ENGINEERING TRIPOS PART IIA

EIETL

MODULE EXPERIMENT 3F3

RANDOM VARIABLES and RANDOM NUMBER GENERATION

H. Zhou

Jesus College

hz324

1 Uniform and normal random variables.

Histogram of Gaussian random numbers overlaid on exact Gaussian curve (scaled):

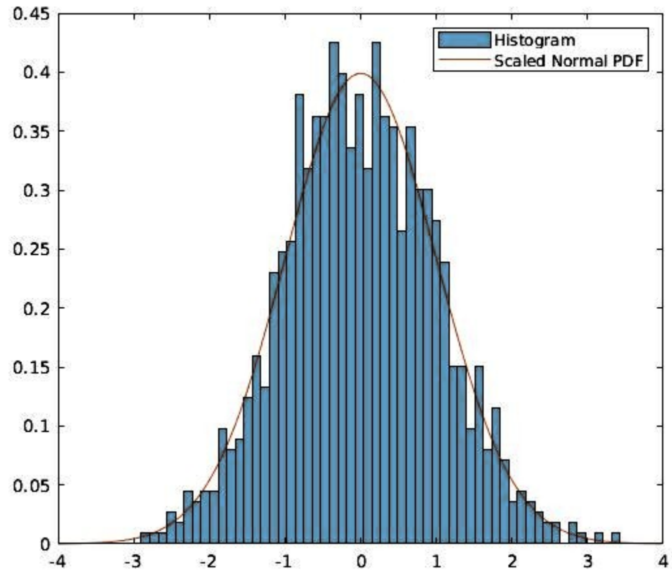

Figure 1: *Histogram of Gaussian random numbers on exact Gaussian curve (scaled)*

Histogram of Uniform random numbers overlaid on exact Uniform curve (scaled):

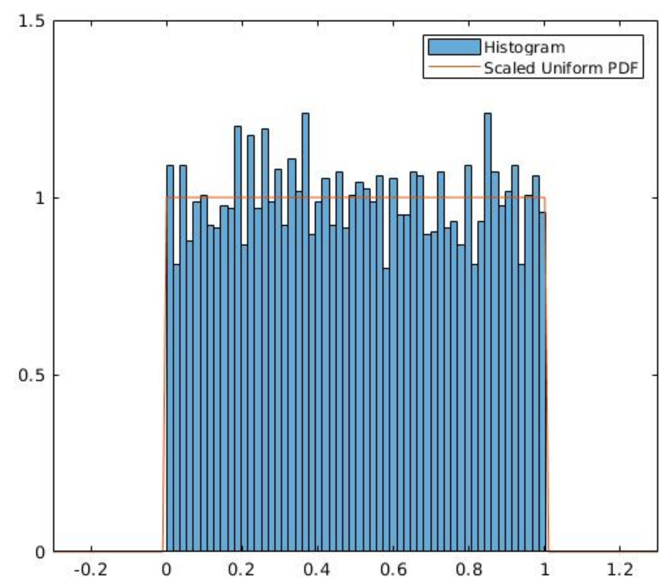

Figure 2: *Histogram of Uniform random numbers on exact Uniform curve (scaled)*

Kernel density estimate for Gaussian random numbers overlaid on exact Gaussian curve:

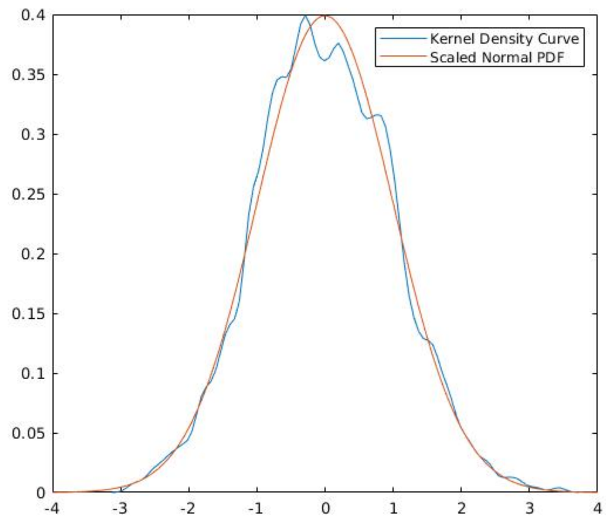

Figure 3: *Kernel density estimate for Gaussian random numbers exact Gaussian curve*

Kernel density estimate for Uniform random numbers overlaid on exact Uniform curve:

