1.3 Floating Point Number Systems

1. Provide the floating point equivalent for each of the following numbers from the floating point number system $\mathbf{F}(10, 4, 0, 4)$. Consider both chopping and rounding. Compute the absolute and relative error in each floating point equivalent.

(a) π **(b)** e
(c) $\sqrt{2}$ **(d)** $1/7$
(e) $\cos 22^\circ$ **(f)** $\ln 10$
(g) $\sqrt[3]{9}$

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

	Chopping		Rounding	
y	$fl(y)$	error	$fl(y)$	error
π	3.141	$\delta = 5.927 \times 10^{-4}$ $\epsilon = 1.886 \times 10^{-4}$	3.142	$\delta = 4.073 \times 10^{-4}$ $\epsilon = 1.297 \times 10^{-4}$
e	2.718	$\delta = 2.818 \times 10^{-4}$ $\epsilon = 1.037 \times 10^{-4}$	2.718	$\delta = 2.818 \times 10^{-4}$ $\epsilon = 1.037 \times 10^{-4}$
$\sqrt{2}$	1.414	$\delta = 2.136 \times 10^{-4}$ $\epsilon = 1.510 \times 10^{-4}$	1.414	$\delta = 2.136 \times 10^{-4}$ $\epsilon = 1.510 \times 10^{-4}$
$1/7$	0.1428	$\delta = 5.714 \times 10^{-4}$ $\epsilon = 4.000 \times 10^{-4}$	0.1429	$\delta = 4.286 \times 10^{-4}$ $\epsilon = 3.000 \times 10^{-4}$
$\cos 22^\circ$	0.9271	$\delta = 8.385 \times 10^{-5}$ $\epsilon = 9.044 \times 10^{-5}$	0.9272	$\delta = 1.615 \times 10^{-5}$ $\epsilon = 1.741 \times 10^{-5}$
$\ln 10$	2.302	$\delta = 5.851 \times 10^{-4}$ $\epsilon = 2.541 \times 10^{-4}$	2.303	$\delta = 4.149 \times 10^{-4}$ $\epsilon = 1.802 \times 10^{-4}$
$\sqrt[3]{9}$	2.080	$\delta = 8.382 \times 10^{-5}$ $\epsilon = 4.030 \times 10^{-5}$	2.080	$\delta = 8.382 \times 10^{-5}$ $\epsilon = 4.030 \times 10^{-5}$

2. Prove the bounds on the absolute and relative roundoff error associated with rounding:

$$|fl_{\text{round}}(y) - y| \leq \frac{1}{2}\beta^{e-k} \quad \text{and} \quad \frac{|fl_{\text{round}}(y) - y|}{|y|} \leq \frac{1}{2}\beta^{1-k}.$$

Consider the floating point system $\mathbf{F}(\beta, k, m, M)$ with rounding. Let y be a real number whose expansion is given by

$$y = \pm(0.d_1d_2d_3 \cdots d_kd_{k+1} \cdots)_\beta \times \beta^e$$

with $d_1 \neq 0$ and $m \leq e \leq M$. If we let d denote $\beta/2$, then a bound on the absolute size of the roundoff error is

$$\begin{aligned} |fl_{\text{round}}(y) - y| &\leq (0.d)_\beta \times \beta^{e-k} \\ &= \frac{1}{2}\beta^{e-k}. \end{aligned}$$

Provided $y \neq 0$, given the restriction on d_1,

$$\begin{aligned} |y| &= (0.d_1 d_2 d_3 \cdots)_\beta \times \beta^e \\ &\geq (0.1)_\beta \times \beta^e = \beta^{e-1}. \end{aligned}$$

Therefore, the relative error in $fl_{\text{round}}(y)$ is bounded by

$$\frac{|fl_{\text{round}}(y) - y|}{|y|} \leq \frac{\frac{1}{2}\beta^{e-k}}{\beta^{e-1}} = \frac{1}{2}\beta^{1-k}.$$

