

### 1.3 Error, Accuracy, and Stability

Although we assume no prior training of the reader in formal numerical analysis, we will need to presume a common understanding of a few key concepts. We will define these briefly in this section.

Computers store numbers not with infinite precision but rather in some approximation that can be packed into a fixed number of *bits* (binary digits) or *bytes* (groups of 8 bits). Almost all computers allow the programmer a choice among several different such *representations* or *data types*. Data types can differ in the number of bits utilized (the *wordlength*), but also in the more fundamental respect of whether the stored number is represented in *fixed-point* (int or long) or *floating-point* (float or double) format.

A number in integer representation is exact. Arithmetic between numbers in integer representation is also exact, with the provisos that (i) the answer is not outside the range of (usually, signed) integers that can be represented, and (ii) that division is interpreted as producing an integer result, throwing away any integer remainder.

In floating-point representation, a number is represented internally by a sign bit  $s$  (interpreted as plus or minus), an exact integer exponent  $e$ , and an exact positive integer mantissa  $M$ . Taken together these represent the number

$$s \times M \times B^{e-E} \quad (1.3.1)$$

where  $B$  is the base of the representation (usually  $B = 2$ , but sometimes  $B = 16$ ), and  $E$  is the *bias* of the exponent, a fixed integer constant for any given machine and representation. An example is shown in Figure 1.3.1.

Several floating-point bit patterns can represent the same number. If  $B = 2$ , for example, a mantissa with leading (high-order) zero bits can be left-shifted, i.e., multiplied by a power of 2, if the exponent is decreased by a compensating amount. Bit patterns that are “as left-shifted as they can be” are termed *normalized*. Most computers always produce normalized results, since these don’t waste any bits of the mantissa and thus allow a greater accuracy of the representation. Since the high-order bit of a properly normalized mantissa (when  $B = 2$ ) is *always* one, some computers don’t store this bit at all, giving one extra bit of significance.

Arithmetic among numbers in floating-point representation is not exact, even if the operands happen to be exactly represented (i.e., have exact values in the form of equation 1.3.1). For example, two floating numbers are added by first right-shifting (dividing by two) the mantissa of the smaller (in magnitude) one, simultaneously increasing its exponent, until the two operands have the same exponent. Low-order (least significant) bits of the smaller operand are lost by this shifting. If the two operands differ too greatly in magnitude, then the smaller operand is effectively replaced by zero, since it is right-shifted to oblivion.

The smallest (in magnitude) floating-point number which, when added to the floating-point number 1.0, produces a floating-point result different from 1.0 is termed the *machine accuracy*  $\epsilon_m$ . A typical computer with  $B = 2$  and a 32-bit wordlength has  $\epsilon_m$  around  $3 \times 10^{-8}$ . (A more detailed discussion of machine characteristics, and a program to determine them, is given in §20.1.) Roughly



Roundoff error is a characteristic of computer hardware. There is another, different, kind of error that is a characteristic of the program or algorithm used, independent of the hardware on which the program is executed. Many numerical algorithms compute “discrete” approximations to some desired “continuous” quantity. For example, an integral is evaluated numerically by computing a function at a discrete set of points, rather than at “every” point. Or, a function may be evaluated by summing a finite number of leading terms in its infinite series, rather than all infinity terms. In cases like this, there is an adjustable parameter, e.g., the number of points or of terms, such that the “true” answer is obtained only when that parameter goes to infinity. Any practical calculation is done with a finite, but sufficiently large, choice of that parameter.

The discrepancy between the true answer and the answer obtained in a practical calculation is called the *truncation error*. Truncation error would persist even on a hypothetical, “perfect” computer that had an infinitely accurate representation and no roundoff error. As a general rule there is not much that a programmer can do about roundoff error, other than to choose algorithms that do not magnify it unnecessarily (see discussion of “stability” below). Truncation error, on the other hand, is entirely under the programmer’s control. In fact, it is only a slight exaggeration to say that clever minimization of truncation error is practically the entire content of the field of numerical analysis!

Most of the time, truncation error and roundoff error do not strongly interact with one another. A calculation can be imagined as having, first, the truncation error that it would have if run on an infinite-precision computer, “plus” the roundoff error associated with the number of operations performed.

Sometimes, however, an otherwise attractive method can be *unstable*. This means that any roundoff error that becomes “mixed into” the calculation at an early stage is successively magnified until it comes to swamp the true answer. An unstable method would be useful on a hypothetical, perfect computer; but in this imperfect world it is necessary for us to require that algorithms be stable — or if unstable that we use them with great caution.

Here is a simple, if somewhat artificial, example of an unstable algorithm: Suppose that it is desired to calculate all integer powers of the so-called “Golden Mean,” the number given by

$$\phi \equiv \frac{\sqrt{5} - 1}{2} \approx 0.61803398 \quad (1.3.3)$$

It turns out (you can easily verify) that the powers  $\phi^n$  satisfy a simple recursion relation,

$$\phi^{n+1} = \phi^{n-1} - \phi^n \quad (1.3.4)$$

Thus, knowing the first two values  $\phi^0 = 1$  and  $\phi^1 = 0.61803398$ , we can successively apply (1.3.4) performing only a single subtraction, rather than a slower multiplication by  $\phi$ , at each stage.

Unfortunately, the recurrence (1.3.4) also has *another* solution, namely the value  $-\frac{1}{2}(\sqrt{5} + 1)$ . Since the recurrence is linear, and since this undesired solution has magnitude greater than unity, any small admixture of it introduced by roundoff errors will grow exponentially. On a typical machine with 32-bit wordlength, (1.3.4) starts

to give completely wrong answers by about  $n = 16$ , at which point  $\phi^n$  is down to only  $10^{-4}$ . The recurrence (1.3.4) is *unstable*, and cannot be used for the purpose stated.

We will encounter the question of stability in many more sophisticated guises, later in this book.

CITED REFERENCES AND FURTHER READING:

- Stoer, J., and Bulirsch, R. 1980, *Introduction to Numerical Analysis* (New York: Springer-Verlag), Chapter 1.
- Kahaner, D., Moler, C., and Nash, S. 1989, *Numerical Methods and Software* (Englewood Cliffs, NJ: Prentice Hall), Chapter 2.
- Johnson, L.W., and Riess, R.D. 1982, *Numerical Analysis*, 2nd ed. (Reading, MA: Addison-Wesley), §1.3.
- Wilkinson, J.H. 1964, *Rounding Errors in Algebraic Processes* (Englewood Cliffs, NJ: Prentice-Hall).