Numerisk Analys
FMNF05

Emil Wihlander
dat15ewi@student.lu.se

22 januari 2018

Kapitel 0: Fundamentals

0.1 Evaluating a Polynomial

0.1.1 a)

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

With nested:

$$6 \cdot 1/3 + 1 = 3$$
$$3 \cdot 1/3 + 5 = 6$$
$$6 \cdot 1/3 + 1 = 3$$
$$3 \cdot 1/3 + 1 = 2$$

Without nested:

$$6 \cdot (1/3)^4 + (1/3)^3 + 5 \cdot (1/3)^2 + 1/3 + 1$$
$$6/81 + 1/27 + 5/9 + 1/3 + 1$$
$$6/81 + 3/81 + 45/81 + 27/81 + 81/81$$
$$162/81 = 2$$

b)

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

With nested:

$$-3 \cdot 1/3 + 4 = 3$$
$$3 \cdot 1/3 + 5 = 6$$
$$6 \cdot 1/3 - 5 = -3$$
$$-3 \cdot 1/3 + 1 = 0$$

Without nested:

$$-3 \cdot (1/3)^4 + 4 \cdot (1/3)^3 + 5 \cdot (1/3)^2 - 5 \cdot (1/3) + 1$$
$$-3/81 + 4/27 + 5/9 - 5/3 + 1$$
$$-3/81 + 12/81 + 45/81 - 135/81 + 81/81$$
$$0/81 = 0$$

c)

$$P(x) = 2x^4 + x^3 - x^2 + 1 = 1 + x(0 + x(-1 + x(1 + 2x)))$$

With nested:

$$2 \cdot 1/3 + 1 = 5/3$$
$$5/3 \cdot 1/3 - 1 = -4/9$$
$$-4/9 \cdot 1/3 = -4/27$$
$$-4/27 \cdot 1/3 + 1 = 77/81$$

Without nested:

$$2 \cdot (1/3)^4 + (1/3)^3 - (1/3)^2 + 1$$
$$2/81 + 1/27 - 1/9 + 1$$
$$2/81 + 3/81 - 9/81 + 81/81$$
$$77/81$$

0.1.2 a)

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

With nested:

$$6 \cdot (-1/2) - 2 = -5$$

$$-5 \cdot (-1/2) - 3 = -1/2$$

$$-1/2 \cdot (-1/2) + 7 = 29/4$$

b)

$$P(x) = 8x^5 - x^4 - 3x^3 + x^2 - 3x + 1 = 1 + x(-3 + x(1 + x(-3 + x(-1 + 8x))))$$

With nested:

$$8 \cdot (-1/2) - 1 = -5$$

$$-5 \cdot (-1/2) - 3 = -1/2$$

$$-1/2 \cdot (-1/2) + 1 = 5/4$$

$$5/4 \cdot (-1/2) - 3 = -29/8$$

$$-29/8 \cdot (-1/2) + 1 = 45/16$$

c)

$$P(x) = 4x^6 - 2x^4 - 2x + 4 = 4 + x(-2 + x(0 + x(0 + x(-2 + x(0 + 4x)))))$$

With nested:

$$4 \cdot (-1/2) = -2$$

$$-2 \cdot (-1/2) - 2 = -1$$

$$-1 \cdot (-1/2) = 1/2$$

$$1/2 \cdot (-1/2) = -1/4$$

$$-1/4 \cdot (-1/2) - 2 = -15/8$$

$$-15/8 \cdot (-1/2) + 4 = 79/16$$

0.1.3

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

With nested:

$$(1/2)^2 - 4 = -15/4$$

$$-15/4 \cdot (1/2)^2 + 2 = 17/16$$

$$17/16 \cdot (1/2)^2 + 1 = 81/64$$

0.1.4 a)

$$P(x) = 1 + x(1/2 + (x - 2)(1/2 + (x - 3)(-1/2)))$$

With nested:

$$-1/2 \cdot (5 - 3) + 1/2 = -1/2$$

$$-1/2 \cdot (5 - 2) + 1/2 = -1$$

$$-1 \cdot 5 + 1 = -4$$

b)

$$P(x) = 1 + x(1/2 + (x-2)(1/2 + (x-3)(-1/2)))$$

With nested:

$$-1/2 \cdot (-1-3) + 1/2 = 5/2$$

$$5/2 \cdot (-1-2) + 1/2 = -7$$

$$-7 \cdot (-1) + 1 = 8$$

0.1.5 a)

