NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

Course Notes for AMATH741/CM750/CS778

Part I: Finite Difference Methods

Hans De Sterck
Paul Ullrich

January 2020

These notes have been funded by

MEF
Mathematics Endowment Fund

University of Waterloo

CONTENTS

Preface **5**

1 Overview of PDEs **9**

1.1 Linear Second Order PDEs with Two Independent Variables 9

1.1.1 A Note About Leibniz and Subscript Notation 10

1.1.2 Classification of Linear Second-Order PDEs 10

1.1.3 Derivation of the Heat Equation 14

1.2 Hyperbolic PDEs with Two Independent Variables 17

1.2.1 The Linear Advection Equation 18

1.2.2 The Wave Equation 19

1.2.3 d'Alembert's Solution for the Wave Equation IVP 20

1.2.4 Domain of Influence and Domain of Dependence 21

1.2.5 Existence and Uniqueness for the IVBVP 22

1.3 Elliptic PDEs with Two Independent Variables 23

1.3.1 The Dirac Delta 24

1.3.2 Domain of Influence 25

1.3.3 Discontinuous Boundary Conditions 28

1.3.4 Existence and Uniqueness 29

1.4 Parabolic PDEs with Two Independent Variables 29

1.4.1 Domain of Influence and Domain of Dependence 30
1.4.2 Discontinuous Initial Conditions 31
1.5 Linear Second Order PDEs with Three Independent Variables 32
1.5.1 A Note About Vector Calculus Notation 33

2 Finite Difference Methods 35
2.1 Finite Difference Methods for Elliptic PDEs 35
2.1.1 1D Elliptic Model Problem 35
2.1.2 2D Elliptic Model Problem 38
2.1.3 Convergence Theory 41
2.2 FD Methods for Hyperbolic PDEs 46
2.2.1 FD Methods for the 1D Linear Advection Equation 47
2.2.2 Stability 53
2.2.3 Dissipation and Dispersion 61
2.2.4 Finite Difference Methods for the Wave Equation 70
2.2.5 Finite Difference Methods in 2D and 3D 73
2.3 Finite Difference Methods for Parabolic PDEs 74
2.4 Finite Difference Convergence Theory for Time-Dependent Problems 76
2.4.1 Actual Error, Truncation Error and Consistency 79
2.4.2 Stability and Convergence: Lax Convergence Theorem 79
2.4.3 2-Norm Convergence 82

A Norms of Vectors, Functions and Operators 85
A.1 Vector and Function Norms 85
A.2 Norms of Grid Functions 87
A.3 Matrix Norms (Operator Norms) 89

Bibliography 93

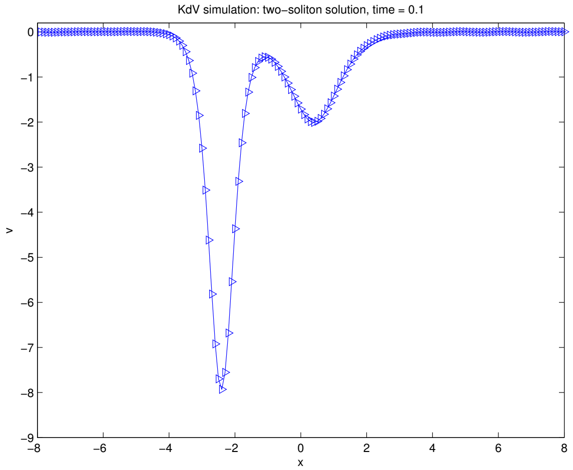
Preface

Mathematical models based on partial differential equations (PDEs) are ubiquitous these days, arising in all areas of science and engineering, and also in medicine and finance. Example application fields include fluid mechanics, general relativity, quantum mechanics, biology, tumour modeling and option pricing. Unfortunately, it is almost always impossible to obtain closed-form solutions of PDE equations, even in very simple cases. Therefore, numerical methods for finding approximate solutions to PDE problems are of great importance: numerical solutions of PDEs on powerful computers allow researchers to push the boundaries of knowledge, and allow companies to increase their competitive edge.

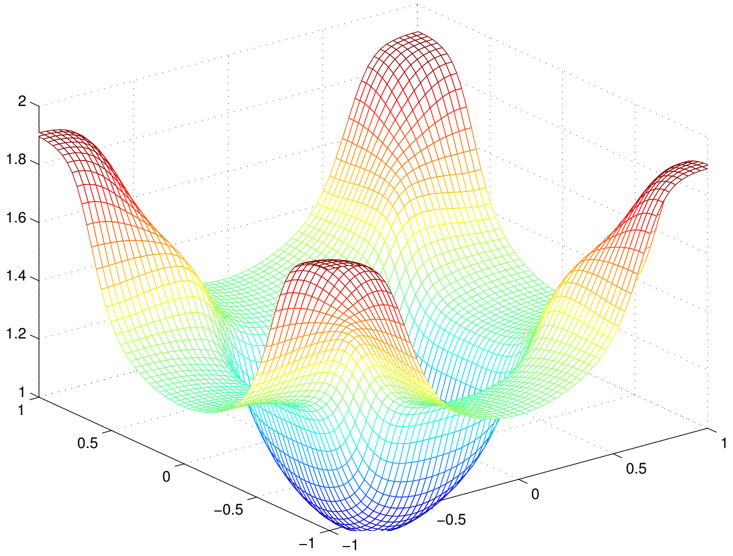
In this course you will learn about three major classes of numerical methods for PDEs, namely, the finite difference (FD), finite volume (FV) and finite element (FE) methods. Some theoretical background will be introduced for these methods, and it will be explained how they can be applied to practical problems.

The examples on the following few pages illustrate the types of problems that may be addressed by the techniques to be learned in this course. (You will also solve problems similar to this in the computational assignments of this course.)

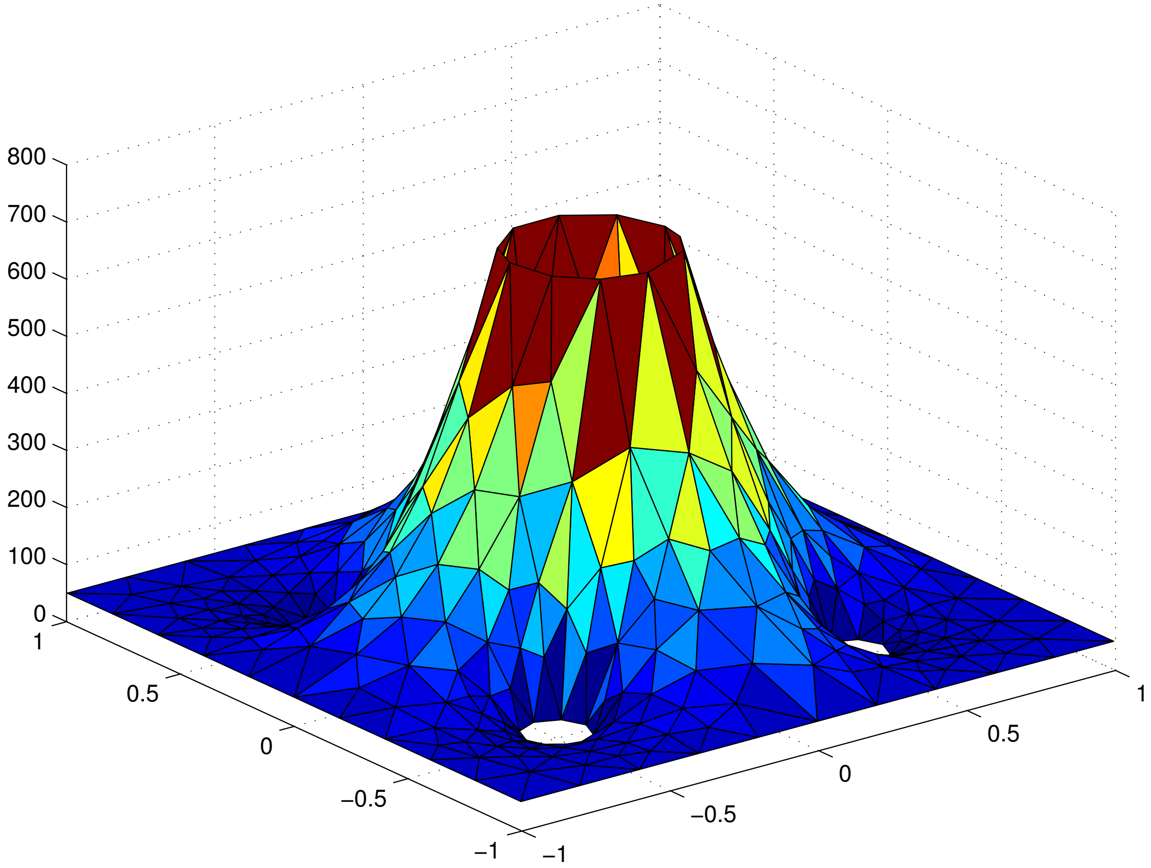
The figure below shows a numerical solution of interacting solitary waves, obtained by a FD method. Solitary waves are wave solutions of nonlinear PDEs that do not change shape, even after overtaking each other. This is a numerical simulation result for the so-called *Korteweg-deVries* PDE, which models the propagation of nonlinear waves in fluids.



The next figure shows a snapshot of a FV simulation of the so-called *shallow water* system of PDEs. The height of the water is shown, for a case where the water is contained in a square box. The water was initially concentrated near the center of the box, but as time progresses, the water spreads out in the box and splashes up against the walls. The solution is symmetric due to a symmetric initial condition.



The figure below shows the temperature distribution in an engine cylinder block with four pipes for cooling. The temperature is highest at the cylinder wall, and lowest at the cooling pipes. This result was obtained by applying the FE method to the stationary heat equation PDE on a so-called unstructured grid composed of triangles.



CHAPTER 1

Overview of PDEs

In this chapter we present a brief overview of partial differential equations and their general properties, focusing on linear second order PDEs with two independent variables.

1.1 Linear Second Order PDEs with Two Independent Variables

We now discuss linear second order PDEs with two independent variables, which are arguably the simplest non-trivial PDEs. Much of the theory of higher order linear PDEs, or those in more than two independent variables, can be derived as a natural extension of the material presented in this section.

Definition 1.1 *A **second-order PDE in two variables** x **and** y is an equation of the form*

* learn binomial expansion

$$F\left(u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \underbrace{\frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial x \partial y}, \frac{\partial^2 u}{\partial y^2}}_{}, x, y\right) = 0. \tag{1.1}$$

In addition, we provide the following definitions:

1. *We say the PDE (1.1) is **linear** if and only if F is linear in u and its partial derivatives. Otherwise, the PDE is **nonlinear**.* ??

2. *We say the PDE is **homogeneous** if and only if it is satisfied by a function which identically vanishes (i.e. $u \equiv 0$). Otherwise, the PDE is **inhomogeneous**.*

??

For example, a PDE of the form

$$a(x,t)\frac{\partial u}{\partial t} + b(x,t)\frac{\partial^2 u}{\partial x^2} = 0 \tag{1.2}$$

is both linear and homogeneous.

We note that any homogeneous PDE satisfies the *superposition principle*. Namely, if $u_1(x,y)$ and $u_2(x,y)$ are two solutions of a homogeneous PDE, then the function $u(x,y)$, defined by

$$u(x,y) = c_1u_1(x,y) + c_2u_2(x,y), \tag{1.3}$$

is also a solution of the homogeneous PDE.

1.1.1 A Note About Leibniz and Subscript Notation

Throughout this text we will interchangably use Leibniz notation and subscript notation to denote differentiation. The following table summarizes these differences.

Leibniz Notation	Subscript Notation
$\frac{\partial u}{\partial x}$	u_x
$\frac{\partial^2 u}{\partial y^2}$	u_{yy}
$\frac{\partial^4 u}{\partial^2 x \partial y \partial z}$	u_{xxyz}

1.1.2 Classification of Linear Second-Order PDEs

There are three general classes of linear second-order PDEs with two independent variables, namely *parabolic*, *hyperbolic* and *elliptic* equations. These classes are defined as follows:

Definition 1.2 *A linear second-order PDE with two independent variables on a domain Ω in the form*

$$A(x,y)u_{xx} + B(x,y)u_{xy} + C(x,y)u_{yy} = W(u,u_x,u_y,x,y) \tag{1.4}$$

is said to be

i) ***parabolic*** *if for all $x,y \in \Omega$, $B^2 - 4AC = 0$,*

ii) ***hyperbolic*** *if for all $x,y \in \Omega$, $B^2 - 4AC > 0$.*

$u_t = u_{xx}$ the change of temperature with time is proportional to the curvature of the temperature profile over x

$u_{tt} = u_{xx}$ the curvatu of the surface in time (frequency) is proportional to the curvatu of the surface in space ($1/\lambda$)

$u_{xx} = -u_{yy}$ the curvatu of the surface over x is opposite and proportional to the curvatu of the surface over y.

iii) ***elliptic*** *if for all* $x, y \in \Omega$, $B^2 - 4AC < 0$,

We now give three important examples of second-order linear PDEs in two variables.

B^2-4AC	Partial Differential Equation	Type	Example Solution
$\emptyset$	$\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0$ (heat equation)	Parabolic	$u(x,t) = \exp(-t)\cos(x)$, $t > 0$
4	$\dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} = 0$ (wave equation)	Hyperbolic	$u(x,t) = \cos(x \pm t)$
-4	$\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0$ (Laplace equation)	Elliptic	$u(x,y) = x + y$

+ as t advances x must recede
for $\cos(x \pm t)$ to stay the same
− as t advances x must advance

The classification of these PDEs can be quickly verified from definition 1.2. These three equations are known as the *prototype equations*, since many homogeneous linear second order PDEs in two independent variables can be transformed into these equations upon making a change of variable. We now discuss each of these equations in general.

Example 1. The 1D Heat Equation (Parabolic Prototype) One of the most basic examples of a PDE is the 1-dimensional heat equation, given by

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad u = u(x,t). \tag{1.5}$$

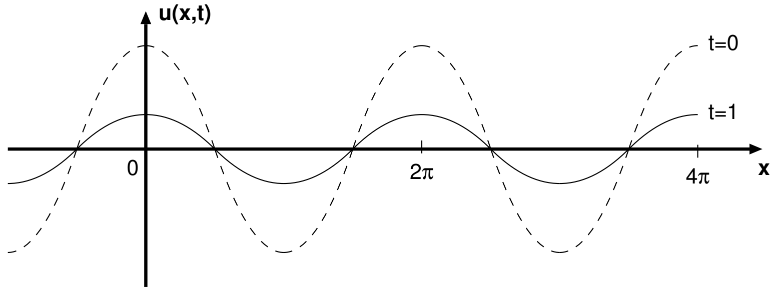
There are many different solutions of this PDE, dependent on the choice of initial conditions and boundary conditions. An example of one such solution is

$$u(x,t) = \exp(-t)\cos(x). \tag{1.6}$$

It can be quickly verified that this solution satisfies (1.5), since

$$\frac{\partial u}{\partial t}(x,t) = -\exp(-t)\cos(x), \quad \text{and} \quad \frac{\partial^2 u}{\partial x^2}(x,t) = -\exp(-t)\cos(x). \tag{1.7}$$

Graphically, this solution is given as follows.



This solution is dissipative (*i.e.* its amplitude decays over time). As we will see later, diffusion is a typical property of parabolic PDEs.

The heat equation (1.5) is often used in models of temperature diffusion, where this equation gets its name, but also in modelling other diffusive processes, such as the spread of pollutants in the atmosphere.

Example 2. The 1D Wave Equation (Hyperbolic Prototype) The 1-dimensional wave equation is given by

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad u = u(x,t). \tag{1.8}$$

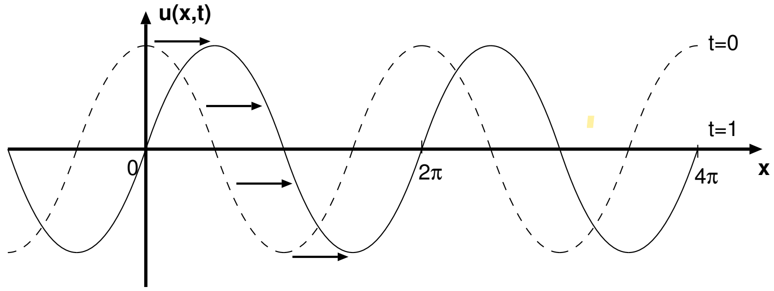
Again, the solution of this DE depends on the choice of initial conditions and boundary conditions. However, in an unbounded domain it can be easily shown (exercise) that any solution of the form

$$u(x,t) = f(x \pm t) \tag{1.9}$$

satisfies the PDE (1.5). We depict this solution below for the choice $u(x,t) = f(x-t) = \cos(x-t)$.

u	$f(x \pm t)$
$\frac{\partial}{\partial t}$	$\pm f'(x \pm t)$
$\frac{\partial}{\partial x}$	$f'(x \pm t)$
$\frac{\partial^2}{\partial t^2}$	$f'(x \pm t)$
$\frac{\partial^2}{\partial x^2}$	$f'(x \pm t)$

$u_{tt} - u_{xx} = 0$
$f' - f' = 0$



Unlike solutions of the heat equation (1.5), solutions of the wave equation (1.8) do not dissipate. This property is typical of hyperbolic PDEs.

The wave equation (1.5) models most types of waves, including water waves and electromagnetic waves.

Example 3. The 2D Laplace Equation (Elliptic Prototype) The 2-dimensional Laplace equation is given by

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \qquad u = u(x,y). \tag{1.10}$$

edge of domain

Normally we consider this equation on a bounded domain $\Omega \in \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$. The solution of this DE then depends on boundary conditions, specified along Γ.

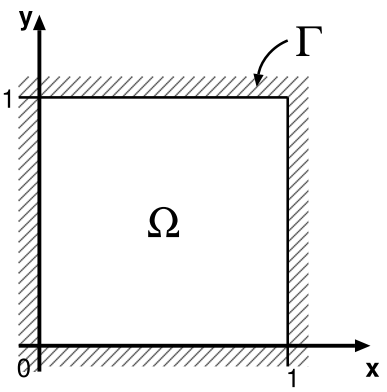
The inhomogeneous form of the Laplace equation is known as the *Poisson equation* and is defined as

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y). \tag{1.11}$$

Consider the following boundary value problem (BVP):

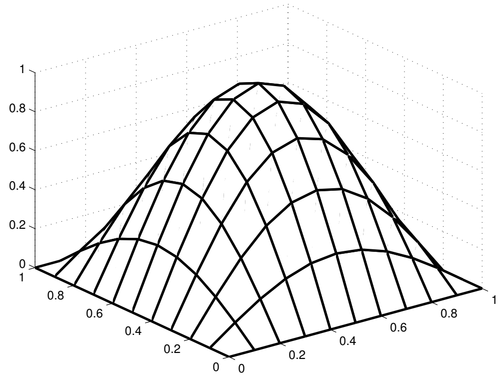
$$\text{BVP} \begin{cases} \Omega = (0,1) \times (0,1), \\ u(x,y) = 0 \text{ on } \Gamma = \partial\Omega, \\ u_{xx} + u_{yy} = -2\pi^2 \sin(\pi x)\sin(\pi y) \quad \text{in } \Omega. \end{cases} \tag{1.12}$$

The domain is the unit square, depicted in the following figure.



It can be shown that the unique solution of this BVP is

$$u(x,y) = \sin(\pi x)\sin(\pi y). \tag{1.13}$$

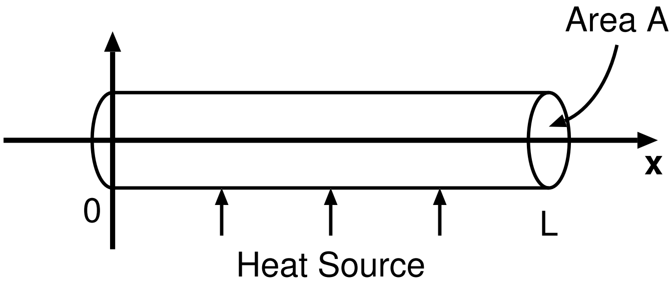


We normally say that a boundary condition of the form $u(x, y) = g$ on Γ is of *Dirichlet type*. Correspondingly, a boundary condition of the form $u(x, y) = 0$ on Γ is of *homogeneous Dirichlet type*.

The Poisson equation (1.11) is used in modelling many physical phenomenon, including elastic membranes, electric potential and steady state temperature distributions.

1.1.3 Derivation of the Heat Equation

In order to motivate the study of the heat equation (1.5), we provide a derivation of this equation from physical principles.



Consider a metal rod of length L and cross-sectional area A that is aligned parallel to the x-axis (see figure). Assuming that the temperature gradient in the y and z directions is negligible, the temperature profile in the rod will be given by $u(x, t)$ for $0 \leq x \leq L$. Then starting with an initial temperature profile $g(x) = u(x, 0)$, we heat the rod in accordance with a heat source function $h(x)$. We then pose the following question:

What is the temperature profile $u(x,t)$ for $0 \leq x \leq L$ and $t \geq 0$?

Appropriate to the subject of this text, we will answer this question by deriving a PDE model describing the physics behind the problem.

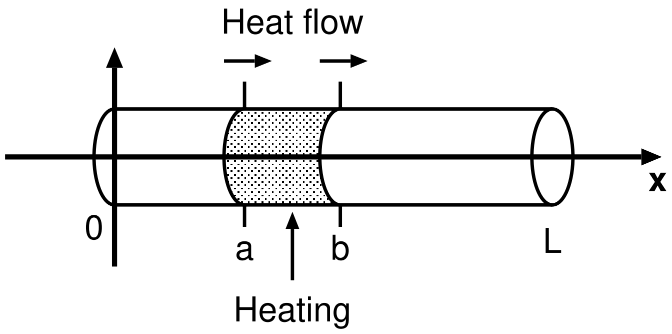
The physical quantities we are interested in are given in the following table.

Quantity	Physical Meaning	Dimensions	Unit
$u(x,t)$	Temperature	temperature	K
$h(x)$	Heat source	$\frac{\text{energy}}{\text{time} \cdot \text{volume}}$	$Js^{-1}m^{-3}$
$\rho(x)$	Mass density	$\frac{\text{mass}}{\text{volume}}$	gm^{-3}
c	Specific heat	$\frac{\text{energy}}{\text{mass} \cdot \text{temperature}}$	$Jg^{-1}K^{-1}$
~~$J(x)$~~ $J(x,t)$	Energy flux	$\frac{\text{energy}}{\text{area} \cdot \text{time}}$	$Jm^{-2}s^{-1}$
$q(x,t)$	Energy density	$\frac{\text{energy}}{\text{volume}}$	Jm^{-3}
κ	Thermal conductivity	$\frac{\text{energy}}{\text{time} \cdot \text{length} \cdot \text{temperature}}$	$Js^{-1}m^{-1}K^{-1}$

Now, in order to derive a physical relationship between these variables, we must rely on physical principles. We consider a volume $\Omega = \{x \mid x \in [a,b]\}$ along the rod (where a and b are constants). Then conservation of energy states that

$$\begin{aligned} \frac{d}{dt}(\text{total energy in } \Omega) &= (\text{energy flux through boundary of } \Omega) \\ &+ (\text{total heat energy added per unit time to } \Omega) \end{aligned} \tag{1.14}$$

(see figure).