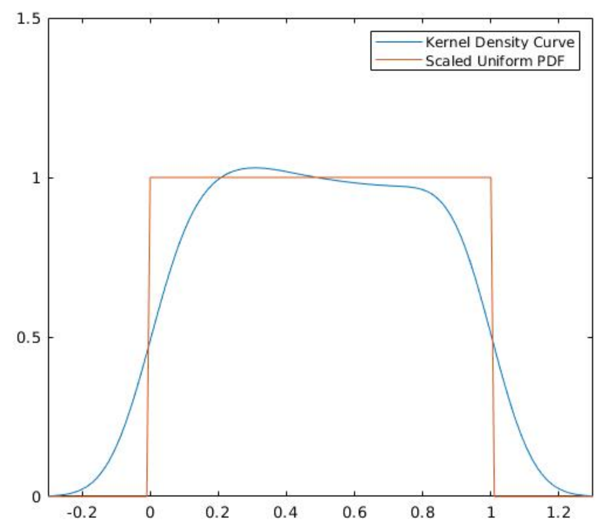

Figure 4: *Kernel density estimate for Uniform random numbers on exact Uniform curve*

Comment on the advantages and disadvantages of the kernel density method compared with the histogram method for estimation of a probability density from random samples:

The histogram has a few advantages. It, first of all, requires relatively small amount of computing power, since the complexity of generating such histogram is $O(n)$. Only a linear search through the sample space seems to be needed.

More importantly, the histogram models a discontinuous step much better than a kernel density curve. The histogram represented the uniform distribution much better than the kernel density method.

The kernel density method, however, is able to model a continuous pdf much better than that of a histogram. We can see from the figures above, that the kernel density estimation gave a much better curve that fitted the gaussian PDF.

Despite it's advantage in continuous PDFs, the kernel density estimation seems to have a complexity of $O(n^2)$, each datapoint in the sample space needs to be computed taking into all other datapoints.

This estimation method also doesn't model a step very well. In the uniform distribution case, a very flat and unrealistic gradient was seen on both of the step of a uniform PDF.

Theoretical mean and standard deviation calculation for uniform density as a function of N:

From multinomial distribution theory, we get

$$\mu = Np_j \text{ and } \sigma = \sqrt{N(1-p_j)p_j},$$

where $p_j$is the probability of a sample lies in bin j. In the uniform case,

$$mathp_j = 1/J$$

, where J is the total bin number.

Thus, the mean and standard variation are

$$\mu = N/J \text{ and } \sigma = \sqrt{N(J-1)/J^2}$$

Explain behaviour as N becomes large:

From the weak law of large numbers, the value of each bin should tend to the theoretical mean. In our case

$$\lim_{N\to\infty} P(|X_i - \mu| \geq \epsilon) = 0$$

Where X_i is the number of random variable in each bin.

This means that the variation is a lot smaller with the histogram fitting the theorectical μ line in this case.

Plot of histograms for $N = 100$, $N = 1000$ and $N = 10000$ with theoretical mean and ± 3 standard deviation lines:

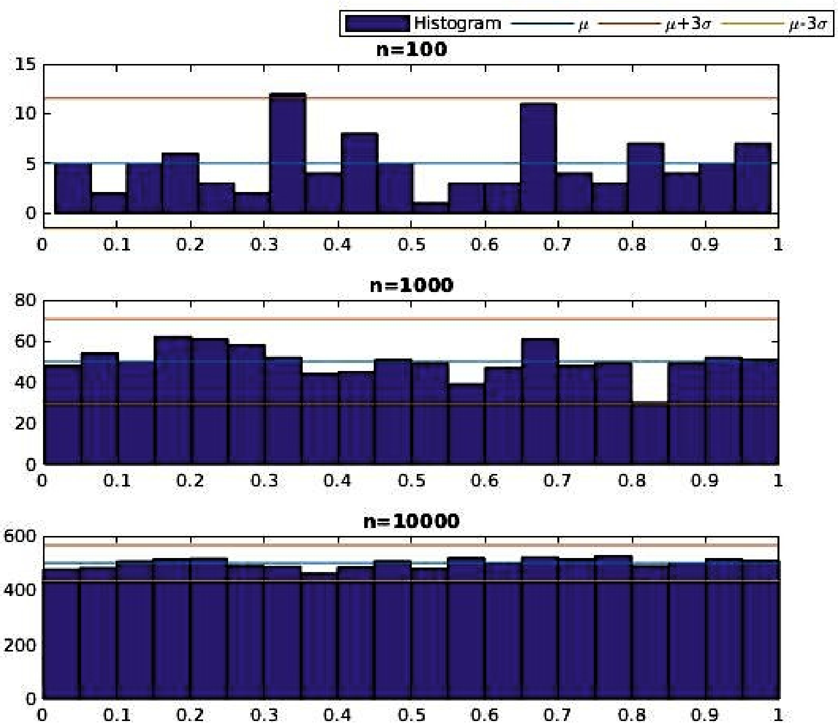

Figure 5: Histograms for $N = 100$, $N = 1000$ and $N = 10000$ with theoretical mean and ± 3 standard deviation lines

Are your histogram results consistent with the multinomial distribution theory?

