Floating Point Representations and Errors

Error Propagation.

1) Compute the following sums on your computer:

$$a)\ \sum_{i=1}^{10000} 0.0001$$

$$b)\ \sum_{i=1}^{100000} 0.000001$$

$$c)\ \sum_{i=1}^{1000000} 0.0000001$$

Subtract exactly 1 from each sum to show the round-off error effect. Explain your results.

2) Compute the following sums:

$$a)\ \sum_{i=1}^{32767} \left(\frac{1}{i^n}\right)$$

$$b)\ \sum_{i=32767}^{1} \left(\frac{1}{i^n}\right)$$

for $n = 1,2,$ and $3.$ Are the values for the two sums done in the opposite direction the same? If not, why not? Which one of the two ways of performing the addition is more precise?

Repeat the experiment computing this time

$$a)\ \prod_{i=1}^{1000} \left(\frac{i}{10}\right)^n$$

$$b)\ \prod_{i=1000}^{1} \left(\frac{i}{10}\right)^n$$

3) When adding very small quantities with a computer, you may eventually have the result $x = x + a.$ Which is the value of n which satisfies the relation

$$1 + 2^{-n} = 1$$

in your computer?

4) Compute the values

$$a(k) = 10^k + 0.3$$

$$b(k) = 10^k + 0.1 + 0.2$$

for $k = 1,2,\ldots,10.$ Compare the absolute errors and the relative errors for each value of $k.$

5) We can compute $f(x) = e^{-x}$ using Taylor polynomials in two ways, either using:

$$e^{-x} = 1 - x + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \ldots$$

or using

$$e^{-x} = \frac{1}{1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \ldots}$$

Use these expressions to obtain $e^{-0.1}$. Which of the two values is the best one?

6) The function $f_1(x,\delta) = \cos(x+\delta) - \cos(x)$ can be transformed into another form, $f_2(x,\delta)$, using the trigonometric formula

$$\cos\phi - \cos\psi = -2\sin\left(\frac{\phi+\psi}{2}\right)\sin\left(\frac{\phi-\psi}{2}\right).$$

Thus, f_1 and f_2 have the same values, in exact arithmetic, for any given argument values x and δ.

a) Show that, analytically, $f_1(x,\delta)/\delta$ or $f_2(x,\delta)/\delta$ are effective approximations of the function $-\sin(x)$ for δ sufficiently small.

b) Derive $f_2(x,\delta)$.

c) Write a MATLAB script which will calculate

$$g_1(x,\delta) = f_1(x,\delta) + \sin(x)$$
$$g_2(x,\delta) = f_2(x,\delta) + \sin(x)$$

for $x = 3$ and $\delta = 10^{-11}$.

d) Explain the difference in the results of the two calculations.

7) One approximation for π consists of inscribing polygons in a circle of radius $1/2$, starting with hexagons, and successively doubling the number of sides. The recurrence formula for the perimeters of such polygons is

$$p_{n+1} = 6 \cdot 2^n \cdot x_n$$

where

$$x_1 = \frac{1}{\sqrt{3}}, \qquad x_{n+1} = \frac{\sqrt{x_n^2+1} - 1}{x_n}, \qquad n = 2,3,4,\ldots$$

Observe that we can rewrite the preceding expression in the mathematically equivalent form:

$$x_1 = \frac{1}{\sqrt{3}}, \qquad x_{n+1} = \frac{x_n}{\sqrt{x_n^2+1} + 1}, \qquad n = 2,3,4,\ldots$$

Compute the first 30 values of these recurrences and compare the results.

8) Use the random number generator to obtain a list of real numbers $a_1, a_2, \ldots, a_n$ and compute the following values for this set:

Arithmetic Mean	$m = \frac{(a_1 + a_2 + \cdots + a_n)}{n}$
Variance	$v = \frac{(a_1 - m)^2 + \cdots + (a_n - m)^2}{n}$
Standard Deviation	$\sigma = \sqrt{n}$

Use a set of values with a very short range and a set of values with a very wide

range. Obtain the following alternative formula to compute the variance and compare the results of both expressions.

$$v = \frac{1}{n}\sum_{i=1}^{n} a_i^2 - m^2$$

Use a random set of numbers and a large n, say $n = 10000$

1) Write a quadratic equation solver.

$$ax^2 + bx + c = 0$$

Your MATLAB script should get a, b, c as input, and accurately compute the roots of the corresponding quadratic equation. Make sure to check end cases such as $a = 0$, and consider ways to avoid an overflow and cancellation errors. Implement your algorithm and demonstrate its performance on the following cases:

a) $a = 1;\ b = -10^5;\ c = 1.$

b) $a = 6 \cdot 10^{30};\ b = 5 \cdot 10^{30};\ c = -4 \cdot 10^{30}.$

c) $a = 10^{-30};\ b = -10^{30};\ c = 10^{30}.$

Show that your algorithm produces better results than the standard formula for computing roots of a quadratic equation.

