PROBLEM SET 2.1

1. What is the unit roundoff error for a binary machine carrying 48-bit mantissas?
2. If the **MARC-32** did not round off numbers correctly but simply dropped excess bits, what would the unit roundoff be?
3. What is the unit roundoff error for a *decimal* machine that allocates 12 decimal places to the mantissa? Such a machine stores numbers in the form $x = \pm r \times 10^n$ with $1/10 \leqq r < 1$.
4. Prove that 4/5 is not representable exactly on the **MARC-32**. What is the closest machine number? What is the relative roundoff error involved in storing this number on the **MARC-32**?
5. What numbers are representable with a finite expression in the binary system but are not finitely representable in the decimal system?
6. What can be said of the relative roundoff error in adding n machine numbers? (Make no assumption about the numbers being positive, as this case is covered by a theorem in the text.)
7. Find a real number x in the range of the **MARC-32** such that $fl(x) = x(1+\delta)$ with $|\delta|$ as large as possible. Can the bound 2^{-24} be attained by $|\delta|$?
8. Show that, under the assumptions made about the **MARC-32**, we shall have $fl(x) = x/(1+\delta)$ with $|\delta| \leqq 2^{-24}$.
9. Show that $fl(x^k) = x^k(1+\delta)^{k-1}$ with $|\delta| \leqq \varepsilon$, if x is a floating-point machine number in a computer with unit roundoff ε.
10. Show by examples that often $fl[fl(xy)z] \neq fl[xfl(yz)]$ for machine numbers x,y, and z. This phenomenon is often described informally by saying *machine multiplication is not associative*.
11. Prove that if x and y are machine numbers in the **MARC-32**, and if $|y| \leqq |x|2^{-25}$, then $fl(x+y) = x$.
12. If 1/10 is correctly rounded to the normalized binary number $(.a_1a_2...a_{24})_2 \times 2^m$, what is the roundoff error? What is the relative roundoff error?
13. (a) If 3/5 is correctly rounded to the binary number $(.a_1a_2\ldots a_{24})_2$, what is the roundoff error?

    (b) Answer the same question for the number 2/7.
14. Is $\frac{2}{3}(1-2^{-24})$ a machine number in the **MARC-32**? Explain.
15. Let $x_1, x_2, \ldots, x_n$ be positive machine numbers in the **MARC-32**. Let S_n denote the sum $x_1 + x_2 + \cdots + x_n$, and let S_n^* be the corresponding sum in the computer. (Assume that the addition is carried out in the order given.) Prove the following: If $x_{i+1} \geqq 2^{-24}S_i$ for each i, then

$$|S_n^* - S_n|/S_n \leqq (n-1)2^{-24}$$

16. Prove this slight improvement of Inequality (7)

$$\left|\frac{x - x^*}{x}\right| < \frac{1}{1+2^{24}}$$

for the representation of numbers in the **MARC-32**.

17. How many normalized machine numbers are there in the **MARC-32**? (Do not count zero.)
18. Prove that each machine number in the **MARC-32** has a unique normalized representation, but that this would not be true without the assumption of normalization.
19. Let $x = (0.111\ldots111000\ldots)_2 \times 2^{17}$, in which the fractional part has 26 ones followed by zeros. For the **MARC-32**, determine $x', x'', fl(x), x - x', x'' - x, x'' - x'$, and $|x - fl(x)|/x$.
20. Let $x = 2^3 + 2^{-19} + 2^{-22}$. Find the machine numbers on the **MARC-32** that are just to the right and just to the left of x. Determine $fl(x)$, the absolute error $|x - fl(x)|$, and the relative error $|x - fl(x)|/|x|$. Verify that the relative error in this case does not exceed 2^{-24}.
21. Find the machine number just to the right of 1/9 in a binary computer with a 43-bit normalized mantissa.
22. What is the exact value of $x^* - x$, if $x = \sum_{n=1}^{26} 2^{-n}$ and x^* is the nearest machine number on the **MARC-32**?
23. Let $S_n = x_1 + x_2 + \cdots + x_n$, where each x_i is a machine number. Let S_n^* be what the machine computes. Then $S_n^* = fl(S_{n-1}^* + x_n)$. Prove that on the **MARC-32**

$$S_n^* \approx S_n + S_2\delta_2 + \cdots + S_n\delta_n \qquad |\delta_k| \leqq 2^{-24}$$

