Fundamentals of Numerical Computation

—

Tobin A. Driscoll & Richard J. Braun

Graham Strickland

March 22, 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) From (a), we see that

$$n \in \{1,2,3,4\} \Rightarrow n \in \mathbb{F},$$

since $3 = 2 + \frac{8}{8}$. But then we have

$$[4,8] \cap \mathbb{F} = \left\{4, 4+\frac{1}{4}, 4+\frac{1}{2}, \cdots 5, 5+\frac{1}{2}, 6, 7, 8\right\}$$

and

$$[8,16] \cap \mathbb{F} = \{8, 10, 12, 14, 16\},$$

so that the the smallest positive integer not in $\mathbb{F}$ is 9.

2. *Proof.* ($\Rightarrow$) First, suppose we have

$$\frac{|\mathrm{fl}(x)-x|}{|x|} \leq \frac{2^{n-d-1}}{2^n} \leq \frac{1}{2}\epsilon_{\mathrm{mach}},$$

so that

$$\begin{aligned}
|\mathrm{fl}(x)-x| &\leq \frac{|x|}{2}\epsilon_{\mathrm{mach}} \\
\Leftrightarrow -\frac{|x|}{2}\epsilon_{\mathrm{mach}} \leq \mathrm{fl}(x)-x &\leq \frac{|x|}{2}\epsilon_{\mathrm{mach}} \\
\Leftrightarrow x-\frac{|x|}{2}\epsilon_{\mathrm{mach}} \leq \mathrm{fl}(x) &\leq x+\frac{|x|}{2}\epsilon_{\mathrm{mach}},
\end{aligned}$$

so that clearly

$$x \leq \mathrm{fl}(x) \leq x\left(1+\frac{1}{2}\epsilon_{\mathrm{mach}}\right)$$

or $\mathrm{fl}(x) = x(1+\epsilon)$ for

$$0 \leq |\epsilon| \leq \frac{1}{2}\epsilon_{\mathrm{mach}}.$$

($\Leftarrow$) Now, suppose that

$$\mathrm{fl}(x) = x(1+\epsilon) \quad \text{for some} \quad |\epsilon| \leq \frac{1}{2}\epsilon_{\mathrm{mach}},$$

so that

$$\begin{aligned} \mathrm{fl}(x) - x &= \epsilon x \\ \Leftrightarrow |\mathrm{fl}(x) - x| &= |\epsilon x| \\ \Leftrightarrow \frac{|\mathrm{fl}(x) - x|}{|x|} &= |\epsilon| \\ &\leq \frac{1}{2}\epsilon_{\mathrm{mach}}. \end{aligned}$$

□

3. (a) We have the absolute accuracy given by

$$\left|\frac{355}{113} - \pi\right| \approx 2.667641894049666 \times 10^{-7},$$

where we have used the estimate

```
julia> abs(355/113 - Float64(π))
2.667641894049666e-7
```

and the relative accuracy is given by

$$\left|\frac{\frac{355}{113} - \pi}{\pi}\right| \approx 8.49136787674061 \times 10^{-8},$$

where we have used the estimate

```
julia> abs(355/113 - Float64(π)) / abs(Float64(π))
8.49136787674061e-8
```

(b) We have the absolute accuracy given by

$$\left|\frac{103638}{32989} - \pi\right| \approx 1.493639878447084 \times 10^{-9},$$

where we have used the estimate

```
julia> abs(103638/32989 - Float64(π))
1.493639878447084e-9
```

and the relative accuracy is given by

$$\left|\frac{\frac{103638}{32989} - \pi}{\pi}\right| \approx 4.754403397080622e \times 10^{-10},$$

where we have used the estimate

```
julia> abs(103638/32989 - Float64(π)) / abs(Float64(π))
4.754403397080622e-10
```