NUMERICAL METHODS
Week-2
19.02.2013

Error Analysis and Number Representations

Asst. Prof. Dr. Berk Canberk

Error Analysis

- **Why** do we measure?
- **What** do we measure?
- **How** do we measure?

Why do we measure errors?

1) To determine the accuracy of numerical results.

2) To develop stopping criteria for iterative algorithms.

What do we measure?
→Two sources of numerical errors

1) Round off error

- It is caused by representing a number approximately.

$$\frac{1}{3} \cong 0.333333 \qquad \sqrt{2} \cong 1.4142...$$

2) Truncation error

- Error caused by truncating or approximating a mathematical procedure.

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ...$$

$$Truncation\ \ Error = e^x - \left(1 + x + \frac{x^2}{2!}\right)$$

How do we measure errors? →Error Types

- True Error
- Relative True Error
- Approximate Error
- Relative Approximate Error

True Error

- The difference between the true value in a calculation and the approximate value found using a numerical method.

True Error = True Value – Approximate Value

Example—True Error

The derivative, $f'(x)$ of a function $f(x)$ can be approximated by the equation,

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

If $f(x) = 7e^{0.5x}$ and $h = 0.3$

a) Find the approximate value of $f'(2)$

b) True value of $f'(2)$

c) True error for part (a)

Example (cont.)

Solution:

a) For $x = 2$ and $h = 0.3$

$$f'(2) \approx \frac{f(2+0.3) - f(2)}{0.3}$$

$$= \frac{f(2.3) - f(2)}{0.3}$$

$$= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3}$$

$$= \frac{22.107 - 19.028}{0.3} = 10.263$$

Example (cont.)

Solution:

b) The exact value of $f'(2)$ can be found by using our knowledge of differential calculus.

$$f(x) = 7e^{0.5x}$$

$$f'(x) = 7 \times 0.5 \times e^{0.5x}$$

$$= 3.5e^{0.5x}$$

So the true value of $f'(2)$ is

$$f'(2) = 3.5e^{0.5(2)}$$

$$= 9.5140$$

True error is calculated as

$$E_t = \text{True Value} - \text{Approximate Value}$$

$$= 9.5140 - 10.263 = -0.722$$

Relative True Error

- Defined as the ratio between the true error, and the true value.

$$\text{Relative True Error } (\in_t) = \frac{\text{True Error}}{\text{True Value}}$$

Example—Relative True Error

Following from the previous example for true error, find the relative true error for $f(x) = 7e^{0.5x}$ at $f'(2)$ with $h = 0.3$

From the previous example,

$$E_t = -0.722$$

Relative True Error is defined as

$$\in_t = \frac{\text{True Error}}{\text{True Value}}$$

$$= \frac{-0.722}{9.5140} = -0.075888$$

as a percentage,

$$\in_t = -0.075888 \times 100\% = -7.5888\%$$

Approximate Error

- What can be done if true values are not known or are very difficult to obtain?
- **Approximate error is defined as the difference between the present approximation and the previous approximation.**

Approximate Error (E_a) = Present Approximation – Previous Approximation

Example—Approximate Error

For $f(x) = 7e^{0.5x}$ at $x = 2$ find the following,

a) $f'(2)$ using $h = 0.3$

b) $f'(2)$ using $h = 0.15$

c) approximate error for the value of $f'(2)$ for part b)

Solution:

a) For $x = 2$ and $h = 0.3$

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

$$f'(2) \approx \frac{f(2+0.3) - f(2)}{0.3}$$

Example (cont.)

Solution: (cont.)

$$= \frac{f(2.3) - f(2)}{0.3}$$

$$= \frac{7e^{0.5(2.3)} - 7e^{0.5(2)}}{0.3}$$

$$= \frac{22.107 - 19.028}{0.3} = 10.263$$

b) For $x = 2$ and $h = 0.15$

$$f'(2) \approx \frac{f(2 + 0.15) - f(2)}{0.15}$$

$$= \frac{f(2.15) - f(2)}{0.15}$$

Example (cont.)

