Dhaka Univ. J. Sci. 68(1): 59-64, 2020 (January)

Rejection Sampling Schemes for Simulating from Arbitrary Probability Densities

Anamul Haque Sajib

Department of Statistics, Dhaka University, Dhaka-1000, Bangladesh

(Received : 14 July2019 ; Accepted : 7 January 2020)

Abstract

Simulating random variates from arbitrary non-normalized probability densities, very often they do not have familiar forms, is an increasingly important requirement in many different fields, especially in Bayesian statistics. Accept-reject algorithm is one of the commonly used methods to simulate random variates from such densities but restriction on choosing proposal density under this framework (heavier tails than the target density) limits its applicability to a larger extent. On the other hand, Markov Chain Monte Carlo (MCMC) method can choose proposal density arbitrary which makes this method applicable to a larger class of target densities[5]. In addition to MCMC method, a more general widely used method known as ratio-of-uniforms (RoU) which requires only two uniform variates to simulate one variates from such densities. However, no empirical comparison among these methods for simulating random variates from such densities was seen in the literature. In this paper, we limit our study only to MCMC and RoU methods to simulate random variates from such densities. Following the generation of random variates from such densities using these two methods, we compare the performance of these two methods based on quality of the generated samples. Finally, we conclude that RoU method performs better than MCMC method as far as quality of the generated sample (randomness) and computational cost are concerned.

Keywords: Accept-Reject method, MCMC, RoU method, Non-normalized density, Statistical computation.

I. Introduction

Simulating random variables $X \sim f_X(x)$, $f_X(x)$ is the density function of X, is required to conduct an empirical study of the random variable X. Furthermore, estimation problem in frequentist approach is treated as simulation problem in Bayesian approach. Therefore, simulation from the posterior is required to estimate the unknown parameters. For example, suppose we have observed sample $\boldsymbol{x} = (x_1, x_2, \cdots x_n)$ from exponential density with rate parameter $\lambda = \lambda_0$. Maximum likelihood estimation method is used to estimate λ_0 in frequentist approach while a point estimate of λ_0 can be defined using the generated sample obtained from the posterior distribution (likelihood × prior distribution of λ_0) of λ_0 in Bayesian approach. There are several methods available in the literature to simulate random variates from the probability densities of random variate X (posterior distribution in Bayesian setting). Most importantly, inverse transformation method and accept-reject algorithm are commonly used to simulate from such probability densities. The choice of simulating $X \sim f_X(x)$ mainly depends on (i) whether the distribution of X, $F_X(x)$, has a closed form or not (ii) simplicity of the chosen algorithm as far as implementation is concerned (iii) computing time[7]. When the distribution function $F_X(x)$ has no closed form then the inverse transformation method is not applicable. For example, inverse transformation method can not be applied to generate random variates from the Gaussian density as its distribution function has no closed form. In such cases, other available methods need to be applied. In Bayesian statistics, we often have to deal with a density function which is known up to normalizing constant and very often even such densities don't have familiar forms. This situation arises because of considering non-conjugate distributions i.e. prior and posterior are not in the same probability distribution family. For example, suppose we are given a data set $\boldsymbol{x} = (x_1, x_2, \cdots x_n)$ that come from the Gaussian density with mean μ and variance $\sigma^2 = 1$, and we are also told to estimate unknown mean μ based on this observed sample. Under Bayesian setting, a prior distribution for unknown mean μ needs to be specified to get the posterior distribution of μ. Suppose, we consider a standard Cauchy density for μ, $p(\mu) \propto (1+\mu^2)^{-1}$, which yields posterior density of μ $\pi(\mu|\boldsymbol{x}) \propto \left[\prod_{i=1}^{n} e^{-0.5*(x_i-\mu)^2}\right] \times \left[\frac{1}{1+\mu^2}\right]; -\infty < \mu < \infty$. For simplicity let's consider we have only one observation in our sample i.e. $n = 1$. Then the form of the above posterior density becomes $\pi(\mu|x) \propto e^{-0.5*(\mu-x)^2} \times (1+\mu^2)^{-1}$, which is known up to normalizing constant and does not have any familiar form. The normalizing constant of this posterior density $\pi(\mu|x)$ is $\int_{-\infty}^{\infty}\left[e^{-0.5*(\mu-x)^2} \times (1+\mu^2)^{-1}\right] d\mu$, which is mathematically intractable i.e. can not solve analytically. Therefore, distribution function of $\pi(\mu|x)$ has no closed form. Inverse transformation method can not be applied to simulate random variates from such types of densities as distribution function has no closed form.

There is a need, therefore, other efficient methods to simulate from such types of density. Accept-Reject algorithm is a well-known method to simulate random variates from an arbitrary probability density which is known up to normalizing constant. A suitable proposal density needs to be chosen to simulate random variates from such densities under this framework. One of the main challenges of accept-reject algorithm is to have a finite upper bound (M) of the ratio of target density (from which we want to simulate) over proposal density as far as implementation is concerned. Proposal density must be a heavier tail than the target density to ensure a finite upper bound[1]. For example, to simulate from $N(0,1)$ one can choose standard Cauchy as a proposal which yields a finite M (standard Cauchy has a heavier tail than standard normal). But the other way around is not possible as $N(0,1)$ has a lighter tail than standard Cauchy, resulting an infinite

*Author for correspondence. e-mail: *sajibstat@du.ac.bd*

M. In addition to have finite value of M, Accept-Reject method also needs to find an efficient proposal density among all possible proposal densities (theoretically infinite numbers of proposal densities) for which M is small. The smaller the M value Accept-Reject has, the more efficient it is. Unfortunately, there are no general method available to analytically find the smallest value of M[5]. There is a need, therefore, a general technique which will choose the proposal density without any restriction and will be free from determining the lowest value of M.