3. Show that machine precision is the smallest floating point number, v, such that $fl(1+v) > 1$.

First consider the floating point number system $\mathbf{F}(\beta, k, m, M)$ with chopping. The number one is represented by the expansion

$$(0.1 \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^1.$$

If we let

$$\begin{aligned} v = u = \beta^{1-k} &= (0.1 \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^{2-k} \\ &= (0. \underbrace{00\cdots00}_{k-1 \text{ zeros}} 1 \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^1, \end{aligned}$$

then

$$1 + v = (0.1 \underbrace{00\cdots00}_{k-2 \text{ zeros}} 1 \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^1$$

and

$$fl_{\text{chop}}(1+v) = 1.00\cdots001 > 1.$$

If we assign to *v* any value smaller than *u*, then the *k*th digit in the mantissa of $1+v$ is zero and $fl_{\text{chop}}(1+v) = 1$. Thus, with chopping, machine precision is the smallest floating point number, *v*, such that $fl(1+v) > 1$.

Now, consider the floating point number system $\mathbf{F}(\beta, k, m, M)$ with rounding. For notational convenience, let d denote $\beta/2$. If we take

$$v = u = \frac{1}{2}\beta^{1-k} = (0.d \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^{1-k}$$

$$= \quad (0.\underbrace{00\cdots00}_{k \text{ zeros}}d\ \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^1,$$

then

$$1+v = (0.1\ \underbrace{00\cdots00}_{k-1 \text{ zeros}}\ d\ \underbrace{00\cdots00}_{k-1 \text{ zeros}})_\beta \times \beta^1$$

and

$$fl_{\text{round}}(1+v) = 1.00\cdots001 > 1.$$

If we assign to v any value smaller than u, then the $(k+1)$st digit in the mantissa of $1+v$ is smaller than $\beta/2$ and $fl_{\text{round}}(1+v) = 1$. Thus, with rounding, machine precision is the smallest floating point number, v, such that $fl(1+v) > 1$.

4. **(a)** Construct an algorithm to determine machine precision and another algorithm to determine the smallest positive number of a floating point number system.

(b) Implement the algorithms from part (a) to determine machine precision and the smallest positive number on your computing system. Consider both single and double precision.

(c) Assuming that your computing system uses $\beta = 2$ and rounding, use the results from part (b) to determine the values for k and m.

(a) Assuming the floating point system uses rounding, here is an algorithm to determine machine precision. Multiplication by β is performed in the output step because the `while` loop terminates when one too many divisions by β have been carried out.

GIVEN: base β

STEP 1: initialize $u = 1/2$
STEP 2: while $(1 + u > 1)$
replace u by u/β
OUTPUT: $\beta \cdot u$

Here is an algorithm to determine the smallest positive number, assuming that underflow is handled by setting the value to zero.

GIVEN: base β

STEP 1: initialize $temp = 1$
STEP 2: while $(temp > 0)$
STEP 3: set $sm = temp$
replace *temp* by $temp/\beta$
OUTPUT: sm

(b) Answers will of course vary. On a SunBlade 100, machine precision in both single and double precision is 2.22045×10^{-16}. The smallest positive number in single precision is 1.4013×10^{-45} and in double precision is 4.94066×10^{-324}.

(c) In general, machine precision with rounding is $\frac{1}{2}\beta^{1-k}$ and the smallest positive number is $(0.1)_\beta \times \beta^m = \beta^{m-1}$. Assuming $\beta = 2$, we solve $2^{-k} = 2.22045 \times 10^{-16}$ to find $k = 52$ in both single and double precision on the SunBlade 100. In single precision, we solve $2^{m-1} = 1.4013 \times 10^{-45}$ to find $m = -148$; in double precision, the equation $2^{m-1} = 4.94066 \times 10^{-324}$ yields $m = -1073$.

5. Determine machine precision, the smallest positive number and the largest positive number for the floating point number system used by your calculator. Assuming the calculator uses $\beta = 10$, determine the values for k, m and M.

