Uniform Random Real Number Generator Discretized into Floating Point Number by Rounding Function

Shotaro Ishida[1] Reiji Suda[1]

Abstract: Many papers discuss various methods to generate integer uniform random numbers on a computer. On the other hand, there are a few method to generate floating point uniform random numbers (or transform integer uniform random numbers to floating point ones). Under these circumstances, we divide integer uniform random numbers by a constant number (e.g., $rand\left(\right)/2^{32}$) in order to obtain floating point uniform random numbers. However, this method can output only specific form floating point numbers and can not generate most of representable floating point numbers. To avoid this problem, Moler proposed a uniform random number generator that can generate all floating point numbers in $\left[2^{-53}, 1-2^{-53}\right]$, and then Thoma expanded its range into $(0, 1)$.

By experimental and theoretical inspection, however, we found that the method proposed by Thoma made strange behavior according to floating point rounding mode. For example, generating probability of a specific floating point number is 3 times as high/low as that of the neighbor one. Moreover, Moler did not mention a method to change the random number generation range and Thoma did not guarantee that we can generate all the floating point numbers in the new range.

Accordingly, this paper aims to propose a modified method without the strange behaviors appearing in Thoma's method, to expand the random number generation range into arbitrary one we desire, and to construct a higher precision floating point uniform random numbers generator than normal IEEE754 numbers. In order to achieve these aims, this paper will discuss what floating point uniform random number is, and then propose one of such generator and prove its correctness, and lastly show its performance of the generator by experiment.

Keywords: Uniform random number, Discretization, IEEE754, Floating point number, Double-Double precision, Arbitrary range, Moler, Thoma

[1] Graduate School of Information Science and Technology, The University of Tokyo

1. Introduction

1.1 Background

Whatever random numbers we want to use on a computer, uniform random number is basically required. So, uniform random number is the most important random number on a computer. In fact, uniform random numbers have been used in quite many situations, such as generating a random number that follows several distributions or calculating high dimension numerical integration by Monte Carlo [24] method. The followings are examples of generating Gaussian random numbers by using uniform random numbers. Box-Muller [4] method generates two independent Gaussian random numbers from two independent uniform random numbers and Polar [2,14] method removes the trigonometric calculations in Box-Muller method. Kabal [11] proposed piecewise linear approximation of the Gaussian distribution by using some triangular distributions. Ziggurat [21] method divides the probability density function of the Gaussian distribution into rectangles with all equal area along the horizontal axis. On the other hand, Monty-Python [20] method divides the probability density function into several pieces by using affine transformation and then embeds them into a rectangle whose area is 1. In addition to those methods, there are several methods, such as Acceptance-Rejection [17] method, Odd-Even [5] method, and Ratio-of-Uniform [13] method.

Under these circumstances, many papers discuss various method to generate integer uniform random numbers on a computer. Middle-Squared [32] method, Linear Congruential Generator [28], Xor-Shift [19], Mersenne Twister [23], and other methods [3,29] are a few of such examples. On the other hand, there are few methods to generate floating point uniform random numbers (or transform integer uniform random numbers to floating point format ones). When floating point uniform random numbers are required, we have often divided integer uniform random number by a constant value (e.g., $rand\left(\right)/2^{32}$) in order to obtain floating point uniform random numbers. However, this method can output only a very small fraction of floating point numbers and can not generate most of representable floating point numbers. Box-Muller method is one of examples where this property has bad influence. The method takes logarithm of a uniform random number u_1 and lets $a = \sqrt{-2\log_e\left(u_1\right)}$. Then the method generates another uniform random number[*1] u_2 and lets $b = 2\pi u_2$. At last, the method outputs two independent Gaussian random numbers $a\sin\left(b\right)$ and $a\cos\left(b\right)$. Here, since $\log\left(u_1\right)$ decides the absolute value of a, the sparser uniform random numbers near 0 are, the smaller the absolute value of the output becomes. Therefore, we can not reproduce each edge of the Gaussian distribution if uniform random number is sparse near 0.

To solve this problem, Moler [25,26] focused on a mantissa of floating point numbers and proposed a floating point uniform random number generator that can output all the double precision floating point number in $\left[2^{-53}, 1-2^{-53}\right]$. This method generates a floating point uniform random number that is a integer multiples of 2^{-53} first. Then, the algorithm generates an additional uniform random integer and takes a bitwise-xor to the mantissa of floating point random number with the integer random number. This means that the algorithm generates the exponent and mantissa of a floating point uniform random number separately. In the concrete, the algorithm first prepares 32 initial random numbers(seed) that is all integer multiples of 2^{-53}, $z_0, z_1, \ldots, z_{31}$, and a borrow flag b. Then, the algorithm generates a floating point uniform random number by using the following recurrence relation[*2].

$$z_i = z_{i+20} - z_{i+5} - b.$$

Here, each subscript, $i, i+20, i+5$, is calculated on module 32. Besides, if the operation makes z_i be negative, then add 1 to z_i and then set $b = 2^{-53}$, the half of the machine epsilon[*3]. Otherwise, set $b = 0$. In practical application, MATLAB version 5 has adopt the algorithm for its floating point uniform random number generator.

Subsequently, Thoma [31] proposed floating point uniform random number generator that could output all the floating point numbers in $(0,1)$. Thoma originally aimed to construct a Gaussian random number generator that could reproduce the tail region of the distribution. In this research, Thoma required uniform random number generator specialized for floating point numbers, which a Gaussian random number generator used.

Experimental and theoretical inspections, however, show that Thoma's method contains strange behaviors in some floating point rounding modes. For example, Thoma's method can not output floating point numbers in the subnormal area (quite close floating point numbers to 0) and the generation probability of a specific floating point number is 3 times as high or low as that of its neighbors. Figure 14 in Section 5 shows the behaviors. Additionally, Thoma's method does not guarantee that all the floating point numbers can be generated when we apply the method to another ranges other than $(0,1)$. Worse, Moler did not mention how to apply the method to another range except $\left[2^{-53}, 1-2^{-53}\right]$.

1.2 Objective

Now, we have 2 problems. One is that Thoma's method shows strange behaviors, and the other is that we can not apply

[*1] "another uniform random number" means a uniform random number that is independent of u_1.
[*2] This algorithm is based on idea by Marsaglia [18,22].
[*3] That is, the half of the difference between 1 and the minimal floating point numbers that is greater than 1.

both Moler's method and Thoma's method to another ranges other than its original range with guaranteeing that all the floating point numbers can be generated. So, the aim of this paper is as follows.

(1) Modifying strange behaviors in Thoma's method.
Remove the strange behaviors in Thoma's method because probability is one of the most important point in random number generation.

(2) Constructing a generator that can output all the values in arbitrary range.
Since Moler's method and Thoma's method are specialized for the given range, this paper will propose a generator that can output all the floating point numbers in $[a, b]$ for arbitrary floating point number a and b.

In order to achieve those aims, (1) this paper will define the concept of ideal uniformness and then construct a uniform random number generator that satisfies the definition and show that the strange behaviors observed in Thoma's method is removed. After that, (2) this paper will propose a generator that can output all the floating point number in an arbitrary range whose edge is a floating point number and will show its performance by experiment.

The organization of this paper is as follows.

Section 1 is the current section.
Section 2 explains IEEE754 floating point numbers as a background.
Section 3 explains Thoma's method and its problem.
Section 4 defines floating point uniform random number generator and calculates its random number generation probability.
Section 5 solves the problem explained in the Section 3.
Section 6 proposes a generator that can output all the floating point numbers in an arbitrary range whose both edge is a floating point number.
Section 7 evaluates the proposed method by experiment.
Section 8 summarizes this paper and gives a future work.

1.3 Notation

After this section, the authors use the following notation if necessary.

- $\mathbb{N}\left(n \in \mathbb{N}\right) = \{k \in \mathbb{N} \mid 0 \leq k \leq 2^n - 1\} \subseteq \mathbb{N}$.
  $\mathbb{N}\left(n\right)$ denotes the set of n-bit unsigned integers.
- $E \in \mathbb{N} \geq 1$.
  E denotes the number of bits of exponent in floating point number.
- $M \in \mathbb{N} \geq 0$.
  M denotes the number of bits of mantissa in floating point number.
- $\mathbb{F} \subset \mathbb{R}$.
  $\mathbb{F}$ denotes the set of floating point numbers.
- $val_{\mathbb{F}} : \mathbb{N}\left(1\right) \times \mathbb{N}\left(E\right) \times \mathbb{N}\left(M\right) \to \mathbb{F}$.
  $val_{\mathbb{F}}\left(s, e, m\right)$ denotes the value of a floating point number where
  $\left(\text{Sign}, \text{Exponent}, \text{Mantissa}\right) = \left(s, e, m\right)$.
- $fl_{\mathbb{F}} : \mathbb{R} \to \mathbb{F}$.
  $fl_{\mathbb{F}}$ denotes a rounding function.
- $URNG_{\mathbb{R}} : \emptyset \to U_{\mathbb{R}}$.
  $URNG_{\mathbb{R}}$ denotes a Uniform Random Number Generator on $\mathbb{R}$.
- $U_{\mathbb{R}} \subset \mathbb{R}$.
  $U_{\mathbb{R}}$ denotes the set of random numbers that $URNG_{\mathbb{R}}$ can output.
- $URNG_{\mathbb{F}} : \emptyset \to U_{\mathbb{F}}$.
  $URNG_{\mathbb{F}}$ denotes a Uniform Random Number Generator on $\mathbb{F}$.
- $U_{\mathbb{F}} \subset \mathbb{F}$.
  $U_{\mathbb{F}}$ denotes the set of random numbers that $URNG_{\mathbb{F}}$ can output.
- $round_{\mathbb{F}} : \mathbb{R} \to \mathbb{F}$.
  $round_{\mathbb{F}}\left(r \in \mathbb{R}\right)$ denotes a sound rounding function.
- $P_{\mathbb{F}} : \mathbb{F} \to \{r \in \mathbb{R} \mid 0 \leq r \leq 1\}$.
  $P_{\mathbb{F}}\left(f \in \mathbb{F}\right)$ denotes the probability that $URNG_{\mathbb{F}}$ generates $f \in \mathbb{F}$. Of course, $P_{\mathbb{F}}\left(f \in \mathbb{F} \setminus U_{\mathbb{F}}\right) = 0$.
- $URNG_{n \in \mathbb{N}} : \emptyset \to \{i \in \mathbb{N} \mid 0 \leq i \leq 2^n - 1\}$.
  $URNG_{n \in \mathbb{N}}$ denotes an n-bit uniform random integer generator.
- $W \in \mathbb{N}$.
  W denotes the number of bits of unsigned integer on the computer.
- $FURNG : \mathbb{R}^2 \times \left(\mathbb{R} \to \mathbb{F}\right) \to U_{\mathbb{F}}$.

$FURNG\left(a, b, round_{\mathbb{F}}\right)$ denotes a $URNG_{\mathbb{F}}$ where $U_{\mathbb{R}} = [a, b]$ and the rounding mode is $round_{\mathbb{F}}$.

- $-round_{\mathbb{F}} : \mathbb{R} \to \mathbb{F}$.
  $-round_{\mathbb{F}}$ denotes a flipped $round_{\mathbb{F}}$ horizontally. This means that

$$-round_{\mathbb{F}} = \begin{cases} \text{Round-to-Nearest} & round_{\mathbb{F}} \text{ is Round-to-Nearest} \\ \text{Toward } -\infty & round_{\mathbb{F}} \text{ is Toward } +\infty \\ \text{Toward } +\infty & round_{\mathbb{F}} \text{ is Toward } -\infty \\ \text{Toward } 0 & round_{\mathbb{F}} \text{ is Toward } 0 \\ \text{Toward } \pm\infty & round_{\mathbb{F}} \text{ is Toward } \pm\infty \end{cases}$$

  Here,

$$round_{\mathbb{F}}\left(r \in \mathbb{R}\right) = f \in \mathbb{F} \Leftrightarrow \left(-round_{\mathbb{F}}\right)\left(-r \in \mathbb{R}\right) = -f \in \mathbb{F}$$

  holds.

2. IEEE754 floating point number

This section aims to explain IEEE754 floating point numbers because the problems of Thoma's method come from format and rounding mode of floating point number.

2.1 Format

One floating point number consists of the bitfields shown in Figure 1.

- Sign is a 1-bit unsigned integer.
- Exponent is an E-bit unsigned integer.
- Mantissa is an M-bit unsigned integer.

The most used pair of (E, M) in IEEE754 is the followings.

$$(E, M) = \begin{cases} (8, 23) & \text{called single precision} \\ (11, 52) & \text{called double precision} \\ (15, 112) & \text{called quadruple precision} \end{cases}.$$

Table A·1 shows more detailed information.

2.2 Value

We define the value of a floating point number whose (sign, exponent, mantissa) is (s, e, m)[*4], $val_{\mathbb{F}}\left(s, e, m\right)$, as follows[*5].

- Case: $e = 0$.
  $(-1)^s \times \left(0 + m \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)}$.
- Case: $1 \leq e \leq 2^E - 2$.
  $(-1)^s \times \left(1 + m \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)}$.
- Case: $e = 2^E - 1,\ m = 0$.
  $(-1)^s \times \infty$.
- Case: $e = 2^E - 1,\ m \neq 0$.
  NaN(Not a Number).

The floating point numbers for each case are called subnormal number, normal number, infinity, NaN (See Table 1). Hereafter, **floating point number, ($\mathbb{F}$), denotes subnormal number, normal number, and infinity**. In addition, we **distinguish** -0 **and** $+0$.

Fig. 1 Bitfield of floating point number.

Sign	Exponent			Mantissa		
s_0	e_0	$\cdots$	e_{E-1}	m_0	$\cdots$	m_{M-1}

Table 1 Classification of floating point numbers.

Class	Exponent(e)	Mantissa(m)
Subnormal numbers	$e = 0$	$0 \leq m \leq 2^M - 1$
Normal numbers	$1 \leq e < 2^E - 1$	$0 \leq m \leq 2^M - 1$
Infinity	$e = 2^E - 1$	$m = 0$
Not a Number(NaN)	$e = 2^E - 1$	$1 \leq m \leq 2^M - 1$

[*4] Note: $s \in \{0, 1\}, 0 \leq e \in \mathbb{N} \leq 2^E - 1, 0 \leq m \leq \mathbb{N} \leq 2^M - 1$

[*5] We have two types of 0, that is, $+0$ and -0.

2.3 Rounding

There are several choices for rounding a real number to a floating point number. The following rounding modes are mainly used as a rounding function $fl_{\mathbb{F}} : \mathbb{R} \to \mathbb{F}$.

- Rounding-to-Nearest
  This function rounds a real number to the closest floating point numbers to the real number. In more detailed, we have the following 3 rounding modes when there are 2 nearest floating point numbers[*6].
  - Ties to Even
    rounds to the floating point number whose mantissa, m, is even.
  - Away from ±∞
    rounds to the floating point number whose absolute value is less than the others.
  - Away from 0
    rounds to the floating point number whose absolute value is greater than the others.
- Directed-Rounding
  This function rounds a real number to a floating point number located on the given side of the real number.
  - Toward $-\infty$
    rounds to the largest floating point number that is not greater than the real number.
  - Toward $+\infty$
    rounds to the smallest floating point number that is not less than the real number.
  - Toward 0
    rounds in the same way as Toward $+\infty$ when the original real number is negative. Otherwise, round in the same way as Toward $-\infty$.
  - Toward ±∞
    rounds in the same way as Toward $-\infty$ when the original real number is negative. Otherwise, round in the same way as $+\infty$.

  Here, we define that a floating point number corresponding to real number $0 \in \mathbb{R}$ is $+0 \in \mathbb{F}$, and that if $0 < r \in \mathbb{R}$ then r is closer to $+0 \in \mathbb{F}$ than $-0 \in \mathbb{F}$, and that if $0 > r \in \mathbb{R}$ then r is closer to $-0 \in \mathbb{F}$ than $+0 \in \mathbb{F}$. Additionally, we sometimes regard $\pm\infty \in \mathbb{F}$ as $\pm 2^{\left(2^{E-1}\right)} \in \mathbb{F}$ in rounding operation.

2.4 Property

IEEE754 floating point number has the following property.

Order

If $s, e, m, s', e', m' \in \mathbb{N}$ satisfies

$$0 \leq s, s' \in \mathbb{N} \leq 1$$
$$0 \leq e, e' \in \mathbb{N} \leq 2^E - 2$$
$$0 \leq m, m' \in \mathbb{N} \leq 2^M - 1$$

then

$$val_{\mathbb{F}}\left(s, e, m\right) \leq val_{\mathbb{F}}\left(s', e', m'\right)$$
$$\Updownarrow$$
$$(-1)^s \times \left(e \times 2^M + m\right) \leq (-1)^{s'} \times \left(e' \times 2^M + m'\right)$$

is satisfied.

This property means that we can find the right/left adjacent floating point number $val_{\mathbb{F}}\left(s', e', m'\right)$ to $val_{\mathbb{F}}\left(s, e, m\right)$ by finding the right/left adjacent integer $(-1)^{s'} \times \left(e' \times 2^M + m'\right)$ to $(-1)^s \times \left(e \times 2^M + m\right)$. The authors use this property in later proofs.

3. Thoma's method

This section aims to explain Thoma's [31] floating point uniform random number generator and shows its problem.

3.1 Algorithm

Thoma's algorithm is for a floating point uniform random number generator that aims to generate all the floating point

[*6] That is, the real number is the center of two adjacent floating point numbers.

numbers in $(0, 1)$. Thoma mentions that we need $M + 1 < W$ to use the algorithm.

The main idea of Thoma's method is to regard a uniform random real number $u \in \mathbb{R}$ on $(0, 1)$ generated by $URNG_{\mathbb{R}}$ as an infinite precision floating point number. Now, let u_k be a k-th bit after the binary point of u written in binary notation$(k = 1, 2, \ldots)$, then we have

$$u = \sum_{k=1}^{\infty} u_k \times 2^{-k}.$$

Here, let u_K be the first non-zero bit of u, that is, let u_K be one of $u_1, u_2, \ldots$ that satisfies

$$\begin{cases} \forall k < K, \ u_k = 0 \\ u_K = 1 \end{cases}.$$

Then, we can write u as the following infinite precision floating point number.

$$\begin{aligned} u &= \sum_{k=1}^{\infty} u_k \times 2^{-k} \\ &= \sum_{k=K}^{\infty} u_k \times 2^{-k} \\ &= \sum_{k=0}^{\infty} u_{K+k} \times 2^{-(K+k)} \\ &= \left(\sum_{k=0}^{\infty} u_{K+k} \times 2^{-k} \right) \times 2^{-K} \\ &= \left(u_K \times 2^{-0} + \sum_{k=1}^{\infty} u_{K+k} \times 2^{-k} \right) \times 2^{-K} \\ &= \left(1 + \sum_{k=1}^{\infty} u_{K+k} \times 2^{-k} \right) \times 2^{-K}. \end{aligned}$$

The main idea by Thoma is to convert u into a floating point number on a computer by truncating the above infinite summation.

3.1.1 Pseudocode

The pseudocode of Thoma's algorithm is as follows.

```
00: Set a floating point number $c$ as the maximal value of random numbers.
      $c = 1$
10: Generate uniform random bits until the first non-zero bit is found.
      do {
        $x = URNG_W\,()$
        $c = c \times 2^{-W}$
      } while $(x \neq 0)$
20: Shift the first non-zero bit to the MSB by left-shift.
      $t = W$
      while $\left(x < 2^{W-1}\right)$ {
        $x = x \times 2$     // This is equivalent to $x = x << 1$.
        $c = c \times \frac{1}{2}$     // Divide $c$ by 2 to make $c \times x$ be constant.
        $t = t - 1$
      }
30: Add uniform random bits if necessary.
      if $(t < M + 1)$ {
        $x = x + \left(URNG_W\,() \times 2^{-t}\right)$
      }
40: Convert to a floating point number.
      return $(c \times x)$
```

3.1.2 Explanation for the pseudocode

The meaning of the pseudocode is as follows. Here, the authors explain only the line from 10 to 30 because the meaning of the line 00 and 40 is obvious.

10: Generate uniform random bits until the first non-zero bit is found.

This operation corresponds to finding u_K roughly. In the concrete, the algorithm judges whether an integer x generated by

$$x = \sum_{k=1}^{W} u_{W\times(n-1)+k} \times 2^{k-1}$$

contains u_K or not in the n-th iteration.

20: Shift the first non-zero bit to the MSB by left-shift.
This operation corresponds to finding where u_K is in x. In the pseudocode, t denotes the number of the bits from u left in x. So t is decremented by 1 every 1-bit left-shift of x. At the end of the operation in this line, the lower $(W-t)$ bits of x is 0 as a result of left-shift.

30: Add uniform random bits if necessary.
If the number of the bits from u left in x is less than the precision of floating point number, $(M+1)$[*7], then the algorithm generates an additional uniform random integer and put it on lower bit of x. Here, we can replace the operation in the if statement with

$$x = x \mid URNG_{W-t}\left(\right)$$

because x is integer.

3.2 Problem

Thoma's algorithm has the following problems based on IEEE754 floating point number. In this section, let $round_{\mathbb{F}}$ be the same as $fl_{\mathbb{F}}$[*8] from the perspective of fairness.

(A) The generation probability of 0 is higher than ideal probability $P_{\mathbb{F}}(0)$[*9].
The algorithm does not output 0 mathematically because both c and x is not 0. However, underflow of floating point number changes the situation. For example, if $URNG_W()$ generates 0 repeatedly $\lceil \frac{(M+2^{E-1})}{W} \rceil$ or more times at the line 10 in the pseudocode, then we have

$$\begin{aligned} c &\le fl_{\mathbb{F}}\left(2^{-W\times\lceil\frac{\left(M+2^{E-1}\right)}{W}\rceil}\right) \\ &\le fl_{\mathbb{F}}\left(2^{-\left(M+2^{E-1}\right)}\right) \\ &= fl_{\mathbb{F}}\left(\frac{1}{4}\times val_{\mathbb{F}}(0,0,1)\right) \\ &= 0 \quad (\text{If the rounding mode is not Toward } +\infty). \end{aligned}$$

So, c can be 0 as a result of underflow. Here, the probability that $URNG_W()$ generates 0 repeatedly $\lceil \frac{(M+2^{E-1})}{W} \rceil$ times or more is

$$\begin{aligned} 2^{-W\times\lceil\frac{\left(M+2^{E-1}\right)}{W}\rceil} &\ge 2^{-W\times\lfloor\frac{\left(M+2^{E-1}\right)}{W}+1\rfloor} \\ &\ge 2^{-\left(M+2^{E-1}+W\right)}. \end{aligned}$$

Therefore, the probability that the algorithm outputs 0 is at least $2^{-\left(M+2^{E-1}+W\right)}$. On the other hand, the ideal probability, $P_{\mathbb{F}}(0)$, is

$$P_{\mathbb{F}}(0) = \begin{cases} 2^{-\left(M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{-\left(M+2^{E-1}-2\right)} & \text{Case: Toward } -\infty \text{ or Toward0} \\ 0 & \text{Case: Toward } +\infty \text{ or Toward } \pm\infty \end{cases}.$$

So, the ratio between the random number generation probability of 0 by Thoma's method is at least 2^{W-2} times as high as the ideal probability $P_{\mathbb{F}}(0)$.

(B) Floating point numbers near 0 do not appear.
By the line 10 in the pseudocode, the algorithm guarantees $2^{W-1} \le x$. Besides, since the minimal positive value of c is greater than or equal to the minimal value of positive floating point number, $val_{\mathbb{F}}(0,0,1) \le c$ holds. Therefore, the

[*7] "+1" in $(M+1)$ comes from economized form in IEEE754 floating point number.

[*8] If $round_{\mathbb{F}}$ is different from $fl_{\mathbb{F}}$, a rounding function used on the computer differs from a rounding function used in the definition of uniformity. Hence, it is not unnatural that the algorithm does not satisfies the definition of uniform.

[*9] The definition of ideal is explained as "uniform in narrow sense" in the Section 4.1.

Table 2 Strange behaviors in Thoma's method. (A): The generation probability of 0 is higher than ideal. (B): Floating point numbers near 0 do not appear. (C): The random number generation probability is not uniform in some ranges.

Rounding mode		(A) occurs?	(B) occurs?	(C) occurs?
Round-to-Nearest	(Ties to Even)	Yes.	Yes.	Yes.
Round-to-Nearest	(Away from $\pm\infty$)	Yes.	Yes.	No.
Round-to-Nearest	(Away from 0)	Yes.	Yes.	No.
Directed-Rounding	(Toward $-\infty$)	Yes.	Yes.	No.
Directed-Rounding	(Toward $+\infty$)	No.	Yes.	No.
Directed-Rounding	(Toward 0)	Yes.	Yes.	No.
Directed-Rounding	(Toward $\pm\infty$)	No.	Yes.	No.

minimal value of positive output by Thoma's method, $c \times x$, is greater than or equal to $2^{W-1} \times val_{\mathbb{F}}(0,0,1)$. This means that the algorithm can not output a positive floating point number that is less than 2^{W-1} times as large as the minimal positive floating point number. This means that Thoma's method does not satisfy its purpose that Thoma's algorithm can output all the floating point number in $(0,1)$.

(C) The random number generation probability is not uniform in some ranges.

Consider the case where the algorithm outputs a floating point uniform random number in $\left(2^{-(W-M-1)}, 2^{-(W-M-2)}\right)$ and the rounding mode is Round-to-Nearest(Ties to Even).

First, $URNG_W$ needs to generate a random integer in $\left[2^{M+1}+2, 2^{M+2}-2\right]$ in the first time at the line 10 in the pseudocode so that the algorithm outputs a floating point number in $\left(2^{-(W-M-1)}, 2^{-(W-M-2)}\right)$. Let X be this random integer generated by $URNG_W$. Since

$$(c,x,t) = \left(2^{M+2-2W}, X \times 2^{W-M-2}, M+2\right)$$

holds at the end of line 20 in the pseudocode, the algorithm skips the if statement at the line 30 in the pseudocode and outputs

$$\begin{aligned} c \times x &= fl_{\mathbb{F}}\left(fl_{\mathbb{F}}(c) \times fl_{\mathbb{F}}(x)\right) \\ &= fl_{\mathbb{F}}\left(fl_{\mathbb{F}}\left(2^{M+2-2W}\right) \times fl_{\mathbb{F}}\left(X \times 2^{W-M-2}\right)\right) \\ &= fl_{\mathbb{F}}\left(fl_{\mathbb{F}}(X) \times 2^{-W}\right) \\ &= fl_{\mathbb{F}}(X) \times 2^{-W} \end{aligned}$$

at the line 40 in the pseudocode. Here, since $2^{M+1}+2 \leq X \leq 2^{M+2}-2$, the rounding function uses the least significant bit of x when rounding x to a floating point number. Then, X is rounded to a floating point number whose mantissa is even when the least significant bit of x is 1 and X is rounded to X when the bit is 0. Thus, the value of X such that the mantissa of $fl_{\mathbb{F}}(X)$ is m is as follows.

- Case: m is even.

$$X = \begin{cases} 2^{M+1} + m \times 2 - 1 \\ 2^{M+1} + m \times 2 \\ 2^{M+1} + m \times 2 + 1 \end{cases}.$$

- Case: m is odd.

$$X = 2^{M+1} + m \times 2.$$

Since $X\,(= URNG_W())$ is uniform random integer, the probability that the mantissa of $fl_{\mathbb{F}}(X)$ is even is 3 times as high as the probability that the mantissa is odd. Since a floating point number whose mantissa is even alternates with a floating point number whose mantissa is odd, this means that the generation probability of a floating point number is 3 times as high or low as that of its adjacent floating point numbers. This is unnatural from the perspective of uniform random number even if the algorithm satisfies the Formula 1 for the definition of uniformity in wide sense.

Table 2 shows which rounding mode causes each strange behavior. Figure 2 shows the probability by Thoma's method and the probability calculated by the Formula 1 where the rounding mode is Round-to-Nearest(Ties to Even) and $(E,M,W) = (4,3,5)$, and Figure 3 is enlarged view near 0. Table A·2 shows the ratio of the probability between Thoma's method and ideal as well. First, the Table 2 shows that all the rounding mode contains at least one problems. Next, we can see the strange behavior (A) and (B) in the Figure 3. The Table A·2 shows that the random number generation probability of 0 by Thoma's method is 32 ($\geq 2^{W-2} = 2^3 = 8$) times as high as the ideal probability $P_{\mathbb{F}}(0)$. Last, the Figure 2 and Figure **??** shows the strange behavior (C).

Fig. 2 Random number generation probability in $[0,1]$ by Thoma's method where the rounding mode is Round-to-Nearest(Ties to Even) and $(E, M, W) = (4, 3, 5)$.

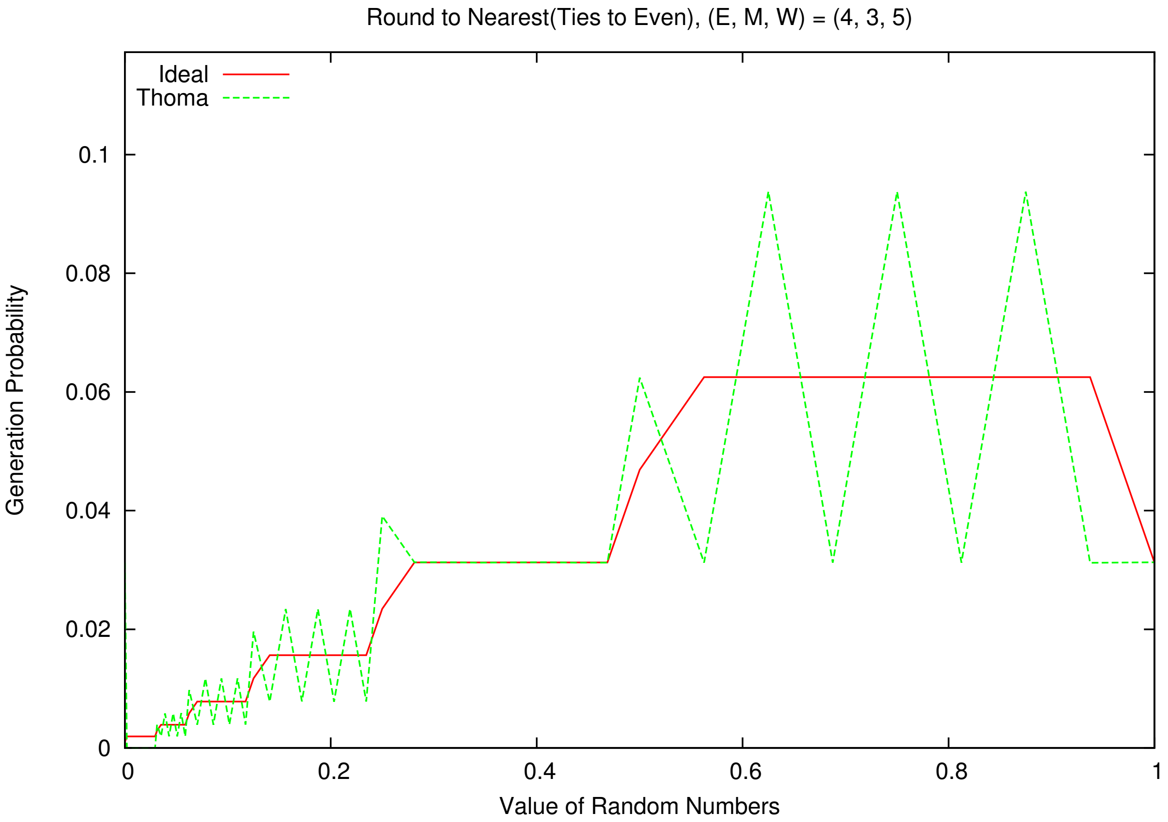
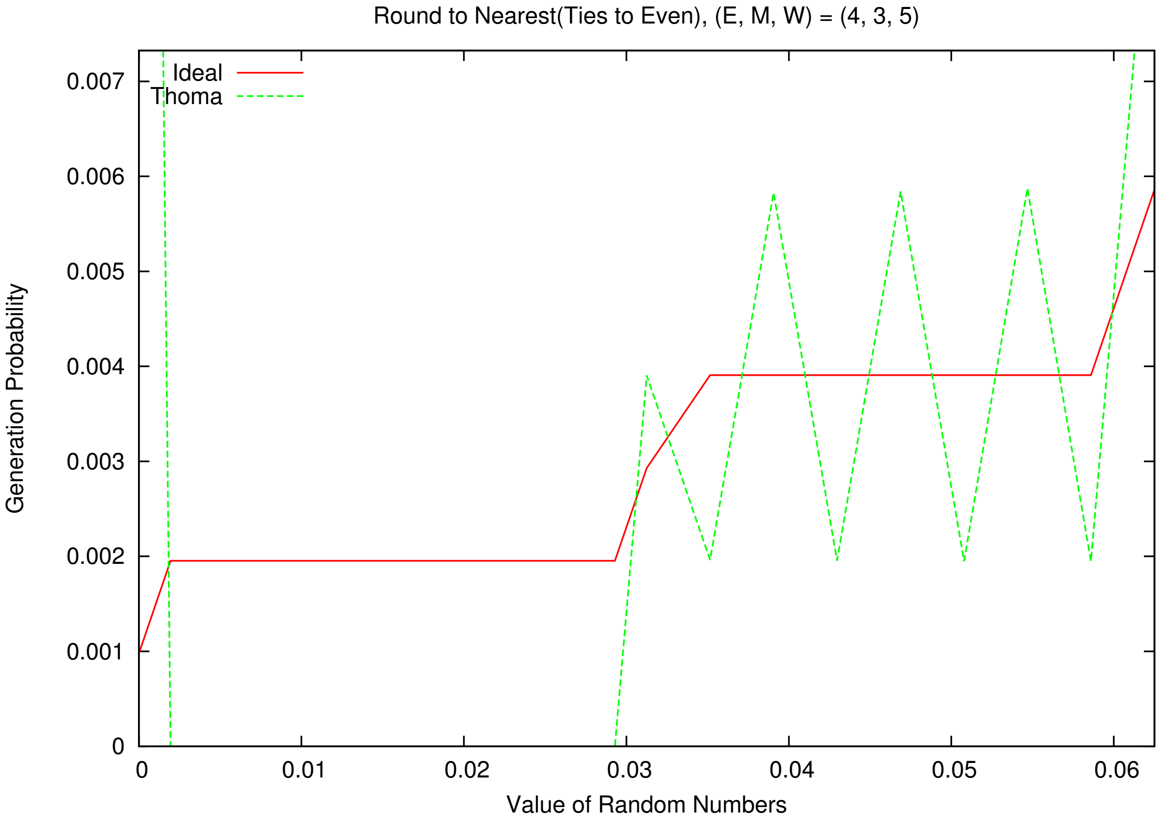


Fig. 3 Random number generation probability in $[0, 2^{-4}]$ by Thoma's method where the rounding mode is Round-to-Nearest(Ties to Even) and $(E, M, W) = (4, 3, 5)$.



4. Definition of floating point uniform random number generator

This section aims to define what a floating point number uniform random number generator is and to calculate random number generation probability of such generator before implementing on a computer.

4.1 Definition of uniform random number generator

4.1.1 Sound rounding function

Let $\mathbb{X}$ be a subset of $\mathbb{R}$. The authors define that a rounding function that rounds $r \in \mathbb{R}$ to $x \in \mathbb{X}$, $round_{\mathbb{X}} : \mathbb{R} \to \mathbb{X}$, is said to be sound if $round_{\mathbb{X}}$ satisfies all the following conditions.

Definition: Sound rounding function

Sound rounding function satisfies all the following conditions.

- Totality
  For arbitrary real number $r \in \mathbb{R}$, there exists a unique element $x \in \mathbb{R}$ that satisfies
$$round_{\mathbb{X}}(r) = x.$$
- Idempotence
  For arbitrary $x \in \mathbb{X}$,
$$round_{\mathbb{X}}(x) = x$$
  holds.
- Monotonicity
  For some real number $p, q \in \mathbb{R}$, if
$$round_{\mathbb{X}}(p) = round_{\mathbb{X}}(q)$$
  holds, then for arbitrary $t \in [0, 1] \subset \mathbb{R}$,
$$round_{\mathbb{X}}(t \times p + (1 - t) \times q) = round_{\mathbb{X}}(p)$$
  holds.

4.1.2 Definition of uniformity

Floating point uniform random number generator, $URNG_{\mathbb{F}}$, is said to be uniform in wide sense if the following condition is satisfied.

Definition: Uniform in wide sense

$URNG_{\mathbb{F}}$ is said to be uniform random number generator in wide sense if there exists a sound rounding function $round_{\mathbb{F}} : \mathbb{R} \to \mathbb{F}$ that satisfies

$$\forall x \in U_{\mathbb{F}},\ Pr\left[URNG_{\mathbb{F}}() = x\right] = Pr\left[round_{\mathbb{F}}\left(URNG_{\mathbb{R}}()\right) = x\right]. \tag{1}$$

This definition is based on the thought that when we implement a function on real number, $f_{\mathbb{R}}$, on a computer(floating point number), the implemented function, $f_{\mathbb{F}}$, should satisfies the following condition.

For arbitrary output of $f_{\mathbb{F}}$ is the same as a rounded value of $f_{\mathbb{R}}$ by a rounding function $round_{\mathbb{F}}$.

Since we can say that $URNG_{\mathbb{F}}$ is $URNG_{\mathbb{R}}$ implemented on a computer, we obtain

$$URNG_{\mathbb{F}}() = round_{\mathbb{F}}\left(URNG_{\mathbb{R}}()\right)$$

by substituting $f_{\mathbb{F}}$ with $URNG_{\mathbb{F}}$ and $f_{\mathbb{R}}$ with $URNG_{\mathbb{R}}$ in the above thought. Therefore, the probability that $URNG_{\mathbb{F}}$ generates $x \in U_{\mathbb{F}}$ satisfies the Formula 1.

Here, we define uniform in narrow sense as the case where $round_{\mathbb{F}}$ in the definition of uniform in wide sense is one of rounding modes introduced in the Section 2.3.

Definition: Uniform in narrow sense

$URNG_{\mathbb{F}}$ is said to be uniform random number generator in narrow sense if

$$\forall x \in U_{\mathbb{F}},\ Pr\left[URNG_{\mathbb{F}}\left(\right) = x\right] = Pr\left[round_{\mathbb{F}}\left(URNG_{\mathbb{R}}\left(\right)\right) = x\right]$$

holds where $round_{\mathbb{F}}$ is one of Round-to-Nearest(Ties to Even/Away from $\pm\infty$/Away from 0) or Directed-Rounding(Toward $-\infty$/Toward $+\infty$/Toward 0/Toward $\pm\infty$).

Hereafter, **"uniform" denotes "uniform in narrow sense"**.

4.2 Random number generation probability of uniform $URNG_{\mathbb{F}}$

4.2.1 Idea for calculation

By letting $U_{\mathbb{R}} = [a, b]$[*10][*11], we can transform the Formula 1 as follows.

$$\begin{aligned} Pr\left[URNG_{\mathbb{F}}\left(\right) = x\right] &= Pr\left[round_{\mathbb{F}}\left(URNG_{\mathbb{R}}\left(\right)\right) = x\right] \\ &= \int_{\{t \in U_{\mathbb{R}} | round_{\mathbb{F}}(t) = x\}} \frac{1}{b-a} dt \\ &= \frac{\sup\{t \in U_{\mathbb{R}} \mid round_{\mathbb{F}}\left(t\right) = x\}}{b-a} - \frac{\inf\{t \in U_{\mathbb{R}} \mid round_{\mathbb{F}}\left(t\right) = x\}}{b-a}. \end{aligned}$$

Thus, we can calculate the random number generation probability of a uniform $URNG_{\mathbb{F}}$ by finding the range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}\left(t\right) = x$ for each $x \in U_{\mathbb{F}}$. Hereafter, let $P_{\mathbb{F}}\left(x\right)$ denote $Pr\left[URNG_{\mathbb{F}}\left(\right) = x\right]$.

4.2.2 How to calculate the probability

Since $x \in U_{\mathbb{F}} \subset [a, b]$, we have the case where $a < x < b$ and the case where $x = a, b$.

(1) Case: $a < x < b$.

We can calculate the value of $P_{\mathbb{F}}\left(x\right)$ as follows when $a < x < b$. First, find the both side of adjacent floating point numbers to x and let x_l be the left one and x_r be the right one. Since $[x_l, x_r] \subseteq U_{\mathbb{R}}$, we can obtain the range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}\left(t\right) = x$ for each rounding modes and then calculate the value of $P_{\mathbb{F}}\left(x\right)$.

(a) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

We have 3 cases according to $round_{\mathbb{F}}$, that is, Ties to Even, Away from $\pm\infty$, and Away from 0.

(i) Case: $round_{\mathbb{F}}$ is Round-to-Nearest(Ties to Even).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}\left(t\right) = x$ is

$$\begin{cases} \frac{x_l + x}{2} < t < \frac{x + x_r}{2} & \text{Case: The mantissa of } x \text{ is odd.} \\ \frac{x_l + x}{2} \leq t \leq \frac{x + x_r}{2} & \text{Case: The mantissa of } x \text{ is even.} \end{cases}.$$

Therefore, we obtain

$$P_{\mathbb{F}}\left(x\right) = \frac{x_r - x_l}{2\left(b - a\right)}$$

in all the cases.

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest(Away from $\pm\infty$).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}\left(t\right) = x$ is

$$\begin{cases} \frac{x_l + x}{2} \leq t < \frac{x + x_r}{2} & \text{Case: } x < 0. \\ \frac{x_l + x}{2} \leq t \leq \frac{x + x_r}{2} & \text{Case: } x = 0. \\ \frac{x_l + x}{2} < t \leq \frac{x + x_r}{2} & \text{Case: } 0 < x. \end{cases}.$$

Therefore, we obtain

$$P_{\mathbb{F}}\left(x\right) = \frac{x_r - x_l}{2\left(b - a\right)}$$

in all the cases.

(iii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest(Away from 0).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}\left(t\right) = x$ is

[*10] This paper just considers the case where $a, b \in \mathbb{F}$ in order to simplify the problem.

[*11] $U_{\mathbb{R}}$ can be 4 patterns, that is, $U_{\mathbb{R}} = [a, b]$ or $[a, b)$ or $(a, b]$ or (a, b). However, the value of the right side of the Formula 1 does not change among these 4 cases. Therefore, we can consider only the case where $U_{\mathbb{R}} = [a, b]$.

$$\begin{cases} \frac{x_l+x}{2} < t \leq \frac{x+x_r}{2} & \text{Case: } x < 0. \\ \frac{x_l+x}{2} < t < \frac{x+x_r}{2} & \text{Case: } x = 0. \\ \frac{x_l+x}{2} \leq t < \frac{x+x_r}{2} & \text{Case: } 0 < x. \end{cases}.$$

Therefore, we obtain

$$P_{\mathbb{F}}(x) = \frac{x_r - x_l}{2(b-a)}$$

in all the cases.

From (i), (ii), (iii), we obtain

$$P_{\mathbb{F}}(x) = \frac{x_r - x_l}{2(b-a)}$$

when $round_{\mathbb{F}}$ is Round-to-Nearest.

(b) Case: $round_{\mathbb{F}}$ is Directed-Rounding.

We have 4 cases according to $round_{\mathbb{F}}$, that is, Toward $-\infty$, Toward $+\infty$, Toward 0, and Toward $\pm\infty$.

(i) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $-\infty$).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}(t) = x$ is

$$x \leq t < x_r.$$

Therefore, we obtain

$$P_{\mathbb{F}}(x) = \frac{x_r - x}{b-a}.$$

(ii) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}(t) = x$ is

$$x_l < t \leq x.$$

Therefore, we obtain

$$P_{\mathbb{F}}(x) = \frac{x - x_l}{b-a}.$$

(iii) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward 0).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}(t) = x$ is

$$\begin{cases} x_l < t \leq x & \text{Case: } x < 0. \\ x_l < t < x_r & \text{Case: } x = 0. \\ x \leq t < x_r & \text{Case: } 0 < x. \end{cases}.$$

Therefore, we obtain

$$P_{\mathbb{F}}(x) = \begin{cases} \frac{x-x_l}{b-a} & \text{Case: } x < 0. \\ \frac{x_r-x_l}{b-a} & \text{Case: } x = 0. \\ \frac{x_r-x}{b-a} & \text{Case: } 0 < x. \end{cases}.$$

(iv) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $\pm\infty$).

The range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}(t) = x$ is

$$\begin{cases} x \leq t < x_r & \text{Case: } x < 0. \\ t = x = 0 & \text{Case: } x = 0. \\ x_l < t \leq x & \text{Case: } 0 < x. \end{cases}.$$

Therefore, we obtain

$$P_{\mathbb{F}}(x) = \begin{cases} \frac{x_r-x}{b-a} & \text{Case: } x < 0. \\ 0 & \text{Case: } x = 0. \\ \frac{x-x_l}{b-a} & \text{Case: } 0 < x. \end{cases}.$$

(2) Case: $x = a, b$.

