Numerical Analysis: Report Question[1]

Xuefeng LIU

October 12, 2016

I am sorry that I cannot teach you directly this time because I have a conference to attend. Please study the pages introduced in this document and finish the report problem. If you have any question, please come to my room next week. I will give some remark in next lecture on Oct. 20.

1 Floating point number

At the beginning of this course, we would like to study one of the fundamental things about scientific computing – the representation of number in computer.

In early days of computer, there is no standard rule for the representation of numbers and the same code will give different results on different computers. To solve this problem, in 1984, the IEEEE-754 standard is published to define how to uniquely represent a number in computer and how to perform the basic number operation like $+, -, /, *$. In these days, almost all the computers are following this standard to do the computing.

Please study the following article and video by yourself and finish the report problem.

1) *IEEE Standard 754 Floating Point Numbers* by Steve Hollasch. `http://steve.hollasch.net/cgindex/coding/ieeefloat.html`

2) *IEEE 754: Intro to the Floating Point Format* by WhatsaCreel, Video on Youtube, 13 minutes, `https://www.youtube.com/watch?v=owtK58XiPGo`

Other references:

1) IEEE floating point. (2016, October 6). In Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/wiki/IEEE_floating_point`

2 Question 1:

Review the definition of single format in IEEE-754 standards:

1	8	23
s (sign)	e (exponent)	f (significand)

Fig. 1

A 32-bit single format number X is divided as shown in Fig 1. The value v of X is inferred from its constituent fields thus

(1) If $e = 255$ and $f \neq 0$, then v is NaN regardless of s

(2) If $e = 255$ and $f = 0$, then $v = (-1)^s \infty$

(3) If $0 < e < 255$, then $v = (-1)^s 2^{e-127}(1 \cdot f)$

(4) If $e = 0$ and $f \neq 0$, then $v = (-1)^s 2^{-126}(0 \cdot f)$ (denormalized numbers)

(5) If $e = 0$ and $f = 0$, then $v = (-1)^s 0$ (zero)

Let's design a new format for floating point number with 6 bits:

1 bit	3 bits	2 bits
sign bit	biased exponent bits: + (3)	significand bits

The range of un-biased exponent number: $E_{min} = -2, E_{max} = 3$. Then

1) Display the binary representation of number 1, 0, -1.25.

2) Display all the binary representation of number inf, NaN.

3) List all possible numbers and there values in decimal mode. If it is denormalized numbers, mark it out. For example "$000010 = 0.125$ (denormalized number)".

4) Draw all the numbers except inf and NaN on a line.

5) What is the minimal gap between two neighbour numbers? How about the minimal gap for 32 bits single precision numbers?