Using the physical quantities to express this conservation principle, we write (1.14) as

$$\frac{d}{dt}\left(\int_a^b q(x,t)A\,dx\right) = J(a,t)A - J(b,t)A + \int_a^b h(x)A\,dx. \tag{1.15}$$

Then by the fundamental theorem of calculus, we have

$$-\int_a^b \frac{\partial J}{\partial x}dx = J(a,t) - J(b,t). \tag{1.16}$$

Upon substituting (1.16) into (1.15) and bringing the derivative on the left-hand side inside the integral,[1] we obtain

$$\int_a^b \left(\frac{\partial q}{\partial t} + \frac{\partial J}{\partial x} - h(x)\right)dx = 0. \tag{1.17}$$

Since this equation must hold for all intervals $[a,b] \in [0,L]$, we must have

$$\frac{\partial q}{\partial t} + \frac{\partial J}{\partial x} - h(x) = 0. \tag{1.18}$$

In order to proceed, we require a mechanism to describe the energy flux in terms of the temperature gradient. We use Fourier's law of heat conduction, which states that heat flows from a warm body to a cold body at a rate proportional to the temperature gradient between the two bodies. Mathematically, we write

positive when flowing from hot to cold

OUTFLOW ←

$$J(x,t) = -\kappa\frac{\partial u}{\partial x}(x,t). \tag{1.19}$$

Further, the energy density q can be written in terms of other physical quantities as

$$q(x,t) = c\rho(x)u(x,t). \tag{1.20}$$

$$c\rho\frac{\partial u}{\partial t} - K\frac{\partial^2 u}{\partial x^2} - h = 0$$

On substituting (1.19) and (1.20) into (1.18) we obtain

$$\frac{\partial u}{\partial t} - \frac{K}{c\rho}\frac{\partial^2 u}{\partial x^2} - \frac{h}{c\rho} = 0$$

$$\frac{\partial u}{\partial t} - \frac{K}{c\rho}\frac{\partial^2 u}{\partial x^2} = \frac{h}{c\rho}$$

$$\frac{\partial u}{\partial t} - \frac{K}{\rho c}\frac{\partial^2 u}{\partial x^2} = \frac{h}{\rho c}. \tag{1.21}$$

We now define the thermal diffusivity D and temperature source f by

$$D(x) = \frac{K}{c\rho(x)}, \quad \text{and} \quad f(x) = \frac{h(x)}{c\rho(x)} \tag{1.22}$$

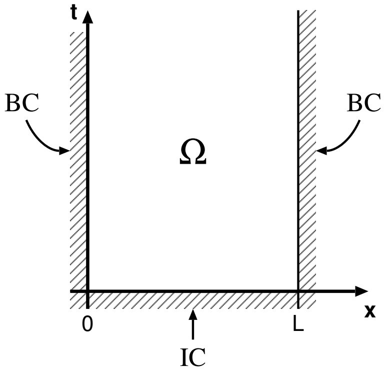
[1] We assume sufficient continuity in order to switch the derivative and the integral operations.

and hence obtain the final form of the heat equation:

$$\frac{\partial u}{\partial t} - D(x)\frac{\partial^2 u}{\partial x^2} = f(x). \tag{1.23}$$

Note that if the mass density of the rod is constant then it follows from (1.22) that $D(x)$ is constant. Further, in the case of $f(x) = 0$ (*i.e.* no external heating) and $D(x) = 1$, this problem simply reduces to the homogeneous heat equation (1.5).

The domain of the heating problem is given by all points satisfying $0 \leq x \leq L$ and $t \geq 0$. Boundary conditions must be imposed at $x = 0$ and $x = L$ and initial conditions imposed at $t = 0$ (see figure).



Using the heat equation (1.23), we can now formulate the heating problem as an initial value boundary value problem (IVBVP), as follows. For simplicity we assume the rod is kept at a constant temperature at either end (constant boundary conditions) and has a constant temperature initially. The IVBVP then reads

$$\begin{cases} \Omega = (0, L) \times (0, \infty) & \text{(domain)}, \\ u(L, t) = u_L & \text{(BC)}, \\ u(x, 0) = C & \text{(IC)}, \\ u_t - Du_{xx} = f(x) & \text{(PDE)}. \end{cases} \tag{1.24}$$

??

Existence and uniqueness for the solution of (1.24) can be shown (see, for example, AM353 course notes), although this result is beyond the scope of these notes. We shall return to the topic of parabolic PDEs in section 1.4.

1.2 Hyperbolic PDEs with Two Independent Variables

We now take a closer look at hyperbolic PDEs. In the following section, we derive the wave equation (1.8) using the linear advection equation and derive a general solution of the wave equation for given initial conditions. We conclude with some general comments about hyperbolic PDEs.

u	$f(x-at)$
$\frac{\partial u}{\partial t}$	$-af'(x-at)$
$\frac{\partial u}{\partial x}$	$f'(x-at)$

1.2.1 The Linear Advection Equation

The world's simplest first-order PDE is the *linear advection equation*, defined by

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \qquad u = u(x,t). \qquad \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \tag{1.25}$$

$$-af' + af' = 0$$

We claim that

$$u(x,t) = f(x-at) \tag{1.26}$$

is a general solution of (1.25) for any function $f(s)$. This result can be proven very easily, as follows.

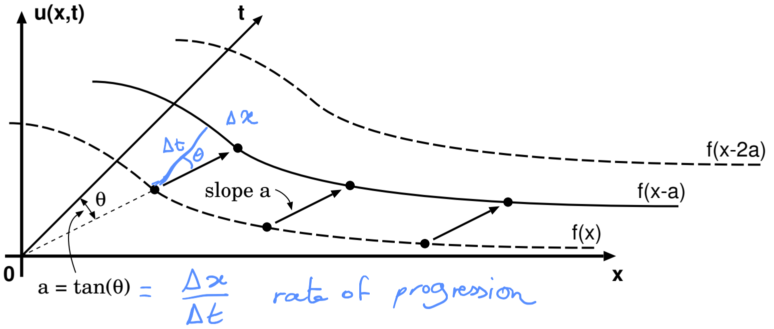
Proof. Define $s = x - at$. Then, by chain rule

$$\frac{\partial u}{\partial x} = \frac{df}{ds}\frac{\partial s}{\partial x} = (-a)\frac{df}{ds}, \quad \text{and} \quad \frac{\partial u}{\partial t} = \frac{df}{ds}\frac{\partial s}{\partial t} = \frac{df}{ds}. \tag{1.27}$$

The result follows upon substituting (1.27) into (1.25). □

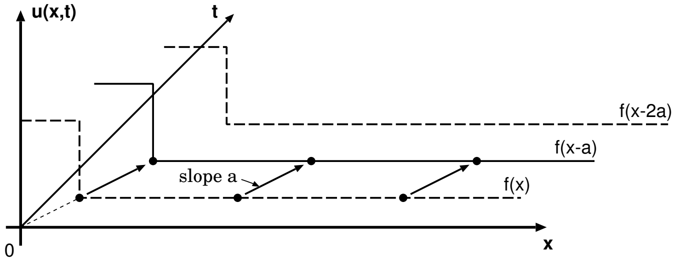
?

The general solution (1.26) has an intuitive meaning. Namely, any given profile is simply advected forward with advection speed a (or backwards, depending on the sign of a) without modifying the initial profile. We depict this effect in the following figure.



Comments: i) Note that for a specific choice of s in $f(s)$, such as $s = 0$, we can track the coordinates of this point as it moves through the domain. For example, if $t = 0$ then $s = x - at = 0$ implies $x = 0$. If $t = 1$ then $s = x - at = 0$ implies $x = a$.

ii) In some generalized sense (to be discussed in more detail later), a discontinuous function $f(s)$ can also be a solution of this PDE, even though the derivative may not be defined at one or more points. In this case, as the profile is advected, the discontinuous profile will remain discontinuous.



iii) The linear advection equation is unidirectional, *i.e.* it defines a preferred direction depending on the sign of a. Namely, if a is positive (negative), profiles will be advected in the positive (negative) x direction. $s = x - at$

1.2.2 The Wave Equation

We now show how the linear wave equation can be derived from the linear advection equation presented in the previous section.

Before proceeding, we must introduct the concept of a *linear differential operator*. In general, a *linear differential operator* L maps a function f to a linear combination of f and its partial derivatives.

Consider the linear advection equation (1.25). This equation can be rewritten using a differential operator L applied to u, *i.e.*

$$L_1 u = 0, \quad \text{where} \quad L_1 = \frac{\partial}{\partial t} + a\frac{\partial}{\partial x}. \tag{1.28}$$

We can also define a second differential operator L_2, which has equal advection speed but opposite direction, and define the PDE

$$L_2 u = 0, \quad \text{where} \quad L_2 = \frac{\partial}{\partial t} - a\frac{\partial}{\partial x}. \tag{1.29}$$

Note that the differential equations $L_1 u = 0$ and $L_2 u = 0$ have general solutions $f(x-at)$ and $f(x+at)$, respectively, for any choice of f. Consider the equation defined by

$$L_1 L_2 u = 0. \tag{1.30}$$

On rewriting (1.30) as a PDE, we obtain

$$\left(\frac{\partial}{\partial t} - a\frac{\partial}{\partial x}\right)\left(\frac{\partial}{\partial t} + a\frac{\partial}{\partial x}\right) u = 0, \tag{1.31}$$

$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}$

$\frac{\partial}{\partial t}$ | $\frac{\partial^2 u}{\partial t^2} + a\frac{\partial^2 u}{\partial x \partial t}$ $\qquad$ $\frac{\partial^2 u}{\partial t^2} + a\frac{\partial^2 u}{\partial x\partial t} - a\left(\frac{\partial^2 u}{\partial x\partial t} + a\frac{\partial^2 u}{\partial x^2}\right)$

$\frac{\partial}{\partial x}$ | $\frac{\partial^2 u}{\partial t\partial x} + a\frac{\partial^2 u}{\partial x^2}$ $\qquad$ $\frac{\partial^2 u}{\partial t^2} - a^2\frac{\partial^2 u}{\partial x^2}$

which simplifies to

$$\left(\frac{\partial^2}{\partial t^2} - a^2\frac{\partial^2}{\partial x^2}\right)u = 0. \tag{1.32}$$

Then (1.32) is exactly the one dimensional wave equation with constant speed a. Note that L_1 and L_2 commute: $L_1L_2u = 0 \iff L_2L_1u = 0$. This means that both $f(x-at)$ and $f(x+at)$ are solutions of (1.30). It can also be shown directly by using the chain rule (exercise) that

implied by commutativity

$$u(x,t) = \underbrace{f(x-at)}_{\substack{\text{right moving wave}\\ \text{with speed } a}} + \underbrace{g(x+at),}_{\substack{\text{left moving wave}\\ \text{with speed } a}} \tag{1.33}$$

is a general solution of (1.32) for arbitrary functions f and g.

1.2.3 d'Alembert's Solution for the Wave Equation IVP

The unbounded initial value problem (IVP) for the one-dimensional wave equation (1.32) with $a \equiv 1$ on an unbounded spatial domain is given by

$$IVP \begin{cases} \Omega : t \in (0,\infty), x \in (-\infty,\infty), \\ u(x,0) = \phi_0(x), u_t(x,0) = \phi_1(x), \\ u_{tt} - u_{xx} = 0. \end{cases} \tag{1.34}$$

We already know that a general solution of this problem is given by (1.33), and so aim to derive an expression for the functions $f(s)$ and $g(s)$ in terms of the initial conditions $\phi_0(x)$ and $\phi_1(x)$.

We substitute the initial conditions (1.34) into (1.33), evaluated at $t = 0$ (or $s = x$), obtaining

$$f(x) + g(x) = \phi_0(x), \quad \text{and} \quad -\frac{df}{dx} + \frac{dg}{dx} = \phi_1(x). \tag{1.35}$$

Integrating the second expression in (1.35) then yields

$$-(f(x) - f(c)) + (g(x) - g(c)) = \int_c^x \phi_1(\tilde{x})d\tilde{x}. \tag{1.36}$$

It follows that (1.35) and (1.36) can be combined to obtain

$$2f(x) = \phi_0(x) - \int_c^x \phi_1(\tilde{x})d\tilde{x} + f(c) - g(c), \tag{1.37}$$

$$2g(x) = \phi_0(x) + \int_c^x \phi_1(\tilde{x})d\tilde{x} - f(c) + g(c). \tag{1.38}$$

$$2f(x-t) = \phi_0(x-t) + \int_c^{x-t} \phi_1(\tilde{x})d\tilde{x} + \cancel{f(c)} - \cancel{g(c)}$$

$$2g(x+t) = \phi(x+t) + \int_c^{x+t} \phi_1(\tilde{x})d\tilde{x} - \cancel{f(c)} + \cancel{g(c)}$$

$$\frac{1}{2}\left(\phi_0(x-t) + \phi(x+t) + \int_{x-t}^{x+t} \phi_1(\tilde{x})d\tilde{x}\right)$$

On substituting (1.37) and (1.38) back into (1.33) and applying a simple identity from calculus, we obtain

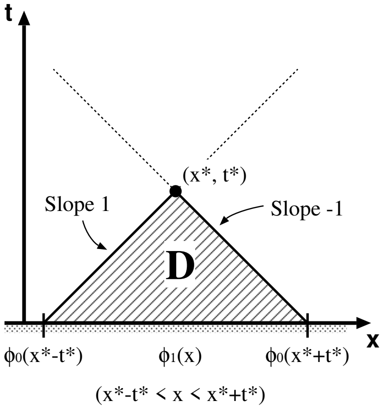
$$u(x,t) = \tfrac{1}{2}\left(\phi_0(x+t) + \phi_0(x-t) + \int_{x-t}^{x+t} \phi_1(\tilde{x})d\tilde{x}\right). \tag{1.39}$$

This equation is known as *d'Alembert's solution of the wave equation.* It can be shown that this solution is the unique solution of (1.34).

Note that the wave equation requires two initial conditions at $t = 0$ to determine a unique solution. However, the heat equation (a parabolic PDE) only requires one initial condition at $t = 0$ (see, for example, eq. (1.24)). ↳ but it also requires one boundary condition

1.2.4 Domain of Influence and Domain of Dependence

We consider the IVP (1.34) and choose some point $(x^*, t^*) \in (-\infty, \infty) \times (0, \infty)$. Then according to d'Alembert's solution (1.39), this point only depends on the value of the functions ϕ_0 and ϕ_1 in the interval $x \in [x^* - t^*, x^* + t^*]$. We can depict the set of points D that influence (x^*, t^*), as follows.

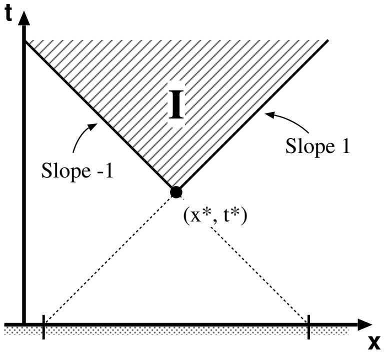


why u & u_t only?
why not u_x, u_{tt}, u_{xt}, or u_{xx}?

??.

The set of points D is called the *domain of dependence of* (x^*, t^*); that is, (x^*, t^*) only depends on the values of u and u_t inside the domain D.

Similarly, we can consider the set I of points that are influenced by the solution at (x^*, t^*). This set is called the *domain of influence of* (x^*, t^*); that is, $(x, t) \in I$ depend on the value of u and u_t at (x^*, t^*). The domain of influence I for some point (x^*, t^*) is depicted as follows.



Comments: i) In general, one can show that both the domain of dependence and the domain of influence for hyperbolic PDEs is finite in space at any given time, *i.e.* along some line $t =$ constant. Hence, one says that hyperbolic PDEs feature propagation of information at a finite speed (which is the *wave speed*).

ii) The solution at (x^*, t^*) in any hyperbolic PDE only depends on the solution at previous times, *i.e.* for $0 < t < t^*$. As a consequence, we can perform time marching as a numerical method (we shall describe this process later).

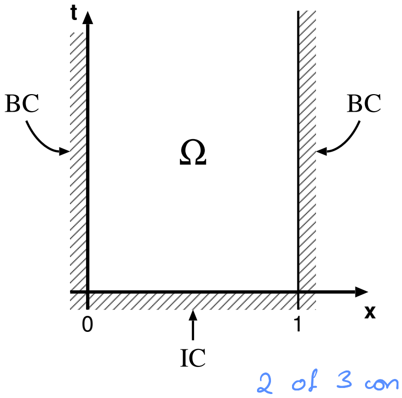
1.2.5 Existence and Uniqueness for the IVBVP

We now consider the wave equation (1.32) with fixed boundaries at $x = a$ and $x = b$ (with $a < b$). One important problem to consider is if we can guarantee existence and uniqueness of the solution, *i.e.* is there exactly one solution which satisfies a given IVBVP?

Consider the general IVBVP for the wave equation with fixed boundaries:

$$IVBVP \left\{ \begin{array}{l} \Omega : (x,t) \in (a,b) \times (0,+\infty), \\ u(x,0) = \phi_0(x), u_t(x,0) = \phi_1(x), \\ u(a,t) = g_1(t), u(b,t) = g_2(t), \\ u_{tt} - u_{xx} = f(x,t). \quad \leftarrow \text{inhomogeneous} \end{array} \right. \tag{1.40}$$

The domain of this problem can be illustrated as follows.



well-posed problems:

- 1) the solution exists
- 2) the solution is unique
- 3) the solution's behavior changes continuously with initial and boundary conditions

2 of 3 conditions for well-posedness

Although we do not prove it in this text, existence and uniqueness of the solution $u(x,t)$ of the IVBVP (1.40) can be shown for "well-behaved" functions f, ϕ_0, ϕ_1, g_1 and g_2. We refer the reader instead to Evans [2002].

One particular example of an initial value boundary value problem (IVBVP) for the wave equation with fixed boundaries is given by

$$IVBVP \begin{cases} \Omega : (x,t) \in (0,1) \times (0,+\infty), \\ u(x,0) = \sin(x), u_t(x,0) = 0, \\ u(0,t) = 0, u(1,t) = 0, \\ u_{tt} - u_{xx} = 0. \end{cases} \tag{1.41}$$

This IVBVP has the unique solution

$$u(x,t) = \sin(\pi x)\cos(\pi t). \tag{1.42}$$

The IVBVP (1.41) describes certain physical phenomena, such as a "standing" sound wave in a closed tube or on a plucked string with both ends fixed.

1.3 Elliptic PDEs with Two Independent Variables

We now turn our attention to elliptic PDEs. In the following section, we will illustrate the domain of dependence and domain of influence of an elliptic PDE and will show that discontinuous boundary conditions are smoothed out within the domain.

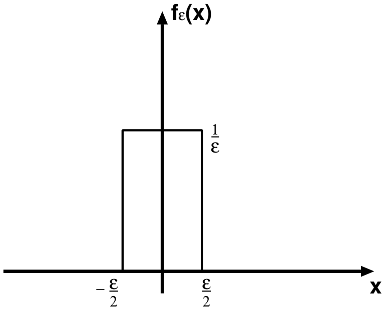
1.3.1 The Dirac Delta

The Dirac delta is a type of “generalized function”, used in mathematical modelling and differential equations to represent a physical impulse in the system. In this case, it will be useful in analyzing the behaviour of elliptic PDEs.

Consider the discontinuous function $f_\epsilon(x)$, defined by

$$f_\epsilon(x) = \begin{cases} 1/\epsilon & |x| < \epsilon/2, \\ 0 & \text{otherwise.} \end{cases} \tag{1.43}$$

This function is depicted in the following plot.



One can clearly see that for any value of $\epsilon > 0$, f_ϵ satisfies

$$\int_{-\infty}^{\infty} f_\epsilon(x) = 1. \tag{1.44}$$

This motivates the following definition

Definition 1.3 *The* ***Dirac delta****, denoted $\delta(x)$ is defined as*

$$\delta(x) = \lim_{\epsilon \to 0} f_\epsilon(x), \tag{1.45}$$

where $f_\epsilon(x)$ is defined by (1.43).

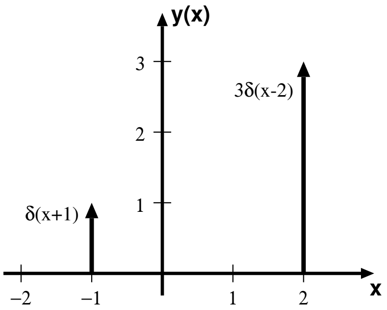
The Dirac delta technically is not a function, but instead fits into a category of operators known as “generalized functions”. It has the following properties:

$$\text{i)} \qquad \delta(x) = \begin{cases} 0 & x \neq 0, \\ +\infty & x = 0. \end{cases} \tag{1.46}$$

$$\text{ii)} \qquad \int_{-\infty}^{+\infty} \delta(x)dx = 1, \tag{1.47}$$

$$\text{iii)} \qquad \int_{-\infty}^{+\infty} f(x)\delta(x)dx = f(0). \tag{1.48}$$

In plotting the Dirac delta, we will generally use arrows, as in the following figure.



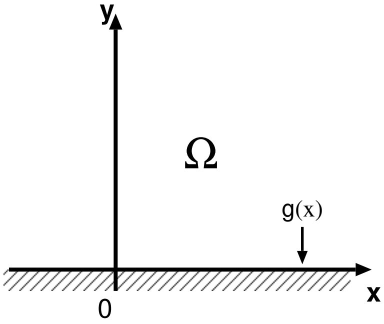
1.3.2 Domain of Influence

We now examine the domain of influence and the domain of dependence for elliptic PDEs, on recalling the results obtained for hyperbolic PDEs in section 1.2.