Apart from Accept-Reject method, there are two other available methods namely Markov Chain Monte Carlo (MCMC) and Ratio-of-Uniforms (RoU) methods those can be used to simulate random variates from such densities. MCMC is a special type of rejection sampling in which proposal density can be chosen arbitrarily[5]. Under this framework, it is not mandatory to choose a proposal density which has a heavier tail than the target density which makes this method applicable to a larger extend. On the other hand, RoU method invented by Kinderman and Monahan method does not need any proposal density to simulate from target density.Instead of choosing any proposal density it requires only two independent uniform numbers (play the role as proposal) to simulate one variate from the target density. The advantage of using the RoU method is to have a unique theoretical acceptance rate[3] (like M^{-1} in Accept-Reject method). The detail about MCMC and RoU methods will be discussed in section 3.

We limited our study to only MCMC and RoU methods to simulate random variates from an arbitrary density known up to normalizing constant. Although both MCMC and RoU methods can be used to simulate random variates from arbitrary density, we are unaware of any empirical comparison of these two methods. In this paper, our aim is to carry out an empirical study between these two methods for simulating random variates from an arbitrary probability density known up to normalizing constant.

We organize the rest of the paper as follows: Section 2 presents the concept of Markov chain with examples along with some other related terminology used in this paper. The MCMC and RoU methods are discussed in section 3 while section 4 and 5 show the implementation of these two methods through simulation study. In the penultimate section, we present the results and discussions which are followed by conclusion and future works presented in section 7.

II. Concept of Markov Chain and Other Related Terminology

In this section, we discuss some key facts about the Markov chain which will be necessary to understand the MCMC technique. Apart from discussing Markov chain, we also discuss here some important terminology such as arbitrary probability density, normalizing constant and mathematically intractable normalizing constant. In this paper, we consider the Markov chain which is the generalized version of original Markov chain. It is noted that original Markov chain was proposed by Metropolis et al.[6] and the generalization is made by Hastings[2]. Hammersley and Handscomb[1] also discussed an introduction to Markov chain methods of sampling proposed by Metropolis in 1964, and we follow their paper to prepare the terminology related to MCMC: Markov chain, transition function, periodicity and irreducibility of Markov chain, stationary distribution and Ergodic theorem.

Markov chain

To define Markov chain, we consider here only discrete time points at which transitions of a chain occurs. Let a process has finite number of states $S_1, S_2, \cdots\cdots$ and at time t it is in X_t. Then X_t be a random variable for which the following conditional probabilities can be defined

$$\Pr\left(X_t = S_{j_t} | X_{(t-1)} = S_{i_{(t-1)}}, \cdots, X_1 = S_{k_1}\right).$$

The above process is called Markov chain if the distribution of X_t depends only on its immediate predecessor $X_{(t-1)}$ i.e

$$\Pr\left(X_t = S_{j_t} | X_{(t-1)} = S_{i_{(t-1)}}\right). \quad (1)$$

The probabilities defined in equation (1) do not depend on time t, and the transition probabilities for the above Markov chain can be defined as

$$p_{ij} = Pr(S_i \to S_j) = Pr\left(X_t = S_{j_t} | X_{(t-1)} = S_{i_{(t-1)}}\right).$$

Transition function

A matrix $\boldsymbol{P}$ which consists of all p_{ij} elements of Markov chain i.e. $\boldsymbol{P} = \{p_{ij}\}$ is called the transition probability matrix. When the Markov chain has continuous state space then $\boldsymbol{P}$ is equivalent to transition function or transition kernel.

Irreducibility and Periodicity of Markov chain

To understand the irreducibility and periodicity of Markov chain, we need to be familiar with n step transition probabilities and first passage probabilities first. The n step transition probabilities, denoted by $p_{ij}^{(n)}$, are defined as $\Pr(X_t = S_j | X_{(t-n)} = S_i)$. Then the first passage probabilities can be defined as

$$f_{ij}^{(n)} = \Pr(X_t = S_j, X_{t-1} \neq S_j, X_{t-n+1} \neq S_j | X_{(t-n)} = S_i),$$

which means Markov chain reaches to state S_j for the first time after n step starting from state S_i. We can define mean first passage time by using the idea of $f_{ij}^{(n)}$. The mean first passage times denoted by m_{ij} can be defined as $m_{ij} = \sum_{n=1}^{\infty} n\, f_{ij}^{(n)}$, provided $\sum_{n=1}^{\infty} f_{ij}^{(n)} = 1$. When $i = j$, mean passage times become mean recurrence times. The state S_i is called recurrent, positive recurrent and null when $\sum_{n=1}^{\infty} f_{ij}^{(n)} = 1$, $m_{ii} < \infty$ and $m_{ii} = \infty$, respectively.

If $p_{ii}^{(n)} \neq 0$ (chain starting from i returns to i again after n step) which occurs only when n is a multiple of d, then d is

called the period of this Markov chain. If $d = 1$, then Markov chain is aperiodic. To be an irreducible Markov chain, all states of Markov chain need to belong to the same class.

Stationary distribution

Suppose $\boldsymbol{\pi} = (\pi_1, \pi_2 \cdots \pi_S)$ be a probability distribution with $\pi_i > 0, \forall\, i,$ then $\boldsymbol{\pi}$ is said to be a stationary distribution of Markov chain with transition probability matrix $\boldsymbol{P}$ if the condition $\pi_j = \sum_i \pi_i\, p_{ij}$ is hold for all j.

Ergodic Theorem

An aperiodic, irreducible Markov chain with transitional probability matrix $\boldsymbol{P}$ and stationary distribution $\boldsymbol{\pi}$ is ergodic, so that the ergodic average $\bar{h}_n = \frac{1}{n}\sum_{i=1}^{n} h(X) \rightarrow E_\pi[h(X)]$ as $n \rightarrow \infty$ where $E_\pi[h(X)] = \sum_{x=1}^{S} h(x)\pi_x$

Arbitrary and Non-normalized Density

By arbitrary density, here we mean the form of the density does not necessarily belong to any known family. In addition, by arbitrary density we don't mean arbitrary measure (reference measure) here. For the density considered here, the reference measure is the Lebesgue measure. The posterior density, considered in the introduction section, is said to be a probability density function if it has the form $\frac{\pi(\mu|\boldsymbol{x})}{\int \pi(\mu|\boldsymbol{x})\, d\mu}$, where $\int \pi(\mu|\boldsymbol{x})\, d\mu$ is the normalizing constant. Without having this normalizing constant, the posterior is said to be non-normalized density (common term used in Bayesian Statistics is density known up to normalizing constant).

