Numerical Solutions of Differential Equations

Waleed A A Ali

Table of contents

# Introduction

This unit will cover many numerical methods and how they can be implemented in MATLAB. The programming language MATLAB is an incredibly versatile and robust platform used in many industries including engineering, aeronautics, astronautics, manufacturing, telecommunication and many more.

# 1. Internal Workings of MATLAB

## 1.1 Floating-Point Arithmetic

Since computers have limited resources, only a finite strict subset of the real numbers can be represented. This set of possible stored values is known as ***Floating-Point Numbers*** and these are characterised by properties that are different from those in , since any real number is – in principle – truncated by the computer, giving rise to a new number denoted by , which does not necessarily coincide with the original number .

A computer represents a real number as a floating-point number in as

where:

* determines the sign of the number;
* is the base;
* is the exponent.
* is the mantissa (or significand). The mantissa has length which is the maximum number of digits that can be stored. Each term in the mantissa must satisfy for all and (to ensure that the same number cannot have different representations). The digits (with ) are often called the first significant digits of .

The set is therefore fully characterised by the basis , the number of significant digits and the range of values that can take which is with and .

A computer typically uses binary representation, meaning that the base is with the available digits (also known as bits) and each digit is the coefficient of a power of 2. Available platforms (like MATLAB and Python) typically use the IEEE754 double precision format for , which uses 64-bits as follows:

* 1 bit for (either 0 or 1) to determine the sign;
* 11 bits for (which can be );
* 52 bits for (since , it has to be equal to 1).

This gives a total of 64-bits with and . For 32-bit storage, the exponent is at most 7 and the mantissa has 23 digits. Note that 0 does not belong to since it cannot be represented in the form shown in [Equation 1.1](#eq-float) and it is therefore handled separately.

If a non-zero real number is replaced by its floating-point representation , then there will inevitably be a round-off error which is given by

where is the machine epsilon and provides the distance between 1 and its closest floating-point number greater than 1. In the binary case, (this can be found using eps(1)). In other words, the number is the maximum relative error that the computer can make while representing a real number by finite arithmetic.

Moreover, since and are finite, one cannot represent numbers whose absolute value is either arbitrarily large or arbitrarily small. Precisely, the smallest and the largest positive real numbers of are given by and , respectively. A positive number smaller than yields under-flow and a positive number greater than yields over-flow. The elements in are more dense near , and less dense while approaching . However, the relative distance is small in both cases. The largest and smallest values that can be represented as floating-point numbers in MATLAB can be found using the realmin and realmax commands.

>> realmin  
ans =  
 2.2251e-308  
>> realmax  
ans =  
 1.7977e308

Since is a strict subset of , elementary algebraic operations on floating-point numbers do not inherit all the properties of analogous operations on . Precisely, commutativity still holds for addition and multiplication, i.e.  and . Associativity is violated whenever a situation of overflow or underflow occurs or, similarly, whenever two numbers with opposite signs but similar absolute values are added, the result may be quite inexact and the situation is referred to as loss of significant digits.

Properly handling floating point computations can be tricky sometimes and, if not correctly done, may have serious consequences. There are many webpages (and books) collecting examples of different disasters caused by a poor handling of computer arithmetic or a bad algorithmic implementation. See, for instance, see [Software Bugs](https://www5.in.tum.de/~huckle/bugse.html) and the [Patriot Missile Fail](http://ta.twi.tudelft.nl/users/vuik/wi211/disasters.html) among others.

## 1.2 Computational Complexity

The ***Computational Complexity*** of a function can be defined as the relationship between the size of the input and the difficulty of running the function to completion. The size (or at least, an attribute of the size) of the input is usually denoted , for instance, for a 1-D array, can be its length.

The difficulty of a problem can be measured in several ways. One suitable way to describe the difficulty of the problem is to count the number of basic, or ***Floating-Point Operations***, such as additions, subtractions, multiplications, divisions and assignments. Floating-point operations, also called *flops*, usually measure the speed of a computer, measured as the maximum number of floating-point operations which the computer can execute in one second. Although each basic operation takes a different amount of time, the number of basic operations needed to complete a function is sufficiently related to the running time to be useful, and it is usually easy to count and less dependent on the specific machine (hardware) that is used to perform the computations.

A common notation for complexity is the ***Big-O*** notation (denoted ), which establishes the relationship in the growth of the number of basic operations with respect to the size of the input as the input size becomes very large. In general, the basic operations grow in direct response to the increase in the size of the input and, as gets large, the highest power dominates. Therefore, only the highest power term is included in Big-O notation; moreover, coefficients are not required to characterise growth and are usually dropped (although this will also depend on the precision of the estimates).

Formally, a function behaves as as tends to infinity if

For example, the polynomial behaves like as tends to infinity since this term will be the fastest to grow. This can be written as as .

|  |
| --- |
| Couting flops |
| Let be given by  This function can be coded as fun in MATLAB as follows:  function [out]=fun(n)  out = 0;  for i=1:1:n   for j=1:1:n   out = out + i\*j;    end   end  end  For example, should perform the overall calculation  so fun(3) should output out=36.  This code requires the following operations:   * assignments:   + 1: out=0;   + : i=1:1:n;   + : for every i, j=1:1:n;   + : for every i, out=out+i\*j; * multiplications: i\*j; * additions: out=out+i\*j.   Therefore, for any , this code will need flops, meaning that the computational complexity is , i.e. the code runs in polynomial time. It is not uncommon to find algorithms that run in exponential time , like some recursive algorithms, or in logarithmic time . |

For more complicated codes, it is important to see where most of the time is spent in a code and how execution can be improved. A rudimentary way of timing can be done by the toc toc:

>> tic;  
>> Run code or code block  
>> toc;

This will produce a simple time in seconds that MATLAB took from tic until toc, so if toc has not been types, then the timer will continue.

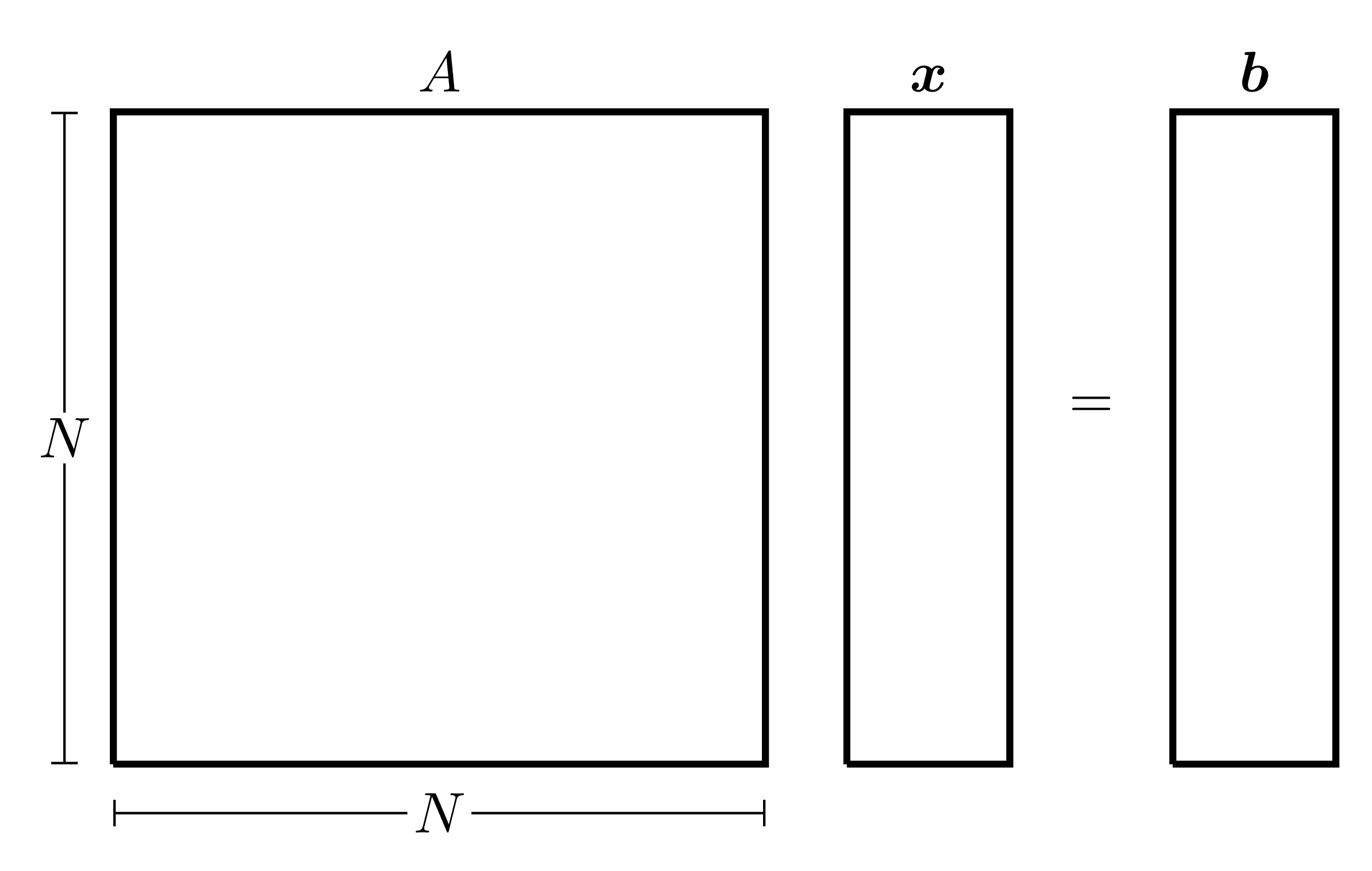
For more advanced analysis, MATLAB uses a **Code Profiler** to analyse code which includes run times for each iteration, times a code has been called and a lot more.

|  |
| --- |
| Code Profiling |
| Suppose that a code needs to be written that finds the Fibonacci number starting the sequence with (1,1). This can be done in two ways:   * **Iteratively** by having a self-contained code that generates all the terms of the sequence up to and displays the last term.   function [F]=Fib\_Iter(N)  S=ones(1,N);  for n=3:1:N   S(n)=S(n-1)+S(n-2);  end  F=S(end);  end   * **Recursively** by have a self-referential code that keeps referring back to itself to generate last term in the sequence from the previous terms.   function [F]=Fib\_Rec(N)  if N<3   F=1;  else   F=Fib\_Rec(N-1)+Fib\_Rec(N-2);  end  end  When running these codes for an input of , the times are very short, of the order of seconds but as gets larger, the recursive code starts to take much longer. Suppose the code efficiency is to be analysed for the input , this can be done using the profiler as follows:  >> profile on >> Fib\_Iter(40); >> profile off >> profile viewer  This will give a full breakdown of how many times every line was run and how much time it took. For Fib\_Iter(40), a total of 38 operations were performed, each taking such a short amount of time that it registers as “0 seconds”.    However, performing the profiler for Fib\_Rec(40) gives a *dramatically* different answer with the code taking nearly 247 seconds and having to call itself more than 102 million times.    This is why it is important to profile longer codes to see which parts take the longest time and which loops are the most time consuming. |

|  |
| --- |
| Good Practice |
| To reduce computational time in general, avoid self-referential codes because these tend to grow in usage exponentially. Another important practice is to use in-built MATLAB syntax, like using sum to add elements in a vector rather than manually hard coding it. This is where being familiar with a lot of the MATLAB syntax is important; MATLAB has a lot of built-in codes and syntaxes which can save a lot of time. |

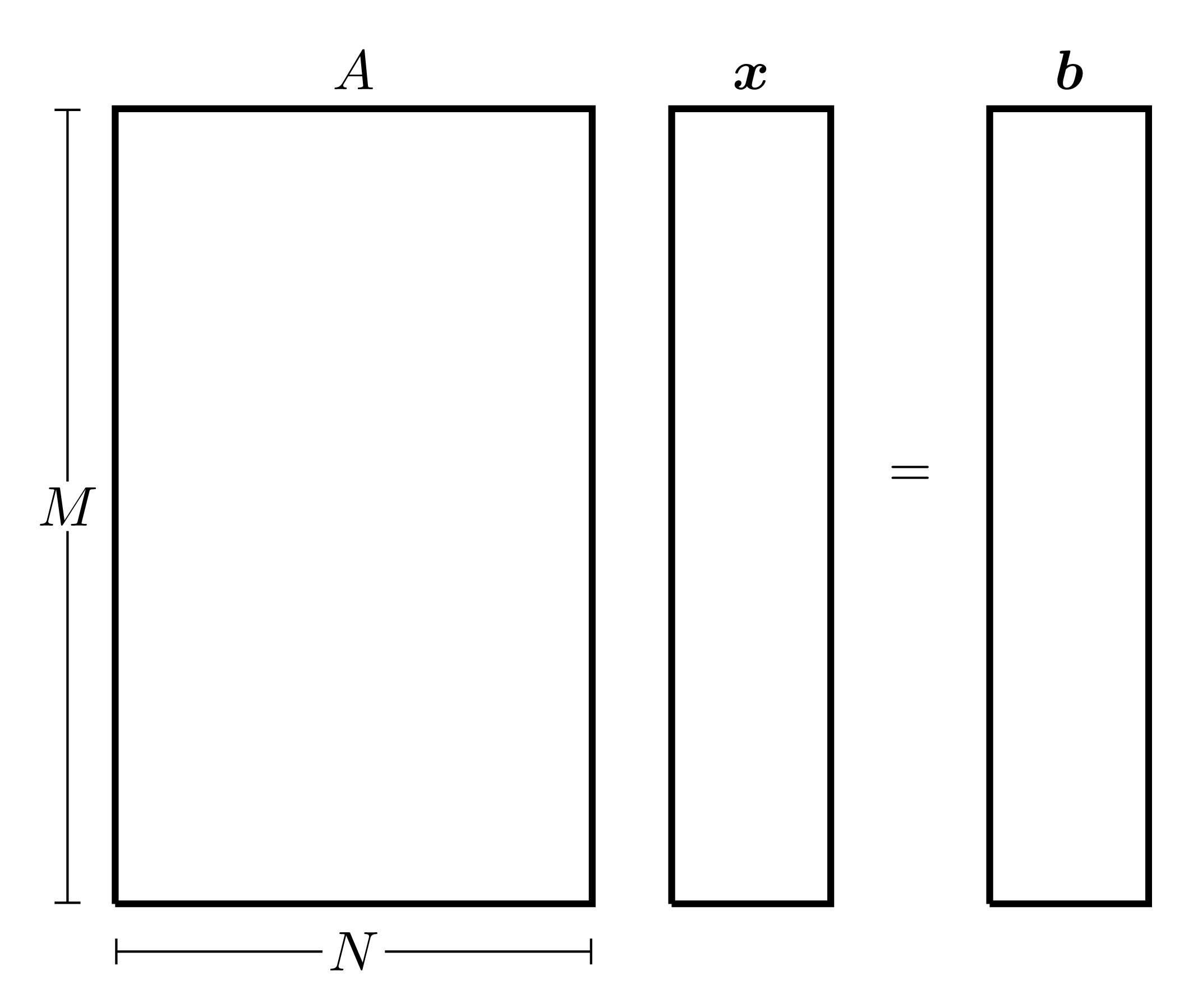
# 2. Linear Algebra

In this unit, the following Numerical Linear Algebra processes will be studies:

* Solving linear systems of the form
* 

This can be done by using a *Direct Method* if the solution of the system can be obtained in a finite number of steps or an *Iterative Method* if the solution, in principle, requires an infinite number of steps.

The choice between direct and iterative methods may depend on several factors, primarily the predicted theoretical efficiency of the scheme, but also the particular type of matrix (such as systems that are sparse, diagonally dominant, tridiagonal and so forth) and the memory storage requirements.

* Solving, or at least understanding, the nature of *overdetermined* problems of the form
* 

This problem often arises when fitting models to a large data sets, where represents parameters in the model and represents the results (or measurements) of some experiments.

* Solving eigenvalue problems of the form
* and the generalisations thereof.

## 2.1 Solving Linear Systems of Equations

The first part of linear algebra is solving the square linear system

This is a situation when the LHS forms a system of equations with a vector of unknowns and the RHS is known.

|  |
| --- |
| Simple Example of Linear System |
| Let and be integers such that: - their sum is equal to 20 - is twice as large as - is bigger than by 10.  These three relationships can be written in equation form as:  This can written in matrix form as: |

There are many way in which this can be done, depending on the form of the matrix:

* Direct Methods:
  + Direct substitution for diagonal systems
  + Forward substitution for lower triangular systems
  + Backward substitution for upper training systems
  + TDMA for tridiagonal systems
  + Cramer’s Rule and Gaussian Elimination for more general matrix systems
* Iterative Methods
  + Jacobi:
  + Gauss-Seidel
* In-built Methods:
  + Backsklash

### 2.1.1 Computational Stability of Linear Systems

Before tackling any linear algebra techniques, it is important to understand ***Computational Stability***.

Consider the linear system

In real-life applications, the matrix is usually fully known and often invertible while the vector may not be known exactly and its measurement may often include rounding errors. Suppose that the vector has a small error , then the solution will also have a small error , meaning that the system will in fact be

Subtracting form [Equation 2.1](#eq-err1) gives , therefore .

For , consider the ratio between the -norm of the error and the -norm of the exact solution :

|  |
| --- |
| Note 2.1: Submultiplicative Property of Matrix Norms |
| For a matrix and a vector , |

Let , then

The quantity is called the ***Condition Number***[[1]](#footnote-53) and it can be regarded as a measure of how sensitive a matrix is to perturbations, in other words, it gives an indication as to the stability of the matrix system. A problem is ***Well-Conditioned*** if the condition number is small, and is ***Ill-Conditioned*** if the condition number is large (the terms “*small*” and “*large*” are somewhat subjective here and will depend on the context). Bear in mind that in practice, calculating the condition number may be computationally expensive since it requires inverting the matrix .

