2.2 Loss of Significance

In this section, we show how loss of significance in subtraction can often be reduced or eliminated by various techniques, such as the use of rationalization, Taylor series, trigonometric identities, logarithmic properties, double precision, and/or range reduction. These are some of the techniques that can be used when one wants to guard against the degradation of precision in a calculation. Of course, we cannot always know when a loss of significance has occurred in a long computation, but we should be alert to the possibility and take steps to avoid it, if possible.

Significant Digits

We first address the elusive concept of **significant digits** in a number. Suppose that x is a real number expressed in normalized scientific notation in the decimal system

$$x = \pm r \times 10^n \qquad \left(\tfrac{1}{10} \leqq r < 1\right)$$

For example, x might be

$$x = 0.37214\,98 \times 10^{-5}$$

The digits 3, 7, 2, 1, 4, 9, 8 used to express r do not all have the same significance because they represent different powers of 10. Thus, we say that 3 is the *most* significant digit, and the significance of the digits diminishes from left to right. In this example, 8 is the *least* significant digit.

If x is a *mathematically exact* real number, then its approximate decimal form can be given with as many significant digits as we wish. Thus, we may write

$$\frac{\pi}{10} \approx 0.31415\,92653\,58979$$

and all the digits given are correct. If x is a *measured quantity*, however, the situation is quite different. Every measured quantity involves an error whose magnitude depends on the nature of the measuring device. Thus, if a meter stick is used, it is not reasonable to measure any length with precision better than 1 millimeter. Therefore, the result of measuring, say, a plate glass window with a meter stick should not be reported as 2.73594 meters. That would be misleading. Only digits that are believed to be correct or in error by at most a few units should be reported. It is a scientific convention that the least significant digit given in a measured quantity should be in error by at most five units; that is, the result is rounded correctly.

Similar remarks pertain to quantities computed from measured quantities. For example, if the side of a square is reported to be $s = 0.736$ meter, then one can assume that the error does not exceed a few units in the third decimal place. The diagonal of that square is then

$$s\sqrt{2} \approx 0.10408\,61182 \times 10^1$$

but should be reported as 0.1041×10^1 or (more conservatively) 0.104×10^1. The infinite precision available in $\sqrt{2}$,

$$\sqrt{2} = 1.41421\,35623\,73095\ldots$$

does *not* convey any more precision to $s\sqrt{2}$ than was already present in s.

Computer-Caused Loss of Significance

Perhaps it is surprising that a loss of significance can occur within the computer. It is essential to understand this process so that blind trust will not be placed in numerical output from a computer. One of the most common causes for a deterioration in precision is the subtraction of one quantity from another nearly equal quantity. This effect is potentially quite serious and can be catastrophic. The closer these two numbers are to each other, the more pronounced is the effect.

To illustrate this phenomenon, let us consider the assignment statement

$$y \leftarrow x - \sin(x)$$

and suppose that at some point in a computer program this statement is executed with an x value of $\frac{1}{15}$. Assume further that our computer works with floating-point numbers that have ten decimal digits. Then

$$\begin{aligned} x &\leftarrow 0.66666\,66667 \times 10^{-1} \\ \sin(x) &\leftarrow 0.66617\,29492 \times 10^{-1} \\ x - \sin(x) &\leftarrow 0.00049\,37175 \times 10^{-1} \\ x - \sin(x) &\leftarrow 0.49371\,75000 \times 10^{-4} \end{aligned}$$

In the last step, the result has been shifted to normalized floating-point form. Three zeros have then been supplied by the computer in the three *least* significant decimal places. We refer to these as **spurious zeros**; they are *not* significant digits. In fact, the ten-decimal-digit correct value is

$$\frac{1}{15} - \sin\frac{1}{15} \approx 0.49371\,74327 \times 10^{-4}$$

Another way of interpreting this is to note that the final digit in $x - \sin(x)$ is derived from the tenth digits in x and $\sin(x)$. When the eleventh digit in either x or $\sin(x)$ is 5, 6, 7, 8, or 9, the numerical values are rounded up to ten digits so that their tenth digits may be altered by plus one unit. Since these tenth digits may be in error, the final digit in $x - \sin(x)$ may also be in error—which it is!