Consider the Elliptic BVP in the half-plane given by

$$BVP \begin{cases} \Omega : (x, y) \in (-\infty, \infty) \times (0, \infty), \\ u(x, 0) = g(x), \ \text{BC} \\ u_{xx} + u_{yy} = 0. \end{cases} \tag{1.49}$$

The domain of this problem is depicted in the following figure.



$$g(a) = \int_{-\infty}^{\infty} \delta(x-a)\, g(s)\, ds$$

when $x = a$, $g(s)$ is sampled

$$\therefore\; g(x) = \int_{-\infty}^{\infty} \delta(x-s)\, g(s)\, ds$$

In order to understand the domain of influence for elliptic PDEs, we need to study the influence of one point of the boundary on the solution in the entire domain. We will require the Dirac delta (1.45) to give us the desired results.

Usually definite integrals result in a final value. Here it results in a function $g(x)$

Observe that using (1.48), $g(x)$ can be written as

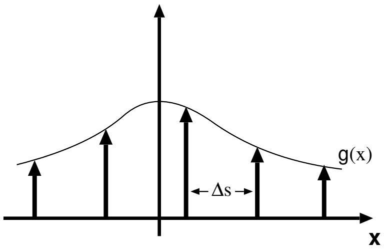
Visually, it looks as though $\delta(x)$ is scanning $g(s)$

$$g(x) = \int_{-\infty}^{\infty} \delta(x-s)g(s)ds. \tag{1.50}$$

looks like a change of variable

This operation is known as the *convolution* of $g(x)$ and $\delta(x)$. Intuitively, in this form we can describe the Dirac delta as "picking out" the value of $g(s)$ when $x = s$. On discretizing this integral as a Riemann sum, we obtain

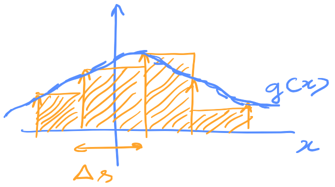
$$g(x) \approx \sum_{i=-\infty}^{\infty} g(s_i)\delta(x-s_i)\Delta s. \tag{1.51}$$



If we integrate (1.51) over x and assume we can switch the order of integration and summation, we obtain

$$\int_{-\infty}^{\infty} g(x)dx \approx \sum_{i=-\infty}^{\infty} g(s_i)\Delta s \cancelto{1}{\int_{-\infty}^{\infty} \delta(x-s_i)dx} = \sum_{i=-\infty}^{\infty} g(s_i)\Delta s, \tag{1.52}$$

as we would expect if we were to directly discretize the integral on the left-hand side of this expression.



$u_t - u_{xx} = 0$

$u_{tt} - u_{xx} = 0$

$u_{yy} + u_{xx} = 0$

In attempting to understand the domain of influence, it is sufficient to look at the influence of one delta function since we can obtain an approximate solution for general $g(x)$ by the principle of superposition and the discretization (1.51). Hence, we let $g(x) = \delta(x)$ and look for a solution of the BVP (1.49) with this choice of boundary condition. We claim that the solution of the BVP with $g(x) = \delta(x)$ is exactly

$\boxed{u_{yy} + u_{xx} = 0}$

$$u(x, y) = \frac{1}{\pi} \frac{y}{x^2 + y^2}. \tag{1.53}$$

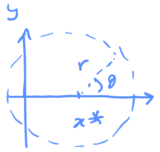
This result can be verified with some effort:

First, we show that (1.53) satisfies the PDE. Upon differentiating (1.53), we obtain

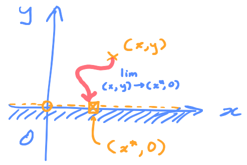
$$u_{yy} = \frac{2y(y^2 - 3x^2)}{\pi(x^2 + y^2)^3}, \quad \text{and} \quad u_{xx} = -\frac{2y(y^2 - 3x^2)}{\pi(x^2 + y^2)^3}, \tag{1.54}$$

which clearly satisfies $u_{yy} + u_{xx} = 0$.

Second, we must verify that the boundary conditions are satisfied by this solution. Consider an arbitrary point $(x^*, 0)$ on the boundary of Ω. If $x^* \neq 0$, it follows by inspection that



$$\lim_{(x,y)\to(x^*,0)} u(x, y) = 0. \tag{1.55}$$



$x = r\cos\theta + x^*$

$y = r\sin\theta$

$(x,y) \to (x^*,0)$

If $x^* = 0$ then we can apply L'Hôpital's rule to obtain

$$\lim_{(x,y)\to(0,0)} u(x, y) = \infty. \tag{1.56}$$

$\frac{y}{x^2+y^2}$

$= \frac{r\sin\theta}{r^2\cos^2\theta + 2rx^*\cos\theta + x^{*2} + r^2\sin^2\theta}$

Also, using integration along any slice $y =$ constant, it can be shown (exercise) that

$$\lim_{y\to0^+} \int_{-\infty}^{\infty} u(x, y)dx = 1. \tag{1.57}$$

$\frac{1}{\pi}\left[\tan^{-1}\left(\frac{x}{y}\right)\Big|_{-\infty}^{\infty}\right] = \frac{1}{\pi}\left[\frac{\pi}{2} + \frac{\pi}{2}\right] = 1$

$\lim_{r\to\emptyset}$ $\div r$

$\frac{\sin\theta}{r\cos^2\theta + 2x^*\cos\theta + \frac{x^{*2}}{r} + r\sin^2\theta}$

$\to \emptyset$

if $x^* \neq \emptyset$

if $x^* = \emptyset$

then

∞

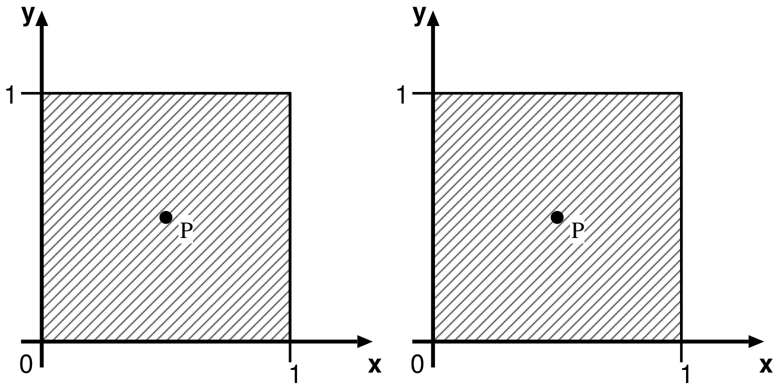
We can thus conclude that (1.53) satisfies the BVP (1.49). The domain of influence of a single point on the boundary is then given by the set of points in Ω where $u(x, y) > 0$. By inspection of (1.53), we note that all points in the domain have this property, and hence conclude that the domain of influence of a single point on the boundary is the entire domain Ω. Since this result implies that all points in the domain instantaneously communicate with one another, one says that elliptic problems have "infinite propagation speed."

If we pose the BVP on the lower half plane, we similarly obtain

$$u(x, y) = \frac{1}{\pi} \frac{|y|}{x^2 + y^2}. \tag{1.58}$$

This means that the Dirac delta function also influences points in the lower half plane.

Thus, we have found that, for an elliptic PDE, any point influences all other points, and hence we cannot use time-marching strategies to solve elliptic problems, *i.e.* we must solve for the whole problem at once. For an elliptic PDE on the unit square, the domain of dependence and domain of influence for a point P are illustrated in the following figure.

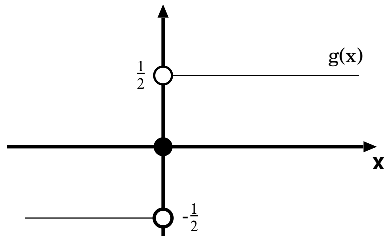


1.3.3 Discontinuous Boundary Conditions

We now examine the effect of discontinuous boundary conditions on the solution of the elliptical BVP (1.49). Consider the boundary condition given by

$$g(x) = \begin{cases} \frac{1}{2} & x > 0, \\ 0 & x = 0, \\ -\frac{1}{2} & x < 0. \end{cases} \tag{1.59}$$

This function is depicted in the following plot.



It can be shown that

$$u(x,y) = \frac{1}{\pi}\arctan(\cancel{y/x}) \tag{1.60}$$

x/y

satisfies the PDE (exercise) and satisfies the boundary condition $u(x,0) = g(x)$ in the limit as $y \to 0$. Further, it is easy to see that $u(x,y)$ is continuous in the domain Ω. Using this example, we hypothesize that, for linear elliptic PDEs, if $g(x)$ has a finite number of discontinuities then they are smoothed out immediately in the domain.

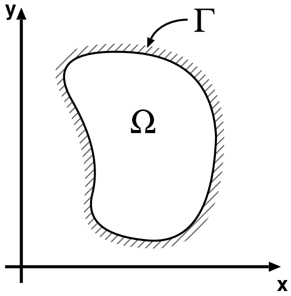
1.3.4 Existence and Uniqueness

We now briefly discuss existence and uniqueness of solutions of the general Poisson BVP. The general Poisson BVP in two variables with boundary conditions of Dirichlet type is given by

inhomogeneous
Laplace equation

$$BVP \begin{cases} \Omega \subset \mathbb{R}^2, \Omega \text{ bounded} \\ u(x,y) = g(x,y) \text{ on } \Gamma = \partial\Omega, \\ u_{xx} + u_{yy} = f(x,y) \text{ in } \Omega. \end{cases} \tag{1.61}$$

A general domain Ω is depicted as follows.



It can be shown that for well-behaved functions f, g and boundary shape the BVP has a unique solution. We refer the reader to Evans [2002] for a proof of this result.

1.4 Parabolic PDEs with Two Independent Variables

We now turn our attention to parabolic PDEs, in particular the heat equation (1.5). In the following section, we will examine the domain of dependence and domain of influence of this equation and examine the effect of discontinuous boundary conditions on the solution.

The homogeneous initial value problem (IVP) for the heat equation on an unbounded spatial domain is given by

$$IVP \begin{cases} \Omega : (x,t) \in \mathbb{R} \times (0,\infty) \\ u(x,0) = g(x), \\ u_t - u_{xx} = 0. \end{cases} \tag{1.62}$$

Note that if the initial condition identically vanishes, *i.e.* $g(x) = 0$, then the unique solution is exactly $u(x,t) = 0$.

1.4.1 Domain of Influence and Domain of Dependence

As with the ~~Poisson~~ wave equation (see section 1.3.2), we now examine the domain of influence and the domain of dependence of a point $(x,t) \in \Omega$ by choosing the boundary condition to be given by a Dirac delta function, *i.e.* $g(x) = \delta(x)$. We claim that the solution of the IVP (1.62) is then given by

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \exp\left(\frac{-x^2}{4t}\right). \tag{1.63}$$

Gaussian

This result can be verified with some effort:

First, upon differentiating (1.63), we obtain

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} = \frac{x^2 - 2t}{8\sqrt{\pi t^3}} \exp\left(\frac{-x^2}{4t}\right). \tag{1.64}$$

Hence, $u(x,t)$ satisfies $u_t - u_{xx} = 0$ on Ω.

Second, we must verify that the boundary conditions are satisfied by this solution. Consider an arbitrary point $(x^*, 0)$ on the boundary of Ω. If $x^* \neq 0$, we can apply L'Hôpital's rule to obtain

$$\lim_{t \to 0^+} u(x^*, t) = 0. \tag{1.65}$$

If $x^* = 0$ then the exponential term is exactly 1 in the limit, and so the limit satisfies

$$\lim_{t \to 0^+} u(0, t) = \infty. \tag{1.66}$$

It now remains to show that

$$\lim_{t \to 0^+} \int_{-\infty}^{+\infty} u(x,t) dx = 1. \tag{1.67}$$

This result is non-trivial, but can be shown after some tedious calculus. We leave the details of this calculation to the reader.

We conclude that (1.63) satisfies the IVP (1.62). As with elliptic problems, the domain of influence of a single point on the boundary is then given by the set of points in Ω where $u(x,y) > 0$. By inspection of (1.63), we note that all points in the domain have this property, and hence conclude that the domain of influence of a *single point on the boundary is the entire domain* Ω. It follows that, as with the Poisson equation, the heat equation exhibits an *infinite propagation speed*.

However, unlike the elliptic BVP (1.49), we note that the heat equation is not *time-reversible*. Consider the initial value problem on the lower half plane

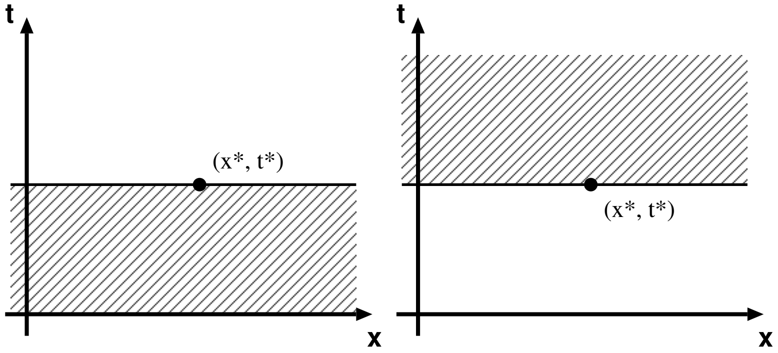
$$IVP \begin{cases} \Omega : (x,t) \in \mathbb{R} \times (-\infty, 0) \\ u(x,0) = g(x), \\ u_t - u_{xx} = 0. \end{cases} \tag{1.68}$$

If we try the function obtained by making the substitution $t \to (-t)$ in (1.63), *i.e.* the function given by

$$\tilde{u}(x,t) = \frac{1}{\sqrt{4\pi(-t)}} \exp\left(-\frac{x^2}{4(-t)}\right), \tag{1.69}$$

we find that $\tilde{u}(x,t)$ does not satisfy (1.68). In fact, the function $\tilde{u}$ instead satisfies the PDE $u_t + u_{xx} = 0$.

Since the heat equation is not time-reversable, the domain of influence for any point is the whole spatial domain for all future times. Similarly, we can obtain that the domain of dependence for any point is the whole spatial domain for all past times. Note that this result allows us to perform time marching for parabolic problems. The figures below depict the domain of dependence (left) and the domain of influence (right) of (x^*, t^*).



1.4.2 Discontinuous Initial Conditions

We now examine the effect of discontinuous initial conditions on the heat equation IVP (1.62). Consider a discontinuous initial condition $g(x)$ defined by

$$g(x) = \begin{cases} \frac{1}{2} & x > 0, \\ 0 & x = 0, \\ -\frac{1}{2} & x < 0. \end{cases} \tag{1.70}$$

It can be shown that

$$u(x,t) = \tfrac{1}{2}\mathrm{erf}\left(\frac{x}{\sqrt{4t}}\right), \quad \text{with} \quad \mathrm{erf}(w) = \frac{2}{\sqrt{\pi}}\int_0^w \exp(-z^2)dz, \tag{1.71}$$

satisfies the PDE (exercise) and satisfies the boundary condition $u(x,0) = g(x)$ in the limit as $t \to 0^+$. Further, it is easy to see that $u(x,t)$ is continuous in the domain Ω. Using this example as motivation, we hypothesize that the IVP for a linear parabolic PDE, with boundary condition $g(x)$ possessing a finite number of discontinuities, has a smooth solution away from the boundary. That is to say, discontinuities in the initial state are smoothed out immediately.

1.5 Linear Second Order PDEs with Three Independent Variables

We now briefly discuss linear second order PDEs with three independent variables.

There exist certain "canonical cases" where classification of linear second order PDEs with three independent variables is straightforward. Consider the canonical form of a linear second order PDE with three independent variables:

$$\lambda_1(x,y,t)u_{tt} + \lambda_2(x,y,t)u_{xx} + \lambda_3(x,y,t)u_{yy} = W(u, u_x, u_y, u_t, x, y, t), \tag{1.72}$$

where W is linear in u, u_x, u_y and u_t. We claim, without proof, that any second-order linear PDE can be transformed into canonical form (1.72) by eliminating cross derivatives with a change of variables. The canonical form of the PDE leads to the following definition:

Definition 1.4 *A second-order linear PDE in canonical form (1.72) is said to be*

i) ***elliptic*** *if and only if all λ_i are the same sign,*

ii) ***hyperbolic*** *if and only if one λ_i has the opposite sign of the other λ_i,*

iii) ***parabolic*** *if and only if one λ_i is zero and the other λ_i have the same sign.*

Note: This classification also holds for more than three independent variables; for example, the PDE

$$u_{tt} - u_{xx} - u_{yy} - u_{zz} = 0 \tag{1.73}$$

is hyperbolic and

$$u_t - u_{xx} - u_{yy} - u_{zz} = 0 \tag{1.74}$$

is parabolic.

1.5.1 A Note About Vector Calculus Notation

PDEs in higher dimensions are often expressed using vector calculus operators, for which we use the notation of the table below. For example, the Laplace equation $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0$ can be written as

$$\triangle u = 0,$$

where $\triangle$ is called the Laplace operator or the *Laplacian*.

In the following table we assume that u denotes a scalar function and $\vec{v}$ denotes a vector with components $(v_1, v_2) \in \mathbb{R}^2$ or $(v_1, v_2, v_3) \in \mathbb{R}^3$.

Operator	Notation	2D Definition	3D Definition
Gradient of u	∇u or $\operatorname{grad}(u)$	$\nabla u = (u_x, u_y)$	$\nabla u = (u_x, u_y, u_z)$
Divergence of $\vec{v}$	$\nabla \cdot \vec{v}$ or $\operatorname{div}(\vec{v})$	$\nabla \cdot \vec{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y}$	$\nabla \cdot \vec{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$
Laplacian of u	$\triangle u$ or $\nabla^2 u$	$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$	$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$

elliptical
2nd order pde

Note that the Laplacian of u, $\nabla^2 u$, can be expressed in terms of the gradient and divergence as

$$\nabla^2 u = \nabla \cdot (\nabla u). \tag{1.75}$$

CHAPTER 2

Finite Difference Methods

In this chapter we focus on *finite difference (FD) methods*, perhaps the most straightforward numerical approach for solving PDEs. We begin in section 2.1 by introducing FD methods for elliptic PDEs and setting up much of the groundwork for further study of FD methods. In sections 2.2 and 2.3 we introduce FD methods for time-dependent problems, focusing primarily on the theory behind numerical methods for hyperbolic and parabolic PDEs. Section 2.4 is a wrap-up of the study of time-dependent problems and focuses on extending the convergence theory for elliptic schemes to hyperbolic and parabolic FD methods.

2.1 Finite Difference Methods for Elliptic PDEs

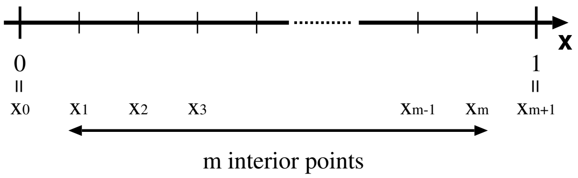
In this section we focus on the finite difference methods for elliptic PDEs, with emphasis placed on the Poisson equation in 1D and 2D. Of particular interest in the theory of numerical methods is *convergence*, *i.e.* in this section we will attempt to explain when a given FD method provides a solution that converges to a solution of the associated PDE problem, in the limit of infinite resolution.

2.1.1 1D Elliptic Model Problem

We start our study of FD methods for elliptic PDEs by considering a simple model problem in 1D:

$$\text{BVP}\begin{cases}\Omega = \{x : x \in (0,1)\},\\ u(0) = \alpha,\ u(1) = \beta,\\ u''(x) = f(x).\end{cases} \tag{2.1}$$

Numerical solutions of this BVP can be obtained by discretizing the domain Ω using $m+2$ distinct points $x_0, x_1, \ldots, x_{m+1}$, yielding $m+1$ intervals. The *boundary points* (at 0 and 1) then consist of x_0 and x_{m+1} and *interior points* consist of x_1 through x_m, inclusive. For simplicity we choose the x_i to be equidistant, *i.e.* $x_i - x_{i-1} = \Delta x = h$ for all $i = 1, \ldots, m+1$.



We denote the exact solution of this BVP by $u(x)$. The values of the solution at each x_i are then given by $u(x_i) = u_i$ for $i = 0, \ldots, m+1$. We denote the derivatives of the solution at each x_i by $u'(x_i) = u_i'$ and similarly for higher derivatives; for example, $u''(x_i) = u_i''$, etc. We then use a *central difference formula*[1] to discretize $u''(x)$, given by

$$u''(x_i) \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}. \tag{2.2}$$

This choice of discretization scheme follows from expanding $u(x)$ in a Taylor series at $i+1$ and $i-1$ according to

$$u_{i+1} = u_i + u_i'h + \tfrac{1}{2}u_i''h^2 + \tfrac{1}{6}u_i'''h^3 + \cdots, \tag{2.3}$$

and

$$u_{i-1} = u_i - u_i'h + \tfrac{1}{2}u_i''h^2 - \tfrac{1}{6}u_i'''h^3 + \cdots. \tag{2.4}$$

Summing these two series then yields

$$u_{i+1} + u_{i-1} = 2u_i + u_i''h^2 + \tfrac{1}{12}u_i''''h^4 + O(h^5), \tag{2.5}$$

which, upon rearranging, gives

$$u_i'' = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - \tfrac{1}{12}u_i^{(4)}h^2 + O(h^3). \tag{2.6}$$

[1] Note that other choices of discretization are possible here.

Thus, to second order in h, we recover (2.2).

We now define a numerical approximation v_i to the exact solution u_i. Using the discretization (2.2), we define the approximate solution v_i associated with the one-dimensional BVP (2.1) to be the unique discrete function v_i satisfying

$$\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} = f_i, \qquad i = 1, \ldots, m, \tag{2.7}$$

where f_i is defined by $f(x_i) = f_i$, and subject to the boundary conditions given by the exact boundary conditions for the BVP, *i.e.*

$$v_0 = \alpha, \quad \text{and} \quad v_{m+1} = \beta. \tag{2.8}$$

Matrix Form of the BVP

We can write (2.7) in matrix form as

$$A^h V^h = F^h, \tag{2.9}$$

where A^h is a matrix and V^h and F^h are vectors. Here V^h is referred to as a *grid function*, *i.e.* a discrete representation or approximation of a continuous function on a grid. Here, the h is a generic superscript that denotes a grid function. On using (2.7) and (2.8), we see that A^h, V^h and F^h are given by

$$A^h = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & & 0 \\ 1 & -2 & 1 & & \\ 0 & 1 & -2 & \ddots & \\ & & \ddots & \ddots & 1 \\ 0 & & & 1 & -2 \end{pmatrix}, \quad V^h = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_m \end{pmatrix}, \quad F^h = \begin{pmatrix} f_1 - \alpha\frac{1}{h^2} \\ f_2 \\ f_3 \\ \vdots \\ f_m - \beta\frac{1}{h^2} \end{pmatrix}. \tag{2.10}$$

Note that A^h is a sparse matrix, *i.e.* the majority of its entries are zero. As a consequence, the linear system (2.9) is generally easy to solve. We will also make use of the grid function U^h, which denotes the vector consisting of the exact solution $u_i = u(x_i)$ evaluated at grid nodes. The matrix form (2.9) is a generic form for linear FD methods applied to elliptic problems. We will make use of this form later for the 2D elliptic PDE.

