3F3 Random Number Generation Report

Investigation on probability distribution visualization and mapping

Xiaoding Lu

xl402
Pembroke College

05.10.2018

1 Overview

Random number generation is a vital part of many engineering applications, this short lab report provides results accompanied with brief explanations and derivations for the four tasks undertaken during the session. Overall four tasks were investigated:

1. Visualize and investigate the advantages and disadvantages of the kernel density method compared with the histogram method.
2. Using the Jacobian formula in order to map a distribution onto another when transforming the random variable.
3. Using the inverse Cumulative Density Function (CFD) method, obtain a target probability distribution by mapping from an uniform distribution.
4. Investigate the effects of parameters of the *Stable Distribution*

2 Discrete Random Variables Visualization and Estimation

When dealing with real world problems, we often need to sample from a random variable to either determine the correctness of our hypothesis or to construct hypothesis. The sampled results can be presented either as a histogram or we can use it to estimate the probability density of a random variable.
Figure 1 below shows plots for both a normal and a uniform distribution for 1000 random variables, with their origional function and probability density estimate overlayed above. The estimation method used is the *Kernel Density* method, it is a smoothing estimate with a kernel typically chosen as a gaussian kernel $\mathcal{K}(x) = \mathcal{N}(x|0, 1)$.

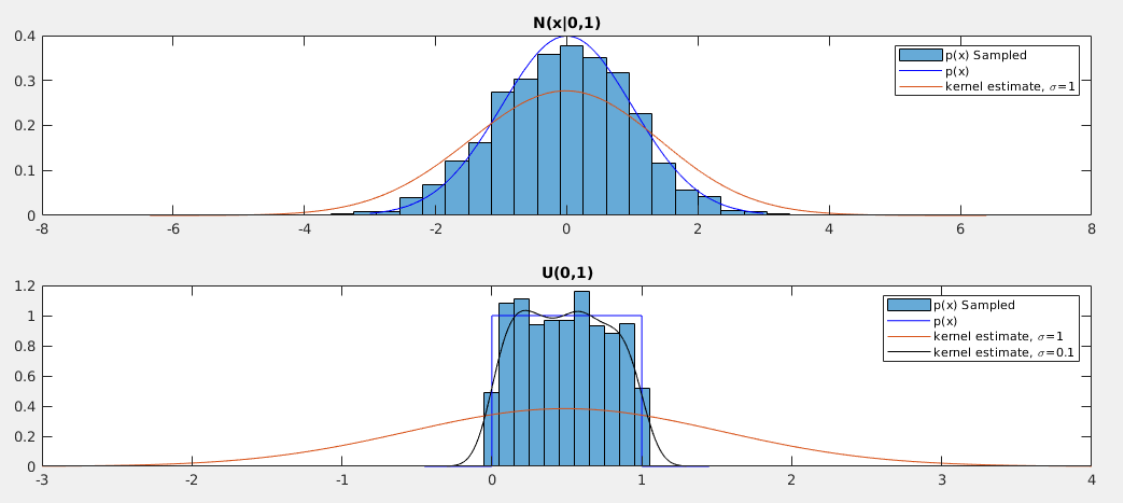

Figure 1: Sampling and Kernel Density estimate demonstration

It is clearly seen, the choice of kernel size (bandwidth) determines the quality of our probability density estimate, the bandwidth of the kernel should neither be too small or too large (underfitting for larger kernels, overfitting for small kernels). The kernel density method also cannot handle discontinouties which is likely to happen on real world models. Histogram on the otherhand provides a unbiased way of visualizing the sampled distribution. However, it is difficult to perform mathematical operations on descrete data, often we need to convert it into a continous variable. The bin size of the histogram is also important, as linear bins may result in missing details for dense regions and no counts for sparse regions (this can however be resolved

through entropy based binning methods).
Intuitively, the more number of samples we have, the more representative (accurate) the histogram is. From the theory of *multinomial distribution*, we are given the mean of the count data in bin j with width δ is Np_j and the standard deviation is $\sqrt{Np_j(1-p_j)}$, where p_j is given by:

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

We can therefore derive that the theoretical mean count within any bins for a uniformaly distributed random variable is $N\delta$, with standard deviation $\sqrt{N\delta(1-\delta)}$. Hence the mean count varies linearly with N with standard deviation only varying in square root. We are therefore expecting less variance within bins (hence more accuracy) as N increases. This is confirmed through experimentation, figure 2 shows how theoretical mean and standard deviation varies with N:

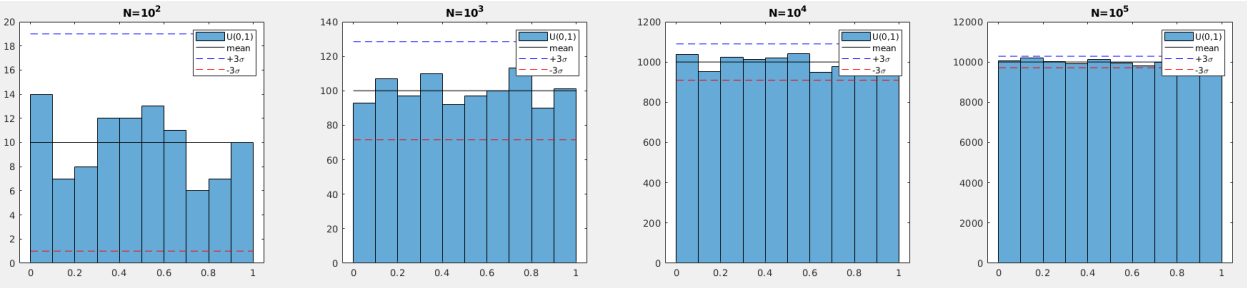

Figure 2: Sampling accuracy for uniform distribution

3 Functions of Random Variables

It is shown that given a random variable x, with density function $p(x)$ and there exists a function which maps x to another variable so that $y = f(x)$, the density function of $p(y)$ can be found using the Jacobian method, note $x_k(y)$ denotes the possible solution for $f^{-1}(y)$, if $f(y)$ is a one-to-one mapping function, then we can ignore the summation.:

$$p(y) = \sum_{k=1}^{K} \frac{p(x)}{|dy/dx|} \mid_{x=x_k(y)} \tag{1}$$

Using this theorum, for normally distributed $\mathcal{N}(x|0,1)$ random variables, take $y = f(x) = ax + b$, we aim to derive the density function $p(y)$. First, we write out the full form of our input distribution:

$$p(x) = \mathcal{N}(x|0,1) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{2}$$

given $y = ax + b$, differentiating we have:

$$\frac{dy}{dx} = a \quad \forall x \tag{3}$$

We also have $f^{-1}(y) = \dfrac{y-b}{a}$, subsituting these into equation 1, we have:

$$p(y) = \frac{p(x)}{a} \mid x(y) = \frac{y-b}{a} \tag{4}$$

$$p(y) = \frac{1}{\sqrt{2\pi a^2}} e^{-(y-b)^2/2a^2} = \mathcal{N}(y|b,a) \tag{5}$$

Therefore, we are expecting a normal distribution centred around b with a standard deviation of a. Left Figure 3 demonstrates exactly this if $a = 2$ and $b = 3$, we indeed see $p(y)$ is a normal distribution with correct mean and standard deviation.

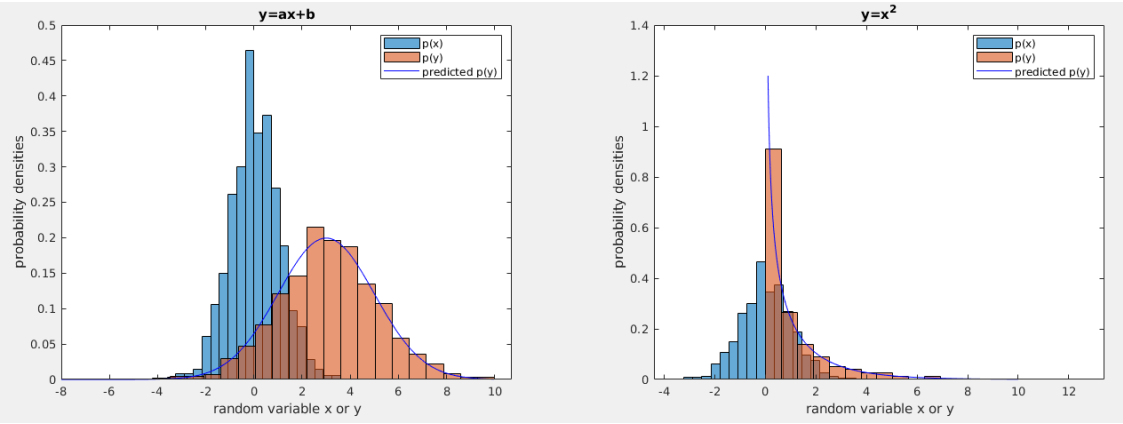

Figure 3: Functions of random variables probability density mapping

Similarly, if instead $y = x^2$, following the same derivation, we arrive at the probability density function below. This is also confirmed by the experimental plot shown on the right of figure 3:

$$p(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \tag{6}$$

4 Inverse Cumulative Frequency Density Method

Transformations of random variables leads to a very general method for non-uniform random variable generation. One of the simplest methods for generation of random variates with specified probability density is the *Cumulative Frequency Density* (CFD) method. It states that for any target probability distribution $p(y)$ that has an invertable CFD denoted by $F^{-1}(x)$, we can achive the target distribution $p(y)$ by transforming random variables sampled from a uniform distribution $p(x) = \mathcal{U}(0, 1)$ through:

$$y = f(x) = F^{-1}(x) \tag{7}$$

For example, given a target exponential distribution $p(y) = \exp(-y)$ we first compute its CFD and Inverse CFD, $F(X)$ and $F^{-1}(x)$ respectively:

$$F(X) = \int_0^X p(y) = 1 - e^{-X} \tag{8}$$

$$F^{-1}(X') = -\ln(1 - X') \tag{9}$$

Using the equation above, random variables sampled from a uniform distribution $p(x) = U(0, 1)$ is passed through the mapping function $F^{-1}(x)$, figure 4 shows histograms and kernel density estimates of this transformation with actual target ditribution overlayed on top. We can see clearly that the inverse CFD method can be used to generate reliable samples.