The condition number derived above follows the assumption that the error only occurs in which then results in an error in . If an error is also committed in , then for sufficiently small , the error bound for the ratio is

An example for which is large is a discretisation matrix of a PDE, in this case, the condition number of can be very large and increases rapidly as the number of mesh points increases. For example, for a PDE with mesh points in 2-dimensions, the condition number is of order and it is not uncommon to have between and . In this case, errors in may be amplified enormously in the solution process. Thus, if is large, there may be difficulties in solving the system reliably, a problem which plagues calculations with partial differential equations.

Moreover, if is large, then the system may be solved using an *iterative method* which generate a sequence of approximations to while ensuring that each iteration is easy to perform and that rapidly tends to , within a certain tolerance, as tends to infinity. If is large, then the number of iterations to reach this tolerance increases rapidly as the size of increases, often being proportional to or even to . Thus not only do errors in accumulate for large , but the number of computation required to find increases as well.

In MATLAB, the condition number can be calculated using the cond(A,p) command where A is the square matrix in question and p is the chosen norm which can only be equal to 1, 2, inf or 'Fro' (when using the Frobenius norm). Also note that cond(A) without the second argument p produces the condition number with the 2-norm by default.

#### Properties of the Condition Number

Let and be invertible matrices, and . The condition number has the following properties:

* ;
* if and only if is an orthogonal matrix, i.e. ;
* ;
* ;
* .

## 2.2 Direct Methods

Direct methods can be used to solve matrix systems in a finite number of steps, although these steps could possibly be computationally expensive.

### 2.2.1 Direct Substitution

Direct substitution is the simplest direct method and requires the matrix to be a diagonal with none of the diagonal terms being 0 (otherwise the matrix will not be invertible).

Consider the matrix system where

and . Direct substitution involves simple multiplication and division:

The solution can be written explicitly as for all . Every step can done independently, meaning that direct substitution lends itself well to parallel computing. In total, direct substitution requires exactly computations (all being division).

|  |
| --- |
| Example of Direct Substituion |
| Consider the system where  Solving the system using direct substitution: |

### 2.2.2 Forward/Backward Substitution

Forward/backward substitution require that the matrix be lower/upper triangular.

Consider the matrix system where

and (so that the determinant is non-zero). The matrix is upper triangular in this case and will require backwards substitution:

Backward substitution involves using the solutions from the later equations to solve the earlier ones, this gives:

This can be written more explicitly as:

A similar version can be obtained for the forward substitution for lower triangular matrices. For any , calculating it requires 1 division, multiplications and subtractions. Therefore cumulatively, require divisions, multiplications and additions with one more division required for , meaning that in total, backward (and forward) substitution requires computations.

|  |
| --- |
| Example of Backward Substituion |
| Consider the system where  This problem can be solved by suing backward substitution: |

### 2.2.3 TDMA Algorithm

The ***TriDiagonal Matrix Algorithm***, abbreviated as ***TDMA*** (also called the ***Thomas Algorithm***) was developed by Llewellyn Thomas which solves tridiagonal matrix systems.

Consider the matrix system where

The terms denote the diagonal elements, denote subdiagonal elements (left of the diagonal terms) and denote the superdiagonal elements (right of the diagonal terms). The TDMA algorithm has explicit formulas for the solution using a forward and backwards sweep that convert the tridiagonal system into an upper triangular system as follows:

The computational complexity can be calculated as follows:

| Term |  |  |  |
| --- | --- | --- | --- |
|  | 0 | 0 | 1 |
|  | 1 | 1 | 1 |
|  |  |  |  |
|  | 1 | 1 | 1 |
|  | 0 | 0 | 1 |
|  | 2 | 2 | 1 |
|  |  |  |  |
|  | 2 | 2 | 1 |
|  | 2 | 2 | 1 |
|  | 1 | 1 | 0 |
|  | 1 | 1 | 0 |
|  |  |  |  |
|  | 1 | 1 | 0 |

This gives a total of computations for , computations for and computations for giving a total of computations.

There are other similar algorithms to the TDMA algorithm, like those the solve pentadiagonal systems and those where the tridiagonal matrix starts or ends with a full row.

### 2.2.4 Cramer’s Rule

***Cramer’s Rule*** is a method that can be used to solve *any* system (of course provided that is non-singular).

Cramer’s rule states that the elements of the vector are given by

where is the matrix obtained from by replacing the column by . This method seems very simple to execute thanks to its very simple formula, but in practice, it can be very computationally expensive.

|  |
| --- |
| Example of Cramer’s Rule |
| Consider the system where  The determinant of is equal to 7. Using Cramer’s rule, the solution can be calculated as: |

Generally, for a matrix of size , the determinant will require computations (other matrix forms or methods may require fewer, of at least). Cramer’s rule requires calculating the determinants of matrices each is size and performing divisions, therefore the computational complexity of Cramer’s rule is . This means that if a machine runs at 1 Gigaflops per second ( flops), then a matrix system of size will require 1620 years to compute.

### 2.2.5 Gaussian Elimination Method

The ***Gaussian Elimination Method*** is an algorithm that transforms the linear system where and into an equivalent upper triangular system after steps, where is an upper triangular matrix and . This uses ***Elementary Row Operations*** (swapping rows, multiplying a row by a constant, adding two rows), after which point, the system can solved by the backward substitution. Note that this method is possible when the elementary row operations are performed on both and simultaneously, so if rows and are swapped in , the rows and must also be swapped in , simialry for the other operations.

The Gaussian elimination method can be performed as follows (the superscripts in brackets will be the step number):

|  |
| --- |
| Parallel Example |
| The algorithm will be explained and an example will be done in parallel to explain the steps with the matrix system where |

1. **Establish the starting matrix**: If , then set and as

* If , then swap the first row with any other row whose first term is not zero and the result will be the starting matrix .

|  |
| --- |
|  |
|  |

1. **Form the multiplier vector**: The desired outcome is to have the matrix be upper triangular, i.e. all the terms below the diagonal should be 0. To achieve this, introduce a vector of multipliers, whose entry is given by

* hence the reason why the assumption must be imposed. Essentially, the vector is the first column of divided the the first element of .

|  |
| --- |
|  |
|  |

1. **Elimination terms in the first column**: For , multiply row 1 by and add it to row to give the new row :

|  |
| --- |
| Row Operations |
|  |

Notice that by the definition of , the first element in every row must be equal to 0, therefore, this set of operation makes all the terms in the first column equal to 0 except the first. Define this new matrix as the second term in the iteration: where for all

|  |
| --- |
|  |
|  |

1. **Modification of the right hand side**: The vector has to also undergo the same operations as , i.e. for , let row of be row 1 multiplied by plus row and the final vector is the vector .

|  |
| --- |
|  |
|  |

1. **Matrix representation of elimination**: This whole procedure can be written as and where

|  |
| --- |
|  |
| To check: |

1. **Repeat for other columns**: The process must now be repeated for the rest of the rows, specifically, those that have non-zero pivot points, i.e. the first point in a row that is non-zero. This process can be done more simply by generating the matrices in the same way as before without going through the starting steps. This process should be reapeated until the last row is reached.

|  |
| --- |
| Multiplier Matrices |
| The matrix can be generated in the same way as , so  To check: |

1. **Solve using backwards substitution**: After repeating for all other columns (a total of times), the final matrix will be an upper triangular matrix with non-zero terms on the diagonal and the system can then be solved by backwards substitution.

|  |
| --- |
| Backwards Substitution |
|  |

The total number of operations in every step is given in the table below (the “steps” here refer to the matrix manipulation step and not exactly to the step numbers of the algorithm):

| Step | Multiplications | Additions | Divisions |
| --- | --- | --- | --- |
| 1 |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

This means that the total number of multiplications is

similarly for the additions. Whereas the total number of divisions is

Therefore the total number of operations is

This means that for large , the Gaussian elimination algorithm requires operations when is a non-sparse matrix. This procedure is computationally expensive even for moderate sized matrices, this also assumes that the pivot points are non-zero, or more specifically, that the matrix has non-zero determinant. As an illustration of this computational complexity, if (which not atypical), then for a computer with the computing power of 1 Gigaflops per second, an system will need 21 years to find a solution. A lot of more modern computational techniques are based on attempting to reduce this computational complexity, either by eliminating terms in some suitable way or chnaging the matrix in a more pallatable form.

Overall, every step of this process can be represented by a matrix transformation . This means that in order to convert the matrix into an upper triangular matrix , the matrix transformations have to be applied reverse order as

This can be written as

Notice that every matrix is lower triangular and this fact will be used later on in **?@sec-LU**.

## 2.3 Iterative Methods

For a large matrix , solving the system directly can be computationally restrictive as seen in the different methods shown in [Section 2.2](#sec-Dir). An alternative would be to use *iterative* methods which generate a sequence of approximations to the exact solution . The hope is that the iterative method converges to the exact solution, i.e.

A possible strategy to realise this process is to consider the following recursive definition

where is a suitable matrix called the ***Iteration Matrix*** (which would generally depend on ) and is a suitable vector (depending on and ). Since the iterations must tend to as tends to infinity, then

{#eqn-Bxg} Next, a sufficient condition needs to be derived; define as the error incurred from iteration , i.e.  and consider the linear systems

Subtracting these gives

For the sake of argument, consider the case when the matrix is symmetric. In order to find a bound for the error, take the 2-norm of the error equation

By the submultiplicative property of matrix norms given in [Note 2.1](#nte-SubMult) and the fact that the 2-norm of a symmetric matrix is equal to the spectral radius (largest eigenvalue in absolute terms), the error can be bounded above

This can be iterated backwards, so for ,

Generally, this means that the error at any iteration can be bounded above by the error at the initial iteration . Therefore, since is arbitrary, if then the set of vectors generated by the iterative scheme will converge to the exact solution which solves , hence giving a sufficient condition for convergence.

### 2.3.1 Constructing an Iterative Method

A general technique to devise an iterative method to solve is based on a “splitting” of the matrix . First, write the matrix as where is a suitable non-singular matrix (somehow linked to and “easy” to invert). Then

Therefore, the vector can be written implicitly as

which is of the form given in **?@eq-Bxg** where and . It would then stand to reason that if the iterative procedure was of the form

(as in [Equation 2.3](#eq-Bxgk)), then the method should converge to the exact solution (provided a suitable choice for ). Of course, for the iterative procedure, the iteration needs an initial vector to start which will be

The choice of the matrix should depend on in some way. So suppose that the matrix is broken down into three parts, where is the matrix of the diagonal entries of , is the strictly lower triangular part or (i.e. not including the diagonal) and is the strictly upper triangular part of .

|  |
| --- |
| Note |
| For example |

The matrix can be chosen as a combination of and in some way to generate the matrix . There are many ways in which this can be done, each with their own advantages:

* ***Jacobi Method***:

The matrix is chosen to be equal to the diagonal part of , then the splitting procedure gives the iteration matrix and the iteration itself is for , which can be written component-wise as

If is strictly diagonally dominant by rows[[2]](#footnote-92), then the Jacobi method converges (i.e. , where ). Note that each component of the new vector is computed independently of the others, meaning that the update is simultaneous which makes this method suitable for parallel programming.

* ***Gauss-Seidel Method***:

The matrix is chosen to be equal to the lower triangular part of , therefore the iteration matrix is given by and the iteration itself is which can be written component-wise as

Contrary to Jacobi method, Gauss-Seidel method updates the components in sequential mode.

There are many other methods that use splitting like:

* Richardson method: where is the identity matrix and
* Damped Jacobi method: for some
* Successive over-relaxation method: for some
* Symmetric successive over-relaxation method: for some

### 2.3.2 Computational Cost & Stopping Criteria

There are essentially two factors contributing to the effectiveness of an iterative method for : the computational cost per iteration and the number of performed iterations. The computational cost per iteration depends on the structure and sparsity of the original matrix and on the choice of the splitting. For both Jacobi and Gauss-Seidel methods, without further assumptions on , the computational cost per iteration is . Iterations should be stopped when one or more stopping criteria are satisfied, as will be discussed below. For both Jacobi and Gauss-Seidel methods, the cost of performing iterations is ; so as long as , these methods are much cheaper than Gaussian elimination.

In theory, iterative methods require an infinite number of iterations to converge to the exact solution of a linear system but in practice, aiming at the exact solution is neither reasonable nor necessary. Indeed, what is actually needed is an approximation for which the error is guaranteed to be lower than a desired tolerance . On the other hand, since the error is itself unknown (as it depends on the exact solution), a suitable *a posteriori* error estimator is needed which predicts the error starting from quantities that have already been computed. There are two natural estimators one may consider:

* The residual at the iteration, i.e. . More precisely, an iterative method can be stopped at the first iteration step for which
* When the above estimate is satisfied, it is guaranteed that
* i.e. the control on the residual is meaningful only for those matrices whose condition number is reasonably small; in this way the relative error will be of the same size as the relative residual.
* The increment at the iteration, i.e. . More precisely, iterative method would stop after the first iteration step for which
* If is symmetric and positive definite, then

Recalling that should be less than 1 in order for the iterative method to converge, we deduce

i.e. the control on the increment is meaningful only if since in that case the error will be of the same size as the increment.

## 2.4 In-Built MATLAB Procedures

Given that MATLAB is well-suited to dealing with matrices, it has a very powerful method of solving linear systems and it is using the ***Backslash Operator***. This is a powerful in-built method that can solve any square linear system regardless of its form. MATLAB does this by first determining the general form of the matrix (sparse, triangular, Hermitian, etc.) before applying the appropriate optimised method.

For the linear system

MATLAB can solve this using the syntax x=A\b.

|  |
| --- |
| Starting Example |
| Returning to the example in the beginning of this section, the matrix system was  This can be solved as follows:  >> A=[1,1,1;1,-2,0;0,1,-1]; >> b=[20;0;10]; >> A\b ans =  15.0000  7.5000  -2.5000 |

|  |
| --- |
| Note |
| The [MATLAB website](https://uk.mathworks.com/help/matlab/ref/double.mldivide.html) shows the following flowcharts for how A\b classifies the problem before solving it.  If the matrix A is full.  If the matrix is full.  If the matrix A is sparse.  If the matrix is sparse. |

# 3. The Euler Method

Consider the first order ordinary differential equation (ODE)

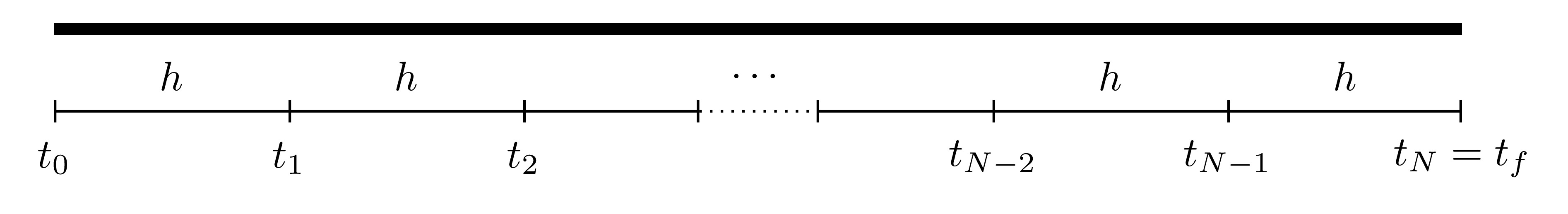
where is a known function, is an initial time and is the final time. An *initial condition* can be prescribed to this differential equation which will assign a “starting value” for the unknown function at the starting time as . The combination of the first order ODE and the initial value gives the ***Initial Value Problem*** (or ***IVP***)

There are many analytic methods for solving first order ordinary differential equations, but they all hold restrictions, like linearity or homogeneity. This chapter will develop the simplest numerical technique for solving any first order ordinary differential equation, this method is called the ***Euler Method***.

Consider the following first order IVP

The function is known and in most cases, is assumed to be “well-behaved” (does not have discontinuities or sharp corners). The term is known as the ***Initial Value*** of the function at the starting time . Solving this initial value problem is essentially finding an unknown curve that starts at the point and ends at time .

The first step in the Euler method (as is the case in most numerical techniques) is to discretise the domain. This changes the domain from the continuous interval to subintervals, each with constant[[3]](#footnote-110) width (sometimes also denoted ), which is known as the ***Stepsize***. The discretised interval will be the set of points



The aim of the numerical procedure is to start from the starting point and progressively find consequent points until the final time is reached.

The Euler method uses the gradient, namely , at the starting point in order to find the value of at the subsequent point which will be labelled . This will, in turn, determine the new gradient at and this process is then continued until the final time is reached. The smaller the value of is, the more points there will be between and resulting in a more accurate final solution to the initial value problem.