EXAMPLE 1 If $x = 0.37214\,48693$ and $y = 0.37202\,14371$, what is the relative error in the computation of $x - y$ in a computer that has five decimal digits of accuracy?

Solution The numbers would first be rounded to $\widetilde{x} = 0.37214$ and $\widetilde{y} = 0.37202$. Then we have $\widetilde{x} - \widetilde{y} = 0.00012$, while the correct answer is $x - y = 0.00012\,34322$. The relative error involved is

$$\frac{|(x-y) - (\widetilde{x} - \widetilde{y})|}{|x - y|} = \frac{0.00000\,34322}{0.00012\,34322} \approx 3 \times 10^{-2}$$

This magnitude of relative error must be judged quite large when compared with the relative error of $\widetilde{x}$ and $\widetilde{y}$. (They cannot exceed $\frac{1}{2} \times 10^{-4}$ by the coarsest estimates, and in this example, they are, in fact, approximately 1.3×10^{-5}.) ■

It should be emphasized that this discussion pertains not to the operation

$$\text{fl}(x - y) \leftarrow x - y$$

but rather to the operation

$$\text{fl}[\text{fl}(x) - \text{fl}(y)] \leftarrow x - y$$

Roundoff error in the former case is governed by the equation

$$\text{fl}(x - y) = (x - y)(1 + \delta)$$

where $|\delta| \leqq 2^{-24}$ on a 32-bit word-length computer, and on a five-decimal-digit computer in the example above $|\delta| \leqq \frac{1}{2} \times 10^{-4}$.

In Example 1, we observe that the computed difference of 0.00012 has only two significant figures of accuracy, whereas in general, one expects the numbers and calculations in this computer to have five significant figures of accuracy.

The remedy for this difficulty is first to anticipate that it may occur and then to reprogram. The simplest technique may be to carry out part of a computation in double- or extended-precision arithmetic (that means roughly twice as many significant digits), but often a slight change in the formulas is required. Several illustrations of this will be given, and the reader will find additional ones among the problems.

Consider Example 1, but imagine that the calculations to obtain x, y, and $x - y$ are being done in double precision. Suppose that single-precision arithmetic is used thereafter. In the computer, all ten digits of x, y, and $x - y$ will be retained, but at the end, $x - y$ will be rounded to its five-digit form, which is 0.12343×10^{-3}. This answer has five significant digits of accuracy, as we would like. Of course, the programmer or analyst must know in advance where the double-precision arithmetic will be necessary in the computation. Programming everything in double precision is very wasteful if it is not needed. This approach has another drawback: There may be such serious cancellation of significant digits that even double precision might not help.

Theorem on Loss of Precision

Before considering other techniques for avoiding this problem, we ask the following question: *Exactly how many significant binary digits are lost in the subtraction $x - y$ when x is close to y?* The closeness of x and y is conveniently measured by $|1 - (y/x)|$. Here is the result:

THEOREM 1 **LOSS OF PRECISION THEOREM**

Let x and y be normalized floating-point machine numbers, where $x > y > 0$. If $2^{-p} \leqq 1 - (y/x) \leqq 2^{-q}$ for some positive integers p and q, then at most p and at least q significant binary bits are lost in the subtraction $x - y$.

