Numerical analysis
Chapter 0: Fundamentals

Purpose of book

Understanding the way computer calculations work

Evaluating polynomials

Evaluate $P(x) = 2x^4 + 3x^3 - 3x^2 + 5x - 1$

$x = \frac{1}{2}$ is stored in memory

Method 1 (naive)

simply substitute x with $\frac{1}{2}$

this will require 10 mults. and 4 adds.

Method 2 (store powers)

First, calculate powers of x, and store them

$x = \frac{1}{2}$

$x^2 = x \cdot x = \frac{1}{4}$

$x^3 = x \cdot x^2 = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$ and so on...

And then substitute

This will require 7 mults (P_0), still 4 adds

Method 3 (nesting / Horner's method)

Rewrite for inside out eval.
this will require 4 mults (?) and 4 adds

> A degree n polynom. can be eval. in n mults. and n adds.

Binary numbers

Binary numbers are represented as ones & zeros
yadda yadda

Decimal to binary

53,7

1. break into integer & fraction
   53 0,7

2. repetitively divide integer part by 2 (see book), round down, note remainder
3. repet. multiply fraction part with 2

Binary to decimal

Integer part

Simply multiply by 2^{p-1} where p = position

For each 1

$$\underset{4}{1}\quad \underset{3}{0}\quad \underset{2}{1}\quad \underset{1}{1}$$

$$2^{4-1} + 0 + 2^{2-1} + 2^{1-1} = 2^3 + 2 + 1 = 11$$

Fractional part

Multiply by $2^{-p} = \frac{1}{2^p}$

$$.\ \underset{1}{1}\quad \underset{2}{0}\quad \underset{3}{1}\quad \underset{4}{1}$$

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

Floating point representation

IEEE 754 in this book

rounding errors are inevitable !?

Floating point formats

Three parts:

1. sign (+ or -)
2. mantissa (the important bits)
3. exponent

Three precisions:

	Sign	Exponent	Mantissa
Single	1	8	23
Double	1	11	52
Long double	1	15	64

Normalized form

$$\pm\ 1.bbb\ldots b \times 2^{p}$$

- $\pm$: sign
- $1.bbb\ldots b$: mantissa
  - 1.: radix point
  - justification (left)
- 2^{p}: exponent in 10_2

Machine epsilon

ϵ_{mach} = distance between 1 + the smallest float and 1

Truncation methods

When a number does not fit into the finite bits available, we have to truncate it

Chopping

Simply disclude any out-of-bounds bits, losing accuracy. Moves the result towards zero

Rounding (IEEE Rounding to Nearest rule)

1. In IEEE 754, if bit $n+1$ is 1: $1.x....x|11xx....$
   add 1 to bit n (round up) $1.x....x + 0.0...01$
2. if $n+1$ is 0: $1.x.....x|0x....$
   do nothing (round down) $1.x.....x$

3. if n is 0, $n+1 = 1$, but $n+2 = 0$: $1.x....x0|10x$
   do nothing (round down) $1.x.....x0$

The goal is to be as close as possible to avoid drifting

Rounding errors

x_c = computed x = actual value

Absolute error

$$|x_c - x|$$

Relative error

$$\frac{|x_c - x|}{|x|}$$

eg. if error small & number big, relative error is negligible

In IEEE this is no more than $\frac{1}{2}\epsilon_{mach}$

Machine representation

In a computer a float is stored as an 8-byte word (64 bits)

s $e_1 \dots e_{11}$ $b_1 \dots b_{52}$

↑ sign | exponent | mantissa

0 = positive 1 = negative

The exponent bits is calculated by adding $2^{10} - 1 = 1023$ (exp. bias) to the exponent

for exponents between -1022 to 1023

The bias is added to convert negative exponents to positive ones (thereby not needing to sign them too)

Addition of floats

While the addition can occur with precision higher than 52 bits, the result after all of the additions will still be rounded to 52 bits, hence

$1 + 2^{-53} = 1$ in comp. arithmetic

Loss of significance

When members of a subtraction have the wrong properties, the result can suffer from loss of significance.

This can happen when subtracting a number that is almost equal to the number subtracted from.

Review of calculus

Intermediate value theorem

If f is continous in $[a, c]$ then
any value $b : a \leq b \leq c$ is
also in the graph ($f(b)$ has a value)

Continuous limits

If x_n approaches x_0 as $n \to \infty$
then $\lim_{n \to \infty} f(x_n) = f(x_0)$

if f is continuous.

(otherwise the limit has to be calculated outside)

Mean value theorem

If f is continuously diff. in $[a, c]$
then it is also diff. in $b : a \leq b \leq c$

Rolle's theorem

If f is cont. diff. in $[a, c]$ and $f(a) = f(c)$
then a turning point b, $f'(b) = 0$, $: a \leq b \leq c$
must exist

Taylor's theorem with remainder

If the remainder is small, Taylor's theorem provides a good approximation.

