INSTRUCTOR'S SOLUTIONS MANUAL

NUMERICAL ANALYSIS
THIRD EDITION

Timothy Sauer

George Mason University

Pearson

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 501 Boylston Street, Boston, MA 02116.

1 17

Pearson

ISBN-13: 978-0-13-469732-1
ISBN-10: 0-13-469732-4

Table of Contents

Chapter 5: Numerical Differentiation and Integration

Chapter 6: Ordinary Differential Equations

Chapter 7: Boundary Value Problems

Chapter 8: Partial Differential Equations

Chapter 9: Random Numbers and Applications

Chapter 10: Trigonometric Interpolation and the FFT

CHAPTER 0
Fundamentals

EXERCISES 0.1 Evaluating a Polynomial

1 (a) $P(x) = 1 + x(1 + x(5 + x(1 + x(6))))$.
$P(\frac{1}{3}) = 6(\frac{1}{3})^4 + (\frac{1}{3})^3 + 5(\frac{1}{3})^2 + \frac{1}{3} + 1 = 1 + \frac{1}{3}(1 + \frac{1}{3}(5 + \frac{1}{3}(1 + \frac{1}{3}(6)))) = 2$.
1 (b) $P(x) = 1 + x(-5 + x(5 + x(4 + x(-3))))$
$P(\frac{1}{3}) = -3(\frac{1}{3})^4 + 4(\frac{1}{3})^3 + 5(\frac{1}{3})^2 - 5(\frac{1}{3}) + 1 = 1 + \frac{1}{3}(-5 + \frac{1}{3}(5 + \frac{1}{3}(4 + \frac{1}{3}(-3)))) = 0$
1 (c) $P(x) = 1 + x(0 + x(-1 + x(1 + x(2))))$
$P(\frac{1}{3}) = 2(\frac{1}{3})^4 + (\frac{1}{3})^3 - (\frac{1}{3})^2 + 1 = 1 + \frac{1}{3}(0 + \frac{1}{3}(-1 + \frac{1}{3}(1 + \frac{1}{3}(2)))) = 77/81$.

2 (a) $P(x) = 7+x(-3+x(-2+x(6)))$; $P(-\frac{1}{2}) = 7+(-\frac{1}{2})(-3+(-\frac{1}{2})(-2+(-\frac{1}{2})(6))) = 29/4$.
2 (b) $P(x) = 1 + x(-3 + x(1 + x(-3 + x(-1 + x(8)))))$;
$P(-\frac{1}{2}) = 1 + (-\frac{1}{2})(-3 + (-\frac{1}{2})(1 + (-\frac{1}{2})(-3 + (-\frac{1}{2})(-1 + (-\frac{1}{2})(8))))) = 45/16$.
2 (c) $P(x) = 4 + x(-2 + x(0 + x(0 + x(-2 + x(0 + x(4))))))$;
$P(-\frac{1}{2}) = 4 + (-\frac{1}{2})(-2 + (-\frac{1}{2})(0 + (-\frac{1}{2})(0 + (-\frac{1}{2})(-2 + (-\frac{1}{2})(0 + (-\frac{1}{2})(4)))))) = 79/16$.

3 $P(\frac{1}{2}) = 1 + (\frac{1}{2})^2(2 + (\frac{1}{2})^2(-4 + (\frac{1}{2})^2(1))) = 81/64$.

4 (a) $P(5) = 1 + 5(\frac{1}{2} + (5-2)(\frac{1}{2} + (5-3)(-\frac{1}{2}))) = -4$
4 (b) $P(-1) = 1 + (-1)(\frac{1}{2} + (-1-2)(\frac{1}{2} + (-1-3)(-\frac{1}{2}))) = 8$

5 (a) $P(\frac{1}{2}) = 4 + \frac{1}{2}(4 + (\frac{1}{2}-1)(1 + (\frac{1}{2}-2)(3 + (\frac{1}{2}-3)(2)))) = 5$
5 (b) $P(-\frac{1}{2}) = 4 - \frac{1}{2}(4 + (-\frac{1}{2}-1)(1 + (-\frac{1}{2}-2)(3 + (-\frac{1}{2}-3)(2)))) = 41/4$

6 (a) $P(x) = a_0 + x^5(a_5 + x^5(a_{10} + x^5 a_{15}))$. The three multiplications $x^2 = x \cdot x, x^4 = x^2 \cdot x^2, x^5 = x^4 \cdot x$ are needed, together with 3 multiplications and 3 additions from the nested multiplication. Total of 6 multiplications and 3 additions.
6 (b) $P(x) = x^7(a_7 + x^5(a_{12} + x^5(a_{17} + x^5(a_{22} + x^5 a_{27}))))$. The four multiplications $x^2 = x \cdot x, x^4 = x^2 \cdot x^2, x^5 = x^4 \cdot x, x^7 = x^5 \cdot x^2$ are needed, together with 5 multiplications and 4 additions from the nested multiplication. Total of 9 multiplications and 4 additions.

7 The degree n polynomial with base points is $P(x) = c_1 + (x - r_1)(c_2 + (x - r_2)(c_3 + (x - r_3)(c_4 + \ldots + (x - r_n)c_{n+1})))$. The operations needed are n multiplications and $2n$ additions.