The accuracy of the Euler method is usually characterised by how small is or how large is. Since the stepsize may not always give an appropriate subdivision (like dividing the interval into subintervals of width ), then the number of subdivisions can be used to find an appropriate by using

## 3.1 Steps of the Euler Method

Consider the IVP

|  |
| --- |
| Parallel Example |
| The steps of the Euler method will be explained theoretically and applied to this IVP in parallel to demonstrate the steps:  In this case, the function on the RHS is . Note that this IVP has the exact solution |

1. Discretise the interval with stepsize to form the set of points

|  |
| --- |
| Inverval Discretisation |
| Suppose that the interval is to be split into subintervals, then and  Therefore the discretised points are  Note that denotes the number of subintervals and **not** the number of points, that would be points since the starting point is . |

1. At the starting point , the gradient is known since

|  |
| --- |
| Gradient at |
| At the initial point,  So the starting gradient is . |

1. The next step is to find the the value of at the subsequent time . For this purpose, consider the Taylor series expansion of at ,

|  |
| --- |
| Note |
| The term simply means that the terms after this point have a common factor of and these terms are regarded as *higher order terms* and can be neglected since they are far smaller than the first terms provided is small. |

Since is assumed to be sufficiently small, then all terms higher order terms, in this case or higher, can be neglected (i.e.  for ). Therefore

Let denote the *approximated value* of the solution at the point , i.e. , so in this case,

This determines the value of which is an approximation to .

|  |
| --- |
| Calcuating |
| The point can be calculated as follows:  This means that the next point is . |

1. This iteration can be continued to find (which is the approximate value of ) for all

|  |
| --- |
| Calculating |
| The values of and can be calculated as follows: |

1. The solution to the IVP can now be approximated by the function that passes through the points

|  |
| --- |
| Solution to the IVP |
| The approximate solution to the IVP  is the function that passes through the points:  This is a good approximation since the *exact* locations, as per the exact solution are, (to 4 decimal places):  which is not bad for such a coarse interval breakdown. |

The Euler method needs steps to complete and every step requires finding and . Of course, the larger is, the smaller becomes, meaning that more steps will be required but the solution will be closer to the exact solution

Notice that the terms on the right hand side of [Equation 3.1](#eq-Euler1) are all known and for this reason, the Euler method is known as an ***Explicit Method***.

## 3.2 Accuracy

Consider the Taylor series expansion for the function at the point ,

Using ***Taylor’s Theorem***[[4]](#footnote-130), this can be written as

for some point between and . The Euler method determines the approximation to the function at the point , particularly,

The ***Local Truncation Error*** at the first step, denoted , is defined as the absolute difference between the exact and approximated values at the first step, and this is given by

This can be done for all the locations to give a list of local truncation errors . Note that *technically*, these errors are hypothetical since the *exact solution* , and thus , are not known but these are put as placeholders to establish the full accuracy of the method. In this case, the local truncation error is said to be of *second order* since .

As the iteration progresses, the errors will accumulate to result in a ***Global Integration Error*** denoted . In this case, the global integration error is

The global integration error has to be at most the accumulation of all the local truncation errors, namely

A bound for the sum needs to be found in order bound the global integration error. To this end, consider the set of the second derivatives in the sum above, i.e.

Since all these terms take a finite value, then at least one of these terms must be larger than all the rest, this is denoted and can be written as

This can also be expressed differently as

Therefore, since

then

Thus, returning back to the expression for in [Equation 3.2](#eq-GIEEuler)

Hence, the global integration error , this means that the Euler method is a ***First Order Method***. This means that both and the global integration error behave linearly to one another, so if is halved, then the global integration error is halved as well.

In conclusion, the local truncation error of the Euler method is while the global integration error when is small.

|  |
| --- |
| Different Stepsizes |
| Returning to the IVP  The Euler method can be repeated for different values of and these can be seen in the figure below.    The table below shows the global integration error for the different values of :   |  |  | | --- | --- | | 0.4 | 0.05399 | | 0.2 | 0.03681 | | 0.1 | 0.02036 | | 0.05 | 0.01060 |   When the value of is halved, the global integration error is approximately halved as well. |

## 3.3 Set of IVPs

SO far, the Euler Method has been used to solve a single IVP, however this can be extended to solving a set of linear IVPs.

Consider the set of linear IVPs defined on the interval : where, for , the functions are unknown, are known constant coefficients and are all known (these can generally depend on ).

This set of initial value problems need to be written in matrix form as

In this case, is the unknown solution vector, is a matrix of constants, is the vector of initial values and is a vector of known terms (possibly depending on ) and is referred to as the ***Inhomogeneity*** or ***Forcing Term***.

The Euler iteration would be performed in a similar way as before. First, the interval needs to be discretised into equally spaced subintervals, each of width to give the set of discrete times where for . Let be the approximation to the function vector at the time , then

subject to the initial values . (Note that if the vector depends on , then .)

|  |
| --- |
| Sets of IVPs |
| Consider the two coupled IVPs on the interval :  Before attempting to solve this set of IVPs, it needs to be written in matrix form as  In this case,  Let , so  The Euler iteration will be  This can be written as  keeping in mind that the vector and : |

## 3.4 Higher Order IVPs

The previous sections solved one first order IVP and a set of first order IVPs. What happens if a higher order IVP is to be solved? Or a set of higher order IVPs? The difference will be minimal, subject to a few manipulations first.

Consider the order linear IVP on the interval

where and is a known function. This IVP is to be solved subject to the initial conditions

This order IVP can be written as a set of first order IVPs. Indeed, let the functions be given by

Notice that Let be the vector of the unknown functions . This means that the IVP in [Equation 3.3](#eq-HOODE) can be written in matrix form as follows:

The initial condition vector will be

The matrix is called the ***Companion Matrix*** and is a matrix with 1 on the super diagonal and the last row is the minus of the coefficients in the higher order IVP, and zeros otherwise. Now that the order IVP has been converted into a set of linear IVPs, it can be solved just as in [Section 3.3](#sec-EulerSys). Note that *any* linear order IVP can always be converted into a set of first order IVPs but the converse is not always possible.

|  |
| --- |
| Higher Order IVPs |
| Consider the following higher order IVP  Let and . The derivatives of and are: Define the vector  The initial condition vector will be  Now the IVP can be solved using the Euler method as before but only the first function is the most relevant, all others have been used as placeholders. |

### 3.4.1 Sets of Higher Order IVPs

The method above can be extended into a set of higher order IVPs.

|  |
| --- |
| Set of Higher Order IVPs |
| Consider the following coupled system of higher order IVPs  In the case of a coupled system, the vector function should consist of all the unknown functions and their derivatives up to but not including their highest order derivative. In other words,  The vector of initial values would be  Now this can be solved just as before with the most relevant terms being the first and third (since those are and ). |

### 3.4.2 Stability of a Set of ODEs

Consider the set of homogeneous ODEs

Let be the eigenvalues of the matrix and be their *distinct* corresponding eigenvectors (distinct for the sake argument). Analytically, the set of differential equations has the general solution

where are constants that can be determined from the initial values.

**Definition 3.1** The initial value problem

is said to be ***Asymptotically Stable*** if as , in other words, all functions in tend to 0 as tends to infinity.

This definition will be important when looking at the long term behaviour of solutions from the eigenvalues to then determine stepsize bounds.

**Theorem 3.1** The initial value problem

is asymptotically stable if **all** the eigenvalues of the matrix have negative real parts. If has at least one eigenvalue with a non-negative real part, then the system is not asymptomatically stable.

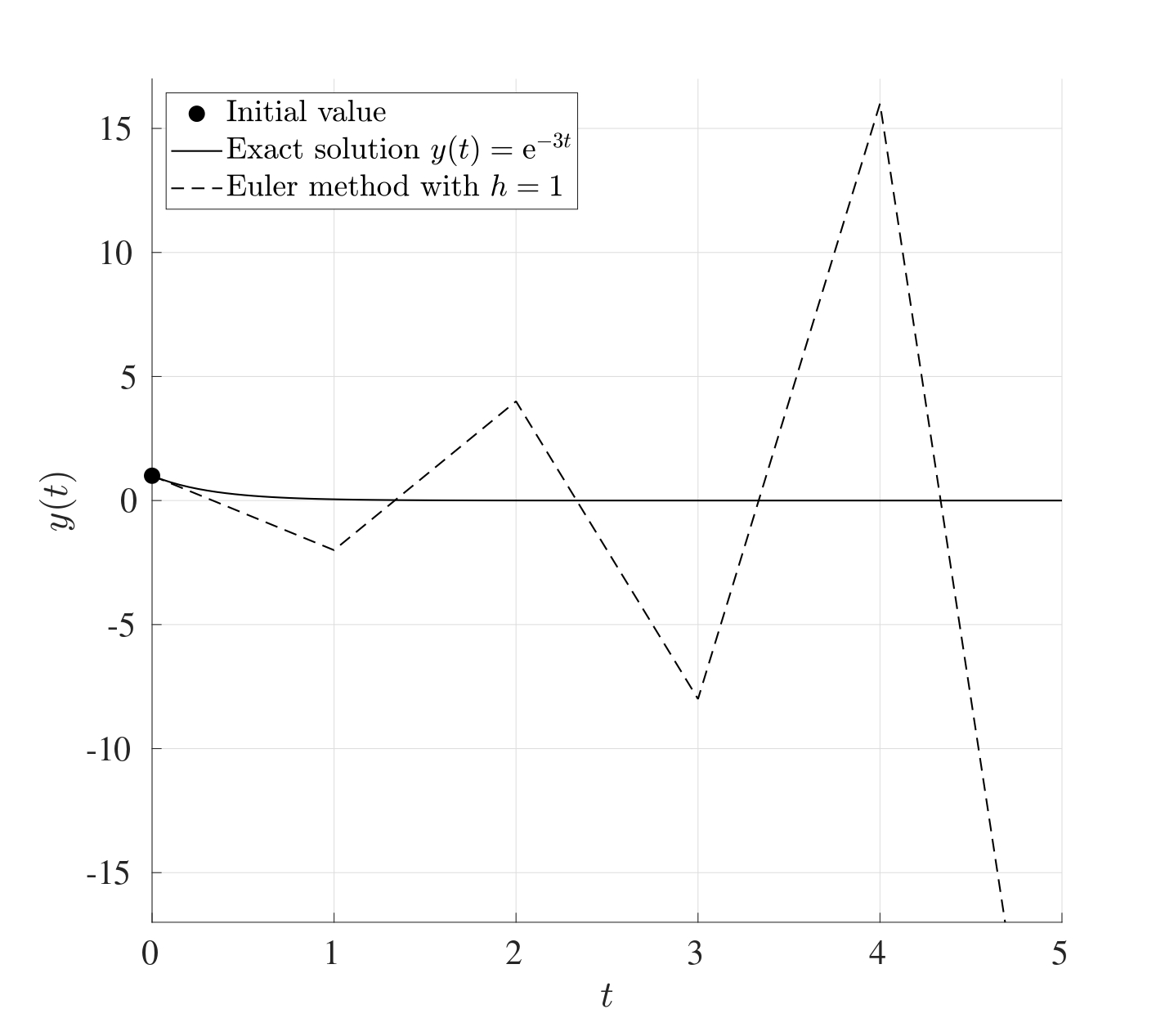
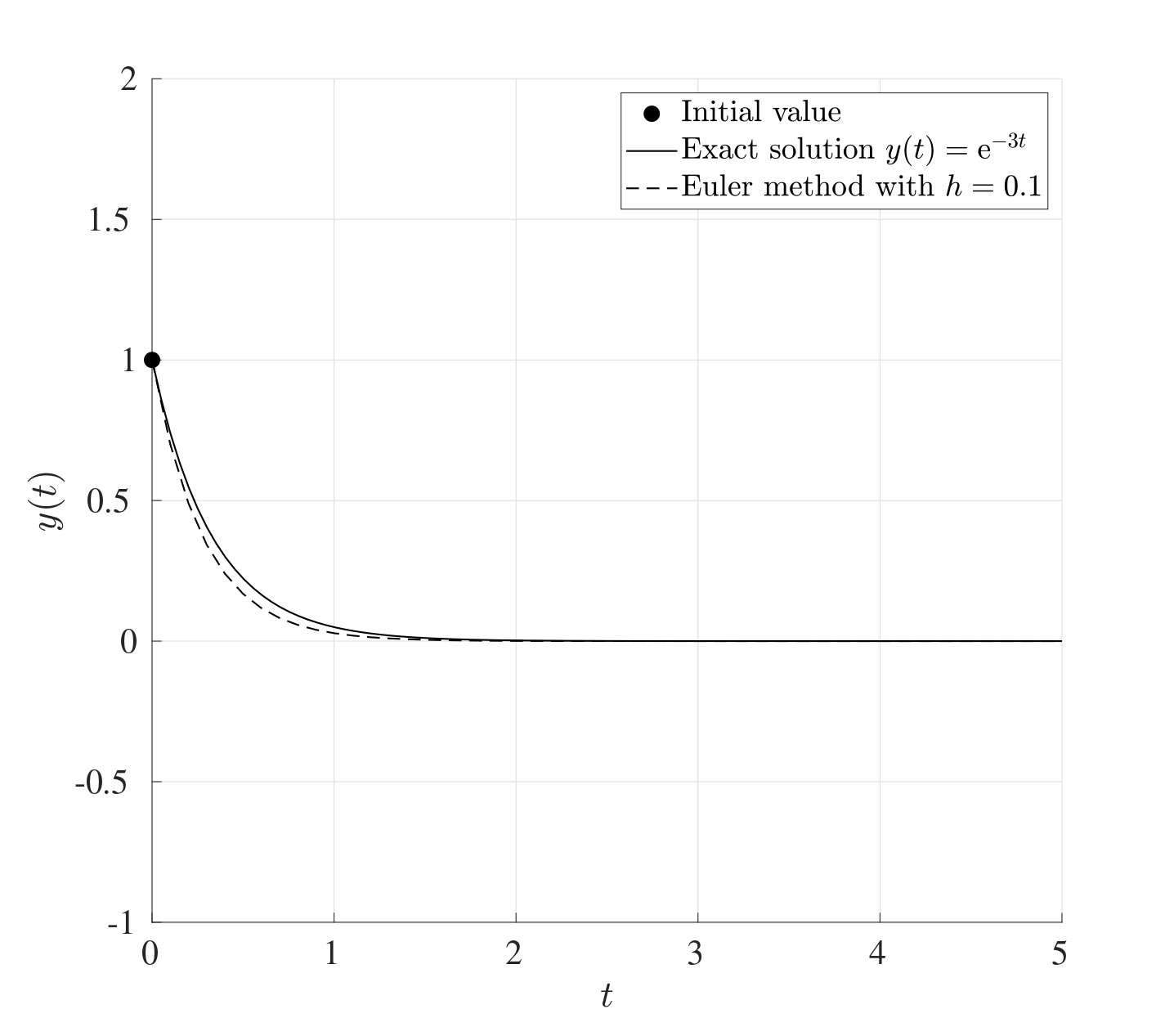
Notice that the stability of a set of ODEs does *not* depend on the forcing term nor does it depend on the initial condition .

## 3.5 Limitations of the Euler Method

In some cases, if the stepsize is taken to be too large, then the Euler method can give misleading results.

For example, consider the initial value problem:

Choosing a large stepsize can render the method ineffective. Case in point, when , the approximate solution oscillates and grows quite rapidly, however choosing a smaller value of , say , gives a very good approximation to the exact solution. These are illustrated in the figures below.

Another situation when the Euler method fails is when the IVP does not have a unique solution. For example, consider the IVP:

This has the exact solution however this is not unique since is also a perfectly valid solution. The Euler method in this case will not be able to capture the first non-trivial solution but will only capture the second trivial solution giving a straight line at 0[[5]](#footnote-160).

### 3.5.1 Bounds on the Stepsize

Consider the initial value problem

If is asymptotically stable, then a maximum bound for the stepsize can be found to ensure that the iterations converge. (This means that asymptotic stability of is a necessary and sufficient condition for the existence of an upper bound such that if , then the Euler iteration converges.)

If the stepsize is too large, then the method may not converge but on the other hand if it is too low, then the iteration will take a considerable amount of time to perform. Therefore an “optimal” stepsize is needed to obtain sufficiently accurate solutions.

|  |
| --- |
| Different Stepsizes |
| Consider the following initial value problem  The figure below shows the Euler method being used to solve the initial value problem in the interval for the stepsizes .    When , the Euler method does not converge. At , the Euler method converges but there clearly is a distinct artefact in the solution that shows a slight oscillation. For less than , this oscillation is no longer observed and the Euler method is convergent. |

### 3.5.2 Exact Bound

Consider the IVP

Let be the eigenvalues of . Suppose that the matrix is asymptotically stable (i.e.  for all ). In order for the Euler iterations to converge, the stepsize needs be less than the threshold stepsize where

In other words, if the initial value problem is asymptotically stable, then the Euler method is stable if an only if . This means that the convergence of the Euler is characterised by the eigenvalue that is furthest away from the origin, also called the ***Dominant Eigenvalue***.

!!!!!DO!!!!!

|  |
| --- |
| Euler Upper Bound |
| Consider the system of differential equations with where  The eigenvalues of the matrix are . Since all the eigenvalues are negative, this system is asymptotically stable. Since all the eigenvalues are real, then the threshold stepsize for a convergent Euler method is  Solutions for different stepsizes are as shown below with the initial values (blue), (red) and (magenta). It can be seen that if , then at least one solution will diverge but if , then all solutions converge to 0. |

### 3.5.3 Estimated Bound

One drawback in attempting to determine the value of using [Equation 3.4](#eq-h0Bound) is that *all* the eigenvalues of the matrix have to be determined before can be found. This can be computationally expensive for especially for very large matrices.

An estimate for the threshold stepsize can be found with far fewer computations using the *sup-norm* (also known as the *infinity norm* or the *Chebyshev norm*). Recall that for a vector , the **sup-norm** of is the maximum absolute value in the vector, i.e.

