

FLOATING-POINT NUMBER SOLUTIONS IN A SIMPLE LINEAR EQUATION WITH ADDITION ALGORITHM

辻 久美子 (Kumiko Tsuji) 九州帝京短期大学経営情報科

Abstract

A model of a floating-point addition $u \oplus v = c$ is established in FORTRAN language. Here the exponent of u is greater than the exponent of v . The two kinds of simple linear equations $y \oplus v = c$ and $u \oplus x = c$ are solved theoretically. Here u, v and c are given floating-point numbers and y and x are unknown floating-point numbers. The affection of round-off error arising from algorithm is analysed to the error of the solutions for two kinds of the linear equations. It is shown that the solution y is machine precision accuracy and x is not always machine precision accuracy.

MODEL OF A FLOATING-POINT ADDITION AND DOMAIN FOR INPUTS

Let $d[e-1, e-t >$ denote a floating-point number with exponent e which has the positions from $e-1$ to $e-t$ with the bit $d(k)$ at k position for $k = e-1, \dots, e-t$. Here $d(e-1) = 1$ and $d(k) = 1$ or 0 for $k = e-2, \dots, e-t$. Let T be a set of floating-point numbers of length $t \geq 3$: $T = \{d[e-1, e-t >; -\infty < e < \infty, d > 0\}$. A model algorithm of $u \oplus v$ is defined as a mapping from $T \times T$ to T for a pair (u, v) in $T \times T$. The algorithm is defined as follows: (1) v is correctly rounded so as to the last significant position is $e(u) - t$ and this value is denoted as v_c : $v_c = v[e(v) - 1, e(u) - t)) + v(e(u) - t - 1)b^{e(u)-t}$. (2) u and v_c are added: $u + v_c =$

$$u[e(u) - 1, e(u) - t > + v[e(v) - 1, e(u) - t)) + v(e(u) - t - 1)b^{e(u)-t}.$$

Here $v[e(v) - 1, e(u) - t))$ is the sub power series in positions from $e(v) - 1$ to $e(u) - t$. Our algorithm is the case that $e(u + v_c) = e(u) + 1$. $u + v_c$ is $t + 1$ length, $u + v_c$ is chopped to t length. The floating-point addition is given as

$$u \oplus v = u[e(u) - 1, e(u) - t + 1)) + v[e(v) - 1, e(u) - t + 1)) \\ + C(u(e(u) - t), v(e(u) - t), v(e(u) - t - 1))b^{e(u)-t+1},$$

where $C(x, y, z)$ is the carry to heigher poisition in $x + y + z$ defined as $xy \vee (x \oplus y)z$. Here xy , $x \vee y$ and $x \oplus y$ denote respectively AND, OR and Exclusive-OR operations in Boolean functions.

The algorithm is introduced for one case such that

$$e(u + v_c) = e(u) + 1, e(v) + t - 2 \geq e(u) \geq e(v) + 1, uv > 0.$$

Let E denote the domain for inputs u and v , and then

$$E = \{(u, v) : e(u + v_c) = e + 1, 2 \leq i \leq t - 1, uv > 0\},$$

putting $e(u) = e$, $e(c) = e + 1$ and $e(v) = e - i + 1$.

ROUND-OFF ERROR OF THE FLOATING-POINT ADDITION

The round-off error of the floating-point addition is defined as $u + v - u \oplus v$ and denoted as $\delta(u, v)$. $\delta(u, v)$ is calculated as

$$\delta(u, v) = v((e - t, e - i + 1 - t > \\ - C(u(e - t), v(e - t), v(e - t - 1))b^{e-t+1} + u(e - t)b^{e-t}.$$

$v((e - t, e - i + 1 - t >$ denote a sub power series of v in positions from $e - t$ to $e - i + 1 - t$. Since $\delta(u, v)$ depends only on $u(e - t)$, $\delta(u, v)$ is also denoted as $\delta'(u(e - t), v)$.

TRANPOSED EQUATION $y \oplus v = c$

The linear equation $y + v = c$ is transposed on a computer as $y \oplus v = c$. The transposed equation $y \oplus v = c$ has two solutions:

$$y[e-1, e-t] \geq c - v[e-i, e-t+1]) + y(e-t)b^{e-t} \\ - C(y(e-t), v(e-t), v(e-t-1))b^{e+1-t},$$

corresponding to the bits $y(e-t) = 0, 1$ in the least significant position of y . Here (v, c) is in the trapezoid

$$S(i, 0) = \{b^{e-i+1} - b^{e-i+1-t} \geq v \geq b^{e-i}, \\ b^e - b^{e+1-t} + G(0, v) \geq c \geq b^e\}.$$

$G(j, v)$ is a rounding step function in v which is a monotone increasing step function with step width $p = b^{e+1-t}$ for $j = 0, 1$:

$$G(j, v) = v[e-i, e-i-t+1]) + C(j, v(e-t), v(e-t-1))b^{e+1-t}.$$

For (v, c) in the set

$$S(i, 1) - S(i, 0) \\ = \{v(e-t) \oplus v(e-t-1) = 1, c = b^e + v[e-i, e-t+1])\},$$

the equation $u \oplus v = c$ has one solution corresponding to $y(e-t) = 1$.