COMPUTER PROBLEMS 0.1

1 The MATLAB command `nest(50,ones(51,1),1.00001)` gives 51.01275208274999, differing from $(x^{51} - 1)/(x - 1)$ with $x = 1.00001$ by 4.76×10^{-12}.

2 The command `nest(99,(-1).^(0:99),1.00001)` gives -0.00050024507964763. The equivalent expression $(1-x^{100})/(1+x)$ for $x = 1.00001$ differs by 1.713×10^{-16}.

EXERCISES 0.2 Binary Numbers

1 (a) $(64)_{10} = (2^6)_{10} = (1000000)_2$
1 (b) $(17)_{10} = (16+1)_{10} = (10001)_2$
1 (c)

$$\begin{aligned}
79 \div 2 &= 39\,\mathrm{R}\,1 \\
39 \div 2 &= 19\,\mathrm{R}\,1 \\
19 \div 2 &= 9\,\mathrm{R}\,1 \\
9 \div 2 &= 4\,\mathrm{R}\,1 \\
4 \div 2 &= 2\,\mathrm{R}\,0 \\
2 \div 2 &= 1\,\mathrm{R}\,0 \\
1 \div 2 &= 0\,\mathrm{R}\,1
\end{aligned}$$

Therefore $(79)_{10} = (1001111)_2$.
1 (d)

$$\begin{aligned}
227 \div 2 &= 113\,\mathrm{R}\,1 \\
113 \div 2 &= 56\,\mathrm{R}\,1 \\
56 \div 2 &= 28\,\mathrm{R}\,0 \\
28 \div 2 &= 14\,\mathrm{R}\,0 \\
14 \div 2 &= 7\,\mathrm{R}\,0 \\
7 \div 2 &= 3\,\mathrm{R}\,1 \\
3 \div 2 &= 1\,\mathrm{R}\,1 \\
1 \div 2 &= 0\,\mathrm{R}\,1
\end{aligned}$$

Therefore $(227)_{10} = (11100011)_2$.

2 (a) $(1/8)_{10} = (2^{-3})_{10} = (0.001)_2$
2 (b) $(7/8)_{10} = (2^{-1}+2^{-2}+2^{-3})_{10} = (0.111)_2$
2 (c) $(35/16)_{10} = (2+3/16)_{10} = (2+1/8+1/16)_{10} = (10.0011)_2$

2 (d)

$$\begin{aligned}
31/64 \times 2 &= 31/32 + 0 \\
31/32 \times 2 &= 15/16 + 1 \\
15/16 \times 2 &= 7/8 + 1 \\
7/8 \times 2 &= 3/4 + 1 \\
3/4 \times 2 &= 1/2 + 1 \\
1/2 \times 2 &= 0 + 1
\end{aligned}$$

Therefore $(31/64)_{10} = (0.011111)_2$.

3 (a) $10.5 = 10 + 0.5$. Integer part: $(10)_{10} = (8+2)_{10} = (1010)_2$. Fractional part: $(0.5)_{10} = (0.1)_2$, so $(10.5)_{10} = (1010.1)_2$.

3 (b)

$$\begin{aligned}
\frac{1}{3} \times 2 &= \frac{2}{3} + 0 \\
\frac{2}{3} \times 2 &= \frac{1}{3} + 1 \\
\frac{1}{3} \times 2 &= \frac{2}{3} + 0 \\
&\vdots
\end{aligned}$$

Therefore $(\frac{1}{3})_{10} = (0.\overline{01})_2$.

3 (c)

$$\begin{aligned}
\frac{5}{7} \times 2 &= \frac{3}{7} + 1 \\
\frac{3}{7} \times 2 &= \frac{6}{7} + 0 \\
\frac{6}{7} \times 2 &= \frac{5}{7} + 1 \\
\frac{5}{7} \times 2 &= \frac{3}{7} + 1 \\
\frac{3}{7} \times 2 &= \frac{6}{7} + 0 \\
&\vdots
\end{aligned}$$

Therefore $(\frac{5}{7})_{10} = (0.\overline{101})_2$.

3 (d) $(12.8)_{10} = (12)_{10} + (0.8)_{10}; (12)_{10} = (1100)_2$.

$$\begin{aligned} 0.8 \times 2 &= 0.6 + 1 \\ 0.6 \times 2 &= 0.2 + 1 \\ 0.2 \times 2 &= 0.4 + 0 \\ 0.4 \times 2 &= 0.8 + 0 \\ 0.8 \times 2 &= 0.6 + 1 \\ &\vdots \end{aligned}$$

Therefore $(12.8)_{10} = (1100.\overline{1100})_2$.

3 (e) $(55.4)_{10} = (55)_{10} + (0.4)_{10}; (55)_{10} = (32 + 16 + 4 + 2 + 1)_{10} = (110111)_2$.

$$\begin{aligned} 0.4 \times 2 &= 0.8 + 0 \\ 0.8 \times 2 &= 0.6 + 1 \\ 0.6 \times 2 &= 0.2 + 1 \\ 0.2 \times 2 &= 0.4 + 0 \\ 0.4 \times 2 &= 0.8 + 0 \\ &\vdots \end{aligned}$$

Therefore $(55.4)_{10} = (110111.\overline{0110})_2$.