Whereas for a matrix , the **sup-norm** of is the maximal absolute row sum. In other words, for a given matrix , take the absolute value of all the terms, take the sum of each row and the sup-norm will be the largest out of these.

|  |
| --- |
| Sup-Norm of Vectors & Matrices |
| Consider the vector and matrix given by  The sup-norm of is simply the largest absolute element which is , therefore .  As for , to find the sup-norm, first take the absolute value of all the terms, then add the rows. The sup-norm is the maximum element that results:  Therefore .  Both of these can be found in MATLAB using norm(x,Inf) and norm(M,Inf). |

**Theorem 3.2** Consider the set of linear IVPs

where is asymptotically stable. Then the Euler method is numerically convergent for any choice of which satisfies

Computing all the eigenvalues of the matrix can be computationally expensive but obtaining the sup-norm is takes far fewer computations, however as a drawback, the resulting value of would be an estimate.

|  |
| --- |
| Stepsize Bound Estimate 1 (Tridiagonal) |
| Consider the differential equation where  To find the upper bound for the stepsize for which the Euler method converges, first evaluate :  To find the sup-norm, take the absolute value of all the terms and find the maximal row sum:  Let and . Since , then , therefore  In order to satisfy the inequality , consider the cases when and separately:   1. If , then :  * Therefore is indeed true.  1. If , then :  * If , then . Simplifying this would result in which contradicts with the assumption that .   From these two cases, it is clear that (since that case leads to a contradiction), therefore . Thus for a convergent Euler method, the stepsize must be less than the threshold stepsize .  This can be compared to the exact bound; the eigenvalues of the matrix are  Therefore  which is a larger bound compared to the one obtained using the sup-norm method. Observe that if the size of the matrix was larger but followed the same theme (i.e.  on the main diagonal and and the sub and super diagonals), then no further calculations are required for the sup-norm method, the outcome will still be . As for the eigenvalue method, all the eigenvalues have to be recalculated again. |

|  |
| --- |
| Stepsize Bound Estimate 2 (Bidiagonal) |
| Consider the differential equation where  To find the upper bound for the stepsize for which the Euler method converges, first evaluate :  To find the sup-norm, take the absolute value of all the terms and find the maximal row sum:  Let and . Clearly since , therefore  In order to satisfy the inequality, , consider the cases when and :   1. If , then :  * therefore is indeed true.  1. If , then :  * If , then , meaning that which contradicts with the assumption that .   This means that for a convergent Euler method, the stepsize must be less than .  This can be compared to the exact upper bound. The eigenvalues of the matrix are just five times, therefore  this shows that the sup-norm method gives a tighter than using eigenvalues. |

The sup-norm method works well when the matrix in question has a diagonal, bidiagonal or tridiagonal structure where the diagonal terms are the same. In general, the sup-norm method might *not* be suitable for any matrix.

|  |
| --- |
| Stepsize Bound Estimate 3 (General) |
| Consider the differential equation where  Find the sup-norm:  Let and . Here, it is not obvious which is larger, or . Therefore, consider the three cases , and .   1. : In this case, and , therefore and , hence . In order to satisfy , this would mean that which contradicts with the fact that . Therefore . 2. : In this case, and , therefore and . This should now be split into two subcases to check which one will lead to a contradiction:    1. Suppose that , then    * which contradicts with    1. Suppose that , then    * not leading to any contradiction. therefore since , then .   In order to satisfy then which contradicts with the fact that . Therefore .   1. : In this case, and , therefore and . Clearly since , so . In order to satisfy then which contradicts with the fact that . This means that .   So in every possible case, there will be a contradiction when using the sup-norm method. This *does not* mean that the system is asymptotically unstable, in fact, the eigenvalues of the matrix are meaning that the system is asymptotically stable and the threshold stepsize is in fact .  This example shows that the sup-norm method cannot be used for any matrix system, but if a matrix has a banded structure, then it would be appropriate and would require fewer computations compared to finding all the eigenvalues. |

## 3.6 MATLAB Code

The following MATLAB code performs the Euler iteration for the following set of IVPs on the interval :

|  |
| --- |
| Linearity |
| Note that this code is built for a general case that *does not* have to be linear even though the entire derivation process was built on the fact that the system is linear. |

function IVP\_Euler  
  
%% Solve a set of first order IVPs using Euler  
  
% This code solves a set of IVP when written explicitly  
% on the interval [t0,tf] subject to the initial conditions  
% y(0)=y0. The output will be the graph of the solution(s)  
% and the vector value at the final point tf. Note that the  
% IVPs do not need to be linear or homogeneous.  
  
%% Lines to change:  
  
% Line 28 : t0 - Start time  
% Line 31 : tf - End time  
% Line 34 : N - Number of subdivisions  
% Line 37 : y0 - Vector of initial values  
% Line 105+ : Which functions to plot, remembering to assign  
% a colour, texture and legend label  
% Line 125+ : Set of differential equations written  
% explicitly. These can also be non-linear and  
% include forcing terms. These equations can  
% also be written in matrix form if the  
% equations are linear.  
  
%% Set up input values  
  
% Start time  
t0=0;  
  
% End time  
tf=1;  
  
% Number of subdivisions  
N=50;  
  
% Column vector initial values y0=y(t0)  
y0=[0;1;0];  
  
%% Set up IVP solver parameters  
  
% T = Vector of times t0,t1,...,tN.  
% This is generated using linspace which splits the  
% interval [t0,tf] into N+1 points (or N subintervals)  
T=linspace(t0,tf,N+1);  
  
% Stepsize  
h=(tf-t0)/N;  
  
% Number of differential equations  
K=length(y0);  
  
%% Perform the Euler iteration  
  
% Y = Solution matrix  
% The matrix Y will contain K rows and N+1 columns. Every  
% row corresponds to a different IVP and every column  
% corresponds to a different time. So the matrix Y will  
% take the following form:  
% y\_1(t\_0) y\_1(t\_1) y\_1(t\_2) ... y\_1(t\_N)  
% y\_2(t\_0) y\_2(t\_1) y\_2(t\_2) ... y\_2(t\_N)  
% ...  
% y\_K(t\_0) y\_K(t\_1) y\_K(t\_2) ... y\_K(t\_N)  
Y=zeros(K,N+1);  
  
% The first column of the vector Y is the initial vector y0  
Y(:,1)=y0;  
  
% Set the current time t to be the starting time t0 and the  
% current value of the vector y to be the strtaing values y0  
t=t0;  
y=y0;  
  
for n=2:1:N+1  
  
 dydt=DYDT(t,y,K); % Find gradient at the current step  
  
 y=y+h\*dydt; % Find y at the current step  
  
 t=T(n); % Update the new time  
  
 Y(:,n)=y; % Replace row n in Y with y  
  
end  
  
%% Setting plot parameters  
  
% Clear figure  
clf  
  
% Hold so more than one line can be drawn  
hold on  
  
% Turn on grid  
grid on  
  
% Setting font size and style  
set(gca,'FontSize',20,'FontName','Times')  
  
% Label the axes  
xlabel('$t$','Interpreter','Latex')  
ylabel('$\mathbf{y}(t)$','Interpreter','Latex')  
  
% Plot the desried solutions. If all the solutions are  
% needed, then consider using a for loop in that case  
plot(T,Y(1,:),'-b','LineWidth',2)  
plot(T,Y(2,:),'-r','LineWidth',2)  
plot(T,Y(3,:),'-k','LineWidth',2)  
  
% Legend labels  
legend('$y\_1(t)$','$y\_2(t)$','$y\_3(t)$')  
set(legend,'Interpreter','Latex')  
  
% Display the values of the vector y at tf  
disp(strcat('The vector y at t=',num2str(tf),' is:'))  
disp(Y(:,end))  
  
end  
  
function [dydt]=DYDT(t,y,K)  
  
% When the equation are written in explicit form  
  
dydt=zeros(K,1);  
  
dydt(1)=2\*y(1)+y(2)+y(3)+cos(t);  
  
dydt(2)=sin(y(1))+exp(-y(2)+y(3));  
  
dydt(3)=y(1)\*y(2)-y(3);  
  
% If the set of equations is linear, then these can be  
% written in matrix form as dydt=A\*y+b(t). For example, if  
% the set of equations is:  
% dudt = 7u - 2v + w + exp(t)  
% dvdt = 2u + 3v - 9w + cos(t)  
% dwdt = 2v + 5w + 2  
% Then:  
% A=[7,-2,1;2,3,-9;0,2,5];  
% b=@(t) [exp(t);cos(t);2];  
% dydt=A\*y+b(t)  
  
end

# 4. The Modified Euler Method

The Euler method can be effective when it comes to solving differential equations numerically but on occasions, the global error of is rather poor. The Euler method can modified and improved to give **Modified** or **Improved Euler Method** (also known as the *Heun Method*, named after Karl Heun).

## 4.1 Steps of the Modified Euler Method

The Modified Euler Method utilises the **Fundamental Theorem of Calculus** which states that for a differentiable function defined on the interval (where for some stepsize ),

In the interval , the derivative may be approximated by the derivative at the leftmost point , this approximation forms the basis of the standard Euler method;

However, if varies substantially then this approximation can lead to some poor predictions. This can be modified so rather than approximating by only, it can be approximated by taking an average between and , namely

Thus

Initially, one might suspect that the derivative can be found from the differential equation itself, namely, but to do that, a **Prediction-Correction** procedure needs to be employed where the Euler method can be used to predict a value of and this is then corrected afterwards. This is done as follows:

|  |
| --- |
| Modified Euler Method |
| Consider the differential equation  This differential equation is non-linear but has a known particular solution which is  and this will be compared to the approximate solutions obtained from the standard and Modified Euler methods.  The figure below shows how the standard and modified Euler methods compare to the exact solution for the same stepsize . This suggests that the Modified Euler method has improved accuracy compared to the Euler method for the same stepsize, however as a consequence, the function on the right hand side of the differential equation has to be calculated twice for every step; once in the prediction stage and once for the correction. However even with this in mind, doubling the number of calculations to improve accuracy can also warrant for a coarser choice of the stepsize to allow for a more efficient use of computational time. |

## 4.2 Accuracy of the Modified Euler Method

In order to asses the accuracy of the Modified Euler method, consider the Taylor series expansion of at the points and about :

Subtracting from gives

The Taylor series expansion can also be done for the derivative of at the points and about in a similar way as above, i.e.

Adding to gives

thus multiplying by and using equation [Equation 4.1](#eq-yhalf) yields

The first step of the Modified Euler method is to predict the value of using the Euler iteration;

Hence

All this information can now be used to obtain the improved update which is the corrected form of . Thus from equation [Equation 4.2](#eq-fullyy),

Equations [Equation 4.3](#eq-Y1) and [Equation 4.2](#eq-fullyy) can be used to find the local truncation error for the Modified Euler method at the first time step which is

Therefore the *local truncation error* meaning that the Modified Euler method is third order accurate which is an improvement over the Euler method.

The global integration error can be obtained just as before to show that the global integration error of the Modified Euler method is meaning that this is a second order method. In particular, if the stepsize is halved, the global integration error will be reduced by a factor of four while the local truncation error will reduce by a factor of eight.

## 4.3 MATLAB Code

The following MATLAB code performs the Modified Euler iteration for the following set of IVPs on the interval :

|  |
| --- |
| Linearity |
| Note that this code is built for a general case that *does not* have to be linear even though the entire derivation process was built on the fact that the system is linear. |

function IVP\_Mod\_Euler  
  
%% Solve a set of first order IVPs using Modified Euler  
  
% This code solves a set of IVP when written explicitly  
% on the interval [t0,tf] subject to the initial conditions  
% y(0)=y0. The output will be the graph of the solution(s)  
% and the vector value at the final point tf. Note that the  
% IVPs do not need to be linear or homogeneous.  
  
%% Lines to change:  
  
% Line 28 : t0 - Start time  
% Line 31 : tf - End time  
% Line 34 : N - Number of subdivisions  
% Line 37 : y0 - Vector of initial values  
% Line 115+ : Which functions to plot, remembering to assign  
% a colour, texture and legend label  
% Line 135+ : Set of differential equations written  
% explicitly. These can also be non-linear and  
% include forcing terms. These equations can  
% also be written in matrix form if the  
% equations are linear.  
  
%% Set up input values  
  
% Start time  
t0=0;  
  
% End time  
tf=1;  
  
% Number of subdivisions  
N=50;  
  
% Column vector initial values y0=y(t0)  
y0=[0;1;0];  
  
%% Set up IVP solver parameters  
  
% T = Vector of times t0,t1,...,tN.  
% This is generated using linspace which splits the  
% interval [t0,tf] into N+1 points (or N subintervals)  
T=linspace(t0,tf,N+1);  
  
% Stepsize  
h=(tf-t0)/N;  
  
% Number of differential equations  
K=length(y0);  
  
%% Perform the Modified Euler iteration  
  
% Y = Solution matrix  
% The matrix Y will contain K rows and N+1 columns. Every  
% row corresponds to a different IVP and every column  
% corresponds to a different time. So the matrix Y will  
% take the following form:  
% y\_1(t\_0) y\_1(t\_1) y\_1(t\_2) ... y\_1(t\_N)  
% y\_2(t\_0) y\_2(t\_1) y\_2(t\_2) ... y\_2(t\_N)  
% ...  
% y\_K(t\_0) y\_K(t\_1) y\_K(t\_2) ... y\_K(t\_N)  
Y=zeros(K,N+1);  
  
% The first column of the vector Y is the initial vector y0  
Y(:,1)=y0;  
  
% Set the current time t to be the starting time t0 and the  
% current value of the vector y to be the strtaing values y0  
t=t0;  
y=y0;  
  
for n=2:1:N+1  
  
 % Prediction Step:  
 % Use the Euler iteration to obtain an appromxation for  
 % the derivatives at the current time step  
  
 dydt=DYDT(t,y,K); % Find gradient at the current step  
 y\_pred=y+h\*dydt; % Predict y at current time step  
  
 % Corrector Step:  
 % Use the Modified Euler to correct y\_pred  
  
 dydt\_pred=DYDT(t,y\_pred,K); % Predict the gradient  
 % from the predicted y  
 y=y+0.5\*h\*(dydt+dydt\_pred); % Find y at the current step  
  
 t=T(n); % Update the new time  
  
 Y(:,n)=y; % Replace row n in Y with y  
  
end  
  
%% Setting plot parameters  
  
% Clear figure  
clf  
  
% Hold so more than one line can be drawn  
hold on  
  
% Turn on grid  
grid on  
  
% Setting font size and style  
set(gca,'FontSize',20,'FontName','Times')  
  
% Label the axes  
xlabel('$t$','Interpreter','Latex')  
ylabel('$\mathbf{y}(t)$','Interpreter','Latex')  
  
% Plot the desried solutions. If all the solutions are  
% needed, then consider using a for loop in that case  
plot(T,Y(1,:),'-b','LineWidth',2)  
plot(T,Y(2,:),'-r','LineWidth',2)  
plot(T,Y(3,:),'-k','LineWidth',2)  
  
% Legend labels  
legend('$y\_1(t)$','$y\_2(t)$','$y\_3(t)$')  
set(legend,'Interpreter','Latex')  
  
% Display the values of the vector y at tf  
disp(strcat('The vector y at t=',num2str(tf),' is:'))  
disp(Y(:,end))  
  
end  
  
function [dydt]=DYDT(t,y,K)  
  
% When the equation are written in explicit form  
  
dydt=zeros(K,1);  
  
dydt(1)=2\*y(1)+y(2)+y(3)+cos(t);  
  
dydt(2)=sin(y(1))+exp(-y(2)+y(3));  
  
dydt(3)=y(1)\*y(2)-y(3);  
  
% If the set of equations is linear, then these can be  
% written in matrix form as dydt=A\*y+b(t). For example, if  
% the set of equations is:  
% dudt = 7u - 2v + w + exp(t)  
% dvdt = 2u + 3v - 9w + cos(t)  
% dwdt = 2v + 5w + 2  
% Then:  
% A=[7,-2,1;2,3,-9;0,2,5];  
% b=@(t) [exp(t);cos(t);2];  
% dydt=A\*y+b(t)  
  
end

# 5. The Runge-Kutta Method

The Modified Euler method extended the Euler method to a two-stage procedure with a global integration error of . This can be extended further to a ***Multi-Stage Method***, also called a ***Runge-Kutta Method*** with stages and a global error integration error of for any arbitrarily large (in this case, the Modified Euler method is known as a second order Runge-Kutta method since it has two stages). For instance, the *fourth order Runge-Kutta method* requires four calculations for every step and has a global integration error of , this is formulated as follows:

Runge-Kutta methods like this are quite versatile and are generally the most used methods for their accuracy since the stepsize does not need to be too small to achieve good results. Even though every step requires four calculations, the value of can be made larger in order to reduce the cost but retain considerable accuracy.

|  |
| --- |
| Runge-Kutta Method |
| Consider the differential equation  The exact solution to this differential equation is known to be    The figure above shows the exact solution to the differential equation (solid line) with the three different methods used to approximate the solution all at the same resolution of . The stepsize is quite coarse but this is merely for the purposes of demonstration. The Euler method is the least accurate for this coarse grid, the Heun method improves the accuracy while the fourth order Runge-Kutta method is the most accurate out of the three even for the same stepsize. |

## 5.1 MATLAB Code

The following MATLAB code performs the fourth order Runge-Kutta iteration for the following set of IVPs on the interval :

|  |
| --- |
| Linearity |
| Note that this code is built for a general case that *does not* have to be linear even though the entire derivation process was built on the fact that the system is linear. |

function IVP\_RK4  
  
%% Solve a set of first order IVPs using RK4  
  
% This code solves a set of IVP when written explicitly  
% on the interval [t0,tf] subject to the initial conditions  
% y(0)=y0. The output will be the graph of the solution(s)  
% and the vector value at the final point tf. Note that the  
% IVPs do not need to be linear or homogeneous.  
  
%% Lines to change:  
  
% Line 28 : t0 - Start time  
% Line 31 : tf - End time  
% Line 34 : N - Number of subdivisions  
% Line 37 : y0 - Vector of initial values  
% Line 109+ : Which functions to plot, remembering to assign  
% a colour, texture and legend label  
% Line 129+ : Set of differential equations written  
% explicitly. These can also be non-linear and  
% include forcing terms. These equations can  
% also be written in matrix form if the  
% equations are linear.  
  