Next is the case where $x = a$ and the case where $x = b$. Here, if $x = a$ then we have

$$x_l < x = a = \inf U_{\mathbb{R}}$$

and if $x = b$ then we have

$$\sup U_{\mathbb{R}} = b = x < x_r.$$

Thus, we need to take the intersection of $U_{\mathbb{F}}$ and the range of $t \in U_{\mathbb{R}}$ that satisfies $round_{\mathbb{F}}(t) = x$ in the case where $a < x < b$. The concrete way is as follows.

(a) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

$$\begin{cases} P_{\mathbb{F}}(a) & = \frac{x_r - a}{2(b-a)} \\ P_{\mathbb{F}}(b) & = \frac{b - x_l}{2(b-a)} \end{cases}$$

in each case where $round_{\mathbb{F}}$ is Round-to-Nearest(Ties to Even, Away from ±∞, Away from 0).

(b) Case: $round_{\mathbb{F}}$ is Directed-Rounding.

We have 4 cases according to $round_{\mathbb{F}}$, that is, Toward $-\infty$, Toward $+\infty$, Toward 0, and Toward $\pm\infty$. Hence, the probability, $P_{\mathbb{F}}(a)$ and $P_{\mathbb{F}}(b)$ is as follows.

(i) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $-\infty$).

$$\begin{cases} P_{\mathbb{F}}(a) & = \frac{x_r - a}{b-a} \\ P_{\mathbb{F}}(b) & = 0 \end{cases}.$$

(ii) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$).

$$\begin{cases} P_{\mathbb{F}}(a) & = 0 \\ P_{\mathbb{F}}(b) & = \frac{b - x_l}{b-a} \end{cases}.$$

(iii) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward 0).

$$\begin{cases} P_{\mathbb{F}}(a) & = \begin{cases} 0 & \text{Case: } a < 0. \\ \frac{x_r - a}{b-a} & \text{Case: } 0 \leq a. \end{cases} \\ P_{\mathbb{F}}(b) & = \begin{cases} \frac{b - x_l}{b-a} & \text{Case: } b \leq 0. \\ 0 & \text{Case: } 0 < b. \end{cases} \end{cases}.$$

(iv) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $\pm\infty$).

$$\begin{cases} P_{\mathbb{F}}(a) & = \begin{cases} \frac{x_r - a}{b-a} & \text{Case: } a < 0. \\ 0 & \text{Case: } a \leq 0. \end{cases} \\ P_{\mathbb{F}}(b) & = \begin{cases} 0 & \text{Case: } b \leq 0. \\ \frac{b - x_l}{b-a} & \text{Case: } 0 < b. \end{cases} \end{cases}.$$

4.2.3 Random number generation probability for $\left[0, 2^N\right]$

Let $(a, b) = \left(0, 2^N\right)$, that is, $U_{\mathbb{R}} = \left[0, 2^N\right]$ for $N \in \mathbb{N}$ that satisfies

$$1 - \left(M + 2^{E-1} - 1\right) \leq N \in \mathbb{N} \leq 2^{E-1}.$$

This section explains the random number generation probability by uniform $URNG_{\mathbb{F}}$ in the 3 cases where $round_{\mathbb{F}}$ is Round-to-Nearest, Directed-Rounding(Toward $-\infty$), or Directed-Rounding(Toward $+\infty$)[*12].

First, consider the case where

$$1 - \left(M + 2^{E-1} - 1\right) \leq N \leq 1 - \left(2^{E-1} - 1\right).$$

(1) Case: $x = 0 = val_{\mathbb{F}}(0, 0, 0)$[*13].

The right adjacent floating point numbers to $x = 0$ is

$$\begin{aligned} x_r &= val_{\mathbb{F}}(0, 0, 1) \\ &= \left(0 + 1 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= 2^{1-\left(M+2^{E-1}-1\right)}. \end{aligned}$$

[*12] Toward 0 is equivalent to Toward $-\infty$ and Toward $\pm\infty$ is equivalent to Toward $+\infty$ because $\inf U_{\mathbb{R}} = 0 \geq 0$, that is, all the numbers in $U_{\mathbb{F}} = \left[0, 2^N\right]$ is not negative. Therefore, we do not need to consider the case where $round_{\mathbb{F}}$ is Toward 0 or Toward $\pm\infty$.

[*13] $x = 0$ is the left edge of $U_{\mathbb{R}} = \left[0, 2^N\right]$.

Therefore, the value of $P_{\mathbb{F}}(x)$ is as follows.

$$P_{\mathbb{F}}(x) = \begin{cases} \frac{x_r - x}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^{N}} & \text{Case: Toward } -\infty \\ 0 & \text{Case: Toward } +\infty \end{cases}$$
$$= \begin{cases} 2^{-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } -\infty \\ 0 & \text{Case: Toward } +\infty \end{cases} .$$

(2) Case: $x = 2^{N}$[*14].
Since $1 - \left(M + 2^{E-1} - 1\right) \leq N \leq 1 - \left(2^{E-1} - 1\right)$, we have

$$\begin{aligned} 0 \leq x &\leq 2^{N} \\ &\leq 2^{1-\left(2^{E-1}-1\right)} \\ &= \left(1 + 0 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= val_{\mathbb{F}}(0,1,0). \end{aligned}$$

This means that the interval of floating point numbers in $\left[0, 2^{N}\right]$ is $val_{\mathbb{F}}(0,0,1)$. Hence, the left adjacent floating point number to x is

$$\begin{aligned} x_l &= x - val_{\mathbb{F}}(0,0,1) \\ &= x - \left(0 + 1 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= x - 2^{1-\left(M+2^{E-1}-1\right)}. \end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}(x)$ is as follows.

$$P_{\mathbb{F}}(x) = \begin{cases} \frac{x - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ 0 & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^{N}} & \text{Case: Toward } +\infty \end{cases}$$
$$= \begin{cases} 2^{-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 0 & \text{Case: Toward } -\infty \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } +\infty \end{cases} .$$

(3) Case: $0 < x < 2^{N}$.
By the same way as (2), we have the fact that the interval of floating point numbers in $\left[0, 2^{N}\right]$ is $val_{\mathbb{F}}(0,0,1)$. Hence, the left adjacent floating point number to x and the right adjacent one to x is

$$\begin{aligned} x_l &= x - val_{\mathbb{F}}(0,0,1) \\ &= x - \left(0 + 1 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= x - 2^{1-\left(M+2^{E-1}-1\right)} \\ x_r &= x + val_{\mathbb{F}}(0,0,1) \\ &= x + 2^{1-\left(M+2^{E-1}-1\right)} \end{aligned}$$

respectively. Therefore, the value of $P_{\mathbb{F}}(x)$ is as follows.

$$P_{\mathbb{F}}(x) = \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^{N}} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^{N}} & \text{Case: Toward } +\infty \end{cases}$$
$$= \begin{cases} 2^{1-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{1-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } -\infty \\ 2^{1-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } +\infty \end{cases} .$$

Next, consider the case where

[*14] $x = 2^{N}$ is the right edge of $U_{\mathbb{R}} = \left[0, 2^{N}\right]$.

$$2-\left(2^{E-1}-1\right)\le N\le 2^{E-1}.$$

(1) Case: $x=0=val_{\mathbb{F}}\left(0,0,0\right)$

We can consider in the same way as the case where

$$1-\left(M+2^{E-1}-1\right)\le N\le 1-\left(2^{E-1}-1\right).$$

(2) Case: $x=2^N=val_{\mathbb{F}}\left(0,N+2^{E-1}-1,0\right)$.

The left adjacent floating point numbers to x is

$$\begin{aligned}x_l&=val_{\mathbb{F}}\left(0,N+2^{E-1}-2,2^M-1\right)\\&=\left(1+\left(2^M-1\right)\times 2^{-M}\right)\times 2^{N-1}\\&=\left(2-2^{-M}\right)\times 2^{N-1}\\&=\left(1-2^{-(M+1)}\right)\times 2^N.\end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}\left(x\right)$ is as follows.

$$\begin{aligned}P_{\mathbb{F}}\left(x\right)&=\begin{cases}\frac{x-x_l}{2^{N+1}} & \text{Case: Round-to-Nearest}\\ 0 & \text{Case: Toward }-\infty\\ \frac{x-x_l}{2^N} & \text{Case: Toward }+\infty\end{cases}\\&=\begin{cases}2^{-(M+2)} & \text{Case: Round-to-Nearest}\\ 0 & \text{Case: Toward }-\infty\\ 2^{-(M+1)} & \text{Case: Toward }+\infty\end{cases}.\end{aligned}$$

(3) Case: $x=val_{\mathbb{F}}\left(0,e,0\right)$.

Since we have already calculated the value of $P_{\mathbb{F}}\left(x\right)$ in the case where $e=0,N+2^{E-1}-1$ in (1) and (2), we consider only the case where $1\le e\le N+2^{E-1}-2$.

(3-1) Case: $e=1$.

In this case, we have

$$\begin{aligned}x&=val_{\mathbb{F}}\left(0,1,0\right)\\&=\left(1+0\times 2^{-M}\right)\times 2^{1-\left(2^{E-1}-1\right)}\\&=2^M\times 2^{-\left(M+2^{E-1}-2\right)}.\end{aligned}$$

Hence, the left adjacent floating point number to x and the right adjacent one to x is

$$\begin{aligned}x_l&=val_{\mathbb{F}}\left(0,0,2^M-1\right)\\&=\left(0+\left(2^M-1\right)\times 2^{-M}\right)\times 2^{1-\left(2^{E-1}-1\right)}\\&=\left(2^M-1\right)\times 2^{-\left(M+2^{E-1}-2\right)}\\x_r&=val_{\mathbb{F}}\left(0,1,1\right)\\&=\left(1+1\times 2^{-M}\right)\times 2^{1-\left(2^{E-1}-1\right)}\\&=\left(2^M+1\right)\times 2^{-\left(M+2^{E-1}-2\right)}.\end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}\left(x\right)$ is as follows.

$$\begin{aligned}P_{\mathbb{F}}\left(x\right)&=\begin{cases}\frac{x_r-x_l}{2^{N+1}} & \text{Case: Round-to-Nearest}\\ \frac{x_r-x}{2^N} & \text{Case: Toward }-\infty\\ \frac{x-x_l}{2^N} & \text{Case: Toward }+\infty\end{cases}\\&=\begin{cases}2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Round-to-Nearest}\\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward }-\infty\\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward }+\infty\end{cases}.\end{aligned}$$

(3-2) Case: $2\le e\le N+2^{E-1}-2$.

In this case, we have

$$\begin{aligned} x &= val_{\mathbb{F}}\left(0, e, 0\right) \\ &= 2^{e-\left(2^{E-1}-1\right)} \\ &= 2^{M+2} \times 2^{e-\left(M+2^{E-1}+1\right)}. \end{aligned}$$

Hence, the left adjacent floating point number to x and the right adjacent one to x is

$$\begin{aligned} x_l &= val_{\mathbb{F}}\left(0, e-1, 2^M - 1\right) \\ &= \left(1 + \left(2^M - 1\right) \times 2^{-M}\right) \times 2^{(e-1)-\left(2^{E-1}-1\right)} \\ &= \left(2 - 2^{-M}\right) \times 2^{(e-1)-\left(2^{E-1}-1\right)} \\ &= \left(2^{M+2} - 2\right) \times 2^{e-\left(M+2^{E-1}+1\right)} \\ x_r &= val_{\mathbb{F}}\left(0, e, 1\right) \\ &= \left(1 + 1 \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ &= \left(2^{M+2} + 4\right) \times 2^{e-\left(M+2^{E-1}+1\right)}. \end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}\left(x\right)$ is as follows.

$$\begin{aligned} P_{\mathbb{F}}\left(x\right) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 3 \times 2^{e-\left(N+M+2^{E-1}+1\right)} & \text{Case: Round-to-Nearest} \\ 4 \times 2^{e-\left(N+M+2^{E-1}+1\right)} & \text{Case: Toward } -\infty \\ 2 \times 2^{e-\left(N+M+2^{E-1}+1\right)} & \text{Case: Toward } +\infty \end{cases}. \end{aligned}$$

(4) Case: $x = val_{\mathbb{F}}\left(0, 0, m\right)$.

Since we have already calculated the value of $P_{\mathbb{F}}\left(x\right)$ in the case where $m = 0$ in (1), we consider only the case where $1 \leq m \leq N + 2^M - 1$.

(4-1) Case: $1 \leq m \leq 2^M - 2$

In this case, we have

$$\begin{aligned} x &= val_{\mathbb{F}}\left(0, 0, m\right) \\ &= \left(m \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= m \times 2^{-\left(M+2^{E-1}-2\right)}. \end{aligned}$$

Hence, the left adjacent floating point number to x and the right adjacent one to x is

$$\begin{aligned} x_l &= val_{\mathbb{F}}\left(0, 0, m-1\right) \\ &= \left((m-1) \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= (m-1) \times 2^{-\left(M+2^{E-1}-2\right)} \\ x_r &= val_{\mathbb{F}}\left(0, 0, m+1\right) \\ &= \left((m+1) \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= (m+1) \times 2^{-\left(M+2^{E-1}-2\right)}. \end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}\left(x\right)$ is as follows.

$$\begin{aligned} P_{\mathbb{F}}\left(x\right) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Round-to-Nearest} \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } -\infty \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } +\infty \end{cases}. \end{aligned}$$

(4-2) Case: $m = 2^M - 1$.
In this case, we have

$$\begin{aligned} x &= val_{\mathbb{F}}\left(0, 0, 2^M - 1\right) \\ &= \left(0 + \left(2^M - 1\right) \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= \left(2^M - 1\right) \times 2^{-\left(M+2^{E-1}-2\right)}. \end{aligned}$$

Hence, the left adjacent floating point number to x and the right adjacent one to x is

$$\begin{aligned} x_l &= val_{\mathbb{F}}\left(0, 0, 2^M - 2\right) \\ &= \left(0 + \left(2^M - 2\right) \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= \left(2^M - 2\right) \times 2^{-\left(M+2^{E-1}-2\right)} \\ x_r &= val_{\mathbb{F}}\left(0, 1, 0\right) \\ &= \left(1 + 0 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= \left(2^M\right) \times 2^{-\left(M+2^{E-1}-2\right)}. \end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}(x)$ is as follows.

$$\begin{aligned} P_{\mathbb{F}}(x) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Round-to-Nearest} \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } -\infty \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } +\infty \end{cases} . \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} P_{\mathbb{F}}(x) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Round-to-Nearest} \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } -\infty \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } +\infty \end{cases} \end{aligned}$$

in all the cases.

(5) Case: $x = val_{\mathbb{F}}(0, e, m)$.
Since we have already calculated the value of $P_{\mathbb{F}}(x)$ in the case where $e = 0$ in (4), we consider only the case where $1 \leq e \leq N + 2^{E-1} - 2$[*15]. Additionally, since we have finished the case where $m = 0$ in (3), we consider only the case where $1 \leq m \leq 2^M - 1$.

(5-1) Case: $1 \leq m \leq 2^M - 2$.
In this case, we have

$$\begin{aligned} x &= val_{\mathbb{F}}(0, e, m) \\ &= \left(1 + m \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ &= \left(2^M + m\right) \times 2^{e-\left(M+2^{E-1}-1\right)}. \end{aligned}$$

Hence, the left adjacent floating point number to x and the right adjacent one to x is

[*15] If $e = N + 2^{E-1} - 1$, then we have $m = 0$ because $x \leq \sup U_{\mathbb{F}} = 2^N = val_{\mathbb{F}}\left(0, N + 2^{E-1} - 1, 0\right)$. This case is the same as (2).

$$\begin{aligned} x_l &= val_{\mathbb{F}}\left(0, e, m-1\right) \\ &= \left(1 + (m-1) \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ &= \left(2^M + m - 1\right) \times 2^{e-\left(M+2^{E-1}-1\right)} \\ x_r &= val_{\mathbb{F}}\left(0, e, m+1\right) \\ &= \left(1 + (m+1) \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ &= \left(2^M + m + 1\right) \times 2^{e-\left(M+2^{E-1}-1\right)}. \end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}(x)$ is as follows.

$$\begin{aligned} P_{\mathbb{F}}(x) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } -\infty \\ 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } +\infty \end{cases} . \end{aligned}$$

(5-2) Case: $m = 2^M - 1$.
In this case, we have

$$\begin{aligned} x &= val_{\mathbb{F}}\left(0, e, 2^M - 1\right) \\ &= \left(1 + \left(2^M - 1\right) \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ &= \left(2^{M+1} - 1\right) \times 2^{e-\left(M+2^{E-1}-1\right)}. \end{aligned}$$

Hence, the left adjacent floating point number to x and the right adjacent one to x is

$$\begin{aligned} x_l &= val_{\mathbb{F}}\left(0, e, 2^M - 2\right) \\ &= \left(1 + \left(2^M - 2\right) \times 2^{-M}\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ &= \left(2^{M+1} - 2\right) \times 2^{e-\left(2^{E-1}-1\right)} \\ x_r &= val_{\mathbb{F}}\left(0, e+1, 0\right) \\ &= 1 \times 2^{(e+1)-\left(2^{E-1}-1\right)} \\ &= \left(2^{M+1}\right) \times 2^{e-\left(M+2^{E-1}-1\right)}. \end{aligned}$$

Therefore, the value of $P_{\mathbb{F}}(x)$ is as follows.

$$\begin{aligned} P_{\mathbb{F}}(x) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } -\infty \\ 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } +\infty \end{cases} . \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} P_{\mathbb{F}}(x) &= \begin{cases} \frac{x_r - x_l}{2^{N+1}} & \text{Case: Round-to-Nearest} \\ \frac{x_r - x}{2^N} & \text{Case: Toward } -\infty \\ \frac{x - x_l}{2^N} & \text{Case: Toward } +\infty \end{cases} \\ &= \begin{cases} 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } -\infty \\ 2^{e-\left(N+M+2^{E-1}-1\right)} & \text{Case: Toward } +\infty \end{cases} \end{aligned}$$

in all the cases.

In summary, we obtain the following result.

(1) Case: $x = val_{\mathbb{F}}(0,0,0) = 0$.

$$P_{\mathbb{F}}(x) = \begin{cases} 2^{-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } -\infty \\ 0 & \text{Case: Toward } +\infty \end{cases}.$$

(2) Case: $val_{\mathbb{F}}(0,0,1) \le x < val_{\mathbb{F}}(0,2,0)$.

$$P_{\mathbb{F}}(x) = 2^{1-\left(N+M+2^{E-1}-1\right)}.$$

(3) Case: $val_{\mathbb{F}}(0,2,0) \le x < val_{\mathbb{F}}\left(0, N+2^{E-1}-1, 0\right)$.

In this case, we have 2 sub cases shown as follows. Here, $2 \le e \le N + 2^{E-1} - 2$ in both cases.

(3-1) Case: $x = val_{\mathbb{F}}(0,e,0)$.

$$P_{\mathbb{F}}(x) = \alpha \times 2^{e-\left(N+M+2^{E-1}-1\right)}$$

where

$$\alpha = \begin{cases} \frac{3}{4} & \text{Case: Round-to-Nearest} \\ 1 & \text{Case: Toward } -\infty \\ \frac{1}{2} & \text{Case: Toward } +\infty \end{cases}.$$

(3-2) Case: $val_{\mathbb{F}}(0,e,1) \le x < val_{\mathbb{F}}(0,e+1,0)$.

$$P_{\mathbb{F}}(x) = 2^{e-\left(N+M+2^{E-1}-1\right)}.$$

(4) Case: $x = 2^N$.

- Case: $1 - \left(M + 2^{E-1} - 1\right) \le N \le 1 - \left(2^{E-1} - 1\right)$.

$$P_{\mathbb{F}}(x) = \begin{cases} 2^{-\left(N+M+2^{E-1}-1\right)} & \text{Case: Round-to-Nearest} \\ 0 & \text{Case: Toward } -\infty \\ 2^{-\left(N+M+2^{E-1}-2\right)} & \text{Case: Toward } +\infty \end{cases}.$$

- Case: $2 - \left(2^{E-1} - 1\right) \le N \le 2^{E-1}$.

$$P_{\mathbb{F}}(x) = \begin{cases} 2^{-(M+2)} & \text{Case: Round-to-Nearest} \\ 0 & \text{Case: Toward } -\infty \\ 2^{-(M+1)} & \text{Case: Toward } +\infty \end{cases}.$$

5. Modified algorithm for Thoma's method

This section aims to modify the problems in Thoma's algorithm, which is a floating point uniform random number generator, and prove that the modified method is uniform in narrow sense defined in the Section 4 and make some experiments in order to show that the modified method solves the problems.

5.1 Modified algorithm

The modified algorithm is a uniform $URNG_{\mathbb{F}}$ in narrow sense for $U_{\mathbb{R}} = \left[0, 2^N\right]$ [*16]. The main idea of the algorithm is as follows. Let $u \in \mathbb{R}$ be a uniform random real number on $\left[0, 2^N\right]$ generated by $URNG_{\mathbb{R}}$. First, find the maximal floating point number $val_{\mathbb{F}}(0,e,m)$ that satisfies

$$val_{\mathbb{F}}(0,e,m) \le u \le val_{\mathbb{F}}(0,e,m+1)$$

[*17][*18]. Next, simulate rounding operation based on $round_{\mathbb{F}}$ for this u, that is, simulate $round_{\mathbb{F}}(u)$.

In the concrete, the algorithm can generates e as a geometric random integer and m as a uniform random integer because u is distributed on $\left[0, 2^N\right]$ uniformly. And then, the algorithm judges whether u, which is distributed on $\left[val_{\mathbb{F}}(0,e,m), val_{\mathbb{F}}(0,e,m+1)\right]$ uniformly, is rounded to $val_{\mathbb{F}}(0,e,m)$ or rounded to $val_{\mathbb{F}}(0,e,m+1)$ according to $round_{\mathbb{F}}$.

[*16] $N \in \mathbb{N}$ must satisfy $1 - \left(M + 2^{E-1} - 1\right) \le N \le 2^{E-1}$.

[*17] If $m = 2^M - 1$, then replace $val_{\mathbb{F}}(0,e,m+1)$ with $val_{\mathbb{F}}(0,e+1,0)$ in the inequality.

[*18] If $u \in \mathbb{F}$, then we can take 2 different floating point numbers that satisfy the inequality. However, since u is a uniform random real number, the probability that u is equal to a specific one value is 0. Therefore, we do not need to consider such a case.

5.1.1 Pseudocode

The pseudocode of the modified algorithm is as follows.

```
00: Set $val_{\mathbb{F}}(0, e_{max}, 0)$ as the maximal value of uniform random numbers.
      $e_{max} = N + 2^{E-1} - 1$
10: Generate a fixed point uniform random number if $e_{max} \leq 1$.
      if ($e_{max} \leq 1$) {
          $e = 0$
          if ($e_{max} \leq 1 - M$) {
              $m = 0$
          } else {
              $m = URNG_{M+e_{max}-1}()$
          }
          goto 40
      }
20: Find $e$.
      $n = 1$
      while ($n < e_{max}$) {
          if ($URNG_1() = 1$) {
              break
          } else {
              $n = n + 1$
          }
      }
      $e = e_{max} - n$
30: Find $m$.
      $m = URNG_M()$
40: Branch according to $round_{\mathbb{F}}$.
  • Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $-\infty$).
    goto 41
  • Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$).
    goto 42
  • Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward 0).
    goto 41
  • Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $\pm\infty$).
    goto 42
  • Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
    goto 43
41: Simulate Directed-Rounding(Toward $-\infty$ and Toward 0).
      goto 50
42: Simulate Directed-Rounding(Toward $+\infty$ and Toward $\pm\infty$).
      if ($m = 2^M - 1$) {
          $m = 0$
          $e = e + 1$
      } else {
          $m = m + 1$
      }
      goto 50
43: Simulate Round-to-Nearest.
      $b = URNG_1()$
      if ($b = 0$) {
          goto 41
      } else {
          goto 42
      }
50: Convert $e$ and $m$ to a floating point number.
      return $val_{\mathbb{F}}(0, e, m)$
```

5.1.2 Explanation for the pseudocode

The meaning of the pseudocode is as follows.

00: Set $val_{\mathbb{F}}(0, e_{max}, 0)$ as the maximal value of uniform random numbers.
Set the maximal value of uniform random numbers by $2^{e_{max}-\left(2^{E-1}-1\right)}$. In the algorithm, e_{max} must satisfy $1-M \leq e_{max}$. If $e_{max} < -M$, then the algorithm behaves in the same way as the case where $e_{max} = 1 - M$.

10: Generate a fixed point uniform random number if $e_{max} \leq 1$.
The algorithm is equivalent to fixed point uniform random number generator if $e_{max} \leq 1$.

20: Find e.
Generate a geometric random integer for e by the Bernoulli trial by using 1-bit uniform random integer. In the pseudocode, n denotes the number of tries until the first non-zero bit is generated.

30: Find m.
Generate an M-bit uniform random integer for m.

40: Branch according to $round_{\mathbb{F}}$.
Judge whether u, which is distributed on $[val_{\mathbb{F}}(0, e, m), val_{\mathbb{F}}(0, e, m+1)]$ uniformly, is rounded to the left edge of the range or the right one. Here, Toward 0 is equivalent to Toward $-\infty$ and Toward $\pm\infty$ is equivalent to Toward $+\infty$ because $U_{\mathbb{R}} = \left[0, 2^N\right]$.

41: Simulate Directed-Rounding(Toward $-\infty$ and Toward 0).
Since u is always rounded to the left edge in this case, select the left one.

42: Simulate Directed-Rounding(Toward $+\infty$ and Toward $\pm\infty$).
Since u is always rounded to the right edge in this case, select the right one.

43: Simulate Round-to-Nearest.
In this case, the probability that a uniform random real number distributed on $[val_{\mathbb{F}}(0, e, m), val_{\mathbb{F}}(0, e, m+1)]$ is rounded to the left edge $val_{\mathbb{F}}(0, e, m)$ is the same as the probability that the random number is rounded to the right edge $val_{\mathbb{F}}(0, e, m+1)$, that is, both are $\frac{1}{2}$. Therefore, generate a 1-bit uniform random integer and then select the left one if the bit is 0 or select the right one if the bit is 1.

5.2 Proof for correctness

This section proves that the random number generation probability of the modified algorithm satisfies the Formula 1. First, calculate the probability that the algorithm outputs $x \in \mathbb{F}$, $P(x)$, for each floating point number. Next, compare $P(x)$ with $P_{\mathbb{F}}(x)$[*19], which is calculated in the Section 4.2.3, and confirm that $P(x) = P_{\mathbb{F}}(x)$ holds. Here, Since toward 0 is equivalent to Toward $-\infty$ and Toward $\pm\infty$ is equivalent to Toward $+\infty$ because $U_{\mathbb{R}} = \left[0, 2^N\right]$, we need to prove only the case where $round_{\mathbb{F}}$ is Round-to-Nearest, Directed-Rounding(Toward $-\infty$), or Directed-Rounding(Toward $+\infty$). In the proof, let

$$Pr\left[\text{"constraint of variables" in "line number in the pseudocode"}\right]$$

be the probability that the constraint is satisfied at the end of the line in the pseudocode.

First, consider the case where

$$1 - \left(M + 2^{E-1} - 1\right) \leq N \leq 1 - \left(2^{E-1} - 1\right).$$

That is,

$$1 - M \leq e_{max} \leq 1.$$

Now, we have 3 cases according to $round_{\mathbb{F}}$, that is, the case where $round_{\mathbb{F}}$ is Directed-Rounding (Toward $-\infty$), the case where $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$), and the case where $round_{\mathbb{F}}$ is Round-to-Nearest.

(i) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $-\infty$).

(1) Case: $val_{\mathbb{F}}(0, 0, 0) \leq x < 2^N$.

In this case, we have

$$\begin{aligned} 2^N &= 2^{e_{max}-\left(2^{E-1}-1\right)} \\ &= \left(0 + 2^{M+e_{max}-1} \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)}. \end{aligned}$$

Hence, we can express x by

$$x = val_{\mathbb{F}}\left(0, 0, m'\right)$$

[*19] Note: $P_{\mathbb{F}}(x)$ is the random number generation probability of uniform $URNG_{\mathbb{F}}$ in narrow sense.

for an integer m' that satisfies $0 \le m' < 2^{M+e_{max}-1}$. Thus we obtain

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = 0, m = m' \text{ in } 50\right] \\
&= Pr\left[e = 0, m = m' \text{ in } 41\right] \\
&= Pr\left[e = 0, m = m' \text{ in } 40\right] \\
&= Pr\left[e = 0, m = m' \text{ in } 10\right] \\
&= Pr\left[m = m' \text{ in } 10\right] \\
&= 2^{-\left(M+e_{max}-1\right)} \\
&= 2^{1-\left(M+e_{max}\right)} \\
&= 2^{1-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}\left(x\right) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P\left(x\right) = P_{\mathbb{F}}\left(x\right)$ holds.

(2) Case: $x = 2^N$.

Consider the case where $e_{max} = 0$ and the other case, that is, the case where $N = 1 - \left(2^{E-1} - 1\right)$ and the case where $N < 1 - \left(2^{E-1} - 1\right)$.

- Case: $N = 1 - \left(2^{E-1} - 1\right)$.
  In this case, we can express x as follows.

$$\begin{aligned}
x &= 2^N \\
&= \left(1 + 0 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\
&= val_{\mathbb{F}}\left(0, 1, 0\right).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = 1, m = 0 \text{ in } 50\right] \\
&= Pr\left[e = 1, m = 0 \text{ in } 41\right] \\
&= Pr\left[e = 1, m = 0 \text{ in } 40\right] \\
&= Pr\left[e = 1, m = 0 \text{ in } 10\right] \\
&= 0.
\end{aligned}$$

- Case: $N < 1 - \left(2^{E-1} - 1\right)$.
  In this case, we can express x as follows.

$$\begin{aligned}
x &= 2^N \\
&= 2^{e_{max}-\left(2^{E-1}-1\right)} \\
&= 2^{e_{max}-1} \times 2^{1-\left(2^{E-1}-1\right)} \\
&= \left(0 + 2^{M+e_{max}-1} \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\
&= val_{\mathbb{F}}\left(0, 0, 2^{M+e_{max}-1}\right).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 50\right] \\
&= Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 41\right] \\
&= Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 40\right] \\
&= Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 10\right] \\
&= 0.
\end{aligned}$$

Thus, we obtain

$$P(x) = 0.$$

in both cases. Besides, we have

$$P_{\mathbb{F}}(x) = 0$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(ii) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$).

(1) Case: $x = 0$.

In this case, we can express x by $x = val_{\mathbb{F}}(0,0,0)$. Hence, we obtain

$$\begin{aligned} P(x) &= Pr\left[e = 0, m = 0 \text{ in } 50\right] \\ &= Pr\left[e = 0, m = 0 \text{ in } 42\right] \\ &= Pr\left[e = -1, m = 2^M - 1 \text{ in } 40\right] \\ &= Pr\left[e = -1, m = 2^M - 1 \text{ in } 10\right] \\ &= 0. \end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 0$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(2) Case: $val_{\mathbb{F}}(0,0,1) \leq x < 2^N$.

In this case, we have

$$\begin{aligned} 2^N &= 2^{e_{max} - \left(2^{E-1} - 1\right)} \\ &= \left(0 + 2^{M + e_{max} - 1} \times 2^{-M}\right) \times 2^{1 - \left(2^{E-1} - 1\right)}. \end{aligned}$$

Hence, we can express x by

$$x = val_{\mathbb{F}}\left(0, 0, m'\right)$$

for an integer m' that satisfies $0 < m' < 2^{M + e_{max} - 1}$. Thus, we obtain

$$\begin{aligned} P(x) &= Pr\left[e = 0, m = m' \text{ in } 50\right] \\ &= Pr\left[e = 0, m = m' \text{ in } 42\right] \\ &= Pr\left[e = 0, m = m' - 1 \text{ in } 40\right] \\ &= Pr\left[e = 0, m = m' - 1 \text{ in } 10\right] \\ &= 2^{-(M + e_{max} - 1)} \\ &= 2^{1 - (M + e_{max})} \\ &= 2^{1 - \left(N + M + 2^{E-1} - 1\right)}. \end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 2^{1 - \left(N + M + 2^{E-1} - 1\right)}$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(3) Case: $x = 2^N$.

Consider the case where $e_{max} = 0$ and the other case, that is, the case where $N = 1 - \left(2^{E-1} - 1\right)$ and the case where $N < 1 - \left(2^{E-1} - 1\right)$.

- Case: $N = 1 - \left(2^{E-1} - 1\right)$.

  In this case, we can express x as follows.

$$\begin{aligned} x &= 2^N \\ &= \left(1 + 0 \times 2^{-M}\right) \times 2^{1 - \left(2^{E-1} - 1\right)} \\ &= val_{\mathbb{F}}(0, 1, 0). \end{aligned}$$

Hence, we have

$$\begin{aligned}P(x) &= Pr\left[e=1, m=0 \text{ in } 50\right]\\ &= Pr\left[e=1, m=0 \text{ in } 42\right]\\ &= Pr\left[e=0, m=2^M-1 \text{ in } 40\right]\\ &= Pr\left[e=0, m=2^M-1 \text{ in } 10\right]\\ &= 2^{-(M+e_{max}-1)}\\ &= 2^{1-(M+e_{max})}\\ &= 2^{1-\left(N+M+2^{E-1}-1\right)}.\end{aligned}$$

- Case: $N < 1-\left(2^{E-1}-1\right)$.
  In this case, we can express x as follows.

$$\begin{aligned}x &= 2^N\\ &= 2^{e_{max}-\left(2^{E-1}-1\right)}\\ &= 2^{e_{max}-1} \times 2^{1-\left(2^{E-1}-1\right)}\\ &= \left(0+2^{M+e_{max}-1} \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)}\\ &= val_{\mathbb{F}}\left(0, 0, 2^{M+e_{max}-1}\right).\end{aligned}$$

  Hence, we have

$$\begin{aligned}P(x) &= Pr\left[e=0, m=2^{M+e_{max}-1} \text{ in } 50\right]\\ &= Pr\left[e=0, m=2^{M+e_{max}-1} \text{ in } 42\right]\\ &= Pr\left[e=0, m=2^{M+e_{max}-1}-1 \text{ in } 40\right]\\ &= Pr\left[e=0, m=2^{M+e_{max}-1}-1 \text{ in } 10\right]\\ &= 2^{-(M+e_{max}-1)}\\ &= 2^{1-(M+e_{max})}\\ &= 2^{1-\left(N+M+2^{E-1}-1\right)}.\end{aligned}$$

Thus, we obtain

$$P(x) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

in both cases. Besides, we have

$$P_{\mathbb{F}}(x) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(iii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

(1) Case: $x = 0$.

In this case, we can express x by $x = val_{\mathbb{F}}(0,0,0)$. Hence, we obtain

$$\begin{aligned}P(x) &= Pr\left[e=0, m=0 \text{ in } 50\right]\\ &= Pr\left[e=0, m=0 \text{ in } 41\right] + Pr\left[e=0, m=0 \text{ in } 42\right]\\ &= Pr\left[b=0, e=0, m=0 \text{ in } 43\right] + Pr\left[b=1, e=-1, m=2^M-1 \text{ in } 43\right]\\ &= \frac{1}{2} \times Pr\left[e=0, m=0 \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e=-1, m=2^M-1 \text{ in } 40\right]\\ &= \frac{1}{2} \times Pr\left[e=0, m=0 \text{ in } 10\right] + \frac{1}{2} \times Pr\left[e=-1, m=2^M-1 \text{ in } 10\right]\\ &= \frac{1}{2} \times 2^{-(M+e_{max}-1)} + \frac{1}{2} \times 0\\ &= \frac{1}{2} \times 2^{1-(M+e_{max})}\\ &= \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}.\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}$$

in the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(2) Case: $val_{\mathbb{F}}(0,0,1) \leq x < 2^N$.

In this case, we have

$$\begin{aligned} 2^N &= 2^{e_{max}-\left(2^{E-1}-1\right)} \\ &= \left(0 + 2^{M+e_{max}-1} \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)}. \end{aligned}$$

Hence, we can express x by

$$x = val_{\mathbb{F}}\left(0,0,m'\right)$$

for an integer m' that satisfies $0 < m' < 2^{M+e_{max}-1}$. Thus, we obtain

$$\begin{aligned} P(x) &= Pr\left[e=0, m=m' \text{ in } 50\right] \\ &= Pr\left[e=0, m=m' \text{ in } 41\right] + Pr\left[e=0, m=m' \text{ in } 42\right] \\ &= Pr\left[b=0, e=0, m=m' \text{ in } 43\right] + Pr\left[b=1, e=0, m=m'-1 \text{ in } 43\right] \\ &= \frac{1}{2} \times Pr\left[e=0, m=m' \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e=0, m=m'-1 \text{ in } 40\right] \\ &= \frac{1}{2} \times Pr\left[e=0, m=m' \text{ in } 10\right] + \frac{1}{2} \times Pr\left[e=0, m=m'-1 \text{ in } 10\right] \\ &= \frac{1}{2} \times 2^{-(M+e_{max}-1)} + \frac{1}{2} \times 2^{-(M+e_{max}-1)} \\ &= 2^{-(M+e_{max}-1)} \\ &= 2^{1-(M+e_{max})} \\ &= 2^{1-\left(N+M+2^{E-1}-1\right)}. \end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(3) Case: $x = 2^N$.

Consider the case where $e_{max} = 0$ and the other case, that is, the case where $N = 1 - \left(2^{E-1}-1\right)$ and the case where $N < 1 - \left(2^{E-1}-1\right)$.

- Case: $N = 1 - \left(2^{E-1}-1\right)$.

  In this case, we can express x as follows.

$$\begin{aligned} x &= 2^N \\ &= \left(1 + 0 \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\ &= val_{\mathbb{F}}(0,1,0). \end{aligned}$$

  Hence, we have

$$\begin{aligned} P(x) &= Pr\left[e=1, m=0 \text{ in } 50\right] \\ &= Pr\left[e=1, m=0 \text{ in } 41\right] + Pr\left[e=1, m=0 \text{ in } 42\right] \\ &= Pr\left[b=0, e=1, m=0 \text{ in } 43\right] + Pr\left[b=1, e=0, m=2^M-1 \text{ in } 43\right] \\ &= \frac{1}{2} \times Pr\left[e=1, m=0 \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e=0, m=2^M-1 \text{ in } 40\right] \\ &= \frac{1}{2} \times Pr\left[e=1, m=0 \text{ in } 10\right] + \frac{1}{2} \times Pr\left[e=0, m=2^M-1 \text{ in } 10\right] \\ &= \frac{1}{2} \times 0 + \frac{1}{2} \times 2^{-(M+e_{max}-1)} \\ &= \frac{1}{2} \times 2^{1-(M+e_{max})} \\ &= \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}. \end{aligned}$$

- Case: $N < 1 - \left(2^{E-1} - 1\right)$.
  In this case, we can express x as follows.

$$\begin{aligned}
x &= 2^{N} \\
&= 2^{e_{max} - \left(2^{E-1}-1\right)} \\
&= 2^{e_{max}-1} \times 2^{1-\left(2^{E-1}-1\right)} \\
&= \left(0 + 2^{M+e_{max}-1} \times 2^{-M}\right) \times 2^{1-\left(2^{E-1}-1\right)} \\
&= val_{\mathbb{F}}\left(0, 0, 2^{M+e_{max}-1}\right).
\end{aligned}$$

  Hence, we have

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 50\right] \\
&= Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 41\right] + Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 42\right] \\
&= Pr\left[b = 0, e = 0, m = 2^{M+e_{max}-1} \text{ in } 43\right] \\
&\quad + Pr\left[b = 1, e = 0, m = 2^{M+e_{max}-1} - 1 \text{ in } 43\right] \\
&= \frac{1}{2} \times Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 40\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = 0, m = 2^{M+e_{max}-1} - 1 \text{ in } 40\right] \\
&= \frac{1}{2} \times Pr\left[e = 0, m = 2^{M+e_{max}-1} \text{ in } 10\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = 0, m = 2^{M+e_{max}-1} - 1 \text{ in } 10\right] \\
&= \frac{1}{2} \times 2^{-(M+e_{max}-1)} + \frac{1}{2} \times 0 \\
&= \frac{1}{2} \times 2^{1-(M+e_{max})} \\
&= \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}$$

Thus, we obtain

$$P\left(x\right) = \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}$$

in both cases. Besides, we have

$$P_{\mathbb{F}}\left(x\right) = \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P\left(x\right) = P_{\mathbb{F}}\left(x\right)$ holds.

Next, consider the case where

$$2 - \left(2^{E-1} - 1\right) \leq N.$$

Now, we have 3 cases according to $round_{\mathbb{F}}$, that is, the case where $round_{\mathbb{F}}$ is Directed-Rounding (Toward $-\infty$), the case where $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$), and the case where

(i) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $-\infty$).

(1) Case: $val_{\mathbb{F}}\left(0,0,0\right) \leq x < val_{\mathbb{F}}\left(0,1,0\right)$.

In this case, we can express x by

$$x = val_{\mathbb{F}}\left(0, 0, m'\right)$$

for an integer m' that satisfies $0 \leq m' < 2^{M}$. Hence, we obtain

$$\begin{aligned}P(x) &= Pr\left[e = 0, m = m' \text{ in } 50\right]\\ &= Pr\left[e = 0, m = m' \text{ in } 41\right]\\ &= Pr\left[e = 0, m = m' \text{ in } 40\right]\\ &= Pr\left[e = 0, m = m' \text{ in } 30\right]\\ &= Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[e = 0, m = m' \text{ in } 30\right]\\ &= Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right]\\ &= Pr\left[n = e_{max} \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right]\\ &= 2^{1-e_{max}} \times 2^{-M}\\ &= 2^{1-\left(N+M+2^{E-1}-1\right)}.\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(2) Case: $val_{\mathbb{F}}\left(0, e', 0\right) \leq x < val_{\mathbb{F}}\left(0, e'+1, 0\right)$ where $\left(1 \leq e' \leq N + 2^{E-1} - 2\right)$.

In this case, we can express x by

$$x = val_{\mathbb{F}}\left(0, e', m'\right)$$

for an integer m' that satisfies $0 \leq m' < 2^M$. Hence, we obtain

$$\begin{aligned}P(x) &= Pr\left[e = e', m = m' \text{ in } 50\right]\\ &= Pr\left[e = e', m = m' \text{ in } 41\right]\\ &= Pr\left[e = e', m = m' \text{ in } 40\right]\\ &= Pr\left[e = e', m = m' \text{ in } 30\right]\\ &= Pr\left[e = e' \text{ in } 20\right] \times Pr\left[e = e', m = m' \text{ in } 30\right]\\ &= Pr\left[e = e' \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right]\\ &= Pr\left[n = e_{max} - e' \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right]\\ &= 2^{e'-e_{max}} \times 2^{-M}\\ &= 2^{e'-\left(N+M+2^{E-1}-1\right)}.\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 2^{e'-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(3) Case: $x = val_{\mathbb{F}}\left(0, N + 2^{E-1} - 1, 0\right) = 2^N$.

In this case, we can express x by $x = val_{\mathbb{F}}(0, e_{max}, 0)$. Hence we obtain

$$\begin{aligned}P(x) &= Pr\left[e = e_{max}, m = 0 \text{ in } 50\right]\\ &= Pr\left[e = e_{max}, m = 0 \text{ in } 41\right]\\ &= Pr\left[e = e_{max}, m = 0 \text{ in } 40\right]\\ &= Pr\left[e = e_{max}, m = 0 \text{ in } 30\right]\\ &= Pr\left[e = e_{max} \text{ in } 20\right] \times Pr\left[e = e_{max}, m = 0 \text{ in } 30\right]\\ &= Pr\left[e = e_{max} \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right]\\ &= Pr\left[n = 0 \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right]\\ &= 0 \times Pr\left[m = 0 \text{ in } 30\right]\\ &= 0.\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 0$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(ii) Case: $round_{\mathbb{F}}$ is Directed-Rounding(Toward $+\infty$).

(1) Case: $x = 0$.

In this case, we can express x by $x = val_{\mathbb{F}}(0,0,0)$. Hence, we obtain

$$\begin{aligned}
P(x) &= Pr\left[e = 0, m = 0 \text{ in } 50\right] \\
&= Pr\left[e = 0, m = 0 \text{ in } 42\right] \\
&= Pr\left[e = -1, m = 2^M - 1 \text{ in } 40\right] \\
&= Pr\left[e = -1, m = 2^M - 1 \text{ in } 30\right] \\
&= Pr\left[e = -1 \text{ in } 20\right] \times Pr\left[e = -1, m = 2^M - 1 \text{ in } 30\right] \\
&= Pr\left[e = -1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= Pr\left[n = e_{max} + 1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= 0 \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= 0.
\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 0$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(2) Case: $val_{\mathbb{F}}(0,0,1) \leq x \leq val_{\mathbb{F}}(0,1,0)$.

Consider the case where the mantissa of x is 0 and the case where the mantissa of x is not 0.

(a) Case: $x = val_{\mathbb{F}}(0,1,0)$.