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

With nested:

$$2 \cdot (1/2 - 3) + 3 = -2$$

$$-2 \cdot (1/2 - 2) + 1 = 4$$

$$4 \cdot (1/2 - 1) + 4 = 2$$

$$2 \cdot (1/2) + 4 = 5$$

b)

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

With nested:

$$2 \cdot (-1/2 - 3) + 3 = -4$$

$$-4 \cdot (-1/2 - 2) + 1 = 11$$

$$11 \cdot (-1/2 - 1) + 4 = -25/2$$

$$-25/2 \cdot (-1/2) + 4 = 41/4$$

0.1.6 a)

$$P(x) = a_0 + a_5x^5 + a_{10}x^{10} + a_{15}x^{15} = a_0 + x^5(a_5 + x^5(a_{10} + x^5(a_{15})))$$

$a_{15}x^5 + a_{10} = b_1$ 5 multiplications and 1 addition

$b_1x^5 + a_5 = b_2$ (since x^5 is calculated) 1 multiplications and 1 addition

$b_2x^5 + a_0 = b_3$ 1 multiplications and 1 addition

$5 + 1 + 1 = 7$ multiplications, $1 + 1 + 1 = 3$ addition.

b)

$$P(x) = a_7x^7 + a_{12}x^{12} + a_{17}x^{17} + a_{22}x^{22} + a_{27}x^{27} = x^7(a_7 + x^5(a_{12} + x^5(a_{17} + x^5(a_{22} + x^5(a_{27})))))$$

$a_{27}x^5 + a_{22} = b_1$ 5 multiplications and 1 addition

$b_1x^5 + a_{17} = b_2$ (since x^5 is calculated) 1 multiplications and 1 addition

$b_2x^5 + a_{12} = b_3$ 1 multiplications and 1 addition

$b_3x^5 + a_7 = b_4$ 1 multiplications and 1 addition

$b_4x^7 = b_4$ 2 multiplications

$5 + 1 + 1 + 1 + 2 = 10$ multiplications, $1 + 1 + 1 + 1 = 4$ addition.

0.1.7 n multiplications, $2n$ addition.

(c) 0.1.1

```
format long
x = 1.00001;
p = nest(50,ones(1,51),x)
q = (x^51-1)/(x-1)
estError = abs(p-q)
```

Output:

`p=51.012752082749991`

`q=51.012752082745230`

`estError=0.000000000004761`

(c) 0.1.2

$$P(x) = 1 - x + x^2 - x^3 + \ldots + x^{98} - x^{99} = 1 - x + x^2(1-x) + \ldots + x^{98}(1-x) = \sum_{k=0}^{49} x^{2k}(1-x) = (1-x)\sum_{k=0}^{49}(x^2)^k = (1-x)\frac{1-(x^2)^{50}}{1-x} = 1 - x^{100}$$

```
format long
x = 1.00001;
p = nest(99,(-1).^(0:99),x)
q = (1-x^100)
estError = abs(p-q)
```

Output:

`p=-0.000500245079648`

`q=-0.001000495161746`

`estError=0.000500250082098`

0.2 Binary Numbers

0.2.1 a)

$64/2 = 32$ R 0

$32/2 = 16$ R 0

$16/2 = 8$ R 0

$8/2 = 4$ R 0

$4/2 = 2$ R 0

$2/2 = 1$ R 0

$1/2 = 0$ R 1

$(64)_{10} = (1000000)_2$

b)

$17/2 = 8$ R 1

$8/2 = 4$ R 0

$4/2 = 2$ R 0

$2/2 = 1$ R 0

$1/2 = 0$ R 1

$(17)_{10} = (10001)_2$

c)

$79/2 = 32 \text{ R } 1$

$39/2 = 19 \text{ R } 1$

$19/2 = 9 \text{ R } 1$

$9/2 = 4 \text{ R } 1$

$4/2 = 2 \text{ R } 0$

$2/2 = 1 \text{ R } 0$

$1/2 = 0 \text{ R } 1$

$(79)_{10} = (1001111)_2$

d)

