1. Determine the form of the single precision floating point arithmetic used in the computers at AIMS. What is the largest number that can be accurately represented? What is the smallest positive number $n_1$? The second smallest positive number $n_2$? Which is larger: the gap between $n_1$ and 0 or the gap between $n_1$ and $n_2$? Discuss.

Solution: According to the IEEE standard, double precision floating point numbers (which is the PYTHON standard form) have 53 bits in their mantissa, of which the leading 1 is not needed so only 52 bits are stored, plus 1 sign bit, leaving 11 bits for the exponent. The range of $2^{11} = 2048$ exponents is, in fact, slightly biased in one direction, and the exponent resides in the range $-1021$ to 1024. (This is done by subtracting 1021 from the exponent bits.)

Thus, the smallest number is $n_1 = .10000\ldots \times 2^{-1021} \approx 4.4501477 \times 10^{-309}$. The next smallest number has a 1 in the 53-rd binary place, and so differs from the smallest by $n_2 - n_1 = 2^{-53} \times 2^{-1021} = 2^{-1074} \approx 4.940656 \times 10^{-324}$. The ratio between the two gaps is huge: $\frac{n_1}{n_2 - n_1} = 2^{53} \approx 9.0072 \times 10^{15}$.

For a detailed discussion of how numbers are represented in floating point, see

http://stevehollasch.com/cgindex/coding/ieeefloat.html

2. Determine the value of each of the following quantities using 4 digit rounding and four digit chopping arithmetic. Find the absolute and relative errors of your approximation. (a) $\pi + e - \cos 22^\circ$, (b) $\frac{e^\pi - \pi^e}{\log_{10} 11}$.

Solution:

(a) The exact value is 4.93269062748205. With 4 digit rounding:

$$\pi + e - \cos(22\,\pi/180) = 3.142 + 2.718 - \cos(69.12/180) = 5.860 - \cos(.3840)$$

$$= 5.860 - .9272 = 4.933,$$

\[\text{\tiny† Although this is slightly cheating, since we used the high precision evaluation of the cosine function in this computation.}\]
with absolute error $3.094 \times 10^{-4}$ and relative error $6.272 \times 10^{-5}$.

With 4 digit chopping:

$$
\pi + e - \cos(22 \pi/180) = 3.141 + 2.718 - \cos(69.10/180) = 5.859 - \cos(.3838)
$$

$$
= 5.859 - .9272 = 4.931,
$$

with absolute error $1.691 \times 10^{-3}$ and relative error $3.427 \times 10^{-4}$.

(b) The exact value is $4.93269062748205$. With 4 digit rounding:

$$
\frac{e^{\pi} - \pi e}{\log_{10} 11} = \frac{23.14 - 22.46}{\log .9091} = \frac{.6800}{-.09530} = -7.135,
$$

with absolute error $1.570 \times 10^{-2}$ and relative error $2.196 \times 10^{-3}$.

With 4 digit chopping:

$$
\frac{e^{\pi} - \pi e}{\log_{10} 11} = \frac{23.14 - 22.45}{\log .9090} = \frac{.6900}{-.09541} = -7.231,
$$

with absolute error $8.030 \times 10^{-2}$ and relative error $1.123 \times 10^{-2}$.

3. (a) To how many significant decimal digits do the numbers $\sqrt{10002}$ and $\sqrt{10001}$ agree? (b) Subtract the two numbers. How many significant decimal digits are lost in the computation? (c) How might you rearrange the computation to obtain a more accurate answer.

Solution: (a) 5 significant digits; (b) 5 lost digits. (c) Use the identity $\sqrt{x} - \sqrt{y} = \frac{x - y}{\sqrt{x} + \sqrt{y}}$, which, in this case, is $\sqrt{10002} - \sqrt{10001} = \frac{1}{\sqrt{10002} + \sqrt{10001}}$.

4. (a) Verify that $f(x) = 1 - \sin x$ and $g(x) = \frac{\cos^2 x}{1 + \sin x}$ are identical functions.

(b) Which function should be used for computations when $x$ is near $\frac{1}{2} \pi$? Why?

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

Solution: (a) $\frac{\cos^2 x}{1 + \sin x} = \frac{1 - \sin^2 x}{1 + \sin x} = 1 - \sin x$. (b) $g(x)$ since $f(x)$ requires subtracting two nearly equal numbers; (c) $f(x)$ since $g(x)$ requires dividing two very small numbers.

\[\text{†}\] Again we are cheating by using high precision exponentiation and logarithms in the computation.
5. Horner’s Method

(a) Suppose $x$ is a real number and $n$ a positive integer. How many multiplications are need to efficiently compute $x^n$? Hint: The answer is not $n - 1$.

(b) Verify the polynomial identity

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = a_0 + x(a_1 + x(a_2 + x(\cdots + x a_n)))$$

Explain why the right hand side is to be preferred when computing the values of the polynomial $p(x)$.

Solution:

(a) We can iteratively compute $x^{2^k} = x^{2^{k-1}} \cdot x^{2^{k-1}}$ by $k$ multiplications. More generally, if $2^k \leq n < 2^{k+1}$, we write $x^n = x^{2^k} \cdot x^m$ where $m < 2^k$ and iteratively evaluate $x^m$ by the same method. The net result is that it requires $k + l - 1$ multiplications, where $l$ is the number of 1’s in the binary expansion of $n$. For example, $x^{43} = x^{32} \cdot x^8 \cdot x^2 \cdot x$ takes $5 + 4 - 1 = 8$ multiplications, where $43 = 101011$ in binary has 4 ones.

(b) The identity is easily established by mutliplying out the terms. The right hand side only requires $n - 1$ multiplications, while the left hand side requires considerably more, even if the efficient algorithm of part (a) is used to evaluated the powers.

6. Let

$$f(x) = e^x - \cos x - x.$$

(a) Using calculus, what should the graph of $f(x)$ look like for $x$ near 0?

(b) Using both single and double precision arithmetic, graph $f(x)$ for $|x| \leq 5 \times 10^{-8}$ and discuss what you observe.

(c) How might you obtain a more realistic graph?

Solution:

(a) Smooth, with a local minimum at 0 and convex.
(c) Use the first few terms in the Taylor expansion:

\[ e^x - \cos x - x = x^2 + \frac{1}{6} x^3 + \frac{1}{120} x^5 + \cdots. \]

7. Consider the linear system of equations

\[
\begin{align*}
1.1x + 2.1y &= a, \\
2x + 3.8y &= b.
\end{align*}
\]

Solve the system for the following right hand sides: (i) \(a = 3.2, \ b = 5.8\); (ii) \(a = 3.21, \ b = 5.79\); (iii) \(a = 3.1, \ b = 5.7\). Discuss the conditioning of this system of equations.

**Solution:** The respective solutions are (i) \(x = 1, \ y = 1\); (ii) \(x = -1.95, \ y = 2.55\); (iii) \(x = 9.5, \ y = -3.5\). The rapid change in the solution under small changes in the right hand sides indicates that this system is ill-conditioned. Further evidence comes from the fact that the determinant of the coefficient matrix is \(-.02\).