ENGINEERING TRIPOS PART II A

EIETL MODULE EXPERIMENT 3F3

RANDOM VARIABLES and RANDOM NUMBER GENERATION
Short Report Template

Name: Naitian Zheng(nz253)

College: Fitzwilliam

Lab Group Number: nz253

1. **Uniform and normal random variables.**

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

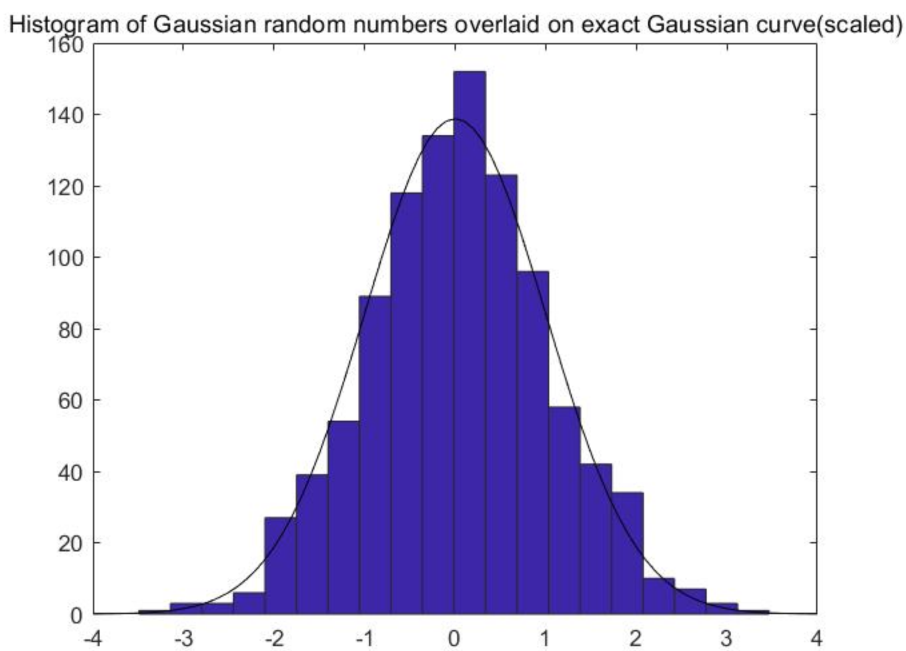

Figure 1: Histogram of Gaussian random numbers with exact Gaussian pdf(scaled)

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

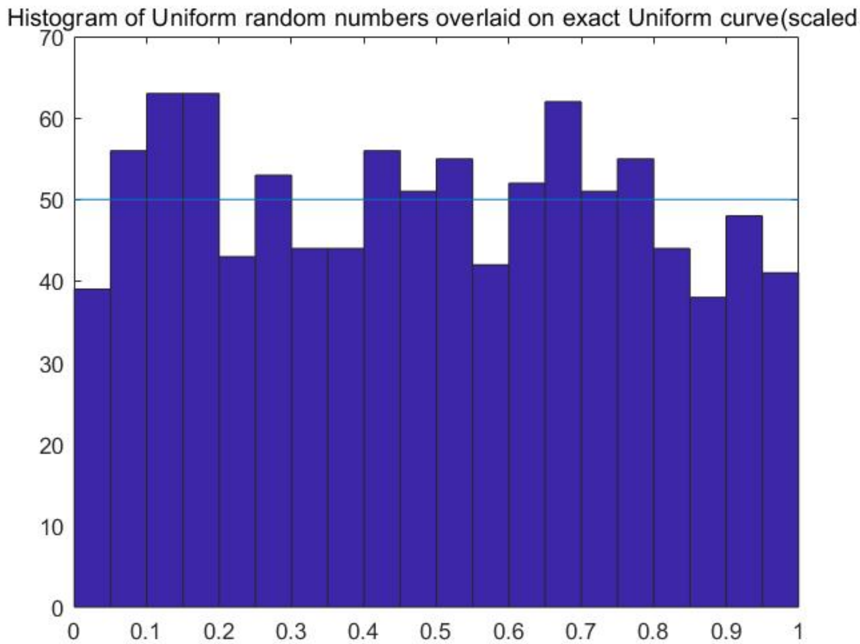

Figure 2: Histogram of Uniform random numbers with exact pdf(scaled)

