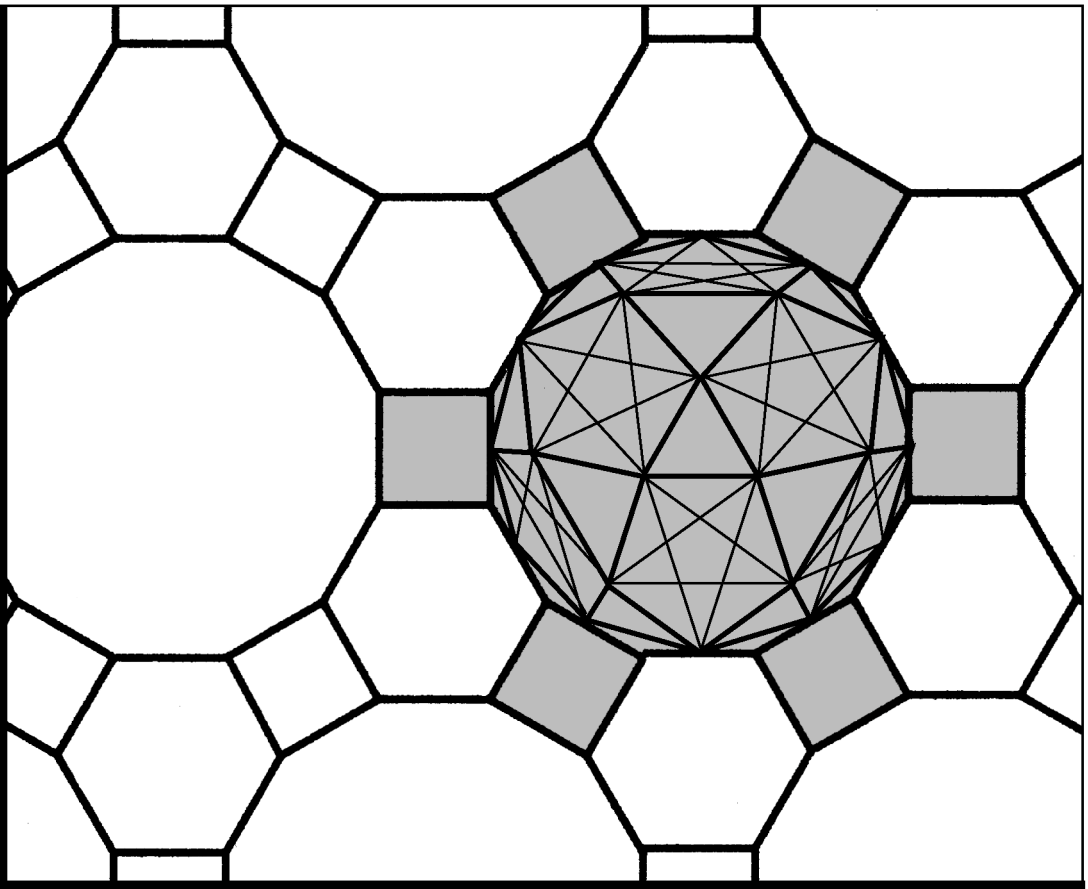
Workbook 30



Introduction to Numerical Methods

HELM®

HELM: Helping Engineers Learn Mathematics

http://helm.lboro.ac.uk

About the HELM Project

HELM (Helping Engineers Learn Mathematics) materials were the outcome of a three-year curriculum development project undertaken by a consortium of five English universities led by Loughborough University, funded by the Higher Education Funding Council for England under the Fund for the Development of Teaching and Learning for the period October 2002 – September 2005, with additional transferability funding October 2005 – September 2006.

HELM aims to enhance the mathematical education of engineering undergraduates through flexible learning resources, mainly these Workbooks.

HELM learning resources were produced primarily by teams of writers at six universities: Hull, Loughborough, Manchester, Newcastle, Reading, Sunderland.

HELM gratefully acknowledges the valuable support of colleagues at the following universities and colleges involved in the critical reading, trialling, enhancement and revision of the learning materials:
Aston, Bournemouth & Poole College, Cambridge, City, Glamorgan, Glasgow, Glasgow Caledonian, Glenrothes Institute of Applied Technology, Harper Adams, Hertfordshire, Leicester, Liverpool, London Metropolitan, Moray College, Northumbria, Nottingham, Nottingham Trent, Oxford Brookes, Plymouth, Portsmouth, Queens Belfast, Robert Gordon, Royal Forest of Dean College, Salford, Sligo Institute of Technology, Southampton, Southampton Institute, Surrey, Teesside, Ulster, University of Wales Institute Cardiff, West Kingsway College (London), West Notts College.

HELM Contacts:
Post: HELM, Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU.
Email: helm@lboro.ac.uk *Web:* http://helm.lboro.ac.uk

HELM Workbooks List

1	Basic Algebra	26	Functions of a Complex Variable
2	Basic Functions	27	Multiple Integration
3	Equations, Inequalities & Partial Fractions	28	Differential Vector Calculus
4	Trigonometry	29	Integral Vector Calculus
5	Functions and Modelling	30	Introduction to Numerical Methods
6	Exponential and Logarithmic Functions	31	Numerical Methods of Approximation
7	Matrices	32	Numerical Initial Value Problems
8	Matrix Solution of Equations	33	Numerical Boundary Value Problems
9	Vectors	34	Modelling Motion
10	Complex Numbers	35	Sets and Probability
11	Differentiation	36	Descriptive Statistics
12	Applications of Differentiation	37	Discrete Probability Distributions
13	Integration	38	Continuous Probability Distributions
14	Applications of Integration 1	39	The Normal Distribution
15	Applications of Integration 2	40	Sampling Distributions and Estimation
16	Sequences and Series	41	Hypothesis Testing
17	Conics and Polar Coordinates	42	Goodness of Fit and Contingency Tables
18	Functions of Several Variables	43	Regression and Correlation
19	Differential Equations	44	Analysis of Variance
20	Laplace Transforms	45	Non-parametric Statistics
21	z-Transforms	46	Reliability and Quality Control
22	Eigenvalues and Eigenvectors	47	Mathematics and Physics Miscellany
23	Fourier Series	48	Engineering Case Study
24	Fourier Transforms	49	Student's Guide
25	Partial Differential Equations	50	Tutor's Guide

Production of this 2015 edition, containing corrections and minor revisions of the 2008 edition, was funded by the **sigma** Network.

sigma Σ
network for excellence in mathematics and statistics support

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

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

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

$\pi = 3.142$ 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.

Gaussian Elimination 30.2

Introduction

In this Section we will reconsider the Gaussian elimination approach discussed in HELM 8, and we will see how rounding error can grow if we are not careful in our implementation of the approach. A method called partial pivoting, which helps stop rounding error from growing, will be introduced.

Prerequisites

Before starting this Section you should ...

- revise matrices, especially matrix solution of equations
- recall Gaussian elimination
- be able to find the inverse of a 2×2 matrix

Learning Outcomes

On completion you should be able to ...

- carry out Gaussian elimination with partial pivoting

1. Gaussian elimination

Recall from HELM 8 that the basic idea with Gaussian (or Gauss) elimination is to replace the matrix of coefficients with a matrix that is easier to deal with. Usually the nicer matrix is of **upper triangular** form which allows us to find the solution by **back substitution**. For example, suppose we have

$$\begin{aligned} x_1 + 3x_2 - 5x_3 &= 2 \\ 3x_1 + 11x_2 - 9x_3 &= 4 \\ -x_1 + x_2 + 6x_3 &= 5 \end{aligned}$$

which we can abbreviate using an **augmented matrix** to

$$\left[\begin{array}{ccc|c} \boxed{1} & 3 & -5 & 2 \\ 3 & 11 & -9 & 4 \\ -1 & 1 & 6 & 5 \end{array}\right].$$

We use the boxed element to eliminate any non-zeros below it. This involves the following row operations

$$\left[\begin{array}{ccc|c} \boxed{1} & 3 & -5 & 2 \\ 3 & 11 & -9 & 4 \\ -1 & 1 & 6 & 5 \end{array}\right] \begin{array}{l} \\ R2 - 3 \times R1 \\ R3 + R1 \end{array} \Rightarrow \left[\begin{array}{ccc|c} \boxed{1} & 3 & -5 & 2 \\ 0 & 2 & 6 & -2 \\ 0 & 4 & 1 & 7 \end{array}\right].$$

And the next step is to use the $\boxed{2}$ to eliminate the non-zero below *it*. This requires the final row operation

$$\left[\begin{array}{ccc|c} 1 & 3 & -5 & 2 \\ 0 & \boxed{2} & 6 & -2 \\ 0 & 4 & 1 & 7 \end{array}\right] \begin{array}{l} \\ \\ R3 - 2 \times R2 \end{array} \Rightarrow \left[\begin{array}{ccc|c} 1 & 3 & -5 & 2 \\ 0 & \boxed{2} & 6 & -2 \\ 0 & 0 & -11 & 11 \end{array}\right].$$

This is the augmented form for an upper triangular system, writing the system in extended form we have

$$\begin{aligned} x_1 + 3x_2 - 5x_3 &= 2 \\ 2x_2 + 6x_3 &= -2 \\ -11x_3 &= 11 \end{aligned}$$

which is easy to solve from the bottom up, by **back substitution**.

Example 5

Solve the system

$$\begin{aligned} x_1 + 3x_2 - 5x_3 &= 2 \\ 2x_2 + 6x_3 &= -2 \\ -11x_3 &= 11 \end{aligned}$$

Solution

The bottom equation implies that $x_3 = -1$. The middle equation then gives us that

$$2x_2 = -2 - 6x_3 = -2 + 6 = 4 \qquad \therefore \quad x_2 = 2$$

and finally, from the top equation,

$$x_1 = 2 - 3x_2 + 5x_3 = 2 - 6 - 5 = -9.$$

Therefore the solution to the problem stated at the beginning of this Section is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -9 \\ 2 \\ -1 \end{bmatrix}.$$

The following Task will act as useful revision of the Gaussian elimination procedure.

Carry out row operations to reduce the matrix

$$\begin{bmatrix} 2 & -1 & 4 \\ 4 & 3 & -1 \\ -6 & 8 & -2 \end{bmatrix}$$

into upper triangular form.

Your solution

Answer
The row operations required to eliminate the non-zeros below the diagonal in the first column are as follows

$$\begin{bmatrix} 2 & -1 & 4 \\ 4 & 3 & -1 \\ -6 & 8 & -2 \end{bmatrix} \begin{matrix} \\ R2 - 2 \times R1 \\ R3 + 3 \times R1 \end{matrix} \Rightarrow \begin{bmatrix} 2 & -1 & 4 \\ 0 & 5 & -9 \\ 0 & 5 & 10 \end{bmatrix}$$

Next we use the 5 on the diagonal to eliminate the 5 below it:

$$\begin{bmatrix} 2 & -1 & 4 \\ 0 & 5 & -9 \\ 0 & 5 & 10 \end{bmatrix} \begin{matrix} \\ \\ R3 - R2 \end{matrix} \Rightarrow \begin{bmatrix} 2 & -1 & 4 \\ 0 & 5 & -9 \\ 0 & 0 & 19 \end{bmatrix}$$

which is in the required upper triangular form.

2. Partial pivoting

Partial pivoting is a refinement of the Gaussian elimination procedure which helps to prevent the growth of rounding error.

