Université Paris Nanterre

English for probability and statistics

Assignment 1

Exercise 1.3

Ayman YAYA

Compute the density f of the distribution of $|X|$ where X be a $\mathcal{N}(0, 1)$ random variable. Express the cumulative distribution function F of $|X|$ in terms of the cumulative distribution function Φ of the standard normal distribution.

Let $g(t) = e^{-t}$ be the density of the standard exponential distribution $\mathcal{E}(1)$. Compute the smallest C such that $f(t) \leq Cg(t)\ \forall t \in [0, +\infty)$.

Draw the graphs of f and Cg. Simulate a 10^3-sample of the distribution of $|X|$ by the rejection algorithm. How many iterations of the algorithm were needed for each simulation? Compare this with the theoretical distribution of the number of trials.

Draw histograms and a graph of the empirical cdf of the simulated distribution with the theoretical one (computed previously) to draw the graph of F).

Solution

Let X be a standard normal distributed random variable with a mean of 0 and a standard deviation of 1. Therefore X is a symmetrical distributions and the cumulative distribution function Φ of X is defined by :

$$\Phi(t) = \int_{-\infty}^{t} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} \mathrm{dx}$$

Let's compute the density f of the distribution of $|X|$.

$$\begin{aligned} P(|X| \leq t) &= P(-t < X \leq t) \\ &= P(X \leq t) - P(X \leq -t). \end{aligned}$$

By definition of the distribution of X, this is equivalent to :

$$P(|X| \leq t) = P(X \leq t) - [1 - P(X \leq t)]$$
$$P(|X| \leq t) = 2 \times P(X \leq t) - 1.$$

The derivative of an integral of a function beeing the function itself we have that density of $|X|$ is $f(x) = \frac{2e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$.

We also have that,

$$F(t) = 2\Phi(t) - 1.$$

Let $g(t) = e^{-t}$ be the density of the standard exponential distribution $\mathcal{E}(1)$, f and g are both density functions on R and we want the smallest C that verify $\forall t \in [0, +\infty)$,

$$f(t) \leq Cg(t).$$

For that, we begin by separating C and t, so that we have an expression that we can work with :

$$f(t) \leq Cg(t)$$
$$\frac{f(t)}{g(t)} \leq C.$$

The last statement is the same as saying that the smallest C that verify $C = \sup_t \frac{f(t)}{g(t)}$.

Let $A(t) = \frac{f(t)}{g(t)}$, then :

$$A(t) = \frac{f(t)}{g(t)} = \frac{\frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}}{e^{-t}}$$
$$= \frac{2e^{-\frac{t^2}{2}+t}}{\sqrt{2\pi}}$$
$$= \frac{2e^{-\frac{1}{2}(t-1)^2+\frac{1}{2}}}{\sqrt{2\pi}} = e^{-\frac{1}{2}(t-1)^2} \times \sqrt{\frac{2}{\pi}}e^{\frac{1}{2}}$$

We have now an expression of A that we can exploit. To find a solution to our original problem we must study $A(t)$:

$$\frac{\partial A}{\partial t} = -(t-1)\sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}(t-1)^2}\, e^{\frac{1}{2}}$$

Hence the function A has one critical point at t = 1, and is increasing $\forall t \in [0, 1)$ and decreasing $\forall t \in (1, +\infty)$. So A has an absolute maximum in t=1. The absolute maximum value of $A(t)$ is : $A(1) = \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}}$.

We can conclude that

$$C = \sup_t A(t) = \sqrt{\frac{2}{\pi}} e^{\frac{1}{2}}$$

We can now draw the graph of f as well as the graph of Cg :