Ljung-Box Test

To test the randomness of a time series Ljung-Box test is widely used in Econometrics and other applications of time series analysis, and this test is jointly developed by Ljung and Box[4]. According to them, the algorithm of Ljung-Box test is: (i) H_0 : the data are independently distributed against H_a : the data possess some serial correlation up to a certain lag h (ii) The quantity $Q = n(n+1)\sum_{k=1}^{h}[(n-k)^{-1} r_k^2]$, which is a function of sample autocorrelation r at lag k and sample size n, denotes the test statistic (iii) $Q \sim \chi^2_{(h)}$ under H_0 and reject the null hypothesis if $Q > \chi^2_{(1-\alpha,h)}$ where $\chi^2_{(1-\alpha,\ h)}$ is the $(1-\alpha)^{th}$ quintile of the χ^2 distribution with h degrees of freedom.

III. MCMC and RoU Methods

In this section, we will discuss the MCMC and the RoU methods in detail by considering suitable examples.

MCMC

MCMC methods provide a way of simulating random variables from an arbitrary density $\pi(\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ needs only be known up to a normalizing constant, and $\boldsymbol{\theta}$ can be a high dimensional. The basis for MCMC methods is the combination of convergence and ergodic properties of a Markov chain. The basic idea is: (i) to sample from distribution $\pi(\boldsymbol{\theta})$ simulate a Markov chain with stationary distribution π (ii) to estimate any function of density π use the ergodic average of the chain. Two most commonly used MCMC techniques are Metropolis-Hastings (MH) and Gibbs sampling algorithms, and both of these techniques can be used to simulate random variables from an arbitrary density (possibly multivariate) known up to normalizing constant.

MH Algorithm

MH algorithm starts working by firstly defining a proposal density $P(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, also known as conditional density of $\boldsymbol{\theta}^*$ given $\boldsymbol{\theta}$**.** Using this conditional density, the algorithm simulates a Markov chain $\{\boldsymbol{\theta}_n\}$ through steps mentioned in Algorithm 1 whose stationary distribution is $\pi(\boldsymbol{\theta})$.

To generate a sample of size n from $\pi(\boldsymbol{\theta})$, Algorithm 1 needs to run n number of times. Algorithm 1 presented here is known as generalized version of the original Metropolis-Hastings algorithm. The original Metropolis-Hastings algorithm invented by Metropolis et al[6]. allows only symmetric proposal ($P(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ is said to be symmetric when $P(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = P(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$). On the other hand, generalized version of it proposed by Hastings[2] allows any normalized probability density as a proposal. Under generalized MH, proposal density does not need to be symmetric. Let's explain this symmetric idea (different from the classical idea of symmetric distribution) by considering an one dimensional problem. Suppose the proportional form of $P(\theta^*|\theta)$ be $\exp(\theta^* - \theta)^2$ then $P(\theta^*|\theta)$ and $P(\theta|\theta^*)$ remain the same form when their positions are swapped. In this case, we can say $P(\theta^*|\theta)$ is symmetric proposal. But if $P(\theta^*|\theta)$ be $\exp(\theta^* - 0.5\theta)^2$, $P(\theta^*|\theta)$ and $P(\theta|\theta^*)$ do not have the same form when their positions are swapped. In this case, $P(\theta^*|\theta)$ is considered a non-symmetrical proposal density.

Algorithm 1: Metropolis-Hastings algorithm

Input: Current value of $\boldsymbol{\theta}$.

Output: Simulated value from $\pi(\boldsymbol{\theta})$.

Begin

1. Propose new state $\boldsymbol{\theta}^*$ from the $P(\cdot\,|\boldsymbol{\theta})$
2. Calculate $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = min\left[1,\ \frac{\pi(\boldsymbol{\theta}^*)\, P(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})\, P(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right]$
3. Generate $U \sim Uniform(0,1)$
4. **IF** $U < \alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta})$**then**
   - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^*$

   Else
   - $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}$

   End If

End Begin

To illustrate how MH works we consider a simple task where we aim to simulate from $\pi(\theta) \propto exp(-2\theta), 0 < \theta < \infty$, using MH algorithm. To simulate from $\pi(\theta)$ using MH first we need to consider a transition kernel (proposal density) by which Markov chain moves to new position θ^* from an initial point θ. By adding a point y, generated from the density f i.e. $y \sim f$, with an initial point θ transition kernel simply moves to a new point θ^* i.e. $\theta^* = \theta + y$ and this kind of kernel is known as random walk kernel. Under the setting of random walk kernel there are many common

choices for f including the uniform distribution on $[-a, a]$ for some $a > 0$, the normal distribution and the $t-$ distribution.The distribution of θ^*depends on the choice of f. Apart from random walk kernel there is another kind of kernel known as independent sampler, which draws new position independently of the current state of the chain. In this paper, we will use random walk kernel, and suppose we consider $f \sim N(0, \sigma^2)$ then the proposal density becomes $Q(\theta^*|\theta) \sim N(\theta, \sigma^2)$. To implement MH here we consider $\sigma^2 = 1$, and the detail discussion regarding how to choose σ^2 for a particular problem will be discussed in section V. Choosing $Q(\theta^*|\theta) \sim N(\theta, 1)$ makes proposal density symmetrical, yielding $\alpha(\theta^*|\theta) = min\left[1, \frac{\pi(\theta^*)}{\pi(\theta)}\right] = min\,[1,\ exp\{-2(\theta^* - \theta\}]$. Figure 1 plots the generated samples from $\pi(\theta) \propto exp(-2\theta)$ where MH is used.