3 (f)

$$\begin{aligned} 0.1 \times 2 &= 0.2 + 0 \\ 0.2 \times 2 &= 0.4 + 0 \\ 0.4 \times 2 &= 0.8 + 0 \\ 0.8 \times 2 &= 0.6 + 1 \\ 0.6 \times 2 &= 0.2 + 1 \\ 0.2 \times 2 &= 0.4 + 0 \\ &\vdots \end{aligned}$$

Therefore $(0.1)_{10} = (0.0\overline{0011})_2$.

4 (a) $11.25 = 11 + 0.25$. Integer part: $(11)_{10} = (8 + 2 + 1)_{10} = (1011)_2$. Fractional part: $(0.25)_{10} = (0.01)_2$, so $(10.25)_{10} = (1011.01)_2$.

4 (b)

$$\begin{aligned}
\frac{2}{3} \times 2 &= \frac{1}{3} + 1 \\
\frac{1}{3} \times 2 &= \frac{2}{3} + 0 \\
\frac{2}{3} \times 2 &= \frac{1}{3} + 1 \\
&\vdots
\end{aligned}$$

Therefore $(\frac{2}{3})_{10} = (0.\overline{10})_2$.

4 (c)

$$\begin{aligned}
\frac{3}{5} \times 2 &= \frac{1}{5} + 1 \\
\frac{1}{5} \times 2 &= \frac{2}{5} + 0 \\
\frac{2}{5} \times 2 &= \frac{4}{5} + 0 \\
\frac{4}{5} \times 2 &= \frac{3}{5} + 1 \\
\frac{3}{5} \times 2 &= \frac{1}{5} + 1 \\
&\vdots
\end{aligned}$$

Therefore $(\frac{3}{5})_{10} = (0.\overline{1001})_2$.

4 (d) $(3.2)_{10} = (3)_{10} + (0.2)_{10}; (3)_{10} = (11)_2$.

$$\begin{aligned}
0.2 \times 2 &= 0.4 + 0 \\
0.4 \times 2 &= 0.8 + 0 \\
0.8 \times 2 &= 0.6 + 1 \\
0.6 \times 2 &= 0.2 + 1 \\
0.2 \times 2 &= 0.4 + 0 \\
&\vdots
\end{aligned}$$

Therefore $(3.2)_{10} = (11.\overline{0011})_2$.

4 (e) $(30.6)_{10} = (30)_{10} + (0.6)_{10}; (30)_{10} = (16+8+4+2)_{10} = (11110)_2.$

$$\begin{aligned} 0.6 \times 2 &= 0.2 + 1 \\ 0.2 \times 2 &= 0.4 + 0 \\ 0.4 \times 2 &= 0.8 + 0 \\ 0.8 \times 2 &= 0.6 + 1 \\ 0.6 \times 2 &= 0.2 + 1 \\ &\vdots \end{aligned}$$

Therefore $(30.6)_{10} = (11110.\overline{1001})_2$.

4 (f) $(99.9)_{10} = (99)_{10} + (0.9)_{10}; (99)_{10} = (64+32+2+1)_{10} = (1100011)_2.$

$$\begin{aligned} 0.9 \times 2 &= 0.8 + 1 \\ 0.8 \times 2 &= 0.6 + 1 \\ 0.6 \times 2 &= 0.2 + 1 \\ 0.2 \times 2 &= 0.4 + 0 \\ 0.4 \times 2 &= 0.8 + 0 \\ 0.8 \times 2 &= 0.6 + 1 \\ &\vdots \end{aligned}$$

Therefore $(99.9)_{10} = (1100011.1\overline{1100})_2$.

5 $(\pi)_{10} = (3)_{10} + (\pi - 3)_{10}$

$$\begin{aligned} 0.14159265 \times 2 &= 0.28318531 + 0 \\ 0.28318531 \times 2 &= 0.56637061 + 0 \\ 0.56637061 \times 2 &= 0.13274123 + 1 \\ 0.13274123 \times 2 &= 0.26548246 + 0 \\ 0.26548246 \times 2 &= 0.53096491 + 0 \\ 0.53096491 \times 2 &= 0.06192983 + 1 \\ 0.06192983 \times 2 &= 0.12385966 + 0 \\ 0.12385966 \times 2 &= 0.24771932 + 0 \\ 0.24771932 \times 2 &= 0.49543864 + 0 \\ 0.49543864 \times 2 &= 0.99087728 + 0 \\ 0.99087728 \times 2 &= 0.98175455 + 1 \\ 0.98175455 \times 2 &= 0.96350910 + 1 \\ 0.96350910 \times 2 &= 0.92701821 + 1 \\ &\vdots \end{aligned}$$

Therefore $(\pi)_{10} = (11.0010010000111\ldots)_2$.

