Revisit to Floating-point Division Algorithm Based on Taylor-Series Expansion

Jianglin Wei, A. Kuwana, H. Kobayashi,
K. Kubo

Division of Electronics and Informatics, Gunma University
Oyama National College of Technology
Japan

群馬大学
GUNMA UNIVERSITY

Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Verification
- Hardware Implementation Tradeoff
- Conclusion

Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Results
- Hardware Implementation Tradeoff
- Conclusion

Research Background

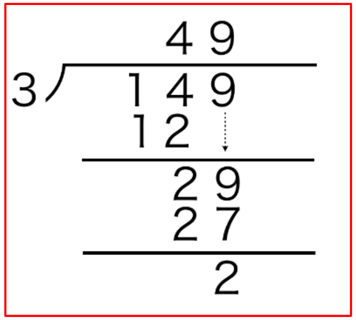
◆High-speed high-precision floating-point arithmetic

→ Embedded systems, mobile applications.

◆Addition / Subtraction → Relatively easy

Multiplication → Modestly complicated

Division → Very tough !



Research Objective

- Floating-point division operation

  -- Simple circuit

  -- High-speed

- Application of Taylor expansion

  to floating-point division arithmetic

- Clarification of its calculation procedure

- Clarification of its tradeoff among

  accuracy, number of operations and LUT size

Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Verification
- Hardware Implementation Tradeoff
- Conclusion

Taylor Series

Idea behind Taylor expansion :

We can re-write every smooth function as an infinite sum of polynomial terms.

Function $f(x)$ for a point $x = a$ is given by :

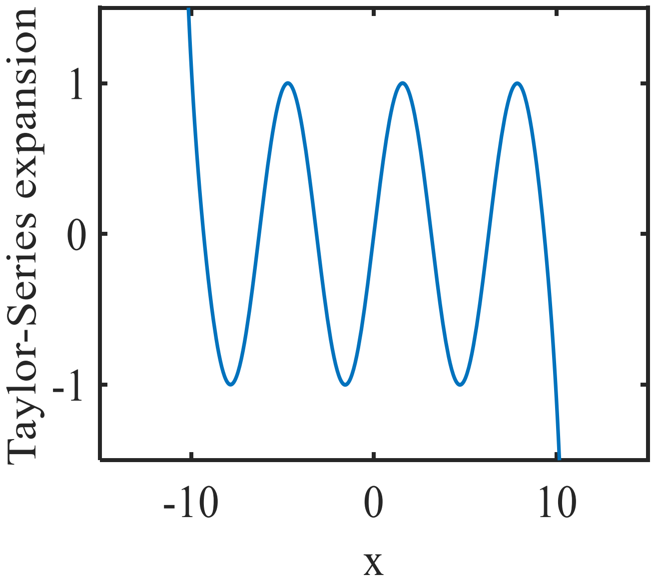
$$f(x) = f(a) + f'(a)(x-a) + \frac{(f)''(a)}{2!}(x-a)^2 + \cdots + \frac{(f)^n(a)}{n!}(x-a)^n + \cdots$$

Convergent range $\alpha < x < \beta$

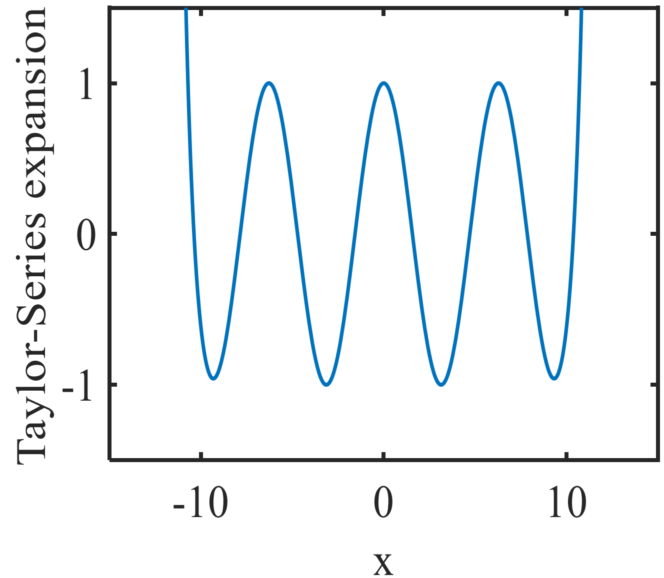
Taylor-Series of sin(x) and cos(x)

Taylor series simulation results of
sin(x) and cos(x)

Convergent range : $-\infty < x < +\infty$



$\sin(x)$



$\cos(x)$

Center value : 0

Number of Taylor-series expansion terms : 20.

Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Verification
- Hardware Implementation Tradeoff
- Conclusion

Normalized Floating-Point Representation

Floating-point representation in binary ：

Mantissa : M　　　　Exponent : E

$M \times 2^{E}$　　　　$a, b, c, d, e, f, \cdots$: 0 or 1

Decimal point

$1.abcdef\cdots \times 2^{E}$

Mantissa　　Exponent

$1 \leqq M < 2$

Ex :1011001 (binary) = 89 (decimal)

Binary representation ： 1.011001×2^{110}

Decimal representation ： $1.390625 \times 2^{6} = 89$

Division of Binary Numbers

Consider calculation of $A = \frac{N}{D}$.

A, N, D are in floating-point representation :

$$A : M_A \times 2^{E_A}$$

$$N : M_N \times 2^{E_N}$$

$$D : M_D \times 2^{E_D}$$

First, calculate reciprocal of mantissa $\frac{1}{M_D}$ $(1 \leq M_D < 2)$ using Taylor-expansion of $f(x) = 1/x$

$$\frac{1}{\mathrm{D}} = \frac{1}{M_D} \times 2^{-E_D}$$

Then, calculate $N \times \frac{1}{\mathrm{D}}$ and obtain A.

