Problem 1: Compute the absolute error and relative error in approximations of $p$ by $p^*$. (Use calculator!)

a) $p = \pi$, $p^* = \frac{22}{7}$;

b) $p = \pi$, $p^* = 3.1416$.

Solution: For this exercise, you can use either calculator or Matlab.

a) Absolute error: $|p - p^*| = |\pi - \frac{22}{7}| = 0.0012645$.

Relative error: $rac{|p - p^*|}{|p|} = \frac{|\pi - \frac{22}{7}|}{\pi} = 4.0250 \times 10^{-4}$.

b) Absolute error: $|p - p^*| = |\pi - 3.1416| = 7.3464 \times 10^{-6}$.

Relative error: $rac{|p - p^*|}{|p|} = \frac{|\pi - 3.1416|}{\pi} = 2.3384 \times 10^{-6}$.

Problem 2: Find the largest interval in which $p^*$ must lie to approximate $\sqrt{2}$ with relative error at most $10^{-5}$ for each value for $p$.

Solution: The relative error is defined as $\frac{|p-p^*|}{|p|}$, where in our case, $p = \sqrt{2}$. We have

$$\frac{|\sqrt{2} - p^*|}{\sqrt{2}} \leq 10^{-5}.$$

Therefore,

$$|\sqrt{2} - p^*| \leq \sqrt{2} \cdot 10^{-5},$$

or

$$-\sqrt{2} \cdot 10^{-5} \leq \sqrt{2} - p^* \leq \sqrt{2} \cdot 10^{-5}.$$

Hence,

$$\sqrt{2} - \sqrt{2} \cdot 10^{-5} \leq p^* \leq \sqrt{2} + \sqrt{2} \cdot 10^{-5}.$$

This interval can be written in decimal notation as $[1.41419942\ldots, 1.41422770\ldots]$. 

✓
Problem 3: Use the 64-bit long real format to find the decimal equivalent of the following floating-point machine numbers.

a) \(0 \ 10000001010 \ 10010011000000 \ldots 0\)

b) \(1 \ 10000001010 \ 01010011000000 \ldots 0\)

Solution:

a) Given a binary number (also known as a machine number)

\[
\begin{array}{c}
0 \\
10000001010 \\
10010011000000 \\
\vdots \\
0
\end{array}
\]

a decimal number (also known as a floating-point decimal number) is of the form:

\[(-1)^s2^{c-1023}(1 + f)\]

Therefore, in order to find a decimal representation of a binary number, we need to find \(s\), \(c\), and \(f\).

The leftmost bit is zero, i.e. \(s = 0\), which indicates that the number is positive.

The next 11 bits, \(10000001010\), giving the characteristic, are equivalent to the decimal number:

\[c = 1 \cdot 2^{10} + 0 \cdot 2^9 + \cdots + 0 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0\]

\[= 1024 + 8 + 2 = 1034.\]

The exponent part of the number is therefore \(2^{1034-1023} = 2^{11}\).

The final 52 bits specify that the mantissa is

\[f = 1 \cdot \left(\frac{1}{2}\right)^1 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^7 + 1 \cdot \left(\frac{1}{2}\right)^8\]

\[= 0.57421875.\]

Therefore, this binary number represents the decimal number

\[(-1)^s2^{c-1023}(1 + f) = (-1)^0 \cdot 2^{1034-1023} \cdot (1 + 0.57421875)\]

\[= 2^{11} \cdot 1.57421875\]

\[= 3224. \checkmark\]
b) Given a binary number
\[
\begin{array}{c|c|c}
\hline
s & 10000001010 & c \\
\hline
f & 01011000000\cdots0 & \\
\hline
\end{array}
\]
a decimal number is of the form:
\[
(-1)^s 2^{c-1023}(1 + f).
\]

Therefore, in order to find a decimal representation of a binary number, we need to find \(s\), \(c\), and \(f\).

The leftmost bit is zero, i.e. \(s = 1\), which indicates that the number is negative.

The next 11 bits, 10000001010, giving the characteristic, are equivalent to the decimal number:
\[
c = 1 \cdot 2^{10} + 0 \cdot 2^9 + \cdots + 0 \cdot 2^4 + 1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 0 \cdot 2^0
\]
\[
= 1024 + 8 + 2 = 1034.
\]

The exponent part of the number is therefore \(2^{1034-1023} = 2^{11}\).

