

Part I — An introduction to partial differential equations

§1 Terminology and some examples

Many of the equations that arise in areas such as mathematical physics, fluid dynamics, mechanics, optics, heat flow, quantum mechanics etc are partial differential equations (PDEs). The derivatives in these equations represent natural physical quantities such as velocity, acceleration, force, flux, and current.

Definition. A PDE is an equation that contains partial derivatives of a single (unknown) function.

This is to be contrasted with ordinary differential equations (ODEs) in which the unknown function, u say, depends on only one variable, for example, $u(x)$.

In a PDE, the unknown function depends on at least two variables. For example:

$$u(x, t), \quad u(x, y, z, t).$$

In these notes, we use the standard notation for partial derivatives in which the subscripts indicate the variable with respect to which we take the partial derivative, that is,

$$u_x \equiv \frac{\partial u}{\partial x}, \quad u_{xx} \equiv \frac{\partial^2 u}{\partial x^2}, \quad u_{xt} \equiv \frac{\partial^2 u}{\partial t \partial x} \quad \text{etc.}$$

Example 1.1. Heat equation in 1D: $\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}$ or $u_t = \kappa u_{xx}$. ⊠

Example 1.2. Heat equation in 3D: $\frac{\partial u}{\partial t} = \kappa \nabla^2 u$, where $\nabla^2 u \equiv u_{xx} + u_{yy} + u_{zz}$. ⊠

Example 1.3. Laplace's equation on a circle in polar coordinates (so u is a function of radius r and angle θ):

$$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0. \quad \text{⊠}$$

Example 1.4. Wave equation in 3D: $u_{tt} = c^2 \nabla^2 u$. ⊠

§2 Solution methods

We are interested in finding the unknown function u . There are a number of solution methods available:

- (a) Separation of variables: a PDE in n variables is reduced to n ODEs.
- (b) Integral transforms: reduces a PDE in n variables to one in $n - 1$ variables.
- (c) Change of coordinates: a PDE is transformed into an ODE or an easier PDE via techniques such as rotation of axes.

- (d) Transformation of dependent variable: for example, $v(x, y) = \log [u(x, y)]$.
- (e) Numerical methods: often the only technique that will work, but get only approximations to the unknown function u .
- (f) Perturbation methods: changes a non-linear problem to a sequence of linear ones that approximate the original problem.
- (g) Impulse-response method: this decomposes the initial/boundary conditions into simple impulses and then finds the response to each impulse. The responses are then superposed—such a technique assumes/requires linearity.
- (h) Integral equations: convert the PDE into an integral equation in which the unknown function appears inside an integral. For example, it may be shown that the PDE

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} + F(x) \sin(\omega t),$$

with boundary conditions $u(0, t) = u(1, t) = 0$ has a solution of the form

$$u(x, t) = X(x) \sin(\omega t),$$

where X satisfies the integral equation

$$\begin{aligned} X(x) &= \int_0^1 k(x, w) \left(\frac{\omega^2}{c^2} X(w) - F(w) \right) dw \\ &= - \int_0^1 k(x, w) F(w) dw + \frac{\omega^2}{c^2} \int_0^1 k(x, w) X(w) dw, \end{aligned}$$

with

$$k(x, w) = \begin{cases} w(1-x), & w \leq x, \\ x(1-w), & x < w. \end{cases}$$

- (i) Calculus of variations: reformulate the PDE as a minimisation problem. The minimum of a certain expression (often the energy) is the solution of the PDE.
- (j) Eigenfunction expansion: the solution is of the form

$$\sum_{n=1}^{\infty} (\text{coeffs})(\text{eigenfunctions}),$$

where the eigenfunctions are found by solving the associated eigenvalue problem for the PDE.

§3 Classification

For PDEs, the general theory and methods of solution usually apply to a given class of equations.

Six basic classifications are:

1. Order: this is the order of the highest partial derivative in the equation. For example, $u_t = u_{xx}$ is second order.
2. Number of independent variables: the PDE $u_t = u_{xx}$ has the independent variables x and t .
3. Linearity: a PDE is linear if u and its partial derivatives appear in a linear fashion. For example, no powers, products, or functions of them such as u_t^2 , uu_{xx} , $\sin(u)$ etc. The most general second order linear PDE in two variables x and y is

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G. \quad (\text{I.1})$$

In (I.1), the coefficients $A(x, y), \dots, G(x, y)$ are continuous functions over some domain Ω of the x - y plane. $u(x, y)$ and/or its derivatives are to satisfy given conditions on the boundary of Ω .

4. Homogeneity: the equation (I.1) is homogeneous if $G(x, y) = 0$ for all $(x, y) \in \Omega$ and is non-homogeneous otherwise.
5. Kinds of coefficients: If the functions A – F in (I.1) are constants, then (I.1) is called a constant coefficient PDE; otherwise it is a variable coefficient PDE.
6. Four basic types of linear equations: all linear PDEs like (I.1) are either parabolic, hyperbolic, elliptic, or mixed.

(a) Parabolic—when the discriminant $B^2 - 4AC = 0$ in Ω . An example is the equation

$$u_t = \alpha^2 u_{xx}, \quad \alpha \in \mathbb{R},$$

which is the heat or diffusion equation. It can be used to model the temperature distribution in a rod or to model the diffusion of gases.

(b) Hyperbolic—when $B^2 - 4AC > 0$ in Ω . An example is the wave equation

$$u_{tt} = \alpha^2 u_{xx}, \quad \alpha \in \mathbb{R}.$$

A more complicated example is the telegraph equation

$$u_{xx} = KLu_{tt} + (KR + LS)u_t + RSu, \quad KL > 0.$$

In this last equation $u(x, t)$ represents the current or potential at time t at a point x from one end of a transmission line which has electrostatic potential K , self-inductance L , resistance R , and leakage conductance S . Hyperbolic PDEs usually arise when waves or vibrations occur in a physical system. Mathematical modelling of such systems usually involves solution of a hyperbolic equation or of a hyperbolic system.

(c) Elliptic—when $B^2 - 4AC < 0$ in Ω . Two examples are Laplace's equation given by

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

and Poisson's equation given by

$$u_{xx} + u_{yy} = G(x, y).$$

Both these equations can be used to model the steady-state or equilibrium temperature distribution in a plate.

(d) Mixed type—an equation can be of different types at different places in Ω . For example,

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

has

$$B^2 - 4AC = -4y \Rightarrow \begin{cases} \text{parabolic} & \text{for } y = 0, \\ \text{hyperbolic} & \text{for } y < 0, \\ \text{elliptic} & \text{for } y > 0. \end{cases}$$

However, mixed types are not common in physical systems.

There are also several grades of non-linearity. For example, consider a first order PDE in two variables x and y :

- linear: $A(x, y)u_x + B(x, y)u_y + C(x, y)u = D(x, y)$.
- semi-linear: $A(x, y)u_x + B(x, y)u_y = E(x, y, u)$.
- quasi-linear: $A(x, y, u)u_x + B(x, y, u)u_y = E(x, y, u)$.

In the above, the first order derivatives appear only to the first power and there are no products, etc. of derivatives. Often if a technique works on a linear equation, it will also work for the semi-linear and quasi-linear forms.

§4 PDEs and physical systems

The PDEs which model physical systems usually have many solutions. To select the single solutions that represents the solution to the physical system requires imposing certain auxiliary conditions specific to the system being modelled. These are of two categories.

(a) Boundary conditions: if a PDE in u holds in a domain Ω with boundary $\partial\Omega$, then at each point of the boundary, one needs to know that u satisfies one of the following:

- (i) $u = g$ (Dirichlet condition)
- (ii) $\frac{\partial u}{\partial \mathbf{n}} = g$ (Neumann or flux condition)
- (iii) $\alpha u + \beta \frac{\partial u}{\partial \mathbf{n}} = g$ (mixed or Robin condition)

Here g , α , and β are known/given functions on $\partial\Omega$ and $\frac{\partial u}{\partial \mathbf{n}} = \mathbf{n} \cdot \nabla u$ with \mathbf{n} the unit normal to the boundary.

(b) Initial conditions: conditions which must be satisfied throughout Ω at initial time $t = 0$.
For example, with $u = u(x, y, t)$:

$$u(x, y, 0) = f(x, y), \quad u_x(x, y, 0) = a(x, y), \quad u_y(x, y, 0) = b(x, y).$$

Definition. *The initial conditions, boundary conditions, and coefficients of the PDE and any non-homogeneous term in it comprise the data of the PDE.*

Definition. *The solution is said to depend continuously upon the data if small changes in the data produce correspondingly small changes in the solution.*

Definition. *A problem is well-posed if:*

- (i) a solution exists;*
- (ii) the solution is unique;*
- (iii) the solution depends continuously on the data.*

Otherwise, it is said to be ill-posed.

For the auxiliary conditions, that together with the PDE, comprise a well-posed problems, there cannot be too many (else the problem will have no solution) and there cannot be too few (else the solution will not be unique). Also, they must be of the correct type (else the solution will not depend continuously on the data).

Part II — Solution of the 1D heat equation

§1 Sturm-Liouville problems

When we consider the heat equation, we shall use the method of separation of variables, a technique seen previously in MATH255 and perhaps also MATH331. Therefore it is appropriate to first consider Sturm-Liouville problems as these arise when using separation of variables. These problems are also relevant when we consider eigenfunction expansions.

Definition. *The general form for a second order Sturm-Liouville problem is given by the differential equation*

$$[p(x)y']' - q(x)y + \mu r(x)y = 0, \quad x \in (a, b), \quad (\text{II.1})$$

along with boundary conditions

$$\alpha_1 y(a) + \alpha_2 y'(a) = 0, \quad \beta_1 y(b) + \beta_2 y'(b) = 0. \quad (\text{II.2})$$

The functions p , p' , q , r are assumed to be continuous on $[a, b]$, and it is further assumed that $p(x) > 0$ and $r(x) > 0$ for $x \in [a, b]$.

To see what linear second order differential equations may be written as Sturm-Liouville differential equations, consider the linear second order differential equation

$$a_0(x)y'' + a_1(x)y' + a_2(x)y = 0.$$

By dividing through by $a_0(x)$, we obtain

$$y'' + P(x)y' + Q(x)y = 0. \quad (\text{II.3})$$

Let

$$p(x) = e^{\int P(x) dx}$$

so that $p'(x) = p(x)P(x)$. Multiplying (II.3) through by $p(x)$ yields

$$p(x)y'' + p(x)P(x)y' + p(x)Q(x)y = 0,$$

or

$$(py')' + R(x)y = 0, \quad \text{where } R(x) = p(x)Q(x).$$

This last form is known as the *self-adjoint* form since the equation is self-adjoint. (An equation is said to be self-adjoint if the adjoint equation is the same as itself. However, the theory of the *adjoint* equation is beyond the scope of this paper.) We then see that if $R(x) = p(x)Q(x)$ is of the form $R = -q + \mu r$, then it is a differential equation of the Sturm-Liouville form. If the boundary conditions are appropriate, then we have a second order Sturm-Liouville problem.

As an example of the theory associated with Sturm-Liouville problems, we give the following theorem.

Theorem II.1. *The differential equation (II.1) with boundary conditions $y(0) = y(\ell) = 0$ has solutions for an infinite sequence of values of μ .*

Proof. Omitted. □

These values of μ are known as the eigenvalues and the corresponding solutions the eigenfunctions.

Example 1.1. For the Sturm-Liouville problem

$$y'' + \mu y = 0, \quad y(0) = y(\ell) = 0,$$

one may show that the eigenfunctions are given by $\varphi_n = \sin(\sqrt{\mu_n}x)$, where $\mu_n = n^2\pi^2/\ell^2$. □

The theory of Sturm-Liouville problems that is of interest to us in the solution of partial differential equations is given in the following theorem.

Theorem II.2. *The eigenfunctions of the Sturm-Liouville problem given by (II.1) and (II.2) are orthogonal on the interval $[a, b]$ with respect to the weight function r .*

Proof. Let φ_m and φ_n be eigenfunctions corresponding to *different* eigenvalues μ_m and μ_n . Hence they satisfy

$$(p\varphi_m')' + (-q + \mu_m r)\varphi_m = 0, \tag{II.4}$$

$$(p\varphi_n')' + (-q + \mu_n r)\varphi_n = 0. \tag{II.5}$$

Then $\varphi_n \times$ (II.4) $- \varphi_m \times$ (II.5) yields

$$\varphi_n(p\varphi_m')' - \varphi_m(p\varphi_n')' + (\mu_m - \mu_n)r\varphi_m\varphi_n = 0.$$

This may be written as

$$[p(\varphi_n\varphi_m' - \varphi_m\varphi_n')]_a^b + (\mu_m - \mu_n)r\varphi_m\varphi_n = 0.$$

Upon integration over $[a, b]$, we find

$$[p(\varphi_n\varphi_m' - \varphi_m\varphi_n')]_a^b + (\mu_m - \mu_n) \int_a^b r\varphi_m\varphi_n \, dx = 0,$$

or

$$p(b)W[\varphi_n(b), \varphi_m(b)] - p(a)W[\varphi_n(a), \varphi_m(a)] + (\mu_m - \mu_n) \int_a^b r\varphi_m\varphi_n \, dx = 0,$$

where $W[\varphi_n(x), \varphi_m(x)]$ is the Wronskian

$$W[\varphi_n(x), \varphi_m(x)] = \begin{vmatrix} \varphi_n(x) & \varphi_m(x) \\ \varphi_n'(x) & \varphi_m'(x) \end{vmatrix}.$$

However, since both φ_m and φ_n are solutions of the Sturm-Liouville problem, then the boundary conditions in (II.2) show that

$$\alpha_1 \varphi_m(a) + \alpha_2 \varphi_m'(a) = 0, \quad \alpha_1 \varphi_n(a) + \alpha_2 \varphi_n'(a) = 0.$$

By using $\varphi_m'(a) = -\alpha_1 \varphi_m(a)/\alpha_2$ and $\varphi_n'(a) = -\alpha_1 \varphi_n(a)/\alpha_2$ in the expression for $W[\varphi_n(a), \varphi_m(a)]$, it is easy to verify that $W[\varphi_n(a), \varphi_m(a)] = 0$. A similar argument shows that $W[\varphi_n(b), \varphi_m(b)] = 0$. Hence we conclude that

$$(\mu_m - \mu_n) \int_a^b r \varphi_m \varphi_n \, dx = 0.$$

Since $\mu_m \neq \mu_n$, the result follows. \square

An orthogonal set of functions is like an orthogonal basis for vectors. It is natural to ask whether we can expand any function out as an (infinite) linear combination of the basis functions, that is, if $\{\varphi_n\}$ is an orthogonal set such as ones that arise from Sturm-Liouville problems, can we expand any function f in the form

$$f(x) = \sum_{n=1}^{\infty} c_n \varphi_n(x)? \quad (\text{II.6})$$

If the answer is ‘yes’, the expansion given in (II.6) is called a *Fourier series*. Assuming (II.6) is valid, we can obtain expressions for the *Fourier coefficients* c_n . Let $\langle \cdot, \cdot \rangle$ be the inner product given by

$$\langle f, g \rangle = \int_a^b r(x) f(x) g(x) \, dx.$$

Then by taking the inner product of (II.6) with φ_m , we get

$$\langle f, \varphi_m \rangle = \sum_{n=1}^{\infty} c_n \langle \varphi_n, \varphi_m \rangle.$$

But $\langle \varphi_n, \varphi_m \rangle = 0$ for $n \neq m$ and hence

$$\langle f, \varphi_m \rangle = c_m \langle \varphi_m, \varphi_m \rangle = c_m \|\varphi_m\|^2.$$

Hence we conclude that

$$c_m = \frac{\langle f, \varphi_m \rangle}{\|\varphi_m\|^2} = \frac{\int_a^b r(x) f(x) \varphi_m(x) \, dx}{\int_a^b r(x) \varphi_m^2(x) \, dx}. \quad (\text{II.7})$$

Example 1.2. For $y'' + \mu y = 0$, $y(0) = y(\ell) = 0$, the eigenvalues are $\mu_n = (n\pi/\ell)^2$, $n \geq 1$, with eigenfunctions $\varphi_n(x) = \sin(n\pi x/\ell)$. The previous theorem then shows that

$$\int_0^\ell \sin(m\pi x/\ell) \sin(n\pi x/\ell) \, dx = 0 \quad \text{for } m \neq n.$$

Moreover, we have

$$f(x) = \sum_{n=1}^{\infty} c_n \sin(n\pi x/\ell),$$

where

$$c_n = \frac{\int_0^\ell f(x) \sin(n\pi x/\ell) dx}{\int_0^\ell \sin^2(n\pi x/\ell) dx} = \frac{2}{\ell} \int_0^\ell f(x) \sin(n\pi x/\ell) dx. \quad \square$$

Example 1.3. For $y'' + \mu y = 0$, $y'(0) = y'(\ell) = 0$, the eigenvalues are $\mu_n = (n\pi/\ell)^2$, $n \geq 0$, with eigenfunctions $\varphi_n(x) = \cos(n\pi x/\ell)$. The previous theorem then shows that

$$\int_0^\ell \cos(m\pi x/\ell) \cos(n\pi x/\ell) dx = 0 \quad \text{for } m \neq n. \quad \square$$

Example 1.4. For the equation $y'' + \mu y = 0$, let the boundary values be $y(0) = 0$ and $y'(\ell) = 0$. It is not hard to verify that the trivial solution is obtained when $\mu \leq 0$. Thus take $\mu = \omega^2 > 0$. Then

$$y(x) = A \sin(\omega x) + B \cos(\omega x).$$

$y(0) = 0 \Rightarrow B = 0$. Since $y'(\ell) = A\omega \cos(\omega\ell) = 0$, we require

$$\omega\ell = (2n - 1)\frac{\pi}{2}, \quad n \geq 1.$$

Thus the eigenvalues are

$$\mu_n = \frac{(2n - 1)^2 \pi^2}{4\ell^2}$$

and the corresponding eigenfunctions are $\sin(\sqrt{\mu_n}x)$. The previous theorem then shows that

$$\int_0^\ell \sin(\sqrt{\mu_m}x) \sin(\sqrt{\mu_n}x) dx = 0 \quad \text{for } m \neq n. \quad \square$$

Example 1.5. For the equation $y'' + \mu y = 0$, let the boundary values be $y'(0) = 0$ and $\beta y(\ell) + y'(\ell) = 0$, $\beta \neq 0$. It is not hard to verify that the trivial solution is obtained when $\mu \leq 0$. Thus take $\mu = \omega^2 > 0$. Then

$$y(x) = A \sin(\omega x) + B \cos(\omega x).$$

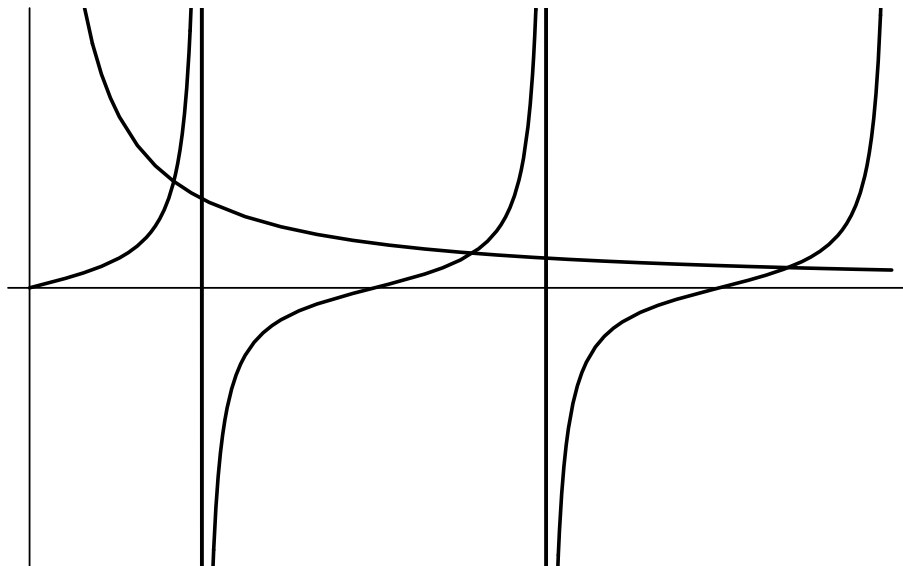
$y'(0) = 0 \Rightarrow A = 0$. Then we require

$$\beta B \cos(\omega\ell) - B\omega \sin(\omega\ell) = 0.$$

This yields

$$\tan(\omega\ell) = \frac{\beta}{\omega}. \quad (\text{II.8})$$

One can see from a graph that there exists an infinite number of values of ω_n satisfying this equation.