Answers will of course vary. On a Casio $fx - 300SA$, machine precision is 5×10^{-10}, the smallest positive number is 10^{-99} and the largest positive number is $9.999999999 \times 10^{99}$. In general, machine precision with rounding is $\frac{1}{2}\beta^{1-k}$, the smallest positive number is $(0.1)_\beta \times \beta^m = \beta^{m-1}$ and the largest positive number is $(1 - \beta^{-k}) \times \beta^M$. Assuming that the calculator uses $\beta = 10$, the values for machine precision, the smallest positive number and the largest positive number on the Casio $fx - 300SA$ determine $k = 10$, $m = -98$ and $M = 100$.

6. Determine the number of significant decimal digits and the number of significant binary digits to which each of the following pairs of numbers agree.

(a) $355/113$ and π

(b) $685/252$ and e

(c) $\sqrt{10002}$ and $\sqrt{10001}$

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

(a) Because

$$\left| \frac{\frac{355}{113} - \pi}{\pi} \right| = 8.491 \times 10^{-8}$$

and

$$10^{-8} < 8.491 \times 10^{-8} \leq 10^{-7},$$

it follows that $\frac{355}{113}$ and π agree to at least 7 and at most 8 decimal digits. Since

$$2^{-24} = 5.960 \times 10^{-8} < 8.491 \times 10^{-8} < 1.192 \times 10^{-7} = 2^{-23},$$

we see that $\frac{355}{113}$ and π agree to at least 23 and at most 24 binary digits.

(b) Because

$$\left| \frac{\frac{685}{252} - e}{e} \right| = 1.025 \times 10^{-5}$$

and

$$10^{-5} < 1.025 \times 10^{-5} \leq 10^{-4},$$

it follows that $\frac{685}{252}$ and e agree to at least 4 and at most 5 decimal digits. Since

$$2^{-17} = 7.629 \times 10^{-6} < 1.025 \times 10^{-5} < 1.526 \times 10^{-5} = 2^{-16},$$

we see that $\frac{685}{252}$ and e agree to at least 16 and at most 17 binary digits.

(c) 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. Since

$$2^{-15} = 3.052 \times 10^{-5} < 4.999 \times 10^{-5} < 6.103 \times 10^{-5} = 2^{-14},$$

we see that $\sqrt{10002}$ and $\sqrt{10001}$ agree to at least 14 and at most 15 binary digits.

(d) Because

$$\left| \frac{\frac{103}{280} - 1/e}{1/e} \right| = 6.061 \times 10^{-5}$$

and

$$10^{-5} < 6.061 \times 10^{-5} \leq 10^{-4},$$

it follows that $\frac{103}{280}$ and $1/e$ agree to at least 4 and at most 5 decimal digits. Since

$$2^{-15} = 3.052 \times 10^{-5} < 6.061 \times 10^{-5} < 6.103 \times 10^{-5} = 2^{-14},$$

we see that $\frac{103}{280}$ and $1/e$ agree to at least 14 and at most 15 binary digits.

7. The ideal gas law states that $PV = nRT$, where P is the pressure of the gas, V is the volume, n is the number of moles, T is the temperature and $R = 0.08206$ atm·m^3/moles·K is the universal gas constant.

(a) Experimentally, it has been determined that $P = 0.750$ atm, $V = 1.15$ m^3 and $T = 294.1$K. Assuming that all values have been rounded to the digits shown, in what range of values does n fall?

(b) Experimentally, it has been determined that $V = 0.331$ m^3, $n = 0.00712$ moles and $T = 264.7$K. Assuming that all values have been rounded to the digits shown, in what range of values does P fall?

(a) With

$$0.7495 \text{ atm} < P < 0.7505 \text{ atm}$$
$$1.145 \text{ m}^3 < V < 1.155 \text{ m}^3$$
$$294.05 \text{ K} < T < 294.15 \text{ K}$$

it follows from the ideal gas law that

$$\frac{(0.7495)(1.145)}{(0.08206)(294.15)} < n < \frac{(0.7505)(1.155)}{(0.08206)(294.05)}$$

or $0.03555 \text{ moles} < n < 0.03592 \text{ moles}$.