Use conventional digital multiplication algorithm.

Proposed Algorithm

Calculate reciprocal of mantissa : $\frac{1}{M_D}$ (1≤ M_D <2)

$x = M_D$

$f(x) = \frac{1}{x}$ by Taylor expansion at $x = a$ $(1 \leq a < 2)$

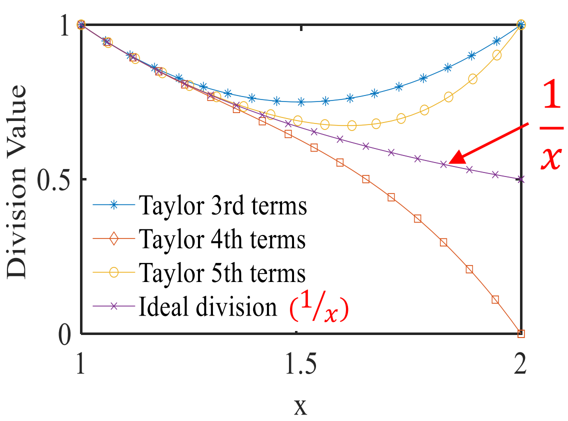
$$f(x) = \frac{1}{a} - (x-a) + (x-a)^2 - (x-a)^3 + \cdots + (-1)^n (x-a)^n + \cdots$$

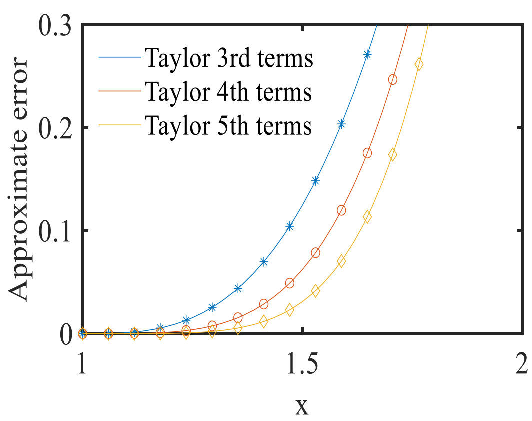
Stored in LUT in advance

Coefficient of each term is +1 or -1

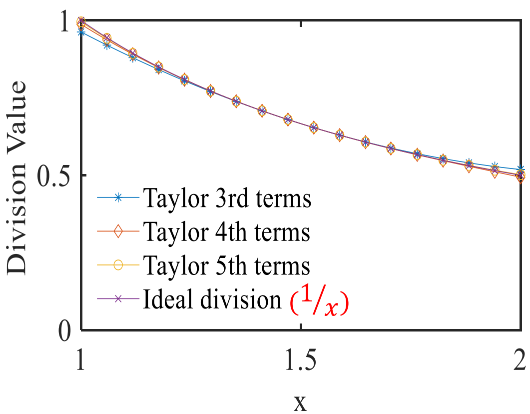
No multiplication required there

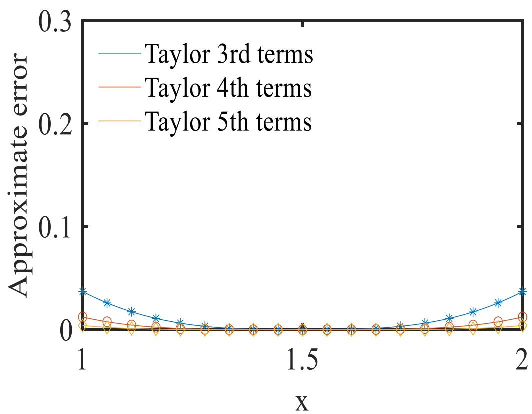
Concept of Proposed Method





Taylor series expansion of $\frac{1}{x}$ at center value a = 1





Taylor series expansion of $\frac{1}{x}$ at center value a = 1.5

$$E = \frac{I - T}{I}$$

E : Approximate error

I : Ideal value

T : Taylor series expansion value

Proposed method :

Region technology

For example :

1 region :

a = 1.5 $1 \le x < 2$

2 regions :

a = 1.25 $1 \le x < 1.5$

a = 1.75 $1.5 \le x < 2$

4 regions :

a = 1.125 $1 \le x < 1.25$

a = 1.375 $1.25 \le x < 1.5$

a = 1.625 $1.5 \le x < 1.75$

a = 1.875 $1.75 \le x < 2$

⋮

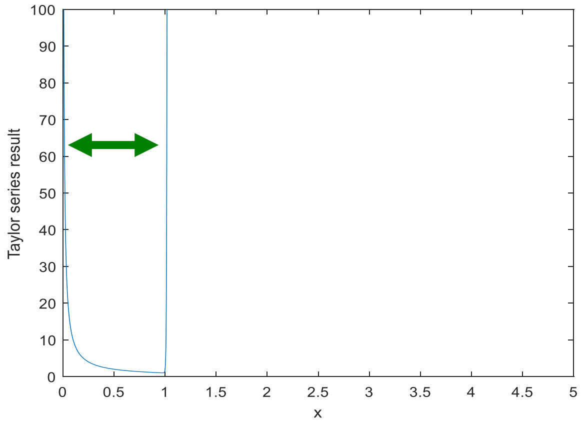
Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Verification
- Hardware Implementation Tradeoff
- Conclusion

