INTRODUCTION TO COMPUTATIONAL PDES

Course Notes for AM 452

H. De Sterck
P. Ullrich
Department of Applied Mathematics
University of Waterloo

July 26, 2007

These notes have been funded by

CONTENTS

Introduction to Computational PDEs

Partial differential equations (PDEs)

A Note About 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
$\dfrac{\partial u}{\partial x}$	u_x
$\dfrac{\partial^2 u}{\partial y^2}$	u_{yy}
$\dfrac{\partial^4 u}{\partial^2 x \partial y \partial z}$	u_{xxyz}

Further, we use vector calculus notation for the following higher-dimensional operators. In the following table we assume 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 $\mathrm{grad}(u)$	$\nabla u = (u_x, u_y)$	$\nabla u = (u_x, u_y, u_z)$
Divergence of $\vec{v}$	$\nabla \cdot \vec{v}$ or $\mathrm{div}(\vec{v})$	$\nabla \cdot \vec{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y}$	$\nabla \cdot \vec{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$
Laplacian of u	$\triangle u$ or $\nabla^2 u$	$\nabla^2 u = \left(\frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}\right)$	$\nabla^2 u = \left(\frac{\partial^2 u}{\partial x^2}, \frac{\partial^2 u}{\partial y^2}, \frac{\partial^2 u}{\partial z^2}\right)$

Note that the Laplacian of u, $\nabla^2 u$, is defined in terms of the gradient and divergence as

$$\nabla^2 u = \nabla \cdot (\nabla u). \tag{1}$$

CHAPTER 1

Overview of PDEs

Partial differential equations arise in almost every scientific discipline, including fluid mechanics, general relativity, quantum mechanics and biology. As a result, we often desire to find solutions of these equations in order to obtain a more thorough understanding of the given problem. Unfortunately, it is almost always impossible to obtain closed-form solutions of these equations, even in very simple cases. Here computational techniques are an immense asset, as they allow us to obtain approximate solutions to PDEs and hence a greater understanding of their behaviour.

In this chapter we present a general 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*

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

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. $u = 0$). 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

$$u(x,y) = c_1 u_1(x,y) + c_2 u_2(x,y), \tag{1.3}$$

is also a solution of the homogeneous PDE.

Classification of linear second-order PDEs with two independent variables

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 $x, y \in \Omega$, $B^2 - 4AC > 0$.*

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

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

Classification	Partial Differential Equation	Example Solution
Parabolic	$\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0$	$u(x,t) = \exp(-t)\cos(x),\ t > 0$
Hyperbolic	$\dfrac{\partial^2 u}{\partial t^2} - \dfrac{\partial^2 u}{\partial x^2} = 0$	$u(x,t) = \cos(x \pm t)$
Elliptic	$\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} = 0$	$u(x,y) = x + y$

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

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

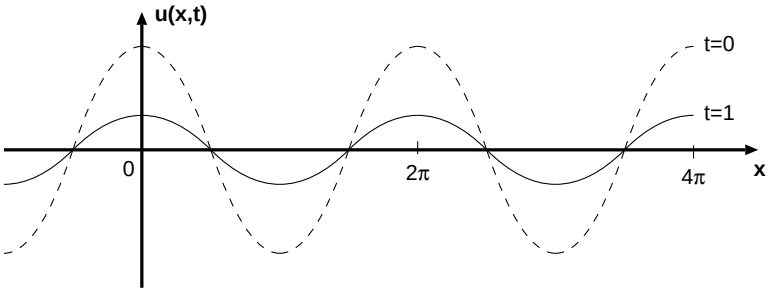
There are many different solutions of this PDE, dependant on the choice of initial conditions and boundary conditions. An example of one such known solution is

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

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 diffuses (*i.e.* spreads out over time) and dissipates (*i.e.* decays in amplitude 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.

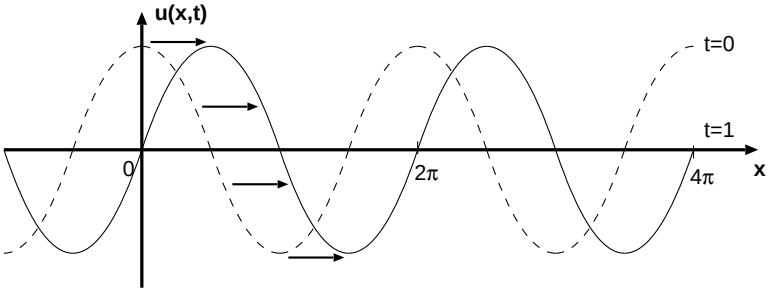
Example 2. The 1D Wave Equation (Hyperbolic Prototype) The 1-dimensional wave equation is given by

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

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 $f(x) = \cos(x)$.



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

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

Normally we consider this equation as applied to 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.

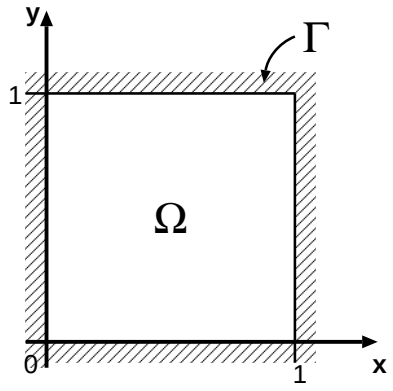
The inhomogeneous form of the Laplace equation is known as the *Poisson equation* and is defined as

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

Consider the following boundary value problem (BVP):

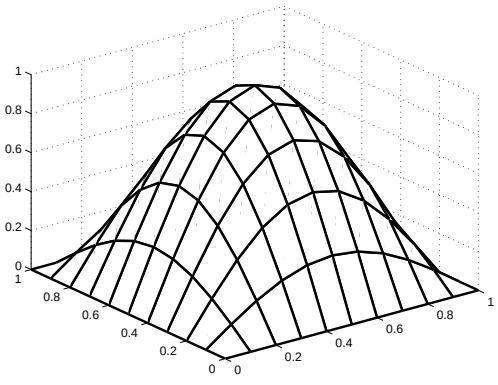
$$\text{BVP} \begin{cases} \Omega = (0,1) \times (0,1), \\ u(x,y) = 0 \text{ on } \Gamma = \partial\Omega, \\ u_{xx} + u_{yy} = -2\pi^2 \sin(\pi x) \sin(\pi y) \quad \text{in } \Omega. \end{cases} \tag{1.12}$$

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



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

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

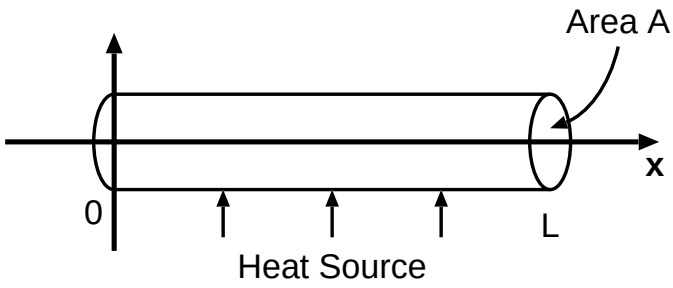


We normally say that a boundary condition of the form $u(x, y) = g$ on Γ is of *Dirichlet type*. Correspondingly, a boundary condition of the form $u(x, y) = 0$ on Γ is of *homogeneous Dirichlet type*.

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

1.1.1 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.

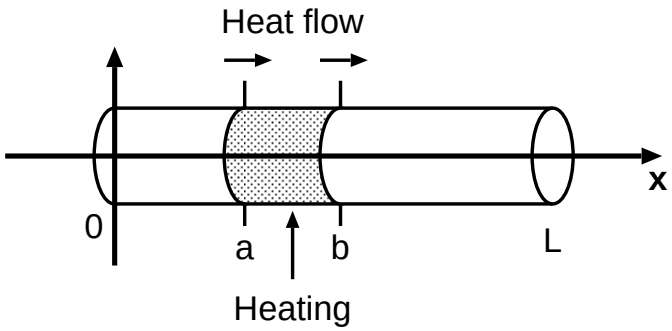
The physical quantities we are interested in are given in the following table.

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	$\frac{\text{energy}}{\text{mass} \cdot \text{temperature}}$	$Jg^{-1}K^{-1}$
$J(x)$	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}
K	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 for this problem, we write (1.14) as

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

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

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

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

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

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

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

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

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

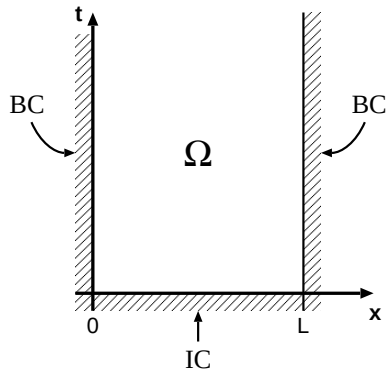
and hence obtain the final form of the heat equation:

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

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 point 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).

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



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

$$\begin{cases} \Omega = (0, L) \times (0, \infty) & \text{(domain)} \\ u(0, t) = u_0(t) & \text{(BC)} \\ u(x, 0) = C & \text{(IC)} \\ u_t - Du_{xx} = f(x) & \text{(PDE)} \end{cases} \tag{1.24}$$

Existence and uniqueness of (1.24) can be shown (see, for example, ??ref.), 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 in Two Independent Variables

We now take a closer look at hyperbolic PDEs. In the following section, we will 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.

1.2.1 The Linear Advection Equation

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

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

We claim that the general solution of (1.25) is

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

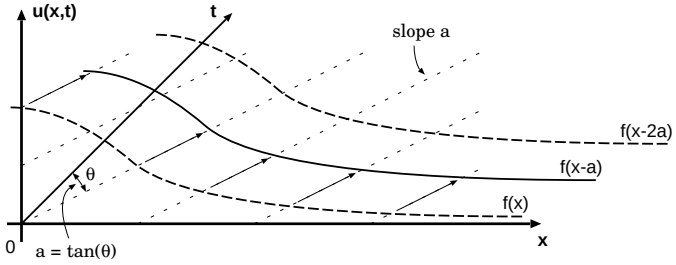
for any function $f(s)$. This result can be proven very easily, as follows.

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

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

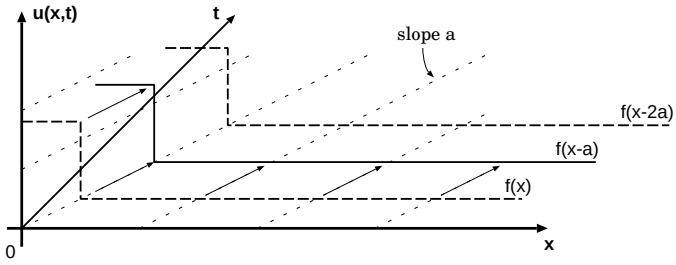
The result follows upon substituting (1.27) into (1.25). □

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, a discontinuous function $f(s)$ is also 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.

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. Our aim is to derive an equation that does not have a preferred direction by composing two linear advection equations of equal advection speed but opposite direction.

Before proceeding, we must introduct the concept of a *linear differential operator*. In general, *differential operator* L is a function of the differentiation operator, *i.e.* given some function f, the operation of applying L to f, which we denote $g = Lf$, is a function composed of the derivatives of f. A special class of differential operators are the *linear differential operators*, which consist of linear combinations of various derivatives. We will focus on linear differential operators in this text, but emphasize that this restriction is not necessary in general.

Consider the linear advection equation (1.25). This equation can be rewritten as a differential operator L applied to some function 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, by

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

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

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

will have both $f(x + at)$ and $f(x - at)$ as solutions. Upon rewriting (1.30) as a PDE, we obtain

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

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. It can be quickly shown (exercise) that the general solution of (1.32) is

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

for arbitrary functions f and g.

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

The unbounded initial value problem (IVP) for the one-dimensional wave equation (1.32) is given by

$$IVP \begin{cases} \Omega : t \in (0,\infty), x \in (-\infty,\infty), \\ u(x,0) = \phi_0(x), u_t(x,0) = \phi_1(x), \\ u_{tt} - u_{xx} = 0 \quad (a=1). \end{cases} \tag{1.34}$$

We already know that the 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 $t = 0$, 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

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

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

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

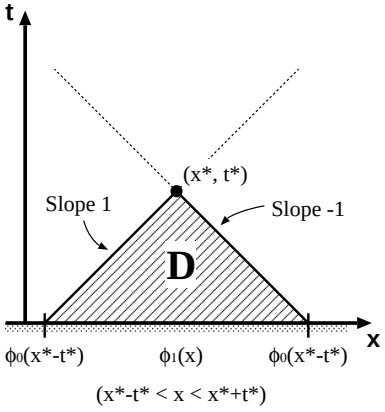
$$u(x,t) = \tfrac{1}{2}\left(\phi_0(x+t) + \phi_1(x-t) + \int_{x-t}^{x+t} \phi_1(\tilde{x}) d\tilde{x}\right). \tag{1.39}$$

This equation is known as *d'Alambert'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)).

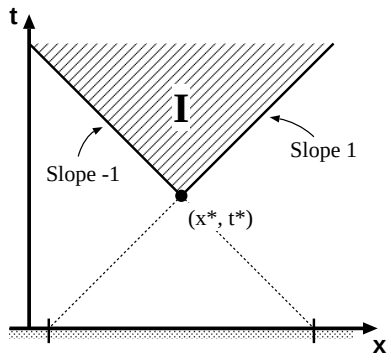
1.2.4 Domain of Influence and Domain of Dependence

We consider the IVP (1.34) and choose some point $(x^*, t^*) \in (-\infty, \infty) \times (0, \infty)$. Then according to d'Alambert'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.



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 the 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 (known as 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).

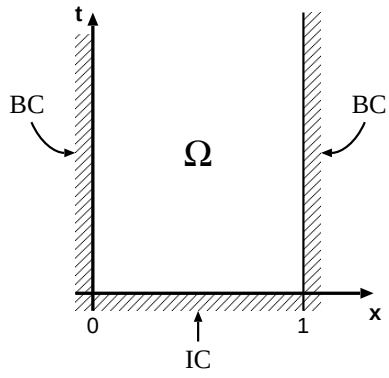
1.2.5 Existence and Uniqueness for the IVBVP

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 gurantee 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:

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

The domain of this problem is given as follows.



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 ??ref.

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

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

Further, 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 phenomenon, such as a sound wave travelling in a closed tube or a plucked string with both ends fixed.

1.3 Elliptic PDEs in Two Independent Variables

We now turn our attention to elliptic PDEs. In the following section, we will derive 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

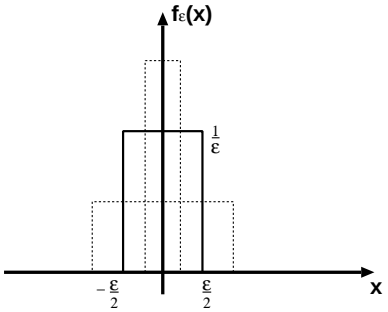
The Dirac delta is a type of "generalized function," used in mathematical modelling and differential equations to represent a physical impulse in the system. In this case, it will be useful in analyzing the behaviour

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 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).

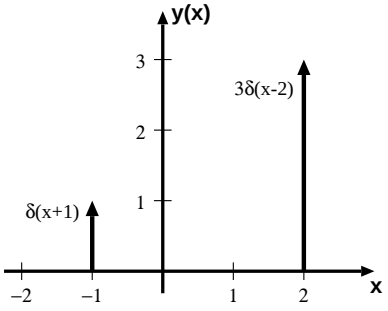
The Dirac delta technically is not a function, but instead fits into a category of operators known as "generalized functions." It has the following properties:

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

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

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

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



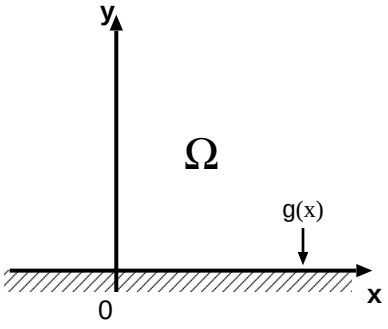
1.3.2 Domain of Influence

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

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

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



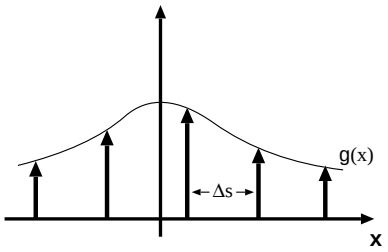
In order to understand the domain of influence for elliptic PDEs, we need to study the influence of one point of the boundary on the solution in the entire domain. We will require the Dirac delta (1.45) to give us the desired results.

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

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

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 switch the order of integration, we obtain

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

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

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 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

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

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

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

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

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

Also, if we integrate along any slice $y =$ constant, it can be shown (exercise) that

$$\lim_{y\to 0^+} \int_{-\infty}^{\infty} u(x,y)dx = 1. \tag{1.57}$$

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."

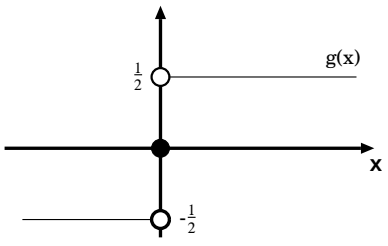
Thus, we have found that any point in an elliptic PDE influences all other points, and hence we cannot use time-marching strategies to solve elliptic problems, *i.e.* we must solve for the whole problem at once.

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

$$g(x) = \begin{cases} \frac{1}{2} & x > 0, \\ 0 & x = 0, \\ -\frac{1}{2} & x < 0. \end{cases} \tag{1.58}$$

This function is depicted in the following plot.



It can be shown that

$$u(x,y) = \tfrac{1}{\pi} \arctan(y/x) \tag{1.59}$$

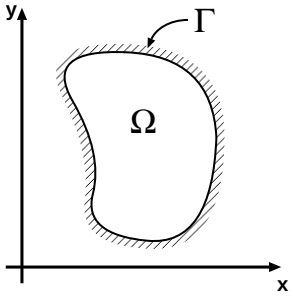
satisfies the PDE (exercise) and satisfies the boundary condition $u(x,0) = g(x)$ in the limit as $y \to 0$. Further, it is easy to see that $u(x,y)$ is continuous in the domain Ω. Using this example, we hypothesize that for linear elliptic PDEs that if $g(x)$ has a finite number of discontinuities then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

$$BVP \begin{cases} \Omega \subset \mathbb{R}^2, \Omega \text{ bounded} \\ u(x,y) = g(x,y) \text{ on } \Gamma = \partial\Omega, \\ u_{xx} + u_{yy} = f(x,y) \text{ in } \Omega. \end{cases} \tag{1.60}$$

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 ??ref for a proof of this result.

1.4 Parabolic PDEs in 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 unbounded initial value problem (IVP) for the heat equation is given by

$$IVP \begin{cases} \Omega : (x,t) \in \mathbb{R} \times (0,\infty) \\ u(x,0) = g(x), \\ u_t - u_{xx} = 0. \end{cases} \tag{1.61}$$

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 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 to be given by a Dirac delta, *i.e.* $g(x) = \delta(x)$. We claim that the solution of the IVP (1.61) is then given by

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \exp\left(\frac{-x^2}{4t}\right). \tag{1.62}$$

This result can be verified with some effort:

First, upon differentiating (1.62), we obtain

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} = \frac{x^2 - 2t}{8\sqrt{\pi t^3}} \exp\left(\frac{-x^2}{4t}\right). \tag{1.63}$$

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

$$\lim_{t \to 0^+} u(x^*, t) = 0. \tag{1.64}$$

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

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

It now remains to show that

$$\lim_{t \to 0^+} \int_{-\infty}^{+\infty} u(x,t)dx = 1. \tag{1.66}$$

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

We conclude that (1.62) satisfies the IVP (1.61). 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 (1.62), 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-reversable*. Consider the time-reversed initial value problem

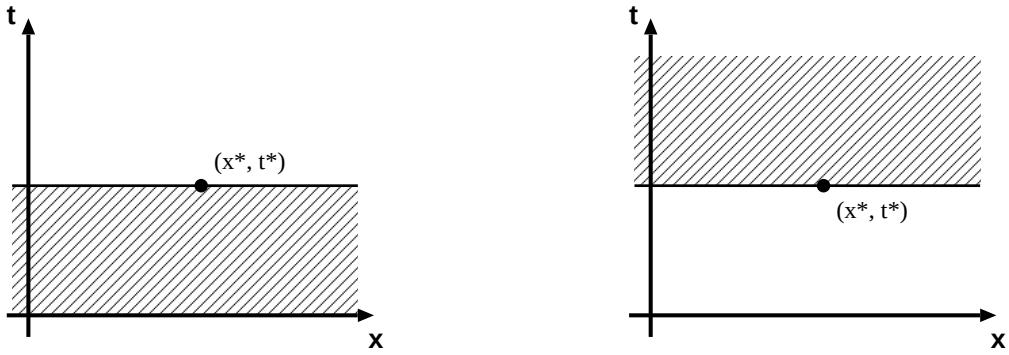
$$IVP \begin{cases} \Omega : (x,t) \in \mathbb{R} \times (-\infty, 0) \\ u(x,0) = g(x), \\ u_t - u_{xx} = 0. \end{cases} \tag{1.67}$$

If we try the function obtained by making the substitution $t \to (-t)$ in (1.62), *i.e.* the function given by

$$\tilde{u}(x,t) = \frac{1}{\sqrt{4\pi(-t)}} \exp\left(-\frac{x^2}{4(-t)}\right), \tag{1.68}$$

we will find that $\tilde{u}(x,t)$ does not satisfy (1.67). 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, but in doing so we must solve for all spatial points simultaneously.