An example to motivate the idea

Consider the example

$$\begin{bmatrix} 10^{-4} & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

First of all let us work out the *exact* answer to this problem

$$\begin{aligned} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 10^{-4} & 1 \\ -1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{2 \times 10^{-4} + 1} \begin{bmatrix} 2 & -1 \\ 1 & 10^{-4} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \frac{1}{2 \times 10^{-4} + 1} \begin{bmatrix} 1 \\ 1 + 10^{-4} \end{bmatrix} = \begin{bmatrix} 0.999800... \\ 0.999900... \end{bmatrix}. \end{aligned}$$

Now we compare this exact result with the output from Gaussian elimination. Let us suppose, for sake of argument, that all numbers are rounded to 3 significant figures. Eliminating the one non-zero element below the diagonal, and remembering that we are only dealing with 3 significant figures, we obtain

$$\begin{bmatrix} 10^{-4} & 1 \\ 0 & 10^{4} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 10^{4} \end{bmatrix}.$$

The bottom equation gives $x_2 = 1$, and the top equation therefore gives $x_1 = 0$. Something has gone seriously wrong, for this value for x_1 is nowhere near the true value 0.9998... found without rounding.The problem has been caused by using a small number (10^{-4}) to eliminate a number much larger in magnitude (-1) below it.

The general idea with partial pivoting is to try to avoid using a small number to eliminate much larger numbers.

Suppose we swap the rows

$$\begin{bmatrix} -1 & 2 \\ 10^{-4} & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and proceed as normal, still using just 3 significant figures. This time eliminating the non-zero below the diagonal gives

$$\begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

which leads to $x_2 = 1$ and $x_1 = 1$, which is an excellent approximation to the exact values, given that we are only using 3 significant figures.

Partial pivoting in general

At each step the aim in Gaussian elimination is to use an element on the diagonal to eliminate all the non-zeros below. In partial pivoting we look at all of these elements (the diagonal and the ones below) and swap the rows (if necessary) so that the element on the diagonal is not very much smaller than the other elements.

Key Point 3

Partial Pivoting

This involves scanning a column from the diagonal down. If the diagonal entry is very much smaller than any of the others we swap rows. Then we proceed with Gaussian elimination in the usual way.

In practice on a computer we swap rows to ensure that the diagonal entry is always the largest possible (in magnitude). For calculations we can carry out by hand it is usually only necessary to worry about partial pivoting if a zero crops up in a place which stops Gaussian elimination working. Consider this example

$$\begin{bmatrix} 1 & -3 & 2 & 1 \\ 2 & -6 & 1 & 4 \\ -1 & 2 & 3 & 4 \\ 0 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \\ 12 \\ 0 \end{bmatrix}.$$

The first step is to use the 1 in the top left corner to eliminate all the non-zeros below it in the augmented matrix

$$\left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 2 & -6 & 1 & 4 & 1 \\ -1 & 2 & 3 & 4 & 12 \\ 0 & -1 & 1 & 1 & 0 \end{array}\right] \begin{array}{l} R2 - 2 \times R1 \\ R3 + R1 \end{array} \Rightarrow \left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & \boxed{0} & -3 & 2 & 9 \\ 0 & -1 & 5 & 5 & 8 \\ 0 & -1 & 1 & 1 & 0 \end{array}\right].$$

What we would *like* to do now is to use the boxed element to eliminate all the non-zeros below it. But clearly this is impossible. We need to apply partial pivoting. We look **down** the column starting

at the diagonal entry and see that the two possible candidates for the swap are both equal to -1. Either will do so let us swap the second and fourth rows to give

$$\left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & \boxed{-1} & 1 & 1 & 0 \\ 0 & -1 & 5 & 5 & 8 \\ 0 & 0 & -3 & 2 & 9 \end{array}\right].$$

That was the partial pivoting step. Now we proceed with Gaussian elimination

$$\left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & \boxed{-1} & 1 & 1 & 0 \\ 0 & -1 & 5 & 5 & 8 \\ 0 & 0 & -3 & 2 & 9 \end{array}\right] \quad R3 - R2 \quad \Rightarrow \left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 4 & 4 & 8 \\ 0 & 0 & -3 & 2 & 9 \end{array}\right].$$

The arithmetic is simpler if we cancel a factor of 4 out of the third row to give

$$\left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & -3 & 2 & 9 \end{array}\right].$$

And the elimination phase is completed by removing the -3 from the final row as follows

$$\left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & \boxed{1} & 1 & 2 \\ 0 & 0 & -3 & 2 & 9 \end{array}\right] \quad R4 + 3 \times R3 \quad \Rightarrow \left[\begin{array}{cccc|c} 1 & -3 & 2 & 1 & -4 \\ 0 & -1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 5 & 15 \end{array}\right].$$

This system is upper triangular so back substitution can be used now to work out that $x_4 = 3$, $x_3 = -1$, $x_2 = 2$ and $x_1 = 1$.

The Task below is a case in which partial pivoting is required.

[For a large system which can be solved by Gauss elimination see Engineering Example 1 on page 62].

Transform the matrix

$$\left[\begin{array}{ccc} 1 & -2 & 4 \\ -3 & 6 & -11 \\ 4 & 3 & 5 \end{array}\right]$$

into upper triangular form using Gaussian elimination (with partial pivoting when necessary).

Your solution

Answer

The row operations required to eliminate the non-zeros below the diagonal in the first column are

$$\begin{bmatrix} 1 & -2 & 4 \\ -3 & 6 & -11 \\ 4 & 3 & 5 \end{bmatrix} \begin{matrix} \\ R2+3\times R1 \\ R3-4\times R1 \end{matrix} \Rightarrow \begin{bmatrix} 1 & -2 & 4 \\ 0 & 0 & 1 \\ 0 & 11 & -11 \end{bmatrix}$$

which puts a zero on the diagonal. We are forced to use partial pivoting and swapping the second and third rows gives

$$\begin{bmatrix} 1 & -2 & 4 \\ 0 & 11 & -11 \\ 0 & 0 & 1 \end{bmatrix}$$

which is in the required upper triangular form.

Key Point 4

When To Use Partial Pivoting

1. When carrying out Gaussian elimination on a computer, we would usually always swap rows so that the element on the diagonal is as large (in magnitude) as possible. This helps stop the growth of rounding error.
2. When doing hand calculations (not involving rounding) there are two reasons we might pivot
   (a) If the element on the diagonal is zero, we **have** to swap rows so as to put a non-zero on the diagonal.
   (b) Sometimes we might swap rows so that there is a "nicer" non-zero number on the diagonal than there would be without pivoting. For example, if the number on the diagonal can be arranged to be a 1 then no awkward fractions will be introduced when we carry out row operations related to Gaussian elimination.

Exercises

1. Solve the following system by back substitution

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 3 \\ 5x_2 + 6x_3 &= -2 \\ 7x_3 &= -14 \end{aligned}$$

2. (a) Show that the exact solution of the system of equations

$$\begin{bmatrix} 10^{-5} & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 10 \end{bmatrix} \text{ is } \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -0.99998 \\ 2.00001 \end{bmatrix}.$$

(b) Working to 3 significant figures, and using Gaussian elimination *without* pivoting, find an approximation to $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Show that the rounding error causes the approximation to x_1 to be a very poor one.

(c) Working to 3 significant figures, and using Gaussian elimination *with* pivoting, find an approximation to $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. Show that the approximation this time is a good one.

3. Carry out row operations (with partial pivoting if necessary) to reduce these matrices to upper triangular form.

$$\text{(a)} \begin{bmatrix} 1 & -2 & 4 \\ -4 & -3 & -3 \\ -1 & 13 & 1 \end{bmatrix}, \qquad \text{(b)} \begin{bmatrix} 0 & -1 & 2 \\ 1 & -4 & 2 \\ -2 & 5 & -4 \end{bmatrix}, \qquad \text{(c)} \begin{bmatrix} -3 & 10 & 1 \\ 1 & -3 & 2 \\ -2 & 10 & -4 \end{bmatrix}.$$

(Hint: before tackling (c) you might like to consider point 2(b) in Key Point 4.)

Answers

1. From the last equation we see that $x_3 = -2$. Using this information in the second equation gives us $x_2 = 2$. Finally, the first equation implies that $x_1 = -3$.

2. (a) The formula $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \dfrac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ can be used to show that

$$x_1 = -\frac{50000}{50001} = -0.99998 \qquad \text{and} \qquad x_2 = \frac{200005}{100002} = 2.00001 \qquad \text{as required.}$$

(b) Carrying out the elimination without pivoting, and rounding to 3 significant figures we find that $x_2 = 2.00$ and that, therefore, $x_1 = 0$. This is a very poor approximation to x_1.

(c) To apply partial pivoting we swap the two rows and then eliminate the bottom left element. Consequently we find that, after rounding the system of equations to 3 significant figures, $x_2 = 2.00$ and $x_1 = -1.00$. These give excellent agreement with the exact answers.

Answers

3.

(a) The row operations required to eliminate the non-zeros below the diagonal in the first column are as follows

$$\begin{bmatrix} 1 & -2 & 4 \\ -4 & -3 & -3 \\ -1 & 13 & 1 \end{bmatrix} \begin{matrix} \\ R2+4\times R1 \\ R3+1\times R1 \end{matrix} \Rightarrow \begin{bmatrix} 1 & -2 & 4 \\ 0 & -11 & 13 \\ 0 & 11 & 5 \end{bmatrix}$$

Next we use the element in the middle of the matrix to eliminate the value underneath it. This gives

$$\begin{bmatrix} 1 & -2 & 4 \\ 0 & -11 & 13 \\ 0 & 0 & 18 \end{bmatrix}$$

which is of the required upper triangular form.

(b) We must swap the rows to put a non-zero in the top left position (this is the partial pivoting step). Swapping the first and second rows gives the matrix

$$\begin{bmatrix} 1 & -4 & 2 \\ 0 & -1 & 2 \\ -2 & 5 & -4 \end{bmatrix}.$$

We carry out one row operation to eliminate the non-zero in the bottom left entry as follows

$$\begin{bmatrix} 1 & -4 & 2 \\ 0 & -1 & 2 \\ -2 & 5 & -4 \end{bmatrix} \begin{matrix} \\ \\ R3+2\times R1 \end{matrix} \Rightarrow \begin{bmatrix} 1 & -4 & 2 \\ 0 & -1 & 2 \\ 0 & -3 & 0 \end{bmatrix}$$

Next we use the middle element to eliminate the non-zero value underneath it. This gives

$$\begin{bmatrix} 1 & -4 & 2 \\ 0 & -1 & 2 \\ 0 & 0 & -6 \end{bmatrix}$$

which is of the required upper triangular form.

(c) If we swap the first and second rows of the matrix then we do not have to deal with fractions. Having done this the row operations required to eliminate the non-zeros below the diagonal in the first column are as follows

$$\begin{bmatrix} 1 & -3 & 2 \\ -3 & 10 & 1 \\ -2 & 10 & -4 \end{bmatrix} \begin{matrix} \\ R2+3\times R1 \\ R3+2\times R1 \end{matrix} \Rightarrow \begin{bmatrix} 1 & -3 & 2 \\ 0 & 1 & 7 \\ 0 & 4 & 0 \end{bmatrix}$$

Next we use the element in the middle of the matrix to eliminate the non-zero value underneath it. This gives

$$\begin{bmatrix} 1 & -3 & 2 \\ 0 & 1 & 7 \\ 0 & 0 & -28 \end{bmatrix}$$

which is of the required upper triangular form.

LU Decomposition

30.3

Introduction

In this Section we consider another direct method for obtaining the solution of systems of equations in the form $AX = B$.

Prerequisites

Before starting this Section you should ...

- revise matrices and their use in systems of equations
- revise determinants

Learning Outcomes

On completion you should be able to ...