$227/2 = 113 \text{ R } 1$

$113/2 = 56 \text{ R } 1$

$56/2 = 28 \text{ R } 0$

$28/2 = 14 \text{ R } 0$

$14/2 = 7 \text{ R } 0$

$7/2 = 3 \text{ R } 1$

$3/2 = 1 \text{ R } 1$

$1/2 = 0 \text{ R } 1$

$(227)_{10} = (11100011)_2$

0.2.2 a)

$1/8 \cdot 2 = 1/4 \text{ R } 0$

$1/4 \cdot 2 = 1/2 \text{ R } 0$

$1/2 \cdot 2 = 0 \text{ R } 1$

$(1/8)_{10} = (.001)_2$

b)

$7/8 \cdot 2 = 3/4 \text{ R } 1$

$3/4 \cdot 2 = 1/2 \text{ R } 1$

$1/2 \cdot 2 = 0 \text{ R } 1$

$(7/8)_{10} = (.111)_2$

c) It's larger than 2, factor it out.

Integer part:

$2/2 = 1 \text{ R } 0$

$1/2 = 56 \text{ R } 1$

Fractional part:

$3/16 \cdot 2 = 3/8 \text{ R } 0$

$3/8 \cdot 2 = 3/4 \text{ R } 0$

$3/4 \cdot 2 = 1/2 \text{ R } 1$

$1/2 \cdot 2 = 0 \text{ R } 1$

$(35/16)_{10} = (10.0011)_2$

d)

$$31/64 \cdot 2 = 31/32 \text{ R } 0$$

$$31/32 \cdot 2 = 15/16 \text{ R } 1$$

$$15/16 \cdot 2 = 7/8 \text{ R } 1$$

$$7/8 \cdot 2 = 3/4 \text{ R } 1$$

$$3/4 \cdot 2 = 1/2 \text{ R } 1$$

$$1/2 \cdot 2 = 0 \text{ R } 1$$

$$(31/64)_{10} = (.011111)_2$$

0.2.3 a) Solve the integer and fractional part separately.

Integer part:

$$10/2 = 5 \text{ R } 0$$

$$5/2 = 2 \text{ R } 1$$

$$2/2 = 1 \text{ R } 0$$

$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.5 \cdot 2 = 0 \text{ R } 1$$

Sum:

$$(10.5)_{10} = (1010.1)_2$$

b)

$$1/3 \cdot 2 = 2/3 \text{ R } 0$$

$$2/3 \cdot 2 = 1/3 \text{ R } 1$$

$$1/3 \cdot 2 = 2/3 \text{ R } 0$$

The period is two.

$$(1/3)_{10} = (.\overline{01})_2$$

c)

$$5/7 \cdot 2 = 3/7 \text{ R } 1$$

$$3/7 \cdot 2 = 6/7 \text{ R } 0$$

$$6/7 \cdot 2 = 5/7 \text{ R } 1$$

$$5/7 \cdot 2 = 3/7 \text{ R } 1$$

The period is three.

$$(5/7)_{10} = (.\overline{101})_2$$

d) Solve the integer and fractional part separately.

Integer part:

$$12/2 = 6 \text{ R } 0$$
$$6/2 = 3 \text{ R } 0$$
$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$

The period is four.

Sum:

$$(12.8)_{10} = (1100.\overline{1100})_2$$

e) Solve the integer and fractional part separately.

Integer part:

$$55/2 = 27 \text{ R } 1$$
$$27/2 = 13 \text{ R } 1$$
$$13/2 = 6 \text{ R } 1$$
$$6/2 = 3 \text{ R } 0$$
$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$

The period is four.

Sum:

$$(55.4)_{10} = (110111.\overline{0110})_2$$

f)

$$.1 \cdot 2 = .2 \text{ R } 0$$

$$.2 \cdot 2 = .4 \text{ R } 0$$

$$.4 \cdot 2 = .8 \text{ R } 0$$

$$.8 \cdot 2 = .6 \text{ R } 1$$

$$.6 \cdot 2 = .2 \text{ R } 1$$

$$.2 \cdot 2 = .4 \text{ R } 0$$

The period is four after first bit.

$$(0.1)_{10} = (0.0\overline{0011})_2$$

0.2.4 a) Solve the integer and fractional part separately.