The eigenvalues are $\mu_n = \omega_n^2$ with eigenfunctions $\cos(\omega_n x)$. The previous theorem then shows that

$$\int_0^\ell \cos(\omega_m x) \cos(\omega_n x) dx = 0 \quad \text{for } m \neq n.$$

Moreover, if we write

$$f(x) = \sum_{n=1}^{\infty} c_n \cos(\omega_n x),$$

then we have from (II.7) that

$$c_n = \frac{\int_0^\ell f(x) \cos(\omega_n x) dx}{\|\cos(\omega_n x)\|^2}.$$

We now obtain an expression for $\|\cos(\omega_n x)\|^2$. We have

$$\|\cos(\omega_n x)\|^2 = \int_0^\ell \cos^2(\omega_n x) dx = \frac{1}{2} \int_0^\ell (1 + \cos(2\omega_n x)) dx = \frac{1}{2} \left[\ell + \frac{\sin(2\omega_n \ell)}{2\omega_n} \right].$$

However,

$$\sin(2\omega_n \ell) = 2 \sin(\omega_n \ell) \cos(\omega_n \ell) = \frac{2 \tan(\omega_n \ell)}{\sec^2(\omega_n \ell)} = \frac{2 \tan(\omega_n \ell)}{1 + \tan^2(\omega_n \ell)} = \frac{2\beta/\omega_n}{1 + \beta^2/\omega_n^2} = \frac{2\beta\omega_n}{\omega_n^2 + \beta^2},$$

where the penultimate step follows from (II.8). Thus

$$\|\cos(\omega_n x)\|^2 = \frac{1}{2} \left[\ell + \frac{\beta}{\omega_n^2 + \beta^2} \right]. \quad \boxtimes$$

§2 Boundary conditions

Recall that the heat equation in 1D is

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}, \quad \kappa > 0.$$

This equation models the temperature distribution of a rod which we shall assume is of length $\ell > 0$. In this equation, x satisfying $0 \leq x \leq \ell$ is the distance from the left-hand end and $t \geq 0$ is time. The parameter κ is called the *thermal diffusivity* of the material of which the rod is composed.

In order to ensure a unique solution u , one usually has boundary conditions at $x = 0$ and $x = \ell$ and an initial condition at $t = 0$ which specifies $u(x, 0)$.

As we have seen earlier, there might be several types of boundary conditions.

- (i) $u(0, t) = g_1(t)$, $u(\ell, t) = g_2(t)$, so the temperature is specified on the boundary.
- (ii) $u_x(0, t) = g_1(t)$, $u_x(\ell, t) = g_2(t)$, so the heat flow across the boundaries are specified.
- (iii) $u_x(0, t) + \gamma u(0, t) = g_1(t)$, $u_x(\ell, t) + \gamma u(\ell, t) = g_2(t)$. This specifies the temperature of the surrounding medium.

§3 Separation of variables

This technique is useful for initial boundary-value problems in which:

- (i) The PDE is linear *and* homogeneous.
- (ii) The boundary conditions are of the form

$$\alpha u_x(0, t) + \beta u(0, t) = 0, \quad \gamma u_x(\ell, t) + \eta u(\ell, t) = 0,$$

where α , β , γ , and η are constants. These boundary conditions are called linear homogeneous boundary conditions.

The basic idea is to assume there exists a solution of the form $u(x, t) = X(x)T(t)$ which satisfies the PDE and the boundary conditions. In fact, there are usually infinitely-many such solutions, $u_n(x, t) = X_n(x)T_n(t)$, known as fundamental solutions. We add these together to form the general solution

$$u(x, t) = \sum_{n=1}^{\infty} A_n X_n(x) T_n(t),$$

with the A_n chosen so that the initial condition is satisfied.

Example 3.1. We solve by separation of variables, the problem

$$u_t = \kappa u_{xx}$$

with conditions

$$u(0, t) = u(\ell, t) = 0, \quad u(x, 0) = \phi(x).$$

We assume there exists a solution of the form $u(x, t) = X(x)T(t)$. Substitution into the PDE yields

$$X(x) \frac{dT}{dt} = \kappa \frac{d^2X}{dx^2} T(t).$$

We then collect all the terms in x on one side and the terms in t on the other side:

$$\frac{1}{\kappa} \frac{\dot{T}(t)}{T(t)} = \frac{X''(x)}{X(x)}.$$

(Here, we use $'$ to denote differentiation with respect to x and $\dot{}$ to denote differentiation with respect to t .)

The left-hand side of this equation is a function of t only while the right-hand side is a function of x only. The only way this can happen is if the functions are equal to a constant, k say. Hence

$$\frac{1}{\kappa} \frac{\dot{T}(t)}{T(t)} = k = \frac{X''(x)}{X(x)}.$$

Rearranging yields the pair of uncoupled ODEs to solve:

$$\dot{T}(t) - k\kappa T(t) = 0, \quad X''(x) - kX(x) = 0.$$

But what value of k should we use. We assume that u , X , T , and κ are all real and not complex. Hence k is real and the three possibilities are $k > 0$, $k = 0$, and $k < 0$.

For $k \neq 0$, the solution of $\dot{T} - k\kappa T = 0$ is $T(t) = T(0)e^{(k\kappa)t}$. For $k > 0$, $T(t) \rightarrow \infty$ as $t \rightarrow \infty$ which is not physical. For $k < 0$, the solution looks plausible.

For $k = 0$, $\dot{T} = 0$ means T is a constant, say T_0 . Then $X''(x) = 0$ means $X(x) = ax + b$. Hence

$$u(x, t) = X(x)T(t) = T_0(ax + b) \equiv Cx + D.$$

Returning back to the case $k < 0$, let us write $k = -\lambda^2$. Then we have

$$T(t) = T(0)e^{-\kappa\lambda^2 t}.$$

Also, the solution of $X''(x) + \lambda^2 X(x) = 0$ is

$$X(x) = A \sin(\lambda x) + B \cos(\lambda x),$$

where A and B are arbitrary constants. Hence

$$u(x, t) = X(x)T(t) = [A \sin(\lambda x) + B \cos(\lambda x)] T(0)e^{-\kappa \lambda^2 t}. \quad (\text{II.9})$$

We see we can take $T(0) = 1$, that is, absorb the value of $T(0)$ into A and B .

We now have an infinity of solutions to the PDE. These are given by either $Cx + D$ or (II.9). We want those that satisfy the boundary conditions. Considering $u(0, t) = 0$, we see that we have

$$D = 0 \quad \text{and} \quad B = 0.$$

Also, $u(\ell, t) = 0$ implies

$$C = 0 \quad \text{and} \quad A \sin(\lambda \ell) = 0.$$

We don't want $A = 0$, otherwise $u(x, t) \equiv 0$, the trivial solution. We then conclude that $\sin(\lambda \ell) = 0$ and hence

$$\lambda = \pm \frac{n\pi}{\ell}, \quad n = 1, 2, 3, \dots$$

So the fundamental solutions that satisfy the boundary conditions are given by

$$u_n(x, t) = A_n e^{-\kappa(n\pi/\ell)^2 t} \sin(n\pi x/\ell).$$

(Note that we take $n > 0$ here; the solutions with $n < 0$ are essentially the same.)

We now choose the A_n to satisfy the initial condition. As the PDE is linear, the sum of the fundamental solutions is also a solution of the PDE and satisfies the boundary conditions. So the general solution of the PDE satisfying the boundary conditions is given by

$$u(x, t) = \sum_{n=1}^{\infty} A_n e^{-\kappa(n\pi/\ell)^2 t} \sin(n\pi x/\ell).$$

Each term in $u(x, t)$ is a sine wave with an exponentially decaying amplitude.

This expression for $u(x, t)$ must hold for all $t \geq 0$. Since $u(x, 0) = \phi(x)$, we have

$$\phi(x) = \sum_{n=1}^{\infty} A_n \sin(n\pi x/\ell).$$

We recognise that we have expressed $\phi(x)$ as a Fourier sine series. From Example 1.2, we conclude that

$$A_n = \frac{2}{\ell} \int_0^{\ell} \phi(x) \sin(n\pi x/\ell) dx. \quad \square$$

Example 3.2. We solve by separation of variables the heat equation as in the previous example, but with conditions

$$u(0, t) = 0, \quad u_x(\ell, t) + \beta u(\ell, t) = 0, \quad u(x, 0) = x,$$

where $\beta\ell \neq -1$. As before, we obtain the pair of uncoupled ODEs given by

$$\dot{T}(t) - k\kappa T(t) = 0, \quad X''(x) - kX(x) = 0.$$

The same reasoning as in the previous example shows we should reject $k > 0$. For $k = 0$, we have $X(x) = Cx + D$. The boundary conditions $u(0, t) = 0$ yields $D = 0$ so that $X(x) = Cx$. The other boundary condition $u_x(\ell, t) + \beta u(\ell, t) = 0$ then yields $C + \beta C\ell = 0$, which shows that $C = 0$.

So we have to take $k < 0$ and we write $k = -\lambda^2$, as before. Again, we obtain the solutions

$$u(x, t) = X(x)T(t) = [A \sin(\lambda x) + B \cos(\lambda x)] e^{-\kappa\lambda^2 t}.$$

We now need those solutions that satisfy the boundary conditions. Considering $u(0, t) = 0$, we see that we have $B = 0$. Moreover, the boundary condition $u_x(\ell, t) + \beta u(\ell, t) = 0$ yields

$$A\lambda \cos(\lambda\ell) + \beta A \sin(\lambda\ell) = 0.$$

We don't want $A = 0$, otherwise $u(x, t) \equiv 0$, the trivial solution. We then conclude that we require

$$\tan(\lambda\ell) = \frac{-\lambda}{\beta}. \quad (\text{II.10})$$

There exists an infinite number of values λ_n satisfying this relationship (compare with Example 1.5). So the fundamental solutions that satisfy the boundary conditions are given by

$$u_n(x, t) = A_n e^{-\kappa\lambda_n^2 t} \sin(\lambda_n x).$$

We now choose the A_n to satisfy the initial condition. The general solution of the PDE satisfying the boundary conditions is given by

$$u(x, t) = \sum_{n=1}^{\infty} A_n e^{-\kappa\lambda_n^2 t} \sin(\lambda_n x).$$

Since $u(x, 0) = x$, we have

$$x = \sum_{n=1}^{\infty} A_n \sin(\lambda_n x).$$

Our work on Sturm-Liouville problems shows that

$$A_n = \frac{\int_0^\ell x \sin(\lambda_n x) dx}{\|\sin(\lambda_n x)\|^2}.$$

We now obtain an expression for $\|\sin(\lambda_n x)\|^2$. We have

$$\begin{aligned}\|\sin(\lambda_n x)\|^2 &= \int_0^\ell \sin^2(\lambda_n x) dx = \frac{1}{2} \int_0^\ell (1 - \cos(2\lambda_n x)) dx = \frac{1}{2} \left[\ell - \frac{\sin(2\lambda_n \ell)}{2\lambda_n} \right] \\ &= \frac{\lambda_n \ell - \sin(\lambda_n \ell) \cos(\lambda_n \ell)}{2\lambda_n}.\end{aligned}$$

Similar to Example 1.5, we can obtain an expression for $\|\sin(\lambda_n x)\|^2$ which does not involve sine and cosine terms. We have

$$\sin(\lambda_n \ell) \cos(\lambda_n \ell) = \frac{\tan(\lambda_n \ell)}{\sec^2(\lambda_n \ell)} = \frac{\tan(\lambda_n \ell)}{1 + \tan^2(\lambda_n \ell)} = \frac{-\lambda_n/\beta}{1 + \lambda_n^2/\beta^2} = \frac{-\beta\lambda_n}{\beta^2 + \lambda_n^2},$$

where the penultimate step follows from (II.10). Thus

$$\|\sin(\lambda_n x)\|^2 = \frac{1}{2} \left[\ell + \frac{\beta}{\beta^2 + \lambda_n^2} \right]. \quad \square$$

§4 Heat equation and non-homogeneous boundary conditions

Suppose we have the heat equation given by $u_t - \kappa u_{xx} = 0$ along with the non-homogeneous boundary conditions given by

$$\alpha_1 u_x(0, t) + \beta_1 u(0, t) = g_1(t), \quad \alpha_2 u_x(\ell, t) + \beta_2 u(\ell, t) = g_2(t).$$

The question we explore in this section is whether we can apply some transformation so that the boundary conditions become homogeneous, that is, have a function $U(x, t)$ for which

$$\alpha_1 U_x(0, t) + \beta_1 U(0, t) = 0, \quad \alpha_2 U_x(\ell, t) + \beta_2 U(\ell, t) = 0.$$

If one thinks about the physical system behind the heat equation, one might expect that as time progresses, the temperature in the rod might reach a steady-state or equilibrium solution. So one might wish to consider

$$u(x, t) = \text{steady-state solution} + \text{transient solution},$$

where the transient solution goes to 0 as $t \rightarrow \infty$.

Example 4.1. We consider $u_t - \kappa u_{xx} = 0$ along with the non-homogeneous boundary conditions given by

$$u(0, t) = g_1(t), \quad u_x(\ell, t) + \beta u(\ell, t) = g_2(t),$$

and initial condition $u(x, 0) = \phi(x)$.

Suppose we assume $u(x, t) = S(x, t) + U(x, t)$, where the transient solution $U(x, t)$ satisfies homogeneous boundary conditions. In this case, we would have

$$U(0, t) = 0, \quad U_x(\ell, t) + \beta U(\ell, t) = 0.$$

It then follows that

$$S(0, t) = g_1(t), \quad S_x(\ell, t) + \beta S(\ell, t) = g_2(t).$$

It is not clear what the form of $S(x, t)$ should be. One form that works is to take

$$S(x, t) = A(t)(1 - x/\ell) + B(t)x/\ell.$$

so that $g_1(t) = S(0, t) = A(t)$ and $g_2(t) = S_x(\ell, t) + \beta S(\ell, t) = -A(t)/\ell + B(t)/\ell + \beta B(t)$. We then have $A(t) = g_1(t)$ and

$$B(t) = \frac{g_2(t) + A(t)/\ell}{1/\ell + \beta} = \frac{g_2(t) + g_1(t)/\ell}{1/\ell + \beta} = \frac{\ell g_2(t) + g_1(t)}{1 + \beta\ell}.$$

Thus $S(x, t)$ is now known. Note that since $S(x, t)$ is linear in x , then $S_{xx}(x, t) = 0$.

Returning to the PDE, $u_t - \kappa u_{xx} = 0$ becomes

$$S_t + U_t - \kappa (S_{xx} + U_{xx}) = 0 \quad \text{or} \quad \dot{A}(t)(1 - x/\ell) + \dot{B}(t)x/\ell + U_t - \kappa U_{xx} = 0.$$

Hence the PDE for $U(x, t)$ is now

$$U_t - \kappa U_{xx} = -S_t,$$

which is generally non-homogeneous (but not always; for example, if g_1 and g_2 were constants, then $S_t = 0$). By construction, U satisfies homogeneous boundary conditions and the (new, but known) initial condition is given by

$$U(x, 0) = u(x, 0) - S(x, 0) = \phi(x) - A(0)(1 - x/\ell) - B(0)x/\ell.$$

If the resulting PDE for $U(x, t)$ is non-homogeneous, then it turns out that the method of separation of variables does not work and we need other techniques to solve the problem. \square

§5 Transforming hard equations into easier ones

The idea is similar to what we did in the previous section in which we transformed non-homogeneous boundary conditions into homogeneous ones by introducing a new function. Here the focus is on simplifying the PDE rather than the boundary conditions. However, we normally need to have a good “guess” to make progress with this method.

Example 5.1. Suppose we have the non-homogeneous heat equation $u_t - \kappa u_{xx} = -\gamma u$ with boundary and initial conditions

$$u(0, t) = u(\ell, t) = 0, \quad u(x, 0) = \phi(x).$$

The physics of the situation indicates that at any point x , the temperature is changing as a result of two phenomenon:

- diffusion of heat within the rod which is represented by u_{xx} .
- heat flow through the sides represented by $-\gamma u$.

Observe that if no diffusion occurred, that is, $\kappa = 0$, then the equation becomes $u_t = -\gamma u$, which has solution

$$u(x, t) = \phi(x)e^{-\gamma t},$$

Based on this, let us try $u(x, t) = w(x, t)e^{-\gamma t}$. Then our original PDE becomes

$$w_t e^{-\gamma t} - w\gamma e^{-\gamma t} - \kappa w_{xx} e^{-\gamma t} = -\gamma w e^{-\gamma t} \quad \text{or} \quad w_t - \kappa w_{xx} = 0.$$

For the function $w(x, t)$, the boundary and initial conditions are the same as those for $u(x, t)$:

$$w(0, t) = w(\ell, t) = 0, \quad w(x, 0) = \phi(x).$$

So we now have a problem which we have already solved previously. ☒

§6 Non-homogeneous PDEs and eigenfunction expansions

When we have a non-homogeneous PDE, then even though the boundary conditions may be homogeneous, then we cannot use the method of separation of variables. Other options include integral transforms (which are quite powerful) and eigenfunction expansions. We now consider the latter technique.