Solution: (cont.)

$$= \frac{7e^{0.5(2.15)} - 7e^{0.5(2)}}{0.15}$$

$$= \frac{20.50 - 19.028}{0.15} = 9.8800$$

c) So the approximate error, E_a is

E_a = Present Approximation – Previous Approximation

$$= 9.8800 - 10.263$$

$$= -0.38300$$

Relative Approximate Error

- Defined as the ratio between the approximate error and the present approximation.

$$\text{Relative Approximate Error } (\in_a) = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$

Example—Relative Approximate Error

For $f(x) = 7e^{0.5x}$ at $x = 2$, find the relative approximate error using values from $h = 0.3$ and $h = 0.15$

Solution:

From Example 3, the approximate value of $f'(2) = 10.263$ using $h = 0.3$ and $f'(2) = 9.8800$ using $h = 0.15$

E_a = Present Approximation – Previous Approximation

$= 9.8800 - 10.263$

$= -0.38300$

Example (cont.)

Solution: (cont.)

$$\in_a = \frac{\text{Approximate Error}}{\text{Present Approximation}}$$

$$= \frac{-0.38300}{9.8800} = -0.038765$$

as a percentage,

$$\in_a = -0.038765 \times 100\% = -3.8765\%$$

Absolute relative approximate errors may also need to be calculated,

$$|\in_a| = |-0.038765| = 0.038765 \text{ or } 3.8765\%$$

Taylor & Maclaurin Series (Revisited)

Brook Taylor(1685-1731)

Colin Maclaurin (1698–1746)

What is a Taylor series?

Some examples of Taylor series which you must have seen

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

What does this mean in plain English?

As Archimedes would have said, "*Give me the value of the function at a single point, and the value of all (first, second, and so on) its derivatives at that single point, and I can give you the value of the function at any other point*"

General Taylor Series

The general form of the Taylor series is given by

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2!}h^2 + \frac{f'''(x)}{3!}h^3 + \cdots$$

provided that all derivatives of f(x) are continuous and exist in the interval [x,x+h]

Example—Taylor Series

Find the value of $f(6)$ given that $f(4)=125,\ f'(4)=74,\ f''(4)=30,\ f'''(4)=6$ and all other higher order derivatives of $f(x)$ at $x=4$ are zero.

Solution:

$$f(x+h)=f(x)+f'(x)h+f''(x)\frac{h^2}{2!}+f'''(x)\frac{h^3}{3!}+\cdots$$

$x=4$

$h=6-4=2$

Example (cont.)

Solution: (cont.)

Since the higher order derivatives are zero,

$$f(4+2) = f(4) + f'(4)2 + f''(4)\frac{2^2}{2!} + f'''(4)\frac{2^3}{3!}$$

$$f(6) = 125 + 74(2) + 30\left(\frac{2^2}{2!}\right) + 6\left(\frac{2^3}{3!}\right)$$

$$= 125 + 148 + 60 + 8$$

$$= 341$$

Note that to find $f(6)$ exactly, we only need the value of the function and all its derivatives at some other point, in this case $x = 4$

Derivation for Maclaurin Series for e^x

Derive the Maclaurin series

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

The Maclaurin series is simply the Taylor series about the point x=0

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2!} + f'''(x)\frac{h^3}{3!} + f''''(x)\frac{h^4}{4} + f'''''(x)\frac{h^5}{5} + \cdots$$

$$f(0+h) = f(0) + f'(0)h + f''(0)\frac{h^2}{2!} + f'''(0)\frac{h^3}{3!} + f''''(0)\frac{h^4}{4} + f'''''(0)\frac{h^5}{5} + \cdots$$

Derivation (cont.)