Actual Error and Convergence

Since the numerical approximation (2.2) is different from the exact formula (2.6), V^h merely provides an approximation of the exact solution U^h. As a result, we are interested in the deviation of V^h from the exact solution U^h.

Definition 2.1 *The **actual error** E^h is*

$$E^h = U^h - V^h, \tag{2.11}$$

where U^h is the grid function associated with the exact solution $u(x)$ and V^h is the approximate solution, obtained by solving (2.9). The elements of E^h are denoted e_i and are given by $e_i = u_i - v_i$.

For any FD method that solves the BVP (2.1), we require *convergence*. Namely, as $h \to 0$, we want $E^h \to 0$ as well, *i.e.* as the distance between grid points becomes infinitesimally small, the actual error introduced due to the numerical scheme goes to zero. For the choice of discretization (2.2), we know from (2.6) that

$$u_i'' = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + O(h^2), \tag{2.12}$$

and so can hope (or expect) that using some norm,[2] the error satisfies

$$\|E^h\| = O(h^2). \tag{2.13}$$

This result will be provided rigorously later on in this chapter.

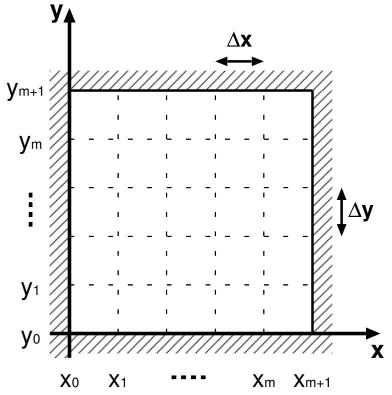
2.1.2 2D Elliptic Model Problem

The 2D elliptic model BVP can be formulated as follows:

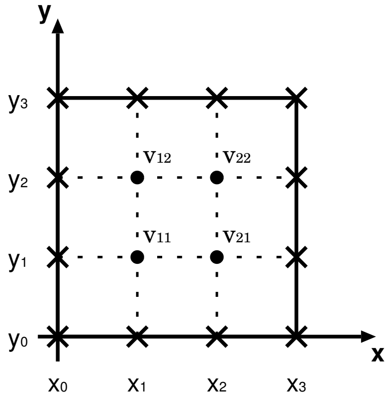
$$\text{BVP} \begin{cases} \Omega : (x,y) \in (0,1)^2, \\ u(x,y) = g \text{ on } \Gamma = \partial\Omega, \\ u_{xx} + u_{yy} = f(x,y) \quad \text{in } \Omega. \end{cases} \tag{2.14}$$

We discretize Ω into square regions of side length $h = \Delta x = \Delta y$, obtaining $m+2$, $m+1$ intervals, and m interior points in each direction. This discretization is then depicted in the following figure.

[2] See Appendix A.



We give an example of this discretization, in the case of $m = 2$. In the following image, interior nodes are depicted as circles and boundary nodes are given as crosses.



We denote the exact solution of the BVP by $u(x, y)$. The associated grid function is then given by $u_{i,j}$, which satisfies

$$u_{i,j} = u(x_i, y_j). \tag{2.15}$$

The source function $f(x_i, y_j)$ can also be evaluated at grid points, leading us to define

$$f_{i,j} = f(x_i, y_j). \tag{2.16}$$

We now require a discretization of the PDE. On recalling the 1D discretization (2.2), we discretize the partial derivatives u_{xx} and u_{yy} as

$$u_{xx} \approx \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2}, \tag{2.17}$$

and

$$u_{yy} \approx \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta y^2}. \tag{2.18}$$

Hence, using the fact that $h = \Delta x = \Delta y$, our discretization of the PDE operator is given by[3]

$$u_{xx} + u_{yy} \approx \frac{u_{i+1,j} + u_{i-1,j} - 4u_{i,j} + u_{i,j+1} + u_{i,j-1}}{h^2}. \tag{2.19}$$

This leads us to define the numerical approximation $v_{i,j}$ as the solution of the system of equations

$$\frac{v_{i+1,j} + v_{i-1,j} - 4v_{i,j} + v_{i,j+1} + v_{i,j-1}}{h^2} = f_{i,j}, \tag{2.20}$$

subject to the boundary conditions

$$v_{i,j} = g(x_i, y_j) \quad \text{for } i, j = 0 \text{ or } m. \tag{2.21}$$

Matrix Form of the BVP

We now formulate this problem in matrix form (2.9). The solution vector V^h consists of all interior points (the unknowns), ordered in any desired manner. For simplicity, we will choose our ordering to be *row-lexicographic ordering*, *i.e.* we group the unknowns by row and first vary over the first index, and then over the second index. For example, in the case of $m = 2$, we obtain

$$V^h = \begin{bmatrix} v_{11} \\ v_{21} \\ v_{12} \\ v_{22} \end{bmatrix}. \tag{2.22}$$

It follows from the system of equations (2.20) that A^h can be written in block-diagonal form as

$$A^h = \frac{1}{h^2} \left[\begin{array}{c|c|c|c} T & I & 0 & 0 \\ \hline I & T & \ddots & 0 \\ \hline 0 & \ddots & \ddots & I \\ \hline 0 & 0 & I & T \end{array} \right], \tag{2.23}$$

[3] Again, it should be emphasized that other choices for the discretization are possible.

where T and I are $m \times m$ matrices given by

$$T = \begin{bmatrix} -4 & 1 & & 0 \\ 1 & -4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -4 \end{bmatrix}, \quad \text{and} \quad I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}. \tag{2.24}$$

The boundary conditions are absorbed into F^h (it is left as an exercise for the reader to give the resulting form of F^h).

Recall that the actual error E^h (see definition 2.1) is given by $E^h = U^h - V^h$. Since the discretization we have used is second order in x and y, we can again hope that the error satisfies $\|E^h\|_2 = O(h^2)$.

2.1.3 Convergence Theory

Having introduced two numerical methods for solving the elliptic BVPs, we have sufficient material to study the convergence of these numerical methods. We will demonstrate the convergence theory in the simplest case, namely for the elliptic BVP in 1D, since the theory can be easily generalized.

Consider the 1D BVP (2.1), with approximate solution v_i given by the system

$$\frac{v_{i+1} - 2v_i + v_{i-1}}{h^2} = f_i, \;\; v_0 = \alpha, \;\; v_{m+1} = \beta \quad \Longleftrightarrow \quad A^h U^h = F^h. \tag{2.25}$$

Definition 2.2 *The* ***truncation error*** T^h *is the error obtained when plugging the exact solution* $u(x)$ *into the discrete formula.*

If we use the general matrix form (2.9), the truncation error assumes the form

$$T^h = A^h V^h - F^h. \tag{2.26}$$

Example: In the 1D case, the truncation error is given by

$$T_i = \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} - f_i. \tag{2.27}$$

On plugging (2.6) into (2.27), we have

$$T_i = u_i'' + \tfrac{1}{12} h^2 u_i^{(4)} - f_i + O(h^3), \tag{2.28}$$

which, on recalling that $u_i'' - f_i = 0$ at any i, simplifies to

$$T_i = \tfrac{1}{12}h^2 u_i^{(4)} + O(h^3) = O(h^2). \tag{2.29}$$

In fact, it can be shown that

$$T_i = \tfrac{1}{12}h^2 u^{(4)}(\gamma(x_i)), \tag{2.30}$$

for some $\gamma(x_i) \in [x_{i-1}, x_{i+1}]$. We note that this result can be derived using Taylor's theorem with a remainder (exercise).

One of the major components of convergence theory is the concept of consistency of a numerical method, defined as follows:

Definition 2.3 *A numerical method $A^h V^h = F^h$ is **consistent** with the linear elliptic PDE $Lu = f$ if*

$$\lim_{h\to 0} T_i = 0. \tag{2.31}$$

*Further, we say that it is **consistent with order** q ($q \in \mathbb{N}$) if $T_i = O(h^q)$.*

Note that it follows from (2.29) that the discretization (2.25) is consistent with order $q = 2$. Further, from $T_i = O(h^2)$, we can deduce that $\|T^h\|_2 = O(h^2)$, as follows: If we assume that $u^{(4)}(x)$ is continuous, and let

$$c_T = \max_{x\in[0,1]} |u^{(4)}(x)|, \tag{2.32}$$

on $[0, 1]$, it follows from (2.30) that

$$T_i \leq \tfrac{1}{12}h^2 c_T. \tag{2.33}$$

Then, on taking the 2-norm, we have

$$\begin{aligned}
\|T^h\|_2 &= \sqrt{h}\sqrt{\sum_{i=1}^{m}(T_i)^2}, \\
&\leq \sqrt{h}\sqrt{m(\tfrac{1}{12}h^2 c_T)^2}, \\
&\leq \sqrt{h}\sqrt{(m+1)(\tfrac{1}{12}h^2 c_T)^2},
\end{aligned}$$

but since $h = 1/(m+1)$, we obtain

$$\|T^h\|_2 \leq \tfrac{1}{12}h^2 c_T. \tag{2.34}$$

The Error Equation

We will now derive an important relation between T^h and E^h. On taking the difference between (2.26) and (2.9), we obtain

$$A^h(U^h - V^h) = T^h, \tag{2.35}$$

which by (2.11) can be written as

$$A^h E^h = T^h. \tag{2.36}$$

On multiplying both sides by $(A^h)^{-1}$ and taking the p-norm of this expression, we obtain

$$\|E^h\|_p = \|(A^h)^{-1}T^h\|_p \leq \|(A^h)^{-1}\|_p \|T^h\|_p. \tag{2.37}$$

If we know that $\|T^h\|_p$ is at least $O(h)$, *i.e.* the numerical method is consistent, convergence then follows if there exists a c independent of h so that $\|(A^h)^{-1}\|_p \leq c$. This result motivates the following definition.

Definition 2.4 *A numerical method $A^hV^h = F^h$ is* ***stable*** *for the linear elliptic PDE $Lu = f$ if and only if there exists c_s so that*

$$\|(A^h)^{-1}\|_p \leq c_s, \tag{2.38}$$

with c_s independent of h.

Lax Convergence Theorem for Elliptic PDEs

As stated previously, one can see that convergence of a numerical method quickly follows from definition 2.3, definition 2.4 and (2.37). This result is the foundation of the so-called *Lax Convergence Theorem for Elliptic PDEs*, which we now state.

Theorem 2.1 (Lax Convergence Theorem) *Consider the linear numerical method $A^hV^h = F^h$ for the linear elliptic PDE $Lu = f$. If the method is consistent with order q in the p-norm,*

$$\|T^h\|_p = O(h^q), \tag{2.39}$$

and stable in the p-norm,

$$\|(A^h)^{-1}\|_p \leq c_s, \tag{2.40}$$

then the method is convergent with order q,

$$\|E^h\|_p = O(h^q). \tag{2.41}$$

Proof: The desired result follows immediately from (2.37), under the assumption of stability and using the definition of convergence. □

Notes: i) This theorem can be extended as follows: Consider a linear method that is consistent with order q. Then the method is stable if and only if it converges with order q, *i.e.* it can be shown that convergence with order q and stability are equivalent (this result is known as the Lax Equivalence Theorem).

ii) Note that the actual error E^h converges with the same order as the truncation error T^h. Hence, rather than trying to estimate $U^h - V^h$ directly, we can instead use the order of the truncation error to obtain the order of convergence for the actual error.

2-Norm Convergence for 1D Elliptic Problems

We now use the Lax convergence theorem to show convergence of the discretization (2.7) and (2.8) for the 1D elliptic BVP in the 2-norm. Recall that we have already shown in (2.34) that

$$\|T^h\|_2 \leq \tfrac{1}{12}h^2 c_T, \tag{2.42}$$

where

$$c_T = \max_{x\in[0,1]} |u^{(4)}(x)| = \max_{x\in[0,1]} |f''(x)|, \tag{2.43}$$

i.e. that our discretization (2.7) of the 1D elliptic PDE BVP (2.1) is consistent, provided $f''(x)$ is continuous on $[0, 1]$.

In order to show stability, and hence demonstrate convergence, we need to find an upper bound on $\|(A^h)^{-1}\|_2$, where A^h is given by (2.10). In order to proceed, we require two important results from linear algebra:

R_1) First, recall that if A^h is symmetric then it follows that $(A^h)^{-1}$ is symmetric as well, *i.e.*

$$A^h = (A^h)^T \implies (A^h)^{-1} = ((A^h)^{-1})^T.$$

Hence, it follows by property P_1 in section A.3 that

$$\|(A^h)^{-1}\|_2 = \rho((A^h)^{-1}). \tag{2.44}$$

R_2) Second, if $A^h \in \mathbb{R}^{m\times m}$ is invertible and has eigenvalues λ_1, λ_2, …, λ_m, it follows that $(A^h)^{-1}$ has eigenvalues λ_1^{-1}, λ_2^{-1}, …, λ_m^{-1}. The proof of this result is straightforward: If λ is an eigenvalue of an invertible matrix A with associated eigenvector $\vec{v}$, then

$$A\vec{v} = \lambda\vec{v} \implies \frac{1}{\lambda}\vec{v} = A^{-1}\vec{v}. \tag{2.45}$$

This implies that λ^{-1} is an eigenvalue of A^{-1}.

Thus, using results R_1) and R_2), we have

$$\|(A^h)^{-1}\|_2 = \rho((A^h)^{-1}) = \max_{1\le i\le m}\left|\frac{1}{\lambda_i}\right| = \left(\min_{1\le i\le m}|\lambda_i|\right)^{-1}, \tag{2.46}$$

i.e. the 2-norm of $(A^h)^{-1}$ is given by the inverse of the smallest eigenvalue of A^h.

It can be shown that for A^h given by (2.10), the eigenvalues are (exercise)

$$\lambda_k = \frac{2}{h^2}(\cos(k\pi h) - 1), \qquad k = 1, \ldots, m, \tag{2.47}$$

where $h = (m+1)^{-1}$. By inspection, the smallest eigenvalue occurs when $k = 1$. Using Taylor's theorem for the $\cos(\pi h)$ term, we obtain

$$\cos(\pi h) = 1 - \tfrac{1}{2}\pi^2 h^2 + \tfrac{1}{24}\pi^4 h^4 \cos(\pi\xi), \tag{2.48}$$

where $\xi \in [0, h]$. Hence, substituting (2.48) into (2.47), we have

$$\lambda_1 \approx \frac{2}{h^2}(-\tfrac{1}{2}\pi^2 h^2 + \tfrac{1}{24}\pi^4 h^4 \cos(\pi\xi)) = -\pi^2 + \tfrac{1}{12}\pi^4 h^2 \cos(\pi\xi). \tag{2.49}$$

Clearly, for h sufficiently small, λ_1 satisfies $\frac{1}{2}\pi^2 \le |\lambda_1| \approx \pi^2$, independent of h. We conclude that

$$\frac{1}{\pi^2} \approx \|(A^h)^{-1}\|_2 = \frac{1}{|\lambda_1|} \le \frac{2}{\pi^2}, \tag{2.50}$$

and so the method is stable. Thus, by the Lax convergence theorem (Theorem 2.1), we have that the method is convergent with order 2 in the 2-norm, *i.e.*

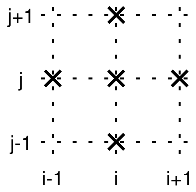
$$\|E^h\|_2 = O(h^2). \tag{2.51}$$

Note that a somewhat more informative error bound can be obtained by substituting (2.34) and (2.50) into (2.37), which gives

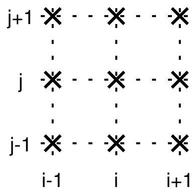
$$\|E^h\|_2 \le \frac{h^2}{6\pi^2}c_T + O(h^3). \tag{2.52}$$

Notes: i) Convergence with order 2 can also be proven for this example for the 1-norm and ∞-norm and further for the 2D BVP (2.14). The method discussed in this section can also be applied to other numerical FD methods.

ii) In order to obtain a higher order of convergence, we must develop a more accurate discretization by using more grid points. For example, the simple 2D discretization used in (2.20) is known as a 5-point central difference discretization as depicted in the following figure.



Other discretizations can be developed, such as the 9-point weighted central difference discretization depicted below.



iii) Finally, finite difference formulas can be derived for non-uniform grid spacings, for example grids which may have more points in regions of rapid change than in regions of slow change.

2.2 FD Methods for Hyperbolic PDEs

We now consider FD methods for hyperbolic PDEs, with emphasis placed on the advection equation in 1D. We give a detailed analysis of six FD methods, focusing on convergence and stability properties and error-propagation. We conclude this section with a discussion of FD methods for the wave equation and extensions of the methods to higher dimensions.

Several concepts required for the study of hyperbolic FD methods generalize directly from elliptic FD methods. In particular, the *actual error*, given by definition 2.1, and the truncation error, given by definition 2.2, are both defined in the same manner as for elliptic FD methods. We will require these two concepts in the analysis in this section.

2.2.1 FD Methods for the 1D Linear Advection Equation

Recall the linear advection equation in 1D, given by

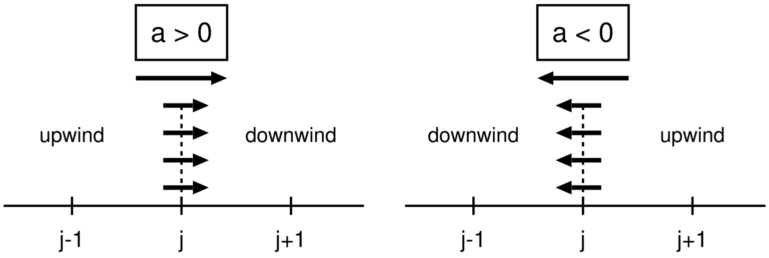
$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \qquad u = u(x,t), \tag{2.53}$$

with general solution

$$u(x,t) = f(x - at), \tag{2.54}$$

which, assuming $a > 0$, describes a right-travelling wave (also see (1.25) and (1.26)).

We now consider three discretizations for the spatial derivative in (2.53). For each grid point j, we can divide the domain into two regions depending on the direction of the "wind", *i.e.* the direction that the PDE carries the state variable in as time advances. For the advection equation with $a > 0$, the "wind" carries the solution from left to right, and so we describe all nodes i that satisfy $i < j$ as "upwind" and all nodes i that satisfy $i > j$ as "downwind" (see graphic below).



Spatial Discretization	Formula
$\boxed{\text{C}}$ Central Difference	$\left.\frac{\partial u}{\partial x}\right\|_j = \frac{u_{j+1} - u_{j-1}}{2\Delta x} + O(\Delta x^2)$
$\boxed{\text{D}}$ Downwind $(a > 0)$	$\left.\frac{\partial u}{\partial x}\right\|_j = \frac{u_{j+1} - u_j}{\Delta x} + O(\Delta x)$
$\boxed{\text{U}}$ Upwind $(a > 0)$	$\left.\frac{\partial u}{\partial x}\right\|_j = \frac{u_j - u_{j-1}}{\Delta x} + O(\Delta x)$

As a first step, we can obtain a *pseudo-discretization* for (2.53) by only discretizing the spatial component of the PDE and leaving the time derivative untouched. In this course, we will not following this approach, but will rather treat the discretization of space and time derivatives in an integrated way.

For example, using the central discretization, we obtain

$$\boxed{\text{C}} \qquad \frac{dv_j}{dt} + a\frac{v_{j+1} - v_{j-1}}{2\Delta x} = 0. \tag{2.55}$$

If, instead, we use an upwind discretization, we obtain

$$\boxed{\text{U}} \qquad \frac{dv_j}{dt} + a\frac{v_j - v_{j-1}}{\Delta x} = 0. \tag{2.56}$$

This pseudo-discretization leads to a set of coupled ODEs in $v_j(t)$, the state variables at each point, which can then be integrated using any standard ODE integration technique. This approach is called the *method of lines*.

Our second step is to add a temporal discretization, for which many options are available. We distinguish between *explicit* schemes and *implicit schemes*. An explicit scheme computes each unknown at time level $n+1$ using the known state of the system at times n, $n-1$, $n-2$, etc. Implicit schemes also use the state of the system at time level $n+1$. Hence, implicit schemes lead to a system of linear equations which must be solved in each timestep. In general, explicit schemes are more memory efficient and computationally cheaper per timestep, whereas implicit schemes are more stable (they allow for larger timesteps).

For simplicity, we focus on three common temporal discretizations:

Temporal Discretization	Formula
$\boxed{\text{FE}}$ Forward Euler (Explicit)	$\dfrac{v_j^{n+1} - v_j^n}{\Delta t} + a \left.\dfrac{\partial u}{\partial x}\right\|_j^n = 0$
$\boxed{\text{BE}}$ Backward Euler (Implicit)	$\dfrac{v_j^{n+1} - v_j^n}{\Delta t} + a \left.\dfrac{\partial u}{\partial x}\right\|_j^{n+1} = 0$
$\boxed{\text{CN}}$ Crank-Nicolson (Implicit)	$\dfrac{v_j^{n+1} - v_j^n}{\Delta t} + \dfrac{a}{2}\left(\left.\dfrac{\partial u}{\partial x}\right\|_j^{n+1} + \left.\dfrac{\partial u}{\partial x}\right\|_j^n\right) = 0$

Here, $\left.\frac{\partial u}{\partial x}\right|_j^n$ denotes the discretized spatial derivative evaluated in node j at time step n. We now present three common numerical methods constructed in this manner.

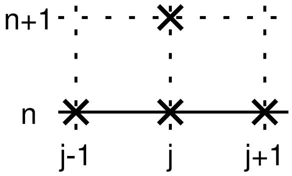
Forward Central Scheme. This scheme uses the spatial central difference formula and forward Euler time discretization. It is written as

$$\boxed{\text{FC}} \qquad \frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} = 0. \tag{2.57}$$

The truncation error of this scheme is

$$T_j^n = O(\Delta t) + O(\Delta x^2). \tag{2.58}$$

The forward central scheme induces a *stencil* on the grid, *i.e.* a set of points that are used in evaluating v_j^{n+1}, as follows.



Although the FC scheme is explicit and hence computationally cheap per timestep, it is also unstable: regardless of our choice of timestep Δt, this method will lead to uncontrolled oscillations that will cause solutions to blow up. This will be explained in detail in section 2.2.2.