(b) With

$$0.3305 \text{ m}^3 < V < 0.3315 \text{ m}^3$$
$$0.007115 \text{ moles} < n < 0.007125 \text{ moles}$$
$$264.65 \text{ K} < T < 264.75 \text{ K}$$

it follows from the ideal gas law that

$$\frac{(0.007115)(0.08206)(264.65)}{0.3315} < P < \frac{(0.007125)(0.08206)(264.75)}{0.3305}$$

or $0.46612 \text{ atm} < P < 0.46836 \text{ atm}$.

8. In a physics laboratory, students measure the mass of a rectangular block to be 243.27 ± 0.005 grams. The length, width and depth of the block are measured to be 7.8 ± 0.05 cm, 3.1 ± 0.05 cm and 4.2 ± 0.05 cm, respectively.

(a) In what range of values does the volume of the block fall?

(b) In what range of values does the density of the block fall? Density is mass per unit volume.

(a) With

$$7.75 \text{ cm} < \text{length} < 7.85 \text{ cm}$$
$$3.05 \text{ cm} < \text{width} < 3.15 \text{ cm}$$
$$4.15 \text{ cm} < \text{depth} < 4.25 \text{ cm}$$

it follows that

$$98.095625 \text{ cm}^3 = (7.75)(3.05)(4.15) < \text{volume} < (7.85)(3.15)(4.25) = 105.091875 \text{ cm}^3.$$

(b) Density is defined as mass per unit volume. It is given that

$$243.265 \text{ grams} < \text{mass} < 243.275 \text{ grams},$$

and we determined in part (a) that $98.095625 \text{ cm}^3 < \text{volume} < 105.091875 \text{ cm}^3$, so

$$2.32 \frac{\text{grams}}{\text{cm}^3} = \frac{243.625}{105.091875} < \text{density} < \frac{243.275}{98.095625} = 2.48 \frac{\text{grams}}{\text{cm}^3}.$$

9. Students are using a pendulum to experimentally determine the acceleration due to gravity, g. They measure the period, T, of the pendulum to be 2.2 seconds, and the length, l, of the pendulum to be 1.15 meters. Assuming that all values are correct to the digits shown, in what range of values does g fall? The variables in this problem are related by the formula $T = 2\pi\sqrt{l/g}$.

Solving $T = 2\pi\sqrt{l/g}$ for g yields $g = 4\pi^2 l/T^2$. With

$$2.15 \text{ s} < T < 2.25 \text{ s} \qquad \text{and} \qquad 1.145 \text{ m} < l < 1.155 \text{ m},$$

it follows that

$$4\pi^2 \frac{1.145}{2.25^2} < g < 4\pi^2 \frac{1.155}{2.15^2},$$

or $8.929 \text{ m/s}^2 < g < 9.864 \text{ m/s}^2$.

10. Determine machine precision, the smallest positive number and the largest positive number in the IEEE standard double precision system. Approximately how many significant decimal digits does the double precision standard supply?

With $\beta = 2$ and $k = 53$, machine precision with rounding is

$$u = \frac{1}{2} 2^{1-53} = 2^{-53} \approx 1.11 \times 10^{-16}.$$

Accordingly, there are between 15 and 16 significant decimal digits available in IEEE standard double precision. The smallest positive number in double precision is

$$(0.1)_2 \times 2^{-1021} = 2^{-1022} \approx 2.23 \times 10^{-308},$$

while the largest positive number is

$$(0.111\cdots1)_2 \times 2^{1024} = (1 - 2^{-53})2^{1024} \approx 1.80 \times 10^{308}.$$

11. In addition to the standard single and double precision floating point systems, Intel microprocessors also have an extended precision system $\mathbf{F}(2, 64, -16381, 16384)$. Determine machine precision, the smallest positive number and the largest positive number for this extended precision system.

