Govher Sapardurdyyeva

4. Determine the single-precision and double-precision machine representation of the following decimal numbers:

a. 1.0, −1.0 **b.** +0.0, −0.0 **c.** −9876.54321

[a]**d.** 0.234375 [a]**e.** 492.78125 **f.** 64.37109375

g. −285.75 **h.** 10^{-2}

Single precision → 23 bits

double precision → 53 bits

Bias rep 8 bits

$2^{n-1} - 1 = 127$

(a) $1.0 \cdot 2^0 = 1.0$

Sign:

positive - 0

negative - 1

Exponent $= 0 + 127 = 127_{10}$

$127/2 = 63\ (1)$

$63/2 = 31\ (1)$

$31/2 = 15\ (1)$

$15/2 = 7\ (1)$

$7/2 = 3\ (1)$

$3/2 = 1\ (1)$

$1/2 = 0\ (1)$

$127_{10} = 01111111_2$

Single precision machine representation of 1.0 and -1.0

Sign	Exponent	Mantissa
0	01111111	00000000000000000000000
1	01111111	00000000000000000000000

Double precision machine representation of 1.0 and -1.0

Sign	Exponent	Mantissa
0	01111111	0000000000000000000000000000000000000000000000000000
1	01111111	0000000000000000000000000000000000000000000000000000

Convert to base 16

0011 1111 1000 0000 0000 0000 0000 0000

3 F 8 0 0 0 0 0

1011 1111 1000 0000 0000 0000 0000 0000

B F 8 0 0 0 0 0

Hex representation:

1.0 = 3F800000

-1.0 = BF800000

(b) $+0.0$ -0.0

Decimal	Sign	Exponent	Mantissa (single precision)	Mantissa (Double precision)
0.0	0	00000000	00000000000000000000000	0000000000000000000000000000000000000000000000000000
-0.0	1	00000000	00000000000000000000000	0000000000000000000000000000000000000000000000000000

Hex value:
0.0 = 00000000
-0.0 = 80000000

c) -9876.54321

$9876_{10} = 0010\ 0110\ 1001\ 0100_2$

$0.54321_{10} = 0.10010110000011111101_2$

$9876.54321 \approx 1.\underbrace{00110100100100101000}_{\text{mantissa}} \times 2^{13}$

Exponent $= 127 + 13 = 140_{10} = 1000\ 1100_2$

Single precision:
1100 0110 0001 1010 0101 0010 0010 1100
Hex value:
C61A522C

Double precision:
1023 + 13 = 1036
Binary = 10000001100
1100 0000 1100 0011 0100 1010 0100 0110 1110 0001 0101 1000 0000 0000 0000 0000
Hex Value:
C0C34A46E1580000

a) $64.37109375 = 1.\underbrace{00000001011111000000000}_{\text{mantissa}} \cdot 2^{6}$

$127 + 6 = 133_{10} = 10000101$ $1023 + 6 = 1029$

Single : 0100 0010 1000 0000 1011 1110 0000 0000
Hex: 4280BE00
Double: 0100 0000 0101 0000 0001 0111 1100 0000 0000 0000 0000 0000 0000 0000 0000 0000
Hex: 405017C000000000

16. Consider a computer that operates in base β and carries n digits in the mantissa of its floating-point number system. Show that the rounding of a real number x to the nearest machine number $\tilde{x}$ involves a relative error of at most $\frac{1}{2}\beta^{1-n}$.
Hint: Imitate the argument in the text.

Assume x is the floating point approximation to a number $\hat{x}$, then according the chapter 1.1, the absolute rounding error would be:

$$|\tilde{x} - x|$$

And the relative error would be: $\dfrac{|\tilde{x} - x|}{|x|}$

The floating point representation of a base β number with n digits in mantissa would be: $x = \pm m \beta^e$

In this case, m is mantissa such that $1 \leq m \leq \beta$
And e is the exponent.

Considering that the difference between floating point numbers is β^{1-n} times the exponent factor of LSB of the mantissa, the absolute rounding error is:

$$|\tilde{x} - x| = \frac{1}{2}\beta^{1-n}|\hat{x}|$$

Now, the relative error is: $\dfrac{|\tilde{x} - x|}{|x|} \leq \dfrac{\frac{1}{2}\beta^{1-n}}{|\bar{x}|}|x| \Rightarrow \dfrac{|\tilde{x} - x|}{|x|} \leq \dfrac{1}{2}\beta^{1-n}$

Since:

$$|\hat{x}| \leq |x| \Rightarrow \frac{|\bar{x} - x|}{|x|} \leq \frac{|\bar{x} - x|}{|\bar{x}|} = \frac{\frac{1}{2}\beta^{1-n}|\bar{x}|}{|\bar{x}|} = \frac{1}{2}\beta^{1-n}$$

We see that the rounding of a real number x to the nearest machine number x has a relative error at most $\frac{1}{2}\beta^{1-n}$