TRANPOSED EQUATION $u \oplus x = c$

The linear equation $u + x = c$ is transposed on a computer as $u \oplus x = c$. The transposed equation $u \oplus x = c$ has at most 2^{i-t} solutions corresponding to the ways of choosing the bits in $x((e-t, e-i+1-t) > :$

$$x[e-i, e-i+1-t] \geq c[e, e+1-t] - u[e-1, e+1-t]) \\ - C(u(e-t), x(e-t), x(e-t-1))b^{e-t+1} + x((e-t, e-i+1-t) > .$$

The following theorem shows the number of solutions x for $u \oplus x = c$.

Theorem 1 *Let i and n be integers such that*

$$c[e, e-t+1] > -u[e-1, e-t+1]) = b^{e-i} + nb^{e-t+1}.$$

Let IN be an integer such that

$$IN = b^{t-2} - u((e-2, e+1-t))b^{-e-1+t} - b^{t-i-1}.$$

Let NUM denote the number of the solutions of $u \oplus x = c$. Then NUM is given as follows.

1. *If $(u, c) \in S'_1(i)$ then $0 \leq n \leq b^{t-i-1}$ and*

$$NUM = (1-j)b^{i-1} + b^{i-2} \text{ for } n = 0;$$

$$NUM = b^i \text{ for } 1 \leq n \leq b^{t-i-1} - 1;$$

$$NUM = b^{i-2} + jb^{i-1} \text{ for } n = b^{t-i-1}.$$

Here $S'_1(i)$ is the trapezoid

$$S'_1(i) = \{b^{e-i+1} \geq c - u[e-1, e-t+1]) \geq b^{e-i}, \\ b^e - b^{e-t} \geq u \geq b^e - b^{e-i}\}.$$

2. If $(u, c) \in S'_2(i)$ then $IN \leq n \leq b^{t-i-1}$ and

$$NUM = b^i \text{ for } b^{t-i-1} - 1 \geq n \geq IN;$$

$$NUM = b^{i-2} + jb^{i-1} \text{ for } n = b^{t-i-1}.$$

Here $S'_2(i)$ is the trapesoid

$$S'_2(i) = \{b^{e-i+1} + u[e-1, e-t+1]) \geq c \geq b^e, \\ b^e - b^{e-i} - b^{e-t} \geq u \geq b^e - b^{e-i+1} + b^{e+1-t}\}.$$

3. If $c = b^e$ and $u = b^e - b^{e-i+1} + jb^{e-t}$ with $j = 0$ or $j = 1$, then $IN = b^{t-i-1}$ and

$$NUM = b^{i-2} + jb^{i-1}.$$

COMPARISON OF ROUND-OFF ERROR FOR TWO SOLUTIONS FOR $u \oplus x = c$ and $y \oplus v = c$

Since $u \oplus x = c$,

$$\varepsilon(u, x) = (c - u) - x = (u \oplus x - u) - x \\ = -\delta'(u, x) = -\delta'(u(e-t), x).$$

Since $y \oplus v = c$,

$$\varepsilon(y, v) = (c - v) - y = (y \oplus v - v) - y \\ = -\delta(y, v) = -\delta'(y(e-t), v).$$

The following theorem shows that the maximum round-off error of the solutions for

$$u \oplus x = c \text{ and } y \oplus v = c,$$

is the same $b^{e-t} + b^{e-t-1} - b^{e-t-i+1}$. Let $D'(t)$ be a set defined as

$$D'(t) = \{v[e-i, e-i+1-t] > \}.$$

Theorem 2 1. The error function $\varepsilon(y, v)$ in v and the error function $\varepsilon(u, x)$ in x are expressed as the same functions $-\delta'(j, \cdot)$ if $y(e-t) = u(e-t) = j$.

2. The function $\delta'(j, v)$ is a piecewise linear periodic function as follows with period $p = b^{e-t+1}$:

(a) $\delta'(j, v)$ is a periodic function with period p :

$$\delta'(j, v + p) = \delta'(j, v).$$

(b) The figure of $\delta'(j, v)$ on the initial half-open interval I is given as follows:

$$\delta'(j, v) = v - b^{e-i} + jb^{e-t} \\ \text{on } \{b^{e-i} \leq v \leq s_1(j, i) - b^{e-i+1-t}\}; \\ \delta'(j, v) = v - s_1(j, i) - b^{e-t-1} \\ \text{on } \{s_1(j, i) \leq v \leq b^{e-i} + p - b^{e-i+1-t}\}.$$

