Numerical Solution of Partial Differential Equations

Andrea Cangiani[1]
Mathematics Area
SISSA
e-mail: Andrea.Cangiani@sissa.it

March 2024

[1]The bulk of these notes were developed together with Emmanuil Georgoulis at the University of Leicester.

Chapter 1

Introduction to Partial Differential Equations

1.1 Introduction

We give some basic definition, classification, and results from the theory of partial differential equation which are of importance when studying numerical methods for their solution. We refer the reader to any classical book on partial differential equations (eg. Evans') for a more complete presentation.

A partial differential equation (PDE) is an equation involving an unknown function of two or more variables with some of its partial derivatives. PDEs are of fundamental importance in applied mathematics and physics, and have recently shown to be useful in as varied disciplines as financial modelling and modelling of biological systems. More specifically, we have the following definition.

Definition 1.1 *Let $\Omega \subset \mathbb{R}^d$ be an open subset of $\mathbb{R}^d$ (called the* domain of definition*), for $d > 1$ a positive integer (called the* dimension*), and denote by $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ a vector in Ω. Let (unknown) function $u : \Omega \to \mathbb{R}$ whose partial derivatives up to order k (for k positive integer) exist. A* partial differential equation of order k in Ω in d dimensions *is an equation of the form:*

$$F(x, u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \ldots, \frac{\partial u}{\partial x_d}, \ldots, \frac{\partial u^k}{\partial x_{d-1} \partial x_d^{k-1}}) = 0, \tag{1.1}$$

where $F : \Omega \times \mathbb{R} \times \mathbb{R}^d \times \cdots \times \mathbb{R}^{d^{k-1}} \times \mathbb{R}^{d^k} \to \mathbb{R}$ is a given function.

Systems of PDEs are analogously defined by considering vector-valued u and F.

We shall be mostly interested in PDEs in two and three dimensions (as these are the ones most often appearing in practical applications), and we shall confine the notation to these cases using (x, y) and (t, x) or (x, y, z) and (t, x, y) to describe two- and three-dimensional vectors respectively (when the notation t is used for an independent variable, this variable should almost always describing "time"). Nevertheless, many properties and ideas described below apply also to the general case of d-dimensions for $d > 3$.

Also, to simplify the notation, we shall often resort to the more compact notation u_x, u_y, u_{xx}, u_{xy}, etc., to signify partial derivatives $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial^2 u}{\partial x^2}$, $\frac{\partial^2 u}{\partial x \partial y}$, etc., respectively.

Definition 1.2 *Consider the notation of Definition 1.1. We call the (classical)* general solution *of the PDE (1.1), the family of functions $u : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ that has continuous partial derivatives up to (and including) order k and that satisfies (1.1).*

There is *no method of solving PDEs that works in general*: different methods work for different families of PDEs. Therefore, it is important to identify such families of PDEs that admit similar properties and, subsequently to describe particular methods of solving PDEs from each such family. Such properties, as we shall see, are also of paramount importance in the selection of the right computational schemes for the numerical solution of the PDE at hand.

1.2 Classification of PDEs

To study PDEs it is often useful to classify them into various families, since PDEs belonging to particular families can be characterised by similar behaviour and properties. There are many and varied classifications for PDEs. Perhaps the most widely accepted and generally useful classification is the distinction between linear and non-linear PDEs. In particular, we have the following definition.

Definition 1.3 *If the PDE (1.1) can be written in the form*

$$\begin{aligned} &a(\mathbf{x})u + b_1(\mathbf{x})u_{x_1} + b_2(\mathbf{x})u_{x_2} + \cdots + b_d(\mathbf{x})u_{x_d} \\ &+ c_1(\mathbf{x})u_{x_1x_1} + \cdots + c_2(\mathbf{x})u_{x_1x_2} + \cdots + c_{d^2}(\mathbf{x})u_{x_dx_d} + \cdots = f(\mathbf{x}), \end{aligned} \tag{1.2}$$

i.e., if the coefficients of the unknown function u and of all its derivatives depend only *on the independent variables $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, then it is called a* linear PDE. *If it is not possible to write (1.1) in the form (1.2), then it is called a* nonlinear PDE.

The family of nonlinear PDEs can be further subdivided into smaller families of PDEs. In particular we have the following definition.

Definition 1.4 *Consider a nonlinear PDE of order k with unknown solution u.*

- *If the coefficients of the k order partial derivatives of u are functions of the independent variables $\mathbf{x} = (x_1, x_2, \ldots, x_d)$* only*, then this is called a* semilinear PDE.
- *If the coefficients of the k order partial derivatives of u are functions of the independent variables $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ and/or of partial derivatives of u of order at most $k-1$ (including u itself) , then this is called a* quasilinear PDE.
- *If a (nonlinear) PDE is not quasilinear, then it is called* fully nonlinear.

Clearly a semilinear PDE is also a quasilinear PDE.

LINEARISATION

Example 1.5 *We give some examples of nonlinear PDEs along with their classifications.*

- *The* reaction-diffusion equation
$$u_t = u_{xx} + u^2,$$
is a semilinear PDE.
- *The* inviscid Burgers' equation
$$u_t + uu_x = 0,$$
is a quasilinear PDE and it is not a semilinear PDE.
- *The* Korteweg-de Vries (KdV) equation
$$u_t + uu_x + u_{xxx} = 0,$$
is a semilinear PDE.
- *The* Monge-Ampère equation
$$u_{xx}u_{yy} - (u_{xy})^2 = 0,$$
is a fully nonlinear PDE.

The above classification of PDEs into linear, semilinear, quasilinear, and fully nonlinear is, roughly speaking, a classification of "increasing difficulty" in terms of studying and solving PDEs. Indeed, the mathematical theory of linear PDEs is now well understood. On the other hand, less is known about semilinear PDEs and quasilinear PDEs, and even less about fully nonlinear PDEs.

1.3 First order linear PDEs

We begin our study of linear PDEs with the case of first order linear PDEs. To simplify the discussion ,we shall only consider equations in 2 dimensions, i.e., for $d = 2$; the case of three or more dimensions can be treated in a completely analogous fashion. We begin with an example.

Example 1.6 *We consider the following linear transport PDE in* $\mathbb{R}^2$*:*

$$u_x + u_y = 0. \tag{1.3}$$

To find its general solution, we perform the following transformation of coordinates (also known as change of variables *in Calculus): we consider new variables* $(\xi, \eta) \in \mathbb{R}^2$ *defined by the transformation of coordinates*

$$(x, y) \to (\xi, \eta) \quad , \text{where } \xi(x, y) = x + y \quad \text{and} \quad \eta(x, y) = y - x.$$

We can also calculate the inverse transformation of coordinates

$$(\xi, \eta) \to (x, y),$$

by solving with respect to x *and* y*, obtaining*

$$x = \frac{1}{2}(\xi - \eta) \quad \text{and} \quad y = \frac{1}{2}(\xi + \eta). \tag{1.4}$$

We write the PDE (1.3) in the new coordinates, using the chain rule from Calculus. Setting $v(\xi, \eta) = u(x(\xi, \eta), y(\xi, \eta))$ *we have, respectively:*

$$u_x = v_\xi \xi_x + v_\eta \eta_x, \qquad u_y = v_\xi \xi_y + v_\eta \eta_y,$$

giving

$$u_x = v_\xi - v_\eta, \qquad u_y = v_\xi + v_\eta.$$

Putting these back to the PDE (1.3), we deduce

$$0 = u_x + u_y = v_\xi - v_\eta + v_\xi + v_\eta = 2v_\xi \quad \text{or} \quad v_\xi = 0. \tag{1.5}$$

Integrating this equation with respect to ξ*, we arrive to*

$$v(\xi, \eta) = f(\eta),$$

for any differentiable function of one variable $f : \mathbb{R} \to \mathbb{R}$*. Using now the inverse transformation of coordinates (1.4), we conclude that the general solution of the PDE (1.3) is given by:*

$$u(x, y) = v(\xi(x, y), \eta(x, y)) = f(\eta(x, y)) = f(y - x).$$

The change of variables $(x, y) \to (\xi, \eta)$ is essentially a clockwise rotation of the axes by an angle $\frac{\pi}{4}$. After rotation, the PDE takes the simpler form (1.5), which can be interpreted geometrically as: v is constant with respect to the variable ξ. In other words, the solution u is constant when $y - x = c$, for any constant $c \in \mathbb{R}$. Hence, the straight lines of the form $y = x + c$ "characterise" the solution of the PDE above; such curves are called *characteristic curves* of a PDE, as we shall see below.

Next, we shall incorporate these ideas into the case of the general first order linear PDE. The general form of a 1st order linear PDE in 2 dimensions can be written as:

$$a(x, y)u_x + b(x, y)u_y + c(x, y)u = g(x, y), \quad \text{for} \quad (x, y) \in \Omega \subset \mathbb{R}^2, \tag{1.6}$$

where a, b, c, g are functions of the independent variables x and y only. We also assume that a, b have continuous first partial derivatives, and that they do *not* vanish simultaneously at any point of the domain of definition Ω. Finally ,we assume that the solution u of the PDE (1.6) has continuous first partial derivatives.

Consider a transformation of coordinates of $\mathbb{R}^2$:

$$(x, y) \leftrightarrow (\xi, \eta),$$

with $\xi = \xi(x, y)$ and $\eta = \eta(x, y)$, which is assumed to be smooth (that is, the functions $\xi(x, y)$ and $\eta(x, y)$ have all derivatives with respect to x and y well-defined) and non-singular, i.e., its Jacobian

$$\frac{\partial(\xi, \eta)}{\partial(x, y)} := \begin{vmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{vmatrix} = \xi_x \eta_y - \xi_y \eta_x \neq 0, \tag{1.7}$$

in Ω; (this requirement ensures that the change of variables is meaningful, in the sense that it is one-to-one and onto). We also denote by $x = x(\xi, \eta)$ and $y = y(\xi, \eta)$ the inverse transformation, as it will be useful below.

We write the PDE (1.6) in the new coordinates, using the chain rule. Setting $v(\xi, \eta) = u(x(\xi, \eta), y(\xi, \eta))$ we have, respectively:

$$u_x = v_\xi \xi_x + v_\eta \eta_x, \qquad u_y = v_\xi \xi_y + v_\eta \eta_y,$$

giving

$$(a\xi_x + b\xi_y)v_\xi + (a\eta_x + b\eta_y)v_\eta + cv = g(x(\xi, \eta), y(\xi, \eta)), \tag{1.8}$$

after substitution into (1.6). To simplify the above equation, we require that the function $\eta(x, y)$ is such that

$$a\eta_x + b\eta_y = 0; \tag{1.9}$$

if this is the case then (1.8) becomes an ordinary differential equation with respect to the independent variable ξ, whose solution can be found by standard separation of variables.

The equation (1.9) is a slightly simpler PDE of first order than the original PDE. To find the required η we are seek to construct curves such that $\eta(x, y) = const$ for any constant; these are called the *characteristic curves* of the PDE (compare this with the straight lines of the example above).

Differentiating this equation with respect to x, we get

$$0 = \frac{\mathrm{d}\, const}{\mathrm{d}x} = \frac{\mathrm{d}\eta(x, y)}{\mathrm{d}x} = \eta_x \frac{\mathrm{d}x}{\mathrm{d}x} + \eta_y \frac{\mathrm{d}y}{\mathrm{d}x} = \eta_x + \eta_y \frac{\mathrm{d}y}{\mathrm{d}x},$$

where in the penultimate equality we made use of the chain rule for functions of two variables; the above equality yields

$$\frac{\eta_x}{\eta_y} = -\frac{\mathrm{d}y}{\mathrm{d}x}, \tag{1.10}$$

assuming, without loss of generality, that $\eta_y \neq 0$ (for otherwise, we argue as above with the rôles of the x and y variables interchanged, and we get necessarily $\eta_x \neq 0$ from hypothesis (1.7)).

Using (1.10) on (1.9), we deduce the *characteristic equation*:

$$-a\frac{\mathrm{d}y}{\mathrm{d}x} + b = 0, \quad \text{or} \quad \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b}{a}, \tag{1.11}$$

assuming, without loss of generality that $a \neq 0$ near the point (x_0, y_0) (for otherwise, we have that necessarily $b \neq 0$ near the point (x_0, y_0), as a, b cannot vanish simultaneously at any point due to hypothesis, and we can apply the same argument as above with x and y interchanged). Equation (1.11) is an ordinary differential equation of first order that can be solved using standard separation of variables to give a solution $f(x, y) = const$, say. Setting $\eta = f(x, y)$ and ξ to be any function for which (1.7) holds, we can easily see that (1.9) holds also. Therefore, the PDE (1.6) can be written as

$$(a\xi_x + b\xi_y)v_\xi + cv = g(x(\xi, \eta), y(\xi, \eta)), \quad \text{or} \quad v_\xi + \frac{c}{(a\xi_x + b\xi_y)}v = \frac{g(x(\xi, \eta), y(\xi, \eta))}{(a\xi_x + b\xi_y)},$$

which is an ordinary differential equation of first order with respect to ξ and can be solved using the (standard) method of multipliers to find $v(\xi, \eta)$. Using the inverse transformation of coordinates, we can now find the solution $u(x, y)$ from $v(\xi, \eta)$. This is the so-called *method of characteristics* in finding the solution to a first order PDE.

1.3.1 Problems

Problem 1.7 *Show that the characteristic curves for the PDE*

$$yu_x - xu_y + yu = xy \quad \text{for } y \neq 0,$$

are concentric circles centred at the origin. Then, use the method of characteristics to show that the general solution is

$$u(x,y) = v(\xi(x,y),\eta(x,y)) = \xi(x,y) - 1 + f(\eta(x,y)) = x - 1 + f\big(\frac{x^2+y^2}{2}\big)\mathrm{e}^{-x},$$

for any differentiable function f of one variable.

Problem 1.8 *Use the method of characteristics to find the general solution of the first order linear PDE:*

$$u_x - 2u_y = 0.$$

1.4 Second order linear PDEs

Next up, we have linear PDEs of 2nd order. Here, for simplicity, we shall consider only equations in 2 dimensions, i.e., for $d = 2$. The general form of a 2nd order linear PDE in 2 dimensions can be written as:

$$au_{xx} + 2bu_{xy} + cu_{yy} + du_x + eu_y + fu = g, \quad \text{for} \quad (x,y) \in \Omega \subset \mathbb{R}^2, \tag{1.12}$$

where a, b, c, d, e, f, g are functions of the independent variables x and y only. We also assume that a, b, c have continuous second partial derivatives, and that they do *not* vanish simultaneously at any point of the domain of definition Ω. Finally, we assume that the solution u of the PDE (1.12) has continuous second partial derivatives. We shall classify PDEs of the form (1.12) in different types, depending on the sign of the *discriminant* defined by

$$\mathcal{D} := b^2 - ac,$$

at each point $(x_0, y_0) \in \Omega$. More specifically, we have the following definition.

Definition 1.9 *Let $\mathcal{D} = b^2 - ac$ be the discriminant of a second order PDE of the form (1.12) in $\Omega \subset \mathbb{R}^2$ and let a point $(x_0, y_0) \in \Omega$.*

- *If $\mathcal{D} > 0$ at the point (x_0, y_0), the PDE is said to be* hyperbolic *at (x_0, y_0).*
- *If $\mathcal{D} = 0$ at the point (x_0, y_0), the PDE is said to be* parabolic *at (x_0, y_0).*
- *If $\mathcal{D} < 0$ at the point (x_0, y_0), the PDE is said to be* elliptic *at (x_0, y_0).*

The equation is said to be hyperbolic, parabolic or elliptic in the domain Ω if it is, respectively, hyperbolic, parabolic or elliptic at all points of Ω.

Example 1.10 *The following are the archetypal examples of linear second order PDEs $\mathbb{R}^2$.*

- *The* wave equation
$$u_{tt} + cu_{xx} = 0,$$
with $c < 0$, is hyperbolic.
- *The* heat *or* diffusion equation
$$u_t + au_{xx} = 0,$$
with $a < 0$, is parabolic.
- *The* Laplace equation
$$\Delta u := u_{xx} + u_{yy} = 0,$$
is elliptic.

Similarly, in 3 dimensions, the wave equation reads $u_{tt} + c\Delta u = 0$ and heat equation reads $u_t + a\Delta u$, with Δ the Laplace operator with respect to the variables x, y. In turns, the Laplace equation reads $\Delta u(x,y,z) = 0$.

The well known character of the Laplace, diffusion, and wave equations reflect the general character of the classes they represent: elliptic equation are generally time-independent, parabolic equations are time-dependent and usually model diffusion phenomena, and hyperbolic equations are also time-dependent but they model transport, wave-like phenomena which are characterised by finite speed of propagation.

Different *physics* can be modelled locally by considering so-called changing type PDEs.

Example 1.11 *The following are examples of equations of changing type.*

- *The* Grušin equation
$$u_{xx} + x^2 u_{yy} = 0,$$
is elliptic in the set $\{(x,y) \in \mathbb{R}^2 : x \neq 0\}$ *and parabolic in the set* $\{(x,y) \in \mathbb{R}^2 : x = 0\}$.

- *The* Tricomi equation
$$yu_{xx} + u_{yy} = 0,$$
is elliptic in the set $\{(x,y) \in \mathbb{R}^2 : y > 0\}$, *parabolic in the set* $\{(x,y) \in \mathbb{R}^2 : y = 0\}$, *and hyperbolic in the set* $\{(x,y) \in \mathbb{R}^2 : y < 0\}$.

The relevance of the above classification stems from the following result, assuring that applying a change of variables will not alter the type of the PDE.

Theorem 1.12 *The sign of the discriminant* $\mathcal{D}$ *of a second order PDE of the form (1.12) in* $\Omega \subset \mathbb{R}^2$ *is invariant under smooth non-singular transformations of coordinates (also known as* change of variables*).*

Proof. Consider a transformation of coordinates of $\mathbb{R}^2$:
$$(x,y) \leftrightarrow (\xi,\eta),$$
with $\xi = \xi(x,y)$ and $\eta = \eta(x,y)$, which is assumed to be smooth (that is, the functions $\xi(x,y)$ and $\eta(x,y)$ have all derivatives with respect to x and y well-defined) and non-singular, i.e., its Jacobian
$$\frac{\partial(\xi,\eta)}{\partial(x,y)} := \begin{vmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{vmatrix} = \xi_x\eta_y - \xi_y\eta_x \neq 0, \tag{1.13}$$
in Ω. We also denote by $x = x(\xi,\eta)$ and $y = y(\xi,\eta)$ the inverse transformation, as it will be useful below.

We write the PDE (1.12) in the new coordinates, using the chain rule. Setting $v(\xi,\eta) = u(x(\xi,\eta), y(\xi,\eta))$ we have, respectively:
$$u_x = v_\xi \xi_x + v_\eta \eta_x, \qquad u_y = v_\xi \xi_y + v_\eta \eta_y, \tag{1.14}$$
giving
$$\begin{aligned} u_{xx} &= v_{\xi\xi}\xi_x^2 + 2v_{\xi\eta}\xi_x\eta_x + v_{\eta\eta}\eta_x^2 + v_\xi \xi_{xx} + v_\eta \eta_{xx}, \\ u_{yy} &= v_{\xi\xi}\xi_y^2 + 2v_{\xi\eta}\xi_y\eta_y + v_{\eta\eta}\eta_y^2 + v_\xi \xi_{yy} + v_\eta \eta_{yy}, \\ u_{xy} &= v_{\xi\xi}\xi_x\xi_y + v_{\xi\eta}(\xi_x\eta_y + \xi_y\eta_x) + v_{\eta\eta}\eta_x\eta_y + v_\xi \xi_{xy} + v_\eta \eta_{xy}. \end{aligned} \tag{1.15}$$

Inserting (1.14) and (1.15) into (1.12), and factorising accordingly, we arrive to
$$Av_{\xi\xi} + 2Bv_{\xi\eta} + Cv_{\eta\eta} + Dv_\xi + Ev_\eta + fv = g, \tag{1.16}$$
where the new coefficients A, B, C, D and E are given by
$$\begin{aligned} A &= a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2, \\ B &= a\xi_x\eta_x + b(\xi_x\eta_y + \xi_y\eta_x) + c\xi_y\eta_y, \\ C &= a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2, \\ D &= a\xi_{xx} + 2b\xi_{xy} + c\xi_{yy} + d\xi_x + e\xi_y, \\ E &= a\eta_{xx} + 2b\eta_{xy} + c\eta_{yy} + d\eta_x + e\eta_y. \end{aligned} \tag{1.17}$$

Thus the discriminant of the PDE in new variables (1.16), is given by
$$\begin{aligned} B^2 - AC &= (a\xi_x\eta_x + b(\xi_x\eta_y + \xi_y\eta_x) + c\xi_y\eta_y)^2 - (a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2)(a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2) \\ &= \cdots = (b^2 - ac)(\xi_x\eta_y - \xi_y\eta_x)^2 = (b^2 - ac)\Big(\frac{\partial(\xi,\eta)}{\partial(x,y)}\Big)^2. \end{aligned} \tag{1.18}$$

This means that the discriminant $B^2 - AC$ of (1.16) has always the same sign as the discriminant $b^2 - ac$ of (1.12), as $\frac{\partial(\xi,\eta)}{\partial(x,y)} \neq 0$ from the hypothesis and, therefore, $\big(\frac{\partial(\xi,\eta)}{\partial(x,y)}\big)^2 > 0$. Since the discriminant of the transformed PDE has always the same sign as the one of the original PDE, the type of the PDE remains invariant. □

Let us now consider some special transformations for PDEs of each type. What we shall see is that, given certain transformation, it is possible to write (1.12) locally in much simpler form, the so-called *canonical form*.

Example 1.13 *Consider the wave equation*

$$u_{xx} - u_{yy} = 0,$$

which as we saw before is hyperbolic in $\mathbb{R}^2$*. Let us also consider the transformation of coordinates of* $\mathbb{R}^2$*:*

$$(x, y) \leftrightarrow (\xi, \eta), \quad \textit{with} \quad \xi = x + y \quad \textit{and} \quad \eta = x - y.$$

It is, of course, smooth as $x + y$ *and* $x - y$ *are infinite times differentiable with respect to* x *and* y*, and it is non-singular. The transformed equation is given by*

$$4v_{\xi\eta} = 0, \quad \textit{or} \quad v_{\xi\eta} = 0.$$

(Note that, indeed, this equation is still hyperbolic.) From this canonical form, we can in fact compute the general solution of the wave equation. Indeed, integrating with respect to η*, we arrive to* $v_\xi = h(\xi)$ *for an arbitrary continuously differentiable function* h*. Integrating now the last equality with respect to* ξ*, we deduce* $v = \int^\xi h(s)\mathrm{d}s + G(\eta)$*. If we set* $F(\xi) := \int^\xi h(s)\mathrm{d}s$*, to simplify the notation, we get* $v(\xi, \eta) = F(\xi) + G(\eta)$*, for an arbitrary twice continuously differentiable function* G*, or equivalently*

$$u(x, y) = F(x + y) + G(x - y),$$

for all twice continuously differentiable functions F *and* G *of one variable: the solution is the sum of the left-travelling function* F *and the right-travelling function* G*. This formula is due to d'Alembert (1717-83).*

We now investigate the following question: is it always possible to find transformations of coordinates that make the general PDE (1.12) "simpler"?

For the general PDE, we employ a geometric argument. We seek functions $\xi(x, y)$ and $\eta(x, y)$ for which we have

$$a\xi_x^2 + 2b\xi_x\xi_y + c\xi_y^2 = 0 \quad \text{and} \quad a\eta_x^2 + 2b\eta_x\eta_y + c\eta_y^2 = 0; \tag{1.19}$$

i.e., $A = C = 0$ for the coefficients of the transformed PDE (1.16). The equations (1.19) are PDEs of first order, for which we are now seeking to construct curves such that $\xi(x, y) = const$ for any constant. When (x, y) are points on a curve, i.e, they are such that $\xi(x, y) = const$, they are dependent. Hence, differentiating this equation with respect to x, we get

$$0 = \frac{\mathrm{d}\, const}{\mathrm{d}x} = \frac{\mathrm{d}\xi(x, y)}{\mathrm{d}x} = \xi_x\frac{\mathrm{d}x}{\mathrm{d}x} + \xi_y\frac{\mathrm{d}y}{\mathrm{d}x} = \xi_x + \xi_y\frac{\mathrm{d}y}{\mathrm{d}x},$$

where in the penultimate equality we made use of the chain rule for functions of two variables; the above equality yields

$$\frac{\xi_x}{\xi_y} = -\frac{\mathrm{d}y}{\mathrm{d}x}, \tag{1.20}$$

assuming, without loss of generality, that $\xi_y \neq 0$. Now, we go back to the desired equations (1.19), and we divide the first equation by ξ_y^2 to obtain

$$a\Big(\frac{\xi_x}{\xi_y}\Big)^2 + 2b\frac{\xi_x}{\xi_y} + c = 0,$$

and, using (1.20), we arrive to

$$a\Big(\frac{\mathrm{d}y}{\mathrm{d}x}\Big)^2 - 2b\frac{\mathrm{d}y}{\mathrm{d}x} + c = 0, \tag{1.21}$$

which is called the *characteristic equation* for the PDE (1.12). This is a quadratic equation for $\frac{\mathrm{d}y}{\mathrm{d}x}$, with discriminant $\mathcal{D} = b^2 - ac$! The roots of the characteristic equation are given by

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{b \pm \sqrt{\mathcal{D}}}{a}. \tag{1.22}$$

Each of the equations above is a first order ordinary differential equation that can be solved using standard separation of variables to give (families of) solutions $f_1(x, y) = const$ and $f_2(x, y) = const$, say. The curves defined by the equations $f_1(x, y) = const$ and $f_2(x, y) = const$ are called the *characteristic curves* of the second order PDE.

Therefore, if the original PDE (1.12) is hyperbolic, i.e., if $\mathcal{D} > 0$, the characteristic equation has two real distinct roots, giving two real distinct characteristics curves for the PDE. If the original PDE (1.12) is parabolic, thereby $\mathcal{D} = 0$, the characteristic equation has one double root, giving one real characteristic curve for the PDE. Finally, if the original PDE (1.12) is elliptic, thereby $\mathcal{D} < 0$, the characteristic equation has no real roots, and therefore the PDE has **no** real characteristic curves, but as we shall see below it has complex characteristic curves. The characteristic curves can be thought as the "natural directions" in which the PDE "communicates information" to different points in its domain of definition Ω. With this statement in mind, it is possible to see that each type of PDE models different phenomena and also admits different properties, rendering the above classification into hyperbolic, parabolic and elliptic PDEs of great importance.

From the above development it is immediate to prove the following theorems characterising the canonical form for those PDEs admitting real characteristic curves, namely those of hyperbolic or parabolic type.

Theorem 1.14 *Let (1.12) be a hyperbolic PDE. Then, for every $(x_0, y_0) \in \Omega$ there exists a transformation of coordinates $(x, y) \leftrightarrow (\xi, \eta)$ in the neighbourhood of (x_0, y_0), such that (1.12) can be written as*

$$v_{\xi\eta} + \cdots = g, \tag{1.23}$$

where "..." are used to signify the terms involving u, u_x, or u_y. This is called the canonical form *of a hyperbolic PDE.*

Theorem 1.15 *Let (1.12) be parabolic PDE. Then, for every $(x_0, y_0) \in \Omega$ there exists a transformation of coordinates $(x, y) \leftrightarrow (\xi, \eta)$ in the neighbourhood of (x_0, y_0), such that (1.12) can be written as*

$$v_{\xi\xi} + \cdots = g, \tag{1.24}$$

where "..." are used to signify the terms involving u, u_x, or u_y. This is called the canonical form *of a parablic PDE.*

These leaves out the elliptic case (when the characteristic equation (1.21) has no real roots). By other means (eg. the theory of analytic functions) it is still possible to prove that also in this case a canonical form always exists.

Theorem 1.16 *Let (1.12) be an elliptic PDE. Then, for every $(x_0, y_0) \in \Omega$ there exists a transformation of coordinates $(x, y) \leftrightarrow (\xi, \eta)$ in the neighbourhood of (x_0, y_0), such that (1.12) can be written as*

$$v_{\xi\xi} + v_{\eta\eta} + \cdots = g, \tag{1.25}$$

where "..." are used to signify the terms involving u, u_x, or u_y. This is called the canonical form *of an elliptic PDE.*

Remark 1.17 *Notice that the whole discussion in this section about linear second order PDEs will still be valid for the case of semilinear second order PDEs too! Indeed, since in second order semilinear PDEs the non-linearities are not present in the coefficients of the second order derivatives, the calculations and the theorems above will still be valid.*

1.4.1 Problems

Problem 1.18 *Consider the PDE:*

$$(1 - M^2)u_{xx} + u_{yy} = 0.$$

(This equation models the potential of the velocity field of a fluid around a planar obstacle; M is called the Mach number.) What is the type of the above second order linear PDE for different values of M? If you know what a "sonic boom" is, can you see a relation to it and the properties of the equation above?

Problem 1.19 *Calculate the characteristic curves of the Tricomi equation*

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

for $y \leq 0$. Show that, when $y < 0$ the Tricomi equation can be written in the canonical form (1.23)

Problem 1.20 *The Black-Scholes equation for a European call option with value $C = C(\tau, s)$ (τ the time variable and s is the asset price), is given by*

$$C_\tau + \frac{\sigma^2}{2} s^2 C_{ss} + rsC_s - rC = 0, \tag{1.26}$$

where r is a positive constant (the interest rate). What type of 2nd order linear PDE is (1.26) and why? Using the following transformation of coordinates of $\mathbb{R}^2$:

$$(\tau, s) \leftrightarrow (t, x), \quad \textit{with} \quad \tau = T - \frac{2t}{\sigma^2}, \quad \textit{and} \quad s = \mathrm{e}^x,$$

where T is a constant (the final time), show that (1.26) can be transformed into the following PDE in canonical form:

$$v_{xx} + (k-1)v_x - v_t - kv = 0, \tag{1.27}$$

where $v(t, x) := C(\tau(t, x), s(t, x)) = C(T - 2t/\sigma^2, \mathrm{e}^x)$, and $k := 2r/\sigma^2$. Setting now

$$v(t, x) = \mathrm{e}^{\alpha x + \beta t} u(t, x),$$

for some function $u = u(t, x)$, show that the transformed equation (1.27) can be written as

$$u_t - u_{xx} = 0,$$

when

$$\alpha = -\frac{1}{2}(k-1), \quad \textit{and} \quad \beta = -\frac{1}{4}(k+1)^2,$$

i.e., the Black-Scholes equation can be transformed into the heat equation!

1.5 The Cauchy problem and well-posedness of PDEs

We have seen, eg. by the method of characteristics, that the general solution of a PDEs contain unknown functions. These take the role of the unknown constants appearing in the general solution of ODEs.

In this section, we shall study some appropriate conditions that will be sufficient to specify the unknown functions and arrive to unique solutions. Hence we shall introduce the notion of *well-posedness* of a PDE problem.

Definition 1.21 *Consider a PDE of the form (1.1), of order k in Ω in d dimensions and let S be a (given) smooth surface on $\mathbb{R}^d$. Let also $n = n(x)$ denote the unit normal vector to the surface S at a point $\mathbf{x} = (x_1, x_2, \dots, x_d) \in S$. Suppose that on any point $\mathbf{x}$ of the surface S the values of the solution u and of all its directional derivatives up to order $k-1$ in the direction of n are given, i.e., we are given functions $f_0, f_1, \dots, f_{k-1} : S \to \mathbb{R}$ such that*

$$u(\mathbf{x}) = f_0(\mathbf{x}), \quad \textit{and} \quad \frac{\partial u}{\partial n}(\mathbf{x}) = f_1(\mathbf{x}), \quad \textit{and} \quad \frac{\partial^2 u}{\partial n^2}(\mathbf{x}) = f_2(\mathbf{x}), \dots, \quad \textit{and} \quad \frac{\partial^{k-1} u}{\partial n^{k-1}}(\mathbf{x}) = f_{k-1}(\mathbf{x}). \tag{1.28}$$

The Cauchy problem *consists of finding the unknown function(s) u that satisfy simultaneously the PDE and the conditions (1.28). The conditions (1.28) are called the* initial conditions *and the given functions $f_0, f_1, \dots, f_{k-1}$, will be referred to as the* initial data.

Example 1.22 *Consider the Cauchy problem for the 1st order transport equation*

$$\begin{cases} u_x + u_y = 0, \\ u(0, y) = \sin y \quad \textit{on } S = \{(x, y) \in \mathbb{R}^2 : x = 0\}. \end{cases} \tag{1.29}$$

The characteristic curves of the PDE are $y = x + c$, $c \in \mathbb{R}$; notice that S intersects all *of them. From Example 1.6 we also know that the general solution to the PDE is $u(x, y) = f(y - x)$, for all $(x, y) \in \mathbb{R}^2$. Using the initial condition we deduce that a solution to the Cauchy problem is given by $u(x, y) = \sin(y - x)$.*

Example 1.23 *Consider the Cauchy problem for the wave equation*

$$\begin{cases} u_{xx} - u_{yy} = 0, \\ u(x, 0) = \sin x, \quad \textit{and} \quad u_y(x, 0) = 0. \end{cases} \tag{1.30}$$

Here the surface S in Definition 1.21 is implicitly given as $S = \{(x, y) \in \mathbb{R}^2 : y = 0\}$. Imposing the initial conditions to the general solution to the wave equation from Example 1.13 we deduce that a solution to the Cauchy problem is

$$u(x, y) = F(x + y) + G(x - y) = \frac{1}{2}(\sin(x + y) + \sin(x - y)).$$

One question that arises is whether the solutions to the Cauchy problems in the previous examples are unique. A partial answer to this question is given by the celebrated Cauchy-Kovalevskaya Theorem, the proof of which can be found in any standard PDE theory textbook.

Theorem 1.24 (The Cauchy-Kovalevskaya Theorem) *Consider the Cauchy problem from Definition (1.21) for the case of a linear PDE of the form (1.2). Let $\mathbf{x}_0$ be a point of the initial surface S, which is assumed to be analytic. Suppose that S is* not *a characteristic surface at the point $\mathbf{x}_0$. Assume that all the coefficients of the PDE (1.2), the right-hand side f, and all the initial data $f_0, f_1, \dots, f_{k-1}$ are analytic functions on a neighbourhood of the point $\mathbf{x}_0$. Then the Cauchy problem has a solution u, defined in the neighbourhood of $\mathbf{x}_0$. Moreover, the solution u is analytic in a neighbourhood of $\mathbf{x}_0$ and it is unique in the class of analytic functions.*

Therefore, according to the Cauchy-Kovalevskaya Theorem (under the analyticity assumptions), the Cauchy problem has a solution which is unique in the space of analytic functions. Even if a PDE problem has a unique solution, this does not necessarily mean that the PDE problem is "well behaved". By well-behaved here we understand if the PDE problem changes "slightly" (e.g., by altering "slightly" some coefficient), then also its solution should change only "slightly" also. In other words, "well behaved" is to be understood as follows: "small" changes in the initial data or the PDE itself should *not* result to arbitrarily "large" changes in the behaviour of the solution to the PDE problem.

Definition 1.25 *A PDE problem is* well-posed *if the following 3 properties hold:*

- *the PDE problem has a solution*
- *the solution is unique*
- *the solution depends continuously on the PDE coefficients and the problem data.*

If a PDE problem is not *well-posed, then we say that it is* ill-posed.

The concept of well-posedness is due to Hadamard[1].

Example 1.26 *The Cauchy problem for the wave equation*

$$\begin{cases} u_{xx} - u_{yy} = 0, \\ u(x,0) = f(x), \quad u_y(x,0) = 0, \end{cases}$$

for some known *initial datum f, is an example of a well posed problem. Indeed, working completely analogously to Example 1.23, we can see that a solution to the above problem is given by*

$$u(x,y) = \frac{1}{2}\big(f(x-y) + f(x+y)\big).$$

The proof of uniqueness of solution is more involved and will be omitted (it is based on the so-called energy property of the wave equation).

Finally, to show the continuity of the solution to the initial data, we consider also the Cauchy problem

$$\tilde{u}_{xx} - \tilde{u}_{yy} = 0, \quad \textit{together with the initial conditions} \quad \tilde{u}(x,0) = \tilde{f}(x), \quad \tilde{u}_y(x,0) = 0,$$

i.e., we consider a different initial condition $\tilde{f}$ for the Cauchy problem, giving a new solution $\tilde{u}$. Working as above, we can immediately see that the solution to this Cauchy problem is given by

$$\tilde{u}(x,y) = \frac{1}{2}\big(\tilde{f}(x-y) + \tilde{f}(x+y)\big).$$

Now, we look at the difference of the solutions of the two Cauchy problems above. We have

$$u(x,y) - \tilde{u}(x,y) = \frac{1}{2}\big(f(x-y)+f(x+y)\big) - \frac{1}{2}\big(\tilde{f}(x-y)+\tilde{f}(x+y)\big) = \frac{1}{2}\Big(\big(f(x-y)-\tilde{f}(x-y)\big)+\big(f(x+y)-\tilde{f}(x+y)\big)\Big).$$

Hence if the difference $f(z) - \tilde{f}(z)$ is small for all $z \in \mathbb{R}$, then the difference $u - \tilde{u}$ will also be small! That is the solution depends continuously on the PDE coefficients and the problem data.

In Chapter 2, we shall consider appropriate conditions for each type of linear second order equations (elliptic, parabolic, hyperbolic), that result to well-posed problems.

1.5.1 Problems

Problem 1.27 *Show that the solution of the Cauchy problem for the wave equation*

$$\begin{cases} u_{xx} - u_{yy} = 0, \\ u(x,0) = f(x), \quad \textit{and} \quad u_y(x,0) = g(x), \end{cases}$$

for some known initial datum f and g is given by d'Alembert's formula

$$u(x,y) = \frac{1}{2}(f(x-y) + f(x+y)) + \frac{1}{2}\int_{x-y}^{x+y} g(s)ds.$$

Problem 1.28 *Find the solution to the Cauchy problem for the wave equation*

$$\begin{cases} u_{xx} - u_{yy} = 0, \\ u(x,0) = 0, \quad u_y(x,0) = g(x), \end{cases}$$

for some known initial datum g. Is this problem well-posed or ill-posed? Why?

[1] Jacques Salomon Hadamard (1865 - 1963), French mathematician

Problem 1.29 *This example is due to Hadamard. Consider the Cauchy problem for the Laplace equation*

$$\begin{cases} u_{xx} + u_{yy} = 0, \quad \textit{for } -\frac{\pi}{2} < x < \frac{\pi}{2}, \quad \textit{and} \quad y > 0, \\ u(x,0) = 0, \quad u_y(x,0) = \mathrm{e}^{-\sqrt{n}} \cos(nx), \quad \textit{for } -\frac{\pi}{2} \le x \le \frac{\pi}{2} \\ u(-\pi/2, y) = 0 = u(\pi/2, y), \quad \textit{for } y \ge 0, \end{cases}$$

for every $n = 1, 3, 5, \ldots$. *The solution to this problem can be found using the method of separation of variables to be*

$$u(x,y) = \frac{\mathrm{e}^{-\sqrt{n}}}{n} \cos(nx) \sinh(ny).$$

Show that this problem is ill-posed by observing that, while the change in the initial condition $u_y(x,0)$ *in function of* n *is exponentially small, the change in the respective solution is exponentially large.*

Chapter 2

Elliptic problems

2.1 The Laplace equation

We begin the discussion with Laplace equation:

$$\Delta u = 0, \quad \text{for} \quad (x_1, x_2, \ldots, x_d) \in \Omega \subset \mathbb{R}^d, \tag{2.1}$$

where $\Delta := (\cdot)_{x_1 x_1} + (\cdot)_{x_2 x_2} + \cdots + (\cdot)_{x_d x_d}$ denotes the so-called *Laplace operator* in d dimensions; in particular, in two dimensions Laplace equation reads:

$$\Delta u = u_{xx} + u_{yy} = 0, \quad \text{for} \quad (x, y) \in \Omega \subset \mathbb{R}^2, \tag{2.2}$$

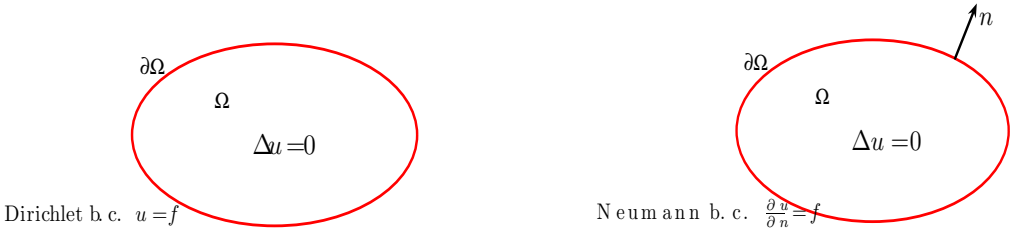
where $\Delta := (\cdot)_{xx} + (\cdot)_{yy}$. The non-homogeneous version of the Laplace equation, namely

$$\Delta u = f \quad \text{in } \Omega \tag{2.3}$$

for some known function $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$, is known as the *Poisson equation*.

Laplace and Poisson equations model predominately phenomena that do *not* evolve in time, typically properties of materials (elasticity, electric or gravitational charge), probability densities of random variables, etc. The solution typically represent the density of some physical quantity at equilibrium. The Laplace equation is also used to define harmonic functions[1], hence it is of central importance for the theory of complex functions.

As we saw in Chapter 1, Laplace (and therefore, Poisson) equation is the most classical example of PDE of elliptic type.



(a) Dirichlet boundary value problem.

(b) Neumann boundary value problem.

Figure 2.1: Dirichlet and Neumann boundary value problems. Change f to g!!!

For the problem to be well posed, we equip the Laplace equation with conditions along the whole of the boundary $\partial\Omega$ of the domain Ω. We shall call these *boundary conditions*[2]. We shall consider two types of boundary conditions, namely the *Dirichlet boundary condition*:

$$u(x, y) = g(x, y), \quad \text{for} (x, y) \in \partial\Omega, \tag{2.4}$$

[1] An harmonic function is a $\mathcal{C}^2$ function satisfying Laplace's equation on the plane.