In this case, we have

$$\begin{aligned}
P(x) &= Pr\left[e = 1, m = 0 \text{ in } 50\right] \\
&= Pr\left[e = 1, m = 0 \text{ in } 42\right] \\
&= Pr\left[e = 0, m = 2^{M-1} \text{ in } 40\right] \\
&= Pr\left[e = 0, m = 2^{M-1} \text{ in } 30\right] \\
&= Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[e = 0, m = 2^M - 1 \text{ in } 30\right] \\
&= Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= Pr\left[n = e_{max} \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= 2^{1-e_{max}} \times 2^{-M} \\
&= 2^{1-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}$$

(b) Case: $x = val_{\mathbb{F}}\left(0,0,m'\right)$ where $\left(1 \leq m' < 2^M\right)$.

In this case, we have

$$\begin{aligned}
P(x) &= Pr\left[e = 0, m = m' \text{ in } 50\right] \\
&= Pr\left[e = 0, m = m' \text{ in } 42\right] \\
&= Pr\left[e = 0, m = m' - 1 \text{ in } 40\right] \\
&= Pr\left[e = 0, m = m' - 1 \text{ in } 30\right] \\
&= Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[e = 0, m = m' - 1 \text{ in } 30\right] \\
&= Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= Pr\left[n = e_{max} \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= 2^{1-e_{max}} \times 2^{-M} \\
&= 2^{1-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}$$

Thus, we obtain

$$P\left(x\right) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

in both cases. Besides, we have

$$P_{\mathbb{F}}\left(x\right) = 2^{1-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P\left(x\right) = P_{\mathbb{F}}\left(x\right)$ holds.

(3) Case: $val_{\mathbb{F}}\left(0, e', 1\right) \leq x \leq val_{\mathbb{F}}\left(0, e'+1, 0\right)$ where $\left(1 \leq e' \leq N + 2^{E-1} - 2\right)$.

Consider the case where the mantissa of x is 0 and the case where the mantissa of x is not 0.

(a) Case: $x = val_{\mathbb{F}}\left(0, e'+1, 0\right)$.

In this case, we have

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = e'+1, m = 0 \text{ in } 50\right] \\
&= Pr\left[e = e'+1, m = 0 \text{ in } 42\right] \\
&= Pr\left[e = e', m = 2^{M-1} \text{ in } 40\right] \\
&= Pr\left[e = e', m = 2^{M-1} \text{ in } 30\right] \\
&= Pr\left[e = e' \text{ in } 20\right] \times Pr\left[e = e', m = 2^{M} - 1 \text{ in } 30\right] \\
&= Pr\left[e = e' \text{ in } 20\right] \times Pr\left[m = 2^{M} - 1 \text{ in } 30\right] \\
&= Pr\left[e = e_{max} - e' \text{ in } 20\right] \times Pr\left[m = 2^{M} - 1 \text{ in } 30\right] \\
&= 2^{e'-e_{max}} \times 2^{-M} \\
&= 2^{e'-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}$$

(b) Case $x = val_{\mathbb{F}}\left(0, e', m'\right)$ where $\left(1 \leq m' < 2^{M}\right)$.

In this case, we have

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = e', m = m' \text{ in } 50\right] \\
&= Pr\left[e = e', m = m' \text{ in } 42\right] \\
&= Pr\left[e = e', m = m' - 1 \text{ in } 40\right] \\
&= Pr\left[e = e', m = m' - 1 \text{ in } 30\right] \\
&= Pr\left[e = e' \text{ in } 20\right] \times Pr\left[e = e', m = m' - 1 \text{ in } 30\right] \\
&= Pr\left[e = e' \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= Pr\left[n = e_{max} - e' \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= 2^{e'-e_{max}} \times 2^{-M} \\
&= 2^{e'-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}$$

Thus, we obtain

$$P\left(x\right) = 2^{e'-\left(N+M+2^{E-1}-1\right)}$$

in both cases. Besides, we have

$$P_{\mathbb{F}}\left(x\right) = 2^{e'-\left(N+M+2^{E-1}-1\right)}$$

by the Section 4.2.3. Therefore, $P\left(x\right) = P_{\mathbb{F}}\left(x\right)$ holds.

(iii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

(1) Case: $x = 0$.

In this case, we can express x by $x = val_{\mathbb{F}}\left(0, 0, 0\right)$. Hence, we obtain

$$
\begin{aligned}
P(x) &= Pr\left[e = 0, m = 0 \text{ in } 50\right] \\
&= Pr\left[e = 0, m = 0 \text{ in } 41\right] + Pr\left[e = 0, m = 0 \text{ in } 42\right] \\
&= Pr\left[b = 0, e = 0, m = 0 \text{ in } 43\right] + Pr\left[b = 1, e = -1, m = 2^M - 1 \text{ in } 43\right] \\
&= \frac{1}{2} \times Pr\left[e = 0, m = 0 \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e = -1, m = 2^M - 1 \text{ in } 40\right] \\
&= \frac{1}{2} \times Pr\left[e = 0, m = 0 \text{ in } 30\right] + \frac{1}{2} \times Pr\left[e = -1, m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[e = 0, m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = -1 \text{ in } 20\right] \times Pr\left[e = -1, m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = -1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[n = e_{max} \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[n = e_{max} + 1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times 2^{1-e_{max}} \times 2^{-M} + \frac{1}{2} \times 0 \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}
$$

Besides, we have

$$
P_{\mathbb{F}}(x) = \frac{1}{2} \times 2^{1-\left(N+M+2^{E-1}-1\right)}
$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(2) Case: $val_{\mathbb{F}}(0, 0, 1) \leq x < val_{\mathbb{F}}(0, 1, 0)$.

In this case, we can express x by

$$
x = val_{\mathbb{F}}\left(0, 0, m'\right)
$$

for an integer m' that satisfies $1 \leq m' < 2^M$. Hence, we obtain

$$
\begin{aligned}
P(x) &= Pr\left[e = 0, m = m' \text{ in } 50\right] \\
&= Pr\left[e = 0, m = m' \text{ in } 41\right] + Pr\left[e = 0, m = m' \text{ in } 42\right] \\
&= Pr\left[b = 0, e = 0, m = m' \text{ in } 43\right] + Pr\left[b = 1, e = 0, m = m' - 1 \text{ in } 43\right] \\
&= \frac{1}{2} \times Pr\left[e = 0, m = m' \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e = 0, m = m' - 1 \text{ in } 40\right] \\
&= \frac{1}{2} \times Pr\left[e = 0, m = m' \text{ in } 30\right] + \frac{1}{2} \times Pr\left[e = 0, m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[e = 0, m = m' \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[e = 0, m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = 0 \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[n = e_{max} \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[n = e_{max} \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times 2^{1-e_{max}} \times 2^{-M} + \frac{1}{2} \times 2^{1-e_{max}} \times 2^{-M} \\
&= 2^{1-\left(N+M+2^{E-1}-1\right)}.
\end{aligned}
$$

Besides, we have

$$
P_{\mathbb{F}}(x) = 2^{1-\left(N+M+2^{E-1}-1\right)}
$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(3) Case: $val_{\mathbb{F}}\left(0, e', 1\right) \leq x < val_{\mathbb{F}}\left(0, e'+1, 0\right)$ where $\left(1 \leq e' \leq N + 2^{E-1} - 2\right)$.
In this case, we can express x by

$$x = val_{\mathbb{F}}\left(0, e', m'\right)$$

for an integer m' that satisfies $1 \leq m' < 2^M$. Hence, we obtain

$$\begin{aligned}
P(x) &= Pr\left[e = e', m = m' \text{ in } 50\right] \\
&= Pr\left[e = e', m = m' \text{ in } 41\right] + Pr\left[e = e', m = m' \text{ in } 42\right] \\
&= Pr\left[b = 0, e = e', m = m' \text{ in } 43\right] + Pr\left[b = 1, e = e', m = m' - 1 \text{ in } 43\right] \\
&= \frac{1}{2} \times Pr\left[e = e', m = m' \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e = e', m = m' - 1 \text{ in } 40\right] \\
&= \frac{1}{2} \times Pr\left[e = e', m = m' \text{ in } 30\right] + \frac{1}{2} \times Pr\left[e = e', m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = e' \text{ in } 20\right] \times Pr\left[e = e', m = m' \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = e' \text{ in } 20\right] \times Pr\left[e = e', m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = e' \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = e' \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[n = e_{max} - e' \text{ in } 20\right] \times Pr\left[m = m' \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[n = e_{max} - e' \text{ in } 20\right] \times Pr\left[m = m' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times 2^{e' - e_{max}} \times 2^{-M} + \frac{1}{2} \times 2^{e' - e_{max}} \times 2^{-M} \\
&= 2^{e' - \left(N + M + 2^{E-1} - 1\right)}.
\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}(x) = 2^{e' - \left(N + M + 2^{E-1} - 1\right)}$$

by the Section 4.2.3. Therefore, $P(x) = P_{\mathbb{F}}(x)$ holds.

(4) Case: $x = val_{\mathbb{F}}\left(0, e', 0\right)$ where $\left(2 \leq e' \leq N + 2^{E-1} - 2\right)$
In this case, we obtain

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = e', m = 0 \text{ in } 50\right] \\
&= Pr\left[e = e', m = 0 \text{ in } 41\right] + Pr\left[e = e', m = 0 \text{ in } 42\right] \\
&= Pr\left[b = 0, e = e', m = 0 \text{ in } 43\right] + Pr\left[b = 1, e = e' - 1, m = 2^M - 1 \text{ in } 43\right] \\
&= \frac{1}{2} \times Pr\left[e = e', m = 0 \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e = e' - 1, m = 2^M - 1 \text{ in } 40\right] \\
&= \frac{1}{2} \times Pr\left[e = e', m = 0 \text{ in } 30\right] + \frac{1}{2} \times Pr\left[e = e' - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = e' \text{ in } 20\right] \times Pr\left[e = e', m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = e' - 1 \text{ in } 20\right] \times Pr\left[e = e', m = 0 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = e' \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = e' - 1 \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[n = e_{max} - e' \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[n = e_{max} - e' + 1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times 2^{e' - e_{max}} \times 2^{-M} + \frac{1}{2} \times 2^{e' - e_{max} - 1} \times 2^{-M} \\
&= \frac{2}{4} \times 2^{e' - e_{max}} \times 2^{-M} + \frac{1}{4} \times 2^{e' - e_{max}} \times 2^{-M} \\
&= \frac{3}{4} \times 2^{e' - \left(N + M + 2^{E-1} - 1\right)}.
\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}\left(x\right) = \frac{3}{4} \times 2^{e' - \left(N + M + 2^{E-1} - 1\right)}$$

by the Section 4.2.3. Therefore, $P\left(x\right) = P_{\mathbb{F}}\left(x\right)$ holds.

(5) Case: $x = val_{\mathbb{F}}\left(0, N + 2^{E-1} - 1, 0\right) = 2^N$.

In this case, we can express x by $x = val_{\mathbb{F}}\left(0, e_{max}, 0\right)$. Hence, we obtain

$$\begin{aligned}
P\left(x\right) &= Pr\left[e = e_{max}, m = 0 \text{ in } 50\right] \\
&= Pr\left[e = e_{max}, m = 0 \text{ in } 41\right] + Pr\left[e = e_{max} - 1, m = 2^M - 1 \text{ in } 42\right] \\
&= Pr\left[b = 0, e = e_{max}, m = 0 \text{ in } 43\right] + Pr\left[b = 1, e = e_{max} - 1, m = 2^M - 1 \text{ in } 43\right] \\
&= \frac{1}{2} \times Pr\left[e = e_{max}, m = 0 \text{ in } 40\right] + \frac{1}{2} \times Pr\left[e = e_{max} - 1, m = 2^M - 1 \text{ in } 40\right] \\
&= \frac{1}{2} \times Pr\left[e = e_{max}, m = 0 \text{ in } 30\right] + \frac{1}{2} \times Pr\left[e = e_{max} - 1, m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = e_{max} \text{ in } 20\right] \times Pr\left[e = e_{max}, m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = e_{max} - 1 \text{ in } 20\right] \times Pr\left[e = e_{max} - 1, m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[e = e_{max} \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[e = e_{max} - 1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 30\right] \\
&= \frac{1}{2} \times Pr\left[n = 0 \text{ in } 20\right] \times Pr\left[m = 0 \text{ in } 30\right] \\
&\quad + \frac{1}{2} \times Pr\left[n = 1 \text{ in } 20\right] \times Pr\left[m = 2^M - 1 \text{ in } 33\right] \\
&= \frac{1}{2} \times 0 \times Pr\left[m = 0 \text{ in } 30\right] + \frac{1}{2} \times 2^{-1} \times 2^{-M} \\
&= 2^{-(M+2)}
\end{aligned}$$

Besides, we have

$$P_{\mathbb{F}}\left(x\right) = 2^{-(M+2)}$$

by the Section 4.2.3. Therefore, $P\left(x\right) = P_{\mathbb{F}}\left(x\right)$ holds.

Table 3 Environment

CPU	Intel® Core$^{\mathrm{TM}}$ i7-4702MQ
OS	Ubuntu 12.04 LTS 64-bit
Kernel	Linux 3.13.4-031304-generic
Compiler	g++ 4.6.3
Source code	https://goo.gl/K1NAnE
Rounding mode	Round-to-Nearest(Ties to Even)

5.3 Experiment for correctness

Now, we have proved that the modified algorithm is uniform in narrow sense. This section confirms that the algorithm can remove the strange behaviors of Thoma's method by some experiments.

5.3.1 Target

In this experiment, the target is the following floating point uniform random number generator.

- Ratio method.
  The floating point uniform random number generator that outputs $\frac{URNG_W()}{2^W}$.
- Moler's method.
  The floating point uniform random number generator proposed by Moler [25].
- Thoma's method.
  The floating point uniform random number generator proposed by Thoma [31].
- Modified method.
  The modified floating point uniform random number generator proposed in the Section 5.1. Let $N = 0$ so that $U_{\mathbb{R}} = [0, 1]$ in the generator.

Here, the authors used Round-to-Nearest(Ties to Even) for $fl_{\mathbb{F}}$ and $round_{\mathbb{F}}$[*20] and used the 32/64-bit Mersenne Twister [23] for $URNG_W$ in each generator.

5.3.2 Environment

Table 6 shows the environment where the experiments was done.

5.3.3 Methodology

The experiment consists of the following 2 parts.

Part 1 Test for all the floating point numbers in $[0, 1]$.

This part measures the random number generation probability for all the floating point numbers where $(E, M) = (5, 4)$ and then compares them with the values of $P_{\mathbb{F}}$ calculated by the Formula (1).

In the concrete, generate 2^{30} floating point uniform random numbers and calculate the generation probability for each floating point number. Then, calculate $P_{\mathbb{F}}$ by the Formula (1) and test the null hypothesis "The random number generation probability is uniform in narrow sense" by χ^2 test[*21][*22]. Here, the authors let $W = 7$ in this part[*23].

Part 2 Test for specific single precision floating point numbers.

This part makes the same experiment as part 1 for single precision[*24] floating point numbers in $\left[2^{-8} - 2^{-25}, 2^{-8} + 2^{-25}\right]$[*25]. In this part, the authors generated 2^{40} floating point uniform random numbers so that about 2^{16} numbers fell within $\left[2^{-8} - 2^{-25}, 2^{-8} + 2^{-25}\right]$ and let $W = 32$ because 1 single precision floating point number consisted of 32 bits.

Here, the authors used Keisan Online Calculator(`http://keisan.casio.jp/exec/system/1161228834`) , which is provided by CASIO COMPUTER CO., LTD., in order to calculate percent points in the χ^2 test.

5.3.4 Result and discussion: Part 1

Table 4 shows the result of the χ^2 value and Figure 4, Figure 6, Figure 8, and Figure 10 shows the random number generation probability of Ratio method, Moler's method, Thoma's method, and the modified method in $[0, 1]$ respectively. Besides, Figure 5, Figure 7, Figure 9, Figure 11 shows the random number generation probability in $\left[0, 2^{3-\left(2^{E-1}-1\right)}\right] = \left[0, 2^{-12}\right]$.

First, the Figure 4 and Figure 8 shows that the probability of Ratio method and Thoma's method waves and is different from the ideal probability denoted by the red line for almost all random numbers. Here, the reason why the probability of Ratio method and Thoma's method is similar to the ideal one in $\left(2^{-3}, 2^{-2}\right)$ is that we can calculate $\frac{URNG_7()}{2^7}$ without any rounding error because $\frac{URNG_7()}{2^7}$ can be expressed by $5\,(= M + 1)$ bits floating point number in this region.

Next, the Figure 6 and Figure 7 shows that the probability of Moler's method is similar to the ideal one in whole random numbers except near 0. The reason why the probability of Moler's method near zero is like a hill is that Moler's method can generate 0 in the middle of the operation and output subnormal numbers by taking xor-mask of a uniform random integer to

[*20] $round_{\mathbb{F}}$ is used for $P_{\mathbb{F}}$, which is ideal probability.
[*21] The degree of freedom is $\left(2^{E-1} - 1\right) \times 2^M + 1 - 1 = 240$.
[*22] Strictly speaking, the authors tested the generation number of each floating point number.
[*23] The authors used $(E, M, W) = (5, 4, 7)$ because more kinds of problem had been detected when E, M, W was coprime each other.
[*24] That is, $(E, M) = (8, 23)$.
[*25] The degree of freedom of the χ^2 test is $\left(2^{E-1} - 1\right) \times 2^M + 1 - 1 = 1065353216$.

the mantissa of this 0.

Last is the modified method. The Figure 10 shows that the probability of the modified method is similar to the ideal one in whole random numbers and the Figure 11 shows that the probability near 0 is also similar to the ideal one.

As if the Table 4 supports these, it shows that the χ^2 value is quite greater than that of 99.9% point in each generator except the modified method and the χ^2 test rejects the null-hypothesis that the random number generation probability is uniform. On the other hand, the χ^2 value of the modified method is less than that of 95% and the test does not reject the null-hypothesis.

5.3.5 Result and discussion: Part 2

Table 5 shows the result of the χ^2 value and Figure 12, Figure 13, Figure 14, and Figure 15 shows the random number generation probability of Ratio method, Moler's method, Thoma's method, and the modified method in $\left[2^{-8} - 2^{-25}, 2^{-8} + 2^{-25}\right]$ respectively.

First, the Figure 12 and Figure 14 shows that Ratio method and Thoma's method takes quite different behavior between the left side of $2^{-8} = 2^{-(W-M-1)}$ and the right side, which is far away from the ideal probability denoted by the red line. The reason of this strange behavior is explained in the Section 3.2, and the result supports the explanation. On the other hand, the reason of the ideal behavior in the left side can be explained in the same way as the Figure 4 and Figure 8. That is, we can calculate $\frac{URNG_W()}{2^W}$ without any rounding error because $\frac{URNG_W()}{2^W}$ can be expressed by $M+1$ bits floating point number in $\left(2^{-(W-M)}, 2^{-(W-M-1)}\right)$.

Next, we can not find any problems of Moler's method and the modified method from the Figure 13 and Figure 15. However, the Table 5 shows that the χ^2 value is quite greater than that of 99.9% point in each generator except the modified method and the χ^2 test rejects the null-hypothesis that the random number generation probability is uniform. On the other hand, the χ^2 value of the modified method is less than that of 95% and the test does not reject the null-hypothesis. Here, the reason why the χ^2 test rejects null-hypothesis of Moler's method even if we can not find any problem from the figure is that the generation probability of subnormal numbers is quite higher than the ideal probability.

5.4 Summary

This section has proposed the modified algorithm for Thoma's method and proved its correctness. However, the modified method has the following disadvantage.

(1) The random number generation range is only $\left[0, 2^N\right]$.
For example, if we multiply the output of the modified method by t in order to obtain uniform random numbers in $\left[0, t \times 2^N\right]$, the algorithm does not guarantee that we can obtain all the floating point numbers in $\left[0, t \times 2^N\right]$.

(2) We can not receive so much advantage on IEEE754 double precision.
For example, Moler's method can generate almost all the floating point numbers[*26] in $\left[2^{-53}, 1 - 2^{-53}\right]$ without any problem[*27][*28]. Thus, we can receive advantages by the modified method only when the generator outputs a floating point number in $\left[0, 2^{-53}\right) \cup \left(1 - 2^{-53}, 1\right]$. However, the probability that we obtain such a floating point number is at most $2^{-53} \times 2 = 2^{-52}$. This probability is negligible for practical use.

Therefore, the next section proposes the method to improve the former[*29].

Table 4 χ^2 value for the random number generation probability where $(E, M, W) = (5, 4, 7)$.

Generator	χ^2 value($\times 10^2$)	P-value
Ratio method	3.4929120×10^8	$< 0.1\%$
Moler's method	1.8232878×10^8	$< 0.1\%$
Thoma's method	1.4334131×10^6	$< 0.1\%$
Modified method	2.2858594	n.s.
Point where P-value is 95.0%.	2.7713765	
Point where P-value is 99.0%.	2.9388810	
Point where P-value is 99.9%.	3.1343690	

6. Arbitrary range floating point random number generator

This section aims to improve the modified method in the Section 5 so that we can change the random number generation range with another floating point number. That is, this section proposes a floating point uniform random number generator that can output all the floating point numbers in arbitrary range whose edge is a floating point number.

[*26] All the floating point numbers except its mantissa is 0.

[*27] The random number generation probability is uniform in narrow sense, that is, the Formula (1) is satisfied by $round_{\mathbb{F}}$ =Round-to-Nearest(Ties to Even).

[*28] The Formula (1) is also satisfied when the mantissa is 0, but $round_{\mathbb{F}}$ is not Round-to-Nearest(Ties to Even). That is, the random number generation probability is uniform in wide sense.

[*29] The latter one is one of the future work.

Fig. 4 Random number generation probability of Ratio method in $[0, 1]$.

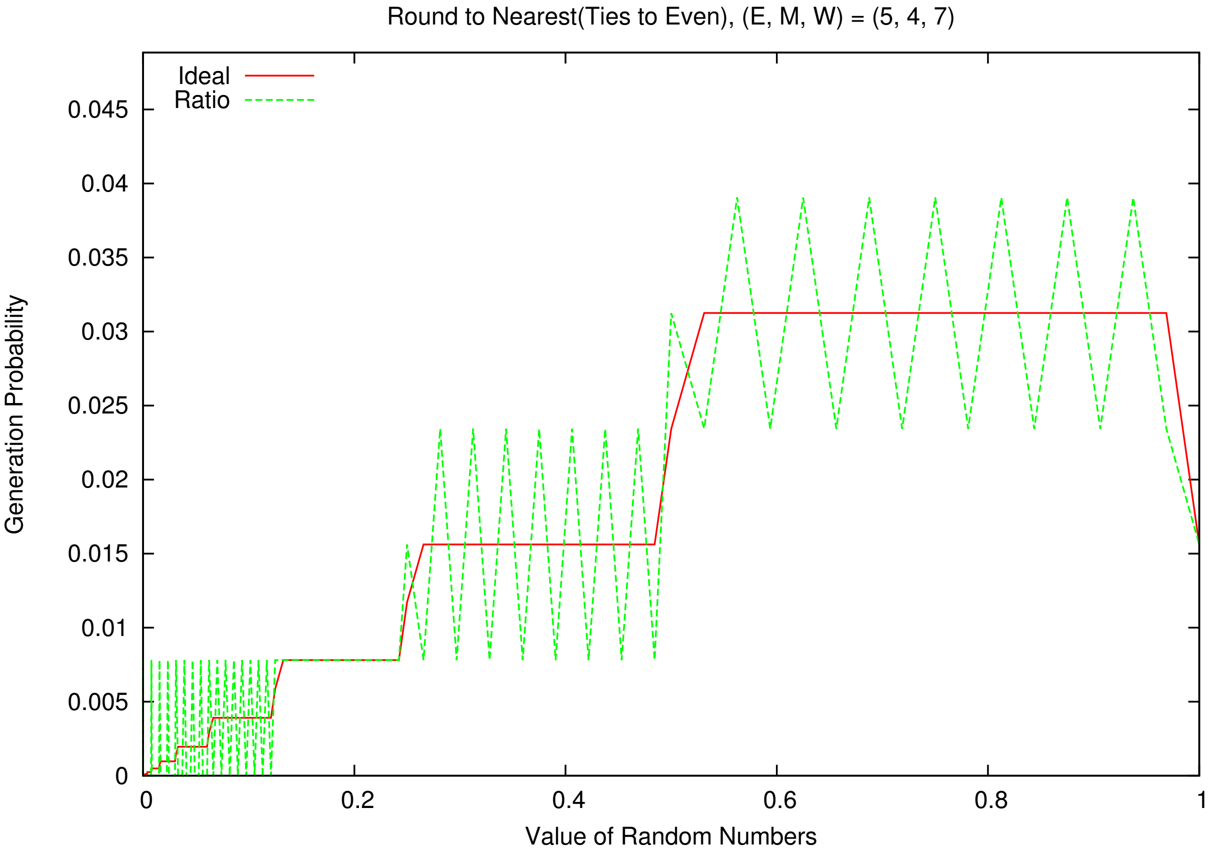


Fig. 5 Random number generation probability of Ratio method in $[0, 2^{-12}]$.

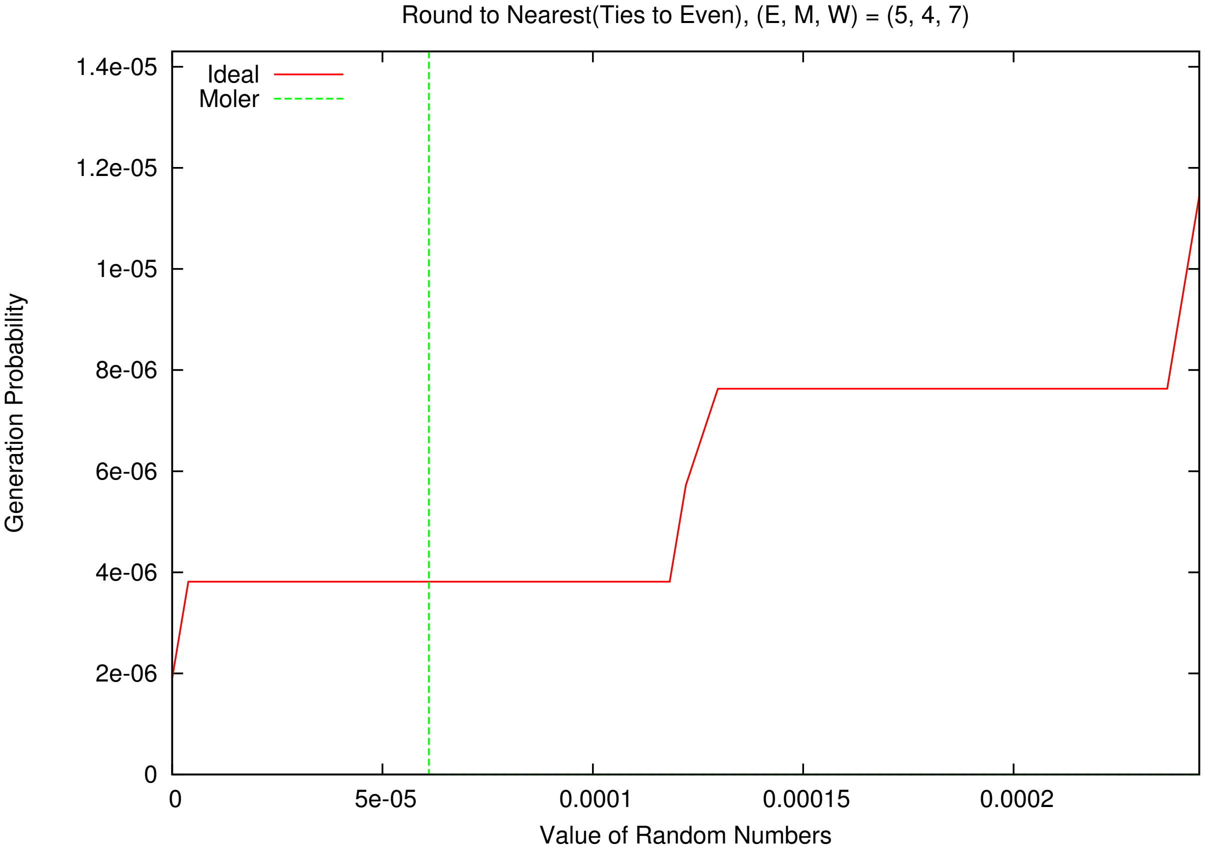


Fig. 6 Random number generation probability of Moler's method in $[0, 1]$.

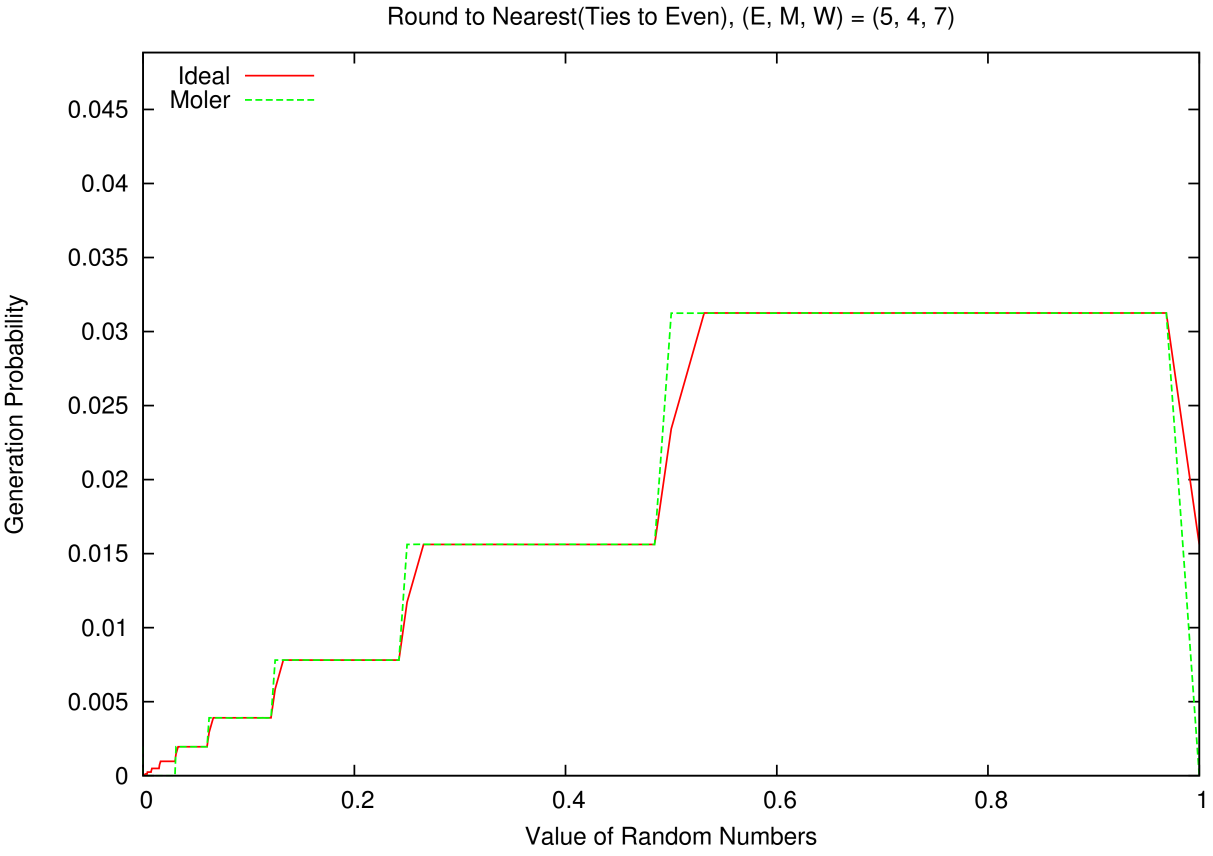


Fig. 7 Random number generation probability of Moler's method in $[0, 2^{-12}]$.

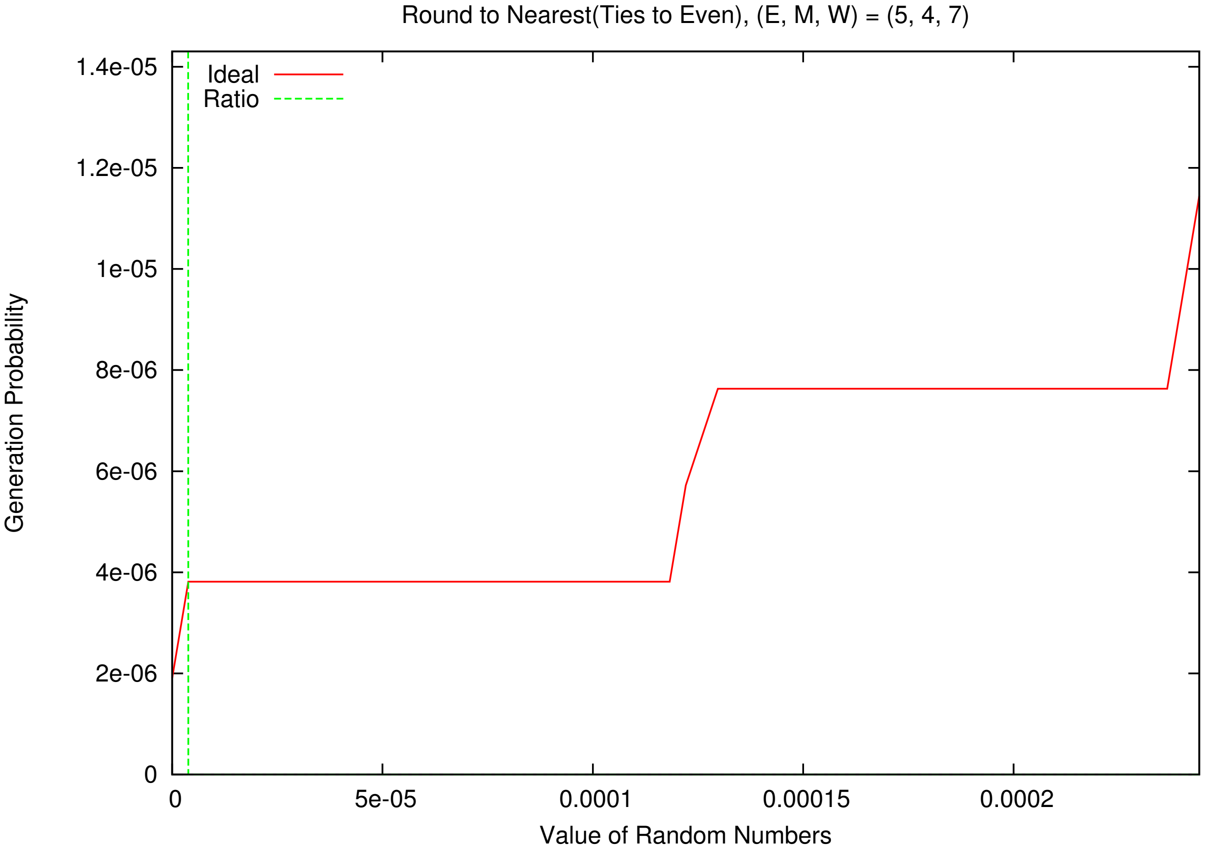


Fig. 8 Random number generation probability of Thoma's method in $[0, 1]$.

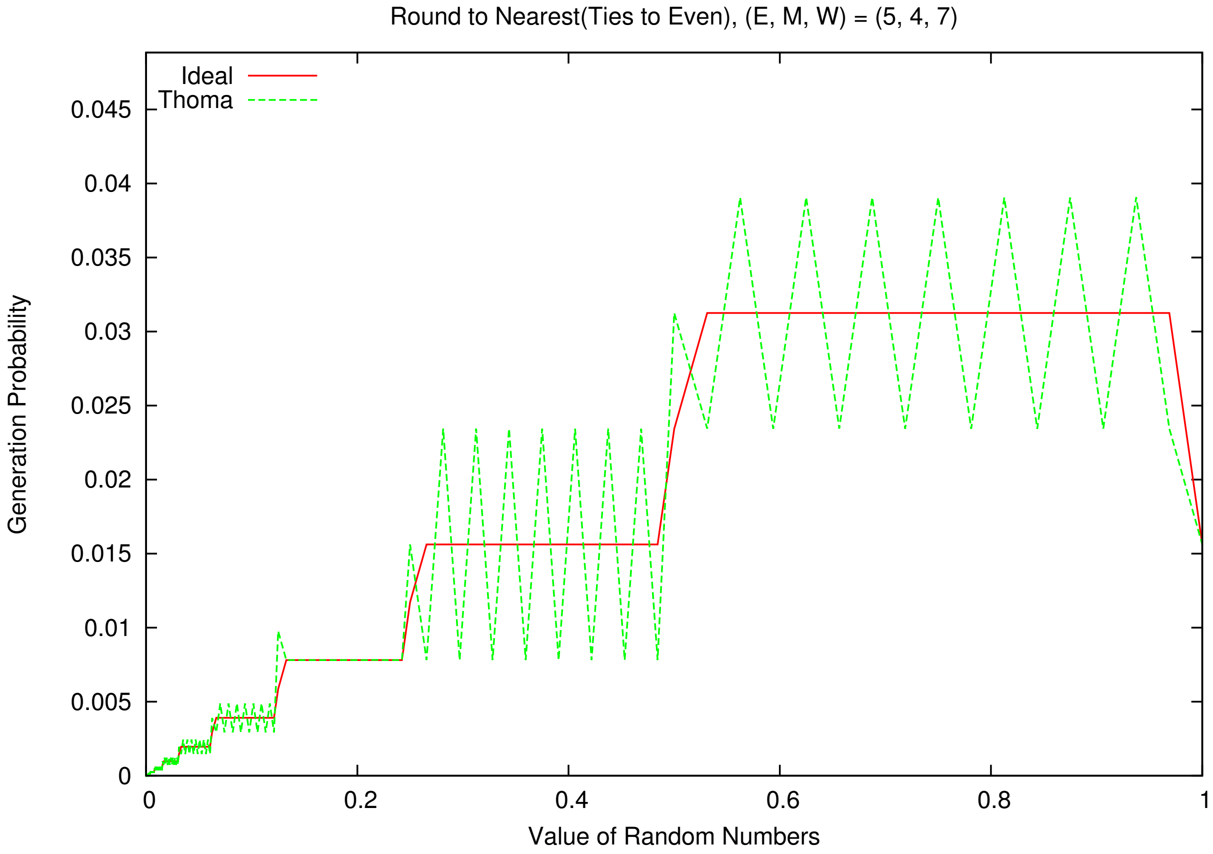


Fig. 9 Random number generation probability of Thoma's method in $[0, 2^{-12}]$.

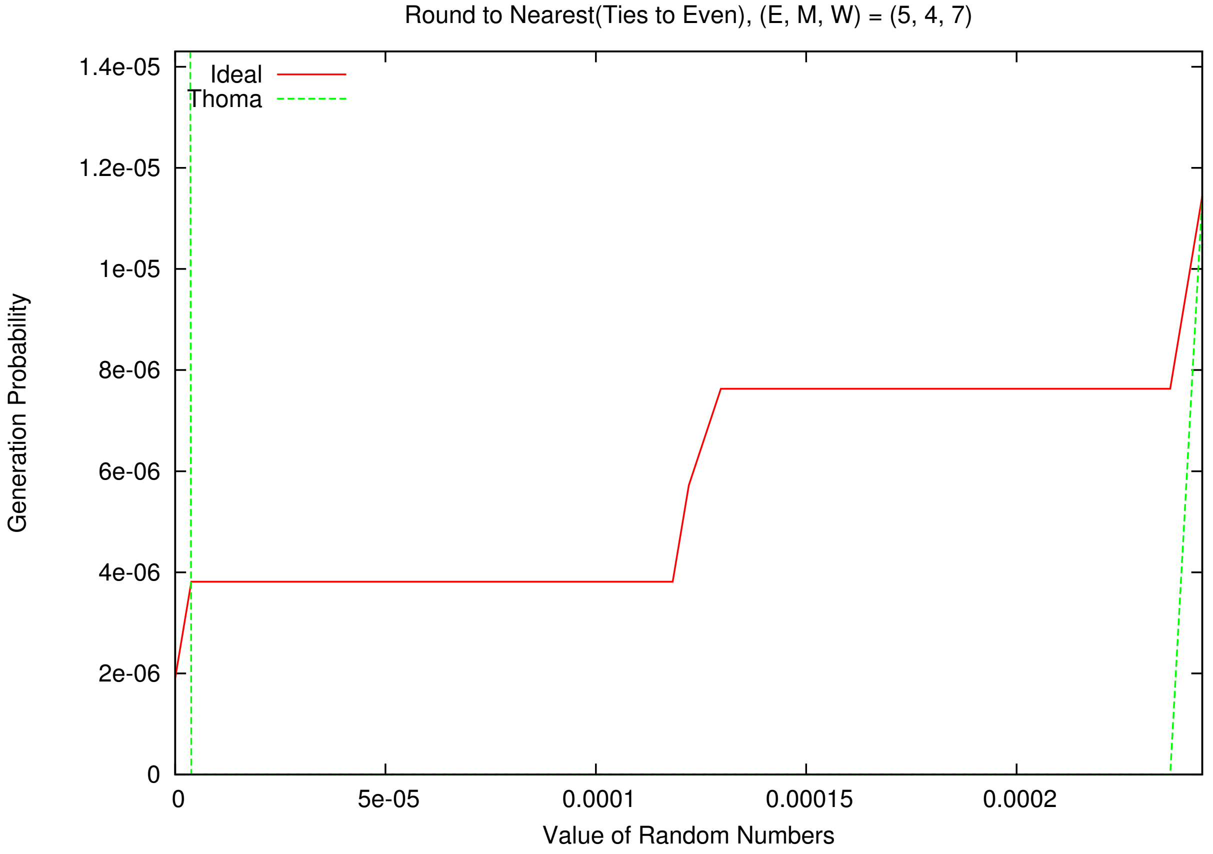


Fig. 10 Random number generation probability of modified method in $[0, 1]$.

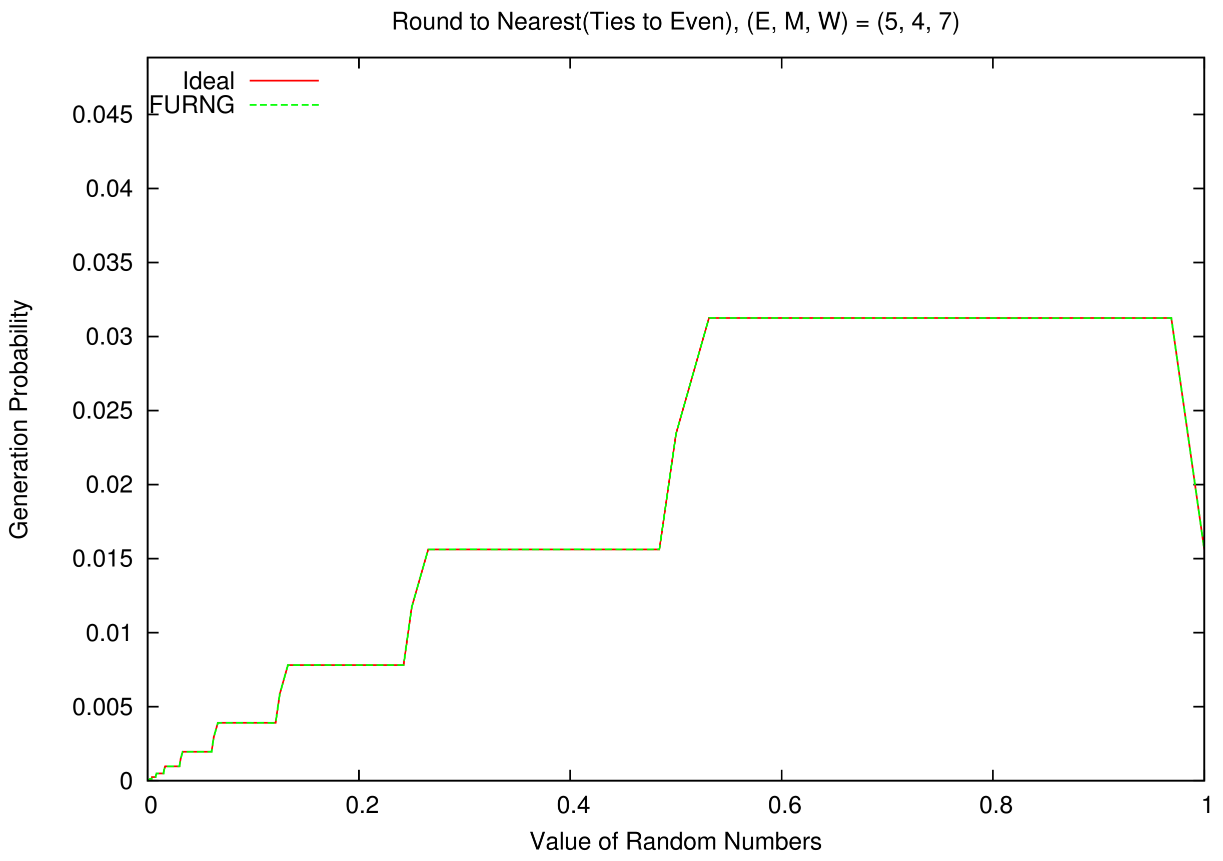


Fig. 11 Random number generation probability of modified method in $[0, 2^{-12}]$.

