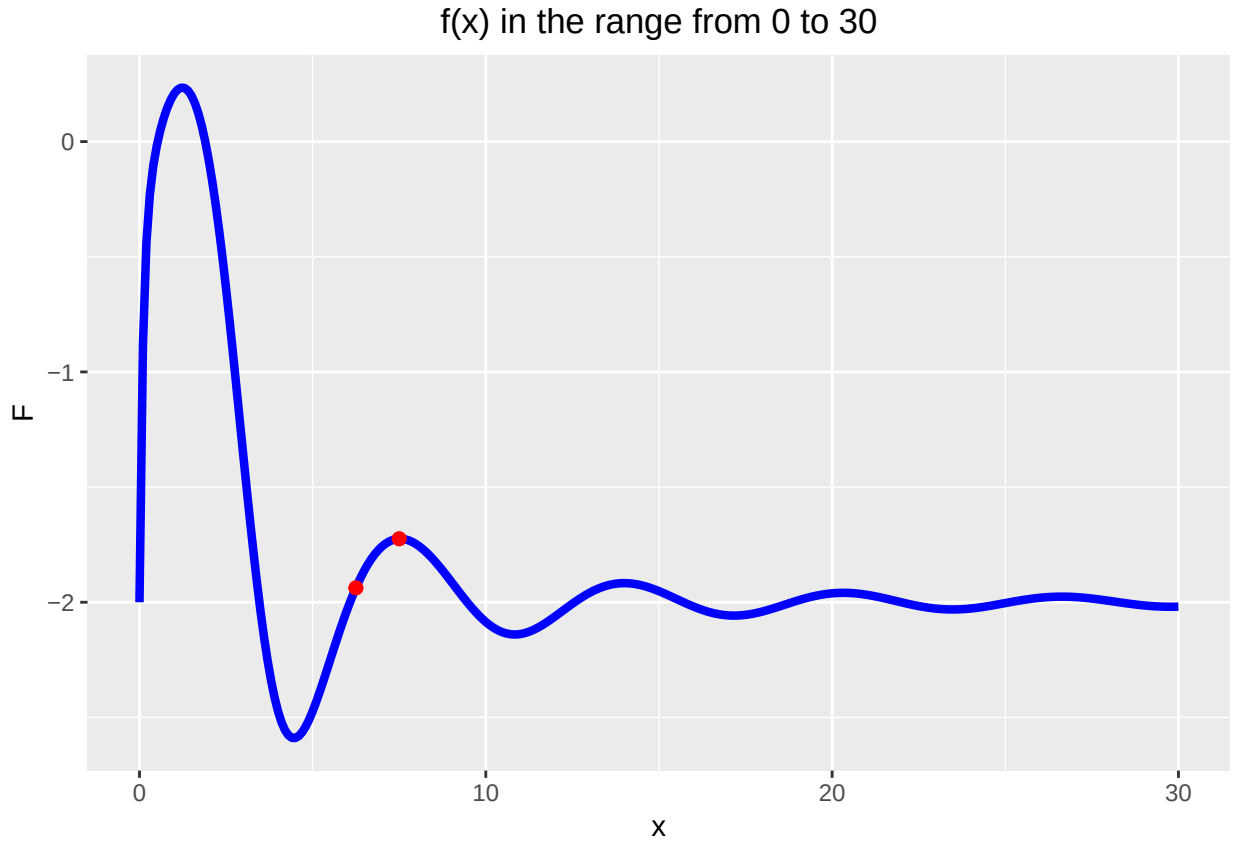
```
ggplot()+geom_line(data = df,mapping = aes(x = x,y = F),color = 'blue',size = 1.5)+
geom_point(data = df2,mapping = aes(x = X,y = Values),
color = 'red',size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```