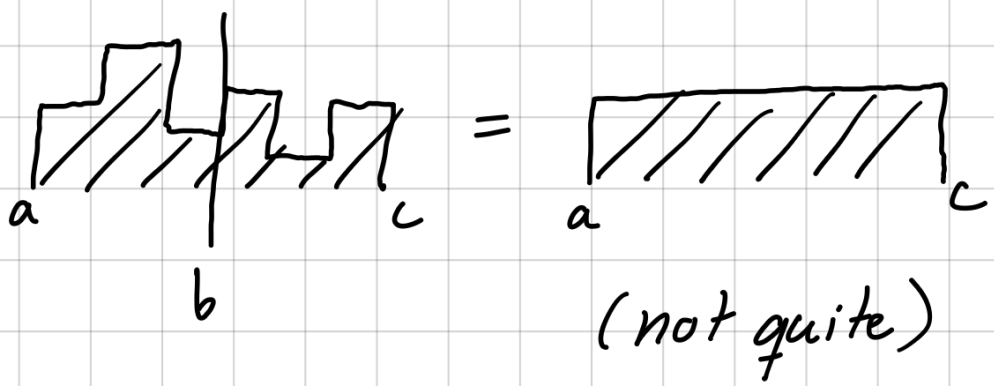
This is good for simplifying poly. eval.

Mean value theorem for integrals

If f is cont. in $[a, c]$ and
g's sign remains static in $[a, c]$
then, for b : $a \leq b \leq c$

$$\int_a^c f(x)g(x)\,dx = f(b)\int_a^c g(x)\,dx$$

in other words, there exist some value b that results in the same area



(not quite)

Exercise solutions

Polynomial evaluation (0.1)

1a) $x = \frac{1}{3}$ $\quad P(x) = 6x^4 + x^3 + 5x^2 + x + 1$

Eval. without:

$$P\left(\tfrac{1}{3}\right) = 6 \cdot \tfrac{1}{3^4} + \tfrac{1}{3^3} + 5 \cdot \tfrac{1}{3^2} + \tfrac{1}{3} + 1$$

$$= \frac{6}{81} + \frac{1}{27} + \frac{5}{9} + \frac{1}{3} + \frac{3}{3}$$

$$= \frac{6+3+45+27+81}{81} = \frac{162}{81} = 2$$

Retrieve nested form:

$$P(x) = 6x^4 + x^3 + 5x^2 + x + 1$$

$$= 1 + x(1 + 5x + x^2 + 6x^3)$$

$$= 1 + x(1 + x(5 + x + 6x^2))$$

$$= 1 + x(1 + x(5 + x(1 + 6x)))$$

Eval with

(1) $\quad 1 + 6x = 1 + 2 = 3$

(2) $\Rightarrow \quad 5 + x(1+6x) = 5 + \frac{1}{3} \cdot 3 = 6$

(3) $\Rightarrow \quad 1 + x(\ldots = 1 + \frac{1}{3} \cdot 6 = 3$

(4) $\Rightarrow \quad 1 + x(\ldots = 1 + \frac{1}{3} \cdot 3 = 2$

b) Eval without

$$P\left(\frac{1}{5}\right) = -3 \cdot \frac{1}{5^4} + 4 \cdot \frac{1}{5^3} + 5\,\frac{1}{5^2} - 5\,\frac{1}{5} + 1$$

$$= -\frac{1}{27} + \frac{4}{27} + \frac{5}{9} - \frac{5}{3} + \frac{3}{3}$$

$$= \frac{-1 + 4 + 15 - 45 + 27}{27} = 0$$

Horner's

$$P(x) = 1 + x(-5 + x(5 + x(4 - 3x)))$$

Eval with

(1) $4 - 3x = 4 - 1 = 3$

(2) $5 + x(\ldots = 5 + \frac{1}{5}\,3 = 6$

(3) $-5 + x(\ldots = -5 + \frac{1}{5}\,6 = -3$

(4) $1 + x(\ldots = 1 + \frac{1}{5} \cdot -3 = 0$

c) Eval without

$$P\left(\frac{1}{3}\right) = 2\,\frac{1}{3^4} + \frac{1}{3^3} - \frac{1}{3^2} + 1$$