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

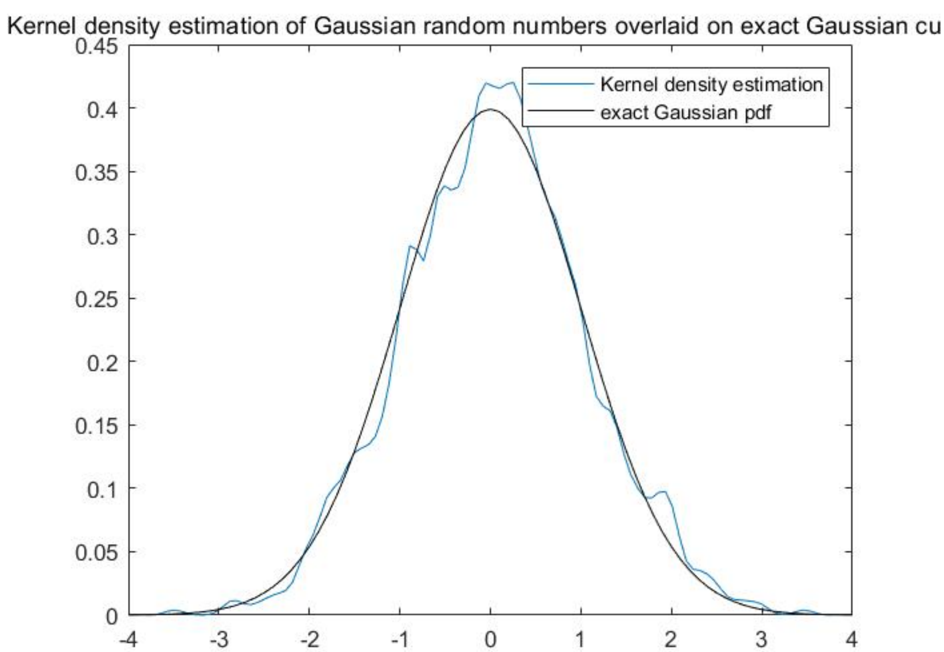

Figure 3: Kernel density estimation of Gaussian random numbers with exact pdf

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

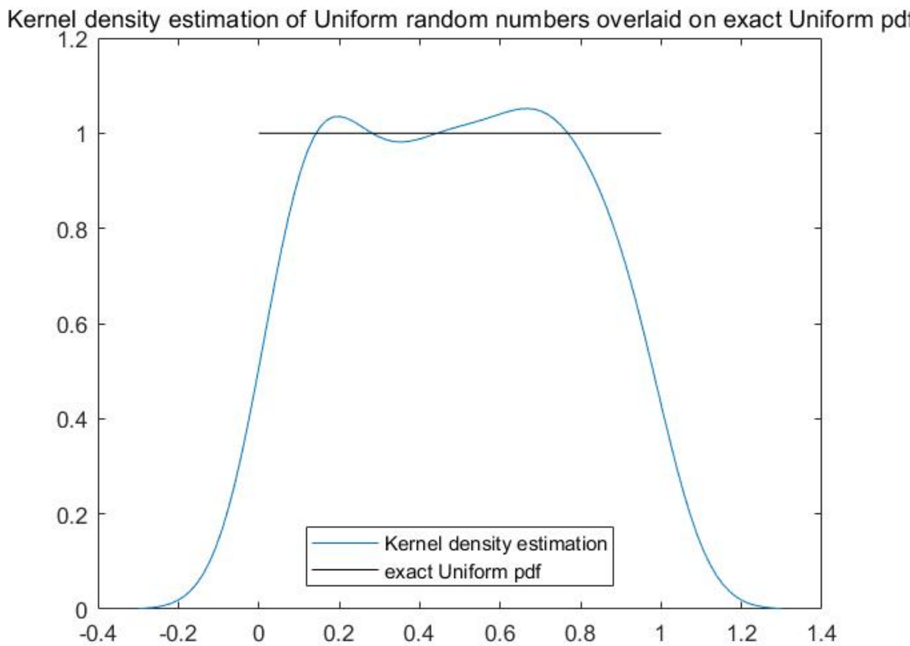

Figure 4: Kernel density estimation of Uniform random numbers with exact pdf

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:

Ans: In can be clearly seen that, though both method fit the real pdf quite well, the kernel density method gives a smoother estimation of the probability density, while histogram gives lots of separate bars. Also, the kernel density method makes use of every single sample whereas histogram divide the samples into groups and sums them, resulting in loss of some information. Furthermore, its performance does not depend on the selection of bin numbers and centres, as in the histogram case.

However, the kernel density method also has obvious disadvantages. Since it makes use of every sample, the complexity of calculation is way higher than that of histogram method. Besides, it introduces distortion when there is a sharp edge in the real distribution. In figure 4, for instance, the kernel density estimation has non-zero value at x larger than 1 and x smaller than 0. Since we used a Gaussian kernel and the range of a Gaussian kernel is infinity, distortion is introduced. Therefore, the performance of kernel density method strongly depends on selection of the filter.

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

Ans: According to the multinomial distribution theory, the mean of the count data in the jth bin is Np_j and the standard deviation is $\sqrt{Np_j(1-p_j)}$. For the uniform distribution, the probability that a sample lies in each bin is the same, say p_j:

$$p_j = \int_{c_j-\delta/2}^{c_j+\delta/2} p(x)dx = \int_{c_j-\delta/2}^{c_j+\delta/2} 1 * dx = \delta = 1/J$$

Where J is the number of bins. Therefore, for uniform distribution, the mean of the count data in each bin is N/J and the standard deviation is $\sqrt{N/J*(1-1/J)}$.

Explain behaviour as N becomes large:

Ans: As the number of bins remains constant but the number of samples N becomes large, both the mean and the standard deviation of the count data increase. However, they have a different order of growth: since the mean is proportional to N while the standard deviation is proportional to $\sqrt{N}$, the standard deviation becomes negligible with respect to the mean at very large N, where the histogram estimation of the uniform distribution looks smoother and more like a rectangular.