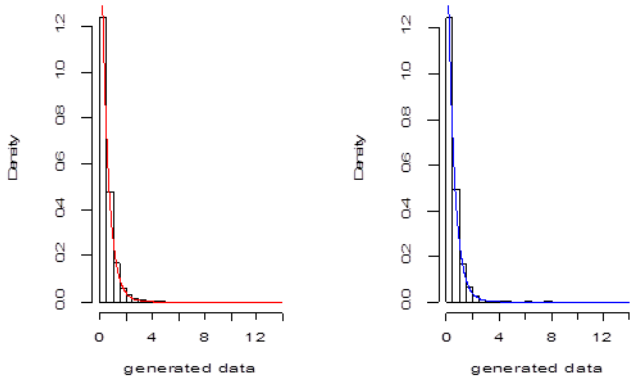

Fig. 1. Overlaying exponential with density (rate=2) lines over histograms of generated samples produced by MH algorithms. Markov chain started from $\theta = 5$ (left plot) and $\theta = 8$ (right plot) in two cases.

Ratio-of-Uniforms Method

Ratio-of-uniforms is one of the random variates generation techniques from an arbitrary probability density, often specified up to proportionality, under acceptance-rejection framework which was proposed by Kinderman and Monahan[3]. Unlike, the conventional acceptance-rejection method, this technique doesn't require any proposal density to sample from an arbitrary probability density. Suppose our aim is to simulate from $\pi(\theta) = \frac{\pi_1(\theta)}{\int \pi_1(\theta)\, d\theta} = \frac{\pi_1(\theta)}{c} \propto \pi_1(\theta)$, where $c = \int \pi_1(\theta)\, d\theta$ is the normalizing constant. Kinderman and Monahan[3] showed that if the joint density of two uniform random variables is uniformly distributed on

$$R = \left\{(u, v): 0 < u \leq \sqrt{\pi_1\left(\frac{v}{u}\right)}\right\} \qquad (1)$$

then the variable $X = \frac{V}{U}$ has probability density function $\pi(x) = \frac{\pi_1(x)}{c} \propto \pi_1(x)$. To generate (U, V) uniformly over R, the boundary of the region R needs to be specified firstly. For a, b_1 and $b_2 \in \mathbb{R}$, Kinderman and Monahan[3] enclosed R in a rectangle $[0,\ a] \times [b_1,\ b_2]$ provided that the following theorem is hold:

Theorem 1: The region R will be enclosed in a rectangle $[0,\ a] \times [b_1,\ b_2]$ subject to the conditions that $\pi_1(x)$ and $x^2\pi_1(x)$ are bounded where $a = sup_x \sqrt{\pi_1(x)}$, $b_1 = inf_{x \leq 0}\, x\sqrt{\pi_1(x)}$ and $b_2 = sup_{x \geq 0}\, x\sqrt{\pi_1(x)}$. We do not consider the proofs of equation 1 and theorem 2 as proofs are available in their paper. The theoretical acceptance probability, P_{accep}, of a point generated in the bounding rectangle under the ratio-of-uniforms method is given by

$$P_{accep} = \frac{Area\ of\ R}{Area\ of\ rectangle} = \frac{c}{2a(b_2 - b_1)} \quad (2)$$

Finally, for symmetrical unimodal densities, Kinderman and Monahan[3] showed that the probability of acceptance P_{accep} is maximized when mode of these densities (μ) is relocated to zero which is stated below in Theorem 2.

Theorem 2: Without loss of generality, mode $(x = \mu)$ of a positive symmetric function $\pi_1(x)$ defined on $\mathbb{R}$ can be rescaled to $x = 0$. Furthermore, provided that $sup_x[\pi_1(x)]^{0.5} < \infty$ and $sup_x\, x^2\, \pi_1(x) < \infty$, then sampling from $\pi_1(x)$ is equivalent to sampling from $\pi_1(x - \mu)$. Under these conditions, P_{accep} is maximized when $\mu = 0$. The proof of the above theorem is not considered here but available in Kinderman and Monahan[3] paper. The detail procedure of ratio-of-uniforms method to simulate a sample of size n from an arbitrary probability density $\pi(x) = \frac{\pi_1(x)}{c} \propto \pi_1(x)$ with bounded $\pi_1(x)$ and $x^2\pi_1(x)$ is summarized in algorithm 2.

Algorithm 2: Algorithm of ratio-of-uniforms method
Input: Bounding constraints a, b_1 and b_2
Output: Produce X from the target density $\pi(x)$
Begin
For $i = 1, 2, \cdots\cdots, n$ **do**
1. Generate $U_1, U_2 \sim uniform(0,1)$
2. Calculate $U = a * U_1$ and $V = b_1 + (b_2 - b_1) * U_2$
3. **If** $U \leq \sqrt{\pi_1\left(\frac{V}{U}\right)}$ **then**
   • $X = \frac{V}{U}$
Else
   • Go back to step 1
End If
End For loop
• Return all $X_1, X_2 \cdots, X_n$ as a desired sample
End Begin

IV. Simulating RV via RoU Method

In this section, we will show how RoU method can be used to simulate random variates from arbitrary densities. Here we aim to simulate from the density $\pi(\mu|x) \propto e^{-0.5*(\mu - x)^2} \times (1 + \mu^2)^{-1}$ (posterior density) considered in section 1.