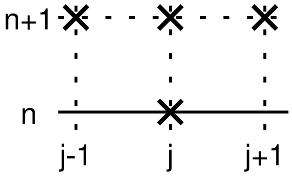
Backward Central Scheme. In order to stabilize the FC scheme, we instead apply a backward time discretization, and hence obtain

$$\boxed{\text{BC}} \qquad \frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_{j+1}^{n+1} - v_{j-1}^{n+1}}{2\Delta x} = 0. \tag{2.59}$$

The truncation error of this scheme is again

$$T_j^n = O(\Delta t) + O(\Delta x^2). \tag{2.60}$$

The stencil is given as follows.



Since this method is implicit, we can rewrite (2.59) in terms of a linear system that is then solved at every time step. Clearly, this approach requires more work per timestep, but it is unconditionally stable, *i.e.* it is stable regardless of the choice of Δt.

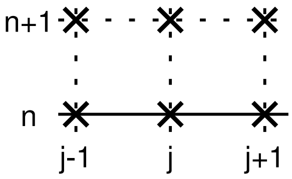
Crank-Nicolson Central Scheme. In order to increase the temporal order of accuracy we can apply the Crank-Nicolson discretization, and hence obtain

$$\boxed{\text{CN}} \qquad \frac{v_j^{n+1} - v_j^n}{\Delta t} + \frac{a}{2}\left(\frac{v_{j+1}^{n+1} - v_{j-1}^{n+1}}{2\Delta x} + \frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x}\right) = 0. \tag{2.61}$$

It can be shown that the truncation error of this scheme is

$$T_j^n = O(\Delta t^2) + O(\Delta x^2), \tag{2.62}$$

and that the stencil is given as follows.



Like the backward central scheme, this method is implicit and unconditionally stable.

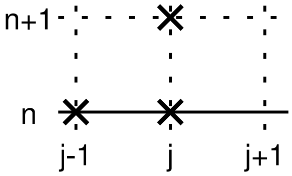
Forward Upwind Scheme. In order to obtain a stable explicit method, we instead apply the spatial upwind discretization, and hence obtain

$$\boxed{\text{FU}} \qquad \frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_j^n - v_{j-1}^n}{\Delta x} = 0 \qquad (a > 0). \tag{2.63}$$

The truncation error of this method is

$$T_j^n = O(\Delta t) + O(\Delta x), \tag{2.64}$$

with stencil given as follows.



This method is explicit and *conditionally stable*, *i.e.* Δt must be chosen sufficiently small in order to guarantee stability.

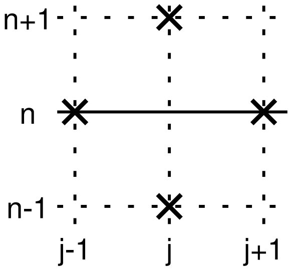
Leapfrog Scheme. We can construct additional temporal discretizations that were not mentioned above, such as the one used by the Leapfrog scheme, which uses a central difference time discretization and central difference space discretization. This scheme is then written as

$$\boxed{\text{LFrog}} \qquad \frac{v_j^{n+1} - v_j^{n-1}}{2\Delta t} + a\frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} = 0. \tag{2.65}$$

It can be shown that the truncation error of this scheme is

$$T_j^n = O(\Delta t^2) + O(\Delta x^2), \tag{2.66}$$

with stencil given as follows.



This method is known as a 3-level scheme, since, when evaluating the state at time $n + 1$, we require knowledge of the state variables v_j at times n and $n - 1$. This is also an example of an explicit high order method. As with other explicit schemes, the leapfrog scheme is conditionally stable.

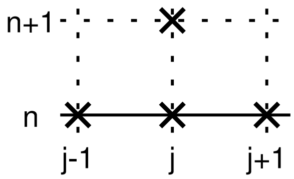
Lax-Wendroff Scheme. The last method we consider is the Lax-Wendroff scheme, given by

$$\boxed{\text{LW}} \qquad \frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x} - \frac{a^2\Delta t}{2}\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2} = 0. \tag{2.67}$$

It can be shown that the truncation error of this method is

$$T_j^n = O(\Delta x^2) + O(\Delta t^2), \tag{2.68}$$

where the stencil is given as follows.



The Lax-Wendroff scheme is a 2-level high order conditionally stable method. It may not be immediately obvious as to why (2.67) is a discretization of (2.53) and so we present the derivation of this scheme:

Recall that the PDE (2.53) allows us to rewrite time derivatives in terms of space derivatives according to

$$u_t = -au_x. \tag{2.69}$$

Hence,

$$u_{tt} = -au_{xt} = (-a)(-a)u_{xx} = a^2 u_{xx}, \tag{2.70}$$

and

$$u_{ttt} = a^2 u_{xxt} = -a^3 u_{xxx}. \tag{2.71}$$

On applying a Taylor series expansion to $u(x, t + \Delta t)$ and using (2.69)-(2.71), we obtain

$$\begin{aligned} u(x, t+\Delta t) &= u(x,t) + u_t \Delta t + u_{tt}\tfrac{1}{2}\Delta t^2 + u_{ttt}\tfrac{1}{6}\Delta t^3 + O(\Delta t^4) \\ &= u(x,t) - au_x\Delta t + a^2 u_{xx}\tfrac{1}{2}\Delta t^2 - a^3 u_{xxx}\tfrac{1}{6}\Delta t^3 + O(\Delta t^4). \end{aligned}$$

Then, on applying the Taylor series of $u(x + \Delta x, t)$ to obtain expressions for u_x and u_{xx} (exercise), we obtain

$$\begin{aligned} u_j^{n+1} &= u_j^n - a\Delta t\left(\frac{u_{j+1} - u_{j-1}}{2\Delta x} - u_{xxx}\tfrac{1}{3}\Delta x^2 + O(\Delta x^3)\right) \\ &\quad + a^2\tfrac{1}{2}\Delta t^2\left(\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} + O(\Delta x^2)\right) \\ &\quad - a^3 u_{xxx}\tfrac{1}{6}\Delta t^3 + O(\Delta t^4). \end{aligned}$$

On taking the difference between this expression and (2.67) for u_j^n, we obtain

$$T_j^n = au_{xxx}\tfrac{1}{3}\Delta x^2 - a^3 u_{xxx}16\Delta t^2 + O(\Delta x^3) + O(\Delta t\Delta x^2) + O(\Delta t^2) = O(\Delta x^2) + O(\Delta t^2). \tag{2.72}$$

The methods introduced in this section only encompass a small fraction of possible FD methods for solving the linear advection equation. It should be emphasized that no single method is the best option for all possible problems. Deciding on the best choice of numerical method for a given problem often requires significant research.

We present a numerical comparison of the five methods introduced in this section in Figures 2.1 and 2.2.

2.2.2 Stability

In this section we examine the conditions for stability of finite difference methods for hyperbolic PDEs.

Consider rewriting the numerical solution obtained using a general FD method as the difference of the exact solution u_j^n and an error term e_j^n, as in

$$\underbrace{v_j^n}_{\text{Numerical solution}} = \underbrace{u_j^n}_{\text{Exact solution}} - \underbrace{e_j^n}_{\text{Actual error at } (x_j,\, t_n)} . \tag{2.73}$$

On substituting (2.73) into the FC scheme (2.57) we obtain

$$\underbrace{\left(\frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}\right)}_{\text{Truncation error } T_j^n}-\underbrace{\left(\frac{e_j^{n+1}-e_j^n}{\Delta t}+a\frac{e_{j+1}^n-e_{j-1}^n}{2\Delta x}\right)}_{\text{Propagation equation for actual error}}=0. \tag{2.74}$$

Note that here the truncation error T_j^n (see Definition 2.2) acts as a source term in the error propagation equation.

We say that a method is *numerically stable* if the actual error e_j^n is bounded as $n \to \infty$. Conversely, a method is *numerically unstable* if the actual error grows without bound (this is a phenomenon known as *numerical instability*).

For simplicity we will only consider the propagation of the error and assume the truncation error is zero. For example, in the FC scheme, this assumption implies the error propagates according to

$$\frac{e_j^{n+1}-e_j^n}{\Delta t}+a\frac{e_{j+1}^n-e_{j-1}^n}{2\Delta x}=0. \tag{2.75}$$

We note that boundedness of e_j^n in (2.75) is a necessary condition for the stability of the FC scheme.

In order to analyze the stability of (2.75), we will decompose the error in terms of Fourier modes or waves with certain wavelengths. This approach is known as the *von Neumann method* of investigating stability.

An error of wave type can be written as

$$e_j^n = \hat{e}^n \exp(ikx_j), \tag{2.76}$$

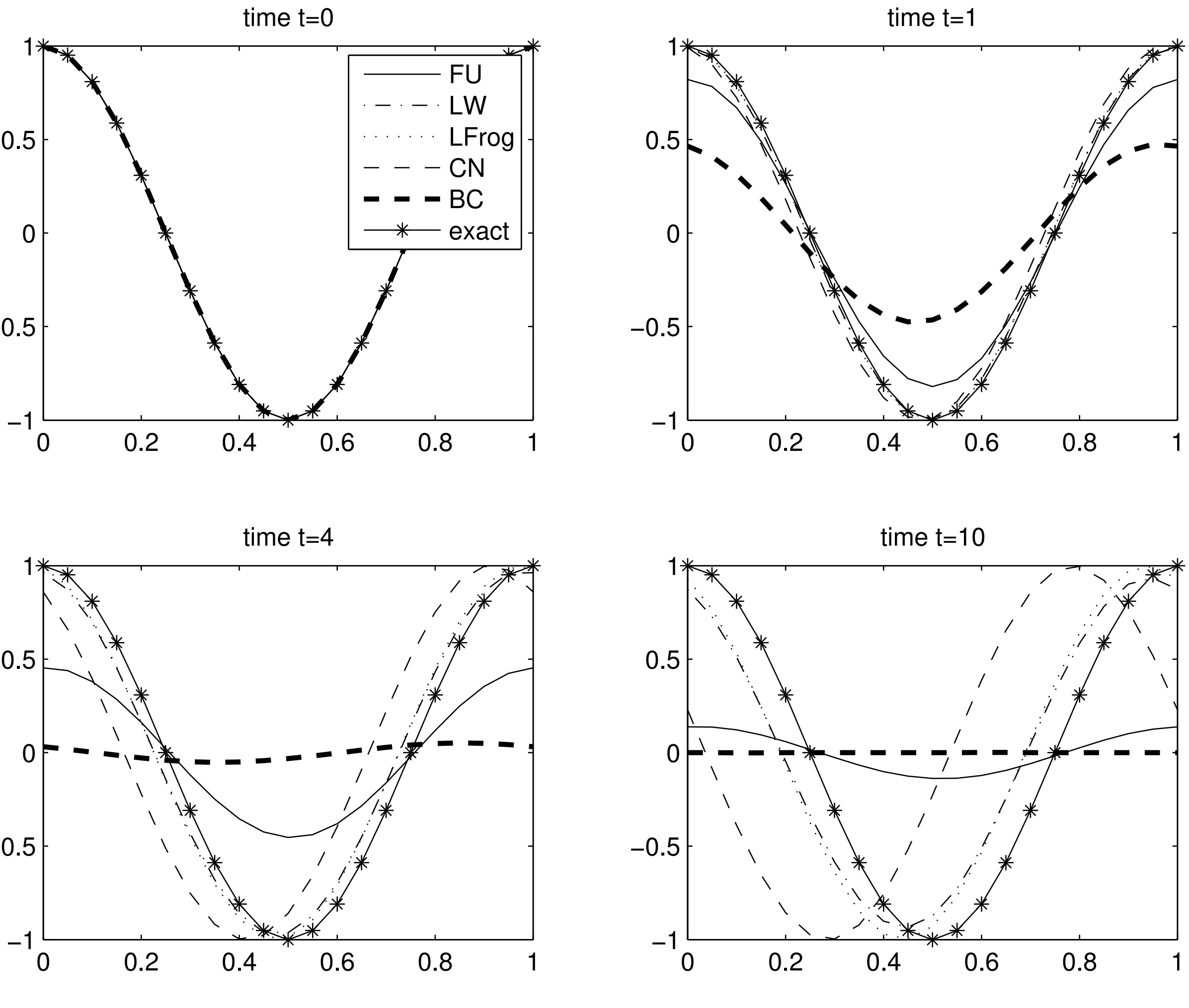


FIGURE 2.1: *A comparison of BC, CN, FU, LFrog and LW numerical schemes for the advection equation with* $a = 1$, *applied to a cosine wave with periodic boundary conditions. FU and BC are very dissipative, and CN and BC are very dispersive (as will be explained later).*

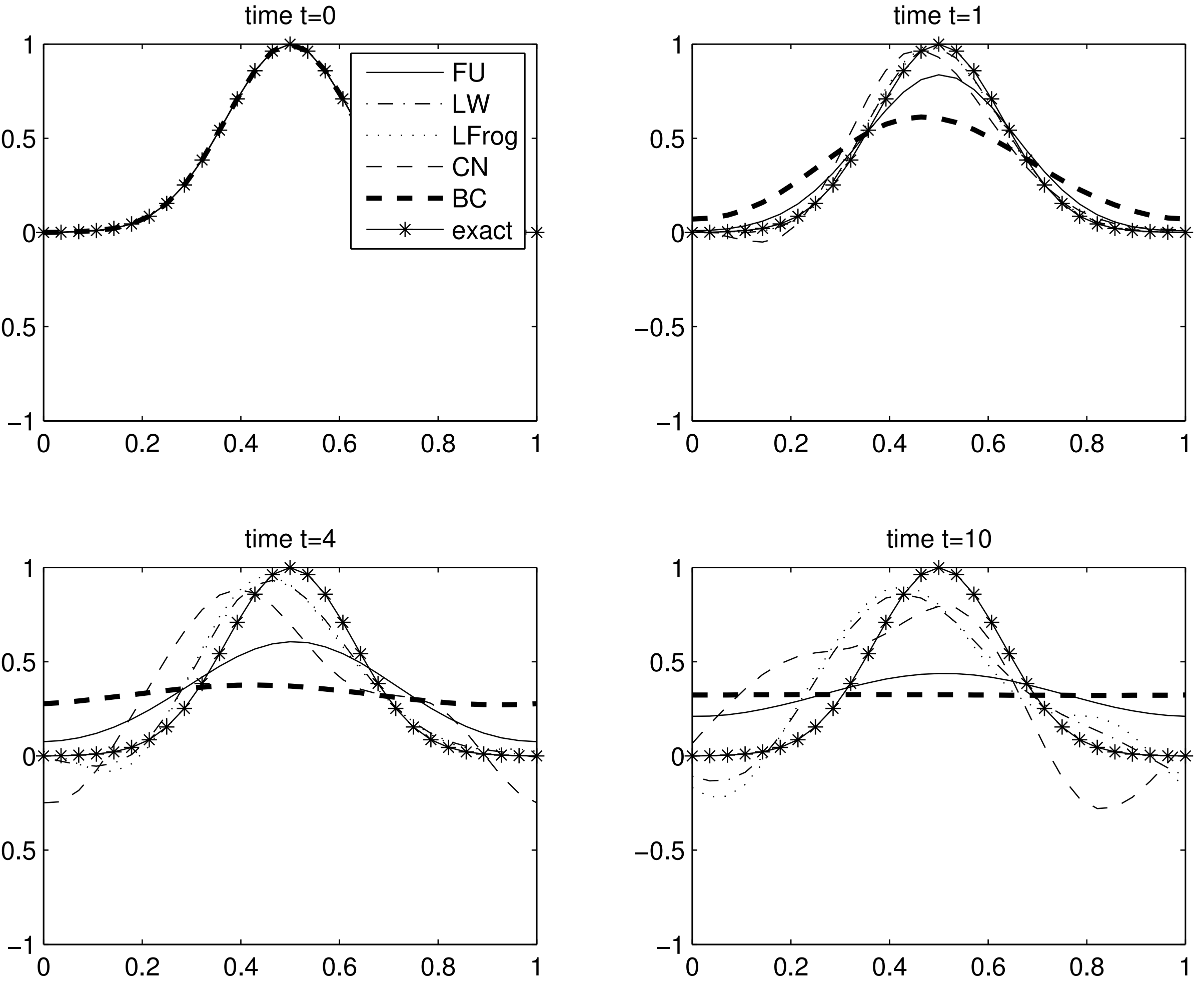


FIGURE 2.2: *A comparison of BC, CN, FU, LFrog and LW numerical schemes for the advection equation with $a = 1$, applied to a Gaussian profile with periodic boundary conditions.*

where $\hat{e}^n$ is the amplitude of the wave at time n and k is the wavenumber. If we assume the grid to be uniform, we can write $x_j = j\Delta x$, and hence obtain

$$e_j^n = \hat{e}^n \exp(ikj\Delta x). \tag{2.77}$$

Since the wavenumber k can be rewritten in terms of the wavelength λ according to $k = 2\pi/\lambda$, the quantity defined by $\theta = k\Delta x$ represents a ratio of the grid spacing and wavelength. Hence, we write

$$e_j^n = \hat{e}^n \exp(ij\theta). \tag{2.78}$$

We are motivated to consider errors of this type since, in general, linear difference operators allow for such wave-like solutions. In particular, we are interested in how the amplitude $\hat{e}^n$ evolves with each time step. Notably, one can observe that if this quantity remains bounded for all θ then the method will be numerically stable. This is so because every initial discrete error can be decomposed as a sum of terms of the form (2.78) and because of linearity.

Returning to our line of reasoning, we substitute (2.78) into (2.75), obtaining

$$e_j^{n+1} = \hat{e}^n \exp(ij\theta) - \tfrac{1}{2}R\left(\hat{e}^n \exp(i(j+1)\theta) - \hat{e}^n \exp(i(j-1)\theta)\right) = \hat{e}^{n+1} \exp(ij\theta), \tag{2.79}$$

where R is shorthand for $a\frac{\Delta t}{\Delta x}$. On dividing by $\exp(ij\theta)$ and rearranging, we obtain

$$\hat{e}^{n+1} = \left[1 - iR\sin\theta\right] \hat{e}^n. \tag{2.80}$$

Equation (2.80) then motivates the following definition:

Definition 2.5 *The* ***symbol*** $S(k)$ *of a two-level finite difference method for the linear advection equation is defined by the ratio*

$$S(k) = \frac{\hat{e}^{n+1}}{\hat{e}^n}. \tag{2.81}$$

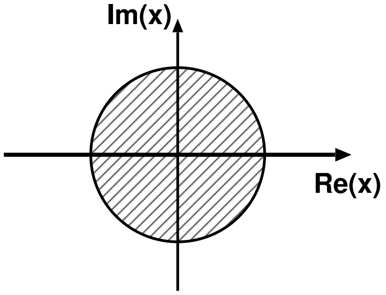
Example As follows from (2.80), the symbol for the forward central scheme (FC) is

$$S(k) = 1 - iR\sin\theta. \tag{2.82}$$

It can be shown that a necessary condition for numerical stability is the *von Neumann stability condition*, given by

$$\boxed{\max_k |S(k)| \leq 1.} \tag{2.83}$$

This condition has an obvious physical meaning in terms of the error amplitudes, as follows from (2.81); namely, the amplitude of any given error mode should not be allowed to grow without bound. Graphically, this condition implies that $S(k)$ must be within a unit circle in $\mathbb{C}$ for all k:



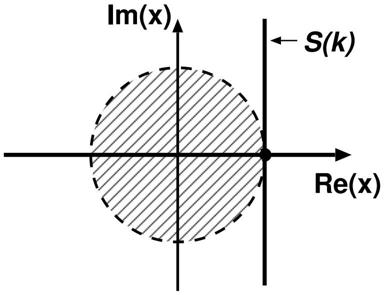
One can easily calculate the value of $|S(k)|$ for the forward central scheme using (2.82), obtaining

$$\boxed{\text{FC}} \quad |S(k)| = \sqrt{1 + R^2 \sin^2 \theta} \geq 1 \; \forall \; \theta. \tag{2.84}$$

Hence,

$$\max_k |S(k)| > 1, \tag{2.85}$$

for any choice of R, and thus, timestep Δt. Hence, we have confirmed our earlier claim that FC is unstable for any choice of Δt. Graphically, the symbol for the FC method is depicted in the following figure.



Example: The Forward Upwind Scheme.

Recall the forward upwind (FU) scheme, given by (2.63). Applying a similar analysis as with the FC scheme, we obtain that the error propagation equation is

$$e_j^{n+1} = e_j^n - R(e_j^n - e_{j-1}^n). \tag{2.86}$$

We substitute (2.78) into (2.86), giving

$$\hat{e}^{n+1}\exp(ij\theta) = \hat{e}^{n}\exp(ij\theta)\left[1 - R(1 - \exp(-i\theta))\right]. \tag{2.87}$$

Hence, the symbol is

$$S(k) = 1 - R(1 - \exp(-i\theta)) = 1 - R + R\cos\theta - iR\sin\theta. \tag{2.88}$$

Evaluating $|S(k)|$, we obtain (exercise)

$$|S(k)|^2 = |1 - 2R(1 - R)(1 - \cos\theta)|. \tag{2.89}$$

We let $C(\theta) = |S(k)|^2$ and take the derivative (recall that $\theta = k\Delta x$), so as to determine extrema of this function. This process leads to

$$\frac{dC(\theta)}{d\theta} = -2R(1 - R)\sin\theta = 0 \quad \Longleftrightarrow \quad \theta = 0, \pm\pi, \pm 2\pi, \ldots \quad \Longleftrightarrow \quad \cos\theta = \pm 1. \tag{2.90}$$

Hence,

$$\begin{aligned} \max|S(k)|^2 &= \max|1 - 2R(1 - R)(1 \pm 1)|, \\ &= \max(1, |1 - 4R(1 - R)|), \\ &= \max(1, |1 - 2R|^2). \end{aligned}$$

We conclude

$$\max|S(k)| = \max(1, |1 - 2R|). \tag{2.91}$$

So, requiring $|1 - 2R| \leq 1$, we obtain $0 \leq R \leq 1$, or

$$\boxed{0 \leq \Delta t \leq \frac{\Delta x}{a}}. \tag{2.92}$$

Thus, in order for FU to be stable we require $\Delta t \leq \frac{\Delta x}{a}$ (a necessary condition). We say that FU is conditionally stable subject to the restriction (2.92). This restriction is known as a *Courant-Friedrichs-Lewy condition (CFL condition).*

Notes: When $a < 0$, we note that the CFL condition (2.92) cannot be satisfied by any Δt, since Δx is positive and a is negative. We conclude that the FU scheme is always unstable when $a < 0$. Conversely, it can be shown that the scheme

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_{j+1}^n - v_j^n}{\Delta x} = 0. \tag{2.93}$$

is always unstable when $a > 0$ and stable when $a < 0$ assuming $0 \leq \Delta t \leq \frac{\Delta x}{-a}$. (Can you think of a physical reason why this might be the case?)