6 $(e)_{10} = (2)_{10} + (e-2)_{10}$

$$\begin{aligned}
0.71828183 \times 2 &= 0.43656366 + 1 \\
0.43656366 \times 2 &= 0.87312731 + 0 \\
0.87312731 \times 2 &= 0.74625463 + 1 \\
0.74625463 \times 2 &= 0.49250926 + 1 \\
0.49250926 \times 2 &= 0.98501851 + 0 \\
0.98501851 \times 2 &= 0.97003702 + 1 \\
0.97003702 \times 2 &= 0.94007404 + 1 \\
0.94007404 \times 2 &= 0.88014809 + 1 \\
0.88014809 \times 2 &= 0.76029617 + 1 \\
0.76029617 \times 2 &= 0.52059234 + 1 \\
0.52059234 \times 2 &= 0.04118468 + 1 \\
0.04118468 \times 2 &= 0.08236937 + 0 \\
0.08236937 \times 2 &= 0.16473874 + 0 \\
&\vdots
\end{aligned}$$

Therefore $(e)_{10} = (10.1011011111100\ldots)_2$.

7 (a) $(1010101)_2 = (2^0 + 2^2 + 2^4 + 2^6)_{10} = (1 + 4 + 16 + 64)_{10} = (85)_{10}$

7 (b) $(1011.101)_2 = (2^3 + 2^1 + 2^0 + 2^{-1} + 2^{-3})_{10} = (11 + \frac{1}{2} + \frac{1}{8})_{10} = (93/8)_{10}$.

7 (c) $(10111.\overline{01})_2 = (2^4 + 2^2 + 2^1 + 2^0)_{10} + (0.\overline{01})_2$. Set $x = (0.\overline{01})_2$. Then $2^2x - x = (01)_2 = 1$ implies $x = \frac{1}{3}$. Therefore $(10111.\overline{01})_2 = (23 + \frac{1}{3})_{10} = (70/3)_{10}$.

7 (d) $(110.\overline{10})_2 = (2^2 + 2^1)_{10} + (0.\overline{10})_2$. Set $x = (0.\overline{10})_2$. Then $2^2x - x = (10)_2$ implies $x = \frac{2}{3}$. Therefore $(110.\overline{10})_2 = (6 + \frac{2}{3})_{10} = (20/3)_{10}$.

7 (e) $(10.\overline{110})_2 = (2)_{10} + (0.\overline{110})_2$. Set $x = (0.\overline{110})_2$. Then $2^3x - x = (110)_2 = 6$ implies $x = 6/7$. Therefore $(10.\overline{110})_2 = (2 + \frac{6}{7})_{10} = (20/7)_{10}$.

7 (f) $(110.1\overline{101})_2 = (6)_{10} + (\frac{1}{2})_{10} + (0.0\overline{101})_2 = (\frac{13}{2} + \frac{x}{2})_{10}$, where $x = (0.\overline{101})_2$. Since $2^3x - x = (101)_2 = 5$, $x = 5/7$. Therefore $(110.1\overline{101})_2 = (\frac{13}{2} + \frac{5}{7}\frac{1}{2})_{10} = (48/7)_{10}$.

7 (g) $(10.010\overline{1101})_2 = (2)_{10} + (\frac{1}{4})_{10} + \frac{1}{8}(0.\overline{1101})_2$. Set $x = (0.\overline{1101})_2$. Then $2^4x - x = (1101)_2 = 13$, implying that $x = \frac{13}{15}$. Therefore $(10.010\overline{1101})_2 = (\frac{9}{4} + \frac{1}{8}\frac{13}{15})_{10} = (283/120)_{10}$.

7 (h) $(111.\overline{1})_2 = (7)_{10} + (0.\overline{1})_2 = (7)_{10} + x$, where $x = (0.\overline{1})_2$. Since $2^1x - x = (1)_2$, $x = 1$, and $(111.\overline{1})_2 = (7 + 1)_{10} = (8)_{10}$.

8 (a) $(11011)_2 = (2^0 + 2^1 + 2^3 + 2^4)_{10} = (1 + 2 + 8 + 16)_{10} = (27)_{10}$

8 (b) $(110111.001)_2 = (2^5 + 2^4 + 2^2 + 2^1 + 2^0 + 2^{-3})_{10} = (55 + \frac{1}{8})_{10}$.

8 (c) $(111.\overline{001})_2 = (2^2 + 2^1 + 2^0)_{10} + (0.\overline{001})_2$. Set $x = (0.\overline{001})_2$. Then $2^3x - x = (001)_2 = 1$ implies $x = 1/7$. Therefore $(111.\overline{001})_2 = (7 + 1/7)_{10}$.

8 (d) $(1010.\overline{01})_2 = (2^3+2^1)_{10} + (0.\overline{01})_2$. Set $x = (0.\overline{01})_2$. Then $2^2x - x = (01)_2$ implies $x = \frac{1}{3}$. Therefore $(1010.\overline{01})_2 = (10 + \frac{1}{3})_{10} = (10 + 1/3)_{10}$.

8 (e) $(10111.1\overline{0101})_2 = (10111.\overline{10})_2 = (2^4+2^2+2^1+2^0)_{10} + (0.\overline{10})_2$. Set $x = (0.\overline{10})_2$. Then $2^2x - x = (10)_2 = 2$ implies $x = 2/3$. Therefore $(10111.1\overline{0101})_2 = (23 + \frac{2}{3})_{10}$.

