Updated on March 5, 2022

Exercises: Preliminaries

Problems marked with ✏ are to be done by hand; those marked with 💻 are to be solved using a computer.

1. (Floating-point number set; **FNC** 1.1.1) Let $\mathbb{F}$ be the set of all floating-point numbers of the form
$$\pm(1 + F) \times 2^E,$$
where E is the exponent and $1 + F$ is the mantissa, in which
$$F = \sum_{i=1}^{d} \frac{b_i}{2^i}, \qquad b_i \in \{0, 1\}.$$
Suppose $d = 4$.
   (a) How many elements of $\mathbb{F}$ are there in the interval $[1/2, 4]$, including the endpoints?
   (b) What is the elements of $\mathbb{F}$ closest to the real number $1/10$?
   (c) What is the smallest positive integer not in $\mathbb{F}$?

2. (Condition numbers; **FNC** 1.2.3) ✏ Calculate the (relative) condition number of each function, and identify all values of x at which $\kappa_f(x) \to \infty$ (including limits as $x \to \pm\infty$).
   (a) $f(x) = \tanh(x)$.
   (b) $f(x) = \dfrac{e^x - 1}{x}$.
   (c) $f(x) = \dfrac{1 - \cos(x)}{x}$.

3. (Catastrophic cancellation; **FNC** 1.3.4)
   (a) ✏ Find the (relative) condition number for $f(x) = (1 - \cos x)/\sin x$.
   (b) ✏ Explain carefully how many digits will be lost to cancellation when computing f directly by the formula in (a) for $x = 10^{-6}$.
   (c) ✏ Show that the mathematically identical formula
$$f(x) = \frac{2\sin^2(x/2)}{\sin(x)}$$
contains no poorly conditioned steps for $|x| < 1$.
   (d) 💻 Using MATLAB, compute and compare the formulas from (a) and (c) numerically at $x = 10^{-6}$.

4. (More catastrophic cancellation; **FNC** 1.3.5) Let $f(x) = (e^x - 1)/x$.

(a) Find the condition number $\kappa_f(x)$. What is the maximum of $\kappa_f(x)$ over $[-1, 1]$?

(b) Use the "obvious" algorithm

```
y = (exp(x)-1) / x;
```

to compute $f(x)$ at 1000 evenly spaced points in the interval $[-1, 1]$.

(c) Use the first 18 terms of the Taylor series

$$f(x) = 1 + \frac{1}{2!}x + \frac{1}{3!}x^2 + \frac{1}{4!}x^3 + \cdots$$

to create a second algorithm, and evaluate it at the same set of points.

(d) Plot the relative difference between the two algorithms as a function of x. Which one do you believe is more accurate, and why?