```
df3 <- GA(10,0.9)
df3
```

```
##      X    Values
## 1 6.25 -1.937529
## 7 7.50 -1.724415
```

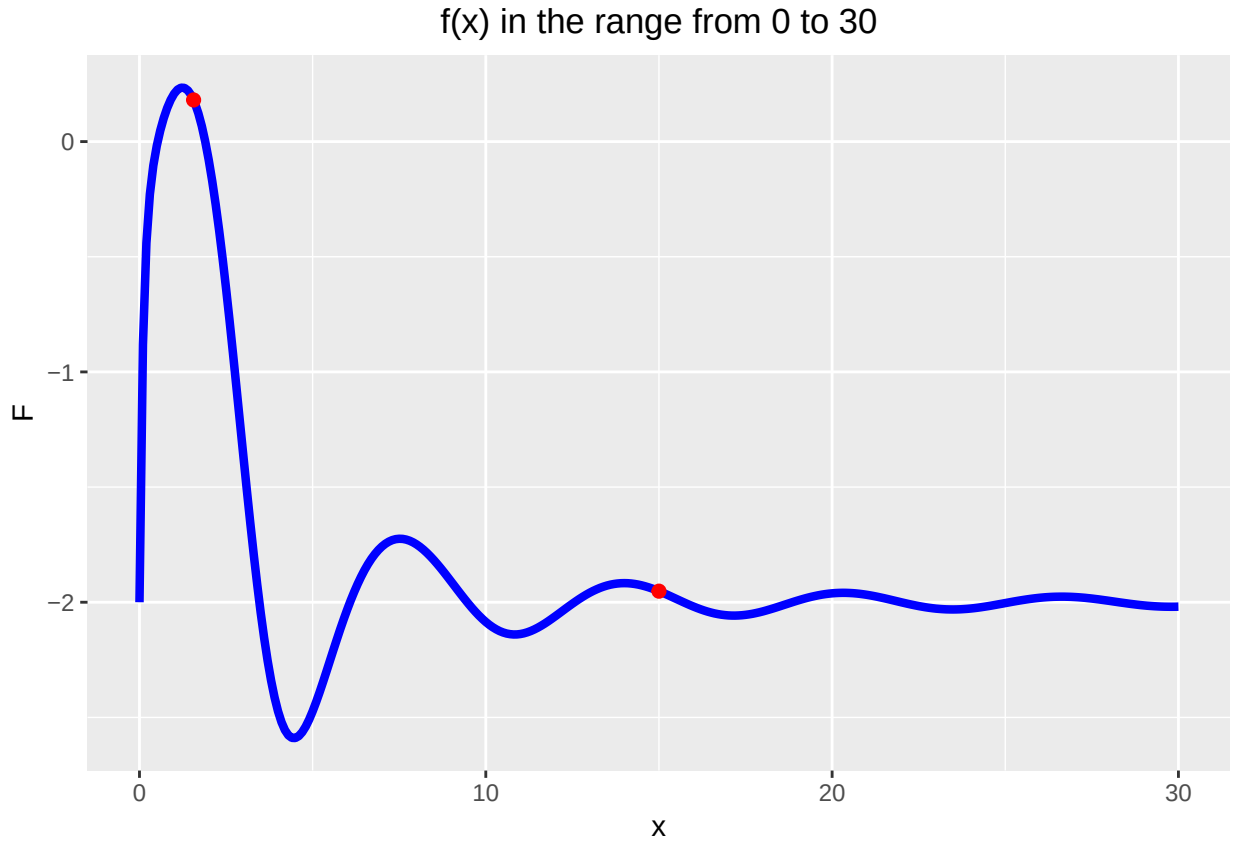
```
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df3,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```




```
df4 <- GA(100,0.1)
df4
```

```
##         X     Values
## 1 15.0000 -1.9519470
## 6  1.5625  0.1806634
```