All bins lied within the $\pm 3\sigma$ lines, hence it was indeed consistent with the theory.

2 Functions of random variables

For normally distributed $\mathcal{N}(x|0, 1)$ random variables, take $y = f(x) = ax + b$. Calculate $p(y)$ using the Jacobian formula:

$$\begin{aligned} p(y) &= \frac{p(x)}{|dy/dx|}\bigg|_{x=f^{-1}(y)} \\ &= \frac{N\,(x|0,1)}{|a|}\bigg|_{x=\frac{y-b}{a}} \\ &= \frac{1}{\sqrt{2\pi}|a|}\exp\left(-\frac{(y-b)^2}{2a^2}\right) \end{aligned}$$

Explain how this is linked to the general normal density with non-zero mean and non-unity variance:

By inspection, the transformed random variable is a normal distribution with $\mu = b$ and $\sigma = |a|$.This can also be verified.

With the random variable as X

$$E[aX + b] = aE[X] + b = bVar[aX + b] = a^2 Var[X] = a^2$$

This agrees with the previous Jacobian result.

Verify this formula by transforming a large collection of random samples $x^{(i)}$ to give $y^{(i)} = f(x^{(i)})$, histogramming the resulting y samples, and overlaying a plot of your formula calculated using the Jacobian:

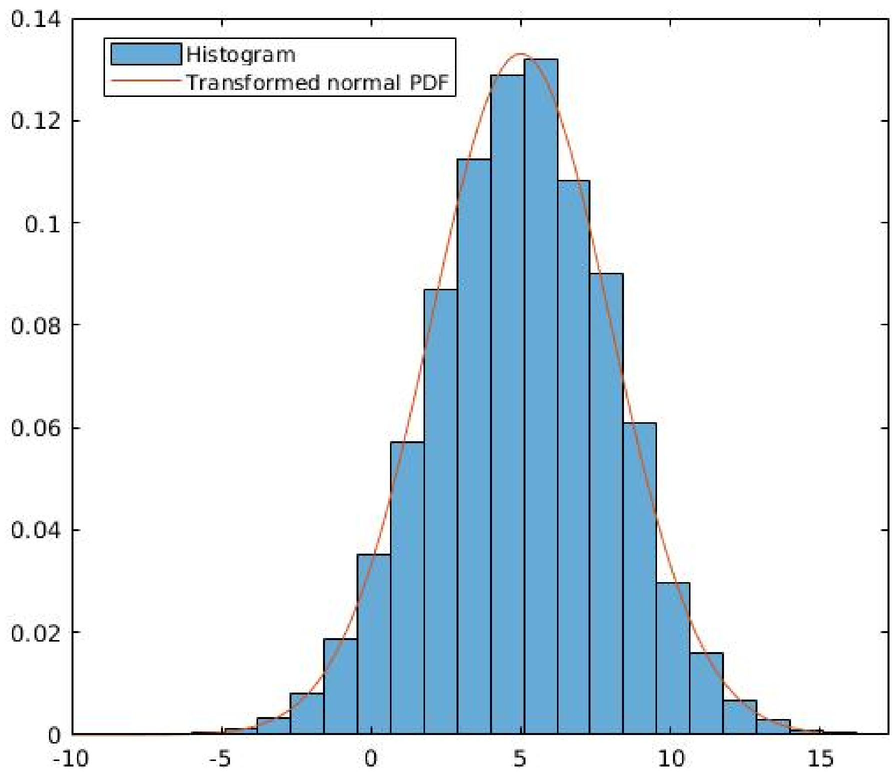

Figure 6: Linear transform on a gaussian variable

Now take $p(x) = \mathcal{N}(x|0,1)$ and $f(x) = x^2$. Calculate $p(y)$ using the Jacobian formula:

The inverse $f^{-1}(y)$ does not have a unique solution this time, so we'll consider all cases:

$$
\begin{aligned}
p(y) &= \sum_{k=1}^{2} \frac{p(x)}{|dy/dx|}\bigg|_{x=f^{-1}(y)} \\
&= \frac{p(x)}{|2x|}\bigg|_{x=\sqrt{y}} + \frac{p(x)}{|2x|}\bigg|_{x=-\sqrt{y}} \\
&= \frac{p(\sqrt{y})}{|\sqrt{y}|} \\
&= \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right)
\end{aligned}
$$

