Contents 30

Introduction to Numerical Methods

Learning outcomes

In this Workbook you will learn about some of the issues involved with using a computer to carry out numerical calculations for engineering problems. For example, the effect of rounding error will be discussed.

Most of this Workbook will consider methods for solving systems of equations. In particular you will see how methods can be adapted so that rounding error becomes less of a problem.

Rounding Error and Conditioning

30.1

Introduction

In this first Section concerning numerical methods we will discuss some of the issues involved with doing arithmetic on a computer. This is an important aspect of engineering. Numbers cannot, in general, be represented exactly, they are typically stored to a certain number of **significant figures**. The associated **rounding error** and its accumulation are important issues which need to be appreciated if we are to trust computational output.

We will also look at ill-conditioned problems which can have an unfortunate effect on rounding error.

Prerequisites

Before starting this Section you should ...

- recall the formula for solving quadratic equations

Learning Outcomes

On completion you should be able to ...

- round real numbers and know what the associated rounding error is
- understand how rounding error can grow in calculations
- explain what constitutes an ill-conditioned problem

1. Numerical methods

Many mathematical problems which arise in the modelling of engineering situations are too difficult, or too lengthy, to tackle by hand. Instead it is often good enough to resort to an approximation given by a computer. Indeed, the process of modelling a "real world" situation with a piece of mathematics will involve some approximation, so it may make things no worse to seek an approximate solution of the theoretical problem.

Evidently there are certain issues here. Computers do not know what a function is, or a vector, or an integral, or a polynomial. Loosely speaking, all computers can do is remember long lists of numbers and then process them (very quickly!). Mathematical concepts must be posed as something **numerical** if a computer is to be given a chance to help. For this reason a topic known as **numerical analysis** has grown in recent decades which is devoted to the study of how to get a machine to address a mathematical problem.

Key Point 1

"Numerical methods" are methods devised to solve mathematical problems on a computer.

2. Rounding

In general, a computer is unable to store every decimal place of a real number. Real numbers are **rounded**. To round a number to n significant figures we look at the $(n+1)^{\text{th}}$ digit in the decimal expansion of the number.

- If the $(n+1)^{\text{th}}$ digit is 0, 1, 2, 3 or 4 then we **round down**: that is, we simply chop to n places. (In other words we neglect the $(n+1)^{\text{th}}$ digit and any digits to its right.)
- If the $(n+1)^{\text{th}}$ digit is 5, 6, 7, 8 or 9 then we **round up**: we add 1 to the n^{th} decimal place and then chop to n places.

For example

$$\frac{1}{3} = 0.3333 \qquad \text{rounded to 4 significant figures,}$$

$$\frac{8}{3} = 2.66667 \qquad \text{rounded to 6 significant figures,}$$

$$\pi = 3.142 \qquad \text{rounded to 4 significant figures.}$$

An alternative way of stating the above is as follows

$\frac{1}{3} = 0.3333$ rounded to 4 decimal places,

$\frac{8}{3} = 2.66667$ rounded to 5 decimal places,

$\pi = 3.142$ rounded to 3 decimal places.

Sometimes the phrases "significant figures" and "decimal places" are abbreviated as "s.f." or "sig. fig." and "d.p." respectively.

Example 1

Write down each of these numbers rounding them to 4 decimal places:
0.12345, -0.44444, 0.5555555, 0.000127351, 0.000005

Solution

0.1235, -0.4444, 0.5556, 0.0001, 0.0000

Example 2

Write down each of these numbers, rounding them to 4 significant figures:
0.12345, -0.44444, 0.5555555, 0.000127351, 25679

Solution

0.1235, -0.4444, 0.5556, 0.0001274, 25680

Write down each of these numbers, rounding them to 3 decimal places:
0.87264, 0.1543, 0.889412, -0.5555

Your solution

Answer

0.873, 0.154, 0.889, -0.556

Rounding error

Clearly, rounding a number introduces an error. Suppose we know that some quantity x is such that

$$x = 0.762143 \qquad 6 \text{ d.p.}$$

Based on what we know about the rounding process we can deduce that

$$x = 0.762143 \quad \pm \quad 0.5 \times 10^{-6}.$$

This is typical of what can occur when dealing with numerical methods. We do not know what value x takes, but we have an **error bound** describing the furthest x can be from the stated value 0.762143. Error bounds are necessarily pessimistic. It is very likely that x is closer to 0.762143 than 0.5×10^{-6}, but we cannot assume this, we have to assume the worst case if we are to be certain that the error bound is safe.

Key Point 2

Rounding a number to n decimal places introduces an error that is no larger (in magnitude) than

$$\frac{1}{2} \times 10^{-n}$$

Note that successive rounding can increase the associated rounding error, for example

$$\begin{aligned} 12.3456 &= 12.3 \ (1 \text{ d.p.}), \\ 12.3456 = 12.346 \ (3 \text{ d.p.}) &= 12.35 \ (2 \text{ d.p.}) = 12.4 \ (1 \text{ d.p.}). \end{aligned}$$

Accumulated rounding error

Rounding error can sometimes grow as calculations progress. Consider these examples.

