

Funamentals of Numerical Computation

—

Tobin A. Driscoll & Richard J. Braun

Graham Strickland

March 13, 2025

1 Introduction

1.1 Floating-point numbers

1. (a) For \mathbb{F} given by the set containing zero and all numbers of the form

$$\pm(1+f) \times 2^n,$$

where $n \in \mathbb{N}$ and

$$f = \sum_{i=1}^4 b_i 2^{-i}, \quad b_i \in \{0, 1\},$$

we have

$$[1/2, 4] \cap \mathbb{F} = \left\{ \frac{1}{2}, \frac{17}{32}, \frac{9}{16}, \dots, 1, 1 + \frac{1}{16}, 1 + \frac{1}{8}, \dots, 2, 2 + \frac{1}{8}, 2 + \frac{1}{4}, \dots, 4 \right\}.$$

- (b) For \mathbb{F} as above, we have the smallest n s.t.

$$\frac{1}{10} \in [2^n, 2^{n+1})$$

given by $n = -4$, so that

$$[2^n, 2^{n+1}) = \left[\frac{1}{16}, \frac{1}{8} \right).$$

Then, we have

$$[2^n, 2^{n+1}) \cap \mathbb{F} = \left\{ \frac{1}{16}, \frac{17}{256}, \frac{5}{64}, \dots, \frac{1}{8} \right\}.$$

From this interval, we see that

$$\frac{25}{256} < \frac{1}{10} < \frac{13}{128}$$

Then, since

$$\left| \frac{1}{10} - \frac{25}{256} \right| \approx 0.002343750000000056$$

and

$$\left| \frac{13}{128} - \frac{1}{10} \right| \approx 0.001562499999999944,$$

clearly $13/128$ is the closest member of \mathbb{F} to the real number $1/10$.

(c) Suppose we have

$$\begin{aligned} & |(1+f) \times 2^{n+1} - (1+f) \times 2^n| > 1 \\ \Leftrightarrow & \left(1 + \sum_{i=1}^4 b_i 2^{-i}\right) \times 2^{n+1} - \left(1 + \sum_{i=1}^4 b_i 2^{-i}\right) \times 2^n > 1 \\ \Leftrightarrow & \sum_{i=1}^4 b_i 2^{-i} \times 2^{n+1} - \sum_{i=1}^4 b_i 2^{-i} \times 2^n > 1 \\ \Leftrightarrow & 2^{n+1} \sum_{i=1}^4 (b_i 2^{-i} - b_i 2^{-(i+1)}) > 1 \\ \Leftrightarrow & \sum_{i=1}^4 (b_i 2^{-i} - b_i 2^{-(i+1)}) > \frac{1}{2^{n+1}} \\ \Leftrightarrow & \sum_{i=1}^4 2^{-i} \left(b_i - \frac{b_i}{2}\right) > \frac{1}{2^{n+1}} \end{aligned}$$