CHAPTER 1 (5 LECTURES)
FLOATING POINT ARITHMETIC AND ERRORS

1. Numerical analysis

Numerical analysis, area of mathematics and computer science that creates, analyzes, and implements algorithms for obtaining numerical solutions to problems involving continuous variables. Such problems arise throughout the natural sciences, social sciences, engineering, medicine, and business. Since the mid 20th century, the growth in power and availability of digital computers has led to an increasing use of realistic mathematical models in science and engineering, and numerical analysis of increasing sophistication is needed to solve these more detailed models of the world. The formal academic area of numerical analysis ranges from quite theoretical mathematical studies to computer science issues. A major advantage for numerical technique is that a numerical answer can be obtained even when a problem has no analytical solution. However, result from numerical analysis is an approximation, in general, which can be made as accurate as desired. For example, to find the approximate values of $\sqrt{2}$, π etc.

Ancient Babylonian method (1500 BC) for approximating a square root of a number was used. This was the perhaps a first algorithm used for approximating $\sqrt{a}$ is known as the Babylonian method, despite there being no direct evidence.

In this approach, let $x = \sqrt{a}$ and e is the error in our estimate such that

$$\begin{aligned} a &= (x+e)^2 \\ &= x^2 + e^2 + 2ex \\ &= x^2 + e(e+2x) \\ e &= \frac{a-x^2}{2x+e} \approx \frac{a-x^2}{2x} \quad \text{as } e << x. \end{aligned}$$

Therefore

$$x + e \approx x + \frac{a-x^2}{2x} = \frac{a}{2x} + \frac{x}{2} = x_{\text{revised}}.$$

Since the computed error was not exact, this becomes our next best guess. The process of updating is iterated until desired accuracy is obtained.

For example, to find square root of 2, we take $a = 2$ and $x = 1$ (initial guess). Then next three values are given as

$$\begin{aligned} x_1 &= \frac{2}{2\times 1} + \frac{1}{2} = 1.5 \\ x_2 &= \frac{2}{2\times 1.5} + \frac{1.5}{2} = 1.4167 \\ x_3 &= \frac{2}{2\times 1.4167} + \frac{1.4167}{2} = 1.4142. \end{aligned}$$

We observe that x_3 is quite close to $\sqrt{2}$.

With the increasing availability of computers, the new discipline of scientific computing, or computational science, emerged during the 1980s and 1990s. The discipline combines numerical analysis, symbolic mathematical computations, computer graphics, and other areas of computer science to make it easier to set up, solve, and interpret complicated mathematical models of the real world.

1.1. **Common perspectives in numerical analysis.** Numerical analysis is concerned with all aspects of the numerical solution of a problem, from the theoretical development and understanding of numerical methods to their practical implementation as reliable and efficient computer programs. Most numerical analysts specialize in small subfields, but they share some common concerns, perspectives, and mathematical methods of analysis. These include the following:

- When presented with a problem that cannot be solved directly, they try to replace it with a "nearby problem" that can be solved more easily. Examples are the use of interpolation in developing numerical integration methods and root-finding methods.
- There is widespread use of the language and results of linear algebra, real analysis, and functional analysis (with its simplifying notation of norms, vector spaces, and operators).
- There is a fundamental concern with error, its size, and its analytic form. When approximating a problem, it is prudent to understand the nature of the error in the computed solution. Moreover, understanding the form of the error allows creation of extrapolation processes to improve the convergence behaviour of the numerical method.
- Numerical analysts are concerned with stability, a concept referring to the sensitivity of the solution of a problem to small changes in the data or the parameters of the problem. Numerical methods for solving problems should be no more sensitive to changes in the data than the original problem to be solved. Moreover, the formulation of the original problem should be stable or well-conditioned.

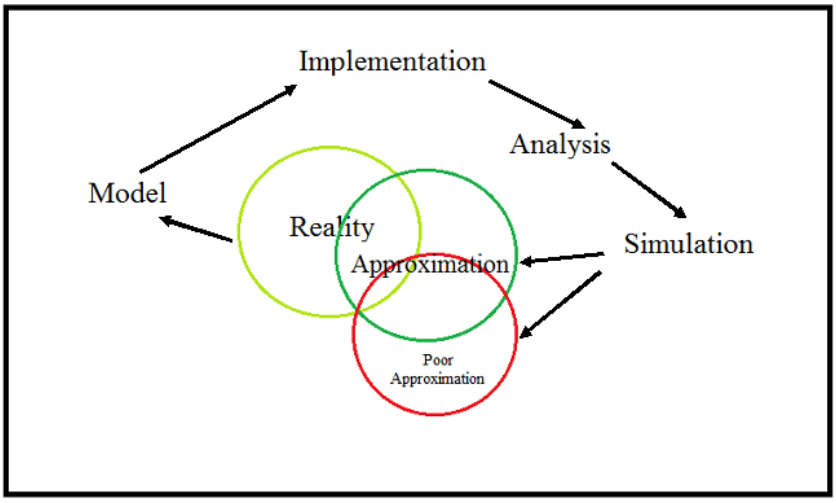

FIGURE 1. Numerical Approximations

In this chapter, we introduce and discuss some basic concepts of numerical computations. We begin with discussion of floating-point representation and then we discuss the most fundamental source of imperfection in numerical computing namely roundoff errors. We also discuss source of errors and then stability of numerical algorithms.

2. Floating-point representation of numbers

Any real number is represented by an infinite sequence of digits. For example

$$\frac{8}{3} = 2.66666\cdots = \left(\frac{2}{10^1} + \frac{6}{10^2} + \frac{6}{10^3} + \dots\right) \times 10^1.$$

This is an infinite series, but computer use an finite amount of memory to represent numbers. Thus only a finite number of digits may be used to represent any number, no matter by what representation method.