The figures above 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 (1.61). Consider a discontinuous initial condition $g(x)$ defined by

$$g(x) = \begin{cases} \frac{1}{2} & x > 0, \\ 0 & x = 0, \\ -\frac{1}{2} & x < 0. \end{cases} \tag{1.69}$$

It can be shown that

$$u(x,t) = \tfrac{1}{2}\text{erf}\left(\frac{x}{\sqrt{4t}}\right), \quad \text{with} \quad \text{erf}(w) = \frac{2}{\sqrt{\pi}}\int_0^w \exp(-z^2)dz, \tag{1.70}$$

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 $g(x)$ 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 possible. Consider the canonical form of a linear second order PDE with three independent variables:

$$\lambda_1(x,y,t)u_{tt} + \lambda_2(x,y,t)u_{xx} + \lambda_3(x,y,t)u_{yy} = W(u,u_x,u_y,u_t,x,y,t), \tag{1.71}$$

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 (1.71) 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 (1.71) is said to be*

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

ii) ***parabolic*** *if and only if exactly two λ_i have the same sign,*

iii) ***hyperbolic*** *if and only if two λ_i have the same sign and one λ_i is zero.*

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

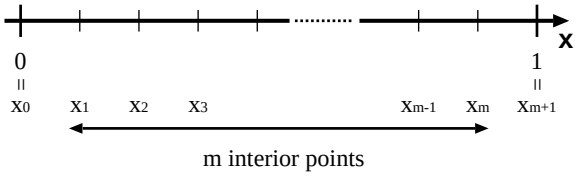
In this section we focus on the finite difference methods for elliptic PDEs, with emphasis placed on the Poisson equation in 1D and 2D. Of particular interest in the theory of numerical methods is *convergence*, *i.e.* in this section we will attempt to explain when a given FD method gives us a solution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

2.1.1 The 1D Elliptic PDE

The normalized 1D elliptic boundary value problem can be formulated as follows.

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

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 $m+1$ distinct 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.* $x_i - x_{i-1} = \Delta x$ for all $i = 1, \ldots, m$.



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 $u''(x)$, according to

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

This choice of discretization scheme follows by 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

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

[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 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 approximation of a continuous function. Here, the h is a generic superscript that denotes a grid function. On using (2.7) and (2.8), we see the elements 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 desire *convergence*. Namely, as $h \to 0$, we want $E^h \to 0$ as well, *i.e.* as the distance between grid points becomes infinitesmally 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 under some norm,[2] the error satisfies

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

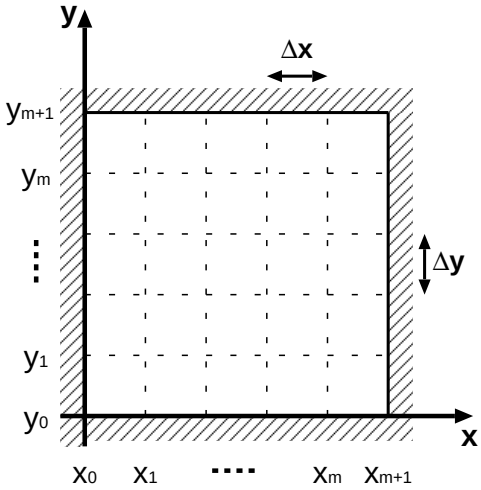
2.1.2 The 2D Elliptic PDE

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

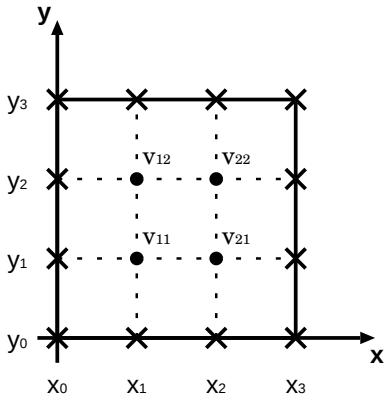
$$\text{BVP} \begin{cases} \Omega : (x, y) \in (0, 1)^2, \\ u(x, y) = g \text{ on } \Gamma = \partial\Omega, \\ u_{xx} + u_{yy} = f(x, y) \quad \text{in } \Omega. \end{cases} \tag{2.14}$$

We discretize Ω into square regions of side length $h = \Delta x = \Delta y$, obtaining $m+2$ points in each direction, $m + 1$ intervals, and m interior points. The result of this discretization is 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 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 first vary over the row index and then over the column 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 that using the system of equations (2.20) allows us to write A^h 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

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

Here we have chosen to absorb the boundary conditions 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 in 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 now introduced two numerical methods for solving the elliptic BVPs, we have sufficient background to study the convergence of 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

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

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 (exercise)

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

for some $\gamma(x_i) \in [x_{i-1}, x_i]$. We note that this result is a consequence of Taylor's remainder theorem.

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** for 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 ($q \in \mathbb{Z}$) 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 let

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

it follows from (2.30) that

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

Then, on taking the 2-norm, we have

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

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

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

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 inverting 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 so that $\|(A^h)^{-1}\|_p \leq c$. This result motivates the following definition.

Definition 2.4 *A numerical method $A^h V^h = F^h$ 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 state as follows.

Theorem 2.1 (Lax Convergence Theorem) *Consider the linear numerical method $A^h V^h = F^h$ 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 calculating $U^h - V^h$, we can instead use the order of the truncation error to derive 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 under 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.

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

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

The proof of this result is left as an exercise for the reader.

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

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

Since this implies that λ^{-1} is an eigenvalue of A, the desired result follows.

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 with the $\cos(\pi h)$ term, we obtain

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

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

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

Clearly, for h sufficiently small, λ_1 satisfies $|\lambda_1| \le \pi^2$ (which is independent of h). We conclude that

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

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.*

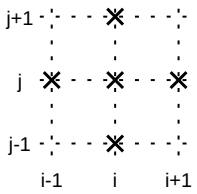
$$\|E^h\|_2 = O(h^2). \tag{2.51}$$

Note that an exact expression can be obtained by substituting (2.34) into (2.37), which gives

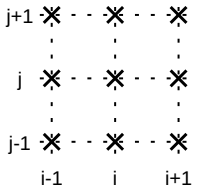
$$\|E^h\|_2 = \frac{h^2}{12\pi^2}c_T. \tag{2.52}$$

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 an arbitrary numerical FD method.

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



Other discretizations can be developed, such as the 9-point weighted central difference discretization depicted as follows. Implemented appropriately, the convergence order of this discretization is $O(h^4)$.



iii) Finally, finite difference formulas can be derived for non-uniform grid spacings, *i.e.* spacing which possibly take advantage of regions of rapid change versus 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 using easily generalized theory, focusing on convergence and stability properties of time-dependent methods and error-propagation. We conclude this section with a discussion of FD methods for the wave equation and extensions of existing 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 with 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

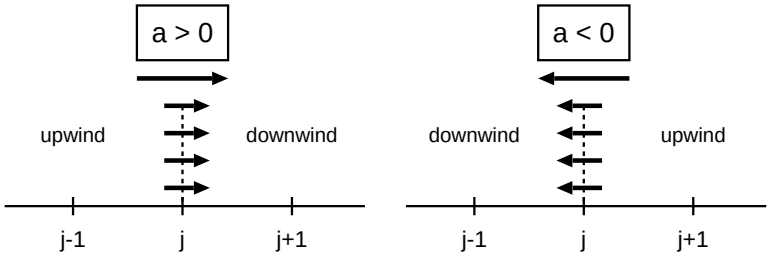
$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \qquad u = u(x,t), \tag{2.53}$$

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 element j, we can divide the state space into two regions depending on the direction of the "wind", *i.e.* the direction that the PDE carries the state variable as time advances. For the advection equation with $a > 0$, the "wind" carries the solution from left to right, and so we describe all elements i that satisfy $i < j$ as "upwind" and all elements i that satisfy $i > j$ as "downwind" (see graphic below).



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

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. This technique is known as the *method of lines.*

For example, using the central discretization, we obtain

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

Instead using an upwind discretization, we obtain

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

The method of lines 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.

Our second step is to add a temporal discretization, of which many options are available. We distinguish between *explicit* schemes and *implicit schemes*. An explicit scheme constructs the state of the system at time level $n+1$ using the known value of the system at time 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 at each timestep. In general, explicit schemes are generally more efficient, *i.e.* they have a smaller memory and computation requirement than implicit schemes, whereas implicit schemes are more stable.

For simplicity, we will focus on three common temporal discretizations:

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

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

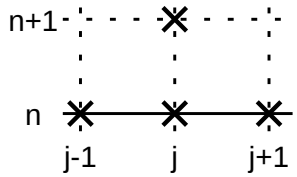
Forward Central Scheme. This scheme uses the central difference formula and forward Euler time discretization. It is written as

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

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}, given as follows:



Although the FC scheme is explicit and hence computationally cheap, it is also unstable. Namely, regardless of our choice of timestep Δt, this method will lead to uncontrolled oscillations that will cause solutions to blow up.

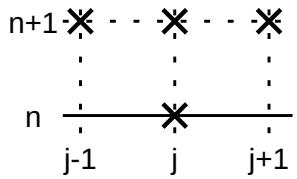
Backward Central Scheme. In order to stabilize the FC scheme, we instead apply a backward time discretization, and hence obtain

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

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, but it is unconditionally stable, *i.e.* it is stable regardless of the choice of Δt.

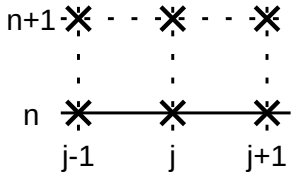
Crank-Nicolson Central Scheme. In order to increase the temporal order we can instead apply the Crank-Nicolson discretization, and hence obtain

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

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

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

and that the stencil is given as follows:



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

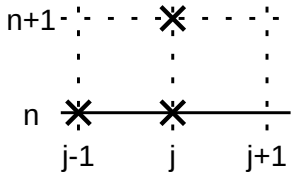
Forward Upwind Scheme. In order to obtain a stable explicit method, we instead apply the spatial upwind discertization, and hence obtain

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

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.

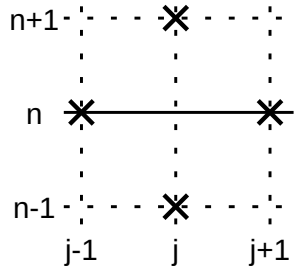
Leapfrog Scheme. We can construct additional temporal discretizations that were not mentioned above, such as the Leapfrog scheme, which uses a central difference time discretization and central difference space discretization. This scheme is then written as

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

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.

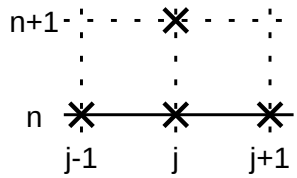
Lax-Wendroff Scheme. The last method we will consider is the Lax-Wendroff scheme, given by

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

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

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

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,

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

and

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

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

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

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

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

On taking the difference between this expression and (2.67), we obtain

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

The methods introduced in this section only encompass a small fraction of possible FD methods for solving PDEs. It should further be emphasized that no single method is the best option for all possible problems. Often simply deciding on the best choice of numerical method for a given problem requires significant research.

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

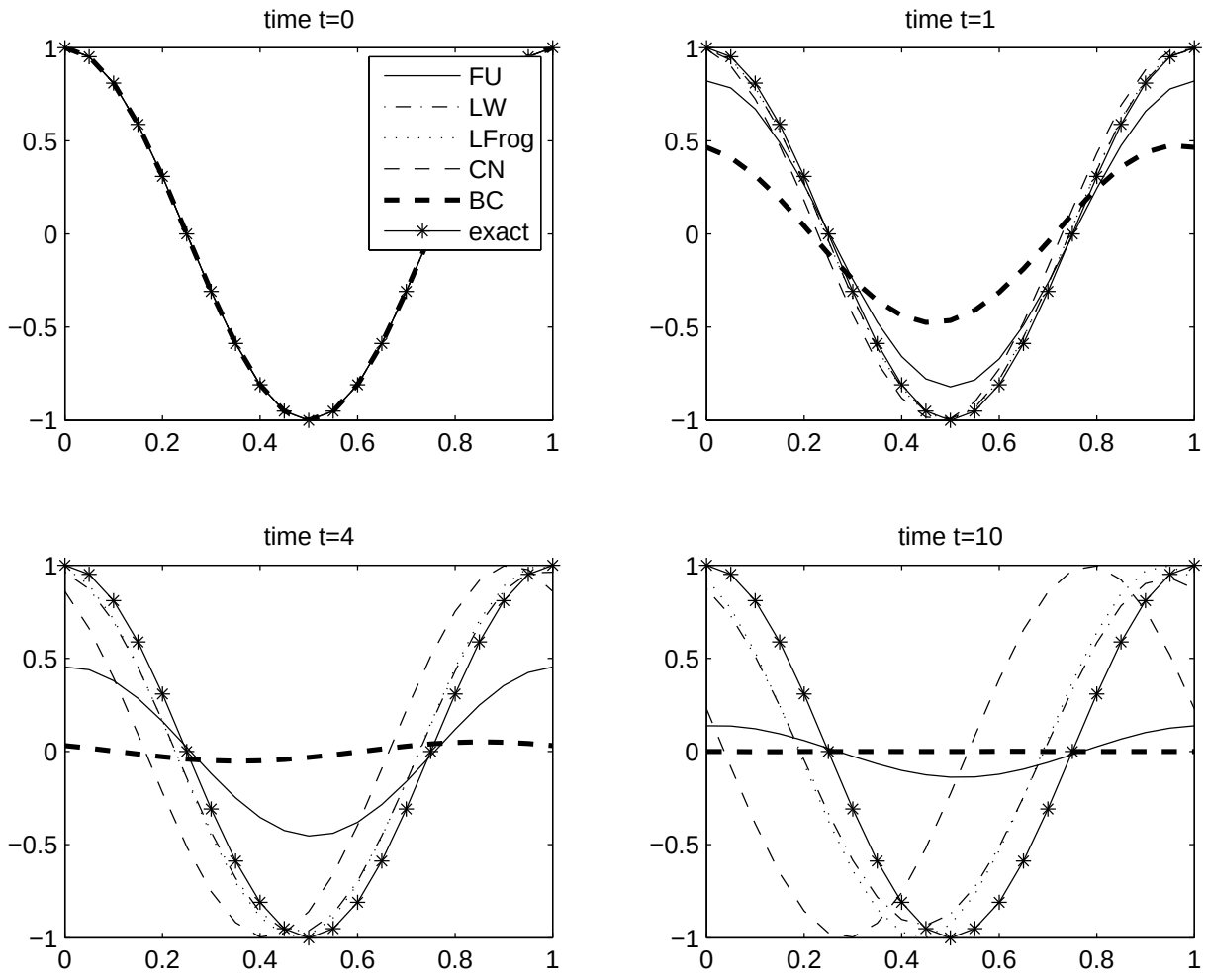


FIGURE 2.1: A comparison of BC, CN, FU, LFrog and LW numerical schemes for the advection equation, applied to a cosine wave with periodic boundary conditions.

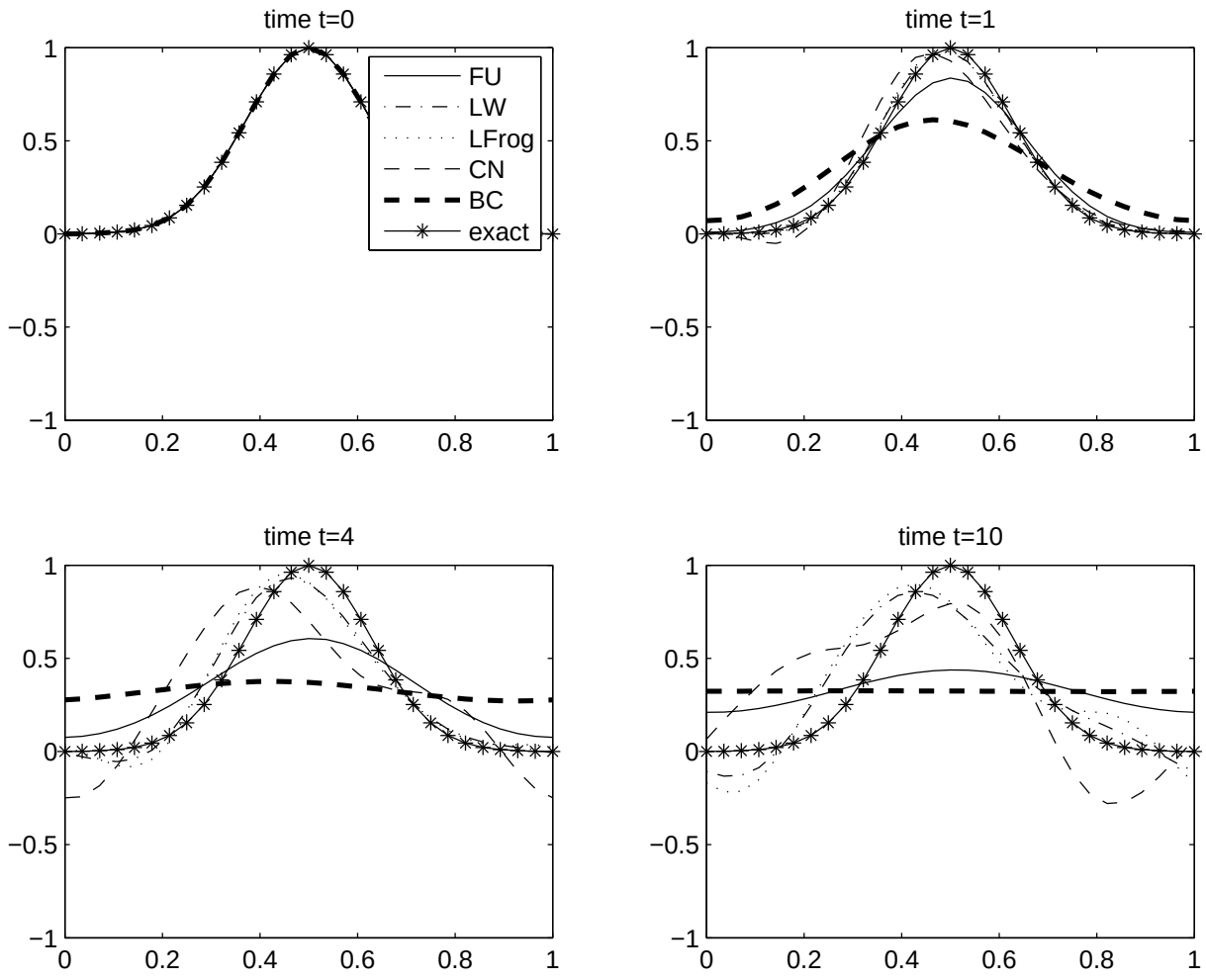


FIGURE 2.2: A comparison of BC, CN, FU, LFrog and LW numerical schemes for the advection equation, applied to a Gaussian profile with periodic boundary conditions.

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 sum 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. 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, as we will show later, stability in this sense is necessary for the stability of the FC scheme.

In order to analyze the stability of (2.75), we will look at an error of wave type. This approach is known as the *von Neumann method* of investigating stability.

An error of wave type can be written as

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

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.

Returning to our example, we substitute (2.78) into (2.75), obtaining

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

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*

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

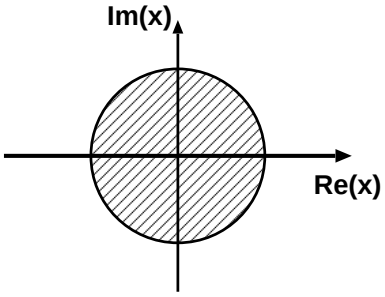
Example As follows from (2.80), the symbol for the forward central scheme (FC) is

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

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. We depict the region of the complex unit circle in the following figure.



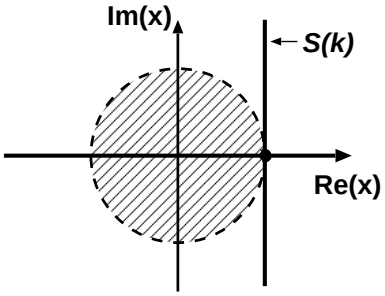
One can easily calculate the value of $|S(k)|$ for the forward central scheme using (2.82), obtaining

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

Hence,

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

for any choice of R. Hence, we have confirmed our earlier result that FC is unstable and further shown that FC will be 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, 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, under the restriction $|1-2R| \leq 1$, 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 Forward Downwind scheme (FD), given by (compare with (2.63))

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

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?)

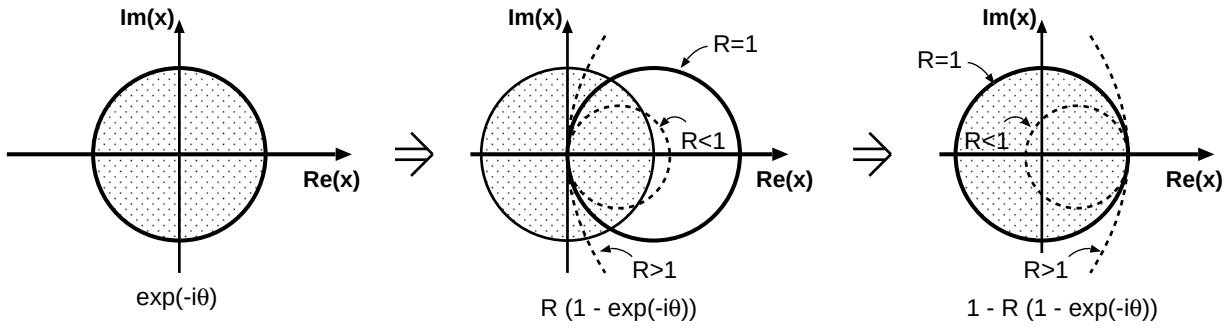
Graphical Techniques for Demonstrating Stability

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 sequence of maps from the real line $\mathbb{R}$ into a 2 dimensional subset of 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.* for any choice of Δt.