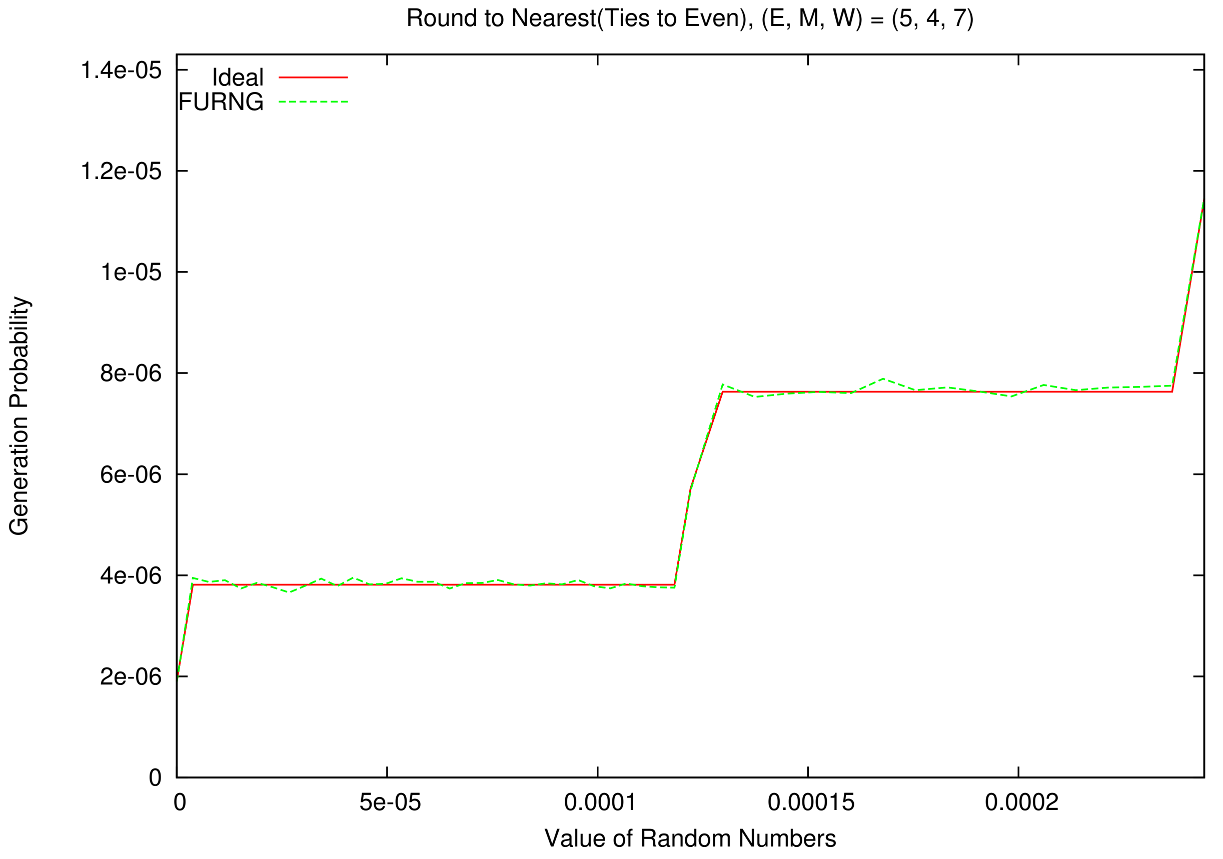


Fig. 12 Random number generation probability of Ratio method on single precision floating point number.

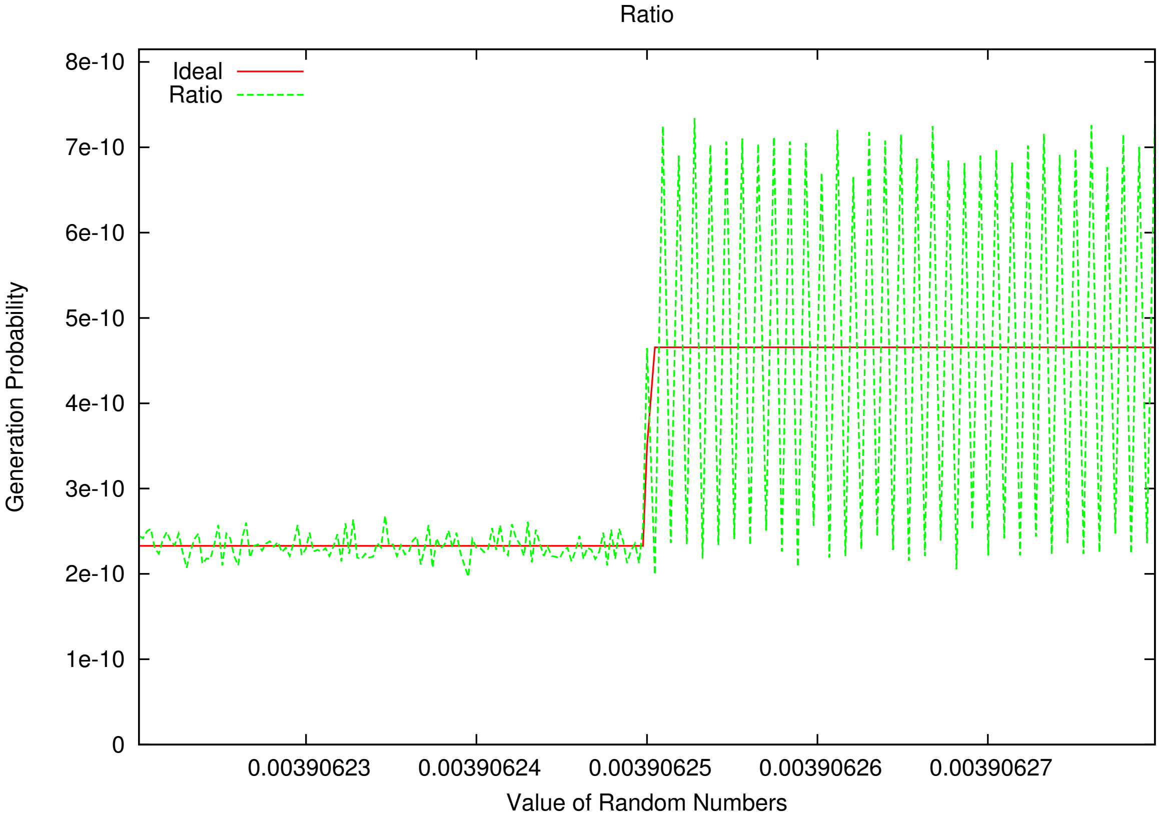


Fig. 13 Random number generation probability of Moler's method on single precision floating point number.

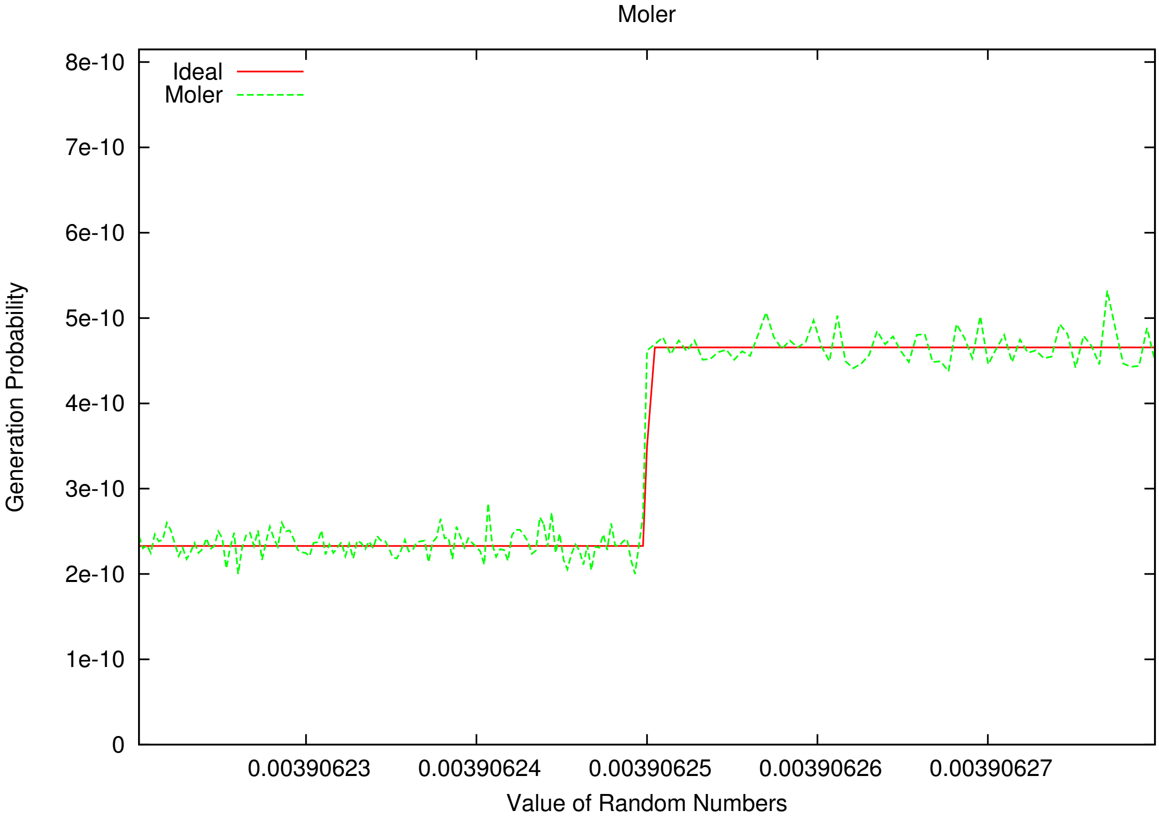


Fig. 14 Random number generation probability of Thoma's method on single precision floating point number.

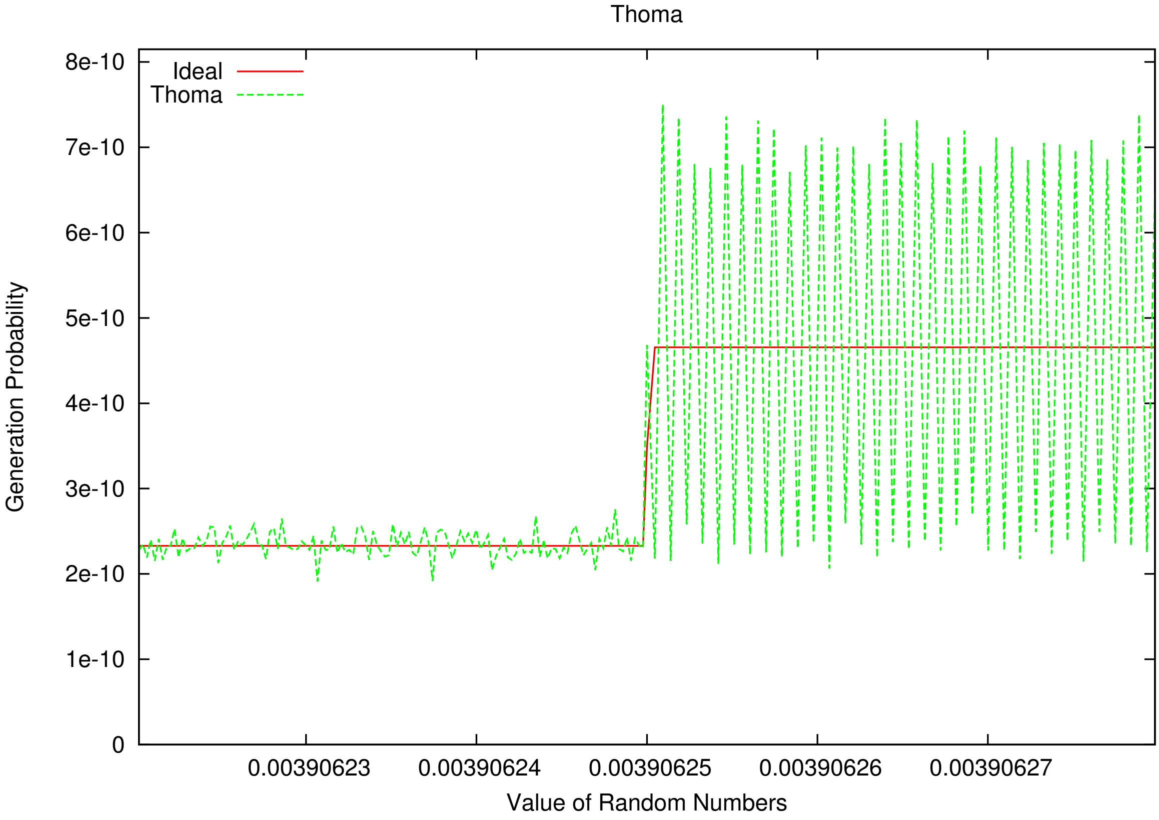


Fig. 15 Random number generation probability of the modified method on single precision floating point number.

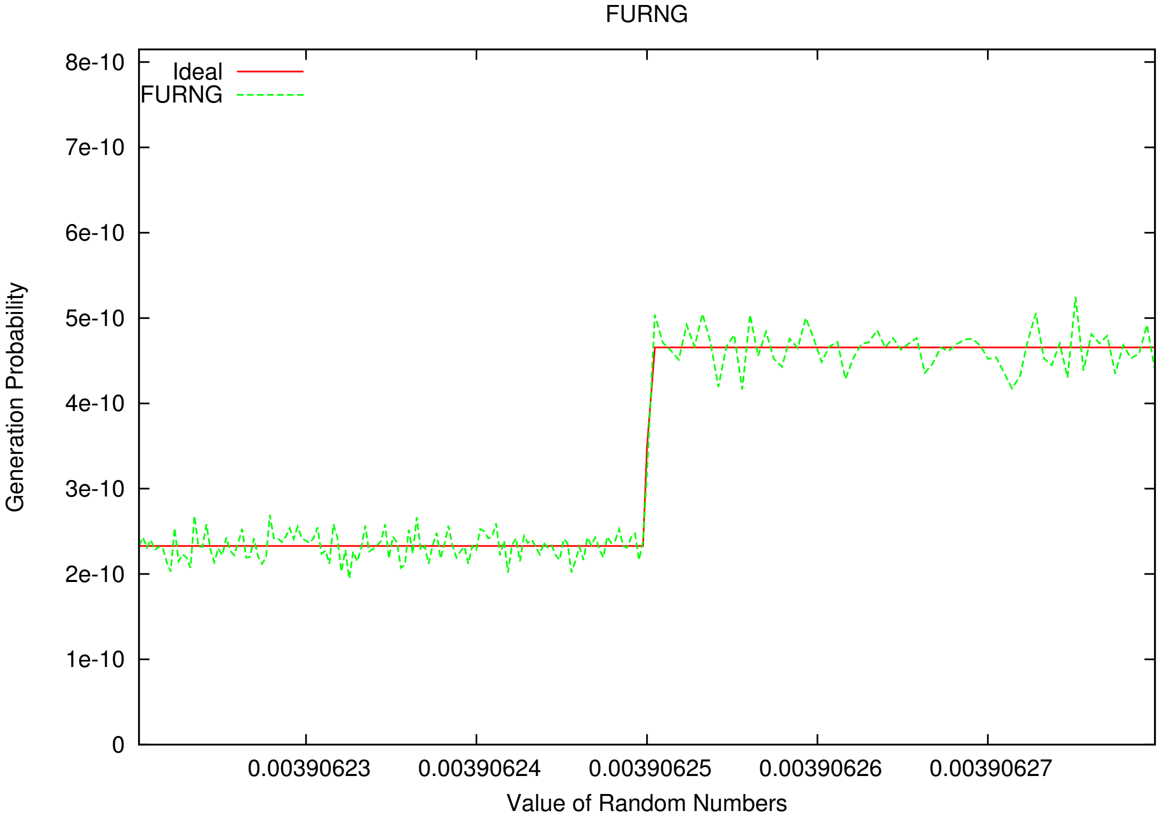


Table 5 χ^2 value for the random number generation probability on single precision floating point numbers($(E, M, W) = (8, 23, 32)$).

Generator	χ^2 value($\times 10^9$)	P-value
Ratio method	7.98368×10^{28}	$< 0.1\%$
Moler's method	4.30389×10^{28}	$< 0.1\%$
Thoma's method	2.49297	$< 0.1\%$
Modified method	0.339975	n.s.
Point where P-value is 95.0%.	1.065429143	
Point where P-value is 99.0%.	1.065460602	
Point where P-value is 99.9%.	1.065495866	

6.1 Algorithm

This section explains algorithm of $FURNG\,(a, b, round_{\mathbb{F}})$[*30]. Here, if $a = b$ then we just output $a\,(= b)$ with the probability 1. Hence, we consider only the case where $a < b$. Beside, if $a < b \le 0$ then we can consider $FURNG\,(-b, -a, -round_{\mathbb{F}})$ instead. Therefore, we need to consider the following 5 cases.

(1) Case: $0 \le a < b \le 2^{1-\left(2^{E-1}-1\right)}$.

In this case, the interval of a floating point number in $[a, b]$ is the same as each other. Besides, that of a floating point number in $[0, (b - a)]$ is also the same. Therefore, we can generate a floating point uniform random number in $[0, (b - a)]$ and output $(x + a)$ in this case. Here, the authors use acceptance-rejection method [33] for generating a floating point uniform random number in $[0, (b - a)]$.

The following is the concrete pseudocode.

```
00: Initialize.
  Find the minimal $k \in \mathbb{N}$ that satisfies $b - a \le 2^k$.
10: Generate a uniform random number.
  Generate a floating point uniform random number $x \in \mathbb{F}$ on $\left[0, 2^k\right]$ by using the modified algorithm in the Section 5.1.
      $x = FURNG\left(0, 2^k, round_{\mathbb{F}}\right).$
20: Judge Acceptance or Rejection.
  Judge whether accept or reject $x$ by the following rules.
  (Reject) Case: $b - a < x$.
      Reject $x$ and go back to 10.
  (Judge) Case: $x = b - a < 2^k$.
      (i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
          Reject $x$ and go back to 10.
      (ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
          Generate a random bit by $URNG_1\,()$. If the bit is 0, then reject $x$ and go back to 10. Otherwise, accept $x$ and go to 30.
      (iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
          Accept $x$ and go to 30.
  (Accept) Case: $x < b - a$ or $x = b - a = 2^k$.
      Accept $x$ and go to 30.
30: Output the result.
  Output $(x + a)$.
```

(2) Case: $2^{n-1} \le a < b \le 2^n$ where $(2 - \left(2^{E-1} - 1\right) \le n \in \mathbb{N} \le 2^{E-1})$.

In this case, the interval of a floating point number in $[a, b]$ is the same as each other. Therefore, we can generate a floating point uniform random number in the same way as (1).

In the concrete, by letting $p, q \in \mathbb{F}$ be

$$\begin{cases} p = \left(a - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \\ q = \left(b - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \end{cases}$$

then we have

$$0 \le p < q \le 2^{1-\left(2^{E-1}-1\right)}.$$

Thus, we can generate a floating point uniform random number $x \in \mathbb{F}$ in $[p, q]$ by the same way as (1) and output

$$x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}.$$

[*30] This paper just considers the case where $a, b \in \mathbb{F}$ in order to simplify the problem. Of course, $a \le b$.

(3) Case: $a < 0 < b$.

In this case, the sign of a is opposite to that of b. Therefore, we can output a floating point uniform random number in $[a, 0]$ with the probability of $\frac{-a}{b-a}$ and output a floating point uniform random number in $[0, b]$ with the probability of $\frac{b}{b-a}$. Here, we need to flip $round_{\mathbb{F}}$ horizontally when we generate a uniform random number in $[a, 0]$ because $a < 0$. The authors use acceptance-rejection method for generating a floating point uniform random number in $[a, 0]$ and $[0, b]$.

The following is the concrete pseudocode.

```
00: Initialize.
 Find the minimal $k \in \mathbb{N}$ that satisfies $\max(-a, b) \leq 2^k$.
10: Select the range.
 Generate a random bit by $URNG_1()$. If the bit is 0, then set $I = [-2^k, 0]$. Otherwise, set $I = [0, 2^k]$.
20: Generate a uniform random number.
 Generate a floating point uniform random number $x \in \mathbb{F}$ in $I$ by using the algorithm in the Section 5.1. Here, if
 $I = [-2^k, 0]$, then generate a floating point uniform random number $x' \in \mathbb{F}$ in $[0, 2^k]$ and then let $x = -x'$.
  • Case: $I = [0, 2^k]$.
        $x = FURNG(0, 2^k, round_{\mathbb{F}})$.
  • Case: $I = [-2^k, 0]$.
        $x = -FURNG(0, 2^k, -round_{\mathbb{F}})$.
30: Judge Acceptance or Rejection.
 (Reject) Case: $x < a$ or $b < x$.
    Reject $x$ and go back to 10.
 (Judge) Case: $x = a > -2^k$.
    (i) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward 0.
       Reject $x$ and go back to 10.
    (ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
       • Case: The mantissa of $a$ is not 0.
         Generate a random bit by $URNG_1()$. If the bit is 0, then reject $x$ and go back to 10. Otherwise, accept
         $x$ and go to 40.
       • Case: The mantissa of $a$ is 0.
         If $a = -2^{1-(2^{E-1}-1)}$, then the operation is the same as the case where the mantissa of $a$ is not 0. Otherwise,
         generate 2 random bits by $URNG_2()$. If the bits is 00, then reject $x$ and go back to 10. If the bits
         is 01 or 10, then accept $x$ and go to 40. If the bits is 11, then generate 2 random bits and judge again.
    (iii) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward $\pm\infty$.
       Accept $x$ and go to 40.
```

(Judge) $x = b < 2^k$

(i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

Reject x and go back to 10.

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

- Case: The mantissa of b is not 0.
  Generate a random bit by $URNG_1()$. If the bit is 0, then reject x and go back to 10. Otherwise, accept x and go to 40.
- Case: The mantissa of b is 0.
  If $b = 2^{1-(2^{E-1}-1)}$, then the operation is the same as the case where the mantissa of b is not 0. Otherwise, generate 2 random bits by $URNG_2()$. If the bits is 00 or 01, then reject x and go back to 10. If the bits is 10, then accept x and go to 40. If the bits is 11, then generate 2 random bits and judge again.

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

Accept x and go to 40.

(Accept) $x = a = -2^k$ or $a < x < b$ or $x = b = 2^k$

Accept x and go to 40.

40: Output the result.

Output x.

(4) Case: $2^{n-2} \leq a < 2^{n-1} < b \leq 2^{n}$ where $(2 - \left(2^{E-1} - 1\right) \leq n \in \mathbb{N} \leq 2^{E-1})$.

In this case, the interval of a floating point number in $\left[a, 2^{n-1}\right]$ is the same as each other. Besides, the interval of a floating point number in $\left[2^{n-1}, b\right]$ is also the same as each other [*31]. Therefore, we can generate a floating point uniform random number in $\left[a, 2^{n-1}\right]$ and $\left[2^{n-1}, b\right]$ in the same way as (1). Here, we need to choose $\left[a, 2^{n-1}\right]$ with the probability of $\frac{2^{n-1}-a}{b-a}$ and choose $\left[2^{n-1}, b\right]$ with the probability of $\frac{b-2^{n-1}}{b-a}$.

The following is the concrete pseudocode.

00: Initialize.

Let $p, q \in \mathbb{F}$ be

$$\begin{cases} p & = \left(a - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \\ q & = \left(b - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \end{cases} .$$

Since $2^{n-1} \leq a < 2^{n-1} < b \leq 2^{n}$, we obtain

$$-2^{1-\left(2^{E-1}-1\right)} \leq p < 0 < q \leq 2^{1-\left(2^{E-1}-1\right)}.$$

Next, find the minimal $k \in \mathbb{N}$ that satisfies $\max\left(-p, q\right) \leq 2^{k}$.

10: Select the range.

Generate 2 random bits by $URNG_2\left(\right)$. If the bits is 00, then set $I = \left[-2^{k}, 0\right]$. If the bits is 01 or 10, then set $I = \left[0, 2^{k}\right]$. If the bits is 11, then generate 2 random bits and judge again.

20: Generate a uniform random number.

Generate a floating point uniform random number $x \in \mathbb{F}$ in I by using the algorithm in the Section 5.1. Here, if $I = \left[-2^{k}, 0\right]$, then generate a floating point uniform random number $x' \in \mathbb{F}$ in $\left[0, 2^{k}\right]$ and then let $x = -x'$.

- Case: $I = \left[0, 2^{k}\right]$.

  $$x = FURNG\left(0, 2^{k}, round_{\mathbb{F}}\right)$$

- Case: $I = \left[-2^{k}, 0\right]$.

  In this case, we need to regard $round_{\mathbb{F}}$ as Toward $-\infty$ when $round_{\mathbb{F}}$ is Toward 0 and regard $round_{\mathbb{F}}$ as Toward $+\infty$ when $round_{\mathbb{F}}$ is Toward $\pm\infty$.

  $$x = -FURNG\left(0, 2^{k}, -round_{\mathbb{F}}\right)$$

30: Judge Acceptance or Rejection.

(Reject) Case: $x < p$ or $q < x$.

Reject x and go back to 10.

(Judge) Case: $x = p > -2^{k}$

- **(i)** Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

  Reject x and go back to 10.

- **(ii)** Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

  Generate a random bit by $URNG_1\left(\right)$. If the bit is 0, then reject x and go back to 10. Otherwise, accept x and go to 30.

- **(iii)** Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

  Accept x and go to 30.

[*31] Here, the interval of a floating point number in $\left[2^{n-1}, b\right]$ is twice as wide as $\left[a, 2^{n-1}\right]$.

(Judge) Case: $x = q < 2^k$.

(i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

Reject x and go back to 10.

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

Generate a random bit by $URNG_1\,()$. If the bit is 0, then reject x and go back to 10. Otherwise, accept x and go to 30.

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

Accept x and go to 30.

(Accept) Case: $x = p = -2^k$ or $p < x < q$ or $x = q = 2^k$.

Accept x and go to 40.

40: Output the result.

- Case: $I = \left[-2^k, 0\right]$.
  Output

$$x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-2} + 2^{n-1}.$$

- Case: $I = \left[0, 2^k\right]$.
  Output

$$x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}.$$

(5) Case: $0 \leq a < 2^{n-2} < 2^{n-1} < b \leq 2^n$ where $(2 - (2^{E-1} - 1) \leq n \in \mathbb{N} \leq 2^{E-1})$.

In this case, we can generate a floating point uniform random number in $[a, b]$ by using acceptance-rejection method. The following is the concrete pseudocode.

10: Generate a uniform random number.

Generate a floating point uniform random number $x \in \mathbb{F}$ in $[0, 2^n]$ by using the modified algorithm in the Section 5.1.

$$x = FURNG\left(0, 2^n, round_{\mathbb{F}}\right).$$

20: Judge Acceptance or Rejection.

- **(Reject)** Case: $x < a$ or $b < x$.
  Reject x and go back to 10.
- **(Judge)** Case: $x = a > 0$.
  - **(i)** Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
    Reject x and go back to 10.
  - **(ii)** Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
    - Case: The mantissa of a is not 0.
      Generate a random bit by $URNG_1()$. If the bit is 0, then reject x and go back to 10. Otherwise, accept x and go to 30.
    - Case: The mantissa of a is 0.
      If $a = 2^{1-(2^{E-1}-1)}$, then the operation is the same as the case where the mantissa of a is not 0. Otherwise, generate 2 random bits by $URNG_2()$. If the bits is 00, then reject x and go back to 10. If the bits is 01 or 10, then accept x and go to 30. If the bits is 11, then generate 2 random bits and judge again.
  - **(iii)** Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
    Accept x and go to 30.
- **(Judge)** Case: $x = b < 2^n$.
  - **(i)** Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
    Reject x and go back to 10.
  - **(ii)** Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
    Generate a random bit by $URNG_1()$. If the bit is 0, then reject x and go back to 10. Otherwise, accept x and go to 30.
  - **(iii)** Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
    Accept x and go to 30.
- **(Accept)** Case: $x = a = 0$ or $a < x < b$ or $x = b = 2^n$.
  Accept x and go to 30.

30: Output the result.

Output x.

6.2 Acceptance ratio

This section calculates the acceptance ratio γ of $FURNG(a, b, round_{\mathbb{F}})$ in the Section 7.1 according to $a, b, round_{\mathbb{F}}$. Of course, we have $0 \leq \gamma \leq 1$.

(1) Case: $0 \leq a < b \leq 2^{2-(2^{E-1}-1)}$.

- Case: $b - a = 2^k$.
  In this case, $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $0 \leq x \leq b - a$ at the line 10 in the pseudocode so that x is accepted. Here, since $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ satisfies the Formula (1), we obtain

$$\begin{aligned}
\gamma &= \sum_{0 \le x \le b-a} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-0} dt \\
&= \sum_{0 \le x \le 2^k} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-0} dt \\
&= \int_{\{t \in [0,2^k] | 0 \le round_{\mathbb{F}}(t) \le 2^k\}} \frac{1}{2^k-0} dt \\
&= \int_0^{2^k} \frac{1}{2^k-0} dt \\
&= \frac{2^k}{2^k} \\
&= \frac{b-a}{2^k}.
\end{aligned}$$

- Case: $b-a < 2^k$.

(i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

In this case, $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $0 \le x < b-a$ at the line 10 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{0 \le x < b-a} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-0} dt \\
&= \int_{\{t \in [0,2^k] | 0 \le round_{\mathbb{F}}(t) < b-a\}} \frac{1}{2^k-0} dt \\
&= \int_0^{b-a} \frac{1}{2^k} dt \\
&= \frac{b-a}{2^k}.
\end{aligned}$$

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

In this case, $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $0 \le x < b-a$ at the line 10 in the pseudocode, or $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b-a$ at the line 10 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 20 in the pseudocode so that x is accepted. Hence, by letting $(b-a)_l$ be the left adjacent floating point number to $(b-a)$ and $(b-a)_r$ be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \sum_{0 \le x < b-a} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-0} dt + \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t)=b-a\}} \frac{1}{2^k-0} dt \times \frac{1}{2} \\
&= \int_{\{t \in [0,2^k] | 0 \le round_{\mathbb{F}}(t) < b-a\}} \frac{1}{2^k-0} dt + \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t)=b-a\}} \frac{1}{2^k-0} dt \times \frac{1}{2} \\
&= \int_0^{\frac{(b-a)_l+(b-a)}{2}} \frac{1}{2^k-0} dt + \int_{\frac{(b-a)_l+(b-a)}{2}}^{\frac{(b-a)+(b-a)_r}{2}} \frac{1}{2^k-0} dt \times \frac{1}{2} \\
&= \frac{(b-a)_l + (b-a)}{2^{k+1}} + \frac{(b-a)_r - (b-a)_l}{2 \times 2^{k+1}} \\
&= \frac{1}{2^{k+1}} \left((b-a) + \frac{(b-a)_l + (b-a)_r}{2} \right).
\end{aligned}$$

Here, since $b-a \neq 2^k$, we have

$$\frac{(b-a)_l + (b-a)_r}{2} = (b-a).$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{k+1}} \left((b-a) + \frac{(b-a)_l + (b-a)_r}{2} \right) \\
&= \frac{2\,(b-a)}{2^{k+1}} \\
&= \frac{b-a}{2^k}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

In this case, $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $0 \leq x \leq b-a$ at the line 10 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}\gamma &= \sum_{0\leq x\leq b-a} \int_{\{t\in[0,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-0}dt \\ &= \int_{\{t\in[0,2^k]\mid 0\leq round_{\mathbb{F}}(t)\leq b-a\}} \frac{1}{2^k-0}dt \\ &= \int_0^{b-a} \frac{1}{2^k}dt \\ &= \frac{b-a}{2^k}.\end{aligned}$$

Therefore, we obtain

$$\gamma = \frac{b-a}{2^k}$$

in all the cases. Here, since $2^{k-1} < b-a \leq 2^k$, we have

$$\frac{1}{2} < \gamma \leq 1.$$

(2) Case: $2^{n-1} \leq a < b \leq 2^n$ where $(2-\left(2^{E-1}-1\right) \leq n \in \mathbb{N} \leq 2^E)$.
In this case, we can calculate γ by substituting a with p and b with q in (1). Therefore, we obtain

$$\gamma = \frac{q-p}{2^k}.$$

Here, since $2^{k-1} < q-p \leq 2^k$, we have

$$\frac{1}{2} < \gamma \leq 1.$$

(3) Case: $a < 0 < b$.
First, calculate the probability that x takes each floating point number at the line 20 in the pseudocode.

- Case: $+0 \in \mathbb{F} \leq x \leq 2^k$.
  In this case, $I = \left[0, 2^k\right]$ is selected at the line 10 in the pseudocode. Here, since $FURNG\left(0, 2^k, round_{\mathbb{F}}\right)$ satisfies the Formula (1), we obtain the probability as

$$\begin{aligned}\frac{1}{2} \times \int_{\{t\in[0,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-0}dt &= \int_{\{t\in[0,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^{k+1}-0}dt \\ &= \int_{\{t\in[0,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{t\in[-2^k,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt.\end{aligned}$$

- Case: $-2^k \leq x \leq -0 \in \mathbb{F}$.
  In this case, $I = \left[-2^k, 0\right]$ is selected at the line 10 in the pseudocode. Here, since $FURNG\left(0, 2^k, -round_{\mathbb{F}}\right)$ satisfies the Formula (1), we obtain the probability as

$$\begin{aligned}\frac{1}{2} \times \int_{\{t\in[0,2^k]\mid(-round_{\mathbb{F}})(t)=-x\}} \frac{1}{2^k-0}dt &= \int_{\{t\in[0,2^k]\mid(-round_{\mathbb{F}})(t)=-x\}} \frac{1}{2^{k+1}-0}dt \\ &= \int_{\{t\in[0,2^k]\mid(-round_{\mathbb{F}})(t)=-x\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{-t\in[0,2^k]\mid(-round_{\mathbb{F}})(-t)=-x\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{-t\in[0,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{t\in[-2^k,0]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{t\in[-2^k,2^k]\mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt.\end{aligned}$$

Therefore, the probability that x takes each floating point number at the line 20 in the pseudocode satisfies the Formula (1).

Next, calculate the acceptance ratio γ according to whether $a = -2^k$ and whether $b = 2^k$.

- Case: $a = -2^k$ and $b = 2^k$.
  In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x \leq b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a \leq x \leq b} \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&= \sum_{-2^k \leq x \leq 2^k} \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{\{t \in [-2^k, 2^k] \mid -2^k \leq round_{\mathbb{F}}(t) \leq 2^k\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{-2^k}^{2^k} \frac{1}{2^k - (-2^k)} dt \\
&= \frac{2^k - \left(-2^k\right)}{2^{k+1}} \\
&= \frac{b - a}{2^{k+1}}.
\end{aligned}$$

- Case: $a = -2^k$ and $b < 2^k$.

  (i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
  In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x < b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a \leq x < b} \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&= \sum_{-2^k \leq x < b} \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{\{t \in [-2^k, 2^k] \mid -2^k \leq round_{\mathbb{F}}(t) < b\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{-2^k}^{b} \frac{1}{2^k - (-2^k)} dt \\
&= \frac{b - \left(-2^k\right)}{2^{k+1}} \\
&= \frac{b - a}{2^{k+1}}.
\end{aligned}$$

  (ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

  – Case: The mantissa of b is not 0.
  In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x < b$ at the line 20 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 20 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 30 in the pseudocode so that x is accepted. Hence, by letting b_l be the left adjacent floating point number to b and b_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \sum_{a \le x < b} \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&\quad + \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = b\}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{2} \\
&= \sum_{-2^k \le x < b} \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&\quad + \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = b\}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{2} \\
&= \int_{\{t \in [-2^k, 2^k] | -2^k \le round_{\mathbb{F}}(t) < b\}} \frac{1}{2^k - (-2^k)} dt \\
&\quad + \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = b\}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{2} \\
&= \int_{-2^k}^{\frac{b_l + b}{2}} \frac{1}{2^k - (-2^k)} dt + \int_{\frac{b_l + b}{2}}^{\frac{b + b_r}{2}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{2} \\
&= \frac{b_l + b - 2 \times \left(-2^k\right)}{2^{k+2}} + \frac{b_r - b_l}{2 \times 2^{k+2}} \\
&= \frac{b_l + b - 2a}{2^{k+2}} + \frac{b_r - b_l}{2 \times 2^{k+2}} \\
&= \frac{1}{2^{k+2}} \left(b - 2a + \frac{b_l + b_r}{2} \right).
\end{aligned}$$

Here, since the mantissa of b is not 0, we have

$$\frac{b_l + b_r}{2} = b.$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{k+2}} \left(b - 2a + \frac{b_l + b_r}{2} \right) \\
&= \frac{2\,(b - a)}{2^{k+2}} \\
&= \frac{b - a}{2^{k+1}}.
\end{aligned}$$

– Case: The mantissa of b is 0.
Since we can calculate γ in the case where $b = 2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of b is not 0, we can consider only the case where $b \neq 2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \le x < b$ at the line 20 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 20 in the pseudocode and $URNG_2\left(\right)$ needs to generate 10 at the line 30 in the pseudocode so that x is accepted. Hence, by letting b_l be the left adjacent floating point number to b and b_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \sum_{a \le x < b} \int_{\{t \in [-2^k, 2^k] | round_{\mho}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&\quad + \int_{\{t \in [-2^k, 2^k] | round_{\mho}(t) = b\}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{3} \\
&= \sum_{-2^k \le x < b} \int_{\{t \in [-2^k, 2^k] | round_{\mho}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&\quad + \int_{\{t \in [-2^k, 2^k] | round_{\mho}(t) = b\}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{3} \\
&= \int_{\{t \in [-2^k, 2^k] | -2^k \le round_{\mho}(t) < b\}} \frac{1}{2^k - (-2^k)} dt \\
&\quad + \int_{\{t \in [-2^k, 2^k] | round_{\mho}(t) = b\}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{3} \\
&= \int_{-2^k}^{\frac{b_l + b}{2}} \frac{1}{2^k - (-2^k)} dt + \int_{\frac{b_l + b}{2}}^{\frac{b + b_r}{2}} \frac{1}{2^k - (-2^k)} dt \times \frac{1}{3} \\
&= \frac{b_l + b - 2 \times \left(-2^k\right)}{2^{k+2}} + \frac{b_r - b_l}{3 \times 2^{k+2}} \\
&= \frac{b_l + b - 2a}{2^{k+2}} + \frac{b_r - b_l}{3 \times 2^{k+2}} \\
&= \frac{1}{2^{k+2}} \left(b - 2a + \frac{2b_l + b_r}{3} \right)
\end{aligned}$$

Here, since the mantissa of b is 0, we have

$$\frac{2b_l + b_r}{3} = b$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{k+2}} \left(b - 2a + \frac{2b_l + b_r}{3} \right) \\
&= \frac{2(b - a)}{2^{k+2}} \\
&= \frac{b - a}{2^{k+1}}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \le x \le b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a \le x \le b} \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&= \sum_{-2^k \le x \le b} \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{\{t \in [-2^k, 2^k] | -2^k \le round_{\mathbb{F}}(t) \le b\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{-2^k}^{b} \frac{1}{2^k - (-2^k)} dt \\
&= \frac{b - \left(-2^k\right)}{2^{k+1}} \\
&= \frac{b - a}{2^{k+1}}.
\end{aligned}$$

- Case: $-2^k < a$ and $b = 2^k$.

  (i) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward 0.
  In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \le b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a<x\leq b} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
&= \sum_{a<x\leq 2^k} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
&= \int_{\{t\in[-2^k,2^k] | a<round_{\mathbb{F}}(t)\leq 2^k\}} \frac{1}{2^k-(-2^k)}dt \\
&= \int_a^{2^k} \frac{1}{2^k-(-2^k)}dt \\
&= \frac{2^k-a}{2^{k+1}} \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

- Case: The mantissa of a is not 0.

  In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 20 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 30 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \leq b$ at the line 20 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a and a_r be the right adjacent one, we have

$$\begin{aligned}
\gamma = &\int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \times \frac{1}{2} \\
&+ \sum_{a<x\leq b} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
&\int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \times \frac{1}{2} \\
&+ \sum_{a<x\leq 2^k} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
&\int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \times \frac{1}{2} \\
&+ \int_{\{t\in[-2^k,2^k] | a<round_{\mathbb{F}}(t)\leq 2^k\}} \frac{1}{2^k-(-2^k)}dt \\
&= \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}} \frac{1}{2^k-(-2^k)}dt \times \frac{1}{2} + \int_{\frac{a+a_r}{2}}^{2^k} \frac{1}{2^k-(-2^k)}dt \\
&= \frac{a_r-a_l}{2\times 2^{k+2}} + \frac{2\times 2^k-(a+a_r)}{2^{k+2}} \\
&= \frac{a_r-a_l}{2\times 2^{k+2}} + \frac{2b-(a+a_r)}{2^{k+2}} \\
&= \frac{1}{2^{k+2}}\left(2b-a-\frac{a_l+a_r}{2}\right)
\end{aligned}$$

  Here, since the mantissa of a is not 0, we have

$$\frac{a_l+a_r}{2} = a.$$

  Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{k+2}}\left(2b-a-\frac{a_l+a_r}{2}\right) \\
&= \frac{2\left(b-a\right)}{2^{k+2}} \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

- Case: The mantissa of a is 0.

  Since we can calculate γ in the case where $a = -2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a \neq -2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$

needs to generate a floating point uniform random number $x = a$ at the line 20 in the pseudocode and $URNG_2\,()$ needs to generate 10 at the line 30 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \leq b$ at the line 20 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a and a_r be the right adjacent one, we have

$$\begin{aligned}
\gamma = & \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \times \frac{2}{3} \\
& + \sum_{a<x\leq b} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
& \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \times \frac{2}{3} \\
& + \sum_{a<x\leq 2^k} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
& \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \times \frac{2}{3} \\
& + \int_{\{t\in[-2^k,2^k]|a<round_{\mathbb{F}}(t)\leq 2^k\}} \frac{1}{2^k-(-2^k)}dt \\
= & \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}} \frac{1}{2^k-(-2^k)}dt \times \frac{2}{3} + \int_{\frac{a+a_r}{2}}^{2^k} \frac{1}{2^k-(-2^k)}dt \\
= & \frac{2\,(a_r-a_l)}{3\times 2^{k+2}} + \frac{2\times 2^k-(a+a_r)}{2^{k+2}} \\
= & \frac{2\,(a_r-a_l)}{3\times 2^{k+2}} + \frac{2b-(a+a_r)}{2^{k+2}} \\
= & \frac{1}{2^{k+2}}\left(2b-a-\frac{2a_l+a_r}{3}\right).
\end{aligned}$$

Here, since the mantissa of a is 0, we have

$$\frac{2a_l+a_r}{3} = a.$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{k+2}}\left(2b-a-\frac{2a_l+a_r}{3}\right) \\
&= \frac{2\,(b-a)}{2^{k+2}} \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward $\pm\infty$.

In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x \leq b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a\leq x\leq b} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
&= \sum_{a\leq x\leq 2^k} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\
&= \int_{\{t\in[-2^k,2^k]|a\leq round_{\mathbb{F}}(t)\leq 2^k\}} \frac{1}{2^k-(-2^k)}dt \\
&= \int_a^{2^k} \frac{1}{2^k-(-2^k)}dt \\
&= \frac{2^k-a}{2^{k+1}} \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

- Case: $-2^k < a$ and $b < 2^k$.