Verify your result by histogramming of transformed random samples:

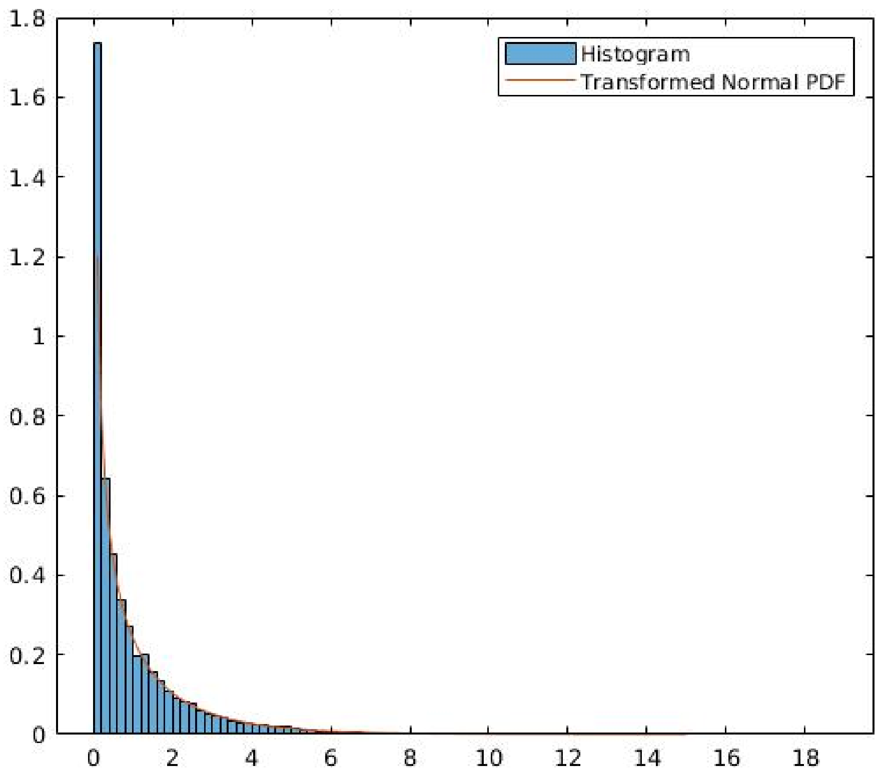

Figure 7: Quadratic transform on a gaussian variable

3 Inverse CDF method

Calculate the CDF and the inverse CDF for the exponential distribution:

$$
\begin{aligned}
x = F(y) &= \int_0^y \exp(-t)\mathrm{d}t \\
&= 1 - exp(-y), y \geq 0. \\
exp(-y) &= 1 - x \\
y &= -ln(1-x) \quad (0 \leq x < 1)
\end{aligned}
$$

Matlab code for inverse CDF method for generating samples from the exponential distribution:

```
% Random variable generation with the calculated cdf:
% 10,000 number of samples
n=10000
% Generating uniform random variable
x1=rand(n,1);
%Convert uniform random variable to  the desired exponential random variable
y=-log(1-x1);
```

Plot histograms/ kernel density estimates and overlay them on the desired exponential density:

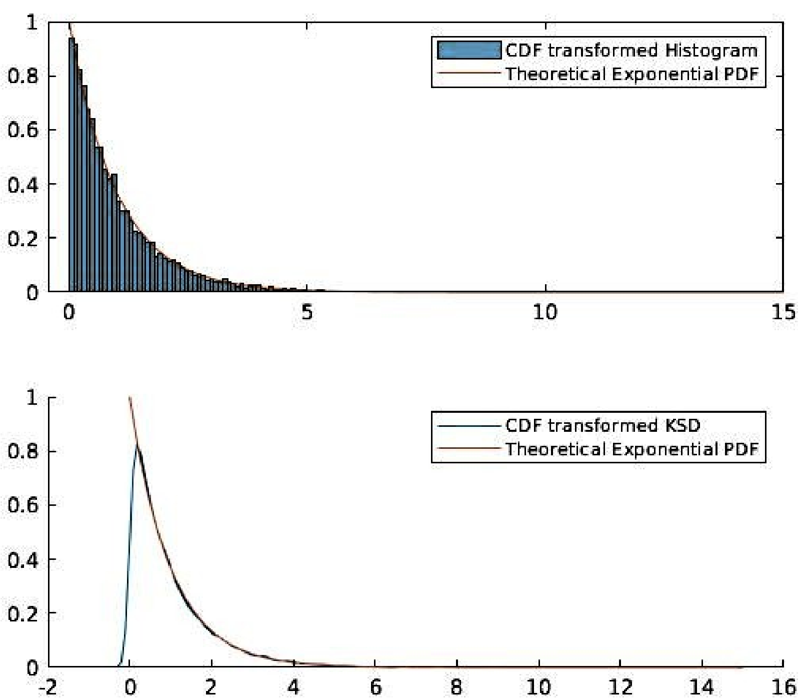

Figure 8: Histograms/kernel density estimates on the desired exponential PDF

4 Simulation from a 'difficult' density.

Matlab code to generate N random numbers drawn from the distribution of X:

```
% Random variable generation with the numerical 'recipe':
% 10,000 number of samples
n=10000;
% Initialising parameter alpha and beta
alpha = 0.5;
beta = 0.9;
% Evaluating b and s
b=(1/alpha)*atan(beta*tan(pi*alpha/2));
s=(1+beta^(2)*(tan(pi*alpha/2))^(2))^(1/(2*alpha));
% Sampling u and v from uniform and exponential PDFs respectively
u=(-pi/2) + rand(n,1)*pi;
v=exprnd(1,n,1);
% Generate x, random variable of interest, from the numerical 'recipe'
x = s*(sin(alpha*(u+b))./(cos(u)).^(1/alpha)).*(cos(u-alpha*(u+b))./v).^((1-alpha)/alpha);
```

Plot some histogram density estimates with $\alpha = 0.5$, 1.5 and several values of β:

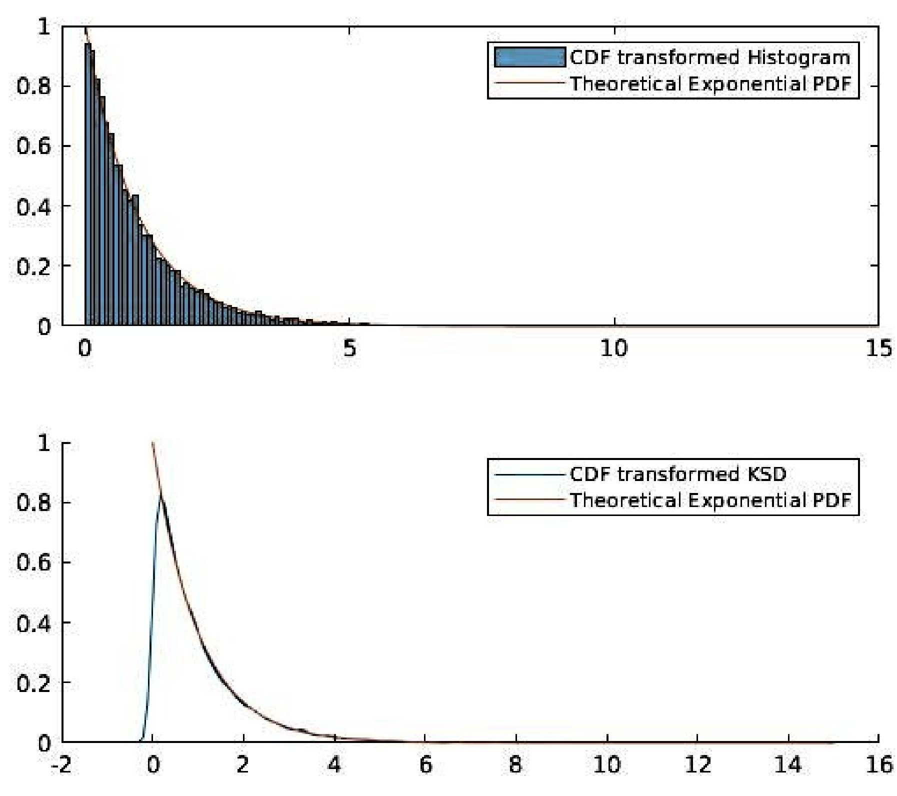

Figure 9: Histograms with various α and β values

Hence comment on the interpretation of the parameters α and $\beta \in [-1, +1]$:

Inferring from the plots, the parameter α changes the slope of the distribution. The curve becomes narrower and steeper as $alpha$ gets smaller. Hence this parameter $alpha$ can be, perhaps treated as a 'characteristic exponential term'.

The parameter β however, seems to have effect on the skewness of the the curve. A characteristic skew term, if you will. When $\beta > 0$, the curve is positive skewed and negatively skewed when $\beta < 0$. The curve seems to symetrical with no skewness at $\beta = 0$