[2] In the previous chapter, we talked about the Cauchy problem consisting of a PDE, together with initial conditions. The term "initial conditions" is used for PDEs that model evolution phenomena (i.e., PDEs for which one variable is "time"), for which the Cauchy problem is well posed. For elliptic PDEs, however, which model phenomena that do *not* evolve in time, it is conventional to use the term "boundary conditions" instead.

where $g : \partial\Omega \to \mathbb{R}$ is a known function fixing the value of u on the boundary, and the *Neumann boundary condition*:

$$\frac{\partial u}{\partial n}(x,y) = g(x,y), \quad \text{for}(x,y) \in \partial\Omega,$$

prescribing, instead, the normal flux. Here, $\frac{\partial u}{\partial n}(x,y)$ is the directional derivative of u in the direction of the unit outward normal vector n at the point (x,y) of the boundary $\partial\Omega$. We shall refer to the Laplace equation together with the Dirichlet boundary condition as the *Dirichlet boundary value problem* and to the Laplace equation together with the Neumann boundary condition as the *Neumann boundary value problem* (see Figure 2.1 for an illustration). Other boundary conditions include the so-called Robin condition which combined the Dirichlet and Neumann:

$$\frac{\partial u}{\partial n}(x,y) + \beta u(x,y) = g(x,y), \quad \text{for}(x,y) \in \partial\Omega,$$

and mixed boundary conditions in which different conditions are imposed on different portions of the boundary.

The problem composed by an equation such as (2.2) and boundary conditions (2.4) is termed a *boundary value problem*.

The question now is to study the well-posedness and further properties of such boundary value problems.In some special cases, it is possible to derive solutions of such boundary value problems. For instance, in Problem2.5 below the *exact* solution can be found by separation of variables.

More in general, a constructive proof of existence of solutions can be obtained by the method of fundamental solutions (eg. Green's functions, however constructing Green's function for general domains is in general difficult). We are not going to show this in this notes and refer the reader to any book on the theory of PDEs. Or perhaps add it from Evans or (more general) Gilbarg-Trudinger? Include Green's functions? However, we shall see below an alternative proof of existence for the *weak formulation* of the boundary value problem.

Regarding uniqueness and continuous dependence on the data, these can be shown based on the following *maximum principle*.

Proof of Maximum principle from Larsson
Exercise: proof of maximum principle from Evans

Example: if u is harmonic and positive on the boundary, then u is positive inside.

Proof of uniqueness from Evans.
Regularity from Evans?

2.1.1 Energy methods

Proof of stability and uniqueness (Larsson, Evans)

Dirichlet Principle (Evans) with proof.

2.1.2 Problems

Problem 2.1 *Let $\Omega = [0,a] \times [0,b] \subset \mathbb{R}^2$. Using the method of separation of variables, calculate the (unique) solution $u : \Omega \to \mathbb{R}$ to the Laplace boundary-value problem*

$$\begin{aligned} \Delta u &= 0 \quad \text{in } \Omega, && (2.5)\\ u(0,y) = u(a,y) = u(x,0) &= 0, \quad \text{for} \quad 0 \le x \le a,\ 0 \le y \le b, && (2.6)\\ u(x,b) &= g(x), \quad \text{for} \quad 0 \le x \le a, && (2.7) \end{aligned}$$

where $g : [0,a] \to \mathbb{R}$ is a known function whose Fourier sine expansion is given by

$$g(x) = \sum_{n=1}^{\infty} G_n \sin\left(\frac{n\pi x}{a}\right),$$

for known $G_n \in \mathbb{R}$.

Problem 2.2 *Let $\Omega = [0,a] \times [0,b] \subset \mathbb{R}^2$. We seek the (unique) solution $u : \Omega \to \mathbb{R}$ to the Laplace boundary-value problem*

$$\begin{aligned} \Delta u &= 0 \quad \text{in } \Omega, \\ u_x(0,y) = u_x(a,y) &= 0, \quad \text{for} \quad 0 \le x \le a,\ 0 \le y \le b, \\ u(x,0) &= 0, \quad \text{for} \quad 0 \le y \le b, \\ u(x,b) &= g(x), \quad \text{for} \quad 0 \le x \le a, \end{aligned}$$

where $g : [0,a] \to \mathbb{R}$ is a known function. Using the method of separation of variables, calculate the solution up to unknown constants, and give a condition that can enable us to calculate these unknown constants.

Find the solution for the case where the boundary condition $g : [0,a] \to \mathbb{R}$ is given by

$$g(x) = \begin{cases} \frac{x}{a}, & \text{if } 0 \le x < \frac{a}{2}; \\ 1 - \frac{x}{a}, & \text{if } \frac{a}{2} \le x \le a. \end{cases}$$

Problem 2.3 *Suppose, we want find the solution to the Poisson equation with Dirichlet boundary conditions*

$$\begin{aligned} \Delta u &= f(x,y) \quad \text{in } \Omega, \\ u(0,y) = u(a,y) = u(x,0) = u(x,b) &= 0, \quad \text{for} \quad 0 \le x \le a,\ 0 \le y \le b. \end{aligned}$$

Explain why the method of separation of variables cannot be applied directly. How would you go about solving this problem?

2.1.3 Variational formulation

functional sspaces

Variational formulation of general elliptic Dirichlet BVP with homog coundary conditions. (Larsson)

Lax-Milgram, well posedness

Dirichlet principle

regularity

Other problems: Nonhomogeneous problem, Neumann

Aggiungere per FEM : tracce, valli: $H^1 1 = H^1(\mathcal{T})$ plus traces are continuous

2.2 Weak derivatives

We give a brief list of the functional analytical notions required to the deifinition of weak problems for PDEs.

The aim of this section is to define a new concept of "differentiation" which will allow us to generalise the notion of a derivative of a function. To do so, we shall need to consider first some elementary concepts.

Definition 2.4 *Consider a function $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d$ open set. We define the* support *of f to be the closure*[3] *of the set $\{\mathbf{x} \in \Omega : f(\mathbf{x}) \neq 0\}$.*

Definition 2.5 *Let $\Omega \subset \mathbb{R}^d$ open set. We denote by $C_0^\infty(\Omega)$, the family of all functions $\phi : \Omega \to \mathbb{R}$ that are infinite times differentiable and have compact support*[4]*.*

The infinitely differentiable functions of compact support play a very important role in the modern theory of functions. In particular, they are utilised in generalising the concept of a derivative of a function.

[3] *Closure* of a set A in $\mathbb{R}^d$ is the smallest closed set containing A.

[4] A set $B \subset \mathbb{R}^d$ is *compact* if it is bounded and closed.

Definition 2.6 *Let $(a,b) \subset \mathbb{R}$ open interval. A function $g : (a,b) \to \mathbb{R}$ is called a* weak derivative *of a function $f : (a,b) \to \mathbb{R}$ if*

$$\int_a^b g(x)\phi(x)\,\mathrm{d}x = -\int_a^b f(x)\phi'(x)\mathrm{d}x \ < +\infty$$

for all *functions $\phi \in C_0^\infty((a,b))$* [5].

Theorem 2.7 *Let $(a,b) \subset \mathbb{R}$ open and $f : (a,b) \to \mathbb{R}$. If f is differentiable in (a,b) then it has a weak derivative g with $g = f'$ almost everywhere*[6].

Proof. Let $\phi \in C_0^\infty((a,b))$. Since ϕ has compact support (i.e., bounded and closed), the endpoints a and b cannot be in the support of ϕ (since the support of ϕ is closed and, therefore, can only be contained strictly in the interval (a,b)). Hence $\phi(a) = \phi(b) = 0$ (or, strictly speaking, $\lim_{x\to a^+} \phi(x) = \lim_{x\to b^-} \phi(x) = 0$). Since f' exists in (a,b), the integration by parts formula implies

$$\int_a^b f'(x)\phi(x)\,\mathrm{d}x = [f(x)\phi(x)]_a^b - \int_a^b f(x)\phi'(x)\mathrm{d}x = -\int_a^b f(x)\phi'(x)\mathrm{d}x,$$

since $\phi(a) = \phi(b) = 0$. Therefore, from Definition 7.5, we have that f' is a weak derivative of f, too. □

The converse of Theorem 7.6 is not true, i.e., there are functions that are not differentiable that have a weak derivative. This somewhat justifies the name: a weak derivative is a "weaker" notion of differentiation that the (classical) derivative.

We conclude this section with some more definitions.

Definition 2.8 *Let $(a,b) \subset \mathbb{R}$ open interval. We define the family of functions*

$$L^2((a,b)) := \Big\{ f : (a,b) \to \mathbb{R} : \int_a^b f^2(x)\,\mathrm{d}x < \infty \Big\},$$

i.e., the family of all square (Lebesgue-)integrable functions. Furthermore, we definite the family

$$H^1((a,b)) := \Big\{ f \in L^2((a,b)) : g \in L^2((a,b)) \text{ for } g \text{ weak derivative of } f \}.$$

Finally, we define

$$H_0^1((a,b)) := \Big\{ f \in H^1((a,b)) : f = 0 \text{ at the endpoints } a \text{ and } b \}.$$

It can be shown that all the above are vector spaces, but this is beyond the scope of these notes. The space $L^2((a,b))$ is an example of the so-called *Lebesgue spaces*, whereas $H^1((a,b))$ and $H_0^1((a,b))$ are examples of the so-called *Sobolev spaces*.

2.2.1 Problems

Problem 2.9 *Calculate the weak derivative of the function $f : (0,2) \to \mathbb{R}$ with*

$$f(x) = \begin{cases} x^2, & \text{for } 0 < x \le 1; \\ 2 - x, & \text{for } 1 < x < 2. \end{cases}$$

Does this function have a classical derivative everywhere in the interval $(0,2)$? Explain. Does this function belong to $L^2((0,2))$? Does it belong to $H^1((0,2))$? Does it belong to $H_0^1((0,2))$? Explain.

[5] Notice that since the function g only appears under the integral sign, it is strictly-speaking not unique, as changing the value of g at finite number of points it will not change the value of the integral!

[6] *Almost everywhere* here means that g is equal to f' at all points in (a,b) up to a set of Lebesgue measure zero.

2.3 The two-point boundary value problem in weak form

We consider again the two-point boundary value problem

$$\text{Find } u : (a,b) \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u(b) = 0. \tag{2.8}$$

where $f(x)$ is a known function.

The first step in defining a finite element method is to rewrite the two-point boundary value problem (7.1) in the so-called weak form, as follows.

Let $\mathcal{H} := H_0^1((a,b))$ defined in the previous section to be the family of functions v, that have a weak derivative and satisfy the Dirichlet boundary conditions $v(a) = 0 = v(b)$. We multiply the equation by a *test function* $v \in \mathcal{H}$, to get

$$-u''(x)v(x) = f(x)v(x),$$

and we integrate over the domain (a,b):

$$-\int_a^b u''(x)v(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x.$$

Now, if we perform an integration by parts to the integral on the left-hand side, we get

$$\int_a^b u'(x)v'(x)\mathrm{d}x - \big[u'(x)v(x)\big]_a^b = \int_a^b f(x)v(x)\mathrm{d}x,$$

for all $v \in \mathcal{H}$. Using the fact that $v(a) = 0 = v(b)$ for all $v \in \mathcal{H}$, we arrive to

$$\int_a^b u'(x)v'(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x,$$

for all $v \in \mathcal{H}$. Hence, the two-point boundary value problem can be transformed to the following problem in *weak form* (also known as *variational form*):

$$\text{Find } u \in \mathcal{H} \text{ s.t. } \int_a^b u'(x)v'(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x, \quad \text{for all } v \in \mathcal{H}. \tag{2.9}$$

Notice that if a function u is a solution to the problem (7.1), then it is also a solution to the problem (7.2). The converse, however, is *not* true, i.e., if a function u is a solution to the problem (7.2), then it is *not* necessarily a solution to the problem (7.1). Indeed, this can be verified by recalling that for $u \in \mathcal{H}$ to be a solution to (7.2), we only require that u has only a weak derivative, whereas for u to be a solution to (7.1), we have to require that u is twice differentiable. In this sense, Problem (7.2) is more general.

Indeed, suppose that a classical solution u exists. As u must be twice continuously differentiable by definition and $-u'' = f$, it follows that f must be continuous as well. So: for problem (7.1) to admit classical solution, we are only allowed to consider data functions f that are continuous. Now we may ask the same question, but for the weak formulation: for *which* functions f is the weak problem (7.2) well-posed? Functions f with (countable) discontinuities are integrable, and hence they certainly belongs to $L^2((a,b))$. It follows that the right-hand side of the weak problem (7.2) is well-defined in the sense that the integral is defined and finite (why?). It turns out that problem (7.2) is well-posed for *every* $f \in L^2((a,b))$! We have indeed the following general well-posedness result.

Lemma 2.10 *Suppose that the function $f \in L^2((a,b))$. Then Problem (7.2) is well-posed.*

Proof. The proof is beyond the scope of these notes. □

Another advantage of Problem (7.2) is that it is more *natural* in the sense that it matches the minimisation problem from which the boundary value problem (7.1) is derived, as we shall now see.

Lemma 2.11 *Consider the quadratic functional $F : \mathcal{H} \to \mathbb{R}$ given by*

$$F(v) = \frac{1}{2}\int_a^b v'(x)v'(x)\,\mathrm{d}x - \int_a^b f(x)v(x)\,\mathrm{d}x.$$

Then u is the solution of Problem (7.2) if and only if it minimises the functional F, that is [7]

$$u = argmin_{v\in\mathcal{H}} F(v).$$

[7] The notation *argmin* means 'the argument that minimizes', that is the minimizer as opposed to the minimum itself.

Proof. Suppose that $u \in \mathcal{H}$ is the solution to the weak problem (7.2). We show that $F(v) \geq F(u)$ for all $v \in \mathcal{H}$, hence u is the minimizer. Indeed, having using twice the fact that u satisfies (7.2), we have

$$\begin{aligned} F(v) - F(u) &= \frac{1}{2}\int_a^b v'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x - \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x + \int_a^b fu \,\mathrm{d}x \\ &= \frac{1}{2}\int_a^b v'v' \,\mathrm{d}x - \int_a^b u'v' \,\mathrm{d}x - \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x + \int_a^b u'u' \,\mathrm{d}x \\ &= \frac{1}{2}\int_a^b v'v' \,\mathrm{d}x - \int_a^b u'v' \,\mathrm{d}x + \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x \\ &= \frac{1}{2}\int_a^b (v-u)'(v-u)' \,\mathrm{d}x \geq 0. \end{aligned}$$

Now assume that u is the minimizer. Then in particular for every $v \in \mathcal{H}$ and $\lambda \in (0, 1]$ we have

$$\begin{aligned} 0 \leq F(u+\lambda v) - F(u) &= \frac{1}{2}\int_a^b (u+\lambda v)'(u+\lambda v)' \,\mathrm{d}x - \lambda \int_a^b fv \,\mathrm{d}x - \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x \\ &= \lambda\left(\int_a^b u'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x\right) + \frac{1}{2}\lambda^2 \int_a^b v'v' \,\mathrm{d}x. \end{aligned}$$

Dividing through by λ and letting $\lambda \to 0$ gives

$$\int_a^b u'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x \geq 0 \qquad \forall v \in \mathcal{H}.$$

Substituting v with $-v$ gives

$$\int_a^b u'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x \leq 0 \qquad \forall v \in \mathcal{H},$$

and we conclude that equality must hold, that is u solves (7.2). □

Problem

37. Write the following two-point boundary value problems in weak form

- find $u : (a, b) \to \mathbb{R}$, such that $-u''(x) + u'(x) = f(x)$ and $u(a) = 0$, $u(b) = 0$;
- find $u : (a, b) \to \mathbb{R}$, such that $-u''(x) = f(x)$ and $u(a) = 0$, $u'(b) = 0$.

Chapter 3

Divided Differences

3.1 Divided Differences

In numerical analysis of differential equations we often approximate/represent derivatives with the so-called *divided differences*.

3.1.1 Divided differences for first derivatives

Definition 3.1 *Let $f : [a, b] \to \mathbb{R}$, for $a < b$, be a bounded function, let a point x, and let some spacing $h > 0$. We define the* forward divided difference of f on x with spacing h *by*

$$\delta_{h,+} f(x) = \frac{f(x+h) - f(x)}{h}, \tag{3.1}$$

provided that $x + h \in [a, b]$. Similarly, we define the backward divided difference of f on x with spacing h *by*

$$\delta_{h,-} f(x) = \frac{f(x) - f(x-h)}{h}, \tag{3.2}$$

provided that $x - h \in [a, b]$.

We also define the central divided difference of f on x with spacing h *by*

$$\delta_h f(x) = \frac{f(x + \frac{h}{2}) - f(x - \frac{h}{2})}{h}, \tag{3.3}$$

provided that $x - \frac{h}{2}, x + \frac{h}{2} \in [a, b]$.

Notice that, if we let $h \to 0$, we obtain

$$\lim_{h\to 0^+} \delta_{h,+} f(x) = f'_+(x), \quad \lim_{h\to 0^+} \delta_{h,-} f(x) = f'_-(x), \text{ and } \lim_{h\to 0^+} \delta_h f(x) = f'(x),$$

where $f'_+(x)$, $f'_-(x)$ and $f'(x)$ denote the right derivative, the left derivative and the (normal) derivative of the function f at the point x, respectively (whenever, these derivatives are well-defined or, equivalently, whenever these limits exist). Hence, it is reasonable to consider $\delta_{h,+} f(x)$, $\delta_{h,-} f(x)$ and $\delta_h f(x)$ as approximations to $f'_+(x)$, $f'_-(x)$ and $f'(x)$, respectively. In these notes, the symbol "$\approx$" should be understood as "approximation of"; for instance, motivated by the above, we can write

$$\delta_{h,+} f(x) \approx f'_+(x), \quad \delta_{h,-} f(x) \approx f'_-(x), \text{ and } \delta_h f(x) \approx f'(x). \tag{3.4}$$

Of course, when the function f is differentiable, we have by definition that $f'_+(x) = f'_-(x) = f'(x)$ and, therefore in that case, the above 3 divided differences constitute 3 different ways of approximating $f'(x)$, viz.,

$$\delta_{h,+} f(x) \approx f'(x), \quad \delta_{h,-} f(x) \approx f'(x), \text{ and } \delta_h f(x) \approx f'(x).$$

Two natural questions arise:

- How good are the above approximations?

- Which of the 3 is the preferable approximation of $f'(x)$ (if any)?

The following result gives an answer to the first question and a partial answer to the second.

Lemma 3.2 *Let $f : [a, b] \to \mathbb{R}$, for $a < b$, be twice differentiable with continuous second derivative, and let a point x along with some spacing $h > 0$ as in Definition 3.1. Then, the following bound holds:*

$$|\delta_{h,+} f(x) - f'(x)| \leq \frac{h}{2} \max_{\xi \in [a,b]} |f''(\xi)|. \tag{3.5}$$

Similarly, we have

$$|\delta_{h,-} f(x) - f'(x)| \leq \frac{h}{2} \max_{\xi \in [a,b]} |f''(\xi)|. \tag{3.6}$$

Assume further that f is three times differentiable with continuous third derivative; then, we have

$$|\delta_h f(x) - f'(x)| \leq \frac{h^2}{24} \max_{\xi \in [a,b]} |f'''(\xi)|. \tag{3.7}$$

Proof. Using Taylor's theorem[1], we have

$$\delta_{h,+} f(x) = \frac{f(x+h) - f(x)}{h} = \frac{f(x) + hf'(x) + \frac{h^2}{2} f''(\xi) - f(x)}{h} = f'(x) + \frac{h}{2} f''(\xi),$$

for some $\xi \in (x, x+h)$, from which the first bound follows, by taking the maximum over $[x, x+h]$; this maximum is well defined as f'' is continuous on the compact set $[x, x+h]$ (this set is a closed subset of $\mathbb{R}$ and, therefore, it is compact). Similarly for the backward difference, we use again Taylor's formula with h replaced by $-h$ this time, to obtain

$$\delta_{h,-} f(x) = \frac{f(x) - f(x-h)}{h} = \frac{f(x) - f(x) + hf'(x) - \frac{h^2}{2} f''(\zeta)}{h} = f'(x) - \frac{h}{2} f''(\zeta),$$

for some $\zeta \in (x-h, x)$; arguing as before, we conclude the second bound. Finally, for the last bound we use Taylor's Theorem with $n = 2$ and with h replaced by $\frac{h}{2}$ and $-\frac{h}{2}$, respectively, to deduce

$$\begin{aligned}
\delta_h f(x) &= \frac{f(x + \frac{h}{2}) - f(x - \frac{h}{2})}{h} \\
&= \frac{1}{h}\big(f(x) + \frac{h}{2} f'(x) + \frac{h^2}{8} f''(x) + \frac{h^3}{48} f'''(\xi_1) - f(x) + \frac{h}{2} f'(x) - \frac{h^2}{8} f''(x) + \frac{h^3}{48} f'''(\zeta_1)\big) \\
&= f'(x) + \frac{h^2}{48}\big(f'''(\xi_1) + f'''(\zeta_1)\big),
\end{aligned}$$

for some $\xi_1 \in (x, x + \frac{h}{2})$ and some $\zeta_1 \in (x - \frac{h}{2}, x)$, and the result follows as before. □

Hence, we see from the above bounds that the error of the approximations (3.4) can be quantified by a constant times powers of h. In particular, we observe that the error of approximation of the derivative of f at the point x by $\delta_{h,+}$ or by $\delta_{h,-}$ decays linearly with respect to h, whereas the error of approximation by δ_h decays quadratically (i.e., like h^2). These observations can be used as an answer to the first question above.

So as we take smaller and smaller h, we should expect the central difference to provide us with a more accurate approximation that either the forward or the backward difference. This could serve as a partial answer to the second question above. However, this is not the end of the story, as both forward and backward differences can be very useful in numerical analysis (recall Euler's methods from elementary Numerical Analysis!). For instance, if we want to approximate $f'(a)$ or $f'(b)$ (i.e., the derivative of f at one of the endpoints of the interval), we are confined in using some form of forward or backward differences

[1] **Taylor's Theorem.** Let $a < b$, n positive integer, $x \in [a, b]$ and $f : [a, b] \to \mathbb{R}$ continuous function such that the derivatives up to order $n + 1$ (inclusive) exist at every point in $[a, b]$. Then, for every $x \in [a, b]$, there exists ξ between x and $x + h$, for some $h \in \mathbb{R}$, with $x + h \in [a, b]$, such that

$$f(x+h) = \sum_{k=0}^{n} \frac{h^k}{k!} f^{(k)}(x) + \frac{h^{n+1}}{(n+1)!} f^{(n+1)}(\xi). \tag{3.8}$$

respectively, as we may not be in possession of function values outside the interval $[a,b]$. Furthermore, and perhaps more importantly, in many cases the problems we are seeking to use approximations of derivatives have preferred directions (say we want to approximate solutions to differential equations modelling a fast flowing river or the air flow around an airborne plane); in such cases it can be of crucial importance to use ideas based on backward or forward differences for reasons of "stability" of the numerical methods, even if we might lose in terms of accuracy. We shall see more of this in the discussion on numerical methods for hyperbolic PDEs.

The decay with respect to the *discretisation parameter* h will be of central importance in what follows; therefore, we shall give some formal definitions and we shall introduce some shorthand notation related to this.

Definition 3.3 *Let $g : [a,b] \to \mathbb{R}$, for $a < b$, be a bounded function and consider some approximation $g_h(x)$ of g at the point $x \in [a,b]$ with discretisation parameter $h > 0$. We say that $g_h(x)$ is* convergent *if*

$$\lim_{h \to 0^+} \big(g_h(x) - g(x)\big) = 0.$$

Now, assume that $g_h(x)$ is convergent. We say that $g_h(x)$ converges to $g(x)$ with order p *(or* is p-th order convergent to $g(x)$*) with respect to h, for some $p > 0$, if*

$$|g_h(x) - g(x)| = \mathcal{O}(h^p),$$

whereby the symbol $\mathcal{O}(\cdot)$ is understood as follows

$$w(h) = \mathcal{O}(h^p) \Leftrightarrow \lim_{h \to 0^+} \frac{w(h)}{h^p} \in \mathbb{R}, \tag{3.9}$$

for some function w defined in a neighbourhood of 0.

In light of this, and in view of Lemmas 3.2 and 3.5, we say that the forward and backward differences $\delta_{h,+}$ and $\delta_{h,-}$ are first order convergent to $f'(x)$ with respect to h, whereas the central difference δ_h is second order convergent to $f'(x)$, for all $x \in [a,b]$.

It sometimes can be very useful to make use of forward or backward divided differences of order higher than one (as it is the case for δ_+ and δ_-). To this end, we define the following divided differences:

$$\begin{aligned} \delta_{h,+,2} f(x) &:= (\delta_{h,+} - \frac{h}{2}\delta_{h,+}^2) f(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}\delta_{h,+}\big(\frac{f(x+h) - f(x)}{h}\big) \\ &= \frac{1}{2h}\big(-f(x+2h) + 4f(x+h) - 3f(x)\big), \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} \delta_{h,-,2} f(x) &:= -(\delta_{h,-} + \frac{h}{2}\delta_{h,-}^2) f(x) = -\frac{f(x) - f(x-h)}{h} - \frac{h}{2}\delta_{h,-}\big(\frac{f(x) - f(x-h)}{h}\big) \\ &= \frac{1}{2h}\big(-f(x-2h) + 4f(x-h) - 3f(x)\big). \end{aligned} \tag{3.11}$$

The following result describes the accuracy of these divided differences.

Lemma 3.4 *Let $f : [a,b] \to \mathbb{R}$, for $a < b$, be three times differentiable with continuous third derivative, and let a point x along with some spacing $h > 0$. Then, the following bound holds:*

$$|\delta_{h,+,2} f(x) - f'(x)| \leq h^2 \max_{\xi \in [a,b]} |f'''(\xi)|, \tag{3.12}$$

provided that $x + 2h \in [a,b]$. Similarly, we have

$$|\delta_{h,-,2} f(x) - f'(x)| \leq h^2 \max_{\xi \in [a,b]} |f'''(\xi)|, \tag{3.13}$$

provided that $x - 2h \in [a,b]$.

Proof. The proof of the second equalities in (3.10) and (3.11), as well as the bounds (3.12) and (3.13) are left as an exercise. □

Notice that the constants in the bounds (3.12) and (3.13) are bigger than the constant in (3.7). Notice also, that for (3.10) we require 3 values of the function f at the points $x, x+h, x+2h$ (and similarly for and (3.11)), as opposed to the only 2 values at $x + \frac{h}{2}, x - \frac{h}{2}$ required by (3.3) for the same order of accuracy.

3.1.2 Divided differences for higher derivatives

We can use these developments to define also divided differences to approximate higher derivatives. One of the most important divided differences, in what follows, is the *second central divided difference of f at a point x with spacing h*, defined as

$$\begin{aligned}\delta_h^2 f(x) &:= \delta_h(\delta_h f(x)) = \delta_h\Big(\frac{f(x+\frac{h}{2})-f(x-\frac{h}{2})}{h}\Big)\\ &= \frac{\frac{f(x+\frac{h}{2}+\frac{h}{2})-f(x-\frac{h}{2}+\frac{h}{2})}{h} - \frac{f(x+\frac{h}{2}-\frac{h}{2})-f(x-\frac{h}{2}-\frac{h}{2})}{h}}{h}\\ &= \frac{f(x+h)-2f(x)+f(x-h)}{h^2},\end{aligned} \tag{3.14}$$

which is a (popular) approximation to $f''(x)$ (where we have adopted the notational convention that powers of divided difference operators are understood as composition). In particular, we have the following result.

Lemma 3.5 *Let $f:[a,b]\to\mathbb{R}$, for $a<b$, be 4 times differentiable with continuous 4th derivative, and let a point x along with some spacing $h>0$ such that $x-h, x+h\in[a,b]$. Then, the following bound holds:*

$$|\delta_h^2 f(x) - f''(x)| \le \frac{h^2}{12}\max_{\xi\in[a,b]}|f^{(4)}(\xi)|. \tag{3.15}$$

Proof. Using Taylor's formula for $n=3$ with h, and with h replaced by $-h$, respectively, we have

$$\delta_h^2 f(x) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} = \cdots = f''(x) + \frac{h^2}{24}\big(f^{(4)}(\xi)+f^{(4)}(\zeta)\big), \tag{3.16}$$

for some $\xi\in(x,x+h)$ and for some $\zeta\in(x-h,x)$. The result follows by taking the maximum as before. □

We observe that the approximation error of $\delta_h^2 f(x)$ to $f''(x)$ decays quadratically with respect to h (i.e., like h^2) as h goes to zero and, therefore, we say that the second central difference δ_h^2 is second order convergent to $f''(x)$, for all $x\in[a,b]$.

Next, we consider one-sided divided differences for the approximation of second derivatives – these could be proven useful in the design of computational methods for second order PDEs near the boundaries of the computational domains. For instance, suppose we want to approximate $f''(a)$ and $f''(b)$ with second order divided differences that do not require any function values outside the interval $[a,b]$. An answer can be given by the following result.

Lemma 3.6 *Let $f:[a,b]\to\mathbb{R}$, for $a<b$, be 4 times differentiable with continuous 4th derivative, and let a point x along with some spacing $h>0$. Consider the divided difference*

$$\delta_{h,right} f(x) := (\delta_{h,+}^2 - h\delta_{h,+}^3)f(x) = \frac{1}{h^2}\big(2f(x)-5f(x+h)+4f(x+2h)-f(x+3h)\big), \tag{3.17}$$

and

$$\delta_{h,left} f(x) := (\delta_{h,-}^2 + h\delta_{h,-}^3)f(x) = \frac{1}{h^2}\big(2f(x)-5f(x-h)+4f(x-2h)-f(x-3h)\big), \tag{3.18}$$

with $x+3h\in[a,b]$ or with $x-3h\in[a,b]$, respectively. Then, the following bounds hold:

$$|\delta_{h,right} f(x) - f''(x)| \le \frac{22}{24}h^2\max_{\xi\in[a,b]}|f^{(4)}(\xi)|, \tag{3.19}$$

and

$$|\delta_{h,left} f(x) - f''(x)| \le \frac{22}{24}h^2\max_{\xi\in[a,b]}|f^{(4)}(\xi)|. \tag{3.20}$$

Proof. The proof of the second equalities in (3.17) and (3.18), as well as the bounds (3.19) and (3.20) are left as an exercise. □

As in the case of first derivatives above, notice that the constants in the bounds (3.19) and (3.20) are bigger than the constant in (3.15). Notice also, that for (3.17) we require 4 values of the function f at the points $x, x+h, x+2h, x+3h$ (and similarly for and (3.18)), as opposed to the only 3 values at $x, x+h, x-h$ required by (3.14) for the same order of accuracy.

3.1.3 Comments and further reading

There is a vast amount of literature devoted to the calculus of divided differences. An relatively short but inspiring exposition of the basic concepts of the calculus of divided differences applied to the approximation of derivatives is given in Chapter 7 of the book by Iserles.

Problems

11. Prove the second equalities in (3.10), (3.11) and Lemma 3.4.

12. Prove the second equalities in (3.17), (3.18) and Lemma 3.6.

3.2 Difference methods for two-point boundary value problems

We shall be concerned with constructing a difference method for the the numerical approximation of the solution of the following boundary value problem:

$$\text{Find } u : [a,b] \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u(b) = 0. \tag{3.21}$$

where $f(x)$ is a known function. Obviously, by integrating the relationship $-u''(x) = f(x)$ twice and using the boundary conditions $u(a) = 0,\ u(b) = 0$ we can find a solution to this problem; but in this instance we suppose that we do not know how to do this, and we seek an approximation to the solution using the following method.

We consider equally distributed points $x_0 < x_1 < \cdots < x_{N+1}$, at distance h between them, such that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N+1} = b.$$

We approximate $u''(x)$ by the second central divided difference at the points x_i, $i = 1, \ldots, n$, i.e.,

$$u''(x_i) \approx \frac{u(x_i + h) - 2u(x_i) + u(x_i - h)}{h^2} = \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}.$$

Motivated by this formal approximation, our aim is to find approximations u_i of $u(x_i)$ (that is, the value of the exact solution u at the point x_i), for all $i = 1, \ldots, N$, such that

$$\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = -f(x_i), \quad i = 1, \ldots, N. \tag{3.22}$$

Notice, that the value of u at the points $x_0 = a$ and $x_{N+1} = b$ is known from the boundary conditions, i.e., we set $u_0 = u(a) = 0$ and $u_{N+1} = u(b) = 0$!

To find the approximations u_i's, $i = 1, \ldots, N$, we solve the linear system of equations (3.22). This is a linear system of N equations with N unknowns; in matrix form it can be written as

$$\underbrace{\begin{pmatrix} -2 & 1 & 0 & 0 & \ldots & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots \\ 0 & \ldots & 0 & 1 & -2 & 1 \\ 0 & \ldots & 0 & 0 & 1 & -2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \end{pmatrix}}_{U} = \underbrace{-h^2 \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \end{pmatrix}}_{rhs}, \tag{3.23}$$

after multiplication by h^2. Solving this linear system we can obtain the approximations u_i to $u(x_i)$, for all $i - 1, \ldots, N$.

Next, we consider the case of non-homogeneous Dirichlet boundary conditions. In particular, we are interested in finding an approximation to the solution of the problem:

$$\text{Find } u : [a,b] \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = c,\ u(b) = d.$$

where $f(x)$ is a known function and $c, d \in \mathbb{R}$ known boundary values. Motivated by the discussion above, we can construct a finite difference method of the form (3.22), with the difference that now we need to be careful with the equations at the nodes $i = 1$ and $i = N$. In particular, for $i = 1$, we now have

$$\frac{u_2 - 2u_1}{h^2} = -f(x_1) - \frac{c}{h^2},$$

and for $i = N$, we get

$$\frac{-2u_N + u_{N-1}}{h^2} = -f(x_N) - \frac{d}{h^2}.$$

Therefore, it is not too hard to see that, after a multiplication by h^2, we arrive to the linear system of N equations with N unknowns:

$$AU = -h^2(f(x_1) + \frac{c}{h^2}, f(x_2), \ldots, f(x_{N-1}), f(x_N) + \frac{d}{h^2})^T,$$

recalling the notation above.

Finally, we consider the somewhat more involved case of Neumann boundary conditions. In particular, we are interested in finding an approximation to the solution of the problem:

$$\text{Find } u : [a, b] \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u'(b) = 0.$$

where $f(x)$ is a known function. To construct a finite difference methods for this problem we work as follows.

The boundary condition at the left endpoint can be treated as above, i.e., by setting $u_0 = 0$. The values u_i at the nodes x_i, $i = 1, \ldots, N-1$ are set to satisfy equations of the form (3.22). In this case also u_{N+1} is unknown, as it is not prescribed by the boundary conditions; hence, at the point x_N, we require

$$\frac{u_{N+1} - 2u_N + u_{N-1}}{h^2} = -f(x_N).$$

Since we cannot represent exactly the Neumann boundary condition $u'(b) = 0$ due to the presence of the derivative, we shall construct a divided difference approximation of u' at the endpoint b. For reasons that will be discussed in the error analysis below, we shall choose the second order central difference with spacing $2h$ (see (3.3) to approximate the first derivative u' at the endpoint b. This means that we need an additional approximation of the exact solution u at the point $b+h$, say u_{N+2}, which is of course unknown. Then, from the second order central difference with spacing $2h$ approximating $u'(b)$, we get the equation

$$\frac{u_{N+2} - u_N}{2h} = 0, \quad \text{which implies} \quad u_{N+2} = u_N. \tag{3.24}$$

In this case, the value u_{N+1} (the approximation of u at the endpoint $b = x_{N+1}$) is no longer given by the boundary conditions, and thus it is also an unknown. Since we have also u_{N+2} at our disposal (due to the implementation of the Neumann boundary conditions described above), we can request that the numerical method also holds at the point $b = x_{N+1}$, i.e., we require

$$\frac{u_{N+2} - 2u_{N+1} + u_N}{h^2} = -f(b),$$

or, using (3.24)

$$-2u_{N+1} + 2u_N = -h^2 f(b) \quad \text{or} \quad -u_{N+1} + u_N = -\frac{h^2}{2} f(b).$$

Therefore, we arrive to the following linear system of $N+1$ equations with $N+1$ unknowns:

$$\begin{pmatrix} -2 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots & \\ 0 & \ldots & 0 & 1 & -2 & 1 & 0 \\ 0 & \ldots & 0 & 0 & 1 & -2 & 1 \\ 0 & \ldots & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{N-1} \\ u_N \\ u_{N+1} \end{pmatrix} = -h^2 \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ \vdots \\ f(x_{N-1}) \\ f(x_N) \\ \frac{f(b)}{2} \end{pmatrix}. \tag{3.25}$$

Problem

13. Construct a difference method for the the numerical approximation of the solution of the following boundary value problem:

$$\text{Find } u : [a, b] \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u(b) = 0.$$

using a non-uniform grid of the form

$$a = x_0,\ x_1 = x_0 + h_1,\ x_2 = x_1 + h_2,\ \ldots,\ x_N = x_{N-1} + h_N,\ x_{N+1} = b,$$

i.e., $h_i = x_i - x_{i-1}$ for $i = 1, \ldots N+1$. Write the resulting linear system. [*Hint: Start by constructing the 2nd divided difference on a non-uniform grid.*]

3.2.1 Error analysis

We have studied a finite difference method for approximating the solutions to the two-point boundary-value problem with various boundary conditions of Dirichlet and Neumann type. We now turn our attention in estimating the *quality* of these approximations. In particular, we are interested in analysing the error of the approximation. To this end, we define the *truncation error* of the finite difference scheme by

$$T(x) := f(x) + \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \tag{3.26}$$

In other words, the truncation error is defined by substituting the exact solution into the finite difference method, thereby representing how much the finite difference method fails to imitate the boundary-value problem. Assuming that the exact solution u is 4 times differentiable, with continuous 4th derivative, and using Taylor's Theorem, we obtain

$$T(x) = f(x) + \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = f(x) + u''(x) - \frac{h^2}{24}\big(u^{(4)}(\xi) + u^{(4)}(\zeta)\big),$$

for some $\xi \in (x, x+h)$ and for some $\zeta \in (x-h, x)$. Using now the differential equation, we arrive to

$$T(x) = -\frac{h^2}{24}\big(u^{(4)}(\xi) + u^{(4)}(\zeta)\big),$$

and, therefore, we deduce the following bound

$$|T(x)| \leq \frac{h^2}{12} \max_{a \leq \xi \leq b} |u^{(4)}(\xi)|. \tag{3.27}$$

The bound (3.27) is potentially good news: it implies that $T(x) \to 0$ as $h \to 0$. This means that the numerical scheme "approximates" the differential equation better and better as $h \to 0$. However, the bound (3.27) is *not* giving any information on how well the values $u(x_i)$ are approximated by the u_i's, for $i = 1, ..., N$ (or, for $i = 1, ..., N+1$ for the case of the Neumann boundary conditions). To derive bounds for the error of the approximation itself, we shall make use of the following discrete maximum principle.

Theorem 3.7 (Discrete Maximum Principle) *Consider the finite difference approximation to the two-point boundary-value problem (3.21) above and let A be the $N+2 \times N+2$ version of the matrix defined in (3.23). Let $v \in \mathbb{R}^{N+2}$ with $v = (v_0, v_1, \ldots, v_N, v_{N+1})^T$. If $(Av)_i \geq 0$ for all $i = 1, 2, \ldots, N$; then we have*

$$\max_{1 \leq i \leq N} v_i \leq \max\{v_0, v_{N+1}, 0\}, \tag{3.28}$$

i.e., v cannot attain a non-negative maximum at an interior point.

Proof. We shall prove the result by contradiction. Suppose that (3.28) is not true, i.e., there exists $1 \leq n \leq N$ such that

$$v_n = \max_{1 \leq i \leq N} v_i \quad \text{and} \quad v_n > v_0 \text{ and } v_n > v_{N+1}, \tag{3.29}$$

i.e., the maximum is attained at an interior point. Then, from the hypothesis of the theorem and the structure of the matrix A, we have

$$0 \leq (Av)_n = v_{n-1} - 2v_n + v_{n+1} \quad \text{or} \quad v_n \leq \frac{1}{2}(v_{n-1} + v_{n+1}).$$

But $v_n \geq v_{n-1}$ and $v_n \geq v_{n+1}$, since v_n is the maximum value; therefore, the previous inequality yields

$$v_n \leq \frac{1}{2}(v_{n-1} + v_{n+1}) \leq v_n \quad \text{or} \quad v_n = \frac{1}{2}(v_{n-1} + v_{n+1}),$$

which readily implies that $v_n = v_{n-1} = v_{n+1}$. Thus also the neighbours of v_n attain the maximum value. Arguing in completely analogous manner, we deduce that the neighbours of the neighbours also attain the maximum value, and so on. So we conclude that $v_i = v_n$ for all $i = 0, 1, \ldots, N, N+1$. Hence (3.29) is false and we arrive to a contradiction. □

We are now ready to prove a bound for the error of the finite difference scheme for the two-point boundary-value problem.

Theorem 3.8 *The finite difference method (3.23), employed to approximate the solution u of the two-point boundary-value problem (3.21), satisfies the error bound:*

$$|u(x_i) - u_i| \leq \frac{h^2}{48}(b-a)^2 \max_{a\leq\xi\leq b} |u^{(4)}(\xi)|. \tag{3.30}$$

Proof. We define

$$T := \max_{a\leq\xi\leq b} |T(\xi)|,$$

the maximum of the truncation error $T(x)$, introduced above, over the interval $[a, b]$. From the definition of the truncation error (3.26) and from the definition of the finite difference method (3.22) we get, respectively,

$$T(x_i) = f(x_i) + \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = -\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2},$$

for $i = 1, 2, \ldots, N$; this can be rearranged to

$$h^2 T(x_i) = (u(x_{i+1}) - u_{i+1}) - 2(u(x_i) - u_i) + (u(x_{i-1}) - u_{i-1}) = e_{i+1} - 2e_i + e_{i+1}, \tag{3.31}$$

upon defining the *nodal errors* $e_i := u(x_i) - u_i$, for $i = 0, 1, \ldots, N, N+1$. We can immediately see that $e_0 = 0 = e_{N+1}$ from to construction of the finite difference method.