Integer part:

$$11/2 = 5 \text{ R } 1$$

$$5/2 = 2 \text{ R } 1$$

$$2/2 = 1 \text{ R } 0$$

$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.25 \cdot 2 = .5 \text{ R } 0$$

$$.5 \cdot 2 = 0 \text{ R } 1$$

Sum:

$$(11.25)_{10} = (1101.01)_2$$

b)

$$2/3 \cdot 2 = 1/3 \text{ R } 1$$

$$1/3 \cdot 2 = 2/3 \text{ R } 0$$

$$2/3 \cdot 2 = 1/3 \text{ R } 1$$

The period is two.

$$(2/3)_{10} = (.\overline{10})_2$$

c)

$$3/5 = 0.6$$

$$.6 \cdot 2 = .2 \text{ R } 1$$

$$.2 \cdot 2 = .4 \text{ R } 0$$

$$.4 \cdot 2 = .8 \text{ R } 0$$

$$.8 \cdot 2 = .6 \text{ R } 1$$

$$.6 \cdot 2 = .2 \text{ R } 1$$

The period is four.

$$(3/5)_{10} = (.\overline{1001})_2$$

d) Solve the integer and fractional part separately.

Integer part:

$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$

The period is four.

Sum:

$$(3.2)_{10} = (11.\overline{0011})_2$$

e) Solve the integer and fractional part separately.

Integer part:

$$30/2 = 15 \text{ R } 0$$
$$15/2 = 7 \text{ R } 1$$
$$7/2 = 3 \text{ R } 1$$
$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$

The period is four.

Sum:

$$(30.6)_{10} = (11110.\overline{1001})_2$$

f) Solve the integer and fractional part separately.

Integer part:

$$99/2 = 49 \text{ R } 1$$
$$49/2 = 24 \text{ R } 1$$
$$24/2 = 12 \text{ R } 0$$
$$12/2 = 6 \text{ R } 0$$

$$6/2 = 3 \text{ R } 0$$
$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.9 \cdot 2 = .8 \text{ R } 1$$
$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$

The period is four after the first bit.

Sum:

$$(99.9)_{10} = (1100011.1\overline{1100})_2$$

0.2.5 Solve the integer and fractional part separately. At least 4 decimal points (3.1416) will give the correct answer.

Integer part:

$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.14159265358979 \cdot 2 = .28318530717958 \text{ R } 0$$
$$.28318530717958 \cdot 2 = .56637061435916 \text{ R } 0$$
$$.56637061435916 \cdot 2 = .13274122871832 \text{ R } 1$$
$$.13274122871832 \cdot 2 = .26548245743664 \text{ R } 0$$
$$.26548245743664 \cdot 2 = .53096491487328 \text{ R } 0$$
$$.53096491487328 \cdot 2 = .06192982974656 \text{ R } 1$$
$$.06192982974656 \cdot 2 = .12385965949312 \text{ R } 0$$
$$.12385965949312 \cdot 2 = .24771931898624 \text{ R } 0$$
$$.24771931898624 \cdot 2 = .49543863797248 \text{ R } 0$$
$$.49543863797248 \cdot 2 = .99087727594496 \text{ R } 0$$
$$.99087727594496 \cdot 2 = .98175455188992 \text{ R } 1$$
$$.98175455188992 \cdot 2 = .96350910377984 \text{ R } 1$$
$$.96350910377984 \cdot 2 = .92701820755968 \text{ R } 1$$

Sum:

$$(\pi)_{10} \approx (11.0010010000111)_2$$

0.2.6 Do it the same way as in the last exercise. At least 4 decimal points (2.7183) will give the correct answer.

$$(e)_{10} \approx (10.1011011111100)_2$$

0.2.7 a)

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

b) Solve the integer and fractional part separately.

Integer part:

$$2^3 + 2^1 + 1 = 8 + 2 + 1 = 11$$

Fractional part:

$$1/2 + 1/8 = 5/8 = .625$$

Sum:

$$(1011.101)_2 = (11.625)_{10}$$

c) Solve the integer and fractional part separately.