Since $f(x) = e^x$, $f'(x) = e^x$, $f''(x) = e^x$, $\ldots$, $f^n(x) = e^x$ and $f^n(0) = e^0 = 1$

the Maclaurin series is then

$$f(h) = (e^0) + (e^0)h + \frac{(e^0)}{2!}h^2 + \frac{(e^0)}{3!}h^3 \ldots$$

$$= 1 + h + \frac{1}{2!}h^2 + \frac{1}{3!}h^3 \ldots$$

So,

$$f(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

Binary Representation

Two conversion types..

- Binary→ Decimal
- Decimal→Binary

How a Decimal Number is Represented

$$257.76 = 2\times10^{2} + 5\times10^{1} + 7\times10^{0} + 7\times10^{-1} + 6\times10^{-2}$$

Base 2

$$(1011.0011)_2 = \left(\begin{array}{l} (1\times 2^3 + 0\times 2^2 + 1\times 2^1 + 1\times 2^0) \\ +(0\times 2^{-1} + 0\times 2^{-2} + 1\times 2^{-3} + 1\times 2^{-4}) \end{array} \right)_{10}$$

$$= 11.1875$$

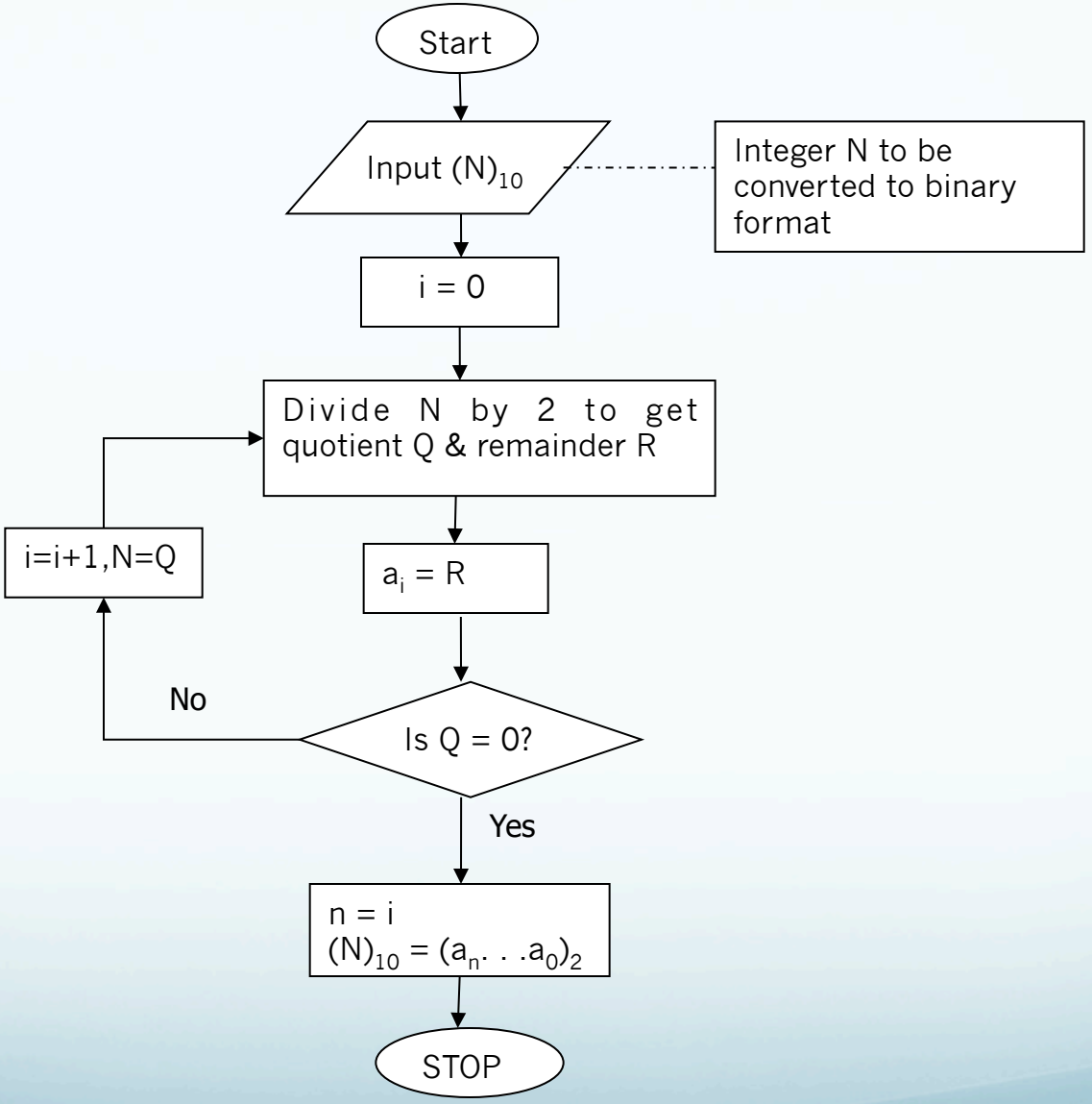
Convert Base 10 Integer to binary representation