We define the auxiliary function $\phi(x) : [a, b] \to \mathbb{R}$, with

$$\phi(x) = \frac{T}{2}\big(x - \gamma\big)^2, \quad \text{where} \quad \gamma = \frac{1}{2}(a+b),$$

We also define a vector $v \in \mathbb{R}^{N+2}$ with $v := (v_0, v_1, \ldots, v_N, v_{N+1})^T$ by setting

$$v_i := e_i + \phi(x_i) \quad \text{for} \quad i = 0, 1, \ldots, N, N+1.$$

We check if the vector v satisfies the hypothesis of Theorem 3.7. Using (3.31), we deduce

$$\begin{aligned}(Av)_i &= v_{i-1} - 2v_i + v_{i+1} = h^2 T(x_i) + \phi(x_{i-1}) - 2\phi(x_i) + \phi(x_{i+1}) \\ &= h^2 T(x_i) + \frac{T}{2}\big(x_i - h - \gamma\big)^2 - T\big(x_i - \gamma\big)^2 + \frac{T}{2}\big(x_i + h - \gamma\big)^2 = h^2(T(x_i) + T) \geq 0,\end{aligned}$$

since T is by definition the maximum (in absolute value) of the truncation error. Hence, from Theorem 3.7, we conclude that

$$v_i \leq \max\{v_0, v_{N+1}, 0\}, \quad \text{for all} \quad i = 1, 2, \ldots, N. \tag{3.32}$$

But

$$v_0 = \phi(a) = \frac{T}{8}(b-a)^2 \quad \text{and} \quad v_{N+1} = \phi(b) = \frac{T}{8}(b-a)^2,$$

since $e_0 = e_{N+1} = 0$, and thus (3.32) yields

$$v_i \leq \frac{T}{8}(b-a)^2,$$

or, using the definition of T and (3.27),

$$u(x_i) - u_i = e_i \leq \frac{T}{8}(b-a)^2 - \phi(x_i) \leq \frac{T}{4}(b-a)^2 \leq \frac{h^2}{48}(b-a)^2 \max_{a\leq\xi\leq b} |u^{(4)}(\xi)|. \tag{3.33}$$

The above is an one-sided bound. Setting

$$\hat{v}_i := -e_i + \phi(x_i) \quad \text{for} \quad i = 0, 1, \ldots, N, N+1,$$

and following the same steps as above, we can deduce also that

$$u_i - u(x_i) \leq \frac{h^2}{48}(b-a)^2 \max_{a\leq\xi\leq b} |u^{(4)}(\xi)|. \tag{3.34}$$

Combining (3.33) and (3.34), we can conclude that (3.30) holds. □

Remark 3.9 *The bound (3.33) can be further improved to $u(x_i) - u_i \leq \frac{T}{8}(b-a)^2$, by noticing that $\phi(x_i) \geq 0$.*

Problem

14. Consider the difference method from Problem 13. Define a suitable truncation error, calculate it, and give a condition for this error to converge to zero.

Chapter 4

Finite Difference Methods for Parabolic Problems

We start our discussion on computational methods for parabolic PDEs by considering the problem of approximating the solution to the *heat equation* (also known as the *diffusion equation*), together with corresponding initial and boundary conditions.

4.1 Explicit Euler method

We begin by considering the initial/boundary value problem for the heat equation with homogeneous Dirichlet boundary conditions. More specifically, consider the space-domain $[a, b] \subset \mathbb{R}$, and the time-domain $[0, T_f]$, for some *final time* $T_f > 0$. We want to find an approximation to the solution of the problem: find a function $u : [0, T_f] \times [a, b] \to \mathbb{R}$ with continuous second derivatives, such that

$$u_t(t, x) = u_{xx}(t, x) \text{ for all } t \in [0, T_f] \text{ and } x \in [a, b], \tag{4.1}$$

subject to the initial condition

$$u(0, x) = u_0(x), \text{ for all } x \in [a, b], \tag{4.2}$$

for some known continuous function $u_0 : [a, b] \to \mathbb{R}$, and subject to homogeneous boundary conditions

$$u(t, a) = u(t, b) = 0, \text{ for all } t \in [0, T_f]. \tag{4.3}$$

Equation (4.1) is the archetype parabolic PDE and it is possible to show that this problem has a unique solution.

We are concerned with the development of *finite difference methods* for the problem (4.1), (4.2), (4.3). Using the ideas presented in the previous chapters, we can approximate the derivatives in (4.1) using divided differences. To this end, we construct a grid as follows: we consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N_x+1}$, at distance h between them, such that

$$a = x_0, \ x_1 = x_0 + h, \ x_2 = x_1 + h, \ \ldots, \ x_{N_x} = x_{N_x-1} + h, \ x_{N_x+1} = b,$$

in the space-direction, and an equally distributed subdivision $t_0 < t_1 < \cdots < t_{N_t}$, at distance τ between them, such that

$$0 = t_0, \ t_1 = t_0 + \tau, \ t_2 = t_1 + \tau, \ \ldots, \ t_{N_t-2} = t_{N_t-1} + \tau, \ t_{N_t} = T_f,$$

in the time-direction (see Figure 4.1); hence, we have $h = (b - a)/(N_x + 1)$ and $\tau = T_f/N_t$.

We formally make the following approximations, using forward difference for the time-derivative and central second difference for the space-derivative

$$u_t(t, x) \approx \delta^t_{\tau,+} u(t, x) = \frac{u(t + \tau, x) - u(t, x)}{\tau},$$

and

$$u_{xx}(t, x) \approx (\delta^x_h)^2 u(t, x) = \frac{u(t, x + h) - 2u(t, x) + u(t, x - h)}{h^2},$$

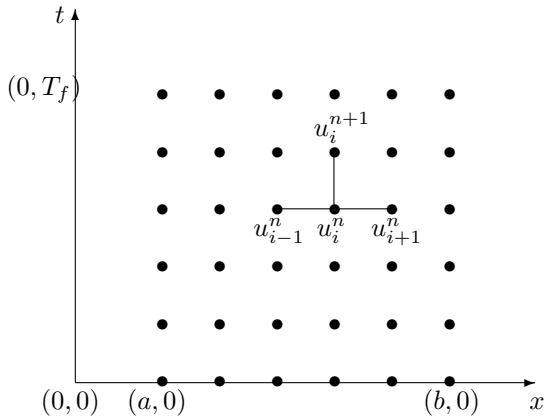


Figure 4.1: The grid for $N_x = 4$ and $N_t = 5$.

where we have adopted the notational convention that the superscript t, denotes that the divided difference operator acts on the time variable t and correspondingly for x, giving

$$\frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} - \frac{u(t_n, x_{i+1}) - 2u(t_n, x_i) + u(t_n, x_{i-1})}{h^2} \approx u_t(t_n, x_i) - u_{xx}(t_n, x_i) = 0, \tag{4.4}$$

using the equation (4.1).

Motivated by this formal reasoning, our aim is to find approximations u_i^n of the function values $u(t_n, x_i)$, we consider the following system of equations:

$$\frac{u_i^{n+1} - u_i^n}{\tau} - \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} = 0,$$

for $n = 0, \ldots, N_t - 1$ and $i = 1, \ldots, N_x$ which becomes, after multiplication by τ and rearrangement

$$u_i^{n+1} = \mu u_{i+1}^n + (1 - 2\mu)u_i^n + \mu u_{i-1}^n, \quad \text{for } n = 0, \ldots, N_t - 1,\ i = 1, \ldots, N_x, \tag{4.5}$$

where we have used the notation $\mu := \tau/h^2$ and we shall refer to this quantity as the *Courant number*. (Notice that here and in the sequel we shall denote the index related to time discretisation as a superscript – i.e., n in u_i^n is *not* an exponent!– and the indices related to space discretisation as subscripts.) For (4.5) to make complete sense, we need to clarify the values on the grid points residing on the boundaries, i.e., which values we should choose for u_i^n when $n = 0$ and $i = 1, \ldots, N_x$ (i.e., the values on the grid points at initial time $\{(0, x) : a \leq x \leq b\}$), and which values we should choose when $i = 0$ or $N_x + 1$ and $n = 0, \ldots, N_t$ (i.e., the values on the Dirichlet boundary grid points $\{(t, a) : 0 \leq t \leq T\}$ and $\{(t, b) : 0 \leq t \leq T\}$). The values of the solution u at initial time can be found from the initial condition (4.2), giving

$$u_i^0 = u_0(x_i) \quad i = 1, \ldots, N_x, \tag{4.6}$$

and the values for the nodes residing on the Dirichlet boundary can be given using the boundary conditions (4.3)

$$u_0^n = 0 = u_{N_x+1}^n \quad n = 0, \ldots, N_t. \tag{4.7}$$

The finite difference method defined by the equations (4.5), (4.6) and (4.7), will be referred to as the *explicit Euler method* (also known as in the literature as the *explicit scheme*, or as the *explicit method*).

The reason for the name can be traced at the following observation: given the initial and boundary values (4.6) and (4.7), respectively, it is straightforward to find the values at each time level n *explicitly*! Indeed, setting $n = 0$ in (4.5), we get

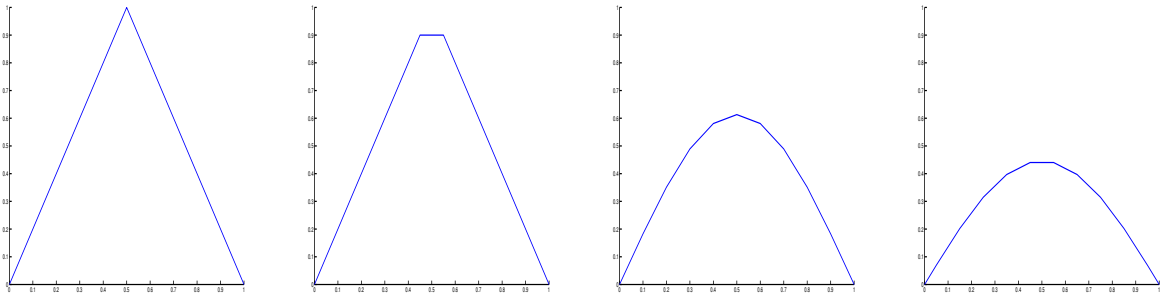
$$u_i^1 = \mu u_{i+1}^0 + (1 - 2\mu)u_i^0 + \mu u_{i-1}^0, \quad \text{for } i = 1, \ldots, N_x,$$

i.e., each u_i^1 (that is, the approximation of the solution at the point x_i in the next time level t_1) can be found as a linear combination of the three *known* values u_{i-1}^0, u_i^0 and u_{i+1}^0 at the previous time level; the same applies when setting $n = 1$ in (4.5), once all the values u_i^1 have been calculated, and so on.

Example 4.1 *We want to approximate the solution of the initial/boundary value problem above using the explicit Euler method when* $a = 0$, $b = 1$ *and*

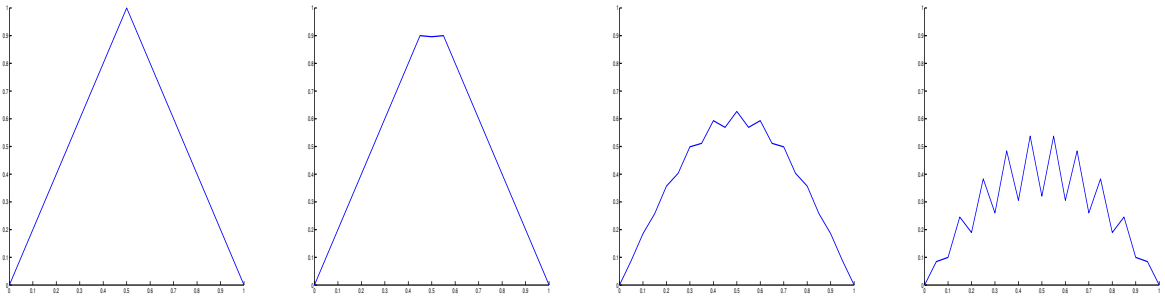
$$u_0(x) = \begin{cases} 2x, & 0 \leq x \leq 1/2; \\ 2 - 2x, & 1/2 < x \leq 1. \end{cases}$$

After implementing the explicit Euler algorithm above in the computer, the approximate solutions can be seen in Figures (4.2) and (4.3)



(a) Approximation at $n = 0$ (b) Approximation at $n = 1$ (c) Approximation at $n = 25$ (d) Approximation at $n = 50$

Figure 4.2: Approximation using explicit Euler method for $N_x = 19$ and $N_t = 800$, resulting to $\mu = 0.5$.



(a) Approximation at $n = 0$ (b) Approximation at $n = 1$ (c) Approximation at $n = 25$ (d) Approximation at $n = 50$

Figure 4.3: Approximation using explicit Euler method for $N_x = 19$ and $N_t = 770$, resulting to $\mu = 0.52$.

We observe that the solution looks reasonable in the first case, but not in the second, where it seems that the approximation is not close to the exact solution. Indeed the solution seems to oscillate in the second case, indicating that the oscillatory behaviour will get worse and worse in the next time-steps. In the sequel, we shall investigate the source of this erratic behaviour of the approximations, leading to "instabilities".

The previous example illustrates that we should always be *very careful* when designing computational methods to approximate solution to PDEs, as it can lead to catastrophically inaccurate results! Therefore, it is important to investigate further what are the causes of such oscillatory behaviour by analysing the method; this will be the content of the next two sections.

Problem

15. Construct an explicit scheme for the approximation to the solution of the problem: find a function $u : [0, T] \times [a, b] \to \mathbb{R}$, such that

$$u_t(t, x) = u_{xx}(t, x) \text{ for all } t \in [0, T] \text{ and } x \in [a, b],$$

subject to the initial condition

$$u(0, x) = u_0(x), \text{ for all } x \in [a, b],$$

for some known continuous function $u_0 : [a, b] \to \mathbb{R}$, and subject to the boundary conditions

$$u(t, a) = 0, \quad \text{and} \quad u_x(t, b) = 0 \quad \text{for all } t \in [0, T].$$

4.1.1 Error analysis of the explicit Euler method

To analyse the error of approximation, we assume that the exact solution u is twice continuously differentiable with respect to the time variables and four times continuously differentiable with respect to the space variable. We define the truncation error for this method

$$T_i^n := \frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} - \frac{u(t_n, x_{i+1}) - 2u(t_n, x_i) + u(t_n, x_{i-1})}{h^2}, \tag{4.8}$$

for $n = 0, \ldots, N_t - 1$, $i = 1, \ldots, N_x$; i.e., the truncation error is given if we substitute the exact solution into the numerical scheme. The following lemma describes how well the truncation error approximates the original problem.

Lemma 4.2 *For the explicit Euler method defined above, we have*

$$|T_i^n| \leq \frac{\tau}{2} M_{tt} + \frac{h^2}{12} M_{xxxx}, \tag{4.9}$$

for all $n = 0, \ldots, N_t - 1$, $i = 1, \ldots, N_x$, where

$$M_{tt} := \max |u_{tt}(t, x)|, \quad \textit{and} \quad M_{xxxx} := \max |u_{xxxx}(t, x)|,$$

and the maxima are taken over all $(t, x) \in [0, T_f] \times [a, b]$.

Proof. Using Taylor's Theorem, we have

$$\frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} = \frac{u(t_n + \tau, x_i) - u(t_n, x_i)}{\tau} = u_t(t_n, x_i) + \frac{\tau}{2} u_{tt}(\rho_n, x_i),$$

for $\rho_n \in (t_n, t_{n+1})$, and

$$\begin{aligned} \frac{u(t_n, x_{i+1}) - 2u(t_n, x_i) + u(t_n, x_{i-1})}{h^2} &= \frac{u(t_n, x_i + h) - 2u(t_n, x_i) + u(t_n, x_i - h)}{h^2} \\ &= \cdots = u_{xx}(t_n, x_i) + \frac{h^2}{24}\big(u_{xxxx}(t_n, \xi_i) + u_{xxxx}(t_n, \zeta_i)\big), \end{aligned}$$

for $\xi_i \in (x_{i-1}, x_i)$ and $\zeta_i \in (x_i, x_{i+1})$, for all $n = 0, \ldots, N_t - 1$, $i = 1, \ldots, N_x$.

Thus, making use of the PDE $u_t = u_{xx}$ at the point (t_n, x_i), we deduce

$$T_i^n = \frac{\tau}{2} u_{tt}(\rho_n, x_i) - \frac{h^2}{24}\big(u_{xxxx}(t_n, \xi_i) + u_{xxxx}(t_n, \zeta_i)\big),$$

which implies that

$$|T_i^n| \leq \frac{\tau}{2}|u_{tt}(\rho_n, x_i)| + \frac{h^2}{24}\big(|u_{xxxx}(t_n, \xi_i)| + |u_{xxxx}(t_n, \zeta_i)|\big) \leq \frac{\tau}{2} M_{tt} + \frac{h^2}{12} M_{xxxx}.$$

□

Notice that the truncation error converges to 0 when $h, \tau \to 0$, i.e., the numerical method approximates the original initial/boundary value problem. For brevity, we shall use the following short-hand notation:

$$\mathcal{T} := \frac{\tau}{2} M_{tt} + \frac{h^2}{12} M_{xxxx};$$

with this notation, (4.9) can be written as $|T_i^n| \leq \mathcal{T}$.

The next theorem describes the error behaviour of the explicit Euler method.

Theorem 4.3 *Consider the explicit Euler method defined defined by the equations (4.5), (4.6) and (4.7). Let u be the exact solution of the initial/boundary value problem (4.1), (4.2), and (4.3). Assume that the Courant number satisfies $0 \leq \mu \leq \frac{1}{2}$ for every τ, h. Then we have the following error bound:*

$$\max_{1 \leq i \leq N_x} |u(t_n, x_i) - u_i^n| \leq T_f\left(\frac{\tau}{2} M_{tt} + \frac{h^2}{12} M_{xxxx}\right), \tag{4.10}$$

for $n = 1, \ldots, N_t$.

Proof. For brevity we shall denote by $e_i^n := u(t_n, x_i) - u_i^n$ the error on each node (t_n, x_i); notice that $e_i^0 = u_0(x_i) - u_0^n = 0$. From the definition of the truncation error (4.8), we obtain

$$\tau T_i^n = u(t_{n+1}, x_i) - u(t_n, x_i) - \mu\big(u(t_n, x_{i+1}) - 2u(t_n, x_i) + u(t_n, x_{i-1})\big),$$

after multiplication of (4.8) by τ and using the definition of the Courant number μ; after rearrangement we get

$$u(t_{n+1}, x_i) = \mu u(t_n, x_{i+1}) + (1 - 2\mu)u(t_n, x_i) + \mu u(t_n, x_{i-1}) + \tau T_i^n. \tag{4.11}$$

for $n = 0, \ldots, N_t - 1,\ i = 1, \ldots, N_x$. Using this identity and the definition of the explicit Euler method (4.5), we can calculate

$$\begin{aligned} e_i^{n+1} &= u(t_{n+1}, x_i) - u_i^{n+1} \\ &= \mu u(t_n, x_{i+1}) + (1 - 2\mu)u(t_n, x_i) + \mu u(t_n, x_{i-1})\big) + \tau T_i^n - \big(\mu u_{i+1}^n + (1 - 2\mu)u_i^n + \mu u_{i-1}^n\big) \\ &= \mu e_{i+1}^n + (1 - 2\mu)e_i^n + \mu e_{i-1}^n + \tau T_i^n, \end{aligned}$$

for $n = 0, \ldots, N_t - 1,\ i = 1, \ldots, N_x$, which implies

$$|e_i^{n+1}| \le \mu|e_{i+1}^n| + |1 - 2\mu|\,|e_i^n| + \mu|e_{i-1}^n| + \tau|T_i^n|, \tag{4.12}$$

Define

$$E^n := \max_{1 \le i \le N_x} \{|e_i^n|\},$$

i.e., the maximum error at the n-th time-step. then taking the maximum with respect to i on the right-hand side of (4.12), we arrive to

$$|e_i^{n+1}| \le \mu E^n + |1 - 2\mu|E^n + \mu E^n + \tau\mathcal{T}.$$

Since $0 \le \mu \le 1/2$, we have $|1 - 2\mu| = 1 - 2\mu$; thus we deduce

$$|e_i^{n+1}| \le E^n + \tau\mathcal{T},$$

or, taking the maximum with respect to i,

$$E^{n+1} \le E^n + \tau\mathcal{T},$$

for $n = 0, \ldots, N_t - 1$. This means that we have inductively

$$E^n \le E^{n-1} + \tau\mathcal{T} \le E^{n-2} + 2\tau\mathcal{T} \le E^{n-3} + 3\tau\mathcal{T} \le \cdots \le E^0 + n\tau\mathcal{T} = n\tau\mathcal{T},$$

as $E^0 = 0$, being the maximum of $e_i^0 = 0$ with respect to i. Next we notice that we used that $n \le N_t$, and thus $n\tau \le \tau N_t = T_f$, so we deduce $E^n \le T_f\mathcal{T}$. The result now follows from (4.9). □

The above theorem tells us that the approximations u_i^n converge to the exact values of the solution $u(t_n, x_i)$ with first order with respect to the time-step τ and with second order with respect to the space grid parameter h, provided that the Courant number $\mu = \tau/h^2 \le \frac{1}{2}$. The restriction in the Courant number implies that $\tau \le \frac{h^2}{2}$. Using this, we can bound the error further to obtain from (4.10)

$$\max_{1 \le i \le N_x} |u(t_n, x_i) - u_i^n| \le \frac{T_f}{4}(M_{tt} + \frac{1}{3}M_{xxxx})\,h^2,$$

i.e., provided that the Courant number μ remains less or equal to $\frac{1}{2}$, the error is second order convergent with respect to the grid parameter h. Notice that, to have $\mu \le \frac{1}{2}$, we must make the time-step smaller and smaller appropriately as $h \to 0$.

Note also the condition $\mu \le \frac{1}{2}$ required in Theorem 4.3 to prove the convergence of the explicit Euler method. Going back to Example 4.1, we can see that in the first case (Figure 4.2) the approximation appeared reasonable, whereas in the second (Figure 4.3 case where the solution was oscillatory (or "unstable"), we had $\mu = 0.52 > \frac{1}{2}$! Hence, the assumption $\mu \le \frac{1}{2}$ appears reasonable and of some significance. In the next section, we shall investigate this issue further using tools from Fourier analysis.

Problem

16. Consider the explicit Euler method, approximating the solution to the initial/boundary value problem

$$\begin{aligned} u_t(t,x) &= u_{xx}(t,x) \quad \text{for all } t \in [0,T_f] \text{ and } x \in [a,b], \\ u(0,x) &= u_0(x), \qquad \text{for all } x \in [a,b], \\ u(t,a) = u(t,b) &= 0, \qquad \text{for all } t \in [0,T_f]. \end{aligned}$$

Show that when the Courant number $\mu = \frac{1}{6}$, the truncation error of the explicit Euler method is of second order with respect to the time-step τ. Conclude that the error of approximation is second order accurate with respect to the time-step τ.

4.1.2 Stability Analysis

To motivate the discussion below, we recall the Definition **??** of a Fourier series expansion. Using Euler's formula

$$\mathrm{e}^{\iota\theta} = \cos\theta + \iota\sin\theta,$$

where $\iota = \sqrt{-1}$ is the imaginary unit, we can rewrite the Fourier series expansion of a function $f : [-L, L] \to \mathbb{R}$ as

$$f(x) \sim \sum_{m=-\infty}^{\infty} c_m \mathrm{e}^{\iota m\pi x/L}, \quad \text{with} \quad c_m := \frac{1}{2L}\int_{-L}^{L} f(x)\mathrm{e}^{-\iota m\pi x/L}\mathrm{d}x,$$

where $a_m = c_m + c_{-m}$ and $b_m = \iota(c_m - c_{-m})$ for a_m, b_m as in Definition **??**; we shall refer to functions of the form $\mathrm{e}^{\iota m\pi x/L}$ as *simple waves*.

In Example 4.1, we saw that the difference method can develop non-physical oscillations, that pollute the quality of the approximation, leading to "instabilities". In particular, we notice that the "amplitude" of these oscillations increases after each time step. We shall attempt to give an explanation for this phenomenon.

Consider the explicit Euler method defined by the recursive relation (4.5). For simplicity we shall consider the problem on the whole real line for the space variable x (as opposed to $x \in [a, b]$ as in the original problem); that way we can make the discussion easier by not considering any boundary conditions at (t, a) and (t, b). Thus, we consider a grid of the form (t_n, x_i), with $t_n = n\tau$ and $x_i = ih$ for $n = 0, \dots N_t$, $i = 0, \pm1, \pm2, \dots$, with $\tau = T_f/N_t$ and $h \in \mathbb{R}$, i.e., a grid with infinite points in the x-direction.

To see how the explicit Euler method propagates information in the time-variable (hoping to understand why instabilities develop), we shall use simple waves as initial conditions, and calculate what happens after each time-step. To this end, we set

$$u_i^0 = \mathrm{e}^{\iota k x_i},$$

for $k \in \mathbb{R}$, to be the initial value at the node $(0, x_i)$ (i.e., the nodes located on the x-axis), for $i = 0, \pm1, \pm2, \dots$. Notice that since $x_i = ih$, we have $\mathrm{e}^{\iota k x_i} = \mathrm{e}^{\iota k i h}$. We now use the recursive relation (4.5) to evolve one time-step; then we get

$$u_i^1 = \mu u_{i+1}^0 + (1-2\mu)u_i^0 + \mu u_{i-1}^0 = \mu\mathrm{e}^{\iota k(i+1)h} + (1-2\mu)\mathrm{e}^{\iota k i h} + \mu\mathrm{e}^{\iota k(i-1)h} = \big(\mu\mathrm{e}^{\iota k h} + 1 - 2\mu + \mu\mathrm{e}^{-\iota k h}\big)\mathrm{e}^{\iota k i h}.$$

Observing now that

$$\mathrm{e}^{\iota k h} - 2 + \mathrm{e}^{-\iota k h} = \big(\mathrm{e}^{\iota k h/2} - \mathrm{e}^{-\iota k h/2}\big)^2 = -4\sin^2(\frac{1}{2}kh), \tag{4.13}$$

we deduce

$$u_i^1 = \big(1 - 4\mu\sin^2(\frac{1}{2}kh)\big)\mathrm{e}^{\iota k i h}. \tag{4.14}$$

This means that, after one time-step, the solution is amplified by the factor $\lambda \equiv \lambda(k) := 1 - 4\mu\sin^2(\frac{1}{2}kh)$. Inserting (4.14) into (4.5) to evolve to the next time-step, we obtain in completely similar fashion

$$u_i^2 = \lambda^2\mathrm{e}^{\iota k i h},$$

and so on; at the n-th time-step, we get

$$u_i^n = \lambda^n\mathrm{e}^{\iota k i h}. \tag{4.15}$$

We shall denote by $|\cdot|$ the size of a complex number. Hence, observing that $|\mathrm{e}^{\iota x}| = \sqrt{\cos^2 x + \sin^2 x} = 1$ for any $x \in \mathbb{R}$, we get

$$|u_i^n| = |\lambda^n|\,|\mathrm{e}^{\iota k i h}| = |\lambda^n| = |\lambda|^n.$$

Going back to the exact solution of the initial/boundary value problem with separation of variables (just above Example (**??**) in Section (**??**)), we can see that, for some constant C, we have $|u(t, x)| \le C$ as $t \to \infty$ for *any* initial condition $f(x)$ which admits a Fourier series expansion, i.e., the solution to the initial/boundary value problem remains bounded.[1].

Hence, since we chose $u_i^0 = \mathrm{e}^{\iota k x_i}$ as initial conditions here (which, of course, admit Fourier series expansion), we expect to observe that also $|u_i^n| \le C$ as $n \to \infty$; when this happens, we say that the numerical method is *stable*. For this to happen we must require $|\lambda|^n \le C$ as $n \to \infty$, which necessarily implies that $|\lambda| \le 1$ is satisfied.

[1] More rigorously, we can say that the solution of the Cauchy problem $u_t = u_{xx}$ with $u(0, x) = f(x)$ for $t \in (0, \infty)$ and $x \in \mathbb{R}$ satisfies $|u(t, x)| \le C$ as $t \to \infty$ for *any* initial condition $f(x)$, provided that $\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x < +\infty$. That way we are considering the relevant initial value problem for the stability analysis, i.e., the problem defined in the domain $[0, T_f] \times \mathbb{R}$, for any $T_f > 0$.

We check validity of the condition $|\lambda| < 1$: we have

$$1 \geq |\lambda| = |1 - 4\mu \sin^2(\frac{1}{2}kh)|, \quad \text{or} \quad 1 \geq 1 - 4\mu \sin^2(\frac{1}{2}kh) \geq -1, \quad \text{or} \quad 0 \leq \mu \sin^2(\frac{1}{2}kh) \leq \frac{1}{2},$$

which implies that $0 \leq \mu \leq 1/2$, since $0 \leq \sin^2(\frac{1}{2}kh) \leq 1$ for any $k \in \mathbb{R}$. Hence, we conclude that *the explicit Euler method is stable if and only if* $0 \leq \mu \leq 1/2$! Notice that this is in accordance with the behaviour we observed in Example (4.1). The above type of stability analysis using tools from Fourier analysis, is some times called *von Neumann* analysis, in the honour of its inventor John von Neumann [2].

We finally remark that it would be possible to arrive to the same answer $0 \leq \mu \leq 1/2$, if we have set directly $u_i^n = \lambda^n \mathrm{e}^{\iota k x_i}$ and carried over the calculation to find λ on the $n+1$-time level.

Problem

17. Consider the explicit Euler method, approximating the solution to the initial/boundary value problem from Problem (4.1) above, with $a = 0$ and $b = 1$. The stability analysis for the explicit Euler method, presented in Section 4.1.2 in the notes, does not take into account the effect of boundary conditions. One way of taking into account the boundary conditions is to set

$$u_i^0 = \sin(m\pi x_i)$$

for $m = 1, 2, 3 \ldots$, to be the initial conditions at node $(0, x_i)$, using the notation from Section 4.1.2. Why is that? What is the restriction on the Courant number μ in this case for stability?

[2] John von Neumann (1903-1957)

4.2 Implicit Methods

Having seen the restrictions in the Courant number μ that the explicit scheme is subject to, we shall now present some "implicit" schemes that are less prone to instabilities. The importance in having less restrictions in the Courant number can be seen by the fact that to maintain $0 \leq \mu \leq 1/2$ in the explicit Euler method, we had to take very small time-step τ compared the to space grid-size h, which is of course very demanding in terms of computations.

4.2.1 Implicit Euler Method

We consider the same initial/boundary value problem (4.1), (4.2) and (4.3).

Again, we shall approximate the derivatives in (4.1) using divided differences. To this end, we construct a grid as follows: we consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N_x+1}$, at distance h between them, such that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_{N_x} = x_{N_x-1} + h,\ x_{N_x+1} = b,$$

in the space-direction, and an equally distributed subdivision $t_0 < t_1 < \cdots < t_{N_t}$, at distance τ between them, such that

$$0 = t_0,\ t_1 = t_0 + \tau,\ t_2 = t_1 + \tau,\ \ldots,\ t_{N_t-2} = t_{N_t-1} + \tau,\ t_{N_t} = T_f,$$

in the time-direction (see Figure 4.1); hence, we have $h = (b-a)/(N_x+1)$ and $\tau = T_f/N_t$.

We formally make the following approximations, using backward difference for the time-derivative and central second difference for the space-derivative

$$u_t(t,x) \approx \delta^t_{\tau,-} u(t,x) = \frac{u(t,x) - u(t-\tau,x)}{\tau},$$

and

$$u_{xx}(t,x) \approx (\delta^x_h)^2 u(t,x) = \frac{u(t,x+h) - 2u(t,x) + u(t,x-h)}{h^2},$$

giving

$$\frac{u(t_n,x_i) - u(t_{n-1},x_i)}{\tau} - \frac{u(t_n,x_{i+1}) - 2u(t_n,x_i) + u(t_n,x_{i-1})}{h^2} \approx u_t(t_n,x_i) - u_{xx}(t_n,x_i) = 0, \tag{4.16}$$

using the equation (4.1).

Motivated by this formal reasoning, our aim is to find approximations u_i^n of the function values $u(t_n,x_i)$, for $n = 1,\ldots,N_t$ and $i = 1,\ldots,N_x$, we consider the following system of equations:

$$\frac{u_i^n - u_i^{n-1}}{\tau} - \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} = 0;$$

changing the numbering of the time-stepping, so that n is replaced by $n+1$, the method becomes

$$\frac{u_i^{n+1} - u_i^n}{\tau} - \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{h^2} = 0,$$

for $n = 0,\ldots,N_t - 1$ and $i = 1,\ldots,N_x$ [3]. After multiplication by τ and rearrangement, this becomes

$$-\mu u_{i+1}^{n+1} + (1+2\mu)u_i^{n+1} - \mu u_{i-1}^{n+1} = u_i^n, \quad \text{for } n = 0,\ldots,N_t-1,\ i = 1,\ldots,N_x, \tag{4.17}$$

where, again $\mu := \tau/h^2$ is the Courant number for this method. For (4.17) to make complete sense, we need to describe the initial condition (4.2) in the method, giving

$$u_i^0 = u_0(x_i) \quad i = 1,\ldots,N_x, \tag{4.18}$$

and the values for the nodes residing on the Dirichlet boundary can be given using the boundary conditions (4.3)

$$u_0^n = 0 = u_{N_x+1}^n \quad n = 0,\ldots,N_t. \tag{4.19}$$

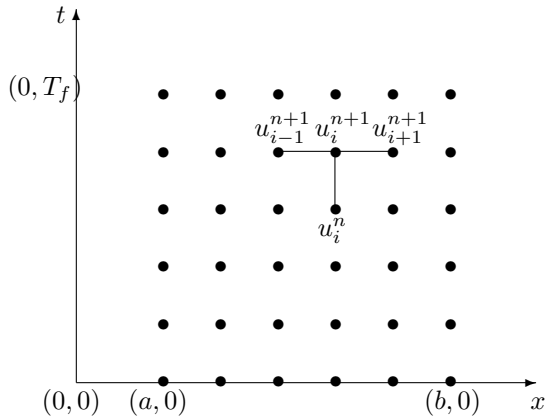


Figure 4.4: Implicit Euler Method.

The finite difference method defined by the equations (4.17), (4.18) and (4.19), will be referred to as the *implicit Euler method* (or the *backward Euler method*). In Figure 4.4, we can see a representation of the implicit Euler method.

The name "implicit" can be justified by observing to compute the approximations u_i^{n+1} we now *need to solve a linear system at each time-step*! Indeed, the u_i^{n+1}'s *cannot* be calculated explicitly from the previously computed u_i^n. Instead, writing the system (4.17) in matrix form (using the boundary conditions (4.19)), we get

$$\underbrace{\begin{pmatrix} 1+2\mu & -\mu & 0 & 0 & \dots & 0 \\ -\mu & 1+2\mu & -\mu & 0 & \dots & 0 \\ 0 & -\mu & 1+2\mu & -\mu & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots \\ 0 & \dots & 0 & -\mu & 1+2\mu & -\mu \\ 0 & \dots & 0 & 0 & -\mu & 1+2\mu \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ \vdots \\ u_{N_x-1}^{n+1} \\ u_{N_x}^{n+1} \end{pmatrix}}_{U^{n+1}} = \underbrace{\begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ \vdots \\ u_{N_x-1}^n \\ u_{N_x}^n \end{pmatrix}}_{U^n}, \tag{4.20}$$

for $n = 0, 1, \dots N_t - 1$; the values on the right-hand side for $n = 0$ are given by the initial condition. Therefore, to compute the approximations u_i^{n+1}, we need to solve the linear system (4.20), since the right-hand side vector contains the know values u_i^n's, computed in the previous time-step.

This method appears to be more computationally demanding that the explicit Euler method. Indeed here, to compute the u_i^{n+1}'s, a linear system has to be solved for each n, as opposed to the few direct calculations needed for the explicit Euler method[4].

In view of the additional computational cost for the implicit Euler method, it is natural to question its relevance in practice. Since the forward and the backward divided differences are both first order accurate (see Lemma (3.2)), we do not expect to achieve higher convergence rather for the implicit Euler method, as opposed to the explicit Euler method. Indeed, the following lemma, shows that the truncation error is of first order with respect to τ and of second order with respect to h.

Lemma 4.4 *We define the truncation error for the implicit Euler method by*

$$T_i^n := \frac{u(t_n, x_i) - u(t_{n-1}, x_i)}{\tau} - \frac{u(t_n, x_{i+1}) - 2u(t_n, x_i) + u(t_n, x_{i-1})}{h^2}.$$

Then, we have

$$|T_i^n| \leq \frac{\tau}{2} M_{tt} + \frac{h^2}{12} M_{xxxx}, \tag{4.21}$$

[3]the change $n \to n+1$ in the time-stepping is merely done for stylistic reasons, so that the comparison with the explicit Euler method described above can be facilitated.

[4]In fact the difference in the computational cost is not as big as it appears initially. Indeed, given that the matrix D is tridiagonal, inverting can be a very fast process. One such method is the so-called *Thomas algorithm* which is described in Morton & Mayers' book (Section 2.9).

for all $n = 0, \ldots, N_t$, $i = 1, \ldots, N_x$, *where*

$$M_{tt} := \max |u_{tt}(t,x)|, \quad \textit{and} \quad M_{xxxx} := \max |u_{xxxx}(t,x)|,$$

and the maxima are taken over all $(t,x) \in [0,T_f] \times [a,b]$.

Proof. The proof is left as an exercise. □

Thus, there seems to be no real benefit in terms of error reduction for any of the two methods [5].

Let us now examine the stability properties of the implicit Euler method. To this end, ignoring the boundary conditions (4.19), we consider a grid of the form (t_n, x_i), with $t_n = n\tau$ and $x_i = ih$ for $n = 0, \ldots N_t$, $i = 0, \pm 1, \pm 2, \ldots$, with $\tau = 1/N_t$ and $h \in \mathbb{R}$, i.e., a grid with infinite points in the x-direction. We set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the approximate solution at the node (t_n, x_i). We now use (4.17) to evolve one time-step; then we get

$$-\mu u_{i+1}^{n+1} + (1+2\mu) u_i^{n+1} - \mu u_{i-1}^{n+1} = u_i^n,$$

or

$$-\mu \lambda^{n+1} \mathrm{e}^{\iota k (i+1) h} + (1+2\mu) \lambda^{n+1} \mathrm{e}^{\iota k i h} - \mu \lambda^{n+1} \mathrm{e}^{\iota k (i-1) h} = \lambda^n \mathrm{e}^{\iota k i h}.$$

Dividing the last equation by $\lambda^n \mathrm{e}^{\iota k i h}$, we deduce

$$-\mu \lambda \mathrm{e}^{\iota k h} + (1+2\mu)\lambda - \mu \lambda \mathrm{e}^{-\iota k h} = 1,$$

which becomes

$$\lambda - \mu \lambda \big(\mathrm{e}^{\iota k h} - 2 + \mathrm{e}^{-\iota k h} \big) = 1.$$

Using (4.13), we get

$$\lambda + 4\mu\lambda \sin^2(\frac{1}{2} k h) = 1,$$

which, finally implies that

$$\lambda = \frac{1}{1 + 4\mu \sin^2(\frac{1}{2} k h)}.$$

For the numerical method to be stable, we must have $|\lambda| \leq 1$, which is the case here for all $\mu \geq 0$, as $0 \leq \sin^2(\frac{1}{2} k h) \leq 1$. Therefore, we conclude that *the implicit Euler method is stable for all* $\mu \geq 0$! A method that is stable for all Courant numbers $\mu \geq 0$ is called *unconditionally stable*. Hence, the additional computational cost to use the implicit Euler method is counterbalanced by its superior stability properties, as now we are allowed to take larger time-steps, leading to smaller N_t; notice, however, that taking larger time-step τ may result to slower convergence. The implicit Euler method is of great relevance in practice.

Problems

18. Write an implicit Euler method, to approximate the solution to the initial/boundary value problem

$$\begin{aligned} u_t(t,x) &= u_{xx}(t,x) \quad \text{for all } t \in [0,T_f] \text{ and } x \in [0,1], \\ u(0,x) &= u_0(x), \quad \text{for all } x \in (0,1), \\ u_x(t,0) = u(t,1) &= 0, \quad \text{for all } t \in [0,T_f]. \end{aligned}$$

19. Prove Lemma 4.4.

[5] We shall not consider the error analysis of the implicit Euler method here, as this requires a relatively involved discrete maximum principle which goes beyond the scope of these notes. We refer the interested reader to Section 2.11 of the book by Morton & Mayers.

4.2.2 The Crank-Nicolson Method

In the previous sections we presented both the explicit and the implicit Euler methods for the solution of parabolic initial/boundary value problems. We saw that the explicit methods is first order accurate with respect to the time-stepping τ (Theorem 4.3), and we have good reasons to believe that this is the case also for the implicit Euler method (as the truncation error for the implicit method is of first order with respect to τ; see Problem 4 in Problem Sheet 5). The crucial difference between the aforementioned methods lies in the superior stability properties of the implicit Euler method, that enables us to take larger time steps without the danger of the development of non-physical oscillations.

Here, we shall combine both methods, hoping to construct a method that is stable and, hopefully, of higher order in τ. Again, we consider the same initial/boundary value problem (4.1), (4.2) and (4.3).

We construct a grid as follows: we consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N_x+1}$, at distance h between them, such that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_{N_x} = x_{N_x-1} + h,\ x_{N_x+1} = b,$$

in the space-direction, and an equally distributed subdivision $t_0 < t_1 < \cdots < t_{N_t}$, at distance τ between them, such that

$$0 = t_0,\ t_1 = t_0 + \tau,\ t_2 = t_1 + \tau,\ \ldots,\ t_{N_t-2} = t_{N_t-1} + \tau,\ t_{N_t} = T_f,$$

in the time-direction (see Figure 4.1); hence, we have $h = (b-a)/(N_x+1)$ and $\tau = T_f/N_t$.

Our aim, as before, is to find approximations u_i^n of the function values $u(t_n, x_i)$, for $n = 1, \ldots, N_t$ and $i = 1, \ldots, N_x$. We define the *Crank-Nicolson method* by the following system of equations:

$$\frac{u_i^{n+1} - u_i^n}{\tau} = \frac{1}{2}\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} + \frac{1}{2}\frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{h^2},$$

for $n = 0, \ldots, N_t - 1$ and $i = 1, \ldots, N_x$; i.e., we can think of the Crank-Nicolson method as derived by taking the average of the explicit and implicit Euler methods. After multiplication by 2τ and rearrangement, this becomes

$$-\mu u_{i+1}^{n+1} + (2+2\mu)u_i^{n+1} - \mu u_{i-1}^{n+1} = \mu u_{i+1}^n + (2-2\mu)u_i^n + \mu u_{i-1}^n, \quad \text{for } n = 0, \ldots, N_t - 1,\ i = 1, \ldots, N_x, \tag{4.22}$$

where, again $\mu := \tau/h^2$ is the Courant number for this method. For (4.17) to make complete sense, we need to describe the initial condition (4.2) in the method, giving

$$u_i^0 = u_0(x_i) \quad i = 1, \ldots, N_x, \tag{4.23}$$

and the values for the nodes residing on the Dirichlet boundary can be given using the boundary conditions (4.3)

$$u_0^n = 0 = u_{N_x+1}^n \quad n = 0, \ldots, N_t. \tag{4.24}$$

In Figure 4.5, we can see a representation of the Crank-Nicolson method.

