Approximations and Round-Off Errors

Definition: The number of *significant figures* or *significant digits* in the representation of a number is the number of digits that can be used with confidence. In particular, for our purposes, the number of significant digits is equal to the number of digits that are known (or assumed) to be correct plus one estimated digit.

Examples:

1. The output from a physical measuring device or sensor is generally known to be correct up to a fixed number of digits. For example, if the temperature is measured with a thermometer that is calibrated in tenths of a degree Fahrenheit, then we might know that the temperature is between 85.6 and 85.7 degrees. In this case, we know that the first three digits of the temperature are 8, 5, and 6, but we do not know the value of any subsequent digits. The first unknown digit is sometimes estimated at half of the value of the calibration size, or 0.05 degrees in our case. If we did this, the temperature would be reported to four significant digits as 85.65 degrees.

2. Alternatively, we could choose to report the temperature to only three significant digits. In this case, we could *truncate* or chop off the unknown digits to give a result of 48.6 degrees, or we could round off the result to the nearest tenth of a degree to give either 48.6 or 48.7 degrees depending on whether the actual reading was more or less than half-way between the two calibrations. Round-off is generally the preferred procedure in this example, but without knowing which technique was adopted, we would really only be confident in the first two digits of the temperature. Hence, if the temperature is reported as 48.6 degrees without any further explanation, we do not know whether the 6 is a correct digit or an estimated digit.

Comments:

1. In general, if a number is reported with *n* digits (excluding leading zeroes), the *n*th digit is assumed to be estimated and we say that there are *n* significant digits.
2. Leading zeroes are never counted as significant digits, even if preceded by a decimal point. Hence, 0.0062 and 0.62 both have only two significant digits.

Definition:

1. The *precision* of a number is an indication of how carefully the number was measured. Precision is generally equated with significant digits; *i.e.*, four digits of precision is the same as four significant digits. If the same quantity is measured repeatedly with high precision, the resulting estimates will all be close together. Hence, precision can be estimated by considering the spread or *standard deviation* of a sequence of measurements.
2. The accuracy of a number is an indication of how close the number is to some (generally unknown) true value. Inaccuracies are generically referred to simply as *errors* and can result from many sources, including insufficient precision in measurements. It is often rather difficult to determine the accuracy of a measurement without prior information, such as a reference value. However, if a reference value is available, the accuracy of a measurement can be determined by averaging a sequence of measurements and comparing the average or *mean value* to the reference value.

As mentioned, there are many different sources of error, for example:

1. **Truncation error**: Technically, this name refers to errors that are caused by *truncating* or stopping a numerical procedure before it has

really *converged* or stopped changing. Many numerical procedures would have to be run forever in order to truly converge if you could record intermediate values with infinite precision. However, in practice there is always some sort of stopping criterion, and the difference between the final intermediate value and the true value represents the truncation error associated with the procedure.

2. **Inherent errors**: Errors resulting from the inaccuracies of the measurement process itself, or possibly from inaccuracies in the mathematical model that the analyst uses to represent the physical world, are lumped together and called inherent errors.

3. **Propagated error**: This is error in succeeding steps of a process that is *propagated forward* from a previous inaccurate result. This can result in a procedure that never converges or one which "blows up." For example, we could try to find the square root of 2 recursively by solving the equation $x^2 = 2$ using the procedure $x_{n+1} = 2/x_n$. However, this means that

$$\begin{aligned} x_{n+1} &= 2/x_n \\ &= \frac{2}{2/x_{n-1}} \\ &= x_{n-1}. \end{aligned}$$

So, if we do not start at exactly the correct answer, which is impossible, since $\sqrt{2}$ is an irrational number, we just cycle between two incorrect answers. On the other hand, if we use the recursion

$$x_{n+1} = 1 + \frac{1}{1 + x_n},$$

the procedure converges very quickly to the correct answer. Hence, one of these procedures suffers from propagation error, and the other does not. Note that if we start with $x_0 = 1$, then the second procedure can be written in the form of a *continued fraction*, *i.e.*,

$$\sqrt{2} = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \cdots}}}}}}.$$

4. **Round-off Error**: Since all computers have a finite word length and use the binary number system, most decimal numbers cannot be represented with complete accuracy in a computer. Note that this is true whether the numbers are stored in fixed-point format or floating-point format. Whether the infinite binary string that represents the true value of the decimal number is truncated (*i.e.*, chopped off) or rounded somehow to produce an internal representation, this type of error is simply called round-off error.