For example, we can chop the infinite decimal representation of $\frac{8}{3}$ after 4 digits,

$$\frac{8}{3} = \left(\frac{2}{10^1} + \frac{6}{10^2} + \frac{6}{10^3} + \frac{6}{10^4}\right) \times 10^1 = 0.2666 \times 10^1.$$

Generalizing this, we say that number has n decimal digits and call this n as precision.

For each real number x, we associate a floating point representation denoted by $fl(x)$, given by

$$fl(x) = \pm(0.a_1a_2\dots a_n)_\beta \times \beta^e,$$

here fraction part is called mantissa with all a_i integers and e is known as exponent. This representation is called $\beta-$based floating point representation of x and we take base $\beta = 10$ in this course.
For example,

$$\begin{aligned} 42.965 &= 4 \times 10^1 + 2 \times 10^0 + 9 \times 10^{-1} + 6 \times 10^{-2} + 5 \times 10^{-3} \\ &= 0.42965 \times 10^2. \\ -0.00234 &= -0.234 \times 10^{-2}. \end{aligned}$$

Number 0 is written as $0.00\ldots0 \times 10^e$. Likewise, we can use for binary number system and any real x can be written

$$x = \pm q \times 2^m$$

with $\frac{1}{2} \leq q \leq 1$ and some integer m. Both q and m will be expressed in terms of binary numbers. For example,

$$\begin{aligned} 1001.1101 &= 1 \times 2^3 + 1 \times 2^0 + 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-4} \\ &= (9.8125)_{10}. \end{aligned}$$

Remark 2.1. *The above representation is not unique.*
For example, $0.2666 \times 10^1 = 0.02666 \times 10^2$ *etc.*

Definition 2.1 (Normal form)**.** *A non-zero floating-point number is in normal form if the values of mantissa lies in* $(-1, -0.1]$ *or* $[0.1, 1)$*.*

Therefore, we normalize the representation by $a_1 \neq 0$. Not only the precision is limited to a finite number of digits, but also the range of exponent is also restricted. Thus there are integers m and M such that $-m \leq e \leq M$.

Definition 2.2 (Overflow and underflow)**.** *An overflow is obtained when a number is too large to fit into the floating point system in use, i.e* $e > M$*. An underflow is obtained when a number is too small, i.e* $e < -m$ *. When overflow occurs in the course of a calculation, this is generally fatal. But underflow is non-fatal: the system usually sets the number to* 0 *and continues. (Matlab does this, quietly.)*

2.1. **Rounding and chopping.** Let x be any real number and $fl(x)$ be its machine approximation. There are two ways to do the "cutting" to store a real number

$$x = \pm(0.a_1a_2\ldots a_na_{n+1}\ldots) \times 10^e, \quad a_1 \neq 0.$$

(1) Chopping: We ignore digits after a_n and write the number as following in chopping

$$fl(x) = (0.a_1a_2\ldots a_n) \times 10^e.$$

(2) Rounding: Rounding is defined as following

$$fl(x) = \begin{cases} \pm(0.a_1a_2\ldots a_n) \times 10^e, \ 0 \leq a_{n+1} < 5 & \text{(rounding down)} \\ \pm[(0.a_1a_2\ldots a_n) + (0.00\ldots01)] \times 10^e, \ 5 \leq a_{n+1} < 10 & \text{(rounding up).} \end{cases}$$

Example 1.

$$fl\left(\frac{6}{7}\right) = \begin{cases} 0.86 \times 10^0 & \text{(rounding)} \\ 0.85 \times 10^0 & \text{(chopping).} \end{cases}$$

S

3. Errors in numerical approximations

Definition 3.1 (Absolute and relative error)**.** *If* $fl(x)$ *is the approximation to the exact value* x*, then the absolute error is* $|x - fl(x)|$*, and relative error is* $\dfrac{|x - fl(x)|}{|x|}$.

Remark: As a measure of accuracy, the absolute error may be misleading and the relative error is more meaningful.

Example 2. *Find the largest interval in which* $fl(x)$ *must lie to approximate* $\sqrt{2}$ *with relative error at most* 10^{-5} *for each value of* x*.*

Sol. We have

$$\left|\frac{\sqrt{2}-fl(x)}{\sqrt{2}}\right| \leq 10^{-5}.$$

Therefore

$$\begin{aligned} |\sqrt{2}-fl(x)| &\leq \sqrt{2}\cdot 10^{-5}, \\ -\sqrt{2}\cdot 10^{-5} &\leq \sqrt{2}-fl(x) \leq \sqrt{2}\cdot 10^{-5} \\ -\sqrt{2}-\sqrt{2}\cdot 10^{-5} &\leq -fl(x) \leq -\sqrt{2}+\sqrt{2}\cdot 10^{-5} \\ \sqrt{2}+\sqrt{2}\cdot 10^{-5} &\geq fl(x) \geq \sqrt{2}-\sqrt{2}\cdot 10^{-5}. \end{aligned}$$

Hence interval (in decimals) is $[1.4141994\cdots,\ 1.4142277\cdots]$.

3.1. **Chopping and Rounding Errors.** Let x be any real number we want to represent in a computer. Let $fl(x)$ be the representation of x in the computer then what is largest possible values of $\frac{|x-fl(x)|}{|x|}$? In the worst case, how much data we are losing due to round-off errors or chopping errors?

