SOLUTIONS

Math 614 Numerical Linear Algebra (Bueler) 18 October 2021

Worksheet: If IEEE 754 had a 12-bit standard ...

A floating point system $\mathbb{F}$ described in Lecture 13 of the textbook (L. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM Press 1997) is, in reality, implemented in bits. The actual IEEE 754 standards for 32-bit single precision and 64-bit double precision representations are cumbersome to play with, so for convenience we pretend here that the standard has a 12-bit version. It might look like this:

s	e_1	e_2	e_3	e_4	b_1	b_2	b_3	b_4	b_5	b_6	b_7

These 12 bits are organized to represent a *nonzero* number:

$$x = (-1)^s \, (1.b_1b_2b_3b_4b_5b_6b_7)_2 \; 2^{(e_1e_2e_3e_4)_2-(0111)_2}$$

Note that $(1.b_1b_2b_3b_4b_5b_6b_7)_2$ is called the *mantissa*. The power on the 2 is the *exponent*. The special offset $(0111)_2$, equal to 7 in base ten, is called the *exponent bias*. We also define some exceptional cases:

- exponent bits $(0000)_2$ define the number zero or subnormal numbers
- exponent bits $(1111)_2$ define the other exceptions: $\pm\infty$ and NaN

We will say nothing further about the $(1111)_2$ exceptions.

(a) What is the largest real number that this system can represent? Show the bits.

0	1	1	1	0	1	1	1	1	1	1	1

s e b

$$x = +(1.1111111)_2 \times 2^{14-7} = \left(2-\frac{1}{2^7}\right)\times 2^7 = 2^8-1 = 255_{10}$$

(b) Not considering subnormal numbers, what is the smallest positive number that this system can represent? (*The first normal number to the right of zero.*) Show the bits.

0	0	0	0	1	0	0	0	0	0	0	0

$$x = +(1.0000000)_2 \times 2^{1-7} = 2^{-6} = 0.015625$$

(c) If we define $\epsilon_{\text{machine}}$ as the gap between 1 and the next representable number greater than 1, what is the value of $\epsilon_{\text{machine}}$ in this system?

$$(1+2^{-7}) - 1 = (1.0000001)_2 - (1.0000000)_2$$

$$= 2^{-7} = \varepsilon_{\text{machine}}$$

(d) What is the representation of zero? Show the bits.

0	0	0	0	0	0	0	0	0	0	0	0

Note: One goal of these standards is that "x==0" has same meaning whether x is integer or floating point.

(e) What is the representation of 4? Show the bits.

0	1	0	0	1	0	0	0	0	0	0	0

$$4 = +(1.0000000)_2 \times 2^2 = (1.0000000)_2 \times 2^{9-7}$$

$$9 = (1001)_2$$

(f) What is the largest representable number which is smaller than 8? Show the bits.

0	1	0	0	1	1	1	1	1	1	1	1

$$x = +(1.1111111)_2 \times 2^2 = (1.1111111)_2 \times 2^{(1001)_2 - 7}$$

$$= \left(2 - \frac{1}{2^7}\right) 2^2 = \left(8 - \frac{1}{32}\right)_{10}$$

(g) In the interval $[4, 8)$, how many numbers can be represented?

From (e) and (f), these numbers have same s and e bits. There are 2^7 possibilities for the b bits.

should be "normal"

(h) Exactly how many distinct ~~non-exceptional~~ numbers can be represented in this system? (*Include the number zero but exclude subnormal numbers and any exceptions using exponent* $(1111)_2$, *i.e.* $\pm\infty$ *and NaN.*)

zero s e b

$$1 + 2 \times (2^4 - 2) \times 2^7 = 1 + 14 \times 2^8 = 3585$$

(i) Show the bits of one subnormal number.

0	0	0	0	0	0	1	0	1	0	1	0

a curiosity only...

(This is subnormal because e bits are $(0000)_2$.

It represents: $+(0.0101010)_2 \times 2^{-6} = \left(\frac{1}{4} + \frac{1}{16} + \frac{1}{64}\right) \times 2^{-6}$ // 0.005127