8 (f) $(1111.010\overline{001})_2 = (15)_{10} + (1/4)_{10} + \frac{1}{8}(0.\overline{001})_2 = (15 + 1/4 + \frac{x}{8})_{10}$, where $x = (0.\overline{001})_2$. Since $2^3x - x = (001)_2 = 5$, $x = 1/7$. Therefore $(1111.010\overline{001})_2 = (15 + 1/4 + \frac{1}{8}\frac{1}{7})_{10} = (15 + 15/56)_{10}$.

EXERCISES 0.3 Floating Point Representation of Real Numbers

1 (a) $(\frac{1}{4})_{10} = (0.01)_2$; $\text{fl}(\frac{1}{4}) = +1.0 \times 2^{-2}$.

1 (b) $(\frac{1}{3})_{10} = (0.\overline{01})_2 =$
$+1.\boxed{0101010101010101010101010101010101010101010101010101}\,0101\ldots \times 2^{-2}$.
The Rounding to Nearest Rule says to round down when the 53rd bit is 0.
$\text{fl}(\frac{1}{3}) = +1.\boxed{0101010101010101010101010101010101010101010101010101} \times 2^{-2}$.

1 (c) $(\frac{2}{3})_{10} = (0.\overline{10})_2 =$
$+1.\boxed{0101010101010101010101010101010101010101010101010101}\,0101\ldots \times 2^{-1}$.
$\text{fl}(\frac{2}{3}) = +1.\boxed{0101010101010101010101010101010101010101010101010101} \times 2^{-1}$.

1 (d) $(0.9)_{10} = (0.1\overline{1100})_2 =$
$+1.\boxed{1100110011001100110011001100110011001100110011001100}\,1100\ldots \times 2^{-1}$.
The Rounding to Nearest Rule says to round up since the 53rd bit is nonzero, and further bits are nonzero.
$\text{fl}(0.9) = +1.\boxed{1100110011001100110011001100110011001100110011001101} \times 2^{-1}$.

2 (a) $(9.5)_{10} = (1001.1)_2$; $\text{fl}(9.5) = 1.0011 \times 2^3$.

2 (b) $(9.6)_{10} = (1001.\overline{1001})_2 = 1.001\overline{1001} \times 2^3 =$
$+1.\boxed{0011001100110011001100110011001100110011001100110011}\,0011\ldots \times 2^3$.
$\text{fl}(9.6) = +1.\boxed{0011001100110011001100110011001100110011001100110011} \times 2^3$.

2 (c) $(100.2)_{10} = (1100100.\overline{0011})_2 = 1.100100\overline{0011} \times 2^6 =$
$+1.\boxed{1001000011001100110011001100110011001100110011001100}\,1100\ldots \times 2^6$.
$\text{fl}(100.2) = +1.\boxed{1001000011001100110011001100110011001100110011001101} \times 2^6$.

2 (d) $\left(\frac{44}{7}\right)_{10} = (6 + \frac{2}{7})_{10} = (110.\overline{010})_2 =$
$+1.\boxed{1001001001001001001001001001001001001001001001001001}\,0010\ldots \times 2^2$.
$\text{fl}\left(\frac{44}{7}\right) = +1.\boxed{1001001001001001001001001001001001001001001001001001} \times 2^2$.

3 Note that $\text{fl}(5) = 1.01 \times 2^2$. Adding 1 as bit $3, 4, \ldots, 52$ of the mantissa will not incur rounding error. These correspond to 2^{-k} for $k = 1, 2, \ldots, 50$.

4 Note that $\text{fl}(19) = 1.0011 \times 2^4$. Adding 1 to bit 52 of the mantissa, corresponding to $19 + 2^{-48}$, will not be rounded away, and so 48 is the largest such k.

5 (a) $1 + (2^{-51} + 2^{-53}) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000010}\,1 \times 2^0$.
$\mathrm{fl}(1 + (2^{-51} + 2^{-53})) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000010} \times 2^0$, using the Rounding to Nearest Rule. Therefore $\mathrm{fl}((1 + (2^{-51} + 2^{-53})) - 1) =$
$.\boxed{0000000000000000000000000000000000000000000000000010}$
$= 1.\boxed{0000000000000000000000000000000000000000000000000000} \times 2^{-51} = 2^{-51}$.

5 (b) $1 + (2^{-51} + 2^{-52} + 2^{-53}) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000011}\,1 \times 2^0$.
$\mathrm{fl}(1 + (2^{-51} + 2^{-52} + 2^{-53})) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000100} \times 2^0$, using the Rounding to Nearest Rule. Therefore $\mathrm{fl}((1 + (2^{-51} + 2^{-52} + 2^{-53})) - 1) =$
$.\boxed{0000000000000000000000000000000000000000000000000100}$
$= 1.\boxed{0000000000000000000000000000000000000000000000000000} \times 2^{-50} = 2^{-50}$.

6 (a) $1 + (2^{-51} + 2^{-52} + 2^{-54})$
$= +1.\boxed{0000000000000000000000000000000000000000000000000011}\,01 \times 2^0$.
$\mathrm{fl}(1 + (2^{-51} + 2^{-52} + 2^{-54})) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000011} \times 2^0$, using the Rounding to Nearest Rule. Therefore $\mathrm{fl}((1 + (2^{-51} + 2^{-52} + 2^{-54})) - 1) =$
$.\boxed{0000000000000000000000000000000000000000000000000011} =$
$1.\boxed{1000000000000000000000000000000000000000000000000000} \times 2^{-51}$
$= 2^{-51} + 2^{-52} = 3\epsilon_{\mathrm{mach}}$.