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

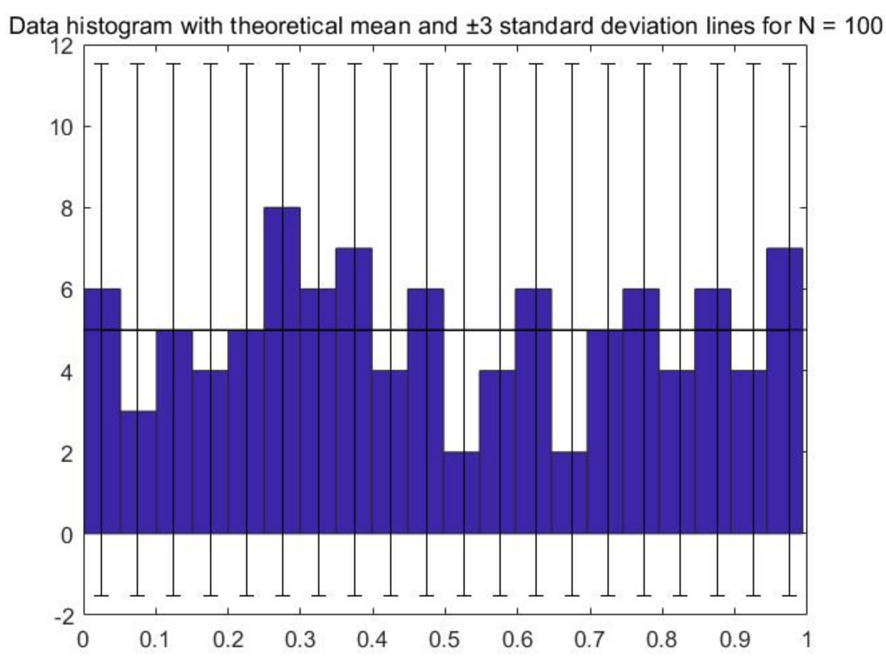

Figure 5: Histogram with theoretical mean and ± 3 standard deviation lines, N=100

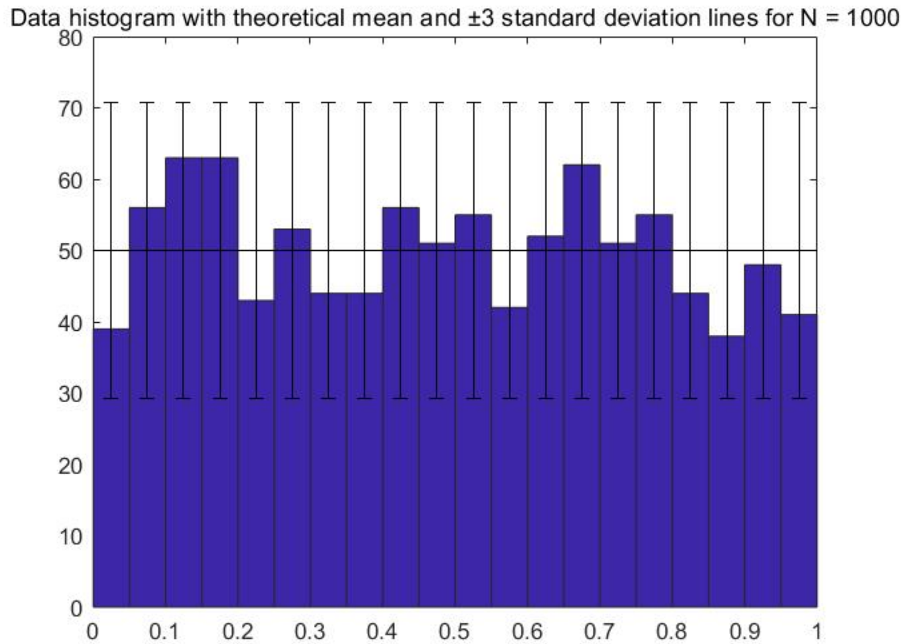

Figure 6: Histogram with theoretical mean and ±3 standard deviation lines, N=1000

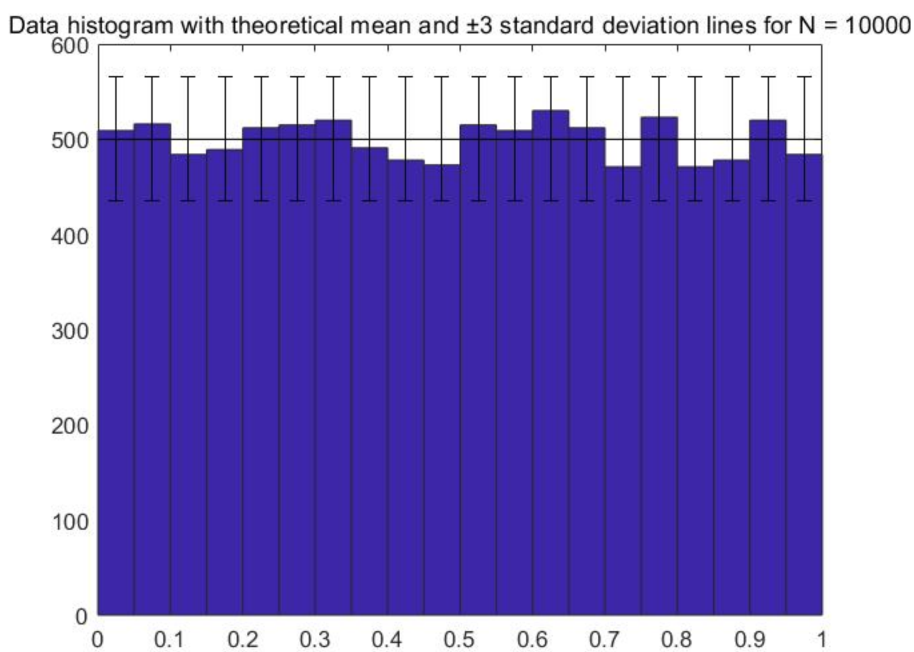

Figure 7: Histogram with theoretical mean and ±3 standard deviation lines, N=10000

Are your histogram results consistent with the multinomial distribution theory?