The final 52 bits specify that the mantissa is
\[
f = 1 \cdot \left(\frac{1}{2}\right)^2 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^7 + 1 \cdot \left(\frac{1}{2}\right)^8
\]
\[
= 0.32421875.
\]

Therefore, this binary number represents the decimal number
\[
(-1)^s 2^{c-1023}(1 + f) = (-1)^1 \cdot 2^{1034-1023} \cdot (1 + 0.32421875)
\]
\[
= -2^{11} \cdot 1.32421875
\]
\[
= -2712. \checkmark
\]
Problem 4: Find the next largest and smallest machine numbers in decimal form for the numbers given in the above problem.

Solution:

a) Consider a binary number (also known as a machine number)

\[ 0 \ 10000001010 \ 10010011000000 \ldots \, 00 \]

• The next largest machine number is

\[ 0 \ 10000001010 \ 10010011000000 \ldots \, 01 \].

From problem 3(a), we know that \( s = 0 \) and \( c = 1034 \). We need to find \( f \):

\[
f = 1 \cdot \left( \frac{1}{2} \right)^1 + 1 \cdot \left( \frac{1}{2} \right)^4 + 1 \cdot \left( \frac{1}{2} \right)^7 + 1 \cdot \left( \frac{1}{2} \right)^8 + 1 \cdot \left( \frac{1}{2} \right)^{52} \\
= 0.57421875 + 2.220446049250313 \ldots \times 10^{-16} \\
= 0.57421875 + 0.000000000000002220446\ldots \\
= 0.5742187500000002220446\ldots \\
\]

Therefore, this binary number (in (1)) represents the decimal number

\[
(-1)^s 2^{c-1023}(1 + f) = (-1)^0 \cdot 2^{1034-1023} \cdot (1 + 0.57421875 + 0.000000000000002220446\ldots) \\
= 2^{11} \cdot (1.57421875 + 2.220446049250313 \ldots \times 10^{-16}) \\
= 3224 + 4.547473508864641 \times 10^{-13} \\
= 3224.0000000000045474735\ldots \, ✓
\]

• The next smallest machine number is

\[ 0 \ 10000001010 \ 10010010111111 \ldots \, 11 \].

From problem 3(a), we know that \( s = 0 \) and \( c = 1034 \). We need to find \( f \):

\[
f = 1 \cdot \left( \frac{1}{2} \right)^1 + 1 \cdot \left( \frac{1}{2} \right)^4 + 1 \cdot \left( \frac{1}{2} \right)^7 + \sum_{n=9}^{52} 1 \cdot \left( \frac{1}{2} \right)^n \\
= \left( \frac{1}{2} \right)^1 + \left( \frac{1}{2} \right)^4 + \left( \frac{1}{2} \right)^7 + \left( \frac{1}{2} \right)^8 - \left( \frac{1}{2} \right)^{52} \\
= 0.57421875 - 2.220446049250313\ldots \times 10^{-16} \\
= 0.57421875 - 0.000000000000002220446\ldots \\
= 0.57421874999999977795539\ldots
\]

\[\text{Note that} \sum_{n=0}^{N} 2^n = 2^{N+1} - 1.\]

The formula above is a specific case of the following more general equation:

\[
\sum_{n=M}^{N} 2^n = 2^{N+1} - 2^M.
\]

Similarly, we also have a formula:

\[
\sum_{n=M}^{N} \left( \frac{1}{2} \right)^n = \left( \frac{1}{2} \right)^{M-1} - \left( \frac{1}{2} \right)^N.
\]

To get some intuition about these formulas, consider an example with \( M = 2 \) and \( N = 5 \), for instance.
Therefore, this binary number (in (2)) represents the decimal number
\[
(-1)^s2^{c-1023}(1 + f) = (-1)^0 \cdot 2^{1034-1023} \cdot (1 + 0.57421875 - 0.0000000000000220446\ldots)
\]
\[
= 2^{11} \cdot (1.57421875 - 2.220446\ldots \times 10^{-16})
\]
\[
= 3224 - 4.547473508 \times 10^{-13}
\]
\[
= 3224 - 0.0000000000004547473508
\]
\[
= 3223.999999999995452527\ldots \checkmark
\]