Table 1 Converting a base-10 integer to binary representation.

	Quotient	**Remainder**
11/2	5	$1 = a_0$
5/2	2	$1 = a_1$
2/2	1	$0 = a_2$
1/2	0	$1 = a_3$

Hence

$$(11)_{10} = (a_3 a_2 a_1 a_0)_2$$
$$= (1011)_2$$


Start
Input $(N)_{10}$
Integer N to be converted to binary format
i = 0
Divide N by 2 to get quotient Q & remainder R
$a_i = R$
Is Q = 0?
No
i=i+1,N=Q
Yes
n = i
$(N)_{10} = (a_n . . . a_0)_2$
STOP

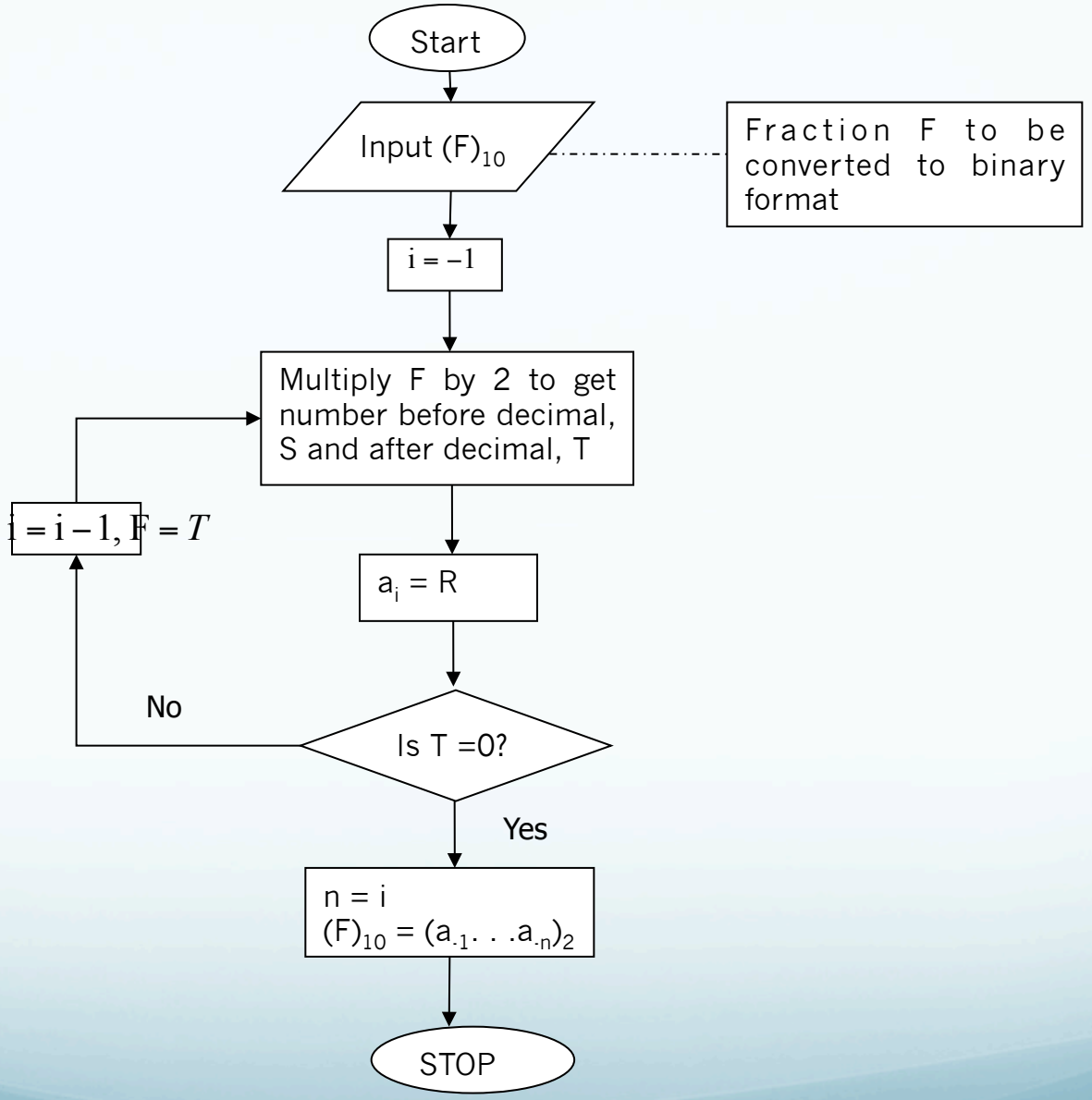
Fractional Decimal Number to Binary

Table 2. Converting a base-10 fraction to binary representation.

	Number	Number after decimal	Number before decimal
0.1875×2	0.375	0.375	$0 = a_{-1}$
0.375×2	0.75	0.75	$0 = a_{-2}$
0.75×2	1.5	0.5	$1 = a_{-3}$
0.5×2	1.0	0.0	$1 = a_{-4}$

Hence

$$(0.1875)_{10} = (a_{-1}a_{-2}a_{-3}a_{-4})_2$$
$$= (0.0011)_2$$


Start
Input $(F)_{10}$
Fraction F to be converted to binary format
i = –1