Ans: Yes. As shown in figures 5, 6 and 7, the multinomial distribution theory successfully predicts the means and all bars lie in the ± 3 *standard deviation lines.*

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:

   Ans: Since f() is a one-to-one, invertible, differentiable function and we know the density function p(x), we can calculate p(y) by the Jacobian formula:

$$dy/dx = |a|$$

$$x = f^{-1}(y) = \frac{y-b}{a}$$

$$p(y) = \frac{p(y)}{|dy/dx|}|_{x=f^{-1}(y)} = \frac{1}{\sqrt{2\pi}|a|} exp(-\frac{(y-b)^2}{2a^2})$$

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

   Ans: p(y) is still a Gaussian distribution with mean b and variance $|a|$*. Therefore, the function* $y = f(x) = ax + b$ *is actually translating the mean of the density function p(x) by b units and enlarging its variance to* $|a|$ *times of its original value.*

   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:

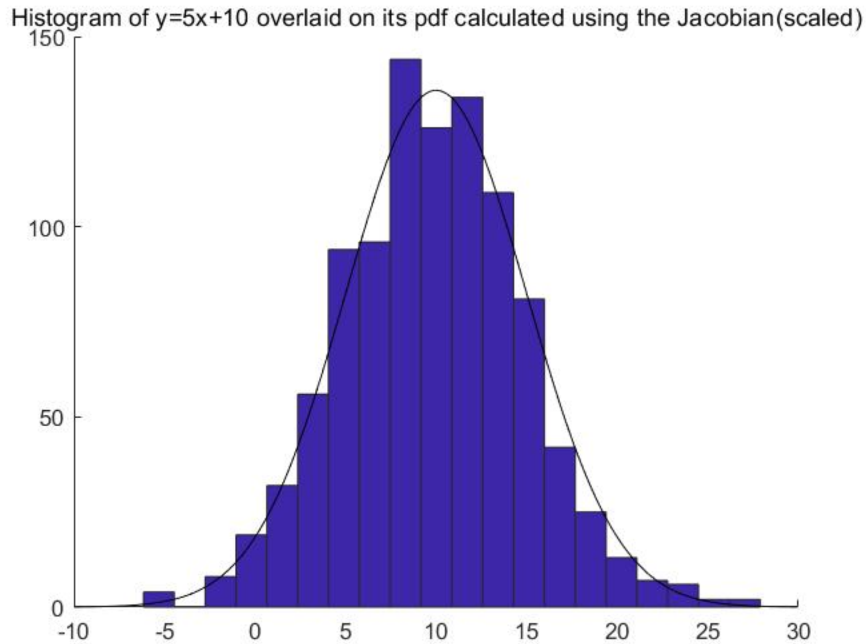

Figure 8: Histogram of y = 5x + 10 with its pdf calculated with Jacobian(scaled)

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

Ans: Note that f() is a non-invertible function, and there are two inverses, i.e. $x_1(y) = \sqrt{y}$, $x_2(y) = -\sqrt{y}$. *Hence the new density function is:*

$$p(y) = \sum_{k=1}^{K} \frac{p(x)}{|dy/dx|}|_{x=x_k(y)} = \frac{p(\sqrt{y})}{2\sqrt{y}} + \frac{p(-\sqrt{y})}{(-2(-\sqrt{y}))}$$

$$= \frac{1}{\sqrt{2\pi y}}, y > 0$$

$$or \qquad 0, y \leq 0$$

Verify your result by histogramming of transformed random samples:

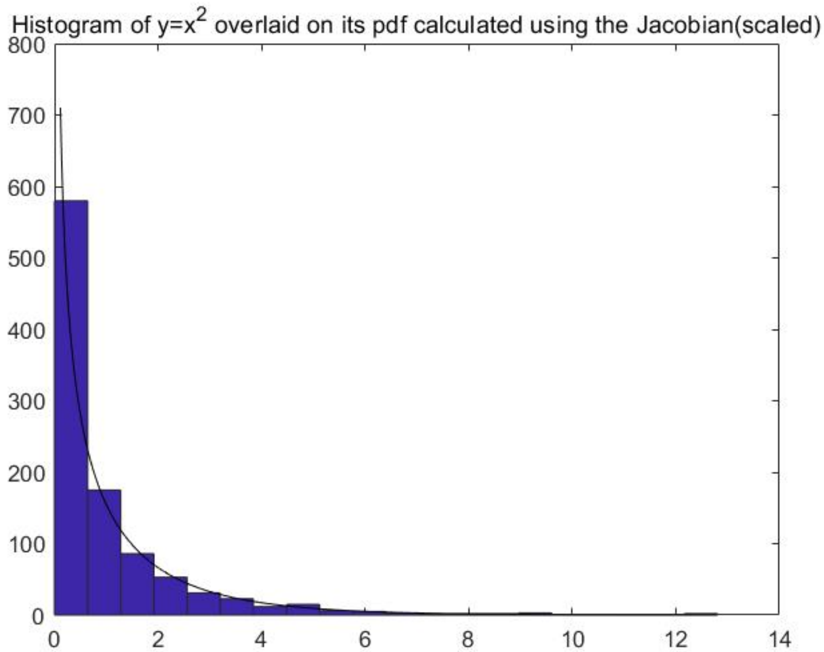

Figure 9: Histogram of $y = x^2$ and its pdf calculated with Jacobian(scaled)

3. **Inverse CDF method**

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

   Ans: Using the functions,

$$x = F_Y(y) = \int_0^y e^{-n} dn = 1 - e^{-y}$$

$$y = F^{-1}(x) = -ln(1 - x)$$

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

   Ans: the following matlab code generate 1000 samples from the exponential distribution:

   x = rand(1000,1);

   y = -log(1-x);