Discussion

In this section we have applied the von Neumann stability analysis to the three most 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 in this analysis are given in the following table:

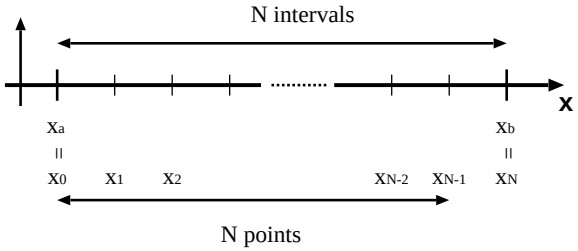
	Scheme	Symbol $S(k)$	Stable?
FC	Forward Central	$1 - iR\sin\theta$	Unstable
BC	Backward Central	$(1 - iR\sin\theta)/(1 + R^2\sin^2\theta)$	Unconditional
FU	Forward Upwind	$1 - R(1 - \exp(-i\theta))$	Conditional (CFL)
CN	Crank-Nicolson	$(2 - iR\sin\theta)/(2 + iR\sin\theta)$	Unconditional
LW	Lax-Wendroff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

Notes: i) The FU scheme 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, as demonstrated:

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

2.2.3 Dissipation and Dispersion

The error introduced by a numerical method typically can be decomposed into *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 potentially rewrite the wave solution (2.103) as

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

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), where

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

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

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

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 three dispersion relations:

$$\omega(k) = ak, \tag{2.107a}$$

$$\omega(k) = ak - iDk^2, \tag{2.107b}$$

$$\omega(k) = ak + \mu k^3. \tag{2.107c}$$

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

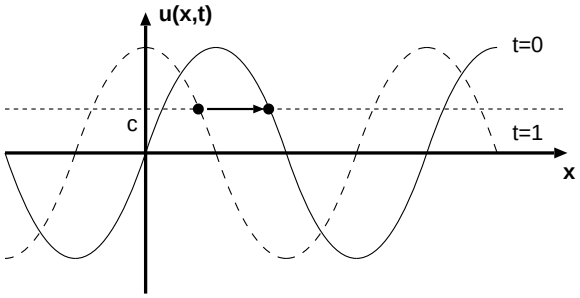
$$\begin{aligned} w(x,t) &= A_0 \exp(i(kx - \alpha(k)t - i\beta(k)t)) \\ &= \underbrace{A_0 \exp(\beta(k)t)}_{A(t)} \exp(i(kx - \alpha(k)t)). \end{aligned} \tag{2.109}$$

This expression clearly distinguishes the role played by the real and imaginary parts of ω: namely, the former, given by $\alpha(k)$ contributes to the speed of the wave. The latter, given by $\beta(k)$ instead 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:

$$L_1 \Rightarrow v_{ph} = \frac{ak}{k} = a, \tag{2.112a}$$

$$L_2 \Rightarrow v_{ph} = k^{-1}\mathrm{Re}(ak - iDk) = a, \tag{2.112b}$$

$$L_3 \Rightarrow v_{ph} = k^{-1}(ak + \mu k^3) = a + \mu k^2. \tag{2.112c}$$

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:

$$L_1 \Rightarrow \beta(k) = 0 \qquad A(t) = A_0, \tag{2.113a}$$

$$L_1 \Rightarrow \beta(k) = -Dk^2 \qquad A(t) = A_0 \exp(-Dk^2 t), \tag{2.113b}$$

$$L_1 \Rightarrow \beta(k) = 0 \qquad A(t) = A_0. \tag{2.113c}$$

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. Then 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 will now show how dissipative and dispersive errors occur 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 $\omega = \omega(k)$. We will return to this fact momentarily.

It turns out that $S(k)$ has all the information we need to determine the existence and strength of numerical dissipation and dispersion. In particular, we seek 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 this 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 then 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 apply $R = a\frac{\Delta t}{\Delta x}$ 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.

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

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

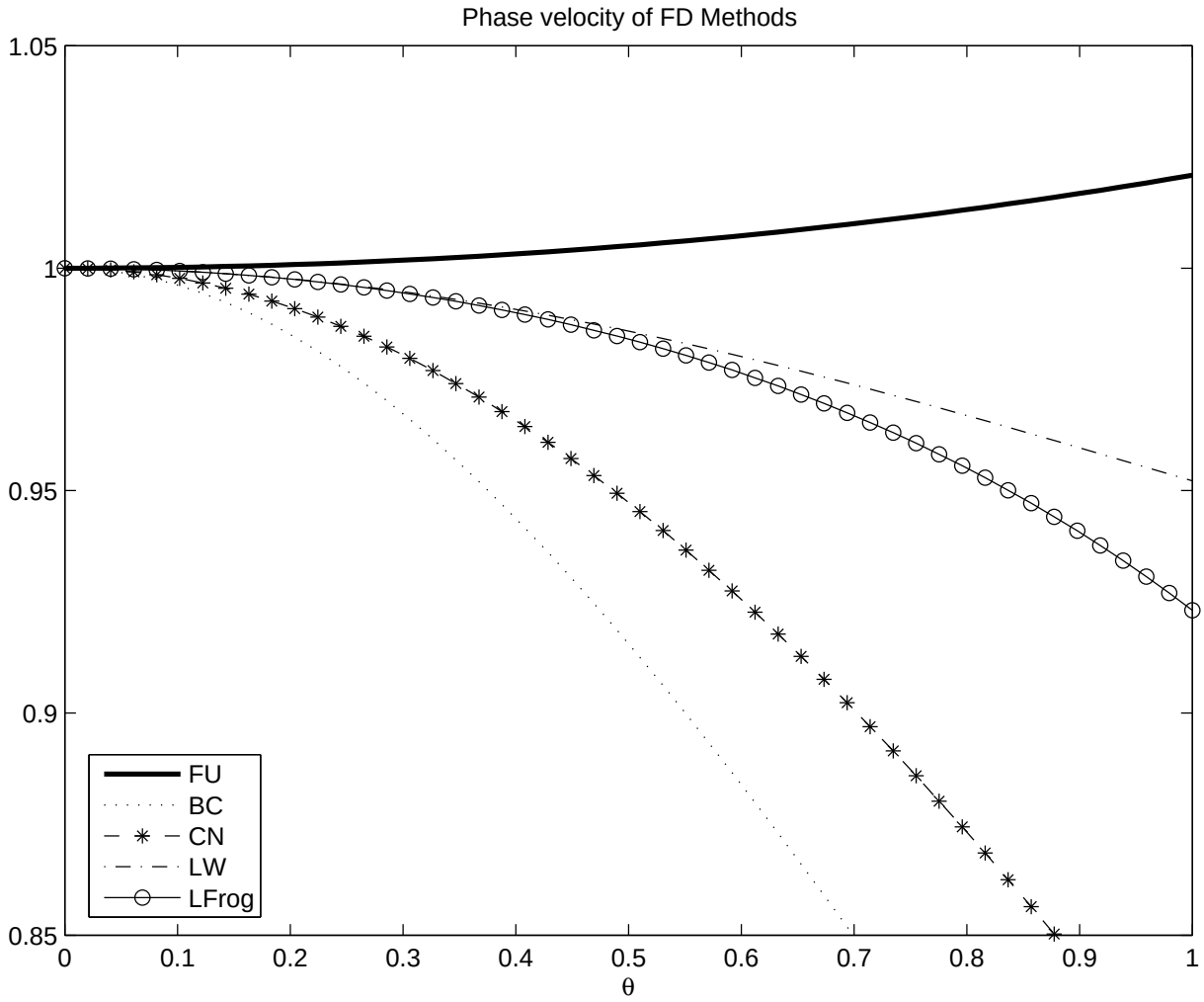


FIGURE 2.3: A comparison of the phase velocity v_{ph} for the BC, CN, FU, LFrog and LW numerical schemes applied to the linear advection equation.

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 element size 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 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 see 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

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

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

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

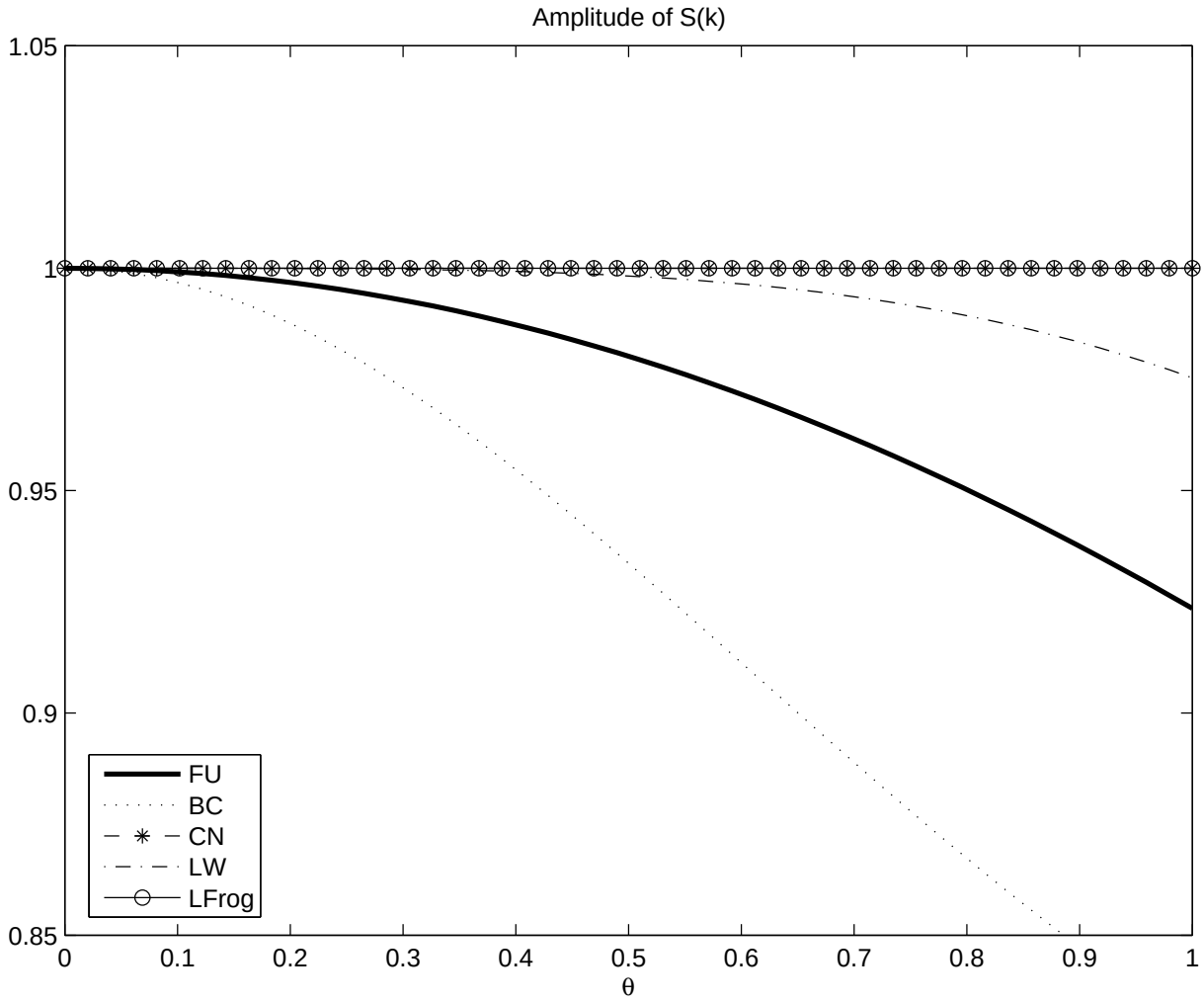


FIGURE 2.4: A comparison of the symbol amplitude $|S|$ for the BC, CN, FU, LFrog and LW numerical schemes applied to the linear advection equation.

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 dominanted by dispersion.

We can also see 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 that the wave equation (1.8) can be written as

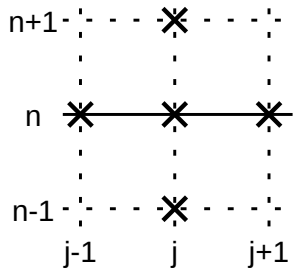
$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \qquad u = u(x,t). \tag{2.130}$$

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 of 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 applyed 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

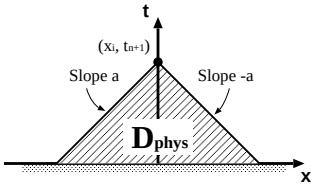
$$\frac{\partial}{\partial t}\begin{bmatrix} u \\ w \end{bmatrix} + \begin{bmatrix} -a & 0 \\ 0 & a \end{bmatrix}\frac{\partial}{\partial x}\begin{bmatrix} u \\ w \end{bmatrix} = \begin{bmatrix} w \\ 0 \end{bmatrix}. \tag{2.136}$$

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

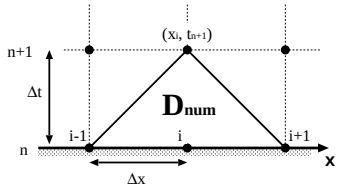
Physical interpretation of CFL condition for explicit methods

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

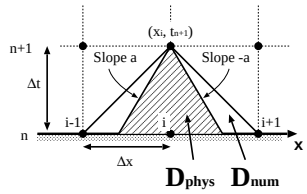
Recall the definition of the (physical) domain of dependence for a hyperbolic PDE, given in section 1.2.4. For the 1D wave equation, the domain of dependence assumes the following form.



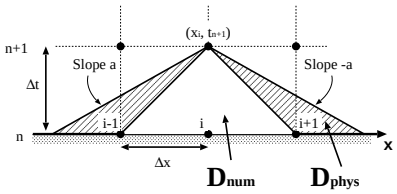
Similar to the physical domain of dependence associated with the PDE, we can also construct a numerical domain of dependence associated with the explicit FD scheme. If the value of v_i^{n+1} only depends on v_{i-1}^n, v_i^n and v_{i+1}^n, then the numerical domain of dependence effectively takes on the 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}. Namely, a given FD method is unstable when the physical evolution of the PDE requires more information than can be obtained from the numerical data.

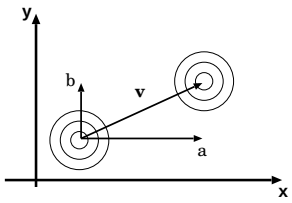
2.2.5 Finite Difference Methods in 2D and 3D

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 will demonstrate this process by example.

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

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

The advection speed is the vector in 2D given by $\mathbf{v} = (a, b)$. Given some initial profile $u(x, y, t = 0)$, we expect that the profile will be 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 the 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 terrible calculation[5], we obtain

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

[5]Thomas, reference needed.

which, on applying some simple identities, reduces to

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

The previous example was a very simple example of extending FD methods to higher dimensions, and clearly allows for 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 the domain of dependence of a hyperbolic PDE is finite in extent at any given time, a parabolic PDE is infinite at any given time. This result suggests that explicit methods are intractable 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. To demonstrate, we 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 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 this 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 increased restrictiveness in simulating them.

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 is unconditionally stable. This method is significantly more practical for simulating the heat problem, but it is implicit and hence requires solving a linear system at 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[6]

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

where the spatial derivative operator L has been detached from the general PDE. We make no assumptions about the dimensionality of the problem.

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

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

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

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 (*i.e.* FU, LW, CN and LW).

Example 1: 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

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

Example 2: 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, this 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.

Example 1: 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: 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 quickly 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*

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

Definition 2.10 *A FD method in the form (2.167) is said to be* ***consistent*** *for 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 ($q_1, q_2 \in \mathbb{Z}$) 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 unbounded 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 (FD) method (2.167) is* ***convergent in the p-norm with order q_1 in time and q_2 in space*** *(to the IVP solution) if and only if*

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

where Δt and Δx may be required to go to 0 in a specific way.

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 are apply Forward Upwind to the linear advection equation, we impose that Δt must satisfy the CFL conditon (2.92).

- If we are applying the Forward Euler discretization to the heat diffusion equation, we impose that Δt 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 (FD) method (2.167) is* ***stable in the*** p***-norm*** *if there exists* c *(independent of* h *and* Δt*) so that*

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

and

$$\|(P_{h,\Delta t}^{-1} Q_{h,\Delta t})^n P_{h,\Delta t}^{-1}\|_p \leq c, \tag{2.176}$$

for all n *and* Δt *so that* $n\Delta t \leq t^*$*, where* Δt *and* Δx *may be required to go to* 0 *in a specific way.*

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) *If an arbitrary FD method of the form (2.167) is consistent in the* p*-norm with order* q_1 *in time and* q_2 *in space, and is stable in the* p*-norm then it 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

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

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

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

We take the p-norm of this result and apply standard identities, obtaining

$$\|E_h^n\| \leq \|(P^{-1}Q)^n\|_p \|E_h^0\|_p + \Delta t \sum_{m=1}^{n} \|(P^{-1}Q)^{n-m} P^{-1}\|_p \|T_h^{m-1}\|_p. \tag{2.182}$$

Stability of the numerical method then implies

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

which leads to

$$\|E_h^n\| \leq c\|E_h^0\|_p + cn\Delta t T_{max,p}. \tag{2.184}$$

Since we choose exact values of $u_0(x)$ at $t = 0$, we have $\|E_h^0\|_p = 0$. Then consistency of the numerical method leads to

$$\|E_h^n\| \leq O(\Delta t^{q_1}) + O(\Delta x^{q_2}), \tag{2.185}$$

as desired. □

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 arbitrary 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 norm identity

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

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 necessary condition for stability in the 2-norm.

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

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 is equivalent to 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 $\Delta t \leq \frac{\Delta x}{a}$. By the Lax convergence theorem, we conclude 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.

CHAPTER 3

Finite Volume Methods for Nonlinear Hyperbolic Conservation Laws

In this chapter we study *finite volume (FV) methods*. These methods stem from the study of physical time-dependent problems and processes, generally in fluid mechanics or gas dynamics. For much of this chapter we will study FV methods applied to the linear advection equation, emphasizing that these methods can be easily generalized to physical systems. In section 3.1 we briefly introduce the notion of characteristic curves of a hyperbolic PDE, as they will be useful later in the study of PDEs. In section 3.2 we introduce conservation laws in 1D, which make up the set of 1D hyperbolic PDEs that are compatible with the finite volume approach. We discuss problems with FD methods in section (3.3), which motivates us to develop the 1D FV methods in section 3.4. Finally, in sections 3.5 and 3.6 we extend conservation laws and the associated FV methods to higher dimensions and consider systems of conservation laws, giving an important example of a physical system that can be solved using the FV approach.

3.1 Characteristic Curves

The characteristic curves of a PDE are important in the study of finite volume methods. In this section we briefly review their theory as applied to the linear advection equation.

Recall that the linear advection equation in 1D takes the form

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \tag{3.1}$$

with right-moving wave solution

$$u(x,t) = f(x - at). \tag{3.2}$$

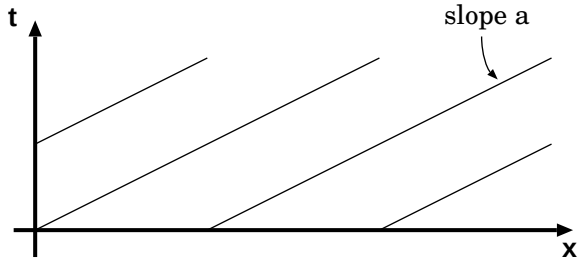
Consider a general curve in the xt-plane described by $x = x(t)$. We are interested in knowing how the exact solution $u(x(t), t)$ changes along this curve. Using the chain rule, we can write

$$\frac{d}{dt}u(x(t),t) = \frac{\partial u}{\partial x}\frac{dx(t)}{dt} + \frac{\partial u}{\partial t}. \tag{3.3}$$

On comparing this equation with (3.1), we are motivated to choose the special curve whose derivative is given by

$$\frac{dx(t)}{dt} = a. \tag{3.4}$$

Observe that (3.4) describes straight lines with slope a. These curves are depicted in the following figure.



On substituting (3.4) into (3.3) we see that

$$\frac{d}{dt}u(x(t),t) = a\frac{\partial u}{\partial x} + \frac{\partial u}{\partial t}, \tag{3.5}$$

which in turn satisfies

$$\frac{d}{dt}u(x(t),t) = 0, \tag{3.6}$$

due to (3.1). The curves (3.4) are called the *characteristic curves* of the linear advection equation.

Formally, a characteristic curve $x(t)$ for a PDE in 1D has the following properties:

1) The solution u is constant along these curves, *i.e.* $\frac{d}{dt}u(x(t),t) = 0$.

2) The PDE changes to an ODE along $x(t)$.

3) Since u is constant, boundary conditions or initial conditions cannot be specified along characteristic

In general, hyperbolic PDEs will have characteristic curves, whereas elliptic PDEs do not. Parabolic PDEs are somewhere in between, *i.e.* in general they will not have a complete set of characteristic curves.

3.2 1D Conservation Laws and the Burgers' Equation

In this section we introduce the concept of a hyperbolic conservation law and derive several theoretical results that relate to PDEs of this form. In particular, we will apply the theory of conservation laws to the study the Burgers' equation, a hyperbolic PDE that originates in the study of waves and is relevant to the study of finite volume methods. We begin this section with the definition of a conservation law.

Definition 3.1 *The differential form of a **conservation law in 1D** for a state variable $u(x,t)$ is a PDE of the form*

$$\boxed{\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0,} \tag{3.7}$$

*where $f(u)$ is called the **flux function** and is an arbitrary function of u.*

Aside from the linear advection equation, which is a trivial example of a conservation law, the *inviscid Burgers' equation* is perhaps the simplest nonlinear case of a conservation law. The Burgers' equation takes the form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0. \tag{3.8}$$

Clearly, (3.8) is a conservation law with flux function

$$f(u) = \tfrac{1}{2}u^2. \tag{3.9}$$

In general, a scalar conservation law (*i.e.* a conservation law in one variable) is always hyperbolic.

3.2.1 Conservation-Integral Forms

We now expand on the theory of conservation laws and introduce the first and second integral form of a conservation law.

One may wonder why an equation of the form (3.7) is called a "conservation law." In order to answer this question, we consider an arbitrary interval $[a,b]$ in 1D. If we integrate the conservation law within this

interval, interchange the integral and derivative operations,[1] and apply the second fundamental theorem of calculus, we obtain

$$\frac{d}{dt}\int_a^b u dx + f(u(b,t)) - f(u(a,t)) = 0. \tag{3.10}$$

If we define

$$Q(t) = \int_a^b u(x)dx, \tag{3.11}$$

then $Q(t)$ is a *conserved quantity* in $[a,b]$, *i.e.* $Q(t)$ will only change when there is a net inflow or outflow through the domain boundaries.[2] This result motivates the following definition.