Multiply F by 2 to get number before decimal, S and after decimal, T
$i = i - 1, F = T$
a_i = R
No
Is T =0?
Yes
n = i
$(F)_{10} = (a_{.1} \ldots a_{.n})_2$
STOP

Decimal Number to Binary

$$(11.1875)_{10} = (\ \ ?.?\ \ )_2$$

Since

$$(11)_{10} = (1011)_2$$

and

$$(0.1875)_{10} = (0.0011)_2$$

we have

$$(11.1875)_{10} = (1011.0011)_2$$

Another Way to Look at Conversion

Convert $(11.1875)_{10}$ to base 2

$$(11)_{10} = 2^3 + 3$$

$$= 2^3 + 2^1 + 1$$

$$= 2^3 + 2^1 + 2^0$$

$$= 1\times 2^3 + 0\times 2^2 + 1\times 2^1 + 1\times 2^0$$

$$= (1011)_2$$

$$(0.1875)_{10} = 2^{-3} + 0.0625$$

$$= 2^{-3} + 2^{-4}$$

$$= 0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4}$$

$$= (.0011)_2$$

$$(11.1875)_{10} = (1011.0011)_2$$

All Fractional Decimal Numbers Cannot be Represented Exactly

Table 3. Converting a base-10 fraction to approximate binary representation.