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

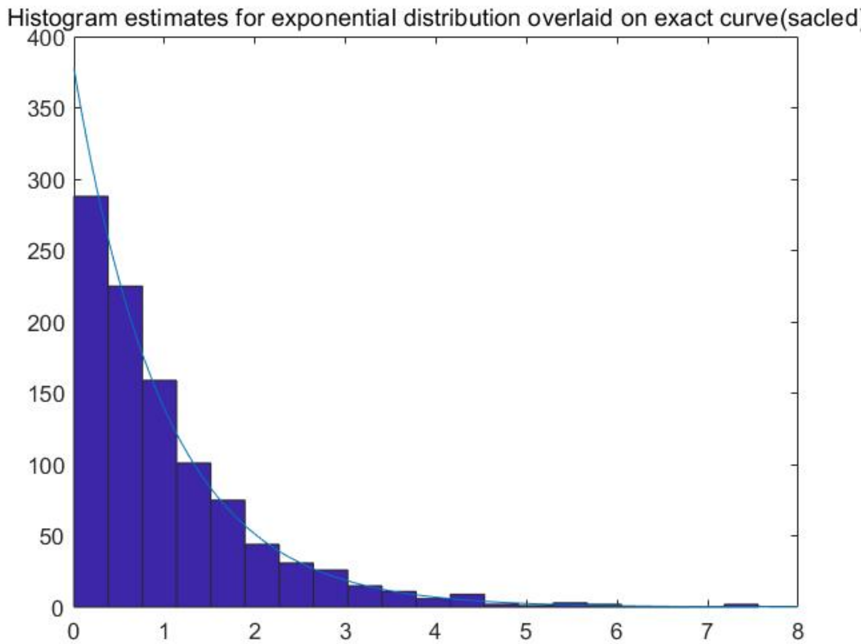

Figure 10: Histogram estimation for exponential distribution with its exact pdf(scaled)

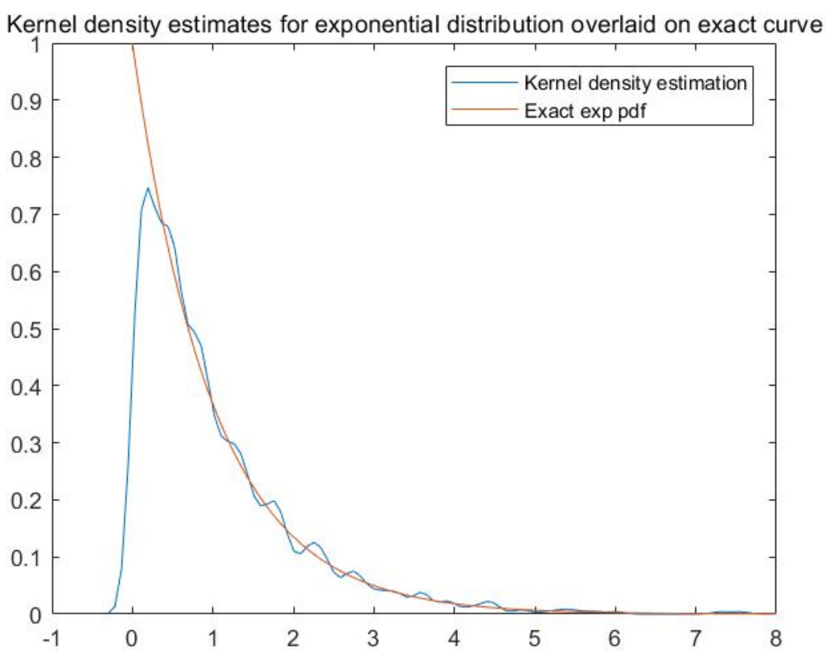

Figure 11: Kernel density estimation for exponential distribution with its exact pdf(scaled)

4. **Simulation from a 'difficult' density.**

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

```
alpha = 0;
beta= 0; %assign values for alpha and beta
b = atan(beta*tan(pi*alpha/2))/alpha;
s = (1+beta^2*tan(pi*alpha/2)^2) ^ (1/(2*alpha)); %calculate the value of b and s
U = rand(1000,1);
U = U.*pi-pi/2; %generate U~U(-pi/2,pi/2)
V = exprnd(1,1000,1); %generate V~E(V|1)
A = s.*(sin(alpha.*(U+b))./cos(U).^(1/alpha));
B = (cos(U-alpha.*(U+b))./V).^(1/alpha-1);
X = A.*B; %calculate X
```