With $\beta = 2$ and $k = 64$, machine precision with rounding is

$$u = \frac{1}{2} 2^{1-64} = 2^{-64} \approx 5.42 \times 10^{-20}.$$

Accordingly, there are between 19 and 20 significant decimal digits available in this extended precision format. The smallest positive number is

$$(0.1)_2 \times 2^{-16381} = 2^{-16382} \approx 3.36 \times 10^{-4932},$$

while the largest positive number is

$$(0.111\cdots1)_2 \times 2^{16384} = (1-2^{-64})2^{16384} \approx 1.19 \times 10^{4932}.$$

12. IBM System/390 mainframes provide three floating point number systems: short precision $\mathbf{F}(16, 6, -64, 63)$, long precision $\mathbf{F}(16, 14, -64, 63)$ and extended precision $\mathbf{F}(16, 28, -64, 63)$. Compare machine precision, the smallest positive number and the largest positive number for each of these number systems.

In the short precision system $\mathbf{F}(16, 6, -64, 63)$, machine precision with rounding is

$$u = \frac{1}{2}16^{1-6} = 2^{-21} \approx 4.77 \times 10^{-7},$$

while machine precision with rounding in the long precision system $\mathbf{F}(16, 14, -64, 63)$ is

$$u = \frac{1}{2}16^{1-14} = 2^{-53} \approx 1.11 \times 10^{-16}.$$

In the extended precision system $\mathbf{F}(16, 28, -64, 63)$, machine precision with rounding is

$$u = \frac{1}{2}16^{1-28} = 2^{-109} \approx 1.54 \times 10^{-33}.$$

Accordingly, the short precision system provides between 6 and 7 significant decimal digits, the long precision system provides between 15 and 16 significant decimal digits and the extended precision system provides between 32 and 33 significant decimal digits. In all three systems, the smallest positive number is

$$(0.1)_{16} \times 16^{-64} = 16^{-65} \approx 5.40 \times 10^{-79}.$$

The largest positive number in short, long and extended precision is

$$\begin{gathered}(1-16^{-6})\cdot 16^{63},\\ (1-16^{-14})\cdot 16^{63}, \text{ and}\\ (1-16^{-28})\cdot 16^{63},\end{gathered}$$

respectively.

13. A common floating point number system used on modern calculators is $\mathbf{F}(10, 10, -98, 100)$. Determine machine precision, the smallest positive number and the largest positive number for this extended precision system.

With $\beta = 10$ and $k = 10$, machine precision with rounding is

$$u = \frac{1}{2}10^{1-10} = 5 \times 10^{-10}.$$

Accordingly, there are between 9 and 10 significant decimal digits available in $\mathbf{F}(10, 10, -98, 100)$. The smallest positive number is

$$(0.1)_{10} \times 10^{-98} = 10^{-99},$$

while the largest positive number is

$$(1 - 10^{-10})10^{100} = 9.999999999 \times 10^{99}.$$

14. **(a)** Show that the number of elements in the set $\mathbf{F}(\beta, k, m, M)$ is given by $1 + 2(\beta - 1)\beta^{k-1}(M - m + 1)$.

(b) How many elements are in the IEEE standard single precision number system?

(c) How many elements are in the IEEE standard double precision number system?

(a) Let's start by counting the number of positive elements in $\mathbf{F}(\beta, k, m, M)$. The only restriction on the mantissa is that the first digit cannot be zero; hence, there are $(\beta - 1)\beta^{k-1}$ different mantissas that can be formed. With a minimum exponent of m and a maximum exponent of M, there are $M - m + 1$ different exponents. Consequently, there are $(\beta - 1)\beta^{k-1}(M - m + 1)$ positive elements in $\mathbf{F}(\beta, k, m, M)$. Because the number system contains a zero element and equally many negative elements as positive elements, it follows that $\mathbf{F}(\beta, k, m, M)$ contains a total of

$$1 + 2(\beta - 1)\beta^{k-1}(M - m + 1) \text{ elements.}$$

(b) IEEE standard single precision is the system $\mathbf{F}(2, 24, -125, 128)$. Therefore, the IEEE standard single precision number system has

$$1 + 2(2 - 1)2^{24-1}(128 - (-125) + 1) = 4,261,412,865$$

elements.