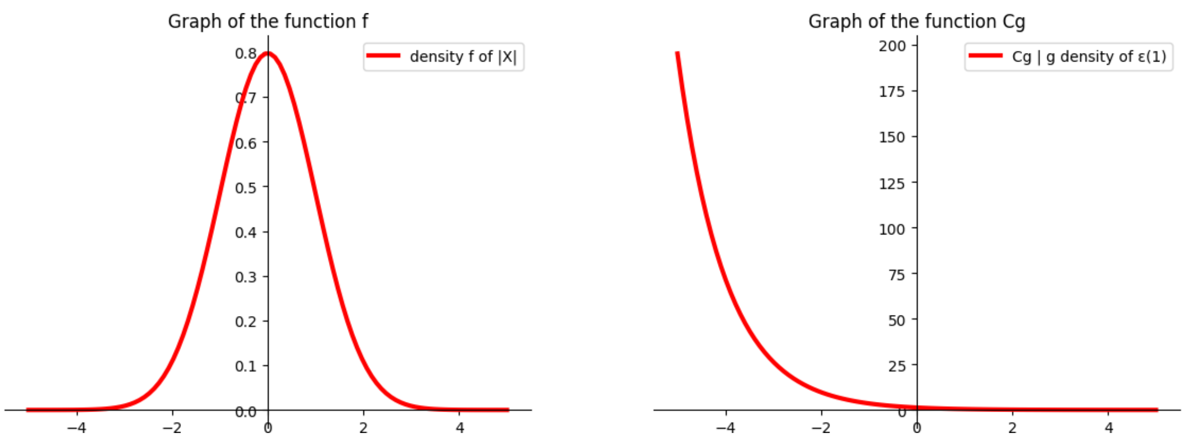

Figure 1: Graphs of the functions f and Cg

As we can see f represents a normal density with mean 0. Cg also represents an exponential density with parameter 1 multiplied by a constant C.

Now we will simulate the distribution of $|X|$ with density $f(t) = \frac{2e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}$ via the rejection method. As an instrumental density, we have the density g on $\mathbb{R}^+$ defined by $g(t) = e^{-t}$. The quantile function of g is thus $Q(s) = -log(1-s)$, therefore, if U uniformly distributed on $[0, 1]$, then

$Q(U) = -log(1 - U)$ have a standard exponential distribution.

Let X_i be a sequence of i.i.d. random variables with density f. Let $\{Y, Y_i, i \geq 1\}$ be a sequence of i.i.d. random variables with density g and, so the density of Y is g. Let $\{U, U_i, i \geq 1\}$ be an independent sequence of i.i.d. random variables uniformly distributed on $[0, 1]$. Define

$$N = \inf\{i \leq 1 : Cg(Y_i)U_i \leq f(Y_i)\}$$

Set $X = Y_N$. Then

$$\begin{aligned} \mathrm{E}[h(X)] = \mathrm{E}[h(Y_N)] &= \sum_{k=1}^{\infty} \mathrm{E}[h(Y)]1\{Cg(X_k)U_k \leq f(X_k)\}(1-p)^{k-1} \\ &= \sum_{k=1}^{\infty} \mathrm{E}[h(Y)\frac{f(Y)}{Cg(Y)}](1-p)^{k-1} \\ &= p^{-1}\mathrm{E}[h(Y)\frac{f(Y)}{Cg(Y)}] = C\int_{-\infty}^{\infty} h(x)\frac{f(x)}{Cg(x)}g(x)\mathrm{d}x = \int_{-\infty}^{\infty} h(x)f(x)\mathrm{d}x \end{aligned}$$

Thus has density f as desired. Furthermore, N has a geometric distribution with parameter p given by

$$p = \mathrm{P}(Cg(Y)U \leq f(Y)) = \mathrm{E}[\frac{f(Y)}{Cg(Y)}] = \frac{1}{C}\int_{-\infty}^{\infty}\frac{f(x)}{g(x)}g(x)\mathrm{d}x = \frac{1}{C}.$$

This means that the mean number of trials per realization is C.

We can now apply the rejection algorithm to simulate $F(\cdot)$.

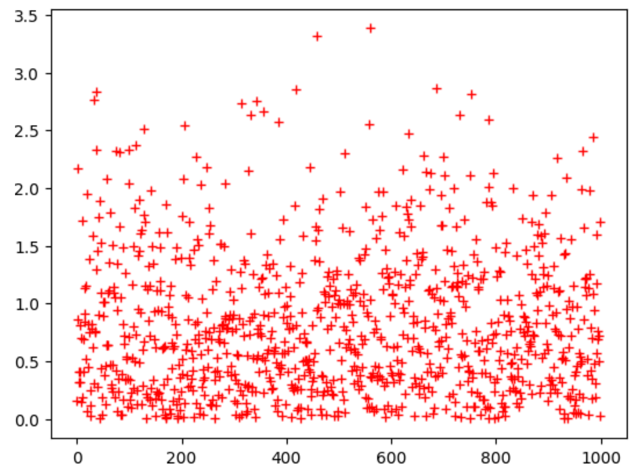

Figure 2: 1000-sample of the distribution of $|X|$, simulated by the rejection algorithm