24. Which of these is not necessarily true on the **MARC-32**? (Here x, y, z are machine numbers and $|\delta| \leqq 2^{-24}$.)
    (a) $fl(xy) = xy(1+\delta)$
    (b) $fl(x+y) = (x+y)(1+\delta)$
    (c) $fl(xy) = xy/(1+\delta)$
    (d) $|fl(xy) - xy| \leqq |xy|2^{-24}$
    (e) $fl(x+y+z) = (x+y+z)(1+\delta)$
25. Use the **MARC-32** for this problem. Determine a bound on the relative error in computing $(a+b)/(c+d)$ for machine numbers a, b, c, d.
26. Which of these is a machine number on the **MARC-32**?
    (i) 10^{40} (ii) $2^{-1} + 2^{-26}$ (iii) 1/5 (iv) 1/3 (v) 1/256
27. Let $x = 2^{16} + 2^{-8} + 2^{-9} + 2^{-10}$. Let x^* be the machine number closest to x in the **MARC-32**. What is $x - x^*$?
28. Criticize the following argument: In combining two machine numbers arithmetically in the **MARC-32**, the relative roundoff error cannot exceed 2^{-24}. Therefore, in combining n such numbers, the relative roundoff error cannot exceed $(n-1)2^{-24}$.
29. Let $x = 2^{12} + 2^{-12}$.
    (a) Find the machine numbers x' and x'' in the **MARC-32** that are just to the left and right of x, respectively.
    (b) For this number show that the relative error between x and $fl(x)$ is no greater than the unit roundoff error in the **MARC-32**.
30. What relative roundoff error is possible in computing the product of n machine numbers in the **MARC-32**? How is your answer changed if the n numbers are not necessarily machine numbers (but are within the range of the machine)?
31. Give examples of real numbers x and y for which $fl(x \odot y) \neq fl(fl(x) \odot fl(y))$. Illustrate all four arithmetic operations, using a five-decimal machine.
32. When we write $\prod_{i=1}^{n}(1+\delta_i) = 1 + \varepsilon$, where $|\delta_i| \leqq 2^{-24}$, what is the range of possible values for ε? Is $|\varepsilon| \leqq n2^{-24}$ a realistic bound?

Evaluation of Functions

There is another situation in which a drastic loss of significant digits will occur. This is in the evaluation of certain functions for very large arguments. Let us illustrate with the cosine function, which has the periodicity property

$$\cos(x + 2n\pi) = \cos x$$

By the use of this property, the evaluation of $\cos x$ for any argument can be effected by evaluating at a *reduced* argument in the interval $[0, 2\pi]$. The library subroutines available on computers exploit this property in a process called **range reduction**. Other properties may also be used, such as

$$\cos(-x) = \cos x = -\cos(\pi - x)$$

For example, the evaluation of $\cos x$ at $x = 33278.21$ proceeds by finding the *reduced* argument

$$y = 33278.21 - 5296 \times 2\pi = 2.46$$

Here we retain only two decimals, for only two decimal places of accuracy are present in the original argument. The reduced argument has three significant figures, although the original argument may have had seven significant figures. The cosine will then have at most three significant figures. We must not be misled into thinking that the infinite precision available in $5296 \times 2\pi$ is conveyed to the reduced argument y. Also, one should not be deceived by the apparent precision in the printed output from a subroutine. If the cosine subroutine is given an argument y with three significant digits, the value $\cos y$ will have no more than three significant figures, even though it may be displayed as

$$\cos(2.46) = -0.7765702835$$

(The reason for this is that the subroutine treats the argument as being accurate to full machine precision, which of course it is not.)

PROBLEM SET 2.2

1. How many bits of precision are lost in a computer when we carry out the subtraction $x - \sin x$ for $x = \frac{1}{2}$?
2. How many bits of precision are lost in the subtraction $1 - \cos x$ when $x = \frac{1}{4}$?
3. [c] Write and execute a program to compute

$$f(x) = \sqrt{x^2 + 1} - 1$$
$$g(x) = x^2/(\sqrt{x^2 + 1} + 1)$$

for a succession of values of x such as $8^{-1}, 8^{-2}, 8^{-3}, \ldots$. Although $f = g$, the computer will produce different results. Which results are reliable and which are not?

4. [c] Write and test a subroutine that accepts a machine number x and returns the value $y = x - \sin x$, with nearly full machine precision.

27. Explain why loss of significance due to subtraction is not serious in using the approximation

$$x - \sin x \approx (x^3/6)(1 - (x^2/20)(1 - x^2/42))$$

28. In computing the sum of an infinite series $\sum_{n=1}^{\infty} x_n$, suppose that the answer is desired with an absolute error less than ϵ. Is it safe to stop the addition of terms when their magnitude falls below ϵ? Illustrate with the series $\sum_{n=1}^{\infty}(0.99)^n$.
29. Repeat Problem 28 under the additional assumptions that the terms x_n are alternately positive and negative and that $|x_n|$ converges monotonically downward to 0. (Use a theorem in calculus about alternating series.)
30. Show that if x is a machine number on the MARC-32 and if $x > \pi \cdot 2^{25}$, then $\cos x$ will be computed with *no* significant digits.
[c]31. An interesting numerical experiment is to compute the dot product of the following two vectors

$$x = [2.718281828,\ -3.141592654,\ 1.414213562,\ 0.5772156649,\ 0.3010299957]$$
$$y = [1486.2497,\ 878366.9879,\ -22.37492,\ 4773714.647,\ 0.000185049]$$

Compute the summation in four ways:

(i) forward order $\sum_{i=1}^{n} x_i y_i$

(ii) reverse order $\sum_{i=n}^{1} x_i y_i$

(iii) largest-to-smallest order (add positive numbers in order from largest to smallest, then add negative numbers in order from smallest to largest, and then add the two partial sums).