Chopping errors: Let

$$\begin{aligned} x &= (0.a_1a_2\ldots a_na_{n+1}\ldots)\times 10^e \\ &= \left(\frac{a_1}{10}+\frac{a_2}{10^2}+\cdots+\frac{a_n}{10^n}+\frac{a_{n+1}}{10^{n+1}}+\cdots\right)\times 10^e. \\ &= \left(\sum_{i=1}^{\infty}\frac{a_i}{10^i}\right)\times 10^e,\ a_1\neq 0, \\ fl(x) &= (0.a_1a_2\ldots a_n)\times 10^e = \left(\sum_{i=1}^{n}\frac{a_i}{10^i}\right)\times 10^e. \end{aligned}$$

Therefore

$$|x-fl(x)| = \left(\sum_{i=n+1}^{\infty}\frac{a_i}{10^i}\right)\times 10^e.$$

Now since each $a_i \leq 9 = 10-1$, therefore,

$$\begin{aligned} |x-fl(x)| &\leq \sum_{i=n+1}^{\infty}\frac{10-1}{10^i}\times 10^e \\ &= (10-1)\left[\frac{1}{10^{n+1}}+\frac{1}{10^{n+2}}+\ldots\right]\times 10^e \\ &= (10-1))\left[\frac{\frac{1}{10^{n+1}}}{1-\frac{1}{10}}\right]\times 10^e \\ &= 10^{e-n}. \end{aligned}$$

Therefore absolute error bound is

$$E_a = |x-fl(x)| \leq 10^{e-n}.$$

Now

$$|x| = (0.a_1a_2\ldots a_n)_{10}\times 10^e \geq 0.1\times 10^e = \frac{1}{10}\times 10^e.$$

Therefore relative error bound is

$$E_r = \frac{|x-fl(x)|}{|x|} \leq \frac{10^{-n}\times 10^e}{10^{-1}\times 10^e} = 10^{1-n}.$$

Rounding errors: For rounding

$$fl(x) = \begin{cases} (0.a_1a_2\ldots a_n)_{10} \times 10^e = \left(\sum_{i=1}^{n} \frac{a_i}{10^i}\right) \times 10^e, \ 0 \le a_{n+1} < 5 \\ (0.a_1a_2\ldots a_{n-1}[a_n+1])_{10}\ 10^e = \left(\frac{1}{10^n} + \sum_{i=1}^{n} \frac{a_i}{10^i}\right) \times 10^e, \ 5 \le a_{n+1} < 10. \end{cases}$$

For $0 < a_{n+1} < 5 = 10/2$,

$$\begin{aligned} |x - fl(x)| &= \sum_{i=n+1}^{\infty} \frac{a_i}{10^i} \times 10^e \\ &= \left[\frac{a_{n+1}}{10^{n+1}} + \sum_{i=n+2}^{\infty} \frac{a_i}{10^i}\right] \times 10^e \\ &\le \left[\frac{10/2-1}{10^{n+1}} + \sum_{i=n+2}^{\infty} \frac{(10-1)}{10^i}\right] \times 10^e \\ &= \left[\frac{10/2-1}{10^{n+1}} + \frac{1}{10^{n+1}}\right] \times 10^e \\ &= \frac{1}{2}10^{e-n}. \end{aligned}$$

For $5 \le a_{n+1} < 10$,

$$\begin{aligned} |x - fl(x)| &= \left|\sum_{i=n+1}^{\infty} \frac{a_i}{10^i} - \frac{1}{10^n}\right| \times 10^e \\ &= \left[\frac{1}{10^n} - \sum_{i=n+1}^{\infty} \frac{a_i}{10^i}\right] \times 10^e \\ &= \left[\frac{1}{10^n} - \frac{a_{n+1}}{10^{n+1}} - \sum_{i=n+2}^{\infty} \frac{a_i}{10^i}\right] \times 10^e \\ &\le \left[\frac{1}{10^n} - \frac{a_{n+1}}{10^{n+1}}\right] \times 10^e \end{aligned}$$

Since $-a_{n+1} \le -10/2$, therefore

$$\begin{aligned} |x - fl(x)| &\le \left[\frac{1}{10^n} - \frac{10/2}{10^{n+1}}\right] \times 10^e \\ &= \frac{1}{2}10^{e-n}. \end{aligned}$$

Thus, for both the cases, absolute error bound is

$$E_a = |x - fl(x)| \le \frac{1}{2}\ 10^{e-n}.$$

Hence relative error bound is

$$E_r = \frac{|x - fl(x)|}{|x|} \le \frac{1}{2}\ \frac{10^{-n} \times 10^e}{10^{-1} \times 10^e} = \frac{1}{2}10^{1-n} = 5 \times 10^{-n}.$$

4. Accuracy and Precision

Accurate to n decimal places means that we can trust n digits to the right of the decimal place. Accurate to n significant digits means that we can trust a total of n digits as being meaningful beginning with the leftmost nonzero digit. Significant digits are digits beginning with the leftmost nonzero digit and ending with the rightmost correct digit, including final zeros that are exact.

Suppose we use a ruler graduated in millimeters to measure lengths. The measurements will be accurate to one millimeter, or 0.001 m, which is three decimal places written in meters. A measurement such as

12.345 m would be accurate to three decimal places. A measurement such as 12.3456789 m would be meaningless, since the ruler produces only three decimal places, and it should be 12.345 m or 12.346 m. If the measurement 12.345 m has five dependable digits, then it is accurate to five significant figures. On the other hand, a measurement such as 0.076 m has only two significant figures.
Mathematically, looking at an approximation 2.75303 to an actual value of 2.75194, we note that the three most significant digits are equal, and therefore one may state that the approximation has three significant digits of accuracy. One problem with simply looking at the digits is given by the following two examples:

(1) 1.9 as an approximation to 1.1 may appear to have one significant digit, but with a relative error of 0.73, this seems unreasonable.
(2) 1.9999 as an approximation to 2.0001 may appear to have no significant digits, but the relative error is 0.00010 which is almost the same relative error as the approximation 1.9239 is to 1.9237.