Example 3

Let $x = \dfrac{22}{7}$ and $y = \pi$. It follows that, to 9 decimal places

$$\begin{aligned} x &= 3.142857143 \\ y &= 3.141592654 \\ x + y &= 6.284449797 \\ x - y &= 0.001264489 \end{aligned}$$

(i) Round x and y to 7 significant figures. Find $x + y$ and $x - y$.

(ii) Round x and y to 3 significant figures. Find $x + y$ and $x - y$.

Solution

(i) To 7 significant figures $x = 3.142857$ and $y = 3.141593$ and it follows that, with this rounding of the numbers

$$\begin{aligned} x + y &= 6.284450 \\ x - y &= 0.001264. \end{aligned}$$

The outputs ($x + y$ and $x - y$) are as accurate to as many decimal places as the inputs (x and y). Notice however that the difference $x - y$ is now only accurate to 4 significant figures.

(ii) To 3 significant figures $x = 3.14$ and $y = 3.14$ and it follows that, with this rounding of the numbers

$$\begin{aligned} x + y &= 6.28 \\ x - y &= 0. \end{aligned}$$

This time we have no significant figures accurate in $x - y$.

In Example 3 there was loss of accuracy in calculating $x-y$. This shows how rounding error can grow with even simple arithmetic operations. We need to be careful when developing numerical methods that rounding error does not grow. What follows is another case when there can be a loss of accurate significant figures.

This Task involves solving the quadratic equation

$$x^2 + 30x + 1 = 0$$

(a) Use the quadratic formula to show that the two solutions of $x^2+30x+1 = 0$ are $x = -15 \pm \sqrt{224}$.

(b) Write down the two solutions to as many decimal places as your calculator will allow.

(c) Now round $\sqrt{224}$ to 4 significant figures and recalculate the two solutions.

(d) How many accurate significant figures are there in the solutions you obtained with the rounded approximation to $\sqrt{224}$?

Your solution

Answer

(a) From the quadratic formula $x = \dfrac{-30 \pm \sqrt{30^2 - 4}}{2} = -15 \pm \sqrt{15^2 - 1} = -15 \pm \sqrt{224}$ as required.

(b) $-15 + \sqrt{224} = -0.03337045291$ is one solution and $-15 - \sqrt{224} = -29.96662955$ is the other, to 10 significant figures.

(c) Rounding $\sqrt{224}$ to 4 significant figures gives

$$-15 + \sqrt{224} = -15 + 14.97 = -0.03 \qquad -15 - \sqrt{224} = -15 - 14.97 = -29.97$$

(d) The first of these is only accurate to 1 sig. fig., the second is accurate to 4 sig. fig.

In the previous Task it was found that rounding to 4 sig. fig. led to a result with a large error for the smaller root of the quadratic equation. Use the fact that for the general quadratic

$$ax^2 + bx + c = 0$$

the product of the two roots is $\dfrac{c}{a}$ to determine the smaller root with improved accuracy.

Your solution

Answer

Here $a = 1,\ b = 30,\ c = 1$ so the product of the roots $= \dfrac{c}{a} = 1$. So starting from the rounded value -29.97 for the larger root we obtain the smaller root to be $\dfrac{1}{-29.97} \approx -0.03337$ with 4 sig. fig. accuracy.

(This indirect method is often built into computer software to increase accuracy.)

3. Well-conditioned and ill-conditioned problems

Suppose we have a mathematical problem that depends on some input data. Now imagine altering the input data by a *tiny* amount. If the corresponding solution always varies by a correspondingly tiny amount then we say that the problem is **well-conditioned**. If a *tiny* change in the input results in a *large* change in the output we say that the problem is **ill-conditioned**. The following Example should help.

Example 4

Show that the evaluation of the function $f(x) = x^2 - x - 1500$ near $x = 39$ is an ill-conditioned problem.

Solution

Consider $f(39) = -18$ and $f(39.1) = -10.29$. In changing x from 39 to 39.1 we have altered it by about 0.25%. But the percentage change in f is greater than 40%. The demonstrates the ill-conditioned nature of the problem.

Work out the derivative $\dfrac{df}{dx}$ for the function used in Example 4 and so explain why the numerical results show the calculation of f to be ill-conditioned near $x = 39$.

Your solution

Answer

We have $f = x^2 - x - 1500$ and $\dfrac{df}{dx} = 2x - 1$. At $x = 39$ the value of f is -18 and, using calculus, the value of $\dfrac{df}{dx}$ is 77. Thus $x = 39$ is very close to a zero of f (i.e. a root of the quadratic equation $f(x) = 0$). The fractional change in f is thus very large even for a small change in x. The given values of $f(38.6)$ and $f(39.4)$ lead us to an estimate of

$$\frac{12.96 - (-48.64)}{39.4 - 38.6}$$

for $\dfrac{df}{dx}$. This ratio gives the value 77.0, which agrees exactly with our result from the calculus. Note, however, that an exact result of this kind is not usually obtained; it is due to the simple quadratic form of f for this example.