Suppose we have the non-homogeneous heat equation given by $u_t - \kappa u_{xx} = f(x, t)$ along with the homogeneous boundary conditions given by

$$\alpha_1 u_x(0, t) + \beta_1 u(0, t) = 0, \quad \alpha_2 u_x(\ell, t) + \beta_2 u(\ell, t) = 0,$$

and initial condition $u(x, 0) = \phi(x)$.

The idea is to find a solution of the form

$$u(x, t) = \sum_{n=1}^{\infty} X_n(x)T_n(t),$$

where the $X_n(x)$ are the eigenfunctions corresponding to the Sturm-Liouville problem arising from solving the homogeneous PDE. As we shall shortly see in an example, the $T_n(t)$ will satisfy

a first order ODE. In order to obtain this ODE, we need to expand $f(x, t)$ into an expansion similar to that for $u(x, t)$, namely

$$f(x, t) = \sum_{n=1}^{\infty} X_n(x) f_n(t).$$

Then similar to the derivation of (II.7), we obtain

$$f_n(t) = \frac{\langle f(\cdot, t), X_n \rangle}{\|X_n\|^2} = \frac{\int_0^\ell r(x) f(x, t) X_n(x) dx}{\int_0^\ell r(x) X_n^2(x) dx}.$$

Example 6.1. Let us consider $u_t - \kappa u_{xx} = f(x, t)$ along with the boundary and initial conditions given by

$$u(0, t) = u(\ell, t) = 0, \quad u(x, 0) = \phi(x).$$

For these homogeneous boundary conditions, we know that the $X_n(x)$ are given by $\sin(n\pi x/\ell)$. Hence,

$$f(x, t) = \sum_{n=1}^{\infty} \sin(n\pi x/\ell) f_n(t),$$

where the $f_n(t)$ are given by

$$f_n(t) = \frac{2}{\ell} \int_0^\ell f(x, t) \sin(n\pi x/\ell) dx.$$

Assuming that

$$u(x, t) = \sum_{n=1}^{\infty} X_n(x) T_n(t) = \sum_{n=1}^{\infty} \sin(n\pi x/\ell) T_n(t),$$

we now substitute this into the PDE. (Note that the boundary conditions $u(0, t) = u(\ell, t) = 0$ are already satisfied.) Hence we obtain

$$\sum_{n=1}^{\infty} \sin(n\pi x/\ell) \dot{T}_n(t) + \kappa \left(\frac{\pi}{\ell}\right)^2 \sum_{n=1}^{\infty} n^2 \sin(n\pi x/\ell) T_n(t) = f(x, t) = \sum_{n=1}^{\infty} \sin(n\pi x/\ell) f_n(t),$$

or

$$\sum_{n=1}^{\infty} \sin(n\pi x/\ell) \left(\dot{T}_n(t) + \kappa \left(\frac{n\pi}{\ell}\right)^2 T_n(t) - f_n(t) \right) = 0.$$

Since this must hold for all x in the domain, we conclude that

$$\dot{T}_n(t) + \kappa \left(\frac{n\pi}{\ell}\right)^2 T_n(t) = f_n(t). \quad (\text{II.11})$$

To satisfy the initial condition $u(x, 0) = \phi(x)$, we require

$$\phi(x) = \sum_{n=1}^{\infty} \sin(n\pi x/\ell) T_n(0) \Rightarrow T_n(0) = \frac{2}{\ell} \int_0^{\ell} \phi(x) \sin(n\pi x/\ell) dx.$$

Solving the initial value problem consisting of the first order ODE given in (II.11) and this initial condition $T_n(0)$ then yields the required functions $T_n(t)$.

In fact, by using an integrating factor on (II.11), we see that it has solution given by

$$T_n(t) = e^{-\kappa n^2 \pi^2 t/\ell^2} \int e^{\kappa n^2 \pi^2 t/\ell^2} f_n(t) dt + C_n e^{-\kappa n^2 \pi^2 t/\ell^2},$$

where C_n is an arbitrary constant. Setting

$$G_n(t) = \int e^{\kappa n^2 \pi^2 t/\ell^2} f_n(t) dt,$$

the initial condition yields $T_n(0) = G_n(0) + C_n$ and hence $C_n = T_n(0) - G_n(0)$. Thus

$$T_n(t) = e^{-\kappa n^2 \pi^2 t/\ell^2} (G_n(t) + T_n(0) - G_n(0)) = e^{-\kappa n^2 \pi^2 t/\ell^2} \left(\int_0^t e^{\kappa n^2 \pi^2 v/\ell^2} f_n(v) dv + T_n(0) \right).$$

Hence $u(x, t)$ is given by

$$\sum_{n=1}^{\infty} \sin(n\pi x/\ell) e^{-\kappa n^2 \pi^2 t/\ell^2} \int_0^t e^{\kappa n^2 \pi^2 v/\ell^2} f_n(v) dv + \sum_{n=1}^{\infty} T_n(0) \sin(n\pi x/\ell) e^{-\kappa n^2 \pi^2 t/\ell^2}.$$

In this expression, we see that the solution is in two parts. The first part arises from the non-homogeneous term $f(x, t)$ (and contributes to the steady state component of the solution) while the second part arises from the initial condition. The second part is transient since it goes to zero as $t \rightarrow \infty$. \boxtimes

Example 6.2. As a special case of the previous example, suppose $f(x, t) = \sin(5\pi x/\ell)$ and $\phi(x) = \sin(2\pi x/\ell)$. Then

$$f_n(t) = \frac{2}{\ell} \int_0^{\ell} \sin(5\pi x/\ell) \sin(n\pi x/\ell) dx = \delta_{n5},$$

where δ_{ij} is the Kronecker delta function, that is,

$$\delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Similarly, $T_n(0) = \delta_{n2}$. It then follows that

$$\begin{aligned} u(x, t) &= \sin(5\pi x/\ell) e^{-25\kappa\pi^2 t/\ell^2} \int_0^t e^{25\kappa\pi^2 v/\ell^2} dv + \sin(2\pi x/\ell) e^{-4\kappa\pi^2 t/\ell^2} \\ &= \sin(5\pi x/\ell) e^{-25\kappa\pi^2 t/\ell^2} \left[\frac{e^{25\kappa\pi^2 t/\ell^2} - 1}{25\kappa\pi^2/\ell^2} \right] + \sin(2\pi x/\ell) e^{-4\kappa\pi^2 t/\ell^2} \\ &= \sin(5\pi x/\ell) \ell^2 \left[\frac{1 - e^{-25\kappa\pi^2 t/\ell^2}}{25\kappa\pi^2} \right] + \sin(2\pi x/\ell) e^{-4\kappa\pi^2 t/\ell^2}. \end{aligned}$$

In this solution we see that as $t \rightarrow \infty$, $u(x, t) \rightarrow \frac{\ell^2 \sin(5\pi x/\ell)}{25\kappa\pi^2}$. Here we see a steady-state part arising from $f(x, t)$ and the transient part arising from the initial condition. \boxtimes

Part III — Classification and characteristics of second order linear PDEs

§1 Classification

As already mentioned previously, the most general second order linear PDE in two independent variables x and y is

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G. \quad (\text{III.1})$$

The coefficients $A(x, y), \dots, G(x, y)$ are continuous functions over some domain Ω of the x - y plane. $u(x, y)$ and/or its derivatives are to satisfy given conditions on the boundary of Ω .

The *type* of this equation is determined by its *principal part* (that is, the terms involving the second order derivatives) and in part by the sign of the discriminant $B^2 - 4AC$.

§2 Characteristics

These are special curves $\Gamma = y(x)$ associated with (III.1). Questions we are interested in are:

Q1. How can a coordinate transform be used to simplify the principal part of (III.1)?

Q2. Along what curves $y(x)$ is a knowledge of u, u_x , and u_y insufficient to uniquely determine the second order derivatives u_{xx}, u_{xy} , and u_{yy} ?

For Q1, suppose there exists a change of variables

$$\phi = \phi(x, y) \quad \text{and} \quad \psi = \psi(x, y)$$

which is locally invertible, that is, $\phi_x \psi_y - \phi_y \psi_x \neq 0$. Under the change of variables, then

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G \rightarrow au_{\phi\phi} + bu_{\phi\psi} + cu_{\psi\psi} + du_{\phi} + eu_{\psi} + Fu = G.$$

Then one can show (as an exercise!) that

$$a = A(\phi_x)^2 + B\phi_x\phi_y + C(\phi_y)^2, \quad (\text{III.2})$$

$$b = 2A\phi_x\psi_x + B(\phi_x\psi_y + \phi_y\psi_x) + 2C\phi_y\psi_y, \quad (\text{III.3})$$

$$c = A(\psi_x)^2 + B\psi_x\psi_y + C(\psi_y)^2, \quad (\text{III.4})$$

$$d = A\phi_{xx} + B\phi_{xy} + C\phi_{yy} + D\phi_x + E\phi_y, \quad (\text{III.5})$$

$$e = A\psi_{xx} + B\psi_{xy} + C\psi_{yy} + D\psi_x + E\psi_y. \quad (\text{III.6})$$

Some algebra yields

$$b^2 - 4ac = (B^2 - 4AC)(\phi_x\psi_y - \phi_y\psi_x)^2. \quad (\text{III.7})$$

The last term is never zero. Therefore the original and transformed discriminant have the same sign.

Moreover, if we choose ϕ and ψ such that $a = c = 0$, the principal part of the transformed equation will be particularly simple. This requires that ϕ and ψ are each solutions of

$$A(z_x)^2 + Bz_xz_y + C(z_y)^2 = 0, \quad z = z(x, y). \quad (\text{III.8})$$

Such a solution defines a surface $z(x, y)$. The level curves or contours $z(x, y) = \text{constant}$ of this surface are called the characteristics of (III.1).

For Q2, suppose u , u_x , and u_y are known along a smooth curve Γ on the x - y plane. If this curve is parametrised by the parameter s so that

$$x = x(s), \quad y = y(s), \quad u = F(s), \quad u_x = G(s), \quad u_y = H(s),$$

then we use the chain rule given by

$$\frac{d}{ds} = \frac{dx}{ds} \frac{\partial}{\partial x} + \frac{dy}{ds} \frac{\partial}{\partial y}$$

to obtain

$$\frac{d}{ds} (u_x) = u_{xx} \frac{dx}{ds} + u_{xy} \frac{dy}{ds} = \frac{dG}{ds}$$

and

$$\frac{d}{ds} (u_y) = u_{yx} \frac{dx}{ds} + u_{yy} \frac{dy}{ds} = \frac{dH}{ds}.$$

Together with the PDE (III.1), there are three linear equations in three unknowns for the second order derivatives. This may be written in matrix form as

$$\begin{bmatrix} A & B & C \\ x'(s) & y'(s) & 0 \\ 0 & x'(s) & y'(s) \end{bmatrix} \begin{bmatrix} u_{xx} \\ u_{xy} \\ u_{yy} \end{bmatrix} = \text{known vector.}$$

This system has a unique solution provided the determinant of the coefficient matrix is not zero. Now this matrix has determinant given by

$$\begin{aligned} A[y'(s)]^2 - Bx'(s)y'(s) + C[x'(s)]^2 &= A \left[\frac{dy}{ds} \right]^2 - B \left[\frac{dx}{ds} \frac{dy}{ds} \right] + C \left[\frac{dx}{ds} \right]^2 \\ &= A \left[\frac{dy}{dx} \frac{dx}{ds} \right]^2 - B \left[\frac{dx}{ds} \frac{dy}{dx} \frac{dx}{ds} \right] + C \left[\frac{dx}{ds} \right]^2 \\ &= [A\eta^2 - B\eta + C] \left[\frac{dx}{ds} \right]^2, \end{aligned}$$

with $\eta = \frac{dy}{dx}$ being the tangent to Γ at $(x, y(x))$. In general $x'(s) \neq 0$ and hence the determinant is zero when

$$A\eta^2 - B\eta + C = 0. \quad (\text{III.9})$$

This is a nonlinear ODE for $\eta = \frac{dy}{dx}$ and determines the curves Γ for which the linear system does not have a unique solution. These curves are the characteristics of the PDE.

We note the following:

- Clearly there are two, one, or zero real solutions to (III.9) depending on the sign of $B^2 - 4AC$.
- In the hyperbolic case of two real solutions, the characteristics define natural coordinates ϕ and ψ in which to study the PDE given in (III.1).
- In the elliptic case, there are zero real solutions. This means that there are no curves along which discontinuities can propagate. Thus solutions of elliptic equations are generally smooth.
- In the parabolic case, there is just one real solution. This means that it is not possible to make both a and c equal to the zero function.
- Related to the indeterminacy/non-uniqueness of second order derivatives along a characteristic is that fact that (physically significant) discontinuities in the solution of (III.1) can propagate ONLY along characteristics. So for physical systems modelled by hyperbolic PDEs, one might expect phenomena such as shocks.

We now give examples of characteristics by considering the wave equation (hyperbolic), Laplace's equation (elliptic), and the heat equation (parabolic).

Example 2.1. The one dimensional wave equation is given by $u_{tt} - c^2 u_{xx} = 0$. Using t as the 'y'-variable, then $\eta = \frac{dt}{dx}$ satisfies the equation given in (III.9). Hence

$$-c^2 \eta^2 - 0 + 1 = 0 \Rightarrow \frac{dt}{dx} = \pm \frac{1}{c}.$$

Thus the characteristics are the curves

$$t = \frac{x}{c} + c_1 \quad \text{and} \quad t = -\frac{x}{c} + c_2,$$

where c_1 and c_2 are constants, or $ct - x = k_1$ and $ct + x = k_2$. ⊠

Example 2.2. The two dimensional Laplace's equation is given by $u_{xx} + u_{yy} = 0$. Then $\eta = \frac{dy}{dx}$ satisfies

$$\eta^2 - 0 + 1 = 0 \Rightarrow \frac{dy}{dx} = \pm i,$$

where $i^2 = -1$. Thus the characteristics are the curves

$$y = ix + c_1 \quad \text{and} \quad y = -ix + c_2,$$

where c_1 and c_2 are constants. The characteristics are not real-valued. ⊠

Example 2.3. The one dimensional heat equation is given by $u_t - \kappa u_{xx} = 0$. Then $\eta = \frac{dt}{dx}$ satisfies

$$-\kappa \eta^2 - 0 + 0 = 0 \Rightarrow \frac{dt}{dx} = 0.$$

Thus the characteristics are the curves $t = \text{constant}$. ⊠

In answering Q1 and Q2, we have introduced characteristics via (III.8) and (III.9). We would expect some connection between them as is indeed the case.

Theorem III.1. *If $z(x, y)$ is a solution of (III.8), then $z(x, y) = \text{constant}$ is a characteristic of (III.1) if and only if $z(x, y) = \text{constant}$ is a solution of (III.9).*

Proof. Assume $z(x, y)$ is a solution of (III.8) with $z_y \neq 0$. Therefore $z(x, y) = K = \text{constant}$ (implicitly) defines at least one single-valued function $y = F(x; K)$. Differentiating $z(x, y(x)) = K$ with respect to x yields

$$\frac{\partial z}{\partial x} + \frac{\partial z}{\partial y} \frac{dy}{dx} = 0 \Rightarrow \frac{dy}{dx} = -\frac{z_x}{z_y}.$$

Dividing (III.8) by $(z_y)^2$ yields

$$A \left(\frac{z_x}{z_y} \right)^2 + B \frac{z_x}{z_y} + C = 0.$$

With $\eta = \frac{dy}{dx}$, we then see that

$$A\eta^2 - B\eta + C = 0,$$

which is precisely (III.9). Therefore $y = F(x; K)$ satisfies (III.9) which implies that $z(x, y) = K$ is an implicit solution of (III.9)

In our proof so far, we have assumed that $z_y \neq 0$. If it is the case that $z_y = 0$, but (III.8) is not identically satisfied, then we must have $z_x \neq 0$ and we can repeat the above argument with the roles of x and y swapped around.

We now prove the converse of the result. Let $z(x, y) = K$ be a general solution of (III.9). We want to show that $z(x, y)$ satisfies (III.8) at an arbitrary point (x_0, y_0) . So let $K_0 = z(x_0, y_0)$ and consider curves $y = G(x; K_0)$.

Along G , (III.9) holds so that

$$A \left(\frac{dy}{dx} \right)^2 - B \frac{dy}{dx} + C = 0.$$

But also $z(x, y) = \text{constant}$ here so that as before

$$\frac{dy}{dx} = -\frac{z_x}{z_y}.$$

Substitution then yields

$$A \left(\frac{z_x}{z_y} \right)^2 + B \frac{z_x}{z_y} + C = 0 \Rightarrow A(z_x)^2 + Bz_x z_y + C(z_y)^2 = 0.$$

which when evaluated at $x = x_0$ is equal to (III.8). So the solutions of (III.8) and (III.9) match. \square

§3 Canonical forms

It is possible to show that if A , B , and C are smooth functions of x and y , then there will always exist a locally one-to-one coordinate transformation

$$\phi = \phi(x, y) \quad \text{and} \quad \psi = \psi(x, y)$$

which transforms the principal part of (III.1) to the following CANONICAL FORMS:

$$\begin{cases} \text{parabolic (heat equation)} & u_{\phi\phi}, \\ \text{hyperbolic (wave equation)} & u_{\phi\psi} \text{ or } u_{\phi\phi} - u_{\psi\psi}, \\ \text{elliptic (steady state)} & u_{\phi\phi} + u_{\psi\psi}. \end{cases}$$

If it happens that A , B , and C are constants, then the transformation turns out to be a linear change of variable. To obtain the appropriate transformation, we can make use of characteristics.

Example 3.1. It may be verified that the PDE

$$2u_{xx} - 4u_{xy} - 6u_{yy} + u_x = 0$$

is hyperbolic. To find the characteristics, we make use of (III.9). For $\eta = \frac{dy}{dx}$, we have $2\eta^2 + 4\eta - 6 = 0$ and hence

$$\frac{dy}{dx} = \frac{-4 \pm \sqrt{16 - (-48)}}{4} = \frac{-4 \pm 8}{4} = -3, 1.$$

Upon integration, we find the characteristic curves are given by

$$y = -3x + c_1 \quad \text{and} \quad y = x + c_2 \quad \text{or} \quad 3x + y = c_1 \quad \text{and} \quad -x + y = c_2.$$

Then the transformations we require are $\phi(x, y) = 3x + y$ and $\psi(x, y) = -x + y$. By making use of (III.2)–(III.6), we find the original PDE is transformed to

$$-32u_{\phi\psi} + 3u_{\phi} - u_{\psi} = 0. \quad \square$$

We consider the situation in which (III.1) is an elliptic equation, that is, $B^2 - 4AC < 0$. Though the characteristics are not real, it is still possible to make use of the characteristics to transform the principal part into the canonical form $u_{\phi\phi} + u_{\psi\psi}$.