Definition 3.2 *The **first integral form of a conservation law** is given by*

$$\boxed{\frac{d}{dt}Q(t) = f(u(a,t)) - f(u(b,t)),} \tag{3.12}$$

where $Q(t)$ is defined according to (3.11).

Motivated by our prior success, we can integrate (3.12) over some time interval $t \in [0,T]$, obtaining

$$\int_0^T \frac{dQ}{dt}(t)dt + \int_0^T \left[f(u(a,t)) - f(u(b,t))\right] dt = 0. \tag{3.13}$$

On again applying the second fundamental theorem of calclus, we obtain the following definition.

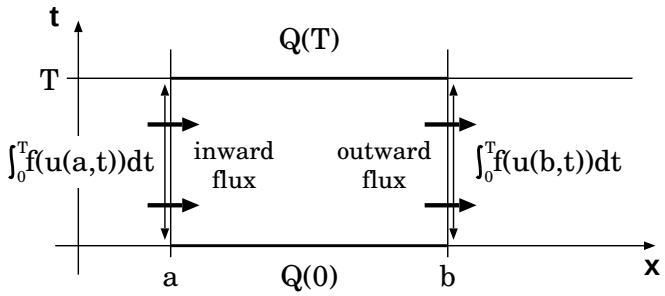
Definition 3.3 *The **second integral form of a conservation law** is given by*

$$\boxed{Q(T) - Q(0) + \int_0^T \left[f(u(a,t)) - f(u(b,t))\right] dt.} \tag{3.14}$$

The second integral form has a straightforward physical interpretation: namely, the total amount of the state variable u in $[a,b]$ between time 0 and T is equal to the difference in the total flux through the boundary from time 0 to time T (see figure).

[1] This interchange is allowed subject to u being C^1 in t.

[2] Recall that in deriving the heat equation in section 1.1.1, we applied a similar technique, obtaining that the total heat energy was a conserved quantity.



3.2.2 Characteristic Curves of the Burgers' Equation

The Burgers' equation (3.8) leads to an interesting phenomenon in its solutions, namely the existence of so-called shock waves and rarefaction waves. In this section we show how the characteristic curves of the Burgers' equation allow us to describe the behaviour of this phenomenon.

Note that on applying the chain rule we can rewrite the Burgers' equation (3.8) as

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0. \tag{3.15}$$

Since characteristic curves $x(t)$ satisfy

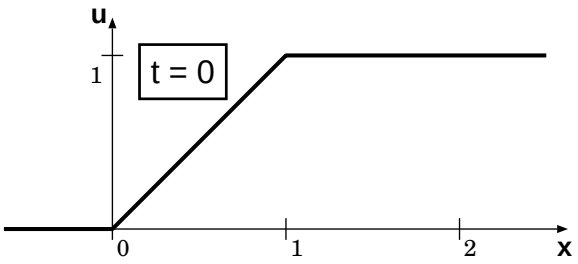
$$\frac{d}{dt}u(x(t),t) = \frac{\partial u}{\partial x}\frac{dx(t)}{dt} + \frac{\partial u}{\partial t} = 0, \tag{3.16}$$

comparing (3.15) with (3.16) leads us to choose

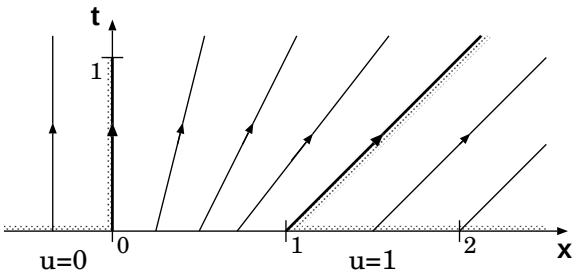
$$\frac{dx(t)}{dt} = u(x(t),t). \tag{3.17}$$

The characteristic curves $x(t)$ of the Burgers' equation must then satisfy this ODE. Since u is constant along the characteristic curves, it follows that the slope of each line must be u, *i.e.* the characteristics are straight lines with slope u (which is constant along each line).

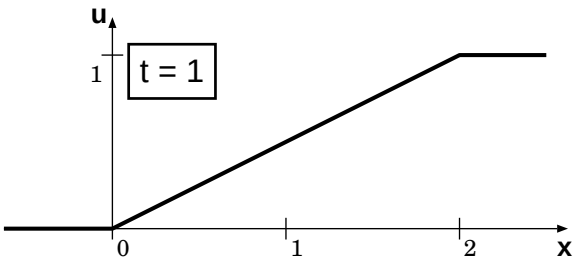
Example 1. Consider an initial profile ($t = 0$) of the following form.



The characteristic curves, when plotted in the xt-plane, are then straight lines whose slope is determined by the initial profile (see figure).

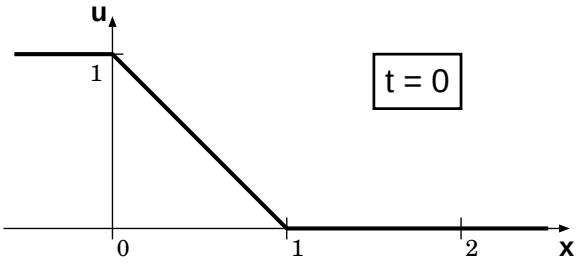


Tracing out the characteristic curves then allows us to draw the profile of u at any future time. For example, at $t = 1$, we obtain the following profile.

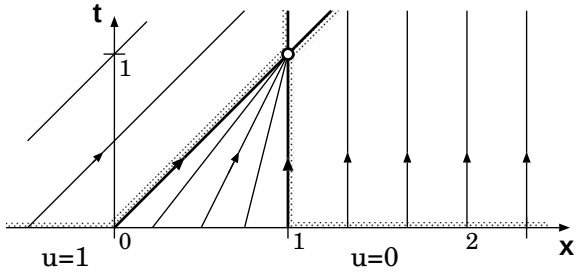


In general this kind of solution is called a *rarefaction wave* solution. This term is drawn from gas dynamics, where a wave of this type is associated with a volume of gas becoming increasingly less dense.

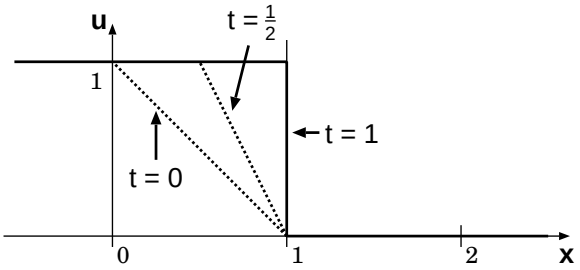
Example 2. Consider an initial profile ($t = 0$) of the following form.



The characteristic curves are depicted in the following figure.



Observe that at $t = 1$ the characteristic curves intersect. Clearly this is problematic, since we expect the solution to have a constant value along each characteristic curve; we can only conclude that the solution takes on all possible values between $[0, 1]$ when $(x, t) = (1, 1)$! On drawing the solution profile up to $t = 1$, we obtain the following sequence of plots.



At $t = 1$ the solution has become discontinuous. This phenomenon is known as *shock formation* and is associated with observable physical behaviour.

This example illustrates an important aspect of nonlinear hyperbolic PDEs, namely that *discontinuities may form from smooth initial conditions in finite time.* Although the differential form of the conservation law (3.7) no longer applies after the formation of a shock, we can use the integral form of the conservation law (3.14) to show that after $t = 1$ the shock will actually travel forward with some shock speed s (this will be shown later).

In general, whenever characteristics intersect, there are two possible outcomes:

1) The solution becomes multivalued. For example, if you have ever been to a beach you have certainly observed *wave breaking* along the shoreline. In this case, nonlinear steepening of the wave causes it to become multivalued, as in the following figure.

This outcome may occur in a physical system, but it is typically very unstable, *i.e.* the multivalued solution will quickly collapse to a single-valued profile.

2) The second possibility is that a *shock wave* forms, *i.e.* a single-valued discontinuous solution. The term "shock wave" is again drawn from fluid mechanics and is essentially what occurs when a supersonic jet passes through the sound barrier. Namely, we obtain a discontinuity in the density of air in front of the shock wave compared to after the shock wave. In this case, the wave appears as in the following profile.

In this text, we will focus on the study of case 2).

3.2.3 Shock Speed: The Rankine-Hugonoit Relation

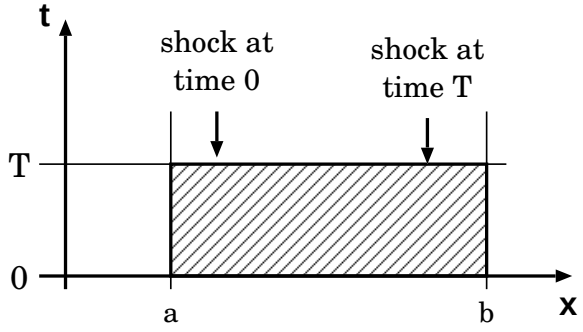
In this section we study the behaviour of a shock wave after it has formed and derive a relationship that gives the shock speed in terms of the shock profile.

Recall that since a shock wave requires a discontinuous profile, the differential form of the DE is no longer valid at the discontinuity. Instead, we must use the integral form.

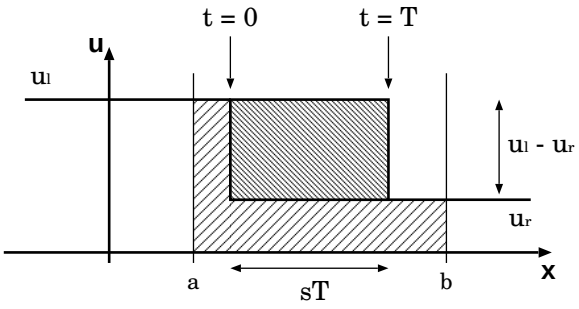
In order to proceed, we must first make some assumptions on the behaviour of the shock wave. We assume that the PDE is governed by a conservation law of the form (3.14) and assume that after $t = 0$ there is a single shock wave propagating rightward with a constant speed s. We define a region Ω in the xt-plane by

$$\Omega = \{(x,t) \in [a,b] \times [0,T]\}\,, \tag{3.18}$$

ensuring that it is sufficiently large to contain the shock for all times $t \in [0,T]$ (see figure).



Further, we use u_ℓ and u_r to denote the state of the system to the left and right of the shock wave, respectively. The shock wave then evolves according to the following figure.



We use $Q(t)$ to denote the amount of material in the interval $[a, b]$ at time t, defined according to (3.11) by $Q(t) = \int_a^b u(x,t)dx$. By inspection, $Q(t)$ must satisfy

$$Q(T) - Q(0) = sT(u_\ell - u_r). \tag{3.19}$$

From (3.14) we also have that

$$Q(T) - Q(0) = \int_0^T [f(u(a,t)) - f(u(b,t))]\, dt = T(f(u_\ell) - f(u_r)), \tag{3.20}$$

since $u(a,t) = u_\ell$ and $u(b,t) = u_r$ for all $t \in [0,T]$. Equating (3.19) and (3.20) and solving for s then leads to the *Rankine-Hugonoit relation* for the shock speed s,

$$\boxed{s = \frac{f(u_\ell) - f(u_r)}{u_\ell - u_r}.} \tag{3.21}$$

Example: Choose $u_\ell = 1$ and $u_r = 0$, subject to the Burgers' equation ($f(u) = \frac{1}{2}u^2$). From the Rankine-Hugonoit relation (3.21) we obtain

$$s = \frac{0 - \frac{1}{2}}{0 - 1} = \frac{1}{2}. \tag{3.22}$$

Thus, the shock propagates rightward with speed $s = \frac{1}{2}$ after $t = 1$.

Notes: i) Characteristic curves are allowed to enter into the shock wave, but cannot emerge from the shock wave. This result is known as the *entropy condition* and is related to the *arrow of time*, *i.e.* entropy increases and information is lost in the shock, not created.

ii) The discontinuous solution is known as a *weak solution* of the PDE. Roughly, this means that it is a solution to the integral form of the PDE (3.14).

3.3 Problems with FD Methods for Hyperbolic Conservation Laws

Although FD methods are perhaps the most straightforward discretization of a PDE, they often introduce several problems when applied to physical systems. In this section we discuss two such problems, the first originating from oscillatory behaviour of some FD methods when applied to discontinuous solutions and the second from an incorrect (non-physical) calculation of shock speeds.

3.3.1 Problem 1: Oscillations when Solution is Discontinuous

Note that the linear advection equation (1.25), given by

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \tag{3.23}$$

can be written as a conservation law (3.7) on making the choice[3]

$$f(u) = au. \tag{3.24}$$

Now consider an initially discontinuous profile on the domain $[0, 1]$, given by

$$u(x) = \begin{cases} 1, \text{ if } x \geq \frac{1}{2}, \\ 0, \text{ if } x < \frac{1}{2}. \end{cases} \tag{3.25}$$

We numerically simulate this profile under the FU and LW finite-difference discretizations, obtaining figure 3.1. Note that the LW scheme introduces what is known as a *Gibbs phenomenon*, *i.e.* an overshoot in the discretization of the profile at the discontinuity.

One can observe that on refining the grid the amplitude of the oscillation at the discontinuity under the LW scheme remains the same. This phenomenon is related to the Fourier series of discontinuous functions, where one can observe a similar result; namely, the Fourier series gives "overshoots" in the presence of a discontinuity.

Recall that for the FU scheme the dominant error term is diffusive, whereas for the LW scheme the dominant error is dispersive. Hence, the discontinuity in the figure is due to interference of wave-like error

[3]Note that a flux function that is linear in u is known as a *linear flux function.*

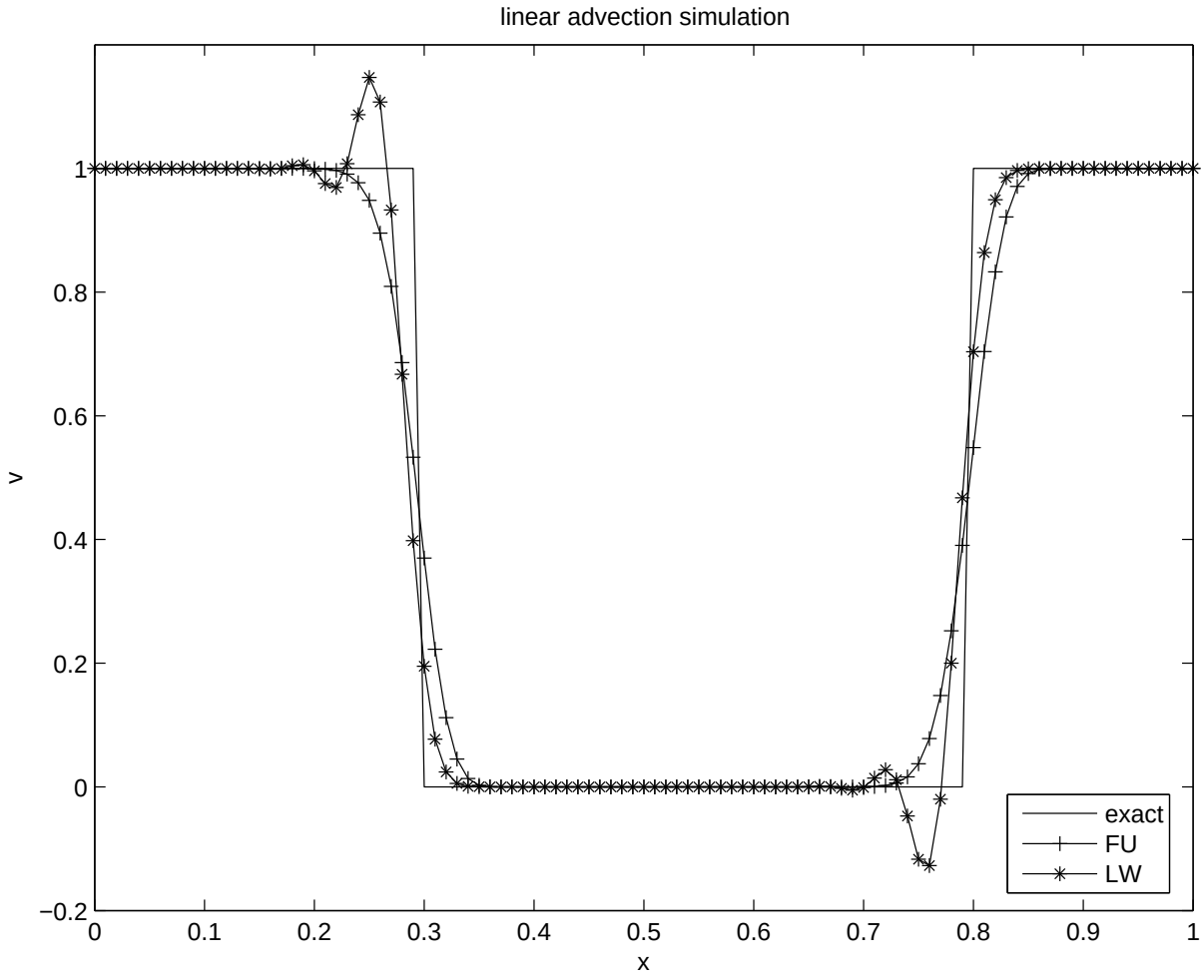


FIGURE 3.1: Advection of an initially discontinuous profile using the forward upwind (FU) and Lax-Wendroff (LW) finite difference methods.

components with different wave lengths. Since standard higher-order methods produce error terms which are dispersive, we conclude that *the standard higher order difference methods must create oscillations at discontinuities.*

Oscillatory effects can be problematic in conservation laws. For example, if u represent the density of a gas, oscillations may cause negative gas densities to be computed.

3.3.2 Problem 2: Standard FD Methods Can Give the Wrong Shock Speeds

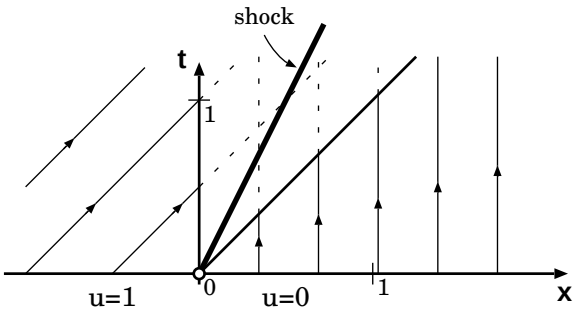
We now turn our attention to FD methods applied to the Burgers' equation (3.8) and consider the emergence of a shock wave. We are interested in an unbounded domain $(x,t) \in (-\infty,\infty) \times (0,\infty)$ with discontinuous initial condition given by

$$u(x) = \begin{cases} 0, \text{ if } x > 0, \\ 1, \text{ if } x \leq 1. \end{cases} \tag{3.26}$$

Recall from the Rankine-Hugonoit relation (3.21) that the shock speed s in this case is given by

$$s = \frac{f(u_r) - f(u_\ell)}{u_r - u_\ell} = \frac{0 - \frac{1}{2}}{0 - 1} = \frac{1}{2}. \tag{3.27}$$

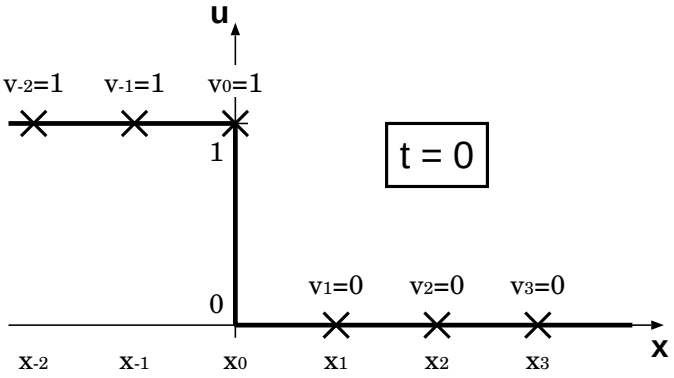
The characteristic curves associated with this initial condition are depicted in the following figure, along with the anticipated location of the shock as it evolves.



We consider a straightforward FD discretization of this PDE given by

$$\frac{v_i^{n+1} - v_i^n}{\Delta t} + v_i^n \frac{v_i^n - v_{i-1}^n}{\Delta x} = 0. \tag{3.28}$$

For theoretical purposes, we assume that the discretization extends infinitely in the spatial direction, with v_0 denoting the value of the numerical solution at $x = 0$. Hence the numerical solution initially satisfies $v_i = 1$ for $i \leq 0$ and $v_i = 0$ for $i > 0$, as depicted below.



We now calculate the discrete solution profile after one time step. On substituting $n = 0$ into (3.28), we obtain

$$v_i^1 = v_i^0 - \frac{\Delta t}{\Delta x} v_i^0 (v_i^0 - v_{i-1}^0), \tag{3.29}$$

and so find that the solution at $t = \Delta t$, denoted v_i^1, satisfies

$$\begin{aligned} i \leq 0 \quad &: \quad v_i^1 = 1 - \tfrac{\Delta t}{\Delta x}(1)(1-1) = 1 = v_i^0, \\ i > 0 \quad &: \quad v_i^1 = 0 - \tfrac{\Delta t}{\Delta x}(0)(\cdots) = 0 = v_i^0. \end{aligned}$$

Observe that at time $t = \Delta t$ the discontinuity has not moved! Note that refining the grid does not help; this FD method simply gives a shock speed of zero.

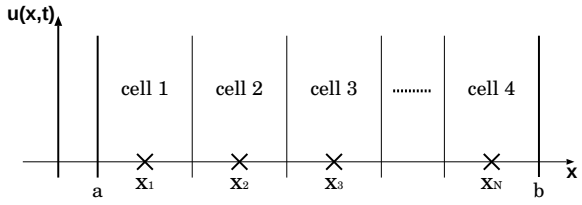
Reason: The FD method (3.28) is based on the differential form of the conservation law (3.7), which is not valid for discontinuous solutions. In order to remedy this problem, we must instead base our numerical methods on the integral form of the conservation law (3.14), as we will now do in deriving the FV methods.