Integer part:

$$2^4 + 2^2 + 2^1 + 1 = 16 + 4 + 2 + 1 = 23$$

Fractional part:

$$x = (.\overline{01})_2$$

$$2^2 x = (01.\overline{01})_2$$

$$(2^2 - 1)x = (01.\overline{01})_2 - (.\overline{01})_2 = (01)_2 = 1 \Leftrightarrow x = \frac{1}{4-1} = 1/3$$

Sum:

$$(10111.\overline{01})_2 = (23 + 1/3)_{10} = (70/3)_{10}$$

d) Solve the integer and fractional part separately.

Integer part:

$$2^2 + 2^1 = 4 + 2 = 6$$

Fractional part:

$$x = (.\overline{10})_2$$

$$2^2 x = (10.\overline{10})_2$$

$$(2^2 - 1)x = (10.\overline{10})_2 - (.\overline{10})_2 = (10)_2 = 2 \Leftrightarrow x = \frac{2}{4-1} = 2/3$$

Sum:

$$(110.\overline{10})_2 = (6 + 2/3)_{10} = (20/3)_{10}$$

e) Solve the integer and fractional part separately.

Integer part:

$$2^1 = 2$$

Fractional part:

$$x = (.\overline{110})_2$$

$$2^3 x = (110.\overline{110})_2$$

$$(2^3 - 1)x = (110.\overline{110})_2 - (.\overline{110})_2 = (110)_2 = 6 \Leftrightarrow x = \frac{6}{8-1} = 6/7$$

Sum:

$$(10.\overline{110})_2 = (2 + 6/7)_{10} = (20/7)_{10}$$

f) Solve the integer and fractional part separately.

Integer part:

$$2^2 + 2^1 = 4 + 2 = 6$$

Fractional part:

$$x = (.1\overline{101})_2$$

$$y = 2x = (1.\overline{101})_2$$
$$z = (.\overline{101})_2$$
$$2^3 z = (101.\overline{101})_2$$
$$(2^3 - 1)z = (101.\overline{101})_2 - (.\overline{101})_2 = (101)_2 = 5 \Leftrightarrow z = \frac{5}{8-1} = 5/7$$
$$y = 1 + 5/7 = 12/7 \Leftrightarrow x = y/2 = 6/7$$

Sum:

$$(110.1\overline{101})_2 = (6 + 6/7)_{10} = (48/7)_{10}$$

g) Solve the integer and fractional part separately.

Integer part:

$$2^1 = 2$$

Fractional part:

$$x = (.010\overline{1101})_2$$

$$y = 2^3 x = (010.\overline{1101})_2$$
$$z = (.\overline{1101})_2$$
$$2^4 z = (1101.\overline{1101})_2$$
$$(2^4 - 1)z = (1101.\overline{1101})_2 - (.\overline{1101})_2 = (1101)_2 = 13 \Leftrightarrow z = \frac{13}{16-1} = 13/15$$
$$y = 2 + 13/15 = 43/15 \Leftrightarrow x = y/8 = 43/120$$

Sum:

$$(10.010\overline{1101})_2 = (2 + 43/120)_{10} = (283/120)_{10}$$

h) Solve the integer and fractional part separately.

Integer part:

$$2^2 + 2^1 + 1 = 4 + 2 + 1 = 7$$

Fractional part:

$$x = (.\overline{1})_2$$

$$2x = (1.\overline{1})_2$$

$$(2-1)x = (1.\overline{1})_2 - (.\overline{1})_2 = (1)_2 = 1 \Leftrightarrow x = \frac{1}{2-1} = 1$$

Sum:

$$(111.\overline{1})_2 = (7 + 1)_{10} = (8)_{10}$$

0.2.8 a)

$$2^4 + 2^3 + 2^1 + 1 = 16 + 8 + 2 + 1 = 27$$

b) Solve the integer and fractional part separately.

Integer part:

$$2^5 + 2^4 + 2^2 + 2^1 + 1 = 32 + 16 + 4 + 2 + 1 = 55$$

Fractional part:

$$1/8 = .125$$

Sum:

$$(110111.001)_2 = (55.125)_{10}$$

c) Solve the integer and fractional part separately.