(i) Case: $round_{\mathbb{F}}$ is Toward $+\infty$.
In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \leq b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a<x\leq b} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)} dt \\
&= \int_{\{t\in[-2^k,2^k]|a<round_{\mathbb{F}}(t)\leq b\}} \frac{1}{2^k-(-2^k)} dt \\
&= \int_a^b \frac{1}{2^k-(-2^k)} dt \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

(ii) Case: $round_{\mathbb{F}}$ is Toward 0.
In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a<x<b} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)} dt \\
&= \int_{\{t\in[-2^k,2^k]|a<round_{\mathbb{F}}(t)<b\}} \frac{1}{2^k-(-2^k)} dt \\
&= \int_a^b \frac{1}{2^k-(-2^k)} dt \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

– Case: The mantissa of a is not 0 and that of b is not 0.
In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 20 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 30 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 20 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 20 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 30 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a, a_r be the right adjacent one, b_l be the left adjacent floating point number to b, and b_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \times \frac{1}{2} \\
&\quad + \sum_{a<x<b} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)} dt \\
&\quad + \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=b\}} \frac{1}{2^k-(-2^k)} dt \times \frac{1}{2} \\
&= \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \times \frac{1}{2} \\
&\quad + \int_{\{t\in[-2^k,2^k]|a<round_{\mathbb{F}}(t)<b\}} \frac{1}{2^k-(-2^k)} dt \\
&\quad + \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=b\}} \frac{1}{2^k-(-2^k)} dt \times \frac{1}{2} \\
&= \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}} \frac{1}{2^k-(-2^k)} dt \times \frac{1}{2} + \int_{\frac{a+a_r}{2}}^{\frac{b_l+b}{2}} \frac{1}{2^k-(-2^k)} dt \\
&\quad + \int_{\frac{b_l+b}{2}}^{\frac{b+b_r}{2}} \frac{1}{2^k-(-2^k)} dt \times \frac{1}{2} \\
&= \frac{a_r-a_l}{2\times 2^{k+2}} + \frac{b-a+b_l-a_r}{2^{k+2}} + \frac{b_r-b_l}{2\times 2^{k+2}} \\
&= \frac{1}{2^{k+2}} \left(b-a+\frac{b_l+b_r}{2}-\frac{a_l+a_r}{2}\right).
\end{aligned}$$

Here, since the mantissa of a is not 0, we have

$$\frac{a_l + a_r}{2} = a.$$

Besides, since the mantissa of b is not 0, we have

$$\frac{b_l + b_r}{2} = b.$$

Thus, we obtain

$$\begin{aligned}\gamma &= \frac{1}{2^{k+2}} \left(b - a + \frac{b_l + b_r}{2} - \frac{a_l + a_r}{2} \right) \\ &= \frac{2\left(b - a\right)}{2^{k+2}} \\ &= \frac{b - a}{2^{k+1}}.\end{aligned}$$

– Case: The mantissa of a is not 0 and that of b is 0.
Since we can calculate γ in the case where $b = 2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of b is not 0, we can consider only the case where $b \neq 2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 20 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 30 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 20 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 20 in the pseudocode and $URNG_2\left(\right)$ needs to generate 10 at the line 30 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a, a_r be the right adjacent one, b_l be the left adjacent floating point number to b, and b_r be the right adjacent one, we have

$$\begin{aligned}\gamma &= \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = a\}} \frac{1}{2^k - \left(-2^k\right)} dt \times \frac{1}{2} \\ &\quad + \sum_{a < x < b} \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = x\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &\quad + \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = b\}} \frac{1}{2^k - \left(-2^k\right)} dt \times \frac{1}{3} \\ &= \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = a\}} \frac{1}{2^k - \left(-2^k\right)} dt \times \frac{1}{2} \\ &\quad + \int_{\{t \in [-2^k, 2^k] \mid a < round_{\mathbb{F}}(t) < b\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &\quad + \int_{\{t \in [-2^k, 2^k] \mid round_{\mathbb{F}}(t) = b\}} \frac{1}{2^k - \left(-2^k\right)} dt \times \frac{1}{3} \\ &= \int_{\frac{a_l + a}{2}}^{\frac{a + a_r}{2}} \frac{1}{2^k - \left(-2^k\right)} dt \times \frac{1}{2} + \int_{\frac{a + a_r}{2}}^{\frac{b_l + b}{2}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &\quad + \int_{\frac{b_l + b}{2}}^{\frac{b + b_r}{2}} \frac{1}{2^k - \left(-2^k\right)} dt \times \frac{1}{3} \\ &= \frac{a_r - a_l}{2 \times 2^{k+2}} + \frac{b - a + b_l - a_r}{2^{k+2}} + \frac{b_r - b_l}{3 \times 2^{k+2}} \\ &= \frac{1}{2^{k+2}} \left(b - a + \frac{2b_l + b_r}{3} - \frac{a_l + a_r}{2} \right).\end{aligned}$$

Here, since the mantissa of a is not 0, we have

$$\frac{a_l + a_r}{2} = a.$$

Besides, since the mantissa of b is 0, we have

$$\frac{2b_l + b_r}{3} = b.$$

Thus, we obtain

$$\begin{aligned}\gamma &= \frac{1}{2^{k+2}}\left(b-a+\frac{2b_l+b_r}{3}-\frac{a_l+a_r}{2}\right)\\ &= \frac{2\left(b-a\right)}{2^{k+2}}\\ &= \frac{b-a}{2^{k+1}}.\end{aligned}$$

– Case: The mantissa of a is 0 and that of b is not 0.
Since we can calculate γ in the case where $a = -2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a \neq -2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 20 in the pseudocode and $URNG_2\left(\right)$ needs to generate 01 or 10 at the line 30 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 20 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 20 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 30 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a, a_r be the right adjacent one, b_l be the left adjacent floating point number to b, and b_r be the right adjacent one, we have

$$\begin{aligned}\gamma &= \int_{\left\{t\in\left[-2^k,2^k\right]\mid round_{\mathbb{F}}(t)=a\right\}}\frac{1}{2^k-\left(-2^k\right)}dt\times\frac{2}{3}\\ &\quad+\sum_{a<x<b}\int_{\left\{t\in\left[-2^k,2^k\right]\mid round_{\mathbb{F}}(t)=x\right\}}\frac{1}{2^k-\left(-2^k\right)}dt\\ &\quad+\int_{\left\{t\in\left[-2^k,2^k\right]\mid round_{\mathbb{F}}(t)=b\right\}}\frac{1}{2^k-\left(-2^k\right)}dt\times\frac{1}{2}\\ &= \int_{\left\{t\in\left[-2^k,2^k\right]\mid round_{\mathbb{F}}(t)=a\right\}}\frac{1}{2^k-\left(-2^k\right)}dt\times\frac{2}{3}\\ &\quad+\int_{\left\{t\in\left[-2^k,2^k\right]\mid a<round_{\mathbb{F}}(t)<b\right\}}\frac{1}{2^k-\left(-2^k\right)}dt\\ &\quad+\int_{\left\{t\in\left[-2^k,2^k\right]\mid round_{\mathbb{F}}(t)=b\right\}}\frac{1}{2^k-\left(-2^k\right)}dt\times\frac{1}{2}\\ &= \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}}\frac{1}{2^k-\left(-2^k\right)}dt\times\frac{2}{3}+\int_{\frac{a+a_r}{2}}^{\frac{b_l+b}{2}}\frac{1}{2^k-\left(-2^k\right)}dt\\ &\quad+\int_{\frac{b_l+b}{2}}^{\frac{b+b_r}{2}}\frac{1}{2^k-\left(-2^k\right)}dt\times\frac{1}{2}\\ &= \frac{2\left(a_r-a_l\right)}{3\times2^{k+2}}+\frac{b-a+b_l-a_r}{2^{k+2}}+\frac{b_r-b_l}{2\times2^{k+2}}\\ &= \frac{1}{2^{k+2}}\left(b-a+\frac{b_l+b_r}{2}-\frac{2a_l+a_r}{3}\right).\end{aligned}$$

Here, since the mantissa of a is 0, we have

$$\frac{2a_l+a_r}{2}=a$$

Besides, since the mantissa of b is not 0, we have

$$\frac{b_l+b_r}{2}=b.$$

Thus, we obtain

$$\begin{aligned}\gamma &= \frac{1}{2^{k+2}}\left(b-a+\frac{2b_l+b_r}{3}-\frac{2a_l+a_r}{3}\right)\\ &= \frac{2\left(b-a\right)}{2^{k+2}}\\ &= \frac{b-a}{2^{k+1}}.\end{aligned}$$

– Case: The mantissa of a is 0 and that of b is 0.
Since we can calculate γ in the case where $a = -2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a \neq -2^{1-\left(2^{E-1}-1\right)}$. Besides, since we can calculate γ in the

case where $b = 2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of b is not 0, we can consider only the case where $b \neq 2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 20 in the pseudocode and $URNG_2\left(\right)$ needs to generate 01 or 10 at the line 30 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 20 in the pseudocode, or $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 20 in the pseudocode and $URNG_2\left(\right)$ needs to generate 10 at the line 30 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a, a_r be the right adjacent one, b_l be the left adjacent floating point number to b, and b_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \int_{\left\{t\in\left[-2^k,2^k\right] \mid round_{\mathbb{F}}(t)=a\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \times \frac{2}{3} \\
&\quad + \sum_{a<x<b} \int_{\left\{t\in\left[-2^k,2^k\right] \mid round_{\mathbb{F}}(t)=x\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \\
&\quad + \int_{\left\{t\in\left[-2^k,2^k\right] \mid round_{\mathbb{F}}(t)=b\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \times \frac{1}{3} \\
&= \int_{\left\{t\in\left[-2^k,2^k\right] \mid round_{\mathbb{F}}(t)=a\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \times \frac{2}{3} \\
&\quad + \int_{\left\{t\in\left[-2^k,2^k\right] \mid a<round_{\mathbb{F}}(t)<b\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \\
&\quad + \int_{\left\{t\in\left[-2^k,2^k\right] \mid round_{\mathbb{F}}(t)=b\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \times \frac{1}{3} \\
&= \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}} \frac{1}{2^k-\left(-2^k\right)}dt \times \frac{2}{3} + \int_{\frac{a+a_r}{2}}^{\frac{b_l+b}{2}} \frac{1}{2^k-\left(-2^k\right)}dt \\
&\quad + \int_{\frac{b_l+b}{2}}^{\frac{b+b_r}{2}} \frac{1}{2^k-\left(-2^k\right)}dt \times \frac{1}{3} \\
&= \frac{2\left(a_r-a_l\right)}{3\times 2^{k+2}} + \frac{b-a+b_l-a_r}{2^{k+2}} + \frac{b_r-b_l}{3\times 2^{k+2}} \\
&= \frac{1}{2^{k+2}}\left(b-a+\frac{2b_l+b_r}{3}-\frac{2a_l+a_r}{3}\right)
\end{aligned}$$

Here, since the mantissa of a is 0, we have

$$\frac{2a_l+a_r}{3} = a.$$

Besides, since the mantissa of b is 0, we have

$$\frac{2b_l+b_r}{3} = b.$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{k+2}}\left(b-a+\frac{2b_l+b_r}{3}-\frac{2a_l+a_r}{3}\right) \\
&= \frac{2\left(b-a\right)}{2^{k+2}} \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

(iv) Case: $round_{\mathbb{F}}$ is Toward $-\infty$.

In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x < b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}
\gamma &= \sum_{a\leq x<b} \int_{\left\{t\in\left[-2^k,2^k\right] \mid round_{\mathbb{F}}(t)=x\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \\
&= \int_{\left\{t\in\left[-2^k,2^k\right] \mid a\leq round_{\mathbb{F}}(t)<b\right\}} \frac{1}{2^k-\left(-2^k\right)}dt \\
&= \int_a^b \frac{1}{2^k-\left(-2^k\right)}dt \\
&= \frac{b-a}{2^{k+1}}.
\end{aligned}$$

(v) Case: $round_{\mathbb{F}}$ is Toward $\pm\infty$.
In this case, $FURNG\left(-2^k, 2^k, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x \leq b$ at the line 20 in the pseudocode so that x is accepted. Hence, we obtain

$$\begin{aligned}\gamma &= \sum_{a\leq x\leq b} \int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=x\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{t\in[-2^k,2^k] \mid a\leq round_{\mathbb{F}}(t)\leq b\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_a^b \frac{1}{2^k-(-2^k)}dt \\ &= \frac{b-a}{2^{k+1}}.\end{aligned}$$

Therefore, we obtain

$$\gamma = \frac{b-a}{2^{k+1}}$$

in all the cases. Here, since $2^{k-1} < \max(-a, b) \leq 2^k$ and $a < 0 < b$, we have

$$\begin{aligned}2^{k-1} &< \max(-a, b) \\ &\leq b + (-a) = b - a \\ &\leq \max(-a, b) + \max(-a, b) \\ &\leq 2^{k+1}.\end{aligned}$$

Therefore, we obtain

$$\frac{1}{4} < \gamma \leq 1.$$

(4) Case: $2^{n-2} \leq a < 2^{n-1} < b \leq 2^n$ where $(2 - \left(2^{E-1} - 1\right) \leq n \in \mathbb{N} \leq 2^{E-1})$.
In this case, we can calculate γ by the similar way as (3). Then, we obtain

$$\begin{aligned}\gamma &= \frac{1}{3} \times \frac{-p}{2^k} + \frac{2}{3} \times \frac{q}{2^k} \\ &= \frac{2q-p}{3\times 2^k}\end{aligned}$$

Here, since $2^{k-1} < \max(-p, q) \leq 2^k$ and $p < 0 < q$, we have

$$\begin{aligned}2^{k-1} &< \max(-p, q) \\ &< q + \max(-p, q) \\ &\leq q + (q + (-p)) = 2q - p \\ &\leq 2 \times \max(-p, q) + \max(-p, q) \\ &\leq 3 \times 2^k.\end{aligned}$$

Therefore, we obtain

$$\frac{1}{6} < \gamma \leq 1.$$

(5) Case: $0 \leq a < 2^{n-2} < 2^{n-1} < b \leq 2^n$ where $(2 - \left(2^{E-1} - 1\right) \leq n \in \mathbb{N} \leq 2^{E-1})$.

- Case: $a = 0$ and $b = 2^n$.
  In this case, $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \leq x \leq b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}
\gamma &= \sum_{a \le x \le b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^n - 0} dt \\
&= \sum_{0 \le x \le 2^n} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^n - 0} dt \\
&= \int_{\{t \in [0,2^n] | 0 \le round_{\mathbb{F}}(t) \le 2^n\}} \frac{1}{2^n - 0} dt \\
&= \int_0^{2^n} \frac{1}{2^n - 0} dt \\
&= \frac{2^n - 0}{2^n} \\
&= \frac{b-a}{2^n}.
\end{aligned}$$

- Case: $a = 0$ and $b < 2^n$.

(i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
In this case, $FURNG\,(0, 2^n, round_{\mathbb{F}})$ needs to generate a floating point uniform random number x that satisfies $a \le x < b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}
\gamma &= \sum_{a \le x < b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^n - 0} dt \\
&= \sum_{0 \le x < b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^n - 0} dt \\
&= \int_{\{t \in [0,2^n] | 0 \le round_{\mathbb{F}}(t) < b\}} \frac{1}{2^n - 0} dt \\
&= \int_0^{b} \frac{1}{2^n} dt \\
&= \frac{b - 0}{2^n} \\
&= \frac{b-a}{2^n}.
\end{aligned}$$

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
In this case, $FURNG\,(0, 2^n, round_{\mathbb{F}})$ needs to generate a floating point uniform random number x that satisfies $a \le x < b$ at the line 10 in the pseudocode, or $FURNG\,(0, 2^n, round_{\mathbb{F}})$ needs to generate a floating point uniform random number $x = b$ at the line 10 in the pseudocode and $URNG_1\,()$ needs to generate 1 at the line 20 in the pseudocode so that x is accepted. Hence, by letting b_l be the left adjacent floating point number to b and b_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \sum_{a \le x < b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n - 0} dt \\
&\quad + \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=b\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \sum_{0 \le x < b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n - 0} dt \\
&\quad + \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=b\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \int_{\{t \in [0,2^n] | 0 \le round_{\mathbb{F}}(t) < b\}} \frac{1}{2^n - 0} dt \\
&\quad + \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=b\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \int_0^{\frac{b_l+b}{2}} \frac{1}{2^n - 0} dt + \int_{\frac{b_l+b}{2}}^{\frac{b+b_r}{2}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \frac{b_l + b - 0}{2^{n+1}} + \frac{b_r - b_l}{2 \times 2^{k+1}} \\
&= \frac{b_l + b - a}{2^{n+1}} + \frac{b_r - b_l}{2 \times 2^{k+1}} \\
&= \frac{1}{2^{n+1}} \left(b - a + \frac{b_l + b_r}{2} \right).
\end{aligned}$$

Here, since the mantissa of b is not 0, we have

$$\frac{b_l + b_r}{2} = b.$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{n+1}} \left(b - a + \frac{b_l + b_r}{2} \right) \\
&= \frac{2(b-a)}{2^{n+1}} \\
&= \frac{b-a}{2^n}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

In this case, $FURNG(0, 2^n, round_{\mathbb{F}})$ needs to generate a floating point uniform random number x that satisfies $a \le x \le b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}
\gamma &= \sum_{a \le x \le b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n - 0} dt \\
&= \sum_{0 \le x \le b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n - 0} dt \\
&= \int_{\{t \in [0,2^n] | 0 \le round_{\mathbb{F}}(t) \le b\}} \frac{1}{2^n - 0} dt \\
&= \int_0^b \frac{1}{2^n} dt \\
&= \frac{b-0}{2^n} \\
&= \frac{b-a}{2^n}.
\end{aligned}$$

- Case: $0 < a$ and $b = 2^n$.

  (i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

  In this case, $FURNG(0, 2^n, round_{\mathbb{F}})$ needs to generate a floating point uniform random number x that satisfies $a \le x \le b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}\gamma &= \sum_{a\le x\le b}\int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}}\frac{1}{2^n-0}dt\\ &= \sum_{a\le x\le 2^n}\int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}}\frac{1}{2^n-0}dt\\ &= \int_{\{t\in[0,2^n]|a\le round_{\mathbb{F}}(t)\le 2^n\}}\frac{1}{2^n-0}dt\\ &= \int_a^{2^n}\frac{1}{2^n}dt\\ &= \frac{2^n-a}{2^n}\\ &= \frac{b-a}{2^n}.\end{aligned}$$

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

- Case: The mantissa of a is not 0.

  In this case, $FURNG\,(0,2^n,round_{\mathbb{F}})$ needs to generate a floating point uniform random number $x=a$ at the line 10 in the pseudocode and $URNG_1\,()$ needs to generate 1 at the line 20 in the pseudocode, or $FURNG\,(0,2^n,round_{\mathbb{F}})$ needs to generate a floating point uniform random number x that satisfies $a<x\le b$ at the line 10 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a and a_r be the right adjacent one, we have

$$\begin{aligned}\gamma &= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}}\frac{1}{2^n-0}dt\times\frac{1}{2}\\ &\quad+\sum_{a<x\le b}\int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}}\frac{1}{2^n-0}dt\\ &= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}}\frac{1}{2^n-0}dt\times\frac{1}{2}\\ &\quad+\sum_{a<x\le 2^n}\int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}}\frac{1}{2^n-0}dt\\ &= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}}\frac{1}{2^n-0}dt\times\frac{1}{2}\\ &\quad+\int_{\{t\in[0,2^n]|a<round_{\mathbb{F}}(t)\le 2^n\}}\frac{1}{2^n-0}dt\\ &= \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}}\frac{1}{2^n-0}dt\times\frac{1}{2}+\int_{\frac{a+a_r}{2}}^{2^n}\frac{1}{2^n-0}dt\\ &= \frac{a_r-a_l}{2\times 2^{n+1}}+\frac{2\times 2^n-(a+a_r)}{2^{n+1}}\\ &= \frac{a_r-a_l}{2\times 2^{n+1}}+\frac{2b-(a+a_r)}{2^{n+1}}\\ &= \frac{1}{2^{n+1}}\left(2b-a-\frac{a_l+a_r}{2}\right).\end{aligned}$$

  Here, since the mantissa of a is not 0, we have

$$\frac{a_l+a_r}{2}=a.$$

  Thus, we obtain

$$\begin{aligned}\gamma &= \frac{1}{2^{n+1}}\left(2b-a-\frac{a_l+a_r}{2}\right)\\ &= \frac{2\,(b-a)}{2^{n+1}}\\ &= \frac{b-a}{2^n}.\end{aligned}$$

- Case: The mantissa of a is 0.

  Since we can calculate γ in the case where $a=2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a\ne 2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\,(0,2^n,round_{\mathbb{F}})$ needs to generate a floating point uniform random number $x=a$ at the line 10 in the pseudocode and $URNG_2\,()$

needs to generate 01 or 10 at the line 20 in the pseudocode, or $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \leq b$ at the line 10 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a and a_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^n-0}dt \times \frac{2}{3} \\
&\quad + \sum_{a<x\leq b} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n-0}dt \\
&= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^n-0}dt \times \frac{2}{3} \\
&\quad + \sum_{a<x\leq 2^n} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n-0}dt \\
&= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^n-0}dt \times \frac{2}{3} \\
&\quad + \int_{\{t\in[0,2^n]|a<round_{\mathbb{F}}(t)\leq 2^n\}} \frac{1}{2^n-0}dt \\
&= \int_{\frac{a_l+a}{2}}^{\frac{a+a_r}{2}} \frac{1}{2^n-0}dt \times \frac{2}{3} + \int_{\frac{a+a_r}{2}}^{2^n} \frac{1}{2^n-0}dt \\
&= \frac{2\left(a_r-a_l\right)}{3\times 2^{n+1}} + \frac{2\times 2^n-\left(a+a_r\right)}{2^{n+1}} \\
&= \frac{2\left(a_r-a_l\right)}{3\times 2^{n+1}} + \frac{2b-\left(a+a_r\right)}{2^{n+1}} \\
&= \frac{1}{2^{n+1}}\left(2b-a-\frac{2a_l+a_r}{3}\right).
\end{aligned}$$

Here, since the mantissa of a is 0, we have

$$\frac{2a_l+a_r}{3} = a.$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{n+1}}\left(2b-a-\frac{2a_l+a_r}{3}\right) \\
&= \frac{2\left(b-a\right)}{2^{n+1}} \\
&= \frac{b-a}{2^n}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

In this case, $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \leq b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}
\gamma &= \sum_{a<x\leq b} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n-0}dt \\
&= \sum_{a<x\leq 2^n} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n-0}dt \\
&= \int_{\{t\in[0,2^n]|a<round_{\mathbb{F}}(t)\leq 2^n\}} \frac{1}{2^n-0}dt \\
&= \int_{a}^{2^n} \frac{1}{2^n}dt \\
&= \frac{2^n-a}{2^n} \\
&= \frac{b-a}{2^n}.
\end{aligned}$$

- Case: $0 < a$ and $b < 2^n$.

  (i) Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

In this case, $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a \le x < b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}\gamma &= \sum_{a \le x < b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^n - 0} dt \\ &= \int_{\{t \in [0,2^n] | a \le round_{\mathbb{F}}(t) < b\}} \frac{1}{2^n - 0} dt \\ &= \int_a^b \frac{1}{2^n} dt \\ &= \frac{b-a}{2^n}.\end{aligned}$$

(ii) Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

– Case: The mantissa of a is not 0.

In this case, $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 10 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 20 in the pseudocode, or $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 10 in the pseudocode, or $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 10 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 20 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a, a_r be the right adjacent one, b_l be the left adjacent floating point number to b, and b_r be the right adjacent one, we have

$$\begin{aligned}\gamma &= \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\ &\quad + \sum_{a < x < b} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = x\}} \frac{1}{2^n - 0} dt \\ &\quad + \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = b\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\ &= \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\ &\quad + \int_{\{t \in [0,2^n] | a < round_{\mathbb{F}}(t) < b\}} \frac{1}{2^n - 0} dt \\ &\quad + \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = b\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\ &= \int_{\frac{a_l + a}{2}}^{\frac{a + a_r}{2}} \frac{1}{2^n - 0} dt \times \frac{1}{2} + \int_{\frac{a + a_r}{2}}^{\frac{b_l + b}{2}} \frac{1}{2^n - 0} dt \\ &\quad + \int_{\frac{b_l + b}{2}}^{\frac{b + b_r}{2}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\ &= \frac{a_r - a_l}{2 \times 2^{n+1}} + \frac{b - a + b_l - a_r}{2^{n+1}} + \frac{b_r - b_l}{2 \times 2^{n+1}} \\ &= \frac{1}{2^{n+1}} \left(b - a + \frac{b_l + b_r}{2} - \frac{a_l + a_r}{2}\right).\end{aligned}$$

Here, since the mantissa of a not is 0, we have

$$\frac{a_l + a_r}{2} = a.$$

Besides, since the mantissa of b is not 0, we have

$$\frac{b_l + b_r}{2} = b.$$

Thus, we obtain

$$\begin{aligned}\gamma &= \frac{1}{2^{n+1}} \left(b - a + \frac{b_l + b_r}{2} - \frac{a_l + a_r}{2}\right) \\ &= \frac{2\left(b - a\right)}{2^{n+1}} \\ &= \frac{b-a}{2^n}.\end{aligned}$$

– Case: The mantissa of a is 0.
Since we can calculate γ in the case where $a = 2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a \neq 2^{1-\left(2^{E-1}-1\right)}$. In this case, $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = a$ at the line 10 in the pseudocode and $URNG_2\left(\right)$ needs to generate 01 or 10 at the line 20 in the pseudocode, or $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x < b$ at the line 10 in the pseudocode, or $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number $x = b$ at the line 10 in the pseudocode and $URNG_1\left(\right)$ needs to generate 1 at the line 20 in the pseudocode so that x is accepted. Hence, by letting a_l be the left adjacent floating point number to a, a_r be the right adjacent one, b_l be the left adjacent floating point number to b, and b_r be the right adjacent one, we have

$$\begin{aligned}
\gamma &= \int_{\left\{t \in [0,2^n] \mid round_{\mathbb{F}}(t) = a\right\}} \frac{1}{2^n - 0} dt \times \frac{2}{3} \\
&\quad + \sum_{a < x < b} \int_{\left\{t \in [0,2^n] \mid round_{\mathbb{F}}(t) = x\right\}} \frac{1}{2^n - 0} dt \\
&\quad + \int_{\left\{t \in [0,2^n] \mid round_{\mathbb{F}}(t) = b\right\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \int_{\left\{t \in [0,2^n] \mid round_{\mathbb{F}}(t) = a\right\}} \frac{1}{2^n - 0} dt \times \frac{2}{3} \\
&\quad + \int_{\left\{t \in [0,2^n] \mid a < round_{\mathbb{F}}(t) < b\right\}} \frac{1}{2^n - 0} dt \\
&\quad + \int_{\left\{t \in [0,2^n] \mid round_{\mathbb{F}}(t) = b\right\}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \int_{\frac{a_l + a}{2}}^{\frac{a + a_r}{2}} \frac{1}{2^n - 0} dt \times \frac{2}{3} + \int_{\frac{a + a_r}{2}}^{\frac{b_l + b}{2}} \frac{1}{2^n - 0} dt \\
&\quad + \int_{\frac{b_l + b}{2}}^{\frac{b + b_r}{2}} \frac{1}{2^n - 0} dt \times \frac{1}{2} \\
&= \frac{2\left(a_r - a_l\right)}{3 \times 2^{n+1}} + \frac{b - a + b_l - a_r}{2^{n+1}} \\
&\quad + \frac{b_r - b_l}{2 \times 2^{n+1}} \\
&= \frac{1}{2^{n+1}}\left(b - a + \frac{b_l + b_r}{2} - \frac{2a_l + a_r}{3}\right).
\end{aligned}$$

Here, since the mantissa of a is 0, we have

$$\frac{2a_l + a_r}{3} = a.$$

Besides, since the mantissa of b is not 0, we have

$$\frac{b_l + b_r}{2} = b.$$

Thus, we obtain

$$\begin{aligned}
\gamma &= \frac{1}{2^{n+1}}\left(b - a + \frac{b_l + b_r}{2} - \frac{2a_l + a_r}{3}\right) \\
&= \frac{2\left(b - a\right)}{2^{n+1}} \\
&= \frac{b - a}{2^n}.
\end{aligned}$$

(iii) Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
In this case, $FURNG\left(0, 2^n, round_{\mathbb{F}}\right)$ needs to generate a floating point uniform random number x that satisfies $a < x \leq b$ at the line 10 in the pseudocode so that x is accepted. Hence, we have

$$\begin{aligned}\gamma &= \sum_{a<x\le b} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=x\}} \frac{1}{2^n-0}dt \\ &= \int_{\{t\in[0,2^n]|a<round_{\mathbb{F}}(t)\le b\}} \frac{1}{2^n-0}dt \\ &= \int_a^b \frac{1}{2^n}dt \\ &= \frac{b-a}{2^n}.\end{aligned}$$

Therefore, we obtain

$$\gamma = \frac{b-a}{2^n}$$

in all the cases. Here, since $0 \le a < 2^{n-2} < 2^{n-1} < b \le 2^n$, we have

$$2^{n-1} - 2^{n-2} < b - a \le 2^n - 0.$$

Therefore, we obtain

$$\frac{1}{4} < \gamma \le 1.$$

In summary, we obtain the following result.

(1) Case: $0 \le a < b \le 2^{2-\left(2^{E-1}-1\right)}$.

$$\frac{1}{2} < \gamma = \frac{b-a}{2^k} \le 1$$

for the minimal $k \in \mathbb{N}$ that satisfies $b - a \le 2^k$.

(2) Case: $2^{n-1} \le a < b \le 2^n$ where $(2 - \left(2^{E-1} - 1\right) \le n \in \mathbb{N} \le 2^{E-1})$.

$$\frac{1}{2} < \gamma = \frac{q-p}{2^k} \le 1$$

for the minimal $k \in \mathbb{N}$ that satisfies $q - p \le 2^k$ where

$$\begin{aligned}p &= \left(a - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \\ q &= \left(a - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}.\end{aligned}$$

(3) Case: $a < 0 < b$.

$$\frac{1}{4} < \gamma = \frac{b-a}{2^{k+1}} \le 1$$

for the minimal $k \in \mathbb{N}$ that satisfies $\max(-a, b) \le 2^k$.

(4) Case: $2^{n-2} \le a < 2^{n-1} < b \le 2^n$ where $(2 - \left(2^{E-1} - 1\right) \le n \in \mathbb{N} \le 2^{E-1})$.

$$\frac{1}{6} < \gamma = \frac{2q-p}{3 \times 2^k} \le 1$$

for the minimal $k \in \mathbb{N}$ that satisfies $q - p \le 2^k$ where

$$\begin{aligned}p &= \left(a - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \\ q &= \left(b - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}.\end{aligned}$$

(5) Case: $0 \le a < 2^{n-2} < 2^{n-1} < b \le 2^n$ where $(2 - \left(2^{E-1} - 1\right) \le n \in \mathbb{N} \le 2^{E-1})$.

$$\frac{1}{4} < \gamma = \frac{b-a}{2^n} \le 1.$$

6.3 Proof for correctness

This section proves that the random number generation probability of the proposed algorithm in the Section 7.1 satisfies the Formula 1. First, calculate the probability that the algorithm outputs $f \in \mathbb{F}$, $P(f)$, for each floating point number. Next, compare $P(f)$ with

$$\begin{aligned} P_{\mathbb{F}}(f) &= Pr\left[round_{\mathbb{F}}\left(URNG_{\mathbb{R}}()\right) = f\right] \\ &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt, \end{aligned}$$

which is calculated by the Formula (1), and confirm that $P(f) = P_{\mathbb{F}}(f)$ holds. In the proof, let

$$Pr\left[\text{"constraint of variables" in "line number in the pseudocode"}\right]$$

be the probability that the constraint is satisfied at the end of the line in the pseudocode.

(1) Case: $0 \leq a < b \leq 2^{1-\left(2^{E-1}-1\right)}$.

(i) Case: $f < a$ or $b < f$.

Since $f - a < 0$ or $b - a < f - a$, we have

$$\begin{aligned} P(f) &= Pr\left[x = f - a \text{ in } 30\right] \\ &= Pr\left[x = f - a \text{ in } 20\right] \\ &= 0. \end{aligned}$$

Next, we have

$$\{t \in [a,b]\,|round_{\mathbb{F}}(t) = f\} = \emptyset$$

because $f < a$ or $b < f$. Hence, we obtain

$$\begin{aligned} P_{\mathbb{F}}(f) &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt \\ &= \int_{\emptyset} \frac{1}{b-a}dt \\ &= 0. \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(ii) Case: $a \leq f < b$.

In this case, since $f - a \in \mathbb{F} < b - a \leq 2^k$, we have

$$\begin{aligned} &\left\{t \in \left[0, 2^k\right] |round_{\mathbb{F}}(t) = f - a\right\} \\ &\quad = \{t \in [0, b-a]\,|round_{\mathbb{F}}(t) = f - a\}. \end{aligned}$$

Besides, since $a, f - a, b$ is a subnormal number, we have

$$\begin{aligned} &\sup\{t \in [0, b-a]\,|round_{\mathbb{F}}(t) = f - a\} - \inf\{t \in [0, b-a]\,|round_{\mathbb{F}}(t) = f - a\} \\ &\quad = \sup\{t \in [a,b]\,|round_{\mathbb{F}}(t) = f\} - \inf\{t \in [a,b]\,|round_{\mathbb{F}}(t) = f\}. \end{aligned}$$

Hence, we obtain

$$\begin{aligned}
P(f) &= Pr\left[x = f - a \text{ in } 30\right] \\
&= Pr\left[x = f - a \text{ in } 20\right] \\
&= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=f-a\}} \frac{1}{2^k - 0} dt \\
&= \frac{1}{\gamma} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=f-a\}} \frac{1}{2^k - 0} dt \\
&= \frac{2^k}{b-a} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=f-a\}} \frac{1}{2^k - 0} dt \\
&= \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=f-a\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[0,b-a] \mid round_{\mathbb{F}}(t)=f-a\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[a,b] \mid round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(iii) Case: $f = b$.

- Case: $b - a = 2^k$.

  In this case, we have

$$\begin{aligned}
P(f) &= P(b) \\
&= Pr\left[x = b - a \text{ in } 30\right] \\
&= Pr\left[x = 2^k \text{ in } 30\right] \\
&= Pr\left[x = 2^k \text{ in } 20\right] \\
&= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k - 0} dt \\
&= \frac{1}{\gamma} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k - 0} dt \\
&= \frac{2^k}{b-a} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k - 0} dt \\
&= \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[0,b-a] \mid round_{\mathbb{F}}(t)=b-a\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[a,b] \mid round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(b) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

  Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $b - a < 2^k$.
  - Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

    In this case, we have

$$\begin{aligned}
P(f) &= P(b) \\
&= Pr\left[x = b - a \text{ in } 30\right] \\
&= Pr\left[x = b - a \text{ in } 20\right] \\
&= 0.
\end{aligned}$$

    Besides, we have

$$\begin{aligned} P_{\mathbb{F}}(f) &= P_{\mathbb{F}}(b) \\ &= 0. \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

  Let $(b-a)_l$ be the left adjacent floating point number to $(b-a)$ and $(b-a)_r$ be the right adjacent one. Since $(b-a)$ is a subnormal number, we have

$$(b-a) - (b-a)_l = (b-a)_r - (b-a).$$

  Here, $(b-a)_r \leq 2^k$ because $(b-a) < 2^k$. Hence, we have

$$\begin{aligned} &\sup\left\{t \in \left[0, 2^k\right] | round_{\mathbb{F}}(t) = b-a\right\} - \inf\left\{t \in \left[0, 2^k\right] | round_{\mathbb{F}}(t) = b-a\right\} \\ &= \frac{(b-a)_r + (b-a)}{2} - \frac{(b-a) + (b-a)_l}{2} \\ &= \frac{(b-a)_r - (b-a)}{2} + \frac{(b-a) - (b-a)_l}{2} \\ &= \frac{(b-a) - (b-a)_l}{2} \times 2 \\ &= \left((b-a) - \frac{(b-a)_l + (b-a)}{2}\right) \times 2 \\ &= (\sup\{t \in [0, b-a] | round_{\mathbb{F}}(t) = b-a\} \\ &\quad - \inf\{t \in [0, b-a] | round_{\mathbb{F}}(t) = b-a\}) \times 2. \end{aligned}$$

  Thus, we obtain

$$\begin{aligned} P(f) &= P(b) \\ &= Pr[x = b-a \text{ in } 30] \\ &= Pr[x = b-a \text{ in } 20] \times \frac{1}{2} \\ &= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{2} \int_{\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = b-a\}} \frac{1}{2^k - 0} dt \\ &= \frac{1}{2\gamma} \int_{\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = b-a\}} \frac{1}{2^k - 0} dt \\ &= \frac{2^k}{2(b-a)} \int_{\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = b-a\}} \frac{1}{2^k - 0} dt \\ &= \frac{1}{2} \int_{\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = b-a\}} \frac{1}{b-a} dt \\ &= \int_{\{t \in [0, b-a] | round_{\mathbb{F}}(t) = b-a\}} \frac{1}{b-a} dt \\ &= \int_{\{t \in [a, b] | round_{\mathbb{F}}(t) = b\}} \frac{1}{b-a} dt \\ &= P_{\mathbb{F}}(b) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

  Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

  In this case, we have

$$\begin{aligned}
P(f) &= P(b)\\
&= Pr\left[x = b-a \text{ in } 30\right]\\
&= Pr\left[x = b-a \text{ in } 20\right]\\
&= \sum_{l=0}^{\infty}(1-\gamma)^l \int_{\{t\in[0,2^k]\,|\,round_{\mathbb{F}}(t)=b-a\}} \frac{1}{2^k-0}dt\\
&= \frac{1}{\gamma}\int_{\{t\in[0,2^k]\,|\,round_{\mathbb{F}}(t)=b-a\}} \frac{1}{2^k-0}dt\\
&= \frac{2^k}{b-a}\int_{\{t\in[0,2^k]\,|\,round_{\mathbb{F}}(t)=b-a\}} \frac{1}{2^k-0}dt\\
&= \int_{\{t\in[0,2^k]\,|\,round_{\mathbb{F}}(t)=b-a\}} \frac{1}{b-a}dt\\
&= \int_{\{t\in[0,b-a]\,|\,round_{\mathbb{F}}(t)=b-a\}} \frac{1}{b-a}dt\\
&= \int_{\{t\in[a,b]\,|\,round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a}dt\\
&= P_{\mathbb{F}}(b)\\
&= P_{\mathbb{F}}(f)\,.
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds in all the cases.

(2) Case: $2^{n-1} \le a < b \le 2^n$ where $(2-\left(2^{E-1}-1\right) \le n \in \mathbb{N} \le 2^{E-1})$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}$, we have

$$\begin{aligned}
P(f) &= P\left(x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}\right)\\
&= \int_{\{t\in[p,q]\,|\,round_{\mathbb{F}}(t)=x\}} \frac{1}{q-p}dt\\
&= \int_{\{t\in[p,q]\,|\,round_{\mathbb{F}}(t)=x\}} \frac{1}{b-a}dt \times \frac{1}{2^{-(n-1)} \times 2^{1-(2^{E-1}-1)}}.
\end{aligned}$$

Here, since all the floating point numbers in $[p,q]$ is a subnormal number, the interval between each floating point number is the same. Besides, the interval of floating point numbers in

$$\left[p \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}, q \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}\right]$$

is also the same as each other and is $\frac{1}{2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1}}$ times as wide as $[p,q]$. Hence, by letting $\alpha = 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1}$, we have

$$\begin{aligned}
\int_{\{t\in[p,q]\,|\,round_{\mathbb{F}}(t)=x\}} dt &= \frac{1}{\alpha}\int_{\{t\in[p\times\alpha+2^{n-1},q\times\alpha+2^{n-1}]\,|\,round_{\mathbb{F}}(t)=x\times\alpha+2^{n-1}\}} dt\\
&= \frac{1}{\alpha}\int_{\{t\in[a,b]\,|\,round_{\mathbb{F}}(t)=f\}} dt.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
P(f) &= \int_{\{t\in[p,q]\,|\,round_{\mathbb{F}}(t)=x\}} \frac{1}{b-a}dt \times \frac{1}{2^{-(n-1)} \times 2^{1-(2^{E-1}-1)}}\\
&= \int_{\{t\in[a,b]\,|\,round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt \times \frac{1}{\alpha} \times \frac{1}{2^{-(n-1)} \times 2^{1-(2^{E-1}-1)}}\\
&= \int_{\{t\in[a,b]\,|\,round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt\\
&= P_{\mathbb{F}}(f)\,.
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(3) Case: $a < 0 < b$.

(i) Case: $f < a$ or $b < f$.

In this case, we have

$$\begin{aligned} P\left(f\right) &= Pr\left[x = f \text{ in } 40\right] \\ &= Pr\left[x = f \text{ in } 30\right] \\ &= 0. \end{aligned}$$

Next, we have

$$\left\{t \in [a, b] \left| round_{\mathbb{F}}\left(t\right) = f\right.\right\} = \emptyset$$

because $f < a$ or $b < f$. Hence, we obtain

$$\begin{aligned} P_{\mathbb{F}}\left(f\right) &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\ &= \int_{\emptyset} \frac{1}{b-a} dt \\ &= 0. \end{aligned}$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

(ii) Case: $a < f < b$.

In this case, we obtain

$$\begin{aligned} P\left(f\right) &= Pr\left[x = f \text{ in } 40\right] \\ &= Pr\left[x = f \text{ in } 30\right] \\ &= \sum_{l=0}^{\infty} \left(1-\gamma\right)^l \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=f\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &= \frac{1}{\gamma} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=f\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &= \frac{2^{k+1}}{b-a} \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=f\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &= \int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\ &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\ &= P_{\mathbb{F}}\left(f\right). \end{aligned}$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

(iii) Case: $f = a$.

- Case: $a = -2^k$.

  In this case, we obtain

$$\begin{aligned}
P(f) &= P(a) \\
&= Pr\left[x = a \text{ in } 40\right] \\
&= Pr\left[x = -2^k \text{ in } 40\right] \\
&= Pr\left[x = -2^k \text{ in } 30\right] \\
&= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = -2^k\}} \frac{1}{2^k - (-2^k)} dt \\
&= \frac{1}{\gamma} \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = -2^k\}} \frac{1}{2^k - (-2^k)} dt \\
&= \frac{2^{k+1}}{b-a} \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = -2^k\}} \frac{1}{2^k - (-2^k)} dt \\
&= \int_{\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = -2^k\}} \frac{1}{b-a} dt \\
&= \int_{\{t \in [a, 2^k] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b-a} dt \\
&= \int_{\{t \in [a, b] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(a) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $-2^k < a$.
  - Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward 0.
    In this case, we have

$$\begin{aligned}
P(f) &= P(a) \\
&= Pr\left[x = a \text{ in } 40\right] \\
&= Pr\left[x = a \text{ in } 30\right] \\
&= 0.
\end{aligned}$$

Besides, we have

$$\begin{aligned}
P_{\mathbb{F}}(f) &= P_{\mathbb{F}}(a) \\
&= 0.
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

  - Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
    - Case: The mantissa of a is not 0.
      In this case, by letting a_l be the left adjacent floating point number to a and a_r be the right one, we have

$$a - a_l = a_r - a.$$

Here, $-2^k \leq a_l$ because $-2^k < a$. Hence, we have

$$\begin{aligned}
&\sup\left\{t \in \left[-2^k, 2^k\right] | round_{\mathbb{F}}(t) = a\right\} - \inf\left\{t \in \left[-2^k, 2^k\right] | round_{\mathbb{F}}(t) = a\right\} \\
&= \frac{a_r + a}{2} - \frac{a + a_l}{2} \\
&= \frac{a_r - a}{2} + \frac{a - a_l}{2} \\
&= \frac{a_r - a}{2} \times 2 \\
&= \left(\frac{a_r + a}{2} - a\right) \times 2 \\
&= \left(\sup\{t \in [a, b] | round_{\mathbb{F}}(t) = a\} - \inf\{t \in [a, b] | round_{\mathbb{F}}(t) = a\}\right) \times 2.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
P(f) &= P(a) \\
&= Pr\left[x = a \text{ in } 40\right] \\
&= Pr\left[x = a \text{ in } 30\right] \times \frac{1}{2} \\
&= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{2} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \\
&= \frac{1}{2\gamma} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \\
&= \frac{2^{k+1}}{2(b-a)} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \\
&= \frac{1}{2} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[a,b] | round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(a) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

$*$ Case: The mantissa of a is 0.

Since we can prove in the case where $a = -2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a \neq -2^{1-\left(2^{E-1}-1\right)}$. In this case, by letting a_l be the left adjacent floating point number to a and a_r be the right one, we have

$$a - a_l = \frac{a_r - a}{2}.$$

Here, $-2^k \leq a_l$ because $-2^k < a$. Hence, we have

$$\begin{aligned}
&\sup\left\{t \in \left[-2^k, 2^k\right] | round_{\mathbb{F}}(t) = a\right\} - \inf\left\{t \in \left[-2^k, 2^k\right] | round_{\mathbb{F}}(t) = a\right\} \\
&= \frac{a_r + a}{2} - \frac{a + a_l}{2} \\
&= \frac{a_r - a}{2} + \frac{a - a_l}{2} \\
&= \frac{a_r - a}{2} \times \frac{3}{2} \\
&= \left(\frac{a_r + a}{2} - a\right) \times \frac{3}{2} \\
&= \left(\sup\left\{t \in [a,b] | round_{\mathbb{F}}(t) = a\right\} - \inf\left\{t \in [a,b] | round_{\mathbb{F}}(t) = a\right\}\right) \times \frac{3}{2}.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
P(f) &= P(a) \\
&= Pr\left[x = a \text{ in } 40\right] \\
&= Pr\left[x = a \text{ in } 30\right] \times \frac{2}{3} \\
&= \sum_{l=0}^{\infty} \frac{2(1-\gamma)^l}{3} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \\
&= \frac{2}{3\gamma} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \\
&= \frac{2 \times 2^{k+1}}{3(b-a)} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)} dt \\
&= \frac{2}{3} \int_{\{t\in[-2^k,2^k] | round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[a,b] | round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(a) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward $\pm\infty$.
  In this case, we obtain

$$\begin{aligned} P(f) &= P(a) \\ &= Pr\left[x = a \text{ in } 40\right] \\ &= Pr\left[x = a \text{ in } 30\right] \\ &= \sum_{l=0}^{\infty}(1-\gamma)^l \int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \\ &= \frac{1}{\gamma}\int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \\ &= \frac{2^{k+1}}{b-a}\int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=a\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a}dt \\ &= \int_{\{t\in[a,b] \mid round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a}dt \\ &= P_{\mathbb{F}}(a) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(iv) Case: $f = b$.