Then $\eta = \frac{dy}{dx}$ is given by the complex values

$$\frac{dy}{dx} = \frac{B \pm i\sqrt{|B^2 - 4AC|}}{2A}.$$

The characteristics are of the form

$$z(x, y(x)) = \phi(x, y) \pm i\psi(x, y) = \text{constant}.$$

We see from (III.8) that

$$\begin{aligned} 0 &= A(z_x)^2 + Bz_xz_y + C(z_y)^2 \\ &= A(\phi_x + i\psi_x)^2 + B(\phi_x + i\psi_x)(\phi_y + i\psi_y) + C(\phi_y + i\psi_y)^2 \\ &= A[(\phi_x)^2 - (\psi_x)^2] + B[\phi_x\phi_y - \psi_x\psi_y] + C[(\phi_y)^2 - (\psi_y)^2] \\ &\quad + i[2A\phi_x\psi_x + B(\phi_x\psi_y + \phi_y\psi_x) + 2C\phi_y\psi_y] \\ &= [A(\phi_x)^2 + B\phi_x\phi_y + C(\phi_y)^2] - [A(\psi_x)^2 + B\psi_x\psi_y + C(\psi_y)^2] \\ &\quad + i[2A\phi_x\psi_x + B(\phi_x\psi_y + \phi_y\psi_x) + 2C\phi_y\psi_y] \\ &= [a - c] + bi, \end{aligned}$$

where we have made use of (III.2)–(III.4). Hence $a = c$ and $b = 0$.

So taking $\phi = \phi(x, y)$ and $\psi = \psi(x, y)$ transforms the principal part $Au_{xx} + Bu_{xy} + Cu_{yy}$ to

$$au_{\phi\phi} + bu_{\phi\psi} + cu_{\psi\psi} = a(u_{\phi\phi} + u_{\psi\psi}).$$

Dividing by a yields the canonical form.

Example 3.2. We find the canonical form of the elliptic equation

$$u_{xx} + 2u_{xy} + 17u_{yy} = 0.$$

Then

$$\frac{dy}{dx} = \frac{2 \pm \sqrt{2^2 - 68}}{2} = 1 \pm 4i \Rightarrow y = x \pm i4x + \text{constant}.$$

So the characteristics are of the form

$$z(x, y) = x - y \pm i4x = \text{constant}.$$

So we can take the transformations to be $\phi(x, y) = x - y$ and $\psi(x, y) = 4x$. Then, as expected, the original equation gets transformed to

$$16u_{\phi\phi} + 16u_{\psi\psi} = 0 \quad \text{or} \quad u_{\phi\phi} + u_{\psi\psi} = 0. \quad \square$$

We now give an example showing how a parabolic equation may be put into canonical form.

Example 3.3. The PDE

$$e^{2x}u_{xx} + 2e^{x+y}u_{xy} + e^{2y}u_{yy} = 0$$

is parabolic since

$$B^2 - 4AC = 4e^{2x+2y} - 4e^{2x}e^{2y} = 0.$$

To find the single real characteristic, we have

$$\frac{dy}{dx} = \frac{2e^{x+y}}{2e^{2x}} = e^{y-x} \Rightarrow e^{-y} dy = e^{-x} dx.$$

Upon integration, we find

$$-e^{-y} = -e^{-x} + \text{constant},$$

so that the characteristics curves are given by $e^{-x} - e^{-y} = \text{constant}$.

By taking $\psi(x, y) = e^{-x} - e^{-y}$, we have $c = 0$. Now in the new coordinates we must still have $b^2 - 4ac = 0$ (see (III.7)). It follows that if $c = 0$, then we must also have $b = 0$.

The choice of ϕ is arbitrary. A convenient choice here is $\phi(x, y) = x$. By making use of (III.2), (III.5), and (III.6) we then obtain the PDE

$$e^{2x} u_{\phi\phi} + \left(e^{2x} e^{-x} + 2e^{x+y} \times 0 - e^{2y} e^{-y} \right) u_{\psi} = 0 \Leftrightarrow u_{\phi\phi} + \left(e^{-x} - e^{y-2x} \right) u_{\psi} = 0.$$

Now we need to change our x and y coordinates into ϕ and ψ coordinates. We can write

$$e^{-x} - e^{y-2x} = -(e^{-x} - e^{-y})e^{y-x} = -\frac{e^{-x} - e^{-y}}{e^{x-y}} = -\frac{e^{-x} - e^{-y}}{1 - (e^{-x} - e^{-y})e^x} = -\frac{\psi}{1 - \psi e^{\phi}}.$$

Hence the final PDE is

$$u_{\phi\phi} - \frac{\psi}{1 - \psi e^{\phi}} u_{\psi} = 0. \quad \square$$

Part IV — First order PDEs and method of characteristics

§1 Introduction

We have been concentrating on second order PDEs. Now we consider first order PDEs in more detail. For more generality, we consider systems of such PDEs.

Definition. *The general quasi-linear system of n first order PDEs in n functions of two independent variables x and y is given by*

$$\sum_{j=1}^n a_{ij} \frac{\partial u_j}{\partial x} + \sum_{j=1}^n b_{ij} \frac{\partial u_j}{\partial y} = c_i, \quad 1 \leq i \leq n, \quad (\text{IV.1})$$

where a_{ij} , b_{ij} , and c_i are functions of x , y , u_1, \dots, u_n .

The system is said to be almost linear if a_{ij} and b_{ij} are independent of u_1, \dots, u_n .

It is said to be linear if in addition, each c_i depends linearly on the u_j .

It is convenient to write the system in matrix-vector notation. So let

$$\mathbf{u} = (u_1, \dots, u_n)^T, \quad \mathbf{c} = (c_1, \dots, c_n)^T, \quad A = (a_{ij}), \quad B = (b_{ij}).$$

Hence the system given in (IV.1) may be written as

$$A\mathbf{u}_x + B\mathbf{u}_y = \mathbf{c}. \quad (\text{IV.2})$$

One may also have a *conservation form* given by

$$\frac{\partial \mathbf{u}}{\partial y} + \frac{\partial}{\partial x} \mathbf{F}(\mathbf{u}) = 0.$$

In such a form, the variable y usually corresponds to time.

If A or B is non-singular, one can usually classify (IV.2) as elliptic, hyperbolic, or parabolic. Let

$$P_n(\lambda) = \det(A - \lambda B) = \det(A^T - \lambda B^T).$$

When B is non-singular, we can write

$$\det(A - \lambda B) = \det(B) \det(B^{-1}A - \lambda I_n),$$

where I_n is the $n \times n$ identity matrix. Then $P_n(\lambda)$ is a polynomial of degree n . The system (IV.2) is classified as:

- Elliptic if $P_n(\lambda)$ has no real roots.
- Hyperbolic if $P_n(\lambda)$ has n real distinct roots *OR* if $P_n(\lambda)$ has n real roots with at least one repeated and the generalised eigenvalue problem $(A^T - \lambda B^T)\mathbf{w} = \mathbf{0}$ yields n linearly independent eigenvectors \mathbf{w} .
- Parabolic if $P_n(\lambda)$ has n real roots with at least one repeated and the generalised eigenvalue problem yields fewer than n linearly independent eigenvectors.

If $P_n(\lambda)$ has both real and complex roots, an exhaustive classification is not possible.

Example 1.1. The Cauchy-Riemann equations which arise in complex analysis are given by $u_x = v_y$ and $u_y = -v_x$. Setting

$$\mathbf{u} = \begin{bmatrix} u \\ v \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

we have $A\mathbf{u}_x + B\mathbf{u}_y = \mathbf{0}$. Then

$$P_2(\lambda) = \det(A - \lambda B) = \det\left(\begin{bmatrix} 1 & \lambda \\ -\lambda & 1 \end{bmatrix}\right) = 1 + \lambda^2.$$

This quadratic has roots $\pm i$ and so the Cauchy-Riemann equations form an elliptic system. \square

Example 1.2. The 1D heat equation is given by $u_t = \kappa u_{xx}$, $\kappa > 0$. If we take y to be t , and set $v = u_x$, then we have $\kappa v_x - u_y = 0$. Setting

$$\mathbf{u} = \begin{bmatrix} u \\ v \end{bmatrix}, \quad A = \begin{bmatrix} 0 & \kappa \\ 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ v \end{bmatrix},$$

we have $A\mathbf{u}_x + B\mathbf{u}_y = \mathbf{c}$.

In our classification of first order systems according to the roots of $P_n(\lambda)$, we have assumed that B was non-singular. However, for the B given above, it clearly has determinant zero and so is singular. (One can verify that $\det(A - \lambda B)$ is independent of λ .)

To get around this problem, we note that A is non-singular, so we can interchange the role of x and y and hence effectively interchange the role of A and B . Thus we may consider

$$P_2(\lambda) = \det(B - \lambda A) = \det\left(\begin{bmatrix} -1 & -\kappa\lambda \\ -\lambda & 0 \end{bmatrix}\right) = -\kappa\lambda^2.$$

This quadratic has the root 0 repeated. Solving the generalised eigenvalue problem $(B^T - \lambda A^T)\mathbf{w} = \mathbf{0}$ yields just the single eigenvector $\mathbf{w} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. So, not unexpectedly, the 1D heat equation forms a parabolic system. \square

§2 Normal form for hyperbolic systems

If the system

$$A\mathbf{u}_x + B\mathbf{u}_y = \mathbf{c}$$

is such that the matrices A and B are related by $A = DB$, for some *diagonal* matrix D , then the system can be written in *compact form*:

$$\sum_{j=1}^n b_{ij} \left(d_{ii} \frac{\partial u_j}{\partial x} + \frac{\partial u_j}{\partial y} \right) = c_i, \quad 1 \leq i \leq n.$$

If the i -th equation involves derivatives in only a single direction, then the system is said to be in *normal form*. In this case, we need to have

$$d_{ii} = \frac{dx}{dy}.$$

To see this, suppose $\alpha\mathbf{i} + \beta\mathbf{j}$ is the unit vector for which $d_{ii} = \alpha/\beta$. Then

$$d_{ii} \frac{\partial u_j}{\partial x} + \frac{\partial u_j}{\partial y} = \frac{\alpha}{\beta} \frac{\partial u_j}{\partial x} + \frac{\partial u_j}{\partial y} = \frac{1}{\beta} \left(\alpha \frac{\partial u_j}{\partial x} + \beta \frac{\partial u_j}{\partial y} \right).$$

Except for the factor $1/\beta$, this is the directional derivative of u_j in the direction $\alpha\mathbf{i} + \beta\mathbf{j}$.

The advantage of systems of PDEs that can be written in such a normal form is that, effectively, the i -th equation depends on just one differential operator. So the theory is closer to that of ODEs and so they can be solved by using techniques for solving ODEs. Before considering how this may be done, we cover some more theory on first order hyperbolic systems.

Definition. Suppose (IV.2) is a hyperbolic system in which $P_n(\lambda)$ has n real distinct roots, say $\lambda_1, \dots, \lambda_n$. Then the characteristics of (IV.2) are the curves in the x - y plane along which

$$\frac{dx}{dy} = \lambda_i, \quad 1 \leq i \leq n.$$

Theorem IV.1. Suppose (IV.2) is a hyperbolic system. Then let D be the $n \times n$ diagonal matrix whose diagonal entries are the λ_i . Then there exists a non-singular $n \times n$ matrix T such that

$$TA = DTB.$$

Proof. For the hyperbolic system, we have

$$\det(A^T - \lambda_i B^T) = 0, \quad 1 \leq i \leq n.$$

Suppose the corresponding eigenvector is \mathbf{w}_i , that is,

$$(A^T - \lambda_i B^T)\mathbf{w}_i = \mathbf{0}.$$

Note that the \mathbf{w}_i form a linearly independent set. Taking the transpose of this last equation yields

$$\mathbf{w}_i^T (A - \lambda_i B) = \mathbf{0}^T.$$

If \mathbf{w}_i has j -th component w_{ij} , then the k -th component of $\mathbf{w}_i^T (A - \lambda_i B) = \mathbf{0}^T$ is given by

$$\sum_{j=1}^n w_{ij} (a_{jk} - \lambda_i b_{jk}) = 0, \quad 1 \leq i \leq n.$$

If $T = (w_{ij})$, then this last equation is simply

$$\begin{aligned}(TA)_{ik} - \lambda_i(TB)_{ik} = 0 &\Leftrightarrow (TA)_{ik} - d_{ii}(TB)_{ik} = 0 \\ &\Leftrightarrow (TA)_{ik} - (DTB)_{ik} = 0.\end{aligned}$$

It then follows that $TA = DTB$. □

A consequence of this theorem is that if (IV.2) is not in normal form, then we can obtain a normal form by using the transformed system

$$TAu_x + TBu_y = Tc.$$

If $A^* = TA$, $B^* = TB$, and $c^* = Tc$, we have

$$A^*u_x + B^*u_y = c^* \quad \text{or} \quad DB^*u_x + B^*u_y = c^*.$$

So it is in compact form. It is in normal form because by construction, $d_{ii} = \lambda_i = \frac{dx}{dy}$, $1 \leq i \leq n$.

When we looked at second order linear PDEs, we saw that we could make use of characteristics to make transformations, which resulted in PDEs of a simpler form. We shall make use of characteristics here to do something similar.

Example 2.1. Let us consider the single quasi-linear PDE

$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u), \quad b(x, y, u) \neq 0.$$

So the matrix A is 1×1 and consists of the single function $a(x, y, u)$. Similarly, the matrix B is 1×1 and consists of the single nonzero function $b(x, y, u)$. Then if $d = a/b$, we have $a = d \times b$, and hence $b(du_x + u_y) = c$, that is, the PDE is in compact form.

Moreover, $\det(A - \lambda B) = a - \lambda b$ and is zero when $\lambda = a/b$. Thus with $d = a/b$, we conclude that $b(du_x + u_y) = c$ is in normal form. Moreover, the characteristic curves satisfy

$$\frac{dx}{dy} = d = \frac{a}{b}.$$

In the method of characteristics, we change from the coordinates x and y to new coordinates s and t such that the PDE becomes an ODE along the characteristic curves in the x - y plane. The variable t will vary along the characteristic curves so that, in a sense, it parametrises the characteristic curves. The ODE obtained is given by

$$\frac{du}{dt} = c.$$

To see this, by the chain rule we have

$$\frac{du}{dt} = \frac{\partial u}{\partial x} \frac{dx}{dt} + \frac{\partial u}{\partial y} \frac{dy}{dt}.$$

However, the PDE is $au_x + bu_y = c$. This then shows that we may obtain the ODE by taking

$$\frac{dx}{dt} = a \quad \text{and} \quad \frac{dy}{dt} = b.$$

With this choice, we indeed have t parametrising the characteristic curves since

$$\frac{dx}{dy} = \frac{\frac{dx}{dt}}{\frac{dy}{dt}} = \frac{a}{b} = d.$$

So along a characteristic curve, we have the equations

$$\frac{dx}{dt} = a, \quad \frac{dy}{dt} = b, \quad \frac{du}{dt} = c. \quad (\text{IV.3})$$

We then see that

$$\frac{du}{dt} = \frac{c}{b} \frac{dy}{dt}, \quad \text{and} \quad \frac{dx}{dt} = \frac{a}{b} \frac{dy}{dt}, \quad (\text{IV.4})$$

which, when rearranged, yields

$$\frac{dx}{a} = \frac{dy}{b} = \frac{du}{c} = \text{constant}. \quad (\text{IV.5})$$

So far we have been relatively silent about the coordinate s . As we shall see with examples later, it turns out that s is used to parametrise the initial curve associated with the initial condition for u . \square

The forms (IV.3) and (IV.5) are convenient ways to find a general solution for such PDEs. This is called the “method of characteristics” for first order PDEs and the equations in (IV.3) are known as the characteristic equations. This method is based on the following theorem.

Theorem IV.2. *A surface S given by $u = f(x, y)$ defines a solution to the quasi-linear first order PDE*

$$au_x + bu_y = c$$

if and only if the characteristic equations given in (IV.3) holds at each point of S .

Proof. If $u = f(x, y)$, then $f(x, y) - u = 0$. Taking the differential of this equation yields

$$\begin{aligned} 0 &= d(f - u) = df(x, y) - du = f_x dx + f_y dy - du \\ &= (f_x, f_y, -1) \cdot (dx, dy, du). \end{aligned}$$

This implies that

$$(f_x, f_y, -1) \cdot \left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{du}{dt} \right) = 0.$$

So if (IV.3) holds, then we have

$$(f_x, f_y, -1) \cdot (a, b, c) = 0 \Rightarrow af_x + bf_y = c,$$

that is, $f(x, y)$ satisfies the PDE.

Let S be defined by a solution $F(x, y, u) \equiv f(x, y) - u = 0$ of the PDE so that $af_x + bf_y = c$. Hence $(f_x, f_y, -1) \cdot (a, b, c) = 0$. In other words, at any point P of S , the vector (a, b, c) is perpendicular to the normal $\nabla F = (f_x, f_y, -1)$ to the surface. So (a, b, c) lies in the tangent plane for S at P .

Now a curve lying in S which passes through P will have (a, b, c) tangent to it. In particular, this curve $(x(t), y(t), u(t))$ satisfies

$$\left(\frac{dx}{dt}, \frac{dy}{dt}, \frac{du}{dt} \right) = (a, b, c).$$

This then yields (IV.3). □

§3 Method of characteristics

We wish to solve the quasi-linear first order PDE

$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u)$$

with initial condition $u = u_0(s)$. This initial condition is associated with the initial curve Γ defined by $x = F(s)$, $y = G(s)$, where for all s ,

$$\frac{F'(s)}{G'(s)} \neq \frac{a(F(s), G(s), u_0(s))}{b(F(s), G(s), u_0(s))} \equiv \frac{a|_{\Gamma}}{b|_{\Gamma}}.$$

In other words, Γ is not tangent to a characteristic of the PDE.

The idea is to “thread” a characteristic through each point on Γ , that is, we construct a characteristic curve emanating from $(F(s), G(s))$. From the previous example, we see that the characteristic equations for this PDE can be written (see (IV.3)) as

$$\frac{dx}{dt} = a, \quad \frac{dy}{dt} = b, \quad \frac{du}{dt} = c,$$

where t is chosen so that $t = 0$ means we are on Γ . These are to be solved subject to the given initial conditions

$$x(s, t = 0) = F(s), \quad y(s, t = 0) = G(s), \quad u(s, t = 0) = u_0(s). \quad (\text{IV.6})$$

Formally, by integration, we have

$$x(t) = \int a dt + X(s) = x(s, t), \quad (\text{IV.7})$$

and similarly

$$y = y(s, t), \quad u = u(s, t). \quad (\text{IV.8})$$

These are parametric equations for a surface S . But by the previous theorem, S is a solution surface for the PDE. Moreover, the initial conditions given in (IV.6) ensure that $\Gamma \times u_0$ lie in S . Hence *if* we can invert (IV.7) and the first equation in (IV.8) to solve for

$$s(x, y) \quad \text{and} \quad t(x, y),$$

we can use the second equation in (IV.8) to find the solution

$$u(x, y) = u(s(x, y), t(x, y)).$$

In fact, we can invert $x = x(s, t)$ and $y = y(s, t)$ in a neighbourhood of Γ because along Γ , the Jacobian

$$J = \left| \frac{\partial(x, y)}{\partial(s, t)} \right| = \begin{vmatrix} x_s & y_s \\ x_t & y_t \end{vmatrix} = x_s y_t - x_t y_s$$

does not vanish. To see this, on Γ , we have

$$J = x_s y_t - x_t y_s = x_s b - a y_s = bF'(s) - aG'(s) = bG' \left(\frac{F'}{G'} - \frac{a}{b} \right) \neq 0,$$

by our assumption.