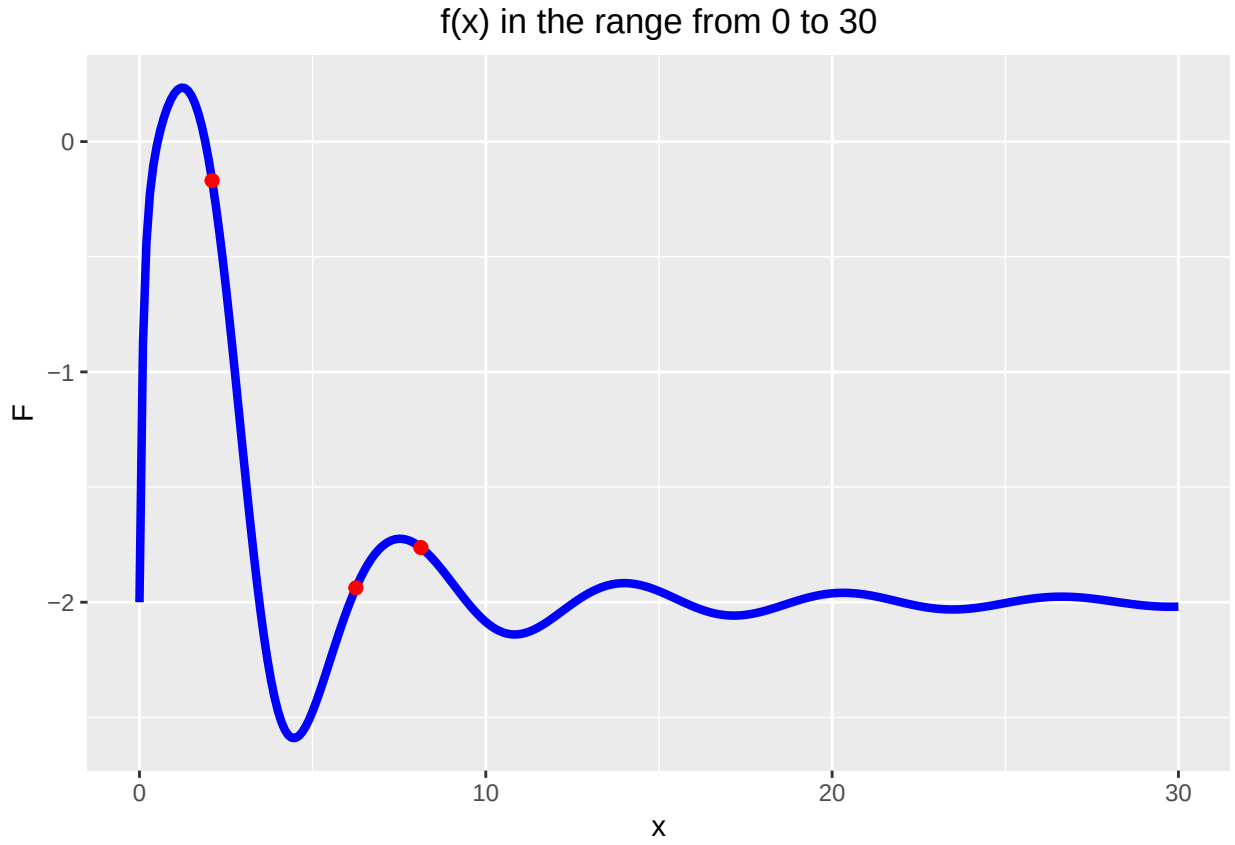
```
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df4,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```