b) Consider a binary number
\[
1\ 1000001010\ 01010011000000\ \cdots \ 0
\]
• The next largest (in magnitude) machine number is
\[
1\ 1000001010\ 01010011000000\ \cdots \ 1
\]
From problem 3(b), we know that \(s = 1\) and \(c = 1034\). We need to find \(f\):
\[
f = 1 \cdot \left(\frac{1}{2}\right)^2 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^7 + \sum_{n=9}^{52} 1 \cdot \left(\frac{1}{2}\right)^n
\]
\[
= 0.32421875 + 2.220446049250313\ldots \times 10^{-16}
\]
\[
= 0.32421875 + 0.0000000000000220446\ldots
\]
\[
= 0.324218750000000220446\ldots
\]
Therefore, this binary number (in (3)) represents the decimal number
\[
(-1)^s2^{c-1023}(1 + f) = (-1)^1 \cdot 2^{1034-1023} \cdot (1 + 0.32421875 + 0.000000000000000220446\ldots)
\]
\[
= -2^{11} \cdot (1.32421875 + 2.220446049250313\ldots \times 10^{-16})
\]
\[
= -2712 - 4.547473508864641 \times 10^{-13}
\]
\[
= -2712 - 0.0000000000004547473508
\]
\[
= -2712.00000000000045474735\ldots \checkmark
\]
• The next smallest (in magnitude) machine number is
\[
1\ 1000001010\ 01010011000000\ \cdots \ 1
\]
From problem 3(b), we know that \(s = 1\) and \(c = 1034\). We need to find \(f\):
\[
f = 1 \cdot \left(\frac{1}{2}\right)^2 + 1 \cdot \left(\frac{1}{2}\right)^4 + 1 \cdot \left(\frac{1}{2}\right)^7 + \sum_{n=9}^{52} 1 \cdot \left(\frac{1}{2}\right)^n
\]
\[
= \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 + \left(\frac{1}{2}\right)^7 + \left(\frac{1}{2}\right)^8 - \left(\frac{1}{2}\right)^{52}
\]
\[
= 0.32421875 - 2.220446049250313\ldots \times 10^{-16}
\]
\[
= 0.32421875 - 0.0000000000000220446\ldots
\]
\[
= 0.32421874999999977795539\ldots
\]
Therefore, this binary number (in (4)) represents the decimal number
\[
(-1)^s2^{c-1023}(1 + f) = (-1)^1 \cdot 2^{1034-1023} \cdot (1 + 0.32421875 - 0.0000000000000220446\ldots)
\]
\[
= -2^{11} \cdot (1.32421875 - 2.220446049250313\ldots \times 10^{-16})
\]
\[
= -2712 + 4.54747308864641 \times 10^{-13}
\]
\[
= -2712 + 0.0000000000004547473508
\]
\[
= -2711.999999999995452527\ldots \checkmark
\]
Problem 5: Use four-digit rounding arithmetic and the formulas to find the most accurate approximations to the roots of the following quadratic equations. Compute the relative error.

a) \( \frac{1}{3} x^2 - \frac{123}{4} x + \frac{1}{6} = 0 \);

b) \( 1.002x^2 + 11.01x + 0.01265 = 0 \).

Solution: The quadratic formula states that the roots of \( ax^2 + bx + c = 0 \) are

\[
 x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. 
\]

a) The roots of \( \frac{1}{3} x^2 - \frac{123}{4} x + \frac{1}{6} = 0 \) are approximately

\[
 x_1 = 92.24457962731231, \quad x_2 = 0.00542037268770.
\]

We use four-digit rounding arithmetic to find approximations to the roots. We find the first root:

\[
 x_1^* = \frac{\frac{123}{4} + \sqrt{\left( -\frac{123}{4} \right)^2 - 4 \cdot \frac{1}{3} \cdot \frac{1}{6}}}{2 \cdot \frac{1}{3}} = \frac{30.75 + \sqrt{30.75^2 - 4 \cdot 0.3333 \cdot 0.1667}}{2 \cdot 0.3333} = \frac{30.75 + \sqrt{945.6 - 1.333 \cdot 0.1667}}{0.6666} = \frac{30.75 + 945.4}{0.6666} = \frac{30.75 + 30.75}{0.6666} = 61.50, \quad 92.26, \quad \checkmark
\]

which has the following relative error:

\[
 \frac{|x_1 - x_1^*|}{|x_1|} = \frac{|92.24457962731231 - 92.26|}{92.24457962731231} = 1.672 \cdot 10^{-4}, \quad \checkmark
\]

b) The roots of \( 1.002x^2 + 11.01x + 0.01265 = 0 \) have the following relative error:

\[
 |x_1 - x_1^*| = |0.00542037268770 - 0| = 0.00542037268770 = 1.0.
\]