Based on the above, we can solve the quasi-linear first order PDE $a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u)$ by using the following solution procedure:

(a) Write down the characteristic equations

$$\frac{dx}{dt} = a, \quad \frac{dy}{dt} = b, \quad \frac{du}{dt} = c \quad \text{or} \quad \frac{dx}{a} = \frac{dy}{b} = \frac{du}{c}.$$

(b) Solve them for $x(s, t)$, $y(s, t)$, and $u(s, t)$.

(c) Invert x and y to give $s(x, y)$ and $t(x, y)$.

(d) Substitute into $u(s, t)$ to give $u(x, y)$.

We now give some examples of this solution procedure.

Example 3.1. Solve

$$xu_x + yu_y = u,$$

given that $u = x^3$ on $xy = 1$, $x > 0$.

The characteristic equations are

$$\frac{dx}{dt} = x, \quad \frac{dy}{dt} = y, \quad \frac{du}{dt} = u.$$

The initial curve is $xy = 1$, so we take the initial conditions to be $x(s, 0) = s$, $y(s, 0) = 1/s$, $u(s, 0) = (x(s, 0))^3 = s^3$.

Upon solving, we find that

$$x(s, t) = c_1(s)e^t, \quad y(s, t) = c_2(s)e^t, \quad u(s, t) = c_3(s)e^t.$$

Applying the initial conditions shows that

$$x(s, t) = se^t, \quad y(s, t) = e^t/s, \quad u(s, t) = s^3e^t.$$

Inverting yields

$$s = (x/y)^{1/2} \quad \text{and} \quad t = \frac{1}{2} \log(xy).$$

Since $u(s, t) = s^3e^t$, the desired solution is then

$$u(x, y) = (x/y)^{3/2}(xy)^{1/2} = x^2y^{-1} = \frac{x^2}{y}. \quad \square$$

Example 3.2. Solve

$$u_x + u_y = -2u,$$

given that $u(x, 0) = \sin(x)$.

The characteristic equations are

$$\frac{dx}{dt} = 1, \quad \frac{dy}{dt} = 1, \quad \frac{du}{dt} = -2u.$$

The initial curve is $y = 0$, so we take the initial conditions to be $x(s, 0) = s$, $y(s, 0) = 0$, $u(s, 0) = \sin(x(s, 0)) = \sin(s)$.

Upon solving, we find that

$$x(s, t) = t + c_1(s), \quad y(s, t) = t + c_2(s), \quad u(s, t) = c_3(s)e^{-2t}.$$

Applying the initial conditions shows that

$$x(s, t) = t + s, \quad y(s, t) = t, \quad u(s, t) = \sin(s)e^{-2t}.$$

Inverting yields

$$s = x - y \quad \text{and} \quad t = y.$$

Since $u(s, t) = \sin(s)e^{-2t}$, the desired solution is then

$$u(x, y) = \sin(x - y)e^{-2y}. \quad \square$$

Example 3.3. Solve

$$xu_x + yuu_y = -xy,$$

given that $u = x^3$ on $xy = 1$, $x > 0$.

The characteristic equations are

$$\frac{dx}{dt} = x, \quad \frac{dy}{dt} = yu, \quad \frac{du}{dt} = -xy.$$

The initial curve is $xy = 1$, so as before, we take the initial conditions to be $x(s, 0) = s$, $y(s, 0) = 1/s$, $u(s, 0) = (x(s, 0))^3 = s^3$.

In this particular problem, though we can solve the first ODE as before, we cannot easily solve the second and third ODEs because we do not know u or the relationship between the variables x and y and the variable t .

However, we can make some progress on this problem by being a bit cleverer. Note that we have

$$\frac{d}{dt}(xy) = \frac{dx}{dt}y + x\frac{dy}{dt} = xy + xyu = -\frac{du}{dt} - \frac{du}{dt}u = -\frac{d}{dt}\left[u + \frac{u^2}{2}\right].$$

This implies that

$$xy = -u - \frac{u^2}{2} + f(s),$$

where $f(s)$ is an arbitrary function of s . The initial conditions show that

$$1 = -s^3 - \frac{s^6}{2} + f(s),$$

and hence

$$f(s) = 1 + s^3 + \frac{s^6}{2}.$$

This is as far as we can go. At this stage, it is not clear how s is related to x and y except we do know that $x(s, t) = se^t$.

Example 3.4. Solve

$$a(u)u_x + u_y = 0,$$

given that $u(x, 0) = u_0(x)$.

The characteristic equations are

$$\frac{dx}{dt} = a(u), \quad \frac{dy}{dt} = 1, \quad \frac{du}{dt} = 0.$$

The initial curve is $y = 0$, so we take the initial conditions to be $x(s, 0) = s$, $y(s, 0) = 0$, $u(s, 0) = u_0(x(s, 0)) = u_0(s)$.

Currently, we do not know u and hence cannot solve the first ODE. However, we can solve the other two ODEs and obtain

$$y(s, t) = t + c_1(s), \quad u(s, t) = c_3(s).$$

Applying the initial conditions shows that

$$y(s, t) = t, \quad u(s, t) = u_0(s).$$

For the first ODE, we now have $\frac{dx}{dt} = a(u_0(s))$. Hence

$$x(s, t) = a(u_0(s))t + f(s).$$

The initial condition shows that $f(s) = s$ and so $x(s, t) = a(u_0(s))t + s$. From above, $y = t$, and so we have $s = x - a(u_0(s))y = x - a(u)y$. We then conclude that $u(x, y) = u_0(s) = u_0(x - a(u)y)$, which is an implicit equation for u .

To get the solution explicitly, we need to be able to solve

$$\Phi(x, y, u) := u - u_0(x - a(u)y) = 0$$

for u as a function of x and y . To do this, we need

$$\frac{\partial \Phi}{\partial u} \neq 0,$$

that is,

$$1 - u'_0(x - a(u)y) \frac{\partial}{\partial u}(x - a(u)y) \neq 0 \Rightarrow 1 + u'_0(x - a(u)y) \left[y \frac{da}{du} \right] \neq 0.$$

This is always true for $|y|$ sufficiently small, and perhaps elsewhere too. \(\square\)

Example 3.5. Continuing the previous example, let $a(u) = u$, so that we have

$$uu_x + u_y = 0. \tag{IV.9}$$

This is a limiting case of Burgers' equation for inviscid flow. From above, we have $u(x, y) = u_0(s) = u_0(x - uy)$.

Now recall that the characteristics are solutions of the equation

$$\frac{dx}{dy} = u(x, y) \quad \text{and that} \quad \frac{du}{dt} = 0.$$

Since this last equation shows that u is constant along a characteristic curve, then we have $y = x/u(x, y) + c$. These are straight lines, but are not parallel. This means that some characteristic curves may cross. Since the solution is constant along each characteristic, a singularity will arise

whenever two characteristics cross; the values of u along the two characteristics are different and hence will become multi-valued at the point of crossing.

This example illustrates a common difficulty in nonlinear hyperbolic equations. The equation is a simple model for the formation of *shocks* in the flow of a gas. Not only does the solution of (IV.9) break down when two characteristics meet, but so does the mathematical model of the situation. Viscosity becomes important then, and the full, viscous Burgers' equation given by

$$uu_x + u_y = \mu u_{xx},$$

where μ is the viscosity, should be used. ⊠

Example 3.6. As a special case of Example 3.4, we consider the advection equation

$$au_x + u_y = 0,$$

where a is a constant. Suppose we have the points $x_0, x_1,$ and x_2 with $x_0 < x_1 < x_2$ and let the initial condition be given by

$$u(x, 0) = u_0(x) = \begin{cases} \phi(x), & x < x_1, \\ \psi(x), & x > x_1, \end{cases}$$

for some given functions ϕ and ψ with $\phi(x_1) \neq \psi(x_1)$. Hence the initial condition has a discontinuity at $x = x_1$.

From Example 3.4, the characteristic curves satisfy $x = at + s$ and $y = t$ so that $y = (x - s)/a$. Setting $s = x_j$ for $j = 0, 1, 2$, then the characteristic through $(x_j, 0)$ is $y = (x - x_j)/a$ and the solution along each characteristic is $u_0(x_j)$. Thus $u(x, (x - x_0)/a) = \phi(x_0)$ while $u(x, (x - x_2)/a) = \psi(x_2)$. If we let $x_0, x_2 \rightarrow x_1$, then we see that the solution must have a discontinuity along the characteristic which goes through $(x_1, 0)$, the point at which there is discontinuity in the initial condition. Hence, as in the previous example, we see that discontinuities can arise across the characteristics for hyperbolic equations. (It can be proved that solutions of parabolic and elliptic equations are analytic even when the boundary or initial conditions have discontinuities.) ⊠

Example 3.7. We solve (for constants $\alpha, \beta,$ and γ)

$$xu_x + yu_y = \alpha u + \beta$$

with initial condition $u(x, x^2) = x^\gamma$. The initial curve is $y = x^2$, so we take the initial conditions to be $x(s, 0) = s, y(s, 0) = s^2, u(s, 0) = (x(s, 0))^\gamma = s^\gamma$. The characteristic equations are

$$\frac{dx}{dt} = x, \quad \frac{dy}{dt} = y, \quad \frac{du}{dt} = \alpha u + \beta.$$

Hence, we have

$$x(s, t) = c_1(s)e^t, \quad y(s, t) = c_2(s)e^t, \quad u(s, t) = (c_3(s)e^{\alpha t} - \beta)/\alpha.$$

Applying the initial conditions shows that $x(s, t) = se^t$, $y(s, t) = s^2e^t$, and

$$u(s, t) = \frac{(\alpha s^\gamma + \beta)e^{\alpha t} - \beta}{\alpha}.$$

From the equations for $x(s, t)$ and $y(s, t)$, we have $s = y/x$ and $t = \log(x/s) = \log(x^2/y)$. Then we conclude that

$$u(x, y) = \frac{(\alpha(y/x)^\gamma + \beta)(x^2/y)^\alpha - \beta}{\alpha} = \frac{\alpha y^{\gamma-\alpha} x^{2\alpha-\gamma} + \beta x^{2\alpha} y^{-\alpha} - \beta}{\alpha}. \quad \square$$

Part V — Second order hyperbolic equations and the method of characteristics

So far, we have considered using characteristics to find exact solutions. However, it is possible to obtain a numerical method based on characteristics. We consider such a method for the second order PDE

$$Au_{xx} + Bu_{xy} + Cu_{yy} = G, \quad (\text{V.1})$$

where A , B , C , and G may be functions of u , u_x , and u_y , but not of u_{xx} , u_{xy} , or u_{yy} . Let us set

$$P = u_x, \quad Q = u_y, \quad R = u_{xx}, \quad S = u_{xy}, \quad T = u_{yy}.$$

Then

$$\frac{dP}{dx} = P_x + P_y \frac{dy}{dx} = R + S \frac{dy}{dx}, \quad (\text{V.2})$$

and

$$\frac{dQ}{dx} = Q_x + Q_y \frac{dy}{dx} = S + T \frac{dy}{dx}, \quad (\text{V.3})$$

while the original equation (V.1) may be written as

$$AR + BS + CT = G.$$

Upon solving for R and T in (V.2) and (V.3) respectively, the original equation then becomes

$$A \left(\frac{dP}{dx} - S \frac{dy}{dx} \right) + BS + C \left(\frac{dQ}{dx} \times \frac{dx}{dy} - S \frac{dx}{dy} \right) = G.$$

Upon multiplying by $-\frac{dy}{dx}$ and rearranging, we obtain

$$S \left[A \left(\frac{dy}{dx} \right)^2 - B \frac{dy}{dx} + C \right] - \left[A \frac{dP}{dx} \frac{dy}{dx} + C \frac{dQ}{dx} - G \frac{dy}{dx} \right] = 0. \quad (\text{V.4})$$

Now let us choose a curve in the x - y plane so that

$$A \left(\frac{dy}{dx} \right)^2 - B \frac{dy}{dx} + C = 0, \quad (\text{V.5})$$

that is, the S term is eliminated. By (V.4), it then follows that

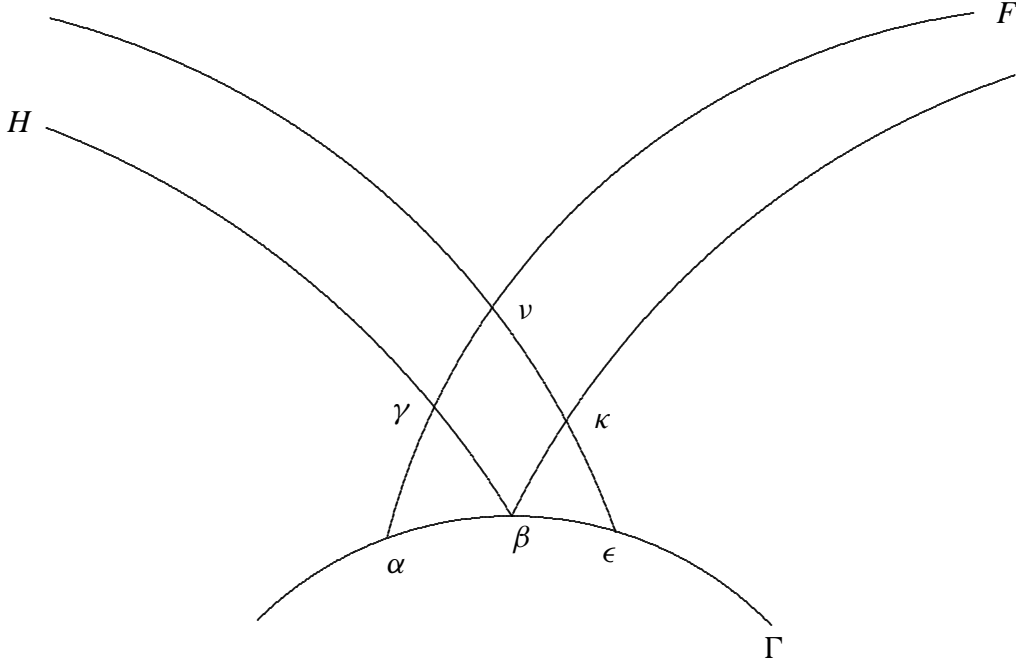
$$A \frac{dP}{dx} \frac{dy}{dx} + C \frac{dQ}{dx} - G \frac{dy}{dx} = 0. \quad (\text{V.6})$$

This shows that there could be up to two directions, given by the roots of the quadratic (V.5), where the relationship (V.6) holds.

Now recall that the characteristics are given by the solution of (V.5). So far, we have made no assumption about (V.1) being a hyperbolic equation. For (V.1) to be hyperbolic, we require

$B^2 - 4AC$ to be positive, that is, the roots of (V.5) are real and distinct. In this case, let us write the roots as F and H . Then the curve whose slope at any point is F is said to be a F -characteristic, and similarly a H -characteristic.

Let Γ be a *non-characteristic* curve along which initial values for u , P , and Q are known. Let $\alpha = (x_\alpha, y_\alpha)$ and $\beta = (x_\beta, y_\beta)$ be points on Γ that are close together and suppose the F -characteristic through α meets the H -characteristic through β at the point $\gamma = (x_\gamma, y_\gamma)$.



Let us treat the arcs $\overline{\alpha\gamma}$ and $\overline{\beta\gamma}$ as straight lines of slope F_α and H_β respectively. These slopes may be found by using (V.5). Then we have

$$y_\gamma - y_\alpha = F_\alpha(x_\gamma - x_\alpha) \quad (\text{V.7})$$

and

$$y_\gamma - y_\beta = H_\beta(x_\gamma - x_\beta), \quad (\text{V.8})$$

which gives equations for the two unknowns x_γ and y_γ . From (V.6) we have the relationships

$$A \frac{dP}{dx} F + C \frac{dQ}{dx} - G \frac{dy}{dx} = 0$$

and

$$A \frac{dP}{dx} H + C \frac{dQ}{dx} - G \frac{dy}{dx} = 0.$$

The first relationship can be approximated along $\overline{\alpha\gamma}$ by

$$A_\alpha \frac{P_\gamma - P_\alpha}{x_\gamma - x_\alpha} F_\alpha + C_\alpha \frac{Q_\gamma - Q_\alpha}{x_\gamma - x_\alpha} - G_\alpha \frac{y_\gamma - y_\alpha}{x_\gamma - x_\alpha} = 0,$$

where $A_\alpha = A(x_\alpha, y_\alpha)$ etc., while the second relationship may be approximated along $\overline{\beta\gamma}$ by

$$A_\beta \frac{P_\gamma - P_\beta}{x_\gamma - x_\beta} H_\beta + C_\beta \frac{Q_\gamma - Q_\beta}{x_\gamma - x_\beta} - G_\beta \frac{y_\gamma - y_\beta}{x_\gamma - x_\beta} = 0.$$

Thus we have

$$A_\alpha(P_\gamma - P_\alpha)F_\alpha + C_\alpha(Q_\gamma - Q_\alpha) - G_\alpha(y_\gamma - y_\alpha) = 0 \quad (\text{V.9})$$

and

$$A_\beta(P_\gamma - P_\beta)H_\beta + C_\beta(Q_\gamma - Q_\beta) - G_\beta(y_\gamma - y_\beta) = 0. \quad (\text{V.10})$$

Once x_γ and y_γ have been calculated from (V.7) and (V.8), these two equations yield P_γ and Q_γ . The value of u at $\gamma = (x_\gamma, y_\gamma)$ (which we write as $u_\gamma = u(x_\gamma, y_\gamma)$) may then be obtained from

$$\frac{du}{dx} = u_x + u_y \frac{dy}{dx} = P + Q \frac{dy}{dx}.$$

This is done by replacing the values of P and Q along $\overline{\alpha\gamma}$ by their average values and approximating this last equation by

$$\frac{u_\gamma - u_\alpha}{x_\gamma - x_\alpha} = \frac{1}{2}(P_\alpha + P_\gamma) + \frac{1}{2}(Q_\alpha + Q_\gamma) \frac{y_\gamma - y_\alpha}{x_\gamma - x_\alpha}$$

or

$$u_\gamma = u_\alpha + \frac{1}{2}(P_\alpha + P_\gamma)(x_\gamma - x_\alpha) + \frac{1}{2}(Q_\alpha + Q_\gamma)(y_\gamma - y_\alpha). \quad (\text{V.11})$$

This first approximation for u_γ can be improved by replacing values of the various coefficients by average values. Thus (V.7) and (V.8) become

$$y_\gamma - y_\alpha = \frac{1}{2}(F_\alpha + F_\gamma)(x_\gamma - x_\alpha) \quad (\text{V.12})$$

and

$$y_\gamma - y_\beta = \frac{1}{2}(H_\beta + H_\gamma)(x_\gamma - x_\beta), \quad (\text{V.13})$$

which yield improved values of x_γ and y_γ . One can consider (V.7) and (V.8) to be application of Euler's method as a predictor while these last two equations may be considered to be application of the trapezoidal method as a corrector. Similarly, we can obtain improved values of P_γ and Q_γ by modifying (V.9) and (V.10) to

$$\frac{1}{2}(A_\alpha + A_\gamma)(P_\gamma - P_\alpha) \frac{1}{2}(F_\alpha + F_\gamma) + \frac{1}{2}(C_\alpha + C_\gamma)(Q_\gamma - Q_\alpha) - \frac{1}{2}(G_\alpha + G_\gamma)(y_\gamma - y_\alpha) = 0 \quad (\text{V.14})$$

and

$$\frac{1}{2}(A_\beta + A_\gamma)(P_\gamma - P_\beta) \frac{1}{2}(H_\beta + H_\gamma) + \frac{1}{2}(C_\beta + C_\gamma)(Q_\gamma - Q_\beta) - \frac{1}{2}(G_\beta + G_\gamma)(y_\gamma - y_\beta) = 0. \quad (\text{V.15})$$

An improved value for $u_\gamma = u(x_\gamma, y_\gamma)$ may then be obtained from (V.11), while the values of $A_\gamma, B_\gamma, C_\gamma$ may be used in (V.5) to find improved values of F_γ and H_γ . An iterative procedure based on (V.12), (V.13), (V.14), (V.15), (V.11), and use of (V.5) to improve F_γ and H_γ may then be carried out until all of $x_\gamma, y_\gamma, P_\gamma, u_\gamma$ converge. If α and β are close together, the number

of iterations required will usually be small. We remark that if A , B , C , and G are constant, then there is no point in doing this improvement process as the values will not change.