	Number	Number after decimal	Number before Decimal
0.3×2	0.6	0.6	$0 = a_{-1}$
0.6×2	1.2	0.2	$1 = a_{-2}$
0.2×2	0.4	0.4	$0 = a_{-3}$
0.4×2	0.8	0.8	$0 = a_{-4}$
0.8×2	1.6	0.6	$1 = a_{-5}$

$$(0.3)_{10} \approx (a_{-1}a_{-2}a_{-3}a_{-4}a_{-5})_2 = (0.01001)_2 = 0.28125$$

What is the true error and appx error??

Floating Point Representation

Floating Decimal Point : Scientific Form

256.78 is written as $+2.5678\times10^{2}$

0.003678 is written as $+3.678\times10^{-3}$

-256.78 is written as -2.5678×10^{2}

Example

The form is

$$\text{sign} \times \text{mantissa} \times 10^{\text{exponent}}$$

or

$$\sigma \times m \times 10^{e}$$

Example:

$$-2.5678 \times 10^{2}$$

$$\sigma = -1$$

$$m = 2.5678$$

$$e = 2$$

Floating Point Format for Binary Numbers

$$y = \sigma \times m \times 2^{e}$$

$$\sigma = \text{sign of number} \quad (0 \text{ for } +,\ 1 \text{ for } -)$$

$$m = \text{mantissa} \left[(1)_2 < m < (10)_2 \right]$$

1 is not stored as it is always given to be 1.

$$e = \text{integer exponent}$$

List all the floating-point numbers that can be expressed in the form

$$x = \pm(0.b_1b_2b_3)_2 \times 2^{\pm k} \qquad (k, b_i \in \{0, 1\})$$

There are two choices for the $\pm$, two choices for b_1, two choices for b_2, two choices for b_3, and three choices for the exponent. Thus, at first, one would expect $2 \times 2 \times 2 \times 2 \times 3 = 48$ different numbers. However, there is some duplication. For example, the nonnegative numbers in this system are as follows:

$$0.000 \times 2^0 = 0 \qquad 0.000 \times 2^1 = 0 \qquad 0.000 \times 2^{-1} = 0$$

$$0.001 \times 2^0 = \frac{1}{8} \qquad 0.001 \times 2^1 = \frac{1}{4} \qquad 0.001 \times 2^{-1} = \frac{1}{16}$$

$$0.010 \times 2^0 = \frac{2}{8} \qquad 0.010 \times 2^1 = \frac{2}{4} \qquad 0.010 \times 2^{-1} = \frac{2}{16}$$

$$0.011 \times 2^0 = \frac{3}{8} \qquad 0.011 \times 2^1 = \frac{3}{4} \qquad 0.011 \times 2^{-1} = \frac{3}{16}$$

$$0.100 \times 2^0 = \frac{4}{8} \qquad 0.100 \times 2^1 = \frac{4}{4} \qquad 0.100 \times 2^{-1} = \frac{4}{16}$$

$$0.101 \times 2^0 = \frac{5}{8} \qquad 0.101 \times 2^1 = \frac{5}{4} \qquad 0.101 \times 2^{-1} = \frac{5}{16}$$

$$0.110 \times 2^0 = \frac{6}{8} \qquad 0.110 \times 2^1 = \frac{6}{4} \qquad 0.110 \times 2^{-1} = \frac{6}{16}$$

$$0.111 \times 2^0 = \frac{7}{8} \qquad 0.111 \times 2^1 = \frac{7}{4} \qquad 0.111 \times 2^{-1} = \frac{7}{16}$$

Altogether there are 31 distinct numbers in the system. The positive numbers obtained are shown on a line in Figure 2.1. Observe that the numbers are symmetrically but unevenly distributed about zero. ■