Integer part:

$$2^2 + 2^1 + 1 = 4 + 2 + 1 = 7$$

Fractional part:

$$x = (.\overline{001})_2$$

$$2^3 x = (001.\overline{001})_2$$

$$(2^3 - 1)x = (001.\overline{001})_2 - (.\overline{001})_2 = (001)_2 = 1 \Leftrightarrow x = \frac{1}{8-1} = 1/7$$

Sum:

$$(111.\overline{001})_2 = (7 + 1/7)_{10} = (50/7)_{10}$$

d) Solve the integer and fractional part separately.

Integer part:

$$2^3 + 2^1 = 8 + 2 = 10$$

Fractional part:

$$x = (.\overline{01})_2$$

$$2^2 x = (01.\overline{01})_2$$

$$(2^2 - 1)x = (01.\overline{01})_2 - (.\overline{01})_2 = (01)_2 = 1 \Leftrightarrow x = \frac{1}{4-1} = 1/3$$

Sum:

$$(1010.\overline{01})_2 = (10 + 1/3)_{10} = (31/3)_{10}$$

e) Solve the integer and fractional part separately.

Integer part:

$$2^4 + 2^2 + 2^1 + 1 = 16 + 4 + 2 + 1 = 23$$

Fractional part:

$$x = (.1\overline{0101})_2$$

$y = 2x = (1.\overline{0101})_2$

$z = (.\overline{0101})_2$

$2^4 z = (0101.\overline{0101})_2$

$(2^4 - 1)z = (0101.\overline{0101})_2 - (.\overline{0101})_2 = (0101)_2 = 5 \Leftrightarrow z = \dfrac{5}{16-1} = 1/3$

$y = 1 + 1/3 = 4/3 \Leftrightarrow x = y/2 = 2/3$

Sum:

$$(10111.1\overline{0101})_2 = (23 + 2/3)_{10} = (71/3)_{10}$$

f) Solve the integer and fractional part separately.

Integer part:

$$2^3 + 2^2 + 2^1 + 1 = 8 + 4 + 2 + 1 = 15$$

Fractional part:

$x = (.010\overline{001})_2$

$y = 2^3 x = (010.\overline{001})_2$

$z = (.\overline{001})_2$

$2^3 z = (001.\overline{001})_2$

$(2^3 - 1)z = (001.\overline{001})_2 - (.\overline{001})_2 = (001)_2 = 1 \Leftrightarrow z = \dfrac{1}{8-1} = 1/7$

$y = 2 + 1/7 = 15/7 \Leftrightarrow x = y/8 = 15/56$

Sum:

$$(1111.010\overline{001})_2 = (15 + 15/56)_{10} = (855/56)_{10}$$

0.3 Floating Point Representation of Real Number

0.3.1 a) Covert decimal to binary.

$1/4 \cdot 2 = 1/2 \text{ R } 0$

$1/2 \cdot 2 = 0 \text{ R } 1$

$(1/4)_{10} = (.01)_2$

Left-justify it by shifting it twice.

$$(1/4)_{10} = 1.000\ldots000 \times 2^{-2}$$

b) Covert decimal to binary.

$1/3 \cdot 2 = 2/3 \text{ R } 0$

$2/3 \cdot 2 = 1/3 \text{ R } 1$

$1/3 \cdot 2 = 2/3 \text{ R } 0$

$(1/3)_{10} = (.\overline{01})_2$

Left-justify it by shifting it twice. Since the 53 bit will be a zero, round down (do nothing).

$$(1/3)_{10} = 1.0101\ldots01 \times 2^{-2}$$

c) Covert decimal to binary.

$$2/3 \cdot 2 = 1/3 \text{ R } 1$$
$$1/3 \cdot 2 = 2/3 \text{ R } 0$$
$$2/3 \cdot 2 = 1/3 \text{ R } 1$$
$$(2/3)_{10} = (.\overline{10})_2$$

Left-justify it by shifting it once. Since the 53 bit will be a zero, round down (do nothing).

$$(1/3)_{10} = 1.0101\ldots01 \times 2^{-1}$$

d) Covert decimal to binary.