In this way we can calculate solution values at the grid points γ and κ (see previous diagram), and then proceed to the grid point ν , and so on. As for one-dimensional hyperbolic equations, discontinuities in initial conditions are propagated as discontinuities into the solution domain along the characteristics. In such a situation, the method of characteristics is probably the best technique. However, if there are no discontinuities, finite difference methods should be satisfactory.

Part VI — Numerical solution of PDEs using finite differences

§1 Finite difference formulas

The techniques we shall now consider are finite difference methods. Such methods differ with the types of the PDE and boundary conditions, but the central feature of each method is the approximation of derivatives by finite differences. We shall concentrate on parabolic equations, but shall have a brief look at hyperbolic equations as well. Here we shall use the independent variable t (for time) rather than y .

If the function f is sufficiently differentiable, then

$$f(x \pm h) = f(x) \pm hf'(x) + \frac{h^2}{2!}f''(x) \pm \frac{h^3}{3!}f^{(3)}(x) + \dots$$

Using this equation, we can derive finite difference formulas such as

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}f''(\xi) \text{ — forward difference,}$$

$$f'(x) = \frac{f(x) - f(x-h)}{h} + \frac{h}{2}f''(\xi) \text{ — backward difference,}$$

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6}f^{(3)}(\xi) \text{ — central difference,}$$

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi) \text{ — central difference.}$$

Thus if u is sufficiently differentiable, we have

$$u_{xx}(x, t) = \frac{u(x+h, t) - 2u(x, t) + u(x-h, t)}{h^2} + O(h^2).$$

§2 Parabolic equations — an explicit method

We now consider the numerical solution of the problem

$$u_t = u_{xx}, \quad 0 \leq x \leq 1, \quad t \geq 0, \quad (\text{VI.1})$$

where $u(x, 0) = f(x)$ and $u(0, t) = u(1, t) = 0$ with $f(0) = f(1) = 0$. Here the boundary conditions are Dirichlet conditions.

Now let $h = 1/M$ for some positive integer M , and let k be the increment in t . Define $x_i = ih$, $0 \leq i \leq M$, and $t_n = nk$ for $n = 0, 1, 2, \dots$. We use finite differences to approximate the solution u at the point (x_i, t_n) of the x - t plane in the region

$$\Omega = \{(x, t) : 0 \leq x \leq 1, t \geq 0\}.$$

Using a central difference approximation for u_{xx} and a forward difference approximation for u_t , we have

$$u_{xx}(x_i, t_n) = \frac{u(x_i + h, t_n) - 2u(x_i, t_n) + u(x_i - h, t_n)}{h^2} - \frac{h^2}{12}u_{xxxx}(x_i + \xi_i h, t_n),$$

and

$$u_t(x_i, t_n) = \frac{u(x_i, t_n + k) - u(x_i, t_n)}{k} - \frac{k}{2}u_{tt}(x_i, t_n + \eta_n k),$$

where $-1 < \xi_i < 1$ and $0 < \eta_n < 1$. Let us denote $u(x_i, t_n)$ by u_i^n . Substitution into (VI.1) yields

$$\frac{u_i^{n+1} - u_i^n}{k} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} + \frac{k}{2}u_{tt}(x_i, t_n + \eta_n k) - \frac{h^2}{12}u_{xxxx}(x_i + \xi_i h, t_n). \quad (\text{VI.2})$$

With the higher order terms neglected, we obtain an approximation y_i^n to u_i^n with a truncation error of $O(h^2) + O(k)$ given by

$$\frac{y_i^{n+1} - y_i^n}{k} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2}. \quad (\text{VI.3})$$

If we set $r = k/h^2$, we obtain

$$y_i^{n+1} = r y_{i+1}^n + (1 - 2r)y_i^n + r y_{i-1}^n. \quad (\text{VI.4})$$

Thus we see that the approximation at time t_{n+1} is dependent on the approximations at time t_n at the x -values x_{i-1} , x_i , and x_{i+1} . Using the boundary conditions, we have

$$u(x, 0) = f(x) \Rightarrow y_i^0 = u(x_i, 0) = f(x_i), \quad 1 \leq i \leq M - 1,$$

and

$$u(0, t) = u(1, t) = 0 \Rightarrow y_0^n = y_M^n = 0, \quad n = 0, 1, 2, \dots$$

Taking $n = 0$ in (VI.4) yields y_i^1 for $1 \leq i \leq M - 1$. Then we can obtain y_i^2 for $1 \leq i \leq M - 1$ etc. Since each y_i^{n+1} is calculated directly from (VI.4) by using known values at the previous t -step, the method is, not surprisingly, called an explicit method. In matrix form, the method is given by $\mathbf{y}^{n+1} = \mathbf{A}\mathbf{y}^n$, where \mathbf{A} is the $(M - 1) \times (M - 1)$ matrix given by

$$\begin{bmatrix} 1 - 2r & r & 0 & \cdots & 0 & 0 & 0 \\ r & 1 - 2r & r & \cdots & 0 & 0 & 0 \\ 0 & r & 1 - 2r & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 - 2r & r & 0 \\ 0 & 0 & 0 & \cdots & r & 1 - 2r & r \\ 0 & 0 & 0 & \cdots & 0 & r & 1 - 2r \end{bmatrix},$$

and

$$\mathbf{y}^n = \begin{bmatrix} y_1^n \\ y_2^n \\ \vdots \\ y_{M-2}^n \\ y_{M-1}^n \end{bmatrix}.$$

We shall shortly see that we cannot expect good results when $r > \frac{1}{2}$. We firstly look at convergence. From (VI.2), we have

$$u_i^{n+1} = ru_{i+1}^n + (1 - 2r)u_i^n + ru_{i-1}^n + k \left[\frac{k}{2}u_{tt}(x_i, t_n + \eta_n k) - \frac{h^2}{12}u_{xxxx}(x_i + \xi_i h, t_n) \right].$$

If $e_i^n = u_i^n - y_i^n$, then subtracting (VI.4) from this last equation yields

$$e_i^{n+1} = re_{i+1}^n + (1 - 2r)e_i^n + re_{i-1}^n + k \left[\frac{k}{2}u_{tt}(x_i, t_n + \eta_n k) - \frac{h^2}{12}u_{xxxx}(x_i + \xi_i h, t_n) \right].$$

If u is sufficiently differentiable, we see that the term in the square brackets is bounded by $M_1 k + M_2 h^2$. If we set $E_n = \max_{0 \leq i \leq M} |e_i^n|$, we see that for $0 < r \leq \frac{1}{2}$ (so that $|1 - 2r| = 1 - 2r$) we have

$$E_{n+1} \leq rE_n + (1 - 2r)E_n + rE_n + k[M_1 k + M_2 h^2] = E_n + k[M_1 k + M_2 h^2].$$

Using induction, it is easy to prove that

$$E_{n+1} \leq E_0 + (n + 1)k[M_1 k + M_2 h^2] = t_{n+1}[M_1 k + M_2 h^2].$$

Letting $(h, k) \rightarrow (0, 0)$, we have $E_n \rightarrow 0$ or $|e_i^n| \rightarrow 0$. Hence the explicit method is convergent for $0 < r \leq \frac{1}{2}$.

It is clear that the errors present in the approximations at $t = t_n$ will affect the accuracy of the approximations for $t = t_{n+1}$. Loosely speaking, the method is said to be stable if errors from whatever source are not magnified as the iteration progresses. Hence they do not accumulate and destroy the accuracy indicated by the truncation errors. Let us assume that \mathbf{e}^0 is the error made in obtaining $\mathbf{y}^0 = [f(x_1), \dots, f(x_{M-1})]^T$, and suppose no other errors are made. Then

$$\mathbf{y}^1 = A(\mathbf{y}^0 + \mathbf{e}^0), \mathbf{y}^2 = A\mathbf{y}^1 = A^2\mathbf{y}^0 + A^2\mathbf{e}^0,$$

and in general

$$\mathbf{y}^n = A^n\mathbf{y}^0 + A^n\mathbf{e}^0.$$

Hence at the n -th step, the error that comes from \mathbf{e}^0 is $A^n\mathbf{e}^0$. For this term to be bounded, we require $\rho_\sigma(A) \leq 1$. (If A has eigenvalues λ_i , $1 \leq i \leq M - 1$, then

$$\rho_\sigma(A) = \max_{1 \leq i \leq M-1} |\lambda_i|.)$$

Since the eigenvalues of A are

$$\lambda_i = 1 - 2r + 2r \cos\left(\frac{i\pi}{M}\right) = 1 - 4r \sin^2\left(\frac{i\pi}{2M}\right), \quad 1 \leq i \leq M-1,$$

we require

$$\left|1 - 4r \sin^2\left(\frac{i\pi}{2M}\right)\right| \leq 1, \quad 1 \leq i \leq M-1.$$

This is equivalent to requiring that

$$0 < 4r \sin^2\left(\frac{i\pi}{2M}\right) \leq 2,$$

which is satisfied when

$$r \leq \frac{1}{2 \sin^2\left(\frac{i\pi}{2M}\right)}.$$

This holds whenever $r \leq \frac{1}{2}$, which is the condition that we imposed earlier on to ensure convergence.

In the analysis of stability, we have used a matrix approach. This is one of the standard ways of investigating the growth of errors. Two other methods are the energy method and the Fourier or von Neumann method. The energy method is more general, but tends to be rather messy to apply. The Fourier method is based on Fourier analysis and we now take a closer look at this technique.

In this method we assume that e_i^0 (the error at $x = x_i, t = t_0 = 0$) is given by

$$e_i^0 = \sum_{m=0}^M \gamma_m e^{i\beta_m x_i},$$

where $i^2 = -1$, the β_m are real numbers, and the γ_m are the Fourier coefficients. For the explicit method being considered, we see that if we ignore the truncation error, then

$$e_i^{n+1} = r e_{i+1}^n + (1 - 2r) e_i^n + r e_{i-1}^n.$$

We see that this finite difference equation is linear. This means that we need consider the propagation of the error due only to a single, typical term. Suppose we consider the typical frequency $|\beta| = |\beta_\ell|$. The Fourier coefficient γ_ℓ is constant and can be neglected.

To study the propagation of this single, typical error as $t \rightarrow \infty$, we want a solution of the finite difference equation above which reduces to $e^{i\beta x_i}$ when $t = 0$. It may be shown that such a solution is given by

$$e_i^n = e^{i\beta x_i} e^{\alpha t_n} = e^{i\beta i h} e^{\alpha n k},$$

where $\alpha = \alpha(\beta)$ is a complex number. For the error not to grow as $t \rightarrow \infty$, we require that for any α , we have

$$|e^{\alpha k}| \leq 1 \text{ for all } n.$$

This is known as *von Neumann's criterion for stability*.

Thus to study stability by using the Fourier method, we substitute $e_i^n = e^{i\beta x_i} e^{\alpha t_n}$ into the difference equation above. This yields

$$e^{i\beta h} e^{\alpha(n+1)k} = r e^{i\beta(i+1)h} e^{\alpha n k} + (1 - 2r) e^{i\beta i h} e^{\alpha n k} + r e^{i\beta(i-1)h} e^{\alpha n k},$$

which is equivalent to

$$e^{\alpha k} = r e^{i\beta h} + (1 - 2r) + r e^{-i\beta h}.$$

Thus to satisfy the von Neumann criterion for stability, we require

$$\left| r e^{i\beta h} + (1 - 2r) + r e^{-i\beta h} \right| = |2r \cos(\beta h) + (1 - 2r)| \leq 1.$$

This leads to us requiring $|1 - 2r + 2r \cos(\beta/M)| = |1 - 4r \sin^2(\beta/2M)| \leq 1$, which is satisfied when

$$r \leq \frac{1}{2 \sin^2\left(\frac{\beta}{2M}\right)}.$$

This certainly holds when $r \leq \frac{1}{2}$ which is the conclusion we came to when we analysed stability using a matrix approach.

From this analysis of convergence and stability just covered, we require $0 < r \leq \frac{1}{2}$. However, this condition on r imposes severe constraints on the explicit method. We recall that the finite difference approximations have truncation error $O(h^2) + O(k)$ so that for good accuracy, we require small values of h and k . Then the condition $0 < r \leq \frac{1}{2}$ means that we require $k/h^2 \leq \frac{1}{2}$ or $k \leq h^2/2$. Thus increments in t may have to be extremely small. So an enormous amount of computation may be required to make any reasonable advance in the t direction. This problem is overcome by implicit methods.

§3 Parabolic equations — an implicit method

In the explicit method looked at in the last section, $u_t(x_i, t_n)$ was approximated by a forward difference. Let us now approximate $u_t(x_i, t_n)$ by a backward difference. In particular, we have

$$u_t(x_i, t_n) = \frac{u(x_i, t_n) - u(x_i, t_n - k)}{k} + \frac{k}{2} u_{tt}(x_i, t_n - \eta_n k).$$

Thus we get

$$\frac{u_i^n - u_i^{n-1}}{k} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} - \frac{k}{2} u_{tt}(x_i, t_n - \eta_n k) - \frac{h^2}{12} u_{xxxx}(x_i + \xi_i h, t_n).$$

Then we get approximations y_i^n satisfying

$$\frac{y_i^n - y_i^{n-1}}{k} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2}. \quad (\text{VI.5})$$

As for the explicit method, the truncation error is $O(h^2) + O(k)$. Setting $r = k/h^2$ again, we have

$$y_i^{n-1} = -ry_{i-1}^n + (1 + 2r)y_i^n - ry_{i+1}^n. \quad (\text{VI.6})$$

In this equation, we see that three approximations at the n -th t -step are related to an approximation at the $(n - 1)$ -th t -step, and so unlike the explicit method, we cannot solve for a single unknown in terms of previously calculated values. Method such as (VI.6) are called implicit methods.

As before, the boundary conditions yield

$$y_i^0 = u(x_i, 0) = f(x_i), \quad 1 \leq i \leq M - 1,$$

and

$$y_0^n = u(0, t_n) = 0 = u(1, t_n) = y_M^n, \quad n = 0, 1, 2, \dots$$

Then we get the system $A\mathbf{y}^n = \mathbf{y}^{n-1}$, where \mathbf{y}^n was as before, and

$$A = \begin{bmatrix} 1 + 2r & -r & 0 & \cdots & 0 & 0 & 0 \\ -r & 1 + 2r & -r & \cdots & 0 & 0 & 0 \\ 0 & -r & 1 + 2r & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + 2r & -r & 0 \\ 0 & 0 & 0 & \cdots & -r & 1 + 2r & -r \\ 0 & 0 & 0 & \cdots & 0 & -r & 1 + 2r \end{bmatrix}.$$

Since A is diagonally dominant, then it may be shown by considering $A\mathbf{x} = \mathbf{0}$ that A is nonsingular. Hence the system $A\mathbf{y}^n = \mathbf{y}^{n-1}$ has a unique solution. So we see that we need to solve a system of $M - 1$ linear equations at each step. This means that the implicit method requires more computation than the explicit method already looked at. However, as we shall see, this is compensated by the fact that the implicit method can use larger values of k .

It may be shown that the implicit method is convergent regardless of the value of r . Also, for stability, we require $\rho_\sigma(A^{-1}) \leq 1$. Now the eigenvalues of A are given by

$$\lambda_i = 1 + 2r - 2r \cos\left(\frac{i\pi}{M}\right) = 1 + 4r \sin^2\left(\frac{i\pi}{2M}\right) \geq 1, \quad 1 \leq i \leq M - 1.$$

Since the eigenvalues of A^{-1} are just $1/\lambda_i$, we see that $\rho_\sigma(A^{-1}) \leq 1$ irrespective of the value of r . Thus the method is stable for all values of r . So the method is convergent and stable for all positive h and k . In this case, the method is said to be *unconditionally stable*.

We remark that the matrix A is symmetric and tridiagonal. Moreover, because the eigenvalues are positive, then A is positive definite. A Cholesky LU decomposition (with $U = L^T$) may then be used to solve the linear equations.

Suppose we now use the Fourier method to look at stability. Then by ignoring the truncation error, it may be shown that (compare (VI.6))

$$e_i^{n-1} = -re_{i-1}^n + (1 + 2r)e_i^n - re_{i+1}^n.$$

Substitution of $e_i^n = e^{i\beta x_i} e^{\alpha t_n}$ yields

$$e^{i\beta i h} e^{\alpha(n-1)k} = -r e^{i\beta(i-1)h} e^{\alpha n k} + (1+2r) e^{i\beta i h} e^{\alpha n k} - r e^{i\beta(i+1)h} e^{\alpha n k},$$

which is equivalent to

$$e^{-\alpha k} = -r e^{-i\beta h} + (1+2r) - r e^{i\beta h}.$$

Thus to satisfy the von Neumann criterion for stability, we require

$$\frac{1}{|-r e^{-i\beta h} + (1+2r) - r e^{i\beta h}|} = \frac{1}{|-2r \cos(\beta h) + (1+2r)|} \leq 1.$$

This is equivalent to $|1+2r-2r \cos(\beta/M)| = |1+4r \sin^2(\beta/2M)| \geq 1$, which is satisfied for all values of r . This is the same conclusion we came to when using the matrix approach.