(c) IEEE standard double precision is the system $\mathbf{F}(2, 53, -1021, 1024)$. Therefore, the IEEE standard single precision number system has

$$\begin{aligned} 1 + 2(2 - 1)2^{53-1}(1024 - (-1021) + 1) &= 18,428,729,675,200,069,633 \\ &\approx 1.84 \times 10^{19} \end{aligned}$$

elements.

15. Consider the function $f(x) = x^2 - 4x + 4$.

(a) What are the zeros of f?

(b) Suppose we were to change the constant term to $4 - 10^{-8}$. What are the zeros of this new function? Relative to the size of the change in the constant term, how big is the change in the zeros of the function?

(c) Now, suppose we were to change the constant term to $4 + 10^{-8}$. What are the zeros of this new function? Relative to the size of the change in the constant term, how big is the change in the zeros of the function?

(a) The function $f(x) = x^2 - 4x + 4 = (x-2)^2$ has a (repeated) zero at $x = 2$.

(b) Now consider the function $f(x) = x^2 - 4x + (4 - 10^{-8})$. By the quadratic formula, the zeros of this new function are

$$\begin{aligned} x &= \frac{4 \pm \sqrt{16 - 4(4 - 10^{-8})}}{2} \\ &= 2 \pm 10^{-4} \\ &= 1.9999, 2.0001 \end{aligned}$$

Observe that the change in the zeros (± 0.0001) is 10,000 times larger than the change in the constant term in the function.

(c) Finally, consider the function $f(x) = x^2 - 4x + (4 + 10^{-8})$. By the quadratic formula, the zeros of this new function are

$$\begin{aligned} x &= \frac{4 \pm \sqrt{16 - 4(4 + 10^{-8})}}{2} \\ &= 2 \pm 0.0001 \cdot \mathrm{i} \end{aligned}$$

Observe that the change in the zeros ($\pm 0.0001 \cdot i$) is 10,000 times larger than the change in the constant term in the function.

16. Consider the linear, first-order differential equation

$$\frac{dx}{dt} + \frac{1}{t}x = \frac{\sin t}{t}.$$

(a) Solve this equation subject to the initial condition $x(\pi/2) = x_0$.

(b) Solve this equation subject to the perturbed initial condition $x(\pi/2) = x_0 + \epsilon$.

(c) By considering the difference between the solutions obtained in parts (a) and (b), comment on the conditioning of this problem.

(a) Multiplying

$$\frac{dx}{dt} + \frac{1}{t}x = \frac{\sin t}{t}$$

by t yields

$$\frac{dx}{dt} + x = \sin t.$$

Note that the terms on the left-hand side of this latter equation are equal to the derivative of the product tx. Integrating both sides of this equation therefore produces

$$tx = -\cos t + C \qquad \text{or} \qquad x(t) = \frac{C - \cos t}{t},$$

where C is a constant of integration. Using the initial condition $x(\pi/2) = x_0$, we determine

$$x_0 = \frac{C - 0}{\pi/2} \qquad \text{or} \qquad C = \frac{\pi x_0}{2}.$$

Hence, the solution of the initial value problem is

$$x(t) = \frac{\pi x_0}{2t} - \frac{\cos t}{t}.$$

(b) The general solution to the differential equation remains

$$x(t) = \frac{C - \cos t}{t},$$

where C is a constant of integration. Using the initial condition $x(\pi/2) = x_0 + \epsilon$, we determine

$$x_0 + \epsilon = \frac{C - 0}{\pi/2} \qquad \text{or} \qquad C = \frac{\pi(x_0 + \epsilon)}{2}.$$

Hence, the solution to the perturbed initial value problem is

$$x(t) = \frac{\pi(x_0 + \epsilon)}{2t} - \frac{\cos t}{t}.$$

(c) The difference between the solutions obtained in parts (a) and (b) is

$$\frac{\pi\epsilon}{2t}.$$

Because this difference decays to zero as $t \to \infty$, this problem is *not* ill-conditioned.