```
df5 <- GA(100,0.5)
df5
```

```
##           X     Values
## 1  6.250000 -1.9375289
## 3  8.125000 -1.7624696
## 20 2.098516 -0.1693645
```

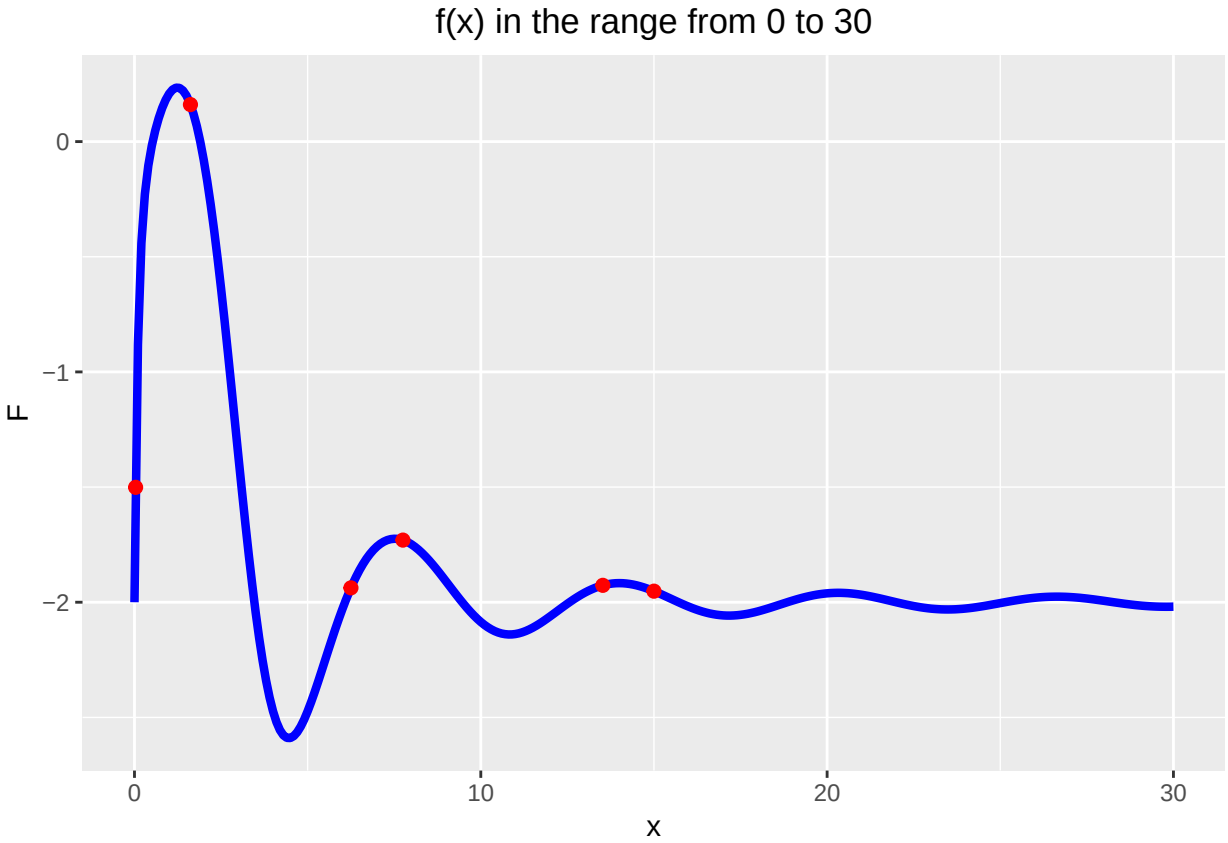
```
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df5,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```




```
df6 <- GA(100,0.9)
df6
```

```
##              X     Values
## 1  15.00000000 -1.9519470
## 7   6.25000000 -1.9375289
## 13 13.52385244 -1.9265684
## 16  7.75131007 -1.7302381
## 47  0.03289043 -1.5011232
## 85  1.61616854  0.1606462
```

```
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df6,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
```



Analysis: Increasing the number of iterations allows the genetic algorithm to converge to the given functions. The effect of probability is to only ensure that some good parent are not lost due to crossover.