- find an LU decomposition of simple matrices and apply it to solve systems of equations
- determine when an LU decomposition is unavailable and when it is possible to circumvent the problem

1. LU decomposition

Suppose we have the system of equations

$$AX = B.$$

The motivation for an LU decomposition is based on the observation that systems of equations involving triangular coefficient matrices are easier to deal with. Indeed, the whole point of Gaussian elimination is to replace the coefficient matrix with one that is triangular. The LU decomposition is another approach designed to exploit triangular systems.
We suppose that we can write

$$A = LU$$

where L is a lower triangular matrix and U is an upper triangular matrix. Our aim is to find L and U and once we have done so we have found an LU decomposition of A.

Key Point 5

An LU decomposition of a matrix A is the product of a lower triangular matrix and an upper triangular matrix that is equal to A.

It turns out that we need only consider lower triangular matrices L that have 1s down the diagonal. Here is an example. Let

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} = LU \text{ where } L = \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix}.$$

Multiplying out LU and setting the answer equal to A gives

$$\begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12}+U_{22} & L_{21}U_{13}+U_{23} \\ L_{31}U_{11} & L_{31}U_{12}+L_{32}U_{22} & L_{31}U_{13}+L_{32}U_{23}+U_{33} \end{bmatrix} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix}.$$

Now we use this to find the entries in L and U. Fortunately this is not nearly as hard as it might at first seem. We begin by running along the top row to see that

$$\boxed{U_{11} = 1}, \qquad \boxed{U_{12} = 2}, \qquad \boxed{U_{13} = 4}.$$

Now consider the second row

$$L_{21}U_{11} = 3 \quad \therefore L_{21} \times 1 = 3 \quad \therefore \boxed{L_{21} = 3},$$

$$L_{21}U_{12} + U_{22} = 8 \quad \therefore 3 \times 2 + U_{22} = 8 \quad \therefore \boxed{U_{22} = 2},$$

$$L_{21}U_{13} + U_{23} = 14 \quad \therefore 3 \times 4 + U_{23} = 14 \quad \therefore \boxed{U_{23} = 2}.$$

Notice how, at each step, the equation being considered has only one unknown in it, and other quantities that we have already found. This pattern continues on the last row

$$L_{31}U_{11} = 2 \quad \therefore L_{31} \times 1 = 2 \quad \therefore \boxed{L_{31} = 2},$$

$$L_{31}U_{12} + L_{32}U_{22} = 6 \quad \therefore 2 \times 2 + L_{32} \times 2 = 6 \quad \therefore \boxed{L_{32} = 1},$$

$$L_{31}U_{13} + L_{32}U_{23} + U_{33} = 13 \quad \therefore (2 \times 4) + (1 \times 2) + U_{33} = 13 \quad \therefore \boxed{U_{33} = 3}.$$

We have shown that

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}$$

and this is an LU decomposition of A.

Find an LU decomposition of $\begin{bmatrix} 3 & 1 \\ -6 & -4 \end{bmatrix}$.

Your solution

Answer

Let

$$\begin{bmatrix} 3 & 1 \\ -6 & -4 \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 \\ L_{21} & 1 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} \end{bmatrix}$$

then, comparing the left and right hand sides row by row implies that $U_{11} = 3$, $U_{12} = 1$, $L_{21}U_{11} = -6$ which implies $L_{21} = -2$ and $L_{21}U_{12} + U_{22} = -4$ which implies that $U_{22} = -2$. Hence

$$\begin{bmatrix} 3 & 1 \\ -6 & -4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & -2 \end{bmatrix}$$

is an LU decomposition of $\begin{bmatrix} 3 & 1 \\ -6 & -4 \end{bmatrix}$.

Task

Find an LU decomposition of $\begin{bmatrix} 3 & 1 & 6 \\ -6 & 0 & -16 \\ 0 & 8 & -17 \end{bmatrix}$.

Your solution

Answer

Using material from the worked example in the notes we set

$$\begin{bmatrix} 3 & 1 & 6 \\ -6 & 0 & -16 \\ 0 & 8 & -17 \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12}+U_{22} & L_{21}U_{13}+U_{23} \\ L_{31}U_{11} & L_{31}U_{12}+L_{32}U_{22} & L_{31}U_{13}+L_{32}U_{23}+U_{33} \end{bmatrix}$$

and comparing elements row by row we see that

$$\begin{array}{lll} U_{11}=3, & U_{12}=1, & U_{13}=6, \\ L_{21}=-2, & U_{22}=2, & U_{23}=-4 \\ L_{31}=0 & L_{32}=4 & U_{33}=-1 \end{array}$$

and it follows that

$$\begin{bmatrix} 3 & 1 & 6 \\ -6 & 0 & -16 \\ 0 & 8 & -17 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2 & -4 \\ 0 & 0 & -1 \end{bmatrix}$$

is an LU decomposition of the given matrix.

2. Using LU decomposition to solve systems of equations

Once a matrix A has been decomposed into lower and upper triangular parts it is possible to obtain the solution to $AX = B$ in a direct way. The procedure can be summarised as follows

- Given A, find L and U so that $A = LU$. Hence $LUX = B$.
- Let $Y = UX$ so that $LY = B$. Solve this triangular system for Y.
- Finally solve the triangular system $UX = Y$ for X.

The benefit of this approach is that we only ever need to solve triangular systems. The cost is that we have to solve two of them.

[Here we solve only small systems; a large system is presented in Engineering Example 1 on page 62.]

Example 6

Find the solution of $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ of the system $\begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 13 \\ 4 \end{bmatrix}$.

Solution

- The first step is to calculate the LU decomposition of the coefficient matrix on the left-hand side. In this case that job has already been done since this is the matrix we considered earlier. We found that

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}.$$

- The next step is to solve $LY = B$ for the vector $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$. That is we consider

$$LY = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 13 \\ 4 \end{bmatrix} = B$$

which can be solved by **forward substitution**. From the top equation we see that $y_1 = 3$. The middle equation states that $3y_1 + y_2 = 13$ and hence $y_2 = 4$. Finally the bottom line says that $2y_1 + y_2 + y_3 = 4$ from which we see that $y_3 = -6$.

Solution (contd.)

- Now that we have found Y we finish the procedure by solving $UX = Y$ for X. That is we solve

$$UX = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ -6 \end{bmatrix} = Y$$

  by using **back substitution**. Starting with the bottom equation we see that $3x_3 = -6$ so clearly $x_3 = -2$. The middle equation implies that $2x_2 + 2x_3 = 4$ and it follows that $x_2 = 4$. The top equation states that $x_1 + 2x_2 + 4x_3 = 3$ and consequently $x_1 = 3$.

Therefore we have found that the solution to the system of simultaneous equations

$$\begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 13 \\ 4 \end{bmatrix} \quad \text{is} \quad X = \begin{bmatrix} 3 \\ 4 \\ -2 \end{bmatrix}.$$

Use the LU decomposition you found earlier in the last Task (page 24) to solve

$$\begin{bmatrix} 3 & 1 & 6 \\ -6 & 0 & -16 \\ 0 & 8 & -17 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \\ 17 \end{bmatrix}.$$

Your solution

Answer

We found earlier that the coefficient matrix is equal to $LU = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2 & -4 \\ 0 & 0 & -1 \end{bmatrix}$.

First we solve $LY = B$ for Y, we have

$$\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 4 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \\ 17 \end{bmatrix}.$$

The top line implies that $y_1 = 0$. The middle line states that $-2y_1 + y_2 = 4$ and therefore $y_2 = 4$. The last line tells us that $4y_2 + y_3 = 17$ and therefore $y_3 = 1$.

Finally we solve $UX = Y$ for X, we have

$$\begin{bmatrix} 3 & 1 & 6 \\ 0 & 2 & -4 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 4 \\ 1 \end{bmatrix}.$$

The bottom line shows that $x_3 = -1$. The middle line then shows that $x_2 = 0$, and then the top line gives us that $x_1 = 2$. The required solution is $X = \begin{bmatrix} 2 \\ 0 \\ -1 \end{bmatrix}$.

3. Do matrices always have an LU decomposition?

No. Sometimes it is impossible to write a matrix in the form "lower triangular" × "upper triangular".

Why not?

An invertible matrix A has an LU decomposition provided that all its **leading submatrices** have non-zero determinants. The k^{th} leading submatrix of A is denoted A_k and is the $k \times k$ matrix found by looking only at the top k rows and leftmost k columns. For example if

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix}$$

then the leading submatrices are

$$A_1 = 1, \qquad A_2 = \begin{bmatrix} 1 & 2 \\ 3 & 8 \end{bmatrix}, \qquad A_3 = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix}.$$

The fact that this matrix A has an LU decomposition can be guaranteed in advance because none of these determinants is zero:

$$|A_1| = 1,$$
$$|A_2| = (1 \times 8) - (2 \times 3) = 2,$$
$$|A_3| = \begin{vmatrix} 8 & 14 \\ 6 & 13 \end{vmatrix} - 2 \begin{vmatrix} 3 & 14 \\ 2 & 13 \end{vmatrix} + 4 \begin{vmatrix} 3 & 8 \\ 2 & 6 \end{vmatrix} = 20 - (2 \times 11) + (4 \times 2) = 6$$

(where the 3×3 determinant was found by expanding along the top row).

Example 7

Show that $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 1 & 3 & 4 \end{bmatrix}$ does not have an LU decomposition.

Solution

The second leading submatrix has determinant equal to

$$\begin{vmatrix} 1 & 2 \\ 2 & 4 \end{vmatrix} = (1 \times 4) - (2 \times 2) = 0$$

which means that an LU decomposition is not possible in this case.

Which, if any, of these matrices have an LU decomposition?

(a) $A = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}$, (b) $A = \begin{bmatrix} 0 & 1 \\ 3 & 2 \end{bmatrix}$, (c) $A = \begin{bmatrix} 1 & -3 & 7 \\ -2 & 6 & 1 \\ 0 & 3 & -2 \end{bmatrix}$.

Your solution

(a)

Answer

$|A_1| = 3$ and $|A_2| = |A| = 3$. Neither of these is zero, so A **does** have an LU decomposition.

Your solution

(b)

Answer

$|A_1| = 0$ so A **does not** have an LU decomposition.

Your solution

(c)

Answer

$|A_1| = 1$, $|A_2| = 6 - 6 = 0$, so A **does not** have an LU decomposition.

Can we get around this problem?

Yes. It is always possible to re-order the rows of an **invertible** matrix so that all of the submatrices have non-zero determinants.

Example 8

Reorder the rows of $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 1 & 3 & 4 \end{bmatrix}$ so that the reordered matrix has an LU decomposition.

Solution

Swapping the first and second rows does not help us since the second leading submatrix will still have a zero determinant. Let us swap the second and third rows and consider

$$B = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 4 \\ 2 & 4 & 5 \end{bmatrix}$$

the leading submatrices are

$$B_1 = 1, \quad B_2 = \begin{bmatrix} 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad B_3 = B.$$

Now $|B_1| = 1$, $|B_2| = 3 \times 1 - 2 \times 1 = 1$ and (expanding along the first row)

$$|B_3| = 1(15 - 16) - 2(5 - 8) + 3(4 - 6) = -1 + 6 - 6 = -1.$$

All three of these determinants are non-zero and we conclude that B does have an LU decomposition.

Reorder the rows of $A = \begin{bmatrix} 1 & -3 & 7 \\ -2 & 6 & 1 \\ 0 & 3 & -2 \end{bmatrix}$ so that the reordered matrix has an LU decomposition.