Proof We prove the second part of the theorem and leave the first as an exercise. To this end, let $x = r \times 2^n$ and $y = s \times 2^m$, where $\frac{1}{2} \leqq r, s < 1$. (This is the normalized binary floating-point form.) Since $y < x$, the computer may have to *shift* y before carrying out the subtraction. In any case, y must first be expressed with the same exponent as x. Hence, $y = (s2^{m-n}) \times 2^n$ and

$$x - y = (r - s2^{m-n}) \times 2^n$$

The mantissa of this number satisfies the equations and inequality

$$r - s2^{m-n} = r\left(1 - \frac{s2^m}{r2^n}\right) = r\left(1 - \frac{y}{x}\right) < 2^{-q}$$

Hence, to normalize the representation of $x - y$, a shift of at least q bits to the left is necessary. Then at least q (spurious) zeros are supplied on the right-hand end of the mantissa. This means that at least q bits of precision have been lost. ■

EXAMPLE 2 In the subtraction 37.59362 1 − 37.58421 6, how many bits of significance will be lost?

Solution Let x denote the first number and y the second. Then

$$1 - \frac{y}{x} = 0.00025\,01754$$

This lies between 2^{-12} and 2^{-11}. These two numbers are 0.00024 4 and 0.00048 8. Hence, at least 11 but not more than 12 bits are lost. ■

Here is an example in decimal form: let $x = .6353$ and $y = .6311$. These are close, and $1 - y/x = .00661 < 10^{-2}$. In the subtraction, we have $x - y = .0042$. There are two significant figures in the answer, although there were four significant figures in x and y.

Avoiding Loss of Significance in Subtraction

Now we take up various techniques that can be used to avoid the loss of significance that may occur in subtraction. Consider the function

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

whose values may be required for x near zero. Since $\sqrt{x^2 + 1} \approx 1$ when $x \approx 0$, we see that there is a potential loss of significance in the subtraction. However, the function can be rewritten in the form

$$f(x) = \left(\sqrt{x^2 + 1} - 1\right)\left(\frac{\sqrt{x^2 + 1} + 1}{\sqrt{x^2 + 1} + 1}\right) = \frac{x^2}{\sqrt{x^2 + 1} + 1} \tag{2}$$

by **rationalizing** the numerator—that is, removing the radical in the numerator. This procedure allows terms to be canceled and thereby removes the subtraction. For example, if we use five-decimal-digit arithmetic and if $x = 10^{-3}$, then $f(x)$ will be computed incorrectly as zero by the first formula but as $\frac{1}{2} \times 10^{-6}$ by the second. If we use the first formula together with double precision, the difficulty is ameliorated but *not* circumvented altogether. For example, in double precision, we have the same problem when $x = 10^{-6}$.

As another example, suppose that the values of

$$f(x) = x - \sin x \tag{3}$$

are required near $x = 0$. A careless programmer might code this function just as indicated in Equation (3), not realizing that serious loss of accuracy will occur. Recall from calculus that

$$\lim_{x \to 0} \frac{\sin x}{x} = 1$$

to see that $\sin x \approx x$ when $x \approx 0$. One cure for this problem is to use the Taylor series for $\sin x$:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

This series is known to represent $\sin x$ for all real values of x. For x near zero, it converges quite rapidly. Using this series, we can write the function f as

$$f(x) = x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} - \cdots \right) = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!} - \cdots \tag{4}$$

We see in this equation where the original difficulty arose; namely, for small values of x, the term x in the sine series is much larger than $x^3/3!$ and thus more important. But when $f(x)$ is formed, this dominant x term disappears, leaving only the lesser terms. The series that starts with $x^3/3!$ is very effective for calculating $f(x)$ when x is small.

In this example, further analysis is needed to determine the range in which Series (4) should be used and the range in which Formula (3) can be used. Using the Theorem on Loss of Precision, we see that the loss of bits in the subtraction of Formula (3) can be limited to at most *one* bit by restricting x so that $\frac{1}{2} \leqq 1 - \sin x/x$. (Here we are considering only the case when $\sin x > 0$.) With a calculator, it is easy to see that x must be at least 1.9. Thus, for $|x| < 1.9$, we use the first few terms in the Series (4), and for $|x| \geqq 1.9$, we use $f(x) = x - \sin x$. One can verify that for the worst case ($x = 1.9$), ten terms in the series give $f(x)$ with an error of at most 10^{-16}. (That is good enough for double precision on a 32-bit word-length computer.)