3.4 Finite Volume Methods

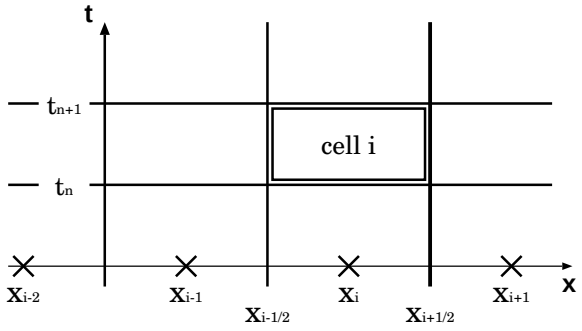
In desiring a set of methods that properly handle conservation laws, we now turn our attention to FV methods. We will now show that the integral form of the conservation law motivates a different approach in developing a method for numerically solving PDEs and give an example of one such approach known as the local Lax-Friedrichs method in 1D.

3.4.1 The Finite Volume Principle

Instead of discretizing the solution at individual points, as with FD methods, we instead divide the spatial domain into *cells of finite volume*. For simplicity, we take the cells to be of equal size Δx, as in the following figure.



This discretization can also be performed in the time domain, leading to the following subdivision of the state space.



Here cell interfaces are denoted by half-integer indices, such as $i + \frac{1}{2}$ or $i - \frac{1}{2}$. The second integral form of the conservation law (3.14) can then be rewritten using the cell from this discretization as

$$Q_i^{n+1} - Q_i^n + \int_{t_n}^{t_{n+1}} \left[f(u(x_{i+\frac{1}{2}}, t)) - f(u(x_{i-\frac{1}{2}}, t)) \right] dt = 0, \tag{3.30}$$

with

$$Q_i^n = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x, t_n) dx. \tag{3.31}$$

We now define $\overline{u}_i^n$, the average value of $u(x, t)$ in cell i at time t_n, by

$$\overline{u}_i^n = \frac{Q_i^n}{\Delta x}, \tag{3.32}$$

and $\overline{f}^{n+\frac{1}{2}}_{i+\frac{1}{2}}$, the average value of $f(u)$ at the interface $i+\frac{1}{2}$ between t_n and t_{n+1} by

$$\overline{f}^{n+\frac{1}{2}}_{i+\frac{1}{2}} = \frac{\int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}},t))dt}{\Delta t}. \tag{3.33}$$

This process then leads to the following definition, on using (3.30).

Definition 3.4 *The **third integral form of the conservation law** is defined by*

$$\boxed{\frac{\overline{u}_i^{n+1} - \overline{u}_i^n}{\Delta t} + \frac{\overline{f}^{n+\frac{1}{2}}_{i+\frac{1}{2}} - \overline{f}^{n+\frac{1}{2}}_{i-\frac{1}{2}}}{\Delta x} = 0.} \tag{3.34}$$

We note that (3.34) is an exact equation, since it simply follows from rewriting the (exact) second integral form.

Equation (3.34) is easily discretized on making the approximations

$$\overline{u}_i^n \approx v_i^n, \qquad \overline{f}^{n+\frac{1}{2}}_{i+\frac{1}{2}} \approx f^*(v_i^n, v_{i+1}^n). \tag{3.35}$$

Here, $f^*(v_i^n, v_{i+1}^n)$ is called the *numerical flux function*, and is often denoted in shorthand by $f^*(v_i^n, v_{i+1}^n) = f^{*n}_{i+\frac{1}{2}}$. Note that this approximation assumes that the flux through the interface $i+\frac{1}{2}$ can be reconstructed using the state of the system in cell i and $i+1$.

We can now construct our difference equation using (3.34) and (3.35), finding that for an explicit FV method, the difference equation takes the form

$$\boxed{\frac{v_i^{n+1} - v_i^n}{\Delta t} + \frac{f^*(v_i^n, v_{i+1}^n) - f^*(v_{i-1}^n, v_i^n)}{\Delta x} = 0.} \tag{3.36}$$

The choice of flux function $f^*(v_i^n, v_{i+1}^n)$ then distinguishes between different FV methods.

Notes: i) Any flux function f^* must satisfy the *consistency requirement*,

$$f^*(v,v) = f(v), \ \forall\, v \in \mathbb{R}. \tag{3.37}$$

ii) An implicit FV method can be obtained by instead evaluating the flux function at time $n+1$. It follows that the implicit difference equation takes the form

$$\frac{v_i^{n+1} - v_i^n}{\Delta t} + \frac{f^*(v_i^{n+1}, v_{i+1}^{n+1}) - f^*(v_{i-1}^{n+1}, v_i^{n+1})}{\Delta x} = 0. \tag{3.38}$$

iii) In order to obtain higher accuracy in our discretization, we use more points to reconstruct the flux function at the interface, *i.e.*

$$\overline{f}^{n+\frac{1}{2}}_{i+\frac{1}{2}} \approx f^*(v^n_{i-1}, v^n_i, v^n_{i+1}, v^n_{i+2}).$$

3.4.2 The Local Lax-Friedrichs Method in 1D

Note that using the chain rule, a 1D conservation law (3.7) can be rewritten in the form

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0 \quad \Longleftrightarrow \quad \frac{\partial u}{\partial t} + \frac{df(u)}{du}\frac{\partial u}{\partial x} = 0. \tag{3.39}$$

We define $\lambda(u) = \frac{df(u)}{du} = f'(u)$, and so write the conservation law as

$$\frac{\partial u}{\partial t} + \lambda(u)\frac{\partial u}{\partial x} = 0. \tag{3.40}$$

Now compare this equation with the linear advection equation (3.1). In (3.40), $\lambda(u)$ plays the role of the linear advection speed a and hence is like a non-linear wave speed. Further, since u is constant along characteristic curves $\lambda(u)$ represents the slope of these characteristics. The non-linear wave speed $\lambda(u)$ motivates us to define the following flux function.

The *local Lax-Friedrichs flux function* is defined by

$$\boxed{f^*(v^n_i, v^n_{i+1}) = \frac{f(v^n_i) + f(v^n_{i+1})}{2} - \tfrac{1}{2}\left|\lambda\left(\frac{v^n_i + v^n_{i+1}}{2}\right)\right|(v^n_{i+1} - v^n_i).} \tag{3.41}$$

Example: When applied to the Burgers' equation ($f(u) = \frac{1}{2}u^2$ and $\lambda(u) = \frac{df(u)}{du} = u$), (3.41) reads

$$f^*(v^n_i, v^n_{i+1}) = \frac{\frac{1}{2}(v^n_i)^2 + \frac{1}{2}(v^n_{i+1})^2}{2} - \tfrac{1}{2}\left|\frac{v^n_i + v^n_{i+1}}{2}\right|(v^n_{i+1} - v^n_i), \tag{3.42}$$

and

$$f^*(v^n_{i-1}, v^n_i) = \frac{\frac{1}{2}(v^n_{i-1})^2 + \frac{1}{2}(v^n_i)^2}{2} - \tfrac{1}{2}\left|\frac{v^n_{i-1} + v^n_i}{2}\right|(v^n_i - v^n_{i-1}), \tag{3.43}$$

Stability, Accuracy and Consistency

Under the Lax-Friedrichs scheme, the stability bound for each FV cell is given by a CFL condition of the form

$$\Delta t_i \leq \frac{\Delta x}{|\lambda(v_i)|}. \tag{3.44}$$

This restriction is comparable to the CFL condition for FD methods (2.92) since $\lambda(u)$ is like a local wave speed. If we require a single time-step for the whole simulation domain, we need to take the maximum Δt calculated from (3.44), *i.e.* we require

$$\Delta t = \min_i \frac{\Delta x}{|\lambda(v_i)|}. \tag{3.45}$$

Note: Although the standard CFL conditions (2.92) and (3.45) are absolute limits on stability, it is usually preferrable to include an additional *safety factor*, especially for non-linear problems. For instance, we may choose

$$\Delta t = c\frac{\Delta x}{|a|}, \tag{3.46}$$

with $c = 0.9$.

The accuracy of the Lax-Friedrichs flux function can be shown to satisfy

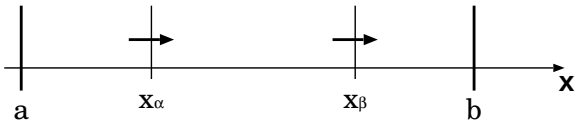
$$T_j^n = O(\Delta t) + O(\Delta x). \tag{3.47}$$

Further, one can easily verify (exercise) that the method is consistent, according to (3.37).

3.4.3 Numerical Conservation

The defining property of FV methods is something known as *numerical conservation*, which we now discuss.

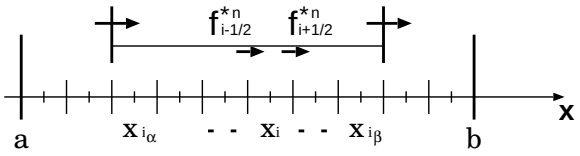
Consider some domain $(x, t) \in [a, b] \times [0, \infty)$ with $[x_\alpha, x_\beta] \subset [a, b]$ any subinterval of the spatial domain (see figure).



Intuitively, conservation in the continuous sense implies that for any $[x_\alpha, x_\beta] \subset [a, b]$ and $[t_n, t_{n+1}] \subset [0, \infty)$, the exact solution must satisfy

$$\int_{x_\alpha}^{x_\beta} u(x, t_{n+1})dx - \int_{x_\alpha}^{x_\beta} u(x, t_n)dx = \int_{t_n}^{t_{n+1}} f(u(x_\alpha, t))dt - \int_{t_n}^{t_{n+1}} f(u(x_\beta, t))dt. \tag{3.48}$$

If we instead consider some partition of the domain into discrete regions (see figure below), we obtain an analogous concept in the discrete case.



We say that a numerical method is conservative in the discrete sense if for all $x_{i_\alpha}, x_{i_\beta}$ it satisfies

$$\sum_{i=i_\alpha}^{i_\beta}(v_i^{n+1}\Delta x) - \sum_{i=i_\alpha}^{i_\beta}(v_i^n \Delta x) = (f^{*n}_{i_\alpha - \frac{1}{2}}\Delta t) - (f^{*n}_{i_\alpha + \frac{1}{2}}\Delta t), \tag{3.49}$$

where $f^{*n}_{i+\frac{1}{2}}$ is some numerical flux function. Namely, this condition implies that in any interval, all discrete fluxes within the domain (interior fluxes) cancel out. It can be shown that any FV method (3.36) exhibits this discrete conservation property (exercise), and hence FV methods are often referred to as *conservative methods*. In this sense, any FV scheme is simply a conservative FD scheme.

Theorem 3.1 (Lax-Wendroff Theorem) *If a conservative FD method (FV method) converges to a solution $v(x,t)$ as the grid is refined, then this $v(x,t)$ has to be a weak solution of the conservation law (with shocks travelling at the right speed).*

In other words, this theorem simply states that conservative FD methods (*i.e.* FV methods) cannot give the wrong shock speeds. This result solves the second problem of FD methods (see section 3.3.2).

3.4.4 FV Methods and the Linear Advection Equation

We now compare the forward upwind (FU) method introduced in section 2.2 and the Lax-Friedrichs FV method when applied to the advection equation. Recall that the exact flux function for the advection equation is given by (3.24).

FD Approach: The forward upwind (FU) method (2.63) can be written in the form

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} + \frac{(av_j^n) - (av_{j-1}^n)}{\Delta x} = 0. \tag{3.50}$$

Comparing this equation with (3.36), we find

$$f^{*n}_{i+\frac{1}{2}} = f^*(v_i^n, v_{i+1}^n) = av_i^n, \tag{3.51}$$

$$f^{*n}_{i-\frac{1}{2}} = f^*(v_{i-1}^n, v_i^n) = av_{i-1}^n, \tag{3.52}$$

and so conclude that the FU method is a conservative FD method for the linear advection equation with flux function

$$f^*(v_i^n, v_{i+1}^n) = v_i^n. \tag{3.53}$$

We refer to (3.53) as the *upwind numerical flux*.

Note: It can be shown that the Lax-Wendroff (LW) scheme is also a conservative FD method, and hence a FV method; however, as we have seen in section 3.3.1, the LW method also creates oscillations at discontinuities and so is generally not used.

FV Approach: On applying the numerical flux function from the Lax-Friedrichs scheme (3.41) to the advection equation, we obtain

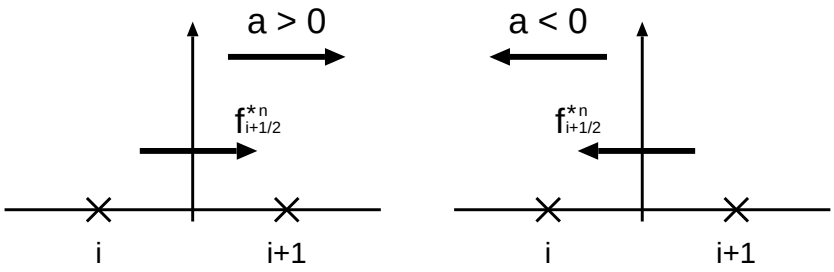
$$f_{i+\frac{1}{2}}^{*n} = \frac{av_i^n + av_{i+1}^n}{2} - \tfrac{1}{2}|a|(v_{i+1}^n - v_i^n). \tag{3.54}$$

Given the sign of a, we can simplify this expression, as follows:

$$\textbf{Case 1: } a > 0 \qquad f_{i+\frac{1}{2}}^{*n} = av_i^n, \tag{3.55}$$

$$\textbf{Case 2: } a < 0 \qquad f_{i+\frac{1}{2}}^{*n} = av_{i+1}^n. \tag{3.56}$$

Graphically, the flux function in each case represents a flow, as depicted in the following figure.



Observe that the LF numerical flux function for the advection equation (3.54) automatically gives the upwind flux regardless of the sign of a, *i.e.* it automatically reduces to the FU method. On substituting (3.54) into (3.36), we obtain[4]

$$\frac{v_i^{n+1} - v_i^n}{\Delta t} + \underbrace{a\frac{v_{i+1}^n - v_{i-1}^n}{2\Delta x}}_{\text{Central discretization}} = \underbrace{\frac{|a|\Delta x}{2}\frac{v_{i+1}^n - 2v_i^n + v_{i-1}^n}{\Delta x^2}}_{\text{Numerical dissipation}}. \tag{3.57}$$

[4]Compare with (2.129).

Similar to (3.54), the first term gives a central discretization and the second term gives numerical dissipation. As we have seen in our analysis of the advection equation, the numerical dissipation term is needed for stability – a fact which generalizes from the linear advection equation to arbitrary flux functions.

Note: Given a conservative FD method in the form (3.57), one can obtain a flux function $f^{*n}_{i+\frac{1}{2}}$ on rewriting the equation and comparing with (3.36). Further, if a given scheme has been applied to the linear advection equation, one can often generalize it to an arbitrary PDE by replacing all av^n_i terms with $f(v^n_i)$ and $|a|$ with $\lambda(\frac{1}{2}(v^n_i + v^n_{i+1}))$.

3.5 Conservation Laws in Higher Dimensions

In this section, we extend the FV methods discussed in section 3.4 to higher dimensions. The fundamental result that allows us to make this generalization is Gauss' divergence theorem, which we discuss in section 3.5.1. The concepts behind generalizing FV methods in 1D to 2D can be intuitively extended to higher dimensions, and so we use conservation laws in 2D as a fundamental example of this technique and leave generalization to the reader.

3.5.1 Gauss' Divergence Theorem

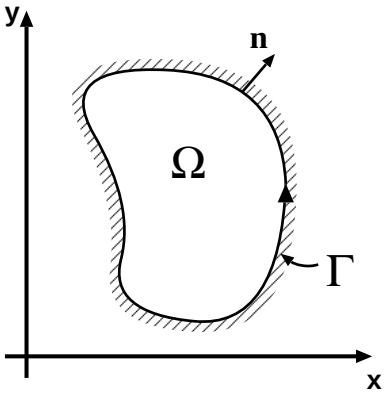
We now state Gauss' divergence theorem for domains of arbitrary dimension and give an example of the theorem in action.

Theorem 3.2 *Let $\Omega \subset \mathbb{R}^n$ be a compact region (closed and bounded) with a piecewise smooth boundary $\Gamma = \partial\Omega$. If $\mathbf{v}$ is a continuously differentiable vector field defined on $\mathbb{R}^n$ then*

$$\int_\Omega \nabla \cdot \mathbf{v} dV = \oint_{\partial\Omega} (\mathbf{v} \cdot \mathbf{n}) d\ell, \tag{3.58}$$

where $\mathbf{n}$ is the outward unit normal of $\partial\Omega$.

For example, one might choose the domain Ω as in the following figure.

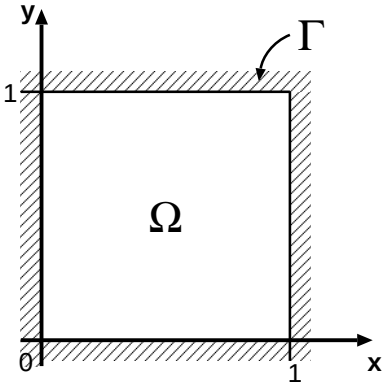


Note: The divergence theorem is a generalization of the second fundamental theorem of calculus,

$$\int_a^b \frac{df}{dx} dx = f(b) - f(a), \tag{3.59}$$

which we used in deriving the integral form of the conservation law in 1D.

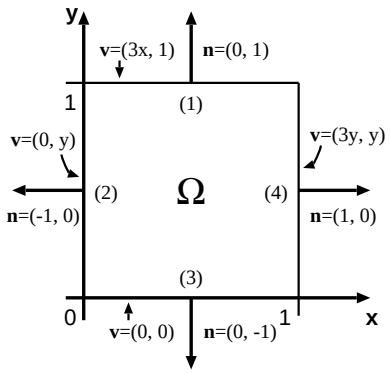
Example: We choose Ω to be the unit square in 2D, depicted as follows.



We choose $\mathbf{v}$ to be a vector field in 2D, defined by $\mathbf{v}(x, y) = (3xy, y)$, with divergence $\nabla \cdot \mathbf{v} = 3y + 1$. On integrating the divergence of $\mathbf{v}$ over the domain Ω, we obtain

$$\begin{aligned} \int_\Omega \nabla \cdot \mathbf{v} dV &= \int_{x=0}^1 \int_{y=0}^1 (3y+1) dy dx \\ &= \left[\tfrac{3}{2}y^2 + y\right]_0^1 \\ &= \tfrac{5}{2}. \end{aligned}$$

To calculate the RHS of the divergence theorem, we require the vector field and unit normal along the boundary $\partial\Omega$. We depict these quantities in the following figure.



The integral around the boundary then reads

$$\begin{aligned}
\oint_{\partial\Omega}(\mathbf{v}\cdot\mathbf{n})d\ell &= \underbrace{\int_0^1 (3x,1)\cdot(0,1)dx}_{(1)} + \underbrace{\int_0^1 (0,y)\cdot(-1,0)dy}_{(2)} \\
&\quad + \underbrace{\int_0^1 (0,0)\cdot(0,-1)dx}_{(3)} + \underbrace{\int_0^1 (3y,y)\cdot(1,0)dx}_{(4)} \\
&= x|_0^1 + 0 + 0 + \tfrac{3}{2}y|_0^1 \\
&= 1 + \tfrac{3}{2} \\
&= \tfrac{5}{2}.
\end{aligned}$$

On comparing $\int_\Omega \nabla\cdot\mathbf{v}dV$ and $\oint_{\partial\Omega}(\mathbf{v}\cdot\mathbf{n})d\ell$, we conclude that the LHS and RHS of (3.58) agree.

3.5.2 Conservation Laws in Higher Dimension

We now introduce the concept of a conservation law in higher dimensions, *i.e.* in $\mathbb{R}^n$ for $n \geq 2$.

It follows by direct generalization of (3.7) that a higher dimensional conservation law can be written in differential form as

$$\boxed{\frac{\partial u}{\partial t} + \nabla\cdot\mathbf{f}(u) = 0,} \tag{3.60}$$

where $\mathbf{f}(u)$ is the flux function (in this case, a vector function $\mathbf{f} : \mathbb{R} \to \mathbb{R}^n$). Given some arbitrary region Ω satisfying the constraints of Gauss' divergence theorem (Theorem 3.2), we can integrate the conservation law so as to obtain

$$\frac{d}{dt}\left(\int_\Omega u(\mathbf{x},t)d\Omega\right) + \int_\Omega \nabla \cdot \mathbf{f}(u)d\Omega = 0. \tag{3.61}$$

We now apply Gauss' divergence theorem to (3.61) and define

$$Q_\Omega(t) = \int_\Omega u(\mathbf{x},t)d\Omega, \tag{3.62}$$

which gives the generalization of the *first integral form of a conservation law*,[5]

$$\frac{d}{dt}Q_\Omega(t) + \oint_{\partial\Omega}(\mathbf{f}(u)\cdot\mathbf{n})d\ell = 0, \tag{3.63}$$

analogous to (3.12). Namely, (3.63) states that the rate of change of $Q_\Omega(t)$ is given by the net influx,

$$-\oint_{\partial\Omega}(\mathbf{f}(u)\cdot\mathbf{n})d\ell, \tag{3.64}$$

of u into the domain Ω through the boundary $\partial\Omega$, as we would expect for a conservation law.

3.5.3 Finite Volume Methods in 2D

We now give some examples of FV methods in 2D, using the notion of the higher-dimension conservation law defined by (3.63).

In restricting to 2D, the flux function $\mathbf{f}(u)$ can be written as

$$\mathbf{f}(u) = (g(u), h(u)). \tag{3.65}$$

The conservation law (3.60) then takes the form

$$\frac{\partial u}{\partial t} + \frac{\partial g(u)}{\partial x} + \frac{\partial h(u)}{\partial y} = 0, \tag{3.66}$$

which, on applying the chain rule, can be rewritten as

$$\frac{\partial u}{\partial t} + \frac{dg}{du}\frac{\partial u}{\partial x} + \frac{dh}{du}\frac{\partial u}{\partial y} = 0. \tag{3.67}$$

The terms $\frac{dg}{du}$ and $\frac{dh}{du}$ represent the wave speed of the system in the x and y directions, respectively.