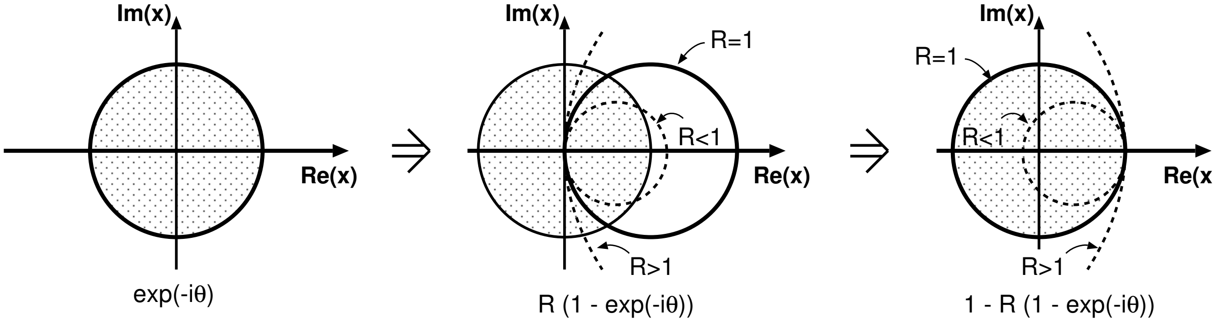
Graphical Techniques for Demonstrating Stability

We can also apply a graphical approach to demonstrate stability of the forward upwind scheme. Recall that the von Neumann stability condition (2.83) is equivalent to stating that $S(k)$ lies within the unit circle for all k. We can view the symbol $S(k)$ (given for the FU scheme in (2.88)) as a map from the real line $\mathbb{R}$ into the complex plane $\mathbb{C}$.

For example, the symbol for the forward upwind scheme, given by

$$S(k) = 1 - R(1 - \exp(-i\theta)), \tag{2.94}$$

gives the following mapping:



We conclude that that the FU scheme is stable if and only if $R \leq 1$, *i.e.*, we again obtain the CFL condition in (2.92).

Example: The Backward Central Scheme

Recall that the backward central scheme is given by (2.59). The error propagation equation is

$$e_j^{n+1} = e_j^n - \tfrac{1}{2}R(e_{j+1}^{n+1} - e_{j-1}^{n+1}). \tag{2.95}$$

On substituting (2.78) into (2.95), it can be shown that the symbol is given by (exercise)

$$S(k) = \frac{1 - iR\sin\theta}{1 + R^2\sin^2\theta}. \tag{2.96}$$

After a short calculation we obtain

$$|S(k)|^2 = \left[1 + R^2\sin^2\theta\right]^{-1/2} \leq 1 \,\forall\, R, \tag{2.97}$$

and so conclude that BC is unconditionally stable, *i.e.*, it is stable for any choice of Δt.

Discussion

In this section we have applied the von Neumann stability analysis to three basic schemes (FC, FU and BC). The von Neumann stability analysis can be easily applied to CN and LW in the same manner, but requires some modification when applied to the 3-level LFrog scheme. The results obtained by this analysis are given in the following table:

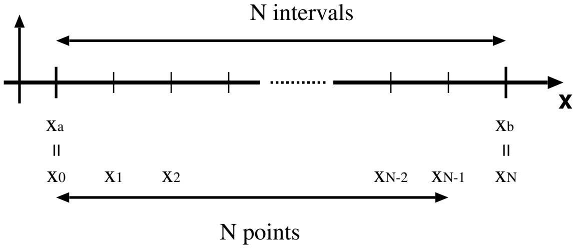
	Scheme	Symbol $S(k)$	Stable?
FC	Forward Central	$1 - iR\sin\theta$	Unstable
BC	Backward Central	$(1 - iR\sin\theta)/(1 + R^2\sin^2\theta)$	Unconditional
FU	Forward Upwind	$1 - R(1 - \exp(-i\theta))$	Conditional (CFL)
CN	Crank-Nicolson	$(2 - iR\sin\theta)/(2 + iR\sin\theta)$	Unconditional
LW	Lax-Wendroff	$1 + R^2(\cos\theta - 1) - iR\sin\theta$	Conditional (CFL)
LFrog	Leapfrog	N/A	Conditional (CFL)

Notes: i) The FU scheme in the form (2.63) is only stable when $a > 0$. LW and LFrog are stable regardless of the sign of a.

ii) Observe that the Crank-Nicolson scheme satisfies $|S(k)| = 1$ for all k. We will show in the next section that this quality is important, since it implies that the CN method does not introduce numerical dissipation.

Link with the Discrete Fourier Transform

Consider a 1D interval of the real line given by (x_a, x_b). We use $N - 1$ interior points to subdivide the interval into N subintervals of equal width, where we label each of the points by $x_a = x_0, x_1, \ldots, x_{N-1}$, $x_N = x_b$. Let $L = x_b - x_a$ denote the length of the interval, with $\Delta x = L/N$ and $x_j = x_a + j\Delta x$ for $j = 0, \ldots, N$ (see figure).



Any function $e(x)$ on (x_a, x_b) then defines a grid function $e[i]$ via $e[i] = e(x_i)$. We further impose periodic boundary conditions so that $e[N] = e[0]$, hence ensuring that $e[i]$ has exactly N degrees of freedom. Now, using the discrete Fourier transform (DFT), any function $e[i]$ can be decomposed in terms of its N Fourier modes

$$e[j] = \sum_{m=0}^{N-1} \hat{e}[m] \exp(i2\pi mj\tfrac{\Delta x}{L}), \tag{2.98}$$

$$\hat{e}[m] = \frac{1}{N}\sum_{j=0}^{N-1} e[j] \exp(-i2\pi mj\tfrac{\Delta x}{L}). \tag{2.99}$$

In performing von Neumann stability analysis, we looked at one mode of this expansion,

$$e_j = \hat{e}\exp(ikx_j) = \hat{e}\exp(ij\theta). \tag{2.100}$$

This result follows since $\exp(i2\pi mj\frac{\Delta x}{L})$ and $\exp(ij\theta)$ are equivalent:

$$\begin{aligned} \exp(i2\pi mj\tfrac{\Delta x}{L}) &= \exp(i2\pi mj\tfrac{1}{N}), && \text{since} \quad \tfrac{\Delta x}{L} = \tfrac{1}{N}, \\ &= \exp(ikj\Delta x), && \text{since} \quad k = 2\pi\tfrac{m}{L}, \\ &= \exp(ij\theta), && \text{since} \quad \theta = k\Delta x = 2\pi\tfrac{m}{N}. \end{aligned}$$

2.2.3 Dissipation and Dispersion

The FD methods produce two types of error, *dissipation* and *dispersion*. We now examine the source of these forms of error and show how dissipation and dispersion terms in PDEs are related to dissipation and dispersion effects in difference formulas.

Dissipation and Dispersion for PDEs

Consider the following three examples of linear PDE operators:

$$L_1 u = \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x}, \tag{2.101a}$$

$$L_2 u = \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} - D\frac{\partial^2 u}{\partial x^2}, \tag{2.101b}$$

$$L_3 u = \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} - \mu\frac{\partial^3 u}{\partial x^3}. \tag{2.101c}$$

Each linear PDE operator generates a linear homogeneous PDE via the equations

$$L_1 u = 0, \qquad L_2 u = 0, \qquad L_3 u = 0. \tag{2.102}$$

We are trying to find wavelike solutions of the form

$$w(x,t) = A_0 \exp(i(kx - \omega t)), \tag{2.103}$$

where A_0 is the amplitude of the wave, k is the wavenumber and ω is the angular frequency. We further define the frequency ν (in oscillations / sec) via $\omega = 2\pi\nu$ and the wavelength (in meters) via $k = 2\pi/\lambda$. The period T (in seconds) of the wave is related to these quantities according to $T = 1/\nu = 2\pi/\omega$. Note that the intervals $x \in [0, \lambda)$ and $t \in [0, T)$ correspond to one full oscillation of the wave in space and time, respectively. Using the variables above, we can rewrite the wave-like solution (2.103) as

$$w(x,t) = A_0 \exp(i2\pi(\tfrac{x}{\lambda} - \nu t)), \quad \text{or} \quad w(x,t) = A_0 \exp(i2\pi(\tfrac{x}{\lambda} - \tfrac{t}{T})). \tag{2.104}$$

By convention we require that A_0 and k are real variables, whereas ω may be complex. We now present the following proposition:

Proposition 2.1 *The wavelike solution (2.103) is an eigenfunction of any linear homogeneous PDE operator in x and t.*

We present no proof to this proposition, instead relying on "proof by example." Consider the three PDE operators presented above (2.101a)-(2.101c), with

$$L_1 w = (-i\omega + aik)w = \lambda_1 w, \tag{2.105a}$$

$$L_2 w = (-i\omega + aik - D(ik)^2)w = \lambda_2 w, \tag{2.105b}$$

$$L_3 w = (-i\omega + aik - \mu(ik)^3)w = \lambda_3 w. \tag{2.105c}$$

The following corollary follows immediately from the proposition.

Corollary 2.1 *Let $Lw = \lambda(\omega, k)w$. Then w is a solution of $Lw = 0$ if and only if ω and k satisfy $\lambda(\omega, k) = 0$.*

The problem of determining wave-like solutions to the PDE operator now reduces to finding solutions to the equation $\lambda(\omega, k) = 0$. Clearly this equation is of importance in the analysis of linear PDE operators, and so it is generally given a name:

Definition 2.6 *The equation*

$$\lambda(\omega, k) = 0, \tag{2.106}$$

is called the ***dispersion relation*** *of the linear homogeneous PDE operator.*

Note that for k real, the dispersion relation (2.106) implicitly defines ω in terms of k, *i.e.* it gives $\omega = \omega(k)$. The three PDE operators (2.101a)-(2.101c) quickly lead to the three dispersion relations:

$$\begin{aligned} \omega(k) &= ak, && (2.107a) \\ \omega(k) &= ak - iDk^2, && (2.107b) \\ \omega(k) &= ak + \mu k^3. && (2.107c) \end{aligned}$$

In general, we find that the dispersion relation will be of the form

$$\omega(k) = \alpha(k) + i\beta(k), \tag{2.108}$$

where $\alpha(k) = \mathrm{Re}(\omega(k))$ and $\beta(k) = \mathrm{Im}(\omega(k))$. Using the dispersion relation in the form (2.108), we rewrite the wave-like solution (2.103) as

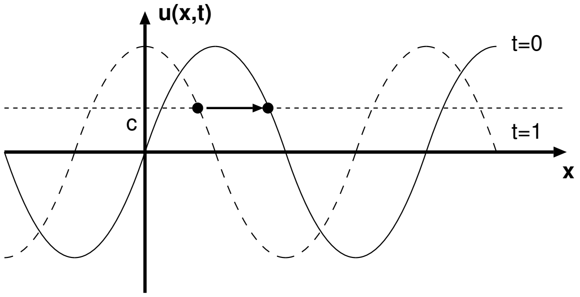
$$\begin{aligned} w(x,t) &= A_0 \exp(i(kx - \alpha(k)t - i\beta(k)t)) \\ &= \underbrace{A_0 \exp(\beta(k)t)}_{A(t)} \exp(i(kx - \alpha(k)t)). && (2.109) \end{aligned}$$

This expression clearly distinguishes the role played by the real and imaginary parts of ω: namely, the former, given by $\alpha(k)$, determines the speed of the wave. The latter, given by $\beta(k)$, affects the amplitude of the wave. The quantity $\alpha(k)$ motivates the following definition.

Definition 2.7 *The* ***phase speed*** *v_{ph} is defined as*

$$v_{ph} = \frac{\mathrm{Re}(\omega(k))}{k} = \frac{\alpha(k)}{k}. \tag{2.110}$$

The physical interpretation of this quantity can be seen as follows: Assume $\beta(k) = 0$ and let c be a constant that implicitly defines x in terms of t via $c = kx - \alpha(k)t$ (we can view c as marking a point along the wave that always remains at a constant value in the wave profile, as in the figure below).



Since c is constant, we have

$$\frac{dc}{dt} = 0 = k\frac{dx}{dt} - \alpha(k) \quad \Longleftrightarrow \quad \frac{dx}{dt} = \frac{\alpha(k)}{k}. \tag{2.111}$$

Hence, the speed of a point in phase with the wave profile is given by the phase speed v_{ph}.

For each of the linear PDE operators, we obtain the following relations for phase speed:

$$\begin{aligned} L_1 \Rightarrow v_{ph} &= \frac{ak}{k} = a, && (2.112a) \\ L_2 \Rightarrow v_{ph} &= k^{-1}\mathrm{Re}(ak - iDk) = a, && (2.112b) \\ L_3 \Rightarrow v_{ph} &= k^{-1}(ak + \mu k^3) = a + \mu k^2. && (2.112c) \end{aligned}$$

Note that for L_3 the profile u experiences *dispersion*, *i.e.* waves of different wavenumber will move at different phase speeds.

We now focus on the amplitude term in (2.109). As a function of time, we obtain the following relations for the amplitude of wave-like solutions for each PDE operator:

$$\begin{aligned} &L_1 \Rightarrow \beta(k) = 0 && A(t) = A_0, && (2.113a) \\ &L_1 \Rightarrow \beta(k) = -Dk^2 && A(t) = A_0 \exp(-Dk^2 t), && (2.113b) \\ &L_1 \Rightarrow \beta(k) = 0 && A(t) = A_0. && (2.113c) \end{aligned}$$

We discover that the amplitude of the wave-like solution is preserved for L_1 and L_3, but decays in time for L_2, *i.e.* the wave-like solution experiences *dissipation*.

Definition 2.8 *Let L be a linear homogeneous PDE operator. We say that L is **dissipative** if and only if $\mathrm{Im}(\omega(k)) < 0$. Further, we say that L is **dispersive** if and only if $\mathrm{Re}(\omega(k))$ is not linear in k.*

In general, for a PDE with a first order time derivative, we require partial spatial derivatives of even order in order to obtain dissipation. On the other hand, in order to obtain dispersion we require partial spatial derivatives of odd order (where the order is at least 3).

Dissipation and Dispersion for Difference Formulas

By approximating a differential equation by a difference formula, we introduce numerical dissipative and dispersive behaviour that is closely related to dissipation and dispersion in linear PDE operators. We now show how dissipative and dispersive errors occur in difference formulas.

Similar to the case of linear homogeneous PDE operators, we consider a wave-like solutions of a finite difference operator given by

$$e_j^n = e_0 \exp(i(kj\Delta x - \omega n\Delta t)), \quad \text{where } x = j\Delta x \text{ and } t = n\Delta t. \tag{2.114}$$

We hence obtain discrete analogues of proposition 2.1 and its corollary:

Proposition 2.2 *The wavelike solution (2.114) is an eigenfunction of any linear homogeneous difference operator.*

Corollary 2.2 *Let $L\mathbf{e} = \lambda(\omega, k)\mathbf{e}$. Then $\mathbf{e}$ is a solution of $L\mathbf{e} = 0$ if and only if ω and k satisfy*

$$\lambda(\omega, k) = 0. \tag{2.115}$$

As in the continuous case, the dispersion relation for a FD operator is again defined by $\lambda(\omega, k) = 0$. By convention we again choose k real. The dispersion relation then implicitly defines $\omega = \omega(k)$.

It turns out that the symbol $S(k)$ has all the information we need to determine the strength of numerical dissipation and dispersion. In particular, we can find a relation between $S(k)$ and $\omega(k)$, as follows. Recall that the symbol $S(k)$ (see Definition 2.5) is given by

$$\hat{e}^{n+1} = S(k)\hat{e}^n. \tag{2.116}$$

If we apply this equation recursively, we obtain the relation

$$\hat{e}^n = (S(k))^n \hat{e}^0, \tag{2.117}$$

where $\hat{e}^0$ denotes the error at some initial time. Using the definition of $\hat{e}$ in the form (2.77), we can rewrite (2.114) as

$$\hat{e}^n = e_0 \exp(-i\omega n \Delta t). \tag{2.118}$$

Then, upon equating (2.117) and (2.118) we obtain

$$e_0(\exp(-i\omega\Delta t))^n = \hat{e}^0(S(k))^n, \quad \forall\, n. \tag{2.119}$$

Clearly (2.118) implies $e_0 = \hat{e}^0$, and so it follows that

$$S(k) = \exp(-i\omega\Delta t), \tag{2.120}$$

where $\omega = \omega(k)$ is the dispersion relation of the FD method. Since $S(k)$ is complex in general, we can write it in polar form as

$$S(k) = |S| \exp(i\phi_S) = \exp(\ln|S|)\exp(i\phi_S), \tag{2.121}$$

where

$$|S| = \sqrt{\mathrm{Re}(S)^2 + \mathrm{Im}(S)^2}, \quad \text{and} \quad \phi_S = \arctan\left(\frac{\mathrm{Im}(S)}{\mathrm{Re}(S)}\right). \tag{2.122}$$

Comparing (2.120) and (2.121) then leads to

$$\boxed{\omega(k) = \frac{-\phi_S + i\ln|S|}{\Delta t}.} \tag{2.123}$$

We now aim to determine the conditions on (2.123) that lead to dispersive and dissipative solutions. The numerical phase speed can be written in terms of (2.110) and (2.123), giving

$$v_{ph} = \frac{1}{k}\mathrm{Re}(\omega) = -\frac{\phi_s}{k\Delta t}. \tag{2.124}$$

In particular, if the phase speed is not constant in k we note that wave-like solutions will be dispersive (see definition 2.8). For the advection equation, we can use $R = a\frac{\Delta t}{\Delta x}$ and $\theta = k\Delta x$ to obtain the relation

$$v_{ph} = \frac{-a\phi_S}{R\theta}. \tag{2.125}$$

The phase velocity for various FD methods is plotted in figure 2.3.

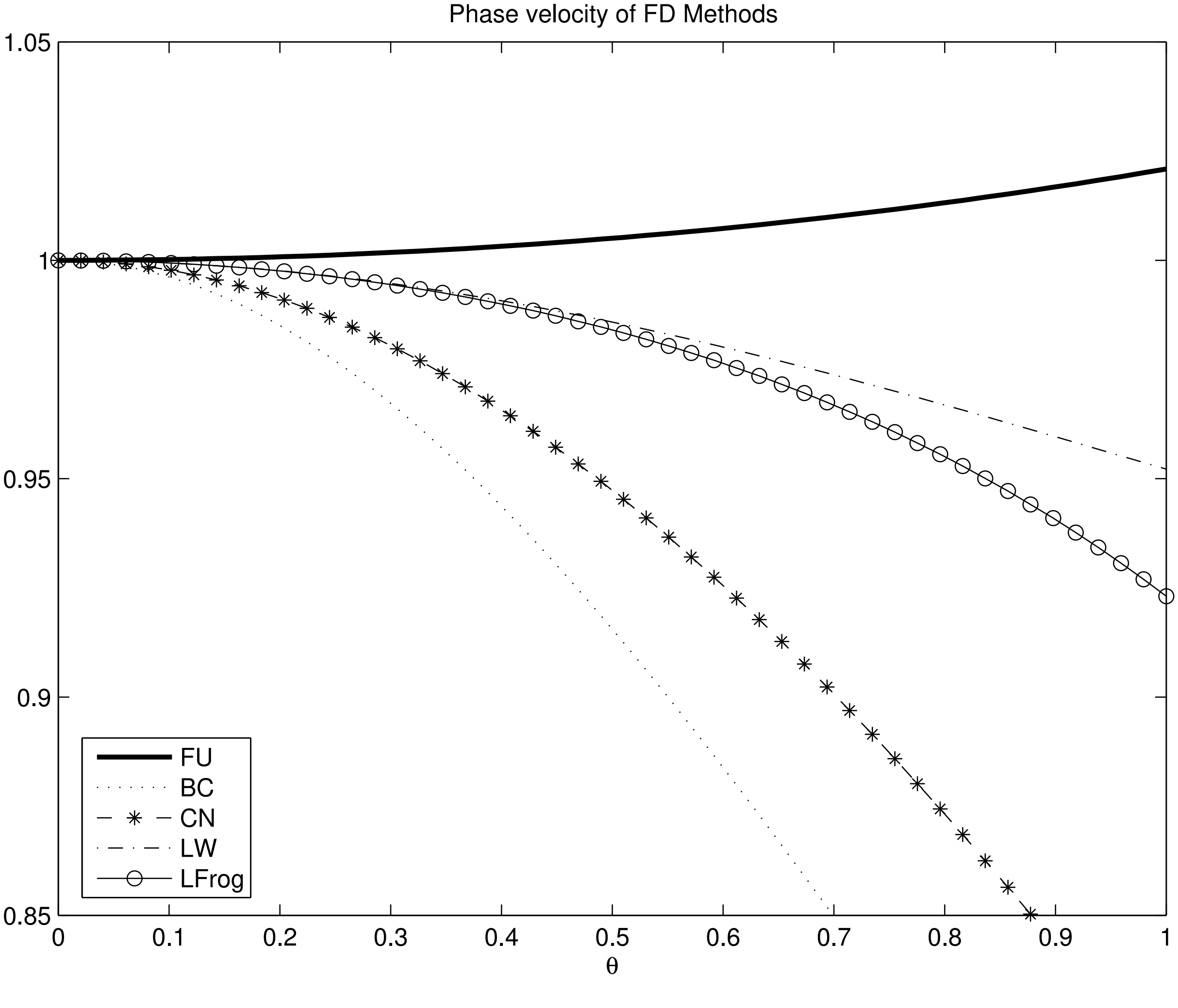


FIGURE 2.3: *A comparison of the phase velocity v_{ph} for the BC, CN, FU, LFrog and LW numerical schemes applied to the linear advection equation.*

Turning our attention to dissipation, it follows from (2.123) that

$$\mathrm{Im}(\omega) = \frac{\ln|S|}{\Delta t}. \tag{2.126}$$

On recalling that dissipation is associated with $\mathrm{Im}(\omega)$ being nonzero, we note that dissipation will be present[4] whenever $|S| < 1$ (again see definition 2.8). The amplitude of the symbol for various FD methods is plotted in figure 2.4.

Note that figure 2.4 suggests we obtain minimal dissipation when $\theta \to 0$. Since $\theta = k\Delta x$, this result has two interpretations: First, if we fix k, decreasing Δx will result in a decrease in θ and so reducing the grid spacing leads to less dispersion and dissipation. Second, if we fix Δx, increasing k will result in an increase in θ. We conclude that waves with higher wave number (and shorter wavelength) will be damped out more quickly by dissipation.