$$= \frac{2}{81} + \frac{1}{27} - \frac{1}{9} + \frac{3}{3}$$

$$= \frac{2 + 3 - 9 + 81}{81} = \frac{77}{81}$$

Horners

$$P(x) = 1 + x(0 + x(-1 + x(1 + 2x)))$$

Eval with

(1) $1+2x = \frac{5}{3}$

(2) $-1+x(\ldots = -1+\frac{5}{9} = \frac{-4}{9}$

(3) $0+x(\ldots = 0-\frac{4}{27} = -\frac{4}{27}$

(4) $1+x(\ldots = 1-\frac{4}{81} = \frac{77}{81}$

2a) Eval w/o

$$P(-\tfrac{1}{2}) = -6\tfrac{1}{2^3} - 2\tfrac{1}{2^2} + 3\tfrac{1}{2} + 7$$

$$= -\frac{3}{4} - \frac{1}{2} + \frac{3}{2} + \frac{28}{4}$$

$$= \frac{-3-2+6+28}{4} = \frac{29}{4}$$

Horner's

$$P(x) = 7 + x(-3 + x(-2+6x))$$

Eval w/

(1) $-2+6x = -2-3 = -5$

(2) $-3+x(\ldots = -3-\frac{1}{2}(-5) = -\frac{1}{2}$

(3) $7+x(\ldots = 7-\frac{1}{2}(-\frac{1}{2}) = 7+\frac{1}{4} = \frac{29}{4}$

b) Eval w/o

$$P(-\tfrac{1}{2}) = -8\tfrac{1}{2^5} - \tfrac{1}{2^4} + 3\tfrac{1}{2^3} + \tfrac{1}{2^2} + \tfrac{3}{2} + 1$$

$$= -\frac{8}{32} - \frac{1}{16} + \frac{3}{8} + \frac{1}{4} + \frac{3}{2} + \frac{2}{2}$$

$$= \frac{-4 - 1 + 6 + 4 + 24 + 16}{16}$$

$$= \frac{45}{16}$$

Horner's

$$P(x) = 1 + x(-3 + x(1 + x(-3 + x(-1 + 8x))))$$

Eval w/

(1) $-1 + 8x = -1 - 4 = -5$

(2) $-3 + x(\ldots = -3 - \frac{1}{2}(-5) = -\frac{1}{2}$

(3) $1 + x(\ldots = 1 - \frac{1}{2}(-\frac{1}{2}) = \frac{5}{4}$

(4) $-3 + x(\ldots = -3 - \frac{1}{2}\frac{5}{4} = \frac{-29}{8}$

(5) $1 + x(\ldots = 1 - \frac{1}{2}(\frac{-29}{8}) = \frac{16 + 29}{16} = \frac{45}{16}$

c) Shipping, getting tedious

3) assume $t = x^2$

means that if $x = \frac{1}{2}$ then $t = \frac{1}{4}$

and $P(t) = t^3 - 4t^2 + 2t + 1$

Eval w/o

$$P\left(\frac{1}{4}\right) = \frac{1}{4^3} - 4\frac{1}{4^2} + 2\frac{1}{4} + 1$$

$$= \frac{1}{64} - \frac{1}{4} + \frac{2}{4} + \frac{4}{4}$$

$$= \frac{1 + 16 + 64}{64} = \frac{81}{64}$$

Horner's

$$P(t) = 1 + t(2 + t(-4 + t))$$

Eval w/

(1) $-4 + t = -4 + \frac{1}{4} = -\frac{15}{4}$

(2) $2 + t(\ldots = 2 + \frac{1}{4}\left(-\frac{15}{4}\right) = \frac{17}{16}$

(3) $1 + t(\ldots = 1 + \frac{1}{4}\left(\frac{17}{16}\right) = \frac{81}{64}$

4a) I'm a little confused what the intent of the question is...

am I supposed to convert it to nested and then evaluate, or just evaluate?

Since it says eval., I guess I'll just eval.

(1) $\frac{1}{2} + (5-3)(-\frac{1}{2}) = \frac{1}{2} - \frac{2}{2} = -\frac{1}{2}$

(2) $\frac{1}{2} + (5-2)(\ldots = \frac{1}{2} - \frac{3}{2} = -1$

(3) $1 + 5(\ldots = -4$

b) (1) $\frac{1}{2} + (-1-3)(-\frac{1}{2}) = \frac{1}{2} + \frac{4}{2} = \frac{5}{2}$

(2) $\frac{1}{2} + (-1-2)\ldots = \frac{1}{2} - \frac{15}{2} = -\frac{14}{2} = -7$

(3) $1 - (\ldots = 8$

5a) (1) $3 + (\frac{1}{2} - 3)(2) = -2$

(2) $1 + (\frac{1}{2} - 2)(\ldots = 4$

(3) $4 + (\frac{1}{2} - 1)(\ldots = 2$

(4) $4 + \frac{1}{2}(\ldots = 5$

b) (1) $3 + (-\frac{1}{2} - 3)(2) = -4$

(2) $1 + (-\frac{1}{2} - 2)(\ldots = 1 + \frac{20}{2} = 11$

(3) $4 + (-\frac{1}{2} - 1)(\ldots = 4 - \frac{33}{2}$

(4) $4 - \frac{1}{2}(\ldots = 4 - 2 + \frac{33}{4} = 2 + \frac{33}{4} = \frac{41}{4}$