Here the initial switching point is

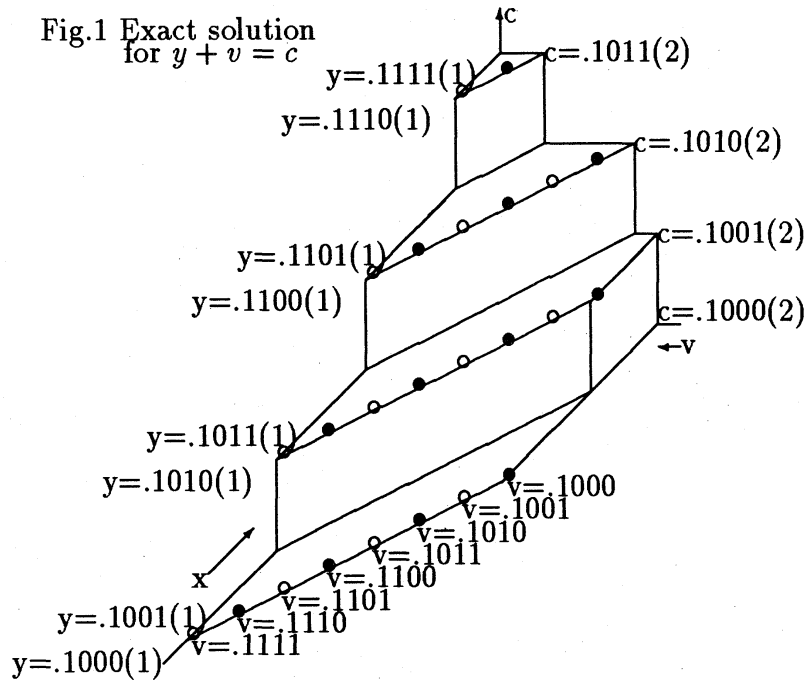
$$s_1(j, i) = b^{e-i} + b^{e-t-1} + jb^{e-t}.$$

3. Let $e(v) = e(x) = e - i$. The maximum of $|\delta'(j, v)|$ in v is given as follows:

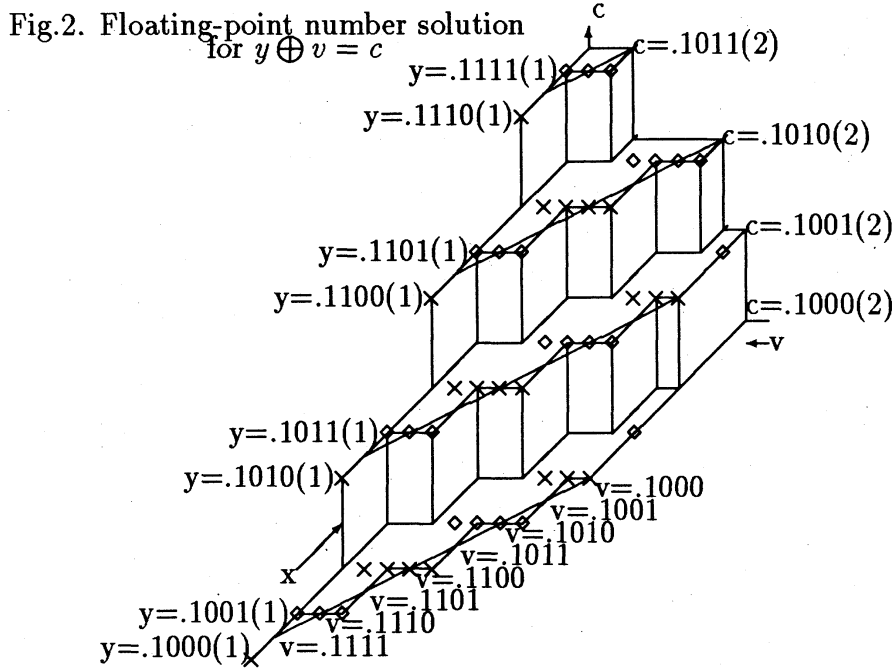
$$\max_{v \in D'(t)} |\delta'(j, v)| = b^{e-t} + b^{e-t-1} - b^{e-t-i+1},$$

which is attained at $v = s_n(j, i) - b^{e-t-i+1}$. Here the switching points are

$$s_n(j, i) = s_1(j, i) + (n-1)p.$$



o: Exact solution, •: Exact solution which coincides with floating-point number solution.



◊ Floating-point number solution (bit is 1 at -3 position), × Floating-point number solution (bit is 0 at -3 position)

By theorem 2, the following results are obtained.

- Theorem 3** 1. The maximum error is $b^{e-t} + b^{e-t-1} - b^{e-t-i+1}$.
2. The maximum error is monotone increasing from b^{e-t} to $b^{e-t} + b^{e-t-1} - b^{e-2t+2}$ as i increases from $i = 2$ to $i - t - 1$.
3. The ISP is $s_1(j, i) = b^{e-i} + b^{e-t-1} + jb^{e-t}$. The distance of ISP and the initial point $v = b^{e-i}$ is independent in i .
4. The period is b^{e-t+1} and is independent in i .
5. The number of oscillations in $D'(t)$ is b^{t-i-1} and decreases as i increases.