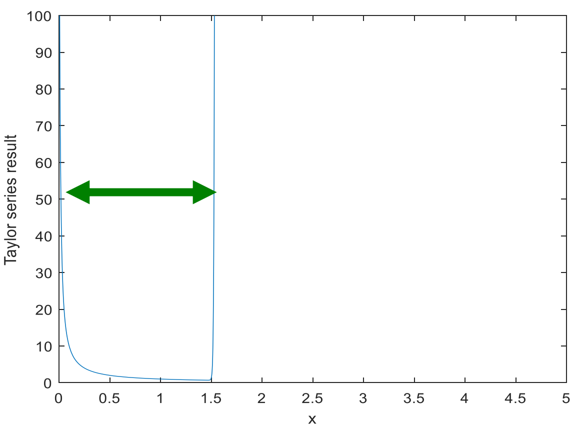
Taylor Expansion Waveform of $1/x$

Taylor expansion waveform of $f(x) = \frac{1}{x}$

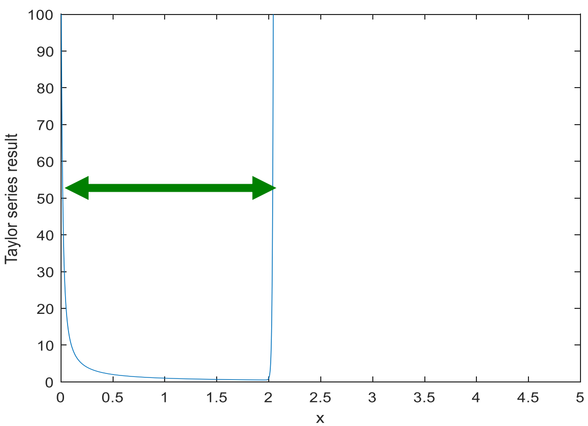
Number of Taylor expansion terms : 25



Taylor expansion at *a* = 0.5
Convergent range : $0 < x < 1$

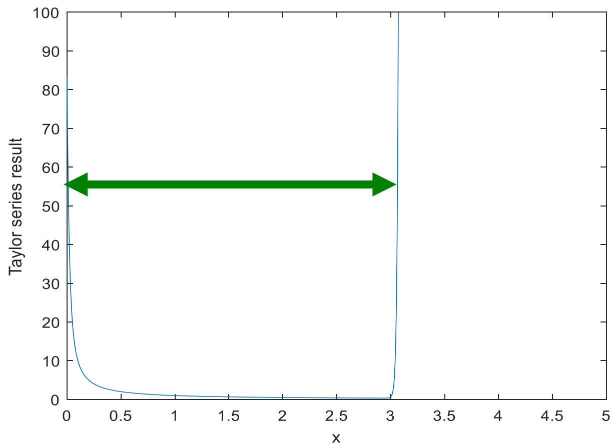


Taylor expansion at *a* = 0.75
Convergent range : $0 < x < 1.5$

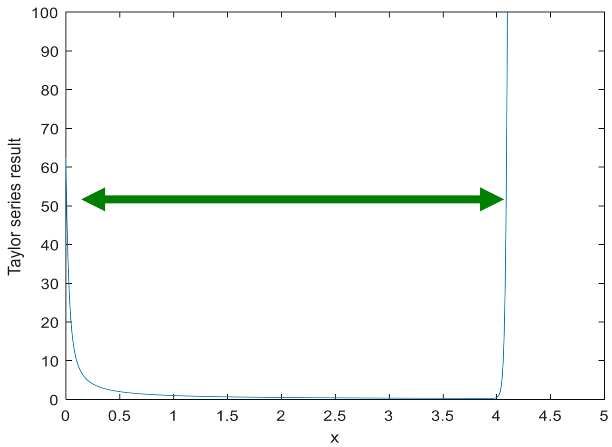


Taylor expansion at *a* = 1
Convergent range : $0 < x < 2$

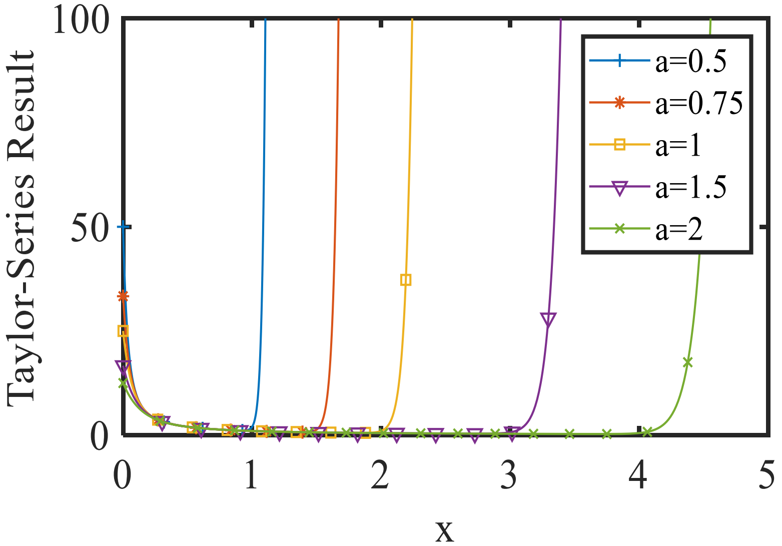
Taylor Expansion of $1/x$: Comparison



Taylor expansion at a = 1.5
Convergent range: $0 < x < 3$



Taylor expansion at a = 2
Convergent range: $0 < x < 4$



Taylor expansion comparisons at different central point convergent ranges.

Definition of Accuracy

Ex : $\frac{1}{2^{16}}$ accuracy

$$max \left| \frac{f(x) - t(n,x)}{f(x)} \right| \le \frac{1}{2^{16}}$$

$f(x)$: Original function (Ideal value)

$t(n,x)$: Taylor expansion up to the minimum of terms n

Simulation Results

Use Taylor series expansion equation :

$$f(x) = \frac{1}{x} \qquad (1 \le x < 2)$$

One-region case :

Mantissa represented by binary decimal point.