6a) In this case we should first assert $t = x^5$, then apply Horner's

$$P(t) = a_0 + a_5 t + a_{10} t^2 + a_{15} t^3$$

$$= a_0 + t(a_5 + t(a_{10} + t(a_{15})))$$

Which would result in 3 mults & 3 adds

b) Here, the difference in powers follows the pattern $n + m \cdot 5$, where $n = 7$

therefore, we can rewrite as

$$P(x) = x^7(a_7 + x^5(a_{12} + x^5(a_{17} + x^5(a_{22} + x^5(a_{27})))))$$

This requires 5 mults and 4 adds

7) n mults, 2n adds

Computer problems

See seperate file for source code.

1) Input: nest(50, ones(50), 1.00001)

Output: 51,0128

Actual value: 51,012752082452

Home → Preferences → Command window → Format → long g to disable rounding

$$\text{Error} = 4{,}76 \cdot 10^{-12}$$

2) Code for array: (-1).^(0:n)

Input: nest(99, (-1).^(0:99), 1.00001)

Output: -0,00050024507964763 2

$$1 - x(1 - x(\ldots = \frac{1 - x^{100}}{1 + x}$$

$$\text{Error} = 1{,}713 \cdot 10^{-16}$$

No way to check if this is correct, lol.

Binary numbers (0.2)

1a) 64:

64 / 2	0
32 / 2	0
16 / 2	0
8 / 2	0
4 / 2	0
2 / 2	0
1 / 2	1

$\Rightarrow 1000000$

b)

17 / 2	1
8 / 2	0
4 / 2	0
2 / 2	0
1 / 2	1

$\Rightarrow 10001$

c)

79 / 2	1
39 / 2	1
19 / 2	1
9 / 2	1
4 / 2	0
2 / 2	0
1 / 2	1

$\Rightarrow 1001111$

d)
227/2	1
113/2	1
56/2	0
28/2	0
14/2	0
7/2	1
3/2	1
1/2	1

$\Rightarrow$ 11100011

2a) $\frac{1}{8} = \frac{1}{2^3} \Rightarrow .0001$

b) $\frac{7}{8} \times 2 = \frac{6}{8} + 1$

$\frac{6}{8} \times 2 = \frac{4}{8} + 1 \Rightarrow .111$

$\frac{4}{8} \times 2 = \frac{0}{8} + 1$

c) $\frac{35}{16} = \frac{32+3}{16} = 2 + \frac{3}{16}$ (integer, Fraction)

Integer

2/2	0
1/2	1

$\Rightarrow$ 10

Fraction

$\frac{3}{16} \times 2 = \frac{6}{16} + 0$

$\frac{6}{16} \times 2 = \frac{12}{16} + 0$

$\frac{12}{16} \times 2 = \frac{8}{16} + 1 \Rightarrow .0011$

$\frac{8}{16} \times 2 = \frac{0}{16} + 1$

Result $\Rightarrow$ 10.0011

d) $\frac{31}{64} \times 2 = \frac{62}{64} + 0$

$\frac{62}{64} \times 2 = \frac{60}{64} + 1$

$\frac{60}{64} \times 2 = \frac{56}{64} + 1$ $\Rightarrow .011111$

$\frac{56}{64} \times 2 = \frac{48}{64} + 1$

$\frac{48}{64}$ -11- $\frac{32}{64} + 1$

$\frac{32}{64}$ -11- $\frac{0}{64} + 1$

3a) Integer

10 / 2	0
5 / 2	1
2 / 2	0
1 / 2	1

$\Rightarrow 1010$

Fraction

$\frac{1}{2} \times 2 = \frac{0}{2} + 1 \Rightarrow .1$

Result $\Rightarrow$ 1010.1

b) $\frac{1}{3} \times 2 = \frac{2}{3} + 0$

$\frac{2}{3} \times 2 = \frac{1}{3} + 1$ $\Rightarrow .\overline{01}$

c) $\frac{5}{7} \times 2 = \frac{3}{7} + 1$

$\frac{3}{7} \times 2 = \frac{6}{7} + 0 \Rightarrow .\overline{101}$

$\frac{6}{7} \times 2 = \frac{5}{7} + 1$

d) Integer

12/2 0
6/2 0 1100
3/2 1
1/2 1

Fraction

$\frac{4}{5} \times 2 = \frac{3}{5} + 1$

$\frac{3}{5} \times 2 = \frac{1}{5} + 1$

$\frac{1}{5} \times 2 = \frac{2}{5} + 0$ $\quad .\overline{1100}$

$\frac{2}{5} \times 2 = \frac{4}{5} + 0$

Result $1100.\overline{1100}$

e) Integer

55/2 1
27/2 1
13/2 1
6/2 0
3/2 1
1/2 1

110111

Fraction

$\frac{2}{5} \times 2 = \frac{4}{5} + 0$

$\frac{4}{5} \times 2 = \frac{3}{5} + 1$

from d) $110 \quad \Rightarrow .\overline{01110}$

Result $110111.\overline{01110}$

f) $\frac{1}{10} \times 2 = \frac{2}{10} + 0$

$\frac{4}{10}$ -11-

$\frac{8}{10}$ -11-

$\frac{6}{10} + 1$

$\frac{2}{10} + 1$

(repeat)