- Case: $b = 2^k$.
  In this case, we obtain

$$\begin{aligned} P(f) &= P(b) \\ &= Pr\left[x = b \text{ in } 40\right] \\ &= Pr\left[x = 2^k \text{ in } 40\right] \\ &= Pr\left[x = 2^k \text{ in } 30\right] \\ &= \sum_{l=0}^{\infty}(1-\gamma)^l \int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k-(-2^k)}dt \\ &= \frac{1}{\gamma}\int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k-(-2^k)}dt \\ &= \frac{2^{k+1}}{b-a}\int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k-(-2^k)}dt \\ &= \int_{\{t\in[-2^k,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{b-a}dt \\ &= \int_{\{t\in[-2^k,b] \mid round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a}dt \\ &= \int_{\{t\in[a,b] \mid round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a}dt \\ &= P_{\mathbb{F}}(b) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

  Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $b < 2^k$.
  - Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
    In this case, we have

$$\begin{aligned} P(f) &= P(b) \\ &= Pr\left[x = b \text{ in } 40\right] \\ &= Pr\left[x = b \text{ in } 30\right] \\ &= 0. \end{aligned}$$

Besides, we have

$$\begin{aligned} P_{\mathbb{F}}\left(f\right) &= P_{\mathbb{F}}\left(b\right) \\ &= 0. \end{aligned}$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

- Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
  - Case: The mantissa of b is not 0.
    In this case, by letting b_l be the left adjacent floating point number to b and b_r be the right one, we have

$$b - b_l = b_r - b$$

Here, $b_r \leq 2^k$ because $b < 2^k$. Hence, we have

$$\begin{aligned} &\sup\left\{t \in \left[-2^k, 2^k\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\} - \inf\left\{t \in \left[-2^k, 2^k\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\} \\ &= \frac{b_r + b}{2} - \frac{b + a_l}{2} \\ &= \frac{b_r - b}{2} + \frac{b - a_l}{2} \\ &= \frac{b - b_l}{2} \times 2 \\ &= \left(b - \frac{b + b_l}{2}\right) \times 2 \\ &= \left(\sup\left\{t \in \left[a, b\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\} - \inf\left\{t \in \left[a, b\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\}\right) \times 2. \end{aligned}$$

Thus, we obtain

$$\begin{aligned} P\left(f\right) &= P\left(b\right) \\ &= Pr\left[x = b \text{ in } 40\right] \\ &= Pr\left[x = b \text{ in } 30\right] \times \frac{1}{2} \\ &= \sum_{l=0}^{\infty} \frac{2\left(1-\gamma\right)^l}{2} \int_{\left\{t \in \left[-2^k, 2^k\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &= \frac{1}{2\gamma} \int_{\left\{t \in \left[-2^k, 2^k\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &= \frac{2^{k+1}}{2\left(b-a\right)} \int_{\left\{t \in \left[-2^k, 2^k\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\}} \frac{1}{2^k - \left(-2^k\right)} dt \\ &= \frac{1}{2} \int_{\left\{t \in \left[-2^k, 2^k\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\}} \frac{1}{b-a} dt \\ &= \int_{\left\{t \in \left[a, b\right] \left| round_{\mathbb{F}}\left(t\right) = b\right.\right\}} \frac{1}{b-a} dt \\ &= P_{\mathbb{F}}\left(b\right) \\ &= P_{\mathbb{F}}\left(f\right). \end{aligned}$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

  - Case: The mantissa of b is 0.
    Since we can prove in the case where $b = 2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of b is not 0, we can consider only the case where $b \neq 2^{1-\left(2^{E-1}-1\right)}$. In this case, by letting b_l be the left adjacent floating point number to b and b_r be the right one, we have

$$b - b_l = \frac{b_r - b}{2}.$$

Here, $b_r \leq 2^k$ because $b < 2^k$. Hence, we have

$$\begin{aligned}
&\sup\left\{t\in\left[-2^k,2^k\right]|round_{\mathbb{F}}(t)=b\right\}-\inf\left\{t\in\left[-2^k,2^k\right]|round_{\mathbb{F}}(t)=b\right\}\\
&=\frac{b_r+b}{2}-\frac{b+b_l}{2}\\
&=\frac{b_r-b}{2}+\frac{b-b_l}{2}\\
&=\frac{b-b_l}{2}\times 3\\
&=\left(b-\frac{b+b_l}{2}\right)\times 3\\
&=\left(\sup\left\{t\in[a,b]\,|round_{\mathbb{F}}(t)=b\right\}-\inf\left\{t\in[a,b]\,|round_{\mathbb{F}}(t)=b\right\}\right)\times 3.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
P(f)&=P(b)\\
&=Pr\left[x=b\text{ in }40\right]\\
&=Pr\left[x=b\text{ in }30\right]\times\frac{1}{3}\\
&=\sum_{l=0}^{\infty}\frac{(1-\gamma)^l}{3}\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=b\}}\frac{1}{2^k-(-2^k)}dt\\
&=\frac{1}{3\gamma}\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=b\}}\frac{1}{2^k-(-2^k)}dt\\
&=\frac{2^{k+1}}{3(b-a)}\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=b\}}\frac{1}{2^k-(-2^k)}dt\\
&=\frac{1}{3}\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=b\}}\frac{1}{b-a}dt\\
&=\int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=b\}}\frac{1}{b-a}dt\\
&=P_{\mathbb{F}}(b)\\
&=P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f)=P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward $\pm\infty$.

  In this case, we obtain

$$\begin{aligned}
P(f)&=P(a)\\
&=Pr\left[x=a\text{ in }40\right]\\
&=Pr\left[x=a\text{ in }30\right]\\
&=\sum_{l=0}^{\infty}(1-\gamma)^l\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}}\frac{1}{2^k-(-2^k)}dt\\
&=\frac{1}{\gamma}\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}}\frac{1}{2^k-(-2^k)}dt\\
&=\frac{2^{k+1}}{b-a}\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}}\frac{1}{2^k-(-2^k)}dt\\
&=\int_{\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=a\}}\frac{1}{b-a}dt\\
&=\int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=a\}}\frac{1}{b-a}dt\\
&=P_{\mathbb{F}}(a)\\
&=P_{\mathbb{F}}(f).
\end{aligned}$$

  Therefore, $P(f)=P_{\mathbb{F}}(f)$ holds.

Therefore, $P(f)=P_{\mathbb{F}}(f)$ holds in all the cases.

(4) Case: $2^{n-2}\leq a<2^{n-1}<b\leq 2^n$ where $(2-\left(2^{E-1}-1\right)\leq n\in\mathbb{N}\leq 2^{E-1})$.

(i) Case: $f<a$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-2} + 2^{n-1}$, we have

$$\begin{aligned} x &< \left(a - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \\ &= p. \end{aligned}$$

Hence, we obtain

$$\begin{aligned} P(f) &= Pr\left[x = \left(f - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \text{ in } 40\right] \\ &\leq Pr\left[x < p \text{ in } 30\right] \\ &= 0. \end{aligned}$$

Next, we have

$$\{t \in [a,b] \,|round_{\mathbb{F}}(t) = f\} = \emptyset$$

because $f < a$. Hence, we obtain

$$\begin{aligned} P_{\mathbb{F}}(f) &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\ &= \int_{\emptyset} \frac{1}{b-a} dt \\ &= 0. \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(ii) Case: $f = a$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-2} + 2^{n-1}$, we have

$$\begin{aligned} x &= \left(a - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \\ &= p. \end{aligned}$$

Now, consider the case where $p = -2^k$ and the case where $-2^k < p$.

- Case: $p = -2^k$.
  In this case, we obtain

$$\begin{aligned} P(f) &= P(a) \\ &= Pr\left[x = p \text{ in } 40\right] \\ &= Pr\left[x = p \text{ in } 30\right] \\ &= Pr\left[x = -2^k \text{ in } 30\right] \\ &= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{2} \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=-2^k\}} \frac{1}{0-(-2^k)} dt \\ &= \frac{1}{3\gamma} \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=-2^k\}} \frac{1}{0-(-2^k)} dt \\ &= \frac{2^k}{2q-p} \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=-2^k\}} \frac{1}{0-(-2^k)} dt \\ &= \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=-2^k\}} \frac{1}{2q-p} dt \\ &= \int_{\{t\in[p,0]|round_{\mathbb{F}}(t)=p\}} \frac{1}{2q-p} dt \\ &= \int_{\{t\in[p,2q]|round_{\mathbb{F}}(t)=p\}} \frac{1}{2q-p} dt \\ &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\ &= P_{\mathbb{F}}(a) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $-2^k < p$.
  - Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
    In this case, we have

$$\begin{aligned} P(f) &= P(a) \\ &= Pr\left[x = p \text{ in } 40\right] \\ &= Pr\left[x = p \text{ in } 30\right] \\ &= 0. \end{aligned}$$

Besides, we have

$$\begin{aligned} P_{\mathbb{F}}(f) &= P_{\mathbb{F}}(b) \\ &= 0. \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

  - Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
    Let p_l be the left adjacent floating point number to p and p_r be the right one. Since p is a subnormal number, we have

$$p - p_l = p_r - p.$$

Here, $-2^k \le p_l$ because $-2^k < p$. Hence, we have

$$\begin{aligned} &\sup\left\{t \in \left[-2^k, 0\right] \mid round_{\mathbb{F}}(t) = p\right\} - \inf\left\{t \in \left[-2^k, 0\right] \mid round_{\mathbb{F}}(t) = p\right\} \\ &= \frac{p_r + p}{2} - \frac{p + p_l}{2} \\ &= \frac{p_r - p}{2} + \frac{p - p_l}{2} \\ &= \frac{p_r - p}{2} \times 2 \\ &= \left(\frac{p_r + p}{2} - p\right) \times 2 \\ &= \left(\sup\left\{t \in [p, 0] \mid round_{\mathbb{F}}(t) = p\right\} - \inf\left\{t \in [p, 0] \mid round_{\mathbb{F}}(t) = p\right\}\right) \times 2. \end{aligned}$$

Thus, we obtain

$$\begin{aligned} P(f) &= P(a) \\ &= Pr\left[x = p \text{ in } 40\right] \\ &= Pr\left[x = p \text{ in } 30\right] \times \frac{1}{2} \\ &= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{6} \int_{\{t \in [-2^k, 0] \mid round_{\mathbb{F}}(t) = p\}} \frac{1}{0 - (-2^k)} dt \\ &= \frac{1}{6\gamma} \int_{\{t \in [-2^k, 0] \mid round_{\mathbb{F}}(t) = p\}} \frac{1}{0 - (-2^k)} dt \\ &= \frac{2^k}{2(2q - p)} \int_{\{t \in [-2^k, 0] \mid round_{\mathbb{F}}(t) = p\}} \frac{1}{0 - (-2^k)} dt \\ &= \frac{1}{2} \int_{\{t \in [-2^k, 0] \mid round_{\mathbb{F}}(t) = p\}} \frac{1}{2q - p} dt \\ &= \int_{\{t \in [p, 0] \mid round_{\mathbb{F}}(t) = p\}} \frac{1}{2q - p} dt \\ &= \int_{\{t \in [p, 2q] \mid round_{\mathbb{F}}(t) = p\}} \frac{1}{2q - p} dt \\ &= \int_{\{t \in [a, b] \mid round_{\mathbb{F}}(t) = a\}} \frac{1}{b - a} dt \\ &= P_{\mathbb{F}}(a) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

  In this case, we obtain

$$\begin{aligned}
P(f) &= P(a) \\
&= Pr\left[x = p \text{ in } 40\right] \\
&= Pr\left[x = p \text{ in } 30\right] \\
&= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{3} \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=p\}} \frac{1}{0-(-2^k)} dt \\
&= \frac{1}{3\gamma} \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=p\}} \frac{1}{0-(-2^k)} dt \\
&= \frac{2^k}{2q-p} \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=p\}} \frac{1}{0-(-2^k)} dt \\
&= \int_{\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=p\}} \frac{1}{2q-p} dt \\
&= \int_{\{t\in[p,0]|round_{\mathbb{F}}(t)=p\}} \frac{1}{2q-p} dt \\
&= \int_{\{t\in[p,2q]|round_{\mathbb{F}}(t)=p\}} \frac{1}{2q-p} dt \\
&= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(a) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

  Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(iii) Case: $a < f < 2^{n-1}$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-2} + 2^{n-1}$, we have

$$\begin{aligned}
P(f) &= Pr\left[x = \left(f - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \text{ in } 40\right] \\
&= Pr\left[x = \left(f - 2^{n-1}\right) \times 2^{-(n-2)} \times 2^{1-\left(2^{E-1}-1\right)} \text{ in } 30\right] \\
&= \sum_{l=0}^{\infty} \frac{(1-\gamma)}{3} \int_{\left\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{0-(-2^k)} dt \\
&= \frac{1}{3\gamma} \int_{\left\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{0-(-2^k)} dt \\
&= \frac{2^k}{2q-p} \int_{\left\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{0-(-2^k)} dt \\
&= \int_{\left\{t\in[-2^k,0]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{2q-p} dt \\
&= \int_{\left\{t\in[-2^k,2^k]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{2q-p} dt \\
&= \int_{\left\{t\in[p,q]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{2q-p} dt \\
&= \int_{\left\{t\in[p,2q]|round_{\mathbb{F}}(t)=(f-2^{n-1})\times 2^{-(n-2)}\times 2^{1-(2^{E-1}-1)}\right\}} \frac{1}{2q-p} dt \\
&= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(iv) Case: $f = 2^{n-1}$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-2} + 2^{n-1}$ or $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}$, we have $x = -0, +0$. Hence, we obtain

$$
\begin{aligned}
P\left(f\right) &= Pr\left[x = -0 \text{ in } 40\right] + Pr\left[x = +0 \text{ in } 40\right] \\
&= Pr\left[x = -0 \text{ in } 30\right] + Pr\left[x = +0 \text{ in } 30\right] \\
&= \sum_{l=0}^{\infty} \frac{\left(1-\gamma\right)}{3} \int_{\left\{t \in [-2^k, 0] | round_{\mathbb{F}}(t) = -0\right\}} \frac{1}{0 - \left(-2^k\right)} dt \\
&\quad + \sum_{l=0}^{\infty} \frac{2\left(1-\gamma\right)}{3} \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = +0\right\}} \frac{1}{2^k - 0} dt \\
&= \frac{1}{3\gamma} \int_{\left\{t \in [-2^k, 0] | round_{\mathbb{F}}(t) = -0\right\}} \frac{1}{0 - \left(-2^k\right)} dt \\
&\quad + \frac{2}{3\gamma} \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = +0\right\}} \frac{1}{2^k - 0} dt \\
&= \frac{2^k}{2q - p} \int_{\left\{t \in [-2^k, 0] | round_{\mathbb{F}}(t) = -0\right\}} \frac{1}{0 - \left(-2^k\right)} dt \\
&\quad + \frac{2^{k+1}}{2q - p} \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = +0\right\}} \frac{1}{2^k - 0} dt \\
&= \int_{\left\{t \in [-2^k, 0] | round_{\mathbb{F}}(t) = -0\right\}} \frac{1}{2q - p} dt + 2 \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = +0\right\}} \frac{1}{2q - p} dt \\
&= \int_{\left\{t \in [p, 0] | round_{\mathbb{F}}(t) = -0\right\}} \frac{1}{2q - p} dt + 2 \int_{\left\{t \in [0, q] | round_{\mathbb{F}}(t) = +0\right\}} \frac{1}{2q - p} dt \\
&= \int_{\left\{t \in [a, 2^{n-1}] | round_{\mathbb{F}}(t) = 2^{n-1}\right\}} \frac{1}{b - a} dt + \int_{\left\{t \in [2^{n-1}, b] | round_{\mathbb{F}}(t) = 2^{n-1}\right\}} \frac{1}{b - a} dt \\
&= \int_{\left\{t \in [a, b] | round_{\mathbb{F}}(t) = 2^{n-1}\right\}} \frac{1}{b - a} dt \\
&= P_{\mathbb{F}}\left(2^{n-1}\right) \\
&= P_{\mathbb{F}}\left(f\right).
\end{aligned}
$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

(v) Case: $2^{n-1} < f < b$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}$, we have

$$
\begin{aligned}
P\left(f\right) &= Pr\left[x = \left(f - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \text{ in } 40\right] \\
&= Pr\left[x = \left(f - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \text{ in } 30\right] \\
&= \sum_{l=0}^{\infty} \frac{2\left(1-\gamma\right)}{3} \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{2^k - 0} dt \\
&= \frac{2}{3\gamma} \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{2^k - 0} dt \\
&= \frac{2^{k+1}}{2q - p} \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{2^k - 0} dt \\
&= 2 \int_{\left\{t \in [0, 2^k] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{2q - p} dt \\
&= 2 \int_{\left\{t \in [-2^k, 2^k] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{2q - p} dt \\
&= 2 \int_{\left\{t \in \left[\frac{p}{2}, q\right] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{2q - p} dt \\
&= \int_{\left\{t \in \left[\frac{p}{2}, q\right] | round_{\mathbb{F}}(t) = (f - 2^{n-1}) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)}\right\}} \frac{1}{q - \frac{p}{2}} dt \\
&= \int_{\left\{t \in [a, b] | round_{\mathbb{F}}(t) = f\right\}} \frac{1}{b - a} dt \\
&= P_{\mathbb{F}}\left(f\right).
\end{aligned}
$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

(vi) Case: $f = b$.
In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}$, we have

$$\begin{aligned} x &= \left(b - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \\ &= q. \end{aligned}$$

Now, consider the case where $q = 2^k$ and the case where $q < 2^k$.

- Case: $q = 2^k$.
  In this case, we obtain

$$\begin{aligned} P(f) &= P(b) \\ &= Pr\left[x = q \text{ in } 40\right] \\ &= Pr\left[x = q \text{ in } 30\right] \\ &= Pr\left[x = 2^k \text{ in } 30\right] \\ &= \sum_{l=0}^{\infty} \frac{2(1-\gamma)^l}{3} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k - 0} dt \\ &= \frac{2}{3\gamma} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k - 0} dt \\ &= \frac{2^{k+1}}{2q-p} \int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2^k - 0} dt \\ &= 2\int_{\{t\in[0,2^k] \mid round_{\mathbb{F}}(t)=2^k\}} \frac{1}{2q-p} dt \\ &= 2\int_{\{t\in[0,q] \mid round_{\mathbb{F}}(t)=q\}} \frac{1}{2q-p} dt \\ &= 2\int_{\{t\in[p,q] \mid round_{\mathbb{F}}(t)=q\}} \frac{1}{2q-p} dt \\ &= \int_{\left\{t\in\left[\frac{p}{2},q\right] \mid round_{\mathbb{F}}(t)=q\right\}} \frac{1}{q-\frac{p}{2}} dt \\ &= \int_{\{t\in[a,b] \mid round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a} dt \\ &= P_{\mathbb{F}}(b) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

  Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $q < 2^k$.
  - Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
    In this case, we have

$$\begin{aligned} P(f) &= P(b) \\ &= Pr\left[x = q \text{ in } 40\right] \\ &= Pr\left[x = q \text{ in } 30\right] \\ &= 0. \end{aligned}$$

    Besides, we have

$$\begin{aligned} P_{\mathbb{F}}(f) &= P_{\mathbb{F}}(q) \\ &= 0. \end{aligned}$$

    Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.
  - Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
    Let q_l be the left adjacent floating point number to q and q_r be the right one. Since q is a subnormal number, we have

$$q - q_l = q_r - q.$$

Here, $q_r \leq 2^k$ because $q < 2^k$. Hence, we have

$$\begin{aligned}
&\sup\left\{t \in \left[0, 2^k\right] |round_{\mathbb{F}}(t) = q\right\} - \inf\left\{t \in \left[0, 2^k\right] |round_{\mathbb{F}}(t) = q\right\}\\
&= \frac{q_r + q}{2} - \frac{q + q_l}{2}\\
&= \frac{q_r - q}{2} + \frac{q - q_l}{2}\\
&= \frac{q - q_l}{2} \times 2\\
&= \left(q - \frac{q + q_l}{2}\right) \times 2\\
&= \left(\sup\left\{t \in [0, q] |round_{\mathbb{F}}(t) = q\right\} - \inf\left\{t \in [0, q] |round_{\mathbb{F}}(t) = q\right\}\right) \times 2.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
P(f) &= P(b)\\
&= Pr[x = q \text{ in } 40]\\
&= Pr[x = q \text{ in } 30] \times \frac{1}{2}\\
&= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{3} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t) = q\}} \frac{1}{2^k - 0} dt\\
&= \frac{1}{3\gamma} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t) = q\}} \frac{1}{2^k - 0} dt\\
&= \frac{2^k}{2q - p} \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t) = q\}} \frac{1}{2^k - 0} dt\\
&= \int_{\{t \in [0,2^k] | round_{\mathbb{F}}(t) = q\}} \frac{1}{2q - p} dt\\
&= 2 \int_{\{t \in [0,q] | round_{\mathbb{F}}(t) = q\}} \frac{1}{2q - p} dt\\
&= 2 \int_{\left\{t \in \left[\frac{p}{2}, q\right] | round_{\mathbb{F}}(t) = q\right\}} \frac{1}{2q - p} dt\\
&= \int_{\left\{t \in \left[\frac{p}{2}, q\right] | round_{\mathbb{F}}(t) = q\right\}} \frac{1}{q - \frac{p}{2}} dt\\
&= \int_{\{t \in [a,b] | round_{\mathbb{F}}(t) = b\}} \frac{1}{b - a} dt\\
&= P_{\mathbb{F}}(b)\\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

– Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
  In this case, we obtain

$$\begin{aligned}
P\left(f\right) &= P\left(b\right) \\
&= Pr\left[x = q \text{ in } 40\right] \\
&= Pr\left[x = q \text{ in } 30\right] \\
&= \sum_{l=0}^{\infty} \frac{2\left(1-\gamma\right)^{l}}{3} \int_{\{t\in[0,2^k]|round_{\mathbb{F}}(t)=q\}} \frac{1}{2^k-0} dt \\
&= \frac{2}{3\gamma} \int_{\{t\in[0,2^k]|round_{\mathbb{F}}(t)=q\}} \frac{1}{2^k-0} dt \\
&= \frac{2^{k+1}}{2q-p} \int_{\{t\in[0,2^k]|round_{\mathbb{F}}(t)=q\}} \frac{1}{2^k-0} dt \\
&= 2 \int_{\{t\in[0,2^k]|round_{\mathbb{F}}(t)=q\}} \frac{1}{2q-p} dt \\
&= 2 \int_{\left\{t\in\left[\frac{p}{2},q\right]|round_{\mathbb{F}}(t)=q\right\}} \frac{1}{2q-p} dt \\
&= \int_{\left\{t\in\left[\frac{p}{2},q\right]|round_{\mathbb{F}}(t)=q\right\}} \frac{1}{q-\frac{p}{2}} dt \\
&= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}\left(b\right) \\
&= P_{\mathbb{F}}\left(f\right).
\end{aligned}$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

(vii) Case: $b < f$.

In this case, since $f = x \times 2^{\left(2^{E-1}-1\right)-1} \times 2^{n-1} + 2^{n-1}$, we have

$$\begin{aligned}
x &> \left(b - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \\
&= q.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
P\left(f\right) &= Pr\left[x = \left(f - 2^{n-1}\right) \times 2^{-(n-1)} \times 2^{1-\left(2^{E-1}-1\right)} \text{ in } 40\right] \\
&\leq Pr\left[x > q \text{ in } 30\right] \\
&= 0.
\end{aligned}$$

Next, we have

$$\left\{t \in [a,b] \left| round_{\mathbb{F}}\left(t\right) = f\right.\right\} = \emptyset$$

because $b < f$. Hence, we obtain

$$\begin{aligned}
P_{\mathbb{F}}\left(f\right) &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a} dt \\
&= \int_{\emptyset} \frac{1}{b-a} dt \\
&= 0.
\end{aligned}$$

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds.

Therefore, $P\left(f\right) = P_{\mathbb{F}}\left(f\right)$ holds in all the cases.

(5) Case: $0 \leq a < 2^{n-2} < 2^{n-1} < b \leq 2^{n}$ where $\left(2 - \left(2^{E-1}-1\right) \leq n \in \mathbb{N} \leq 2^{E-1}\right)$.

(i) Case: $f < a$ or $b < f$.

In this case, we have

$$\begin{aligned}
P\left(f\right) &= Pr\left[x = f \text{ in } 30\right] \\
&= Pr\left[x = f \text{ in } 20\right] \\
&= 0.
\end{aligned}$$

Next, we have

$$\{t \in [a,b] \,|round_{\mathbb{F}}(t) = f\} = \emptyset$$

because $f < a$ or $b < f$. Hence, we obtain

$$\begin{aligned} P_{\mathbb{F}}(f) &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt \\ &= \int_{\emptyset} \frac{1}{b-a}dt \\ &= 0. \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(ii) Case: $a < f < b$.

In this case, we obtain

$$\begin{aligned} P(f) &= Pr[x = f \text{ in } 30] \\ &= Pr[x = f \text{ in } 20] \\ &= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=f\}} \frac{1}{2^n-0}dt \\ &= \frac{1}{\gamma} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=f\}} \frac{1}{2^n-0}dt \\ &= \frac{2^n}{b-a} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=f\}} \frac{1}{2^n-0}dt \\ &= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt \\ &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=f\}} \frac{1}{b-a}dt \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(iii) Case: $f = a$.

- Case: $a = 0$.

  In this case, we obtain

$$\begin{aligned} P(f) &= P(a) \\ &= Pr[x = a \text{ in } 30] \\ &= Pr[x = 0 \text{ in } 30] \\ &= Pr[x = 0 \text{ in } 20] \\ &= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=0\}} \frac{1}{2^n-0}dt \\ &= \frac{1}{\gamma} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=0\}} \frac{1}{2^n-0}dt \\ &= \frac{2^n}{b-a} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=0\}} \frac{1}{2^n-0}dt \\ &= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=0\}} \frac{1}{b-a}dt \\ &= \int_{\{t\in[a,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a}dt \\ &= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a}dt \\ &= P_{\mathbb{F}}(a) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

  Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $0 < a$.
  - Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.

In this case, we have

$$\begin{aligned} P(f) &= P(a) \\ &= Pr\left[x = a \text{ in } 30\right] \\ &= Pr\left[x = a \text{ in } 20\right] \\ &= 0. \end{aligned}$$

Besides, we have

$$\begin{aligned} P_{\mathbb{F}}(f) &= P_{\mathbb{F}}(a) \\ &= 0. \end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Round-to-Nearest.
  - Case: The mantissa of a is not 0.

    In this case, by letting a_l be the left adjacent floating point number to a and a_r be the right one, we have

    $$a - a_l = a_r - a.$$

    Here, $0 \leq a_l$ because $0 < a$. Hence, we have

    $$\begin{aligned} &\sup\left\{t \in [0, 2^n] \left| round_{\mathbb{F}}(t) = a\right.\right\} - \inf\left\{t \in [0, 2^n] \left| round_{\mathbb{F}}(t) = a\right.\right\} \\ &= \frac{a_r + a}{2} - \frac{a + a_l}{2} \\ &= \frac{a_r - a}{2} + \frac{a - a_l}{2} \\ &= \frac{a_r - a}{2} \times 2 \\ &= \left(\frac{a_r + a}{2} - a\right) \times 2 \\ &= \left(\sup\left\{t \in [a, b] \left| round_{\mathbb{F}}(t) = a\right.\right\} - \inf\left\{t \in [a, b] \left| round_{\mathbb{F}}(t) = a\right.\right\}\right) \times 2. \end{aligned}$$

    Thus, we obtain

    $$\begin{aligned} P(f) &= P(a) \\ &= Pr\left[x = a \text{ in } 30\right] \\ &= Pr\left[x = a \text{ in } 20\right] \times \frac{1}{2} \\ &= \sum_{l=0}^{\infty} \frac{(1-\gamma)^l}{2} \int_{\{t \in [0, 2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt \\ &= \frac{1}{2\gamma} \int_{\{t \in [0, 2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt \\ &= \frac{2^n}{2(b-a)} \int_{\{t \in [0, 2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt \\ &= \frac{1}{2} \int_{\{t \in [0, 2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b - a} dt \\ &= \int_{\{t \in [a, b] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b - a} dt \\ &= P_{\mathbb{F}}(a) \\ &= P_{\mathbb{F}}(f). \end{aligned}$$

    Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.
  - Case: The mantissa of a is 0.

    Since we can prove in the case where $a = 2^{1-\left(2^{E-1}-1\right)}$ by the same way as the case where the mantissa of a is not 0, we can consider only the case where $a \neq 2^{1-\left(2^{E-1}-1\right)}$. In this case, by letting a_l be the left adjacent floating point number to a and a_r be the right one, we have

    $$a - a_l = \frac{a_r - a}{2}.$$

Here, $0 \leq a_l$ because $0 < a$. Hence, we have

$$\begin{aligned}
&\sup\left\{t \in [0, 2^n] \,|\, round_{\mathbb{F}}(t) = a\right\} - \inf\left\{t \in [0, 2^n] \,|\, round_{\mathbb{F}}(t) = a\right\}\\
&\quad = \frac{a_r + a}{2} - \frac{a + a_l}{2}\\
&\quad = \frac{a_r - a}{2} + \frac{a - a_l}{2}\\
&\quad = \frac{a_r - a}{2} \times \frac{3}{2}\\
&\quad = \left(\frac{a_r + a}{2} - a\right) \times \frac{3}{2}\\
&\quad = \left(\sup\left\{t \in [a, b] \,|\, round_{\mathbb{F}}(t) = a\right\} - \inf\left\{t \in [a, b] \,|\, round_{\mathbb{F}}(t) = a\right\}\right) \times \frac{3}{2}.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
P(f) &= P(a)\\
&= Pr\left[x = a \text{ in } 30\right]\\
&= Pr\left[x = a \text{ in } 20\right] \times \frac{2}{3}\\
&= \sum_{l=0}^{\infty} \frac{2(1-\gamma)^l}{3} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt\\
&= \frac{2}{3\gamma} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt\\
&= \frac{2 \times 2^n}{3(b-a)} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt\\
&= \frac{2}{3} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b-a} dt\\
&= \int_{\{t \in [a,b] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b-a} dt\\
&= P_{\mathbb{F}}(a)\\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.
  In this case, we obtain

$$\begin{aligned}
P(f) &= P(a)\\
&= Pr\left[x = a \text{ in } 30\right]\\
&= Pr\left[x = a \text{ in } 20\right]\\
&= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt\\
&= \frac{1}{\gamma} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt\\
&= \frac{2^n}{b-a} \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{2^n - 0} dt\\
&= \int_{\{t \in [0,2^n] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b-a} dt\\
&= \int_{\{t \in [a,b] | round_{\mathbb{F}}(t) = a\}} \frac{1}{b-a} dt\\
&= P_{\mathbb{F}}(a)\\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

(iv) Case: $f = b$.

- Case: $b = 2^n$.

In this case, we obtain

$$\begin{aligned}
P(f) &= P(b)\\
&= Pr\left[x = b \text{ in } 30\right]\\
&= Pr\left[x = 2^n \text{ in } 30\right]\\
&= Pr\left[x = 2^n \text{ in } 20\right]\\
&= \sum_{l=0}^{\infty}(1-\gamma)^l \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=2^n\}} \frac{1}{2^n-0}dt\\
&= \frac{1}{\gamma}\int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=2^n\}} \frac{1}{2^n-0}dt\\
&= \frac{2^n}{b-a}\int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=2^n\}} \frac{1}{2^n-0}dt\\
&= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=2^n\}} \frac{1}{b-a}dt\\
&= \int_{\{t\in[0,b]|round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a}dt\\
&= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a}dt\\
&= P_{\mathbb{F}}(b)\\
&= P_{\mathbb{F}}(f)\,.
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $b < 2^n$.
  - Case: $round_{\mathbb{F}}$ is Toward $-\infty$ or Toward 0.

    In this case, we have

$$\begin{aligned}
P(f) &= P(b)\\
&= Pr\left[x = b \text{ in } 30\right]\\
&= Pr\left[x = b \text{ in } 20\right]\\
&= 0.
\end{aligned}$$

    Besides, we have

$$\begin{aligned}
P_{\mathbb{F}}(f) &= P_{\mathbb{F}}(b)\\
&= 0.
\end{aligned}$$

    Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.
  - Case: $round_{\mathbb{F}}$ is Round-to-Nearest.

    In this case, by letting b_l be the left adjacent floating point number to b and b_r be the right one, we have

$$b - b_l = b_r - b$$

    Here, $b_r \leq 2^k$ because $b < 2^k$. Hence, we have

$$\begin{aligned}
&\sup\left\{t\in[0,2^n]\,|round_{\mathbb{F}}(t)=b\right\} - \inf\left\{t\in[0,2^n]\,|round_{\mathbb{F}}(t)=b\right\}\\
&= \frac{b_r+b}{2} - \frac{b+a_l}{2}\\
&= \frac{b_r-b}{2} + \frac{b-a_l}{2}\\
&= \frac{b-b_l}{2}\times 2\\
&= \left(b - \frac{b+b_l}{2}\right)\times 2\\
&= \left(\sup\left\{t\in[a,b]\,|round_{\mathbb{F}}(t)=b\right\} - \inf\left\{t\in[a,b]\,|round_{\mathbb{F}}(t)=b\right\}\right)\times 2.
\end{aligned}$$

    Thus, we obtain

$$\begin{aligned}
P(f) &= P(b) \\
&= Pr[x = b \text{ in } 30] \\
&= Pr[x = b \text{ in } 20] \times \frac{1}{2} \\
&= \sum_{l=0}^{\infty} \frac{2(1-\gamma)^l}{2} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=b\}} \frac{1}{2^n} dt \\
&= \frac{1}{2\gamma} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=b\}} \frac{1}{2^n} dt \\
&= \frac{2^n}{2(b-a)} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=b\}} \frac{1}{2^n} dt \\
&= \frac{1}{2} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=b\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(b) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

- Case: $round_{\mathbb{F}}$ is Toward $+\infty$ or Toward $\pm\infty$.
  In this case, we obtain

$$\begin{aligned}
P(f) &= P(a) \\
&= Pr[x = a \text{ in } 30] \\
&= Pr[x = a \text{ in } 20] \\
&= \sum_{l=0}^{\infty} (1-\gamma)^l \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^n - 0} dt \\
&= \frac{1}{\gamma} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^n - 0} dt \\
&= \frac{2^n}{b-a} \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{2^n - 0} dt \\
&= \int_{\{t\in[0,2^n]|round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= \int_{\{t\in[a,b]|round_{\mathbb{F}}(t)=a\}} \frac{1}{b-a} dt \\
&= P_{\mathbb{F}}(a) \\
&= P_{\mathbb{F}}(f).
\end{aligned}$$

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds.

Therefore, $P(f) = P_{\mathbb{F}}(f)$ holds in all the cases.

7. Experiment

The Section 6 proved that the proposed algorithm satisfies the condition about uniform in narrow sense, which is defined in the Section 4.1.2. Then, this section aims to confirm that the proposed algorithm is uniform in practical use and to show its performance. For this aims, the section contains the following 2 experiments.

Experiment 1 Random number generation probability.

This experiment aims to confirm that the random number generation probability of the proposed method is uniform.

Experiment 2 Performance and acceptance ratio.

This experiment aims to show performance of the proposed method by comparison with that of the existing method and to confirm that the acceptance ratio of the proposed method calculated in the Section 6.2 is correct.

7.1 Target and environment

In this experiment, the target is the following floating point uniform random number generator.

- Ratio method.
  The floating point uniform random number generator that outputs

Table 6 Environment

CPU	Intel® Core^{TM} i7-4702MQ
OS	Ubuntu 12.04 LTS 64-bit
Kernel	Linux 3.13.4-031304-generic
Compiler	g++ 4.6.3
Source code	https://goo.gl/K1NAnE
Rounding mode	Round-to-Nearest(Ties to Even)

$$a + (b - a) \times \frac{URNG_W\,()}{2^W}.$$

- Moler's method.
  The floating point uniform random number generator proposed by Moler [25]. The authors generate a uniform random number x by Moler's method and then outputs $a + (b - a) \times x$ so that we can obtain a uniform random number in $[a, b]$.
- Thoma's method.
  The floating point uniform random number generator proposed by Thoma [31]. The authors let $c = (b - a)$ and generates a uniform random number x by Thoma's method and then outputs $a+x$ so that we can obtain a uniform random number in $[a, b]$.
- Proposed method.
  The modified floating point uniform random number generator proposed in the Section . Let $N = 0$ so that $U_{\mathbb{R}} = [0, 1]$ in the generator.

Here, the authors used Round-to-Nearest(Ties to Even) for $fl_{\mathbb{F}}$ and $round_{\mathbb{F}}$[*32] and used the 32/64-bit Mersenne Twister [23] for $URNG_W$ in each generator.

Table 6 shows the environment where the experiments was done.

7.2 Experiment 1: Probability

7.2.1 Methodology

This part measures the random number generation probability for all the floating point numbers where $(E, M) = (5, 4)$ and then compares them with the values of $P_{\mathbb{F}}$ calculated by the Formula (1). In the concrete, generate 2^{30} floating point uniform random numbers and calculate the generation probability for each floating point number. Here, the authors let $W = 7$ in this part[*33].

In this experiment, the authors selected the following pairs for (a, b).

(1) $(a, b) = \left(\pi \times 10^{-5}, \frac{3}{2}\pi \times 10^{-5}\right)$.
This range corresponds to the case (1) in the Section 7.1.

(2) $(a, b) = \left(\frac{3}{2}\pi, 2\pi\right)$.
This range corresponds to the case (2) in the Section 7.1.

(3) $(a, b) = (-\pi, \pi)$.
This range corresponds to the case (3) in the Section 7.1.

(4) $(a, b) = (\pi, 2\pi)$.
This range corresponds to the case (4) in the Section 7.1.

(5) $(a, b) = (0, 2\pi)$.
This range corresponds to the case (5) in the Section 7.1.

7.2.2 Result and discussion

Figure 16, Figure 17, Figure 18, Figure 19 shows the result of (1), that is, the random number generation probability of Ratio method, Moler's method, Thoma's method, and the proposed method where $(a, b) = \left(\pi \times 10^{-5}, \frac{3}{2}\pi \times 10^{-5}\right)$ respectively. Figure 20, Figure 21, Figure 22, Figure 23 shows the result of (2). Figure 24, Figure 25, Figure 26, Figure 27 shows the result of (3). Figure 28, Figure 29, Figure 30, Figure 31 shows the result of (4). Figure 32, Figure 33, Figure 34, Figure 35 shows the result of (5).

Here, all the Figures show that only the proposed method can satisfy the Formula 1. Here, the reason why another method can not satisfy the Formula 1 is the following rounding errors. The first is Ratio method. In this case, first rounding error occurs in $\frac{URNG_W()}{2^W}$. This is shown in the Section 5.3.4. Then, $b - a$ also causes a rounding error and the error is enlarged by $(b - a) \times \frac{URNG_W()}{2^W}$. Besides, $+a$ operation in $a + (b - a) \times \frac{URNG_W()}{2^W}$ causes additional rounding errors. The next is Moler's method. This case is almost same as Ratio method. The only one difference is that x in $a + (b - a) \times x$ does not contain rounding errors because the exponent and mantissa of x is generated separately. The last is Thoma's method. In this case, first rounding error occurs in $c \times x$ at the line 40 in the pseudocode of Thoma's method. Besides, $+a$ operation in $a + x$ causes additional rounding errors.

[*32] $round_{\mathbb{F}}$ is used for $P_{\mathbb{F}}$, which is ideal probability.

[*33] The authors used $(E, M, W) = (5, 4, 7)$ because more kinds of problem had been detected when E, M, W was coprime each other.

7.3 Experiment 2: Performance

7.3.1 Methodology

This part measures the time to generate 2^{30} double precision[*34] floating point uniform random numbers 16 times for each generator and then calculates the average and standard deviation for the generation time and speed[*35]. In addition, the acceptance ratio of the proposed method is also measured.

In this experiment, the authors selected the following pairs for (a, b).

(1) $(a, b) = \left(\pi \times 10^{-309}, \frac{3}{2}\pi \times 10^{-309}\right)$.
This range corresponds to the case (1) in the Section 7.1.

(2) $(a, b) = \left(\frac{3}{2}\pi, 2\pi\right)$.
This range corresponds to the case (2) in the Section 7.1.

(3) $(a, b) = (-\pi, \pi)$.
This range corresponds to the case (3) in the Section 7.1.

(4) $(a, b) = (\pi, 2\pi)$.
This range corresponds to the case (4) in the Section 7.1.

(5) $(a, b) = (0, 2\pi)$.
This range corresponds to the case (5) in the Section 7.1.

(6) $(a, b) = (0, 1)$.
This range corresponds to the Section 5.1.1 where $N = 1$.

7.3.2 Result and discussion

Table 7 shows the acceptance ratio of the proposed method. Table 8, Table 9, Table 10, Table 11, Table 12, Table 13 shows the measured time and speed of random number generation where $(a, b) = \left(\pi \times 10^{-309}, \frac{3}{2}\pi \times 10^{-309}\right)$, $\left(\frac{3}{2}\pi, 2\pi\right)$, $(-\pi, \pi)$, $(\pi, 2\pi)$, $(0, 2\pi)$, $(0, 1)$ respectively.

First, the Table 7 shows that the acceptance ratio calculated by the experiment is quite similar to γ in the Section 6.2. In addition, the generation speed is slower when the acceptance ratio is lower. This is quite natural because the iteration in the algorithm wastes more time when the acceptance ratio is lower. Next, the Table 8, ..., Table 13 shows that the proposed method kept the generation speed from 21.9% to 80.3% of Thoma's method or kept the speed from 9.07% to 33.4% of 64-bit Mersenne Twister. So we can say that the proposed method has at least the minimal speed for practical use but it is not enough.

Therefore, we should increase the acceptance ratio in order to decrease the execution time for a future work.

Table 7 Acceptance ratio of the proposed method on double precision floating point number.

Generation range		Acceptance Ratio(%)	
a	b	Experiment	Calculated γ
$\pi \times 10^{-309}$	$\frac{3}{2}\pi \times 10^{-309}$	0.564763 ± 0.000012	0.564762
$\frac{3}{2}\pi$	2π	0.785401 ± 0.000009	0.785398
$-\pi$	π	0.785393 ± 0.000011	0.785398
π	2π	0.523597 ± 0.000008	0.523599
0	2π	0.785399 ± 0.000011	0.785398

Table 8 Random number generation time and speed on double precision floating point number where $(W, a, b) = \left(64, \pi \times 10^{-309}, \frac{3}{2}\pi \times 10^{-309}\right)$.

Generator	Time(nsec / cnt)	Speed(10^8cnt / sec)
Mersenne Twister	6.117 ± 0.087	1.635 ± 0.022
Ratio method	6.097 ± 0.013	1.640 ± 0.004
Moler's method	17.159 ± 0.179	0.583 ± 0.006
Thoma's method	14.853 ± 0.243	0.673 ± 0.011
Proposed method	35.883 ± 0.179	0.279 ± 0.001

Table 9 Random number generation time and speed on double precision floating point number where $(W, a, b) = \left(64, \frac{3}{2}\pi, 2\pi\right)$.

Generator	Time(nsec / cnt)	Speed(10^8cnt / sec)
Mersenne Twister	6.098 ± 0.062	1.640 ± 0.016
Ratio method	6.099 ± 0.016	1.640 ± 0.004
Moler's method	17.068 ± 0.096	0.586 ± 0.003
Thoma's method	14.795 ± 0.053	0.676 ± 0.002
Proposed method	22.762 ± 0.097	0.439 ± 0.002

[*34] $(E, M) = (11, 52)$.

[*35] "speed" means how many random numbers are generated every seconds, that is, 2^{30} divided by the generation time.

Table 10 Random number generation time and speed on double precision floating point number where $(W, a, b) = (64, -\pi, \pi)$.

Generator	Time(nsec / cnt)	Speed(10^8cnt / sec)
Mersenne Twister	6.118 ± 0.069	1.635 ± 0.018
Ratio method	6.102 ± 0.069	1.639 ± 0.018
Moler's method	17.077 ± 0.136	0.586 ± 0.005
Thoma's method	14.764 ± 0.036	0.677 ± 0.002
Proposed method	43.885 ± 0.365	0.228 ± 0.002

Table 11 Random number generation time and speed on double precision floating point number where $(W, a, b) = (64, \pi, 2\pi)$.

Generator	Time(nsec / cnt)	Speed(10^8cnt / sec)
Mersenne Twister	6.164 ± 0.102	1.623 ± 0.027
Ratio method	6.249 ± 0.332	1.604 ± 0.079
Moler's method	17.006 ± 0.063	0.588 ± 0.002
Thoma's method	14.848 ± 0.114	0.674 ± 0.005
Proposed method	67.917 ± 0.264	0.147 ± 0.001

Table 12 Random number generation time and speed on double precision floating point number where $(W, a, b) = (64, 0, 2\pi)$.

Generator	Time(nsec / cnt)	Speed(10^8cnt / sec)
Mersenne Twister	6.123 ± 0.107	1.634 ± 0.028
Ratio method	6.125 ± 0.084	1.633 ± 0.022
Moler's method	16.971 ± 0.043	0.589 ± 0.001
Thoma's method	14.816 ± 0.118	0.675 ± 0.005
Proposed method	36.673 ± 0.244	0.273 ± 0.002

Table 13 Random number generation time and speed on double precision floating point number where $(W, a, b) = (64, 0, 1)$.

