Fundamentals 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

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

we have

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

(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

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

From this interval, we see that

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

Then, since

$$\left|\frac{1}{10} - \frac{25}{256}\right| \approx 0.0023437500000000056$$

and

$$\left|\frac{13}{128} - \frac{1}{10}\right| \approx 0.0015624999999999944,$$

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

(c) Suppose we have

$$\begin{aligned}
&|(1+f)\times 2^{n+1} - (1+f)\times 2^n| > 1 \\
\Leftrightarrow \left(1+\sum_{i=1}^{4} b_i 2^{-i}\right)\times 2^{n+1} - \left(1+\sum_{j=1}^{4} b_j 2^{-j}\right)\times 2^n > 1 \\
\Leftrightarrow \sum_{i=1}^{4} b_i 2^{-i}\times 2^{n+1} - \sum_{j=1}^{4} b_j 2^{-j}\times 2^n > 1 \\
\Leftrightarrow 2^n \sum_{i,j=1}^{4}\left(b_i 2^{1-i} - b_j 2^{-j}\right) > 1 \\
\Leftrightarrow 2^n > \frac{1}{\sum_{i,j=1}^{4}\left(b_i 2^{1-i} - b_j 2^{-j}\right)} \\
\Leftrightarrow n > \log_2\left[\frac{1}{\sum_{i,j=1}^{4}\left(b_i 2^{1-i} - b_j 2^{-j}\right)}\right]
\end{aligned}$$