

## Answers to Sept. 18 Exercises

- (a)  $5.77216 \times 10^{-1}$ ; (b)  $5.77215 \times 10^{-1}$ ; (c) same as (a); (d) same as (b).
- (a) absolute error:  $3.35 \times 10^{-7}$ ; relative error:  $5.81 \times 10^{-7}$ . (b) absolute error:  $6.65 \times 10^{-7}$ ; relative error:  $1.15 \times 10^{-6}$ .
- (a) 0.335 ulp; (b) 0.665 ulp.
- (a)  $1.0_{10} = 1.\underbrace{0\cdots0}_{23 \text{ bits}} \times 2^0$ . (b) Changing  $a_{23}$  from 0 to 1 changes the value by  $2^{-23}$ . (c)  $2.0_{10} = 1.\underbrace{0\cdots0}_{23 \text{ bits}} \times 2^1$ ;  $0.5_{10} = 1.\underbrace{0\cdots0}_{23 \text{ bits}} \times 2^{-1}$ ;  $1.125_{10} = 1.001\underbrace{0\cdots0}_{20 \text{ bits}} \times 2^0$ . (d) a 1 ulp error in 2.0 changes the value by  $2^{-22}$ ; in 0.5, by  $2^{-24}$ ; and in 1.125, by  $2^{-23}$ .
- Since  $\frac{1}{10} = 1.1001100 \times 2^{-4}$ , the rounded IEEE single-precision representation is

$$\begin{aligned} & 1.10011001100110011001101 \times 2^{-4} \\ &= 2^{-4} \times \left( \frac{3}{2} + \frac{3}{2} \times 2^{-4} + \frac{3}{2} \times 2^{-8} + \frac{3}{2} \times 2^{-12} + \frac{3}{2} \times 2^{-16} + \frac{13}{8} \times 2^{-20} \right) \\ &= \frac{13,421,773}{134,217,728}, \end{aligned}$$

which exceeds  $\frac{1}{10}$  by about  $1.49 \times 10^{-9}$ .

- The absolute error in the floating-point approximation of  $\sin x$  is largest when  $|x|$  is largest in the given interval. Hence the largest absolute error is when  $|x| = 1/\sqrt{2}$ . Since

$$\frac{1}{2} < \frac{1}{\sqrt{2}} < 1,$$

a 1 ulp error in the floating-point representation of  $1/\sqrt{2}$  is  $2^{-24}$ , since the exponent must be  $-1$ .

- (a) The smallest normalized representable value is  $2^{-126} \approx 1.175 \times 10^{-38}$ . (b) The largest normalized representable value is

$$1.\underbrace{1\cdots1}_{23 \text{ bits}} \times 2^{127} = 2^{128} - 2^{128-23} \approx 3.403 \times 10^{38}.$$

8. (a) The smallest normalized representable value is  $2^{-1022} \approx 2.225 \times 10^{-308}$ .  
 (b) The largest normalized representable value is

$$1.\underbrace{1 \cdots 1}_{52 \text{ bits}} \times 2^{1023} = 2^{1023} - 2^{1023-52} \approx 8.988 \times 10^{308}.$$

9. Theorem 2 states that the maximum absolute error when rounding to nearest never exceeds  $\frac{1}{2}$  ulp and never exceeds 1 ulp in any other rounding mode.

*Proof.* Base- $b$  floating-point numbers are represented in the form

$$a_0.a_1a_2 \cdots a_d \times b^e$$

for some fixed value of  $d$ . The largest absolute error when rounding the value

$$x = a_0.a_1a_2 \cdots a_d a_{d+1} a_{d+2} \cdots \times b^e$$

occurs when  $a_{d+1} = b/2$  and  $a_{d+2} = a_{d+3} = \cdots = 0$ . Thus, this error is at most  $\frac{1}{2}$  ulp.

When rounding toward 0, the largest absolute error occurs when  $a_{d+1} = a_{d+2} = \cdots a_K = b - 1$  for an arbitrarily large  $K$ . This value is bounded by 1 ulp. Similar arguments apply when rounding toward  $\pm\infty$ .

10.  $\frac{1}{3} = 1.010\overline{10} \times 2^{-2} \models 1.\underbrace{0101 \cdots 01}_2 1 \times 2^{-2}$  when rounding to nearest. The rounded value equals

$$\frac{1}{4} \times (1 + 2^{-2} + 2^{-4} + \cdots + 2^{-22} + 2^{-23}) = \frac{11,184,811}{33,554,432},$$

which corresponds to an absolute error of approximately  $9.93 \times 10^{-9}$  and a relative error of approximately  $2.98 \times 10^{-8}$ .

Under chopped rounding, the floating-point value is

$$\frac{1}{4} \times (1 + 2^{-2} + 2^{-4} + \cdots + 2^{-22}) = \frac{5,592,405}{16,777,216}$$

which corresponds to an absolute error of approximately  $1.99 \times 10^{-8}$  and a relative error of approximately  $5.96 \times 10^{-8}$ .