One reason that this matters is because of rounding error. Suppose that, in the Example above, we know is that x is equal to 39 to 2 significant figures. Then we have no chance at all of evaluating f with confidence, for consider these values

$$\begin{aligned} f(38.6) &= -48.64 \\ f(39) &= -18 \\ f(39.4) &= 12.96. \end{aligned}$$

All of the arguments on the left-hand sides are equal to 39 to 2 significant figures so all the values on the right-hand sides are contenders for $f(x)$. The ill-conditioned nature of the problem leaves us with some serious doubts concerning the value of f.

It is enough for the time being to be aware that ill-conditioned problems exist. We will discuss this sort of thing again, and how to combat it in a particular case, in a later Section of this Workbook.

Exercises

1. Round each of these numbers to the number of places or figures indicated

   (a) 23.56712 (to 2 decimal places).

   (b) -15432.1 (to 3 significant figures).

2. Suppose we wish to calculate

   $$\sqrt{x+1} - \sqrt{x},$$

   for relatively large values of x. The following table gives values of y for a range of x-values

x	$\sqrt{x+1} - \sqrt{x}$
100	0.04987562112089
1000	0.01580743742896
10000	0.00499987500625
100000	0.00158113487726

   (a) For each x shown in the table, and working to 6 significant figures evaluate $\sqrt{x+1}$ and then $\sqrt{x}$. Find $\sqrt{x+1} - \sqrt{x}$ by taking the difference of your two rounded numbers. Are your answers accurate to 6 significant figures?

   (b) For each x shown in the table, and working to 4 significant figures evaluate $\sqrt{x+1}$ and then $\sqrt{x}$. Find $\sqrt{x+1} - \sqrt{x}$ by taking the difference of your two rounded numbers. Are your answers accurate to 4 significant figures?

3. The larger solution of the quadratic equation

   $$x^2 + 168x + 1 = 0$$

   is $-84 + \sqrt{7055}$ which is equal to -0.0059525919 to 10 decimal places. Round the value $\sqrt{7055}$ to 4 significant figures and then use this rounded value to calculate the larger solution of the quadratic equation. How many accurate significant figures does your answer have?

4. Consider the function

   $$f(x) = x^2 + x - 1975$$

   and suppose we want to evaluate it for some x.

   (a) Let $x = 20$. Evaluate $f(x)$ and then evaluate f again having altered x by just 1%. What is the percentage change in f? Is the problem of evaluating $f(x)$, for $x = 20$, a well-conditioned one?

   (b) Let $x = 44$. Evaluate $f(x)$ and then evaluate f again having altered x by just 1%. What is the percentage change in f? Is the problem of evaluating $f(x)$, for $x = 44$, a well-conditioned one?

   (Answer: the problem in part (a) is well-conditioned, the problem in part (b) is ill-conditioned.)

Answers

1. $23.57, \quad -15400.$

2. The answers are tabulated below. The 2^{nd} and 3^{rd} columns give values for $\sqrt{x+1}$ and $\sqrt{x}$ respectively, rounded to 10 decimal places. The 4^{th} column shows the values of $\sqrt{x+1}-\sqrt{x}$ also to 10 decimal places. Column (a) deals with part (a) of the question and finds the difference after rounding the numbers in the 2^{nd} and 3^{rd} columns to 6 significant figures. Column (b) deals with part (b) of the question and finds the difference after rounding the numbers in the 2^{nd} and 3^{rd} columns to 4 significant figures.

x	$\sqrt{x+1}$	$\sqrt{x}$		(a)	(b)
100	10.0498756211	10.0000000000	0.0498756211	0.0499	0.0500
1000	31.6385840391	31.6227766017	0.0158074374	0.0158	0.0200
10000	100.0049998750	100.0000000000	0.0049998750	0.0050	0.0000
100000	316.2293471517	316.2277660168	0.0015811349	0.0010	0.0000

Clearly the answers in columns (a) and (b) are not accurate to 6 and 4 figures respectively. Indeed the last two figures in column (b) are accurate to no figures at all!

3. $\sqrt{7055} = 83.99$ to 4 significant figures. Using this value to find the larger solution of the quadratic equation gives

$$-84 + 83.99 = -0.01 \ .$$

The number of accurate significant figures is 0 because the accurate answer is 0.006 and '1' is not the leading digit (it is '6').

4. (a) $f(20) = -1555$ and $f(20.2) = -1546.76$ so the percentage change in f on changing $x = 20$ by 1% is

$$\frac{-1555-(-1546.76)}{-1555} \times 100\% = 0.53\%$$

to 2 decimal places.

(b) $f(44) = 5$ and $f(44.44) = 44.3536$ so the percentage change in f on changing $x = 44$ by 1% is

$$\frac{5-44.3536}{5} \times 100\% = -787.07\%$$

to 2 decimal places.

Clearly then, the evaluation of $f(20)$ is well-conditioned and that of $f(44)$ is ill-conditioned.