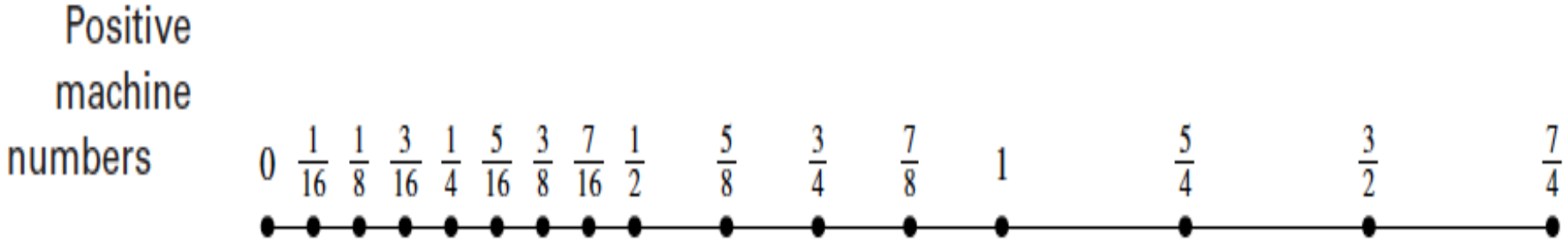

Positive
machine
numbers
0
1/16
1/8
3/16
1/4
5/16
3/8
7/16
1/2
5/8
3/4
7/8
1
5/4
3/2
7/4

Figure 2.1

Example

9 bit-hypothetical word

- the first bit is used for the sign of the number,
- the second bit for the sign of the exponent,
- the next four bits for the mantissa, and
- the next three bits for the exponent

$$(54.75)_{10} = (110110.11)_2 = (1.1011011)_2 \times 2^5$$
$$\cong (1.1011)_2 \times (101)_2$$

We have the representation as

Machine Epsilon

Defined as the measure of accuracy and found by difference between 1 and the next number that can be represented

Relative Error and Machine Epsilon

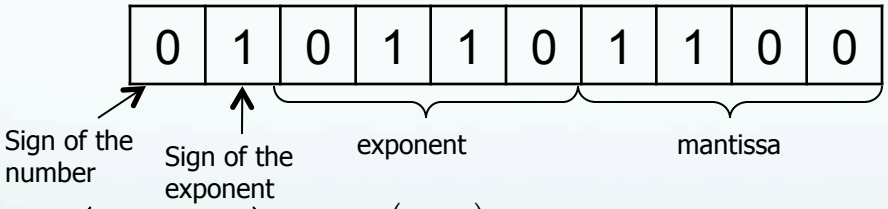
The absolute relative true error in representing a number will be less then the machine epsilon

Example

$$(0.02832)_{10} \cong (1.1100)_2 \times 2^{-5}$$

$$= (1.1100)_2 \times 2^{-(0110)_2}$$

10 bit word (sign, sign of exponent, 4 for exponent, 4 for mantissa)



$$(1.1100)_2 \times 2^{-(0110)_2} = 0.0274375$$

$$\in_a = \left| \frac{0.02832 - 0.0274375}{0.02832} \right|$$

$$= 0.034472 < 2^{-4} = 0.0625$$

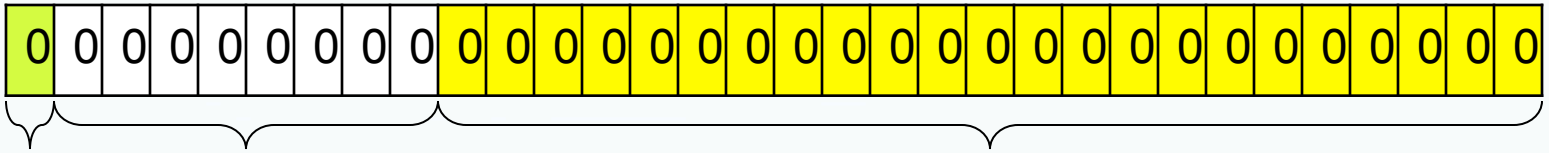
IEEE 754 Standards for Single Precision Representation (Appendix-C in Textbook)

IEEE-754 Floating Point Standard

- Standardizes representation of floating point numbers on different computers in single and double precision.
- Standardizes representation of floating point operations on different computers.