Your solution

Answer

Let us swap the second and third rows and consider

$$B = \begin{bmatrix} 1 & -3 & 7 \\ 0 & 3 & -2 \\ -2 & 6 & 1 \end{bmatrix}$$

the leading submatrices are

$$B_1 = 1, \quad B_2 = \begin{bmatrix} 1 & -3 \\ 0 & 3 \end{bmatrix}, \quad B_3 = B$$

which have determinants 1, 3 and 45 respectively. All of these are non-zero and we conclude that B does indeed have an LU decomposition.

Exercises

1. Calculate LU decompositions for each of these matrices

   (a) $A = \begin{bmatrix} 2 & 1 \\ -4 & -6 \end{bmatrix}$ (b) $A = \begin{bmatrix} 2 & 1 & -4 \\ 2 & 2 & -2 \\ 6 & 3 & -11 \end{bmatrix}$ (c) $A = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 8 & 5 \\ 1 & 11 & 4 \end{bmatrix}$

2. Check each answer in Question 1, by multiplying out LU to show that the product equals A.

3. Using the answers obtained in Question 1, solve the following systems of equations.

   (a) $\begin{bmatrix} 2 & 1 \\ -4 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

   (b) $\begin{bmatrix} 2 & 1 & -4 \\ 2 & 2 & -2 \\ 6 & 3 & -11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ 11 \end{bmatrix}$

   (c) $\begin{bmatrix} 1 & 3 & 2 \\ 2 & 8 & 5 \\ 1 & 11 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}$

4. Consider $A = \begin{bmatrix} 1 & 6 & 2 \\ 2 & 12 & 5 \\ -1 & -3 & -1 \end{bmatrix}$

   (a) Show that A does not have an LU decomposition.

   (b) Re-order the rows of A and find an LU decomposition of the new matrix.

   (c) Hence solve

$$\begin{aligned} x_1 + 6x_2 + 2x_3 &= 9 \\ 2x_1 + 12x_2 + 5x_3 &= -4 \\ -x_1 - 3x_2 - x_3 &= 17 \end{aligned}$$

Answers

1. (a) We let

$$\begin{bmatrix} 2 & 1 \\ -4 & -6 \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 \\ L_{21} & 1 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} \end{bmatrix}.$$

Comparing the left-hand and right-hand sides row by row gives us that $U_{11} = 2$, $U_{12} = 1$, $L_{21}U_{11} = -4$ which implies that $L_{21} = -2$ and, finally, $L_{21}U_{12} + U_{22} = -6$ from which we see that $U_{22} = -4$. Hence

$$\begin{bmatrix} 2 & 1 \\ -4 & -6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & -4 \end{bmatrix}$$

is an LU decomposition of the given matrix.

(b) We let

$$\begin{bmatrix} 2 & 1 & -4 \\ 2 & 2 & -2 \\ 6 & 3 & -11 \end{bmatrix} = LU = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} & L_{21}U_{13} + U_{23} \\ L_{31}U_{11} & L_{31}U_{12} + L_{32}U_{22} & L_{31}U_{13} + L_{32}U_{23} + U_{33} \end{bmatrix}.$$

Looking at the top row we see that $U_{11} = 2$, $U_{12} = 1$ and $U_{13} = -4$. Now, from the second row, $L_{21} = 1$, $U_{22} = 1$ and $U_{23} = 2$. The last three unknowns come from the bottom row: $L_{31} = 3$, $L_{32} = 0$ and $U_{33} = 1$. Hence

$$\begin{bmatrix} 2 & 1 & -4 \\ 2 & 2 & -2 \\ 6 & 3 & -11 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -4 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

is an LU decomposition of the given matrix.

(c) We let

$$\begin{bmatrix} 1 & 3 & 2 \\ 2 & 8 & 5 \\ 1 & 11 & 4 \end{bmatrix} = LU = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12} + U_{22} & L_{21}U_{13} + U_{23} \\ L_{31}U_{11} & L_{31}U_{12} + L_{32}U_{22} & L_{31}U_{13} + L_{32}U_{23} + U_{33} \end{bmatrix}.$$

Looking at the top row we see that $U_{11} = 1$, $U_{12} = 3$ and $U_{13} = 2$. Now, from the second row, $L_{21} = 2$, $U_{22} = 2$ and $U_{23} = 1$. The last three unknowns come from the bottom row: $L_{31} = 1$, $L_{32} = 4$ and $U_{33} = -2$. Hence

$$\begin{bmatrix} 1 & 3 & 2 \\ 2 & 8 & 5 \\ 1 & 11 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 1 \\ 0 & 0 & -2 \end{bmatrix}$$

is an LU decomposition of the given matrix.

2. Direct multiplication provides the necessary check.

Answers

3.

(a) We begin by solving

$$\begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Clearly $y_1 = 1$ and therefore $y_2 = 4$. The values y_1 and y_2 appear on the right-hand side of the second system we need to solve:

$$\begin{bmatrix} 2 & 1 \\ 0 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

The second equation implies that $x_2 = -1$ and therefore, from the first equation, $x_1 = 1$.

(b) We begin by solving the system

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ 11 \end{bmatrix}.$$

Starting with the top equation we see that $y_1 = 4$. The second equation then implies that $y_2 = -4$ and then, from the third equation, $y_3 = -1$. These values now appear on the right-hand side of the second system

$$\begin{bmatrix} 2 & 1 & -4 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ -4 \\ -1 \end{bmatrix}.$$

The bottom equation shows us that $x_3 = -1$. Moving up to the middle equation we obtain $x_2 = -2$. The top equation yields $x_1 = 1$.

(c) We begin by solving the system

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}.$$

Starting with the top equation we see that $y_1 = 2$. The second equation then implies that $y_2 = -1$ and then, from the third equation, $y_3 = 2$. These values now appear on the right-hand side of the second system

$$\begin{bmatrix} 1 & 3 & 2 \\ 0 & 2 & 1 \\ 0 & 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 2 \end{bmatrix}.$$

The bottom equation shows us that $x_3 = -1$. Moving up to the middle equation we obtain $x_2 = 0$. The top equation yields $x_1 = 4$.

Answers

4.

(a) The second leading submatrix has determinant $1 \times 12 - 6 \times 2 = 0$ and this implies that A has no LU decomposition.

(b) Swapping the second and third rows gives $\begin{bmatrix} 1 & 6 & 2 \\ -1 & -3 & -1 \\ 2 & 12 & 5 \end{bmatrix}$. We let

$$\begin{bmatrix} 1 & 6 & 2 \\ -1 & -3 & -1 \\ 2 & 12 & 5 \end{bmatrix} = LU = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ L_{21}U_{11} & L_{21}U_{12}+U_{22} & L_{21}U_{13}+U_{23} \\ L_{31}U_{11} & L_{31}U_{12}+L_{32}U_{22} & L_{31}U_{13}+L_{32}U_{23}+U_{33} \end{bmatrix}.$$

Looking at the top row we see that $U_{11} = 1$, $U_{12} = 6$ and $U_{13} = 2$. Now, from the second row, $L_{21} = -1$, $U_{22} = 3$ and $U_{23} = 1$. The last three unknowns come from the bottom row: $L_{31} = 2$, $L_{32} = 0$ and $U_{33} = 1$. Hence

$$\begin{bmatrix} 1 & 6 & 2 \\ -1 & -3 & -1 \\ 2 & 12 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 6 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

is an LU decomposition of the given matrix.

(c) We begin by solving the system

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 17 \\ -4 \end{bmatrix}.$$

(Note that the second and third rows of the right-hand side vector have been swapped too.) Starting with the top equation we see that $y_1 = 9$. The second equation then implies that $y_2 = 26$ and then, from the third equation, $y_3 = -22$. These values now appear on the right-hand side of the second system

$$\begin{bmatrix} 1 & 6 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 26 \\ -22 \end{bmatrix}.$$

The bottom equation shows us that $x_3 = -22$. Moving up to the middle equation we obtain $x_2 = 16$. The top equation yields $x_1 = -43$.

Matrix Norms

30.4

Introduction

A matrix norm is a number defined in terms of the entries of the matrix. The norm is a useful quantity which can give important information about a matrix.

Prerequisites

Before starting this Section you should ...

- be familiar with matrices and their use in writing systems of equations
- revise material on matrix inverses, be able to find the inverse of a 2×2 matrix, and know when no inverse exists
- revise Gaussian elimination and partial pivoting
- be aware of the discussion of ill-conditioned and well-conditioned problems earlier in Section 30.1

Learning Outcomes

On completion you should be able to ...

- calculate norms and condition numbers of small matrices
- adjust certain systems of equations with a view to better conditioning

1. Matrix norms

The norm of a square matrix A is a non-negative real number denoted $\|A\|$. There are several different ways of defining a matrix norm, but they all share the following properties:

1. $\|A\| \geq 0$ for any square matrix A.
2. $\|A\| = 0$ *if and only if* the matrix $A = 0$.
3. $\|kA\| = |k|\ \|A\|$, for any scalar k.
4. $\|A + B\| \leq \|A\| + \|B\|$.
5. $\|AB\| \leq \|A\|\ \ \|B\|$.

The norm of a matrix is a measure of how large its elements are. It is a way of determining the "size" of a matrix that is not necessarily related to how many rows or columns the matrix has.

Key Point 6

Matrix Norm

The **norm** of a matrix is a real number which is a measure of the magnitude of the matrix.

Anticipating the places where we will use norms later, it is sufficient at this stage to restrict our attention to matrices with only real-valued entries. There is no need to consider complex numbers at this stage.

In the definitions of norms below we will use this notation for the elements of an $n \times n$ matrix A where

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}$$

The subscripts on a have the row number first, then the column number. The fact that

$$a_{rc}$$

is reminiscent of the word "arc" may be a help in remembering how the notation goes.

In this Section we will define three commonly used norms. We distinguish them with a subscript. All three of them satisfy the five conditions listed above, but we will not concern ourselves with verifying that fact.

The 1-norm

$$\|A\|_1 = \max_{1\leq j\leq n}\left(\sum_{i=1}^{n}|a_{ij}|\right)$$

(the maximum absolute column sum). Put simply, we sum the absolute values down each column and then take the biggest answer.

Example 9

Calculate the 1-norm of $A = \begin{bmatrix} 1 & -7 \\ -2 & -3 \end{bmatrix}$.

Solution

The absolute column sums of A are $1 + |-2| = 1 + 2 = 3$ and $|-7| + |-3| = 7 + 3 = 10$. The larger of these is 10 and therefore $\|A\|_1 = 10$.

Example 10

Calculate the 1-norm of $B = \begin{bmatrix} 5 & -4 & 2 \\ -1 & 2 & 3 \\ -2 & 1 & 0 \end{bmatrix}$.

Solution

Summing down the columns of B we find that

$$\begin{aligned} \|B\|_1 &= \max(5+1+2, 4+2+1, 2+3+0) \\ &= \max(8, 7, 5) \\ &= 8 \end{aligned}$$

Key Point 7

The 1-norm of a square matrix is the maximum of the absolute column sums.
(A useful reminder is that "1" is a tall, thin character and a column is a tall, thin quantity.)

The infinity-norm

$$\|A\|_\infty = \max_{1\le i\le n}\left(\sum_{j=1}^{n}|a_{ij}|\right)$$

(the maximum absolute row sum). Put simply, we sum the absolute values along each row and then take the biggest answer.

Example 11

Calculate the infinity-norm of $A = \begin{bmatrix} 1 & -7 \\ -2 & -3 \end{bmatrix}$.

Solution

The absolute row sums of A are $1 + |-7| = 8$ and $|-2| + |-3| = 5$. The larger of these is 8 and therefore $\|A\|_\infty = 8$.

Example 12

Calculate the infinity-norm of $B = \begin{bmatrix} 5 & -4 & 2 \\ -1 & 2 & 3 \\ -2 & 1 & 0 \end{bmatrix}$.