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

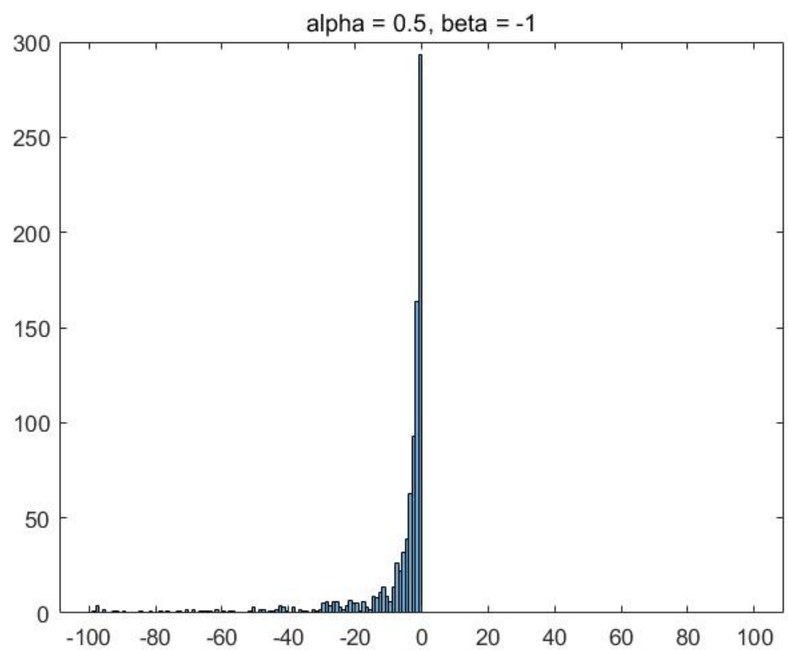

Figure 12a: Histogram of X with alpha = 0.5, beta = -1

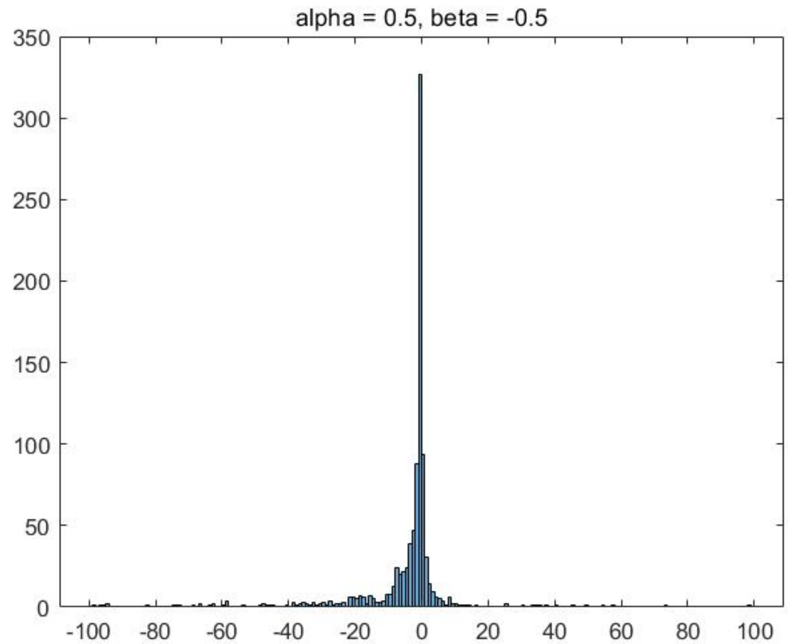

Figure 12b: Histogram of X with alpha = 0.5, beta = -0.5

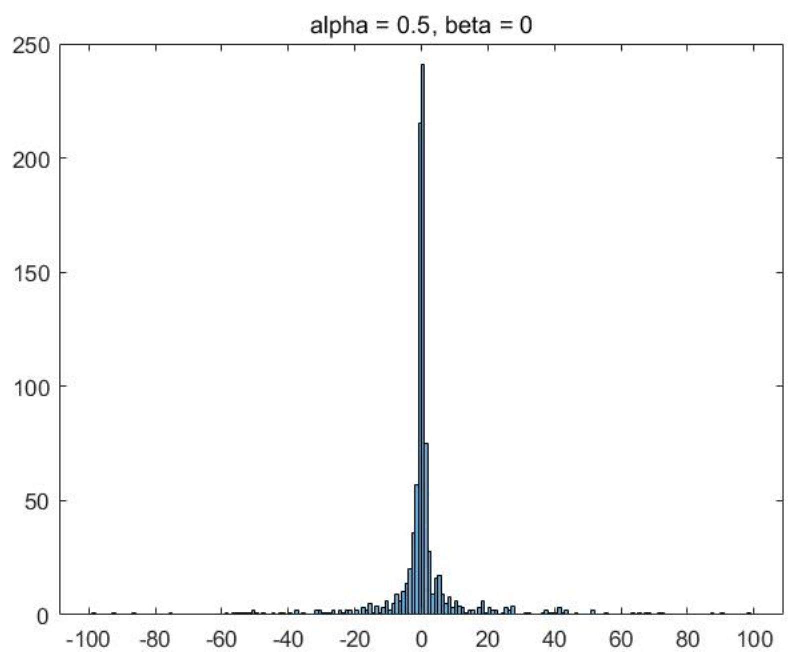

Figure 12c: Histogram of X with alpha = 0.5, beta = 0

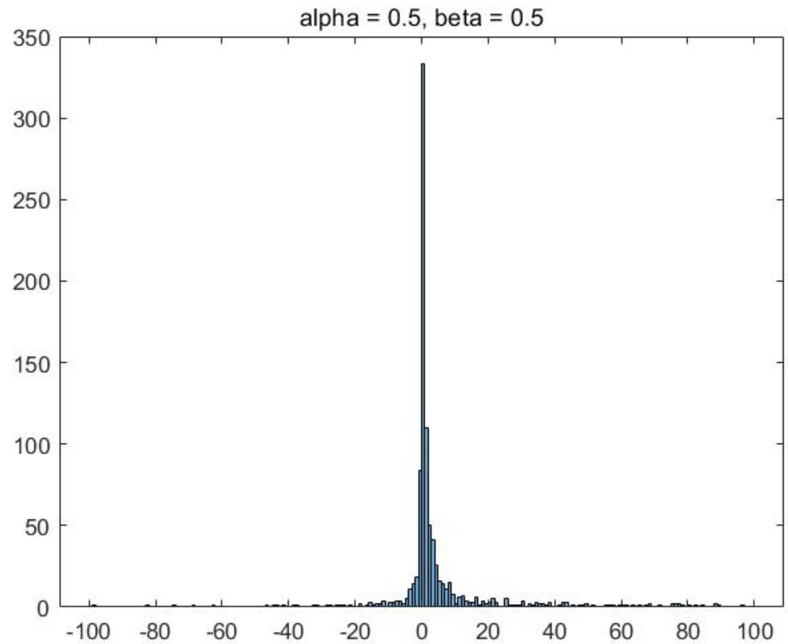

Figure 12d: Histogram of X with alpha = 0.5, beta = 0.5

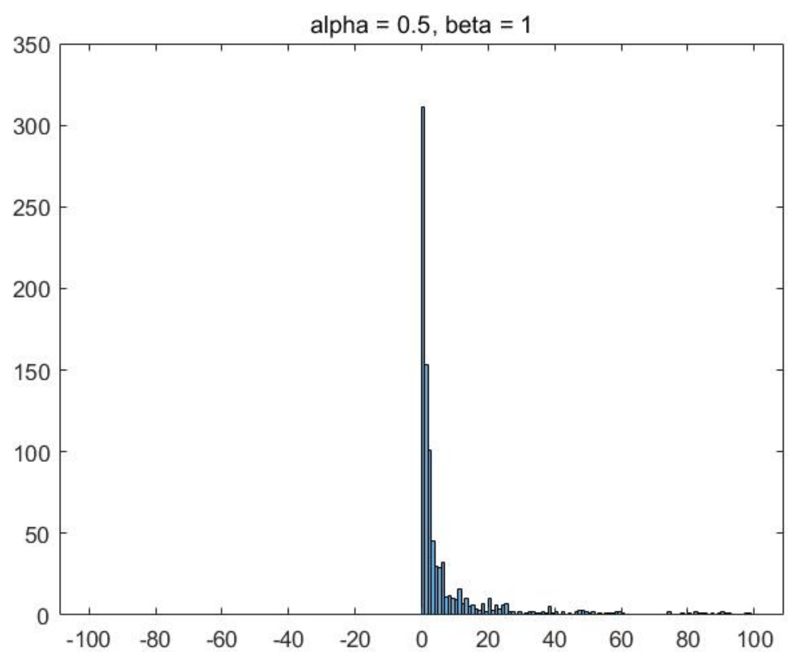

Figure 12e: Histogram of X with alpha = 0.5, beta = 1