Figure 2 shows the 1000 simulated random variables that were made using the rejection algorithm described previously. As we can see, a large majority of them are valued between 0 and 1. We can then draw the histrogram representing our simulation of a 1000 sample of $|X|$:

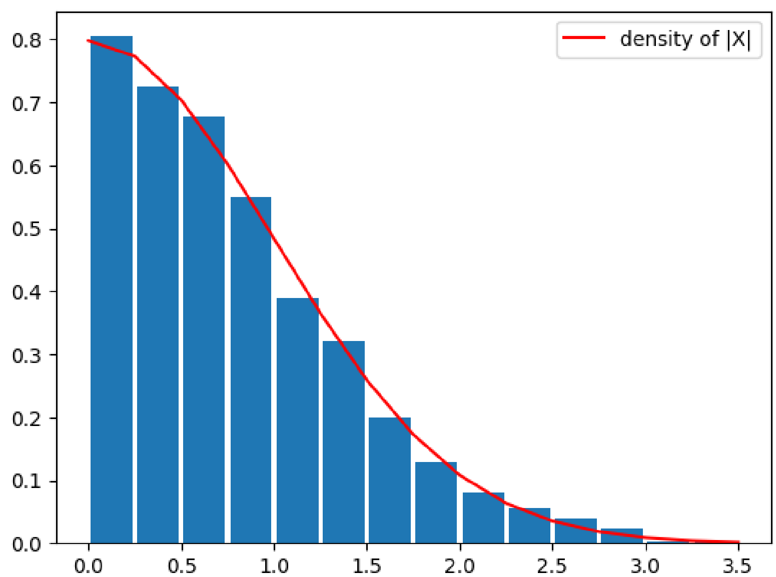

Figure 3: Histogram of of the simulated distribution of $|X|$

In this histogram we can see that our simulation seems to have worked properly as the red graph that represents the density of the random variable $|X|$ follows the same decreasing trajectory of the histogram.

Here we have the histogram that shows how many iterations of the algorithm were needed for each simulation:

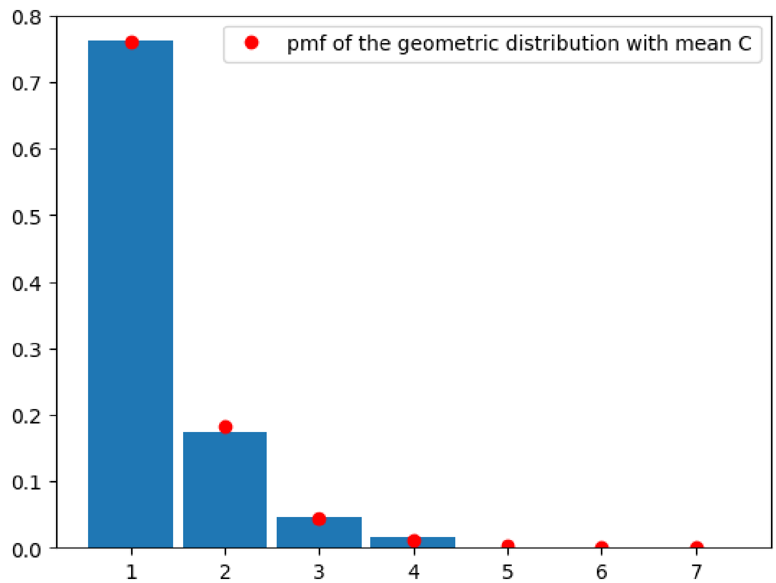

Figure 4: Number of iterations of the algorithm for each simulation

This tells us that in around 75% of the simulations, only one iteration of the algorithm was needed. In around 18% to 19% of the simultaion 2 iterations were needed, 3 iterations in around 5% of the simulations and a very small percentage of the simulations (around 2%) require up to 4 iterations. No simulation needed more than 4 iterations however. This more or less corresponds to the theoretical distribution of the number of trials represented by the red dots that represent the probability mass function of a geometric distribution with mean C. We know that the the mean number of trials per realization is C= $\sqrt{\frac{2}{\pi}} e^{\frac{1}{2}} \approx 1,32$.

On top of that we can also define the empirical cumulative distribution function of the simulated distribution. It is defined as

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \leq t}$$

Finally we know that: $F(t) = P(|X| \leq t) = 2 \times P(X \leq t) - 1$ Where $P(X \leq t)$ is the cumulative distribution function of a $\mathcal{N}(0, 1)$ random variable. But we already know the density of f wich we previously computed so F is simply the primitive of f.