Thus, we need a more mathematical definition of the number of significant digits. Let the number x and approximation x^* be written in decimal form. The number of significant digits tells us to about how many positions x and x^* agree. More precisely, we say that x^* has m significant digits of x if the absolute error $|x - x^*|$ has zeros in the first m decimal places, counting from the leftmost nonzero (leading) position of x, followed by a digit from 0 to 5.
Examples:
5.1 has 1 significant digit of 5: $|5 - 5.1| = \mathbf{0}.1$.
0.51 has 1 significant digits of 0.5: $|0.5 - 0.51| = 0.\mathbf{0}1$.
4.995 has 3 significant digits of 5: $5 - 4.995 = \mathbf{0.00}5$.
4.994 has 2 significant digits of 5: $5 - 4.994 = \mathbf{0.0}06$.
0.57 has all significant digits of 0.57.
1.4 has 0 significant digits of 2: $2 - 1.4 = 0.6$. In the terms of relative errors, the number x^* is said to approximate x to m significant digits (or figures) if m is the largest nonnegative integer for which

$$\frac{|x - x^*|}{|x|} \leq 0.5 \times 10^{-m}.$$

If the relative error is greater than 0.5, then we will simply state that the approximation has zero significant digits.
For example, if we approximate π with 3.14 then relative errors is

$$E_r = \frac{|\pi - 3.14|}{\pi} \approx 0.00051 \leq 0.005 = 0.5 \times 10^{-2},$$

and therefore it is correct to two significant digits.
Also 4.994 has 2 significant digits of 5 as relative errors is $(5 - 4.994)/5 = 0.0012 = 0.12 \times 10^{-2} \leq 0.5 \times 10^{-2}$.
Some numbers are exact because they are known with complete certainty. Most exact numbers are integers: exactly 12 inches are in a foot, there might be exactly 23 students in a class. Exact numbers can be considered to have an infinite number of significant figures.

5. Rules for mathematical operations

In carrying out calculations, the general rule is that the accuracy of a calculated result is limited by the least accurate measurement involved in the calculation. In addition and subtraction, the result is rounded off so that it has the same number of digits as the measurement having the fewest decimal places (counting from left to right). For example, $100 + 23.643 = 123.643$, which should be rounded to 124.
Let the floating-point representations $fl(x)$ and $fl(y)$ are given for the real numbers x and y and that the symbols $\oplus$, $\ominus$, $\otimes$ and $\oslash$ represent machine addition, subtraction, multiplication, and division operations, respectively. We will assume a finite-digit arithmetic given by

$$x \oplus y = fl(fl(x) + fl(y)), \qquad x \ominus y = fl(fl(x) - fl(y)),$$
$$x \otimes y = fl(fl(x) \times fl(y)), \qquad x \oslash y = fl(fl(x) \div fl(y)).$$

This arithmetic corresponds to performing exact arithmetic on the floating-point representations of x and y and then converting the exact result to its finite-digit floating-point representation.

Example 3. *Suppose that* $x = \frac{5}{7}$ *and* $y = \frac{1}{3}$. *Use five-digit chopping for calculating* $x+y$, $x-y$, $x\times y$, *and* $x \div y$.

Sol. Here $x = \frac{5}{7} = 0.714285\cdots$ and $y = \frac{1}{3} = 0.33333\cdots$.
Using the five-digit chopping values of x and y are

$$fl(x) = 0.71428 \times 10^0 \quad \text{and} \quad fl(y) = 0.33333 \times 10^0.$$

Thus,

$$x \oplus y = fl(fl(x) + fl(y)) = fl(0.71428 \times 10^0 + 0.33333 \times 10^0) = fl(1.04761 \times 10^0) = 0.10476 \times 10^1.$$

The true value is $x + y = \frac{5}{7} + \frac{1}{3} = \frac{22}{21}$, so we have
Absolute Error $E_a = |\frac{22}{21} - 0.10476 \times 10^1| = 0.190 \times 10^{-4}$.
Relative Error $E_r = \dfrac{0.190 \times 10^{-4}}{\frac{22}{21}} = 0.182 \times 10^{-4}$.
Similarly we can perform other calculations.
Further we show some examples of arithmetic with different exponents.

Example 4. (1) *Add the following floating-point numbers* $0.4546e3$ *and* $0.5433e7$.
(2) *Subtract the following floating-point numbers:* $0.5424e-99$ *from* $0.5452e-99$.
(3) *Multiply the following floating point numbers:* $0.1111e74$ *and* $0.2000e80$.

Sol.

(1) This problem contains unequal exponent. To add these floating-point numbers, take operands with the largest exponent as,

$$0.5433e7 + 0.0000e7 = 0.5433e7.$$

(Because $0.4546e3$ changes in the same operand as $0.0000e7$).
(2) On subtracting we get $0.0028e-99$. Again this is a floating-point number but not in the normalized form. To convert it in normalized form, shift the mantissa to the left. Therefore we get $0.28e-101$. This condition is called an underflow condition.
(3) On multiplying we obtain $0.1111e74 \times 0.2000e80 = 0.2222e153$. This shows overflow condition of normalized floating-point numbers.

We can use derivatives to find errors as shown in the next example.