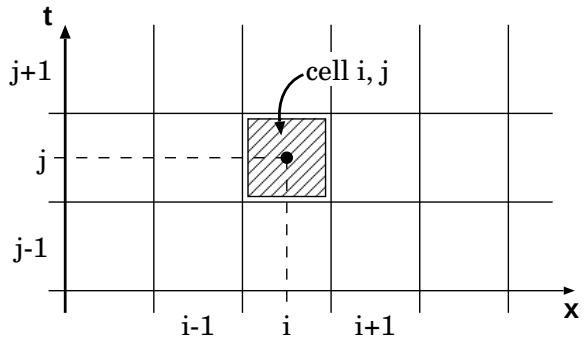
[5]Compare with 3.12.

Example: Recall that the linear advection equation in 2D takes the form (2.137), which we restate as

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} + b\frac{\partial u}{\partial y} = 0. \tag{3.68}$$

The wave speed, or advection velocity vector, is then given by $\mathbf{v} = (a, b)$.

We discretize the domain into square elements labelled by indices i and j, as depicted below.



Similar to the 1D case, discussed in section 3.4.1, we average the value of u over a FV cell and hence obtain from (3.66) the general form for an explicit FV method in 2D,

$$\boxed{\frac{v_{i,j}^{n+1} - v_{i,j}^{n}}{\Delta t} + \frac{g^*(v_{i,j}^n, v_{i+1,j}^n) - g^*(v_{i-1,j}^n, v_{i,j}^n)}{\Delta x} + \frac{h^*(v_{i,j}^n, v_{i,j+1}^n) - h^*(v_{i,j-1}^n, v_{i,j}^n)}{\Delta y}}. \tag{3.69}$$

By defining intermediate flux functions via

$$g^{*n}_{i+\frac{1}{2},j} = g^*(v_{i,j}^n, v_{i+1,j}^n), \quad \text{and} \quad h^{*n}_{i+\frac{1}{2},j} = h^*(v_{i,j}^n, v_{i+1,j}^n), \tag{3.70}$$

we can also write (3.69) as

$$\frac{v_{i,j}^{n+1} - v_{i,j}^{n}}{\Delta t} + \frac{g^{*n}_{i+\frac{1}{2},j} - g^{*n}_{i-\frac{1}{2},j}}{\Delta x} + \frac{h^{*n}_{i+\frac{1}{2},j} - h^{*n}_{i-\frac{1}{2},j}}{\Delta y}. \tag{3.71}$$

Hence, all that remains to define a explicit FV method is to specify numerical flux functions for g^* and h^*. A common technique for higher-dimensional FV methods, known as the *dimension-by-dimension approach* is to simply use the 1D Lax-Friedrichs numerical flux function (3.41) for g^* and h^*, *i.e.*

$$g^*(v_{i,j}^n, v_{i+1,j}^n) = \frac{g(v_{i,j}^n) + g(v_{i+1,j}^n)}{2} - \tfrac{1}{2}\left|(1,0)\cdot\lambda\left(\frac{v_{i,j}^n + v_{i+1,j}^n}{2}\right)\right|(v_{i+1,j}^n - v_{i,j}^n). \tag{3.72}$$

3.6 Systems of Conservation Laws

We now study FV approaches for systems of conservation laws in 1D.

If we consider a system of 1D conservation laws, each of the form (3.7), at each point $x \in \Omega$ we can define a *state vector* $\mathbf{U}(x)$. The system of conservation laws then takes the form

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = 0. \tag{3.73}$$

On applying the chain rule from vector calculus, we can write this equation as

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{d\mathbf{F}(\mathbf{U})}{d\mathbf{U}} \frac{\partial \mathbf{U}}{\partial x} = 0, \tag{3.74}$$

where $\frac{d\mathbf{F}(\mathbf{U})}{d\mathbf{U}}$ is the Jacobian matrix of the flux function $\mathbf{F}$. For example, if $\mathbf{u}$ is of dimension 2, then the state vector and flux function can be written in the form

$$\mathbf{U} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad \mathbf{F}(\mathbf{U}) = \begin{bmatrix} f_1(u_1, u_2) \\ f_2(u_1, u_2) \end{bmatrix}. \tag{3.75}$$

The Jacobian matrix is given by

$$\frac{d\mathbf{F}(\mathbf{U})}{d\mathbf{U}} = \begin{bmatrix} \frac{df_1}{du_1} & \frac{df_1}{du_2} \\ \frac{df_2}{du_1} & \frac{df_2}{du_2} \end{bmatrix}. \tag{3.76}$$

The notion of a hyperbolic PDE extends to conservation laws in the following manner:

Definition 3.5 *A system of conservation laws (3.73) is* ***hyperbolic*** *if and only if all eigenvalue of* $\frac{d\mathbf{F}(\mathbf{U})}{d\mathbf{U}}$ *are real.*

If a system of conservation laws is hyperbolic, we can construct a FV method which solves the system in the usual manner, obtaining the discretization

$$\frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\Delta t} + \frac{\mathbf{F}^{*n}_{i+\frac{1}{2}} - \mathbf{F}^{*n}_{i-\frac{1}{2}}}{\Delta x} = 0. \tag{3.77}$$

We can then use the usual Lax-Friedrichs flux function (3.41), which assumes the vector form

$$\mathbf{F}^{*n}_{i+\frac{1}{2}} = \frac{\mathbf{F}(V_i^n) + \mathbf{F}(V_{i+1}^n)}{2} - \tfrac{1}{2} \left| \lambda \left(\frac{\mathbf{U}_{i+1}^n - \mathbf{U}_i^n}{2} \right) \right|_{max} (\mathbf{V}_{i+1}^n - \mathbf{V}_i^n). \tag{3.78}$$

Example: The *shallow water equations* are a system of two conservation laws which can be written in vector form as

$$\frac{\partial}{\partial t}\begin{bmatrix} h \\ m \end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix} m \\ \frac{m^2}{h} + \frac{1}{2}gh^2 \end{bmatrix} = 0. \tag{3.79}$$

Here h represents the water height, $m = hu$ where u is the 1D velocity of the flow, and g is the gravitational constant. The flux function is then given by

$$\mathbf{F}(\mathbf{U}) = \begin{bmatrix} m \\ \frac{m^2}{h} + \frac{1}{2}gh^2 \end{bmatrix}, \tag{3.80}$$

which we can use to quickly calculate the Jacobian,

$$\frac{d\mathbf{F}(\mathbf{U})}{d\mathbf{U}} = \begin{bmatrix} 0 & 1 \\ (\frac{m}{h})^2 + gh & 2\frac{m}{h} \end{bmatrix}. \tag{3.81}$$

A short calculation then gives the eigenvalues,

$$\lambda_{\pm} = \frac{m}{h} \pm \sqrt{gh} = u \pm \sqrt{gh}. \tag{3.82}$$

A FV method for solving this PDE using the LF flux function then follows from (3.77), (3.78), (3.80) and (3.82).

CHAPTER 4

Finite Element Methods for Elliptic Problems

The last approaches we will consider for the problem of numerically solving PDEs are *finite element (FE) methods*. Finite element methods were originally developed in mechanical engineering to solve problems related to material stresses, *i.e.* they were used to determine how to best support a structure so as to prevent it from collapsing. These problems are generally elliptic in nature as they deal with the steady state conditions of a system. There is no general form of a FE method as with FV methods. Instead FE methods must be developed based on the problem (PDE) being analyzed.

In section 4.1 we begin with an introductory example outlining the techniques for deriving one such FE method. In sections 4.2 and 4.3 we develop FE methods for a 1D and 2D model problem, respectively. Lastly, in section 4.4 we briefly consider Neumann boundary conditions in the context of developing FE methods.

4.1 An Introductory Example

Consider the first order ODE IVP given by

$$IVP \begin{cases} x \in (0,2) \\ u(0) = 1 \\ u'(x) = 2x \end{cases} \tag{4.1}$$

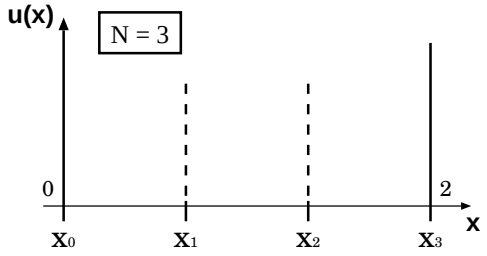
By inspection, the general solution of the ODE is

$$u(x) = x^2 + C. \tag{4.2}$$

Further, in order to satisfy the IC, we choose $C = 1$, giving a particular solution that satisfies the IVP.

We compose a **finite element (FE) method** to solve this IVP in four steps:

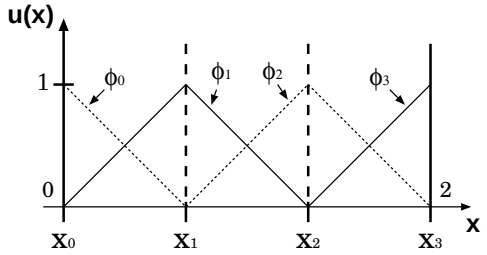
1) First, we must choose *discrete domain* Ω. We choose $N + 1$ discrete points $x_j \in (0, 2)$ so that $x_0 = 0$ and $x_N = 2$. These points are not necessarily equidistant, but are chosen in increasing order so as to define N elements via $[x_{j-1}, x_j]$ for $j = 1, \ldots, N$ (see figure below).



2) Secondly, we choose $N + 1$ basis functions $\phi_j(x)$, with $j = 0, \ldots, N$. For the type of methods we are interested in, the basis functions must satisfy two conditions:

   - $\phi_j(x_j) = \delta_{ij}$ (basis functions which satisfy this condition are called *nodal basis functions*)
   - $\sum_{j=0}^{N} \phi_j(x) = 1$ for all $x \in \Omega$.

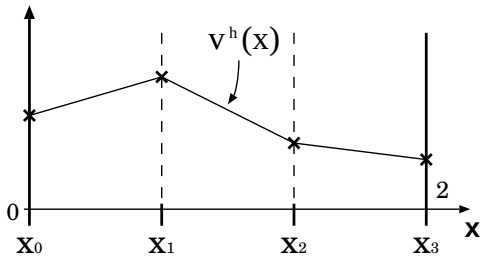
   One common choice in this regard are the so-called *tent functions*, which are depicted in the following figure.



3) Third, we seek a discrete approximation of the exact solution to the PDE of the form

$$v^h(x) = \sum_{j=0}^{N} c_j \phi_j(x), \tag{4.3}$$

where the c_j represent $N+1$ unknowns that must be determined. The grid function $v^h(x)$ is then a continuous approximation of the exact solution $u(x)$. In particular, if we choose the basis functions to be given by the tent functions (as above), $v^h(x)$ gives a piecewise linear approximation of the exact solution (see figure). Further, we know that $v^h(x_i) = c_i$ for all i, because $\phi_j(x_i) = \delta_{ij}$. In order to determine the $N+1$ unknowns c_j we will now need $N+1$ equations.



4) Finally, we formulate $N+1$ Galerkin equations in order to determine the constants c_j. We can integrate (4.1) over the interval $(0, 2)$ to obtain

$$\int_0^2 u' \phi_i(x) dx = \int_0^2 2x \phi_i(x) dx, \tag{4.4}$$

for $i = 0, \ldots, N$. Applying the numerical approximation $u \approx v^h$, we obtain the *Galerkin equations*,

$$\int_0^2 (v^h)' \phi_i(x) dx = \int_0^2 2x \phi_i(x) dx, \tag{4.5}$$

with $i = 0, \ldots, N$, which can be expanded using (4.3) to give

$$\sum_{j=0}^{N} c_j \left(\int_0^2 \phi_j'(x) \phi_i(x) dx \right) = \int_0^2 2x \phi_i(x) dx. \tag{4.6}$$

The equations (4.6) then form a $N+1 \times N+1$ linear system for the unknowns c_j. In matrix form, we write

$$AC = B, \tag{4.7}$$

where

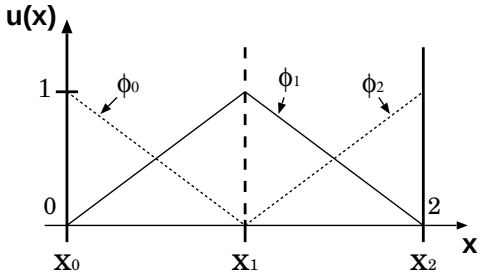
$$A = [a_{ij}], \qquad a_{ij} = \int_0^2 \phi_j'(x)\phi_i(x)dx, \tag{4.8}$$

$$B = [b_i], \qquad b_i = \int_0^2 2x\phi_i(x)dx, \tag{4.9}$$

$$C = [c_i]. \tag{4.10}$$

We can then solve the linear system (4.7), subject to the IC $c_0 = 1$ and hence obtain the approximate solution v^h.

Example: Consider $N = 2$, with $x_i = [0, 1, 2]$, as depicted in the following figure.



The numerical approximation $v^h(x)$ then assumes the form

$$v^h(x) = c_0\phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x), \tag{4.11}$$

where the ϕ_i are piecewise tent functions given by the following table.

	$x \in [0, 1]$	$x \in [1, 2]$
ϕ_0	$1 - x$	0
ϕ_1	x	$2 - x$
ϕ_2	0	$-1 + x$
ϕ_0'	-1	0
ϕ_1'	1	-1
ϕ_2'	0	1

After some computation, it can be shown that

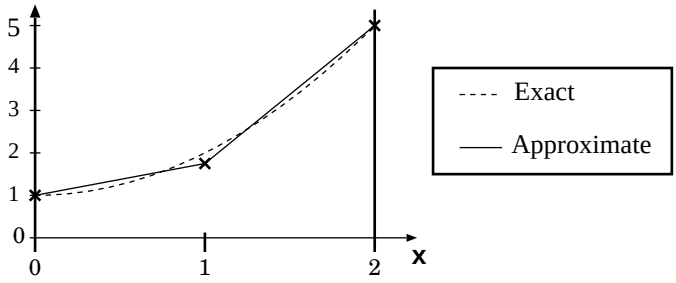
$$A = \begin{bmatrix} -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \qquad B = \begin{bmatrix} \frac{1}{3} \\ 2 \\ \frac{5}{3} \end{bmatrix}. \tag{4.12}$$

A short calculation shows that $\det(A) = 0$ and so the system of equations $AC = B$ has either no solutions or infinitely many solutions. In this case, it can be quickly shown that we have infinitely many solutions, *i.e.* there are only N linearly independent solutions. The extra equation instead follows from the initial conditions, *i.e.* on examining (4.1), we obtain $c_0 = 1$. If we let $c_0 = d$ be arbitrary, we will actually find that c_1 and c_2 are given by (exercise)

$$c_1 = c_0 + \tfrac{2}{3}, \qquad c_2 = c_0 + 4. \tag{4.13}$$

We now compare the exact and approximate solution, obtaining the following table.

x	$u(x)$	$v^h(x)$
0	1	1
1	2	$\frac{5}{3}$
2	5	5



4.2 The 1D Model Problem

In this text we will focus on finite element methods for second order elliptic PDEs, as FE methods were originally developed for PDEs of this form. In particular, we consider the 1D model problem (ODE) given by

$$BVP \begin{cases} \Omega : x \in (a, b), \\ u(a) = 0, u(b) = 0, \quad \text{(homogeneous BCs)} \\ -u''(x) + q(x)u(x) = f(x) \ \text{ in } \Omega. \end{cases} \tag{4.14}$$

Note that the ODE can also be written as a linear operator of the form $Lu = f$. Hereafter, we will refer to the ODE in the form (4.14) as the *strong form* of the ODE. The strong form of the ODE motivates the following definition:

Definition 4.1 *Let $Lu = f$ denote an arbitrary ODE of the form (4.14), where $f(x) \in C(a,b)$. Then a function $u(x) \in C^2(a,b)$ that satisfies the strong form of the ODE is called a* ***classical solution of the ODE****.*

Note that here $C(a,b)$ denotes the set of continuous functions and $C^2(a,b)$ denotes the set of functions that are twice continuously differentiable (*i.e.* continuous first and second derivatives).

4.2.1 Weighted Residual Form and Weak Form

In general, we will find that solutions obtained from FE methods are not classical solutions, *i.e.* they do not necessarily satisfy the strong form of the DE. In this section we will introduce an alternative formulation known as the weak form of the ODE that will form the basis for our study of FE methods.

FE methods are rooted heavily in functional analysis, and so we must first briefly introduce some concepts from this field. We begin with the notion of the L_2 scalar product, which induces the L_2 function space on (a,b).

Definition 4.2 *The L_2* ***scalar product*** *(or* ***inner product****) is an operator on functions $f(x)$ and $g(x)$ of the form*

$$(f(x), g(x)) = \int_a^b f(x)g(x)dx. \tag{4.15}$$

The L_2 scalar product induces the L_2 *norm*, according to

$$\|f(x)\|_2 = \sqrt{(f(x), f(x))} = \left(\int_a^b f(x)^2 dx \right)^{1/2}. \tag{4.16}$$

Further, the L_2 scalar product induces a set of functions according to the following definition.

Definition 4.3 *Let $(a,b) \subset \mathbb{R}$ be an open subset of the real line. Then $\mathbf{L_2}(\mathbf{a},\mathbf{b})$ is the set of functions on (a,b) satisfying*

$$L_2(a,b) := \{f(x) \mid \|f(x)\|_2 < \infty\}. \tag{4.17}$$

Since $L_2(a,b)$ is a vector space, we also obtain the notion of orthogonality of two functions analogous to orthogonality of two vectors.

Definition 4.4 *Let $f(x) \in L_2(a,b)$ and $g(x) \in L_2(a,b)$. Then $f(x)$* ***is orthogonal to*** *$g(x)$ with respect to $(\cdot,\cdot)$ if and only if $(f,g) = 0$.*

In order to proceed with the study of FE methods, we require two additional definitions:

Definition 4.5 *The **set of test functions on** $[a, b]$, denoted W_0 is defined by*

$$W_0 := \{w(x) \mid w'(x) \text{ is piecewise continuous on } [a, b] \text{ and } w(a) = 0,\ w(b) = 0. \tag{4.18}$$

Finally, we require some notion of closeness between an approximate solution $v(x)$ and an exact solution $u(x)$ of a given DE without necessarily knowing the exact solution. A linear DE operator $Lu = f$, such as one of the form (4.14), effectively determines a map between two functions in a given function space as follows.

Definition 4.6 *The **residual** of the DE $Lu = f$ is the function $r(x)$, defined by*

$$r(x) = Lv(x) - f(x), \tag{4.19}$$

for any function $v(x)$.

The Weighted Residual Form of the ODE

Definitions 4.5 and 4.6 motivate what is known as the *weighted residual form of the ODE*, given by

$$\boxed{(Lu - f, w) = 0 \,\forall\, w \in W_0.} \tag{4.20}$$

It can be shown that the form (4.20) and (4.14) are equivalent (exercise). Note that if a function $u(x)$ satisfies the weighted residual form, it follows by definition 4.4 that $r(x)$ is orthogonal to w for all $w \in W_0$.

The Weak Form of the ODE

The weighted residual form (4.20) can be rewritten using (4.14) and (4.15) as

$$\int_a^b (-u'' + qu - f)w dx = 0 \quad \forall\, w \in W_0. \tag{4.21}$$

On applying integration by parts ($\int u''w dx = u'w - \int u'w'$) and definition 4.5, we obtain

$$\underbrace{u'w|_a^b}_{=0} + \int_a^b (u'w' + quw - fw)dx = 0 \quad \forall\, w \in W_0. \tag{4.22}$$

Rewriting this expression in terms of the scalar product, we obtain the so-called *weak form of the ODE*, given by

$$\boxed{(u', w') + (qu, w) = (f, w) \quad \forall\, w \in W_0.} \tag{4.23}$$

Note that the weak form is also called a *variational form* (from the study of calculus of variations). Namely, for some ODEs the process of minimizing the integral leads to the weak form.

The Difference Between the Forms of the ODE

An observant reader will note that if $u(x) \in C^2(a,b)$ then the strong form, the weighted residual form and the weak form of the ODE are all equivalent. However, the weak form of the ODE only requires that a solution $u(x) \in C^1(a,b)$, and hence allows for solutions that are not classical solutions (see definition 4.1). In general, we will refer to solutions of the weak form as *weak solutions*. An advantage of the FE method for solving the ODE (4.14) is that it is based on the weak form of the ODE (4.23) and so will allow us to approximate weak solutions $u(x) \in C^1(a,b)$.

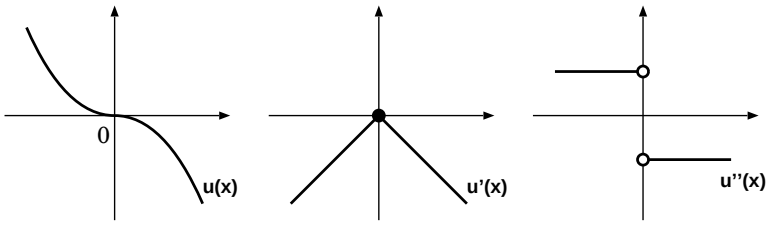
Example: We choose $q(x) = 0$ and choose $f(x)$ to be given by the discontinous function

$$f(x) = \begin{cases} -1 & x \leq 0, \\ 1 & x > 0. \end{cases} \tag{4.24}$$

It can then be shown that $u(x)$, defined by

$$u(x) = \begin{cases} \frac{1}{2}x^2 & x \leq 0, \\ -\frac{1}{2}x^2 & x > 0, \end{cases} \tag{4.25}$$

solves the ODE in its weak form (4.23). We depict $u(x)$ and its derivatives in the following figure.



Clearly, $u(x)$ is not a classical solution of the ODE (4.14) since $u(x) \notin C^2(a,b)$.; however, $u(x)$ satisfies the weak form and hence is a weak solution.