Case 1: $x = 0$ observed

$e^{-0.5*\mu^2}(1 + \mu^2)^{-1} / \int e^{-0.5*\mu^2}(1 + \mu^2)^{-1}\, d\mu$, where $\pi_1(x) = e^{-0.5*\mu^2}(1 + \mu^2)^{-1}$ and $c = \int \pi_1(x)\, d\mu = 1.645$. The value of the normalizing constant $c = 1.645$ is obtained by

approximating this intractable integral using the Monte Carlo integration[1].To implement RoU method, we need to find the values of a, b_1, b_2 and P_{accep}, which can be calculated using the Theorem 1 and the equation 2. (i) First find $a = sup_{\mu}\sqrt{\pi_1(\mu)} = sup_{\mu}(1+\mu^2)^{-0.5}\, e^{-\mu^2/4}$. After taking natural logs in both sides, we have $\log a = sup_{\mu}[-0.5\log(1+\mu^2) - \mu^2/4]$. Maximizing $[-0.5\log(1+\mu^2) - \mu^2/4]$ with respect to μ requires first and second derivatives of $\log a$ which are $-\mu/(1+\mu^2) - \mu/2$ and $-(1-\mu^2)/(1+\mu^2)^2 - 1/2$ respectively. Solving $-\mu/(1+\mu^2) - \mu/2 = 0$ yeilds $\mu = 0$ at which a is maximized as the value of second derivative is $-3/2$, which is negative. Hence $a = sup_x(1+\mu^2)^{-0.5}\, e^{-\mu^2/4} = 1$. (ii) To find b_1 and b_2, we need to minimize and maximize $\mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4}$ for $\mu \le 0$ and for $\mu \ge 0$ respectively. From the graph of $y = \mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4}$ (shown in Figure 2), we can see that all the y values are negative when $\mu \le 0$ (negative function). From calculus theory, we know that minimizing of a negative function is equivalent to maximize it. Here we will use this idea to minimize any negative function.

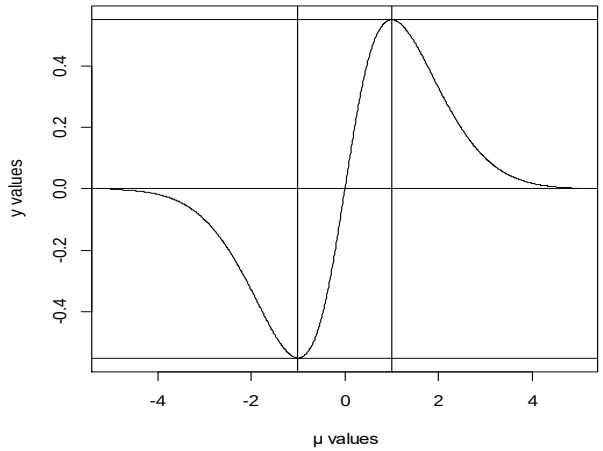

Fig. 2. The graph of $\mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4}$ for $-5 \le \mu \le 5$, vertical and horizontal lines are drawn at ± 1 and at ± 0.550 respectively.

Let's start with finding the value of $b_2 = sup_{\mu\ge 0}\,\mu\sqrt{\pi_1(\mu)} = \sup_{\mu\ge 0}\,\mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4}$. Like earlier, the first and second derivatives of $\log(\mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4})$ with respect to μ are $\left(\frac{1}{\mu} - \frac{\mu}{2} - \frac{\mu}{1+\mu^2}\right)$ and $\left(-\frac{1}{\mu^2} - \frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}\right)$ respectively. Solving for $\left(\frac{1}{\mu} - \frac{\mu}{2} - \frac{\mu}{1+\mu^2}\right) = 0$ yields $\mu = 1$, which is done using *R* package named rootSolve, for which the value of the second derivative is $-1.5 < 0$. Therefore, $\mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4}$ has a maximum at $\mu = 1$, and the maximum value is $b_2 = 0.550$.

Over the domain $\mu \le 0$ all the values of y are negative, so y is negative function. Therefore, finding the minimum value (b_1) of y is equivalent to finding the maximum value of y. Solving the first derivative of function y over the domain $\mu \le 0$ yields $\mu = -1$ for which the value of the second derivative be negative $(-1.5 < 0)$. Therefore, $\mu\,(1+\mu^2)^{-0.5}e^{-\mu^2/4}$ has a maximum at $\mu = -1$, and the maximum value b_1 be -0.550.

Finally, plugging the values of a, b_1 and b_2 into equation 2 yields the theoretical acceptance rate of a point generated in the bounding rectangle which is $P_{accep} = \frac{c}{2a(b_2 - b_1)} = \frac{1.645}{2*1*(2*0.550)} = 0.747$.

Table 1. The values of a, b_1 and b_2 along with required first and second derivatives to find them for different observed values of x are shown in columns 2-3 respectively. Columns 4-5 present the theoretical acceptance rate and mode of the target density (value for which a is maximized).* indicates values for relocated density.

x	a (1st & 2nd derivatives)	b_1 & b_2 (1st & 2nd derivatives)	$P_{acp.}$	Mode (target)
2	$a = \frac{e^{-0.25}}{\sqrt{2}}$ $-\frac{\mu}{1+\mu^2} - \frac{(\mu-2)}{2}$ & $-\frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}$	$b_1 = -0.95$ & $b_2 = 0.90$ $\frac{1}{\mu} - \frac{\mu}{1+\mu^2} - \frac{(\mu-2)}{2}$ & $-\frac{1}{\mu^2} - \frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}$	.653	1
4	$a = 0.258$ $-\frac{\mu}{1+\mu^2} - \frac{(\mu-4)}{2}$ & $-\frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}$	$b_1 = -0.003$ & $b_2 = 0.97$ $\frac{1}{\mu} - \frac{\mu}{1+\mu^2} - \frac{(\mu-4)}{2}$ & $-\frac{1}{\mu^2} - \frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}$	.357	3.46
8	$a = 0.125$ $-\frac{\mu}{1+\mu^2} - \frac{(\mu-8)}{2}$ & $-\frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}$	$b_1 \approx 0$ & $b_2 = 0.992$ $\frac{1}{\mu} - \frac{\mu}{1+\mu^2} - \frac{(\mu-8)}{2}$ & $-\frac{1}{\mu^2} - \frac{1-\mu^2}{(1+\mu^2)^2} - \frac{1}{2}$	.162	7.74
8	$a^* = 0.125$	$b_1^* = -0.11$ & $b_2^* = 0.11$	.733*	0*

Case II: Other Scenarios

In case I, we have shown the procedure for finding the quantities a, b_1, b_2 and P_{accep} in detail when $x = 0$ is observed. Here we consider other different scenarios (observed different x values) and find the values of same quantities for each of the cases which are presented in Table 1.