A generalization of the explicit method and implicit method considered so far is the weighted average approximation

$$\frac{y_i^n - y_i^{n-1}}{k} = \frac{1}{h^2} \left[\theta (y_{i+1}^n - 2y_i^n + y_{i-1}^n) + (1-\theta) (y_{i+1}^{n-1} - 2y_i^{n-1} + y_{i-1}^{n-1}) \right].$$

This may be rewritten as

$$\begin{aligned} -r\theta y_{i-1}^n + (1+2\theta r)y_i^n - r\theta y_{i+1}^n \\ = r(1-\theta)y_{i-1}^{n-1} + (1-2(1-\theta)r)y_i^{n-1} + r(1-\theta)y_{i+1}^{n-1}. \end{aligned}$$

When $\theta = 0$, we get the explicit method we had before, while $\theta = 1$ gives the fully implicit method just considered. When $\theta = \frac{1}{2}$, we get the Crank-Nicolson method to be considered in the next section. This weighted average approximation is unconditionally stable for $\frac{1}{2} \leq \theta \leq 1$, but for $0 \leq \theta < \frac{1}{2}$, we require

$$r \leq \frac{1}{2(1-2\theta)}$$

for stability.

§4 Parabolic equations — the Crank-Nicolson method

The implicit method looked at in the previous section had a truncation error of $O(h^2) + O(k)$. In this section, we look at two implicit methods which have a truncation error of $O(h^2) + O(k^2)$.

The first method approximates $u_t(x_i, t_n)$ by a central difference approximation, namely

$$u_t(x_i, t_n) = \frac{u(x_i, t_n + k) - u(x_i, t_n - k)}{2k} - \frac{k^2}{6} u_{ttt}(x_i, t_n + \gamma_n k),$$

where $-1 < \gamma_n < 1$. Then we have the method

$$\frac{y_i^{n+1} - y_i^{n-1}}{2k} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n}{h^2}. \quad (\text{VI.7})$$

For this method, it is not hard to see that the truncation error is $O(h^2) + O(k^2)$. Unfortunately, this method (sometimes called Richardson's method) has stability problems. In fact, it is unstable for all $r > 0$.

To overcome this problem, the forward difference at $t = t_n$ given by (VI.3) is averaged with the backward difference at $t = t_{n+1}$ given by (VI.5) (with $n + 1$ replacing n) to yield the Crank-Nicolson method given by

$$\frac{y_i^{n+1} - y_i^n + y_i^{n+1} - y_i^n}{2k} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n + y_{i+1}^{n+1} - 2y_i^{n+1} + y_{i-1}^{n+1}}{2h^2},$$

or

$$\frac{y_i^{n+1} - y_i^n}{k} = \frac{y_{i+1}^n - 2y_i^n + y_{i-1}^n + y_{i+1}^{n+1} - 2y_i^{n+1} + y_{i-1}^{n+1}}{2h^2}. \quad (\text{VI.8})$$

This method can be shown to have truncation error $O(h^2) + O(k^2)$. This $O(k^2)$ component comes from the fact that $(y_i^{n+1} - y_i^n)/k$ may be considered to be the central difference approximation at $t = t_n + k/2$.

With $r = k/h^2$, these above equations become

$$-\frac{r}{2}y_{i-1}^{n+1} + (1+r)y_i^{n+1} - \frac{r}{2}y_{i+1}^{n+1} = \frac{r}{2}y_{i-1}^n + (1-r)y_i^n + \frac{r}{2}y_{i+1}^n.$$

With \mathbf{y}^0 again given by the initial conditions at $t = 0$, we have the system $A\mathbf{y}^{n+1} = B\mathbf{y}^n$, where A and B are both tridiagonal matrices. They are given by

$$A = \begin{bmatrix} 1+r & -\frac{r}{2} & 0 & \cdots & 0 & 0 & 0 \\ -\frac{r}{2} & 1+r & -\frac{r}{2} & \cdots & 0 & 0 & 0 \\ 0 & -\frac{r}{2} & 1+r & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+r & -\frac{r}{2} & 0 \\ 0 & 0 & 0 & \cdots & -\frac{r}{2} & 1+r & -\frac{r}{2} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{r}{2} & 1+r \end{bmatrix}$$

and

$$B = \begin{bmatrix} 1-r & \frac{r}{2} & 0 & \cdots & 0 & 0 & 0 \\ \frac{r}{2} & 1-r & \frac{r}{2} & \cdots & 0 & 0 & 0 \\ 0 & \frac{r}{2} & 1-r & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1-r & \frac{r}{2} & 0 \\ 0 & 0 & 0 & \cdots & \frac{r}{2} & 1-r & \frac{r}{2} \\ 0 & 0 & 0 & \cdots & 0 & \frac{r}{2} & 1-r \end{bmatrix}.$$

The matrix A is diagonally dominant, and hence nonsingular. So we have $\mathbf{y}^{n+1} = A^{-1}B\mathbf{y}^n$. For stability, we require $\rho_\sigma(A^{-1}B) \leq 1$. Now we see that $B = 2I_{M-1} - A$, where I_{M-1} is the $(M-1) \times (M-1)$ identity matrix, and so $A^{-1}B = 2A^{-1} - I_{M-1}$. This means that the

eigenvalues of $A^{-1}B$ are given by $\lambda_j = (2 - \mu_j)/\mu_j$, where μ_j , $1 \leq j \leq M - 1$, are the eigenvalues of A . Since

$$\mu_j = 1 + r - r \cos\left(\frac{j\pi}{M}\right) = 1 + 2r \sin^2\left(\frac{j\pi}{2M}\right),$$

we have

$$\lambda_j = \frac{1 - 2r \sin^2(j\pi/(2M))}{1 + 2r \sin^2(j\pi/(2M))}, \quad 1 \leq j \leq M - 1.$$

Clearly $|\lambda_j| \leq 1$. Thus the method is stable for all values of h and k . Using the Fourier method leads to the same conclusion.

§5 Derivative boundary conditions

Typically in heat conduction problems, one finds boundary conditions that reflect Newton's law of cooling. So instead of the boundary conditions $u(0, t) = u(1, t) = 0$ that we've considered so far, we have boundary conditions like

$$u_x(0, t) = \alpha u(0, t) + \beta \quad \text{and} \quad u_x(1, t) = \gamma u(1, t) + \delta, \quad (\text{VI.9})$$

where $\alpha, \beta, \gamma, \delta$ are constants.

A fairly simple procedure may be used to incorporate conditions of the form (VI.9) into any of the three methods looked at so far. We first introduce two fictitious points $x_{-1} = -h$ and $x_{M+1} = 1 + h$. Then assume that the equation given by either (VI.3), (VI.5), or (VI.8) holds for $i = 0$ and $i = M$ as well. As an example, let us consider the method given by (VI.4) which is equivalent to (VI.3). Then

$$y_0^{n+1} = r y_1^n + (1 - 2r) y_0^n + r y_{-1}^n \quad (\text{VI.10})$$

and

$$y_M^{n+1} = r y_{M+1}^n + (1 - 2r) y_M^n + r y_{M-1}^n. \quad (\text{VI.11})$$

On the assumption that u can be extended to be sufficiently differentiable in the region $\tilde{\Omega} = \{(x, t) : -h \leq x \leq 1 + h, t \geq 0\}$, then

$$u_x(0, t_n) = \frac{u(x_1, t_n) - u(x_{-1}, t_n)}{2h} - \frac{h^2}{6} u_{xxx}(\xi_0 h, t_n).$$

Dropping the truncation error, we have

$$\frac{y_1^n - y_{-1}^n}{2h} = \alpha u(0, t_n) + \beta = \alpha y_0^n + \beta,$$

which leads to

$$y_{-1}^n = y_1^n - 2h\alpha y_0^n - 2h\beta.$$

Similarly,

$$\frac{y_{M+1}^n - y_{M-1}^n}{2h} = \gamma u(1, t_n) + \delta = \gamma y_M^n + \delta,$$

which gives

$$y_{M+1}^n = y_{M-1}^n + 2h\gamma y_M^n + 2h\delta.$$

Substituting these into (VI.10) and (VI.11) yields

$$\begin{aligned} y_0^{n+1} &= ry_1^n + (1 - 2r)y_0^n + r(y_1^n - 2h\alpha y_0^n - 2h\beta) \\ &= (1 - 2r - 2hr\alpha)y_0^n + 2ry_1^n - 2hr\beta, \end{aligned}$$

while

$$\begin{aligned} y_M^{n+1} &= r(y_{M-1}^n + 2h\gamma y_M^n + 2h\delta) + (1 - 2r)y_M^n + ry_{M-1}^n \\ &= 2ry_{M-1}^n + (1 - 2r + 2hr\gamma)y_M^n + 2hr\delta. \end{aligned}$$

Then we get the linear system $\mathbf{y}^{n+1} = \mathbf{A}\mathbf{y}^n + \mathbf{b}$, where \mathbf{A} is the $(M + 1) \times (M + 1)$ matrix given by

$$\begin{bmatrix} 1 - 2r(1 + h\alpha) & 2r & 0 & \cdots & 0 & 0 & 0 \\ r & 1 - 2r & r & \cdots & 0 & 0 & 0 \\ 0 & r & 1 - 2r & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 - 2r & r & 0 \\ 0 & 0 & 0 & \cdots & r & 1 - 2r & r \\ 0 & 0 & 0 & \cdots & 0 & 2r & 1 - 2r(1 - h\gamma) \end{bmatrix},$$

$$\mathbf{y}^n = \begin{bmatrix} y_0^n \\ y_1^n \\ \vdots \\ y_{M-1}^n \\ y_M^n \end{bmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} -2hr\beta \\ 0 \\ \vdots \\ 0 \\ 2hr\delta \end{bmatrix}.$$

Suppose the initial starting vector is $\mathbf{y}^0 + \mathbf{e}^0$, where \mathbf{e}^0 is some initial error. Then $\mathbf{y}^1 = \mathbf{A}(\mathbf{y}^0 + \mathbf{e}^0) + \mathbf{b}$, $\mathbf{y}^2 = \mathbf{A}^2(\mathbf{y}^0 + \mathbf{e}^0) + \mathbf{A}\mathbf{b} + \mathbf{b}$. In general,

$$\mathbf{y}^n = \mathbf{A}^n(\mathbf{y}^0 + \mathbf{e}^0) + \mathbf{A}^{n-1}\mathbf{b} + \mathbf{A}^{n-2}\mathbf{b} + \cdots + \mathbf{A}\mathbf{b} + \mathbf{b}.$$

Then we see that the propagation of the initial error is dependent only on \mathbf{A} . For the given \mathbf{A} , there does not appear to be an explicit expression for the eigenvalues. Using Gerschgorin's Theorem we see that the eigenvalues λ_j (assuming they are real) must satisfy $-2r < \lambda_j - (1 - 2r) < 2r$, as well as

$$-2r < \lambda_j - (1 - 2r(1 + h\alpha)) < 2r \quad \text{and} \quad -2r < \lambda_j - (1 - 2r(1 - h\gamma)) < 2r.$$

Hence we deduce that a sufficient condition for stability is

$$r \leq \min \left(\frac{1}{2}, \frac{1}{2 + h\alpha}, \frac{1}{2 - h\gamma} \right).$$

§6 A more general parabolic equation

The derivation of the heat equation $u_t = u_{xx}$ makes the assumption that the properties of the rod such as heat conductivity and cross-section are uniform in x . However, it is more likely that some of these properties will depend on x . For example, a dependence on x is often used to model the flow of heat in a thin bar whose cross-section depends on x . Thus we consider the numerical solution of the equation

$$u_t = a(x, t)u_{xx}$$

with appropriate boundary and initial conditions. We shall assume that the function a is strictly positive.

The explicit method given in Section 2 is extended in an obvious way to

$$y_i^{n+1} = ra_i^n y_{i+1}^n + (1 - 2ra_i^n)y_i^n + ra_i^n y_{i-1}^n,$$

where $a_i^n = a(x_i, t_n)$. The implementation is as before and the analysis of the error is similar. The stability condition is replaced by

$$\frac{k}{h^2}a(x, t) \leq \frac{1}{2}$$

for all x and t in the region of interest.

The weighted average approximation may also be generalized. One way of doing it is to use

$$\frac{y_i^n - y_i^{n-1}}{k} = \frac{a^*}{h^2} \left[\theta(y_{i+1}^n - 2y_i^n + y_{i-1}^n) + (1 - \theta)(y_{i+1}^{n-1} - 2y_i^{n-1} + y_{i-1}^{n-1}) \right],$$

where a^* is some value to be chosen. However, it is not clear what value of a^* should be used. One choice is to choose $a^* = a(x_i, t_n - k/2)$. If it is awkward to calculate $a(x_i, t_n - k/2)$, then an obvious alternative is to use

$$a^* = \frac{a(x_i, t_n) + a(x_i, t_{n-1})}{2}.$$

§7 Self-adjoint parabolic equations

Sometimes a parabolic equation may appear in the self-adjoint form

$$u_t = \frac{\partial}{\partial x}(p(x, t)u_x),$$

where p is assumed to be positive. This may be written as

$$u_t = pu_{xx} + p_x u_x$$

and it would be possible to use a finite difference method to solve this last equation. However, it is usually better to apply such methods to the original self-adjoint form. We have

$$p(x_i + h/2, t_n)u_x(x_i + h/2, t_n) \approx p(x_i + h/2, t_n)\frac{u_{i+1}^n - u_i^n}{h}$$

and also

$$p(x_i - h/2, t_n)u_x(x_i - h/2, t_n) \approx p(x_i - h/2, t_n)\frac{u_i^n - u_{i-1}^n}{h}.$$

Thus if

$$w(x, t) = \frac{\partial}{\partial x}(p(x, t)u_x),$$

then

$$w(x_i, t_n) \approx p(x_i + h/2, t_n)\frac{u_{i+1}^n - u_i^n}{h^2} - p(x_i - h/2, t_n)\frac{u_i^n - u_{i-1}^n}{h^2}.$$

This then yields the explicit scheme

$$\frac{y_i^{n+1} - y_i^n}{k} = p(x_i + h/2, t_n)\frac{y_{i+1}^n - y_i^n}{h^2} - p(x_i - h/2, t_n)\frac{y_i^n - y_{i-1}^n}{h^2},$$

which may be expressed as

$$y_i^{n+1} = [1 - r(p(x_i + h/2, t_n) + p(x_i - h/2, t_n))]y_i^n + rp(x_i + h/2, t_n)y_{i+1}^n + rp(x_i - h/2, t_n)y_{i-1}^n.$$

An error analysis along the lines of the ones that we have done previously shows that the method will converge if

$$2rP \leq 1,$$

where P is an upper bound for p in the region of interest.

§8 Finite difference methods for first order hyperbolic equations

In 1928, Courant, Friedrichs, and Lewy formulated a necessary condition, now known as the *CFL condition*, for the convergence of a finite difference approximation in terms of the concept of a *domain of dependence*. Let us consider the advection equation

$$u_t + au_x = 0, \tag{VI.12}$$

and for the moment assume that a is a constant. From Example 3.6 in Part IV, if the initial condition is $u(x, 0) = u_0(x)$, then the solution is $u(x, t) = u_0(x - at)$. Thus the solution at the point (x_i, t_{n+1}) is $u_0(x_i - at_{n+1})$. Also, recall from that example that the characteristics are straight lines and that the solution is constant along each characteristic. This means that the

value of the solution along the characteristic which goes through the points $(x_i - at_{n+1}, 0)$ and (x_i, t_{n+1}) is just $u_0(x_i - at_{n+1})$.

Suppose we consider a finite difference approximation to (VI.12). The simplest one is an explicit method so we get the method

$$\frac{y_i^{n+1} - y_i^n}{k} + a \frac{y_i^n - y_{i-1}^n}{h} = 0.$$

This may be rewritten as

$$\begin{aligned} y_i^{n+1} &= y_i^n - a \frac{k}{h} (y_i^n - y_{i-1}^n) \\ &= (1 - a\mu)y_i^n + a\mu y_{i-1}^n, \end{aligned}$$

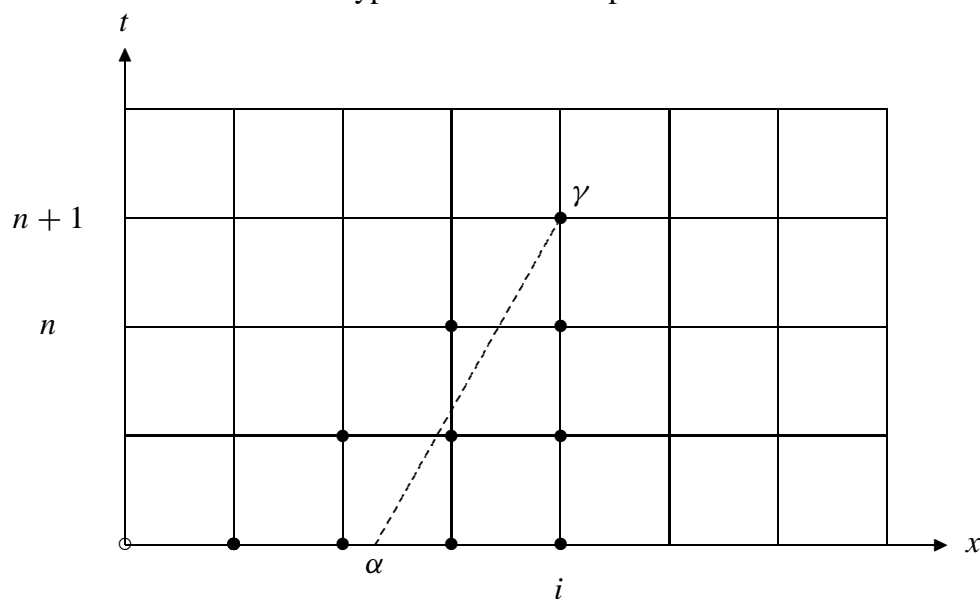
where $\mu = k/h$. The value of y_i^{n+1} depends on the values of y_i^n and y_{i-1}^n , that is, on the approximations at two points on the previous time level t_n . In turn, each of these two approximations depends on the approximations at the time level t_{n-1} etc. As illustrated in the diagram below, the value of y_i^{n+1} depends on data given in the triangle with vertex (x_i, t_{n+1}) . Ultimately, this value depends on the initial values at the $t_0 = 0$ line at the points

$$x_{i-n-1}, x_{i-n}, \dots, x_{i-1}, x_i.$$

This triangle is called the *domain of dependence* of y_i^{n+1} , or of the point (x_i, t_{n+1}) , for this particular numerical method.