Specified accuracy

Taylor-series expansion \ precision		$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
(i) M_D = 1.xxxxxx··· $1 \le M_D < 2$	$a = 1.5$	6	11	13	16	21

Taylor series expansion at center value $a = 1.5$

Number of Taylor expansion terms to meet specified accuracy.

Two-Region Case

Use Taylor series expansion equation :

$$f(x) = \frac{1}{x} \qquad (1 \le x < 2)$$

In two-region case, we check the value (0 or 1) of the first decimal place of Mantissa.

Taylor-series expansion \ precision		$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
(i) $M_D = 1.0xxxxx\cdots$ $1 \le M_D < 1.5$	$a = 1.25$	4	7	9	11	14
(ii) $M_D = 1.1xxxxx\cdots$ $1.5 \le M_D < 2$	$a = 1.75$	3	6	8	9	12

Four-Region Case

Use Taylor series expansion equation :

$$f(x) = \frac{1}{x} \qquad (1 \le x < 2)$$

In four-region case, we check the values (00, 01, 10 or 11) of the first two decimal places of Mantissa.

Taylor-series expansion \ precision		$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
(i) M = 1.00xxxx··· $1 \le M_D < 1.25$	a=1.125	3	6	7	8	11
(ii) M = 1.01xxxx··· $1.25 \le M_D < 1.5$	a=1.375	3	5	6	7	10
(iii) M = 1.10xxxx··· $1.5 \le M_D < 1.75$	a=1.625	3	5	6	7	9
(iv) M = 1.11xxxx··· $1.75 \le M_D < 2$	a=1.875	3	5	6	7	9

Eight-Region Case

In eight-region case, we check the values (000, 001,…, 111) of the first three decimal places of Mantissa.

Taylor-series expansion \ precision		$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
(i) M = 1.000xxxx··· $1 \leq M_D < 1.125$	a = 1.0625	2	4	5	6	8
(ii) M = 1.001xxxx··· $1.125 \leq M_D < 1.25$	a = 1.1875	2	4	5	6	8
(iii) M = 1.010xxxx··· $1.25 \leq M_D < 1.375$	a=1.3125	2	4	5	6	8
(iv) M = 1.011xxxx··· $1.375 \leq M_D < 1.5$	a=1.4375	2	4	5	6	8
(v) M = 1.100xxxx··· $1.5 \leq M_D < 1.625$	a=1.5625	2	4	5	6	7
(vi) M = 1.101xxxx··· $1.625 \leq M_D < 1.75$	a=1.6875	2	4	5	6	7
(vii) M = 1.110xxxx··· $1.75 \leq M_D < 1.875$	a=1.8125	2	4	5	5	7
(viii) M = 1.111xxxx··· $1.875 \leq M_D < 2$	a=1.9375	2	4	5	5	7

Comparison of Two Decimal Point Position Cases

Mantissa: 1.xxxx

Mantissa : 0.1xxxx

Numbers of Taylor expansion terms for specified accuracy are the same.

To obtain 20-bit accuracy

1-region case ➡ 13 terms
2-region case ➡ 9 terms
4-region case ➡ 7 terms
8-region case ➡ 5 terms

Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Verification
- Hardware Implementation Consideration
- Conclusion

Calculation Complexity

- Ex: Taylor expansion 4 terms :

$$g_4 = \frac{1}{a} - (\mathrm{x} - \mathrm{a}) + (\mathrm{x} - \mathrm{a})^2 - (\mathrm{x} - \mathrm{a})^3$$

- $\frac{1}{a}$ value : Stored in memory and read.

  y = x−a Subtraction: 1 time z = y^2 Multiplication: 1 time

$$g_4 = \frac{1}{a} - y + y^2 - y^3 = \frac{1}{a} - y + z - yz$$

Multiplication: 1 time
Addition / Subtraction: 3 times

Total : Multiplication: 2 times
Addition / Subtraction: 4 times

- Ex: Taylor expansion 5 terms :

$$g_5 = \frac{1}{a} - (\mathrm{x} - \mathrm{a}) + (\mathrm{x} - \mathrm{a})^2 - (\mathrm{x} - \mathrm{a})^3 + (\mathrm{x} - \mathrm{a})^4$$

Total : Multiplication: 2 times
Addition / Subtraction: 4 times

Number of Operations

Number of terms versus number of operations in Taylor expansion

Taylor expansion of $f(x) = \frac{1}{x}$ can be calculated with a relatively small number of operations.

Terms of Taylor expansion	multiplication	Addition or subtraction
3	1	3
4	2	4
5	2	4
6	3	5
7	3	5
8	4	6

LUT size

$$f(x) = \boxed{\frac{1}{a}} - (x - a) + (x - a)^2 - (x - a)^3 + \ldots..$$

Stored in LUT

LUT: Look-Up Table

Ex: 4-region case → LUT size is 4 words

Address	LUT
00	inverse of a = 1.125
01	inverse of a = 1.375
10	inverse of a = 1.625
11	inverse of a = 1.875

M = 1.xx

Select by 1 and 2 decimal places of the mantissa.

Outline

- Research Background and Objective
- Taylor-Series Expansion
- Proposed Algorithm
- Simulation Verification
- Hardware Implementation Consideration
- Conclusion

Conclusion

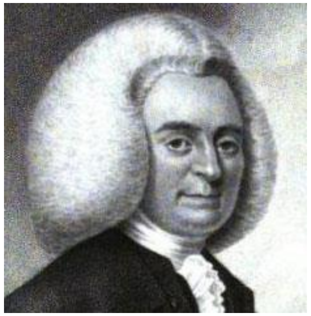
- Algorithm for inverse calculation of mantissa part in binary floating format using Taylor expansion has been investigated.
- Relationship between accuracy and number of terms in Taylor expansion was obtained.
- Number of divided regions becomes larger
  - ➢Number of Taylor expansion terms ➡ small
  - ➢LUT size ➡ big

Designer can choose the best compromised design.