Solution

Summing along the rows of B we find that

$$\begin{aligned} \|B\|_\infty &= \max(5+4+2, 1+2+3, 2+1+0) \\ &= \max(11, 6, 3) \\ &= 11 \end{aligned}$$

Key Point 8

The infinity-norm of a square matrix is the maximum of the absolute row sums.
(A useful reminder is that "∞" is a short, wide character and a row is a short, wide quantity.)

The Euclidean norm

$$\|A\|_E = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij})^2}$$

(the square root of the sum of all the squares). This is similar to ordinary "Pythagorean" length where the size of a vector is found by taking the square root of the sum of the squares of all the elements.

Example 13

Calculate the Euclidean norm of $A = \begin{bmatrix} 1 & -7 \\ -2 & -3 \end{bmatrix}$.

Solution

$$\begin{aligned} \|A\|_E &= \sqrt{1^2 + (-7)^2 + (-2)^2 + (-3)^2} \\ &= \sqrt{1 + 49 + 4 + 9} \\ &= \sqrt{63} \approx 7.937. \end{aligned}$$

Example 14

Calculate the Euclidean norm of $B = \begin{bmatrix} 5 & -4 & 2 \\ -1 & 2 & 3 \\ -2 & 1 & 0 \end{bmatrix}$.

Solution

$$\begin{aligned} \|B\|_E &= \sqrt{25 + 16 + 4 + 1 + 4 + 9 + 4 + 1 + 0} \\ &= \sqrt{64} \\ &= 8. \end{aligned}$$

Key Point 9

The Euclidean norm of a square matrix is the square root of the sum of all the squares of the elements.

Calculate the norms indicated of these matrices

$$A = \begin{bmatrix} 2 & -8 \\ 3 & 1 \end{bmatrix} \quad \text{(1-norm)}, \qquad B = \begin{bmatrix} 3 & 6 & -1 \\ 3 & 1 & 0 \\ 2 & 4 & -7 \end{bmatrix} \quad \text{(infinity-norm)},$$

$$C = \begin{bmatrix} 1 & 7 & 3 \\ 4 & -2 & -2 \\ -2 & -1 & 1 \end{bmatrix} \quad \text{(Euclidean-norm)}.$$

Your solution

Answer

$\|A\|_1 = \max(2+3, 8+1) = 9,$

$\|B\|_\infty = \max(3+6+1, 3+1+0, 2+4+7) = 13,$

$$\begin{aligned} \|C\|_E &= \sqrt{1^2+7^2+3^2+4^2+(-2)^2+(-2)^2+(-2)^2+(-1)^2+1^2} \\ &= \sqrt{89} \approx 9.434 \end{aligned}$$

Other norms

Any definition you can think of which satisifes the five conditions mentioned at the beginning of this Section is a definition of a norm. There are many many possibilities, but the three given above are among the most commonly used.

2. Condition numbers

The condition number of an invertible matrix A is defined to be

$$\kappa(A) = \|A\| \ \ \|A^{-1}\|.$$

This quantity is always bigger than (or equal to) 1.
We must use the same type of norm twice on the right-hand side of the above equation. Sometimes the notation is adjusted to make it clear which norm is being used, for example if we use the infinity norm we might write

$$\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty.$$

Example 15

Use the norm indicated to calculate the condition number of the given matrices.

(a) $A = \begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix}$; 1-norm. (b) $A = \begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix}$; Euclidean norm.

(c) $B = \begin{bmatrix} -3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$; infinity-norm.

Solution

(a) $\|A\|_1 = \max(2+1, 3+1) = 4$,

$$A^{-1} = \frac{1}{-2-3}\begin{bmatrix} -1 & -3 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{5} & \frac{3}{5} \\ \frac{1}{5} & \frac{-2}{5} \end{bmatrix}$$

$$\therefore \quad \|A^{-1}\|_1 = \max(\tfrac{1}{5}+\tfrac{1}{5}, \tfrac{3}{5}+\tfrac{2}{5}) = 1.$$

Therefore $\kappa_1(A) = \|A\|_1 \ \|A^{-1}\|_1 = 4 \times 1 = 4$.

(b) $\|A\|_E = \sqrt{2^2+3^2+1^2+(-1)^2} = \sqrt{15}$. We can re-use A^{-1} from above to see that

$$\|A^{-1}\|_E = \sqrt{\left(\frac{1}{5}\right)^2 + \left(\frac{3}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{-2}{5}\right)^2} = \sqrt{\frac{15}{25}}.$$

Therefore $\kappa_E(A) = \|A\|_E \ \|A^{-1}\|_E = \sqrt{15} \times \sqrt{\frac{15}{25}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3.$

(c) $\|B\|_\infty = \max(3, 4, 2) = 4$.

$$B^{-1} = \begin{bmatrix} -\frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$$

so $\|B^{-1}\|_\infty = \max(\frac{1}{3}, \frac{1}{4}, \frac{1}{2}) = \frac{1}{2}$. Therefore $\kappa_\infty(B) = \|B\|_\infty \ \|B^{-1}\|_\infty = 4 \times \frac{1}{2} = 2.$

Calculate the condition numbers of these matrices, using the norm indicated

$$A = \begin{bmatrix} 2 & -8 \\ 3 & 1 \end{bmatrix} \quad \text{(1-norm)}, \qquad B = \begin{bmatrix} 3 & 6 \\ 1 & 0 \end{bmatrix} \quad \text{(infinity-norm)}.$$

Your solution

Answer

$$A^{-1} = \frac{1}{2+24}\begin{bmatrix} 1 & 8 \\ -3 & 2 \end{bmatrix} \quad \text{so} \quad \kappa_1(A) = \|A\|_1\|A^{-1}\|_1 = \max(5,9) \times \max(\tfrac{4}{26}, \tfrac{10}{26}) = 9 \times \tfrac{10}{26} = \tfrac{45}{13}.$$

$$B^{-1} = \frac{1}{0-6}\begin{bmatrix} 0 & -6 \\ -1 & 3 \end{bmatrix} \quad \text{so} \quad \kappa_\infty(B) = \|B\|_\infty\|B^{-1}\|_\infty = \max(9,1) \times \max(1, \tfrac{4}{6}) = 9.$$

Condition numbers and conditioning

As the name might suggest, the **condition number** gives us information regarding how well-conditioned a problem is. Consider this example

$$\begin{bmatrix} 1 & 10^4 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10^4 \\ 1 \end{bmatrix}.$$

It is not hard to verify that the exact solution to this problem is

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \dfrac{10000}{10002} \\ \dfrac{10001}{10002} \end{bmatrix} = \begin{bmatrix} 0.999800... \\ 0.999900... \end{bmatrix}.$$

Example 16

Using the 1-norm find the condition number of $\begin{bmatrix} 1 & 10^4 \\ -1 & 2 \end{bmatrix}$.

Solution

Firstly, $\|A\|_1 = 2 + 10^4$. Also

$$A^{-1} = \frac{1}{2+10^4}\begin{bmatrix} 2 & -10^4 \\ 1 & 1 \end{bmatrix} \quad \therefore \quad \|A^{-1}\|_1 = \frac{1}{2+10^4}(1+10^4). \quad \text{Hence } \kappa_1(A) = 1 + 10^4 = 10001.$$

The fact that this number is large is the indication that the problem involving A is an ill-conditioned one. Suppose we consider finding its solution by Gaussian elimination, using 3 significant figures throughout. Eliminating the non-zero in the bottom left corner gives

$$\begin{bmatrix} 1 & 10^4 \\ 0 & 10^4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10^4 \\ 10^4 \end{bmatrix}.$$

which implies that $x_2 = 1$ and $x_1 = 0$. This is a poor approximation to the true solution and partial pivoting will not help. We have altered the problem by a relatively tiny amount (that is, by neglecting the fourth significant figure) and the result $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ has changed by a large amount. In other words the problem is ill-conditioned.

One way that systems of equations can be made better conditioned is to fix things so that all the rows have largest elements that are about the same size. In the matrix $A = \begin{bmatrix} 1 & 10^4 \\ -1 & 2 \end{bmatrix}$ the first row's largest element is 10^4, the second row has largest element equal to 2. This is not a happy situation.

If we divide the first equation through by 10^4 then we have

$$\begin{bmatrix} 10^{-4} & 1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

then the top row has largest entry equal to 1, and the bottom row still has 2 as its largest entry. These two values are of comparable size.

The solution to the system was found via pivoting (using 3 significant figures) in the Section concerning Gaussian elimination to be $x_1 = x_2 = 1$, a pretty good approximation to the exact values. The matrix in this second version of the problem is much better conditioned.

Example 17

Using the 1-norm find the condition number of $\begin{bmatrix} 10^{-4} & 1 \\ -1 & 2 \end{bmatrix}$.

Solution

The 1-norm of A is easily seen to be $\|A\|_1 = 3$. We also need

$$A^{-1} = \frac{1}{2 \times 10^{-4} + 1} \begin{bmatrix} 2 & -1 \\ 1 & 10^{-4} \end{bmatrix} \qquad \therefore \qquad \|A^{-1}\|_1 = \frac{3}{2 \times 10^{-4} + 1}.$$

Hence

$$\kappa_1(A) = \frac{9}{2 \times 10^{-4} + 1} \approx 8.998$$

This condition number is much smaller than the earlier value of 10001, and this shows us that the second version of the system of equations is better conditioned.

Exercises

1. Calculate the indicated norm of the following matrices

   (a) $A = \begin{bmatrix} 2 & -2 \\ 1 & -3 \end{bmatrix}$; 1-norm.

   (b) $A = \begin{bmatrix} 2 & -2 \\ 1 & -3 \end{bmatrix}$; infinity-norm.

   (c) $B = \begin{bmatrix} 2 & -3 \\ 1 & -2 \end{bmatrix}$; Euclidean norm.

   (d) $C = \begin{bmatrix} 1 & -2 & 3 \\ 1 & 5 & 6 \\ 2 & -1 & 3 \end{bmatrix}$; infinity-norm.

   (e) $C = \begin{bmatrix} 1 & -2 & 3 \\ 1 & 5 & 6 \\ 2 & -1 & 3 \end{bmatrix}$; 1-norm.

2. Use the norm indicated to calculate the condition number of the given matrices.

   (a) $D = \begin{bmatrix} 4 & -2 \\ 6 & 0 \end{bmatrix}$; 1-norm.

   (b) $E = \begin{bmatrix} -1 & 5 \\ 4 & 2 \end{bmatrix}$; Euclidean norm.

   (c) $F = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}$; infinity-norm.

3. Why is it not sensible to ask what the condition number of $\begin{bmatrix} -1 & 3 \\ 2 & -6 \end{bmatrix}$ is?

4. Verify that the inverse of $G = \begin{bmatrix} 2 & 4 & -1 \\ 2 & 5 & 2 \\ -1 & -1 & 1 \end{bmatrix}$ is $\frac{1}{5}\begin{bmatrix} -7 & 3 & -13 \\ 4 & -1 & 6 \\ -3 & 2 & -2 \end{bmatrix}$.

   Hence find the condition number of G using the 1-norm.

5. (a) Calculate the condition number (use any norm you choose) of the coefficient matrix of the system

   $$\begin{bmatrix} 1 & 10^4 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

   and hence conclude that the problem as stated is ill-conditioned.

   (b) Multiply one of the equations through by a suitably chosen constant so as to make the system better conditioned. Calculate the condition number of the coefficient matrix in your new system of equations.