%% Set up input values  
  
% Start time  
t0=0;  
  
% End time  
tf=1;  
  
% Number of subdivisions  
N=50;  
  
% Column vector initial values y0=y(t0)  
y0=[0;1;0];  
  
%% Set up IVP solver parameters  
  
% T = Vector of times t0,t1,...,tN.  
% This is generated using linspace which splits the  
% interval [t0,tf] into N+1 points (or N subintervals)  
T=linspace(t0,tf,N+1);  
  
% Stepsize  
h=(tf-t0)/N;  
  
% Number of differential equations  
K=length(y0);  
  
%% Perform the RK4 iteration  
  
% Y = Solution matrix  
% The matrix Y will contain K rows and N+1 columns. Every  
% row corresponds to a different IVP and every column  
% corresponds to a different time. So the matrix Y will  
% take the following form:  
% y\_1(t\_0) y\_1(t\_1) y\_1(t\_2) ... y\_1(t\_N)  
% y\_2(t\_0) y\_2(t\_1) y\_2(t\_2) ... y\_2(t\_N)  
% ...  
% y\_K(t\_0) y\_K(t\_1) y\_K(t\_2) ... y\_K(t\_N)  
Y=zeros(K,N+1);  
  
% The first column of the vector Y is the initial vector y0  
Y(:,1)=y0;  
  
% Set the current time t to be the starting time t0 and the  
% current value of the vector y to be the strtaing values y0  
t=t0;  
y=y0;  
  
for n=2:1:N+1  
  
 % Determine the coefficients of RK4  
  
 K1=DYDT(t,y,K);  
 K2=DYDT(t+h/2,y+h\*K1/2,K);  
 K3=DYDT(t+h/2,y+h\*K2/2,K);  
 K4=DYDT(t+h,y+h\*K3,K);  
 y=y+(h/6)\*(K1+2\*K2+2\*K3+K4);  
  
 t=T(n); % Update the new time  
  
 Y(:,n)=y; % Replace row n in Y with y  
  
end  
  
%% Setting plot parameters  
  
% Clear figure  
clf  
  
% Hold so more than one line can be drawn  
hold on  
  
% Turn on grid  
grid on  
  
% Setting font size and style  
set(gca,'FontSize',20,'FontName','Times')  
  
% Label the axes  
xlabel('$t$','Interpreter','Latex')  
ylabel('$\mathbf{y}(t)$','Interpreter','Latex')  
  
% Plot the desried solutions. If all the solutions are  
% needed, then consider using a for loop in that case  
plot(T,Y(1,:),'-b','LineWidth',2)  
plot(T,Y(2,:),'-r','LineWidth',2)  
plot(T,Y(3,:),'-k','LineWidth',2)  
  
% Legend labels  
legend('$y\_1(t)$','$y\_2(t)$','$y\_3(t)$')  
set(legend,'Interpreter','Latex')  
  
% Display the values of the vector y at tf  
disp(strcat('The vector y at t=',num2str(tf),' is:'))  
disp(Y(:,end))  
  
end  
  
function [dydt]=DYDT(t,y,K)  
  
% When the equation are written in explicit form  
  
dydt=zeros(K,1);  
  
dydt(1)=2\*y(1)+y(2)+y(3)+cos(t);  
  
dydt(2)=sin(y(1))+exp(-y(2)+y(3));  
  
dydt(3)=y(1)\*y(2)-y(3);  
  
% If the set of equations is linear, then these can be  
% written in matrix form as dydt=A\*y+b(t). For example, if  
% the set of equations is:  
% dudt = 7u - 2v + w + exp(t)  
% dvdt = 2u + 3v - 9w + cos(t)  
% dwdt = 2v + 5w + 2  
% Then:  
% A=[7,-2,1;2,3,-9;0,2,5];  
% b=@(t) [exp(t);cos(t);2];  
% dydt=A\*y+b(t)  
  
end

# 6. MATLAB’s In-Built Procedures

So far, the three main iterative methods have been developed that solve IVPs numerically. MATLAB, however, has its own built-in procedures that can solve IVPs with a combination of several methods. The two main ones are ode23 (which uses a combination of a second and third order RK methods) and ode45 (which uses a combination of a fourth and fifth order RK methods).

Both ode45 and ode23 are hybrid methods and use adaptive meshing, this means that the time span grid is not necessarily uniform, but it changes depending on the gradients; if the gradient is large at some point, then the stepsize will be small to capture these drastic changes.

The following MATLAB code solves the following set of IVPs on the interval using ode45:

function IVP\_InBuilt  
  
%% Solve a set of first order IVPs using In-Built codes  
  
% This code solves a set of IVP when written explicitly  
% on the interval [t0,tf] subject to the initial conditions  
% y(0)=y0. The output will be the graph of the solution(s)  
% and the vector value at the final point tf. Note that the  
% IVPs do not need to be linear or homogeneous.  
  
%% Lines to change:  
  
% Line 28 : t0 - Start time  
% Line 31 : tf - End time  
% Line 43 : T\_Span - Time span for evaluation  
% Line 46 : y0 - Vector of initial values  
% Line 85+ : Which functions to plot, remembering to assign  
% a colour, texture and legend label  
% Line 105+ : Set of differential equations written  
% explicitly. These can also be non-linear and  
% include forcing terms. These equations can  
% also be written in matrix form if the  
% equations are linear.  
  
%% Set up input values  
  
% Start time  
t0=0;  
  
% End time  
tf=1;  
  
% Time span  
% In-built methods tend to use adaptive meshing; decreasing  
% the stepsize near locations with drastic derivative  
% changes and increasing near small derivative changes.  
% Sometimes this is not desired but a uniform meshing is  
% requiredfrom the start time t0 to the end time tf being  
% split into N equal sub intervals. This can be changed  
% here:  
% Adaptive meshing: T\_Span=[t0 tf]  
% Specific meshing: T\_Span=linspace(t0,tf,N)  
T\_Span=[t0 tf];  
  
% Column vector initial values y0=y(t0)  
y0=[0;1;0];  
  
%% Set up IVP solver parameters  
  
% Number of differential equations  
K=length(y0);  
  
%% Use solver  
  
% Set the solver tolerance  
tol=odeset('RelTol',1e-6);  
  
% Solve the IVP using ode45 or ode23  
[T,Y]=ode45(@(t,y) DYDT(t,y,K),T\_Span,y0,tol);  
  
% Convert T and Y to columns for consistency  
T=T';  
Y=Y';  
  
%% Setting plot parameters  
  
% Clear figure  
clf  
  
% Hold so more than one line can be drawn  
hold on  
  
% Turn on grid  
grid on  
  
% Setting font size and style  
set(gca,'FontSize',20,'FontName','Times')  
  
% Label the axes  
xlabel('$t$','Interpreter','Latex')  
ylabel('$\mathbf{y}(t)$','Interpreter','Latex')  
  
% Plot the desried solutions. If all the solutions are  
% needed, then consider using a for loop in that case  
plot(T,Y(1,:),'-b','LineWidth',2)  
plot(T,Y(2,:),'-r','LineWidth',2)  
plot(T,Y(3,:),'-k','LineWidth',2)  
  
% Legend labels  
legend('$y\_1(t)$')  
set(legend,'Interpreter','Latex')  
  
% Display the values of the vector y at tf  
disp(strcat('The vector y at t=',num2str(tf),' is:'))  
disp(Y(:,end))  
  
end  
  
function [dydt]=DYDT(t,y,K)  
  
% When the equation are written in explicit form  
  
dydt=zeros(K,1);  
  
dydt(1)=2\*y(1)+y(2)+y(3)+cos(t);  
  
dydt(2)=sin(y(1))+exp(-y(2)+y(3));  
  
dydt(3)=y(1)\*y(2)-y(3);  
  
% If the set of equations is linear, then these can be  
% written in matrix form as dydt=A\*y+b(t). For example, if  
% the set of equations is:  
% dudt = 7u - 2v + w + exp(t)  
% dvdt = 2u + 3v - 9w + cos(t)  
% dwdt = 2v + 5w + 2  
% Then:  
% A=[7,-2,1;2,3,-9;0,2,5];  
% b=@(t) [exp(t);cos(t);2];  
% dydt=A\*y+b(t)  
  
end

# 7. Implicit IVP Solvers

In some cases, IVPs can be difficult to solve because of the non-linearily of its terms, this is where ***Implicit Methods*** should be used to accomodate for these issues.

## 7.1 Backwards Euler Method

Consider the Euler method at the starting time . The value of the function at is approximated by

and this gives an upper bound for a stable stepsize of

in order to ensure that the Euler method is computationally stable. However, suppose that this modified slightly by using the gradient at rather than at , in other words, suppose that the value of at is approximated by

This approach is known as the ***Backwards Euler Method*** and is an implicit procedure since the value of is not known to begin with.

The general formulation is as follows: Consider the system of differential equations

Discretise the interval into equal subintervals, each with width . At the time step , the backwards Euler method is

This can be rearranged to give

Rearranging further fives the basis for the Backwards Euler iteration which is

whereas the standard Euler method in matrix form is

The Euler method requires *explicit calculations* using matrix multiplications but the backwards Euler method requires matrix inversion instead.

## 7.2 Stability of the Backwards Euler Method

Consider the initial value problem in its scalar form

The backwards Euler method at the time gives

This initial condition can be perturbed by adding a small parameter to give the perturbed differential equation

The backwards Euler then yields

The differential equations in and can be subtracted to give a perturbation term where

Notice that once again, the forcing function has been eliminated and therefore does not affect the stability of the backwards Euler method. The differential equation for will have the initial condition . This expression can be used to represent in terms of recursively as:

This means that the method is stable for stepsizes that satisfy and since for an asymptotically stable system, then this inequality is *always* satisfied. This means that the backwards Euler method is stable *for all* stepsizes , no matter how large.

|  |
| --- |
| Backwards Euler Method |
| Consider the differential equation  In this case, meaning that this differential equation is asymptotic stable. The maximum allowable stepsize for the Euler method is . However, the backwards Euler method is stable for any stepsize as seen below (very large stepsizes will still converge but they will not give any useful information). |

The formulation presented above also holds for sets of differential equations in the same way with one difference. Instead of having , the procedure for systems will require the matrix inverse or the MATLAB backslash operator can be used instead.

## 7.3 Order of Accuracy

The backwards Euler method is numerically stable for all values the stepsize and has the same order of accuracy as the Euler method, i.e. the local truncation error is of while the global integration error is of . However, this increased stability comes at a cost, the backwards Euler methods requires double the computational cost compared to the Euler method.

## 7.4 Stiff Differential Equations

Stiff sets of differential equations with a large value of the total computational cost can be very difficult to solve numerically using explicit methods but implicit methods can work very well. MATLAB has its very own built-in stiff differential equation solver under the command ode15s and this can be implemented exactly as ode45. This solves sets of differential equations implicitly using numerical differentiation of orders 1 to 5.

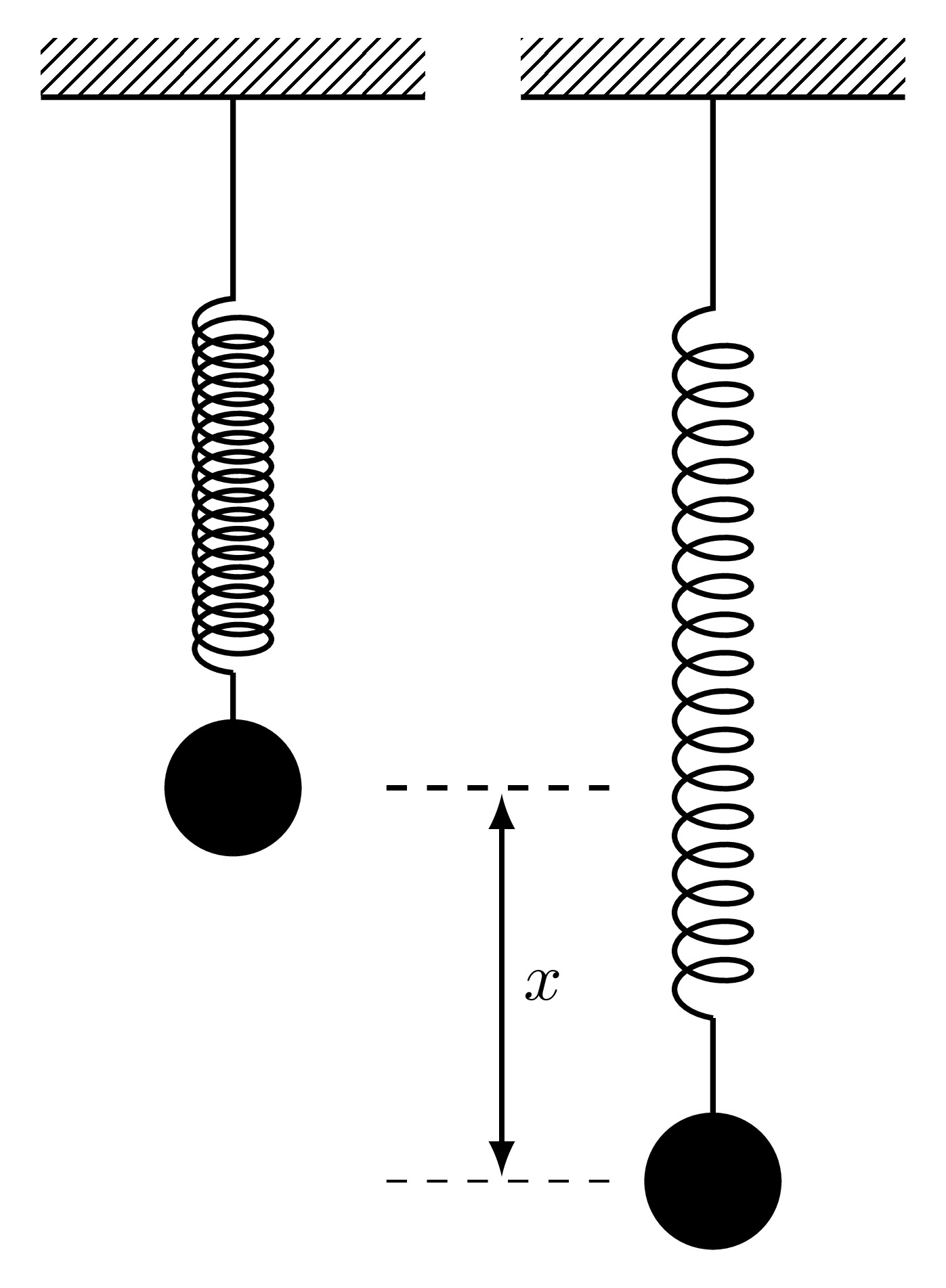
|  |
| --- |
| Stiff IVPs |
| Consider the set of differential equations on the interval  This is a very stiff set of differential equations, solving this using ode45 takes upwards of 92 seconds while solving using the stiff solver ode15s requires a mere 0.233 seconds (depending on you machine). The result of solving this differential equation is shown below for only since takes very large values and this distorts the graphical interpretation.    Using the stiff solver optimises the stepsizes for stiff regions. Particularly, if a region is deemed to be considerably “stiff”, the ode15s will use smaller stepsizes to solve the problem but if there is a region where the differential is not “stiff”, then larger stepsizes will be used. Therefore, ode15s usually requires fewer grid points overall, for instance to solve the above set of differential equations, ode15s only requires 1,836 grid points while ode45 requires 7,820,485 grid points, that is over 4,200 times more grid points than ode15s. This just goes to show that stiff differential need implicit methods, even though the cost for every step is greater than that of an explicit method, fewer steps are required in total.  An alternative stiff differential equation solver is ode23s which achieves that same outcome as ode15s but with a lower accuracy and more grid points using only second and third order methods. |

# 8. Boundary Value Problems

***Boundary Value Problems*** (BVPs) are similar in many ways to initial value problems in the sense that a set of differential equations are given that are to be solved subject to certain conditions. In initial value problems, these conditions are imposed at the starting time but in boundary value problems, they are imposed at particular locations.

One of the most important differences when it comes to solving BVPs versus IVPs is the existence of solutions. Solutions to initial value problems always *exist* and are *unique* (subject to certain restriction on the right hand side), this is as a consequence of the *Picard-Lindelöf* theorem. The same cannot be said for boundary value problems; the solution to BVPs could exist and be unique, exist and not be unique or not exist at all.

## 8.1 Example of Boundary Value Problems

Consider a mass hanging from a spring with spring constant . Suppose that the spring is extended (by pulling the mass) by a distance as seen below. 

Then by *Hooke’s Law*, the force pulling the mass back to its equilibrium position is given by

As the mass is released, it will accelerate upwards with an acceleration and the force responsible for this acceleration is given by *Newton’s Second Law of Motion*

The acceleration is the second derivative of the displacement with respect to time and since it acts in a direction opposite to the extension, then

Equating the two expressions for the force from Newton’s Second Law and Hooke’s Law will give

This differential equation represents the simple harmonic motion of a mass hanging on a frictionless massless spring which oscillates with a frequency . Since this is a second order differential equation, two conditions need to be imposed:

* *Initial conditions* can be imposed at the starting time, specifically and which prescribe the initial position and initial speed,
* *Boundary conditions* can be imposed at different times, say and which prescribe the location at time and time .

## 8.2 Finite Difference Method for Boundary Value Problems

Consider the general second order boundary value problem

where the functions and are known functions of . Boundary value problems like this are solved using an incredibly versatile method known as the ***Finite Difference Method***. This procedure essentially changes a differential equation into a set of difference equations by using approximations to the derivatives.

## 8.3 Existence & Uniqueness of Solutions to BVPs

Consider the differential equation for the undamped simple harmonic oscillator with frequency 1, namely

This differential equation has the general analytic solution

where and are constants of integration which will be determined form the boundary conditions.