this doesn't ↓ ← this repeats

$0.\overline{00011}$

4a) Integer

11/2 1

5/2 1 1011

2/2 0

1/2 1

Fraction

$\frac{1}{4} \times 2 = \frac{2}{4} + 0$

$\frac{2}{4} \times 2 = \frac{0}{4} + 1$

Result 1011.01

b) $\frac{2}{3} \times 2 = \frac{1}{3} + 1$

$\frac{1}{3} \times 2 = \frac{2}{3} + 0$

$0.\overline{10}$

c) $\frac{3}{5} \times 2 = \frac{1}{5} + 1$

$\frac{1}{5} \times 2 = \frac{2}{5} + 0$

$\frac{2}{5} \times 2 = \frac{4}{5} + 0$

$\frac{4}{5} \times 2 = \frac{3}{5} + 1$

$\circ\overline{1001}$

d) Integer

3/2 1

1/2 1

11

Fraction

from c) $\circ\overline{0011}$

Result $11\circ\overline{0011}$

e) Integer

30/2 0

15/2 1

7/2 1

3/2 1

1/2 1

11110

Fraction

from c) $\circ\overline{1001}$ => Result: $11110\circ\overline{1001}$

f) Integer

99	1
49	1
24	0
12	0
6	0
3	1
1	1

1100011

Fraction

$0,9 \times 2 = 0,8 + 1$
$0,8 \times 2 = 0,6 + 1$
$0,6 \times 2 = 0,2 + 1$
$0,2 \times 2 = 0,4 + 0$
$0,4 \times 2 = 0,8 + 0$

$0,\overline{11100}$

Result: $1100011,1\overline{1100}$

5) don't care

6) -//-

7a) 1010101

$2^6 + 2^4 + 2^2 + 1 = 64 + 16 + 4 + 1 = 85$

b) 1011

$2^3+2+1 = 11$

0.101

$\frac{1}{2} + \frac{1}{2^3} = \frac{2^2+1}{2^3} = \frac{5}{8}$

$11 + \frac{5}{8} = \frac{88+5}{8} = \frac{93}{8}$

c) 10111

$2^4+2^2+2+1 = 23$

$x = 0.\overline{01}$

$2^2x = 01.\overline{01}$ apply shift

$2^2x - x = 01 = 1_{10}$ determine shifted dec.

$x(2^2-1) = 1 \iff x = \frac{1}{2^2-1} = \frac{1}{3}$ calc x

Result $23 + \frac{1}{3} = \frac{23 \cdot 3 + 1}{3} = \frac{70}{3}$

d) $110 = 2^2+2 = 6$

$x = \overline{10}$

Shift

$2^2x = 10.\overline{10}$

Determine shifted dec.

$x(2^2 - 1) = 10 = 2_{10}$

$x = \frac{2}{2^2-1} = \frac{2}{3}$ calc. x

Result $6 + \frac{2}{3} = \frac{20}{3}$

e) $10 = 2$

$x = \overline{110}$

$2^3 x = 110.\overline{110}$

$x(2^3 - 1) = 110 = 6$

$x = \frac{6}{7}$

$2 + \frac{6}{7} = \frac{20}{7}$

f) $110 = 6$

$x = .\overline{1101}$

$y = 2x = 1.\overline{101}$

$z = \overline{101} \Rightarrow 2^3 z = 101.\overline{101}$

$\Rightarrow z(2^3 - 1) = 101 = 5$

$$z = \frac{5}{7}$$

$$\Rightarrow \eta = 1 + \frac{5}{7} = \frac{12}{7}$$

$$\Rightarrow x = \frac{6}{7}$$

$$\Rightarrow \text{Result} = 6 + \frac{6}{7} = \frac{48}{7}$$

g) Integer $10 = 2$

Fraction

$$x = .0101\overline{1101}$$

$$\eta = 2^3 x = 010.\overline{1101}$$

$$z = .\overline{1101} \Rightarrow 2^4 z = 1101.\overline{1101}$$

$$z(2^4 - 1) = 1101 = 13_{10}$$

$$z = \frac{13}{15}$$

$$\eta = 2 + \frac{13}{15} = \frac{43}{15}$$

$$\Rightarrow x = \frac{43}{2^3 \cdot 15} = \frac{43}{120}$$

$$\text{Result} = 2 + \frac{43}{120} = \frac{283}{120}$$

h) Integer = 111 = 7

Fraction

$x = 0.\overline{1}$

$2x = 1.\overline{1}$

$x(2-1) = 1_2 = 1_{10}$

$x = \frac{1}{1} = 1$

Result = 7 + 1 = 8

why? because $0.\overline{1} = 0.111\ldots1$

for infinity which is equal to

$\frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} \ldots + \frac{1}{2^\infty}$