Example 5. *The error in the measurement of area of a circle is not allowed to exceed* 0.5%. *How accurately the radius should be measured.*

Sol. Area of the circle is $A = \pi r^2$ (say).

$$\therefore \frac{dA}{dr} = 2\pi r.$$

Relative error (in percentage) in area is

$$\begin{aligned} \frac{dA}{A} \times 100 &\leq 0.5 \\ \implies dA &\leq \frac{0.5 \times A}{100} = \frac{0.5\pi r^2}{100} = \frac{1}{200}\pi r^2. \end{aligned}$$

Relative error (in percentage) in radius is therefore

$$\begin{aligned} \frac{dr}{r} \times 100 &= \frac{100}{r}\frac{dA}{\frac{dA}{dr}} \\ &\leq \frac{100}{r} \times \frac{\pi r^2}{200 \times 2\pi r} = 0.25. \end{aligned}$$

6. Loss of Accuracy

Round-off errors are inevitable and difficult to control. Other types of errors which occur in computation may be under our control. The subject of numerical analysis is largely preoccupied with understanding and controlling errors of various kinds.
One of the most common error-producing calculations involves the cancellation of digits due to the subtractions nearly equal numbers (or the addition of one very large number and one very small number or multiplication of a small number with a quite large number). The loss of accuracy due to round-off error can often be avoided by a reformulation of the calculations, as illustrated in the next example.

Example 6. *Use four-digit rounding arithmetic and the formula for the roots of a quadratic equation, to find the most accurate approximations to the roots of the following quadratic equation. Compute the absolute and relative errors.*

$$1.002x^2 + 11.01x + 0.01265 = 0.$$

Sol. The quadratic formula states that the roots of $ax^2 + bx + c = 0$ are

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Using the above formula, the roots of given eq. $1.002x^2 + 11.01x + 0.01265 = 0$ are approximately (using long format)

$$x_1 = -0.00114907565991, \quad x_2 = -10.98687487643590.$$

We use four-digit rounding arithmetic to find approximations to the roots. We write the approximations of root as x_1^* and x_2^*. These approximations are given by

$$\begin{aligned} x_{1,2}^* &= \frac{-11.01 \pm \sqrt{(-11.01)^2 - 4 \cdot 1.002 \cdot 0.01265}}{2 \cdot 1.002} \\ &= \frac{-11.01 \pm \sqrt{121.2 - 0.05070}}{2.004} \\ &= \frac{-11.01 \pm 11.00}{2.004}. \end{aligned}$$

Therefore, we find the first root:

$$x_1^* = -0.004990,$$

which has the absolute error $|x_1 - x_1^*| = 0.00384095$ and relative error $\frac{|x_1 - x_1^*|}{|x_1|} = 3.34265968$ (very high). We find the second root

$$x_2^* = \frac{-11.01 - 11.00}{2.004} = -10.98,$$

which has the following absolute error

$$|x_2 - x_2^*| = 0.006874876,$$

and relative error

$$\frac{|x_2 - x_2^*|}{|x_2|} = 0.000626127.$$

This quadratic formula for the calculation of first root, encounter the subtraction of nearly equal numbers and cause loss of accuracy. Here b and $\sqrt{b^2 - 4ac}$ become two equal numbers. Therefore, we use the alternate quadratic formula by rationalize the expression to calculate first root and approximation is given by

$$\begin{aligned} x_1^* &= \frac{-b + \sqrt{b^2 - 4ac}}{2a} \, \frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}} \\ &= \frac{-2c}{b + \sqrt{b^2 - 4ac}} \\ &= -0.001149, \end{aligned}$$

which has the following very small relative error

$$\frac{|x_1 - x_1^*|}{|x_1|} = 6.584 \times 10^{-5}.$$

Remark 6.1. *However, if rationalize the numerator in x_2 to get*

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}.$$

The use of this formula results not only involve the subtraction of two nearly equal numbers but also division by the small number. This would cause degrade in accuracy.

Remark 6.2. *Since product of the roots for a quadratic is c/a. Thus we can find the approximation of the first root from the second as*

$$x_1^* = \frac{c}{ax_2^*}.$$

Example 7. *The quadratic formula is used for computing the roots of equation $ax^2 + bx + c = 0,\ a \neq 0$ and roots are given by*

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Consider the equation $x^2 + 62.10x + 1 = 0$ and discuss the numerical results.

Sol. Using quadratic formula and 8-digit rounding arithmetic, we obtain two roots

$$x_1 = -0.01610723$$

$$x_2 = -62.08390.$$

We use these values as "exact values". Now we perform calculations with 4-digit rounding arithmetic. We have $\sqrt{b^2 - 4ac} = \sqrt{62.10^2 - 4.000} = \sqrt{3856 - 4.000} = 62.06$ and

$$fl(x_1) = \frac{-62.10 + 62.06}{2.000} = -0.02000.$$

The relative error in computing of x_1 is

$$\frac{|fl(x_1) - x_1|}{|x_1|} = \frac{|-0.02000 + .01610723|}{|-0.01610723|} = 0.2417.$$

In calculating of x_2,

$$fl(x_2) = \frac{-62.10 - 62.06}{2.000} = -62.10.$$

The relative error in computing x_2 is

$$\frac{|fl(x_2) - x_2|}{|x_2|} = \frac{|-62.10 + 62.08390|}{|-62.08390|} = 0.259 \times 10^{-3}.$$

In this equation since $b^2 = 62.10^2$ is much larger than $4ac = 4$. Hence b and $\sqrt{b^2 - 4ac}$ become two equal numbers. Calculation of x_1 involves the subtraction of nearly two equal numbers but x_2 involves the addition of the nearly equal numbers which will not cause serious loss of significant figures.
To obtain a more accurate 4-digit rounding approximation for x_1, we change the formulation by rationalizing the numerator, that is,

$$x_1 = \frac{-2c}{b + \sqrt{b^2 - 4ac}}.$$

Then

$$fl(x_1) = \frac{-2.000}{62.10 + 62.06} = -2.000/124.2 = -0.01610.$$

The relative error in computing x_1 is now reduced to 0.62×10^{-3}.