Three qualitatively different sets of boundary conditions will be investigated:

* and : The constants and can be found as:
* Therefore the analytic solution to the boundary value problem subject to these conditions is
* and this is captured by the finite difference approximation. In this case, the solution to the boundary value problem *exists* and is *unique*.
* and : The constants and can be found as:
* These two conditions provide an expressions for the constant only and not , therefore the particular solution will be
* which is valid for *any* value of . Therefore in this case, the solution *exists* but is *not unique*.
* and : The constants and can be found as:
* In this case, the boundary values have resulted in a contradiction and therefore the solution *does not exist* when subject to these boundary conditions.

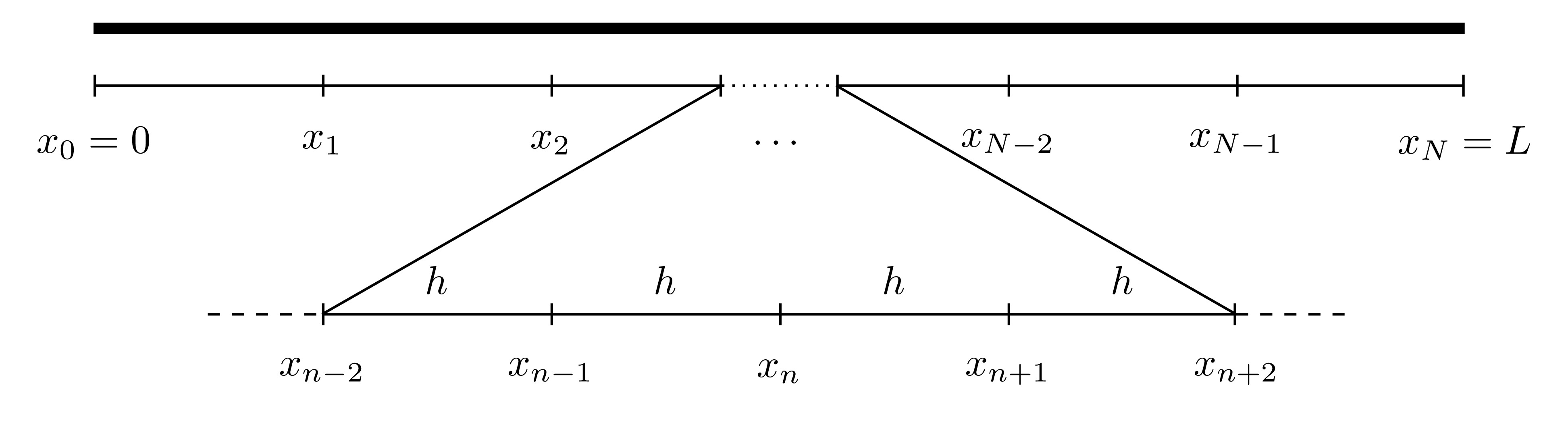
This final case is when the solution to a boundary value problem does *not exist*.

### 8.3.1 Finite Difference Approximations to the Derivatives

The term *finite difference approximation* refers to how derivatives can be approximated using linear expressions. For instance, the derivative of some function at a given point can be approximated as the gradient of between two points around , for example

There are many other ways in which these approximations can be made depending on the way in which the grid has been set up or on the context of the problem.

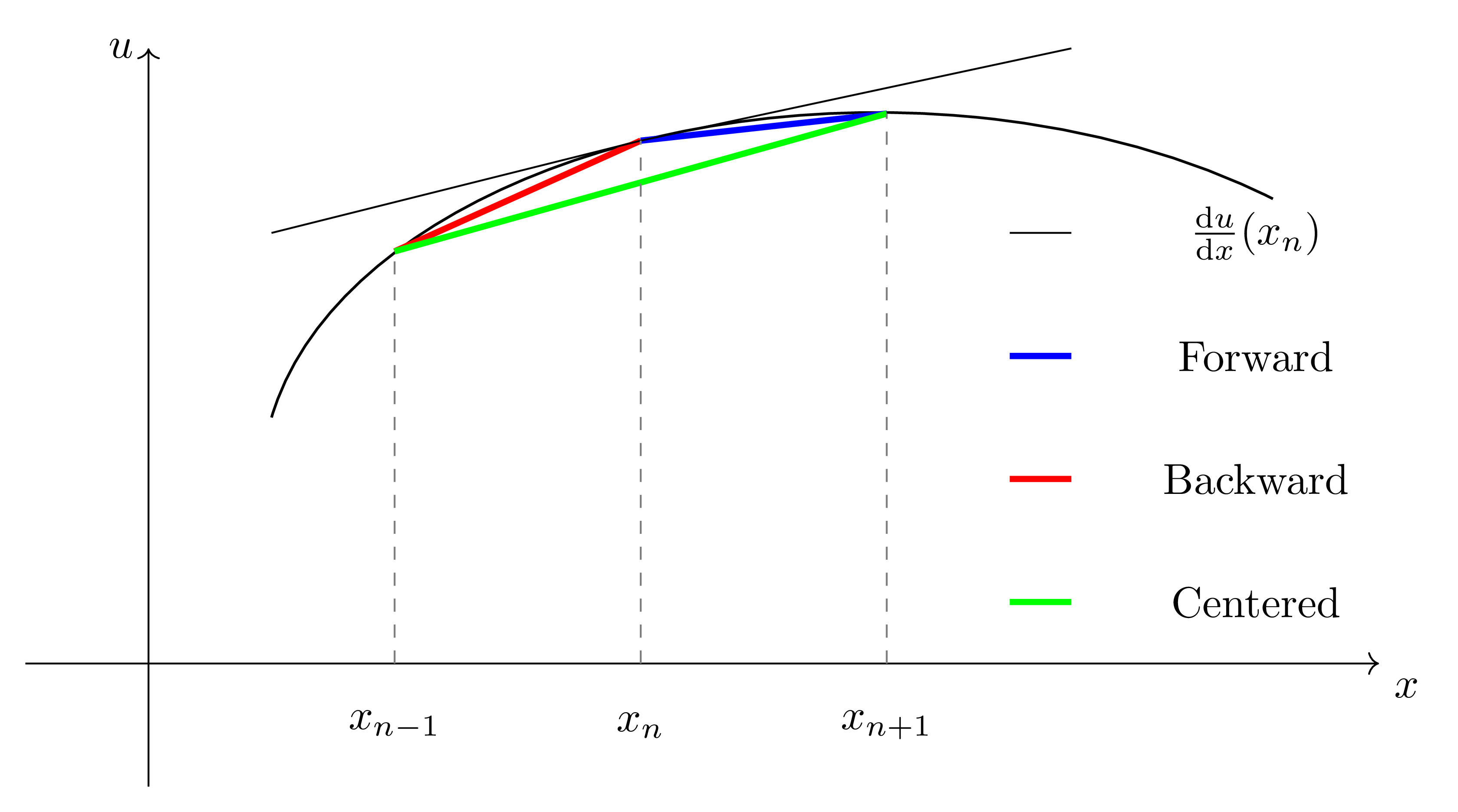
Consider a general unknown function defined on where and are given (as boundary conditions). First, split the interval into equally sized sections, each of width , and label the points as where .



For first and second derivatives, there are three main approximations that are most widely used:

* **Forward Difference**:
* **Backward Difference**:
* **Centred Difference**:

The graphical interpretation of the approximations to the first derivatives are shown below.



|  |
| --- |
| Second Derivative Expression |
| To show how the second derivative expressions are obtained, consider the centred difference approximation  To derive the expression for the second derivative, introduce two fictitious points (which is half-way between and ) and (which is half-way between and ). Then  The derivation of the second derivative approximations for the forward and backward differences can be done in a very similar way but without the need for half steps. |

Any of these three approximations can be used to approximate the derivatives of the function at the point . Denote the approximation of at the point by , i.e. , then

* **Forward Difference**:
* **Backward Difference**:
* **Centred Difference**:

These approximations will form the basis for solving the BVP.

### 8.3.2 Discretisation of the Differential Equation

Returning to the differential equation

Evaluate this equation at for some , then

For now, suppose the *centred differencing approximation* is used to approximate the derivatives. Replacing the approximations of the derivatives of at gives

This can be simplified by collecting the terms resulting in:

This expression will hold for all the values of (otherwise there will be points and which are outside the domain ). Therefore, this mean that there will be equations in unknowns which are .

This system may seem to be undetermined however, there are two boundary conditions that have not been taken into consideration yet, namely and . Since these are known, the approximations and have defined values, i.e.  and . This eliminates two of the unknowns giving equations in unknowns.

At , the approximation to the differential equation is

and since is already known, then it can be taken to the right hand side to give

Similarly, at , the approximation is

and since is known, this can be rewritten as

For , the approximation is

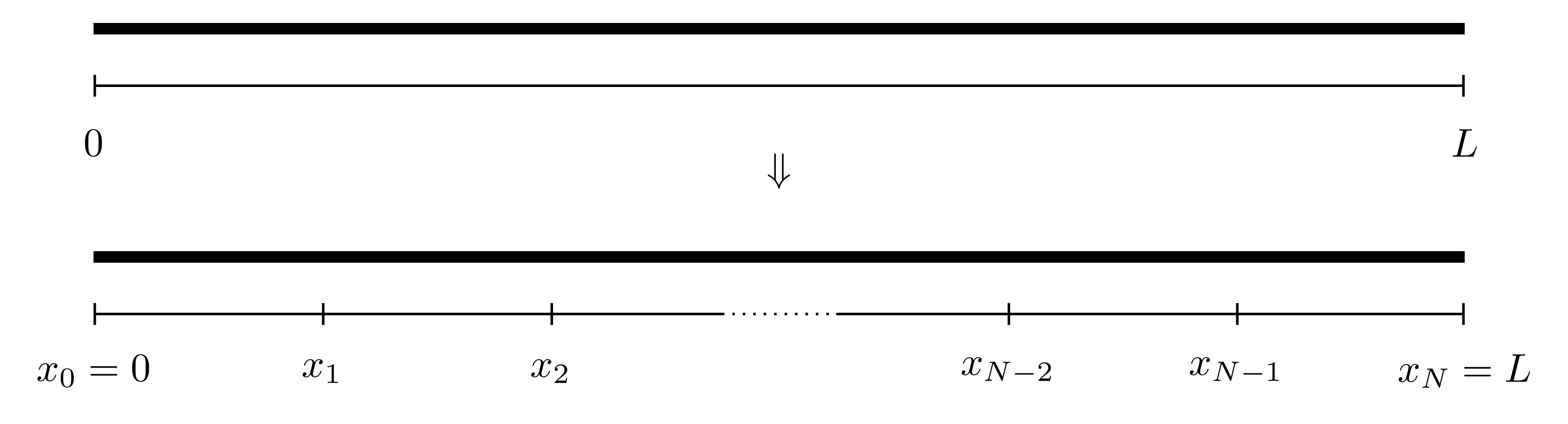
where and are al unknown. In summary, all of these equations are: These can be written in matrix form as , namely The matrix is of size all of whose terms are known, the vector of size also has terms that are all known. The unknown vector here is and it can be found by inverting to give .

Carrying out matrix inversions by hand can become increasingly cumbersome if is larger than and therefore this process should be done computationally. This can be solved in MATLAB by using either U=inv(A)\*g or U=A\g. The *backslash* method is faster than explicit matrix inversion if the matrix is of a large size.

The same process can be done for the forward and backward differencing approximations as well.

### 8.3.3 Steps of The Finite Difference Method

In summary, these are the steps of the finite difference method:

1. Divide the interval into equally sized sections, each of width and label the points as where . 
2. The values of the function are to be found at all the locations . Denote the approximation to the function at the point by , i.e.  for all .
3. Evaluate the differential equation at all the points where the derivatives are replaced by their *finite difference approximations*.
4. This will result in a set of linear equations in unknowns, namely, .
5. The values for and are known from the boundary conditions, since and and no approximation is needed since the exact values are known.
6. Write the whole system of equations in the matrix form and solve using MATLAB’s backlash operator.

|  |
| --- |
| BVP Example |
| Consider the boundary value problem  The differential equation itself can be solved analytically to give  This example will be used for the purposes of demonstration and comparison between the numerically obtained solution and the exact solution.  Suppose the interval is to be divided into equally sized sections, therefore and .  The functions and in this interval are:  The matrix values are  These can be used to obtain expressions for the matrix and the vector as  This system can be solved using U=inv(A)\*g or U=A\g. The numerical solution is compared to exact solution below.  The advantage of using this boundary value solver is that the computations are in no way taxing on MATLAB. The system that results is composed entirely of linear equations and this system is solvable (provided the boundary value problem does indeed have a solution which may not always be possible). MATLAB’s backslash operator is very effective in dealing with matrices, especially owing to the fact that the matrix is a tridiagonal matrix. |

## 8.4 MATLAB Code

Below is the MATLAB code that solves the BVP

using the centred differencing method with .

function BVP\_CD  
  
%% Solve BVPs using centered differences  
  
% The bvp is written in the form  
% a(x) u'' + b(x) u' + c(x) u = f(x) on x in [x0,L]  
% with the boundary conditions u(x0)=ul and u(L)=ur.  
% After the centered difference approximation is  
% used, the system will be written in the form AU=g.  
  
%% Lines to change:  
  
% Line 26 : x0 - Start point  
% Line 29 : L - End point  
% Line 32 : N - Number of subdivisions  
% Line 35 : xl - Left boundary value  
% Line 38 : xr - Right boundary value  
% Line 119 : Expression for the function a(x)  
% Line 127 : Expression for the function b(x)  
% Line 135 : Expression for the function c(x)  
% Line 143 : Expression for the function f(x)  
  
%% Set up input values  
  
% Start point  
x0=0;  
  
% End point  
L=10;  
  
% Number of subdivisions  
N=50;  
  
% Boundary value at x=x0  
ul=1;  
  
% Boundary value at x=L  
ur=-1;  
  
%% Set up BVP solver parameters  
  
% Interval width  
h=(L-x0)/N;  
  
% X = Vector of locations  
% (x1, x2, x3, ..., xN) (notice the start is x1 NOT x0)  
X=x0+h:h:L;  
  
% Evaluate the functions a(x), b(x), c(x) and f(x) at X  
aX=a(X);  
bX=b(X);  
cX=c(X);  
fX=f(X);  
  
% Find the expressions for alpha, beta and gamma at X  
alpha=aX/(h^2)-bX/(2\*h);  
beta=-2\*aX/(h^2)+cX;  
gamma=aX/(h^2)+bX/(2\*h);  
  
% Set up the vector g on the right hand side  
g=zeros(N-1,1);  
g(1)=fX(1)-alpha(1)\*ul;  
g(N-1)=fX(N-1)-gamma(N-1)\*ur;  
for j=2:1:N-2  
 g(j)=fX(j);  
end  
  
% Set up the matrix A on the left hand side (LHS\_A is  
% to avoid confusion with the function a(x))  
A=zeros(N-1,N-1);  
A(1,1)=beta(1);  
A(1,2)=gamma(1);  
A(N-1,N-1)=beta(N-1);  
A(N-1,N-2)=alpha(N-1);  
for j=2:1:N-2  
 A(j,j-1)=alpha(j);  
 A(j,j)=beta(j);  
 A(j,j+1)=gamma(j);  
end  
  
% Solve for the unknown vector U (it is then readjusted  
% from a column vector to a row vector for plotting)  
U=A\g;  
U=U';  
  
% Add the missing term x0 to the start of the vector x  
X=[x0,X];  
  
% Add the left and right boundary values to the vector U  
U=[ul,U,ur];  
  
%% Setting plot parameters  
  
% Clear figure  
clf  
  
% Hold so more than one line can be drawn  
hold on  
  
% Turn on grid  
grid on  
  
% Setting font size and style  
set(gca,'FontSize',20,'FontName','Times')  
  
% Label the axes  
xlabel('$t$','Interpreter','Latex')  
ylabel('$u(t)$','Interpreter','Latex')  
  
% Plot solution  
plot(X,U,'-k','LineWidth',2)  
  
end  
  
function [A]=a(X)  
A=zeros(size(X));  
for i=1:1:length(X)  
 x=X(i);  
 A(i)=1;  
end  
end  
  
function [B]=b(X)  
B=zeros(size(X));  
for i=1:1:length(X)  
 x=X(i);  
 B(i)=2;  
end  
end  
  
function [C]=c(X)  
C=zeros(size(X));  
for i=1:1:length(X)  
 x=X(i);  
 C(i)=exp(-x);  
end  
end  
  
function [F]=f(X)  
F=zeros(size(X));  
for i=1:1:length(X)  
 x=X(i);  
 F(i)=sin(x);  
end  
end

## 8.5 Comparison Between Forward, Backward & Centred Difference Approximations

The main difference between the different differencing schemes if the order of accuracy. Indeed, the error of the forward and backward differencing methods are whereas the error for the centred differencing is . This means that if the stepsize was reduced by a factor of 10, then the error for the forward and backward finite difference approximations would also reduce by a factor of 10 while the centred would reduce by a factor of 100.

|  |
| --- |
| Comparison |
| Consider the BVP  This has the exact solution  Below are the plots for the numerical solution to this boundary value problem using the forward (red), backward (blue) and centred (green) difference approximations compared to the exact solution when .  It can be seen that even for this relatively crude interval subdivision of , the centred approximation has yielded a far more favourable result compared to the other two methods. The following table shows the 2-norm error between the exact solution and the approximation for different values of :   | Method |  |  |  |  | | --- | --- | --- | --- | --- | | Forward | 4.1444 | 3.0875 | 1.9823 | 1.4048 | | Backward | 4.7243 | 3.8535 | 2.0939 | 1.4251 | | Centred | 0.5226 | 0.1677 | 0.0413 | 0.0146 |   It can be seen that even when , the 2-norm error has still not reduced below 1 for the forward and backward difference approximations but the centred has already achieved that even at . This is just a demonstration to show that how a simple change in the way in which derivatives are approximated can have such a drastic effect on the final solution. |

## 8.6 MATLAB’s In-Built Procedures

MATLAB has an in-built mechanism that can also solve second (or even higher order) BVPs, this is done using the bvp4c command.