4.2.2 Discrete Weak Form

In this section we turn our attention to discretizing the weak form of the ODE (4.23). This form, known as the discrete weak form, will then form the foundation for FE methods.

We choose m basis functions $\phi_j(x)$ $(1 \leq j \leq m)$ that are linearly independent and satisfy $\phi_j(a) = 0$ and $\phi_j(b) = 0$ for all j. As with the example problem discussed in section 4.1, we seek an approximation

$v^h(x)$ of the exact solution $u(x)$ that is a linear combination of the basis functions $\phi_j(x)$, *i.e.*

$$v^h(x) = \sum_{j=1}^{m} c_j \phi_j(x), \quad c_j \in \mathbb{R}. \tag{4.26}$$

The complete set of functions of the form (4.26) is then defined by

$$V_0^h = \{v^h(x) \mid v^h(x) = \sum_{j=1}^{m} c_j \phi_j(x),\ c_j \in \mathbb{R}\}. \tag{4.27}$$

The set V_0^h is then an m-dimensional vector space which is spanned by the basis vectors

$$B_0^h = \{\phi_j(x)\}_{j=1}^m. \tag{4.28}$$

We now desire to find $v^h(x)$ so that it satisfies the weak form of the ODE (4.23), *i.e.* we want

$$(v^h, w') + (qv^h, w) = (f, w)\ \forall\ w \in W_0. \tag{4.29}$$

The restriction (4.29) imposes an infinite number of conditions on v^h, since it must be satisfied for all possible test functions. Since v^h is a member of a finite-dimensional vector space, it is in general impossible to find v^h so that it satisfies all possible conditions. As a consequence, we instead must "prune" the number of conditions imposed by (4.29) by discretizing the space of test functions W_0.

Several solutions exist to this problem, all requiring us to make a specific choice of the set of test functions. For the FE methods we are interested in, we choose *Galerkin approach*: Namely, instead of using the whole set of test functions W_0, instead we choose a discrete set of test functions W_0^h given by the set of basis functions V_0^h (*i.e.* $W_0^h = V_0^h$). This approach leads us to the so-called *discrete weak form of the ODE*, given by

$$((v^h)', (w^h)') + (qv^h, w^h) = (f, w^h)\ \forall\ w^h \in V_0^h. \tag{4.30}$$

The following more applicable expression can be quickly obtained from the discrete weak form (exercise).

$$((v^h)', \phi_i') + (qv^h, \phi_i) = (f, \phi_i)\ \forall\ \phi_i \in B_0^h\ (i = 1, \ldots, m). \tag{4.31}$$

The discrete ODE problem is then stated as follows:

Problem: Find $v^h \in V_0^h$ such that v^h satisfies (4.31).

Since v^h can be expanded in terms of the basis functions ϕ_j according to (4.26), the discrete ODE problem then reduces to an algebraic problem from (4.31), namely:

Problem: Find c_j such that

$$\sum_{j=1}^{m} c_j \left[(\phi_j', \phi_i') + (q\phi_j, \phi_i)\right] = (f, \phi_i), \quad i = 1, \ldots, m. \tag{4.32}$$

Matrix Form of the Discrete ODE Problem

If we assume $q(x) = q = \text{const.}$, (4.32) can be written in the form

$$(K + qM)C = L, \tag{4.33}$$

where $C = [c_i]$ and the matrices K, M and L are defined in the following table.

Load vector:	$L = [\ell_i]$	$\ell_i = (f, c_i)$,	(4.34)
Stiffness matrix:	$K = [k_{ij}]$	$k_{ij} = (\phi_j', \phi_i')$,	(4.35)
Mass matrix:	$M = [m_{ij}]$	$m_{ij} = (\phi_j, \phi_i)$.	(4.36)

The names *load vector*, *stiffness matrix* and *mass matrix* originate from mechanical engineering, where FE methods were originally developed.

Note: Observe that the stiffness matrix K and mass matrix M are symmetric. Symmetric matrices are particularly nice to deal with, since it is easy to prove existence-uniqueness of (4.33). Further, efficient methods exist to solve symmetric linear systems.

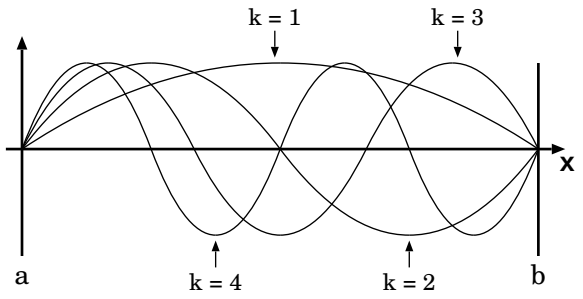
4.2.3 Choice of Basis Functions

There are several possible choices of basis functions, with each choice effectively giving a different method for solving PDEs. Two common choices are *model basis functions* and *nodal basis functions*, which we discuss in this section.

1) A *modal basis function* seperates the solution into a set of "modes" associated with each basis function. For example, we could make a choice of basis of the form

$$B_0^h = \left\{ \sin\left(k\pi \frac{(x-a)}{(b-a)} \right),\ k = 1, 2, \ldots, m \right\}. \tag{4.37}$$

The first few modes in this case are depicted in the following figure.



This choice of basis functions actually gives what is known as a *spectral method*, so named because it decomposes the solution into a spectrum of waves depending on their frequency.

In general, modal basis functions are nonzero over the entire domain and often lead to methods which are high order accurate. However, the accuracy comes at a cost, since K and M will not necessarily be sparse matrices. Further, certain choices of basis functions require regular geometry or periodic boundary conditions.

2) A *nodal basis function* satisfies $\phi_j(x_i) = \delta_{ij}$ and has so-called *compact support*, *i.e.* it is nonzero only on a small portion of the domain. This choice of basis function leads to a set of methods which we will refer to as *finite element methods*.

We have already encountered one example of a nodal basis, namely the tent functions introduced in section 4.1. Basis functions of this type will form the foundation of our analysis for the remainder of this chapter.

Note that the requirement of compact support leads to sparse K and M matrices (*i.e.* the majority of the entries in these matrices will be zero). This result means several efficient methods exist for solving the linear system (4.33).

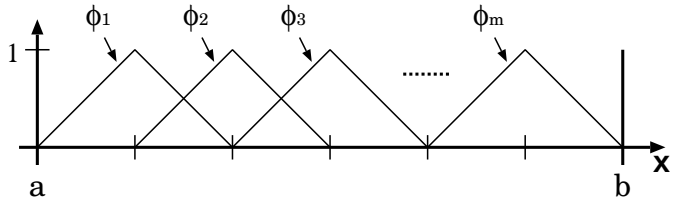
The Tent Functions as Basis Functions

We now consider in detail the so-called tent functions as the choice of basis functions for a finite element method.

For simplicity, we discretize the interval (a, b) into m distinct intervals by choosing x_j equidistant with $x_0 = a$ and $x_{m+1} = b$. The basis functions $\phi_j(x)$ (with $j = 1, \ldots, m$) are then chosen to be linear in each interval and satisfy $\phi_j(x_i) = \delta_{ij}$. The distance between adjacent points is denoted h and given by

$$h = \frac{b-a}{m+1}. \tag{4.38}$$

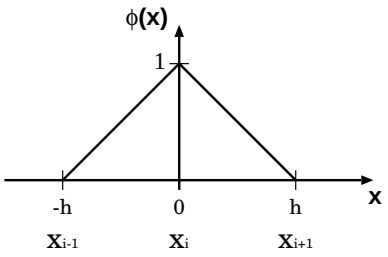
A depiction of this discretization is then given as follows.



Note that $\phi_0(x)$ and $\phi_{m+1}(x)$ are not retained since we want $\phi_i(a) = 0$ and $\phi_i(b) = 0$ for all i. Mathematically, we obtain the following expressions for $\phi_i(x)$:

$$\phi_i(x) = \begin{cases} \frac{1}{h}(x_i - x_{i-1}), & x \in [x_{i-1}, x_i] \text{ (in element } i\text{)}, \\ \frac{1}{h}(x_{i+1} - x_i), & x \in [x_i, x_{i+1}] \text{ (in element } i+1\text{)}, \\ 0 & \text{elsewhere.} \end{cases} \tag{4.39}$$

We can now use (4.39) in conjunction with (4.34)-(4.36) to calculate the components of the linear system (4.33). In order to simplify the calculations, we shift the origin so that $x_i = 0$, $x_{i-1} = -h$ and $x_{i+1} = h$ (see figure).



The tent functions are then described by the following table.

	Element i	Element $i+1$
ϕ_i	$\frac{1}{h}(x+h)$	$\frac{1}{h}(h-x)$
ϕ_{i+1}	0	$\frac{1}{h}x$
ϕ_{i-1}	$-\frac{1}{h}x$	0
ϕ_i'	$\frac{1}{h}$	$-\frac{1}{h}$
ϕ_{i+1}'	0	$\frac{1}{h}$
ϕ_{i-1}'	$-\frac{1}{h}$	0

The mass matrix M has components $m_{i,j} = (\phi_i, \phi_j)$. Following a short calculation (exercise), we obtain

$$m_{i,i} = \tfrac{2}{3}h, \qquad m_{i,i+1} = \tfrac{1}{6}h, \qquad m_{i,i-1} = \tfrac{1}{6}h, \tag{4.40}$$

with all other $m_{i,j} = 0$. The stiffness matrix K has components $k_{i,j} = (\phi_i', \phi_j')$. Another short calculation (exercise) gives

$$k_{i,i} = \tfrac{2}{h}, \qquad k_{i,i+1} = -\tfrac{1}{h}, \qquad k_{i,i-1} = \tfrac{1}{h}, \tag{4.41}$$

with all other $k_{i,j} = 0$. The load vector L with components $\ell_i = (f, \phi_i)$ is obtained by computing

$$\ell_i = \int_{x_{i-1}}^{x_i} f(x)\tfrac{1}{h}(x_i - x_{i-1})dx + \int_{x_i}^{x_{i+1}} f(x)\tfrac{1}{h}(x_{i+1} - x_i)dx. \tag{4.42}$$

If f is a constant, then it can be shown that $\ell_i = fh$.

In summary, the DE (4.14) with q and f constant can be written in matrix form as (4.33), with K, M and F given by

$$K = \frac{1}{h}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}, M = \frac{h}{6}\begin{bmatrix} 4 & 1 & & 0 \\ 1 & 4 & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 4 \end{bmatrix}, F = h\begin{bmatrix} f \\ f \\ \vdots \\ f \end{bmatrix}. \tag{4.43}$$

Note that if $q = 0$, this method is identical to the FD method for the 1D elliptic BVP (2.10).

4.3 The 2D Model Problem

In this section we turn our attention to developing a FE method for solving the extension of the 1D model problem (4.14) to 2D. As before, we begin by introducing the necessary theoretical framework, developing a weak form of the PDE and then discretizing to obtain the FE method.

The 2D model problem is given as follows.

$$BVP \begin{cases} \Omega : (x, y) \in (0, 1)^2, \\ u = g(x, y) \text{ on } \partial\Omega, \\ -\nabla^2 u = f \text{ in } \Omega. \end{cases} \tag{4.44}$$

The *strong form of the PDE* then reads $Lu = f$, where $L = -\nabla^2$. The strong form of the ODE again motivates the definition of a classical solution:

Definition 4.7 *Let $Lu = f$ denote an arbitrary ODE of the form (4.44), where $f(x, y) \in C(\Omega)$. Then a function $u(x)$ with $\frac{\partial^2 u}{\partial x^2}$ and $\frac{\partial^2 u}{\partial y^2}$ continuous that satisfies the strong form of the ODE is called a* ***classical solution of the ODE****.*

Note that here $C(\Omega)$ denotes the set of continuous functions on Ω.

4.3.1 Weak Form

In order to proceed, we must again introduce some concepts from functional analysis that are a direct extension of the definitions in section 4.2.1.

Definition 4.8 *The L_2* ***scalar product*** *on Ω (or* ***inner product****) is an operator on functions $f(\mathbf{x})$ and $g(\mathbf{x})$ of the form*

$$(f, g) = \int_0^1 \int_0^1 f(\mathbf{x}) g(\mathbf{x}) dxdy. \tag{4.45}$$

The definition of the scalar product can be extended to apply to vector fields, as follows.

Definition 4.9 *The L_2* ***scalar product*** *on Ω (or* ***inner product****) is an operator on vector fields $f(\mathbf{x})$ and $g(\mathbf{x})$ of the form*

$$(\mathbf{f}, \mathbf{g}) = \int_0^1 \int_0^1 \mathbf{f}(\mathbf{x}) \cdot \mathbf{g}(\mathbf{x}) dxdy. \tag{4.46}$$

The L_2 scalar product induces the L_2 *norm*, according to

$$\|f(\mathbf{x})\|_2 = \sqrt{(f(\mathbf{x}), f(\mathbf{x}))} = \left(\int_0^1 \int_0^1 f(\mathbf{x})^2 dxdy \right)^{1/2}. \tag{4.47}$$

Further, the L_2 scalar product induces a set of functions according to the following definition.

Definition 4.10 *Let $\Omega \subset \mathbb{R}^2$ be an open set. Then $\mathbf{L_2(\Omega)}$ is the set of functions on Ω satisfying*

$$L_2(\Omega) := \{f(\mathbf{x}) \mid \|f(\mathbf{x})\|_2 < \infty\}. \tag{4.48}$$

We again must define the set of test functions on Ω, given as follows.

Definition 4.11 *The* ***set of test functions on*** *Ω, denoted W_0 is defined by*

$$W_0 := \{w(\mathbf{x}) \mid \tfrac{\partial w}{\partial x} \text{ and } \tfrac{\partial w}{\partial y} \text{ are piecewise continuous and bounded on } \Omega \text{ and } w(\mathbf{x}) = 0 \text{ on } \partial\Omega\}. \tag{4.49}$$

It can be shown that the strong form of the ODE (4.44) is equivalent to the *weighted residual form*, given by

$$(-\nabla^2 u, w) = (f, w) \; \forall \; w \in W_0, \quad u = g \text{ on } \partial\Omega. \tag{4.50}$$

We now apply corollary B.1 to (4.50) in order to obtain

$$\iint_\Omega -\nabla \cdot (w \nabla u) dx dy + (\nabla w, \nabla u) = (f, w) \; \forall w \in W_0, \quad u = g \text{ on } \partial\Omega. \tag{4.51}$$

Then, on applying Gauss' divergence theorem (theorem 3.2), (4.51) can be written as

$$-\oint_{\partial\Omega} (\nabla u w) \cdot \mathbf{n} d\ell + (\nabla u, \nabla w) = (f, w), \tag{4.52}$$

but since $w = 0$ on $\partial\Omega$, the first term is identically zero. Hence, we obtain the so-called *weak form* of the ODE, analogous to the weak form (4.23) for the 1D model problem,

$$\boxed{(\nabla u, \nabla w) = (f, w) \; \forall w \in W_0, \quad u = g \text{ on } \partial\Omega.} \tag{4.53}$$

For any classical solution u it can be easily shown that (4.50), (4.51) and (4.53) are equivalent statements; however, as with the 1D model problem, the weak form (4.53) admits solutions that do not have continuous partial second derivatives. This result motivates the following definition.

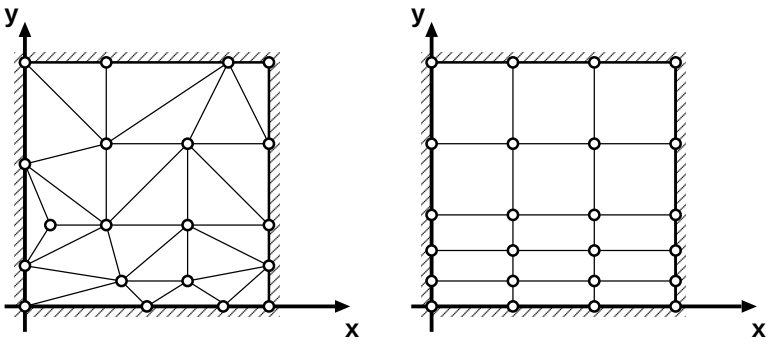
Definition 4.12 *A function $u(x, y)$ that is a solution of the weak form (4.53) but is not a solution of the strong form (4.44) is called a* ***weak solution****.*

Observe that the weak form (4.53) only requires that a solution $u(x, y)$ have $\frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y}$ piecewise continuous and bounded in Ω.

4.3.2 Discrete Weak Form

In order to develop a discrete weak form that will eventually be the foundation for the FE method, we must first partition the domain into a set of discrete elements. Notably, one of the main strengths of the FE method is that there are no requirements of structure on the discretization.

We discretize the grid into m total nodes x_j, $j = 1, \ldots, m$, with n interior nodes and $m - n$ boundary nodes. Two possible discretizations are depicted in the following figure.



From the set of nodes $\{x_j\}$ we define a set of nodal basis functions $\phi_j(\mathbf{x})$ with $j = 1, \ldots, m$, that satisfies

$$\phi_j(\mathbf{x_i}) = \delta_{ij}, \qquad \sum_{j=1}^{m} \phi_j(\mathbf{x}) = 1 \; \forall \; \mathbf{x} \in \Omega. \tag{4.54}$$

The set of discrete candidate solutions is then given by

$$V_g^h = \{v^h \mid v^h(\mathbf{x}) = \sum_{j=1}^{m} c_j \phi_j(\mathbf{x}), \text{with } c_j = g(\mathbf{x}_j) \text{ for boundary nodes}\}. \tag{4.55}$$

Note that the set V_g^h forms a vector space spanned by the nodal basis functions ϕ_j.

In order to obtain the discrete weak form, we must also consider the set of solutions with zero boundary conditions, given by

$$V_0^h = \{v^h \mid v^h(\mathbf{x}) = \sum_{j=1}^{m} c_j \phi_j(\mathbf{x}), \text{with } c_j = 0 \text{ for boundary nodes}\}. \tag{4.56}$$

This vector space is then spanned by the set of basis vectors

$$B_0^h = \{\phi_j \mid j \text{ is not a boundary node}\}. \tag{4.57}$$

Note that B_0^h consists of exactly n basis functions, implying that V_0^h is a vector space of dimension n.

We now have sufficient background in order to define the *discrete weak form of the 2D model problem*, given by

$$(\nabla v^h, \nabla w^h) = (f, w^h) \; \forall \; w^h \in V_0^h. \tag{4.58}$$

The following more applicable expression can be quickly obtained from the discrete weak form (exercise).

$$(\nabla v^h, \nabla \phi_i) = (f, \phi_i) \; \forall \; \phi_i \in B_0^h. \tag{4.59}$$

The discrete ODE problem is then stated as follows:

Problem: Find $v^h \in V_g^h$ such that v^h satisfies (4.31).

As with the 1D problem, (4.58) and (4.59) are intractable. Instead, we expand v^h in terms of the basis functions, obtaining the following statement of the problem.

Problem: Find c_j such that

$$\sum_{j=1}^{m} c_j (\nabla \phi_j, \nabla \phi_i) = (f, \phi_i) \; \forall \; \phi_i \in B_0^h. \tag{4.60}$$

Matrix Form of the Discrete PDE Problem

The discrete PDE problem (4.60) can be written in matrix form as

$$KC = L, \tag{4.61}$$

where K is the *stiffness matrix* and L is the *load vector*. Special attention must be paid to the boundary nodes, where the c_j are predetermined by the boundary conditions. There are two possible viewpoints one can consider when attempting to solve the linear system, which we now consider.

1) We can incorporate the boundary conditions in the linear system and solve for the associated c_j as if they corresponded to interior nodes. The stiffness matrix then takes the form

$$K = \begin{bmatrix} I & 0 \\ K_b & K_i \end{bmatrix}, \tag{4.62}$$

where I is the identity matrix and

$$K_b = \begin{bmatrix} (\nabla \phi_1, \nabla \phi_{m-n+1}) & \cdots & (\nabla \phi_{m-n}, \nabla \phi_{m-n+1}) \\ \vdots & & \vdots \\ (\nabla \phi_1, \nabla \phi_m) & \cdots & (\nabla \phi_{m-n}, \nabla \phi_m) \end{bmatrix}, \tag{4.63}$$

and

$$K_i = \begin{bmatrix} (\nabla\phi_{m-n+1}, \nabla\phi_{m-n+1}) & \cdots & (\nabla\phi_m, \nabla\phi_{m-n+1}) \\ \vdots & & \vdots \\ (\nabla\phi_{m-n+1}, \nabla\phi_m) & \cdots & (\nabla\phi_m, \nabla\phi_m) \end{bmatrix}. \tag{4.64}$$

The C and L vectors are then given by

$$C = \begin{bmatrix} c_1 \\ \vdots \\ c_{m-n} \\ \hline c_{m-n+1} \\ \vdots \\ c_m \end{bmatrix}, \quad L = \begin{bmatrix} g_1 \\ \vdots \\ g_{m-n} \\ \hline (f, \phi_{m-n+1}) \\ \vdots \\ (f, \phi_m) \end{bmatrix}. \tag{4.65}$$

2) In placing the boundary nodes within the stiffness matrix, we are potentially reducing the efficiency of solving the linear system. Instead, consider decomposing (4.62) according to

$$KC = L \quad \Longleftrightarrow \quad \left[\begin{array}{cc} I & 0 \\ \hline K_b & K_i \end{array}\right] \left[\begin{array}{c} C_b \\ \hline C_i \end{array}\right] = \left[\begin{array}{c} G_b \\ \hline F_i \end{array}\right]. \tag{4.66}$$

The upper block of equations then simply reduces to $C_b = G_b$, as expected. The lower block ($K_bC_b + K_iC_i = F_i$) can be rewritten as

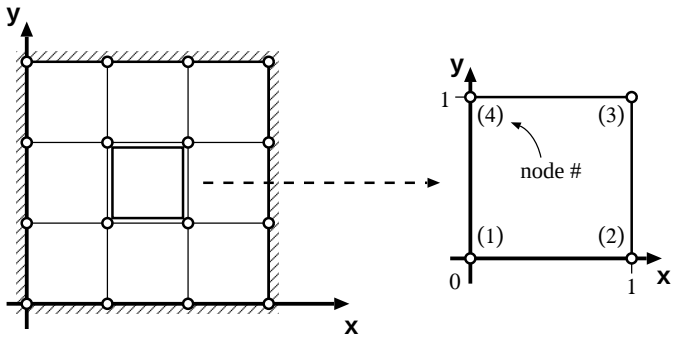
$$K_iC_i = F_i - K_bC_b \quad \Longleftrightarrow \quad K_iC_i = H_i, \tag{4.67}$$

where $H_i = F_i - K_bC_b$. Treating the problem in this manner means we only need to solve a $n \times n$ linear system, as opposed to a $m \times m$ linear system.