Answers

1. (a) $\|A\|_1 = \max(2+1, 2+3) = 5.$

   (b) $\|A\|_\infty = \max(2+-2, 1+3) = 4.$

   (c) $\|B\|_E = \sqrt{4+9+1+4} = \sqrt{18}$

   (d) $\|C\|_\infty = \max(1+2+3, 1+5+6, 2+1+3) = 12.$

   (e) $\|C\|_1 = \max(1+1+2, 2+5+1, 3+6+3) = 12.$

2. (a) To work out the condition number we need to find

$$D^{-1} = \frac{1}{12}\begin{bmatrix} 0 & 2 \\ -6 & 4 \end{bmatrix}.$$

   Given this we work out the condition number as the product of two norms as follows

$$\kappa_1(D) = \|D\|_1\|D^{-1}\|_1 = 10 \times \tfrac{1}{2} = 5.$$

   (b) To work out the condition number we need to find

$$E^{-1} = \frac{1}{-22}\begin{bmatrix} 2 & -5 \\ -4 & -1 \end{bmatrix}.$$

   Given this we work out the condition number as the product of two norms as follows

$$\kappa_E(E) = \|E\|_E\|E^{-1}\|_E = 6.782330 \times 0.308288 = 2.090909.$$

   to 6 decimal places.

   (c) Here $F^{-1} = \begin{bmatrix} \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & 1 \end{bmatrix}$ so that $\kappa_\infty(F) = \|F\|_\infty\|F^{-1}\|_\infty = 6 \times 1 = 6.$

3. The matrix is not invertible.

Answers

4. Verification is done by a direct multiplication to show that $GG^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

Using the 1-norm we find that $\kappa_1(G) = \|G\|_1\|G^{-1}\|_1 = 10 \times \frac{21}{5} = 42$.

5.

(a) The inverse of the coefficient matrix is

$$\frac{1}{3 - 2 \times 10^4} \begin{bmatrix} 3 & -10^4 \\ -2 & 1 \end{bmatrix} = \frac{-1}{19997} \begin{bmatrix} 3 & -10000 \\ -2 & 1 \end{bmatrix}.$$

Using the 1-norm the condition number of the coefficient matrix is

$$(3 + 10^4) \times \frac{1}{19997}(1 + 10^4) = 5002.75$$

to 6 significant figures. This is a large condition number, and the given problem is not well-conditioned.

(b) Now we multiply the top equation through by 10^{-4} so that the system of equations becomes

$$\begin{bmatrix} 10^{-4} & 1 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

and the inverse of this new coefficient matrix is

$$\frac{1}{3 \times 10^{-4} - 2} \begin{bmatrix} 3 & -1 \\ -2 & 10^{-4} \end{bmatrix} = \frac{-1}{1.9997} \begin{bmatrix} 3 & -1 \\ -2 & .0001 \end{bmatrix}.$$

Using the 1-norm again we find that the condition number of the new coefficient matrix is

$$4 \times \frac{1}{1.9997}(5) = 10.0015$$

to 6 significant figures. This much smaller condition number implies that the second problem is better conditioned.

Iterative Methods for Systems of Equations

30.5

Introduction

There are occasions when direct methods (like Gaussian elimination or the use of an LU decomposition) are not the best way to solve a system of equations. An alternative approach is to use an iterative method. In this Section we will discuss some of the issues involved with iterative methods.

Prerequisites

Before starting this Section you should ...

- revise matrices, especially the material in HELM 8
- revise determinants
- revise matrix norms

Learning Outcomes

On completion you should be able to ...

- approximate the solutions of simple systems of equations by iterative methods
- assess convergence properties of iterative methods

1. Iterative methods

Suppose we have the system of equations

$$AX = B.$$

The aim here is to find a sequence of approximations which gradually approach X. We will denote these approximations

$$X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(k)}, \ldots$$

where $X^{(0)}$ is our initial "guess", and the hope is that after a short while these successive **iterates** will be so close to each other that the process can be deemed to have **converged** to the required solution X.

Key Point 10

An **iterative** method is one in which a sequence of approximations (or **iterates**) is produced. The method is successful if these iterates converge to the true solution of the given problem.

It is convenient to split the matrix A into three parts. We write

$$A = L + D + U$$

where L consists of the elements of A strictly below the diagonal and zeros elsewhere; D is a diagonal matrix consisting of the diagonal entries of A; and U consists of the elements of A strictly above the diagonal. **Note that L and U here are not the same matrices as appeared in the LU decomposition! The current L and U are much easier to find.**
For example

$$\underbrace{\begin{bmatrix} 3 & -4 \\ 2 & 1 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix}}_{L} + \underbrace{\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}}_{D} + \underbrace{\begin{bmatrix} 0 & -4 \\ 0 & 0 \end{bmatrix}}_{U}$$

and

$$\underbrace{\begin{bmatrix} 2 & -6 & 1 \\ 3 & -2 & 0 \\ 4 & -1 & 7 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 3 & 0 & 0 \\ 4 & -1 & 0 \end{bmatrix}}_{L} + \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 7 \end{bmatrix}}_{D} + \underbrace{\begin{bmatrix} 0 & -6 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{U}$$

and, more generally for 3×3 matrices

$$\underbrace{\begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ \bullet & 0 & 0 \\ \bullet & \bullet & 0 \end{bmatrix}}_{L} + \underbrace{\begin{bmatrix} \bullet & 0 & 0 \\ 0 & \bullet & 0 \\ 0 & 0 & \bullet \end{bmatrix}}_{D} + \underbrace{\begin{bmatrix} 0 & \bullet & \bullet \\ 0 & 0 & \bullet \\ 0 & 0 & 0 \end{bmatrix}}_{U}.$$

The Jacobi iteration

The simplest iterative method is called **Jacobi iteration** and the basic idea is to use the $A = L + D + U$ partitioning of A to write $AX = B$ in the form

$$DX = -(L + U)X + B.$$

We use this equation as the motivation to define the iterative process

$$DX^{(k+1)} = -(L + U)X^{(k)} + B$$

which gives $X^{(k+1)}$ as long as D has no zeros down its diagonal, that is as long as D is invertible. This is Jacobi iteration.

Key Point 11

The **Jacobi iteration** for approximating the solution of $AX = B$ where $A = L + D + U$ is given by

$$X^{(k+1)} = -D^{-1}(L + U)X^{(k)} + D^{-1}B$$

Example 18

Use the Jacobi iteration to approximate the solution $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ of

$$\begin{bmatrix} 8 & 2 & 4 \\ 3 & 5 & 1 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix}.$$

Use the initial guess $X^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$.

Solution

In this case $D = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ and $L + U = \begin{bmatrix} 0 & 2 & 4 \\ 3 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix}$.

First iteration.

The first iteration is $DX^{(1)} = -(L+U)X^{(0)} + B$, or in full

$$\begin{bmatrix} 8 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ -3 & 0 & -1 \\ -2 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} + \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix} = \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix},$$

since the initial guess was $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$.

Taking this information row by row we see that

$$\begin{aligned} 8x_1^{(1)} &= -16 \quad \therefore \boxed{x_1^{(1)} = -2} \\ 5x_2^{(1)} &= 4 \quad \therefore \boxed{x_2^{(1)} = 0.8} \\ 4x_3^{(1)} &= -12 \quad \therefore \boxed{x_3^{(1)} = -3} \end{aligned}$$

Thus the first Jacobi iteration gives us $X^{(1)} = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} -2 \\ 0.8 \\ -3 \end{bmatrix}$ as an approximation to X.

Second iteration.

The second iteration is $DX^{(2)} = -(L+U)X^{(1)} + B$, or in full

$$\begin{bmatrix} 8 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ -3 & 0 & -1 \\ -2 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} + \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix}.$$

Taking this information row by row we see that

$$\begin{aligned} 8x_1^{(2)} &= -2x_2^{(1)} - 4x_3^{(1)} - 16 = -2(0.8) - 4(-3) - 16 = -5.6 \quad \therefore \boxed{x_1^{(2)} = -0.7} \\ 5x_2^{(2)} &= -3x_1^{(1)} - x_3^{(1)} + 4 = -3(-2) - (-3) + 4 = 13 \quad \therefore \boxed{x_2^{(2)} = 2.6} \\ 4x_3^{(2)} &= -2x_1^{(1)} - x_2^{(1)} - 12 = -2(-2) - 0.8 - 12 = -8.8 \quad \therefore \boxed{x_3^{(2)} = -2.2} \end{aligned}$$

Therefore the second iterate approximating X is $X^{(2)} = \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} -0.7 \\ 2.6 \\ -2.2 \end{bmatrix}$.

Solution (contd.)

Third iteration.

The third iteration is $DX^{(3)} = -(L+U)X^{(2)} + B$, or in full

$$\begin{bmatrix} 8 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ -3 & 0 & -1 \\ -2 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} + \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix}$$

Taking this information row by row we see that

$$8x_1^{(3)} = -2x_2^{(2)} - 4x_3^{(2)} - 16 = -2(2.6) - 4(-2.2) - 16 = -12.4 \quad \therefore \boxed{x_1^{(3)} = -1.55}$$

$$5x_2^{(3)} = -3x_1^{(2)} - x_3^{(2)} + 4 = -3(-0.7) - (2.2) + 4 = 8.3 \quad \therefore \boxed{x_2^{(3)} = 1.66}$$

$$4x_3^{(3)} = -2x_1^{(2)} - x_2^{(2)} - 12 = -2(-0.7) - 2.6 - 12 = -13.2 \quad \therefore \boxed{x_3^{(3)} = -3.3}$$

Therefore the third iterate approximating X is $X^{(3)} = \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix} = \begin{bmatrix} -1.55 \\ 1.66 \\ -3.3 \end{bmatrix}$.

More iterations ...

Three iterations is plenty when doing these calculations by hand! But the repetitive nature of the process is ideally suited to its implementation on a computer. It turns out that the next few iterates are

$$X^{(4)} = \begin{bmatrix} -0.765 \\ 2.39 \\ -2.64 \end{bmatrix}, \quad X^{(5)} = \begin{bmatrix} -1.277 \\ 1.787 \\ -3.215 \end{bmatrix}, \quad X^{(6)} = \begin{bmatrix} -0.839 \\ 2.209 \\ -2.808 \end{bmatrix},$$

to 3 d.p. Carrying on even further $X^{(20)} = \begin{bmatrix} x_1^{(20)} \\ x_2^{(20)} \\ x_3^{(20)} \end{bmatrix} = \begin{bmatrix} -0.9959 \\ 2.0043 \\ -2.9959 \end{bmatrix}$, to 4 d.p. After about 40 iterations successive iterates are equal to 4 d.p. Continuing the iteration even further causes the iterates to agree to more and more decimal places. The method converges to the exact answer

$$X = \begin{bmatrix} -1 \\ 2 \\ -3 \end{bmatrix}.$$

The following Task involves calculating just two iterations of the Jacobi method.

Carry out two iterations of the Jacobi method to approximate the solution of

$$\begin{bmatrix} 4 & -1 & -1 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

with the initial guess $X^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Your solution

First iteration:

Answer

The first iteration is $DX^{(1)} = -(L+U)X^{(0)} + B$, that is,

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

from which it follows that $X^{(1)} = \begin{bmatrix} 0.75 \\ 1 \\ 1.25 \end{bmatrix}$.

Your solution

Second iteration:

Answer

The second iteration is $DX^{(1)} = -(L+U)X^{(0)} + B$, that is,

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

from which it follows that $X^{(2)} = \begin{bmatrix} 0.8125 \\ 1 \\ 1.1875 \end{bmatrix}$.