We obtained a very large relative error, since the calculation for \( x_2^* \) involved the subtraction of nearly equal numbers. In order to get a more accurate approximation to \( x_2^* \), we need to use an alternate quadratic formula, namely

\[
 x_{1,2} = \frac{-2c}{b \pm \sqrt{b^2 - 4ac}}.
\]

Using four-digit rounding arithmetic, we obtain:

\[
 x_2^* = \frac{-2 \cdot \frac{1}{6}}{-\frac{123}{4} - \sqrt{\left( -\frac{123}{4} \right)^2 - 4 \cdot \frac{1}{3} \cdot \frac{1}{6}}} = fl(0.00541951) = 0.005420, \quad \checkmark
\]

which has the following relative error:

\[
 \frac{|x_2 - x_2^*|}{|x_2|} = \frac{|0.00542037268770 - 0.005420|}{0.00542037268770} = 6.876 \cdot 10^{-5}, \quad \checkmark
\]
b) The roots of \( 1.002x^2 + 11.01x + 0.01265 = 0 \) are approximately

\[
x_1 = -0.00114907565991, \quad x_2 = -10.98687487643590.
\]

We use four-digit rounding arithmetic to find approximations to the roots.

If we use the generic quadratic formula for the calculation of \( x_1^* \), we will encounter the subtraction of nearly equal numbers (you may check). Therefore, we use the alternate quadratic formula to find \( x_1^* \):

\[
x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}} = \frac{-2 \cdot 0.01265}{11.01 - \sqrt{11.01^2 + 4 \cdot 1.002 \cdot 0.01265}} = \frac{-2 \cdot 0.01265}{22.01} = -0.001149, \quad \checkmark
\]

which has the following relative error:

\[
\frac{|x_1 - x_1^*|}{|x_1|} = \left| \frac{-0.00114907565991 - (-0.001149)}{-0.00114907565991} \right| = 6.584 \cdot 10^{-5}. \quad \checkmark
\]

We find the second root using the generic quadratic formula:

\[
x_2^* = \frac{-11.01 - \sqrt{(-11.01)^2 - 4 \cdot 1.002 \cdot 0.01265}}{2 \cdot 1.002} = \frac{-11.01 - \sqrt{121.2 - 0.05070}}{2.004} = \frac{-11.01 - \sqrt{121.1}}{2.004} = \frac{-11.01 - 11.00}{2.004} = -10.98, \quad \checkmark
\]

which has the following relative error:

\[
\frac{|x_2 - x_2^*|}{|x_2|} = \left| \frac{-10.98687487643590 - (-10.98)}{-10.98687487643590} \right| = 6.257 \cdot 10^{-4}. \quad \checkmark
\]
Similar Problem

The roots of \( 1.002x^2 - 11.01x + 0.01265 = 0 \) are approximately

\[ x_1 = 10.986887487643590, \quad x_2 = 0.00114907565991. \]

We use four-digit rounding arithmetic to find approximations to the roots. We find the first root:

\[
x_1^* = \frac{11.01 + \sqrt{(-11.01)^2 - 4 \cdot 1.002 \cdot 0.01265}}{2 \cdot 1.002} = \frac{11.01 + \sqrt{121.2 - 4.008 \cdot 0.01265}}{2.004}
\]

\[
= \frac{11.01 + \sqrt{121.2 - 0.50070}}{2.004} = \frac{11.01 + \sqrt{121.1}}{2.004} = \frac{11.01 + 11.00}{2.004}
\]

\[
= \frac{22.01}{2.004} = 10.98, \quad \checkmark
\]

which has the following relative error:

\[
\frac{|x_1 - x_1^*|}{|x_1|} = \frac{|10.986887487643590 - 10.98|}{10.98687} = 6.257 \cdot 10^{-4}, \quad \checkmark
\]