Example 8. *Suppose that the values of $f(x) = x - \sin x$ are required near $x = 0$. Suggest a way to calculate these function values.*

Sol. One cure for this problem is to use the Taylor series for $\sin x$:

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

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

$$f(x) = x - \sin x = x - \left[x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots\right] = \frac{x^3}{3!} - \frac{x^5}{5!} + \cdots$$

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.

Nested Arithmetic: Accuracy loss due to round-off error can also be reduced by rearranging calculations, as shown in the next example. Polynomials should always be expressed in nested form before performing an evaluation, because this form minimizes the number of arithmetic calculations. One way to reduce round-off error is to reduce the number of computations.

Example 9. *Evaluate* $f(x) = x^3 - 6.1x^2 + 3.2x + 1.5$ *at* $x = 4.71$ *using three-digit arithmetic directly and with nesting.*

Sol. The exact result of the evaluation is (by taking more digits):
Exact: $f(4.71) = 4.71^3 - 6.1 \times 4.71^2 + 3.2 \times 4.71 + 1.5 = -14.263899$.
Now using three-digit chopping then

$$\begin{aligned} f(4.71) &= 4.71^3 - 6.1 \times 4.71^2 + 3.2 \times 4.71 + 1.5 \\ &= 22.1 \times 4.71 - 6.1 \times 22.1 + 15.0 + 1.5 \\ &= 104 - 134 + 15.0 + 1.5 = -13.5. \end{aligned}$$

Relative error is

$$\left|\frac{-14.263899 + 13.5}{-14.263899}\right| \approx 0.05.$$

Similarly if we use three-digit rounding arithmetic, we obtain

$$\begin{aligned} f(4.71) &= 4.71^3 - 6.1 \times 4.71^2 + 3.2 \times 4.71 + 1.5 \\ &= 22.2 \times 4.71 - 6.1 \times 22.2 + 15.1 + 1.5 \\ &= 105 - 135 + 15.1 + 1.5 = -13.4. \end{aligned}$$

The relative error in case is

$$\left|\frac{-14.263899 + 13.4}{-14.263899}\right| \approx 0.06.$$

As an alternative approach, we write the polynomial $f(x)$ in a nested manner as

$$f(x) = ((x - 6.1)x + 3.2)x + 1.5.$$

Using three-digit chopping arithmetic now produces

$$\begin{aligned} f(4.71) &= ((4.71 - 6.1)4.71 + 3.2)4.71 + 1.5 \\ &= (-1.39)4.71 + 3.2)4.71 + 1.5 \\ &= (-6.54 + 3.2)4.71 + 1.5 \\ &= (-3.34)4.71 + 1.5 \\ &= -15.7 + 1.5 = -14.2. \end{aligned}$$

In a similar fashion, we can obtain a three-digit rounding and answer is -14.3.
The relative error in case of three-digit (chopping) is

$$\left|\frac{-14.263899 + 14.2}{-14.263899}\right| \approx 0.0045,$$

and for three-digit (rounding) is

$$\left|\frac{-14.263899+14.3}{-14.263899}\right| \approx 0.0025.$$

Nesting has reduced the relative errors for both the approximations. Moreover in original form, there are 7 multiplications while nested form has only 2 multiplications. Thus nested form reduce errors as well as number of calculations.

7. Algorithms and Stability

An algorithm is a procedure that describes, in an unambiguous manner, a finite sequence of steps to be performed in a specified order. The object of the algorithm is to implement a procedure to solve a problem or approximate a solution to the problem. One criterion we will impose on an algorithm whenever possible is that small changes in the initial data produce correspondingly small changes in the final results. An algorithm that satisfies this property is called stable; otherwise it is unstable. Some algorithms are stable only for certain choices of initial data, and are called conditionally stable. The words condition and conditioning are used to indicate how sensitive the solution of a problem may be to small changes in the input data. A problem is well-conditioned if small changes in the input data can produce only small changes in the results. On the other hand, a problem is ill-conditioned if small changes in the input data can produce large changes in the output.
For a certain types of problems, a condition number can be defined. If that number is large (greater than one), it indicates an ill-conditioned problem. In contrast, if the number is modest (up to one), the problem is recognized as a well-conditioned problem.
The condition number can be calculated in the following manner:

$$\begin{aligned}\kappa &= \frac{\text{relative change in output}}{\text{relative change in input}}\\ &= \frac{\left|\dfrac{f(x)-f(x^*)}{f(x)}\right|}{\left|\dfrac{x-x^*}{x}\right|}\\ &\approx \left|\frac{xf'(x)}{f(x)}\right|.\end{aligned}$$

For example, if $f(x)=\dfrac{10}{1-x^2}$, then the condition number can be calculated as

$$\kappa = \left|\frac{xf'(x)}{f(x)}\right| = \frac{2x^2}{|1-x^2|}.$$

Condition number can be quite large for $|x| \approx 1$. Therefore, the function is ill-conditioned.

Example 10. *Compute and interpret the condition number for*
(a) $f(x)=\sin x$ *for* $x=0.51\pi$.
(b) $f(x)=\tan x$ *for* $x=1.7$.

Sol. (a) The condition number is given by

$$\kappa = \left|\frac{xf'(x)}{f(x)}\right|.$$

For $x=0.51\pi$, $f'(x)=\cos(0.51\pi)=-0.03141$, $f(x)=\sin(0.51\pi)=0.99951$.

$$\therefore \kappa = 0.05035 < 1.$$

Since, the condition number is < 1, we conclude that the relative error is attenuated.
(b) $f(x)=\tan x$, $f(1.7)=-7.6966$, $f'(x)=1/\cos^2 x$, $f'(a)=1/\cos^2(1.7)=60.2377$.

$$\kappa = 13.305 >> 1.$$

Thus, the function is ill-conditioned.