which equates to 1

8a)

$11011 = 2^4 + 2^3 + 2 + 1 = 16 + 8 + 2 + 1$

$= 27$

b) Integer = $110111 = 2^5 + 2^4 + 2^2 + 2 + 1$

$= 32 + 16 + 4 + 3 = 55$

Fraction = $0.001 = \frac{1}{2^3} = \frac{1}{8}$

Result = $55 + \frac{1}{8} = \frac{441}{8}$

c) Integer = $111 = 2^2 + 2 + 1 = 7$

Fraction

$x = 0.\overline{001} \Rightarrow x(2^3 - 1) = 001 = 1$

$\Rightarrow x = \frac{1}{7}$

Result $= 7 + \frac{1}{7} = \frac{50}{7}$

d) Integer = $1010 = 2^3 + 2 = 10$

Fraction

$x = 0.\overline{01} \Rightarrow x(2^2 - 1) = 01 = 1$

$x = \frac{1}{3}$

Result $= 10 + \frac{1}{5} = \frac{31}{3}$

e) Integer = $10111 = 2^4 + 2^2 + 2 + 1 = 23$

Fraction

$x = 0.1\overline{0101} = 0.1\overline{01}$

$\eta = 2x = 1.\overline{01}$

$z = 0.\overline{01} = \frac{1}{3}$ (see d))

$\eta = 1 + \frac{1}{3} = \frac{4}{3} \Rightarrow x = \frac{2}{3}$

Result $= 23 + \frac{2}{3} = \frac{71}{3}$

f) Integer = $1111 = 2^3 + 2^2 + 2 + 1 = 15$

Fraction

$x = 0.010\overline{001}$

$\eta = 2^3 x = 010.\overline{001}$

$z = 0.\overline{001} \Rightarrow z(2^3 - 1) = 001 = 1$

$z = \frac{1}{7} \Rightarrow \eta = 2 + \frac{1}{7} = \frac{15}{7}$

$\Rightarrow x = \frac{15}{2^3 \cdot 7} = \frac{15}{56}$

Result $= 15 + \frac{15}{56} = \frac{855}{56}$

Floating point representation (0.3)

1a) $\frac{1}{4} \times 2 = \frac{2}{4} + 0$

$\frac{2}{4} \times 2 = \frac{0}{4} + 1 \quad \Rightarrow .01$

Left justify: $\frac{1}{4}_{10} = 2^2 \times 0.01 \times 2^{-2}$

$= 01.0 \times 2^{-2}$

gives

$+ 1.00\ldots.0 \times 2^{-2}$

b) from 0.2.36: $.\overline{01}$

Left justify: $+ 1.0101\ldots01 \times 2^{-2}$

c) from 0.2.46: $.\overline{10}$

$+ 1.0101\ldots01 \times 2^{-1}$

d) from 0.2.4f: $.1\overline{1100}$

$+ 1.11001100\ldots1101 \times 2^{-1}$

↑
rounding rule
(if bit n+1 is 1 round up)

2a) Convert to binary

Integer

9/2 1
4/2 0
2/2 0 => 1001
1/2 1

Fraction

$\frac{1}{2} \times 2 = \frac{0}{2} + 1 \Rightarrow .1$

Result 1001.1

Convert to IEEE 754

Justify

$1001.1 = 1.0011 \times 2^3$

Pad

$+ 1.00110\ldots0 \times 2^3$

result $\underset{0}{\nabla}$!!

b) Integer is 1001 from a)
Fraction is:
0,6 x 2 = 0,2 + 1

$0,2 \times 2 = 0,4 + 0$
$0,4 \times 2 = 0,8 + 0$
$0,8 \times 2 = 0,6 + 1$
repeat => $0\overline{1001}$