(iv) smallest-to-largest (reverse order of adding in the previous method)

Use both single and double precision for a total of eight answers. Compare the results with the *correct* value to seven decimal places, 1.006571×10^{-9}. Explain your results.

[c]32. (Continuation) Repeat the previous problem but drop the final 9 from x_4 and the final 7 from x_5. What effect does this small change have on the results?
33. By using the error term in Taylor's Theorem, show that at least seven terms are required in the series of Example 2, if the error is not to exceed 10^{-9}.

2.3 Stable and Unstable Computations; Conditioning

In this section we introduce another theme that occurs repeatedly in numerical analysis: the distinction between numerical processes that are *stable* and those that are not. Closely related are the concepts of *well-conditioned* problems and *badly-conditioned* problems.

Numerical Instability

Speaking informally, we say that a numerical process is **unstable** if small errors made at one stage of the process are magnified in subsequent stages and seriously degrade the accuracy of the overall calculation.

An example will help to explain this concept. Consider the sequence of real

computer the numerical solution will probably contain no significant figures. These matters will be discussed in more detail in Chapter 4.

The Hilbert matrix arises in least-squares approximation when we attempt to minimize the expression

$$\int_0^1 \left[\sum_{j=0}^{n} a_j x^j - f(x) \right]^2 dx$$

Upon differentiating this expression with respect to a_i and setting the result equal to zero, we obtain the *normal equations*,

$$\sum_{j=0}^{n} a_j \int_0^1 x^i x^j dx = \int_0^1 x^i f(x) dx \qquad (0 \leqq i \leqq n)$$

Since the integral on the left is

$$\frac{x^{i+j+1}}{i+j+1}\bigg|_0^1 = \frac{1}{i+j+1}$$

the coefficient matrix in the normal equations is the Hilbert matrix of order $n+1$. The functions $x \to x^i$ form a very badly conditioned basis for the space of polynomials of degree n. For this problem, a good basis can be provided by an orthogonal set of polynomials. This will be discussed in Section 6.8.

PROBLEM SET 2.3

1. Find analytically the solution of this difference equation with the given initial values:

$$x_0 = 1 \qquad x_1 = 0.9 \qquad x_{n+1} = -0.2x_n + 0.99x_{n-1}$$

Without computing the solution recursively, predict whether such a computation would be stable or not.

[c]2. Let sequences $[A_n]$ and $[B_n]$ be generated as follows:

$$A_0 = 0 \qquad A_1 = 1 \qquad A_n = nA_{n-1} + A_{n-2}$$
$$B_0 = 1 \qquad B_1 = 1 \qquad B_n = nB_{n-1} + B_{n-2}$$

What is $\lim_{n\to\infty}(A_n/B_n)$?

[c]3. The Bessel functions Y_n satisfy the same recurrence formula that the functions J_n satisfy. (See Section 1.3.) However, they use *different starting values*. For $x = 1$, they are

$$Y_0(1) = 0.08825\ 69642 \qquad Y_1(1) = -0.78121\ 28213$$

Compute $Y_2(1), Y_3(1), \ldots, Y_{20}(1)$ using the recurrence formula. Try to decide whether the results are reliable or not.

4. The exponential integrals are the functions E_n defined by

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

These functions satisfy the equation

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

If $E_1(x)$ is known, can this equation be used to compute $E_2(x), E_3(x), \ldots$ accurately?

5. The condition number of the function $f(x) = x^\alpha$ is independent of x. What is it?
6. What are the condition numbers of the following functions? Where are they large?

(a) $(x-1)^\alpha$ (c) $\sin x$ (e) $x^{-1}e^x$

(b) $\ln x$ (d) e^x (f) $\cos^{-1} x$

7. Consider the example in the text for which $y_{n+1} = e - (n+1)y_n$. How many decimals of accuracy should be used in computing $y_1, y_2, \ldots, y_{20}$ if y_{20} is to be accurate to five decimals?
8. There is a function f of the form

$$f(x) = \alpha x^{12} + \beta x^{13}$$

for which $f(0.1) = 6.06 \times 10^{-13}$ and $f(0.9) = 0.03577$. Determine α and β, and assess the sensitivity of these parameters to slight changes in the values of f.

9. 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 the recurrence relation a good way to compute x_n from all initial values x_0 and x_1?

10. The **Fibonacci** sequence is generated by the formulæ

$$r_0 = 1 \qquad r_1 = 1 \qquad r_{n+1} = r_n + r_{n-1}$$

The sequence therefore starts out 1,1,2,3,5,8,13,21,34,.... Prove that the sequence $[2r_n/r_{n-1}]$ converges to $1+\sqrt{5}$. Is the convergence linear, superlinear, quadratic?

11. (Continuation) If the recurrence relation in the preceding problem is used with starting values $r_0 = 1$ and $r_1 = (1-\sqrt{5})/2$, what is the theoretically correct value of r_n $(n \geqq 2)$? Will the recurrence relation provide a stable means for computing r_n in this case?

[c]12. (Dahlquist) 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 $n = 1, 2, \ldots$. Compute $x_0, x_1, \ldots, x_{10}$ using this recurrence formula and estimate the accuracy of x_{10}.