Notice that at each iteration the first thing we do is get a new approximation for x_1 and then we continue to use the old approximation to x_1 in subsequent calculations for that iteration! Only at the *next* iteration do we use the new value. Similarly, we continue to use an old approximation to x_2 even after we have worked out a new one. And so on.

Given that the iterative process is supposed to improve our approximations why not use the better values straight away? This observation is the motivation for what follows.

Gauss-Seidel iteration

The approach here is very similar to that used in Jacobi iteration. The only difference is that we use new approximations to the entries of X as soon as they are available. As we will see in the Example below, this means rearranging $(L+D+U)X = B$ slightly differently from what we did for Jacobi. We write

$$(D+L)X = -UX + B$$

and use this as the motivation to define the iteration

$$(D+L)X^{(k+1)} = -UX^{(k)} + B.$$

Key Point 12

The **Gauss-Seidel iteration** for approximating the solution of $AX = B$ is given by

$$X^{(k+1)} = -(D+L)^{-1}UX^{(k)} + (D+L)^{-1}B$$

Example 19 which follows revisits the system of equations we saw earlier in this Section in Example 18.

Example 19

Use the Gauss-Seidel iteration to approximate the solution $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ of $\begin{bmatrix} 8 & 2 & 4 \\ 3 & 5 & 1 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix}$. Use the initial guess $X^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$.

Solution

In this case $D + L = \begin{bmatrix} 8 & 0 & 0 \\ 3 & 5 & 0 \\ 2 & 1 & 4 \end{bmatrix}$ and $U = \begin{bmatrix} 0 & 2 & 4 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$.

First iteration.

The first iteration is $(D + L)X^{(1)} = -UX^{(0)} + B$, or in full

$$\begin{bmatrix} 8 & 0 & 0 \\ 3 & 5 & 0 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} + \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix} = \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix},$$

since the initial guess was $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = 0$.

Taking this information row by row we see that

$$\begin{aligned} 8x_1^{(1)} &= -16 && \therefore \boxed{x_1^{(1)} = -2} \\ 3x_2^{(1)} + 5x_2^{(1)} &= 4 \therefore 5x_2^{(1)} = -3(-2) + 4 && \therefore \boxed{x_2^{(1)} = 2} \\ 2x_1^{(1)} + x_2^{(1)} + 4x_3^{(1)} &= -12 \therefore 4x_3^{(1)} = -2(-2) - 2 - 12 && \therefore \boxed{x_3^{(1)} = -2.5} \end{aligned}$$

(Notice how the new approximations to x_1 and x_2 were used immediately after they were found.)

Thus the first Gauss-Seidel iteration gives us $X^{(1)} = \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \\ -2.5 \end{bmatrix}$ as an approximation to X.

Solution

Second iteration.

The second iteration is $(D+L)X^{(2)} = -UX^{(1)} + B$, or in full

$$\begin{bmatrix} 8 & 0 & 0 \\ 3 & 5 & 0 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} + \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix}$$

Taking this information row by row we see that

$$\begin{aligned} 8x_1^{(2)} &= -2x_2^{(1)} - 4x_3^{(1)} - 16 \quad \therefore \boxed{x_1^{(2)} = -1.25} \\ 3x_1^{(2)} + 5x_2^{(2)} &= -x_3^{(1)} + 4 \quad \therefore \boxed{x_2^{(2)} = 2.05} \\ 2x_1^{(2)} + x_2^{(2)} + 4x_3^{(2)} &= -12 \quad \therefore \boxed{x_3^{(2)} = -2.8875} \end{aligned}$$

Therefore the second iterate approximating X is $X^{(2)} = \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} -1.25 \\ 2.05 \\ -2.8875 \end{bmatrix}$.

Third iteration.

The third iteration is $(D+L)X^{(3)} = -UX^{(2)} + B$, or in full

$$\begin{bmatrix} 8 & 0 & 0 \\ 3 & 5 & 0 \\ 2 & 1 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix} = \begin{bmatrix} 0 & -2 & -4 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} + \begin{bmatrix} -16 \\ 4 \\ -12 \end{bmatrix}.$$

Taking this information row by row we see that

$$\begin{aligned} 8x_1^{(3)} &= -2x_2^{(2)} - 4x_3^{(2)} - 16 \quad \therefore \boxed{x_1^{(3)} = -1.0687} \\ 3x_1^{(3)} + 5x_2^{(3)} &= -x_3^{(2)} + 4 \quad \therefore \boxed{x_2^{(3)} = 2.0187} \\ 2x_1^{(3)} + x_2^{(3)} + 4x_3^{(3)} &= -12 \quad \therefore \boxed{x_3^{(3)} = -2.9703} \end{aligned}$$

to 4 d.p. Therefore the third iterate approximating X is

$$X^{(3)} = \begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \end{bmatrix} = \begin{bmatrix} -1.0687 \\ 2.0187 \\ -2.9703 \end{bmatrix}.$$

More iterations ...

Again, there is little to be learned from pushing this further by hand. Putting the procedure on a computer and seeing how it progresses is instructive, however, and the iteration continues as follows:

$$X^{(4)} = \begin{bmatrix} -1.0195 \\ 2.0058 \\ -2.9917 \end{bmatrix}, \quad X^{(5)} = \begin{bmatrix} -1.0056 \\ 2.0017 \\ -2.9976 \end{bmatrix}, \quad X^{(6)} = \begin{bmatrix} -1.0016 \\ 2.0005 \\ -2.9993 \end{bmatrix},$$

$$X^{(7)} = \begin{bmatrix} -1.0005 \\ 2.0001 \\ -2.9998 \end{bmatrix}, \quad X^{(8)} = \begin{bmatrix} -1.0001 \\ 2.0000 \\ -2.9999 \end{bmatrix}, \quad X^{(9)} = \begin{bmatrix} -1.0000 \\ 2.0000 \\ -3.0000 \end{bmatrix}$$

(to 4 d.p.). Subsequent iterates are equal to $X^{(9)}$ to this number of decimal places. The Gauss-Seidel iteration has converged to 4 d.p. in 9 iterations. It took the Jacobi method almost 40 iterations to achieve this!

Carry out two iterations of the Gauss-Seidel method to approximate the solution of

$$\begin{bmatrix} 4 & -1 & -1 \\ -1 & 4 & -1 \\ -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

with the initial guess $X^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$.

Your solution

First iteration

Answer

The first iteration is $(D+L)X^{(1)} = -UX^{(0)} + B$, that is,

$$\begin{bmatrix} 4 & 0 & 0 \\ -1 & 4 & 0 \\ -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ x_3^{(0)} \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

from which it follows that $X^{(1)} = \begin{bmatrix} 0.75 \\ 0.9375 \\ 1.1719 \end{bmatrix}$.

Your solution

Second iteration

Answer

The second iteration is $(D+L)X^{(1)} = -UX^{(0)} + B$, that is,

$$\begin{bmatrix} 4 & 0 & 0 \\ -1 & 4 & 0 \\ -1 & -1 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

from which it follows that $X^{(2)} = \begin{bmatrix} 0.7773 \\ 0.9873 \\ 1.1912 \end{bmatrix}$.

2. Do these iterative methods always work?

No. It is not difficult to invent examples where the iteration fails to approach the solution of $AX = B$. The key point is related to matrix norms seen in the preceding Section.

The two iterative methods we encountered above are both special cases of the general form

$$X^{(k+1)} = MX^{(k)} + N.$$

1. For the Jacobi method we choose $M = -D^{-1}(L+U)$ and $N = D^{-1}B$.
2. For the Gauss-Seidel method we choose $M = -(D+L)^{-1}U$ and $N = (D+L)^{-1}B$.

The following Key Point gives the main result.

Key Point 13

For the iterative process $X^{(k+1)} = MX^{(k)} + N$ the iteration will converge to a solution if **the norm of M is less than 1**.

Care is required in understanding what Key Point 13 says. Remember that there are lots of different ways of defining the norm of a matrix (we saw three of them). If you can find a norm (*any norm*) such that the norm of M is less than 1, then the iteration will converge. It doesn't matter if there are other norms which give a value greater than 1, all that matters is that there is one norm that is less than 1.

Key Point 13 above makes no reference to the starting "guess" $X^{(0)}$. The convergence of the iteration is independent of where you start! (Of course, if we start with a really bad initial guess then we can expect to need lots of iterations.)

Show that the Jacobi iteration used to approximate the solution of

$$\begin{bmatrix} 4 & -1 & -1 \\ 1 & -5 & -2 \\ -1 & 0 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

is certain to converge. (Hint: calculate the norm of $-D^{-1}(L+U)$.)

Your solution

Answer

The Jacobi iteration matrix is

$$\begin{aligned} -D^{-1}(L+U) &= \begin{bmatrix} 4 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 2 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.25 & 0 & 0 \\ 0 & -0.2 & 0 \\ 0 & 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 2 \\ 1 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0.25 & 0.25 \\ -0.2 & 0 & 0.4 \\ 0.5 & 0 & 0 \end{bmatrix} \end{aligned}$$

and the infinity norm of this matrix is the maximum of $0.25+0.25$, $0.2+0.4$ and 0.5, that is

$$\| -D^{-1}(L+U)\|_\infty = 0.6$$

which is less than 1 and therefore the iteration will converge.

Guaranteed convergence

If the matrix has the property that it is **strictly diagonally dominant**, which means that the diagonal entry is larger in magnitude than the absolute sum of the other entries on that row, then both Jacobi and Gauss-Seidel are guaranteed to converge. The reason for this is that if A is strictly diagonally dominant then the iteration matrix M will have an infinity norm that is less than 1.

A small system is the subject of Example 20 below. A large system with slow convergence is the subject of Engineering Example 1 on page 62.

Example 20

Show that $A = \begin{bmatrix} 4 & -1 & -1 \\ 1 & -5 & -2 \\ -1 & 0 & 2 \end{bmatrix}$ is strictly diagonally dominant.

Solution

Looking at the diagonal entry of each row in turn we see that

$$\begin{aligned} 4 &> |-1| + |-1| = 2 \\ |-5| &> 1 + |-2| = 3 \\ 2 &> |-1| + 0 = 1 \end{aligned}$$

and this means that the matrix is strictly diagonally dominant.

Given that A above *is* strictly diagonally dominant it is certain that both Jacobi and Gauss-Seidel will converge.

What's so special about strict diagonal dominance?

In many applications we can be certain that the coefficient matrix A will be strictly diagonally dominant. We will see examples of this in HELM 32 and HELM 33 when we consider approximating solutions of differential equations.

Exercises

1. Consider the system

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$$

   (a) Use the starting guess $X^{(0)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ in an implementation of the Jacobi method to show that $X^{(1)} = \begin{bmatrix} 1.5 \\ -3 \end{bmatrix}$. Find $X^{(2)}$ and $X^{(3)}$.

   (b) Use the starting guess $X^{(0)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ in an implementation of the Gauss-Seidel method to show that $X^{(1)} = \begin{bmatrix} 1.5 \\ -3.25 \end{bmatrix}$. Find $X^{(2)}$ and $X^{(3)}$.

   (Hint: it might help you to know that the exact solution is $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ -4 \end{bmatrix}$.)

2. (a) Show that the Jacobi iteration applied to the system

$$\begin{bmatrix} 5 & -1 & 0 & 0 \\ -1 & 5 & -1 & 0 \\ 0 & -1 & 5 & -1 \\ 0 & 0 & -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 7 \\ -10 \\ -6 \\ 16 \end{bmatrix}$$

   can be written

$$X^{(k+1)} = \begin{bmatrix} 0 & 0.2 & 0 & 0 \\ 0.2 & 0 & 0.2 & 0 \\ 0 & 0.2 & 0 & 0.2 \\ 0 & 0 & 0.2 & 0 \end{bmatrix} X^{(k)} + \begin{bmatrix} 1.4 \\ -2 \\ -1.2 \\ 3.2 \end{bmatrix}.$$

   (b) Show that the method is certain to converge and calculate the first three iterations using zero starting values.