Thank you for listening !

Appendix

Maclaurin Series



Maclaurin
(Scottish mathematician)

Special case :

Taylor series for $a = 0$ $\rightarrow$ Maclaurin series

$$f(x) = \mathrm{f}(0) + f'(0)x + \frac{(\mathrm{f})''(0)}{2!}x^2 + \cdots + \frac{(\mathrm{f})^n(0)}{\mathrm{n}!}x^{\mathrm{n}}$$

Convergent range $\gamma < x < \delta$

Newton’s method

Newton’s method step:

Fist, Start with an initial approximation x_0 close to c.

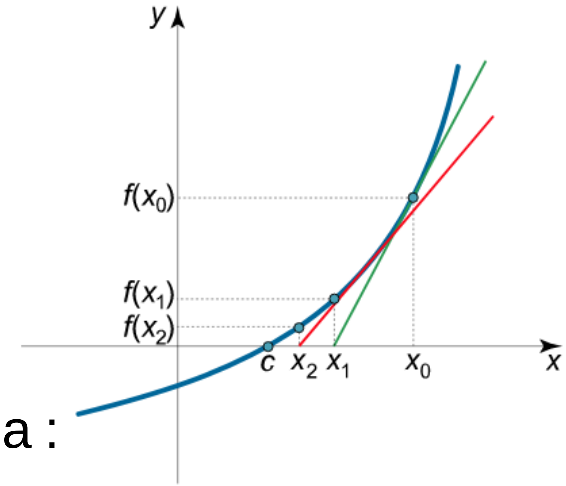
Second, Determine the next approximation by the formula :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Third, Continue the iterative process using the formula :

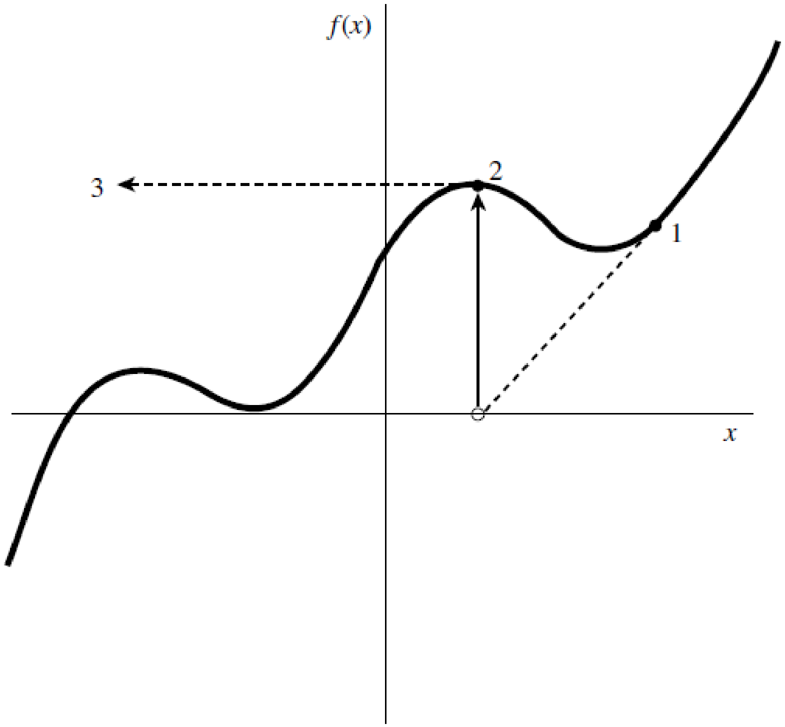
$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

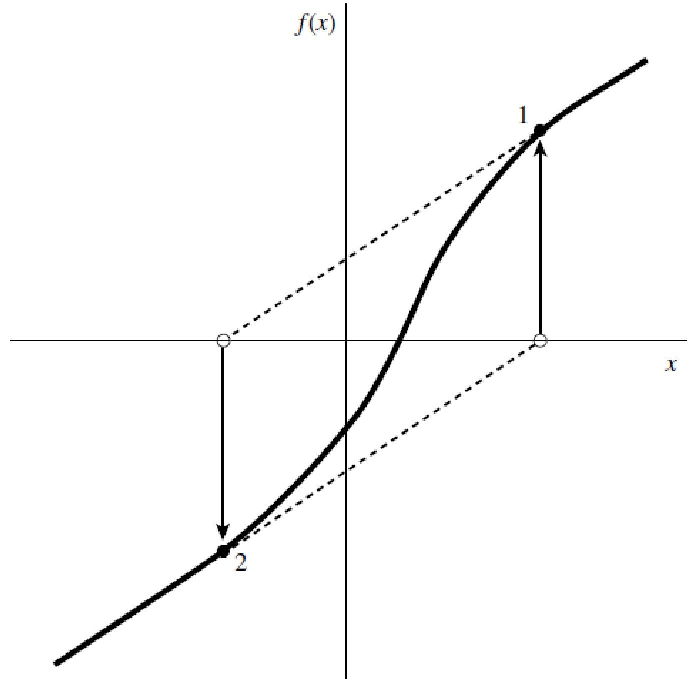
Last, until the root is found to the desired accuracy.



- Poor global convergence properties
- Dependent on initial guess
  - May be too far from local root
  - May encounter a zero derivative
  - May loop indefinitely

Examples of disadvantages





On the left, we have Newton's Method finding a local maxima, in such cases the method will shoot off into negative infinity.

Newton's Method has entered an infinite cycle. Better initial guesses may be able to alleviate this problem.