Below is the MATLAB code that solves the BVP

using bvp4c.

function BVP\_InBuilt  
  
%% Solve BVPs using bvp4c  
  
% The bvp is written in the form  
% a(x) u'' + b(x) u' + c(x) u = f(x) on x in [x0,L]  
% with the boundary conditions u(x0)=ul and u(L)=ur.  
  
%% Lines to change:  
  
% Line 24 : x0 - Start point  
% Line 27 : L - End point  
% Line 30 : N - Number of spatial points  
% Line 33 : xl - Left boundary value  
% Line 36 : xr - Right boundary value  
% Line 44 : Expression for the function a(x)  
% Line 45 : Expression for the function b(x)  
% Line 46 : Expression for the function c(x)  
% Line 47 : Expression for the function f(x)  
  
%% Set up input values  
  
% Start point  
x0=0;  
  
% End point  
L=10;  
  
% Number of spatial points  
N=50;  
  
% Boundary value at x=x0  
ul=1;  
  
% Boundary value at x=L  
ur=-1;  
  
%% Set up BVP solver parameters  
  
% Set up solving space  
X=linspace(x0,L,N);  
  
% Define the functions in the BVP  
a= @(x) 1;  
b= @(x) 2;  
c= @(x) exp(-x);  
f= @(x) sin(x);  
  
%% Set up BVP solving parameters  
  
% First, write the second order ODE as a set of first order  
% ODEs:  
% U'=V  
% V'=(-b(x)\*V-c(x)\*U+f(x))/a(x)  
  
% Second order BVPs can have more than one solution  
% and vector v is the initialising vector of solutions.  
% It can be kept as a vector of zeros  
v=[0 0];  
  
% Initialise vectors for space and v  
init=bvpinit(X,v);  
  
% Solve the bvp subject to the boundary values and  
% inital guesses  
sol=bvp4c(@(x,u) DUDT(x,u,a,b,c,f),@(x0,L) BCs(x0,L,ul,ur),init);  
  
% Evaluate the solution at the grid points  
U=deval(sol,X);  
  
% Convert U to columns for consistency  
U=U';  
  
%% Setting plot parameters  
  
% Clear figure  
clf  
  
% Hold so more than one line can be drawn  
hold on  
  
% Turn on grid  
grid on  
  
% Setting font size and style  
set(gca,'FontSize',20,'FontName','Times')  
  
% Label the axes  
xlabel('$t$','Interpreter','Latex')  
ylabel('$u(t)$','Interpreter','Latex')  
  
% Plot solution  
plot(X,U(:,1),'-k','LineWidth',2)  
  
end  
  
function [dudx]=DUDT(x,u,a,b,c,f)  
  
dudx(1)=u(2);  
  
dudx(2)=(-b(x)\*u(2)-c(x)\*u(1)+f(x))/(a(x));  
  
end  
  
function res=BCs(x0,L,ul,ur)  
% The boundary conditions are written as  
% u(x0)=ul  
% x(L)=ur  
  
res=[x0(1)-ul;L(1)-ur];  
  
end

# 9. Mixed Value Problems

Initial and boundary value problems are not the only two ways in which conditions can be expressed. Sometimes these conditions can be presented in a *mixed form* where the condition on one or both boundaries may depend on the derivative of the solution function. For instance, consider the steady-state convection-diffusion equation on a bar on length with density , convective velocity , specific heat capacity , thermal conductivity and heat source :

where is the temperature at . This set of conditions are known as ***Mixed Conditions***: the first means that the temperature at the location is , the second means that at the location , there is no heat *flux*. This can be quite useful if say, a metal rod is being heated to C on one side an insulated on the other.

The method to solving MVPs is the same as boundary value problems subject to a few modifications.

## 9.1 Finite Difference Method for MVPs

Consider the differential equation

as before. The interval will be split into equally sized sections each of width and the grid points are labelled for . This differential equation can be discretised using the centred difference approximation just as before to give

This gives a set of equations in unknowns, namely (recall that for ).

When the differential equation is subjected to two boundary conditions, say

then expressions for and are provided which gives equations in unknowns, hence resulting in a well-defined system which can be solved as before.

However, suppose that a set of mixed conditions is given as

In this case, only is explicitly known, meaning that there will be equations in unknowns since is not known giving an under-determined system (a system with more unknowns than equations). So either one more equation is needed or one more unknown needs to be removed. All the unknowns are certainly needed, otherwise the solution will be incomplete, so the alternative is to find another equation to add to the set of equations.

The set of equations is: All these come from the discretisation

Evaluating this at gives

Initially, this may seem to be quite strange since there is a point which is the approximation to the solution at the point which is certainly out of the range of consideration. This point is considered to be *an artificial grid point* that will act as a placeholder in meantime.

Consider the condition at the start point

Using the centred finite difference approximation on the derivative gives

This approximation can be manipulated to provide an expression for the artificial point as

Replacing this into the equation [Equation 9.1](#eq-n0) will eliminate completely giving an equation in terms of and only, namely

Therefore, another equation has been found which now completes the system of equations in unknowns. Thus the system of equations is: This can be written in matrix form as where This can once again be solved on MATLAB using U=inv(A)\*g or U=A\g.

If, on the other hand, the mixed conditions were instead

then the artificial point will be located at but the same procedure can be done give the matrix system where

|  |
| --- |
| Mixed Value Problem |
| Consider the differential equation for a damped harmonic oscillator  with the mixed conditions  This MVP is to determine the trajectory of the mass if the launching speed at the start is , which is , and after seconds, the mass reaches its equilibrium state, which is . Notice that there is no restriction on the starting location, only the starting speed, so the mass can start anywhere as long as it is launched with a velocity .  The starting location here happens to be at but that is no restricted by the mixed conditions as long as the gradient at the start is . |

# 10. Symmetric Boundary Conditions

The use of symmetric boundary conditions arises in many cases where conditions at the ends are not known explicitly but they are related. For instance, consider the ODE representing the conduction problem

where is the material’s conductivity and is the heat transfer. Symmetric boundary conditions can be imposed as

for some constant . This problem can be interpreted as an insulated metal rod of length that has been heated all the way through and then as it cools, it loses heat equally from both ends (which is the condition ), and that this heat loss at is proportional to the temperature gradient between the rod and the air (which is the second condition ). The issue with this type of problems is that the temperature at both boundaries are not explicitly known, but it is known that they are the same.

## 10.1 Finite Difference Method for Symmetric Boundary Value Problems

This problem can be tackled in a very similar way to BVPs and MVPs. Consider the differential equation

The interval will be split into equally sized sections each of width and the grid points are labelled for . This differential equation can be discretised using the centred difference approximation (just as in [Section 8.2](#sec-FDM)) to give

This gives a set of equations in unknowns, namely . In this case, neither nor are explicitly known, therefore none of the unknowns can be eliminated from the boundary conditions *per se*.

Suppose the given conditions are

where and are some constants. The first condition is the symmetric boundary condition which represents the fact that the value of the unknown solution at both ends is the same, then , even though neither is explicitly known. The term can be eliminated since determining automatically determines , this reduces the number of unknowns to .

Consider the discretisation at , namely

since , this can be rewritten in terms of instead as

The discretised form of the differential equation at is

Just as in the case with the MVPs, an artificial point is introduced which is the solution approximated at the point which is beyond the computational domain.

To find an expression for , first consider the second condition

The LHS can be rewritten in terms of its centred differencing approximation as

Combining these two can give an expression for as:

Replacing this into [Equation 10.1](#eq-UNp1) gives

thus providing the last equation to complete the set. Finally, this system can be written in matrix form as where This can then be solved in MATLAB but bearing in mind that which determines the function at and .

|  |
| --- |
| Symmetric Boundary Value Problem |
| Consider the conduction problem  with the conditions |

# 11. Heat Equation

Ordinary differential equations have been the main focus of this course so far but this will now be extended to *partial differential equations*. The differential equations that will be studied here are the *1-Dimensional Heat (or Diffusion) Equation* and the *1-Dimensional Advection (or Convection) Equation*.

The 1-dimensional *heat* (or *diffusion*) *equation* is a partial differential equation that represents the heat transfer across a rod and is given by

where is the temperature at location at time and is the thermal diffusivity[[6]](#footnote-281). This equation represents the flow of heat along the length of a rod of length .

This partial differential equation has three derivatives in total, two derivatives in and one derivative in , this means that three conditions are needed, two on and one on :

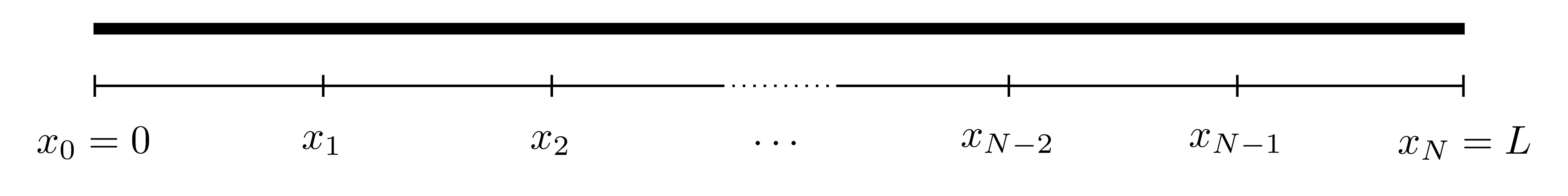
* for : Initial heat distribution across the rod;
* for : The temperature at the left end of the rod;
* for : The temperature at the right end of the rod.

This set of conditions along with the differential equation are known collectively as an ***Initial-Boundary Value Problem*** and can be solved using the ***Method of Lines***.

## 11.1 The Method of Lines for the Heat Equation

The outline of the method of lines for the heat equation is as follows:

1. Divide the spatial interval into equally sized sections and label the points as where and the spatial interval width is .



1. **Left Hand Side**: For each point , define the approximation . Therefore the left hand side of the heat equation can be written as

* and this holds for since and are already known from the boundary conditions. Notice that the derivative of is an ordinary derivative since is a function of only.

1. **Right Hand Side**: Use the finite difference approximation to approximate the *spatial* derivative in the differential equation. Here, the centred difference approximation for the second derivative will be used, namely

* Therefore the right hand side of the heat equation will become
* This holds for bearing in mind, once again, that and are known beforehand.

1. These can be combined to give the discretised form of the heat equation

* for all where . This means that the partial differential equation has been split into ordinary differential equations.

1. This entire system of equations can now be written in matrix form as where subject to the *initial condition*

* This system can now be solved using any of the IVP solvers with a temporal stepsize .

In essence, the *Method of Lines* has converted a PDE into a set of ODEs using the same techniques as BVPs and will be solved in the same way as IVPs.

|  |
| --- |
| Heat Equation |
| Consider an iron rod (of thermal diffusivity ) of length 1 where the middle section of length 0.2 has been heated to a temperature of 1 while the rest is at 0. The ends of the rod have been kept at a constant temperature of 2. This system can be represented by the IBVP  First, divide the interval into five equal sections (which will be of width ).  This system can be discretised using the centred difference method and written in matrix form as where  The differential equation  can be solved using the Euler method with the initial condition  subject to a time stepsize . Below are the plots of the heat distribution at for () and ().    At the beginning, the temperature at the ends is 2 and the middle section is at a temperature of 1. As time progresses, the heat evens out across the iron bar until eventually, the whole bar will be the same temperature. |

## 11.2 Linear Advection Equation

The heat equation deals with heat transfer through diffusion throughout a material. Another way in which heat transfer can be achieved by advection (or convection) and this is given by

where is the temperature at location at time and is the flow speed.

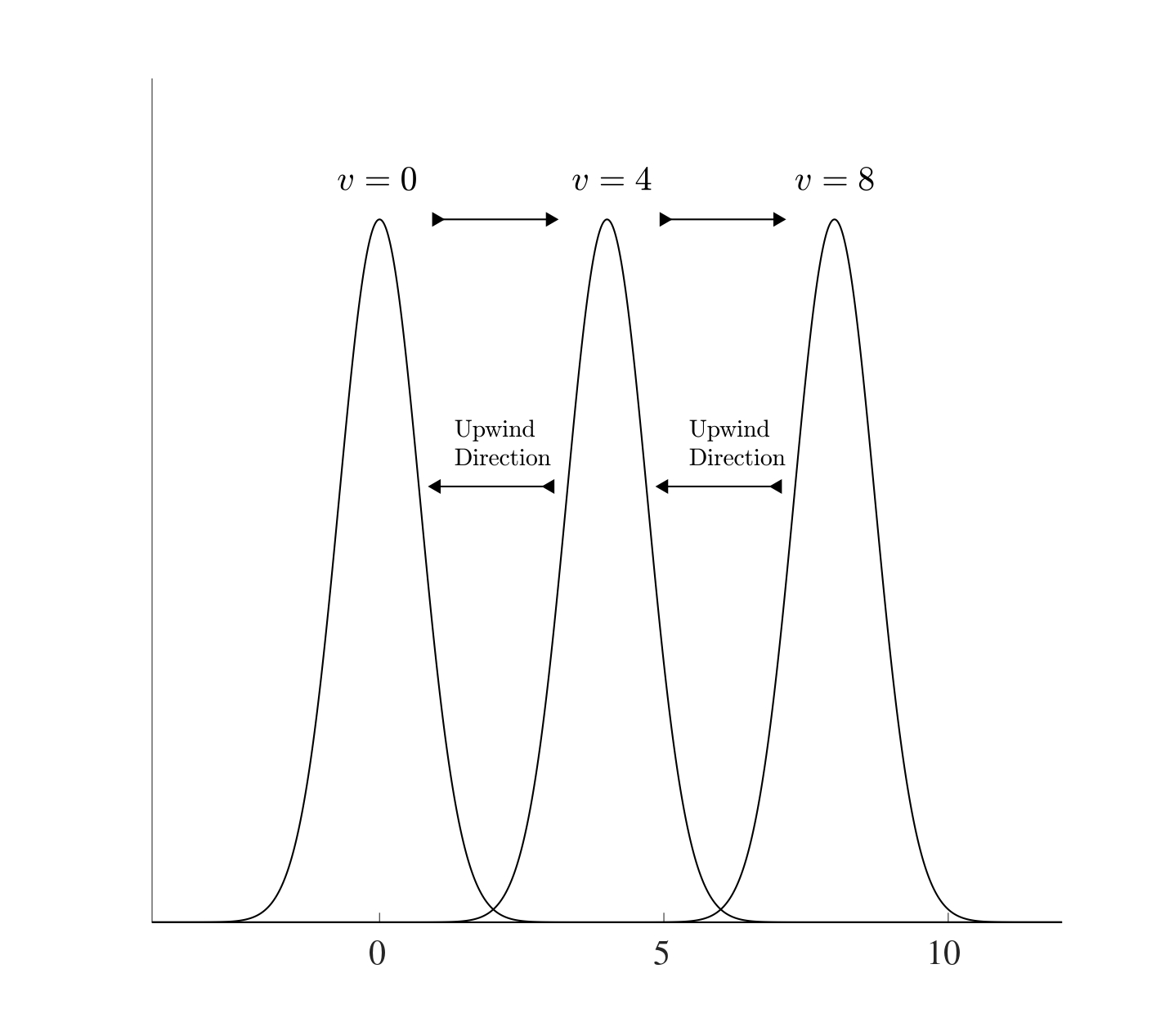
This partial differential equation has two derivatives in total, one in and one in , this means that two conditions are needed, one spatial and one temporal:

* for : Initial heat distribution across the rod;
* for : The temperature at the left end of the rod.

Consider the PDE along with the initial condition only, namely for . The exact solution to this differential equation is given by

this can be verified from the partial differential equation as follows:

This means that if the initial heat profile takes the form of , then after time , the profile will look exactly the same but shifted to the right by a distance .



The “information” moves from left to right so if the finite differences are to be used, the centred differencing approach would not be suitable since the information on the right is not known yet. Therefore the backwards differencing approximation will be the most suitable. This is known as an *upwind/upstream scheme* (i.e. against the direction of the wind/stream) if . Therefore using the convention where is the discretisation of the spatial points for , the backward differencing approximation to the spatial derivative is

Therefore is discretised advection equation is

and this can be solved subject to the initial condition

and boundary condition

to give the discretised set of equations in the form where and the initial condition is

.

## 11.3 Convection-Diffusion Equation

The heat (or diffusion) equation dictates the spread of heat across a length of material while on the other hand, the advection (or convection) equation dictates the flow of heat in a certain direction. The combination of these two effects gives rise to the ***Convection-Diffusion Equation*** which takes the form

Just as in the heat equation, this partial differential equation has three derivatives in total, two derivatives in and one derivative in , this means that three conditions are needed, two on and one on , these will be as follows:

* for : Initial heat distribution across the rod;
* for : The temperature at the left end of the rod;
* for : The temperature at the right end of the rod.

In order to discretise this system, a finite difference approximation needs to be chosen first. The centred difference approximation was used for the heat equation and the backwards difference approximation for the advection. Here, the combination of both will be used. Even though this might initially seem like an inconsistency, but in fact, this will allow the system to present a distinct stable advantage as will be seen in the next section.

This system can be discretised in exactly the same way as before

This system can be written in the form where

and this system can be solved using an Euler iteration subject to the initial condition .