Justify

$1,0011\overline{1001} \times 2^3$

Full length

$+ 1,00111001\ldots 1010 \times 2^3$
↑ rounding
1001**1**...
add 1 gives
1010

c) Integer

100 / 2	0
50	0
25	1
12	0
6	0
3	1
1	1

=> 1100100

Fraction

$0,2 \times 2 = 0,4 + 0$

$0,4 \times 2 = 0,8 + 0$

$0,8 \times 2 = 0,6 + 1$

$0,6 \times 2 = 0,2 + 1$

repeat $.\overline{0011}$

Justify

$1.1001000\overline{0011} \times 2^6$

Full

$+ 1.100100011\ldots01 \times 2^6$

↑ rounding

0011

~~aAA~~ 1

01

d) $\frac{44}{7} = 6 + \frac{2}{7}$

Integer : 110

Fraction

$\frac{2}{7} \times 2 = \frac{4}{7} + 0$

$\frac{4}{7} \times 2 = \frac{1}{7} + 1$

$\frac{1}{7} \times 2 = \frac{2}{7} + 0$

repeat => $.\overline{010}$

Justify

$1.10\overline{010} \times 2^2$

Full

$+1.100100010\ldots001 \times 2^2$

repeating $\frac{52}{3} = 17.3$ times

=> ends with 00100 (out of bounds)

=> no rounding

3) Integer 101

If $h > 0$

Fraction $0.0\ldots1\ldots0$

↑

hth position

case : no overlap

fraction (2^{-4})

$1.0100\ldots1\ldots0 \times 2^2$

Since 3 bits of the mantissa are occupied

h can be as high as 49

case : overlap

one extra position can be used

so 50

If $h < -2$

addition is simply

$010 + 0 \ldots 1 \ldots 0$

↑ hth position

$= 0 \ldots 1 \ldots 010$

which will be justified $k+3$ positions

If $-2 \leq h \leq 0$

justify by h position

So: $-49 \leq h \leq 50$

4) Integer

19/2	1
9/2	1
4/2	0
2/2	0
1/2	1

$\Rightarrow 10011$

Fraction

$0.0 \ldots 1 \ldots 0$

↑ hth position

Justify

$1.0011 \ldots 1 \ldots 0 \times 2^4$

h can be any between 0 and 53

53 will trigger rounding up, so result will still be larger. after 53

$1 + 2^{-h} = 1$

5a) Evaluating innermost gives:

1: $2^{-51} + 2^{-53} = 2^{-51} + 2^{-2} \cdot 2^{-51}$

$$\begin{aligned} & 1.0\ldots.0 \times 2^{-51} \\ + \; & 0.010\ldots.0 \times 2^{-51} \\ = \; & 1.010\ldots.0 \times 2^{-51} \end{aligned}$$

2: 1 + (previous)

$$\begin{aligned} & 1.0\ldots\ldots.00|0\ldots. \times 2^{0} \\ + \; & 0.0\ldots\ldots.10|10\ldots. \times 2^{0} \quad \text{(rule 3 !)} \\ = \; & 1.0\ldots\ldots.10 \end{aligned}$$

3: (previous) − 1

$$\begin{aligned} & 1.0\ldots\ldots.10 \times 2^{0} \\ - \; & 1.0\ldots\ldots.00 \times 2^{0} \\ = \; & 0.0\ldots\ldots.10 \times 2^{0} \\ = \; & 1.0\ldots\ldots.0 \times 2^{-51} = 2\epsilon_{mach} \end{aligned}$$

b) 1° $2^{-51} + 2^{-52} + 2^{-53}$

$1.0\ldots\ldots\ldots 0 \times 2^{-51}$
$+\ 0.010\ldots\ldots 0 \times 2^{-51}$
$+\ 0.0010\ldots\ldots 0 \times 2^{-51}$
$= 1.1110\ldots\ldots 0 \times 2^{-51}$

2° 1 + (previous)

$1.0\ldots\ldots\ldots 00|00\ldots \times 2^{0}$
$+\ 0.0\ldots\ldots\ldots 11|10\ldots \times 2^{0}$
$= 1.0\ldots\ldots\ldots 100 \times 2^{0}$

3° (previous) − 1

$1.0\ldots\ldots 100 \times 2^{0}$
$-\ 1.0\ldots\ldots 000 \times 2^{0}$
$= 0.0\ldots\ldots 100 \times 2^{0}$
$= 1.0\ldots\ldots\ldots 000 \times 2^{-50} = 4\epsilon_{mach}$

6a) 1° $2^{-51} + 2^{-52} + 2^{-54}$

$$
\begin{aligned}
&1.0\ldots\ldots0 \times 2^{-51} \\
+\ &0.10\ldots\ldots0 \times 2^{-51} \\
+\ &0.0010\ldots0 \times 2^{-51} \\
=\ &1.1010\ldots\ldots0 \times 2^{-51}
\end{aligned}
$$

2° 1 + (previous)

$$
\begin{aligned}
&1.0\ldots\ldots\ 000|000\ldots \times 2^{0} \\
+\ &0.0\ldots\ldots\ 011|010\ldots \times 2^{0} \\
=\ &1.0\ldots\ldots011 \times 2^{0}
\end{aligned}
$$