IEEE-754 Format Single Precision

32 bits for single precision



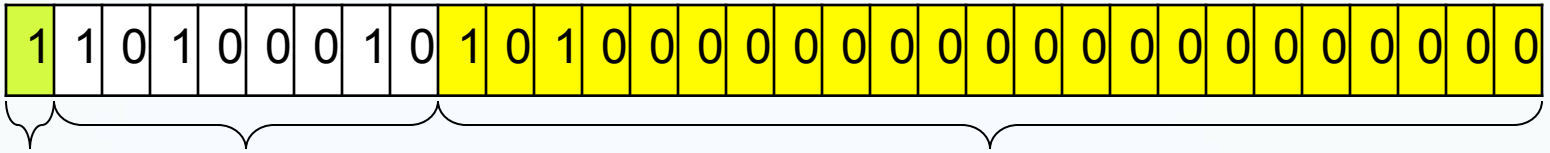
Sign (s)

Biased Exponent (e’)

Mantissa (m)

$$\text{Value} = (-1)^s \times (1.m)_2 \times 2^{e'-127}$$

Example#1



Sign (s)

Biased Exponent (e')

Mantissa (m)

$$\text{Value} = (-1)^s \times (1.m)_2 \times 2^{e'-127}$$

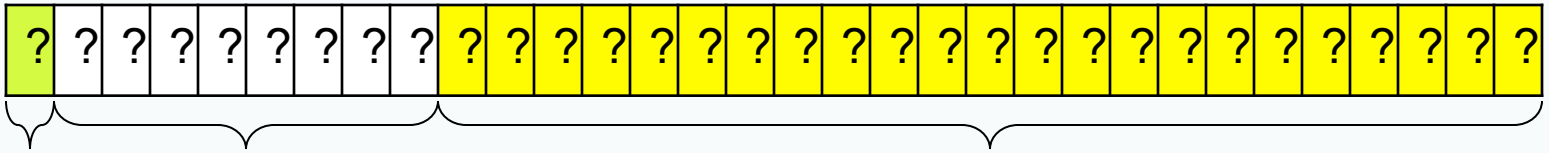
$$= (-1)^1 \times (1.10100000)_2 \times 2^{(10100010)_2 - 127}$$

$$= (-1) \times (1.625) \times 2^{162-127}$$

$$= (-1) \times (1.625) \times 2^{35} = -5.5834 \times 10^{10}$$

Example#2

Represent -5.5834×10^{10} as a single precision floating point number.



Sign (s)

Biased Exponent (e')

Mantissa (m)

$$-5.5834 \times 10^{10} = (-1)^{1} \times (1.?) \times 2^{\pm ?}$$

Exponent for 32 Bit IEEE-754

8 bits would represent

$$0 \leq e' \leq 255$$

Bias is 127; so subtract 127 from representation

$$-127 \leq e \leq 128$$

Exponent for Special Cases

Actual range of e'

$$1 \le e' \le 254$$

$e' = 0$ and $e' = 255$ are reserved for special numbers

Actual range of e

$$-126 \le e \le 127$$

Special Exponents and Numbers

$e' = 0$ — all zeros

$e' = 255$ — all ones

s	e'	m	Represents
0	all zeros	all zeros	0
1	all zeros	all zeros	-0
0	all ones	all zeros	∞
1	all ones	all zeros	$-\infty$
0 or 1	all ones	non-zero	NaN

IEEE-754 Format

The largest number by magnitude

$$(1.1.....1)_2 \times 2^{127} = 3.40 \times 10^{38}$$

The smallest number by magnitude

$$(1.00.....0)_2 \times 2^{-126} = 2.18 \times 10^{-38}$$

Machine epsilon

$$\varepsilon_{mach} = 2^{-23} = 1.19 \times 10^{-7}$$