## 11.4 Asymptotic Stability

The method of lines is essentially a hybrid method that makes use of a combination between a finite difference approximation and the Euler method and is very effective at solving partial differential equations, as seen from solving the heat, advection and convection-diffusion equations. The derivation of the method of lines for the different methods builds on the very same principle and the codes can be adapted quite easily. One main issue that arises here is the choice for the stepsizes for both the spatial and temporal discretisations, i.e. the choice of and respectively. When both methods are combined, there needs to be a restriction on both stepsizes.

The first issue that needs to be addressed is the asymptotic stability of the heat equation and the advection equation. For arbitrarily large matrices, it may not be simple to determine if all the eigenvalues are negative since it may be computationally restrictive to do so. However, a result can be used to see if all the eigenvalues are negative without explicitly calculating them.

**Theorem 11.1 (Gershgorin Circle Theorem)** Let be an given by

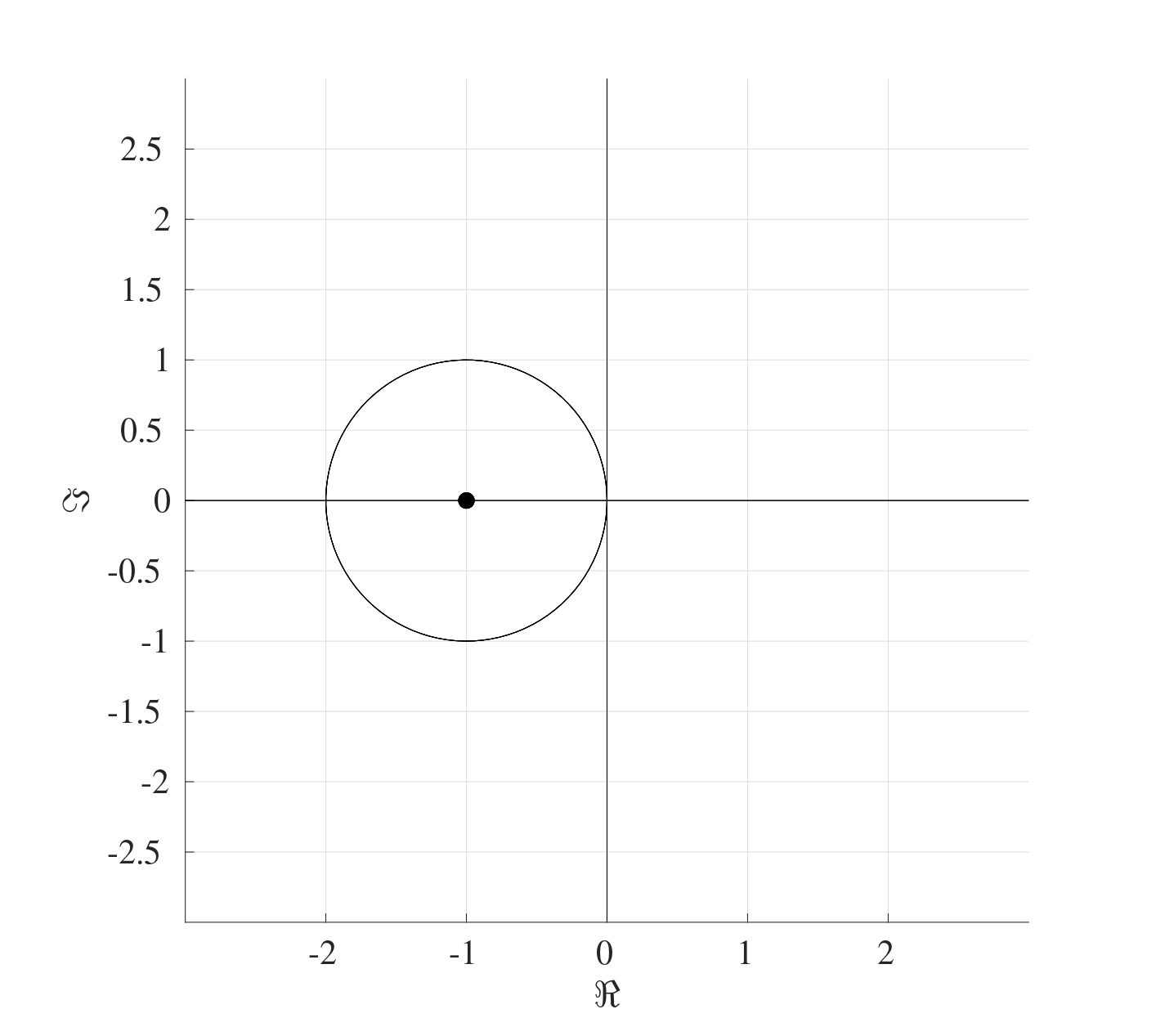
On the complex plane, consider closed discs, each centred at the locations for (the diagonal terms) where the disc centred at has a radius where

Then all the eigenvalues of the matrix will have to lie in at least one of these discs. In other words, every eigenvalues of satisfies

|  |
| --- |
| Gershgorin Circle Theorem Exapmple |
| Consider the matrix  Following the steps of the theorem:   1. Indicate the locations of the diagonal terms (namely ) on the complex plane. 2. Find the radii which are equal to the row sum of the absolute terms *without* the diagonal terms, in other words, 3. Draw a circle around with radius , a circle around with radius and so on. 4. All the eigenvalues of the matrix must lie in at least one of the circles indicated. Indeed, the following figure shows the diagonal terms each with circles around them with the appropriate radius. The eigenvalues are given in red and the blue circles are those which contain all said eigenvalues. |

### 11.4.1 Stability of the Euler Method for the Advection Equation

Consider the matrix of size from the advection equation

Following the steps of the Gershgorin theorem, the centres of all the circles on the complex plane will be located at the diagonal terms, all of which are . The radii of these circles are the row sums of the matrix without the diagonal terms, which means that all the radii will be 1. The figure below shows the circle that results on the complex plane. Therefore regardless of what the eigenvalues might be, it is known that they will always have negative real parts and therefore the advection matrix forms an asymptotically stable system. 

Since the advection equation is asymptotically stable, a bound for the temporal stepsize needs to be found. Consider the advection equation after the discretisation where . The Euler method is numerically stable if the time step satisfies

First calculate :

where . Now taking the absolute value of all the terms and taking the row sums gives:

The row sums of the absolute terms of this matrix are

Since it is assumed that , then therefore, . Consider the two cases when and $ 1. If , then :

Therefore if , then .

1. If , then :

* therefore in this case, if needs to be less than or equal to , then
* which contradicts with the assumption that .

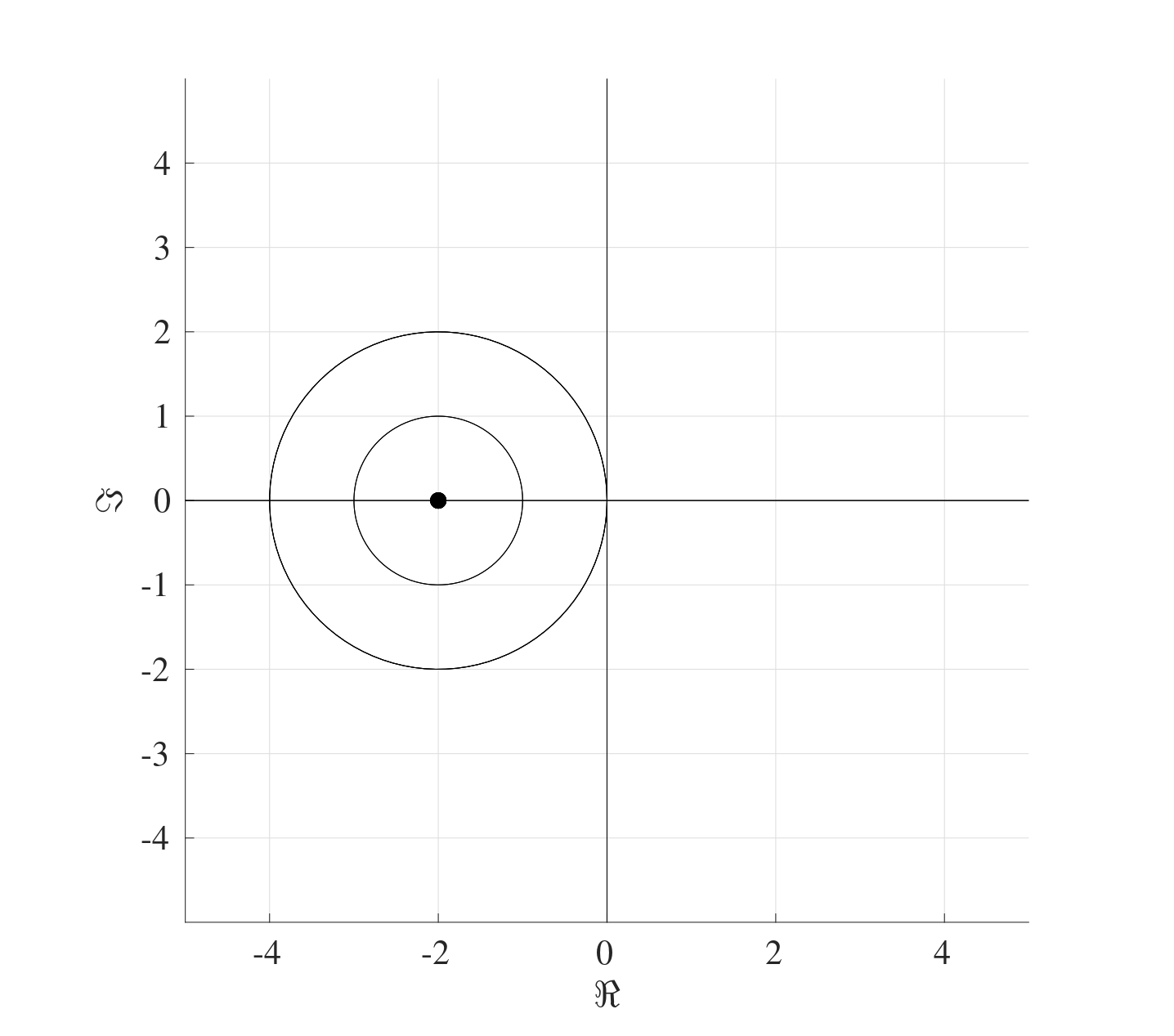
Therefore, the Euler method will produce a convergent solution if

In terms of number of spatial and temporal points and respectively, this restriction would be

So for a fixed velocity , if the time step is to be halved, then the spatial step would also need to be halved as well.

### 11.4.2 Stability of the Euler Method for the Heat Equation

Consider the matrix of size from the heat equation

The steps of the Gershgorin theorem can be followed to produce the following figure on the complex plane.  Once again, this shows that all the eigenvalues will have negative real parts even though their explicit values are not known.

To determine the bound on the stepsize, consider the heat equation after the discretisation, which is where . The Euler method is numerically stable if the time step satisfies

First calculate :

where . Now taking the absolute value of all the terms and taking the row sums gives:

The row sums of the absolute terms of this matrix are

Since , then and therefore, . Consider the two cases and .

1. If , then :

* therefore .

1. If , then :

* therefore in this case, if needs to be less than or equal to , then
* which contradicts with the assumption that .

This means that the Euler method produces a stable convergent solution if

In terms of number of spatial and temporal points and respectively, this restriction would be

So for a fixed diffusivity , if the time step is to be halved, then the spatial step would should be quartered.

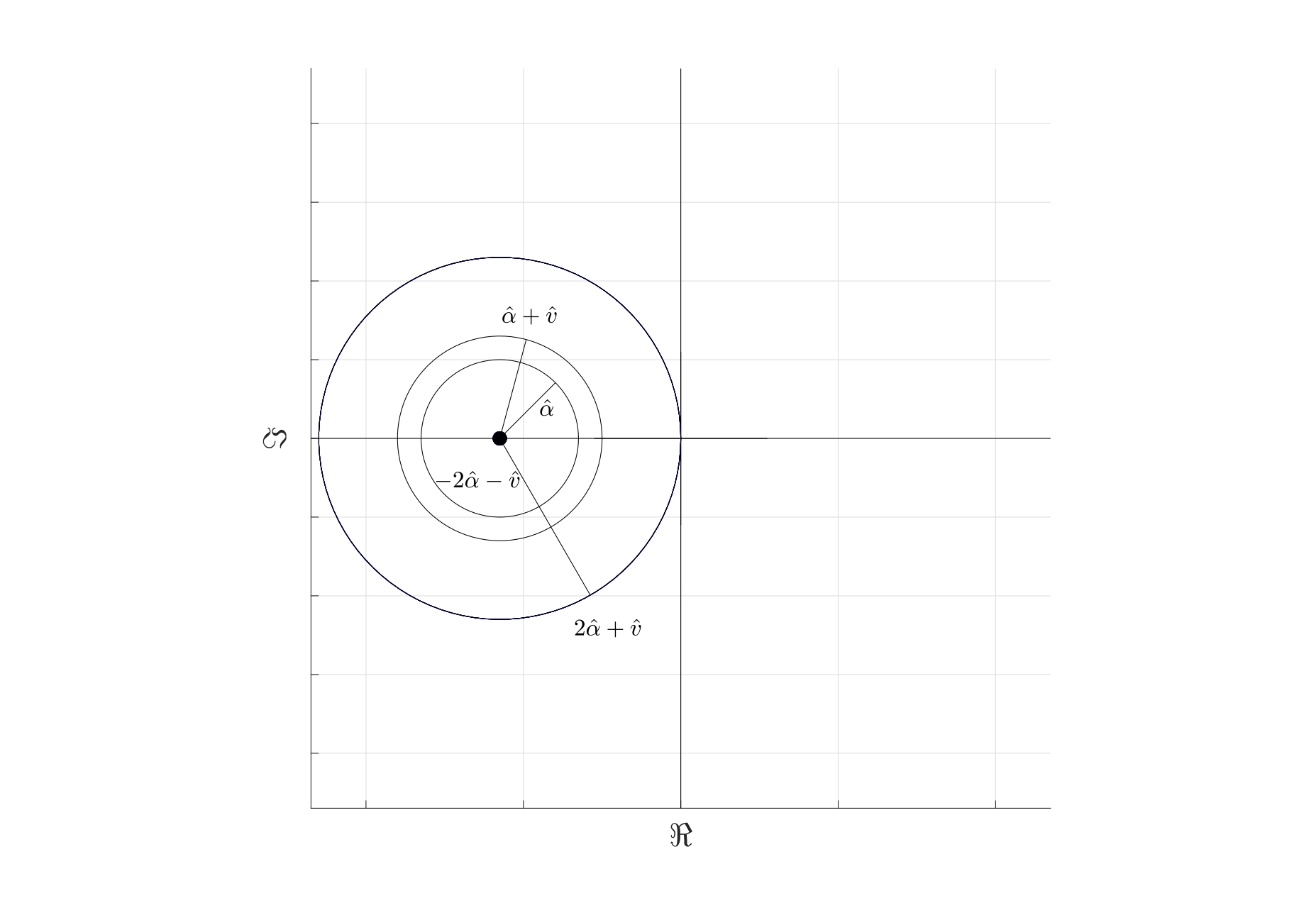
## 11.5 Stability of the Convection-Diffusion Equation

Now that it has been established that both the heat and advection equations are asymptotically stable and the stepsize bounds have been found, it is time to combine both cases to tackle the convection-diffusion equation.

When discretised, the convection-diffusion equation can be written as where the matrix is given by

The Gershgorin theorem can be applied to the matrix to show that all the eigenvectors have negative real parts. Indeed,

where and . By the Gershgorin theorem, the centres of the circles will be located at the diagonal terms, namely at with the radii , and . The largest radius is which means that all the eigenvalues will be negative as shown below. Therefore the convection-diffusion equation is asymptotically stable.



To find the bound for the stepsizes, consider the convection-diffusion equation after the discretisation where . The Euler method is numerically stable if the time step satisfies

Calculating :

where and . Taking the absolute value of all the terms and adding the rows gives

The row sums of the absolute terms of this matrix are

Since and , then , therefore, . Consider the two cases and .

1. If , then :

* therefore .

1. If , then :

* therefore in this case, if needs to be less than or equal to , then
* which contradicts with the assumption that .

This means that the Euler method will produce a stable convergent solution if

This means that a choice can be made with regards to the bounds of the different components, for instance, the values of and can be chosen such that

or any combination thereof provided that the choices satisfy the inequality .

|  |
| --- |
| Bound for Convection-Diffusion |
| Consider the convection-diffusion equation  This can be discretised to give where  subject to  As yet, the value of has not been put forward since the stepsizes need to be established first. For a stable Euler method, the stepsizes and need to satisfy  If and (which corresponds to and ), then the Euler method will be stable. |

1. Note that exists only if is non-singular, meaning that the condition number number only exists if is non-singular. [↑](#footnote-ref-53)
2. A matrix is ***Diagonally Dominant*** if every diagonal entry is larger in absolute value than the sum of the absolute value of all the other terms in that row. More formally

   The matrix is ***Strictly Diagonally Dominant*** if the inequality is strict. [↑](#footnote-ref-92)
3. In most cases, the interval width is constant but more advanced numerical techniques have different subinterval widths. [↑](#footnote-ref-110)
4. ***Taylor’s Theorem*** states that for a function that is at least times differentiable in the open interval (or ), then

   for some point between and . [↑](#footnote-ref-130)
5. In general, according to the ***Picard-Lindelöf Theorem***, an IVP of the form with has a *unique solution* if the function is continuous in and uniformly Lipschitz continuous in . In this example shown above, the function does not satisfy the aforementioned conditions and therefore the initial value problem does not have a unique solution. These concepts of continuity are far beyond the realms of this course and no further mention of them will be made. [↑](#footnote-ref-160)
6. The thermal diffusivity will always be regarded as a constant and usually takes the form where is the thermal conductivity, is the density of the material and is the specific heat capacity. [↑](#footnote-ref-281)