6 (b) $1 + (2^{-51} + 2^{-52} + 2^{-60}) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000011}\,00000001 \times 2^0$.
$\mathrm{fl}(1 + (2^{-51} + 2^{-52} + 2^{-60})) =$
$+1.\boxed{0000000000000000000000000000000000000000000000000011} \times 2^0$, using the Rounding to Nearest Rule. Therefore $\mathrm{fl}((1 + (2^{-51} + 2^{-52} + 2^{-60})) - 1) =$
$.\boxed{0000000000000000000000000000000000000000000000000011} =$
$1.\boxed{1000000000000000000000000000000000000000000000000000} \times 2^{-51}$
$= 2^{-51} + 2^{-52} = 3\epsilon_{\mathrm{mach}}$.

7 (a) $(8)_{10} = (1000.)_2 = 1.0 \times 2^3$. The biased exponent is $3 + 1023 = 1026$, which is $2^{10} + 2$. The sign is 0 (positive), so the sign/exponent is represented by the binary string 0100 0000 0010. The mantissa is 52 zeros, so the machine representation is the 64 bits

$\boxed{0100\ 0000\ 0010\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000}$

or `4020000000000000` in hex format.

7 (b) $(21)_{10} = (10101.)_2 = 1.0101 \times 2^4$. The biased exponent is $4 + 1023 = 1027 = 2^{10} + 3$, represented by 100 0000 0011. The machine representation is

$\boxed{0100\ 0000\ 0011\ 0101\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000}$

or `4035000000000000` in hex format.

7 (c) $(1/8)_{10} = 1.0 \times 2^{-3}$. The biased exponent is $-3 + 1023 = 1020 = 2^{10} - 4$, represented by 011 1111 1100. The machine representation is

$\boxed{0011\ 1111\ 1100\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000\ 0000}$

or `3fc0000000000000` in hex format.

7 (d) $(1/3)_{10} = 1.\overline{01} \times 2^{-2}$, and after rounding down, $\mathrm{fl}(1/3) = 1.0101\ldots0101 \times 2^{-2}$. The biased exponent is $-2 + 1023 = 1021 = 2^{10} - 3$, represented by 011 1111 1101. The machine representation is

$\boxed{0011\ 1111\ 1101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101}$

or `3fd5555555555555` in hex format.

7 (e) $(2/3)_{10} = 1.\overline{01} \times 2^{-1}$, and after rounding down, $\mathrm{fl}(1/3) = 1.0101\ldots0101 \times 2^{-1}$. The biased exponent is $-1 + 1023 = 1022 = 2^{10} - 2$, represented by 011 1111 1110. The machine representation is

$\boxed{0011\ 1111\ 1110\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101\ 0101}$

or `3fe5555555555555` in hex format.

7 (f) $(0.1)_{10} = 1.\overline{1001} \times 2^{-4}$, and after rounding up, $\mathrm{fl}(0.1) = 1.1001\ldots1001\ 1010 \times 2^{-4}$. The biased exponent is $-4 + 1023 = 1019 = 2^{10} - 5$, represented by 011 1111 1011. The machine representation is

$\boxed{0011\ 1111\ 1011\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1010}$

or `3fb999999999999a` in hex format.

7 (g) $(-0.1)_{10} = -1.\overline{1001} \times 2^{-4}$, and after rounding, $\mathrm{fl}(-0.1) = -1.1001\ldots1001\ 1010 \times 2^{-4}$. The biased exponent is $-4 + 1023 = 1019 = 2^{10} - 5$, represented by 011 1111 1011. The machine representation is

$\boxed{1011\ 1111\ 1011\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1010}$

or `bfb999999999999a` in hex format.

7 (h) $(-0.2)_{10} = -1.\overline{1001} \times 2^{-3}$, and after rounding, $\mathrm{fl}(-0.2) = -1.1001\ldots1001\ 1010 \times 2^{-3}$. The biased exponent is $-3 + 1023 = 1020 = 2^{10} - 4$, represented by 011 1111 1100. The machine representation is

$\boxed{1011\ 1111\ 1100\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1001\ 1010}$

or `bfc999999999999a` in hex format.

8 Yes. Yes. No, under chopping, $1/3 + 2/3 = 1 - \epsilon_{\text{mach}}$.

9 (a) $(7/3)_{10} = 1.00\overline{10} \times 2^1$, and after rounding, $\mathrm{fl}(7/3) = 1.0010\ldots1010\ 1011 \times 2^1$. $(4/3)_{10} = 1.\overline{01} \times 2^0$, and after rounding, $\mathrm{fl}(4/3) = 1.01\ldots0101\ 0101 \times 2^0$. Subtracting gives

$$\begin{array}{rl} & 1.\boxed{0010101010101010101010101010101010101010101010101011}0 \times 2^1 \\ - & 0.\boxed{1010101010101010101010101010101010101010101010101010}1 \times 2^1 \\ \hline = & 0.\boxed{1000000000000000000000000000000000000000000000000000}1 \times 2^1 \end{array}$$

that is normalized to

$$1.\boxed{0000000000000000000000000000000000000000000000000001} \times 2^0,$$

which is $1 + \epsilon_{\text{mach}}$. After subtracting 1, the result is that the double precision floating point version of $(7/3 - 4/3) - 1$ is ϵ_{mach}.