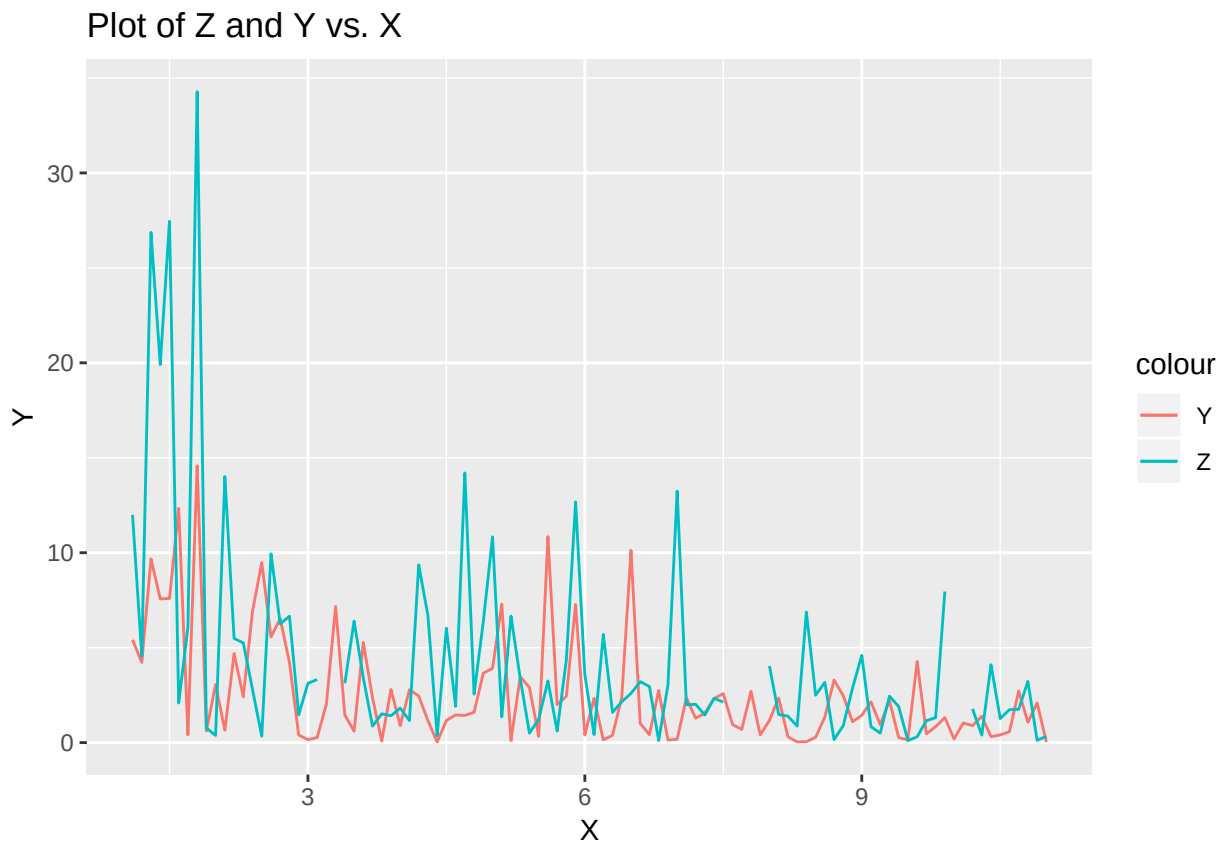
2. EM algorithm

The data file physical.csv describes a behavior of two related physical processes Y = Y (X) and Z = Z(X).

1. Make a time series plot describing dependence of Z and Y versus X. Does it seem that two processes are related to each other? What can you say about the variation of the response values with respect to X?

```
data <- read.csv(file="physical1.csv")

ggplot(data=data,aes(x=X, group=1)) +
  geom_line(aes(y = Y, colour = "Y")) +
  geom_line(aes(y = Z, colour = "Z")) +
  ggtitle("Plot of Z and Y vs. X")
```



Analysis: There is a difference in the amplitude of 'Y' and 'Z' with respect to 'X.' Both of the series vary similarily and show similar patterns of decayed oscillation, however 'Z' has faster decay than 'Y'.

2. Note that there are some missing values of Z in the data which implies problems in estimating models by maximum likelihood. Use the following model

$$Y_i \approx exp(\frac{X_i}{\lambda}), \quad Z_i \approx exp(\frac{X_i}{2*\lambda})$$

Where λ is an unknown parameters. The goal is to derive the EM algorithm that estimates λ.

$$\begin{aligned} L(\lambda|Y,Z) &= \prod_{i=1}^{n} f(Y) \times \prod_{i=1}^{n} f(Z) \\ &= \prod_{i=1}^{n} \frac{X_i}{\lambda} \cdot e^{-\frac{X_i}{\lambda}Y_i} \times \prod_{i=1}^{n} \frac{X_i}{2\lambda} \cdot e^{-\frac{X_i}{\lambda}Z_i} \\ &= \frac{X_1 \cdot \ldots \cdot X_n}{\lambda^n} \times e^{-\frac{1}{\lambda}\sum_1^n X_i Y_i} \times \frac{X_1 \cdot \ldots \cdot X_n}{(2\lambda)^n} \times e^{-\frac{1}{2\lambda}\sum_1^n X_i Z_i} \\ lnL(\lambda|Y,Z) &= \sum_{i=1}^{n} ln(X_i) - nln(\lambda) - \frac{1}{\lambda}\sum_{i=1}^{n} X_i Y_i \ + \sum_{i=1}^{n} ln(X_i) - nln(2\lambda) - \frac{1}{2\lambda}\sum_{i=1}^{n} X_i Z_i \end{aligned}$$