$$0.9 \cdot 2 = 0.8 \text{ R } 1$$
$$0.8 \cdot 2 = 0.6 \text{ R } 1$$
$$0.6 \cdot 2 = 0.2 \text{ R } 1$$
$$0.2 \cdot 2 = 0.4 \text{ R } 0$$
$$0.4 \cdot 2 = 0.8 \text{ R } 0$$
$$0.8 \cdot 2 = 0.6 \text{ R } 1$$
$$(0.9)_{10} = (.1\overline{1100})_2$$

Left-justify it by shifting it once. Since the 53 bit will be a one and has following non-zero bits, round up.

$$(1/3)_{10} = 1.1100\ldots1101 \times 2^{-1}$$

0.3.2 a) Covert decimal to binary. Solve the integer and fractional part separately.

Integer part:

$$9/2 = 4 \text{ R } 1$$
$$4/2 = 2 \text{ R } 0$$
$$2/2 = 1 \text{ R } 0$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.5 \cdot 2 = 0 \text{ R } 1$$

Sum:

$$(9.5)_{10} = (1001.1)_2$$

Left-justify it by shifting it three times then pad with zeros.

$$(9.5)_{10} = 1.00110\ldots00 \times 2^3$$

b) Covert decimal to binary. Solve the integer and fractional part separately.

Integer part:

$$9/2 = 4 \text{ R } 1$$
$$4/2 = 2 \text{ R } 0$$
$$2/2 = 1 \text{ R } 0$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$

Sum:

$$(9.6)_{10} = (1001.\overline{1001})_2$$

Left-justify it by shifting it three times. Since the 53 bit will be a zero, round down (do nothing).

$$(9.6)_{10} = 1.0011\ldots0011 \times 2^3$$

c) Covert decimal to binary. Solve the integer and fractional part separately.

Integer part:

$$100/2 = 50 \text{ R } 0$$
$$50/2 = 25 \text{ R } 0$$
$$25/2 = 12 \text{ R } 1$$
$$12/2 = 6 \text{ R } 0$$
$$6/2 = 3 \text{ R } 0$$
$$3/2 = 1 \text{ R } 1$$
$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$.2 \cdot 2 = .4 \text{ R } 0$$
$$.4 \cdot 2 = .8 \text{ R } 0$$
$$.8 \cdot 2 = .6 \text{ R } 1$$
$$.6 \cdot 2 = .2 \text{ R } 1$$
$$.2 \cdot 2 = .4 \text{ R } 0$$

Sum:

$$(100.2)_{10} = (1100100.\overline{0011})_2$$

Left-justify it by shifting it 6 times. Since the 53 bit will be a one and has following non-zero bits, round up.

$$(100.2)_{10} = 1.1001000011001100\ldots11001101 \times 2^6$$

d) Covert decimal to binary. Solve the integer and fractional part separately.

Integer part:

$$6/2 = 3 \text{ R } 0$$

$$3/2 = 1 \text{ R } 1$$

$$1/2 = 0 \text{ R } 1$$

Fractional part:

$$2/7 \cdot 2 = 4/7 \text{ R } 0$$

$$4/7 \cdot 2 = 1/7 \text{ R } 1$$

$$1/7 \cdot 2 = 2/7 \text{ R } 0$$

$$2/7 \cdot 2 = 4/7 \text{ R } 0$$

Sum:

$$(44/7)_{10} = (110.\overline{010})_2$$

Left-justify it by shifting it twice. Since the 53 bit will be a zero, round down (do nothing).

$$(44/7)_{10} = 1.100100100\ldots001001 \times 2^2$$

0.3.3 Since $(5)_{10} = (101)_2$ and $(2^{-k})_{10} = (0.00\ldots001)_2$ where the number of zeros is equal to k, the sum will be $101.\underbrace{00\ldots00}_{k-1 \text{ zeros}}1$. In the IEEE format the right-most 1 will be at the $(k+2)$th bit. For the number to be represented exactly in double precision $k+2 \le 52 \Leftrightarrow k \le 50$. Since k is a positive integer, $1 \le k \le 50$.