9) Let $p = 1/2.$ Consider the mathematical equivalent sums

$$\begin{aligned} 1 &= \sum_{k\geq 1} \frac{1}{k^p} - \frac{1}{(k+1)^p} \\ &= \sum_{k\geq 1} \frac{(k+1)^p - k^p}{k^p (k+1)^p} \\ &= \sum_{k\geq 1} \frac{1}{k^p (k+1)^p ((k+1)^p + k^p)} \end{aligned}$$

Which of these is the most accurate to evaluate in floating-point using naive recursive summation? Why?

10) Evaluate the function

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

precisely within the interval $[-5 \times 10^{-8}, 5 \times 10^{-8}]$. Compare the values obtained with the direct expression with the values obtained with the Taylor expansion.

Recurrences

1) Show that the recurrence

$$\begin{cases} x_0 = 1, \qquad x_1 = \frac{1}{3} \\ x_{n+1} = \frac{13}{3} x_n - \frac{4}{3} x_{n-1} \qquad (n \geq 1) \end{cases}$$

generates the sequence:

$$x_n = \left(\frac{1}{3}\right)^n$$

Use this recurrence to compute the 30 first values and compare your results with the expected values.

2) Show that the recurrence relation

$$x_n = 2x_{n-1} + x_{n-2}$$

has a general solution of the form

$$x_n = A\lambda^n + B\mu^n.$$

Is this recurrence relation a good way to compute x_n from arbitrary initial values x_0 and x_1? Compute the first 30 values when $x_0 = 1$ and $x_1 = 1$.

3) Consider the recurrence relation

$$x_n = 2(x_{n-1} + x_{n-2}).$$

Show that the general solution is

$$x_n = \alpha\left(1 + \sqrt{3}\right)^n + \beta\left(1 - \sqrt{3}\right)^n$$

Show that the solution with starting values $x_1 = 1$ and $x_2 = 1 - \sqrt{3}$ corresponds to $\alpha = 0$ and $\beta = \left(1 - \sqrt{3}\right)^{-1}$

Compute the particular solution in the preceeding problem in the following three ways

a) x_n directly from the recurrence relation

b) $y_n = \beta\left(1 - \sqrt{3}\right)^n$

c) $z_n = \alpha\left(1 + \sqrt{3}\right)^n + \beta\left(1 - \sqrt{3}\right)^n$, where α is chosen as the smallest roundoff error of your computer. This is, if your work with double precision $\alpha = 10^{-16}$. (MATLAB normally works with a precision of $\alpha = 10^{-16}$).

4) Find a recursive relation to compute the integral values

$$I_k = \int_0^1 x^k \sin(\pi x)dx, \qquad (k = 2,4,\ldots,40,\ldots)$$

Is this recursive relation stable?

5) Prove that the values

$$p_n = \int_0^1 x^n e^x dx$$

satisfy the relation $p_1 > p_2 > \cdots > 0$. Show that

$$p_{n+1} = e - (n+1)p_n.$$

Use this recursion to compute $p_2,\ldots,p_{20}$ from $p_1 = 1$. Explain the results.

6) Find a recurrence relation to obtain the values of the integrals

$$I_k = \int_0^1 \frac{x^k}{x+10}dx \qquad k = 1,\ldots,20$$

Study how to compute these integral values with a precision of $\epsilon < 1/2 \times 10^{-8}$.

7) Define

$$x_n = \int_0^1 t^n (t+5)^{-1} dt$$

Show that $x_0 = \ln 1.2$ and that $x_n = n^{-1} - 5x_{n-1}$ for $n \geq 1$. Compute $x_0, x_1, \ldots, x_{10}$ and estimate the accuracy of x_{10}.

Using the approximate value $x_{20} \approx 7.997523 \times 10^{-3}$, use the recurrence backward to get $x_{19}, x_{18}, \ldots, x_0$. Is x_0 correct? What about the other x_n? Does the recurrence relation behave differently when used backward, and if so, why?

8) The Exponential Integrals are the functions E_n defined by

$$E_n(x) = \int_1^\infty (e^{xt}t^n)^{-1} dt \qquad (n \geq 0,\ x > 0).$$

Show that these functions satisfy the equation

$$nE_{n+1}(x) = e^{-x} - xE_n(x).$$

Is $E_n(x)$ an increasing or decreasing function with n? Suppose that you know $E_1(x)$. Can this equation be used to compute $E_2(x), E_3(x), \ldots$ accurately? Compare these values with the expression

$$E_n(x) = x^{n-1}\Gamma(1-n,x)$$

where $\Gamma(1-n,x)$ is the incomplete Gamma function.