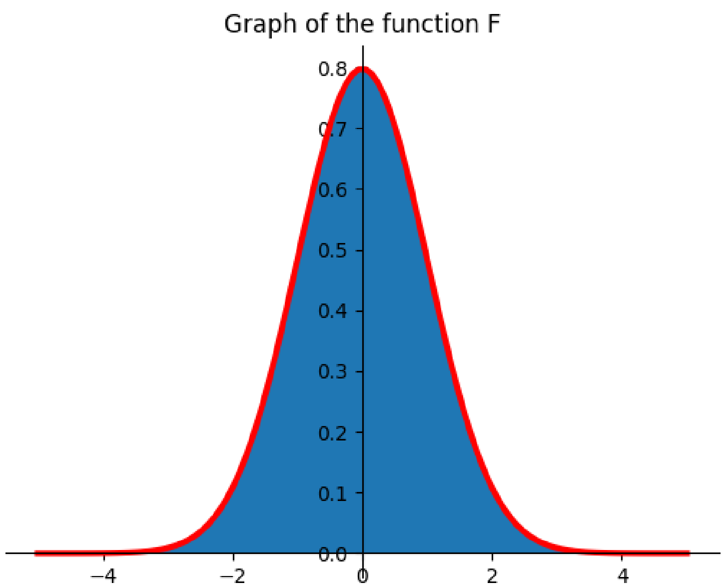

Figure 5: Graph of the cdf function F (colored in blue)

Appendix: python code

```
import matplotlib.pyplot as plt

import numpy as np

    ##Creation of the graph of the function f

x = np.linspace(-5,5,100)

y = (2*np.exp(-(x**2)/2)/np.sqrt(2*np.pi))

fig = plt.figure()

ax = fig.add_subplot(1, 1, 1)

ax.spines['left'].set_position('center')

ax.spines['bottom'].set_position('zero')

ax.spines['right'].set_color('none')

ax.spines['top'].set_color('none')

ax.xaxis.set_ticks_position('bottom')

ax.yaxis.set_ticks_position('left')

plt.plot(x,y, 'r',linewidth = 3)

plt.legend(["density f of |X|"], loc=0, frameon=True)

plt.title('Graph of the function f')

plt.show()


    ###Creation of the graph of the function Cg


z=np.linspace(-5,5,100)

t=np.sqrt(2/np.pi)*np.exp(1/2)*np.exp(-z)

fig2 = plt.figure()

ax2 = fig2.add_subplot(1, 1, 1)
```

```
ax2.spines['left'].set_position('center')
ax2.spines['bottom'].set_position('zero')
ax2.spines['right'].set_color('none')
ax2.spines['top'].set_color('none')
ax2.xaxis.set_ticks_position('bottom')
ax2.yaxis.set_ticks_position('left')
plt.plot(z,t, 'r',linewidth = 3)
plt.legend(["Cg | g density of (1)"], loc=0, frameon=True)
plt.title('Graph of the function Cg')
plt.show()


    ###Simulation of a 1000 sample of the distribution of |X| by the rejection algorithm
def norm_a(x, mu=0, sigma=1):
    return (2 / (sigma * (2 * np.pi)**0.5)) * np.exp(-0.5 * ((x - mu)**2 / sigma**2))
for i in range(K):
    counter[i]=1
    u=npr.random()
    y=-np.log(1-npr.random())
    while u>(np.exp(-0.5*(y-1)**2)):
        counter[i]=counter[i]+1
        u=npr.random()
        y=-np.log(1-npr.random())
    x[i]=y
```

###sample of 1000 of the distribution of |X|

```
plt.plot(x,"r+")
plt.show()
```

###histogram of the simulated sample

```
bb=np.arange(start=min(x),step=.25,stop=max(x))
plt.hist(x,bins=bb,density=True,rwidth=.9)
plt.plot(bb,norm_a(bb),"r",label="density of |X|")
plt.legend()
```

###histogram of the number of trials

```
from scipy.stats import geom
p=1/(np.sqrt(2/np.pi)*np.exp(1/2))
bns=np.arange(start=.5,step=1,stop=max(counter)+1)
plt.hist(counter,bins=bns,density=True,rwidth=.9)
nns=np.arange(1,max(counter)+1)
plt.plot(nns,geom.pmf(nns,p),"ro",label=r"pmf of the geometric distribution with mean C")
plt.legend()
```