7.1. **Creating Algorithms.** Another theme that occurs repeatedly in numerical analysis is the distinction between numerical algorithms are stable and those that are not. Informally speaking, a numerical process is unstable if small errors made at one stage of the process are magnified and propagated in subsequent stages and seriously degrade the accuracy of the overall calculation.
An algorithm can be thought of as a sequence of problems, i.e., a sequence of function evaluations. In this case we consider the algorithm for evaluating $f(x)$ to consist of the evaluation of the sequence $x_1, x_2, \cdots, x_n$. We are concerned with the condition of each of the functions $f_1(x_1), f_2(x_2), \cdots, f_{n-1}(x_{n-1})$ where $f(x) = f_i(x_i)$ for all i. An algorithm is unstable if any f_i is ill-conditioned, i.e. if any $f_i(x_i)$ has condition much worse than $f(x)$. In the following, we study an example to create a stable algorithm.

Example 11. *Write an algorithm to calculate the expression $f(x) = \sqrt{x+1} - \sqrt{x}$, when x is quite large. By considering the condition number κ of the subproblem of evaluating the function, show that such a function evaluation is not stable. Suggest a modification which makes it stable.*

Sol. Consider

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

so that there is potential loss of significance when x is large. Taking $x = 12345$ as an example, one possible algorithm is

$$\begin{aligned}
x_0 : &= x = 12345 \\
x_1 : &= x_0 + 1, \qquad \kappa(x_0) = \left| \frac{x_0 \cdot 1}{x_0 + 1} \right| < 1 \\
x_2 : &= \sqrt{x}_1, \qquad \kappa(x_1) = \left| \frac{x_1 \cdot \frac{1}{2\sqrt{x_1}}}{\sqrt{x}_1} \right| = \frac{1}{2} \\
x_3 : &= \sqrt{x}_0, \qquad \kappa(x_0) = \frac{1}{2} \\
x_4 : &= x_2 - x_3.
\end{aligned}$$

The loss of significance occurs with the final subtraction. We can rewrite the last step in the form $f(x_3) = x_2 - x_3$ to show how the final answer depends on x_3. As $f'(x_3) = -1$, we have the condition number

$$\kappa(x_3) = \left| \frac{x_3 f'(x_3)}{f(x_3)} \right| = \left| \frac{x_3}{x_2 - x_3} \right| \approx 24690.5.$$

Note that this is the condition of a subproblem arrived at during the algorithm.
To find an alternative algorithm, we write

$$f(x) = (\sqrt{x+1} - \sqrt{x}) \frac{\sqrt{x+1} + \sqrt{x}}{\sqrt{x+1} + \sqrt{x}} = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

This suggests the algorithm

$$\begin{aligned}
x_0 : &= x = 12345 \\
x_1 : &= x_0 + 1 \\
x_2 : &= \sqrt{x}_1 \\
x_3 : &= \sqrt{x}_0 \\
x_4 : &= x_2 + x_3, \qquad \kappa(x_3) = \left| \frac{x_3 \cdot 1}{x_2 + x_3} \right| < 1 \\
x_5 : &= 1/x_4, \qquad \kappa(x_4) = \left| \frac{x_4 \cdot \frac{-1}{x_4^2}}{\frac{1}{x_4}} \right| = 1.
\end{aligned}$$

Thus first algorithm is unstable and second is stable for large values of x. In general such analyses are not usually so straightforward but, in principle, stability can be analysed by examining the condition of a sequence of subproblems.

Example 12. *Write an algorithm to calculate the expression* $f(x) = \sin(a+x) - \sin a$, *when* $x = 0.0001$. *By considering the condition number* κ *of the subproblem of evaluating the function, show that such a function evaluation is not stable. Suggest a modification which makes it stable.*

Sol. Let $x = 0.0001$ and let $a = 2$.

$$\begin{aligned}
x_0 &= 0.0001 \\
x_1 &= a + x_0, \qquad \kappa(x_0) = \left|\frac{x_0 \cdot 1}{a + x_0}\right| < 1 \\
x_2 &= \sin x_1, \qquad \kappa(x_1) = \left|\frac{x_1 \cdot \cos(x_1)}{\sin(x_1)}\right| \approx 1 \\
x_3 &= \sin a \\
x_4 &= x_2 - x_3, \qquad \kappa(x_3) = \left|\frac{x_3 \cdot (-1)}{x_2 - x_3}\right| \approx 21848.
\end{aligned}$$

We obtain a very larger condition number, which shows that the last step is not stable.
Thus we need to modify the above algorithm and write the equivalent form

$$f(x) = \sin(a+x) - \sin a = 2\sin(x/2)\cos(a + x/2).$$

The modified algorithm is the following

$$\begin{aligned}
x_0 &= 0.0001 \\
x_1 &= x_0/2, \qquad \kappa(x_0) = \left|\frac{x_0 \cdot 1/2}{x_0/2}\right| = 1 \\
x_2 &= \sin x_1 \\
x_3 &= \cos(a + x_1), \qquad \kappa(x_1) = \left|\frac{x_1 \cdot (-\sin(a+x_1))}{\cos(a+x_1)}\right| = 0.0001 \\
x_4 &= 2x_2x_3, \qquad \kappa(x_3) = \left|\frac{x_3 \cdot 2x_2}{2x_2x_3}\right| = 1.
\end{aligned}$$

Thus the condition number is one, so this form is acceptable.