0.3.4 Since $(19)_{10} = (10011)_2$ and $(2^{-k})_{10} = (0.00\ldots001)_2$ where the number of zeros is equal to k the sum will be $10011.\underbrace{00\ldots00}_{k-1 \text{ zeros}}1$. In the IEEE format the right-most 1 will be at the $(k+4)$th bit. If the 1 is at a position further away than the 52 bit it will be rounded down. This means if $k+4 > 52 \Leftrightarrow k > 48$ then $\text{fl}(19 + 2^{-k}) = \text{fl}(19)$. The largest possible value of k therefore is 48.

0.3.5 a)

$$\begin{aligned}
&(1 + (2^{-51} + 2^{-53})) - 1 = \\
=&(1 + (1.\boxed{0\ldots0} \cdot 2^{-51} + 1.\boxed{0\ldots0} \cdot 2^{-53})) - 1 = \\
=&(1.\boxed{0\ldots0} \cdot 2^{0} + 1.\boxed{010\ldots0} \cdot 2^{-51}) - 1 = \\
=&1.\boxed{0\ldots010} * 2^{0} - 1.\boxed{0\ldots0} \cdot 2^{0} = \\
=&1.\boxed{0\ldots0} * 2^{-51}
\end{aligned}$$

To test in Matlab (0 and 1 is representing `false` and `true` respectively):

```
>> x1 = 2^(-51);
>> x2 = x1 + 2^(-53);
>> x3 = x2 + 1;
>> x4 = x3 - 1;
>> disp(x2 == x4);
   0
>> disp(x1 == x4);
   1
```

b)

$$
\begin{aligned}
&(1+(2^{-51}+2^{-52}+2^{-53}))-1=\\
=&(1+(1.\boxed{0\ldots0}\cdot2^{-51}+1.\boxed{0\ldots0}\cdot2^{-52}+1.\boxed{0\ldots0}\cdot2^{-53}))-1=\\
=&(1.\boxed{0\ldots0}\cdot2^{0}+1.\boxed{110\ldots0}\cdot2^{-51})-1=\\
=&1.\boxed{0\ldots0100}*2^{0}-1.\boxed{0\ldots0}\cdot2^{0}=\\
=&1.\boxed{0\ldots0}*2^{-50}
\end{aligned}
$$

The tricky part is the rounding. $1.\boxed{0\ldots0}\cdot2^0+1.\boxed{110\ldots0}\cdot2^{-51}=1.\boxed{0\ldots011}10\ldots\cdot2^0$. Since the first following bit is a 1 and the rest are 0's the special rule is applied, round the 52nd bit to zero. In this case round up which equals $1.\boxed{0\ldots0100}\cdot2^0$.

To test in Matlab (0 and 1 is representing `false` and `true` respectively):

```
>> x1 = 2^(-51);
>> x2 = x1 + 2^(-52);
>> x3 = x2 + 2^(-53);
>> x4 = x3 + 1;
>> x5 = x4 - 1;
>> disp(x3 == x5);
   0
>> disp(x5 == 2^(-50));
   1
```

0.3.6 a)

$$
\begin{aligned}
&(1+(2^{-51}+2^{-52}+2^{-54}))-1=\\
=&(1+(1.\boxed{0\ldots0}\cdot2^{-51}+1.\boxed{0\ldots0}\cdot2^{-52}+1.\boxed{0\ldots0}\cdot2^{-54}))-1=\\
=&(1.\boxed{0\ldots0}\cdot2^{0}+1.\boxed{1010\ldots0}\cdot2^{-51})-1=\\
=&1.\boxed{0\ldots011}*2^{0}-1.\boxed{0\ldots0}\cdot2^{0}=\\
=&1.\boxed{10\ldots0}*2^{-51}
\end{aligned}
$$

b)

$$
\begin{aligned}
&(1+(2^{-51}+2^{-52}+2^{-60}))-1=\\
=&(1+(1.\boxed{0\ldots0}\cdot2^{-51}+1.\boxed{0\ldots0}\cdot2^{-52}+1.\boxed{0\ldots0}\cdot2^{-60}))-1=\\
=&(1.\boxed{0\ldots0}\cdot2^{0}+1.\boxed{1000000010\ldots0}\cdot2^{-51})-1=\\
=&1.\boxed{0\ldots011}*2^{0}-1.\boxed{0\ldots0}\cdot2^{0}=\\
=&1.\boxed{10\ldots0}*2^{-51}
\end{aligned}
$$