17. Consider the linear, first-order differential equation

$$\frac{dx}{dt} - \frac{1}{t}x = t \sin t.$$

(a) Solve this equation subject to the initial condition $x(\pi/2) = x_0$.

(b) Solve this equation subject to the perturbed initial condition $x(\pi/2) = x_0 + \epsilon$.

(c) By considering the difference between the solutions obtained in parts (a) and (b), comment on the conditioning of this problem.

(a) Multiplying

$$\frac{dx}{dt} - \frac{1}{t}x = t\sin t$$

by t^{-1} yields

$$\frac{dx}{dt} - \frac{1}{t^2}x = \sin t.$$

Note that the terms on the left-hand side of this latter equation are equal to the derivative of the product $t^{-1}x$. Integrating both sides of this equation therefore produces

$$\frac{x}{t} = -\cos t + C \qquad \text{or} \qquad x(t) = t(C - \cos t),$$

where C is a constant of integration. Using the initial condition $x(\pi/2) = x_0$, we determine

$$x_0 = \frac{\pi}{2}(C - 0) \qquad \text{or} \qquad C = \frac{2x_0}{\pi}.$$

Hence, the solution of the initial value problem is

$$x(t) = t\left(\frac{2x_0}{\pi} - \cos t\right).$$

(b) The general solution to the differential equation remains

$$x(t) = t(C - \cos t),$$

where C is a constant of integration. Using the initial condition $x(\pi/2) = x_0 + \epsilon$, we determine

$$x_0 + \epsilon = \frac{\pi}{2}(C - 0) \qquad \text{or} \qquad C = \frac{2(x_0 + \epsilon)}{\pi}.$$

Hence, the solution to the perturbed initial value problem is

$$x(t) = t\left(\frac{2(x_0 + \epsilon)}{\pi} - \cos t\right).$$

(c) The difference between the solutions obtained in parts (a) and (b) is

$$\frac{2\epsilon t}{\pi}.$$

Because this difference tends toward infinity as $t \to \infty$, meaning that a small change in input data can result in a large change in the output, this problem is ill conditioned.

18. Consider the linear system of equations

$$\begin{bmatrix} 1.1 & 2.1 \\ 2 & 3.8 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{b}.$$

(a) Solve the system for the right-hand side vector $\mathbf{b} = \begin{bmatrix} 3.2 & 5.8 \end{bmatrix}^T$.

(b) Solve the system for the right-hand side vector $\mathbf{b} = \begin{bmatrix} 3.21 & 5.79 \end{bmatrix}^T$.

(c) Solve the system for the right-hand side vector $\mathbf{b} = \begin{bmatrix} 3.1 & 5.7 \end{bmatrix}^T$.

(d) By considering the difference between the solutions obtained in parts (a), (b), and (c), comment on the conditioning of this problem.

(a) The system of equations is

$$\begin{aligned} 1.1x + 2.1y &= 3.2 \\ 2x + 3.8y &= 5.8 \end{aligned}$$

If we multiply the first equation by 2 and the second equation by -1.1 and then add, we obtain $0.02y = 0.02$. Thus, $y = 1$. Back substituting into either of the original equations yields $x = 1$.

(b) Working as we did in part (a), we find the solution corresponding to the right-hand side vector $\mathbf{b} = \begin{bmatrix} 3.21 & 5.79 \end{bmatrix}^T$ is $x = -1.95$ and $y = 2.55$.

(c) Working as we did in part (a), we find the solution corresponding to the right-hand side vector $\mathbf{b} = \begin{bmatrix} 3.1 & 5.7 \end{bmatrix}^T$ is $x = 9.5$ and $y = -3.5$.

(d) Given that small changes to the right-hand side vector resulted in relatively large changes to the solution vector, it appears that this problem is ill conditioned