If we use the generic quadratic formula for the calculation of \( x_2^* \), we will encounter the subtraction of nearly equal numbers. Therefore, we use the alternate quadratic formula to find \( x_2^* \):

\[
x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}} = \frac{-2 \cdot 0.01265}{-11.01 - \sqrt{(-11.01)^2 - 4 \cdot 1.002 \cdot 0.01265}}
\]

\[
= \frac{-0.02530}{-11.01 - 11.00} = \frac{-0.02530}{-22.01} = 0.001149, \quad \checkmark
\]

which has the following relative error:

\[
\frac{|x_2 - x_2^*|}{|x_2|} = \frac{|0.00114907565991 - 0.001149|}{0.00114907565991} = 6.584 \cdot 10^{-5}, \quad \checkmark
\]
Problem 6: Suppose that \( fl(y) \) is a \( k \)-digit rounding approximation to \( y \). Show that
\[
\left| \frac{y - fl(y)}{y} \right| \leq 0.5 \times 10^{-k+1}.
\]
(Hint: If \( d_{k+1} < 5 \), then \( fl(y) = 0.d_1 \ldots d_k \times 10^n \).
If \( d_{k+1} \geq 5 \), then \( fl(y) = 0.d_1 \ldots d_k \times 10^n + 10^{n-k} \).

Solution: We have to look at two cases separately.

Case ①: \( d_{k+1} < 5 \).

\[
\left| \frac{y - fl(y)}{y} \right| = \left| \frac{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n - 0.d_1 \ldots d_k \times 10^n}{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n} \right| = \left| \frac{0.0 \ldots 0 d_{k+1} \ldots \times 10^n}{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n} \right| = \frac{0.d_{k+1}d_{k+2} \ldots \times 10^{-k}}{0.d_1d_2 \ldots \times 10^n} \times 10^{-k} \leq \frac{0.d_{k+1}d_{k+2} \ldots \times 10^{-k}}{0.1} \times 10^{-k} \quad \text{since} \ d_1 \geq 1, \ \text{so} \ |0.d_1d_2\ldots| \geq 0.1
\]
\[
\leq \frac{0.5}{0.1} \times 10^{-k} \quad \text{since} \ d_{k+1} \leq 5, \ \text{by assumption}
\]
\[
= 5 \times 10^{-k} = 0.5 \times 10^{-k+1}. \quad \checkmark
\]

Case ②: \( d_{k+1} \geq 5 \).

\[
\left| \frac{y - fl(y)}{y} \right| = \left| \frac{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n - (0.d_1 \ldots d_k \times 10^n + 10^{n-k})}{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n} \right| = \left| \frac{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n - (0.d_1 \ldots d_k + 10^{-k}) \times 10^n}{0.d_1d_2 \ldots \times 10^n} \right| = \frac{0.d_1 \ldots d_k d_{k+1} \ldots \times 10^n - 0.d_1 \ldots \delta_k \times 10^n}{0.d_1d_2 \ldots \times 10^n} \quad \text{where} \ \delta_k = d_k + 1
\]
\[
= \left| \frac{0.d_1 \ldots d_k d_{k+1} \ldots - 0.d_1 \ldots \delta_k}{0.d_1d_2 \ldots} \right| = \frac{|0.0 \ldots 0 d_k d_{k+1} \ldots - 0.0 \ldots 0 \delta_k|}{0.d_1d_2 \ldots} \quad \text{Note that} \ 0.0 \ldots 0 \delta_k > 0.0 \ldots 0 d_k d_{k+1} \ldots,
\]
and \( \delta_k = d_k + 1, \ d_{k+1} \geq 5 \).

For example, \( d_k = 1, \ d_{k+1} = 6, \ \delta_k = 2 \).

\[
\leq \frac{0.0 \ldots 0.05}{0.d_1d_2 \ldots} = \frac{0.5}{0.d_1d_2 \ldots} \times 10^{-k} \quad \text{since} \ d_1 \geq 1, \ \text{so} \ 0.d_1d_2\ldots \geq 0.1
\]
\[
\leq \frac{0.5}{0.1} \times 10^{-k} = 5 \times 10^{-k} = 0.5 \times 10^{-k+1}. \quad \checkmark
\]