Another Decimal Point Position

Change the decimal point position of the mantissa

Mantissa: M　　　Exponent: E

$M \times 2^{E}$

Original decimal point

$1.abcdef \cdots \times 2^{E}$

Mantissa　Exponent

$1 \leqq M < 2$

$M \times 2^{E}$

New decimal point

$0.1abcdef \cdots \times 2^{E}$

Mantissa　Exponent

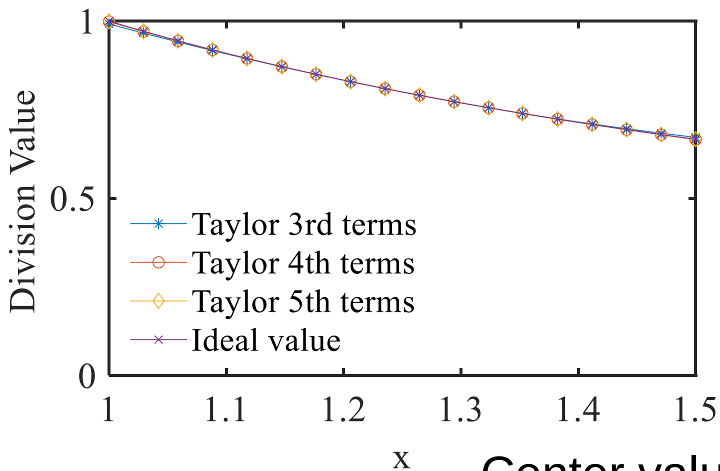
$0.5 \leqq M < 1$

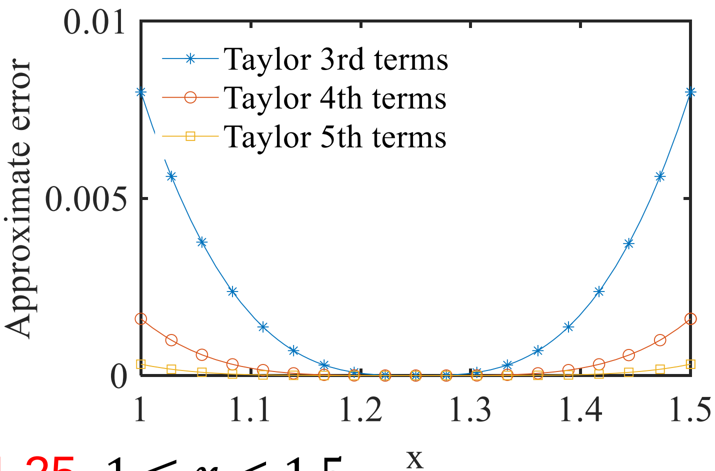
Ex : 1011001 (binary) = 89 (decimal)

Binary representation : 0.1011001×2^{111}

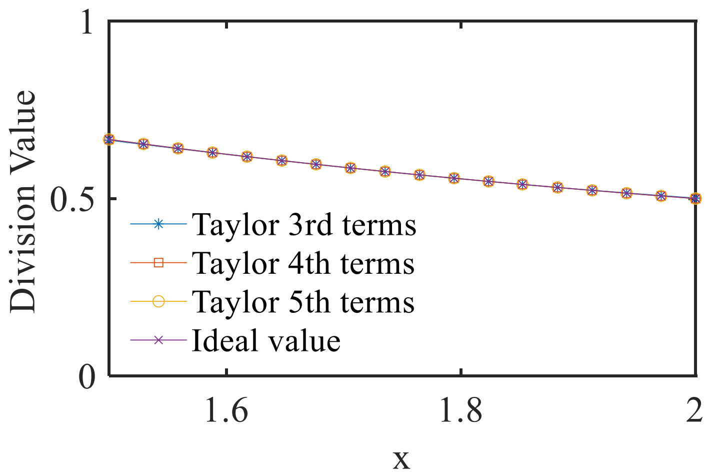
Decimal representation : $0.6953125 \times 2^{7} = 89$