MACHIN PRECISION ACCURACY

The machine precision accuracy means that the error takes the positions less than the least significant position $e(X) - t + 1$ for the exact solution X .

Definition 1 The error ε of the solution is called "machine precision accuracy", if the error ε satisfies

$$|\varepsilon| < b^{e(X)-t+1} \text{ for the exact solution } X.$$

- Theorem 4** 1. The error of the solution y is machine precision accuracy for any v and c : $\max_{x \in D'(t)} |\varepsilon(j, v)| < b^{e-t+1}$.
2. The error of the solution x is not always machine precision accuracy for given c and u . The maximum of the round-off error is

$$\max_{x \in D'(t)} |\varepsilon(j, x)| \geq b^{e-i-t+2} = b^{e(\hat{x})-t+1}$$

for the exact solution $\hat{x}$.

MAXIMUM RELATIVE ERROR

The relative error function $r(y, v)$ in v is defined as

$$r(y, v) = \frac{-\varepsilon(y, v)}{c - v}.$$

The relative error function $r(u, x)$ in x is defined as

$$r(u, x) = \frac{-\varepsilon(u, x)}{c - u}.$$

In order to use the linearity of the function $\delta'(j, x)$ in x , $r(u, x)$ is rewritten as

$$r(u, x) = \frac{\delta'(j, x)}{x - \delta'(j, x)}.$$

Theorem 5 1. The relative error function $r(y, v)$ has the following properties:

- (a) $r(y, v_1) < r(y, v_2)$ for $v_2 = v_1 + p$.
- (b) On $I(n)$, the relative error function $r(y, v)$ is a piecewise monotone increasing convex function given as

$$r(y, v) = \frac{v - (n-1)p - b^{e-i} + jb^{e-t}}{c - v} \text{ on } I_0^j(n);$$

$$r(y, v) = \frac{v - s_n(j, i) - b^{e-t-1}}{c - v} \text{ on } I_1^j(n);$$

2. The relative error function $r(u, x)$ has the following properties:

(a) $r(u, x_1) > r(u, x_2)$ for $x_2 = x_1 + p$.

(b) On $I(n)$, the relative error function $r(u, x)$ is a piecewise increasing linear function given as

$$r(u, x) = \frac{x - (n-1)p - b^{e-i} + jb^{e-t}}{(n-1)p + b^{e-i} - jb^{e-t}} \text{ on } I_0^j(n);$$

$$r(u, x) = \frac{x - s_n(j, i) - b^{e-t-1}}{s_n(j, i) + b^{e-t-1}} \text{ on } I_1^j(n);$$

The maximum of the relative error function $r(y, v)$ with respect to v is defined as

$$mr(y) = \max_{v \in D'(t)} |r(y, v)|.$$

The maximum of relative error function $r(u, x)$ with respect to x is defined as

$$mr(u) = \max_{x \in D'(t)} |r(u, x)|.$$

In the following theorem, two maximum of relative errors $mr(y)$ and $mr(u)$ are compared.

Theorem 6 1. The maximum relative error of the solution y is attained at $v = s_N(j, i) - b^{e-i+1-t}$. Here $s_N(j, i)$ is the last switching point and $N = b^{t-i-1}$.

2. The maximum relative error of y is evaluated as

$$\begin{aligned} mr(y) &= r(y, s_N - b^{e-i+1-t}) \\ &= \frac{b^{e-t} + b^{e-t-1} - b^{e-t-i+1}}{c - s_1(j, i) - (N-1)p + b^{e-i+1-t}} \\ &< b^{-t+1}(1 - b^{-t+j}) \text{ for } j = 0, 1. \end{aligned}$$

3. The maximum relative error of x is attained at $x = s_1(j, i) - b^{e-i+1-t}$ for the initial switching point $s_1(j, i)$.

4. The maximum relative error of x is evaluated as

$$mr(u) = r(u, s_1 - b^{e-i+1-t}) = \frac{b^{i-t}(1 + b^{-1}) - b^{-t+1}}{1 - jb^{i-t}}.$$

The following theorem shows that the maximum relative error $mr(u)$ is monotone increasing as the difference $i-1$ of exponents $e(u) = e$ and $e(x) = e - i + 1$ increases.