Generator	Time(nsec / cnt)	Speed(10^8cnt / sec)
Mersenne Twister	6.145 ± 0.095	1.628 ± 0.025
Ratio method	6.134 ± 0.124	1.631 ± 0.032
Moler's method	16.976 ± 0.048	0.589 ± 0.002
Thoma's method	14.789 ± 0.113	0.676 ± 0.005
Proposed method	18.420 ± 0.067	0.543 ± 0.002

Fig. 16 Random number generation probability of Ratio method in $\left[\pi \times 10^{-5}, \frac{3}{2}\pi \times 10^{-5}\right]$.

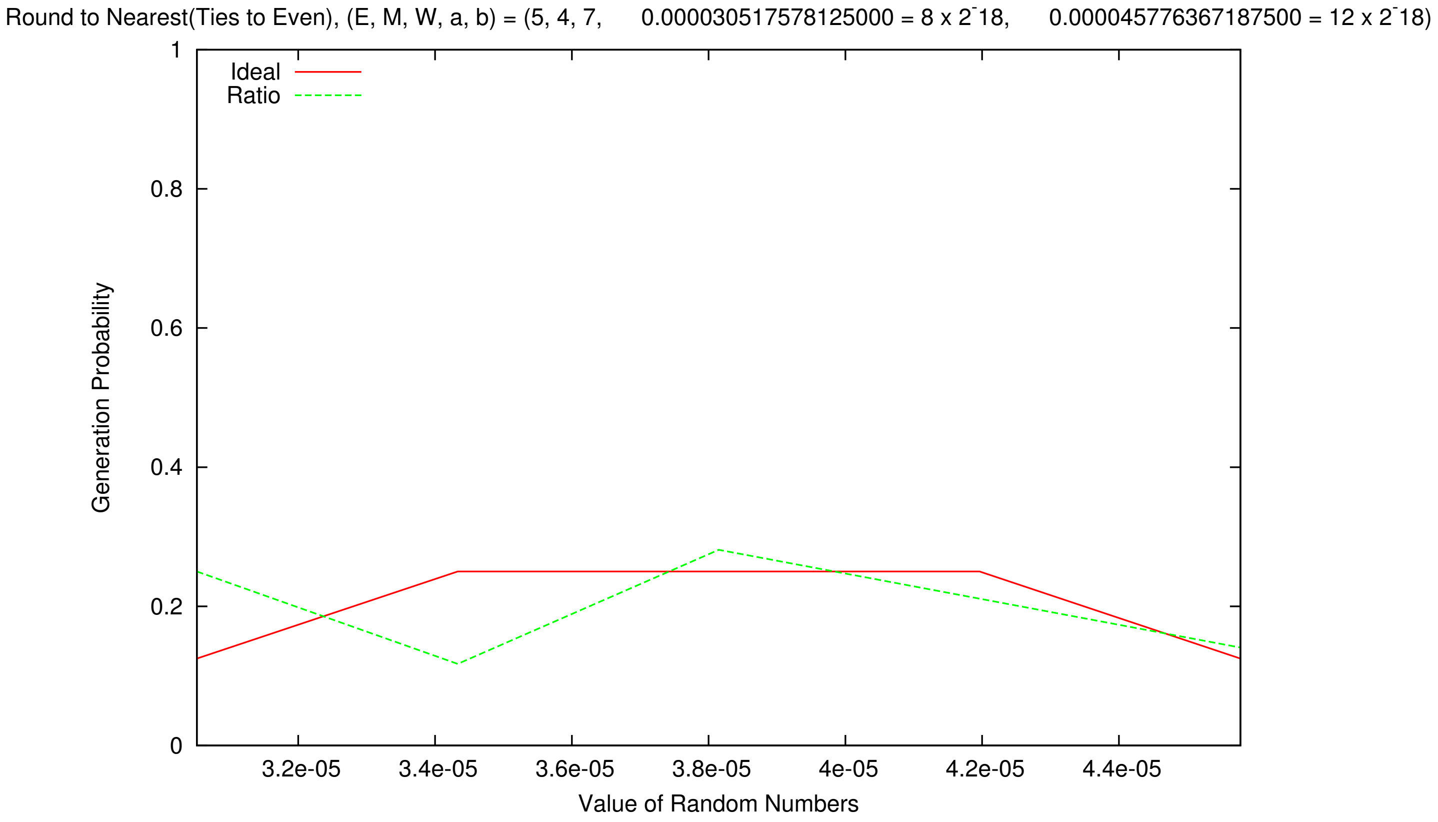


Fig. 17 Random number generation probability of extended Moler's method in $\left[\pi \times 10^{-5}, \frac{3}{2}\pi \times 10^{-5}\right]$.

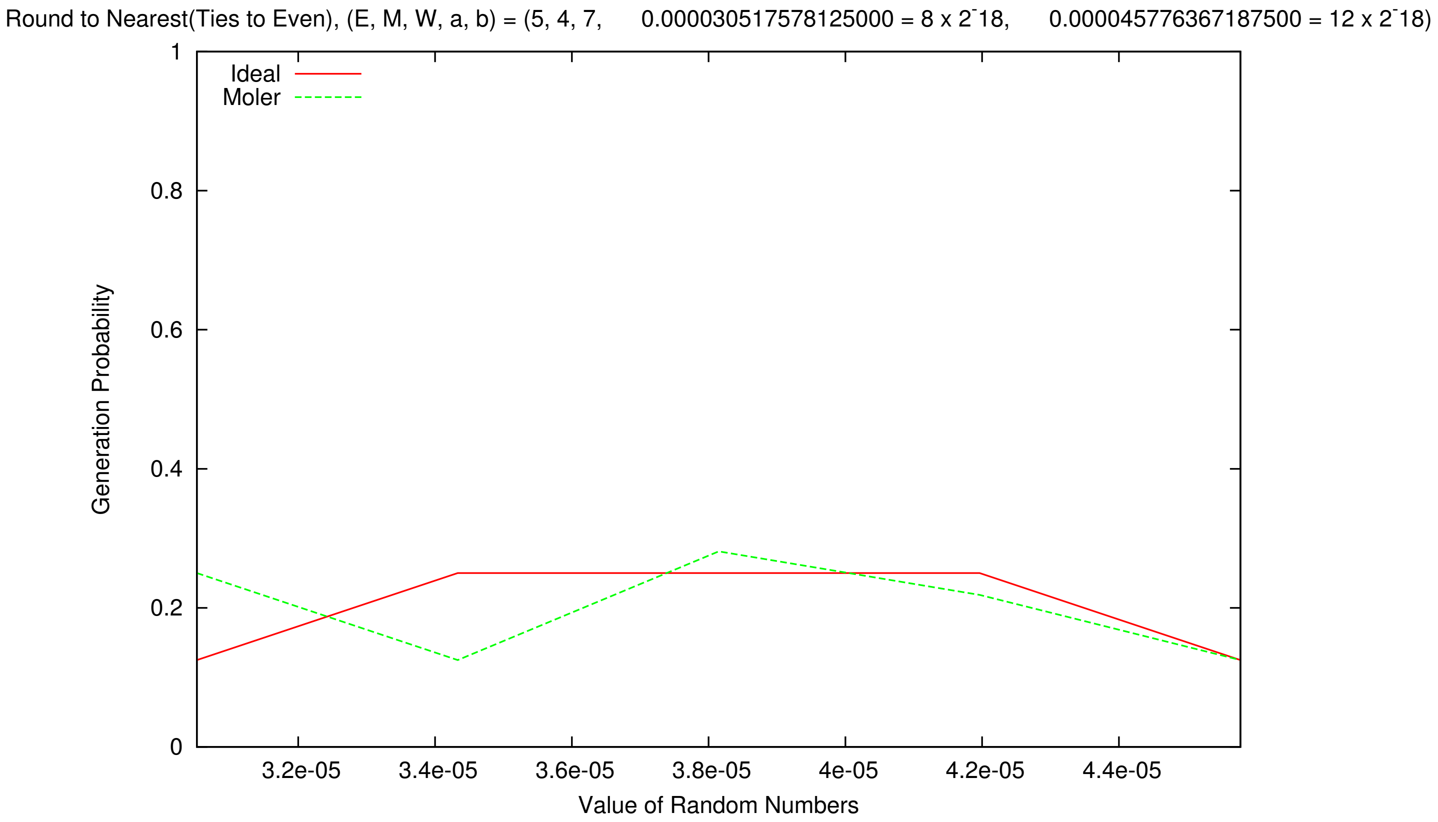


Fig. 18 Random number generation probability of extended Thoma's method in $\left[\pi \times 10^{-5}, \frac{3}{2}\pi \times 10^{-5}\right]$.

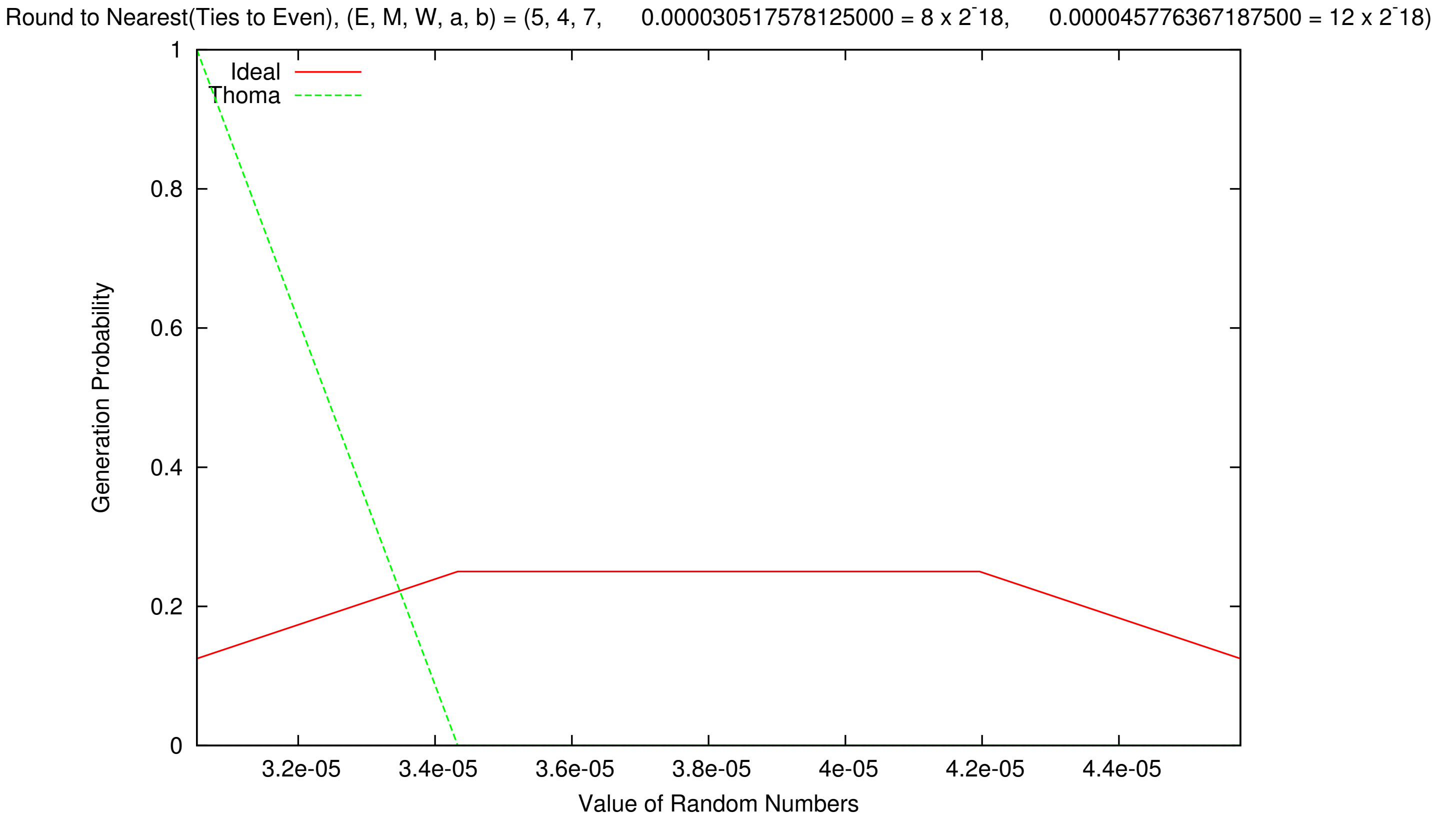


Fig. 19 Random number generation probability of the proposed method in $\left[\pi \times 10^{-5}, \frac{3}{2}\pi \times 10^{-5}\right]$.

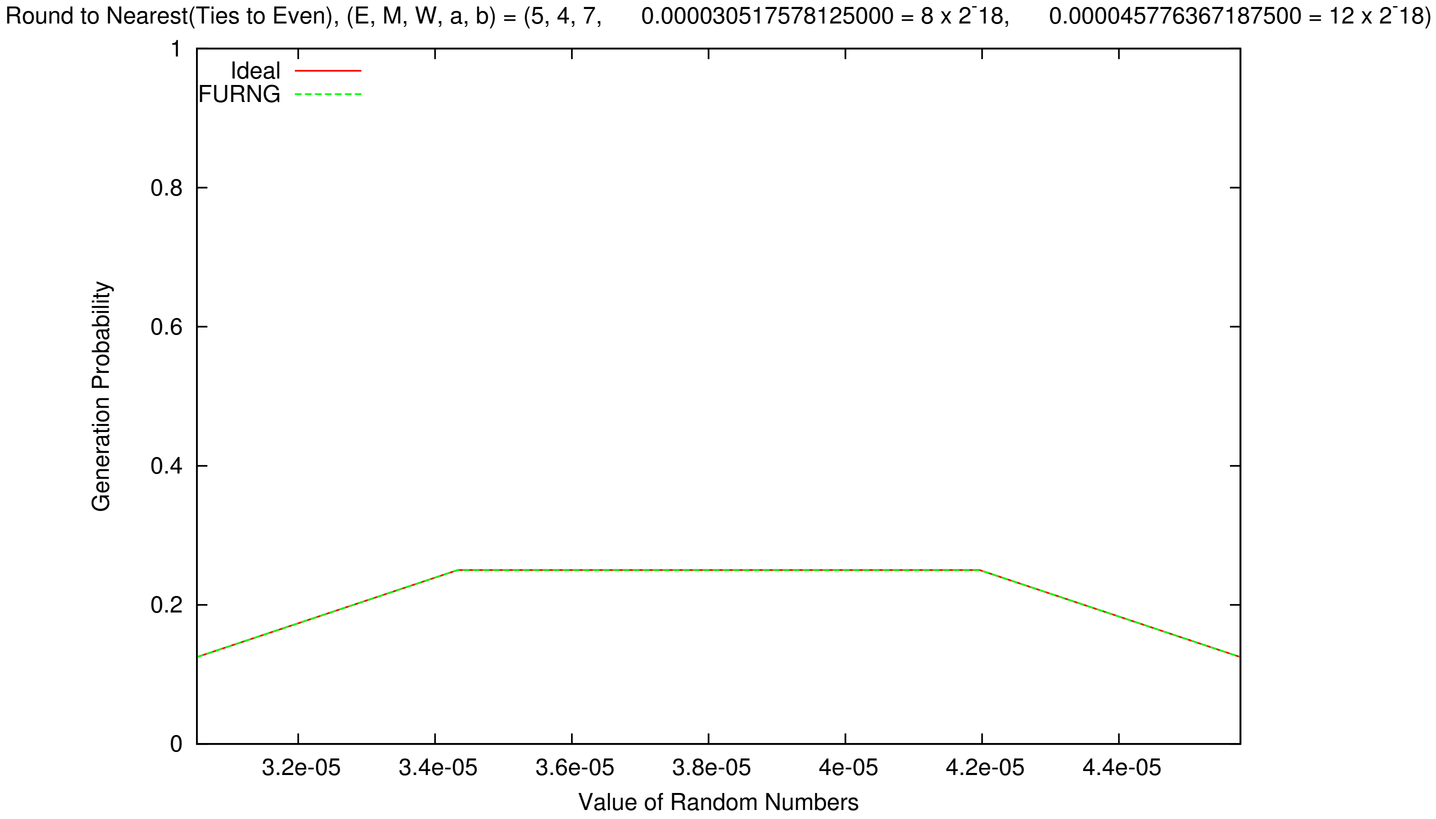


Fig. 20 Random number generation probability of Ratio method in $\left[\frac{3}{2}\pi, 2\pi\right]$.

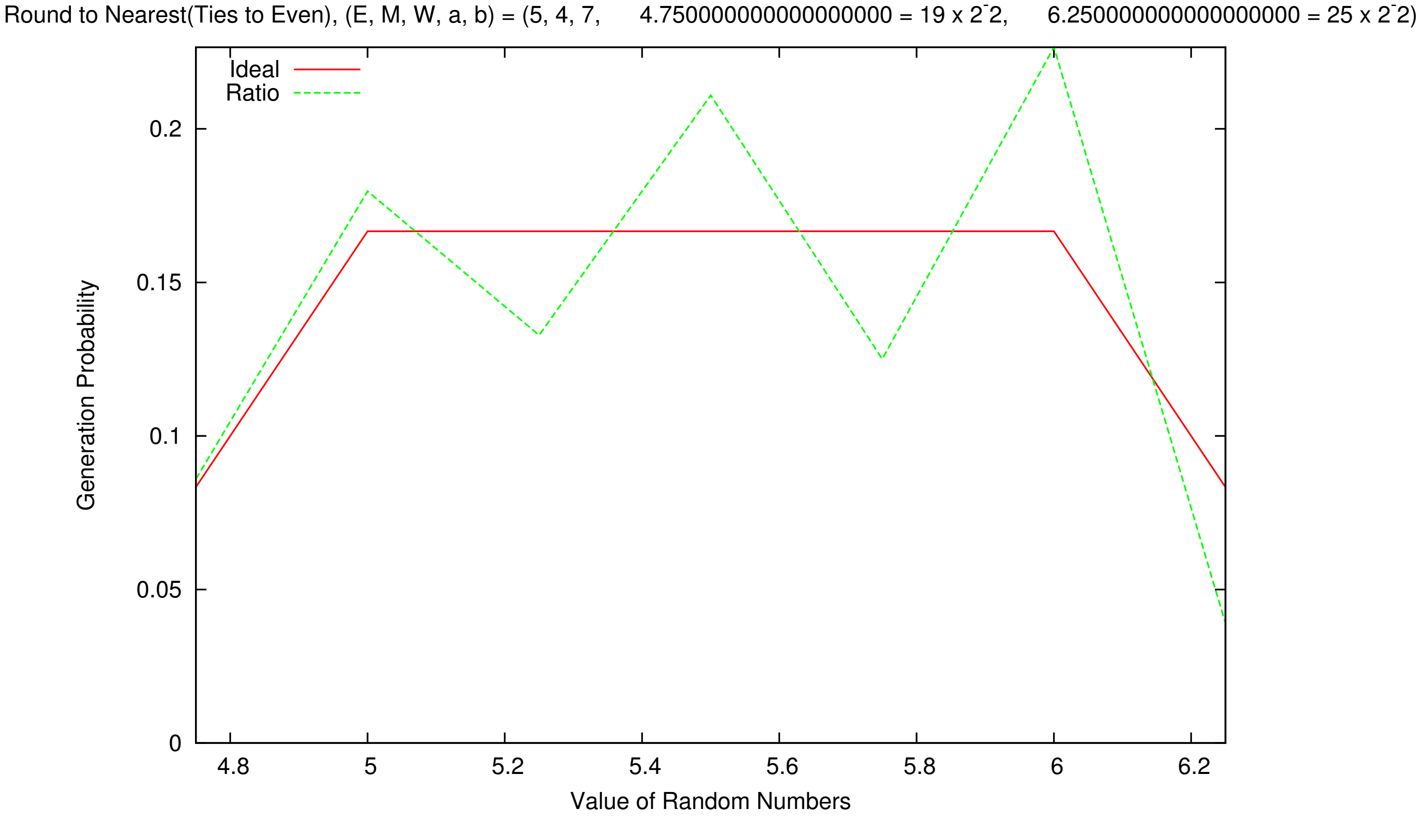


Fig. 21 Random number generation probability of extended Moler's method in $\left[\frac{3}{2}\pi, 2\pi\right]$.

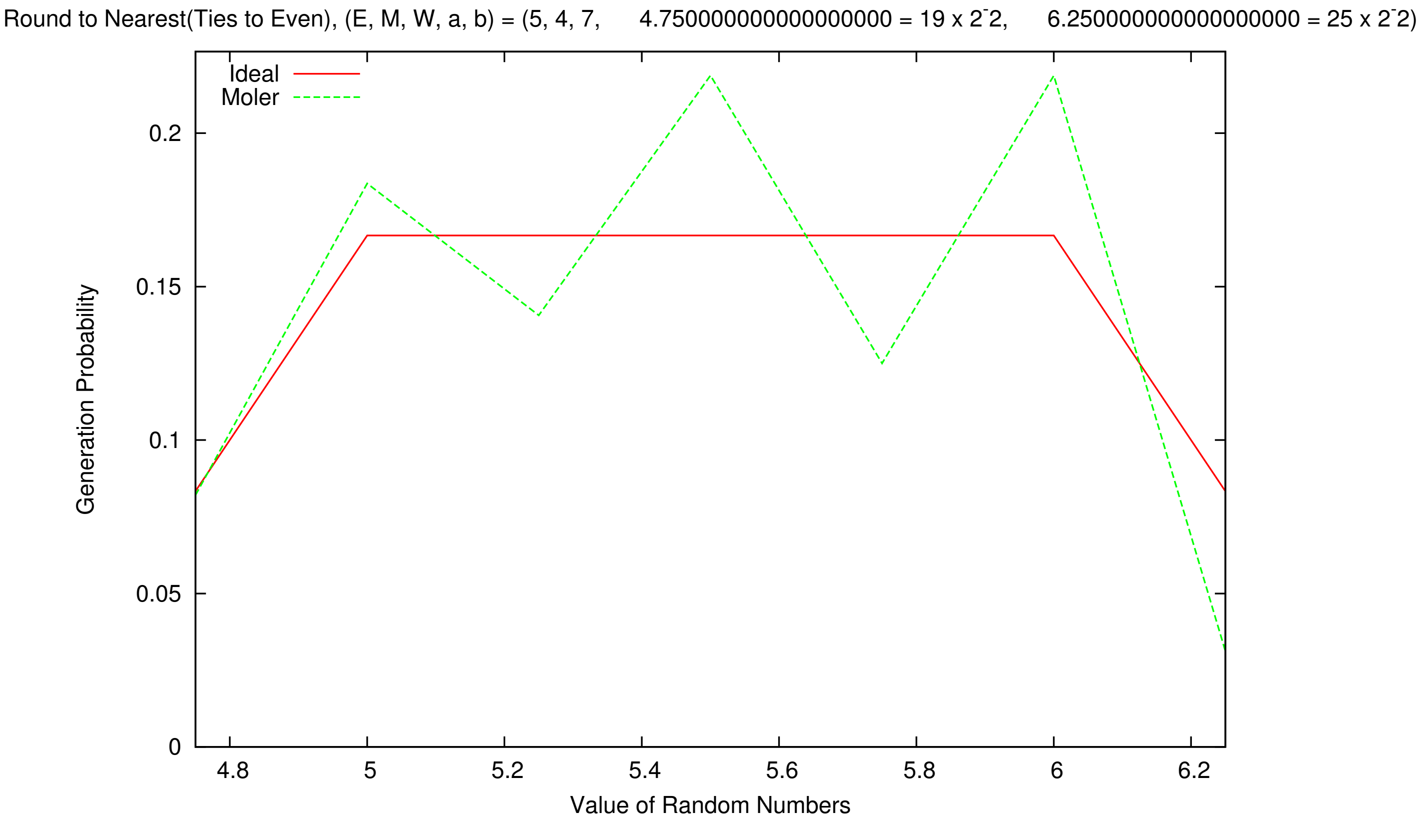


Fig. 22 Random number generation probability of extended Thoma's method in $\left[\frac{3}{2}\pi, 2\pi\right]$.

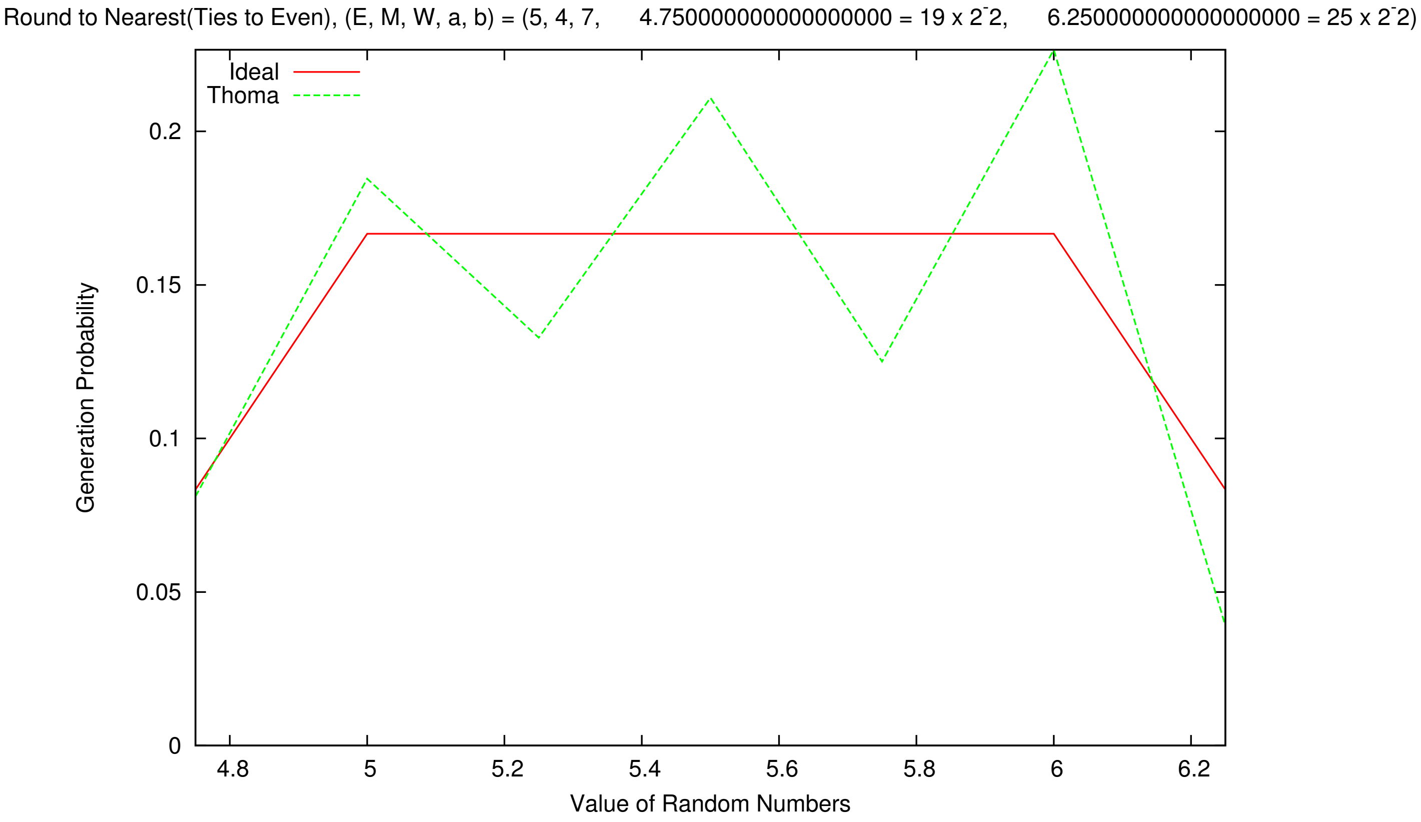


Fig. 23 Random number generation probability of the proposed method in $\left[\frac{3}{2}\pi, 2\pi\right]$.

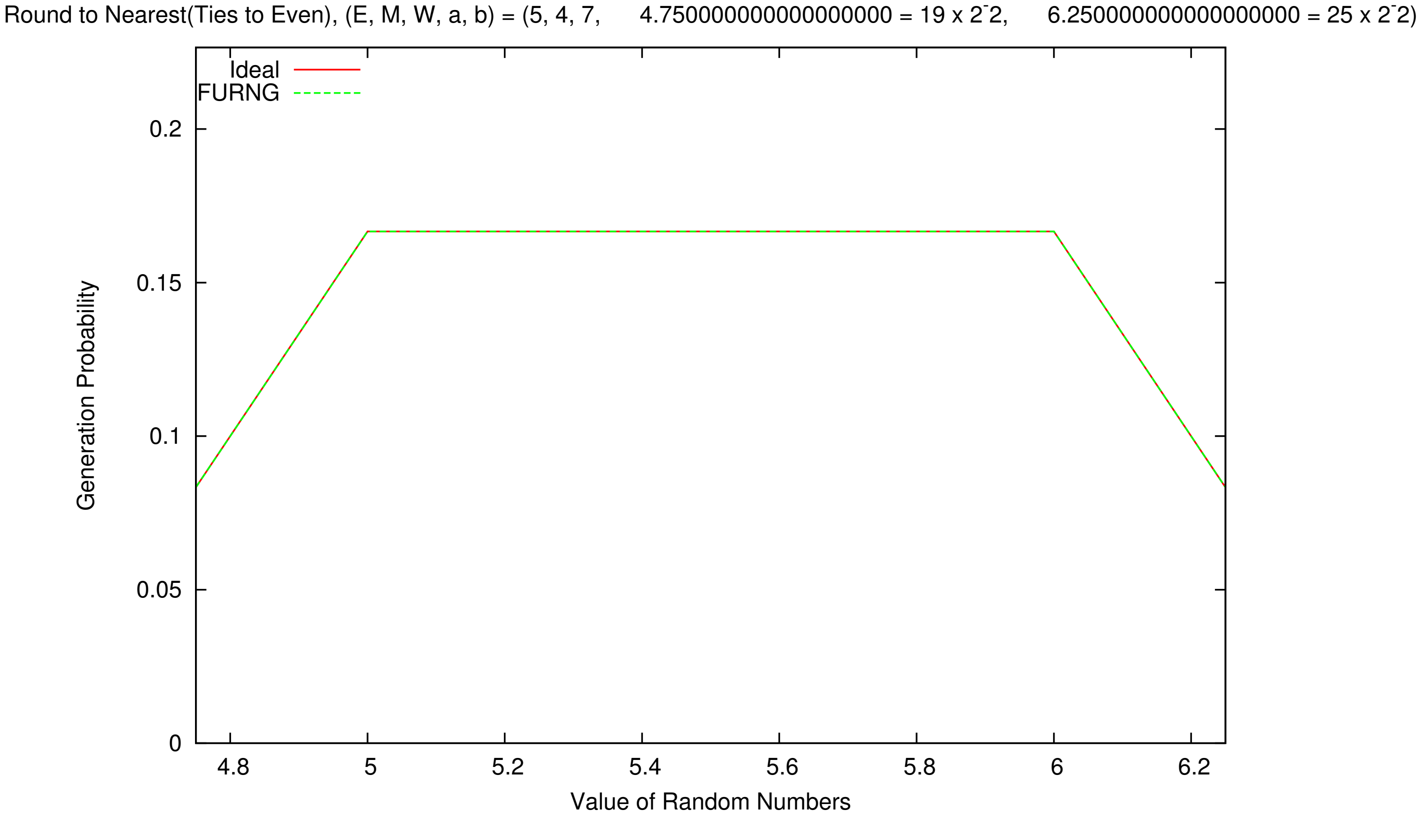


Fig. 24 Random number generation probability of Ratio method in $[-\pi, \pi]$.

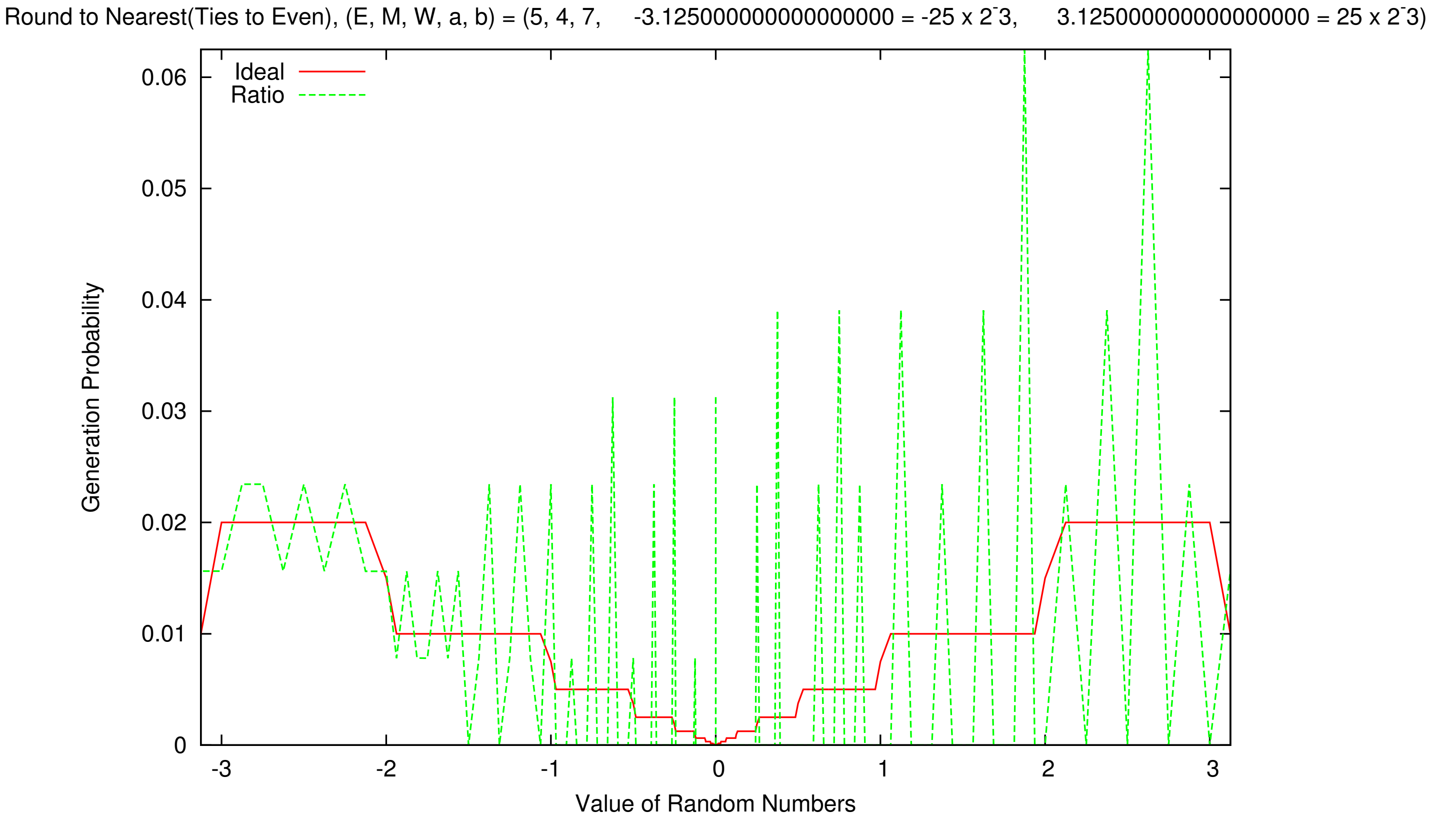


Fig. 25 Random number generation probability of extended Moler's method in $[-\pi, \pi]$.

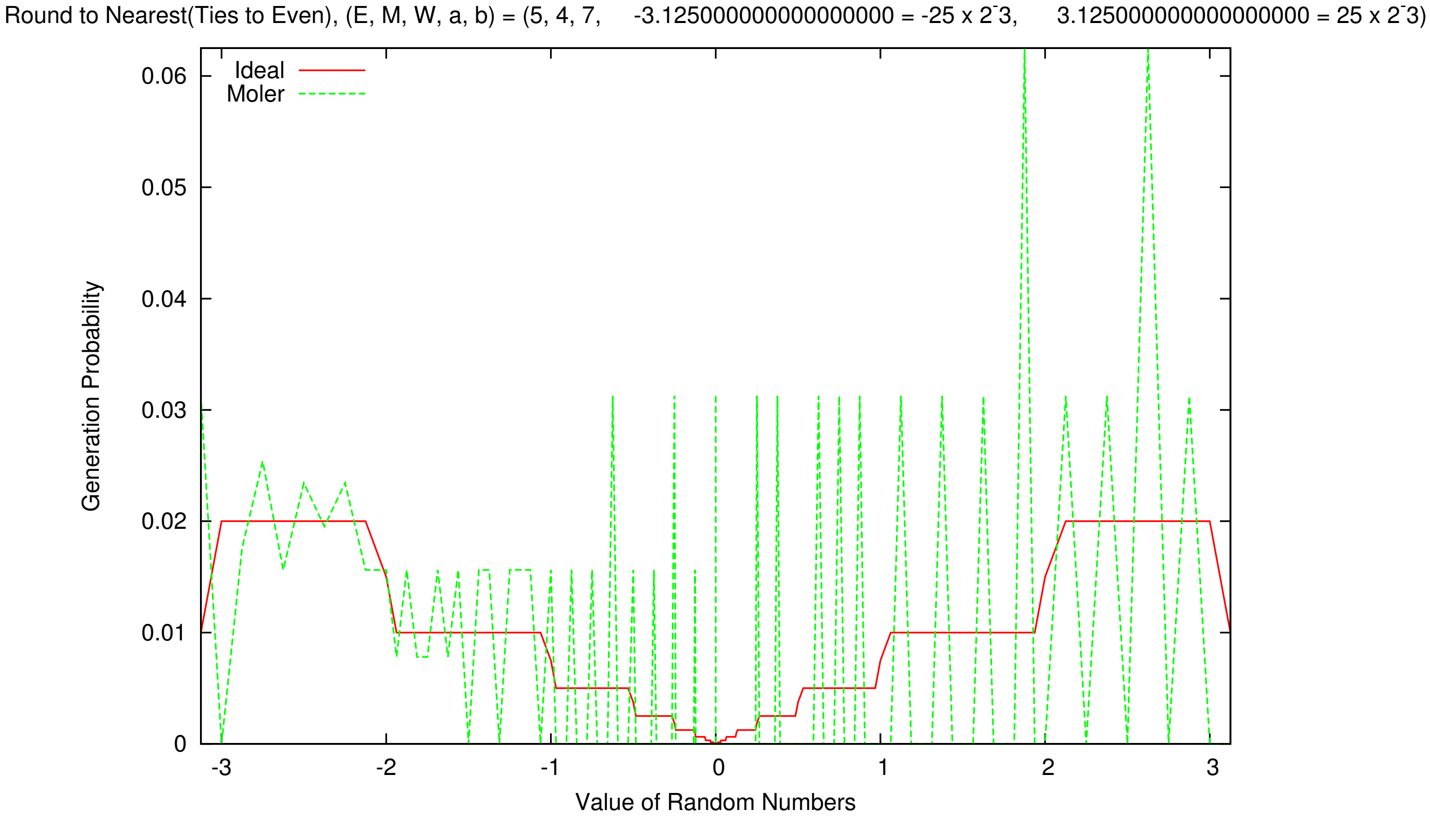
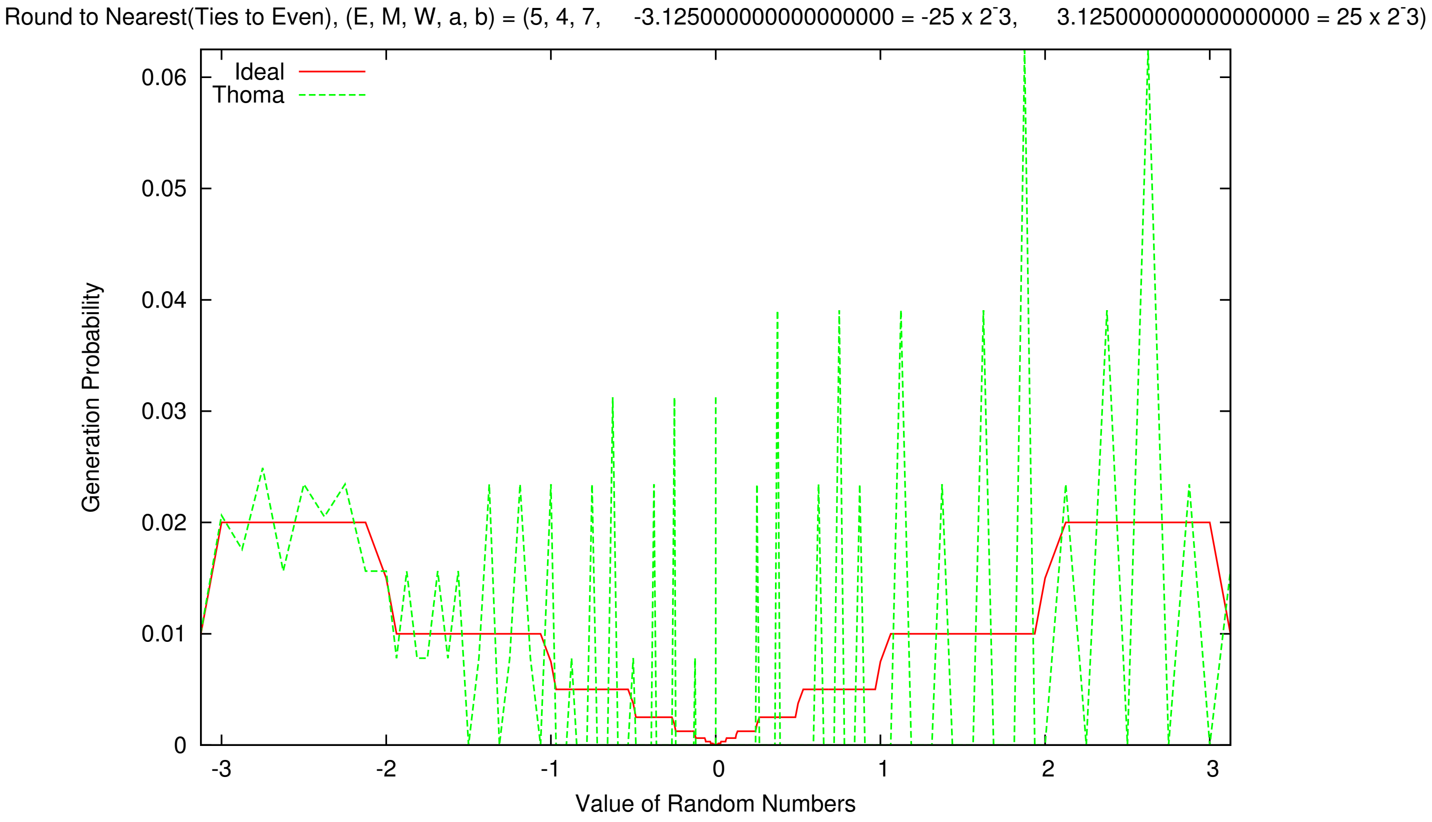


Fig. 26 Random number generation probability of extended Thoma's method in $[-\pi, \pi]$.



8. Conclusion

In this paper, the authors defined what uniform meant and proposed such a generator, proved its correctness, and shows

Fig. 27 Random number generation probability of the proposed method in $[-\pi, \pi]$.