From Table 1, we observe that theoretical acceptance rate (P_{accep}) decreases when mode of the target density moves away from zero. As the target density considered here is symmetric, we can apply Theorem 2 to simulate from this density i.e. relocating the mode of the target density to zero and apply appropriate transformation on simulated data to get the sample from the desired density. Here we will show how relocating of the mode of an arbitrary symmetrical density to zero yields a higher acceptance rate by considering the last case of Table 1. The relocated version of $\pi_{\mu=7.74}(\mu|x) \propto e^{-0.5*(\mu-8)^2} \times (1+\mu^2)^{-1}$ is $\pi_{\mu=0}(\mu|x) \propto$

$e^{-0.5*(\mu+7.74-8)^2} \times (1+(\mu+7.74)^2)^{-1}$, where suffix in π indicates the mode of the density. For relocated density $\pi_{\mu=0}(\mu|x)$, we have $a = 0.125,\ b_1 = -1.4,\ b_2 = 1.4$ and normalizing constant $c = .0403$ which altogether produce $P_{accep} = 0.733$. We have not provided here detail calculation of a, b_1 and b_2 but Figure 3 justifies the values of b_1 and b_2 which are mentioned here.

V. Simulating RV via MCMC Method

In this section, we apply MH technique to generate data from the arbitrary density considered in section 4. We have already described in detail how MCMC technique can be used to generate data from an arbitrary density in section 3. To simulate data from $\pi(\mu|x) \propto e^{-0.5*(\mu-x)^2} \times (1+\mu^2)^{-1}$, we use a random walk kernel to move from an initial point to a new point where proposal density is $Q(\mu^*|\mu) = (2\pi\sigma^2)^{-1} e^{-0.5\left(\frac{\mu^*-\mu}{\sigma}\right)^2} \sim N(\mu, \sigma^2)$. As $Q(\mu^*|\mu) = Q(\mu|\mu^*)$, their ratio $Q(\mu|\mu^*)/Q(\mu^*|\mu)$ becomes 1. Therefore, acceptance probability $\alpha(\mu^*|\mu)$ defined in step 2 of Algorithm 1 simply boils down to $min\left[1, \frac{\pi(\mu)}{\pi(\mu^*)}\right]$. After simplification the ratio $\pi(\mu)/\pi(\mu^*)$ becomes $\frac{1+\mu^{*2}}{1+\mu^2} \times e^{-0.5(\mu^2-\mu^{*2}+2\,\mu^* x-2\,\mu\, x)}$. To calculate the value of this expression we need to have the values of μ^*, μ and x. For a particular problem, the value of x will be observed while μ and μ^* are the initial and proposed values respectively.

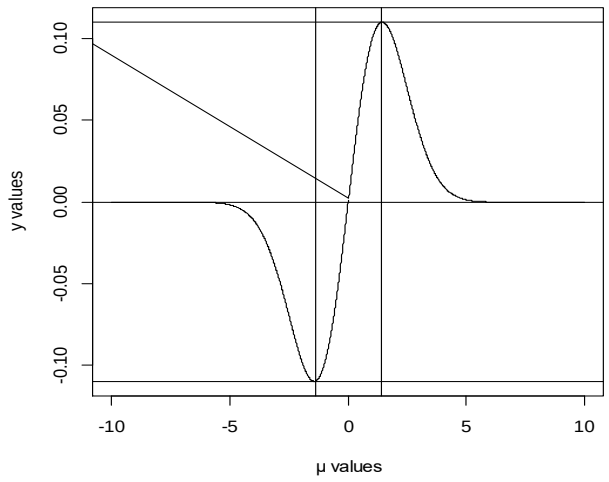

Fig. 3. The graph of $\mu\, e^{-0.25*(\mu+7.74-8)^2} \times (1+(\mu+7.74)^2)^{-0.5}$ for $-10 \le \mu \le 10$, vertical and horizontal lines are drawn at ± 1.4 and at ± 0.110, respectively.

The quantity σ^2 tells how much the proposed μ^* will be far from an initial point μ. Big jump from the current point μ requires large σ^2, which is necessary for any Markov chain to explore the parameter space quickly. However, choosing large value of σ^2 may finally have low acceptance rate because of rejection of too many proposals. As a consequence, the Markov chain often stays in the same place. On the other hand, maintaining high acceptance rate in a Markov chain demands a small jump from the current point (small value of σ^2) but it invites another problem called slow exploration of the parameter space (require large number of iterations to explore the whole parameter space). Unfortunately, having a Markov chain which possess these two criteria together (high acceptance rate and quick exploration) is quite challenging, and most of the time a trade-off is made between these two criteria. Therefore, we consider four different values of σ^2 i.e. $\sigma^2 = 0.5, 1, 1.5$ and 2 in our random walk Markov chain.

VI. Results and Discussions

In this section, both theoretical and simulation results obtained under RoU and MH methods are presented along with discussions. Here all numerical computations are computed in R on a Samsung XI machine with an Intel (R) Core (TM) i7-4900 (single) processor.

Table 2 shows the theoretical acceptance rate (P_{accep}) of a point generated in the bounding rectangle under RoU method for different values of x. From this table, it is clear that P_{accep} decreases when mode of the target density moves away from zero, and P_{accep} is maximized when mode is zero.

Table 2. Theoretical (P_{accep}) and simulated $(\widehat{P}_{accep})$ acceptance rates for different values of x under RoU method.