We can write (4.22) and (4.24) in matrix form as follows:

$$\underbrace{\begin{pmatrix} 2+2\mu & -\mu & 0 & \ldots & 0 \\ -\mu & 2+2\mu & -\mu & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ldots \\ 0 & \ldots & -\mu & 2+2\mu & -\mu \\ 0 & \ldots & 0 & -\mu & 2+2\mu \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \\ \vdots \\ u_{N_x-1}^{n+1} \\ u_{N_x}^{n+1} \end{pmatrix}}_{U^{n+1}} = \underbrace{\begin{pmatrix} 2-2\mu & \mu & 0 & \ldots & 0 \\ \mu & 2-2\mu & \mu & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ldots \\ 0 & \ldots & \mu & 2-2\mu & \mu \\ 0 & \ldots & 0 & \mu & 2-2\mu \end{pmatrix}}_{E} \underbrace{\begin{pmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_{N_x-1}^n \\ u_{N_x}^n \end{pmatrix}}_{U^n},$$

for $n = 0, 1, \ldots N_t - 1$; the values on the right-hand side for $n = 0$ are given by the initial condition.

We observe that the Crank-Nicolson method is also of implicit type, since to calculate the approximations u_i^{n+1} we need to solve a linear system at each time-step; more specifically, to calculate the vector U^{n+1}, we have to solve the linear system $DU^{n+1} = f$, where $f = EU^n$, the product of the matrix E with the vector U^n, which is known from the previous time-level (or, if $n = 0$, from the initial condition).

The error analysis of the Crank-Nicolson method will not be presented here, as it is quite involved (it is based on a discrete maximum principle). We refer the interested reader to the book by Morton & Mayers for a proof. Nevertheless, we can get an idea of the convergence rates by truncation error analysis.

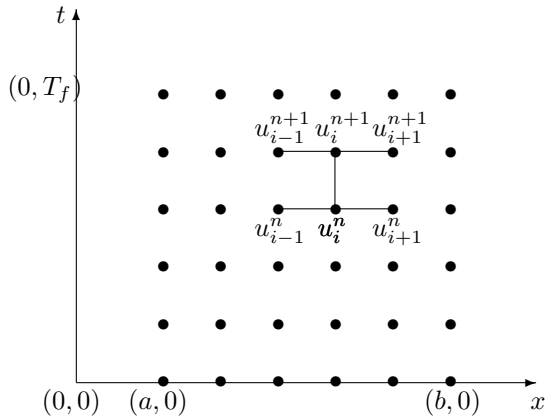


Figure 4.5: The Crank-Nicolson method.

Lemma 4.5 *We define the truncation error for the Crank-Nicolson method by*

$$T_i^{n+\frac{1}{2}} := \frac{u(t_{n+1},x_i)-u(t_n,x_i)}{\tau} - \frac{1}{2}\frac{u(t_{n+1},x_{i+1})-2u(t_{n+1},x_i)+u(t_{n+1},x_{i-1})}{h^2} - \frac{1}{2}\frac{u(t_n,x_{i+1})-2u(t_n,x_i)+u(t_n,x_{i-1})}{h^2}.$$

Then, we have

$$|T_i^{n+\frac{1}{2}}| \le \frac{\tau^2}{12}M_{ttt} + \frac{h^2}{12}M_{xxxx}, \tag{4.25}$$

for all $n = 0,\dots,N_t-1$, $i = 1,\dots,N_x$, where

$$M_{ttt} := \max|u_{ttt}(t,x)|, \quad \textit{and} \quad M_{xxxx} := \max|u_{xxxx}(t,x)|,$$

and the maxima are taken over all $(t,x) \in [0,T_f]\times[a,b]$, whenever M_{ttt} and M_{xxxx} are finite.

Proof. The proof is left as an exercise. □

Therefore, it appears that the Crank-Nicolson method is of second order with respect to the time-step τ. Therefore, we could choose $\tau = \mathcal{O}(h)$ and still achieve second order convergence of the truncation error (and, hopefully, of the error itself). Hence, we can choose larger time-steps and still obtain second order accuracy!

We now examine the stability properties of the Crank-Nicolson method. To this end, ignoring the boundary conditions (4.24), we consider a grid of the form (t_n,x_i), with $t_n = n\tau$ and $x_i = ih$ for $n = 0,\dots N_t$, $i = 0,\pm1,\pm2,\dots$, with $\tau = 1/N_t$ and $h \in \mathbb{R}$, i.e., a grid with infinite points in the x-direction. We set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the value at the node (t_n,x_i). We now use (4.22) to evolve one time-step; then we get

$$-\mu u_{i+1}^{n+1} + (2+2\mu)u_i^{n+1} - \mu u_{i-1}^{n+1} = \mu u_{i+1}^n + (2-2\mu)u_i^n + \mu u_{i-1}^n$$

or

$$-\mu\lambda^{n+1}\mathrm{e}^{\iota k(i+1)h} + (2+2\mu)\lambda^{n+1}\mathrm{e}^{\iota kih} - \mu\lambda^{n+1}\mathrm{e}^{\iota k(i-1)h} = \mu\lambda^n\mathrm{e}^{\iota k(i+1)h} + (2-2\mu)\lambda^n\mathrm{e}^{\iota kih} + \mu\lambda^n\mathrm{e}^{\iota k(i-1)h}.$$

Dividing the last equation by $\lambda^n\mathrm{e}^{\iota kih}$, we deduce

$$-\mu\lambda\mathrm{e}^{\iota kh} + (2+2\mu)\lambda - \mu\lambda\mathrm{e}^{-\iota kh} = \mu\mathrm{e}^{\iota kh} + 2 - 2\mu + \mu\mathrm{e}^{-\iota kh},$$

which becomes

$$2\lambda - \mu\lambda\big(\mathrm{e}^{\iota kh} - 2 + \mathrm{e}^{-\iota kh}\big) = 2 + \mu\big(\mathrm{e}^{\iota kh} - 2 + \mathrm{e}^{-\iota kh}\big).$$

Using (4.13), we get

$$2\lambda + 4\mu\lambda \sin^2(\frac{1}{2}kh) = 2 - 4\mu \sin^2(\frac{1}{2}kh),$$

which, finally implies that

$$\lambda = \frac{1 - 2\mu \sin^2(\frac{1}{2}kh)}{1 + 2\mu \sin^2(\frac{1}{2}kh)}.$$

For the numerical method to be stable, we must have $|\lambda| \leq 1$, which is the case here for all $\mu \geq 0$, as $0 \leq \sin^2(kh) \leq 1$. Therefore, we conclude that *the Crank-Nicolson method is stable for all* $\mu \geq 0$, i.e., it is unconditionally stable. Hence, the additional computational cost to use the Crank-Nicolson method is counterbalanced by its superior stability properties *and* convergence properties, as now we are allowed to take larger time-steps and still retain second order accuracy, without any stability loss.

Problem

20. Prove Lemma 4.5. *[Hint: use Taylor's theorem, expanding about the point* $(t_{n+1/2}, x_i)$*, where* $t_{n+1/2} := t_n + \frac{\tau}{2}$*.]*

4.3 Extensions to problems with non-constant coefficients

In the previous sections, we have seen various finite difference methods for the heat equation and the corresponding initial/boundary value problem. Here we shall consider the case when the PDE has variable coefficients. In particular, we seek an approximation to the solution of the problem: find a function $u : [0, T_f] \times [a, b] \to \mathbb{R}$ with continuous second derivatives, such that

$$u_t(t, x) = a(t, x)u_{xx}(t, x) - c(t, x)u(t, x) \quad \text{for all } t \in [0, T_f] \text{ and } x \in [a, b], \tag{4.26}$$

subject to the initial condition

$$u(0, x) = u_0(x), \text{ for all } x \in [a, b], \tag{4.27}$$

for some known continuous function $u_0 : [a, b] \to \mathbb{R}$, and subject to homogeneous boundary conditions

$$u(t, a) = u(t, b) = 0, \text{ for all } t \in [0, T_f]. \tag{4.28}$$

Throughout this section, we assume that the functions $a, c, f : [0, T_f] \times [a, b] \to \mathbb{R}$ are continuous and that $a(t, x) > 0$ and $c(t, x) \geq 0$, for all $(t, x) \in [0, T_f] \times [a, b]$.

4.3.1 Explicit methods

We begin by considering an extension of the explicit Euler method for this problem. We construct a grid as follows: we consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N_x+1}$, at distance h between them, such that

$$a = x_0, \ x_1 = x_0 + h, \ x_2 = x_1 + h, \ \ldots, \ x_{N_x} = x_{N_x-1} + h, \ x_{N_x+1} = b,$$

in the space-direction, and an equally distributed subdivision $t_0 < t_1 < \cdots < t_{N_t}$, at distance τ between them, such that

$$0 = t_0, \ t_1 = t_0 + \tau, \ t_2 = t_1 + \tau, \ \ldots, \ t_{N_t-2} = t_{N_t-1} + \tau, \ t_{N_t} = T_f,$$

in the time-direction (see Figure 4.1); hence, we have $h = (b - a)/(N_x + 1)$ and $\tau = T_f/N_t$.

Our aim is to find approximations u_i^n of the function values $u(t_n, x_i)$ for $n = 0, \ldots, N_t - 1$ and $i = 1, \ldots, N_x$, we consider the following system of equations:

$$\frac{u_i^{n+1} - u_i^n}{\tau} = a_i^n \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} - c_i^n u_i^n,$$

where $a_i^n := a(t_n, x_i)$ and $c_i^n := c(t_n, x_i)$, which becomes, after multiplication by τ and rearrangement

$$u_i^{n+1} = \mu a_i^n u_{i+1}^n + (1 - 2\mu a_i^n - \tau c_i^n)u_i^n + \mu a_i^n u_{i-1}^n, \quad \text{for } n = 0, \ldots, N_t - 1, \ i = 1, \ldots, N_x, \tag{4.29}$$

where we have used the usual notation for the Courant number $\mu := \tau/h^2$. For (4.29) to make complete sense, we need to describe the initial condition (4.27) in the method, giving

$$u_i^0 = u_0(x_i) \quad i = 1, \ldots, N_x, \tag{4.30}$$

and the values for the nodes residing on the Dirichlet boundary can be given using the boundary conditions (4.28)

$$u_0^n = 0 = u_{N_x+1}^n \quad n = 0, \ldots, N_t. \tag{4.31}$$

In an analogous fashion to the case of $a(t, x) = 1$, $c(t, x) = 0$ (i.e., the heat equation), presented in Section 4.1.1, one can prove the following result.

Theorem 4.6 *Consider the explicit Euler method (4.29). Let u be the exact solution of the initial/boundary value problem (4.26), (4.27), and (4.28). Assume that the Courant number satisfies*

$$\mu a_i^n + \frac{\tau}{2} c_i^n \leq \frac{1}{2},$$

for every τ, h and for every $n = 0, \ldots, N_t - 1$ and $i = 1, 2, \ldots, N_x$. Then we have the following error bound:

$$\max_{1 \leq i \leq N_x} |u(t_n, x_i) - u_i^n| \leq T_f \left(\frac{\tau}{2} M_{tt} + \frac{h^2}{12} A M_{xxxx}\right), \tag{4.32}$$

for $n = 1, \ldots, N_t$, where M_{tt} and M_{xxxx} defined as before, and $A := \max |a(t, x)|$, where the max is taken over $[0, T_f] \times [a, b]$.

Proof. The proof is left as an exercise. □

Let us now examine the stability properties of the explicit Euler method. To this end, ignoring the boundary conditions (4.31), we consider a grid of the form (t_n, x_i), with $t_n = n\tau$ and $x_i = ih$ for $n = 0, \dots N_t$, $i = 0, \pm 1, \pm 2, \dots$, with $\tau = 1/N_t$ and $h \in \mathbb{R}$, i.e., a grid with infinite points in the x-direction. We set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the approximate solution at the node (t_n, x_i). We now use (4.29) to evolve one time-step; then we get

$$\lambda^{n+1} \mathrm{e}^{\iota k i h} = \mu a_i^n \lambda^n \mathrm{e}^{\iota k (i+1) h} + (1 - 2\mu a_i^n - \tau c_i^n) \lambda^n \mathrm{e}^{\iota k i h} + \mu \lambda^n \mathrm{e}^{\iota k (i-1) h}.$$

Dividing the last equation by $\lambda^n \mathrm{e}^{\iota k i h}$, we deduce

$$\lambda = \mu a_i^n \mathrm{e}^{\iota k h} + 1 - 2\mu a_i^n - \tau c_i^n + \mu \mathrm{e}^{-\iota k h}.$$

Using (4.13), we get

$$\lambda = 1 - 4\mu a_i^n \sin^2(\frac{1}{2} k h) - \tau c_i^n.$$

For stability, we require $|\lambda| \leq 1$, which implies

$$-1 \leq 1 - 4\mu a_i^n \sin^2(\frac{1}{2} k h) - \tau c_i^n \leq 1, \quad \text{or} \quad \mu a_i^n + \frac{\tau}{4} c_i^n \leq \frac{1}{2}.$$

Hence, this method is stable if and only if $\mu a_i^n + \frac{\tau}{4} c_i^n \leq \frac{1}{2}$. In some cases, this can be restrictive in practice.

Problem

21. Prove Theorem 4.6

4.3.2 Implicit methods

Due to restrictions in the Courant number stemming from both the convergence and the stability requirements, the use of implicit methods can be more relevant in practical computations. To this end, we now present some extensions of the Crank-Nicolson method for the initial/boundary value problem (4.26), (4.27), and (4.28).

Our aim is to find approximations u_i^n of the function values $u(t_n, x_i)$ for $n = 0, \dots, N_t - 1$ and $i = 1, \dots, N_x$, using the Crank-Nicolson method:

$$\frac{u_i^{n+1} - u_i^n}{\tau} = \frac{1}{2} \alpha_i^n \Big(\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} + \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{h^2} \Big) - \frac{1}{2} \gamma_i^n \big(u_i^{n+1} + u_i^n \big),$$

where

$$\alpha_i^n := \frac{1}{2} \big(a(t_n, x_i) + a(t_{n+1}, x_i) \big), \quad \text{and} \quad \gamma_i^n := \frac{1}{2} \big(c(t_n, x_i) + c(t_{n+1}, x_i) \big). \tag{4.33}$$

This choice can be motivated by the fact that, for Crank-Nicolson-type methods it is preferable to expand the truncation error about the point $(t_{n+1/2}, x_i)$, where $t_{n+1/2} := t_n + \tau/2$; indeed, assuming the the coefficients a and c are twice continuously differentiable with respect to t and applying Taylor's theorem about the point $(t_{n+1/2}, x_i)$, we get

$$\alpha_i^n = a(t_{n+1/2}, x_i) + \frac{\tau^2}{8} a_{tt} a(\rho_n, x_i) + \frac{\tau^2}{8} a_{tt} a(\sigma_n, x_i),$$

for some $\rho_n \in (t_n, t_{n+1/2})$ and $\rho_n \in (t_{n+1/2}, t_{n+1})$, and similarly for γ_i^n.

After multiplication by 2τ and rearrangement, we get

$$-\mu \alpha_i^n u_{i+1}^{n+1} + (2 + 2\mu \alpha_i^n + \tau \gamma_i^n) u_i^{n+1} - \mu \alpha_i^n u_{i-1}^{n+1} = \mu \alpha_i^n u_{i+1}^n + (2 - 2\mu \alpha_i^n - \tau \gamma_i^n) u_i^n + \mu \alpha_i^n u_{i-1}^n, \tag{4.34}$$

for $n = 0, \dots, N_t - 1$, $i = 1, \dots, N_x$, together with (4.30) and (4.31).

We now examine the stability properties of the above method for this problem. To this end, ignoring the boundary conditions (4.24), we consider a grid of the form (t_n, x_i), with $t_n = n\tau$ and $x_i = ih$ for $n = 0, \dots N_t$, $i = 0, \pm 1, \pm 2, \dots$, with $\tau = 1/N_t$ and $h \in \mathbb{R}$, i.e., a grid with infinite points in the x-direction. We set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the value at the node (t_n, x_i). We now use (4.34) to evolve one time-step; then we get

$$\begin{aligned} & -\mu\alpha_i^n \lambda^{n+1} \mathrm{e}^{\iota k(i+1)h} + (2 + 2\mu\alpha_i^n + \tau\gamma_i^n)\lambda^{n+1}\mathrm{e}^{\iota kih} - \mu\alpha_i^n \lambda^{n+1}\mathrm{e}^{\iota k(i-1)h} \\ = \quad & \mu\alpha_i^n \lambda^{n} \mathrm{e}^{\iota k(i+1)h} + (2 - 2\mu\alpha_i^n - \tau\gamma_i^n)\lambda^{n}\mathrm{e}^{\iota kih} + \mu\alpha_i^n \lambda^{n}\mathrm{e}^{\iota k(i-1)h}. \end{aligned}$$

Dividing the last equation by $\lambda^n \mathrm{e}^{\iota kih}$, we deduce

$$-\mu\alpha_i^n \lambda \mathrm{e}^{\iota kh} + (2 + 2\mu\alpha_i^n + \tau\gamma_i^n)\lambda - \mu\alpha_i^n \lambda \mathrm{e}^{-\iota kh} = \mu\alpha_i^n \mathrm{e}^{\iota kh} + 2 - 2\mu\alpha_i^n - \tau\gamma_i^n + \mu\alpha_i^n \mathrm{e}^{-\iota kh},$$

which becomes

$$(2 + \tau\gamma_i^n)\lambda - \mu\alpha_i^n \lambda\big(\mathrm{e}^{\iota kh} - 2 + \mathrm{e}^{-\iota kh}\big) = 2 - \tau\gamma_i^n + \mu\alpha_i^n\big(\mathrm{e}^{\iota kh} - 2 + \mathrm{e}^{-\iota kh}\big).$$

Using (4.13), we get

$$(2 + \tau\gamma_i^n)\lambda + 4\mu\alpha_i^n \lambda \sin^2(\frac{1}{2}kh) = 2 - \tau\gamma_i^n - 4\mu\alpha_i^n \sin^2(\frac{1}{2}kh),$$

which, finally implies that

$$\lambda = \frac{2 - \tau\gamma_i^n - 4\mu\alpha_i^n \sin^2(\frac{1}{2}kh)}{2 + \tau\gamma_i^n + 4\mu\alpha_i^n \sin^2(\frac{1}{2}kh)}.$$

For the numerical method to be stable, we must have $|\lambda| \leq 1$, which is the case here for all $\mu \geq 0$. Therefore, we conclude that *the method is stable for all* $\mu \geq 0$, i.e., it is unconditionally stable.

A more "natural" extension to the Crank-Nicolson method for the initial/boundary value problem (4.26), (4.27), and (4.28), is given by the following. Our aim is to find approximations u_i^n of the function values $u(t_n, x_i)$ for $n = 0, \ldots, N_t - 1$ and $i = 1, \ldots, N_x$, using the Crank-Nicolson method:

$$\frac{u_i^{n+1} - u_i^n}{\tau} = \frac{1}{2}\Big(a_i^{n+1}\frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{h^2} - c_i^{n+1}u_i^{n+1}\Big) + \frac{1}{2}\Big(a_i^{n}\frac{u_{i+1}^{n} - 2u_i^{n} + u_{i-1}^{n}}{h^2} - c_i^{n}u_i^{n}\Big),$$

where $a_i^n := a(t_n, x_i)$ and $c_i^n := c(t_n, x_i)$. That is, we consider "half" of the "differential operator" on the n-th time step while the other "half" on the $n+1$-st time step. It is an interesting exercise to give a bound on the truncation error for this method, as well as perform stability analysis.

Problem

22. Find a bound for the truncation error of the Crank-Nicolson method for the initial/bounary value problem with non-constant coefficients, described above. *[Hint: use Taylor's theorem, expanding about the point* $(t_{n+1/2}, x_i)$*, where* $t_{n+1/2} := t_n + \frac{\tau}{2}$*.]*

Chapter 5

Finite Difference Methods for Elliptic Problems

We shall now consider finite difference methods for elliptic problems. In particular, we shall be concerned with approximating the solution to the Dirichlet problem for the Poisson equation on a (bounded, simply connected, open) domain $\Omega \subset \mathbb{R}^2$, which reads: find a twice differentiable function $u : \Omega \to \mathbb{R}$, such that

$$\Delta u(x, y) = f(x, y) \text{ for all } (x, y) \in \Omega, \text{ and } u(x, y) = 0 \text{ for all } (x, y) \in \partial\Omega, \tag{5.1}$$

for some known bounded function $f : \Omega \to \mathbb{R}$, and with $\partial\Omega$ denoting the boundary of the domain Ω. We shall refer to Ω as the *computational domain*, We recall the definition of the Laplace operator (also known as *Laplacian*)

$$\Delta \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}.$$

The PDE in (5.1) is the archetype elliptic PDE. It is possible to show that this problem has a unique solution, provided that the domain Ω is "nice enough". (We prefer at this point not to give further details regarding the relation of the smoothness of the solution u with the "smoothness" of the boundary of Ω, as it is outside the scope of these notes.) However, in most cases it is very difficult or even impossible to find the solution u exactly, especially when the geometry of Ω is complicated. Therefore, the need to calculate accurate approximations of u is evident, considering that the Poisson problem, or its generalisations, are often met in many problems in mathematical modelling of various disciplines/phenomena.

To make matters simpler, suppose for the moment that $\Omega = (0, 1)^2 \subset \mathbb{R}^2$ is the unit square in $\mathbb{R}^2$. When designing a finite difference method, the first step is to decide upon a finite number of points $(x_i, y_j) \in \Omega$, with $i = 0, 1, 2, \ldots, N_x + 1$ and $j = 0, 1, 2, \ldots, N_y + 1$, on which we shall be seeking approximations $u_{i,j}$ to the exact values $u(x_i, y_j)$. The set $\{(x_i, y_j) : i = 0, 1, 2, \ldots, N_x + 1, j = 0, 1, 2, \ldots, N_y + 1\}$ will be referred to, as the *grid* (also known in the literature as the *mesh*), and the points (x_i, y_j) will be referred to as the *grid points* or as the *nodes*.

5.1 The five-point scheme

Going back to the problem (5.1) on $\Omega = (0, 1)^2$, we set $N_x = N_y$ and we construct a grid as follows. We consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N+1}$, at distance h between them, such that

$$0 = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N+1} = 1,$$

in the x-direction, and an equally distributed subdivision $y_0 < y_1 < \cdots < y_{N+1}$, at distance h between them, such that

$$0 = y_0,\ y_1 = y_0 + h,\ y_2 = y_1 + h,\ \ldots,\ y_N = y_{N-1} + h,\ y_{N+1} = 1,$$

in the y-direction; hence, we have $h = 1/(N + 1)$. We note that it is easy to generalise the ideas presented below when different number of nodes are used in each space direction; this is not done here for simplicity of the presentation.

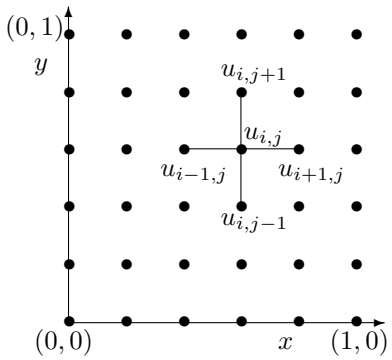


Figure 5.1: The five-point scheme.

Inspired by the discussion on divided differences approximating derivatives from Section 3.1, we can formally make the following approximations

$$\frac{\partial^2 u}{\partial x^2}(x,y) \equiv u_{xx}(x,y) \approx (\delta_h^x)^2 u(x,y), \quad \text{and} \quad \frac{\partial^2 u}{\partial y^2}(x,y) \equiv u_{yy}(x,y) \approx (\delta_h^y)^2 u(x,y),$$

where we have adopted the notational convention that the superscript x, denotes that the divided difference operator acts on the first variable and correspondingly for y, giving

$$\Delta u(x,y) \approx \big((\delta_h^x)^2 + (\delta_h^y)^2\big)u(x,y) = \frac{1}{h^2}\big(u(x+h,y) + u(x-h,y) + u(x,y+h) + u(x,y-h) - 4u(x,y)\big),$$

that is Δu at any point $(x,y) \in \Omega$ can be approximated by 5 neighbouring values of u. Restricting this on the nodes (x_i, y_j), we have formally

$$\begin{aligned} \Delta u(x_i,y_j) &\approx \frac{1}{h^2}\big(u(x_i+h,y_j) + u(x_i-h,y_j) + u(x_i,y_j+h) + u(x_i,y_j-h) - 4u(x_i,y_j)\big) \\ &= \frac{1}{h^2}\big(u(x_{i+1},y_j) + u(x_{i-1},y_j) + u(x_i,y_{j+1}) + u(x_i,y_{j-1}) - 4u(x_i,y_j)\big), \end{aligned}$$

for all $i,j = 1,\ldots,N$. Making use of the governing equation (5.1) applied to the grid points, we have

$$\frac{1}{h^2}\big(u(x_{i+1},y_j) + u(x_{i-1},y_j) + u(x_i,y_{j+1}) + u(x_i,y_{j-1}) + 4u(x_i,y_j)\big) \approx \Delta u(x_i,y_j) = f(x_i,y_j).$$

Motivated by this formal approximation, and recalling that our aim is to find approximations $u_{i,j}$ of the function values $u(x_i,y_j)$ for $i,j = 1,\ldots,N$, we consider the following system of equations:

$$\frac{1}{h^2}\big(u_{i,j+1} + u_{i+1,j} - 4u_{i,j} + u_{i-1,j} + u_{i,j-1}\big) = f(x_i,y_j), \quad 1 \le i,j \le N,$$

which becomes, after multiplication by h^2,

$$u_{i,j+1} + u_{i+1,j} - 4u_{i,j} + u_{i-1,j} + u_{i,j-1} = h^2 f(x_i,y_j), \quad 1 \le i,j \le N; \tag{5.2}$$

(see Figure 5.1). Notice that the approximations $u_{i,j}$ are *not* the same as exact values $u(x_i,y_j)$. For (5.2) to make complete sense, we need to determine the values on the grid points residing on the boundary (notice that when, for instance $j = n$, the first term on the left-hand side of (5.2) becomes $u_{i,N+1}$ which cannot be calculated from neighbouring values). But, this is not a problem, as we do *not* need to to find approximation of u at the grid points residing on the boundary $\partial\Omega$; the values of the solution u on these points are known from the boundary condition $u = 0$ on $\partial\Omega$! So, since we have from the boundary condition

$$u(x_0,y_j) = u(x_{N+1},y_j) = u(x_i,y_0) = u(x_i,y_{N+1}) = 0,$$

for all $i, j = 0, \dots N+1$, we can set

$$u_{0,j} = u_{N+1,j} = u_{i,0} = u_{i,N+1} = 0, \tag{5.3}$$

for all $i, j = 0, \dots, N+1$. Now, it is not hard to see that (5.2) together with the conditions (5.3) give rise to an algebraic system of N^2 equations with N^2 unknowns: the values $u_{i,j}$ for $i, j = 1, \dots N$. Solving this algebraic linear system, we can calculate the approximations $u_{i,j}$ to the exact values $u(x_i, y_j)$ for $0 \leq i, j \leq N$. The method (5.2) is often called the *five point scheme*.

From PDE theory (i.e., the fact that the problem (5.1) is well-posed and has unique solution, we understand that the value of the solution u at each (internal) point $(x, y) \in \Omega$ is completely determined by the values of u on a neighbourhood of each point and, inductively, on the known values at the boundary $\partial\Omega$. Hence, when designing a finite difference method for this problem, we aimed at imitating this principle; indeed, the five-point scheme imitates the dependence of the value of of the solution u on its neighbours, and its neighbours depend on their own neighbours, and so on, until the known boundary values are reached!

Let us study in greater detail the linear system of N^2 equations with N^2 unknowns, our aim being to write the system in matrix form $AU = h^2 F$, where A is an $N^2 \times N^2$ matrix, U is an N^2-vector having components the unknown values $u_{i,j}$, $i, j = 1, \dots N$, and F is an N^2-vector having components the values of the *forcing function* on the grid points $f(x_i, y_j) =: f_{i,j}$, for $i, j = 1, \dots N$. To do so, we should choose a way of sorting the unknown values $u_{i,j}$, $i, j = 1, \dots N$ in the vector U. Let us agree for the moment, that we sort $u_{i,j}$ by rows, from the bottom to the top of the grid, i.e., we choose

$$U = (u_{1,1}, u_{2,1}, \dots, u_{N,1}, u_{1,2}, u_{2,2}, \dots, u_{N,2}, \dots, u_{1,N}, u_{2,N}, \dots, u_{N,N})^T,$$

and, correspondingly, we let

$$F = (f_{1,1}, f_{2,1}, \dots, f_{N,1}, f_{1,2}, f_{2,2}, \dots, f_{N,2}, \dots, f_{1,N}, f_{2,N}, \dots, f_{N,N})^T;$$

then, we can write the method in matrix form $AU = h^2 F$, with

$$\underbrace{\begin{pmatrix} B & I & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ I & B & I & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & I & B & I & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & I & B & I \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & I & B \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} u_{1,1} \\ u_{2,1} \\ \vdots \\ u_{N,1} \\ u_{1,2} \\ \vdots \\ u_{N,2} \\ \vdots \\ u_{1,N} \\ \vdots \\ u_{N,N} \end{pmatrix}}_{U} = h^2 \underbrace{\begin{pmatrix} f_{1,1} \\ f_{2,1} \\ \vdots \\ f_{N,1} \\ f_{1,2} \\ \vdots \\ f_{N,2} \\ \vdots \\ f_{1,N} \\ \vdots \\ f_{N,N} \end{pmatrix}}_{F} \tag{5.4}$$

where $\mathbf{0}$ is the $N \times N$ zero matrix, I is the $N \times N$ identity matrix, and B is $N \times N$ matrix

$$B = \begin{pmatrix} -4 & 1 & 0 & 0 & \dots & 0 \\ 1 & -4 & 1 & 0 & \dots & 0 \\ 0 & 1 & -4 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots \\ 0 & \dots & 0 & 1 & -4 & 1 \\ 0 & \dots & 0 & 0 & 1 & -4 \end{pmatrix}.$$

Notice that the discretisation parameter h does not appear explicitly on the left-hand side of (7.27), but it appears on the right-hand side. This might seem strange at first sight, but a closer look reveals that the matrix A is implicitly dependent on h: indeed, as $h = 1/(N+1)$, we can see that the size of the matrix A can be parametrised by h, as $N = 1/h - 1$. More importantly, we see that as we make h smaller the size of the linear system (7.27) grows and, therefore, the use of computers becomes a necessity: suggestively, we can see that with a modest 100×100 grid, we would need to solve a system of $10,000$ equations with $10,000$ unknowns, which is far beyond the human capabilities.

5.2 Error analysis

For a numerical method to to be deemed useful, we must be confident enough that it provides us with good approximations of the exact solution to our problem. This confidence usually comes by analysing the convergence properties of the method. Generally speaking, by convergence of a numerical method we mean that the approximation should converge to the exact solution (in an appropriate way of defining distance), as the magnitude of the discretisation parameter(s) (e.g., the parameter h above) decreases to zero.

Let us begin, by trying to get a feeling on how well the numerical method (5.2) imitates the original problem (5.1). A way of doing so is to estimate the *truncation error*, which in this case is defined as

$$T_{i,j} := \frac{1}{h^2}\big(u(x_{i+1},y_j)+u(x_{i-1},y_j)+u(x_i,y_{j+1})+u(x_i,y_{j-1})-4u(x_i,y_j)\big)-f(x_i,y_j).$$

The truncation error of a numerical method is defined by substituting the exact solution into the numerical method, thereby representing how much the numerical method fails to imitate the exact boundary-value problem.

Lemma 5.1 *Let u, the exact solution to (5.1), be smooth enough so that the quantities*

$$M_{xxxx} := \max_{(x,y)\in\bar{\Omega}} |u_{xxxx}(x,y)|, \quad \textit{and} \quad M_{yyyy} := \max_{(x,y)\in\bar{\Omega}} |u_{yyyy}(x,y)|,$$

are finite, and let the truncation error $T_{i,j}$ be defined as above, for all $i,j=1,\ldots,n$, where $\bar{\Omega}$ defines the closure of Ω (i.e., $\bar{\Omega}=\Omega\cup\partial\Omega$). Then we have the bound:

$$|T_{i,j}| \le \frac{h^2}{12}\big(M_{xxxx}+M_{yyyy}\big), \tag{5.5}$$

for all $i,j=1,\ldots,N$.

Proof. We use Taylor's Theorem:

$$\begin{aligned} T_{i,j} &= \frac{1}{h^2}\big(u(x_i+h,y_j)+u(x_i-h,y_j)+u(x_i,y_j+h)+u(x_i,y_j-h)+4u(x_i,y_j)\big)-f(x_i,y_j)\\ &= u_{xx}(x_i,y_j)+\frac{h^2}{24}\big(u_{xxxx}(\xi_1,y_j)+u_{xxxx}(\zeta_1,y_j)\big)\\ &\quad +u_{yy}(x_i,y_j)+\frac{h^2}{24}\big(u_{yyyy}(x_i,\xi_2)+u_{yyyy}(x_i,\zeta_2)\big)-f(x_i,y_j), \end{aligned}$$

for some $\xi_1,\zeta_1\in[x_{i-1},x_{i+1}]$, and for some $\xi_2,\zeta_2\in[y_{i-1},y_{i+1}]$, using (3.16) for the first and second variable, respectively. Now, since u is the exact solution to (5.1), we can use the PDE $\Delta u(x_i,y_j)=f(x_i,y_j)$ on the point (x_i,y_j) to obtain

$$T_{i,j}=\frac{h^2}{24}\big(u_{xxxx}(\xi_1,y_j)+u_{xxxx}(\zeta_1,y_j)\big)+\frac{h^2}{24}\big(u_{yyyy}(x_i,\xi_2)+u_{yyyy}(x_i,\zeta_2)\big),$$

which can be finally bounded from above:

$$|T_{i,j}|\le\frac{h^2}{12}\big(\max_{(x,y)\in\bar{\Omega}}|u_{xxxx}(x,y)|+\max_{(x,y)\in\bar{\Omega}}|u_{yyyy}(x,y)|\big),$$

which gives the result. □

The above result is potentially good news: the bound (5.5) says that $T_{i,j}\to 0$ as $h\to 0$, i.e., that the numerical scheme approximates the PDE as $h\to 0$! Unfortunately, though, this does *not* imply also that the numerical approximation $u_{i,j}$ converges to $u(x_i,y_j)$ also, as $h\to 0$. Nevertheless, this is also the case, as the following result reveals.

Theorem 5.2 *Let Ω and the grid as above. Consider the five-point scheme approximation $u_{i,j}$ at the point (x_i,y_j) of the exact solution $u(x_i,y_j)$ of the problem (5.1), and suppose that the forcing function f is smooth enough. Then, the following bound holds*

$$|u_{i,j}-u(x_i,y_j)|\le\frac{h^2}{96}\big(M_{xxxx}+M_{yyyy}\big), \tag{5.6}$$

for all $i,j=1,\ldots,N$, with M_{xxxx} and M_{yyyy} as in Lemma 5.1.

Proof. The proof is beyond the scope of these notes; a proof (using the discrete maximum principle) for general domains Ω can be found in Morton and Mayers (Section 6.2). Another proof (for the case $\Omega = (0,1)^2$) can be found in Iserles (Theorem 7.2) using the eigenvalue/eigenvector decomposition of the matrix A above.

□

The above theorem implies that $u_{i,j} \to u(x_i, y_j)$, as $h \to 0$. Moreover, it says that $u_{i,j} \to u(x_i, y_j)$ like $\mathcal{O}(h^2)$, i.e., every time we half the grid-size h, we should expect the error on the left-hand side of (5.6) to decrease about 4 times.

5.3 Finite difference methods for general elliptic problems

We want to construct a finite difference method for the general 2nd order linear elliptic PDE in 2 dimensions:

$$au_{xx} + 2bu_{xy} + cu_{yy} = f, \quad \text{for} \quad (x,y) \in \Omega \subset \mathbb{R}^2 \text{ and } u = 0 \text{ on } \partial\Omega,$$

where a, b, c, f are functions of the independent variables x and y only, such that $\mathcal{D} = b^2 - ac < 0$. In the previous section, we considered the case $a = 1 = c$, $b = 0$. The partial derivatives u_{xx} and u_{yy} can be approximated by second central divided differences as done in the previous section. Therefore, the remaining challenge is to construct a finite difference approximation of the mixed derivative u_{xy}.

Let us first revisit the elementary divided differences considered in Section 3.1. The second central divided difference

$$\delta_h^2 f(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2},$$

for a function of one variable $f : \mathbb{R} \to \mathbb{R}$ was defined through two successive applications of the first central difference δ_h (see equation (3.14) for details). Hence, we can construct a second central divided difference for $u_{xy} = (u_x)_y$ by successive application of first central difference of spacing $2h$ in the x-direction, followed by a first central difference of spacing $2h$ in the y-direction, i.e.,

$$\begin{aligned}
\delta_{2h}^y \delta_{2h}^x u(x,y) &= \delta_{2h}^y\big(\delta_{2h}^x u(x,y)\big) = \delta_{2h}^y\Big(\frac{u(x+h,y) - u(x-h,y)}{2h}\Big) = \frac{1}{2h}\big(\delta_{2h}^y u(x+h,y) - \delta_{2h}^y u(x-h,y)\big) \\
&= \frac{1}{2h}\Big(\frac{u(x+h,y+h) - u(x+h,y-h)}{2h} - \frac{u(x-h,y+h) - u(x-h,y-h)}{2h}\Big) \\
&= \frac{1}{4h^2}\big(u(x+h,y+h) - u(x+h,y-h) - u(x-h,y+h) + u(x-h,y-h)\big). \qquad (5.7)
\end{aligned}$$

Hence, we can make the following approximation

$$u_{xy} \approx \delta_{2h}^y \delta_{2h}^x u(x,y),$$

which, together with the known approximations for u_{xx} and u_{yy}

$$u_{xx}(x,y) \approx (\delta_h^x)^2 u(x,y), \quad \text{and} \quad u_{yy}(x,y) \approx (\delta_h^y)^2 u(x,y)$$

(considered in the previous section), gives formally

$$f = au_{xx} + 2bu_{xy} + cu_{yy} \approx a(\delta_h^x)^2 u(x,y) + 2b\delta_{2h}^y \delta_{2h}^x u(x,y) + c(\delta_h^y)^2 u(x,y). \qquad (5.8)$$

We now consider a grid, assuming for simplicity that $\Omega = (0,1)^2$. We construct a grid as follows: we consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N+1}$, at distance h between them, such that

$$0 = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N+1} = 1,$$

in the x-direction, and an equally distributed subdivision $y_0 < y_1 < \cdots < y_{N+1}$, at distance h between them, such that

$$0 = y_0,\ y_1 = y_0 + h,\ y_2 = y_1 + h,\ \ldots,\ y_N = y_{N-1} + h,\ y_{N+1} = 1,$$

in the y-direction, giving hence, we have $h = 1/(N+1)$. Therefore, the formal approximation (5.8) motives the following finite difference method: find approximations $u_{i,j}$ of the exact solution $u(x_i, y_j)$, such that

$$\frac{a_{i,j}}{h^2}\big(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}\big) + \frac{b_{i,j}}{2h^2}\big(u_{i+1,j+1} - u_{i+1,j-1} - u_{i-1,j+1} + u_{i-1,j-1}\big) + \frac{c_{i,j}}{h^2}\big(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}\big) = f_{i,j},$$

for all $1 \le i, j \le N$, where $a_{i,j} := a(x_i, y_j)$, $b_{i,j} := b(x_i, y_j)$, $c_{i,j} := c(x_i, y_j)$ and $f_{i,j} := f(x_i, y_j)$; after multiplication by h^2 and rearrangement, this becomes

$$\begin{aligned}
& \frac{b_{i,j}}{2} u_{i-1,j-1} + c_{i,j} u_{i,j-1} - \frac{b_{i,j}}{2} u_{i+1,j-1} \\
+ \quad & a_{i,j} u_{i-1,j} - 2(a_{i,j} + c_{i,j}) u_{i,j} + a_{i,j} u_{i+1,j} \qquad (5.9) \\
- \quad & \frac{b_{i,j}}{2} u_{i-1,j+1} + c_{i,j} u_{i,j+1} + \frac{b_{i,j}}{2} u_{i+1,j+1} = h^2 f_{i,j}. \qquad (5.10)
\end{aligned}$$

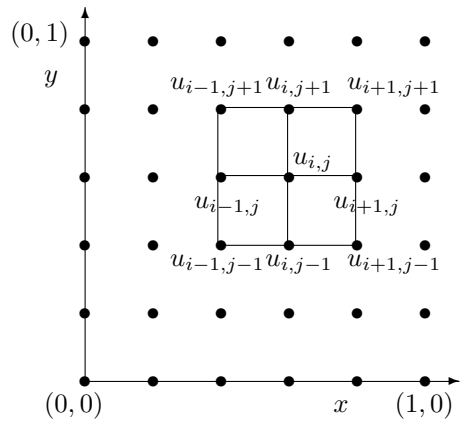


Figure 5.2: The finite difference method for the general 2nd order elliptic problem.

The schematic representation of this method is given in Figure 5.2. For (5.9) to make complete sense, we need to include the homogenous Dirichlet boundary conditions, by setting

$$u_{0,j} = u_{N+1,j} = u_{i,0} = u_{i,N+1} = 0, \tag{5.11}$$

for all $i, j = 0, \ldots, N+1$.