To construct a function procedure for $f(x)$, notice that the terms in the series can be obtained inductively by the algorithm

$$\begin{cases} t_1 = \dfrac{x^3}{6} \\ t_{n+1} = \dfrac{-t_n x^2}{(2n+2)(2n+3)} \qquad (n \geqq 1) \end{cases}$$

Then the partial sums can be obtained inductively by

$$\begin{cases} s_1 = t_1 \\ s_{n+1} = s_n + t_{n+1} \qquad (n \geqq 1) \end{cases}$$

so that

$$s_n = \sum_{k=1}^{n} t_k = \sum_{k=1}^{n} (-1)^{k+1} \left[\frac{x^{2k+1}}{(2k+1)!} \right]$$

Suitable pseudocode for a function is given here:

```
real function $f(x)$
integer $i, n \leftarrow 10$;    real $s, t, x$
if $|x| \geqq 1.9$ then
    $s \leftarrow x - \sin x$
    else
    $t \leftarrow x^3/6$
    $s \leftarrow t$
    for $i = 2$ to $n$ do
        $t \leftarrow -tx^2/[(2i+2)(2i+3)]$
        $s \leftarrow s + t$
    end for
end if
$f \leftarrow s$
end function $f$
```

EXAMPLE 3 How can accurate values of the function

$$f(x) = e^x - e^{-2x}$$

be computed in the vicinity of $x = 0$?

Solution Since e^x and e^{-2x} are both equal to 1 when $x = 0$, there will be a loss of significance because of subtraction when x is close to zero. Inserting the appropriate Taylor series, we obtain

$$\begin{aligned} f(x) &= \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots\right) - \left(1 - 2x + \frac{4x^2}{2!} - \frac{8x^3}{3!} + \cdots\right) \\ &= 3x - \frac{3}{2}x^2 + \frac{3}{2}x^3 - \cdots \end{aligned}$$

An alternative is to write

$$\begin{aligned} f(x) &= e^{-2x}\left(e^{3x} - 1\right) \\ &= e^{-2x}\left(3x + \frac{9}{2!}x^2 + \frac{27}{3!}x^3 + \cdots\right) \end{aligned}$$

By using the Theorem on Loss of Precision, we find that at most one bit is lost in the subtraction $e^x - e^{-2x}$ when $x > 0$ and

$$\frac{1}{2} \leqq 1 - \frac{e^{-2x}}{e^x}$$

This inequality is valid when $x \geqq \frac{1}{3}\ln 2 = 0.23105$. Similar reasoning when $x < 0$ shows that for $x \leqq -0.23105$, at most one bit is lost. Hence, the series should be used for $|x| < 0.23105$. ■

EXAMPLE 4 Criticize the assignment statement

$$y \leftarrow \cos^2(x) - \sin^2(x)$$

Solution When $\cos^2(x) - \sin^2(x)$ is computed, there will be a loss of significance at $x = \pi/4$ (and other points). The simple trigonometric identity

$$\cos 2\theta = \cos^2\theta - \sin^2\theta$$

should be used. Thus, the assignment statement should be replaced by

$$y \leftarrow \cos(2x)$$ ■

EXAMPLE 5 Criticize the assignment statement

$$y \leftarrow \ln(x) - 1$$

Solution If the expression $\ln x - 1$ is used for x near e, there will be a cancellation of digits and a loss of accuracy. One can use elementary facts about logarithms to overcome the difficulty. Thus, we have $y = \ln x - 1 = \ln x - \ln e = \ln(x/e)$. Here is a suitable assignment statement

$$y \leftarrow \ln\left(\frac{x}{e}\right)$$ ■

Range Reduction

Another cause of loss of significant figures is the evaluation of various library functions with very large arguments. This problem is more subtle than the ones previously discussed. We illustrate with the sine function.

