Numerical Analysis Homework #6

due 2020 MAY 19, 9:50 a.m.

1 Assignments

Caution:

- To get full credit, *you must write down sufficient intermediate steps*, only giving the final answer earns you no credit!
- Please make sure that your handwriting is recognizable, otherwise you only get partial credit for the recognizable part.

I. Convert the decimal integer 477 to a normalized FPN with $\beta = 2$.

II. Convert the decimal fraction 3/5 to a normalized FPN with $\beta = 2$.

III. Let $x = \beta^e$, $e \in \mathbb{Z}$, $L < e < U$ be a normalized FPN in $\mathbb{F}$ and $x_L, x_R \in \mathbb{F}$ the two normalized FPNs adjacent to x such that $x_L < x < x_R$. Prove $x_R - x = \beta(x - x_L)$.

IV. By reusing your result of II, find out the two normalized FPNs adjacent to $x = 3/5$ under the IEEE 754 single-precision protocol. What is $\mathrm{fl}(x)$ and the relative roundoff error?

V. If the IEEE 754 single-precision protocol did not round off numbers to the nearest, but simply dropped excess bits, what would the unit roundoff be?

VI. How many bits of precision are lost in the subtraction $1 - \cos x$ when $x = \frac{1}{4}$?

VII. Suggest at least two ways to compute $1 - \cos x$ to avoid catastrophic cancellation caused by subtraction.

The above eight questions weigh 3, 4, 7, 6, 3, 3, 4 points, respectively, totaling 30 points.

2 `C++` programming

(A) (10 points) By programming in `C++`, print values of the functions in (1) at 101 equally spaced points covering the interval $[0.99, 1.01]$. Calculate each function in a straightforward way without rearranging or factoring. Note that the three functions are theoretically the same, but the computed values might be very different. Plot these functions near 1.0 using a magnified scale for the function values to see the variations involved. Discuss what you see. Which one is the most accurate? Why?

(B) (10 points) Consider a normalized FPN system $\mathbb{F}$ with the characterization $\beta = 2, p = 3, L = -1, U = +1$. Answer the following by *programming* in `C++`

- compute UFL($\mathbb{F}$) and OFL($\mathbb{F}$) and output them as decimal numbers;
- enumerate all numbers in $\mathbb{F}$ and verify the corollary on the cardinality of $\mathbb{F}$ in the summary handout;
- plot $\mathbb{F}$ on the real axis;
- enumerate all the subnormal numbers of $\mathbb{F}$;
- plot the *extended* $\mathbb{F}$ on the real axis.

Thus the total point of this homework is 50.

3 Extra credits

Additional 10% credits will be given to you if you typeset your solutions in LaTeX. You are welcome to use the LaTeX template available on my webpage. You can also get partial extra credit for typesetting solutions of *some* problems.

Note: If you choose to typeset your solutions in LaTeX, you still need to turn in a hard copy in class. In addition, please upload your latex source (.tex), supporting files, and `C++` program in a single zip file (**format**: `YourName_Homework6.zip`) to the course email `NumApproximation@163.com`.

$$f(x) = x^8 - 8x^7 + 28x^6 - 56x^5 + 70x^4 - 56x^3 + 28x^2 - 8x + 1 \tag{1a}$$

$$g(x) = (((((((x-8)x+28)x-56)x+70)x-56)x+28)x-8)x+1 \tag{1b}$$

$$h(x) = (x-1)^8 \tag{1c}$$