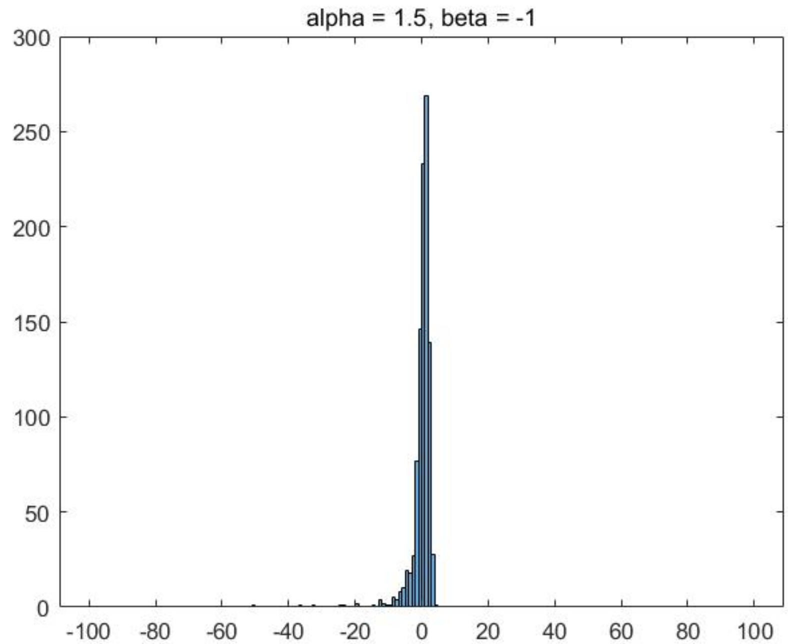

Figure 12f: Histogram of X with alpha = 1.5, beta = -1

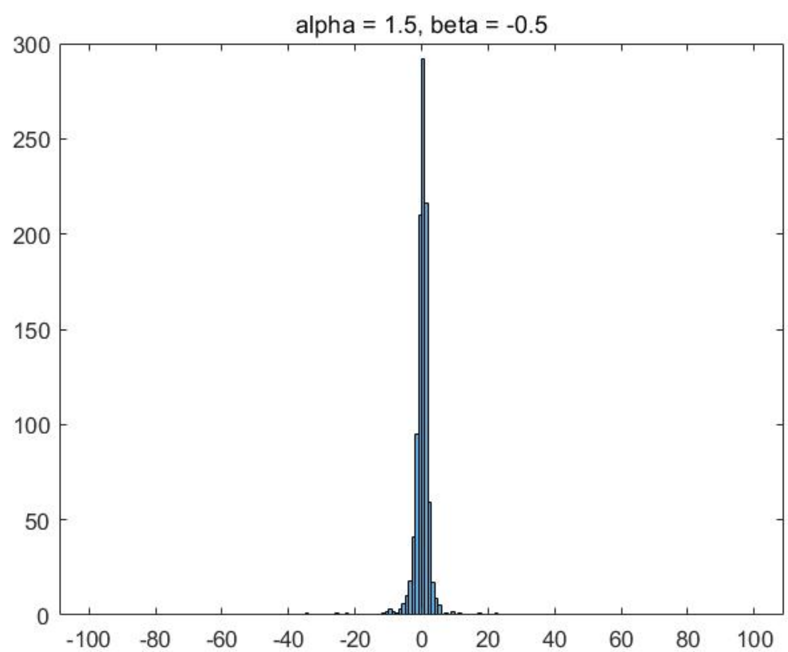

Figure 12g: Histogram of X with alpha = 1.5, beta = -0.5

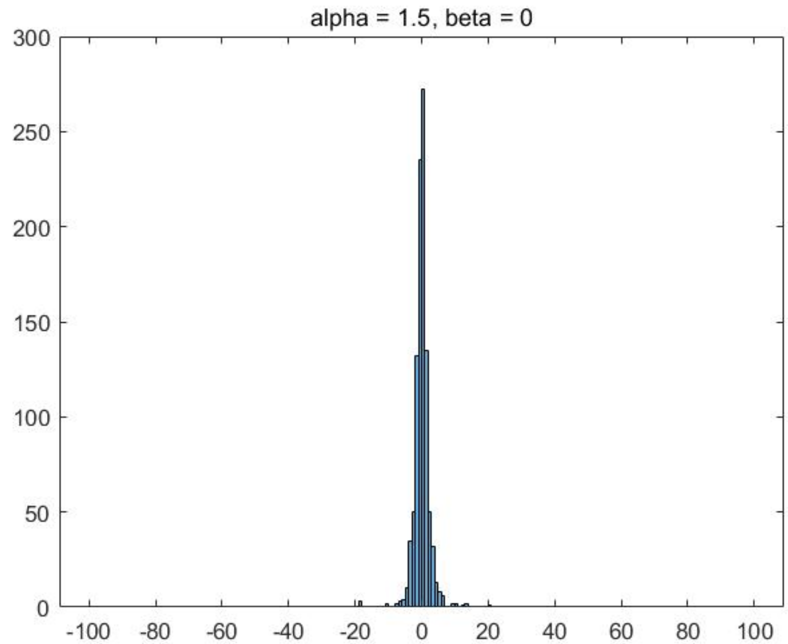

Figure 12h: Histogram of X with alpha = 1.5, beta = 0

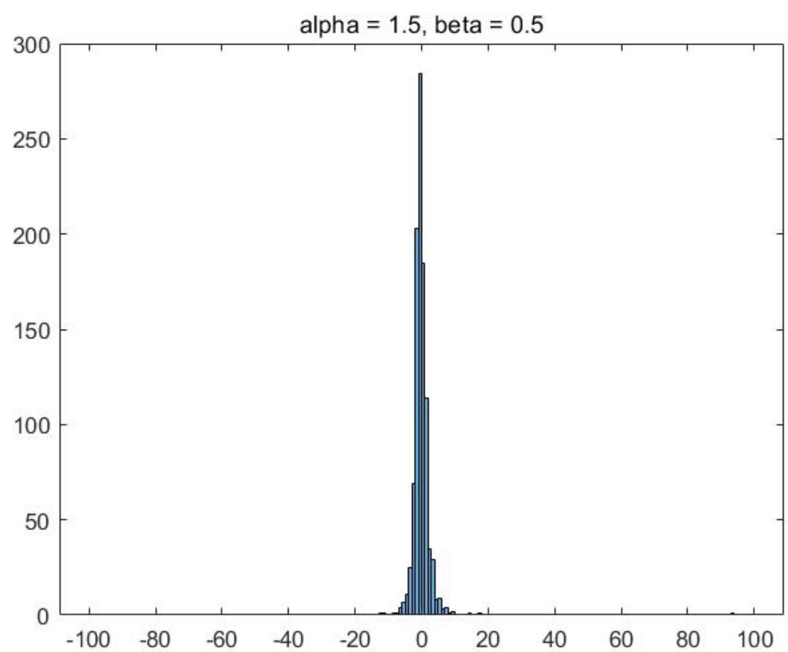

Figure 12i: Histogram of X with alpha = 1.5, beta = 0.5

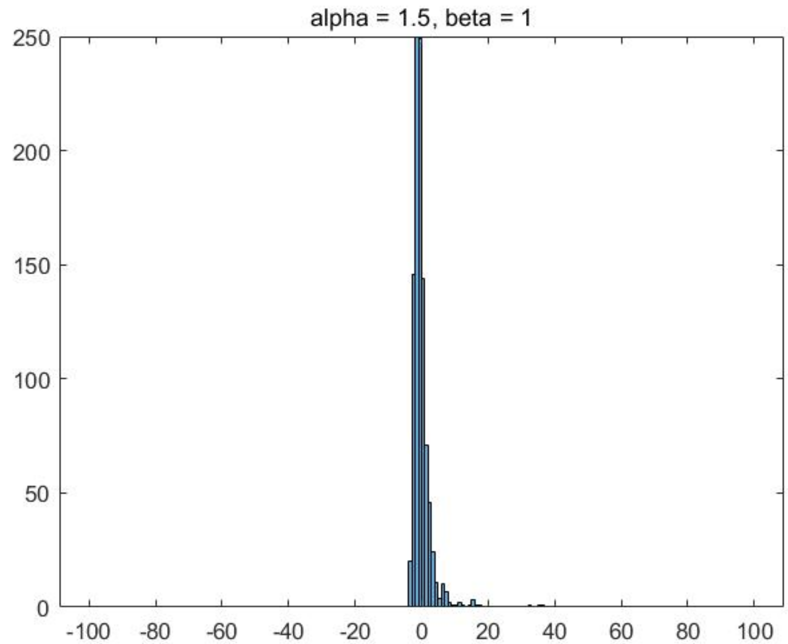

Figure 12j: Histogram of X with alpha = 1.5, beta = 1.5

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

Ans: First, we keep alpha constant and vary beta. As it can be seen in figures 12a-12e(and also 12f-12j), the histogram of X is almost symmetric as beta=0, and it drifts more and more towards one side as the absolute value of beta grows. Hence beta can be considered as a measure of asymmetry of the distribution.

Next, we keep beta constant and vary alpha. Within each pair (figure 12a and 12f, 12b and 12g etc.), the histogram with small alpha has more samples concentrated around 0 than the other, thus alpha can be a measure of concentration of the distribution.