Theorem 7 Put the maximum relative error of x as mr

$$mr = \frac{b^{i-t}(1 + b^{-1}) - b^{-t+1}}{1 - jb^{i-t}}.$$

Then mr is monotone increasing from

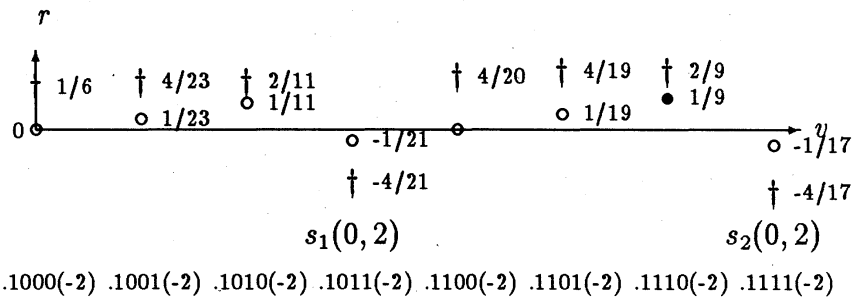
$$\frac{b^{2-t}}{1 - jb^{2-t}} \text{ to } \frac{b^{-1} + b^{-2} - b^{-t+1}}{1 - jb^{-1}}$$

as i increases from $i = 2$ to $i = t - 1$.

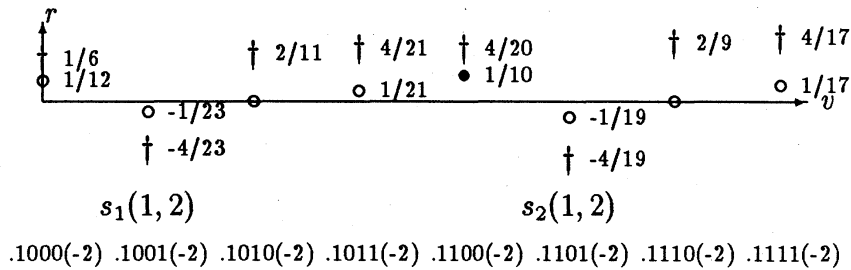
FIGURES OF RELATIVE ERROR FUNCTION

The following figures Fig.3, Fig.4, Fig.5 and Fig. 6 show the relative error functions $r(j, v)$ of the solution y for the equation $y \oplus v = c$ for $j = 0, 1$, $i = 2, 3$ and $t = 4$. The results in theorems 4, 5, and 6 are visualized by the figures.

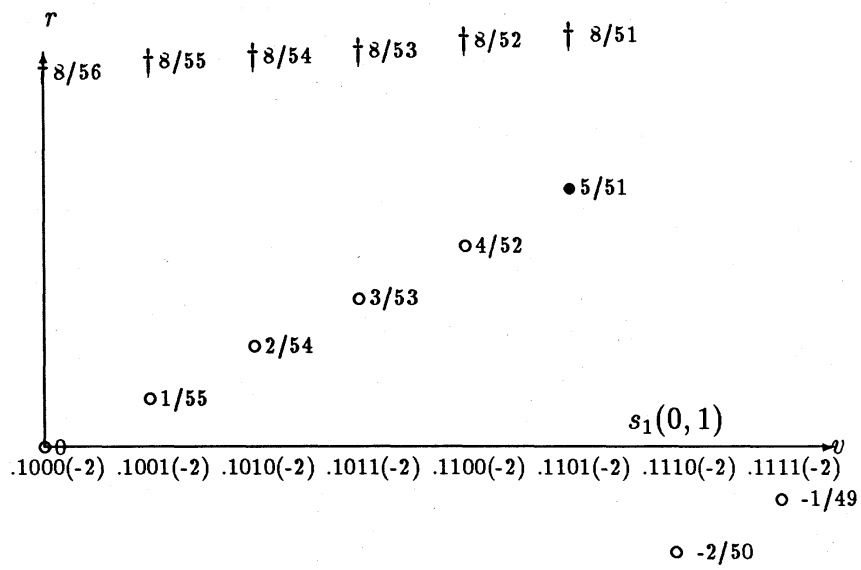
1. The relative error functions are the piecewise monotone increasing convex functions.
2. The switching points are coincide with those of the round-off error functions (see the figures of round-off error functions in [?]).
3. The maximum of relative error is taken in the last interval $I(N)$ with period p . Here the point of the maximum relative error is denoted by "•".
4. The point which attains the maximum relative error is the left side point adjacent to the switching point $s_N(i, j)$.
5. The round-off errors are all machine precision accuracy. Here the relative error such that the round-off error is machine precision accuracy, is denoted by "†".



Equation $y \oplus v = c$; Period $p = b^{-3}$; $N = 2$
Fig.3 Relative error function $r(0, v)$ for $i = 2$ and $t = 4$



y : solution of $y \oplus v = c$; $p = b^{-3}$; $N = 2$
Fig.4 Relative error function $r(1, v)$ for $i = 2$ and $t = 4$