3° (previous) - 1

$$
\begin{aligned}
&1.0\ldots\ldots011 \times 2^{0} \\
-\ &1.0\ldots\ldots000 \times 2^{0} \\
=\ &0.0\ldots\ldots011 \times 2^{0} \\
=\ &1.10\ldots\ldots000 \times 2^{-51} \\
=\ &110.0\ldots\ldots000 \times 2^{-53} \\
=\ &3\epsilon_{mach}
\end{aligned}
$$

b) Result will be same as a), as anything $< \epsilon_{mach}$ gets chopped off

only step 1 will be different

$\Rightarrow$ result ° $3\epsilon_{mach}$

7a) Integer

8|2 0
4|2 0
2|2 0
1|2 1

$\Rightarrow 1000_2 = 1 \times 2^3$

$= 1.0\ldots0 \times 2^1$

Assemble

Sign is 0 (positive)

exponent is $3 + 1023 = 1026_{10} = 1024_{10} + 2_{10}$

$= 10000000010_2$

In binary, we then get

0 10....010 00.....00000

↑ sign | exponent | mantissa

Segmented into 4-bits we have

0100|....|0010|0...|(12 times)....

sign & exponent | mantissa

giving: 4020000000000000

b) Integer

21	1
10	0
5	1
2	0
1	1

$\Rightarrow 1.0101 \times 2^4$

Assemble

sign 0

exponent 4 $\Rightarrow$ $4 + 1023 = 1024 + 2 + 1$

$= \underbrace{10\ldots011}_{11 \text{ bits}}$

0100|0000|0011|0101|.... 12 zeroes

$= 40350\ldots.0_{16}$

c) $1/8_{10} = 0.001_2 = 1.0\ldots0 \times 2^{-3}$

(from 0.2.2a))

$-3 + 1023 = 1020$

1020	0
510	0
255	1
127	1
63	1
31	1
15	1
7	1
3	1

$\Rightarrow 1111111100 = 1.1111111100 \times 2^9$

1 1

0011|1111|1100| (13 zeros)

$= 3FC0\ldots0_{16}$

d) $\frac{1}{3} \times 2 = \frac{2}{3} + 0$

$\frac{2}{3} \times 2 = \frac{1}{3} + 1$

$\Rightarrow 0.\overline{01}$

$= 1.\overline{01} \times 2^{-2}$

$-2 + 1023 = 1021$

$= 1.1111111101 \times 2^{a}$ (see c))

0011|1111|1101|0101|... repeat 12 times

$= 3FD55\ldots5_{16}$

shipping rest... i get it !!

8) $\frac{2}{3} = 0.\overline{10}_2 = 1.0\overline{10} \times 2^{-1}$

$0.10101\ldots01|01\ldots \times 2^{-1}$

$+ 1.01010\ldots10|10\ldots \times 2^{-1}$

$= 10.0\ldots00 \times 2^{-1}$

$= 1.00\ldots00 \times 2^{0} = 1_{10}$

If rounding wasn't used the result would be different

9a) Because $(7/3 - 4/3) - 1 = \epsilon_{mach}$

b) No

10a) $1.0\ldots0 \times 2^0$

$0.0\ldots0\,|\,10\ldots \times 2^0$

$= 1.0\ldots0$

no

b) no

11) no

12a) $\frac{1}{3}$ is $3FD55\ldots5_{16}$ (from 7d))

whereas it would actually be

$3FD55\ldots5\ldots_{16}$ (infinite 5's)

The difference would be

$-\ 3FD00\ldots05\ldots_{16}$

Sign swap $= 0.00\ldots0\,|\,0101\ldots \times 2^{-2}$

$= 1.\overline{01} \times 2^{-56}$

$\epsilon_{mach}/2 = 2^{-53}$

so the diff $< \epsilon_{mach}/2$, I think...

b) c) Shipping due to lack of guidance

13) Shipping

14a) $4.3 \Rightarrow \left\{ \begin{array}{l} 4_{10} = 100 \\ .3_{10} = 0\overline{1001} \end{array} \right\} 100.0\overline{1001}$

$= 1.0001\overline{1001} \times 2^2$

$3.3_{10} = 1.101\overline{1001} \times 2^1$

don't know how to perform this subtraction

Shipping...

Shipping rest of this section ∇_0

Loss of significance (0.4)

1a) not familiar with sec function

shipping

b) near 0, where it becomes closer to $\frac{1-\sim 1}{x}$

not sure how to fix as there are no instructions

c) loss of significance near 0

again, no idea how to fix...

Shipping rest of this section as it lacks guidance

Review of calculus (0.5)

Skipping, as this has been repeated ad nauseam in previous courses.