We study in greater detail the linear system of N^2 equations with N^2 unknowns, our aim being to write the system in matrix form $AU = h^2F$, where A is an $N^2 \times N^2$ matrix, U is an N^2-vector having components the unknown values $u_{i,j}$, $i, j = 1, \ldots N$, and F is the right-hand side N^2-vector. To do so, we should choose a way of sorting the unknown values $u_{i,j}$, $i, j = 1, \ldots N$ in the vector U. Let us agree for the moment, that we sort $u_{i,j}$ by rows, from the bottom to the top of the grid, i.e., we choose

$$U = (u_{1,1}, u_{2,1}, \ldots, u_{N,1}, u_{1,2}, u_{2,2}, \ldots, u_{N,2}, \ldots, u_{1,N}, u_{2,N}, \ldots, u_{N,N})^T,$$

and, correspondingly, we let

$$F = (f_{1,1}, f_{2,1}, \ldots, f_{N,1}, f_{1,2}, f_{2,2}, \ldots, f_{N,2}, \ldots, f_{1,N}, f_{2,N}, \ldots, f_{N,N})^T;$$

then, we can write the method in matrix form $AU = h^2F$, with

$$\underbrace{\begin{pmatrix} B_1 & D_1 & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ C_2 & B_2 & D_2 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & C_3 & B_3 & D_3 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \ldots & \mathbf{0} & C_{N-1} & B_{N-1} & D_{N-1} \\ \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & C_N & B_N \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} u_{1,1} \\ u_{2,1} \\ \vdots \\ u_{N,1} \\ u_{1,2} \\ \vdots \\ u_{N,2} \\ \vdots \\ u_{1,N} \\ \vdots \\ u_{N,N} \end{pmatrix}}_{U} = h^2 \underbrace{\begin{pmatrix} f_{1,1} \\ f_{2,1} \\ \vdots \\ f_{N,1} \\ f_{1,2} \\ \vdots \\ f_{N,2} \\ \vdots \\ f_{1,N} \\ \vdots \\ f_{N,N} \end{pmatrix}}_{F} \tag{5.12}$$

where $\mathbf{0}$ is the $N \times N$ zero matrix, B_j for $j = 1, 2, \ldots, N$, C_j for $j = 2, 3, \ldots, N$, and D_j for $j = 1, \ldots, N-1$, are the $N \times N$ matrices, defined by

$$B_j = \begin{pmatrix} -2(a_{1,j} + c_{1,j}) & a_{1,j} & 0 & \ldots & 0 & 0 \\ a_{2,j} & -2(a_{2,j} + c_{2,j}) & a_{2,j} & 0 & \ldots & 0 \\ 0 & a_{3,j} & -2(a_{3,j} + c_{3,j}) & a_{3,j} & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots \\ 0 & \ldots & 0 & a_{N-1,j} & -2(a_{N-1,j} + c_{N-1,j}) & a_{N-1,j} \\ 0 & \ldots & 0 & 0 & a_{N,j} & -2(a_{N,j} + c_{N,j}) \end{pmatrix},$$

$$C_j = \begin{pmatrix} c_{1,j} & -\frac{b_{1,j}}{2} & 0 & 0 & \dots & 0 \\ \frac{b_{2,j}}{2} & c_{2,j} & -\frac{b_{2,j}}{2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots \\ 0 & \dots & 0 & \frac{b_{N-1,j}}{2} & c_{N-1,j} & -\frac{b_{N-1,j}}{2} \\ 0 & \dots & 0 & 0 & \frac{b_{N,j}}{2} & c_{N,j} \end{pmatrix},$$

and

$$D_j = \begin{pmatrix} c_{1,j} & \frac{b_{1,j}}{2} & 0 & 0 & \dots & 0 \\ -\frac{b_{2,j}}{2} & c_{2,j} & \frac{b_{2,j}}{2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots \\ 0 & \dots & 0 & -\frac{b_{N-1,j}}{2} & c_{N-1,j} & \frac{b_{N-1,j}}{2} \\ 0 & \dots & 0 & 0 & -\frac{b_{N,j}}{2} & c_{N,j} \end{pmatrix}.$$

Finally, we remark that it is possible to show that the divided difference approximation for the mixed derivative u_{xy} described above, converges with second order with respect to the grid parameter h (see problem below).

Problem

23. Show that the divided difference approximation for the mixed derivative u_{xy}

$$\delta_{2h}^y \delta_{2h}^x u(x,y)$$

converges with second order with respect to the grid parameter h.

Chapter 6

Finite Difference Methods for Hyperbolic Problems

Having considered finite difference methods for parabolic and for elliptic problems, we now focus to the case of hyperbolic PDEs, together with the corresponding initial/boundary value problems.

The usual example of a 2nd order hyperbolic PDE is the wave equation, which in two dimensions reads

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

considered together with some initial and/or boundary conditions.

Instead, let $v = v(t, x)$ and consider the following system of equations

$$u_t + v_x = 0, \quad \text{and} \quad u_x + v_t = 0.$$

Differentiating the first equation with respect to t, and using the second equation, we deduce, respectively:

$$0 = u_{tt} + v_{xt} = u_{tt} + v_{tx} = u_{tt} + (v_t)_x = u_{tt} + (-u_x)_x = u_{tt} - u_{xx},$$

i.e., u satisfies the wave equation! (The same applies for v, too.)

It is possible to rewrite the above system of equations in matrix form. Defining $U := (u, v)^T$, and using the convention that partial derivatives of U are understood as the vector of the same partial derivatives of the components, we have

$$U_t + AU_x = 0, \quad \text{where} \quad A := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Before arriving to systems of PDEs, it is natural to consider the simpler case of single (scalar) PDEs. Therefore it is of interest to consider the first order *advection equation*

$$u_t + u_x = 0, \tag{6.1}$$

equipped with initial condition

$$u(0, x) = u_0(x), \tag{6.2}$$

for some $u_0 : \mathbb{R} \to \mathbb{R}$ known function. We have seen in Section 1.5 that this Cauchy problem is well posed. Clearly it is possible to find the exact solution to this problem using the method of characteristics, as described in Example 1.22; the exact solution then reads

$$u(t, x) = u_0(x - t),$$

i.e., the solution is constant along each line $t = x + const$. In other words the characteristic curves "carry" the initial value $u(0, x_0) = u_0(x_0)$ along the line $t = x - x_0$.

Even though the exact solution is available, we shall construct finite difference methods for this problem, as studying this problem can give crucial insights on the design of computational methods for hyperbolic problems.

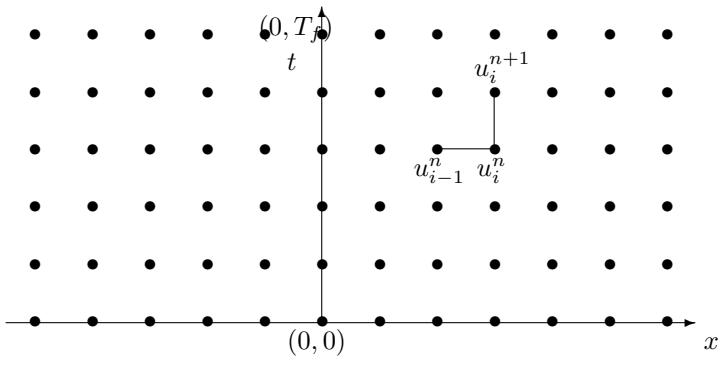


Figure 6.1: The finite difference method (6.4).

6.1 The CFL condition

To make matters simple, suppose that $(t, x) \in [0, T_f] \times \mathbb{R}$, i.e., we do *not* include any boundary conditions. When designing a finite difference method, the first step is to decide upon a finite number of points $(t_n, x_i) \in [0, T_f] \times \mathbb{R}$, with $t_n = n\tau$ for $n = 0, 1, \ldots, N_t$, where $\tau = 1/N_t$ and $x_i = ih$, $i = 0, \pm 1, \pm 2, \ldots$, for some $h \in \mathbb{R}$. On this (infinite in the x-direction) grid, we shall be seeking approximations u_i^n to the exact values $u(t_n, x_i)$.

We formally make the following approximations, using forward difference for the time-derivative and backward difference for the space-derivative

$$u_t(t, x) \approx \delta^t_{\tau,+} u(t, x) = \frac{u(t + \tau, x) - u(t, x)}{\tau},$$

and

$$u_x(t, x) \approx \delta^x_{h,-} u(t, x) = \frac{u(t, x) - u(t, x - h)}{h},$$

where we have adopted the notational convention that the superscript t, denotes that the divided difference operator acts on the time variable t and correspondingly for x, giving

$$\frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} + \frac{u(t_n, x_i) - u(t_n, x_{i-1})}{h} \approx u_t(t_n, x_i) + u_x(t_n, x_i) = 0, \tag{6.3}$$

using the equation (6.1).

Motivated by this formal reasoning, we consider the following system of equations:

$$\frac{u_i^{n+1} - u_i^n}{\tau} + \frac{u_i^n - u_{i-1}^n}{h} = 0,$$

which becomes, after multiplication by τ and rearrangement

$$u_i^{n+1} = (1 - \nu) u_i^n + \nu u_{i-1}^n, \quad \text{for } n = 0, \ldots, N_t - 1,\ i = 0, \pm 1, \pm 2, \ldots, \tag{6.4}$$

where we have used the notation $\nu := \tau/h$ and we shall refer to this quantity as the *Courant number* for this problem. The values of the solution u at initial time can be found from the initial condition (6.2), giving

$$u_i^0 = u_0(x_i) \quad i = 0, \pm 1, \pm 2, \ldots. \tag{6.5}$$

The schematic representation of the method (6.4) is shown in Figure 6.1.

Hence, the value of the approximation u_i^{n+1} depends only on the values u_i^n and u_{i-1}^n, which in turn they depend only on the values u_i^{n-1}, u_{i-1}^{n-1} and u_{i-2}^{n-1}, and so on. If we continue this down to $n = 0$, we conclude that the value of the approximation u_i^{n+1} depends on the initial values u_j^0 for $j = i - n, i - n + 1, \ldots, i$.

In Figure 6.2, we have drawn two line segments: one starting from u_i^{n+1} and finishing at u_{i-n}^0 and one starting from u_i^{n+1} and finishing at u_i^0. The values of the approximation at the nodes located in and inside

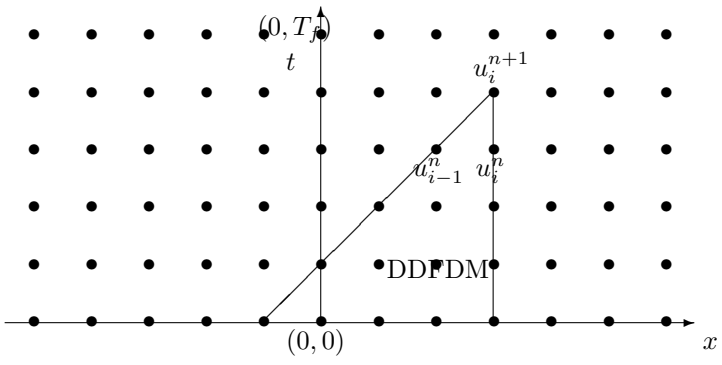


Figure 6.2: The domain of dependence of the finite difference method (DDFDM).

the area between the two line segments are the values that played a role in calculating u_i^{n+1}. This area is known as the *domain of dependence of the finite difference method*. This in turn means that the value of the approximation u_i^{n+1} depends only on the values of the initial condition $u_0(x)$ when x is in the interval $[x_{i-n}, x_i]$!

Let us know consider the characteristic curve running through the point (t_{n+1}, x_i); we saw before that the value of the exact solution $u(t_{n+1}, x_i)$ will be equal to $u_0(x_i - t_{n+1})$; the characteristic curve running through the point (t_{n+1}, x_i) will be referred to as the *domain of dependence of the problem*.

To ensure that the information is "carried over" correctly by our finite difference method from the initial condition to the point u_i^{n+1}, we should make sure that the characteristic curve lies inside the domain of dependence of the finite difference method. In other words: *the domain of dependence of the problem must lie within the domain of dependence of the finite difference method*; this is the infamous *CFL condition*, named after its inventors Courant, Friedrichs and Lewy. Notice that if the CFL condition holds then we have $u(t_{n+1}, x_i) = u_0(x_i - t_{n+1}) \in [x_{i-n}, x_i]$, i.e., the value of the initial condition that determines the value of the PDE at the point (t_{n+1}, x_i) is included in the interval $[x_{i-n}, x_i]$; the value of the initial condition in the interval $[x_{i-n}, x_i]$ determines the value of the approximation u_i^{n+1}!

Recall that the slope of the all the characteristic curves $t = x + const$ is equal to 1, including the slope of the (interesting for us) characteristic curve passing through the point (t_{n+1}, x_i). On the other hand, notice that the slope of the "left" line segment in Figure 6.2 is equal to $\nu = \tau/h$. Assume for the moment that $\nu \leq 1$. Hence, if $\nu \leq 1$, the characteristic curve of interest will pass between the two line segments described above. Therefore, the domain of dependence of the problem lies within the domain of dependence of the finite difference method, and the CFL condition is satisfied. Notice that this is *not* the case when $\nu > 1$.

The CFL condition is an easy method of checking if a finite difference method we design has any chance of being "good', i.e., stable; this is done just be comparing the the domains of dependence of problem and of the finite difference method. Unfortunately, it is only a *necessary* condition for stability, as the following example shows.

On the same semi-infinite grid as above, we formally make the following approximations, using forward difference for the time-derivative and first central difference for the space-derivative

$$u_t(t, x) \approx \delta^t_{\tau,+} u(t, x) = \frac{u(t + \tau, x) - u(t, x)}{\tau},$$

and

$$u_x(t, x) \approx \delta^x_{2h} u(t, x) = \frac{u(t, x + h) - u(t, x - h)}{2h},$$

giving

$$\frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} + \frac{u(t_n, x_{i+1}) - u(t_n, x_{i-1})}{2h} \approx u_t(t_n, x_i) + u_x(t_n, x_i) = 0, \tag{6.6}$$

using the equation (6.1). Motivated by this formal reasoning, we consider the following system of equations:

$$\frac{u_i^{n+1} - u_i^n}{\tau} + \frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0,$$

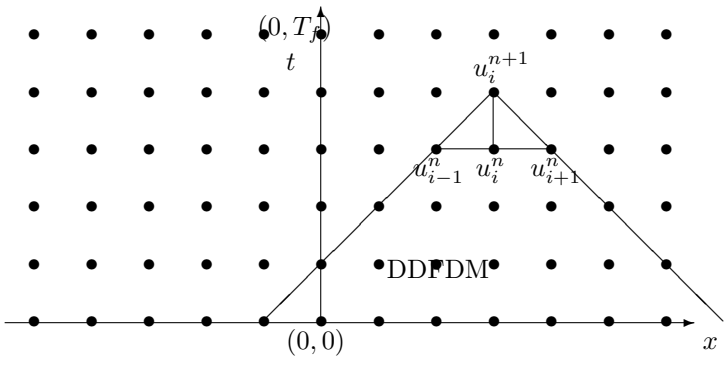


Figure 6.3: The finite difference method (6.7).

which becomes, after multiplication by τ and rearrangement

$$u_i^{n+1} = \frac{\nu}{2} u_{i-1}^n + u_i^n - \frac{\nu}{2} u_{i+1}^n, \quad \text{for } n = 0, \ldots, N_t - 1,\ i = 0, \pm 1, \pm 2, \ldots, \tag{6.7}$$

where, again, we have used the notation $\nu := \tau/h$. The values of the solution u at initial time can be found as in (6.5). The schematic representation of the method (6.7) is shown in Figure 6.3.

The CFL condition for this problem is satisfied for $\nu \leq 1$, as the domain of dependence of this finite difference method contains the domain of dependence of the finite difference method (6.4). To check the stability of this method, we set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the approximate solution at the node (t_n, x_i). We now use (6.7) to evolve one time-step; then we get

$$\lambda^{n+1} \mathrm{e}^{\iota k i h} = \frac{\nu}{2} \lambda^n \mathrm{e}^{\iota k (i-1) h} + \lambda^n \mathrm{e}^{\iota k i h} - \frac{\nu}{2} \lambda^n \mathrm{e}^{\iota k (i+1) h}.$$

Dividing the last equation by $\lambda^n \mathrm{e}^{\iota k i h}$, we deduce

$$\lambda = \frac{\nu}{2} \mathrm{e}^{-\iota k h} + 1 - \frac{\nu}{2} \mathrm{e}^{\iota k h} = 1 - \nu \iota \sin(kh).$$

Hence

$$|\lambda| = \sqrt{1^2 + (-\nu \sin(kh))^2} = \sqrt{1 + \nu^2 \sin^2(kh)},$$

which implies that $|\lambda| > 1$ for all values of $kh \neq m\pi$, for some integer m, i.e., the method is always unstable!

Therefore, we conclude that, even if the CFL condition is satisfied, we have no guarantee for stability. On the other hand, if the CFL condition is not satisfied, we have no hope for stability!

Problem

24. Construct a stable finite difference method that satisfies the CFL condition for the problem

$$\begin{aligned} u_t - u_x &= 0 \quad \text{for all } t \in [0, T_f] \text{ and } x \in [0, 1], \\ u(0, x) &= u_0(x), \qquad \text{for all } x \in [0, 1], \\ u(t, 1) &= 0, \qquad\quad \text{for all } t \in [0, T_f], \end{aligned}$$

where $u_0 : [0, 1] \to \mathbb{R}$ is a known function.

6.2 The upwind method

Let us now consider the general advection Cauchy problem

$$u_t + au_x = 0, \tag{6.8}$$

equipped with initial condition

$$u(0,x) = u_0(x), \tag{6.9}$$

for some $a \equiv a(t,x) : [0,T_f] \times \mathbb{R} \to \mathbb{R}$, with $a \neq 0$, and for some $u_0 : \mathbb{R} \to \mathbb{R}$ known functions. The coefficient a is sometimes referred to as the *wind* or the *drift*. We have seen in Section 1.5 that this Cauchy problem is well posed, and as before it is possible to find the exact solution to this problem using the method of characteristics, which is given by

$$u(t,x) = u_0(x - at),$$

i.e., the solution is constant along each curve $x - at = const$. Even though the exact solution is available, we shall construct finite difference methods for this problem.

For simplicity, suppose that $(t,x) \in [0,T_f] \times \mathbb{R}$, i.e., we do *not* include any boundary conditions. When designing a finite difference method, the first step is to decide upon a finite number of points $(t_n, x_i) \in [0,T_f] \times \mathbb{R}$, with $t_n = n\tau$ for $n = 0,1,\ldots,N_t$, where $\tau = 1/N_t$ and $x_i = ih$, $i = 0, \pm 1, \pm 2, \ldots$, for some $h \in \mathbb{R}$. On this (infinite in the x-direction) grid, we shall be seeking approximations u_i^n to the exact values $u(t_n, x_i)$.

In the previous section, we saw that, for a finite difference method the CFL condition is necessary for stability. We also verified that the finite difference method defined in (6.4) for the case $a(t,x) = 1$, satisfies the CFL condition; this is due to the choice we made of using the first backward difference to approximate u_x. Indeed, if we had used the first forward difference instead, the CFL condition would *not* be satisfied, as in this case the characteristic curve going through the grid point under consideration will have slope equal to $1/a$ which is positive for positive a; thus, the domain of dependence of the finite difference method with forward difference for u_x would only include lines of non-positive slope going though the grid point under consideration, which violates the CFL condition. Similarly if $a < 0$, we should use the forward difference to approximate u_x, for the CFL condition to be satisfied.

Hence, the slopes of the characteristic curves dictate the choice of the type of first divided difference that should be used to approximate u_x. We therefore, formally make the following approximations:

$$u_t(t,x) \approx \delta^t_{h,+} u(t,x), \quad \text{and} \quad u_x(t,x) \approx \begin{cases} \delta^x_{\tau,-} u(t,x), & \text{if } a > 0; \\ \delta^x_{\tau,+} u(t,x), & \text{if } a < 0. \end{cases}$$

giving

$$\frac{u(t_{n+1},x_i) - u(t_n,x_i)}{\tau} + a(t_n,x_i)\frac{u(t_n,x_i) - u(t_n,x_{i-1})}{h} \approx u_t(t_n,x_i) + u_x(t_n,x_i) = 0, \tag{6.10}$$

if $a > 0$ and

$$\frac{u(t_{n+1},x_i) - u(t_n,x_i)}{\tau} + a(t_n,x_i)\frac{u(t_n,x_{i+1}) - u(t_n,x_i)}{h} \approx u_t(t_n,x_i) + u_x(t_n,x_i) = 0, \tag{6.11}$$

if $a < 0$, using the equation (6.8).

Motivated by this formal reasoning, we consider the following system of equations:

$$\frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_i^n - u_{i-1}^n}{h} = 0,$$

with $a_i^n := a(t_n, x_i)$, if $a > 0$, and

$$\frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_i^n}{h} = 0,$$

if $a < 0$. After multiplication by τ and rearrangement, we can write the above in the more compact form

$$u_i^{n+1} = \begin{cases} (1 - a_i^n\nu)u_i^n + a_i^n\nu u_{i-1}^n, & \text{if } a > 0; \\ (1 + a_i^n\nu)u_i^n - a_i^n\nu u_{i+1}^n, & \text{if } a < 0, \end{cases} \tag{6.12}$$

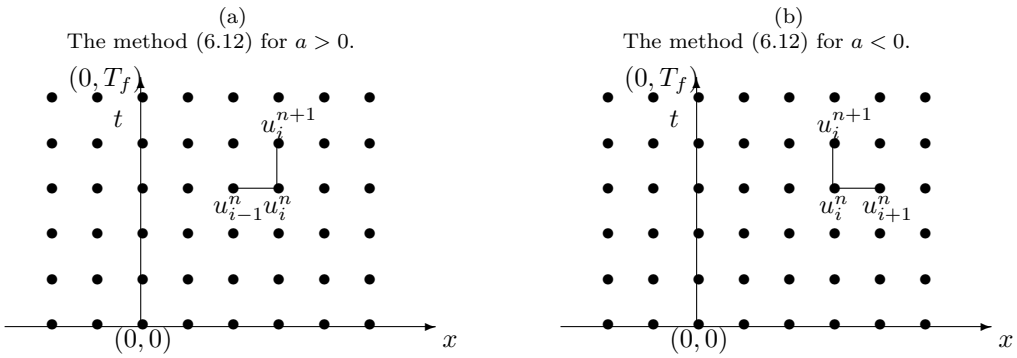


Figure 6.4: The finite difference method (6.12).

for $n = 0, \ldots, N_t - 1$, $i = 0, \pm 1, \pm 2, \ldots$, where we have used the notation $\nu := \tau/h$ for the Courant number for this problem. The values of the solution u at initial time can be found from the initial condition (6.2), giving

$$u_i^0 = u_0(x_i) \quad i = 0, \pm 1, \pm 2, \ldots. \tag{6.13}$$

We shall refer to the finite difference method (6.12), (6.13) as the *upwind method*, the name stemming from the fact that the method follows the direction of the wind a. The schematic representation of the method (6.12) is shown in Figure 6.4.

Of course, to be able to use the method on a computer, we have to consider only a bounded interval for the x-variable to reside in (so that we end up with a finite number of nodes in our grid). Defining boundary conditions for the advection Cauchy problem is somewhat different to the cases of second order elliptic and parabolic problems encountered in the previous chapters: the location of the boundary conditions for the advection Cauchy problem depend on the direction of the characteristic curves.

More specifically, suppose for the moment that the coefficient a is constant and positive; then we saw that the characteristic curves have slope $1/a > 0$. We want to compute an approximation to the solution of the problem (6.8), (6.9), for $0 \leq x \leq 1$. A natural question that arises is: do we need boundary conditions along the lines $x = 0$ and $x = 1$? To answer this, we resort (as usual) to the characteristic curves, which are drawn in Figure 6.7 in the case of constant a for $a > 0$ and $a < 0$, respectively.

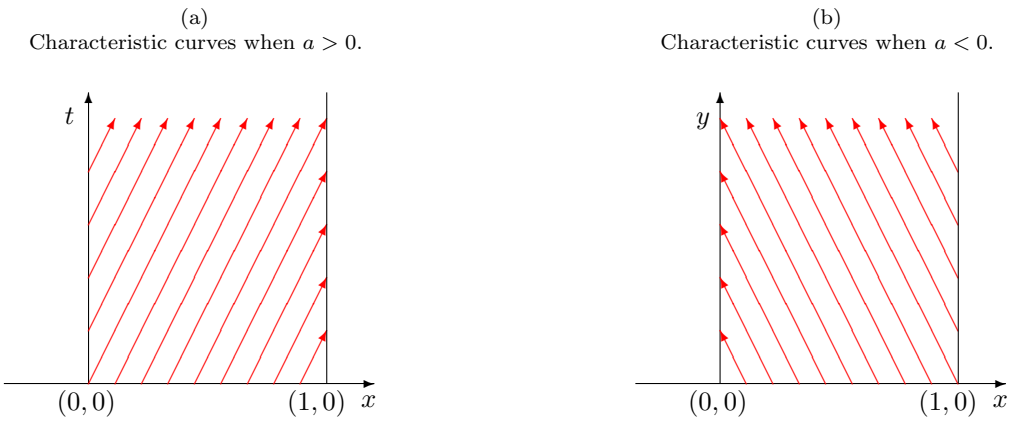


Figure 6.5: Characteristic curves for the advection equation for constant wind.

As the values of the solution "travel" along characteristic curves, we must supply with initial and boundary conditions at the parts of the boundary where the characteristic curves "start". Therefore, in the case of $a > 0$, we should supply with boundary condition along the line $x = 0$ *only*, as the characteristic curves "exit" across the line $x = 1$. Similarly, when $a < 0$, we should supply with boundary condition along the line $x = 1$ *only*, as the characteristic curves "exit" across the line $x = 0$. The part of the boundary where we have to supply boundary conditions is often called the *inflow boundary* and the part of the boundary that do not enforce any boundary conditions is often called the *outflow boundary*.

For simplicity we considered $x \in [0,1]$, the case $x \in [a,b]$ can be treated completely analogously. The initial/boundary value problem for $a > 0$ becomes: find $u : [0,T_f] \times [0,1] \to \mathbb{R}$, such that

$$\begin{aligned} u_t + a u_x &= 0, \quad \text{for } 0 < t \le T_f,\ 0 \le x \le 1, \\ u(0,x) &= u_0(x), \quad \text{for } 0 \le x \le 1, \\ u(t,0) &= u_1(t), \quad \text{for } 0 \le t \le T_f, \end{aligned} \tag{6.14}$$

for some $u_1 : [0,T_f] \to \mathbb{R}$ known function; for the corresponding initial/boundary value problem when $a < 0$, we replace the last equation in (6.14) with $u(t,1) = u_1(t)$, for $0 \le t \le T_f$.

Now, we shall include the boundary conditions in the upwind method. We consider the grid $(t_n, x_i) \in \mathbb{R} \times [0,T_f]$, with $t_n = n\tau$ for $n = 0,1,\ldots,N_t$, where $\tau = 1/N_t$ and $x_i = ih$, $i = 0,1,\ldots,N_x$, for some $h = 1/N_x$. Then the upwind method reads:

$$\begin{aligned} u_i^0 &= u_0(x_i), \quad i = 0,1,\ldots,N_x \\ \left.\begin{array}{ll} u_0^n, & \text{if } a > 0 \\ u_{N_x}^n, & \text{if } a < 0 \end{array}\right\} &= u_1(t_n), \quad n = 1,\ldots,N_t, \\ u_i^{n+1} &= \begin{cases} (1 - a_i^n \nu) u_i^n + a_i^n \nu u_{i-1}^n, & \text{if } a > 0 \text{ for } i = 1,\ldots,N_x; \\ (1 + a_i^n \nu) u_i^n - a_i^n \nu u_{i+1}^n, & \text{if } a < 0 \text{ for } i = 0,\ldots,N_x - 1, \end{cases} \quad n = 0,\ldots,N_t - 1. \end{aligned} \tag{6.15}$$

Example 6.1 *We shall use the upwind method (6.15) to approximate the solution to the advection initial/boundary value problem*

$$\begin{aligned} u_t + 2u_x &= 0, \quad \textit{for} \quad 0 \le t \le 1,\ 0 \le x \le 1 \\ u(0,x) &= u_0(x) := \begin{cases} 10^4 (0.1 - x)^2 (0.2 - x)^2, & \textit{if } 0.1 < x < 0.2; \\ 0, & \textit{otherwise.} \end{cases} \\ u(t,0) &= 0, \quad \textit{for} \quad 0 \le t \le 1. \end{aligned}$$

After implementing the upwind method for this problem (noting that here $a = 2 > 0$), the approximate solution, together with the exact solution, are shown in Figure 6.6. The exact solution can be found by the method of characteristics to be $u(t,x) = u_0(x - 2t)$ and is drawn as is the black lines. The finite difference approximations at various times using the upwind method are drawn as the blue lines. The left column of plots was computed using $N_x = 100$ and $N_t = 250$, resulting to $\nu = 0.4$ and the right column of plots was computed using $N_x = 100$ and and $N_t = 200$, resulting to $\nu = 0.5$.

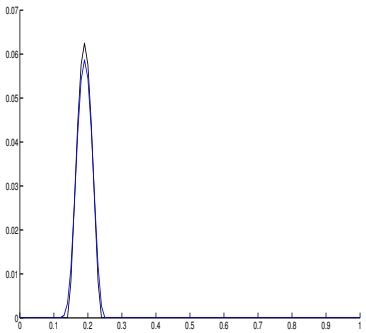
Going back to the example in Figure 6.6, we observe that on the right column the blue and the black lines are overlapping, indicating that the upwind method is extremely accurate for $\nu = 0.5$; in fact, as we shall explain below, the upwind method in this case gives the *exact* solution! This might appear somewhat surprising in the first instance, as for the upwind method on the left column we used more grid points in the t-direction than for the computation shown on the right column; therefore, we should expected that the results on the left column should have been more accurate.

The explanation to this phenomenon can be traced into the upwind method itself. When $\nu = 1/2$ in the above example, where $a = a_i^n = 2$, giving $a_i^n \nu = 1$, the upwind method becomes

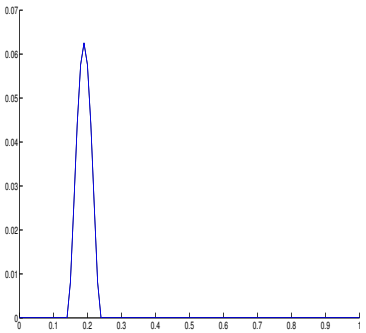
$$u_i^{n+1} = (1 - a_i^n \nu) u_i^n + a_i^n \nu u_{i-1}^n = u_{i-1}^n,$$

i.e., the value of u_i^{n+1} only depends on u_{i-1}^n! Observing now that the characteristic curves for this problem are straight lines with slope $1/a = 1/2$, and noting that the slope of the line segment connecting u_{i-1}^n with u_i^{n+1} has slope $\nu = 1/2$ also, we conclude that the characteristic curve that passes through the point u_i^{n+1}, passes through the point u_{i-1}^n also, and therefore, the upwind method which in this case is $u_i^{n+1} = u_{i-1}^n$, is imitating *exactly* the PDE, which is constant along each characteristic curve! Hence, in this special case where $a_i^n \nu = 1$, the upwind method is exact. Of course, the above observation is of very limited value, as in general may be not easy or desirable to ensure that $a_i^n \nu = 1$ (e.g., in the case where a is not constant, e.t.c.).

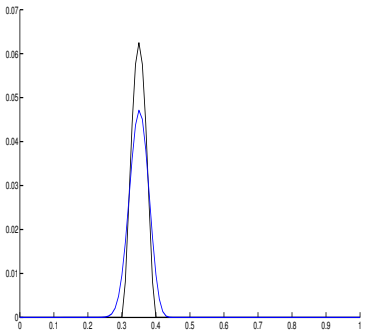
Finally we observe that the solution for the case $\nu = 0.4$ becomes worse and worse after each time-step, at least in the sense that the maximum value of the exact solution minus the one of the finite difference approximation becomes bigger after each time-step. Indeed, we also notice that the finite difference approximation is non-zero in a bigger region after each time-step, which is not the case for the exact solution, which only transports after each time-step. This phenomenon is often referred to as *numerical dissipation*.



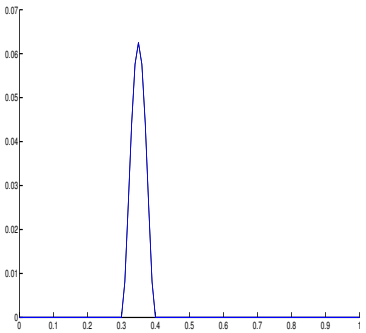
(a) $t = 0.02$



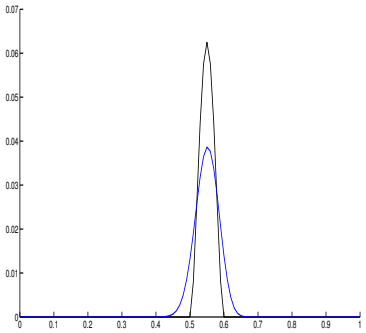
(b) $t = 0.02$



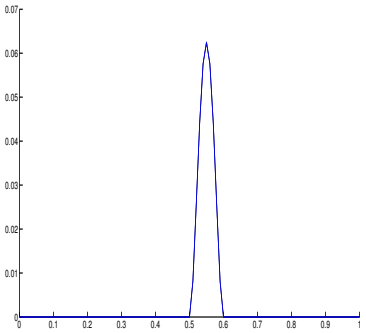
(c) $t = 0.1$



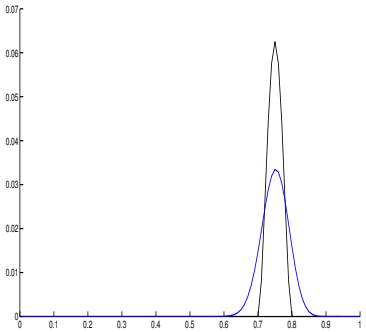
(d) $t = 0.1$



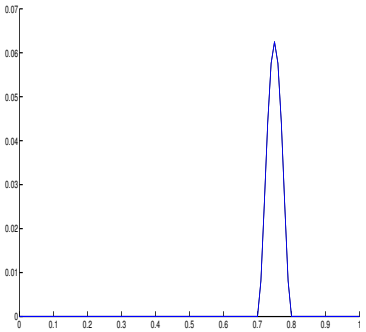
(e) $t = 0.2$



(f) $t = 0.2$



(g) $t = 0.3$



(h) $t = 0.3$

Figure 6.6: The exact solution is the black lines and the finite difference approximation using the upwind method is the blue lines. The left column of plots was computed using $N_x = 100$ and $N_t = 250$, resulting to $\nu = 0.4$ and the right column of plots was computed using $N_x = 100$ and and $N_t = 200$, resulting to $\nu = 0.5$.

In Example 6.1, we witnessed that the upwind method suffers from excessive numerical dissipation. However, we also saw that in the special case where design the grid so that $|a_i^n|\nu = 1$ for all grid points, then the upwind method is exact. We just mention in passing that this exactness of the upwind method for $|a_i^n|\nu = 1$ is very hard or in some instances impossible to be satisfied in the case of nonlinear problems, which are, after all the problems of interest. Therefore, in what follows, we shall not give further consideration to this special case.

Next, we attempt to shed some light on why the upwind method appears to be stable, at least from some values of ν, and the method (6.7) is always unstable, despite both of them satisfying the CFL condition. We start by observing that the upwind method can be rewritten as follows: for $a > 0$, we have

$$\begin{aligned} 0 &= \frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_i^n - u_{i-1}^n}{h} = \frac{u_i^{n+1} - u_i^n}{\tau} + \frac{a_i^n}{2h}\big(2u_i^n - u_{i-1}^n - u_{i-1}^n - u_{i+1}^n + u_{i+1}^n\big) \\ &= \frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_{i-1}^n}{2h} - \frac{a_i^n h}{2} \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}, \end{aligned}$$

and, similarly, for $a < 0$, we have

$$\begin{aligned} 0 &= \frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_i^n}{h} = \frac{u_i^{n+1} - u_i^n}{\tau} + \frac{a_i^n}{2h}\big(u_{i+1}^n + u_{i+1}^n - 2u_i^n - u_{i-1}^n + u_{i-1}^n\big) \\ &= \frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_{i-1}^n}{2h} + \frac{a_i^n h}{2} \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}. \end{aligned}$$

Noting that $|a_i^n| = a_i^n$ if $a > 0$ and $-|a_i^n| = a_i^n$ if $a < 0$, the above two formulas can be combined to give

$$0 = \frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_{i-1}^n}{2h} - \frac{|a_i^n| h}{2} \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}, \tag{6.16}$$

or

$$\frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_{i-1}^n}{2h} = \frac{|a_i^n| h}{2} \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}.$$

On the other hand, if we consider the generalisation of the method (6.7) for the general advection problem, this can be written as

$$\frac{u_i^{n+1} - u_i^n}{\tau} + a_i^n \frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0.$$

Hence, the upwind method can be viewed as the unstable method (6.7) equipped with an additional "stabilisation" term which mimics the second derivative in space, i.e.,

$$\frac{|a_i^n| h}{2} \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} \approx \frac{|a| h}{2} u_{xx},$$

i.e., we add some "numerical diffusion". In other words, the upwind method from (6.16) can be thought as a finite difference method for the parabolic PDE

$$u_t + a u_x - \frac{|a| h}{2} u_{xx} = 0,$$

where we note that as $h \to 0$, the PDE degenerates to the original advection equation $u_t + a u_x = 0$!

Therefore, the additional "numerical diffusion" added to the unstable method (6.7) seems to work well in stabilising the upwind method. On the other hand, however, at least a part of the additional "numerical diffusion" seems to result to numerical dissipation for the upwind method, as observed in Example (6.1). This undoubtedly affects the accuracy of the approximation for the upwind method. In Section 6.3, we shall investigate a method that produces substantially less numerical dissipation, as it is constructed in a way of "adding the right amount of numerical diffusion".

6.2.1 Error analysis

To analyse the error of approximation for the upwind method for $a > 0$, we define the truncation error by

$$T_i^n := \frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} + a(t_n, x_i) \frac{u(t_n, x_i) - u(t_n, x_{i-1})}{h}, \tag{6.17}$$

for $n = 0, \ldots, N_t - 1$, $i = 1, \ldots, N_x$, and for $a < 0$, we have

$$T_i^n := \frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} + a(t_n, x_i)\frac{u(t_n, x_{i+1}) - u(t_n, x_i)}{h},$$

for $n = 0, \ldots, N_t - 1$, $i = 0, \ldots, N_x - 1$. The following lemma describes how well the finite difference method approximates the original problem.

Lemma 6.2 *For the upwind method defined above, we have*

$$|T_i^n| \leq \frac{1}{2}\big(\tau M_{tt} + hAM_{xx}\big), \tag{6.18}$$

for all $n = 0, \ldots, N_t - 1$, $i = 1, \ldots, N_x$, where

$$M_{tt} := \max|u_{tt}(t,x)|, \quad \textit{and} \quad M_{xx} := \max|u_{xx}(t,x)|, \quad \textit{and} \quad A := \max|a(t,x)|,$$

and the maxima are taken over all $(t, x) \in [0, T_f] \times [a, b]$, which are assumed to be finite.

Proof. The proof is left as a exercise. □

For brevity, we shall use the short-hand notation: $\mathcal{T} := 1/2(\tau M_{tt} + hM_{xx})$; with this notation, (6.18) can be written as $|T_i^n| \leq \mathcal{T}$. The next theorem describes the error behaviour of the upwind method.

Theorem 6.3 *Consider the upwind method defined defined by the system (6.15) above. Let u be the exact solution of the initial/boundary value problem (6.14). Assume that the Courant number satisfies $0 \leq |a_i^n|\nu \leq 1$. Then we have the following error bound:*

$$\max_{1 \leq i \leq N_x} |u(t_n, x_i) - u_i^n| \leq \frac{T_f}{2}(\tau M_{tt} + hAM_{xx}), \tag{6.19}$$

for $n = 1, \ldots, N_t$.

Proof. The proof is left as a exercise. □

The above theorem tells us that the approximations u_i^n converge to the exact values of the solution $u(t_n, x_i)$ with first order with respect to both the time-step τ and to the space-step h, provided that the Courant number $0 \leq \nu \leq 1/|a_i^n|$.

6.2.2 Stability analysis

Let us now examine the stability properties of the upwind method. Consider the case $a > 0$. Ignoring the boundary condition, we consider a grid of the form (t_n, x_i), with $t_n = n\tau$ and $x_i = ih$ for $n = 0, \ldots N_t$, $i = 0, \pm 1, \pm 2, \ldots$, with $\tau = 1/N_t$ and $h \in \mathbb{R}$. We set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the approximate solution at the node (t_n, x_i). We use (6.4) to evolve one time-step; then we get

$$\lambda^{n+1}\mathrm{e}^{\iota k i h} = (1 - a_i^n\nu)\lambda^n \mathrm{e}^{\iota k i h} + a_i^n \nu \lambda^n \mathrm{e}^{\iota k (i-1) h}.$$

Dividing the last equation by $\lambda^n \mathrm{e}^{\iota k i h}$, we deduce

$$\lambda = 1 - a_i^n\nu + a_i^n\nu \mathrm{e}^{-\iota k h} = 1 - a_i^n\nu + a_i^n\nu\big(\cos(-kh) + \iota \sin(-kh)\big) = \big(1 - a_i^n\nu + a_i^n\nu\cos(kh)\big) - \iota\big(a_i^n\nu\sin(kh)\big).$$

Hence

$$\begin{aligned}
|\lambda|^2 &= \big(1 - a_i^n\nu + a_i^n\nu\cos(kh)\big)^2 + \big(-a_i^n\nu\sin(kh)\big)^2 \\
&= (1 - a_i^n\nu)^2 + 2(1 - a_i^n\nu)a_i^n\nu\cos(kh) + (a_i^n)^2\nu^2\cos^2(kh) + (a_i^n)^2\nu^2\sin^2(kh) \\
&= (1 - a_i^n\nu)^2 + 2(1 - a_i^n\nu)a_i^n\nu\cos(kh) + (a_i^n)^2 \\
&= 1 - 2a_i^n\nu + 2(a_i^n\nu)^2 + 2(1 - a_i^n\nu)a_i^n\nu\cos(kh) \\
&= 1 - 2a_i^n\nu(1 - a_i^n\nu)(1 - \cos(kh)).
\end{aligned}$$

Clearly $|\lambda| \leq 1$ if and only if $|\lambda|^2 \leq 1$, which is true if and only if the second term on the right-hand side above is non-negative, that is if and only if $1 - a_i^n \nu \geq 0$, as $a_i^n > 0$ in this case. Hence, the method is stable if and only if $a_i^n \nu \leq 1$, for $a > 0$. If $a < 0$, we have similarly, using the method (6.4) to evolve one time-step:

$$\lambda^{n+1} \mathrm{e}^{\iota k i h} = (1 + a_i^n \nu) \lambda^n \mathrm{e}^{\iota k i h} - a_i^n \nu \lambda^n \mathrm{e}^{\iota k (i+1) h}.$$

Dividing the last equation by $\lambda^n \mathrm{e}^{\iota k i h}$, we deduce

$$\lambda = 1 + a_i^n \nu - a_i^n \nu \mathrm{e}^{\iota k h} = 1 + a_i^n \nu - a_i^n \nu \big(\cos(kh) + \iota \sin(kh) \big) = \big(1 + a_i^n \nu - a_i^n \nu \cos(kh) \big) - \iota \big(a_i^n \nu \sin(kh) \big).$$

Hence

$$\begin{aligned}
|\lambda|^2 &= \big(1 + a_i^n \nu - a_i^n \nu \cos(kh) \big)^2 + \big(- a_i^n \nu \sin(kh) \big)^2 \\
&= (1 + a_i^n \nu)^2 - 2(1 + a_i^n \nu) a_i^n \nu \cos(kh) + (a_i^n)^2 \nu^2 \cos^2(kh) + (a_i^n)^2 \nu^2 \sin^2(kh) \\
&= (1 + a_i^n \nu)^2 - 2(1 + a_i^n \nu) a_i^n \nu \cos(kh) + (a_i^n)^2 \\
&= 1 + 2 a_i^n \nu + 2 (a_i^n \nu)^2 - 2(1 + a_i^n \nu) a_i^n \nu \cos(kh) \\
&= 1 + 2 a_i^n \nu (1 + a_i^n \nu)(1 - \cos(kh)).
\end{aligned}$$

Thus, we have $|\lambda|^2 \leq 1$ if and only if the second term on the right-hand side above is negative, that is if and only if $1 + a_i^n \nu \geq 0$, as $a_i^n < 0$ in this case. Hence, the method is stable if and only if $-a_i^n \nu \leq 1$, or $|a_i^n| \nu \leq 1$ for $a < 0$.