More About Errors

Regardless of the source of inaccuracy, inaccurate measurements result in an error. Mathematically, this is represented as

$$\text{True value} = \text{approximation} + \text{error},$$

or

$$E_t = \text{true value} - \text{approximation},$$

where E_t represents the true value of the error. Actually, we are generally only concerned with the *absolute error*, which is just the *magnitude* or absolute value of E_t, *i.e.*, $|E_t|$. Furthermore, a more useful notion than the absolute error is often *relative absolute error*, which computes the error as a fraction of the magnitude of the true value. That is,

$$\varepsilon_t = \frac{|\text{true error}|}{|\text{true value}|},$$

where ε_t represents the relative absolute error. Note that your book actually defines ε_t slightly differently, but the two definitions are not much different, and this is the form we will use.

Finally, since we generally do not actually know either the true error or the true value of the number we are estimating, we are often forced to estimate the relative absolute error itself. The approximate relative absolute error is defined as

$$\varepsilon_a = \frac{|\text{approximate error}|}{|\text{approximation}|},$$

where the definition of “approximate error” is problem specific. For example, if we are using an iterative procedure to produce our estimates, we will often define ε_a as

$$\varepsilon_a = \frac{|\text{current approximation} - \text{previous approximation}|}{|\text{current approximation}|}.$$

Since numbers in computers cannot be stored exactly, the order in which operations are performed sometimes influences the accuracy of the answer. Following are some simple rules to remember to avoid difficulties when performing numerical procedures on a computer.

1. Try to avoid subtracting two nearly equal numbers whenever possible. Sometimes this cannot be avoided, and sometimes it does not cause any problems, but if a very small number results from subtracting two nearly equal large numbers, the relative error of the result can be large. If the result is later multiplied again by a large number, the absolute error can become significant. For example, consider the two roots of the equation $ax^2 + bx + c = 0$, where $a = 1$, $b = 0.4002$, and $c = 0.00008$. If we use four digits of precision and compute the smaller of the two roots in the usual way, *i.e.*,

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a},$$

then we get $x = 2.500 \times 10^{-4}$. On the other hand, if we compute the smaller root using the equivalent formula

$$x = \frac{-2c}{b + \sqrt{b^2 - 4ac}},$$

then we get $x = 2.000 \times 10^{-4}$, which is the true answer correct to four decimal places.

2. If you are forced to subtract two nearly equal numbers, factor out common factors first. For example, consider the computation of $ab - ac$, where $a = 684.2$, $b = 0.5685$, and $c = 0.5641$. If we compute this directly using four digits of precision, we get $ab - ac = 3.000$; whereas, if we compute $a(b - c)$, we get $ab - ac = 3.010$, which is the true answer correct to four decimal places.

3. When adding up a long sequence of numbers, sort the sequence into increasing order, and add up the smaller numbers first. This reduces the propagated error in the result. For example, using four digits of precision, we get

$$\begin{aligned} 739 &= 991.1 + 327.6 + 225.0 + 85.67 + 75.61 \\ &\quad + 25.54 + 3.712 + 1.543 + 0.5112 + 0.1001, \end{aligned}$$

   but

$$\begin{aligned} 736 &= 0.1001 + 0.5112 + 1.543 + 3.712 + 25.54 \\ &\quad + 75.61 + 85.67 + 225.0 + 327.6 + 991.1. \end{aligned}$$

   which is much closer to the correct answer.

4. When adding up a long sequence of nearly equal numbers, try adding them in pairs sequentially. For example, if we add the number 0.100 20,000 times using four digits of precision, we get

$$0.100 + 0.100 + 0.100 + \ldots + 0.100 = 1000;$$

   whereas, if we use the following procedure:

$$\begin{aligned} 0.100+0.100+0.100+\ldots+0.100 &= (0.100+0.100)+\ldots+(0.100+0.100) \\ &= (0.200+0.200)+\ldots+(0.200+0.200) \\ &= (0.400+0.400)+\cdots+(0.400+0.400) \\ &\vdots \\ &= (500.0+500.0)+(500.0+500.0) \\ &= 1000+1000 \\ &= 2000, \end{aligned}$$

we get the correct answer.

Taylor Series

Taylor's Theorem: If the function *f* and its first $n+1$ derivatives are continuous on the interval $[a,b]$, then for any $x \in [a,b]$, we have

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(x),$$

where

$$R_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-a)^{n+1},$$

for some $\xi \in [a,x]$. Notice that this implies that

$$|R_n(x)| \le \frac{\max\limits_{\xi\in[a,x]} \left| f^{(n+1)}(\xi) \right|}{(n+1)!}(x-a)^{n+1}$$

Since $(n+1)!$ becomes quite large as $n \to \infty$, the remainder term becomes quite small for *n* sufficiently large, and the function can be well approximated by a polynomial of some order for all $x \in [a,b]$.