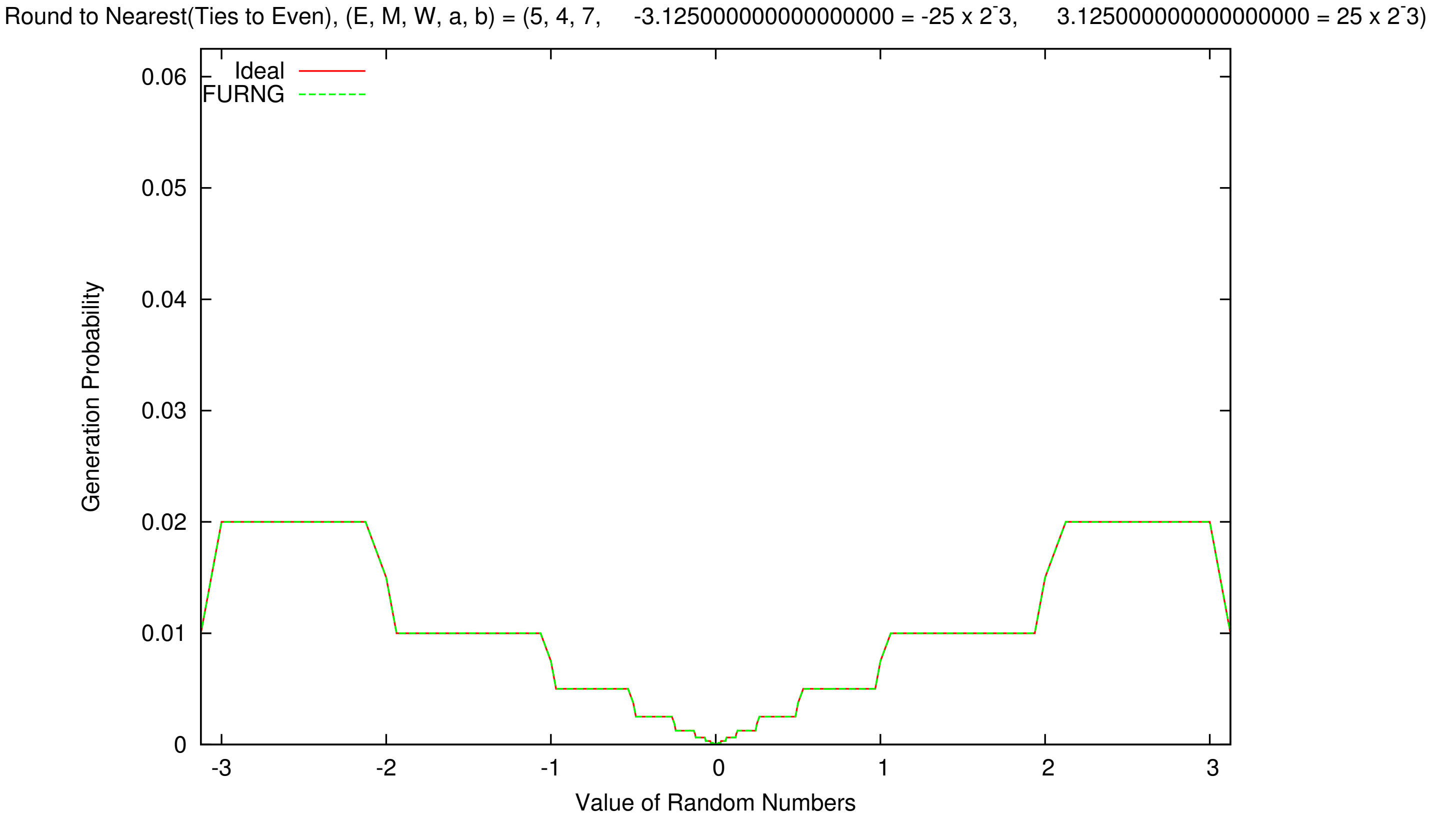
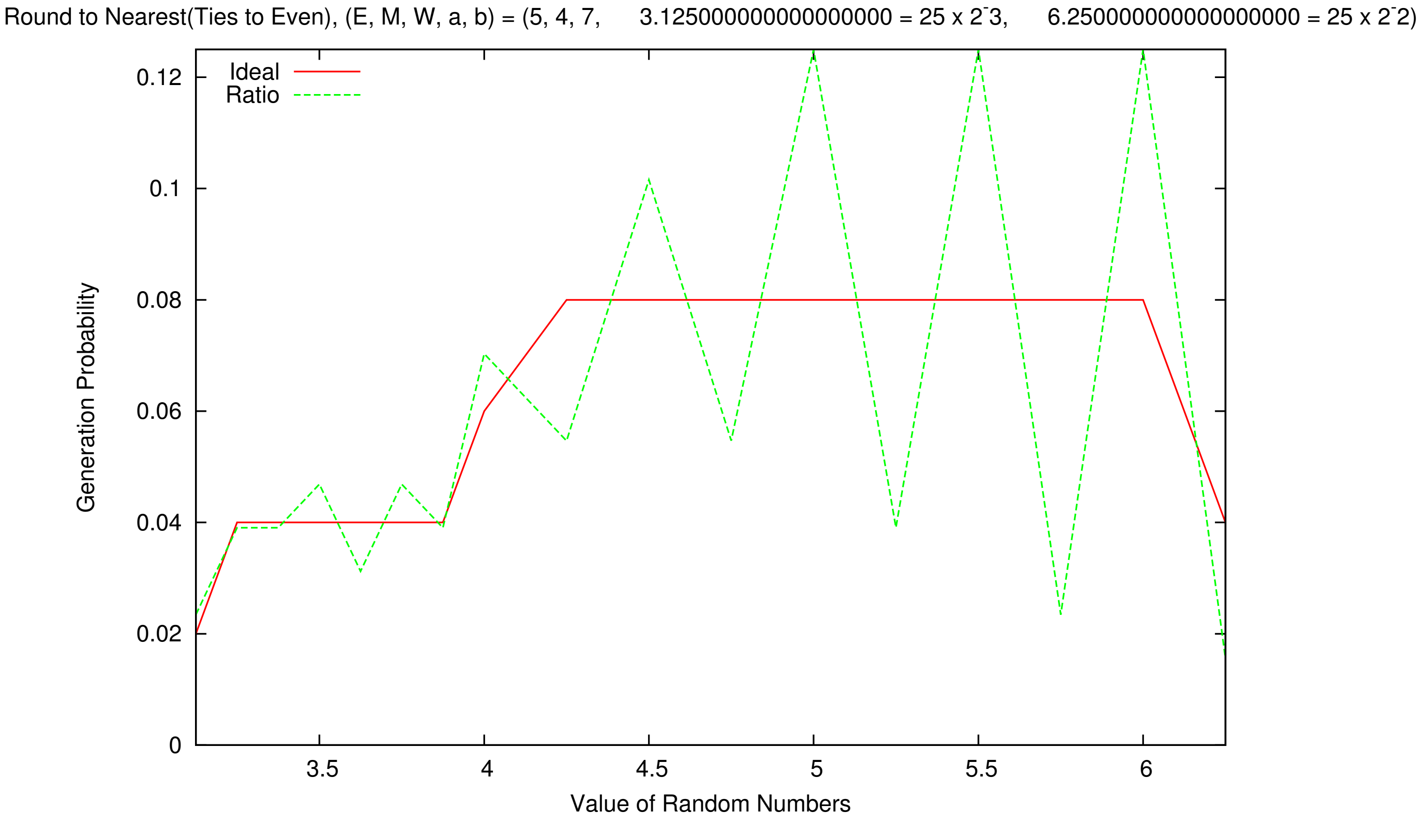


Fig. 28 Random number generation probability of Ratio method in $[\pi, 2\pi]$.



some experiments about its correctness and performance in order to modify the problem of Thoma's method and construct a generator that can output all the floating point numbers in arbitrary range. The advantages of the proposed method is that

Fig. 29 Random number generation probability of extended Moler's method in $[\pi, 2\pi]$.

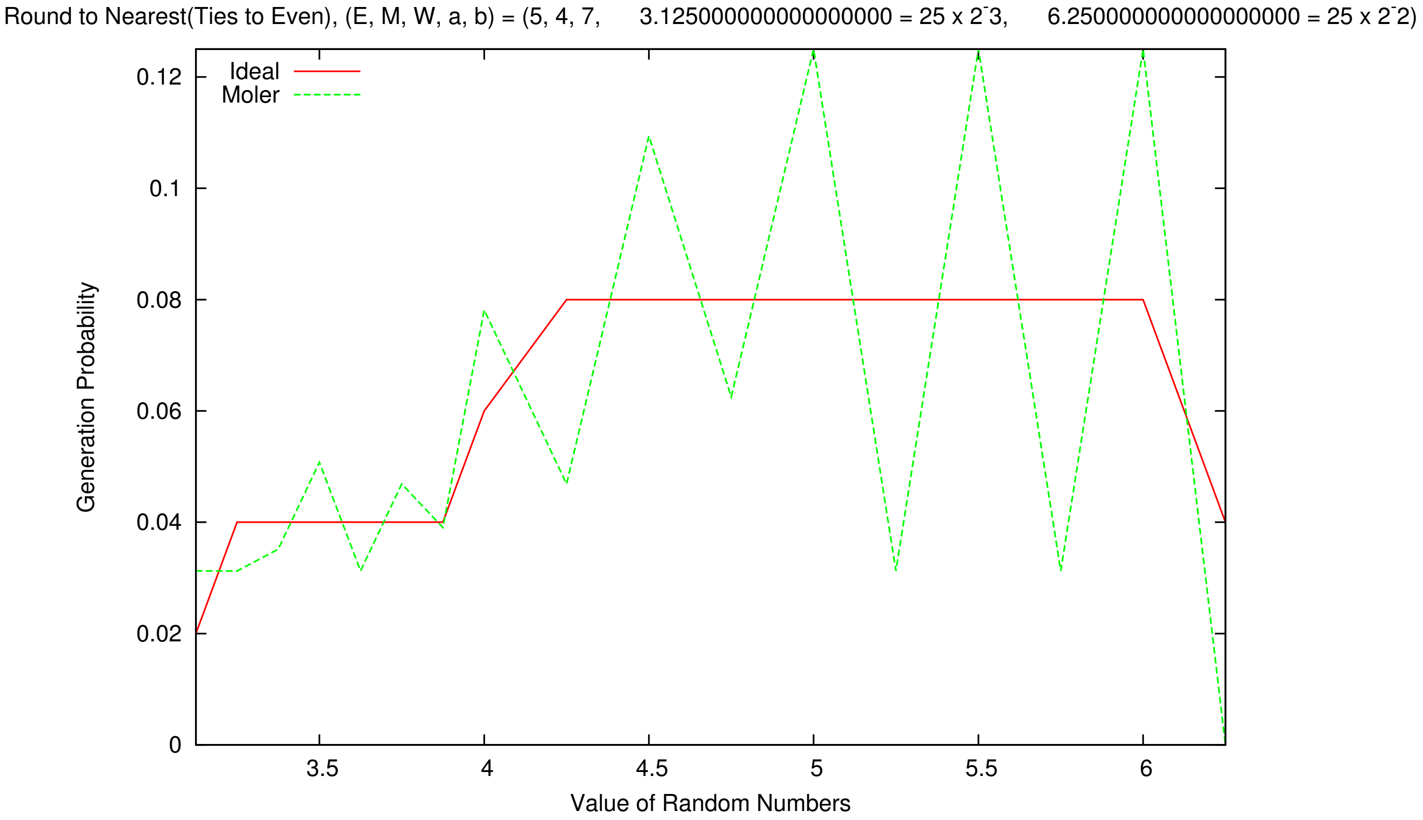
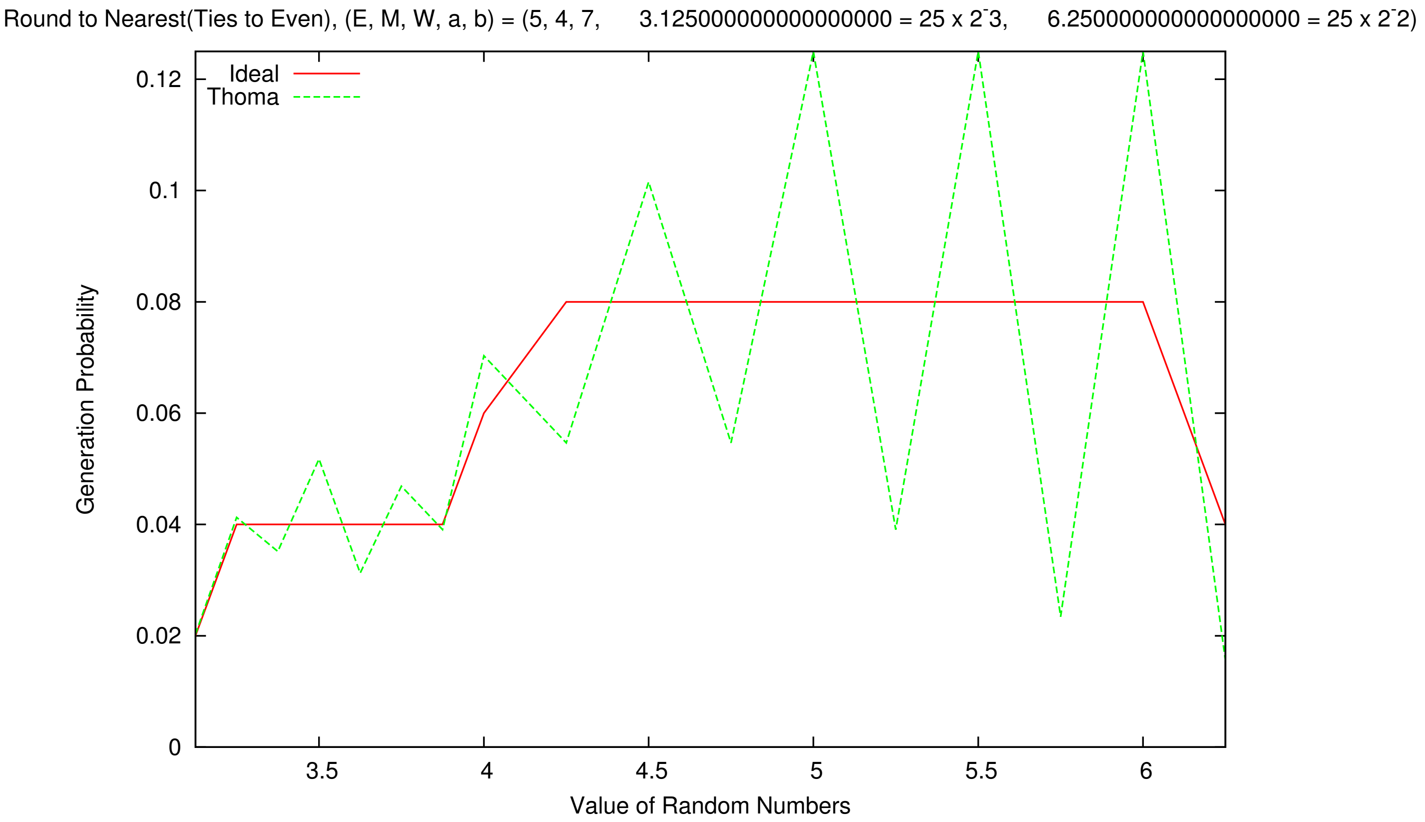


Fig. 30 Random number generation probability of extended Thoma's method in $[\pi, 2\pi]$.



the proposed method can output all the floating point in arbitrary range whose edge is a floating point number, and that FPU does not affect the random number generation probability by the proposed method, and that we can apply the proposed

Fig. 31 Random number generation probability of the proposed method in $[\pi, 2\pi]$.

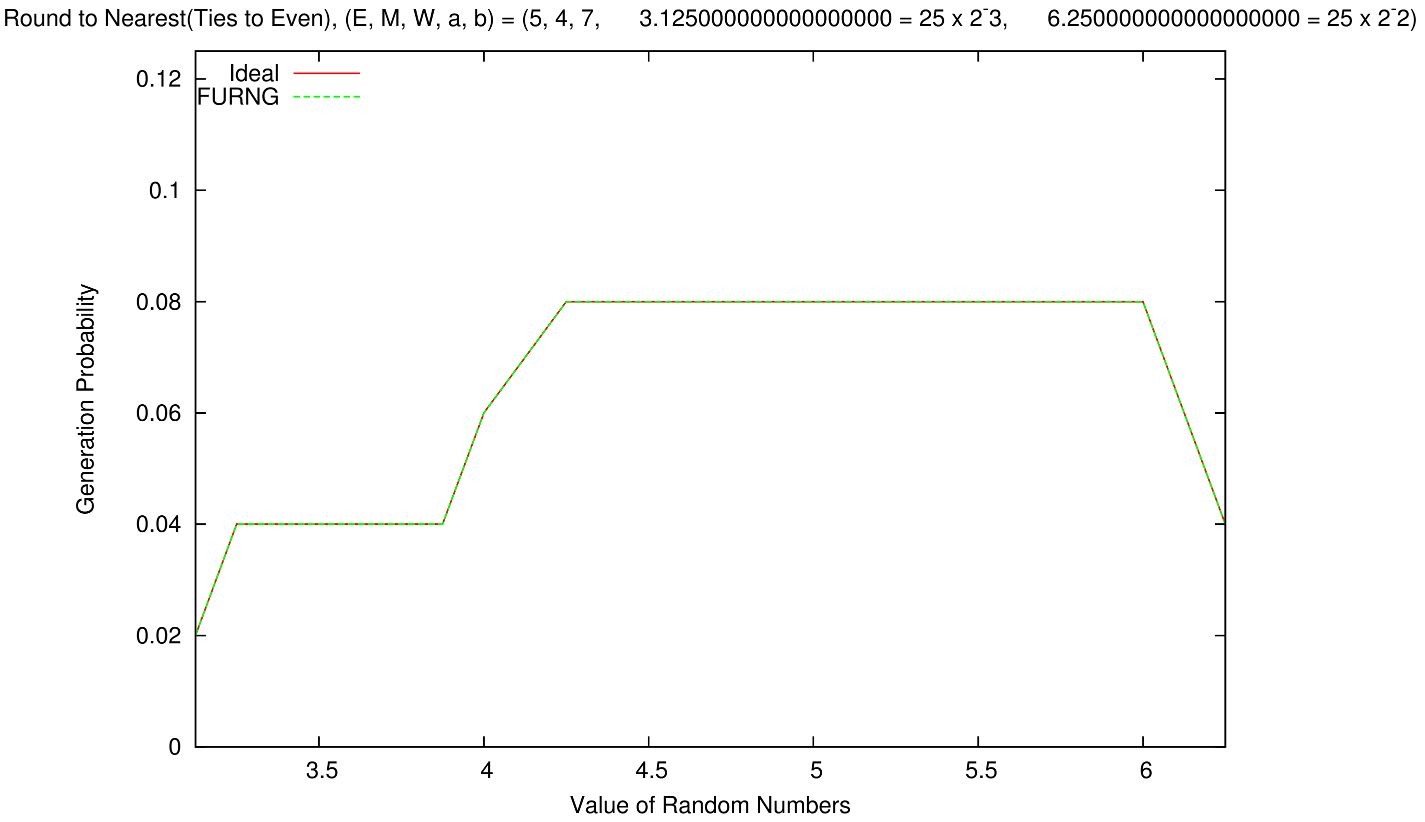
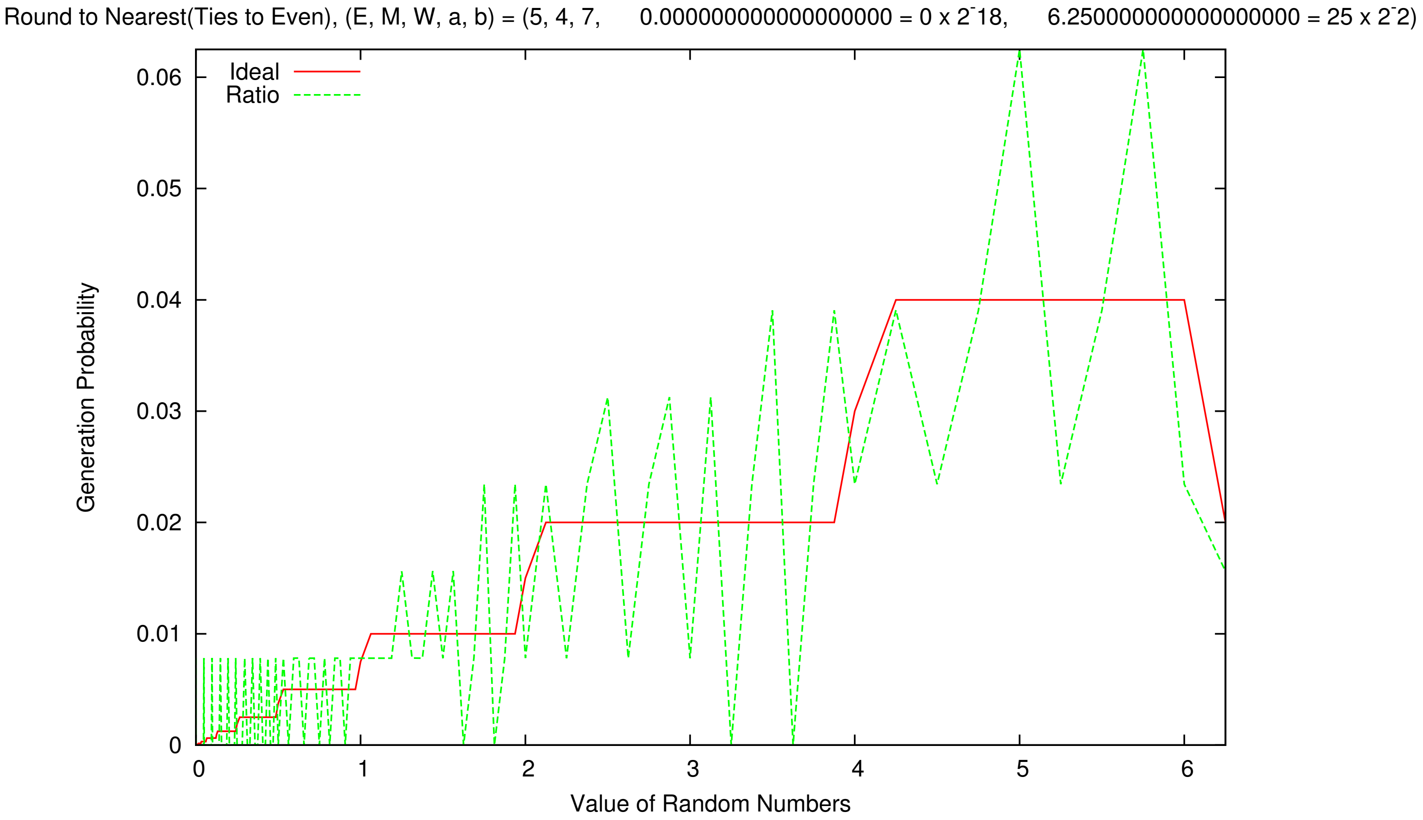


Fig. 32 Random number generation probability of Ratio method in $[0, 2\pi]$.



method to another precision floating point number by changing (E, M).

However, the research also has the following limitations.

Fig. 33 Random number generation probability of extended Moler's method in $[0, 2\pi]$.

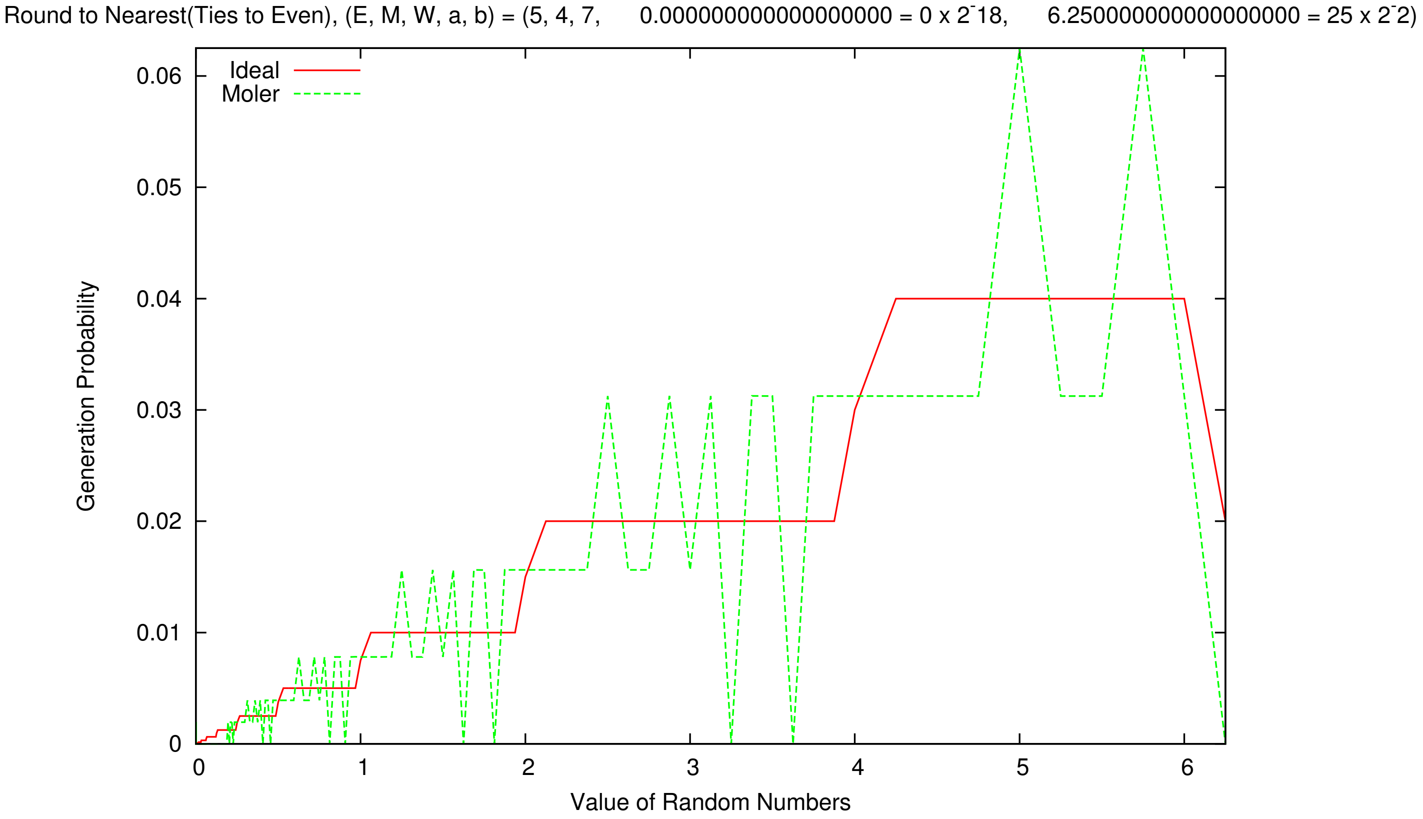
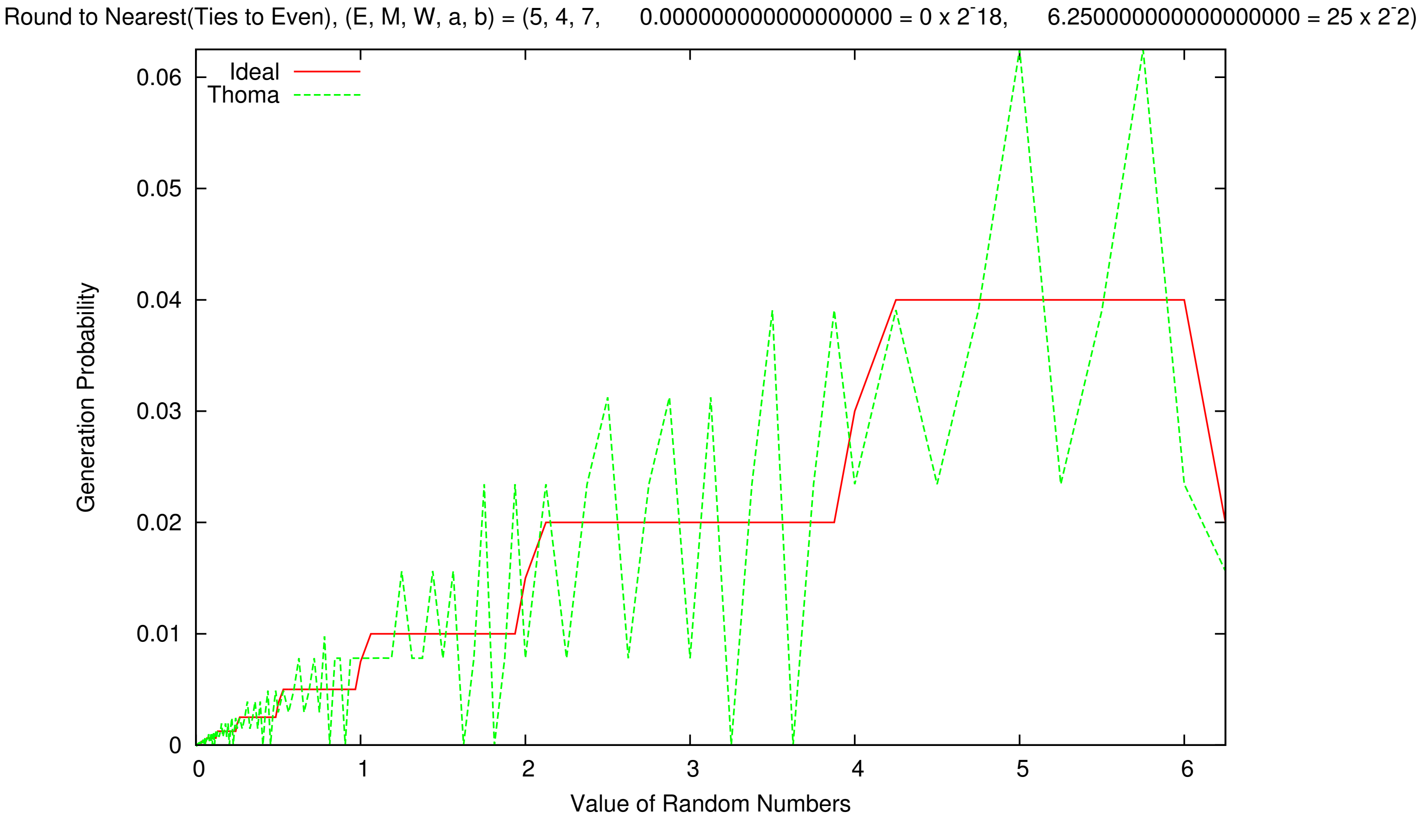
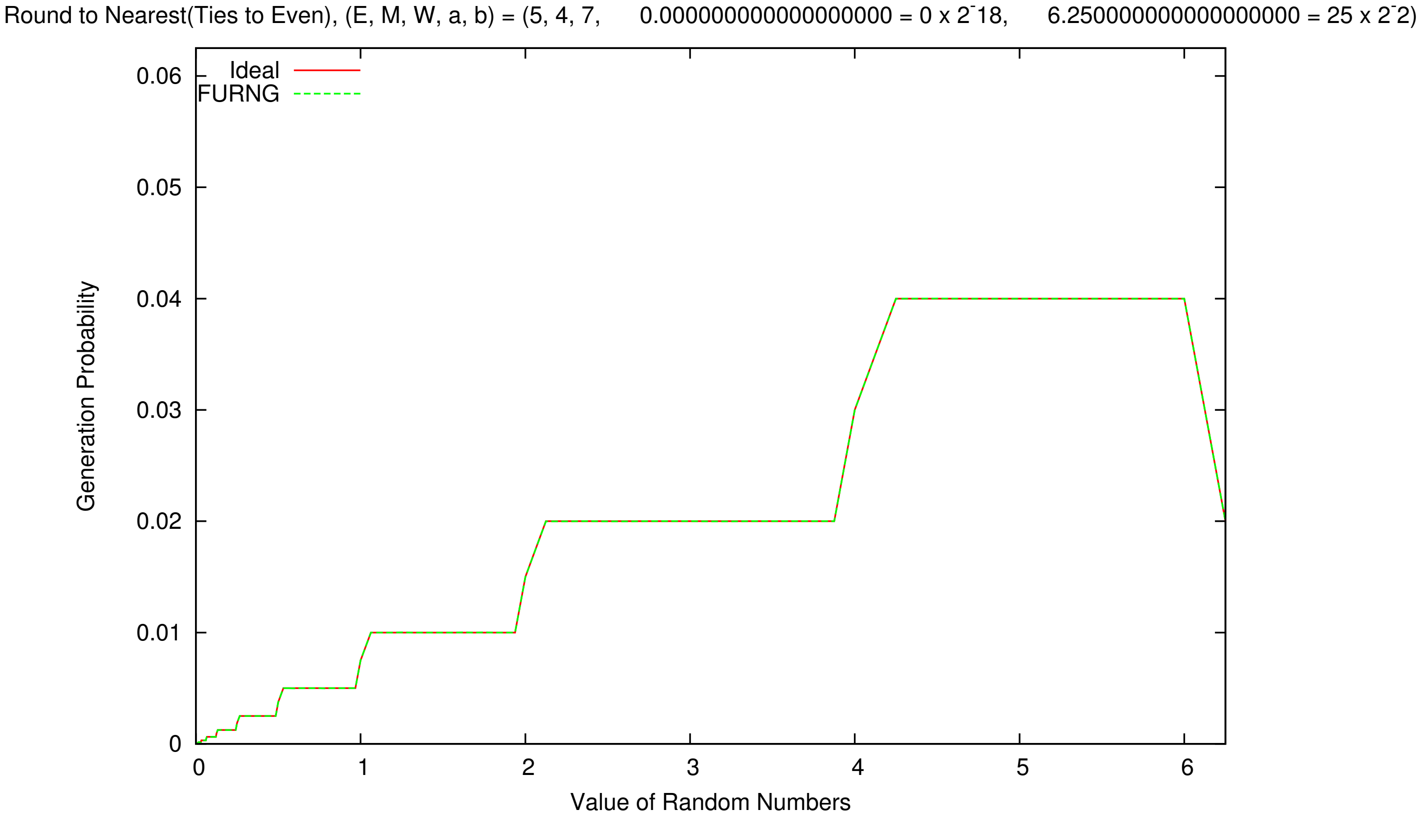


Fig. 34 Random number generation probability of extended Thoma's method in $[0, 2\pi]$.



The first example is that we can not receive so much advantage on IEEE754 double precision. For example, Moler's method

Fig. 35 Random number generation probability of the proposed method in $[0, 2\pi]$.



can generate almost all the floating point numbers[*36] in $\left[2^{-53}, 1-2^{-53}\right]$ without any problem[*37][*38]. Thus, we can receive advantages by the modified method only when the generator outputs a floating point number in $\left[0, 2^{-53}\right) \cup \left(1-2^{-53}, 1\right]$. However, the probability that we obtain such a floating point number is at most $2^{-53} \times 2 = 2^{-52}$. This probability is negligible for practical use.

The next example is that the rejection ratio of the proposed method can be about $\frac{5}{6}$. Now, consider the case where

$$\begin{aligned} U_{\mathbb{F}} &= [a, b] \\ &= \left[val_{\mathbb{F}}\left(0, 2^{E-1}-2, 2^{M-1}-1\right), val_{\mathbb{F}}\left(0, 2^{E-1}-1, 1\right)\right] \\ &= \left[0.75 - 2^{-(M+1)}, 1 + 2^{-M}\right]. \end{aligned}$$

In this case, the acceptance ratio γ is

$$\begin{aligned} \gamma &= \frac{2\left(b-a\right)}{3} \\ &= \frac{2\left(0.25 + 3 \times 2^{-(M+1)}\right)}{3} \\ &= \frac{1}{6} + 2^{-M}. \end{aligned}$$

Therefore, about 5 out of 6 iterations in the algorithm just wastes the execution time.

The last example is that no test other than χ^2 has been done. The authors confirmed that the proposed method passed the χ^2 test, but did not guarantee that the method also passed another test for uniform random number generator.

Now, our future work is as follows.

The first one is to apply the algorithm to double-double precision. One double-double [8] precision floating point number consists of two double precision number, hi and lo, and expresses a value by $hi + lo$. Hence, both quadruple precision and double-double precision requires 128 bits to express one value. Double-double precision, however, has a property that a value near a double precision number has quite higher precision than quadruple precision. For example, the quantity $1 - 2^{-1000}$ is rounded to 1 on quadruple precision. On the other hand, double-double precision can express this by $(hi, lo) = \left(1, 2^{-1000}\right)$. This property is useful when we would like to increase precision of uniform random number because we need to take some calculation, such as the inverse transformation method, to the random number.

The next future work is to increase the acceptance ratio or decreasing the execution time. As the result shown in the Section 7.3.2, the proposed method has at least the minimal speed for practical use but it can be more faster by increasing the acceptance ratio. However, too complex operation for increasing the acceptance ratio might take long execution time and then generation speed can be slower. So, we need to consider the best trade-off between increase of the acceptance ratio and simplification of the algorithm.

The third future work is to inspect the proposed method by another text than χ^2. There are several tests [6, 9, 10] for uniform random number generator. The followings are examples.

- Kolmogorov-Smirnov test [15, 30]
- Run-Length test [1]
- Autocorrelation test
- High-dimensionally equidistribution test
- Collision test
- Random-Walk test

The last future work is to find a concrete application of the proposed method. The proposed method can be useful for transforming uniform random number into random number that follows another distribution [12, 16, 27] and for generating random number from the tail region[*39] [7].

References

[1] James H. Andrews, Alex Groce, Melissa Weston, and Ru-Gang Xu. Random test run length and effectiveness. In *23rd IEEE/ACM International Conference on Automated Software Engineering (ASE 2008), 15-19 September 2008, L'Aquila, Italy*, pages 19–28, 2008.

[2] James R. Bell. Algorithm 334: Normal random deviates. *Commun. ACM*, 11(7):498–, Jul 1968.

[3] L Blum, M Blum, and M Shub. A simple unpredictable pseudo random number generator. *SIAM J. Comput.*, 15(2):364–383, May

[*36] All the floating point numbers except its mantissa is 0.

[*37] The random number generation probability is uniform in narrow sense, that is, the Formula (1) is satisfied by $round_{\mathbb{F}}$ =Round-to-Nearest(Ties to Even).

[*38] The Formula (1) is also satisfied when the mantissa is 0, but $round_{\mathbb{F}}$ is not Round-to-Nearest(Ties to Even). That is, the random number generation probability is uniform in wide sense.

[*39] The tail of the distribution denotes each edge of the distribution.

1986.

[4] G. E. P. Box and Mervin E. Muller. A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611, 06 1958.

[5] Richard P. Brent. Algorithm 488: A gaussian pseudo-random number generator. *Commun. ACM*, 17(12):704–706, Dec 1974.

[6] Donald E.Knuth. *The Art of Computer Programming Volume 2 Seminumerical Algorithms Third Edition.* Addison-Wesley, 1997.

[7] Daniel Fulger, Enrico Scalas, and Guido Germano. Random numbers from the tails of probability distributions using the transformation method. *CoRR*, abs/0902.3207, 2009.

[8] Yozo Hida, Xiaoye S. Li, and David H. Bailey. Library for double-double and quad-double arithmetic. 2007.

[9] Information Technology Laboratory, National Institute of Standards and Technology. *Federal Information Processing Standards Publication 140-2. Security requirements for cryptographic modules*, 2001.

[10] Japanese Standards Association. *JIS Z 9031:2012. Procedure for random number generation and randomization*, 1956.

[11] Peter Kabal. Generating gaussian pseudo-random deviates. Technical report, Department of Electrical and Computer Engineering McGill University, 2000.

[12] Tanaka Ken ' ichiro and Alexis Akira Toda. Discrete approximations of continuous distributions by maximum entropy. *Economics Letters*, 118(3):445–450, 2013.

[13] A. J. Kinderman and J. F. Monahan. Computer generation of random variables using the ratio of uniform deviates. *ACM Trans. Math. Softw.*, 3(3):257–260, Sep 1977.

[14] R. Knop. Remark on algorithm 334 [g5]: Normal random deviates. *Commun. ACM*, 12(5):281–, May 1969.

[15] Andrey N Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell ' Istituto Italiano degli Attuari*, 4(1):83–91, 1933.

[16] A. Luceño. Discrete approximations to continuous univariate distributions-an alternative to simulation. *Journal of the Royal Statistical Society Series B*, 61(2):345–352, 1999.

[17] G. Marsaglia and T. A. Bray. A convenient method for generating normal variables. *SIAM Review*, 6(3):pp. 260–264, 1964.

[18] George Marsaglia. Random numbers fall mainly in the planes. Report, August 1963.

[19] George Marsaglia. Xorshift rngs. *Journal of Statistical Software*, 08(i14), 2003.

[20] George Marsaglia and Wai Wan Tsang. The monty python method for generating random variables. *ACM Trans. Math. Softw.*, 24(3):341–350, Sep 1998.

[21] George Marsaglia and Wai Wan Tsang. The ziggurat method for generating random variables. *Journal of Statistical Software*, 5(8):1–7, 10 2000.

[22] George Marsaglia and Arif Zaman. A new class of random number generators. 1(3):462–480, August 1991.

[23] Makoto Matsumoto and Takuji Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30, Jan 1998.

[24] Nicholas Metropolis and Stanislaw M. Ulam. The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341, September 1949.

[25] Cleve B. Moler. Random thoughts: 10^{435} years is a very long time. Technical note, inst-MATHWORKS, inst-MATHWORKS:adr, Fall 1995.

[26] Cleve B. Moler. *Numerical Computing with MATLAB*, chapter 9. Society for Industrial and Applied Mathematics, 2004.

[27] Harry H. Panjer. *Operational Risk: Modeling Analytics*, pages 411–414. Wiley-Interscience, 2006.

[28] W. H. Payne, J. R. Rabung, and T. P. Bogyo. Coding the lehmer pseudo-random number generator. *Commun. ACM*, 12(2):85–86, February 1969.

[29] John K. Salmon, Mark A. Moraes, Ron O. Dror, and David E. Shaw. Parallel random numbers: As easy as 1, 2, 3. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 16:1–16:12, New York, NY, USA, 2011. ACM.

[30] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. 19(2):279–281, June 1948.

[31] David B. Thoma, Wayne Luk, Philip H.W. Leong, and John D. Villasenor. Gaussian random number generators. *ACM Comput. Surv.*, 39(4), Nov 2007.

[32] John von Neumann. 13. various techniques used in connection with random digits. *Journal of Research of the National Bureau of Standards. Applied Mathematics Series*, 12:36–38, 1951.

[33] John von Neumann. Various techniques used in connection with random digits. *J. Res. Nat. Bur. Stand.*, 12:36–38, 1951.

Appendix

A.1 Additional table

A.1.1 For Section 2.1

Table A·1 shows the number of bits of exponent and mantissa, minimal value of normal number[*40], minimal value of positive number, and maximal value for the most used IEEE754 format floating point number.

Table A·1 Detailed information about IEEE754 floating point number

Precision	Single	Double	Quadruple
#bit of sign	1	1	1
#bit of exponent(E)	8	11	15
#bit of mantissa(M)	23	52	112
Total #bit	32	64	128
Minimal value of normal number $2^{1-\left(2^{E-1}-1\right)}$	2^{-126} $\sim 10^{-38}$	2^{-1022} $\sim 10^{-308}$	2^{-16382} $\sim 10^{-4932}$
Minimal value of positive number $2^{1-\left(M+2^{E-1}-1\right)}$	2^{-149} $\sim 10^{-45}$	2^{-1074} $\sim 10^{-324}$	2^{-16494} $\sim 10^{-4966}$
Maximal value $\left(1-2^{-M}\right)\times 2^{2^{E-1}-1}$	$\sim 2^{127}$ $\sim 10^{38}$	$\sim 2^{1023}$ $\sim 10^{307}$	$\sim 2^{16383}$ $\sim 10^{4931}$

[*40] This value divides floating point numbers into subnormal numbers and normal numbers.

A.1.2 For Section 3.2

Table A·2 shows the random number generation probability of Thoma's method, that of uniform generator defined in Section 4, and the ratio of these.

Table A·2 Random number generation probability of Thoma's method and that of ideal generator where rounding mode is Round-to-Nearest(Ties to Even) and $(E, M, W) = (4, 3, 5)$.

Value of random number	Thoma's method	Ideal	Ratio
0.000×2^{-6}	32/1024	1/1024	32.0
0.125×2^{-6}	0/1024	2/1024	0.00
⋮	⋮	⋮	⋮
1.875×2^{-6}	0/1024	2/1024	0.00
1.000×2^{-5}	4/1024	3/1024	1.33
1.125×2^{-5}	2/1024	4/1024	0.50
1.250×2^{-5}	6/1024	4/1024	1.50
1.375×2^{-5}	2/1024	4/1024	0.50
1.500×2^{-5}	6/1024	4/1024	1.50
1.625×2^{-5}	2/1024	4/1024	0.50
1.750×2^{-5}	6/1024	4/1024	1.50
1.875×2^{-5}	2/1024	4/1024	0.50
1.000×2^{-4}	10/1024	6/1024	1.67
1.125×2^{-4}	4/1024	8/1024	0.50
1.250×2^{-4}	12/1024	8/1024	1.50
1.375×2^{-4}	4/1024	8/1024	0.50
1.500×2^{-4}	12/1024	8/1024	1.50
1.625×2^{-4}	4/1024	8/1024	0.50
1.750×2^{-4}	12/1024	8/1024	1.50
1.875×2^{-4}	4/1024	8/1024	0.50
1.000×2^{-3}	20/1024	12/1024	1.67
1.125×2^{-3}	8/1024	16/1024	0.50
1.375×2^{-3}	8/1024	16/1024	0.50
1.500×2^{-3}	24/1024	16/1024	1.50
1.625×2^{-3}	8/1024	16/1024	0.50
1.750×2^{-3}	24/1024	16/1024	1.50
1.875×2^{-3}	8/1024	16/1024	0.50
1.000×2^{-2}	40/1024	24/1024	1.67
1.125×2^{-2}	32/1024	32/1024	1.00
⋮	⋮	⋮	⋮
1.875×2^{-2}	32/1024	32/1024	1.00
1.000×2^{-1}	64/1024	48/1024	1.33
1.125×2^{-1}	32/1024	64/1024	0.50
1.250×2^{-1}	96/1024	64/1024	1.50
1.375×2^{-1}	32/1024	64/1024	0.50
1.500×2^{-1}	96/1024	64/1024	1.50
1.625×2^{-1}	32/1024	64/1024	0.50
1.750×2^{-1}	96/1024	64/1024	1.50
1.875×2^{-1}	32/1024	64/1024	0.50
1.000×2^{-0}	32/1024	32/1024	1.00