4.3.3 Simple Finite Elements in 2D

Perhaps the most common choices of elements to use in discretizing the region are either rectangular elements or triangular elements. We now discuss the techniques used to calculate the stiffness matrix K and the load vector L corresponding to these elements.

Rectangular Elements: Consider a partition of Ω as follows.



A reference element is then given by removing one element from the rectangular mesh, translating to the origin and rescaling so as to give the unit square. Nodes are numbered locally in a counter-clockwise sense, as depicted in the previous figure.

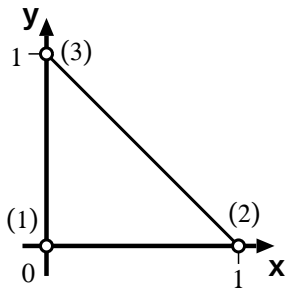
The nodal basis functions associated with this element are then given as follows:

$$\phi_1(x,y) = (1-x)(1-y), \qquad \phi_3(x,y) = xy, \tag{4.68}$$

$$\phi_2(x,y) = x(1-y), \qquad \phi_4(x,y) = (1-x)y. \tag{4.69}$$

Note that a function of this type is called *bilinear*, since it is linear in x and y.

Triangular Elements: On translating and scaling a triangular element, we obtain a reference element as depicted in the following figure.



The nodal basis functions associated with this element are then given as follows:

$$\phi_1(x,y) = 1-(x+y), \qquad \phi_2(x,y) = x, \qquad \phi_3(x,y) = y. \tag{4.70}$$

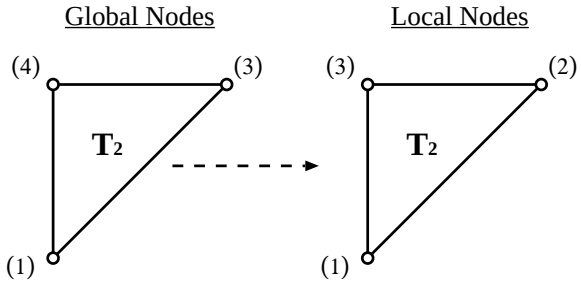
We now discuss how to calculate the stiffness matrix K and load vector L for a general triangular elements. The process we now employ can be easily generalized to elements of arbitrary shape with some effort.

Assuming that a given element has no boundary nodes, the stiffness matrix K and load vector L are given by

$$K_{ij} = (\nabla\phi_j, \nabla\phi_i), \quad \text{and} \quad L_i = (f, \phi_i). \tag{4.71}$$

In order to calculate the elements K_{ij} and L_i, we will require expressions for the $\phi_i(x, y)$ of an arbitrary triangular element. We will use two tricks in order to simplify our calculations:

1. First, we use local node numbering so as to treat each element without concern for global node numbers. The global stiffness matrix K and load vector L can then be reconstructed from a local stiffness matrix K_e and load vector L_e defined individually for each element (as we will show later).



2. Second, we shift the coordinate system so that in the local numbering scheme $(x_1, y_1) = (0, 0)$.

The basis functions ϕ_i then take the general form

$$\phi_i = \alpha_i + \beta_i x + \gamma_i y, \qquad i = 1, 2, 3, \tag{4.72}$$

where the coefficients α_i, β_i and γ_i are functions of $(x_1, y_1) = (0, 0)$, (x_2, y_2) and (x_3, y_3). One can quickly observe that the equations we must solve are

$$\begin{aligned}
&\phi_1(x_1, y_1) = 1, && \phi_2(x_1, y_1) = 0, && \phi_3(x_1, y_1) = 0,\\
&\phi_1(x_2, y_2) = 0, && \phi_2(x_2, y_2) = 1, && \phi_3(x_2, y_2) = 0,\\
&\phi_1(x_3, y_3) = 0, && \phi_2(x_3, y_3) = 0, && \phi_3(x_3, y_3) = 1.
\end{aligned} \tag{4.73}$$

A short calculation reveals (exercise)

$$\alpha_1 = 1, \qquad \alpha_2 = 0, \qquad \alpha_3 = 0, \tag{4.74}$$

$$\beta_1 = \frac{y_2 - y_3}{\Delta}, \qquad \beta_2 = \frac{y_3}{\Delta}, \qquad \beta_3 = -\frac{y_2}{\Delta}, \tag{4.75}$$

$$\gamma_1 = \frac{x_3 - x_2}{\Delta}, \qquad \gamma_2 = -\frac{x_3}{\Delta}, \qquad \gamma_3 = \frac{x_2}{\Delta}, \tag{4.76}$$

where

$$\Delta = x_2 y_3 - x_3 y_2. \tag{4.77}$$

The elements of stiffness matrix can then be calculated:

$$\begin{aligned} K_{ij} &= (\nabla\phi_j, \nabla\phi_i) \\ &= ((\beta_j, \gamma_j), (\beta_i, \gamma_i)) \\ &= \int_{T_e} (\beta_j\beta_i + \gamma_j\gamma_i) d\Omega, \end{aligned}$$

and since $(\beta_j\beta_i + \gamma_j\gamma_i)$ is constant, we obtain

$$K_{ij} = A_e(\beta_j\beta_i + \gamma_j\gamma_i), \tag{4.78}$$

where A_e is the area of element T_e. It can be shown that A_e satisfies

$$A_e = \frac{|\Delta|}{2}. \tag{4.79}$$

The elements of the load vector can then be written as

$$\begin{aligned} L_i &= (f, \phi_i) \\ &= \int_{T_e} f(\alpha_i + \beta_i x + \gamma_i y) d\Omega. \end{aligned} \tag{4.80}$$

At this point, numerical integration is generally necessary. If f is constant, this expression can be evaluated immediately, giving (exercise)

$$L_i = f A_e \left(\alpha_i + \tfrac{1}{3} \sum_{i=1}^{3} x_i + \tfrac{1}{3} \sum_{i=1}^{3} y_i \right). \tag{4.81}$$

Constructing the Global Stiffness Matrix and Load Vector

Now, using (4.78) and (4.80) we have the tools necessary to assemble the global stiffness matrix K and load vector L.

Recall that the discrete weak form of the PDE problem (4.59) requires that we find $v^h \in V_g^h$ such that

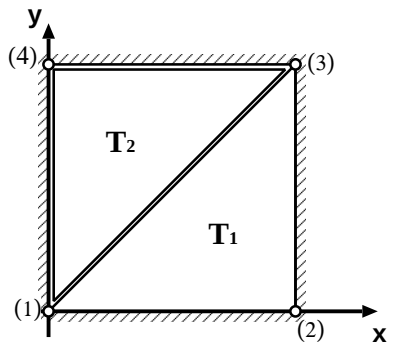
$$(\nabla v^h, \nabla \phi_i) = (f, \phi_i) \; \forall \phi_i \in B_0^h.$$

We can expand the inner products as integrals, and then write the total integral over Ω as the sum over the integral of each element. On applying this process, we instead must find $v^h \in V_g^h$ such that

$$\sum_{e=1}^{n_e} (\nabla v^h, \nabla \phi_i) = \sum_{e=1}^{n_e} (f, \phi_i) \; \forall \; \phi_i \in B_0^h, \tag{4.82}$$

where n_e is the total number of discrete elements.

Example: Consider the following discretization of the unit square into two elements T_1 and T_2.



Observe that T_2 only contributes to the submatrix consisting of the first, third and fourth column of the stiffness matrix and the first, third and fourth rows of the load matrix. In particular, the contribution to the stiffness matrix by T_2 is given by the local stiffness matrix,

$$K_2 = \begin{bmatrix} (\nabla\phi_1, \nabla\phi_1) & (\nabla\phi_3, \nabla\phi_1) & (\nabla\phi_4, \nabla\phi_1) \\ (\nabla\phi_1, \nabla\phi_3) & (\nabla\phi_3, \nabla\phi_3) & (\nabla\phi_4, \nabla\phi_3) \\ (\nabla\phi_1, \nabla\phi_4) & (\nabla\phi_3, \nabla\phi_4) & (\nabla\phi_4, \nabla\phi_4) \end{bmatrix}. \tag{4.83}$$

The global stiffness matrix for this system is a 4×4 matrix, which is constructed from K_1, indicated by 'o', and K_2, indicated by '$\times$' as follows:

$$K = \begin{bmatrix} \times + o & o & \times + o & \times \\ o & o & o & \\ \times + o & o & \times + o & \times \\ \times & & \times & \times \end{bmatrix}. \tag{4.84}$$

The global load vector has dimension 4, and contains contributions indicated as follows:

$$L = \begin{bmatrix} \times + o \\ o \\ \times + o \\ \times \end{bmatrix}. \tag{4.85}$$

Pseudo-Code

The pseudo-code for the method we have discussed in this section for the 2D test problem is given as follows. Note that for simplicity, we choose to use method $1)$ from section 4.3.2 in order to construct the stiffness matrix and load vector.

```
For e = 1 : n_e
  Calculate α_i, β_i, γ_i  (i = 1, 2, 3)
  Build K_e  (3 × 3 matrix)
  Build L_e  (3 × 1 vector)
  Assemble K_e into K
  Assemble L_e into L
End For
For i = 1 : m − n
  K(i, :) = 0,   K(i, i) = 1,   L(i) = g_i
End For
Solve KC = L for C
```

4.4 Neumann Boundary Conditions

In this section we discuss Neumann boundary conditions for elliptic PDEs, which are inevitably necessary for many physical problems. So far we have restricted our attention to an elliptic BVP with Dirichlet boundary conditions, *i.e.* a BVP of the form

$$BVP(1)\begin{cases} \Omega \text{ an open, bounded domain,} \\ u = g \text{ on } \Gamma = \partial\Omega \;\longrightarrow\; \textbf{Dirichlet BCs}, \\ -\nabla^2 u = f \text{ on } \Omega. \end{cases} \tag{4.86}$$

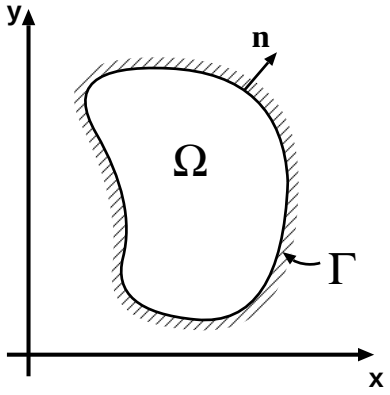
We now wish to consider Neumann boundary conditions, which instead impose a constraint on the *derivative of the function* u. One such BVP with Neumann BCs is

$$BVP(2)\begin{cases} \Omega \text{ an open, bounded domain,} \\ \frac{\partial u}{\partial n} = h \text{ on } \Gamma = \partial\Omega \;\longrightarrow\; \textbf{Neumann BCs}, \\ -\nabla^2 u = f \text{ on } \Omega. \end{cases} \tag{4.87}$$

Here $\frac{\partial u}{\partial n}$ is shorthand for the directional derivative

$$\frac{\partial u}{\partial n} = \nabla u \cdot \mathbf{n}, \tag{4.88}$$

as in the following figure.



Example: Consider the stationary heat diffusion problem, an elliptic PDE that takes the form

$$-D\nabla^2 T(x,y) = f, \quad \text{or} \quad -D\nabla \cdot \nabla T(x,y) = f. \tag{4.89}$$

Recall that the Fourier law of heat conduction states that the heat flux vector (the direction of heat flow) is proportional to ∇T. Hence, if we were to surround a region Ω with an insulating wall that does not permit heat flow, mathematically we are imposing that

$$\frac{\partial T}{\partial n} = 0. \tag{4.90}$$

4.4.1 Compatibility Between h and f

We now demonstrate two important results about the Neumann BVP (4.87).

On integrating the Neumann BVP (4.87) over Ω, we obtain

$$\int_\Omega -\nabla \cdot (\nabla u) d\Omega = \int_\Omega f d\Omega. \tag{4.91}$$

The divergence theorem (Theorem 3.2) then implies that

$$\oint_{\partial\Omega} \nabla u \cdot \mathbf{n} d\ell = \int_\Omega f d\Omega, \tag{4.92}$$

which from (4.87) and (4.88) is simply

$$\boxed{\oint_{\partial\Omega} h d\ell = \int_\Omega f d\Omega.} \tag{4.93}$$

Since (4.93) is interchangable with the PDE (4.87), we conclude that h and f must satisfy this condition - otherwise there will be no solution to the BVP. Observe further that if u is a solution of the BVP (4.87) then $u + c$ is also a solution for any $c \in \mathbb{R}$. These two results are summed up in the following theorem.

Theorem 4.1 *If f and h satisfy the compatibility condition (4.93) then the Neumann BVP has a unique solution up to an additive constant.*

Note: The compatibility condition is similar to the condition on $\mathbf{b}$ when solving a linear system $A\mathbf{x} = \mathbf{b}$ with $\det(A) = 0$. Namely, if $\mathbf{b} \in \text{Range}(A)$ then there is an entire family of solutions. Conversely, if $\mathbf{b} \notin \text{Range}(A)$ then the linear system has no solutions.

4.4.2 Weak Form

In this section we apply the technique used in section 4.3.1 to derive a weak form for (4.86) in order to derive the weak form of the Neumann BVP (4.87).

Consider a classical solution u of the Neumann BVP, *i.e.* a solution $u \in C^2(\Omega)$ that satisfies

$$-\nabla^2 u = f, \quad \tfrac{\partial u}{\partial n} = h \text{ on } \partial\Omega. \tag{4.94}$$

It can be shown that the classical form (4.94) can be rewritten in a weighted residual form as

$$(-\nabla^2 u, w) = (f, w) \; \forall \; w \in W, \tag{4.95}$$

where W is defined analogous to (4.49) to be

$$W := \{w(\mathbf{x}) \mid \tfrac{\partial w}{\partial x} \text{ and } \tfrac{\partial w}{\partial y} \text{ are piecewise continuous and bounded on } \Omega\}. \tag{4.96}$$

On applying corollary B.1, we obtain

$$(\nabla u, \nabla w) - \int_\Omega \nabla \cdot (w \nabla u) d\Omega = (f, w) \; \forall \; w \in W \text{ and } \tfrac{\partial u}{\partial n} = h \text{ on } \partial\Omega. \tag{4.97}$$

Then by Gauss' divergence theorem (theorem 3.2), this equation can be rewritten as

$$(\nabla u, \nabla w) - \oint w \nabla u \cdot \mathbf{n} d\ell = (f, w) \; \forall \; w \in W \text{ and } \tfrac{\partial u}{\partial n} = h \text{ on } \partial\Omega. \tag{4.98}$$

Finally, we apply (4.87) and (4.88) in order to obtain the *weak form of the BVP*,

$$\boxed{(\nabla u, \nabla w) - \oint_{\partial\Omega} w h d\ell = (f, w) \; \forall w \in W.} \tag{4.99}$$

The discrete PDE problem is then stated as follows:

Problem: Find $u \in U$ so that (4.99) is satisfied.

Note that the set of candidate functions U is exactly the set of test functions$W$, as given by (4.96).

Discrete Weak Form

We now discretize U analogous to (4.55), obtaining

$$V^h = \{v^h \mid v^h = \sum_{j=1}^{n} c_j \phi_j(\mathbf{x})\}, \tag{4.100}$$

and

$$B^h = \{\phi_j(\mathbf{x})\}. \tag{4.101}$$

Note that due to the nature of the Neumann BVP the c_j remain unspecified at the boundaries. The discrete weak form then becomes

$$(\nabla v^h, \nabla w^h) - \oint w^h h d\ell = (f, w^h) \; \forall \; w^h \in V^h. \tag{4.102}$$

On expanding v^h in terms of ϕ_j and substituting ϕ_i for w^h, we obtain the desired form of this expression,

$$\sum_{j=1}^{n} c_j (\nabla \phi_j, \nabla \phi_i) - \oint_{\partial\Omega} \phi_i h d\ell = (f, \phi_i) \; \forall \; \phi_i \in B_0^h. \tag{4.103}$$

In order to solve for v^h, we must then solve the $m \times m$ linear system given by (4.103). As anticipated, if everything has been computed successfully we obtain that $\det(A) = 0$ with $L \in \mathrm{Range(A)}$. This result corresponds to the infinite family of solutions $v^h + c$ with $c \in \mathbb{R}$. Fortunately, we can fix c by choosing $v^h(x_i)$ explicitly in one node and hence obtain a solvable linear system.

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* $\vec{x} \in V, \alpha \in \mathbb{R}$*,*

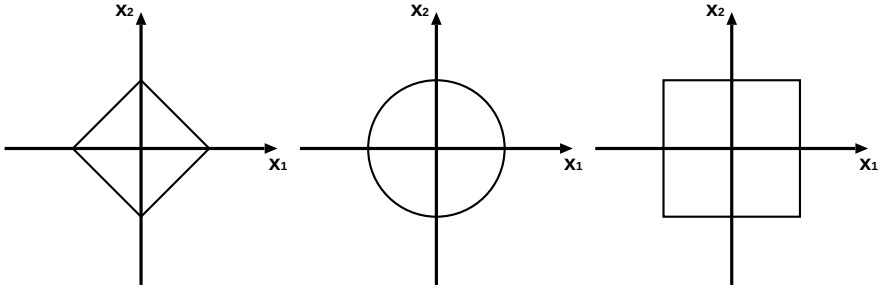
3) $\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$ *for all* $\vec{x}, \vec{y} \in V$*.*

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:

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

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:

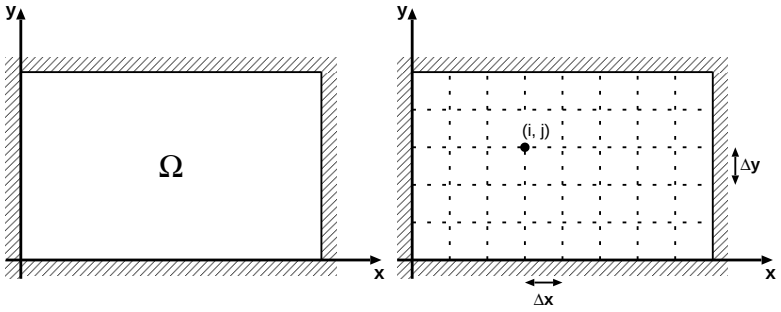
$$\begin{aligned}
\|u\|_2 &= \sqrt{\textstyle\int_a^b u(x)^2 dx} && \text{(2-norm)} \\
\|u\|_1 &= \textstyle\int_a^b |u(x)| dx && \text{(1-norm)} \\
\|u\|_\infty &= \operatorname*{ess\,sup}_{[a,b]} |u(x)| && (\infty\text{-norm}) \\
\|u\|_p &= \left(\textstyle\int_a^b |u(x)|^p\right)^{\frac{1}{p}} && (p\text{-norm})
\end{aligned}$$

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:

$$\begin{aligned}
\|u\|_2 &= \sqrt{\textstyle\iint_\Omega u(x,y)^2 dx} && \text{(2-norm)} \\
\|u\|_1 &= \textstyle\iint_\Omega |u(x,y)| dx && \text{(1-norm)} \\
\|u\|_\infty &= \operatorname*{ess\,sup}_{[a,b]} |u(x,y)| && (\infty\text{-norm}) \\
\|u\|_p &= \left(\textstyle\iint_\Omega |u(x,y)|^p\right)^{\frac{1}{p}} && (p\text{-norm})
\end{aligned}$$

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 $i = 0, \ldots, m$ and $j = 0, \ldots, n$. Then the interpolating grid function $u_{i,j}$ approximating $u(x, y)$ on the grid (x_i, y_j) is defined by

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

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

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

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.

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)

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 $\mathbb{R}^m$ ($1 \leq p \leq \infty$). 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*

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

where $\lambda_1, \lambda_2, \ldots, \lambda_m$ 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

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

[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.

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

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

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

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

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

P_6) The matrix norm satisfies the triangle inequality,

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

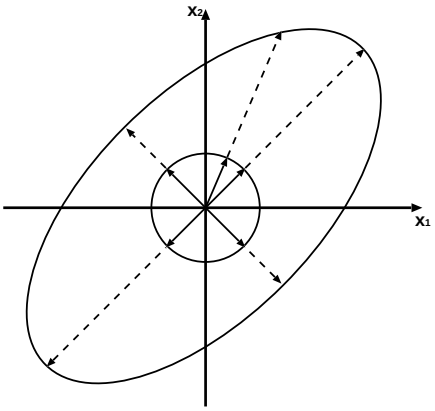
Example Consider the matrix $A \in \mathbb{R}^{2\times 2}$ defined by

$$A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}. \tag{A.21}$$

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 $\|A\|_2 = 4$.

APPENDIX B

Extension of Integration by Parts to 2D

Proposition B.1 *Let f and $\mathbf{g}$ be, respectively, a continuously differentiable scalar field and vector field defined on some set $\Omega \in \mathbf{R}^n$. Then,*

$$\nabla \cdot (f\mathbf{g}) = \nabla f \cdot \mathbf{g} + f(\nabla \cdot \mathbf{g}). \tag{B.1}$$

Proof: (to be written)

Corollary B.1 *Let f and g be, respectively, once and twice continuously differentiable scalar fields defined on some set $\Omega \in \mathbf{R}^n$. Then,*

$$\nabla \cdot (f\nabla g) = \nabla f \cdot \nabla g + f\nabla^2 g. \tag{B.2}$$

Proof: This result follows immediately from proposition B.1.