On comparing and contrasting the various FD methods so far examined in this section, we obtain the following results:

	Scheme	Order	Dispersion	Dissipation
BC	Backward Central	1	large	large
FU	Forward Upwind	1	small	large
CN	Crank-Nicolson	2	large	0
LW	Lax-Wendroff	2	small	small
LFrog	Leapfrog	2	small	0

We can also analyze the dissipative and dispersive effects of numerical schemes using an alternative approach. For example, consider the forward upwind method. The truncation error in this case is given by

$$T_j^n = (R-1)a\tfrac{1}{2}\Delta x u_{xx} + O(\Delta t^2) + O(\Delta x^2) + O(\Delta t \Delta x). \tag{2.127}$$

The dominant error term here is proportional to u_{xx}, which has a dissipative effect. We conclude that the error in the forward upwind method is dominated by dissipation.

Consider instead the Lax-Wendroff method, with truncation error given by

$$T_j^n = a u_{xxx}\tfrac{1}{3}\Delta x^2 - a^3 u_{xxx}\tfrac{1}{6}\Delta t^2 + \text{ h.o.t.} \tag{2.128}$$

The dominant error term here is proportional to u_{xxx}, which has a dispersive effect.

[4] Recall further that the method is unstable if $|S| > 1$.

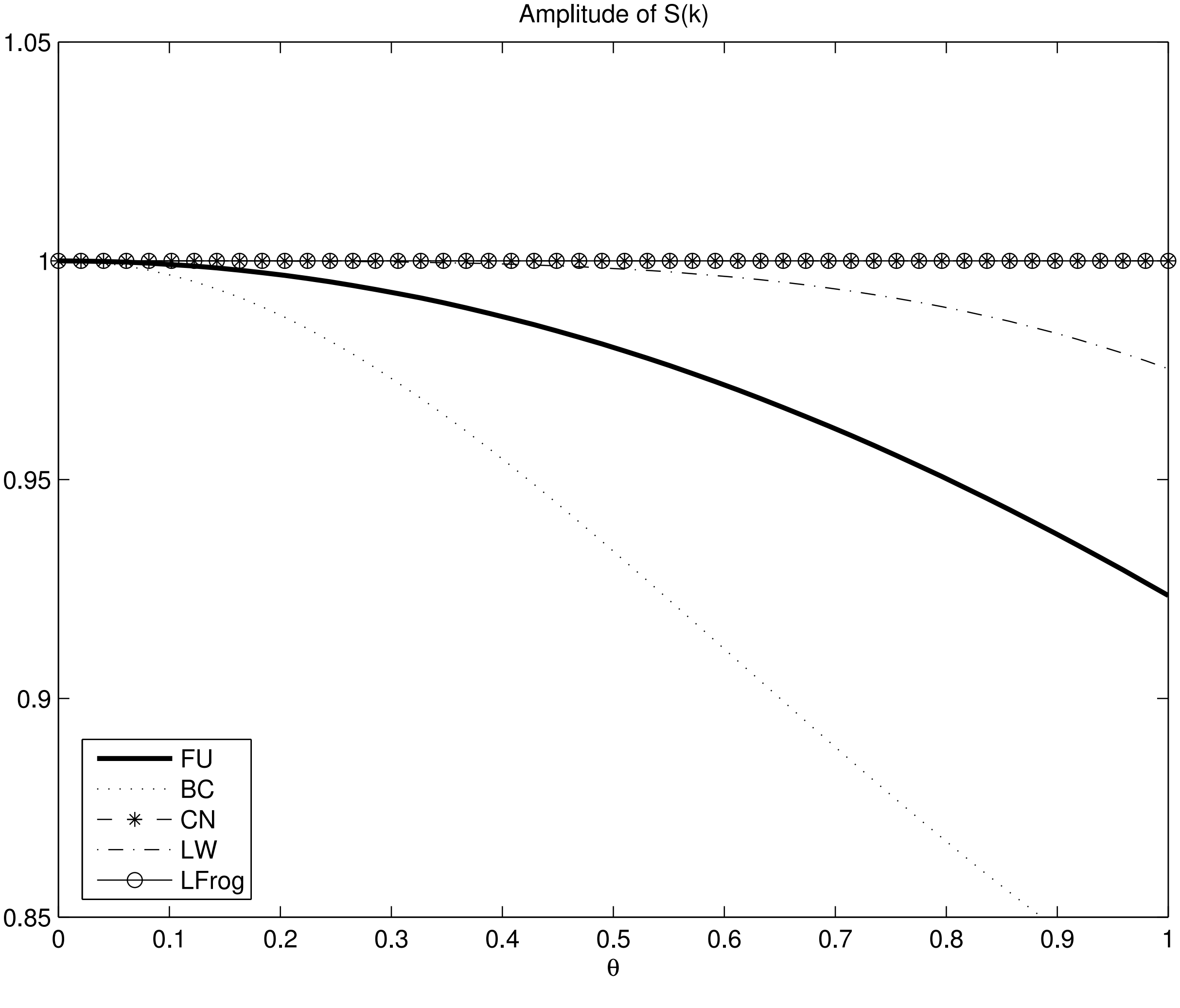


FIGURE 2.4: *A comparison of the symbol amplitude $|S|$ for the BC, CN, FU, LFrog and LW numerical schemes applied to the linear advection equation with $a = 1$.*

Note: In general, first order accurate methods have a dominant error term that is proportional to u_{xx}, and so will have an error dominated by dissipation. Similarly, second order accurate methods have a dominant error term that is proportional to u_{xxx}, and so will have an error dominated by dispersion.

We can also analyze the dissipative term in the FU method by rewriting the difference equation (2.63) as

$$\underbrace{\frac{v_j^{n+1}-v_j^n}{\Delta t}}_{\frac{\partial u}{\partial t}} + a\underbrace{\frac{v_{j+1}^n - v_{j-1}^n}{2\Delta x}}_{\frac{\partial u}{\partial x}} = \frac{a\Delta x}{2}\underbrace{\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2}}_{\frac{\partial^2 u}{\partial x^2}}. \tag{2.129}$$

If we consider this difference equation to be a discretization of (2.101b), the term on the right-hand side will behave much like a diffusion term. We conclude that the error in this discretization is governed by diffusion.

2.2.4 Finite Difference Methods for the Wave Equation

We now extend our analysis of FD methods for hyperbolic PDEs to the 1D wave equation.

Recall from (1.8) that the wave equation is given by

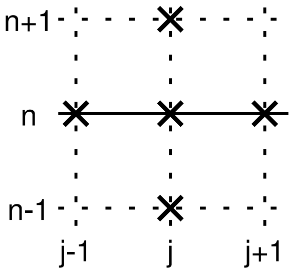
$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad u = u(x,t). \tag{2.130}$$

Of interest is the fact that we have a second t derivative in this equation, which prevents us from directly applying the time discretizations so far considered in this chapter. As a result, we now present two common methods for solving this DE.

Method 1: We can discretize this DE in much the same manner as we did in section 2.2.1 with the linear advection equation. One such discretization is the central difference scheme, originally applied in (2.2), which gives

$$\frac{v_i^{n+1} - 2v_i^n + v_i^{n-1}}{\Delta t^2} = a^2\frac{v_{i+1}^n - 2v_i^n + v_{i-1}^n}{\Delta x^2}. \tag{2.131}$$

The stencil for this method is depicted as follows.



It can be shown that this discretization has truncation error

$$T_i^n = O(\Delta t^2) + O(\Delta x^2), \tag{2.132}$$

and that the usual direction independent CFL condition,

$$\Delta t \leq \frac{1}{|a|}\Delta x, \tag{2.133}$$

is required for stability.

Method 2: Recall that (2.130) can be rewritten as $L_1 L_2 u = 0$, where L_1 and L_2 are linear PDE operators given by

$$L_1 = \frac{\partial}{\partial t} + a\frac{\partial}{\partial x}, \quad \text{and} \quad L_2 = \frac{\partial}{\partial t} - a\frac{\partial}{\partial x}. \tag{2.134}$$

On defining $w = L_2 u$, we obtain the system of equations

$$L_2 u = w, \quad L_1 w = 0, \tag{2.135}$$

which can be rewritten in matrix form as

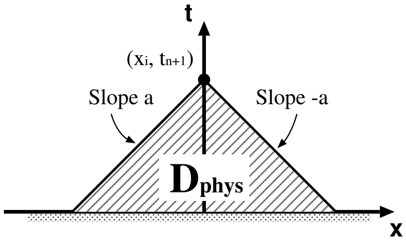
$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ w \end{bmatrix} + \begin{bmatrix} -a & 0 \\ 0 & a \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} w \\ 0 \end{bmatrix}. \tag{2.136}$$

This system can then be solved using generalized versions of FU, BC, LW, CN or $LFrog$ for a coupled system of equations.

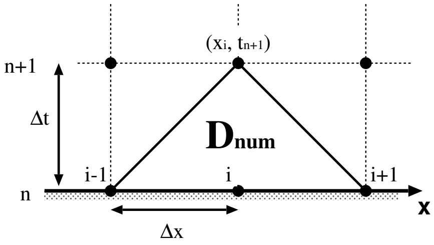
Physical interpretation of CFL condition for explicit methods

We now give a physical interpretation of the CFL condition that follows from the 1D wave equation.

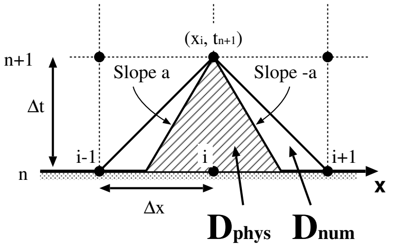
Recall the definition of the (physical) domain of dependence for a hyperbolic PDE, given in section 1.2.4. For the 1D wave equation, the domain of dependence assumes the following form.



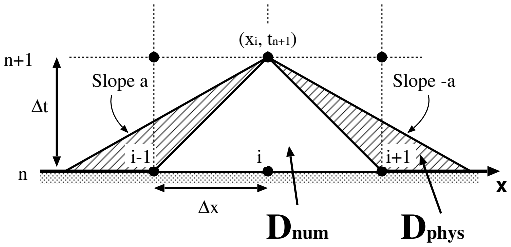
Similar to the physical domain of dependence associated with the PDE, we can also construct a numerical domain of dependence associated with the explicit FD scheme. If the value of v_i^{n+1} only depends on v_{i-1}^n, v_i^n and v_{i+1}^n, then the numerical domain of dependence effectively takes on the following shape.



We now assume that Δt satisfies the CFL condition, *i.e.* $\Delta t \leq \frac{\Delta x}{a}$. The numerical domain of dependence in this case falls outside the physical domain of dependence, as follows.



If Δt does not satisfy the CFL condition, *i.e.* $\Delta t > \frac{\Delta x}{a}$, the numerical domain of dependence instead falls within the physical domain of dependence, as seen below.



Hence we claim that a FD method is stable if and only if the numerical domain of dependence, D_{num} contains the physical domain of dependence, D_{phys}. In other words, a given FD method is unstable when the physical evolution of the PDE requires more information than can be obtained from the numerical data.

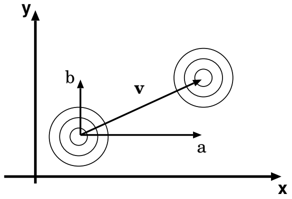
2.2.5 Finite Difference Methods in 2D and 3D

In this section we briefly discuss extending our general FD methods for 1D PDEs to 2D and 3D. In particular, the simplest way of handling this problem is to perform a so-called dimension-by-dimension extension of 1D methods. We demonstrate this process by example.

Example: The linear advection equation in 2D is given by

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} = 0, \qquad (\text{assume } a, b > 0). \tag{2.137}$$

The advection speed is the vector in 2D given by $\mathbf{v} = (a, b)$. Given some initial profile $u(x, y, t = 0)$, the profile is advected in the direction of $\mathbf{v}$, as in the following figure.



We discretize this equation on a standard Cartesian grid with grid points labelled by (x_i, y_j) with $i = 0, \ldots, n$ and $j = 0, \ldots, m$. The numerical solution is then a grid function given by v_{ij}^n. Using a natural generalization of the forward upwind scheme, we discretize the PDE as

$$\frac{v_{i,j}^{n+1} - v_{i,j}^n}{\Delta t} + a\frac{v_{i,j}^n - v_{i-1,j}^n}{\Delta x} + b\frac{v_{i,j}^n - v_{i,j-1}^n}{\Delta y} = 0. \tag{2.138}$$

This method is explicit, with truncation error given by

$$T_{ij}^n = O(\Delta t) + O(\Delta x) + O(\Delta y). \tag{2.139}$$

As with the 1D scheme, we can apply von Neumann stability analysis in order to obtain a condition for the stability of (2.138). In this case, the general wave-like error takes the form

$$e_{j_1 j_2}^n = \hat{e}^n \exp(i(j_1 k_1 \Delta x + j_2 k_2 \Delta y)). \tag{2.140}$$

We substitute (2.140) into the discretization and solve for the symbol $S(k_1, k_2)$, imposing

$$\max_{k_1, k_2} |S(k_1, k_2)| \leq 1, \tag{2.141}$$

for stability. Following a tedious calculation, we obtain

$$0 \leq a\frac{\Delta t}{\Delta x} + b\frac{\Delta t}{\Delta x} \leq 1, \quad a\frac{\Delta t}{\Delta x} \geq 0, \quad b\frac{\Delta t}{\Delta x} \geq 0, \tag{2.142}$$

which, on applying some simple identities, reduces to

$$\Delta t \leq \frac{1}{\frac{a}{\Delta x} + \frac{b}{\Delta y}} \leq \tfrac{1}{2}\min\left(\frac{\Delta x}{a}, \frac{\Delta y}{b}\right). \tag{2.143}$$

The previous example was a very simple example of extending FD methods to higher dimensions, and clearly suggests more interesting generalizations outside the scope of this text. The problem of discretizing PDEs in high dimensions continues to be a research area of significant interest.

2.3 Finite Difference Methods for Parabolic PDEs

Parabolic PDEs differ from hyperbolic PDEs in one important respect: whereas a point's domain of dependence is finite in space at any given time for a hyperbolic PDE, a parabolic PDE has an infinitely large domain of dependence at any given time. This suggests that explicit methods may not be very suitable for parabolic problems. However, in this section we will see that simply capturing the "majority" of information is sufficient to ensure stability.

We now consider two methods for solving the heat diffusion problem in 1D (a parabolic PDE). Recall that the 1D heat equation (1.23) with source term $f(x)$ is given by

$$\frac{\partial u}{\partial t} - D(x)\frac{\partial^2 u}{\partial x^2} = f(x). \tag{2.144}$$

We discretize the spatial derivative using the central-difference method given in (2.2). The time derivative can be discretized using the schemes we have derived in section 2.2 for the advection equation. For example, we can choose the Forward Euler and Crank-Nicolson discretizations, which, when applied to (2.144), give

$$\boxed{\text{FE}} \quad \frac{v_j^{n+1} - v_j^n}{\Delta t} = D\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2} + f(x_j), \tag{2.145}$$

and

$$\boxed{\text{CN}} \quad \frac{v_j^{n+1} - v_j^n}{\Delta t} = \frac{D}{2}\left(\frac{v_{j+1}^{n+1} - 2v_j^{n+1} + v_{j-1}^{n+1}}{\Delta x^2} + \frac{v_{j+1}^{n+1} - 2v_j^n + v_{j-1}^{n+1}}{\Delta x^2}\right) + f(x_j). \tag{2.146}$$

The truncation error for the FE discretization is given by

$$T_j^n = O(\Delta t) + O(\Delta x), \tag{2.147}$$

and for Crank-Nicolson by

$$T_j^n = O(\Delta t^2) + O(\Delta x^2). \tag{2.148}$$

Stability

In order to analyze the stability of the methods (2.145) and (2.146), we now apply the von Neumann method.

On rewriting (2.145) in terms of the actual error e_j, we obtain

$$T_j^n - \left(\frac{e_j^{n+1} - e_j^n}{\Delta t} \right) = -D \left(\frac{e_{j+1}^n - 2e_j^n + e_{j-1}^n}{\Delta x^2} \right). \tag{2.149}$$

We consider only propagation of error (and so set $T_j^n = 0$) and assume a wave-like solution of the form

$$e_j^n = \hat{e}^n \exp(ijk\Delta x). \tag{2.150}$$

After a short calculation, we obtain

$$\hat{e}^{n+1} = (1 + D\tfrac{\Delta t}{\Delta x^2}(2\cos(k\Delta x) - 2))\hat{e}^n, \tag{2.151}$$

and so conclude that the symbol $S(k)$ takes the form

$$S(k) = 1 + D\tfrac{\Delta t}{\Delta x^2}(2\cos(k\Delta x) - 2). \tag{2.152}$$

On noting that the trigonometric term $(2\cos\theta - 2)$ takes values in the interval $[-4, 0]$, we conclude that $D\frac{\Delta t}{\Delta x^2}$ must satisfy

$$D\tfrac{\Delta t}{\Delta x^2} \leq \tfrac{1}{2} \tag{2.153}$$

for stability, *i.e.* we require

$$\boxed{0 < \Delta t \leq \frac{\Delta x^2}{2D}} \tag{2.154}$$

(compare with the CFL condition for the advection equation (2.92)).

Notes: i) The timestep restriction (2.154), being quadratic in Δx, is stricter than the CFL condition for the advection equation. Hence, we will require a very small grid size to ensure stability of this method when solving the parabolic DE.

ii) Note that a large diffusion constant D leads to a small timestep. This result simply reflects the physical interpretation of the diffusion constant, namely the speed at which material "spreads out ." If material spreads out more quickly, more timesteps are required in order to ensure this information is propagated appropriately.

iii) Clearly, the presence of an infinite propagation speed does not prevent the application of an explicit method for discretizing the parabolic PDE. However, because of the strict timestep restriction, we conclude that the FE method is not practical for the heat equation except in the case of a small diffusion parameter D.

iv) We often say that parabolic problems are "more stiff" than hyperbolic problems as a consequence of the more stringent timestep restrictions in simulating them using explicit methods.

A similar calculation for the Crank-Nicolson discretization (2.146) leads to (exercise)

$$S(k) = \frac{1 + \frac{D\Delta t}{2\Delta x^2}(2\cos(k\Delta x) - 2)}{1 - \frac{D\Delta t}{2\Delta x^2}(2\cos(k\Delta x) - 2)}. \tag{2.155}$$

We note that $S(k)$ satisfies $|S(k)| \leq 1$ for all k, and hence CN is unconditionally stable. This method is significantly more practical for simulating the heat equation, but it is implicit and hence requires solving a linear system in each timestep.

2.4 Finite Difference Convergence Theory for Time-Dependent Problems

In this section we discuss convergence theory for finite difference methods applied to time-dependent problems, *i.e.* PDEs of parabolic or hyperbolic type.

We focus on linear PDEs of the form[5]

$$\frac{\partial u}{\partial t} - Lu = f, \tag{2.156}$$

where L is the spatial part of the PDE operator. We make no assumptions about the dimensionality of the problem.

[5] Note that a PDE with a second order time derivative $\frac{\partial^2 u}{\partial t^2}$ can be treated similarly.

Example 1 (Linear Advection Equation in 1D): The PDE takes the form (2.156) with

$$Lu = -a\frac{\partial u}{\partial x}, \quad \text{and} \quad f = 0. \tag{2.157}$$

Example 2 (Diffusion Equation in 2D): The PDE takes the form (2.156) with

$$Lu = D\nabla^2 u, \quad \text{and} \quad f = f(x, y). \tag{2.158}$$

We further restrict our considerations to FD discretizations (spatial and temporal) with exactly two levels in time. This restriction allows for all hyperbolic and parabolic methods discussed in this chapter except for the Leapfrog scheme.

Example 1 (continued): The PDE is discretized as

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_j^n - v_{j-1}^n}{\Delta x} = 0. \tag{2.159}$$

We discretize the derivative operator $\frac{\partial}{\partial x}$ as a matrix A_h and write the numerical solution v_j at an arbitrary timestep as a vector denoted by V_h. Hence, the operation $-Lu$ is discretized by the product $A_h V_h$.

$$\frac{V_h^{n+1} - V_h^n}{\Delta t} + aA_h V_h^n = 0, \tag{2.160}$$

where

$$A_h = \frac{1}{\Delta x}\begin{bmatrix} 1 & 0 & & -1 \\ -1 & 1 & 0 & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix}. \tag{2.161}$$

The -1 in the upper right-corner of the matrix is chosen so as to lead to periodic boundary conditions. We collect terms evaluated at timestep $n+1$ on the left hand side and terms evaluated at timestep n on the right hand side, obtaining

$$V_h^{n+1} = (I - \Delta t a A_h)V_h^n. \tag{2.162}$$

Example 2 (continued): We apply the 5-point weighted discretization of the Laplacian $\nabla^2 u$, given by (2.20),

$$\nabla^2 u \approx \frac{v_{i+1,j} + v_{i-1,j} - 4v_{i,j} + v_{i,j+1} + v_{i,j-1}}{h^2}. \tag{2.163}$$

In matrix form, the discrete operator takes on the block-diagonal form

$$H_h = \frac{1}{h^2}\left[\begin{array}{c|c|c|c} T & I & 0 & 0 \\ \hline I & T & I & 0 \\ \hline 0 & \ddots & \ddots & \ddots \\ \hline 0 & 0 & I & T \end{array}\right], \quad \text{where} \quad T = \begin{bmatrix} -4 & 1 & & 0 \\ 1 & -4 & 1 & \\ & \ddots & \ddots & \ddots \\ 0 & & 1 & -4 \end{bmatrix}. \tag{2.164}$$

Using the Crank-Nicolson time discretization, we have

$$\frac{V_h^{n+1} - V_h^n}{\Delta t} = D\left(\frac{H_h V_h^{n+1} + H_h V_h^n}{2}\right) + F_h. \tag{2.165}$$

We collect terms evaluated at timestep $n+1$ on the left hand side and terms evaluated at timestep n on the right hand side, obtaining

$$(I - \tfrac{1}{2}D\Delta t H_h)V_h^{n+1} = (I + \tfrac{1}{2}D\Delta t H_h)V_h^n + F_h\Delta t. \tag{2.166}$$

On closer examination, we observe that (2.162) and (2.166) can be written in a single unified form given by

$$\boxed{P_{h,\Delta t}V_h^{n+1} = Q_{h,\Delta t}V_h^n + F_h\Delta t.} \tag{2.167}$$

The evolution equation (2.167) is called the *discrete evolution equation* and generalizes all 2-level linear FD methods for time-dependent problems.

Example 1 (continued): We define the matrices $P_{h,\Delta t}$ and $Q_{h,\Delta t}$ by

$$P_{h,\Delta t} = I, \quad \text{and} \quad Q_{h,\Delta t} = I - \Delta a A_h. \tag{2.168}$$

Note that for any explicit method, $P_{h,\Delta t}$ will be the identity matrix. Further, any homogeneous equation will satisfy $F_h = 0$.