   (Hint: the exact solution to the stated problem is $\begin{bmatrix} 1 \\ -2 \\ 1 \\ 3 \end{bmatrix}$.)

Answers

1. (a) $2x_1^{(1)} = 2 - 1x_2^{(0)} = 2$

and therefore $x_1^{(1)} = 1.5$

$$2x_2^{(1)} = -5 - 1x_1^{(0)} = -6$$

which implies that $x_2^{(1)} = -3$. These two values give the required entries in $X^{(1)}$. A second and third iteration follow in a similar way to give

$$X^{(2)} = \begin{bmatrix} 2.5 \\ -3.25 \end{bmatrix} \quad \text{and} \quad X^{(3)} = \begin{bmatrix} 2.625 \\ -3.75 \end{bmatrix}$$

(b) $2x_1^{(1)} = 2 - 1x_2^{(0)} = 3$

and therefore $x_1^{(1)} = 1.5$. This new approximation to x_1 is used straight away when finding a new approximation to $x_2^{(1)}$.

$$2x_2^{(1)} = -5 - 1x_1^{(1)} = -6.5$$

which implies that $x_2^{(1)} = -3.25$. These two values give the required entries in $X^{(1)}$. A second and third iteration follow in a similar way to give

$$X^{(2)} = \begin{bmatrix} 2.625 \\ -3.8125 \end{bmatrix} \quad \text{and} \quad X^{(3)} = \begin{bmatrix} 2.906250 \\ -3.953125 \end{bmatrix}$$

where $X^{(3)}$ is given to 6 decimal places

2. (a) In this case $D = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$ and therefore $D^{-1} = \begin{bmatrix} 0.2 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix}$.

So the iteration matrix $M = D^{-1}\begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.2 & 0 & 0 \\ 0.2 & 0 & 0.2 & 0 \\ 0 & 0.2 & 0 & 0.2 \\ 0 & 0 & 0.2 & 0 \end{bmatrix}$

and that the Jacobi iteration takes the form

$$X^{(k+1)} = MX^{(k)} + M^{-1}\begin{bmatrix} 7 \\ -10 \\ -6 \\ 16 \end{bmatrix} = \begin{bmatrix} 0 & 0.2 & 0 & 0 \\ 0.2 & 0 & 0.2 & 0 \\ 0 & 0.2 & 0 & 0.2 \\ 0 & 0 & 0.2 & 0 \end{bmatrix} X^{(k)} + \begin{bmatrix} 1.4 \\ -2 \\ -1.2 \\ 3.2 \end{bmatrix}$$

as required.

Answers

2(b)

Using the starting values $x_1^{(0)} = x_2^{(0)} = x_3^{(0)} = x_4^{(0)} = 0$, the first iteration of the Jacobi method gives

$$\begin{aligned} x_1^1 &= 0.2x_2^0 + 1.4 = 1.4 \\ x_2^1 &= 0.2(x_1^0 + x_3^0) - 2 = -2 \\ x_3^1 &= 0.2(x_2^0 + x_4^0) - 1.2 = -1.2 \\ x_4^1 &= 0.2x_3^0 + 3.2 = 3.2 \end{aligned}$$

The second iteration is

$$\begin{aligned} x_1^2 &= 0.2x_2^1 + 1.4 = 1 \\ x_2^2 &= 0.2(x_1^1 + x_3^1) - 2 = -1.96 \\ x_3^2 &= 0.2(x_2^1 + x_4^1) - 1.2 = -0.96 \\ x_4^2 &= 0.2x_3^1 + 3.2 = 2.96 \end{aligned}$$

And the third iteration is

$$\begin{aligned} x_1^3 &= 0.2x_2^2 + 1.4 = 1.008 \\ x_2^3 &= 0.2(x_1^2 + x_3^2) - 2 = -1.992 \\ x_3^3 &= 0.2(x_2^2 + x_4^2) - 1.2 = -1 \\ x_4^3 &= 0.2x_3^2 + 3.2 = 3.008 \end{aligned}$$

Engineering Example 1

Detecting a train on a track

Introduction

One means of detecting trains is the 'track circuit' which uses current fed along the rails to detect the presence of a train. A voltage is applied to the rails at one end of a section of track and a relay is attached across the other end, so that the relay is energised if no train is present, whereas the wheels of a train will short circuit the relay, causing it to de-energise. Any failure in the power supply or a breakage in a wire will also cause the relay to de-energise, for the system is fail safe. Unfortunately, there is always leakage between the rails, so this arrangement is slightly complicated to analyse.

Problem in words

A 1000 m track circuit is modelled as ten sections each 100 m long. The resistance of 100 m of one rail may be taken to be 0.017 ohms, and the leakage resistance across a 100 m section taken to be 30 ohms. The detecting relay and the wires to it have a resistance of 10 ohms, and the wires from the supply to the rail connection have a resistance of 5 ohms for the pair. The voltage applied at the supply is $4V$. See diagram below. What is the current in the relay?

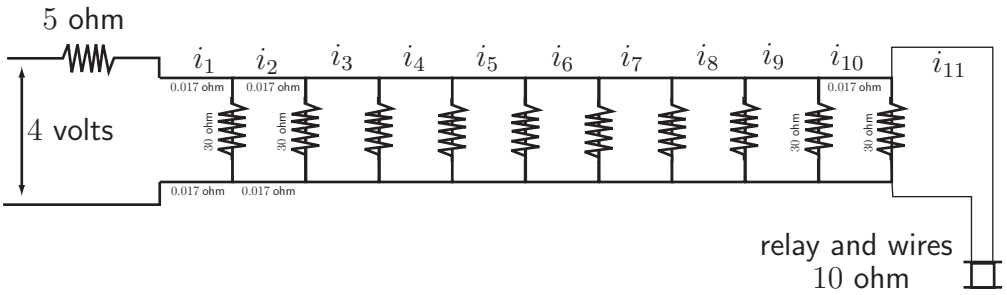


Figure 1

Mathematical statement of problem

There are many ways to apply Kirchhoff's laws to solve this, but one which gives a simple set of equations in a suitable form to solve is shown below. i_1 is the current in the first section of rail (i.e. the one close to the supply), i_2 , $i_3, \ldots i_{10}$, the current in the successive sections of rail and i_{11} the current in the wires to the relay. The leakage current between the first and second sections of rail is $i_1 - i_2$ so that the voltage across the rails there is $30(i_1 - i_2)$ volts. The first equation below uses this and the voltage drop in the feed wires, the next nine equations compare the voltage drop across successive sections of track with the drop in the (two) rails, and the last equation compares the voltage drop across the last section with that in the relay wires.

$$\begin{aligned}
30(i_1 - i_2) + (5.034)i_1 &= 4 \\
30(i_1 - i_2) &= 0.034i_2 + 30(i_2 - i_3) \\
30(i_2 - i_3) &= 0.034i_2 + 30(i_3 - i_4) \\
&\vdots \\
30(i_9 - i_{10}) &= 0.034i_{10} + 30(i_{10} - i_{11}) \\
30(i_{10} - i_{11}) &= 10i_{11}
\end{aligned}$$

These can be reformulated in matrix form as $A\underline{i} = \underline{v}$, where $\underline{v}$ is the 11×1 column vector with first entry 4 and the other entries zero, $\underline{i}$ is the column vector with entries $i_1, i_2, \ldots, i_{11}$ and A is the matrix

$$A = \begin{bmatrix} 35.034 & -30 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -30 & 60.034 & -30 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -30 & 60.034 & -30 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -30 & 60.034 & -30 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -30 & 60.034 & -30 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -30 & 60.034 & -30 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -30 & 60.034 & -30 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -30 & 60.034 & -30 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -30 & 60.034 & -30 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -30 & 60.034 & -30 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -30 & 40 \end{bmatrix}$$

Find the current i_1 in the relay when the input is $4V$, by Gaussian elimination or by performing an L-U decomposition of A.

Mathematical analysis

We solve $A\underline{i} = \underline{v}$ as above, although actually we only want to know i_{11}. Letting M be the matrix A with the column $\underline{v}$ added at the right, as in Section 30.2, then performing Gaussian elimination on M, working to four decimal places gives

$$M = \begin{bmatrix} 35.0340 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4.0000 \\ 0 & 34.3447 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.4252 \\ 0 & 0 & 33.8291 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.9919 \\ 0 & 0 & 0 & 33.4297 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 2.6532 \\ 0 & 0 & 0 & 0 & 33.1118 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 2.3810 \\ 0 & 0 & 0 & 0 & 0 & 32.8534 & -30.0000 & 0 & 0 & 0 & 0 & 2.1572 \\ 0 & 0 & 0 & 0 & 0 & 0 & 32.6396 & -30.0000 & 0 & 0 & 0 & 1.9698 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32.4601 & -30.0000 & 0 & 0 & 1.8105 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32.3077 & -30.0000 & 0 & 1.6733 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32.1769 & -30.0000 & 1.5538 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12.0296 & 1.4487 \end{bmatrix}$$

from which we can calculate that the solution $\underline{i}$ is

$$\underline{i} = \begin{bmatrix} 0.5356 \\ 0.4921 \\ 0.4492 \\ 0.4068 \\ 0.3649 \\ 0.3234 \\ 0.2822 \\ 0.2414 \\ 0.2008 \\ 0.1605 \\ 0.1204 \end{bmatrix}$$

so the current in the relay is 0.1204 amps, or 0.12 A to two decimal places.

You can alternatively solve this problem by an L-U decomposition by finding matrices L and U such that $M = LU$. Here we have

$$L = \begin{bmatrix} 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.8563 & 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.8735 & 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.8868 & 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.8974 & 1.0000 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.9060 & 1.0000 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.9131 & 1.0000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.9191 & 1.0000 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.9242 & 1.0000 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.9286 & 1.0000 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.9323 & 1.0000 \end{bmatrix}$$

and

$$U = \begin{bmatrix} 35.0340 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 34.3447 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 33.8291 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 33.4297 & -30.0000 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 33.1118 & -30.0000 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 32.8534 & -30.0000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 32.6395 & -30.0000 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32.4601 & -30.0000 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32.3076 & -30.0000 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 32.1768 & -30.0000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12.0295 \end{bmatrix}$$

Therefore $U\underline{i} = \begin{bmatrix} 4.0000 \\ 3.4240 \\ 2.9892 \\ 2.6514 \\ 2.3783 \\ 2.1547 \\ 1.9673 \\ 1.8079 \\ 1.6705 \\ 1.5519 \\ 1.4464 \end{bmatrix}$ and hence $\underline{i} = \begin{bmatrix} 0.5352 \\ 0.4917 \\ 0.4487 \\ 0.4064 \\ 0.3644 \\ 0.3230 \\ 0.2819 \\ 0.2411 \\ 0.2006 \\ 0.1603 \\ 0.1202 \end{bmatrix}$ and again the current is found to be 0.12 amps.

Mathematical comment

You can try to solve the equation $A\underline{i} = \underline{v}$ by Jacobi or Gauss-Seidel iteration but in both cases it will take very many iterations (over 200 to get four decimal places). Convergence is very slow because the norms of the relevant matrices in the iteration are only just less than 1. Convergence is nevertheless assured because the matrix A is diagonally dominant.

Index for Workbook 30

Workbook 30

HELM®

HELM: Helping Engineers Learn Mathematics

http://helm.lboro.ac.uk