Equation $y \oplus v = c$; $p = b^{-2}$; $N = 1$

$r(0, v)$ for $i = 3$ and $t = 4$

Fig.5 Relative error function

† -8/50 † -8/49

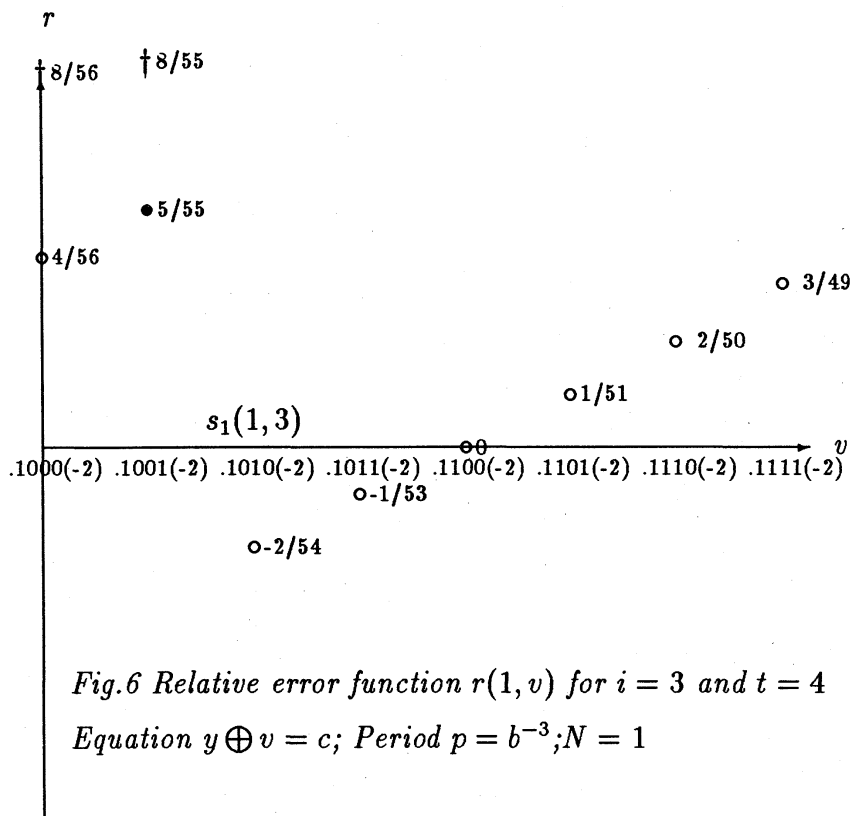


Fig.6 Relative error function $r(1, v)$ for $i = 3$ and $t = 4$
Equation $y \oplus v = c$; Period $p = b^{-3}$; $N = 1$

$$\dagger -8/54 \quad \dagger -8/53 \quad \dagger -8/52 \quad \dagger -8/51 \quad \dagger -8/50 \quad \dagger -8/49$$

The following figures Fig.7, Fig.8, Fig. 9 and Fig. 10 show the relative error functions $r(j, x)$ of the solution x for the equation $x \oplus v = c$ for $j = 0, 1$, $i = 2, 3$ and $t = 4$. The results in theorems 4, 5, 6 and 7, are visualized by the figures.

1. The relative error functions are the piecewise increasing linear functions.
2. The switching points are coincide with those of the round-off error functions (see the figures of round-off error functions in [?]).
3. The maximum of relative error is taken in the initial interval $I(1)$ with period p .
4. The point which attains the maximum relative error is the point of left side adjacent to the initial switching point $s_1(i, j)$ by one.
5. The round-off errors are not always machine precision accuracy. Here the relative error such that the round-off error is machine precision accuracy, is expressed by the line "-----".
6. The maximum relative error $mr(u)$ is monotone increasing as the difference $i - 1$ increases. The maximum relative error $1/4$ in Fig. 7 increases to $5/8$ in Fig. 9 as i increases from 2 to 3. The maximum relative error $1/3$ in Fig. 8 increases to $5/4$ in Fig. 10 as i increases from 2 to 3.

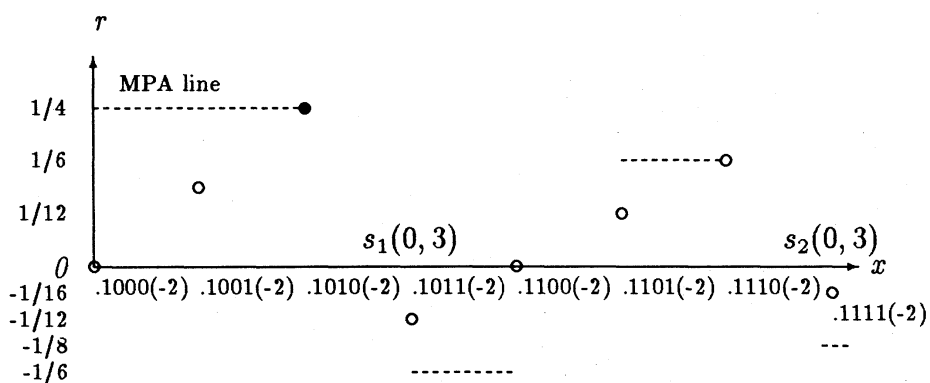


Fig.7 Relative error function $r(0, x)$ for $i = 2$ and $t = 4$

x : solution of $u \oplus x = c$; $p = b^{-3}$; $N = 2$

The above graph shows that 6 points are machine precision accuracy and the other are not.