We, therefore, conclude that the upwind method is stable if and only if $|a_i^n| \nu \leq 1$, which is also the same requirement for convergence, as shown in the previous section.

Problems

25. Let the PDE

$$u_t + u_x = 0,$$

along with some suitable initial and boundary conditions (which are not relevant to the discussion at this point). We consider the finite difference method

$$u_i^{n+1} = \alpha u_{i-1}^n + \beta u_i^n + \gamma u_{i+1}^n$$

for some $\alpha, \beta, \gamma \in \mathbb{R}$. Determine the coefficients α, β, γ so that the corresponding truncation error is of as high an order as possible.

26. Prove Lemma 6.2.

27. Prove Lemma 6.3.

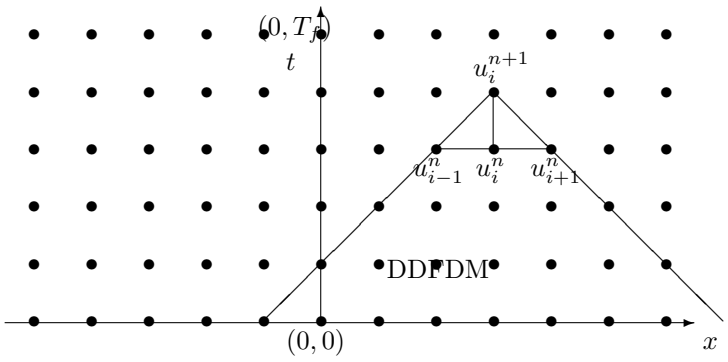


Figure 6.7: The Lax-Wendroff method.

6.3 The Lax-Wendroff method

In Section 6.2 we studied the upwind method for the advection initial/boundary value problem, which is convergent and stable when $|a|\nu \leq 1$ but, nonetheless, suffered from excessive numerical dissipation. We also saw that the upwind method can be viewed as a "stabilised" version of the unstable method (6.7), through the addition of a "numerical diffusion" term.

The amount of "numerical diffusion" added seems to contribute towards the numerical dissipation observed in the upwind method which, in turn, affects the accuracy of the approximation. Here, we shall investigate a method that produces substantially less numerical dissipation, as it is constructed in a way of "adding the right amount of numerical diffusion".

We consider the advection initial/boundary value problem

$$\begin{aligned} u_t + au_x &= 0, \quad \text{for } 0 < t \leq T_f,\ 0 \leq x \leq 1, \\ u(0,x) &= u_0(x), \quad \text{for } 0 \leq x \leq 1, \\ u(t,0) &= u_1(t), \quad \text{for } 0 \leq t \leq T_f, \end{aligned} \tag{6.20}$$

for some $u_1 : [0, T_f] \to \mathbb{R}$ known function; for the corresponding initial/boundary value problem when $a < 0$, we replace the last equation in (6.20) with $u(t, 1) = u_1(t)$, for $0 \leq t \leq T_f$. For simplicity, in this section we shall only consider the case $0 \neq a \in \mathbb{R}$, i.e., when the wind a is a non-zero constant.

We consider the grid $(t_n, x_i) \in \mathbb{R} \times [0, T_f]$, with $t_n = n\tau$ for $n = 0, 1, \ldots, N_t$, where $\tau = 1/N_t$ and $x_i = ih$, $i = 0, 1, \ldots, N_x$, for some $h = 1/N_x$.

Noting that here $a_i^n = a$, as a is constant, we shall study the finite difference method for approximating the solution to the problem (6.20):

$$\frac{u_i^{n+1} - u_i^n}{\tau} + a\frac{u_{i+1}^n - u_{i-1}^n}{2h} - \frac{a^2\tau}{2}\frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} = 0, \tag{6.21}$$

for $n = 0, 1, \ldots, N_t$, $i = 0, 1, \ldots, N_x$, which after multiplication by τ and rearrangement yields

$$u_i^{n+1} = \frac{a\nu}{2}(1 + a\nu)u_{i-1}^n + (1 - a^2\nu^2)u_i^n - \frac{a\nu}{2}(1 - a\nu)u_{i+1}^n, \tag{6.22}$$

where, as per normal $\nu := \tau/h$. The initial and boundary conditions are completely analogous to the case of the upwind method (6.15) and, therefore omitted here[1].This is the infamous *Lax-Wendroff method*. The schematic representation of the method (6.22) is shown in Figure 6.7.

It is easy to see that the CFL condition for this problem is satisfied for $\nu \leq 1/|a|$. To check the stability of this method, we set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

[1]Note however that, unless $1 - a\nu = 0$, the scheme also requires the values $u_{N_x+1}^n$, $n = 1, \ldots, N_t$, which fall outside of the space domain and are not available.

for $k \in \mathbb{R}$, to be the approximate solution at the node (t_n, x_i). We now use (6.22) to evolve one time-step; then we get

$$\lambda^{n+1} e^{\iota k i h} = \frac{a\nu}{2}(1 + a\nu)\lambda^n e^{\iota k(i-1)h} + (1 - a^2\nu^2)\lambda^n e^{\iota k i h} - \frac{a\nu}{2}(1 - a\nu)\lambda^n e^{\iota k(i+1)h}.$$

Dividing the last equation by $\lambda^n e^{\iota k i h}$, we deduce

$$\begin{aligned} \lambda &= \frac{a\nu}{2}(1 + a\nu)e^{-\iota k h} + (1 - a^2\nu^2) - \frac{a\nu}{2}(1 - a\nu)e^{\iota k h} \\ &= 1 + \frac{a^2\nu^2}{2}\big(e^{-\iota k h} - 2 + e^{\iota k h}\big) - \frac{a\nu}{2}\big(e^{\iota k h} - e^{-\iota k h}\big) \\ &= 1 - 2a^2\nu^2 \sin^2(\frac{1}{2}kh) - \iota a\nu \sin(kh). \end{aligned}$$

Hence

$$\begin{aligned} |\lambda|^2 &= \big(1 - 2a^2\nu^2 \sin^2(\frac{1}{2}kh)\big)^2 + \big(-a\nu \sin(kh)\big)^2 \\ &= 1 - 4a^2\nu^2\big(1 - \cos^2(\frac{1}{2}kh)\big)\sin^2(\frac{1}{2}kh) + 4a^4\nu^4 \sin^4(\frac{1}{2}kh) \\ &= 1 - 4a^2\nu^2(1 - a^2\nu^2)\sin^4(\frac{1}{2}kh), \end{aligned}$$

which implies that $|\lambda| \leq 1$ if and only if $|4a^2\nu^2(1 - a^2\nu^2)| \leq 1$, which is true if and only if $|a|\nu \leq 1$, i.e, the Lax-Wendroff method is stable if and only if $\nu \leq 1/|a|$.

Comparing (6.21) with (6.16), we notice that they differ only due to the coefficient of the second divided difference, which in the case of the upwind method is $|a|h/2$, as opposed to $a^2\tau/2$ for the case of the Lax-Wendroff method. Hence the amount of "numerical diffusion" imposed by each method differs. In particular, recalling that for both the upwind and the Lax-Wendroff methods we need $|a|\nu \leq 1$ for stability, we have

$$\frac{|a|h}{2} = \frac{|a|\tau}{2\nu} = \frac{a^2\tau}{2|a|\nu} \geq \frac{a^2\tau}{2},$$

i.e., the amount of "numerical diffusion" added by the Lax-Wendroff method is smaller than the corresponding for the upwind method. To illustrate the comparison of numerical dissipation properties between the upwind method and the Lax-Wendroff method, we use the latter to approximate the solution to the initial/boundary value problem from Example 6.1. The results are shown in Figure 6.8. Comparing these results with the corresponding results from the first column of Figure 6.6, we indeed verify that the Lax-Wendroff method does not suffer from excessive numerical dissipation and appears to more accurate results, compared to the upwind method.

Therefore, the Lax-Wendroff method appears to contain the "right amount" of "numerical diffusion". One may wonder on the specific choice of the coefficient $a^2\tau/2$ present in the Lax-Wendroff method, which governs the amount of "numerical diffusion" added. The particular choice of the value $a^2\tau/2$ is related to the accuracy of the method. Indeed, as we saw in Problem 25, the Lax-Wendroff method has the highest possible order of convergence from all the explicit methods involving the 3 values u_{i-1}^n, u_i^n and u_{i+1}^n! In particular, we have the following bound.

Lemma 6.4 *Let $a \in \mathbb{R}$ with $a \neq 0$. For the Lax-Wendroff method defined above, the truncation error is defined by*

$$T_i^n := \frac{u(t_{n+1}, x_i) - u(t_n, x_i)}{\tau} + a\frac{u(t_n, x_{i+1}) - u(t_n, x_{i-1})}{2h} - \frac{a^2\tau}{2}\frac{u(t_n, x_{i+1}) - 2u(t_n, x_i) + u(t_n, x_{i-1})}{h^2}.$$

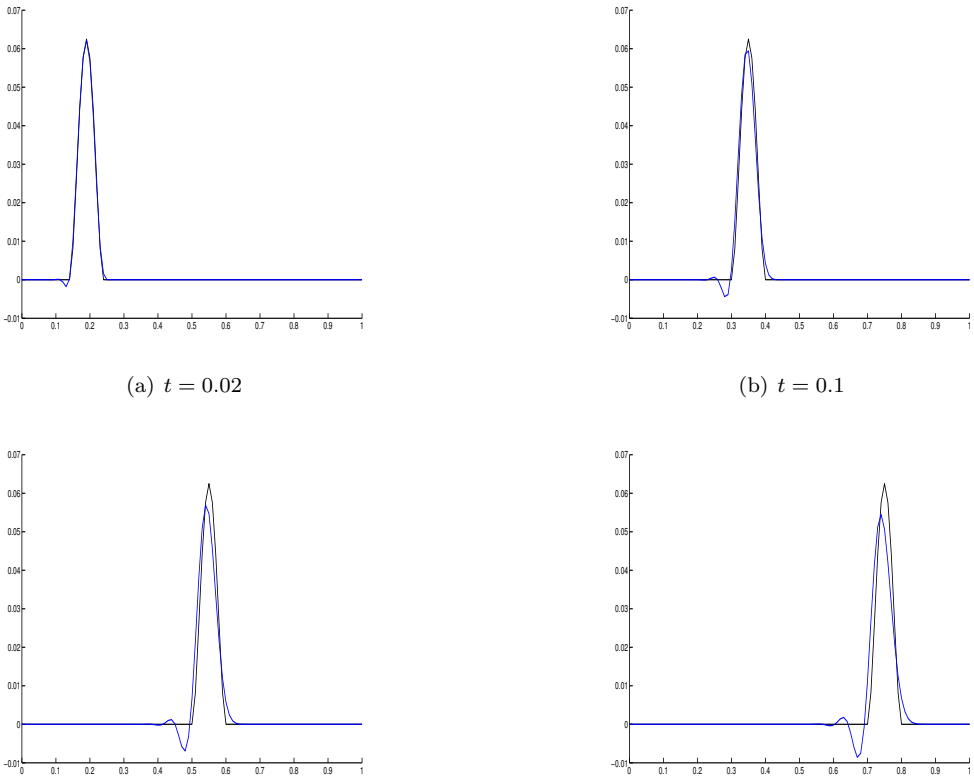
Then, assuming that $|a|\nu \leq 1$, we have

$$|T_i^n| \leq \frac{\tau^2}{6}M_{ttt} + |a|\frac{h^2}{3}M_{xxx}, \tag{6.23}$$

for all $n = 0, \ldots, N_t - 1$, $i = 1, \ldots, N_x$, where

$$M_{ttt} := \max|u_{ttt}(t, x)|, \quad \textit{and} \quad M_{xxx} := \max|u_{xxx}(t, x)|,$$

and the maxima are taken over all $(t, x) \in [0, T_f] \times [0, 1]$, which are assumed to be finite.



(a) $t = 0.02$ (b) $t = 0.1$

(c) $t = 0.2$ (d) $t = 0.3$

Figure 6.8: The exact solution is the black lines and the finite difference approximation using the Lax-Wendroff method is the blue lines, computed using $N_x = 100$ and $N_t = 250$, resulting to $\nu = 0.4$.

Proof. We use Taylor's Theorem, to obtain

$$u(t_{n+1}, x_i) = u(t_n, x_i) + \tau u_t(t_n, x_i) + \frac{\tau^2}{2} u_{tt}(t_n, x_i) + \frac{\tau^3}{6} u_{ttt}(\rho_n, x_i),$$

for some $\rho_n \in (t_n, t_{n+1})$, and

$$\begin{aligned} u(t_n, x_{i+1}) &= u(t_n, x_i) + h u_x(t_n, x_i) + \frac{h^2}{2} u_{xx}(t_n, x_i) + \frac{h^3}{6} u_{xxx}(t_n, \xi_i), \\ u(t_n, x_{i-1}) &= u(t_n, x_i) - h u_x(t_n, x_i) + \frac{h^2}{2} u_{xx}(t_n, x_i) - \frac{h^3}{6} u_{xxx}(t_n, \zeta_i), \end{aligned}$$

for some $\xi_i \in (x_i, x_{i+1})$, $\zeta_i \in (x_{i-1}, x_i)$. Inserting all the above into the truncation error, and using the PDE $u_t + a u_x = 0$, along with the fact $u_{tt} = (u_t)_t = (-a u_x)_t = -a(u_t)_x = a^2 u_{xx}$, we deduce

$$T_i^n = \frac{\tau^2}{6} u_{ttt}(\rho_n, x_i) + \frac{ah}{12}(h - a\tau) u_{xxx}(t_n, \xi_i) + \frac{ah}{12}(h + a\tau) u_{xxx}(t_n, \zeta_i).$$

Now, taking absolute values and the triangle inequality, we arrive to

$$|T_i^n| = \frac{\tau^2}{6} |u_{ttt}(\rho_n, x_i)| + \frac{|a|h}{12} |h - a\tau| \, |u_{xxx}(t_n, \xi_i)| + \frac{|a|h}{12} |h + a\tau| \, |u_{xxx}(t_n, \zeta_i)|,$$

which gives (6.23), by observing that $|a|\nu \leq 1$ implies $|h - a\tau| \leq 2h$ and $|h + a\tau| \leq 2h$. □

Hence the Lax-Wendroff method is second order accurate with respect to both τ and h.

Problem

28. Motivated by the idea of Problem 25, i.e., that the Lax-Wendroff method is the highest order explicit method involving u_{i-1}^n, u_i^n and u_{i+1}^n when a is constant, construct a version of the Lax-Wendroff method for the case where the wind $a = a(t, x)$ is variable.

6.4 Finite difference methods for the 2nd order wave equation: the leap-frog method

Having considered finite difference methods for the advection equation (also known as the one-sided wave equation – see the discussion at the beginning of this chapter), we shall now briefly focus on presenting a finite difference method for the 2nd order wave equation.

To this end, consider the familiar wave initial/boundary value problem: we seek the (unique) solution $u : [0, T_f] \times [0, 1] \to \mathbb{R}$ to the initial/boundary-value problem

$$\begin{aligned} u_{tt}(t,x) &= u_{xx}(t,x) \quad \text{in } (0,T_f] \times [0,1], \\ u(0,x) &= u_0(x), \quad \text{for} \quad 0 \le x \le 1 \\ u_t(0,x) &= v_0(x), \quad \text{for} \quad 0 \le x \le 1 \\ u(t,0) = u(t,1) &= 0, \quad \text{for} \quad 0 < t \le T_f, \end{aligned} \tag{6.24}$$

where $u_0, v_0 : [0, 1] \to \mathbb{R}$ are known functions.

We construct a grid as follows: we consider equally distributed subdivision $x_0 < x_1 < \cdots < x_{N_x+1}$, at distance h between them, such that

$$0 = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_{N_x} = x_{N_x-1} + h,\ x_{N_x+1} = 1,$$

in the space-direction, and an equally distributed subdivision $t_0 < t_1 < \cdots < t_{N_t}$, at distance τ between them, such that

$$0 = t_0,\ t_1 = t_0 + \tau,\ t_2 = t_1 + \tau,\ \ldots,\ t_{N_t-2} = t_{N_t-1} + \tau,\ t_{N_t} = T_f,$$

in the time-direction (see Figure 6.9); hence, we have $h = 1/(N_x + 1)$ and $\tau = T_f/N_t$.

Our aim, as before, is to find approximations u_i^n of the function values $u(t_n, x_i)$, for $n = 1, \ldots, N_t$ and $i = 1, \ldots, N_x$. An idea could be to use second divided differences to approximate both u_{tt} and u_{xx}.

We define the *leap-frog method* by the following system of equations:

$$\frac{u_i^{n+1} - 2u_i^n + u_i^{n-1}}{\tau^2} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2}, \tag{6.25}$$

for $n = 1, \ldots, N_t - 1$ and $i = 1, \ldots, N_x$. Notice that now the value u_i^{n+1} at the new time t^{n+1} depends on the values of the approximations at *two* previous times! (Such a method is called a *two-step method*.)

Also, as per normal, we impose the first initial condition and the boundary conditions by setting:

$$u_i^0 = u_0(x_i) \quad i = 0, 1, \ldots, N_x + 1, \tag{6.26}$$

and

$$u_0^n = 0 = u_{N_x+1}^n \quad n = 1, \ldots, N_t. \tag{6.27}$$

We observe that the leap-frog method (6.25) is still not well defined as, setting $n = 1$ on (6.25), we deduce

$$\frac{u_i^2 - 2u_i^1 + u_i^0}{\tau^2} = \frac{u_{i+1}^1 - 2u_i^1 + u_{i-1}^1}{h^2},$$

that is, the first values we can compute using the leap-frog method is u_i^2's! Moreover, to compute the u_i^2's we must assume knowledge of the u_i^1's, which are currently *not* defined.

Note, however, that we still have not imposed the second initial condition (3rd line of (6.24)). To do so, we shall make use of the fictitious node idea for imposing Neumann boundary conditions, we have previously seen. In that vein, we extend the grid by the (fictitious) nodes (t_{-1}, x_i), $i = 0, 1, \ldots, N_x + 1$ with $t_{-1} := -\tau$, and we consider the respective approximations u_i^{-1}. To impose the second initial condition, we require

$$\frac{u_i^1 - u_i^{-1}}{2\tau} = v_0(x_i) \quad i = 0, 1, \ldots, N_x + 1,$$

i.e., we make use of the first central difference with spacing 2τ of to approximate u_t at $t = 0$. This implies

$$u_i^{-1} = u_i^1 - 2\tau v_0(x_i), \quad i = 0, 1, \ldots, N_x + 1. \tag{6.28}$$

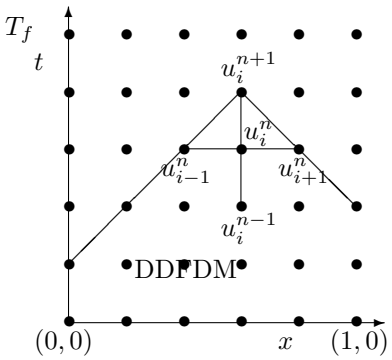


Figure 6.9: The leap-frog method.

Extending also the validity of (6.25) for $n = 0$ (as the u_i^{-1}'s are now defined), we get

$$\frac{u_i^1 - 2u_i^0 + u_i^{-1}}{\tau^2} = \frac{u_{i+1}^0 - 2u_i^0 + u_{i-1}^0}{h^2},$$

which becomes

$$\frac{2u_i^1 - 2u_i^0 - 2\tau v_0(x_i)}{\tau^2} = \frac{u_{i+1}^0 - 2u_i^0 + u_{i-1}^0}{h^2},$$

using (6.28), or, multiplying by $\tau^2/2$ and rearranging,

$$u_i^1 = \frac{\nu^2}{2}u_{i+1}^0 + (1-\nu^2)u_i^0 + \frac{\nu^2}{2}u_{i-1}^0 + \tau v_0(x_i), \tag{6.29}$$

for $\nu = \tau/h$, the Courant number. So to start the leap-frog method (6.25), we first compute the u_i^1's using (6.29). Then, after multiplication by τ^2 and rearrangement, (6.25) gives

$$u_i^{n+1} = \nu^2 u_{i-1}^n + 2(1-\nu^2)u_i^n + \nu^2 u_{i+1}^n - u_i^{n-1}, \quad \text{for } n = 1, \ldots, N_t - 1,\ i = 1, \ldots, N_x, \tag{6.30}$$

where, again $\nu := \tau/h$ is the Courant number for this method. In Figure 6.9, we can see a representation of the leap-frog method.

Notice that when $\nu = 1$, i.e., when $\tau = h$, the leap-frog method becomes

$$u_i^{n+1} = u_{i-1}^n + u_{i+1}^n - u_i^{n-1}, \quad \text{for } n = 1, \ldots, N_t - 1,\ i = 1, \ldots, N_x,$$

i.e., the term u_i^n dissapears, thereby justifying the name "leap-frog"!

The two families of characteristics are given by the equations $x + t = const$ and $x - t = const$, i.e, they are lines of slopes ± 1 passing from the point (t, x). The domain of dependence of the leap-frog method is shown in Figure (6.9). It is evident that the CFL condition for this method then implies $\nu \leq 1$.

We conclude this section by performing stability analysis for the leap-frog method, in order to get a sufficient condition for stability on the Courant number, too.

Ignoring the effect of boundary conditions, we set

$$u_i^n = \lambda^n \mathrm{e}^{\iota k x_i} = \lambda^n \mathrm{e}^{\iota k i h},$$

for $k \in \mathbb{R}$, to be the approximate solution at the node (t_n, x_i). We now use (6.30) to evolve one time-step; then we get

$$\lambda^{n+1}\mathrm{e}^{\iota k i h} = \nu^2\lambda^n\mathrm{e}^{\iota k(i-1)h} + 2(1-\nu^2)\lambda^n\mathrm{e}^{\iota k i h} + \nu^2\lambda^n\mathrm{e}^{\iota k(i+1)h} - \lambda^{n-1}\mathrm{e}^{\iota k i h}.$$

Dividing the last equation by $\lambda^n\mathrm{e}^{\iota k i h}$, we deduce

$$\begin{aligned}
\lambda &= \nu^2\mathrm{e}^{-\iota k h} + 2(1-\nu^2) + \nu^2\mathrm{e}^{\iota k h} - \lambda^{-1} \\
&= 2 + \nu^2\big(\mathrm{e}^{-\iota k h} - 2 + \mathrm{e}^{\iota k h}\big) - \lambda^{-1} \\
&= 2 - 4\nu^2 \sin^2(\frac{1}{2}kh) - \lambda^{-1}.
\end{aligned}$$

Hence, setting $a := 1 - 2\nu^2 \sin^2(\frac{1}{2}kh)$, to simplify the notation, and multiplying by λ, we deduce, upon rearrangement,

$$\lambda^2 - 2a\lambda + 1 = 0.$$

The roots of this quadratic equation are given by $\lambda = a \pm \sqrt{a^2 - 1}$. Since $a \in \mathbb{R}$, it is not too hard to see (see Problem 29) that for $|\lambda| \leq 1$ we must have $|a| \leq 1$. This implies that $\nu \leq 1$, i.e, the leap-frog method is stable if and only if $\nu \leq 1$.

Problem

29. For

$$\lambda^2 - 2a\lambda + 1 = 0,$$

show that $|\lambda| \leq 1$ if and only if $|a| \leq 1$.

Chapter 7

The Finite Element Method for Elliptic Problems

7.1 Introduction

In the previous chapters, we considered finite difference methods for the approximation of solutions to initial and boundary value problems of parabolic, elliptic and hyperbolic type. Finite difference methods are constructed by replacing the partial derivatives of the PDEs by divided differences defined on a grid. So far, we considered PDEs whose domain of definition was a rectangular domain (in, either "(x, y)", or "(t, x)" variables). Of course in practice one may want to solve PDEs on more complicated domains than rectangles, which renders the construction of a grid quite troublesome.

In this chapter, we shall be concerned with the *finite element method* (FEM) for elliptic problems. One of the main advantages of the finite element method compared to the finite difference method is its ability to cope with complicated domains of definition of the underlying PDEs. (There are also other advantages of FEM compared to finite differences, but these topics are outside of the scope of these notes.)

7.2 Weak derivatives

Before embarking with constructing finite element methods, we shall need to understand a bit more various concepts of differentiation. We know from basic Analysis that not every continuous function is differentiable. The aim of this section is to define a new concept of "differentiation" which will allow us to generalise the notion of a derivative of a function. To do so, we shall need to consider first some elementary concepts.

Definition 7.1 *Consider a function $f : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d$ open set. We define the* support *of f to be the closure*[1] *of the set $\{\mathbf{x} \in \Omega : f(\mathbf{x}) \neq 0\}$.*

In other words, the support of a function f is the smallest closed set containing the pre-images of all non-zero values of f.

Example 7.2 *The support of the function $f : \mathbb{R} \to \mathbb{R}$ with*

$$f(x) = \begin{cases} 1 + x, & \text{for } -1 < x < 0; \\ 1 - x, & \text{for } 0 < x < 1; \\ 0, & \text{otherwise,} \end{cases}$$

is the closed interval $[-1, 1]$.

Definition 7.3 *Let $\Omega \subset \mathbb{R}^d$ open set. We denote by $C_0^\infty(\Omega)$, the family of all functions $\phi : \Omega \to \mathbb{R}$ that are infinite times differentiable and have compact support*[2]*.*

[1] *Closure* of a set A in $\mathbb{R}^d$ is the smallest closed set containing A.
[2] A set $B \subset \mathbb{R}^d$ is *compact* if it is bounded and closed.

It is not hard to show that, in fact, $C_0^\infty(\Omega)$ is a (infinite dimensional) vector space, but this is beyond the scope of these notes.

A natural question is whether infinitely differentiable functions $\phi : \Omega \to \mathbb{R}$ with compact support do exist at all, i.e., is $C_0^\infty(\Omega)$ empty or not?

Example 7.4 *Consider the function $f : \mathbb{R} \to \mathbb{R}$ with*

$$f(x) = \begin{cases} \mathrm{e}^{\frac{1}{x^2-1}}, & \text{for } -1 < x < 1; \\ 0, & \text{otherwise.} \end{cases}$$

This function is infinitely differentiable for all $x \neq \pm 1$. At $x = 1$, we have

$$f'_+(1) = \lim_{h\to 0^+} \frac{f(1+h) - f(1)}{h} = \lim_{h\to 0^+} \frac{0-0}{h} = 0,$$

and

$$f'_-(1) = \lim_{h\to 0^+} \frac{f(1) - f(1-h)}{h} = \lim_{h\to 0^+} \frac{0 - \mathrm{e}^{\frac{1}{(1-h)^2-1}}}{h} = \ldots = 0.$$

Hence f is also differentiable at $x = 1$ (and completely analogously for $x = -1$). The case of higher derivatives follows analogously. Notice that this function's support in the closed interval $[-1, 1]$ (which is a bounded and closed set).

Therefore, at least for the case of functions defined on a open set $\Omega \subset \mathbb{R}$ containing the closed interval $[-1, 1]$ the family $C_0^\infty(\Omega)$ is non-trivial.

The infinitely differentiable functions of compact support play a very important role in the modern theory of functions. In particular, they are utilised in generalising the concept of a derivative of a function.

Definition 7.5 *Let $(a, b) \subset \mathbb{R}$ open interval. A function $g : (a, b) \to \mathbb{R}$ is called a* weak derivative *of a function $f : (a, b) \to \mathbb{R}$ if*

$$\int_a^b g(x)\phi(x)\,\mathrm{d}x = -\int_a^b f(x)\phi'(x)\mathrm{d}x \ < +\infty$$

for all *functions $\phi \in C_0^\infty((a, b))$* [3].

Theorem 7.6 *Let $(a, b) \subset \mathbb{R}$ open and $f : (a, b) \to \mathbb{R}$. If f is differentiable in (a, b) then it has a weak derivative g with $g = f'$ almost everywhere*[4].

Proof. Let $\phi \in C_0^\infty((a, b))$. Since ϕ has compact support (i.e., bounded and closed), the endpoints a and b cannot be in the support of ϕ (since the support of ϕ is closed and, therefore, can only be contained strictly in the interval (a, b)). Hence $\phi(a) = \phi(b) = 0$ (or, strictly speaking, $\lim_{x\to a^+} \phi(x) = \lim_{x\to b^-} \phi(x) = 0$). Since f' exists in (a, b), the integration by parts formula implies

$$\int_a^b f'(x)\phi(x)\,\mathrm{d}x = [f(x)\phi(x)]_a^b - \int_a^b f(x)\phi'(x)\mathrm{d}x = -\int_a^b f(x)\phi'(x)\mathrm{d}x,$$

since $\phi(a) = \phi(b) = 0$. Therefore, from Definition 7.5, we have that f' is a weak derivative of f, too. □

The converse of Theorem 7.6 is not true, i.e., there are functions that are not differentiable that have a weak derivative. This somewhat justifies the name: a weak derivative is a "weaker" notion of differentiation that the (classical) derivative.

Example 7.7 *Consider the function $f : \mathbb{R} \to \mathbb{R}$ with*

$$f(x) = \begin{cases} 1 + x, & \text{for } -1 < x \leq 0; \\ 1 - x, & \text{for } 0 < x < 1; \\ 0, & \text{otherwise.} \end{cases}$$

[3] Notice that since the function g only appears under the integral sign, it is strictly-speaking not unique, as changing the value of g at finite number of points it will not change the value of the integral!

[4] *Almost everywhere* here means that g is equal to f' at all points in (a, b) up to a set of Lebesgue measure zero.

This function is not *differentiable at the points* $-1, 0$ *and* 1*. To see if* f *has a weak derivative, we consider* $\phi \in C_0^\infty(\mathbb{R})$*; then*

$$\begin{aligned}
-\int_{-\infty}^{-\infty} f(x)\phi'(x)\,\mathrm{d}x &= -\int_{-1}^{0}(1+x)\phi'(x)\,\mathrm{d}x - \int_{0}^{1}(1-x)\phi'(x)\,\mathrm{d}x \\
&= -[(1+x)\phi(x)]_{-1}^{0} + \int_{-1}^{0}(1+x)'\phi(x)\,\mathrm{d}x - [(1-x)\phi(x)]_{0}^{1} + \int_{0}^{1}(1-x)'\phi(x)\,\mathrm{d}x \\
&= -\phi(0) + \int_{-1}^{0}\phi(x)\,\mathrm{d}x + \phi(0) + \int_{0}^{1}(-1)\phi(x)\,\mathrm{d}x, \\
&= \int_{-1}^{0}\phi(x)\,\mathrm{d}x + \int_{0}^{1}(-1)\phi(x)\,\mathrm{d}x,
\end{aligned}$$

after integrating by parts each term above. Hence, we have

$$-\int_{-\infty}^{-\infty} f(x)\phi'(x)\,\mathrm{d}x = \int_{-\infty}^{-\infty} g(x)\phi(x)\,\mathrm{d}x$$

where

$$g(x) = \begin{cases} 1, & \text{for } -1 < x < 0; \\ -1, & \text{for } 0 < x < 1; \\ 0, & \text{otherwise.} \end{cases}$$

The function g *is a weak derivative of* f*. (Notice that, as before, the value of* g *at a finite number of points and, in particular, at the points* 0 *and* ± 1 *is irrelevant!)*

We conclude this section with some more definitions.

Definition 7.8 *Let* $(a, b) \subset \mathbb{R}$ *open interval. We define the family of functions*

$$L^2((a,b)) := \Big\{ f : (a,b) \to \mathbb{R} : \int_a^b f^2(x)\,\mathrm{d}x < \infty \Big\},$$

i.e., the family of all square (Lebesgue-)integrable functions. Furthermore, we definite the family

$$H^1((a,b)) := \Big\{ f \in L^2((a,b)) : g \in L^2((a,b)) \text{ for } g \text{ weak derivative of } f \}.$$

Finally, we define

$$H_0^1((a,b)) := \Big\{ f \in H^1((a,b)) : f = 0 \text{ at the endpoints } a \text{ and } b \}.$$

It can be shown that all the above are vector spaces, but this is beyond the scope of these notes. The space $L^2((a,b))$ is an example of the so-called *Lebesgue spaces*, whereas $H^1((a,b))$ and $H_0^1((a,b))$ are examples of the so-called *Sobolev spaces*.

Problem

36. Calculate the weak derivative of the function $f : (0, 2) \to \mathbb{R}$ with

$$f(x) = \begin{cases} x^2, & \text{for } 0 < x \le 1; \\ 2 - x, & \text{for } 1 < x < 2. \end{cases}$$

Does this function have a classical derivative everywhere in the interval $(0, 2)$? Explain. Does this function belong to $L^2((0,2))$? Does it belong to $H^1((0,2))$? Does it belong to $H_0^1((0,2))$? Explain.

7.3 The two-point boundary value problem in weak form

We consider again the two-point boundary value problem

$$\text{Find } u : (a,b) \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u(b) = 0. \tag{7.1}$$

where $f(x)$ is a known function.

The first step in defining a finite element method is to rewrite the two-point boundary value problem (7.1) in the so-called weak form, as follows.

Let $\mathcal{H} := H_0^1((a,b))$ defined in the previous section to be the family of functions v, that have a weak derivative and satisfy the Dirichlet boundary conditions $v(a) = 0 = v(b)$. We multiply the equation by a *test function* $v \in \mathcal{H}$, to get

$$-u''(x)v(x) = f(x)v(x),$$

and we integrate over the domain (a,b):

$$-\int_a^b u''(x)v(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x.$$

Now, if we perform an integration by parts to the integral on the left-hand side, we get

$$\int_a^b u'(x)v'(x)\mathrm{d}x - \big[u'(x)v(x)\big]_a^b = \int_a^b f(x)v(x)\mathrm{d}x,$$

for all $v \in \mathcal{H}$. Using the fact that $v(a) = 0 = v(b)$ for all $v \in \mathcal{H}$, we arrive to

$$\int_a^b u'(x)v'(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x,$$

for all $v \in \mathcal{H}$. Hence, the two-point boundary value problem can be transformed to the following problem in *weak form* (also known as *variational form*):

$$\text{Find } u \in \mathcal{H} \text{ s.t. } \int_a^b u'(x)v'(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x, \quad \text{for all } v \in \mathcal{H}. \tag{7.2}$$

Notice that if a function u is a solution to the problem (7.1), then it is also a solution to the problem (7.2). The converse, however, is *not* true, i.e., if a function u is a solution to the problem (7.2), then it is *not* necessarily a solution to the problem (7.1). Indeed, this can be verified by recalling that for $u \in \mathcal{H}$ to be a solution to (7.2), we only require that u has only a weak derivative, whereas for u to be a solution to (7.1), we have to require that u is twice differentiable. In this sense, Problem (7.2) is more general.

Indeed, suppose that a classical solution u exists. As u must be twice continuously differentiable by definition and $-u'' = f$, it follows that f must be continuous as well. So: for problem (7.1) to admit classical solution, we are only allowed to consider data functions f that are continuous. Now we may ask the same question, but for the weak formulation: for *which* functions f is the weak problem (7.2) well-posed? Functions f with (countable) discontinuities are integrable, and hence they certainly belongs to $L^2((a,b))$. It follows that the right-hand side of the weak problem (7.2) is well-defined in the sense that the integral is defined and finite (why?). It turns out that problem (7.2) is well-posed for *every* $f \in L^2((a,b))$! We have indeed the following general well-posedness result.

Lemma 7.9 *Suppose that the function $f \in L^2((a,b))$. Then Problem (7.2) is well-posed.*

Proof. The proof is beyond the scope of these notes. □

Another advantage of Problem (7.2) is that it is more *natural* in the sense that it matches the minimisation problem from which the boundary value problem (7.1) is derived, as we shall now see.

Lemma 7.10 *Consider the quadratic functional $F : \mathcal{H} \to \mathbb{R}$ given by*

$$F(v) = \frac{1}{2}\int_a^b v'(x)v'(x)\,\mathrm{d}x - \int_a^b f(x)v(x)\,\mathrm{d}x.$$

Then u is the solution of Problem (7.2) if and only if it minimises the functional F, that is [5]

$$u = argmin_{v\in\mathcal{H}} F(v).$$

[5] The notation *argmin* means 'the argument that minimizes', that is the minimizer as opposed to the minimum itself.

Proof. Suppose that $u \in \mathcal{H}$ is the solution to the weak problem (7.2). We show that $F(v) \geq F(u)$ for all $v \in \mathcal{H}$, hence u is the minimizer. Indeed, having using twice the fact that u satisfies (7.2), we have

$$\begin{aligned} F(v) - F(u) &= \frac{1}{2}\int_a^b v'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x - \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x + \int_a^b fu \,\mathrm{d}x \\ &= \frac{1}{2}\int_a^b v'v' \,\mathrm{d}x - \int_a^b u'v' \,\mathrm{d}x - \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x + \int_a^b u'u' \,\mathrm{d}x \\ &= \frac{1}{2}\int_a^b v'v' \,\mathrm{d}x - \int_a^b u'v' \,\mathrm{d}x + \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x \\ &= \frac{1}{2}\int_a^b (v-u)'(v-u)' \,\mathrm{d}x \geq 0. \end{aligned}$$

Now assume that u is the minimizer. Then in particular for every $v \in \mathcal{H}$ and $\lambda \in (0, 1]$ we have

$$\begin{aligned} 0 \leq F(u+\lambda v) - F(u) &= \frac{1}{2}\int_a^b (u+\lambda v)'(u+\lambda v)' \,\mathrm{d}x - \lambda \int_a^b fv \,\mathrm{d}x - \frac{1}{2}\int_a^b u'u' \,\mathrm{d}x \\ &= \lambda\left(\int_a^b u'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x\right) + \frac{1}{2}\lambda^2 \int_a^b v'v' \,\mathrm{d}x. \end{aligned}$$

Dividing through by λ and letting $\lambda \to 0$ gives

$$\int_a^b u'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x \geq 0 \qquad \forall v \in \mathcal{H}.$$

Substituting v with $-v$ gives

$$\int_a^b u'v' \,\mathrm{d}x - \int_a^b fv \,\mathrm{d}x \leq 0 \qquad \forall v \in \mathcal{H},$$

and we conclude that equality must hold, that is u solves (7.2). □

Problem

37. Write the following two-point boundary value problems in weak form

- find $u : (a, b) \to \mathbb{R}$, such that $-u''(x) + u'(x) = f(x)$ and $u(a) = 0$, $u(b) = 0$;
- find $u : (a, b) \to \mathbb{R}$, such that $-u''(x) = f(x)$ and $u(a) = 0$, $u'(b) = 0$.

7.4 The finite element method for the two-point boundary value problem

To define the finite element method for the boundary value problem (7.1) above, we consider an approximation to the problem (7.2). In contrast with the finite difference method, here we shall not approximate any derivatives directly; instead, we shall restrict the family of eligible solutions to a smaller family of functions, $\mathcal{H}_h \subset \mathcal{H} := H_0^1((a, b))$ say[6].

To construct a suitable family $\mathcal{H}_h$, we consider equally distributed points $x_0 < x_1 < \cdots < x_{N+1}$, at distance h between them, such that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N+1} = b.$$

We then choose the family $\mathcal{H}_h$ to be the family of continuous functions $v_h \in \mathcal{H}_h$ that are linear (i.e., straight lines) at each interval $[x_{i-1}, x_i]$, for $i = 1, \ldots, N + 1$ and $v_h(x_0) = 0 = v_h(x_{N+1})$; see Figure 7.1 for an illustration. Notice that such functions have a weak derivative on (a, b) and, moreover, they belong to $H_0^1((a, b))$ (why?).

[6]For the linear algebra lovers, $\mathcal{H}$ is an infinite dimensional vector space (why?), and $\mathcal{H}_h$ will be a finite-dimensional subspace, conveniently chosen.

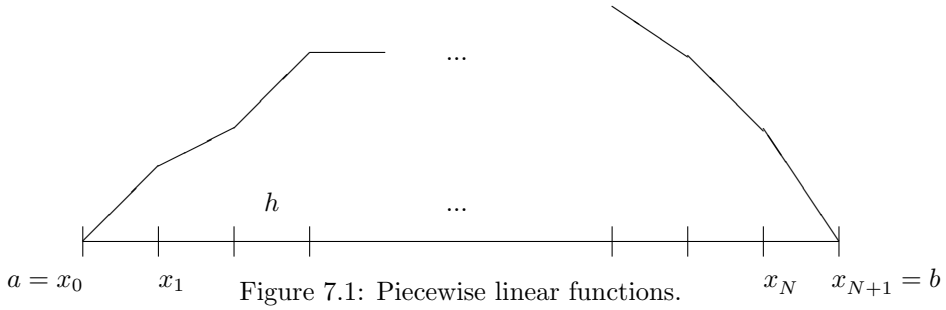


Figure 7.1: Piecewise linear functions.

We notice that for such piecewise linear functions, we only need to determine their value at each node x_i (i.e., we need to determine their values on a finite number of points); then all the intermediate values on each (x_{i-1}, x_i) are fully determined, since they are the values on the straight lines connecting the values at the nodes. Hence, it is possible to consider a finite number simpler functions, whose linear combinations give us all piecewise linear functions on the given grid. Indeed, for every x_i, $i = 1, \ldots, N$, we consider the "hat" functions $\phi_i : [a, b] \to \mathbb{R}$ with

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h}, & \text{if } x_{i-1} \leq x \leq x_i; \\ \frac{x_{i+1} - x}{h}, & \text{if } x_i \leq x \leq x_{i+1}; \\ 0, & \text{otherwise,} \end{cases}$$

i.e., $\phi_i \in \mathcal{H}_h$ is the piecewise linear function which is equal to 1 on the node x_i and is equal to zero on all other nodes; see Figure (7.2). We shall refer to this set of functions as *basis functions*.

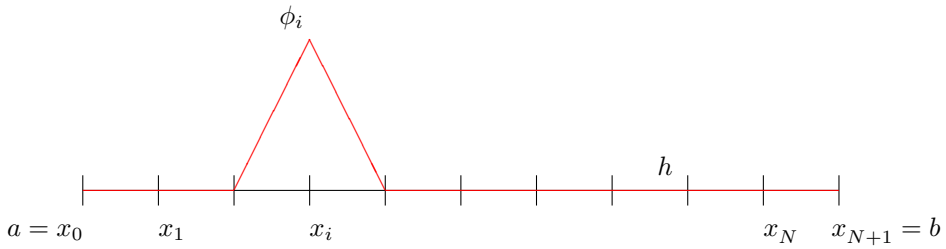


Figure 7.2: "Hat" function ϕ_i.