$X: x$	0	2	4	8
Mode	0	1	3.46	7.74
P_{accep}	0.747	0.653 -	0.357 -	0.162 0.733*
$\widehat{P}_{accep}$	0.743	-	-	0.729

Table 2 also shows the simulated acceptance rate $(\widehat{P}_{accep})$ for $x = 0$ and 8^*which is very close to theoretical P_{accep}. The simulated acceptance rate is calculated based on a sample of size ten thousand (10,000). We have used random seed number to produce $\hat{P}_{accep}$, and we also observed that using different seed numbers produce approximately similar results. From the above discussions, it is observed that acceptance rate of a point generated in the bounding rectangle decreases when mode of the target density moves away from 0, and it is maximized when mode is zero. From Table 2, it is observed that theoretical acceptance rate drops from 0.747 to 0.162 when mode moves from zero (0) to 7.70. However, relocation of the mode of $\pi_{\mu=7.74}(\mu|x)$ to zero i.e. $\pi_{\mu=0}(\mu|x)$ increases the theoretical acceptances rate from 0.162 to 0.733^*. Our simulation study also confirms this acceptance rate which is shown in Table 2. After generating sample from the relocated density, $\pi_{\mu=0}(\mu|x)$, we need to transform (add 7.74 with every element) this sample such that sample comes from $\pi_{\mu=7.74}(\mu|x)$. Figure 4 presents the simulated results: generating sample from $\pi_{\mu=7.74}(\mu|x)$ (right plot) through $\pi_{\mu=0}(\mu|x)$ (left plot). We haven't considered the other cases for relocating as all the approaches are the same.

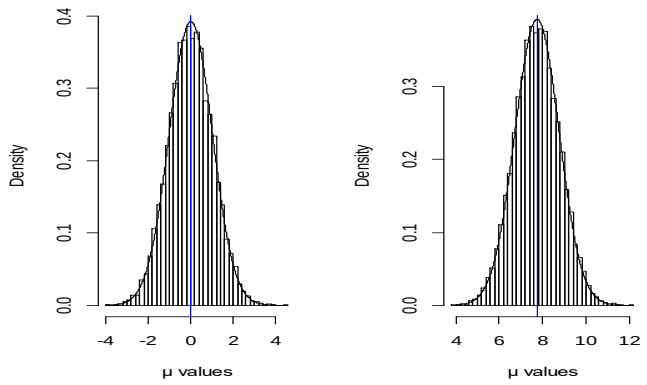

Fig. 4. Overlaying $\pi_{\mu=0}(\mu|x)$ (left) and $\pi_{\mu=7.74}(\mu|x)$ (right) density lines over histograms of generated samples (size 10^4) respectively. Vertical lines are drawn at $\mu = 0$ and 7.74 points.

Before starting evaluating the performance of MH sampler for simulating sample from $\pi(\mu|x)$, firstly we discuss here some of the important issues of MCMC implementation. First of all sample produced by Markov chain are no longer independent as successive observations are correlated which needs to take into account for making any valid inference. As a possible remedy of successive correlation sample are thinned (retaining only every k^{th}observation) so that resulting sample is close to independent. Secondly, we have the convergence issue of a Markov chain that we need to address. Suppose $\{\mu^{(1)}, \mu^{(2)}, \cdots \mu^{(b-1)}, \mu^{(b)}, \mu^{(b+1)}, \cdots, \mu^{(n+1)}, \mu^{(n)}\}$ be the generated sample produced by a Markov chain which has converged after b iterations then any inference regarding any unknown parameter should be made based on $\{\mu^{(b+1)}, \cdots, \mu^{(n+1)}, \mu^{(n)}\}$. Points up to b^{th} iterations are discarded which are known as burn in period, and Monte Carlo estimator based on remaining points still considered as unbiased estimator[1].

From Table 3, we observe that with the increases of σ^2 values the acceptance rate of MH sampler decreases whilst Markov chain converges quickly (require less number of iterations). The later one is also supported by the Figure 5. These findings are consistent with the theory. For large σ^2 Markov chain jumps far away from initial value resulting in quick convergence as well as low acceptance rate. From Table 3 and Figure 6 (left plot), it is also observed that generated sample are highly correlated (lag are significant up to 5000 lag, p-value of Ljung-Box test is very small and shown in parenthesis) for all cases. For practical implementation generated sample should be close to independent while a trade-off between acceptance rate and convergence time is necessary to made. The problem we considered in this paper is relatively easy that's why MH sampler only needs 20-70 iterations to converge but for a complicated problem MH sampler may take long time to converge.

Convergence time of a Markov chain varies when it has started from different initial points (not shown here) but we have observed similar type of pattern under different σ^2 values.

Table 3. Simulated acceptance rate, significant lag order and burn-in period of generated samples obtained by MH over different σ^2 values.

σ^2 value	$\hat{P}_{accep}$	Sig. Lag.Order	Burn-in period
0.5	0.789	5000 (<2.2e-16)	70
1.0	0.703	5000	40
1.5	0.655	5000	20
2.0	0.604	5000	20

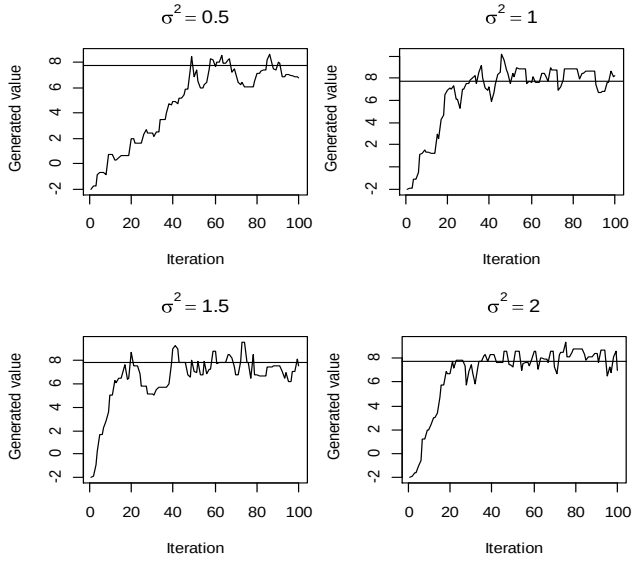

Fig. 5. Trace plots of μ values where first 100 points are considered (visualize clearly) to plot. For all case, Markov chain has started at initial point $\mu = -2$ while horizontal lines are drawn at median point (7.74).