A basic property of the function $\sin x$ is its **periodicity**:

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

for all real values of x and for all integer values of n. Because of this relationship, one needs to know only the values of $\sin x$ in some fixed interval of length 2π to compute $\sin x$ for arbitrary x. This property is used in the computer evaluation of $\sin x$ and is called **range reduction**.

Suppose now that we want to evaluate sin(12532.14). By subtracting integer multiples of 2π, we find that this equals sin(3.47) if we retain only two decimal digits of accuracy. From $\sin(12532.14) = \sin(12532.14 - 2k\pi)$, we want $12532 = 2k\pi$ and $k = 3989/2\pi \approx 1994$. Consequently, we obtain $12532.14 - 2(1994)\pi = 3.49$ and $\sin(12532.14) \approx \sin(3.49)$. Thus, although our original argument 12532.14 had seven significant figures, the reduced argument has only three. The remaining digits disappeared in the subtraction of 3988π. Since 3.47 has only three significant figures, our computed value of sin(12532.14) will have *no more than* three significant figures. This decrease in precision is unavoidable if there is no way of increasing the precision of the original argument. If the original argument (12532.14 in this example) can be obtained with more significant figures, these additional figures will be present in the *reduced* argument (3.47 in this example). In some cases, double- or extended-precision programming will help.

EXAMPLE 6 For $\sin x$, how many binary bits of significance are lost in range reduction to the interval $[0, 2\pi)$?

Solution Given an argument $x > 2\pi$, we determine an integer n that satisfies the inequality $0 \leqq x - 2n\pi < 2\pi$. Then in evaluating elementary trigonometric functions, we use

$f(x) = f(x - 2n\pi)$. In the subtraction $x - 2n\pi$, there will be a loss of significance. By the Theorem on Loss of Precision, at least q bits are lost if

$$1 - \frac{2n\pi}{x} \leqq 2^{-q}$$

Since

$$1 - \frac{2n\pi}{x} = \frac{x - 2n\pi}{x} < \frac{2\pi}{x}$$

we conclude that at least q bits are lost if $2\pi/x \leqq 2^{-q}$. Stated otherwise, at least q bits are lost if $2^q \leqq x/2\pi$. ■

Summary

(1) To avoid loss of significance in subtraction, one may be able to reformulate the expression using rationalizing, series expansions, or mathematical identities.

(2) If x and y are positive normalized floating-point machine numbers with

$$2^{-p} \leqq 1 - \frac{y}{x} \leqq 2^{-q}$$

then at most p and at least q significant binary bits are lost in computing $x - y$. Note that it is permissible to leave out the hypothesis $x > y$ here.

Additional References

For supplemental study and reading of material related to this chapter, see Appendix B as well as the following references: Acton [1996], Bornemann, Laurie, Wagon, and Waldvogel [2004], Goldberg [1991], Higham [2002], Hodges [1983], Kincaid and Cheney [2002], Overton [2001], Salamin [1976], Wilkinson [1963], and others listed in the Bibliography.

Problems 2.2

1. How can values of the function $f(x) = \sqrt{x+4} - 2$ be computed accurately when x is small?

2. Calculate $f(10^{-2})$ for the function $f(x) = e^x - x - 1$. The answer should have five significant figures and can easily be obtained with pencil and paper. Contrast it with the straightforward evaluation of $f(10^{-2})$ using $e^{0.01} \approx 1.0101$.

3. What is a good way to compute values of the function $f(x) = e^x - e$ if full machine precision is needed? *Note:* There is some difficulty when $x = 1$.

[a]**4.** What difficulty could the following assignment cause?

$$y \leftarrow 1 - \sin x$$

Circumvent it without resorting to a Taylor series if possible.