Any function $w_h \in \mathcal{H}_h$ that is continuous in $[a, b]$ and linear on each interval $[x_i, x_{i+1}]$ can be therefore written as

$$w_h = \sum_{i=1}^{N} W_i \phi_i,$$

for $W_i \in \mathbb{R}$ being the "height" of the function at the node x_i[7]. Hence, instead of solving the problem (7.2), we solve the approximate problem

$$\text{Find } u_h \in \mathcal{H}_h \text{ s.t. } \int_a^b u_h'(x) v_h'(x) \mathrm{d}x = \int_a^b f(x) v_h(x) \mathrm{d}x, \quad \text{for all } v_h \in \mathcal{H}_h;$$

This is the *finite element method*. Each grid, together with the associated basis functions ϕ at the nodes is called the *the finite element*. Of course, different choices of grid and different choices of basis functions lead to different finite element methods.

Thus, since every $v_h \in \mathcal{H}_h$ can be written as a linear combination of "hat" functions, we can instead equivalently ask

$$\text{Find } u_h \in \mathcal{H}_h \text{ s.t. } \int_a^b u_h'(x) \phi_i'(x) \mathrm{d}x = \int_a^b f(x) \phi_i(x) \mathrm{d}x, \quad \text{for all } i = 1, \ldots, N.$$

[7]The functions ϕ_i constitute a basis of the finite dimensional vector space $\mathcal{H}_h$ and, therefore, each function $u_h \in \mathcal{H}_h$ can be written as a linear combination of the basis

Also, since $u_h \in \mathcal{H}_h$, too, we have $u_h = \sum_{j=1}^N U_j \phi_j$, for some $U_j \in \mathbb{R}$ and, therefore the problem becomes

$$\text{Find } U_j, j = 1, \ldots, N, \text{ s.t. } \sum_{j=1}^N U_j \int_a^b \phi_j'(x)\phi_i'(x)\mathrm{d}x = \int_a^b f(x)\phi_i(x)\mathrm{d}x, \quad \text{for all } i = 1, \ldots, N.$$

This is a linear system of N equations with N unknowns and can be written as a linear system $A\mathbf{U} = \mathbf{F}$, for $A = [a_{ij}]_{i,j=1}^N$, $\mathbf{U} = (U_1, \ldots, U_N)^T$ and $\mathbf{F} = (F_1, \ldots, F_N)^T$, where

$$a_{ij} = \int_a^b \phi_j'(x)\phi_i'(x)\mathrm{d}x, \quad \text{and} \quad F_i = \int_a^b f(x)\phi_i(x)\mathrm{d}x.$$

We now calculate the entries a_{ij} of the matrix A, recalling that differentiation of ϕ_i's is understood to be taking place only inside each interval (x_{i-1}, x_i). Then ϕ_i' is given by

$$\phi_i'(x) = \begin{cases} \frac{1}{h}, & \text{if } x_{i-1} < x < x_i; \\ -\frac{1}{h}, & \text{if } x_i < x < x_{i+1}; \\ 0, & \text{otherwise,} \end{cases}$$

To calculate a_{ij}, we consider 4 cases:

case 1: $j = i$. We have

$$a_{ii} = \int_a^b \phi_i'(x)\phi_i'(x)\mathrm{d}x = \int_{x_{i-1}}^{x_{i+1}} \frac{1}{h^2}\mathrm{d}x = \frac{2}{h};$$

case 2: $j = i + 1$. We have

$$a_{i(i+1)} = \int_a^b \phi_{i+1}'(x)\phi_i'(x)\mathrm{d}x = \int_{x_i}^{x_{i+1}} \frac{1}{h}\big(-\frac{1}{h}\big)\mathrm{d}x = -\frac{1}{h};$$

case 3: $j = i - 1$. We have

$$a_{i(i-1)} = \int_a^b \phi_{i-1}'(x)\phi_i'(x)\mathrm{d}x = \int_{x_{i-1}}^{x_i} \big(-\frac{1}{h}\big)\frac{1}{h}\mathrm{d}x = -\frac{1}{h};$$

case 4: $|j - i| \geq 2$. In this case ϕ_i and ϕ_j are *not* simultaneously nonzero at any point in $[a, b]$, hence

$$a_{ij} = 0.$$

Therefore, we conclude that the linear system arising from the finite element method above is of the form

$$\begin{pmatrix} \frac{2}{h} & -\frac{1}{h} & 0 & 0 & \ldots & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 & \ldots & 0 \\ 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots \\ 0 & \ldots & 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ 0 & \ldots & 0 & 0 & -\frac{1}{h} & \frac{2}{h} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ \vdots \\ U_{N-1} \\ U_N \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ \vdots \\ F_{N-1} \\ F_N \end{pmatrix}, \tag{7.3}$$

which after multiplication by $-h$ becomes

$$\begin{pmatrix} -2 & 1 & 0 & 0 & \ldots & 0 \\ 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ldots \\ 0 & \ldots & 0 & 1 & -2 & 1 \\ 0 & \ldots & 0 & 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ \vdots \\ U_{N-1} \\ U_N \end{pmatrix} = -h \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ \vdots \\ F_{N-1} \\ F_N \end{pmatrix}. \tag{7.4}$$

Comparing this system with (3.23) which is the linear system arising from the finite different approximation of the two-point boundary value problem, we can see that the matrices in both cases are identical, although

they emerged from completely different procedures! On the other hand the right-hand sides of the systems (3.23) and (7.8) are different. Considering for the moment that the known function f is constant, say $f(x) = C$, then the i-th entry of the right-hand side of (3.23) is equal to $-h^2C$. Calculating now the i-th entry of the right-hand side of (7.8), we have

$$-hF_i = -h\int_a^b f(x)\phi_i(x)\mathrm{d}x = -hC\int_a^b \phi_i(x)\mathrm{d}x = -hC\int_{x_{i-1}}^{x_{i+1}} \phi_i(x)\mathrm{d}x = -hCh = -h^2C,$$

which is identical to what we found for the right-hand side of (3.23)! More interesting cases arise when f is not constant; then the two right-hand sides are, in general, not equal anymore.

Remark 7.11 *If one wishes to implement the above finite element method in the computer for general right-hand side function f, should resort to numerical integration/quadrature for the computation of the integrals $\int_a^b f(x)\phi_i(x)\mathrm{d}x$.*

We shall now consider the case of Neumann boundary condition at the endpoint b. We consider the problem

$$\text{Find } u : (a,b) \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u'(b) = 0. \tag{7.5}$$

where $f(x)$ is a known function.

We write the problem (7.5) in variational form, as follows. Here, we shall define a different family of test functions (and of eligible solutions), to conform with the new boundary conditions. We define

$$\tilde{\mathcal{H}} := \{v \in H^1((a,b)) : v(a) = 0\},$$

i.e., the family of functions $v \in H^1((a,b))$ which satisfy the Dirichlet boundary condition $v(a) = 0$ *only*.

Next, we multiply the PDE by a *test function* $v \in \tilde{\mathcal{H}}$ and we integrate over the domain $[a,b]$:

$$-\int_a^b u''(x)v(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x.$$

Now, if we perform and integration by parts to the integral on the left-hand side, we get

$$\int_a^b u'(x)v'(x)\mathrm{d}x - \big[u'(x)v(x)\big]_a^b = \int_a^b f(x)v(x)\mathrm{d}x,$$

for all $v \in \tilde{\mathcal{H}}$. Using the fact that $v(a) = 0$ and that $u'(b) = 0$ for all $v \in \tilde{\mathcal{H}}$, we arrive to

$$\int_a^b u'(x)v'(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x,$$

for all $v \in \tilde{\mathcal{H}}$. Hence, the two-point boundary value problem can be transformed to the following problem in *weak form*:

$$\text{Find } u \in \tilde{\mathcal{H}} \text{ s.t. } \int_a^b u'(x)v'(x)\mathrm{d}x = \int_a^b f(x)v(x)\mathrm{d}x, \quad \text{for all } v \in \tilde{\mathcal{H}}. \tag{7.6}$$

The second step in defining the finite element method is to restrict the family of functions where we want to solve to the problem (7.2).

To construct a suitable family $\tilde{\mathcal{H}}_h$, we consider equally distributed points $x_0 < x_1 < \cdots < x_{N+1}$, at distance h between them, such that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N+1} = b.$$

We then choose the family $\tilde{\mathcal{H}}_h$ to be the family of continuous functions $v_h \in \mathcal{H}_h$ that are linear (i.e., straight lines) at each interval $[x_{i-1}, x_i]$, for $i = 1, \ldots, N+1$ and $v_h(x_0) = 0$.

Notice that, in contrast with the case of Dirichlet boundary conditions, here we do *not* prescribe any value for the node $x_{N+1} = b$. Instead, this is now an unknown and will be treated in a special fashion as we shall see immediately.

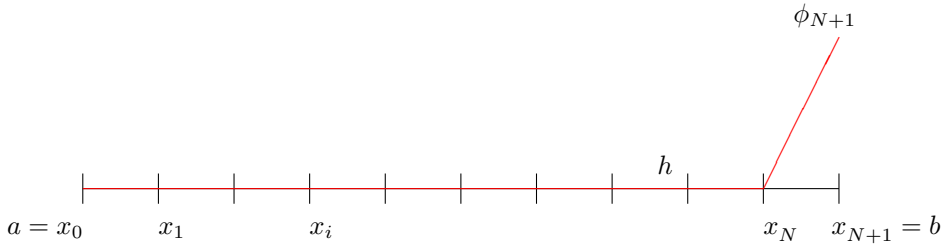


Figure 7.3: "Hat" function ϕ_{N+1}.

For every x_i, $i = 1, \ldots, N$, we consider the basis functions to be the "hat" functions $\phi_i : [a, b] \to \mathbb{R}$ as before; see Figure (7.2). We now also need a basis function for the node x_{N+1}, which is simply given by

$$\phi_{N+1}(x) = \begin{cases} \frac{x - x_N}{h}, & \text{if } x_N \leq x \leq x_{N+1}; \\ 0, & \text{otherwise;} \end{cases}$$

see Figure (7.3)

Any function $w_h \in \tilde{\mathcal{H}}_h$ that is continuous in $[a, b]$ and linear on each interval $[x_i, x_{i+1}]$ can be therefore written as

$$w_h = \sum_{i=1}^{N+1} W_i \phi_i.$$

Hence, instead of solving the problem (7.6), we solve the approximate problem

$$\text{Find } u_h \in \tilde{\mathcal{H}}_h \text{ s.t. } \int_a^b u_h'(x) v_h'(x) \mathrm{d}x = \int_a^b f(x) v_h(x) \mathrm{d}x, \quad \text{for all } v_h \in \tilde{\mathcal{H}}_h;$$

This is the *finite element method* with linear elements for this problem.

Thus, since every $v_h \in \mathcal{H}_h$ can be written as a linear combination of "hat" functions, we can instead equivalently ask

$$\text{Find } u_h \in \tilde{\mathcal{H}}_h \text{ s.t. } \int_a^b u_h'(x) \phi_i'(x) \mathrm{d}x = \int_a^b f(x) \phi_i(x) \mathrm{d}x, \quad \text{for all } i = 1, \ldots, N, N+1.$$

Also, since $u_h \in \tilde{\mathcal{H}}_h$, too, we have $u_h = \sum_{j=1}^{N+1} U_j \phi_j$, for some $U_j \in \mathbb{R}$ and, therefore the problem becomes

$$\text{Find } U_j, j = 1, \ldots, N+1, \text{ s.t. } \sum_{j=1}^{N+1} U_j \int_a^b \phi_j'(x) \phi_i'(x) \mathrm{d}x = \int_a^b f(x) \phi_i(x) \mathrm{d}x, \quad \text{for all } i = 1, \ldots, N, N+1.$$

This is a linear system of $N + 1$ equations with $N + 1$ unknowns and can be written as a linear system $A\mathbf{U} = \mathbf{F}$, for $A = [a_{ij}]_{i,j=1}^{N+1}$, $\mathbf{U} = (U_1, \ldots, U_N, U_{N+1})^T$ and $\mathbf{F} = (F_1, \ldots, F_N, F_{N+1})^T$, where

$$a_{ij} = \int_a^b \phi_j'(x) \phi_i'(x) \mathrm{d}x, \quad \text{and} \quad F_i = \int_a^b f(x) \phi_i(x) \mathrm{d}x.$$

The entries a_{ij} of the matrix A for $i, j = 1, \ldots, N$ can be calculated as above. For the entries $a_{i(N+1)}$ and $a_{(N+1)j}$, we shall need

$$\phi_{N+1}'(x) = \begin{cases} \frac{1}{h}, & \text{if } x_N < x < x_{N+1}; \\ 0, & \text{otherwise.} \end{cases}$$

To calculate $a_{i(N+1)}$ and $a_{(N+1)j}$, we consider 3 cases:

case 1: $i = N + 1$ (or, equivalently, $j = N + 1$). We have

$$a_{(N+1)(N+1)} = \int_a^b \phi_{N+1}'(x) \phi_{N+1}'(x) \mathrm{d}x = \int_{x_N}^{x_{N+1}} \frac{1}{h^2} \mathrm{d}x = \frac{1}{h};$$

case 2: $i = N$ (or, equivalently, $j = N$). We have

$$a_{N(N+1)} = a_{(N+1)N} = \int_a^b \phi_N'(x)\phi_{N+1}'(x)\mathrm{d}x = \int_{x_N}^{x_{N+1}} \big(-\frac{1}{h}\big)\frac{1}{h}\mathrm{d}x = -\frac{1}{h};$$

case 3: $|N+1-i| \geq 2$ (or, equivalently, $|N+1-j| \geq 2$). In this case ϕ_i and ϕ_j are *not* simultaneously nonzero at any point in $[a, b]$, hence

$$a_{(N+1)j} = 0 = a_{i(N+1)}.$$

Therefore, we conclude that the linear system arising from the finite element method above is of the form

$$\begin{pmatrix} \frac{2}{h} & -\frac{1}{h} & 0 & 0 & \dots & 0 & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 & \dots & 0 & 0 \\ 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots & \\ 0 & \dots & 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 \\ 0 & \dots & 0 & 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ 0 & \dots & 0 & 0 & 0 & -\frac{1}{h} & \frac{1}{h} \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ \vdots \\ U_{N-1} \\ U_N \\ U_{N+1} \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ \vdots \\ F_{N-1} \\ F_N \\ F_{N+1} \end{pmatrix}, \tag{7.7}$$

which after multiplication by $-h$ becomes

$$\begin{pmatrix} -2 & 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots & 0 \\ 0 & \dots & 0 & 1 & -2 & 1 & 0 \\ 0 & \dots & 0 & 0 & 1 & -2 & 1 \\ 0 & \dots & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ \vdots \\ U_{N-1} \\ U_N \\ U_{N+1} \end{pmatrix} = -h \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ \vdots \\ F_{N-1} \\ F_N \\ F_{N+1} \end{pmatrix}. \tag{7.8}$$

Comparing this system with (3.25) which is the linear system arising from the finite different approximation of the two-point boundary with Neumann boundary condition, we can see that the matrix arising from the finite difference method with the fictitious node considered in (3.25) and the of the system (7.7) are identical!

Nevertheless, we again have different right-hand sides but, as discussed above, if we consider $f(x) = C$ for some constant C, the right-hand side entries of the systems (3.25) and (7.7) coincide for $i = 1, \dots, N$. For $i = N + 1$, we calculate

$$-hF_{N+1} = -hC \int_{x_N}^{x_{N+1}} \phi_{N+1}(x)\mathrm{d}x = -\frac{1}{2}h^2 C,$$

which coincides with the corresponding value in (3.25)!

Problem

38. Construct a finite element method for the two-point boundary value problem

$$\text{Find } u : (a, b) \to \mathbb{R} \text{ function, such that } -u''(x) = f(x) \text{ and } u(a) = 0,\ u(b) = 0.$$

using a nonuniform grid of the form

$$a = x_0,\ x_1 = x_0 + h_1,\ x_2 = x_1 + h_2,\ \dots,\ x_N = x_{N-1} + h_N,\ x_{N+1} = b,$$

i.e., $h_i = x_i - x_{i-1}$ for $i = 1, \dots N + 1$. Write the resulting linear system.

7.4.1 Error analysis of the finite element method

Here we analyse the finite element method for the two-point boundary value problem (3.21).

Recall Lemma 7.9 stating that the problem (3.21) is well-posed for every $f \in L^2((a,b))$. We did not prove that result, as the proof requires some advanced results in functional analysis. On the contrary, we are now going to prove that the finite element method is well-posed for any such f. This is an easier task because the finite element problem is posed on a finite dimensional setting.

Theorem 7.12 *Let $f \in L^2((a,b))$. Then the finite element problem*

$$\text{Find } u_h \in \mathcal{H}_h \text{ s.t. } \int_a^b u_h'(x)v_h'(x)\mathrm{d}x = \int_a^b f(x)v_h(x)\mathrm{d}x, \quad \text{for all } v_h \in \mathcal{H}_h; \tag{7.9}$$

with $\mathcal{H}_h$ as in 7.4 admits a unique solution.

Proof. We have seen that problem (7.9) is equivalent to the square linear system $A\mathbf{U} = \mathbf{F}$. Now, for such a square linear system uniqueness is equivalent to existence. Hence we just need to prove uniqueness, and we do this by contradiction. Assume that $\mathbf{U}_1$ and $\mathbf{U}_2$ are two distinct solutions. Then $\mathbf{W} = \mathbf{U}_1 - \mathbf{U}_2 \neq \mathbf{0}$ satisfies $A\mathbf{W} = \mathbf{0}$. Now let $w_h = \sum_{i=1}^N V_i\phi_i$ the associated finite element function. Going back to the weak formulation $A\mathbf{W} = \mathbf{0}$ is equivalent to $\int_a^b w_h'(x)v_h'(x)\mathrm{d}x = 0$ for all $v_h \in \mathcal{H}_h$. In particular, testig with $v_h = w_h$ yields $\int_a^b (w_h'(x))^2\mathrm{d}x = 0$, hence $w_h' = 0$, that is w_h is constant. But, as $w_h \in \mathcal{H}_h$ we have $w_h(a) = w_h(b) = 0$ thus $w_h = 0$. Hence $\mathbf{W} = \mathbf{0}$, which is a contradiction. □

Let us now come to the question of measuring how close is u_h to u. To start with, we need to decide how the distance between elements of $\mathcal{H}$ is going to be measured. The nature of $\mathcal{H}$ suggests the use of integral norms. In particular, we shall consider:

$$\|v\|_{\mathcal{H}} := \left(\int_a^b (v'(x))^2 dx\right)^{1/2} \qquad \forall v \in \mathcal{H}. \tag{7.10}$$

This is indeed a norm. You can verify that it is a seminorm using the properties of the integral and the so called Minkowski's inequality (which proofs the triangle inequality in this context). Now suppose that $\|v\|_{\mathcal{H}} = 0$. Then by (7.10) we have $v' = 0$ and thus $v = 0$ follows as in the proof of Theorem 7.12, hence $\|\cdot\|_{\mathcal{H}}$ is a norm. It is somehow the natural norm for problem (7.2): we call it the *energy* norm of the problem[8].

It is also a norm defined by an inner product, as we now show. Consider on $\mathcal{H}$ the product

$$w \cdot v := \int_a^b w'(x)v'(x)dx \qquad \forall w, v \in \mathcal{H}. \tag{7.11}$$

This is an inner product for $\mathcal{H}$ as a vector space on $\mathbb{R}$ as the properties of the integral easily implies that for all $v, w, z \in \mathcal{H}$ and $a, b \in \mathbb{R}$ we have:

- Symmetry: $w \cdot v = v \cdot w$;
- Linearity: $(aw + bz) \cdot v = a(w \cdot v) + b(z \cdot v)$
- Positivity: $v \cdot v \geq 0$ and $v \cdot v = 0$ if and only if $v = 0$.

And our norm on $\mathcal{H}$ is indeed just given by evaluating the inner product with the same argument:

$$\|v\|_{\mathcal{H}} = (v \cdot v)^{1/2}.$$

Knowing this gives us a crucial tool: the Schwarz inequality.

Theorem 7.13 (Cauchy-Schwarz Inequality) *If $(V, \cdot)$ is an inner product space on $\mathbb{R}$ (that is, a real vector space with an inner product), then*

$$|w \cdot v| \leq (w \cdot w)^{1/2}(v \cdot v)^{1/2} \tag{7.12}$$

[8] It is by all means possible to bound the error at the grid nodes as was done in the previous chapters for finite difference methods. But pointwise error estimates are more difficult to obtain in general (not in this case, though).

Proof. Notice that the theorem is general. For instance, you should already know that $\mathbb{R}^n$ with its usual inner product is an inner product space (in this case, the product yields the Euclidean norm!).

If either of w and v is 0, the inequality holds trivially. Let $w, v \neq 0$. For all $t \in \mathbb{R}$,

$$0 \leq (w - tv) \cdot (w - tv) = w \cdot w - 2tw \cdot v + t^2 v \cdot v.$$

This, for $t = (w \cdot v)/(v \cdot v)$, gives

$$0 \leq w \cdot w - \frac{(w \cdot v)^2}{v \cdot v}.$$

which is equivalent to (7.12). □

In our case, recalling the definition of the $\mathcal{H}$ inner product and norm, we see that the Cauchy-Schwarz inequality reads:

$$\left| \int_a^b w'(x)v'(x)dx \right| \leq \|w\|_{\mathcal{H}} \|v\|_{\mathcal{H}} \qquad \forall w, v \in \mathcal{H}. \tag{7.13}$$

We are now ready to analyse our finite element method. As the space $\mathcal{H}_h$ is a subspace of $\mathcal{H}$, the equation (7.2) is satisfied, in particular, for every $v = v_h \in \mathcal{H}_h$, thus

$$\int_a^b u'(x)v_h'(x)\mathrm{d}x = \int_a^b f(x)v_h(x)\mathrm{d}x, \quad \text{for all } v_h \in \mathcal{H}_h$$

and, subtracting (7.9) to this equation we get

$$\int_a^b (u'(x) - u_h'(x))v_h'(x)\mathrm{d}x = 0 \quad \text{for all } v_h \in \mathcal{H}_h. \tag{7.14}$$

This is a crucial property of the finite element solution. It is known as **Galerkin orthogonality**. It is all we need in order to prove the following *optimality* result.

Theorem 7.14 *Among all functions in the space $\mathcal{H}_h$, the finite element solution u_h given by* (7.9) *is the closest to the weak solution u. That is,*

$$\|u - u_h\|_{\mathcal{H}} = \min_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}. \tag{7.15}$$

Proof. We start by considering the square of the norm of the error and use twice Galerkin orthogonality:

$$\begin{aligned}
\|u - u_h\|_{\mathcal{H}}^2 &= \int_a^b (u'(x) - u_h'(x))(u'(x) - u_h'(x))\mathrm{d}x \\
&= \int_a^b (u'(x) - u_h'(x))u'(x)\mathrm{d}x \\
&= \int_a^b (u'(x) - u_h'(x))(u'(x) - v_h'(x))\mathrm{d}x \qquad \forall v_h \in \mathcal{H}_h \\
&\leq \|u - u_h\|_{\mathcal{H}} \|u - v_h\|_{\mathcal{H}} \qquad \forall v_h \in \mathcal{H}_h,
\end{aligned}$$

with the inequality being due to the Schwarz inequality (7.13).

If $\|u - u_h\|_{\mathcal{H}} = 0$ the theorem's statement is trivially verified; otherwise, dividing through by $\|u - u_h\|_{\mathcal{H}}$ we conclude that

$$\|u - u_h\|_{\mathcal{H}} \leq \|u - v_h\|_{\mathcal{H}} \qquad \forall v_h \in \mathcal{H}_h.$$

As the inequality is true for every $v_h \in \mathcal{H}_h$ it follows that

$$\|u - u_h\|_{\mathcal{H}} \leq \inf_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}$$

But, by definition of inf, we also have

$$\inf_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}} \leq \|u - u_h\|_{\mathcal{H}},$$

thus the inf is a min attained at u_h. □

The theorem's result is quite meaningful: it is saying that the finite element method is (at least in this case!) *optimal* in the sense that the error is as good as it can be given that the approximate solution is chosen out of the subspace $\mathcal{H}_h$. Such results are known as Cèa Lemmas. In view of the theorem, the question is now how close is the finite element space to the exact solution! An estimation of such distance can be found as follows.

Definition 7.15 *Given $v \in C^0((a,b))$, the* interpolant v_I *of v is the function in $\mathcal{H}_h$ that is equal to v at the nodes*[9]*. Thus,*

$$v_I := \sum_{i=1}^{N} v(x_i)\phi_i.$$

The following result gives an estimate of the distance between a function and its interpolant.

Theorem 7.16 *Let $u \in \mathcal{H}$. If the second weak derivative u'' of u is in $L^2((a,b))$ then*[10]

$$\|u - u_I\|_{\mathcal{H}} \leq Ch\|u'\|_{\mathcal{H}}, \tag{7.16}$$

for some constant C that does not depend on h.

Proof. Once more, it is convenient to start with the square of the norm we need to bound:

$$\|u - u_I\|_{\mathcal{H}}^2 = \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} ((u - u_I)'(x))^2 dx.$$

Now the idea is to bound each integral under the summation sign separately. Indeed, if we can prove that

$$\int_{x_i}^{x_{i+1}} ((u - u_I)'(x))^2 dx \leq ch^2 \int_{x_i}^{x_{i+1}} (u''(x))^2 dx, \tag{7.17}$$

then the desired bound follows by summing up all the bounds.

To show (7.17), let us first name the interpolation error as $e := u - u_I$. As u_I is linear on the interval $[x_i, x_{i+i}]$ its second derivative is identically zero. Hence, $e'' = u'' - u_I'' = u''$. Thus we can rewrite (7.17) as

$$\int_{x_i}^{x_{i+1}} (e'(x))^2 dx \leq ch^2 \int_{x_i}^{x_{i+1}} (e''(x))^2 dx. \tag{7.18}$$

Further, introducing the change of variables $x = x_i + h\hat{x}$ for $\hat{x} \in [0,1]$ and defining $\hat{e}(\hat{x}) := e(x_i + h\hat{x})$, the above bound is equivalent to

$$\int_0^1 (\hat{e}'(x))^2 d\hat{x} \leq c \int_0^1 (\hat{e}''(x))^2 d\hat{x}. \tag{7.19}$$

We have $\hat{e}(0) = e(x_i) = 0$ and $\hat{e}(1) = e(x_{i+1}) = 0$ by definition of the interpolant. As $\hat{e}$ is continuous, the Theorem of Rolle implies that there exists $\xi \in (0,1)$ such that $\hat{e}'(\xi) = 0$. Thus,

$$\hat{e}'(\hat{x}) = \int_\xi^{\hat{x}} \hat{e}''(y) dy \qquad \forall \hat{x} \in [0,1].$$

Now, using *another* Schwarz inequality,

$$\begin{aligned} |\hat{e}'(\hat{x})| &= \left| \int_\xi^{\hat{x}} \hat{e}''(y) dy \right| \\ &= \left| \int_\xi^{\hat{x}} 1 \cdot \hat{e}''(y) dy \right| \\ &\leq \left| \int_\xi^{\hat{x}} 1 dy \right|^{1/2} \left| \int_\xi^{\hat{x}} (\hat{e}''(y))^2 dy \right|^{1/2} \\ &\leq |\hat{x} - \xi|^{1/2} \left(\int_0^1 (\hat{e}''(y))^2 dy \right)^{1/2}. \end{aligned}$$

[9] The reason why we need to restrict the definition to continuous function is that point-values are not necessarily well defined for L^2 functions...

[10] Under these hypothesis, the function u is continuous, hence its interpolant is defined. The proof of this fact is beyond the scope of these notes.

Hence,

$$\int_0^1 |\hat{e}'(\hat{x})|^2 d\hat{x} \leq \left(\int_0^1 |\hat{x} - \xi| \, d\hat{x} \right) \left(\int_0^1 (\hat{e}''(y))^2 dy \right) \leq \frac{1}{2} \int_0^1 (\hat{e}''(\hat{x}))^2 dy.$$

□

We are now finally ready to give a convergence result for the finite element method.

Theorem 7.17 *If the second weak derivative of the exact solution u is in $L^2((a,b))$ then*

$$\|u - u_h\|_{\mathcal{H}} \leq Ch\|u'\|_{\mathcal{H}},$$

for some constant C that does not depend on h.

Proof. Applying Cèa's Lemma and then using the interpolation error bound we get:

$$\|u - u_h\|_{\mathcal{H}} = \min_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}} \leq \|u - u_I\|_{\mathcal{H}} \leq Ch\|u'\|_{\mathcal{H}}.$$

□

Problem

39. Show that the right-hand side of problem (7.2) is finite whenever $f \in L^2((a,b))$ for every $v \in \mathcal{H}$. *[Hint: show that $\|v\|_{L^2((a,b))} := \left(\int_a^b (v(x))^2 \, dx \right)^{1/2}$ defines a norm associated to a (which?) inner product.]*

7.5 Finite element method in two and three dimensions

So far we have considered the case of one-dimensional problems, whereby we witnessed the conceptual easiness of constructing finite element methods for elliptic problems. The full flexibility of the finite element method, however, can be realised when applied to problems in 2 or 3 dimensions, as we shall see below.

Before starting our discussion on the construction of the finite element method in 2 or 3 dimensions, we note that the concepts of weak derivative can be extended to partial derivatives, along with the families of functions L^2, H^1 and H^1_0. We refrain from giving the explicit definitions here for reasons of brevity[11], noting, however, that all partial derivatives discussed below are understood as weak derivatives.

Let us consider the Poisson problem with homogeneous Dirichlet boundary conditions over an open bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$:

$$\begin{aligned} -\Delta u &= f, && \text{in } \Omega \\ u &= 0, && \text{on } \partial\Omega, \end{aligned} \tag{7.20}$$

where $f : \Omega \to \mathbb{R}$ is a known function. For simplicity, we shall assume that the boundary $\partial\Omega$ of the domain Ω is composed by consecutive line segments (i.e., $\partial\Omega$ is a polygon).

The first step in defining a finite element method is to rewrite the problem (7.20) in weak form, using the *divergence theorem*[12]. We define $\mathcal{H}$ to be the family of square-integrable functions $v : \Omega \to \mathbb{R}$, which are weakly differentiable in Ω (their weak derivative is also square integrable in Ω) and satisfy the Dirichlet boundary condition $v = 0$ on $\partial\Omega$ (this family is usually denoted by $H^1_0(\Omega)$).

Next, we multiply the PDE by a *test function* $v \in \mathcal{H}$, to get

$$-\Delta u\, v = fv,$$

and we integrate over the domain Ω:

$$-\int_\Omega \Delta u\, v \mathrm{d}V = \int_\Omega fv \mathrm{d}V.$$

Now, if we perform and integration by parts in d dimensions (see (7.21)) to the integral on the left-hand side, we get

$$\int_\Omega \nabla u \cdot \nabla v \mathrm{d}V - \int_{\partial\Omega} v \nabla u \cdot \vec{n} \mathrm{d}S = \int_\Omega fv \mathrm{d}V.$$

for all $v \in \mathcal{H}$, where $\vec{n}$ denotes the unit outward normal vector to the boundary $\partial\Omega$. Using the fact that $v = 0$ on $\partial\Omega$ for all $v \in \mathcal{H}$, we arrive to

$$\int_\Omega \nabla u \cdot \nabla v \mathrm{d}V = \int_\Omega fv \mathrm{d}V.$$

for all $v \in \mathcal{H}$. Hence, the Poisson problem with homogeneous Dirichlet boundary conditions can be transformed to the following problem in *weak form*

$$\text{Find } u \in \mathcal{H} \text{ s.t. } \int_\Omega \nabla u \cdot \nabla v \mathrm{d}V = \int_\Omega fv \mathrm{d}V, \quad \text{for all } v \in \mathcal{H}. \tag{7.22}$$

[11] The interested reader can find out more about calculus of functions of several variables, e.g., in the book of Adams and Fournier, Sobolev Spaces, Academic Press (2003).

[12] The divergence theorem, also known as Gauss' theorem, or Green's theorem (when applied to two-dimensional domains) states that, given an open bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, and a differentiable vector field $\vec{F}(x, y) = (F_1(x, y), F_2(x, y))$ if $d = 2$, or $\vec{F}(x, y, z) = (F_1(x, y, z), F_2(x, y, z))$ if $d = 3$, then

$$\int_\Omega \nabla \cdot \vec{F} \mathrm{d}V = \int_{\partial\Omega} \vec{F} \cdot \vec{n} \mathrm{d}S,$$

where $\vec{n}$ denotes the unit outward normal vector to the boundary $\partial\Omega$, $\mathrm{d}V$ the infinitesimal area if $d = 2$ or the infinitesimal volume if $d = 3$, and $\mathrm{d}S$ the infinitesimal arc if $d = 2$ or the infinitesimal surface if $d = 3$. If we apply the divergence theorem to the vector field $\vec{F} = v\nabla u$, we get

$$\int_\Omega \nabla \cdot (v\nabla u) \mathrm{d}V = \int_{\partial\Omega} v \nabla u \cdot \vec{n} \mathrm{d}S.$$

Using the formula $\nabla \cdot (v\nabla u) = \nabla v \cdot \nabla u + v \cdot \nabla \cdot \nabla u$ and recalling that $\nabla \cdot \nabla u = \Delta u$, we deduce the *integration by parts formula* in d dimensions

$$\int_\Omega \Delta u\, v \mathrm{d}V = -\int_\Omega \nabla v \cdot \nabla u \mathrm{d}V + \int_{\partial\Omega} v \nabla u \cdot \vec{n} \mathrm{d}S. \tag{7.21}$$

The second step in defining the finite element method is to consider an approximation to the problem (7.22). For simplicity, we shall consider only the case $d = 2$, i.e., when Ω is a bounded set on the plane; the case $d = 3$ is analogous. To this end, we shall restrict the family of eligible solutions to a smaller family of functions, $\mathcal{H}_h$. To construct a suitable family $\mathcal{H}_h$, we split the domain Ω into triangles $T \in \mathcal{T}$ (from now on we shall refer to these triangles as *elements*), where $\mathcal{T}$ denotes the set of elements to which we shall refer to as the *mesh*, as shown in Figure 7.4. Each vertex of a triangle will be referred to as a *node*. Further, we define the mesh parameter h to be the size of the largest triangle in the mesh:

$$h = \max_{T \in \mathcal{T}} \{\operatorname{diam} T\},$$

(the diameter of a triangle is the length of its longest side). This takes the role of the length h of the subintervals used to analyse one-dimensional problems.

Suppose the triangulation contains N nodes that are not on the boundary $\partial\Omega$. We can then choose (arbitrarily) a numbering between $1, \ldots, N$ for the nodes.

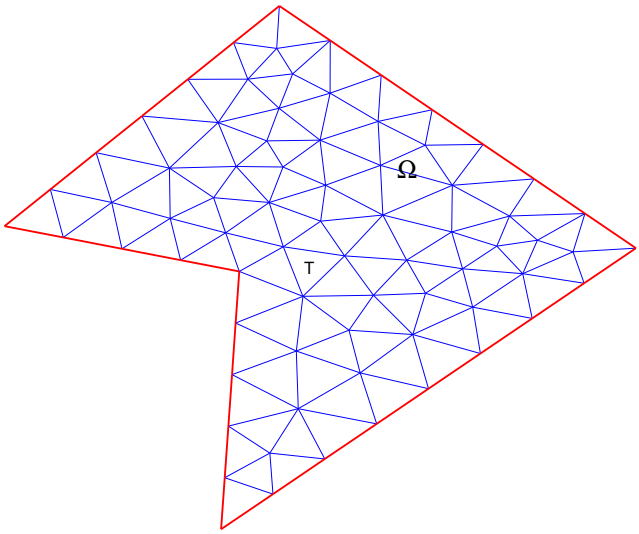


Figure 7.4: A mesh in two dimensions

We consider the family $\mathcal{H}_h$ to consist of the functions that are continuous and linear on every element $T \in \mathcal{T}$, see Figure 7.5 for an illustration of such function.

Hence, any function $w_h \in \mathcal{H}_h$ that is continuous in Ω and linear on each element $T \in \mathcal{T}$ can be therefore written as

$$w_h = \sum_{i=1}^{N} W_i \phi_i,$$

for $W_i \in \mathbb{R}$ being the "height" of the function at the node i. Therefore, instead of solving the problem (7.22), we solve the approximate problem

$$\text{Find } u_h \in \mathcal{H}_h \text{ s.t. } \int_\Omega \nabla u_h \cdot \nabla v_h \mathrm{d}V = \int_\Omega f\, v_h \mathrm{d}V, \quad \text{for all } v_h \in \mathcal{H}_h; \tag{7.23}$$

This is the *finite element method*. Each mesh, together with the associated basis functions ϕ_i at the nodes is called the *the finite element*. Of course, different choices of meshes and different choices of basis functions lead to different finite element methods.

Thus, since every $v_h \in \mathcal{H}_h$ can be written as a linear combination of "pyramid" functions, we can instead equivalently ask

$$\text{Find } u_h \in \mathcal{H}_h \text{ s.t. } \int_\Omega \nabla u_h \cdot \nabla \phi_i \mathrm{d}V = \int_\Omega f\, \phi_i \mathrm{d}V, \quad \text{for all } i = 1, \ldots, N.$$

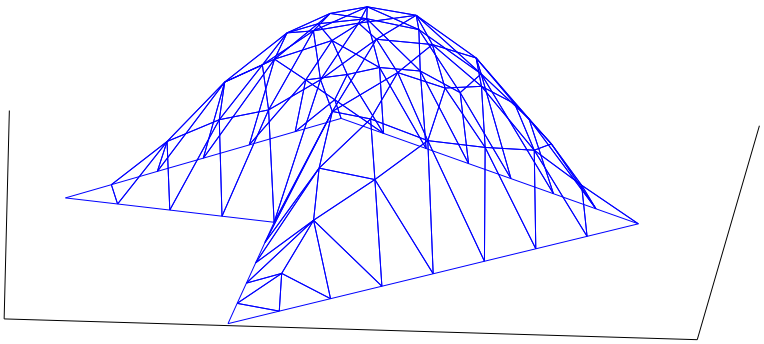


Figure 7.5: A function belonging to the family $\mathcal{H}_h$ subject to the mesh from Figure 7.4

Also, since $u_h \in \mathcal{H}_h$, too, we have $u_h = \sum_{j=1}^{N} U_j \phi_j$, for some $U_j \in \mathbb{R}$ and, therefore the problem becomes

$$\text{Find } U_j, j = 1, \ldots, N, \text{ s.t. } \sum_{j=1}^{N} U_j \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \mathrm{d}V = \int_\Omega f\, \phi_i \mathrm{d}V, \quad \text{for all } i = 1, \ldots, N.$$

This is a linear system of N equations with N unknowns and can be written as a linear system $A\mathbf{U} = \mathbf{F}$, for $A = [a_{ij}]_{i,j=1}^{N}$, $\mathbf{U} = (U_1, \ldots, U_N)^T$ and $\mathbf{F} = (F_1, \ldots, F_N)^T$, where

$$a_{ij} = \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \mathrm{d}V, \quad \text{and} \quad F_i = \int_\Omega f \phi_i \mathrm{d}V. \tag{7.24}$$

Example 7.18 *We use the finite element method to approximate the solution to the Poisson problem with homogeneous Dirichlet boundary conditions:*

$$\begin{aligned} -\Delta u &= 100 \sin(\pi x), && \text{in } \Omega \\ u &= 0, && \text{on } \partial\Omega, \end{aligned} \tag{7.25}$$

where Ω is given by the domain in Figure 7.4, along with the mesh used in the approximation. The finite element approximation is shown in Figure 7.6.

Problem

40. Consider the Poisson problem with mixed boundary conditions

$$\begin{aligned} -\Delta u(x,y) &= f(x,y), && (x,y) \in (0,1)^2 \\ u(x,0) = u(0,y) = u(x,1) &= 0, && 0 \le x, y \le 1 \\ u_x(1,y) &= 1, && 0 \le y \le 1. \end{aligned}$$

Write the problem in weak form and comment on how to construct a finite element method for it.

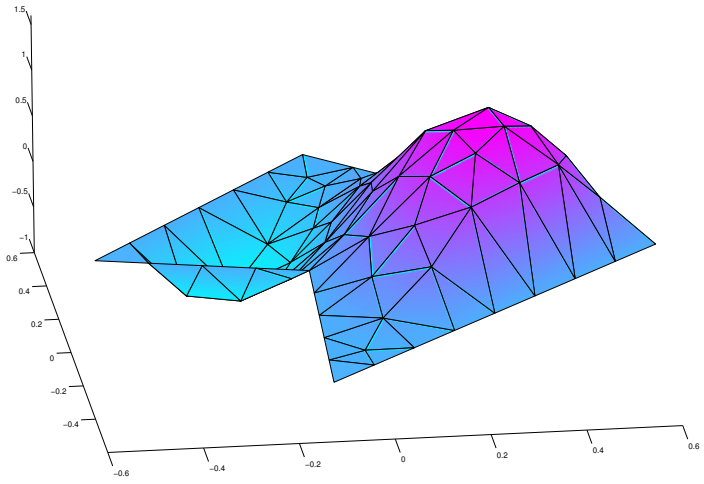


Figure 7.6: Finite element approximation to the problem (7.25).

7.6 Calculation of the FEM system for a simple problem

To make the discussion concrete we shall calculate the system emerging from the finite element method for a simple geometry in two dimensions. In particular, we consider $\Omega = (0,1) \times (0,1)$, which we subdivide into triangles, with nodes (x_i, y_j) such that

$$0 = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N+1} = 1,$$

and

$$0 = y_0,\ y_1 = y_0 + h,\ y_2 = y_1 + h,\ \ldots,\ y_N = y_{N-1} + h,\ y_{N+1} = 1,$$

hence, we have $h = 1/(N+1)$, as illustrated in Figure 7.7.

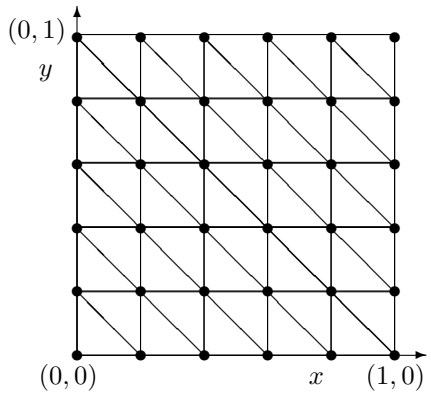


Figure 7.7: Triangulation of $\Omega = (0,1) \times (0,1)$.