In order to use Taylor's theorem to approximate an arbitrary function *f* with a polynomial at some point *x*, we have to pick a "center point" *a* to expand around. In order for *a* to be a good choice for the center of the expansion, it should satisfy the following properties:

1. The function *f* and all of its derivatives must be continuous at the point *a*.

2. We must be able to evaluate *f* and all of its derivatives at the point *a*.

3. The point *a* must be close enough to *x* so that the approximation is good but the order of the polynomial is not unreasonably large.

Example: Suppose we wish to evaluate the function $f(x) = \ln(x)$ for some value of $x > 0$. We choose to use $a = 1$ since the derivatives of $\ln(x)$ exist and are relatively easy to evaluate at $a = 1$. By computing the first few derivatives, it is easy to see that

$$f^{(n)}(x) = \frac{(-1)^{n+1}(n-1)!}{x^n}.$$

Hence, we have

$$\begin{aligned} f(x) &= \ln(1) + \frac{(-1)^2(0!)}{1!}(x-1) + \frac{(-1)^3(1!)}{2!}(x-1)^2 + \ldots + \frac{(-1)^{n+1}(n-1)!}{n!}(x-1)^n + R_n(x) \\ &= (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \ldots + \frac{(-1)^{n+1}(x-1)^n}{n} + R_n(x), \end{aligned}$$

and

$$\left|R_n(x)\right| \le \max \frac{n!}{(n+1)!\left|\xi^{n+1}\right|}\left|x-1\right|^{n+1}$$

$$= \max \frac{\left|x-1\right|^{n+1}}{(n+1)\left|\xi^{n+1}\right|}$$

$$= \begin{cases} \dfrac{(x-1)^{n+1}}{(n+1)}, & x > 1, \\ \dfrac{(1-x)^{n+1}}{(n+1)x^{n+1}}, & x < 1. \end{cases}$$

Example: Suppose we want to approximate $\ln(1.2)$ correct to four decimal places (after rounding). In this case, we have $x = 1.2 > 1$, so

$$\left|R_n(x)\right| \le \frac{(x-1)^{n+1}}{(n+1)} = \frac{(0.2)^{n+1}}{(n+1)},$$

and we need to choose *n* large enough that

$$\frac{(0.2)^{n+1}}{(n+1)} < 0.5 \times 10^{-4}.$$

By checking the first few terms, it is easy to see that

$$(0.2)^6 \big/ 6 = .1067 \times 10^{-4} < 0.5 \times 10^{-4},$$

so we can use a 5th degree polynomial for the approximation. The final answer is

$$\begin{aligned}\ln(1.2) &\approx (1.2-1) - \frac{(1.2-1)^2}{2} + \frac{(1.2-1)^3}{3} - \frac{(1.2-1)^4}{4} + \frac{(1.2-1)^5}{5} \\ &= 0.2 + \frac{(1.2-1)^2}{2} + \frac{(0.2)^3}{3} - \frac{(0.2)^4}{4} + \frac{(0.2)^5}{5} \\ &\approx 0.1823.\end{aligned}$$

Notice that if we let $h = x - a$, then we can rewrite the Taylor series expansion as

$$f(x) = f(a) + f'(a)h + \frac{f''(a)}{2!}h^2 + \ldots + \frac{f^{(n)}(a)}{n!}h^n + R_n(x),$$

where

$$\left|R_n(x)\right| \le \frac{\max\limits_{\xi\in[a,x]}\left|f^{(n+1)}(\xi)\right|}{(n+1)!}h^{n+1}.$$

Since the truncation error (remainder term) for particular values of *x* and *n* decreases at a rate proportional to h^{n+1} as $h \to 0$, we say that the Taylor series expansion of order *n* has a truncation error of *order* $n+1$. This is generally denoted by $O\left(h^{n+1}\right)$.

Finally, note that the special case of the Taylor series expansion around $a = 0$ is often called the *MacLaurin series expansion.* Clearly, the MacLaurin series expansion takes the form

$$f(x) = f(0) + f'(0)h + \frac{f''(0)}{2!}h^2 + \ldots + \frac{f^{(n)}(0)}{n!}h^n + R_n(x).$$