E-step : Derive Q function

Obtaining the expected values for the missing data using an initial parameter estimate.

$$
\begin{aligned}
Q(\theta, \theta^k) &= E[\ loglik(\lambda|Y,Z) \mid \lambda^k, (Y,Z)] \\
&= \sum_{i=1}^{n} ln(X_i) - nln(\lambda) - \frac{1}{\lambda}\sum_{i=1}^{n} X_i Y_i \ + \sum_{i=1}^{n} ln(X_i) - nln(2\lambda) \\
&\quad - \frac{1}{2\lambda}\left[\sum_{i=1}^{n} X_i Z_i \ + m \cdot X_i \cdot \frac{2\lambda_{k-1}}{X_i}\right]
\end{aligned}
$$

Here, we are taking expectation on the missing values in Z, so we need to seperate the Z_{obs} and Z_{miss}. Here we are assuming there are 'm' missing Z values. λ_k is the lambda value from the previous iteration.

M-step

Obtain the maximum likelihood estimate of the parameters by taking the derivative with respect to λ. Repeat till estimate converges.

$$
\begin{aligned}
-\frac{n}{\lambda} - \frac{n}{\lambda} \ + \frac{\sum_{i=1}^{n} X_i Y_i}{\lambda^2} \ + \frac{\sum_{i}^{m} X_i Z_i \ + m \cdot 2\lambda_{k-1}}{2\lambda^2} &:= 0 \\
-2\lambda(2n) + 2\sum_{i=1}^{n} X_i Y_i \ + \sum_{i=1}^{n} X_i Z_i + m \cdot 2\lambda_{k-1} &:= 0
\end{aligned}
$$

$$
\lambda = \frac{\sum_{i=1}^{n} X_i Y_i + \frac{1}{2}\sum_{i=1}^{n} X_i Z_i + m \cdot \lambda_{k-1}}{2n}
$$

3. Implement this algorithm in R, use $\lambda = 100$ and convergence criterion "stop if the change in λ is less than 0.001". What is the optimal λ and how many iterations were required to compute it?

```
my_EM <- function(data,eps,kmax,lamb_0){

        X <- data$X
        Y <- data$Y
        Z <- data$Z

        Xobs <- X[!is.na(Z)]
        Zobs <- Z[!is.na(Z)]
        Zmiss <- Z[is.na(Z)]

        n <- length(X)
        m <- length(Zmiss)

        k <<- 0
        llvalprev <- 0
        llvalcurr <- lamb_0

        print(c(llvalprev,llvalcurr,k))

        while ((abs(llvalprev-llvalcurr)>eps) && (k<(kmax+1))){
```

```
                llvalprev <- llvalcurr
                llvalcurr <- (sum(X*Y)+sum(Xobs*Zobs)/2+m*llvalprev)/(2*n)

                k <<- k+1
        }

        print(c(llvalprev,llvalcurr,k))
}
```

```
my_EM(data,0.001,50,100)
```

```
## [1]   0 100   0
## [1] 10.69587 10.69566  5.00000
```

Analysis: The result indicates that the optimal lambda is 10.69566 and after 5 iteration are needed for convergence.

4. Plot E[Y] and E[Z] versus X in the same plot as Y and Z versus X. Comment whether the computed λ seems to be reasonable.

Following the given model for Y and Z, we can easily derive the mean value with obtained.

$$E[Y] = \frac{\lambda}{X_i} \ , \quad E[Z] = \frac{2\lambda}{X_i}$$

```
lambda <- 10.69566

new_data <- data
new_data$E_Y <- lambda/data$X
new_data$E_Z <- 2*lambda/data$X

ggplot(data=new_data,aes(x=X, group=1)) +
  geom_line(aes(y = Y, colour = "Y")) +
  geom_line(aes(y = Z, colour = "Z")) +
  geom_line(aes(y = E_Y, colour = "E_Y")) +
  geom_line(aes(y = E_Z, colour = "E_Z")) +
  ggtitle("Plot of Y,Z and their expected value vs. X")
```

Analysis: From the plot above, we can see that each E[Z] and E[Y] captures the flow of Z and Y on X respectively. So we can say that our computed lambda is reasonable enough.