For each node $(x_i i, y_j)$, we associate a basis (pyramid) function $\phi_{i,j}$ which is linear on every element and continuous, taking the value 1 at the node (x_i, y_j) and the value 0 at all other nodes $(x_k, y_l) \neq (x_i, y_j)$. The support of the basis function $\phi_{i,j}$ is illustrated in Figure 7.8.

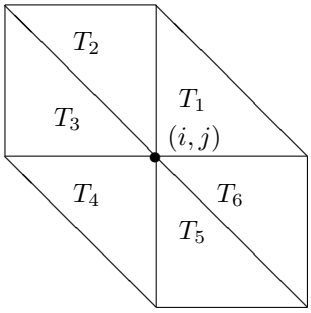


Figure 7.8: The support of $\phi_{i,j}$.

If we calculate the pyramid function $\phi_{i,j}$ explicitly, we find

$$\phi_{i,j}(x,y) = \begin{cases} 1 - \frac{x-x_i}{h} - \frac{y-y_j}{h}, & \text{if } (x,y) \in T_1; \\ 1 - \frac{y-y_j}{h}, & \text{if } (x,y) \in T_2; \\ 1 + \frac{x-x_i}{h}, & \text{if } (x,y) \in T_3; \\ 1 + \frac{x-x_i}{h} + \frac{y-y_j}{h}, & \text{if } (x,y) \in T_4; \\ 1 + \frac{y-y_j}{h}, & \text{if } (x,y) \in T_5; \\ 1 - \frac{x-x_i}{h}, & \text{if } (x,y) \in T_6; \\ 0, & \text{otherwise.} \end{cases}$$

We are now in position to calculate $\nabla\phi_{i,j} = ((\phi_{i,j})_x, (\phi_{i,j})_y)$, so that we can, in turn, calculate the entries (7.24) of the matrix A in the linear system of the finite element method described above. It easy to see that

$$(\phi_{i,j})_x(x,y) = \begin{cases} -\frac{1}{h}, & \text{if } (x,y) \in T_1; \\ 0, & \text{if } (x,y) \in T_2; \\ \frac{1}{h}, & \text{if } (x,y) \in T_3; \\ \frac{1}{h}, & \text{if } (x,y) \in T_4; \\ 0, & \text{if } (x,y) \in T_5; \\ -\frac{1}{h}, & \text{if } (x,y) \in T_6; \\ 0, & \text{otherwise.} \end{cases}, \quad \text{and} \quad (\phi_{i,j})_y(x,y) = \begin{cases} -\frac{1}{h}, & \text{if } (x,y) \in T_1; \\ -\frac{1}{h}, & \text{if } (x,y) \in T_2; \\ 0, & \text{if } (x,y) \in T_3; \\ \frac{1}{h}, & \text{if } (x,y) \in T_4; \\ \frac{1}{h}, & \text{if } (x,y) \in T_5; \\ 0, & \text{if } (x,y) \in T_6; \\ 0, & \text{otherwise.} \end{cases}$$

It is now not too hard to see (just need to check every single case) that the resulting matrix

$$A = -\frac{1}{h^2} \begin{pmatrix} B & I & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ I & B & I & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & I & B & I & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & I & B & I \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & I & B \end{pmatrix}, \tag{7.26}$$

where $\mathbf{0}$ is the $N \times N$ zero matrix, I is the $N \times N$ identity matrix, and B is $N \times N$ matrix

$$B = \begin{pmatrix} -4 & 1 & 0 & 0 & \dots & 0 \\ 1 & -4 & 1 & 0 & \dots & 0 \\ 0 & 1 & -4 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \dots \\ 0 & \dots & 0 & 1 & -4 & 1 \\ 0 & \dots & 0 & 0 & 1 & -4 \end{pmatrix}.$$

The corresponding linear system for this case reads (after multiplication of the equation by $-h^2$) $-h^2 AU =$

$-h^2F$:

$$\begin{pmatrix} B & I & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ I & B & I & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & I & B & I & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & I & B & I \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & I & B \end{pmatrix} \begin{pmatrix} U_{1,1} \\ U_{2,1} \\ \vdots \\ U_{N,1} \\ U_{1,2} \\ \vdots \\ U_{N,2} \\ \vdots \\ U_{1,N} \\ \vdots \\ U_{N,N} \end{pmatrix} = -h^2 \begin{pmatrix} F_{1,1} \\ F_{2,1} \\ \vdots \\ F_{N,1} \\ F_{1,2} \\ \vdots \\ F_{N,2} \\ \vdots \\ F_{1,N} \\ \vdots \\ F_{N,N} \end{pmatrix}, \tag{7.27}$$

where

$$F_{i,j} = \int_\Omega f \phi_{i,j} \mathrm{d}V. \tag{7.28}$$

Comparing this system with (7.27) which is the linear system arising from the five point difference method for the same problem, we can see that the matrix arising from the five point difference method and the one above are identical! Nevertheless, we again have different right-hand sides.

Hence, for this simple geometry Ω and for this particular triangulation described in Figure **??**, the five point method and the finite element method are the same, up to the right-hand side!

7.6.1 Error analysis of the finite element method

We shall now briefly consider the extension of the analysis given in Section 7.4.1 to the Poisson problem (7.20) for $\Omega \subset \mathbb{R}^d$, $d = 2, 3$. The only difficulty in extending the error analysis of the finite element method from the 1-dimensional Poisson problem to the general d-dimensional Poisson problem is in proving the relevant interpolation error bound replacing that of Theorem 7.16.

Consider problem (7.20) and the corresponding finite element problem (7.23). Both problems are well posed. We analyse the error between the exact weak solution u and its finite element approximation u_h using the norm

$$\|v\|_{\mathcal{H}} := \left(\int_a^b (\nabla v)^2 dV \right)^{1/2} \qquad \forall v \in \mathcal{H}. \tag{7.29}$$

Hence our goal will be to quantify $\|u - u_h\|_{\mathcal{H}}$. To this end, we just need to reproduce the steps followed to analyse the 1-dimensional problem. The proof is left as an exercise.

First, the *Galerkin Orthogonality* reads:

$$\int_\Omega \nabla(u - u_h) \cdot \nabla v_h \, dV = 0 \quad \text{for all } v_h \in \mathcal{H}_h, \tag{7.30}$$

from which the finite element *optimality* result follows:

Theorem 7.19 *Among all functions in the space $\mathcal{H}_h$, the finite element solution u_h given by* (7.9) *is the closest to the weak solution u. That is,*

$$\|u - u_h\|_{\mathcal{H}} = \min_{v_h \in \mathcal{H}_h} \|u - v_h\|_{\mathcal{H}}. \tag{7.31}$$

Proof. See Exercise 41 below. □

Now, we define the finite element interpolant.

Definition 7.20 *Given $v \in C^0(\Omega)$, the* interpolant *v_I of v is the function in $\mathcal{H}_h$ that is equal to v at the nodes. Thus,*

$$v_I := \sum_{i=1,j}^{N} v(x_i, y_j)\phi_{i,j}.$$

The following result gives an estimate of the distance between a function and its interpolant.

Theorem 7.21 *Let $u \in \mathcal{H}$. If all second weak partial derivatives of u belong to $L^2(\Omega)$ then*

$$\|u - u_I\|_{\mathcal{H}} \leq Ch(\|u_x\|_{\mathcal{H}}^2 + \|u_y\|_{\mathcal{H}}^2)^{1/2}, \tag{7.32}$$

for some constant C that does not depend on h.

Proof. The proof of this fact is beyond the scope of these notes. □

As n easy consequence, we now have the following a priori error bound.

Theorem 7.22 *If all second weak partial derivatives of the exact solution u belong to $L^2(\Omega)$ then*

$$\|u - u_h\|_{\mathcal{H}} \leq Ch(\|u_x\|_{\mathcal{H}}^2 + \|u_y\|_{\mathcal{H}}^2)^{1/2},$$

for some constant C that does not depend on h.

Proof. See Exercise 41 below. □

Problem

41. Prove the Galerkin Orthogonality relation (7.30), Theorem 7.19, and Theorem 7.22.

Chapter 8

The Finite Element Method for Parabolic Problems

Having derived finite element methods for elliptic problems, we are now in position to consider FEM for parabolic problems, by combining the ideas of finite difference time-stepping and finite element analysis for the "space" variable(s).

8.1 Finite element method in one space dimensions

For simplicity, we analyse parabolic problems in one space dimension first.

Consider the heat equation with a heat source with solution $u(t,x) : [0,T_f] \times \Omega \to \mathbb{R}$, with $\Omega = [a,b]$, $a < b$, hence

$$\begin{aligned} u_t(t,x) - u_{xx}(t,x) &= f(t,x) \quad \text{for all } t \in [0,T_f] \text{ and } x \in (a,b), && (8.1)\\ u(t,a) = u(t,b) &= 0, \qquad\quad \text{for all } t \in [0,T_f], && (8.2)\\ u(0,x) &= u_0(x), \qquad \text{for all } x \in (a,b), && (8.3) \end{aligned}$$

where $f : (0,T_f) \times (a,b) \to \mathbb{R}$ and $u_0 : (0,T_f) \to \mathbb{R}$ are known functions.

The first step in defining a finite element method is to rewrite the problem (8.1) in weak form, using integration by parts. Quite similarly to the case of elliptic problems, we define $\mathcal{H}$ to be the family of functions $v \in \mathcal{H}$, which are differentiable in (a,b) and satisfy the Dirichlet boundary condition $v(a) = v(b) = 0$.

Next, we multiply the PDE by a *test function* $v \in \mathcal{H}$, to get[1]

$$u_t v - u_{xx}\, v = fv.$$

Now we integrate over the domain Ω:

$$\int_a^b u_t v \,\mathrm{d}x - \int_a^b u_{xx}\, v \,\mathrm{d}x = \int_a^b fv \,\mathrm{d}x.$$

and perform integration by parts to the second integral on the left-hand side, to get

$$\int_a^b u_t v \,\mathrm{d}x + \int_a^b u_x v_x \,\mathrm{d}x - [u_x(t,x)v(x)]_a^b = \int_a^b fv \,\mathrm{d}x.$$

for all $v \in \mathcal{H}$. Using the fact that $v(a) = v(b) = 0$, we arrive to

$$\int_a^b u_t v \,\mathrm{d}x + \int_a^b u_x v_x \,\mathrm{d}x = \int_a^b fv \,\mathrm{d}x.$$

for all $v \in \mathcal{H}$. Hence, the heat problem can be transformed to the following problem in *weak form*

$$\text{For each } t \in (0,T_f], \text{ find } u(t,\cdot) \in \mathcal{H} \text{ s.t.} \int_a^b u_t v \,\mathrm{d}x + \int_a^b u_x v_x \,\mathrm{d}x = \int_a^b fv \,\mathrm{d}x, \quad \text{for all } v \in \mathcal{H}. \tag{8.4}$$

[1] Note that the solution $u = u(t,x)$ while the test function $v = v(x)$ is function of x only. Here and in the following we shall omit the dependence on t and x to simplify the notation.

The second step in defining the finite element method is to consider an approximation to the problem (8.4).

We do this in two steps. First, we use the linear finite element method of the previous chapter to discretise the space variable. The resulting method will be "discrete in space" but continuous in time. The second step is to use any of our favourite finite difference approximation for the time variable.

Space discretisation. To discretise the space variable, we shall consider the piecewise linear finite element space already considered in Section 7.4. To construct a suitable family $\mathcal{H}_h$, we consider equally distributed points $x_0 < x_1 < \cdots < x_{N_x+1}$, at distance h between them, such that

$$a = x_0,\ x_1 = x_0 + h,\ x_2 = x_1 + h,\ \ldots,\ x_N = x_{N-1} + h,\ x_{N_x+1} = b.$$

We then choose the family $\mathcal{H}_h$ to be the family of continuous functions $v_h \in \mathcal{H}_h$ that are linear (i.e., straight lines) at each interval $[x_{i-1}, x_i]$, for $i = 1, \ldots, N_x + 1$ and $v_h(x_0) = 0 = v_h(x_{N_x+1})$.

Hence we write the following *spatially semidiscrete finite element method*:

$$\begin{aligned}&\text{For each } t \in (0, T_f], \text{ find } u_h(t,\cdot) \in \mathcal{H}_h \text{ s.t.}\\ &\int_a^b (u_h)_t v_h \,\mathrm{d}x + \int_a^b (u_h)_x (v_h)_x \,\mathrm{d}x = \int_a^b f v_h \,\mathrm{d}x, \quad \text{for all } v_h \in \mathcal{H}_h,\end{aligned} \tag{8.5}$$

starting from[2] $u_h(0, x) = u_0(x)$. Note that problem (8.5) is simply the restriction of problem (8.4) from the space $\mathcal{H}$ to the space $\mathcal{H}_h$.

Time discretisation. In view of obtaining a fully discrete method we now discretise the time variable too. One way of doing this is to consider a finite difference approximation to the "time"-derivative u_t, by backward Euler method say. To this end, we consider an equally distributed subdivision $t_0 < t_1 < \cdots < t_{N_t}$, at distance τ between them, such that

$$0 = t_0,\ t_1 = t_0 + \tau,\ t_2 = t_1 + \tau,\ \ldots,\ t_{N_t-2} = t_{N_t-1} + \tau,\ t_{N_t} = T_f,$$

in the "time"-direction; hence, we have $\tau = 1/N_t$.

We formally make the following approximations, using backward difference for the time-derivative

$$u_t(t,x) \approx \delta^t_{\tau,-} u(t,x) = \frac{u(t,x) - u(t-\tau,x)}{\tau},$$

and for $t = t_n$, we set $u_h^n(x)$ to be the approximations of $u(t_n, x)$ for $x \in \Omega$. Hence, the heat problem is "approximated in space and time" by the family of problems in weak form:

$$\begin{aligned}&\text{For each } n = 1, \ldots, N_t, \text{ find } u_h^n \in \mathcal{H}_h \text{ s.t.}\\ &\int_a^b \frac{u_h^n - u_h^{n-1}}{\tau} v_h \,\mathrm{d}x + \int_a^b (u_h^n)_x (v_h)_x \,\mathrm{d}x = \int_a^b f v_h \,\mathrm{d}x, \quad \text{for all } v_h \in \mathcal{H}_h,\end{aligned}$$

starting from $u_h^0(x) = u_0(x)$. This is the *finite element method with backward Euler time-stepping*. Of course, different choices of divided differences for the time-derivative, different meshes and different choices of basis functions lead to different finite element methods. For each time-level n, the approximation u_h^{n-1} will be known (having been calculated in the previous step). Hence we can rearrange the approximate problem as

$$\begin{aligned}&\text{Set } u_h^0(x) = u_0(x) \text{ and for each } n = 1, \ldots, N_t, \text{ find } u_h^n \in \mathcal{H}_h \text{ s.t.}\\ &\int_a^b u_h^n v_h \,\mathrm{d}x + \tau \int_a^b (u_h^n)_x (v_h)_x \,\mathrm{d}x = \tau \int_a^b f v_h \,\mathrm{d}x + \int_a^b u_h^{n-1} v_h \,\mathrm{d}x, \quad \text{for all } v_h \in \mathcal{H}_h.\end{aligned}$$

In order to write the finite element method in algebraic form, we now introduce a basis for the space $\mathcal{H}_h$. For every $i = 1, \ldots, N_x$, we define the "hat" function ϕ_i as the unique function in $\mathcal{H}_h$ such that $\phi_i(x_i) = 1$ and $\phi_i(x_j) = 0$, for all $j \neq i$. Hence, any function $v_h \in \mathcal{H}_h$ can be written as

$$v_h = \sum_{i=1}^{N} V_i \phi_i,$$

[2] In practice some form of approximation of u_0 is used, either by interpolation or projection.

for $V_i \in \mathbb{R}$ being the "height" of the function at the node i. Since also $u_h^n \in \mathcal{H}_h$, we have $u_h^n = \sum_{j=1}^N U_j^n \phi_j$, for some $U_j^n \in \mathbb{R}$ and, therefore the problem becomes

$$\text{For each } n = 1, \ldots, N_t, \text{ find } U_j^n, j = 1, \ldots, N_x, \text{ s.t.}$$
$$\sum_{j=1}^N U_j^n \int_a^b \phi_j \phi_i \mathrm{d}x + \tau \sum_{j=1}^N U_j^n \int_a^b (\phi_j)_x (\phi_i)_x \mathrm{d}x = \int_a^b (\tau f + u_h^{n-1})\, \phi_i \mathrm{d}x,$$

for all $i = 1, \ldots, N_x$. For each time-level n, the approximation u_h^{n-1} will be known (having been calculated in the previous step). Thus, to find u_h^n for every $n = 1, \ldots, N_t$, we have to solve the linear system of N_x equations with N_x unknowns

$$(M + \tau A)\mathbf{U}^n = \mathbf{G}^n \tag{8.6}$$

for the unknown vector $\mathbf{U}^n = (U_1^n, \ldots, U_{N_x}^n)^T$, with respect to the matrices $M = [m_{ij}]_{i,j=1}^{N_x}$,$A = [a_{ij}]_{i,j=1}^{N_x}$, and the right-hand side vector $\mathbf{G} = (G_1^n, \ldots, G_N^n)^T$, where

$$m_{i,j} = \int_a^b \phi_j \phi_i \mathrm{d}x, \quad a_{i,j} = \int_a^b (\phi_j)_x (\phi_i)_x \mathrm{d}x, \quad G_i^n = \int_a^b (\tau f + u_h^{n-1})\, \phi_i \mathrm{d}x.$$

By comparison with the situation seen for the elliptic two-points boundary value problem of Section 7.4, note that the novelty here is in the appearance of the so called *mass matrix* M. The details are left as an exercise, cf. Problem 36 below. The above system can be solved for each time step given the solution at the previous time step, staring from the initial condition.

Problem

42. Arguing similarly to Section 7.4, compute the entries of the mass matrix M appearing in the linear system of equations (8.6).

43. Consider the initial-boundary value problem with mixed boundary conditions

$$\begin{aligned} u_t(t,x) - u_{xx}(t,x) = &\quad f(t,x), \quad t \in [0,T_f], \in (a,b) \\ u_x(t,a) = &\quad 0, \qquad \text{for all } t \in [0,T_f], \\ u(t,b) = &\quad 0, \qquad \text{for all } t \in [0,T_f], \\ u(0,x) = &\quad u_0(x), \qquad \text{for all } x \in (a,b), \end{aligned}$$

Write the problem in weak form and comment on how to construct a finite element method for it based on Crank-Nicolson time-stepping.

8.2 Finite element method in two and three space dimensions

As we saw, the treatment of elliptic problems in two and three dimensions is completely analogous to the treatment of the two-point boundary value problem. Therefore, in the discussion below, we shall adopt the general case of the heat initial/bounary value problem in two or three dimensions in "space", plus the "time" variable. (In facts, the one-space dimension case we just worked out is also pretty similar!)

Let us consider once again the heat equation with a heat source with solution $u(t,x) : [0,T_f] \times \Omega \to \mathbb{R}$, with, this time, $\Omega \subset \mathbb{R}^d$ open bounded domain, $d = 2, 3$:

$$\begin{aligned} u_t - \Delta u =& f, && \text{for } t \in (0,T_f] \text{ and } x \in \Omega \\ u(t,x) =& 0, && \text{for } t \in (0,T_f] \text{ and } x \in \partial\Omega, \\ u(0,x) =& u_0(x), && \text{for } x \in \Omega, \end{aligned} \tag{8.7}$$

where $f, u_0 : \Omega \to \mathbb{R}$ known functions. For simplicity, we shall assume (as in the case of elliptic problems) that the boundary $\partial\Omega$ of the domain Ω is composed by consecutive line segments (i.e., $\partial\Omega$ is a polygon).

The first step in defining a finite element method is to rewrite the problem (8.7) in weak form, using the *divergence theorem*. We define $\mathcal{H}$ to be the family of functions $v \in \mathcal{H}$, which are differentiable in Ω and satisfy the Dirichlet boundary condition $v = 0$ on $\partial\Omega$.

Next, we multiply the PDE by a *test function* $v \in \mathcal{H}$, to get

$$u_t v - \Delta u\, v = f v,$$

and we integrate over the domain Ω:

$$\int_\Omega u_t v z - \int_\Omega \Delta u\, v \mathrm{d}V = \int_\Omega f v \mathrm{d}V.$$

Now, if we perform and integration by parts in d–dimensions (see (7.21)) to the second integral on the left-hand side, we get

$$\int_\Omega u_t v \mathrm{d}V + \int_\Omega \nabla u \cdot \nabla v \mathrm{d}V - \int_{\partial\Omega} v \nabla u \cdot \vec{n} \mathrm{d}S = \int_\Omega f v \mathrm{d}V.$$

for all $v \in \mathcal{H}$, where $\vec{n}$ denotes the unit outward normal vector to the boundary $\partial\Omega$. Using the fact that $v = 0$ on $\partial\Omega$ for all $v \in \mathcal{H}$, we arrive to

$$\int_\Omega u_t v \mathrm{d}V + \int_\Omega \nabla u \cdot \nabla v \mathrm{d}V = \int_\Omega f v \mathrm{d}V.$$

for all $v \in \mathcal{H}$. Hence, the heat problem can be transformed to the following problem in *weak form*

$$\text{For each } t \in (0, T_f], \text{ find } u(t,\cdot) \in \mathcal{H} \text{ s.t.} \int_\Omega u_t v \mathrm{d}V + \int_\Omega \nabla u \cdot \nabla v \mathrm{d}V = \int_\Omega f v \mathrm{d}V, \quad \text{for all } v \in \mathcal{H}. \tag{8.8}$$

The second step in defining the finite element method is to consider an approximation to the problem (8.8). For simplicity, we shall consider only the case $d = 2$, i.e., when Ω is a bounded set on the plane; the case $d = 3$ is analogous. As with the one-dimensional problem, we shall proceed in two steps.

Space discretisation. To approximate the problem in "space", we shall restrict the family of eligible solutions to a smaller family of functions, $\mathcal{H}_h$. We split the domain Ω into elements $T \in \mathcal{T}$ in complete analogy to the case of elliptic problems, see, e.g., Figure 7.4. We consider the family $\mathcal{H}_h$ to consist of the functions that are continuous and linear on every element $T \in \mathcal{T}$, see Figure 7.5 for an illustration of such function. Suppose the triangulation contains N_x nodes that are not on the boundary $\partial\Omega$. We can then choose (arbitrarily) a numbering between $1, \ldots, N_x$ for the nodes, and we recall that "pyramid" basis function ϕ_i for each node $i = 1, \ldots, N_x$ described in the previous section.

Having the discrete solution space $\mathcal{H}_h$ in our hands, we proceed with the discretisation of the space variable by restriction of the weak problem (8.8):

$$\begin{aligned} &\text{For each } t \in (0, T_f], \text{ find } u_h(t,\cdot) \in \mathcal{H}_h \text{ s.t.} \\ &\int_\Omega (u_h)_t v_h \mathrm{d}V + \int_\Omega \nabla u_h \cdot \nabla v_h \mathrm{d}V = \int_\Omega f v_h \mathrm{d}V, \quad \text{for all } v_h \in \mathcal{H}_h. \end{aligned} \tag{8.9}$$

where we set[3] $u^0(x) = u_0(x)$. This is the *spatially semidiscrete finite element method.*

Time discretisation We shall use once again the backward Euler method to approximate the "time"-derivative u_t on N_t equally spaced time-steps of lenght τ. As before, for $t = t_n$, we set $u^n(x)$ to be the approximations of $u(t_n, x)$ for $x \in \Omega$. Hence, the heat problem is "approximated in both time and space" by the family of problems in weak form

$$\begin{aligned} &\text{For each } n = 1, \ldots, N_t, \text{ find } u_h^n \in \mathcal{H}_h \text{ s.t.} \\ &\int_\Omega \frac{u_h^n - u_h^{n-1}}{\tau} v_h \mathrm{d}V + \int_\Omega \nabla u_h^n \cdot \nabla v_h \mathrm{d}V = \int_\Omega f v_h \mathrm{d}V, \quad \text{for all } v_h \in \mathcal{H}_h, \end{aligned}$$

where $u_h^0(x) = u_0(x)$. This is the *finite element method with backward Euler time-stepping.* Of course, different choices of divided differences for the time-derivative, different meshes and different choices of basis functions lead to different finite element methods. For each time-level n, the approximation u_h^{n-1} will be known (having been calculated in the previous step). We can rearrange the approximate problem then as

$$\begin{aligned} &\text{For each } n = 1, \ldots, N_t, \text{ find } u_h^n \in \mathcal{H}_h \text{ s.t.} \\ &\int_\Omega u_h^n v_h \mathrm{d}V + \tau \int_\Omega \nabla u_h^n \cdot \nabla v_h \mathrm{d}V = \tau \int_\Omega f v_h \mathrm{d}V + \int_\Omega u_h^{n-1} v_h \mathrm{d}V, \quad \text{for all } v_h \in \mathcal{H}_h. \end{aligned}$$

[3] As in the one-dimensional case, in practice some form of approximation of u_0 is used, either by interpolation or projection.

Thus, since every $v_h \in \mathcal{H}_h$ can be written as a linear combination of "pyramid" functions, we can instead equivalently ask

For each $n = 1, \ldots, N_t$, find $u_h^n \in \mathcal{H}_h$ s.t.

$$\int_\Omega u_h^n \phi_i \mathrm{d}V + \tau \int_\Omega \nabla u_h^n \cdot \nabla \phi_i \mathrm{d}V = \tau \int_\Omega f \phi_i \mathrm{d}V + \int_\Omega u_h^{n-1} \phi_i \mathrm{d}V, \quad \text{for all } i = 1, \ldots, N_x.$$

Also, since $u_h^n \in \mathcal{H}_h$, too, we have $u_h^n = \sum_{j=1}^N U_j^n \phi_j$, for some $U_j^n \in \mathbb{R}$ and, therefore the problem becomes

For each $n = 1, \ldots, N_t$, find $U_j^n, j = 1, \ldots, N_x,$ s.t.

$$\sum_{j=1}^N U_j^n \int_\Omega \phi_j \phi_i \mathrm{d}V + \tau \sum_{j=1}^N U_j^n \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \mathrm{d}V = \int_\Omega (\tau f + u_h^{n-1}) \, \phi_i \mathrm{d}V, \quad \text{for all } i = 1, \ldots, N_x.$$

Thus, to find u_h^n for every $n = 1, \ldots, N_t$, we have to solve a linear system of N_x equations with N_x unknowns and can be written in matrix form as $(M + \tau A)\mathbf{U}^n = \mathbf{G}^n$, for $M = [m_{ij}]_{i,j=1}^{N_x}$ and $A = [a_{ij}]_{i,j=1}^{N_x}$, $\mathbf{U}^n = (U_1^n, \ldots, U_N^n)^T$ and $\mathbf{G} = (G_1^n, \ldots, G_N^n)^T$, where

$$m_{ij} = \int_\Omega \phi_j \phi_i \mathrm{d}V, \quad a_{ij} = \int_\Omega \nabla \phi_j \cdot \nabla \phi_i \mathrm{d}V, \quad \text{and} \quad G_i^n = \int_\Omega (\tau f + u_h^{n-1}) \phi_i \mathrm{d}V.$$

Example 8.1 *We use the finite element method to approximate the solution to the heat problem with homogeneous Dirichlet boundary conditions:*

$$\begin{aligned} u_t - \Delta u &= 0, && \text{for } t \in (0,1] \text{ and } x \in \Omega \\ u(t,x) &= 0, && \text{for } t \in (0,1] \text{ and } x \in \partial\Omega, \\ u(0,x) &= 1, && \text{for } x \in \Omega, \end{aligned} \tag{8.10}$$

where Ω is given by the domain in Figure 7.4, along with the mesh used in the approximation. The finite element approximation (using $N_t = 100$, for different times t is shown in Figure 8.1.

Problem

44. Consider the heat equation problem with mixed boundary conditions

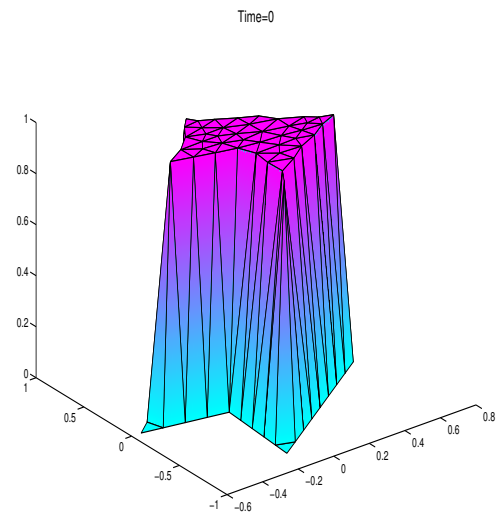
$$\begin{aligned} u_t - \Delta u + u &= f, && t \in (0, T_f],\ (x,y) \in (0,1)^2 \\ u(t,x,0) = u(t,0,y) = u(t,x,1) &= 0, && t \in (0, T_f],\ 0 \le x, y \le 1 \\ u_x(t,1,y) &= 1, && t \in (0, T_f],\ 0 \le y \le 1 \\ u(0,x,y) &= u_0(x,y), && 0 \le x, y \le 1. \end{aligned}$$

Write the problem in weak form and comment on how to construct a finite element method with forward Euler time-stepping.

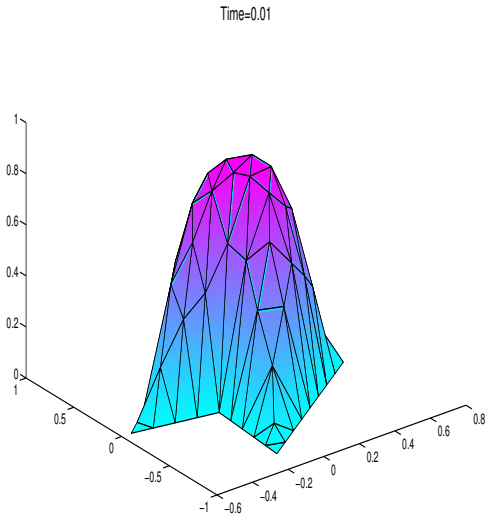
8.3 Error analysis of the finite element method

We analyse the finite element method for the solution of parabolic problems seen in the previous sections. Recall that the finite element method was only used to discretise the space variable. Our analysis will concentrate on the error due to such approximation. In other words we shall only analyse the spatially semidiscrete method. The error for the fully discrete method (discrete in both time and space) depends on the particular choice of finite difference discretisation of the time variable. The analysis of the fully discrete method is beyond the scope of these notes, but typical orders of convergence should be expected.

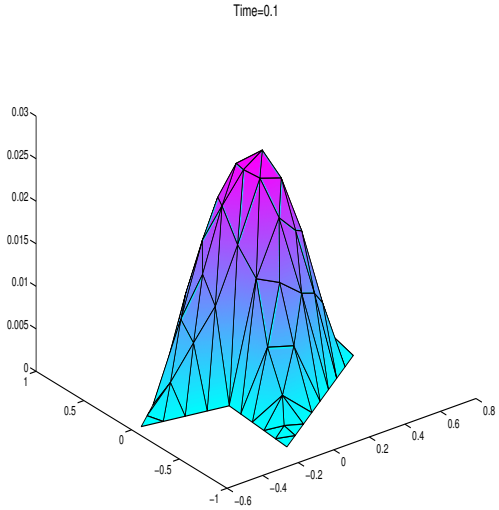
We shall here consider together the formulations in one, two, and three space dimensions: the only difference between the various situations being in the notations! Hence we consider both problem (8.1) and problem (8.7) and the corresponding spatially semidiscrete finite element problems (8.5) and (8.9). Both problems are well posed. We analyse the error between the exact weak solution u and its finite element approximation u_h by bounding the norm $\|u - u_h\|_{\mathcal{H}}$ given by either (7.10) or (7.29). To this end, it is



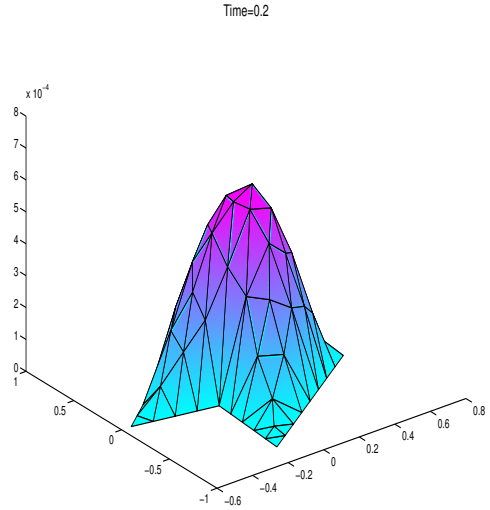
(a) $t = 0$



(b) $t = 0.01$



(c) $t = 0.1$



(d) $t = 0.2$

Figure 8.1: Finite element approximation to the heat problem

practical as well as instructive to write down the problems with a common notation irrespective of the space dimension. The heat problem with heat force in weak form for $d = 1, 2, 3$ can be rewritten as:

$$\text{For each } t \in (0, T_f], \text{ find } u(t,\cdot) \in \mathcal{H} \text{ s.t.} \quad (u_t, v) + a(u, v) = (f, v), \quad \text{for all } v \in \mathcal{H}. \tag{8.11}$$

Here the symbol $(\cdot,\cdot)$ denotes the $L^2(\Omega)$–inner product, with $\Omega \subset \mathbb{R}^d$ the space domain of the problem. For instance, if $d = 1$, we have simply $\Omega = (a, b)$. The $L^2(\Omega)$–inner product is defined as

$$(w, v) = \int_\Omega w\, v \,\mathrm{d}V, \qquad \forall w, v \in L^2(\Omega).$$

It induces the L^2-norm

$$\|v\|_{L^2} := (v, v)^{1/2},$$

which is also a norm most commonly used to analyse finite element methods. The symbol $a(\cdot,\cdot)$ stands for the *bilinear form*[4] given by the spatial operator of the weak problem. In our case, for $d = 2, 3$, it is

$$a(u, v) = \int_\Omega \nabla w \cdot \nabla v \,\mathrm{d}V \qquad \forall w, v \in \mathcal{H},$$

and similarly for $d = 1$ the bilinear form is defined by the second term in the left hand side of (8.4).

Recall that well-posedness of the initial-boundary value problem for the heat equation was discussed in Section **??**. We are not going to prove here that the associated weak problem (8.11) is well-posed as well. We will limit ourselves to verify that the solution of the weak problem depends continuously on the data. We do this by deriving an *energy stimate*. Assume that u is the solution of (8.11). Then, in particular, the equation must be satisfied for $v = u_t(t,\cdot)$. Hence,

$$(u_t, u_t) + a(u, u_t) = (f, u_t). \tag{8.12}$$

Now we have:

$$a(u, u_t) = \int_\Omega \nabla u \cdot \nabla u_t \,\mathrm{d}V = \frac{1}{2}\int_\Omega (|\nabla u|^2)_t \,\mathrm{d}V = \frac{1}{2}\frac{d}{dt}\int_\Omega |\nabla u|^2 \,\mathrm{d}V = \frac{1}{2}\frac{d}{dt}\|u\|^2_{\mathcal{H}},$$

and, by definition of the L^2-norm,

$$(u_t, u_t) = \|u_t\|^2.$$

Finally, since $(\cdot,\cdot)$ is an inner product, the Cauchy-Schwarz inequality holds, hence

$$(f, u_t) \le (f, f)^{1/2}(u_t, u_t)^{1/2} = \|f\|\,\|u_t\| \le \frac{1}{2}(\|f\|^2 + \|u_t\|^2).$$

Using these results into (8.12) we get, for all $t \in (0, T_f]$,

$$\|u_t\|^2 + \frac{1}{2}\frac{d}{dt}\|u\|^2_{\mathcal{H}} \le \frac{1}{2}(\|f\|^2 + \|u_t\|^2),$$

yielding

$$\|u_t\|^2 + \frac{d}{dt}\|u\|^2_{\mathcal{H}} \le \|f\|^2.$$

Hence, integrating this last inequality over $(0, t)$ we get

$$\int_0^t \|u_t\|^2 \,\mathrm{d}t + \|u\|^2_{\mathcal{H}} \le \|u_0\|^2_{\mathcal{H}} + \int_0^t \|f\|^2 \,\mathrm{d}t. \tag{8.13}$$

In particular, we have that

$$\|u\|^2_{\mathcal{H}} \le \|u_0\|^2_{\mathcal{H}} + \int_0^t \|f\|^2 \,\mathrm{d}t.$$

That is to say: the $\mathcal{H}$-norm (or energy norm) of the solution is bounded by appropriate norms of the problem's data.

[4] A bilinear form is a function of two variables which is linear with respect to both variables

Now, the spatially semidiscrete finite element discretisation of problem (8.11) reads:

$$\text{For each } t \in (0, T_f], \text{ find } u_h(t,\cdot) \in \mathcal{H}_h \text{ s.t.} \quad ((u_h)_t, v_h) + a(u_h, v_h) = (f, v_h), \quad \text{for all } v_h \in \mathcal{H}_h. \tag{8.14}$$

Arguing exactly as before we could infer a bound identical to (8.13) for the finite element solution u_h. In other words, the finite element solution depends continuously on the data....this is encouraging!

The goal of this section is to quantify the error $u - u_h$.

Recall that we used the *same* discretisation method here as the one we used for the elliptic problems studied in the previous chapter. Let u be the solution of the weak problem (8.11). For any fixed $t \in [0, T_f]$ we may consider the finite element elliptic problem:

$$\text{Find } w_h \in \mathcal{H}_h \text{ s.t.} \quad a(w_h, v_h) = a(u(t,\cdot), v_h), \quad \text{for all } v_h \in \mathcal{H}_h. \tag{8.15}$$

In other words w_h is the finite element approximation of the weak elliptic problem in $\mathcal{H}$ whose exact solution is $u(t,\cdot)$ itself as trivially $a(u(t,\cdot), v) = a(u(t,\cdot), v)$ for all $v \in \mathcal{H}$. Hence, applying our error bounds of Theorem 7.17 if $d = 1$ and Theorem 7.22 if $d > 1$ we have

$$\|u(t,\cdot) - w_h\|_{\mathcal{H}} \leq Ch \left(\sum_{i=1}^{d} \|u_{x_i}(t)\|_{\mathcal{H}}^2 \right)^{1/2}, \tag{8.16}$$

for all $t \in [0, T_f]$. The idea is indeed to proceed with a divide-and-conquer strategy based on splitting the error into two using the triangle inequality

$$\|u - u_h\|_{\mathcal{H}} \leq \|u - w_h\|_{\mathcal{H}} + \|w_h - u_h\|_{\mathcal{H}}, \tag{8.17}$$

for all times $t \in [0, T_f]$. Notice that the first term we already quantify with (8.16). It remains to quantify the second term $\|w_h - u_h\|$. To this end, we consider the quantity

$$\begin{aligned}
((u_h - w_h)_t, v_h) + a(u_h - w_h, v_h) &= ((u_h)_t, v_h) + a(u_h, v_h) - ((w_h)_t, v_h) - a(w_h, v_h) \\
&= (f, v_h) - ((w_h)_t, v_h) - a(w_h, v_h) \\
&= (f, v_h) - ((w_h)_t, v_h) - a(u, v_h) \\
&= (f, v_h) - ((w_h)_t, v_h) - (f, v_h) + (u_t, v_h) \\
&= ((u - w_h)_t, v_h),
\end{aligned}$$

for all $v_h \in \mathcal{H}_h$, having used (8.14), (8.16), and (8.11) one after the other. Using the above equality with $v_h = (u_h - w_h)_t$ we get

$$\begin{aligned}
((u_h - w_h)_t, (u_h - w_h)_t) + a(u_h - w_h, (u_h - w_h)_t) &= ((u - w_h)_t, (u_h - w_h)_t) \\
&\leq ((u - w_h)_t, (u - w_h)_t)^{1/2} ((u_h - w_h)_t, (u_h - w_h)_t)^{1/2} \\
&\leq \frac{1}{2} \left(((u - w_h)_t, (u - w_h)_t) + ((u_h - w_h)_t, (u_h - w_h)_t) \right),
\end{aligned}$$

by the Cauchy-Schwarz applied to the L^2-inner product $(\cdot,\cdot)$. It now follows that

$$\frac{1}{2}((u_h - w_h)_t, (u_h - w_h)_t) + \frac{1}{2}\frac{d}{dt}[a(u_h - w_h, (u_h - w_h))] \leq \frac{1}{2}((u - w_h)_t, (u - w_h)_t),$$

and in particular

$$\frac{d}{dt}\|u_h - w_h\|_{\mathcal{H}}^2 = \frac{d}{dt}[a(u_h - w_h, (u_h - w_h))] \leq ((u - w_h)_t, (u - w_h)_t) = \|(u - w_h)_t\|_{L^2}^2.$$

Integrating in time between 0 and any $t \in (0, T_f]$ gives

$$\|(u_h - w_h)(t)\|_{\mathcal{H}}^2 \leq \|(u_h - w_h)(0)\|_{\mathcal{H}}^2 + \int_0^t \|(u - w_h)_t(t)\|_{L^2}^2 \, \mathrm{d}t. \tag{8.18}$$

It turns out that the error quantity $\|(u - w_h)_t\|_{L^2}$ can be bounded as follows (the proof of this result is beyond the scope of this notes; we just point out that this is just another quantification of the difference $u - w_h$):

$$\|(u - w_h)_t\|_{L^2} \leq Ch\|u_t(t)\|_{\mathcal{H}}. \tag{8.19}$$

Assume for simplicity that $u(0,\cdot) = u_h(0,\cdot) = w_h(0,\cdot)$ so that the initial error $\|(u_h - w_h)(0)\|_{\mathcal{H}} = 0$ appearing in (8.18) vanishes. In this case, using (8.19) in (8.18) we get

$$\|(u_h - w_h)(t)\|_{\mathcal{H}} \leq Ch \left(\int_0^t \|u_t(t)\|_{\mathcal{H}}^2 \, \mathrm{d}t \right)^{1/2}. \tag{8.20}$$

Using (8.20) and (8.16) in (8.17) we promptly obtain the following result.

Theorem 8.2 *Assume that $u_h(0,\cdot) = w_h(0,\cdot) = u(0,\cdot)$. If u is regular enough, then for all $t \in [0, T_f]$*

$$\|(u - u_h)(t)\|_{\mathcal{H}} \leq Ch \left[\left(\sum_{i=1}^{d} \|u_{x_i}(t)\|_{\mathcal{H}}^2 \right)^{1/2} + \left(\int_0^t \|u_t(t)\|_{\mathcal{H}}^2 \, \mathrm{d}t \right)^{1/2} \right],$$

for some constant C that does not depend on h.