Example 2 (continued): We define the matrices $P_{h,\Delta t}$ and $Q_{h,\Delta t}$ by

$$P_{h,\Delta t} = I - \tfrac{1}{2}D\Delta t H_h, \quad \text{and} \quad Q_{h,\Delta t} = I + \tfrac{1}{2}D\Delta t H_h. \tag{2.169}$$

2.4.1 Actual Error, Truncation Error and Consistency

One can easily extend the definitions of truncation error (Definition 2.2) and consistency (Definition 2.3) to time-dependent PDEs in the form (2.167). For convenience, we present these definitions here.

Definition 2.9 *The* ***truncation error*** T_h *of a time-dependent numerical method of the form (2.167) satisfies*

$$P_{h,\Delta t}U_h^{n+1} = Q_{h,\Delta t}U_h^n + F_h\Delta t + T_h n\Delta t. \tag{2.170}$$

Definition 2.10 *A FD method in the form (2.167) for the PDE (2.156) is said to be* ***consistent*** *if and only if*

$$\lim_{\Delta t\to 0,\Delta x\to 0} T_i = 0. \tag{2.171}$$

Further, we say that it is consistent with order q_1 *in time and order* q_2 *in space (*$q_1, q_2 \in \mathbb{N}_0$*) if and only if*

$$T_j^n = O(\Delta t^{q_1}) + O(\Delta x^{q_2}). \tag{2.172}$$

2.4.2 Stability and Convergence: Lax Convergence Theorem

We now have all the necessary tools to derive a convergence theorem for parabolic and hyperbolic PDEs similar to the Lax convergence theorem for elliptic PDEs (Theorem 2.1).

Consider a general time-dependent IVP of the form

$$IVP \begin{cases} \Omega : (x,t) \in \mathbb{R}\times[0,t^*], \\ u(x,0) = u_0(x), \\ u_t - Lu = f \text{ on } \Omega. \end{cases} \tag{2.173}$$

The notion of convergence to the exact solution of this PDE is essentially the same as with elliptic PDEs; namely, as we refine the grid in space and time we expect that the numerical solution will converge to the exact solution. The difference in this case is that we must consider the time and space dependence separately.

Definition 2.11 *A finite difference method (2.167) is* ***convergent in the*** p***-norm with order*** q_1 ***in time and*** q_2 ***in space*** *if and only if*

$$\max_{n,n\Delta t\le t^*} \|E_h^n\|_p = O(\Delta t^{q_1}) + O(\Delta x^{q_2}), \tag{2.174}$$

where Δt *and* h *may be required to satisfy a stability condition.*

The last clause in this definition may lead to some confusion. This restriction prevents us from arbitrarily refining the time and space components of the mesh without consideration to something like a CFL condition. For example, in the following circumstances and many others, we are required to impose a constraint on the limit:

- If we apply Forward Upwind to the linear advection equation, we impose that Δt and h must satisfy the CFL condition (2.92).
- If we apply the Forward Euler discretization to the heat diffusion equation, we impose that Δt and h must satisfy (2.154).

The notion of stability of a time-dependent FD method is an extension of Definition 2.4.

Definition 2.12 *A finite difference method (2.167) is* ***stable in the*** p***-norm*** *if there exists* c *(independent of* h *and* Δt*) so that*

$$\|(P_{h,\Delta t}^{-1}Q_{h,\Delta t})^n P_{h,\Delta t}^{-1}\|_p \leq c, \tag{2.175}$$

for all n *and* Δt *so that* $n\Delta t \leq t^*$*, where* Δt *and* h *may be required to go to satisfy a stability condition.*

Together, Definitions 2.10, 2.11 and 2.12 lead to the Lax convergence theorem for time-dependent PDEs, which we now state.

Theorem 2.2 (Lax Convergence Theorem) *Let FD method (2.167) be consistent in the* p*-norm with order* q_1 *in time and* q_2 *in space such that*

$$T_{max,p} = \max_{n,\ n\Delta t \leq t^*} \|T_h^n\|_p = O(\Delta t^{q_1}) + O(\Delta x^{q_2}). \tag{2.176}$$

Further, let the FD method be stable in the p*-norm. Then the FD method is convergent in the* p*-norm with order* q_1 *in time and* q_2 *in space.*

Proof: For sake of brevity, let $P = P_{h,\Delta t}$ and $Q = Q_{h,\Delta t}$. The numerical method (2.167) then takes the form

$$PV_h^{n+1} = QV_h^n + F_h\Delta t. \tag{2.177}$$

By definition of the truncation error (2.170), we also have

$$PU_h^{n+1} = QU_h^n + F_h\Delta t + T_h^n\Delta t. \tag{2.178}$$

Taking the difference between (2.177) and (2.178) and applying the definition of the actual error (2.11) then yields

$$PE_h^{n+1} = QE_h^n - T_h^n \Delta t, \tag{2.179}$$

or equivalently

$$E_h^{n+1} = P^{-1}QE_h^n - P^{-1}T_h^n \Delta t \tag{2.180}$$

(it is a consequence of stability that P is invertible; see Definition 2.12). Applying this formula recursively then gives (exercise)

$$E_h^n = (P^{-1}Q)^n E_h^0 + \Delta t \sum_{m=1}^{n} (P^{-1}Q)^{n-m} P^{-1} T_h^{m-1}. \tag{2.181}$$

Note that $E_h^0 = U_h^0 - V_h^0 = 0$. We take the p-norm of this result and apply standard inequalities, obtaining

$$\|E_h^n\|_p \leq \Delta t \sum_{m=1}^{n} \|(P^{-1}Q)^{n-m} P^{-1}\|_p \|T_h^{m-1}\|_p. \tag{2.182}$$

Stability of the numerical method then implies

$$\|E_h^n\|_p \leq c\|E_h^0\|_p + c\Delta t \sum_{m=1}^{n} \|T_h^{m-1}\|_p. \tag{2.183}$$

which leads to

$$\|E_h^n\|_p \leq cn\Delta t T_{max,p}. \tag{2.184}$$

Then consistency of the numerical method leads to

$$\|E_h^n\| = O(\Delta t^{q_1}) + O(\Delta x^{q_2}) \quad \forall n \text{ s.t. } n\Delta t \leq t^*, \tag{2.185}$$

which implies convergence. □

Notes: i) As in the case of elliptic PDEs, the Lax Convergence theorem can be generalized to the Lax Equivalence theorem, which states:

Theorem 2.3 (Lax Equivalence Theorem) *Consider an FD method of the form (2.167) that is consistent in the p-norm with order q_1 in time and q_2 in space. Then the FD method is stable in the p-norm if and only if it is convergent in the p-norm with order q_1 in time and q_2 in space.*

ii) It can be shown that the restrictions

$$\|P^{-1}Q\|_p \leq 1, \quad \text{and} \quad \|P^{-1}\|_p \leq c_p, \tag{2.186}$$

are sufficient for stability of a numerical method. This result follows from applying the inequality

$$\|(P^{-1}Q)^n\|_p \leq \|P^{-1}Q\|^n \quad (\|AB\| \leq \|A\|\|B\|). \tag{2.187}$$

2.4.3 2-Norm Convergence

So far in this chapter we have considered two types of stability: von Neumann stability and stability in the p-norm. We now link these concepts by showing that von Neumann stability is a sufficient condition for stability in the 2-norm for the case of periodic problems.

Theorem 2.4 *Consider a linear FD method of the form* $\frac{\partial u}{\partial t} - Lu = f$ *with L a linear PDE operator with constant coefficients. Then for any IVBVP with periodic BCs,*

$$\|P^{-1}Q\|_2 = \max_k |S(k)|. \tag{2.188}$$

This result can be surprising at first, but consider the following: we have already shown that $\hat{e}^n \exp(ijk\Delta x)$ is an eigenfunction of any linear FD operator with constant coefficients (see Proposition 2.1). It turns out that $S(k)$ is the eigenvalue!

Sketch of Proof: Recall that if A is normal, *i.e.* $AA^T = A^TA$, then $\|A\|_2 = \rho(A)$. Since $P^{-1}Q$ is always normal when the BCs are periodic, the result then follows from knowing that $S(k)$ gives the eigenvalues of $P^{-1}Q$. □

Recall that von Neumann stability requires that $\|S(k)\|_2 \leq 1$. It then follows from (2.186) and Theorem 2.4 that von Neumann stability implies 2-norm stability, subject to $\|P^{-1}\|_2 \leq c_p$.

Example 1: We use the forward upwind discretization for the linear advection problem in 1D with periodic BCs. We have already shown that this method is consistent and von Neumann stable for $\Delta t \leq \frac{\Delta x}{a}$. The method is thus also 2-norm stable, and by the Lax convergence theorem we conclude that this scheme converges in the 2-norm.

Example 2: We use the Crank-Nicolson discretization for the heat diffusion problem in 1D with periodic BCs. Again, we have shown that this method is consistent and always von Neumann stable. By the Lax convergence theorem, we conclude this scheme converges in the 2-norm.

APPENDIX A

Norms of Vectors, Functions and Operators

The study of numerical methods for solving PDEs requires a mathematical tools for measuring the relative size of vectors, functions and operators.

A.1 Vector and Function Norms

Intuitively, we say that a *norm* is a function which can be applied to elements of a vector space in order to introduce a notion of "size" and "distance."

Definition A.1 *Let V be a vector space. Then a* ***norm*** *on V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies*

1) $\|\vec{x}\| \geq 0$ *for all* $\vec{x} \in V$ *and* $\|\vec{x}\| = 0$ *if and only if* $\vec{x} = 0$*,*

2) $\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|$ *for all* $\vec{x} \in V, \alpha \in \mathbb{R}$*,*

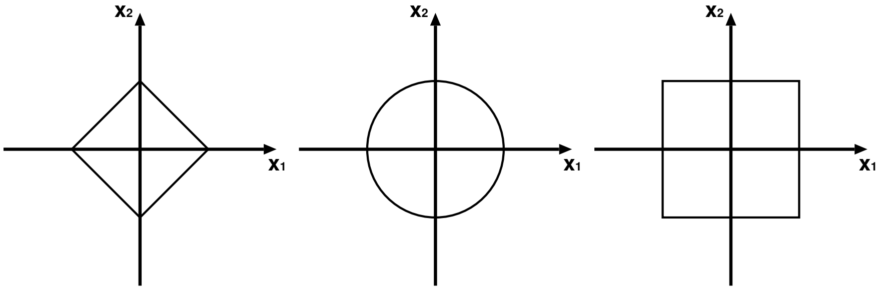
3) $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ *for all* $\vec{x}, \vec{y} \in V$*.*

We now present some common examples of norms.

Example 1 Consider the simple case of $V = \mathbb{R}^2$. It can be verified that for any vector $\vec{x} = (x_1, x_2)$, all of the following functions satisfy the properties of a norm:

$$\|\vec{x}\|_2 = \sqrt{x_1^2 + x_2^2} \qquad \text{(2-norm)}$$
$$\|\vec{x}\|_1 = |x_1| + |x_2| \qquad \text{(1-norm)}$$
$$\|\vec{x}\|_\infty = \max(|x_1|, |x_2|) \qquad (\infty\text{-norm})$$
$$\|\vec{x}\|_p = (|x_1|^p + |x_2|^p)^{\frac{1}{p}} \qquad (p\text{-norm})$$

The choice of norm can significantly change the notion of distance. In the following figure, we depict the *unit circle* in $\mathbb{R}^2$, defined by $\|\vec{x}\|_p = 1$. Here p is given by $p = 1, 2, \infty$ from left to right.



Example 2 Consider the space of real-valued functions $u(x) : [a, b] \to \mathbb{R}$. It can again be verified that all of the following functions satisfy the properties of a norm:

$$\|u\|_2 = \sqrt{\int_a^b u(x)^2 dx} \qquad \text{(2-norm)}$$
$$\|u\|_1 = \int_a^b |u(x)| dx \qquad \text{(1-norm)}$$
$$\|u\|_\infty = \underset{[a,b]}{\text{ess sup}} \; |u(x)| \qquad (\infty\text{-norm})$$

essential supremum

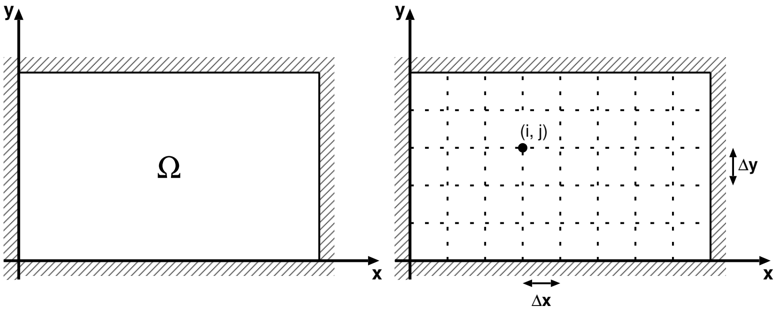
$$\|u\|_p = \left(\int_a^b |u(x)|^p\right)^{\frac{1}{p}} \qquad (p\text{-norm})$$

Example 3 Analogous norms can then be defined as in Examples 2 and 3. For the space of 2-dimensional real-valued functions $u(x, y) : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ over some domain Ω:

$$\|u\|_2 = \sqrt{\iint_\Omega u(x,y)^2 dx} \qquad \text{(2-norm)}$$
$$\|u\|_1 = \iint_\Omega |u(x,y)| dx \qquad \text{(1-norm)}$$
$$\|u\|_\infty = \underset{[a,b]}{\text{ess sup}} \; |u(x,y)| \qquad (\infty\text{-norm})$$
$$\|u\|_p = \left(\iint_\Omega |u(x,y)|^p\right)^{\frac{1}{p}} \qquad (p\text{-norm})$$

A.2 Norms of Grid Functions

Recall that a grid function is a discrete representation or approximation of a continuous function on a grid. As such, it can be represented as a vector, but it also behaves like a function. For example, consider a general 2D function $u(x, y)$ defined on some rectangular region Ω.



We define a grid (x_i, y_j) by $x_i = x_0 + i\Delta x$ and $y_j = y_0 + j\Delta y$, for $i = 0, \dots, m$ and $j = 0, \dots, n$. Then the interpolating grid function $u_{i,j}$ approximating $u(x, y)$ on the grid (x_i, y_j) is defined by

$$u_{i,j} = u(x_i, y_j), \qquad i = 0, \dots, m, \quad j = 0, \dots, n. \tag{A.1}$$

Recall that we have defined the norm $\|u\|_2$ as

$(m+1) \times (n+1)$

$(mn + m + n + 1)$

$$\|u\|_2 = \sqrt{\iint_\Omega u(x, y)^2 dx dy}. \tag{A.2}$$

Using a Riemann sum and applying (A.1), we obtain the approximate formula

$$\|u\|_2 \approx \sqrt{\sum_{i=0}^{n} \sum_{j=0}^{m} u_{i,j}^2 \Delta x \Delta y}. \tag{A.3}$$

This equation then motivates the definition of the 2-norm of our 2-dimensional grid function, given by

$$\|u^h\|_2 = \sqrt{\Delta x \Delta y} \sqrt{\sum_{i=0}^{n} \sum_{j=0}^{m} u_{i,j}^2}. \tag{A.4}$$

It can be easily shown (exercise) that this relation defines a norm on the space of 2-dimensional grid functions on this regular Cartesian grid.

let $M = m+1$, $N = n+1$

map $(i, j) \to k = 0, \dots, M \times N$

then use Cauchy Schwarz

$\left.\sqrt{\Delta x \Delta y}\sqrt{\sum_{k=0}^{M\times N} u_k^2}\ \right\}$ this is the 2-norm of a vector

Analogous to the definition of the 1, 2, ∞ and p norm for functions, we obtain the following expressions for these norms over the space of grid functions on regular Cartesian grids:

Norms of Grid Functions

1D Grid Function Norms:

$$\text{(2-norm)} \qquad \|u^h\|_2 = \sqrt{\Delta x}\sqrt{\sum_{i=0}^{n} u_i^2}, \tag{A.5}$$

$$\text{(1-norm)} \qquad \|u^h\|_1 = \Delta x \left(\sum_{i=0}^{n} |u_i|\right), \tag{A.6}$$

$$(\infty\text{-norm}) \qquad \|u^h\|_\infty = \max_i |u_i|, \tag{A.7}$$

$$\text{(p-norm)} \qquad \|u^h\|_p = \left(\sum_{i=0}^{n} |u_i|^p \Delta x\right)^{\frac{1}{p}}. \tag{A.8}$$

2D Grid Function Norms:

$$\text{(2-norm)} \qquad \|u^h\|_2 = \sqrt{\Delta x \Delta y}\sqrt{\sum_{i=0}^{n}\sum_{j=0}^{m} u_{i,j}^2}, \tag{A.9}$$

$$\text{(1-norm)} \qquad \|u^h\|_1 = \Delta x \Delta y \left(\sum_{i=0}^{n}\sum_{j=0}^{m} |u_{i,j}|\right), \tag{A.10}$$

$$(\infty\text{-norm}) \qquad \|u^h\|_\infty = \max_{i,j} |u_{i,j}|, \tag{A.11}$$

$$\text{(p-norm)} \qquad \|u^h\|_p = \left(\sum_{i=0}^{n}\sum_{j=0}^{m} |u_{i,j}|^p \Delta x \Delta y\right)^{\frac{1}{p}}. \tag{A.12}$$

A.3 Matrix Norms (Operator Norms)

think matrix when considering
linear operator

We now introduce operator norms, which are used in order to quantify the “size” of a linear operator. We concentrate specifically on matrix operators, *i.e.* operators which can be represented in matrix form and applied to vectors in $\mathbb{R}^n$.

Definition A.2 *Let $A \in \mathbb{R}^{m\times m}$ and $\vec{x} \in \mathbb{R}^m$, with associated vector norm $\|\vec{x}\|_p$ on $\mathbb{R}^m$ ($1 \le p \le \infty$). Then the **natural or induced matrix norm** is*

$$\|A\|_p = \max_{\vec{x}\in\mathbb{R}^m} \frac{\|A\vec{x}\|_p}{\|\vec{x}\|_p} = \max_{\vec{x}\in\mathbb{R}^m, \|\vec{x}\|_p=1} \|A\vec{x}\|_p. \tag{A.13}$$

The 2-norm of a matrix A can also be characterized in terms of the spectral radius of the matrix A.

Definition A.3 *Let $A \in \mathbb{R}^{m\times m}$. The **spectral radius of** A, denoted $\rho(A)$, is given by*

$$\rho(A) = \max_{1\le i\le m} |\lambda_i|, \tag{A.14}$$

Think PCA !!

where λ_1, λ_2, $\dots$, λ_m denote the m eigenvalues of A[1]

It can then be shown that the following result holds. Its proof is beyond the scope of this text.

Proposition A.1 *Let $A \in \mathbb{R}^{m\times m}$ with induced matrix norm $\|A\|_2$. Then*

$$\|A\|_2 = \sqrt{\rho(AA^T)} = \sqrt{\rho(A^TA)}. \tag{A.15}$$

The induced matrix norm has the following useful properties:

P_1) If $A = A^T$ then $\|A\|_2 = \rho(A)$.

P_2) The 1-norm $\|A\|_1$ is given by the maximum absolute column sum, *i.e.* if the elements of A are given by a_{ij}, then

$$\|A\|_1 = \max_{1\le j\le m} \left(\sum_{i=1}^{m} |a_{ij}| \right). \tag{A.16}$$

[1] Note that the λ_i may be complex numbers.

P_3) The ∞-norm $\|A\|_\infty$ is given by the maximum absolute row sum, i.e.

$$\|A\|_\infty = \max_{1\le i\le m}\left(\sum_{j=1}^{m}|a_{ij}|\right). \tag{A.17}$$

P_4) For any $1 \le p \le \infty$,

$$\|A\|_p \ge \rho(A). \tag{A.18}$$

P_5) For any $\vec{x} \in \mathbb{R}^m$ we have

$$\|A\vec{x}\|_p \le \|A\|_p\|\vec{x}\|_p. \tag{A.19}$$

P_6) The matrix norm satisfies the triangle inequality,

$$\|A+B\|_p \le \|A\|_p + \|B\|_p. \tag{A.20}$$

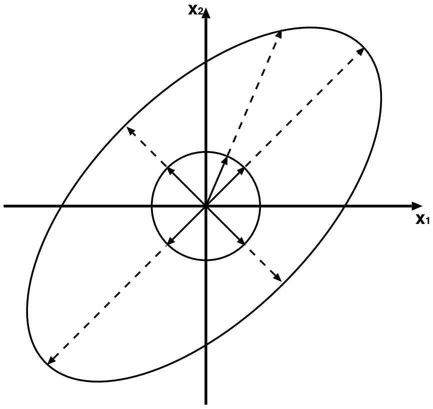
Example Consider the matrix $A \in \mathbb{R}^{2\times 2}$ defined by

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}. \tag{A.21}$$

Using properties P_2 and P_3, it can be quickly shown that $\|A\|_1 = \|A\|_\infty = 4$. Let us now focus on the matrix 2-norm, $\|A\|_2$. It can be quickly verified that the eigenvalues of A are $\lambda_1 = 2$ and $\lambda_2 = 4$, with associated eigenvectors

$$v_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad v_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

In order to determine $\|A\vec{x}\|_2$ for all $\|\vec{x}\|_2 = 1$, we note that $\|\vec{x}\|_2 = 1$ is simply the equation for the unit circle in 2D. Using the eigenvectors and eigenvalues as a guide, it can be shown that under the influence of A, the unit circle is transformed into an ellipse (see figure).



The largest possible stretch factor in this case occurs along v_2 and is given by λ_2. That is, in this example $\|A\|_2 = 4$.

BIBLIOGRAPHY

[1] *Partial Differential Equations*, Evans, Providence, American Mathematical Society, 2002. (PDE theory)

[2] *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*, Leveque, SIAM, 2007. (Excellent text on FD methods)

[3] *Finite volume methods for hyperbolic problems*, Leveque, Cambridge, 2002. (Excellent text on FV methods)

[4] *An introduction to the finite element method*, Reddy, McGraw-Hill, 1993. (FE method, comprehensive introduction with engineering applications)

[5] *The mathematical theory of finite element methods*, Brenner and Scott, Springer, 1994. (FE method theory)

[6] *Finite elements : theory, fast solvers, and applications in solid mechanics*, Braess, Cambridge University Press, 2001. (FE method theory)

[7] *A first course in the numerical analysis of differential equations*, Iserles, Cambridge University Press, 1997. (FD and FE methods)