Appendix

```
knitr::opts_chunk$set(echo = TRUE)
options(scipen=999)
options(stringsAsFactors = FALSE)
library(dplyr)
library(ggplot2)
library(knitr)
set.seed(12345)

f <- function(x){
((x^2)/exp(x))-2*exp(-9*sin(x)/(x^2+x+1))
}

crossover <- function(x,y){
(x+y)/2
}


mutate <- function(x){
x^2%%30
```

```
}

df = data.frame(x = seq(0,30,0.1),F = sapply(X = seq(0,30,0.1),FUN = f))
ggplot(data = df)+
geom_line(mapping = aes(x = x,y = F),
color = 'black',
size = 1.5)+
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))

cat("The maximum values of the function is at x = 2 and the value is =",df[which.max(df$F),2])
GA <- function(maxiter,mutprob){
population <- data.frame(X = seq(0,30,5),Values = sapply(X = seq(0,30,5),FUN = f))
maximums <- data.frame(X = 0,Values = 0)
for(i in 1:maxiter){
parents <- sample(x = population$X,size = 2)
victim <- population[which.min(population$Values),]
cross_kid <- crossover(parents[1],parents[2])
new_member <- ifelse(runif(1)<= mutprob,mutate(cross_kid),cross_kid)
population[which(population$X == victim$X)[1],] <- c(new_member,f(new_member))
maximums[i,] <- population[which.max(population$Values),]
}
unique(maximums)
}

df1 <- GA(10,0.1)
df1

ggplot()+
geom_line(data = df ,mapping = aes(x = x,y = F),color = 'blue',size = 1.5)+
geom_point(data = df1 ,mapping = aes(x = X,y = Values),color = 'red', size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
df2 <- GA(10,0.5)
df2
ggplot()+geom_line(data = df,mapping = aes(x = x,y = F),color = 'blue',size = 1.5)+
geom_point(data = df2,mapping = aes(x = X,y = Values),
color = 'red',size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
df3 <- GA(10,0.9)
df3
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df3,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
df4 <- GA(100,0.1)
df4
```

```
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df4,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
df5 <- GA(100,0.5)
df5
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df5,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))
df6 <- GA(100,0.9)
df6
ggplot()+
geom_line(data = df,mapping = aes(x = x,y = F),
color = 'blue',
size = 1.5)+
geom_point(data = df6,mapping = aes(x = X,y = Values),
color = 'red',
size = 2) +
ggtitle('f(x) in the range from 0 to 30') +
theme(plot.title = element_text(hjust = 0.5))

data <- read.csv(file="physical1.csv")

ggplot(data=data,aes(x=X, group=1)) +
  geom_line(aes(y = Y, colour = "Y")) +
  geom_line(aes(y = Z, colour = "Z")) +
  ggtitle("Plot of Z and Y vs. X")

my_EM <- function(data,eps,kmax,lamb_0){

        X <- data$X
        Y <- data$Y
        Z <- data$Z

        Xobs <- X[!is.na(Z)]
        Zobs <- Z[!is.na(Z)]
        Zmiss <- Z[is.na(Z)]

        n <- length(X)
        m <- length(Zmiss)

        k <<- 0
```

```
        llvalprev <- 0
        llvalcurr <- lamb_0

        print(c(llvalprev,llvalcurr,k))

        while ((abs(llvalprev-llvalcurr)>eps) && (k<(kmax+1))){
                llvalprev <- llvalcurr
                llvalcurr <- (sum(X*Y)+sum(Xobs*Zobs)/2+m*llvalprev)/(2*n)

                k <<- k+1
        }

        print(c(llvalprev,llvalcurr,k))
}

my_EM(data,0.001,50,100)
lambda <- 10.69566

new_data <- data
new_data$E_Y <- lambda/data$X
new_data$E_Z <- 2*lambda/data$X

ggplot(data=new_data,aes(x=X, group=1)) +
  geom_line(aes(y = Y, colour = "Y")) +
  geom_line(aes(y = Z, colour = "Z")) +
  geom_line(aes(y = E_Y, colour = "E_Y")) +
  geom_line(aes(y = E_Z, colour = "E_Z")) +
  ggtitle("Plot of Y,Z and their expected value vs. X")
```