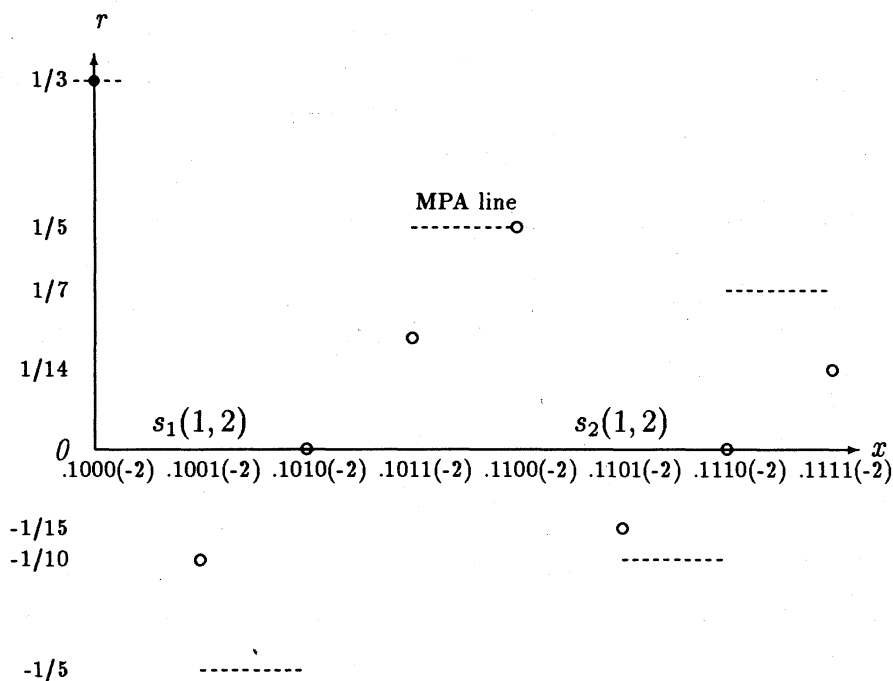
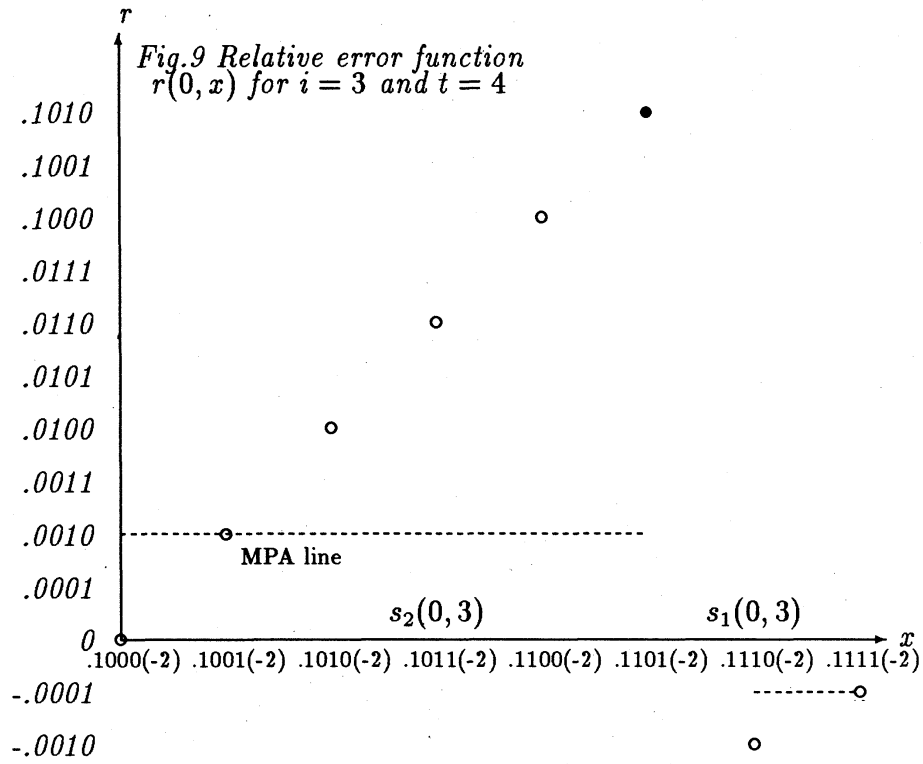