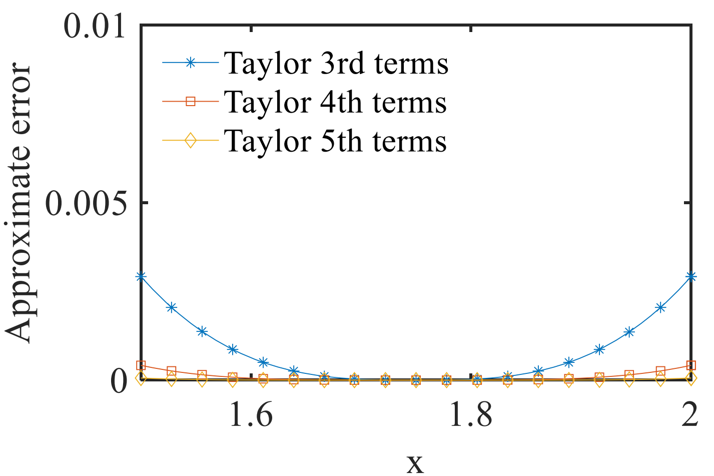
2 region case





Center value a=1.25 $1 \leq x < 1.5$

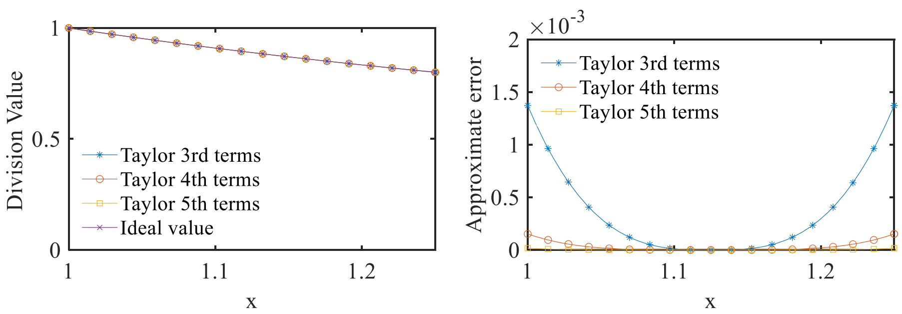




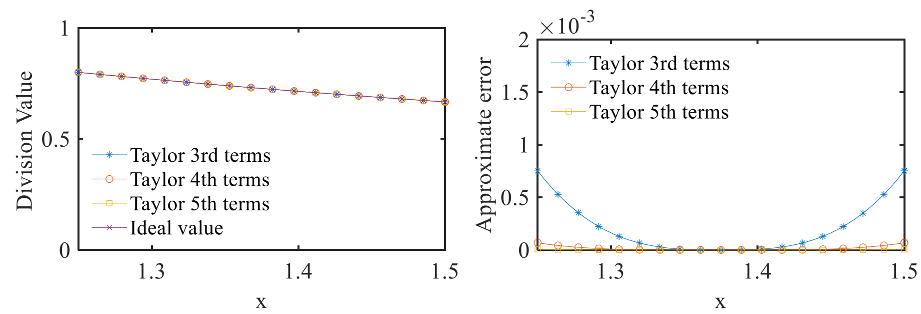
Center value a=1.75 $1.5 \leq x < 2$

Taylor series expansion of $\frac{1}{x}$

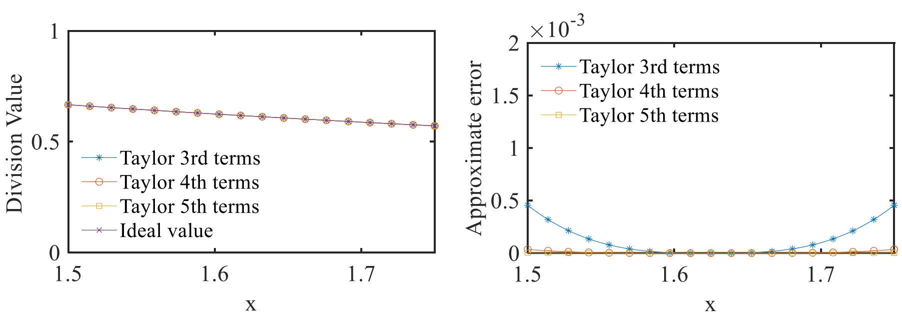
4 region case



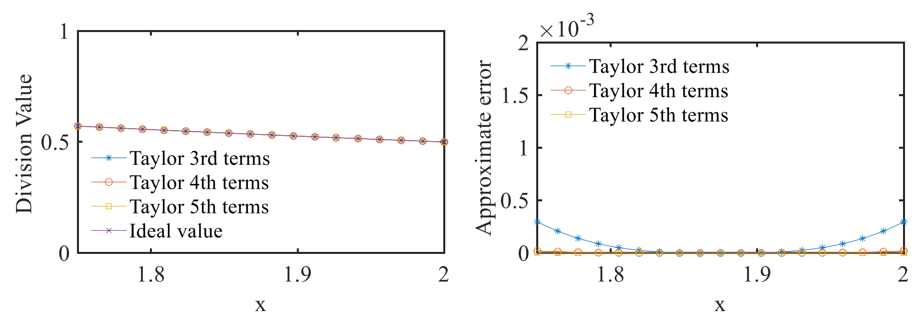
Center value a=1.125 $1 \le x < 1.25$



Center value a=1.375 $1.25 \le x < 1.5$



Center value a=1.625 $1.5 \le x < 1.75$



Center value a=1.875 $175 \le x < 2$

Taylor series expansion of $\frac{1}{x}$

Simulation Results

Use Taylor series expansion equation :

$$f(x) = \frac{1}{x} \qquad (0.5 < x \leq 1)$$

One-region case

Taylor series expansion \ Precision	$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
[1] M = 0.1xxxx··· $0.5 \leqq M_D < 1$ x=0.75	6	11	13	16	21

Two-region case

Taylor series expansion \ Precision	$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
(i) M = 0.10xxxx··· $0.5 \leqq M_D < 0.75$ x=0.625	4	7	9	11	14
(ii) M = 0.11xxxx··· $0.75 \leqq M_D < 1.0$ x=0.875	3	6	8	9	12

Four-region case

Taylor series expansion \ Precision	$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
(i) M = 0.100xxxx··· $0.5 \leqq M_D < 0.625$ x=0.5625	3	6	7	8	11
(ii) M = 0.101xxxx··· $0.625 \leqq M_D < 0.75$ x=0.7125	3	6	7	8	11
(iii) M = 0.110xxxx··· $0.75 \leqq M_D < 0.875$ x=0.8125	3	5	6	7	9
(iv) M = 0.111xxxx··· $0.875 \leqq M_D < 1.0$ x=0.9375	3	5	6	7	9

$f(x) = \frac{1}{x}$

Taylor series expansion of $f(x)$

$1 \leqq M_D < 2$

16 regions case :