9 (b) $(4/3)_{10} = 1.\overline{01} \times 2^0$, and after rounding, fl$(4/3) = 1.01\ldots0101\ 0101 \times 2^0$. $(1/3)_{10} = 1.\overline{01} \times 2^{-2}$, and after rounding, fl$(1/3) = 1.01\ldots0101\ 0101 \times 2^{-2}$. Subtracting gives

$$\begin{array}{rl} & 1.\boxed{0101010101010101010101010101010101010101010101010101}00 \times 2^0 \\ - & 0.\boxed{0101010101010101010101010101010101010101010101010101}01 \times 2^0 \\ \hline = & 0.\boxed{1111111111111111111111111111111111111111111111111111}11 \times 2^0 \end{array}$$

that normalizes to

$$1.\boxed{1111111111111111111111111111111111111111111111111111}1 \times 2^{-1}$$

and rounds to

$$10.\boxed{0000000000000000000000000000000000000000000000000000} \times 2^{-1}$$

which is 1.0×2^0. After subtracting 1, the result is machine zero, not ϵ_{mach}.

10 (a) No.
10 (b) Yes.

11 The associative law of addition fails for floating point addition with the Rounding to Nearest Rule, for example, because $1 + (\epsilon_{\text{mach}}/2 + \epsilon_{\text{mach}}/2) = 1 + \epsilon_{\text{mach}} > 1$, while $(1 + \epsilon_{\text{mach}}/2) + \epsilon_{\text{mach}}/2 = 1$, because $1 + \epsilon_{\text{mach}}/2 = 1$.

12 (a) fl $(1/3) = 1.0101\ldots01 \times 2^{-2}$, with relative rounding error of $2^{-54} < \epsilon_{\text{mach}}/2 = 2^{-53}$.
12 (b) fl $(3.3) = 1.101001100110\ldots0110 \times 2^1$, $3.3 -$ fl $(3.3) = 0.4 \times 2^{-51}$ with relative rounding error of $8\epsilon_{\text{mach}}/33$.
12 (c) fl $(9/7) = 1.010010\ldots0100101 \times 2^0$, fl$(9/7) - 9/7 = 3\epsilon_{\text{mach}}/7$, with relative rounding error of $\epsilon_{\text{mach}}/3$.

13 (a) 2, represented by $010\ldots0$ (b) 2^{-511}, represented by $0010\ldots0$ (c) 0, represented by $10\ldots0$. When bit 4 through 12 is the nonzero bit, the floating point number is positive but less than 2^{-511}. When bit 13 through 64 is the nonzero bit, the number is positive and subnormal, so less than 2^{-511}.

14 (a) 0 (b) 2^{-51} (c) 2^{-51}

15(a) $(8.3)_{10} = 1.0000\overline{1001} \times 2^3$, and rounded, $\text{fl}(8.3) = 1.0000\ 1001\ 1001 \ldots 1001\ 1010 \times 2^3$. $(7.3)_{10} = 1.110\overline{1001} \times 2^2$, and rounded, $\text{fl}(7.3) = 1.1101\ 0011\ 0011 \ldots 0011\ 0011 \times 2^2$. Subtracting gives

$$
\begin{array}{rl}
 & 1.\boxed{0000100110011001100110011001100110011001100110011010}\ 0 \times 2^3 \\
- & 0.\boxed{1110100110011001100110011001100110011001100110011001}\ 1 \times 2^3 \\
\hline
= & 0.\boxed{0010000000000000000000000000000000000000000000000000}\ 1 \times 2^3
\end{array}
$$

that is normalized to

$$= \quad 1.\boxed{0000000000000000000000000000000000000000000000000100} \times 2^0,$$

which is $1 + 2^{-50}$. After subtracting 1, the result is that the double precision floating point version of $(8.3 - 7.3) - 1$ is 2^{-50}.

15(b) $(8.4)_{10} = 1.0000\overline{110} \times 2^3$, and rounded, $\text{fl}(8.4) = 1.0000\ 1100\ 1100 \ldots 1100\ 1101 \times 2^3$. $(7.4)_{10} = 1.11\overline{0110} \times 2^2$, and rounded, $\text{fl}(7.4) = 1.1101\ 1001\ 1001 \ldots 1001\ 1010 \times 2^2$. Subtracting gives

$$
\begin{array}{rl}
 & 1.\boxed{0000110011001100110011001100110011001100110011001101} \times 2^3 \\
- & 0.\boxed{1110110011001100110011001100110011001100110011001101} \times 2^3 \\
\hline
= & 0.\boxed{0010000000000000000000000000000000000000000000000000} \times 2^3
\end{array}
$$

which is 1. After subtracting 1, the result is that the double precision floating point version of $(8.4 - 7.4) - 1$ is 0.