Fig.8 Relative error function $r(1, x)$ for $i = 2$ and $t = 4$

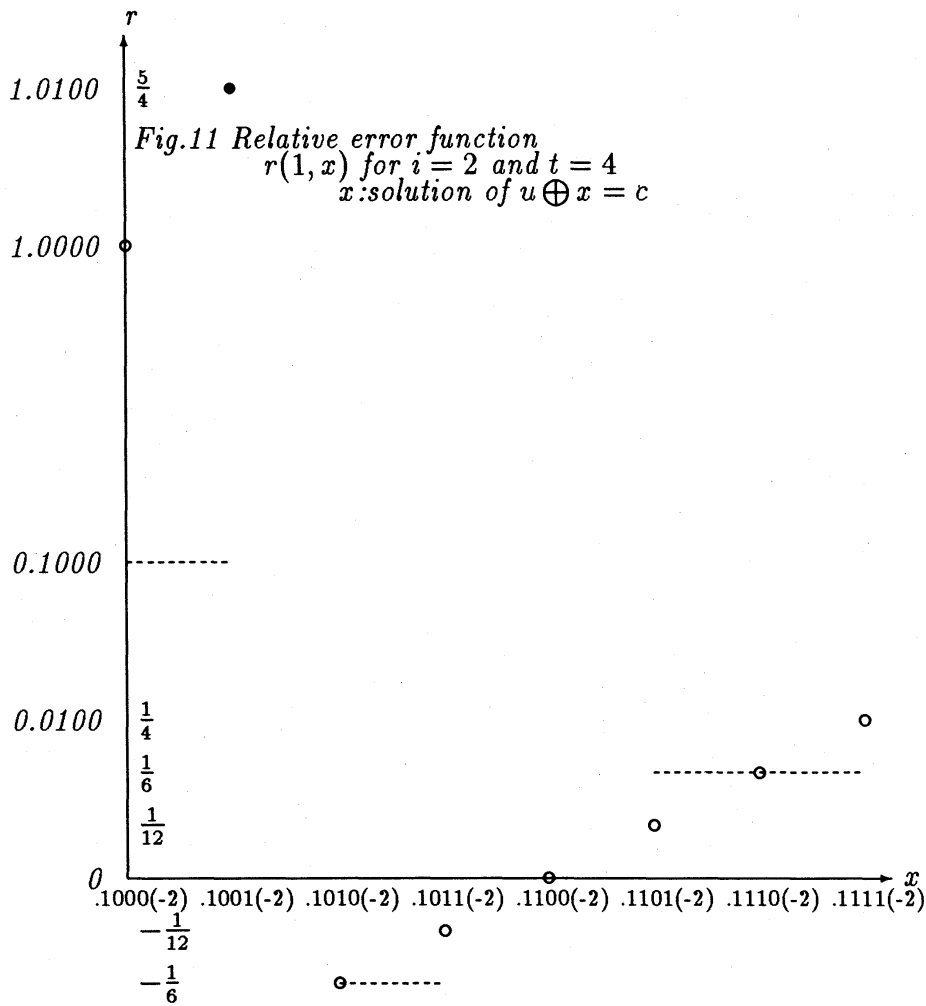
x : solution of $u \oplus x = c$. $p = b^{-3}$, $N = 2$

The above graph shows that 6 points are machine precision accuracy and the other are not.



x : solution of $u \oplus x = c$. $p = b^{-2}$, $N = 1$

The above graph shows that only initial point is machine precision accuracy and the other are not.



In Fig.10, the maximum relative error is the value $5/4$ which is more than 1. This phenomenon is analysed as follows. In this case, for given

$$(1) \quad u[-1, -4] \geq .1111 \text{ and } c[0, -3] \geq 1.000,$$

$u \oplus x = c$ is solved as

$$x[-3, -6] \geq .1001(-2),$$

since

$$\begin{aligned} u \oplus x &= u[-1, -3] + x[-3, -6] + C(u(-4), x(-4), x(-5))b^{-3} \\ &= .111 + .100(-2) + C(1, 0, 0)b^{-3} = 1.000. \end{aligned}$$

For given u and c in (1), $u + x = c$ is solved as $\hat{x} = b^{-4}$. The error is $\hat{x} - x[-3, -6] \geq -.101(-3)$ and the relative error is $\frac{.101(-3)}{b^{-4}} = 5/4$. The exact solution $\hat{x}$ is extraordinarily

small. In the calculation of $c - u$, the catastrophic cancellation occurs. In the calculation of $u + x$, the carry propagates from the least significant position to the leading position of c . Thus the round-off error becomes more than the exact solution.

References

- [1] Tsuji, K., Rounding step function in a floating-point number system, *Abstract of Applied Math. in Math. Soc. Japan*, (1990), 167-170.
- [2] Tsuji, K., Carry function in a floating-point number system, *Abstract of Applied Math. in Math. Soc. Japan*, (1991), 78-81.
- [3] Tsuji, K., Carry function in a floating-point number system, *RMC 66-06 Kyushu Univ.*, (1991).
- [4] Tsuji, K., Error to solutions of a linear equation in a floating-point number system; *First Conf. Pro. JIAM*, (1991), 89-90.
- [5] Tsuji, K., Floating-point number solutions in a simple linear equation with addition algorithm, *Abstract of Applied Math. in Math. Soc. Japan*, (1992), 39-42.
- [6] Tsuji, K., Floating-point number solutions in a simple linear equation with addition algorithm Part 1; Round-off error of floating-point addition, *RMC 67-08 Kyushu Univ.*, (1992).