Table 4. Ljung-Box test statistic and their corresponding *P* values in parenthesis at different lags

Thinning (every n^{th}points)	Lag (1)	Lag (3)	Lag (5)	Lag (8)
$n = 10$	9.688 (0.00185)	12.427 (0.006054)	13.384 (0.02003)	14.379 (0.0724)
$n = 20$	1.0575 (0.3038)	-	-	-

Any statistical inference should be made based on independent sample. From Table 3 and Figure 6 (left plot), we have seen that sample produced by MH sampler is highly correlated (lag is significant even after 5000 lag) irrespective of σ^2 values. To draw independent sample using MH sampler we need to thin the generated sample. Table 4 shows the test of independence of thinned sample where thinning is made by keeping only every 10th and 20th observations. From Table 4, we have seen that even thinned sample (where every 10th observation is kept) exhibits some sort of correlation (lag is significant up to lag 5, p value (.02) < 0.05) but after that lag is insignificant. On the other hand, thinned sample (where every 20th observation is kept)

does not have any correlation, lag is insignificant at lag 1 (p-value 0.3038< 0.05). This is also supported by acf plot shown in Figure 6 (right).

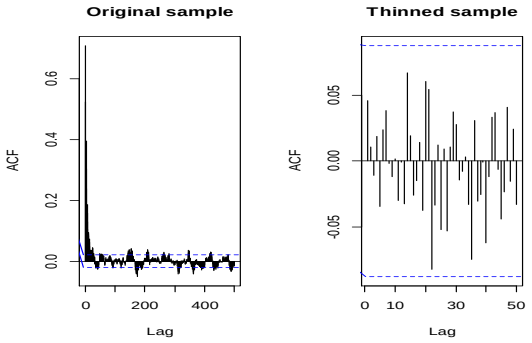

Fig. 6. ACF plots of generated sample produced by MH algorithm: (i) left plot is drawn based on original sample where up to 500 lag shown (ii) right plotis drawn based on thinned sample (every 20th observation) where up to 50 lag shown.

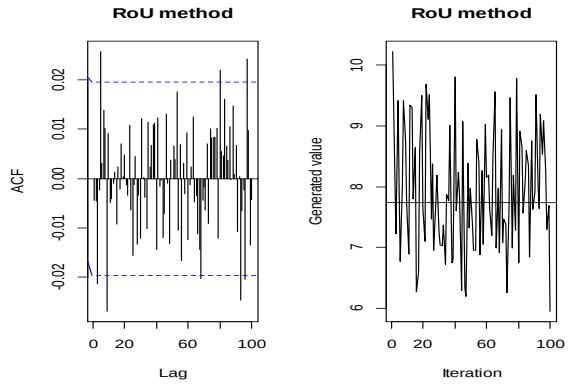

Fig. 7. ACF and trace plots of generated sample produced by RoU method respectively: (i) left plot (acf) is drawn based on all sample observations where up to 100 lag shown (ii) right plot (trace) is drawn based on only first 100 values (visualize clearly).

On the other hand, sample produced by RoU method doesn't have any correlation which is examined again by using Ljung-Box test (p value for testing lag 1 is 0.4659) but result is not shown here. This is shown in Figure 7 (left plot) while right plot show the trace plot of generated value which converges to mode from the very beginning.

Finally, we can say that both MH sampler and RoU method can be used to simulate from any arbitrary density which is known up to normalizing constant. However, using MH sampler one needs to address properly the MCMC issues (correlated sample, acceptance rate and convergence time). On the other hand, implementing RoU method does not require to check these sort of issues. The only one concern of RoU method is to implement it for non-symmetrical target density as relocating via the mode does not help too much as far as acceptance rate is concerned. To draw an independent sample of size 1000, MH sampler requires 20,000 observations which is 20 times higher than number of observations required in RoU method in the context of problem considered in this paper. Considering a more complicated density MH sampler may require even 100 times more observation than that of RoU method require.

VII. Conclusion and Future Works

In this paper, we have illustrated how MCMC and RoU method can be used to simulate random variates from an arbitrary densities, and compared their performance. From our simulation study, we have established that RoU method performs better than MCMC methods as far as quality of the generated sample (randomness) and computational context (to draw 1000 independent observations MH sampler requires 20,000 observations which is 20 times higher than number of observations required in RoU). However, for non-symmetrical density how RoU method will be relocated to increase the acceptance rate still unexplored. Finding the optimal value by which relocating non-symmetrical density yields higher acceptance rate is still ongoing research and one can take it as a future research. In addition to that, one can also try to see how these two methods perform when target density is multi-modal.

References

1. Hammersley, J. M. and D. C. Handscomb, 1964. Monte Carlo Methods. London: Methuen, 113-114.
2. Hastings, W., 1970. Monte Carlo Sampling Methods using Markov chains and Their Applications. Biometrika. 57(1), 97–109.
3. Kinderman, A. J., and J. F. Monahan, 1977. Computer Generation of Random Variables Using the Ratio of Uniform Deviates. ACM Transactions on Mathematical Software 3 (3), 257–260.
4. Ljung, G. M. and G. E. P. Box, 1978. On aMeasureof Lack of Fit in Time Series Models. Biometrika.65, 297–303.
5. Martino, L. and J. Míguez, 2010. A rejection sampling scheme for posterior probability distributions via the ratio-of-uniforms method.18th European Signal Processing Conference, Aalborg. 17-178.
6. Metropolis, N., A.W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, andE. Teller, 1953. Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics. 21(6), 1087–1092.
7. Okwuokenye, M and KE. Peace, 2016 A comparison of inverse transform and composition methods of data simulation from the Lindley distribution.Communications for Statistical Applications and Methods. 23(6), 517-529.
8. Wakefield, J. C., A. E. Gelfand and A. F. M. Smith, 1991. Efficient Generation of Random Variates via the Ratio-of-Uniforms Method. Statistics and Computing 1, 129–133.