15(c) $(8.8)_{10} = 1.000\overline{1100} \times 2^3$, and rounded, $\text{fl}(8.8) = 1.0001\ 1001\ 1001 \ldots 1001\ 1010 \times 2^3$. $(7.8)_{10} = 1.11\overline{1100} \times 2^2$, and rounded, $\text{fl}(7.8) = 1.1111\ 0011\ 0011 \ldots 0011\ 0011 \times 2^2$. Subtracting gives

$$
\begin{array}{rl}
 & 1.\boxed{0001100110011001100110011001100110011001100110011010}\ 0 \times 2^3 \\
- & 0.\boxed{1111100110011001100110011001100110011001100110011001}\ 1 \times 2^3 \\
\hline
= & 0.\boxed{0010000000000000000000000000000000000000000000000000}\ 1 \times 2^3
\end{array}
$$

that is normalized to

$$= \quad 1.\boxed{0000000000000000000000000000000000000000000000000100} \times 2^0,$$

which is $1 + 2^{-50}$. After subtracting 1, the result is that the double precision floating point version of $(8.8 - 7.8) - 1$ is 2^{-50}.

16 (a) $\text{fl}\,(11/4) = 1.011 \times 2^1$, with rounding error of 0.

16 (b) $\text{fl}\,(2.7) = 1.010110011001\ldots100110010 \times 2^1$, $\text{fl}\,(2.7) - 2.7 = 4\epsilon_{\text{mach}}/5$ with relative rounding error of $8\epsilon_{\text{mach}}/27$

16 (c) $\text{fl}\,(10/3) = 1.1010\ldots1011 \times 2^1$, $\text{fl}(10/3) - 10/3 = 2\epsilon_{\text{mach}}/3$, with relative rounding error of $\epsilon_{\text{mach}}/5$.

EXERCISES 0.4 Loss of Significance

1 (a) For x near $2\pi n$ for integer n, $\sec x \approx 1$, and the numerator exhibits subtraction of nearly equal numbers. An algebraically equivalent expression avoids the difficulty:

$$\begin{aligned}
\frac{1 - 1/\cos x}{\tan^2 x} &= \frac{\cos x - 1}{\cos x \tan^2 x} \\
&= \frac{\cos x - 1}{\sec x \sin^2 x} \cdot \frac{\cos x + 1}{\cos x + 1} \\
&= \frac{\cos^2 - 1}{\sec x \sin^2 x(\cos x + 1)} \\
&= -\frac{1}{1 + \sec x}
\end{aligned}$$

1 (b) For x near 0, the numerator subtracts nearly equal numbers. Simplifying to

$$\frac{1 - (1 - x)^3}{x} = \frac{1 - (1 - 3x + 3x^2 - x^3}{x} = 3 - 3x + x^2$$

eliminates the loss of significance.

1 (c) For x near 0, there is subtraction of nearly equal numbers. Using common denominators eliminates the problem:

$$\frac{1}{1+x} - \frac{1}{1-x} = \frac{1 - x - (1 + x)}{(1 + x)(1 - x)} = \frac{2x}{x^2 - 1}$$

2 $-3.000; 7.579 \times 10^{-14}$

3 Since b is positive, the roots should be calculated as in (0.13):

$$\begin{aligned}
x_1 &= -\frac{b + \sqrt{b^2 + 4 \times 10^{-12}}}{2} \\
x_2 &= \frac{2 \times 10^{-12}}{b + \sqrt{b^2 + 4 \times 10^{-12}}}
\end{aligned}$$

4 8.5

5 -0.125

COMPUTER PROBLEMS 0.4

1 (a) Compare the original expression to the revised version $-1/(1+\sec x)$ from Exercise 1(a).

x	original	revised
0.10000000000000	−0.49874791371143	−0.49874791371143
0.01000000000000	−0.49998749979096	−0.49998749979166
0.00100000000000	−0.49999987501429	−0.49999987499998
0.00010000000000	−0.49999999362793	−0.49999999875000
0.00001000000000	−0.50000004133685	−0.49999999998750
0.00000100000000	−0.50004445029084	−0.49999999999987
0.00000010000000	−0.51070259132757	−0.50000000000000
0.00000001000000	0	−0.50000000000000
0.00000000100000	0	−0.50000000000000
0.00000000010000	0	−0.50000000000000
0.00000000001000	0	−0.50000000000000
0.00000000000100	0	−0.50000000000000
0.00000000000010	0	−0.50000000000000
0.00000000000001	0	−0.50000000000000

1 (b) Compare the original expression to the revised version $3-3x+x^2$ from Exercise 1(b).

x	original	revised
0.10000000000000	2.71000000000000	2.71000000000000
0.01000000000000	2.97010000000001	2.97010000000000
0.00100000000000	2.99700100000000	2.99700100000000
0.00010000000000	2.99970000999905	2.99970001000000
0.00001000000000	2.99997000008379	2.99997000010000
0.00000100000000	2.99999700015263	2.99999700000100
0.00000010000000	2.99999969866072	2.99999970000001
0.00000001000000	2.99999998176759	2.99999997000000
0.00000000100000	2.99999991515421	2.99999999700000
0.00000000010000	3.00000024822111	2.99999999970000
0.00000000001000	3.00000024822111	2.99999999997000
0.00000000000100	2.99993363483964	2.99999999999700
0.00000000000010	3.00093283556180	2.99999999999970
0.00000000000001	2.99760216648792	2.99999999999997

2 (a) $p=8$
2 (b) $p=5$