Taylor-series expansion \ precision	$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
[1] M = 1.0000xxxx⋯ ($1 \leqq M_D < 1.0625$) x=1.03125	2	4	4	5	7
[2] M = 1.0001xxxx⋯ ($1.0625 \leqq M_D < 1.125$) x=1.09375	2	4	4	5	7
[3] M = 1.0010xxxx⋯ ($1.125 \leqq M_D < 1.1875$) x=1.15625	2	4	4	5	7
[4] M = 1.0011xxxx⋯ ($1.1875 \leqq M_D < 1.25$) x=1.21875	2	4	4	5	7
[5] M = 1.0100xxxx⋯ ($1.25 \leqq M_D < 1.3125$) x=1.28125	2	3	4	5	6
[6] M = 1.0101xxxx⋯ ($1.3125 \leqq M_D < 1.375$) x=1.34375	2	3	4	5	6
[7] M = 1.0110xxxx⋯ ($1.375 \leqq M_D < 1.4375$) x=1.40625	2	3	4	5	6
[8] M = 1.0111xxxx⋯ ($1.4375 \leqq M_D < 1.5$) x=1.46875	2	3	4	5	6
[9] M = 1.1000xxxx⋯ ($1.5 \leqq M_D < 1.5625$) x=1.53125	2	3	4	5	6
[10] M = 1.1001xxxx⋯ ($1.5625 \leqq M_D < 1.625$)x=1.59375	2	3	4	5	6
[11] M = 1.1010xxxx⋯ ($1.625 \leqq M_D < 1.6875$=1.65625	2	3	4	5	6
[12] M = 1.1011xxxx⋯ ($1.6875 \leqq M_D < 1.75$) x=1.71875	2	3	4	5	6
[13] M = 1.1100xxxx⋯ ($1.75 \leqq M_D < 1.8125$) x=1.78125	2	3	4	5	6
[14] M = 1.1101xxxx⋯ ($1.8125 \leqq M_D < 1.875$)x=1.84375	2	3	4	5	6
[15] M = 1.1110xxxx⋯ ($1.875 \leqq M_D < 1.9375$)x=1.90625	2	3	4	5	6
[16] M = 1.1111xxxx⋯ ($1.9375 \leqq M_D < 2$) x=1.96875	2	3	4	5	6

$f(x) = \frac{1}{x}$

Taylor series expansion of $f(x)$

32 regions case :

$1 \leq M_D < 2$

Taylor series expansion \ precision	$\frac{1}{2^8}$	$\frac{1}{2^{16}}$	$\frac{1}{2^{20}}$	$\frac{1}{2^{24}}$	$\frac{1}{2^{32}}$
[1] M = 1.00000xxxx··· (1 ≦M_D <1.03125) x=1.015625	2	3	4	4	6
[2] M = 1.00001xxxx··· (1.03125≦M_D <1.0625) x=1.046875	2	3	4	4	6
[3] M = 1.00010xxxx··· (1.0625≦M_D <1.09375) x=1.078125	2	3	4	4	6
[4] M = 1.00011xxxx··· (1.09375≦M_D <1.125) x=1.109375	2	3	4	4	6
[5] M = 1.00100xxxx··· (1.125≦M_D <1.15625) x=1.140625	2	3	4	4	6
[6] M = 1.00101xxxx··· (1.15625≦M_D <1.1875) x=1.171875	2	3	4	4	6
[7] M = 1.00110xxxx··· (1.1875≦M_D <1.21875) x=1.203125	2	3	4	4	6
[8] M = 1.00111xxxx··· (1.21875≦M_D <1.25) x=1.234375	2	3	4	4	6
[9] M = 1.01000xxxx··· (1.25≦M_D <1.28125) x=1.265625	2	3	4	4	6
[10] M = 1.01001xxxx··· (1.28125≦M_D <1.3125)x=1.296875	2	3	4	4	6
[11] M = 1.01010xxxx··· (1.3125≦M_D <1.34375)x=1.328125	2	3	4	4	5
[12] M = 1.01011xxxx··· (1.34375≦M_D <1.375) x=1.359375	2	3	4	4	5
[13] M = 1.01100xxxx··· (1.375≦M_D <1.40625) x=1.390625	2	3	4	4	5
[14] M = 1.01101xxxx··· (1.40625≦M_D <1.4375)x=1.421875	2	3	4	4	5
[15] M = 1.01110xxxx··· (1.4375≦M_D <1.46875)x=1.453125	2	3	4	4	5
[16] M = 1.01111xxxx··· (1.46875≦M_D <1.5) x=1.484375	2	3	4	4	5
[17] M = 1.10000xxxx··· (1.5≦M_D <1.53125) x=1.515625	2	3	4	4	5
[18] M = 1.10001xxxx··· (1.53125≦M_D <1.5625)x=1.546875	2	3	4	4	5
[19] M = 1.10010xxxx··· (1.5625≦M_D <1.59375)x=1.578125	2	3	4	4	5
[20] M = 1.10011xxxx··· (1.59375≦M_D <1.625) x=1.609375	2	3	3	4	5
[21] M = 1.10100xxxx··· (1.625≦M_D <1.65625) x=1.640625	2	3	3	4	5
[22] M = 1.10101xxxx··· (1.65625≦M_D <1.6875)x=1.671875	2	3	3	4	5
[23] M = 1.10110xxxx··· (1.6875≦M_D <1.71875)x=1.703125	2	3	3	4	5
[24] M = 1.10111xxxx··· (1.71875≦M_D <1.75) x=1.734375	2	3	3	4	5
[25] M = 1.11000xxxx··· (1.75≦M_D <1.78125) x=1.765625	2	3	3	4	5
[26] M = 1.11001xxxx··· (1.78125≦M_D <1.8125)x=1.796875	2	3	3	4	5
[27] M = 1.11010xxxx··· (1.8125≦M_D <1.84375)x=1.828125	2	3	3	4	5
[28] M = 1.11011xxxx··· (1.84375≦M_D <1.875) x=1.859375	2	3	3	4	5
[29] M = 1.11100xxxx··· (1.875≦M_D <1.90625) x=1.890625	2	3	3	4	5
[30] M = 1.11101xxxx··· (1.90625≦M_D <1.9375)x=1.921875	2	3	3	4	5
[31] M = 1.11110xxxx··· 1.9375≦M_D <1.96875) x=1.953125	2	3	3	4	5
[32] M = 1.11111xxxx··· (1.96875≦M_D <2) x=1.984375	2	3	3	4	5