5 Stable Distribution Simulation

Sometimes random variables cannot be simulated using inverse CDF or other convenient means. One such case is the stable distribution which is fundimental in the field of communications data, sigal processing and

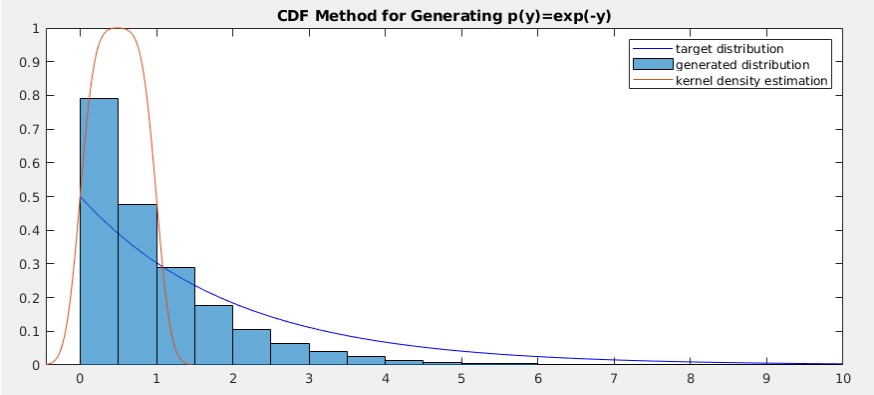

Figure 4: Sampling accuracy for uniform distribution

Internet of Things where interfering noise can be far from the Gaussian assumptions imposed.
There are two parameters controlling the distribution, $\alpha \in (0, 2)$ $(\alpha \neq 1)$ and $\beta \in [-1, +1]$. The distribution is drawn from two standard random variables, a uniform distribution: $U \sim \mathcal{U}(-\pi/2, +\pi/2)$ and a exponential distribution: $V \sim \mathcal{E}(V|1)$. The formula for the alpha stable distribution is quiet complex and is ommited here. Figure 5 depicts the effects of parameters α and β on the distribution. It is seen that the α parameter controls the sharpness of the rolloff of the distribution. The higher it is the more it seems to have a gaussian like behaviour (less skewed towards center). The parameter β controls the left/right skewness of the distribution, with -1 meaning mean shifting more towards right, and $+1$ meaning mean shifting more towars the left.

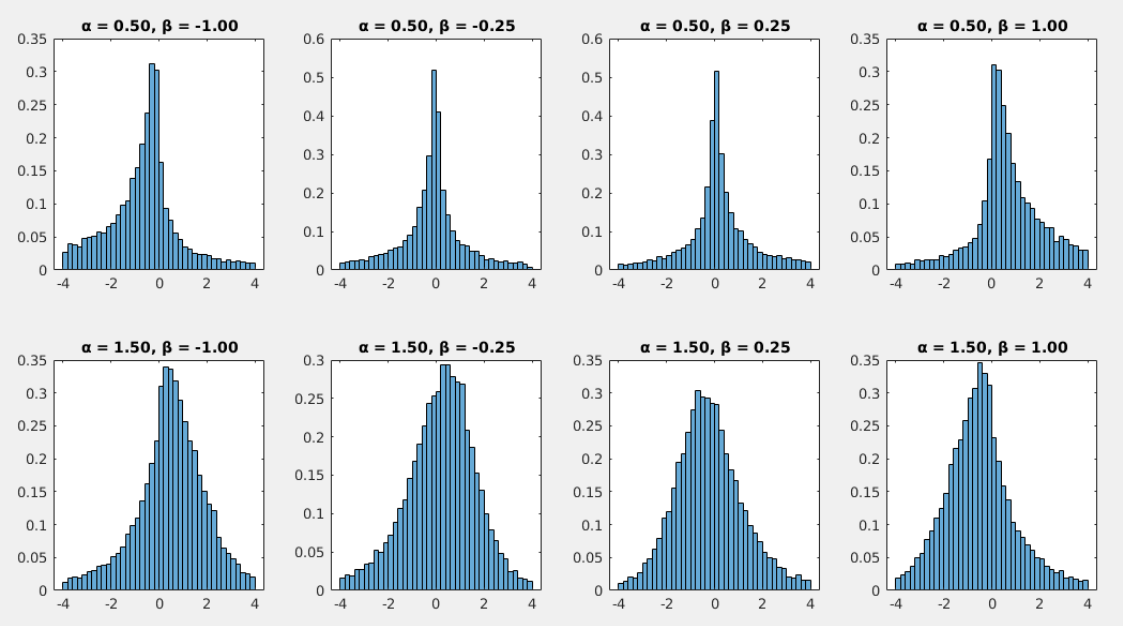

Figure 5: Effects of parameters on Alpha-Stable Distribution