Remarks

(1) Accuracy tells us the closeness of computed solution to true solution of problem. Accuracy depends on conditioning of problem as well as stability of algorithm.
(2) Stability alone does not guarantee accurate results. Applying stable algorithm to well-conditioned problem yields accurate solution. Inaccuracy can result from applying stable algorithm to ill-conditioned problem or unstable algorithm to well-conditioned problem.

Exercises

(1) Compute the absolute error and relative error in approximations of x by x^*.
   (a) $x = \pi,\ x^* = 22/7$.
   (b) $x = \sqrt{2},\ x^* = 1.414$.
   (c) $x = 8!,\ x^* = 39900$.
(2) Find the largest interval in which x^* must lie to approximate x with relative error at most 10^{-4} for each value of x.
   (a) π.
   (b) e.
   (c) $\sqrt{3}$.
   (d) $\sqrt[3]{7}$.
(3) A rectangular parallelepiped has sides of length 3 cm, 4 cm, and 5 cm, measured to the nearest centimeter. What are the best upper and lower bounds for the volume of this parallelepiped? What are the best upper and lower bounds for the surface area?
(4) Use three-digit rounding arithmetic to perform the following calculations. Compute the absolute error and relative error with the exact value determined to at least five digits.
   (a) $\sqrt{3} + (\sqrt{5} + \sqrt{7})$.
   (b) $(121 - 0.327) - 119$.

(c) $-10\pi + 6e - \dfrac{3}{62}$.

(d) $\dfrac{\pi - 22/7}{1/17}$.

(5) Use four-digit rounding arithmetic and the formula to find the most accurate approximations to the roots of the following quadratic equations. Compute the absolute errors and relative errors.

$$\frac{1}{3}x^2 - \frac{123}{4}x + \frac{1}{6} = 0.$$

(6) Find the root of smallest magnitude of the equation $x^2 - 1000x + 25 = 0$ using quadratic formula. Work in floating-point arithmetic using a four-decimal place mantissa.

(7) The derivative of $f(x) = \dfrac{1}{(1-3x^2)}$ is given by $\dfrac{6x}{(1-3x^2)^2}$. Do you expect to have difficulties evaluating this derivative at $x = 0.577$? Try it using 3- and 4-digit arithmetic with chopping.

(8) Suppose two points (x_0, y_0) and (x_1, y_1) are on a straight line with $y_1 \neq y_0$. Two formulas are available to find the x-intercept of the line:

$$x = \frac{x_0 y_1 - x_1 y_0}{y_1 - y_0}, \text{ and } x = x_0 - \frac{(x_1 - x_0)y_0}{y_1 - y_0}.$$

(a) Show that both formulas are algebraically correct.

(b) Use the data $(x_0, y_0) = (1.31, 3.24)$ and $(x_1, y_1) = (1.93, 4.76)$ and three-digit rounding arithmetic to compute the x-intercept both ways. Which method is better and why?

(9) Verify that the functions $f(x)$ and $g(x)$ are identical functions.

$$f(x) = 1 - \sin x, \quad g(x) = \frac{\cos^2 x}{1 + \sin x}.$$

(a) Which function should be used for computations when x is near $\pi/2$? Why?

(b) Which function should be used for computations when x is near $3\pi/2$? Why?

(10) Consider the identity

$$\int_0^x \sin(xt)dt = \frac{1 - \cos(x^2)}{x}.$$

Explain the difficulty in using the right-hand fraction to evaluate this expression when x is close to zero. Give a way to avoid this problem and be as precise as possible.

(11) Assume 3-digit mantissa with rounding

(a) Evaluate $y = x^3 - 3x^2 + 4x + 0.21$ for $x = 2.73$.

(b) Evaluate $y = [(x-3)x + 4]x + 0.21$ for $x = 2.73$.

Compare and discuss the errors obtained in part (a) and (b).

(12) How many multiplications and additions are required to determine a sum of the form

$$\sum_{i=1}^{n} \sum_{j=1}^{i} a_i \, b_j \; ?$$

Modify the sum to an equivalent form that reduces the number of computations.

(13) Let $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial, and let x_0 be given. Construct an algorithm to evaluate $P(x_0)$ using nested multiplication.

(14) Construct an algorithm that has as input an integer $n \geq 1$, numbers $x_0, x_1, \cdots, x_n$, and a number x and that produces as output the product $(x - x_0)(x - x_1)\cdots(x_0 - x_n)$.

(15) Consider the stability (by calculating the condition number) of $\sqrt{1+x} - 1$ when x is near 0. Rewrite the expression to rid it of subtractive cancellation.

(16) Show that the computation of

$$f(x) = \frac{e^x - 1}{x}$$

is unstable for small value of x. Rewrite the expression to make it stable.

(17) Suppose that a function $f(x) = \ln(x+1) - \ln(x)$, is computed by the following algorithm for large values of x using six digit rounding arithmetic

$$\begin{aligned} x_0 : &= x = 12345 \\ x_1 : &= x_0 + 1 \\ x_2 : &= \ln x_1 \\ x_3 : &= \ln x_0 \\ f(x) := x_4 : &= x_2 - x_3. \end{aligned}$$

By considering the condition $\kappa(x_3)$ of the subproblem of evaluating the function, show that such a function evaluation is not stable. Also propose the modification of function evaluation so that algorithm will become stable.

Bibliography

[Burden] Richard L. Burden, J. Douglas Faires, and Annette Burden, "Numerical Analysis," Cengage Learning, 10th edition, 2015.

[Cheney] E. Ward Cheney and David R. Kincaid, "Numerical Mathematics and Computing", Cengage Learning, 7th edition, 2012.