Decimal Machine Numbers

The use of binary digits tends to conceal the computational difficulties that occur when a finite collection of machine numbers is used to represent all the real numbers. To examine these problems, we will use more familiar decimal numbers instead of binary representation. Specifically, we assume that machine numbers are represented in the normalized *decimal* floating-point form

$$\pm 0.d_1d_2 \dots d_k \times 10^n, \quad 1 \le d_1 \le 9, \quad \text{and} \quad 0 \le d_i \le 9,$$

for each $i = 2, \dots, k$. Numbers of this form are called k-digit *decimal machine numbers*.

Any positive real number within the numerical range of the machine can be normalized to the form

$$y = 0.d_1d_2 \dots d_kd_{k+1}d_{k+2} \dots \times 10^n.$$

The error that results from replacing a number with its floating-point form is called **round-off error** regardless of whether the rounding or chopping method is used.

The floating-point form of y, denoted $fl(y)$, is obtained by terminating the mantissa of y at k decimal digits. There are two common ways of performing this termination. One method, called **chopping**, is to simply chop off the digits $d_{k+1}d_{k+2} \dots$. This produces the floating-point form

$$fl(y) = 0.d_1d_2 \dots d_k \times 10^n.$$

The other method, called **rounding**, adds $5 \times 10^{n-(k+1)}$ to y and then chops the result to obtain a number of the form

$$fl(y) = 0.\delta_1\delta_2 \dots \delta_k \times 10^n.$$

For rounding, when $d_{k+1} \ge 5$, we add 1 to d_k to obtain $fl(y)$; that is, we *round up*. When $d_{k+1} < 5$, we simply chop off all but the first k digits; so we *round down*. If we round down, then $\delta_i = d_i$, for each $i = 1, 2, \dots, k$. However, if we round up, the digits (and even the exponent) might change.

Example 1 Determine the five-digit **(a)** chopping and **(b)** rounding values of the irrational number π.

Solution The number π has an infinite decimal expansion of the form $\pi = 3.14159265\dots$. Written in normalized decimal form, we have

$$\pi = 0.314159265\dots \times 10^1.$$

The relative error is generally a better measure of accuracy than the absolute error because it takes into consideration the size of the number being approximated.

(a) The floating-point form of π using five-digit chopping is

$$fl(\pi) = 0.31415 \times 10^1 = 3.1415.$$

(b) The sixth digit of the decimal expansion of π is a 9, so the floating-point form of π using five-digit rounding is

$$fl(\pi) = (0.31415 + 0.00001) \times 10^1 = 3.1416.$$ ■

The following definition describes two methods for measuring approximation errors.

Definition 1.15 Suppose that p^* is an approximation to p. The **absolute error** is $|p - p^*|$, and the **relative error** is $\dfrac{|p - p^*|}{|p|}$, provided that $p \ne 0$. ■

Consider the absolute and relative errors in representing p by p^* in the following example.