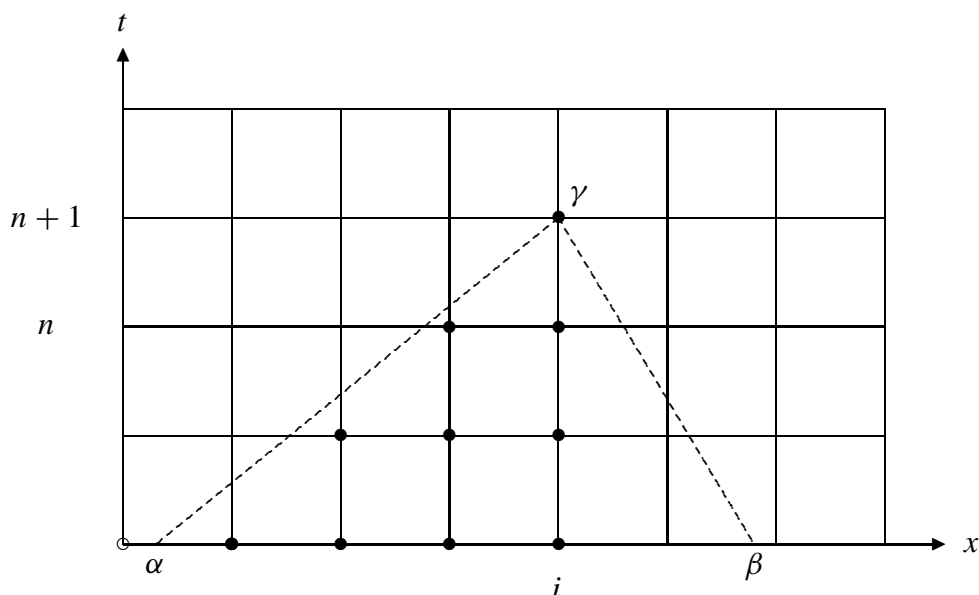
The corresponding domain of dependence of the PDE is the characteristic which goes through the point (x_i, t_{n+1}) (the line $\overline{\alpha\gamma}$ in the diagram below). The CFL condition then states that for a convergent scheme the domain of dependence of the the PDE must lie within the domain of dependence of the numerical scheme. In terms of the diagram, this means that the characteristic line $\overline{\alpha\gamma}$ must lie inside the triangle formed by the \bullet .

Typical domain of dependence



The next diagram shows two situations in which the CFL condition is not satisfied. Suppose $\overline{\alpha\gamma}$ and $\overline{\beta\gamma}$ are characteristic lines (one for $a > 0$ and one for $a < 0$). Clearly they lie outside the triangle. To see why the numerical method is not convergent when the CFL condition is not satisfied, suppose we alter the given initial conditions in some region around the point α . Suppose also that we were to reduce h and k in such a way that the ratio μ was constant; this ensures that the triangular domain of dependence remains the same. This change in the initial condition at α will change the solution of the PDE at γ since the solution is constant along the characteristic $\overline{\alpha\gamma}$. However, the numerical solution at γ is unchanged since the numerical data used to construct the approximations remains unchanged. Thus the numerical solution cannot converge to the required result at γ . A similar argument applies to the characteristic $\overline{\beta\gamma}$.

Violation of the CFL condition



Recalling from Example 3.6 in Part IV that the slope of the characteristic line is $1/a$, we see that the CFL condition cannot be satisfied for (VI.12) when $a < 0$, since the characteristic line would be like $\overline{\beta\gamma}$. For $a > 0$, we need to impose a condition on μ . Recall that at $t_0 = 0$, the left-most point of the domain of dependence is $(x_{i-n-1}, 0)$ and that the point α is $(x_i - at_n, 0)$. Thus for the CFL condition to be satisfied, we require $x_{i-n-1} \leq x_i - at_{n+1}$. Some algebra then shows that this requirement is equivalent to

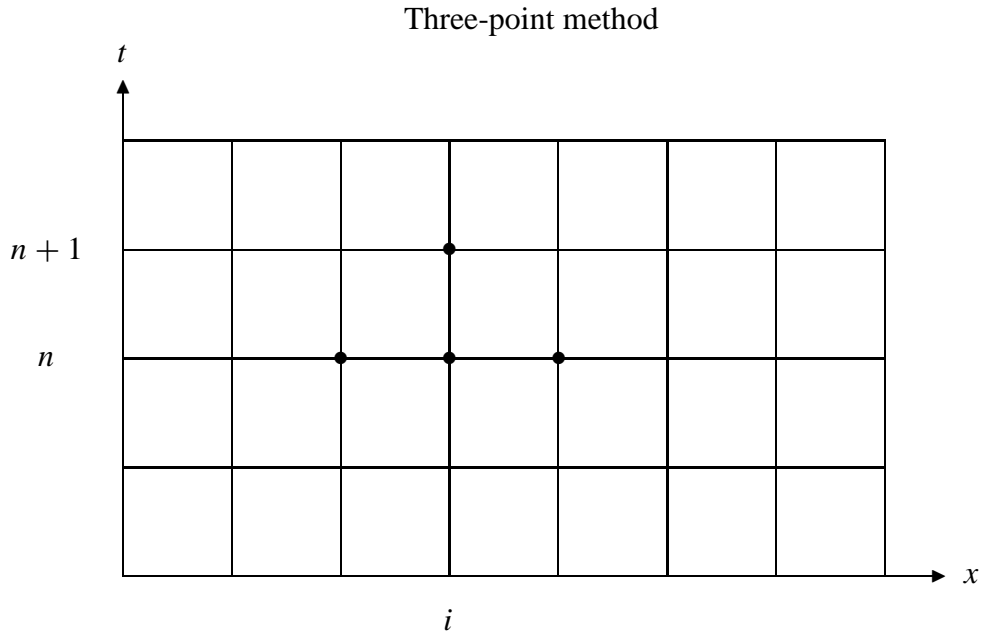
$$a\mu \leq 1.$$

As mentioned earlier, the CFL condition is necessary for convergence and hence (according to Lax's equivalence theorem) necessary for stability. We shall shortly see that it is not sufficient for stability. However, the CFL condition does allow us to reject schemes which does not satisfy the CFL condition as they will not be convergent or stable. Methods satisfying the CFL condition may then be tested further for stability.

We now consider approximating (VI.12) by a more general explicit method using three symmetrically placed points at the old time level (see diagram on the next page). It is then not hard

to see that the CFL condition for such a method is

$$|a|\mu \leq 1.$$



If $a > 0$, the difference method should use both y_{i-1}^n and y_i^n to obtain y_i^{n+1} . If $a < 0$, then it should use y_i^n and y_{i+1}^n . To cover both situations, an obvious method is to use a central difference in space combined with a forward difference in time to yield the method

$$\frac{y_i^{n+1} - y_i^n}{k} + a \frac{y_{i+1}^n - y_{i-1}^n}{2h} = 0.$$

Provided $|a|\mu \leq 1$, the CFL condition is satisfied. However, as mentioned earlier, this condition is not sufficient to guarantee stability. To investigate stability, let us use the Fourier method. Setting $e_i^n = e^{i\beta x_i} e^{\alpha t_n}$, we then have

$$e^{i\beta x_i} e^{\alpha t_{n+1}} - e^{i\beta x_i} e^{\alpha t_n} + \frac{1}{2} a \mu (e^{i\beta x_{i+1}} e^{\alpha t_n} - e^{i\beta x_{i-1}} e^{\alpha t_n}) = 0,$$

which reduces to

$$e^{\alpha k} - 1 + \frac{1}{2} a \mu (e^{i\beta h} - e^{-i\beta h}) = 0.$$

Some algebra then yields

$$e^{\alpha k} = 1 - ia\mu \sin(\beta h).$$

It is clear that $|e^{\alpha k}| \geq 1$ (with equality only when βh is an integer multiple of π). Thus the method is not stable even though it satisfies the CFL condition.

§9 An upwind scheme

A three point method which is stable is the *upwind* scheme in which a backward difference is used if a is positive and a forward difference when a is negative. This method is given by

$$y_i^{n+1} = \begin{cases} y_i^n - a\mu(y_{i+1}^n - y_i^n), & \text{if } a < 0; \\ y_i^n - a\mu(y_i^n - y_{i-1}^n), & \text{if } a > 0. \end{cases} \quad (\text{VI.13})$$

If a is not a constant, but a function of x and t , we need to specify in (VI.13) the point at which a is to be evaluated. At the moment let us assume that we use $a(x_i, t_n)$. The CFL condition is satisfied when $|a|\mu \leq 1$. For the $a > 0$ case, it may be shown that the amplification factor is given by

$$e^{\alpha k} = 1 - a\mu(1 - e^{-i\beta h}).$$

This leads to

$$|e^{\alpha k}|^2 = 1 - 4a\mu(1 - a\mu) \sin^2(\beta h/2).$$

This quantity is not more than 1 when $0 < a\mu \leq 1$, the same condition as the CFL condition. When $a < 0$, the stability condition becomes $|a|\mu \leq 1$.

The upwind scheme of (VI.13) may be rewritten as

$$y_i^{n+1} = \begin{cases} (1 + a\mu)y_i^n - a\mu y_{i+1}^n, & \text{if } a < 0; \\ (1 - a\mu)y_i^n + a\mu y_{i-1}^n, & \text{if } a > 0. \end{cases}$$

This has the following interpretation. In the diagram on the next page for the case $a > 0$, the characteristic through the point $\gamma = (x_i, t_{n+1})$ meets the line $t = t_n$ at the point α , which by the CFL condition must lie between the point $\beta = (x_{i-1}, t_n)$ and the point $\kappa = (x_i, t_n)$. Recalling that the exact solution along this characteristic is constant, we see that $u(\alpha) = u(\gamma)$. If we know an approximation at all the points on the line $t = t_n$, we can interpolate the value of $u(\alpha)$ and use it to obtain the approximation y_i^{n+1} . If we use linear interpolation, then we have

$$y_i^{n+1} \approx u(\alpha) \approx y_{i-1}^n + \frac{\delta}{h}(y_i^n - y_{i-1}^n),$$

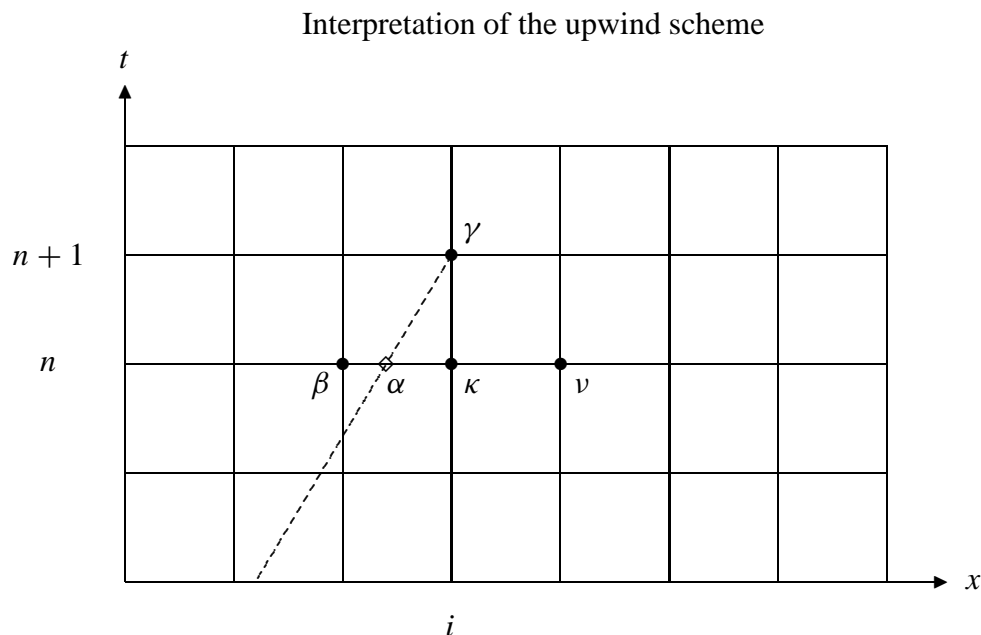
where δ is the distance on the $t = t_n$ line from β to α . When a is a constant, the slope of the characteristic line is $1/a$. It then follows from the diagram on the next page that

$$\frac{k}{h - \delta} = \frac{1}{a}.$$

We then obtain $\delta = h - ak$ so that $\delta/h = 1 - a\mu$, and hence

$$y_i^{n+1} \approx y_{i-1}^n + (1 - a\mu)(y_i^n - y_{i-1}^n) = (1 - a\mu)y_i^n + a\mu y_{i-1}^n,$$

that is, the approximation of the upwind scheme. If a is not a constant, but varies smoothly, we would still expect a good approximation.



§10 The Lax-Wendroff method

In the previous section we showed that the upwind scheme could be interpreted as a method in which the approximations at time $t = t_{n+1}$ are obtained from appropriate approximations at time $t = t_n$ by linear interpolation. One might expect that one could obtain better approximations by using quadratic interpolation instead. This leads to the Lax-Wendroff method. The quadratic interpolation is carried out at the points β , κ , and ν of the diagram above. Such a derivation yields the method

$$y_i^{n+1} = \frac{1}{2}a\mu(1 + a\mu)y_{i-1}^n + (1 - a^2\mu^2)y_i^n - \frac{1}{2}a\mu(1 - a\mu)y_{i+1}^n, \quad (\text{VI.14})$$

where we have assumed that a is a constant. The usual Fourier method shows that the amplification factor is

$$e^{\alpha k} = 1 - ia\mu \sin(\beta h) - 2a^2\mu^2 \sin^2(\beta h/2).$$

After separating the real and imaginary parts and doing some algebra, we obtain

$$|e^{\alpha k}|^2 = 1 - 4a^2\mu^2(1 - a^2\mu^2) \sin^4(\beta h/2).$$

Thus the method is stable if $|a|\mu \leq 1$, the same requirement as the CFL condition.

Suppose now that a is not a constant, but a function of x and t . Then the analogous method is obtained by first writing

$$u(x, t + k) = u(x, t) + ku_t(x, t) + \frac{1}{2}k^2u_{tt}(x, t) + O(k^3). \quad (\text{VI.15})$$

Since the PDE is $u_t + au_x = 0$ or

$$u_t = -au_x,$$

we have $u_{tt} = -a_t u_x - au_{xt}$ and $u_{tx} = (-au_x)_x$. Thus

$$u_{tt} = -a_t u_x + a(au_x)_x.$$

The expression for u_t and this last expression for u_{tt} may then be substituted into (VI.15). By approximating each of these x -derivatives by central differences, we obtain

$$\begin{aligned} y_i^{n+1} = & y_i^n - ka(x_i, t_n) \frac{y_{i+1}^n - y_{i-1}^n}{2h} - \frac{1}{2}k^2 a_t(x_i, t_n) \frac{y_{i+1}^n - y_{i-1}^n}{2h} \\ & + \frac{1}{2}k^2 a(x_i, t_n) \frac{a(x_i + h/2, t_n)(y_{i+1}^n - y_i^n) - a(x_i - h/2, t_n)(y_i^n - y_{i-1}^n)}{h^2}. \end{aligned}$$

If a is a constant, then $a_t(x_i, t_n) = 0$. Some algebra then shows that the resulting method is identical to the method given in (VI.14). The method may be simplified by replacing $a(x_i, t_n) + (k/2)a_t(x_i, t_n)$ by $a(x_i, t_n + k/2)$.

§11 The Lax-Wendroff method for conservation forms

In practical situations, one often obtains the PDE

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0. \quad (\text{VI.16})$$

If we take $b = b(u) = f_u$, then we may write (VI.16) as the hyperbolic equation $u_t + bu_x = 0$. Rather than using the Lax-Wendroff method for this latter equation, it is convenient to derive the Lax-Wendroff method directly for the conservation form (VI.16) (a reason for calling (VI.16) the ‘conservation form’ will be given later). The function f does not involve x or t implicitly, but is a function of u only. An example of such an equation is (IV.9), the limiting case of Burgers’ equation for inviscid flow.

Now we have $u_t = -f_x$ and

$$u_{tt} = -f_{xt} = -f_{tx} = -(f_t)_x = -(f_u u_t)_x = -(bu_t)_x = (bf_x)_x.$$

The derivation of the previous section then yields the method

$$\begin{aligned} y_i^{n+1} = & y_i^n - k \frac{f(y_{i+1}^n) - f(y_{i-1}^n)}{2h} \\ & + \frac{1}{2}k^2 \frac{f_u(y_{i+1/2}^n) (f(y_{i+1}^n) - f(y_i^n)) - f_u(y_{i-1/2}^n) (f(y_i^n) - f(y_{i-1}^n))}{h^2}. \end{aligned}$$

As expected, this reduces to (VI.14) when $f(u) = au$, where a is a constant. The method may be rewritten as

$$\begin{aligned} y_i^{n+1} = & y_i^n - \frac{1}{2}\mu \left[(1 - \mu f_u(y_{i+1/2}^n))(f(y_{i+1}^n) - f(y_i^n)) \right. \\ & \left. + (1 + \mu f_u(y_{i-1/2}^n))(f(y_i^n) - f(y_{i-1}^n)) \right], \end{aligned}$$

where $\mu = k/h$. In this method we see that we need to evaluate $f_u(y_{i-1/2}^n)$ and $f_u(y_{i+1/2}^n)$. To calculate these two quantities, it is usual to set

$$y_{i\pm 1/2}^n = \frac{1}{2}(y_i^n + y_{i\pm 1}^n).$$

One of the great strengths of the Lax-Wendroff method is that it can be extended quite easily to systems of equations. Thus in (VI.16), we can replace u and f with vectors \mathbf{u} and \mathbf{f} . This results in the system

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \mathbf{f}(\mathbf{u})}{\partial x} = \mathbf{0}.$$

Such systems arise in the theory of fluid flow when the equations of motion, continuity, and of energy are combined into one conservation equation (\mathbf{u} and \mathbf{f} will each have three components). Hence the reason why (VI.16) is called the conservation form. The corresponding equation $u_t + bu_x = 0$, where $b = f_u$, then becomes the hyperbolic system

$$\frac{\partial \mathbf{u}}{\partial t} + B \frac{\partial \mathbf{u}}{\partial x} = \mathbf{0},$$

where B is the Jacobian matrix.

We shall not consider the details of the Lax-Wendroff method for systems, but just mention that the wave equation $u_{tt} = \alpha^2 u_{xx}$ may be solved by such a method since it may be written as the first-order equations

$$u_t + \alpha v_x = 0 \text{ and } v_t + \alpha u_x = 0.$$

This may be expressed as the system

$$\begin{bmatrix} u_t \\ v_t \end{bmatrix} + \begin{bmatrix} \alpha v_x \\ \alpha u_x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} u_t \\ v_t \end{bmatrix} + \begin{bmatrix} 0 & \alpha \\ \alpha & 0 \end{bmatrix} \begin{bmatrix} u_x \\ v_x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Part VII — Weighed residual methods

§1 Some material in analysis

In the last Part, we considered finite difference methods in which the partial derivatives have been approximated by finite difference formulas. In this Part, we shall consider methods in which the solution is approximated directly by suitably-chosen approximating functions. The general principles of the methods that we consider may be presented in a fairly simple manner, but details for specific problems can require quite a sophisticated analysis.

We first present some ideas in analysis that we shall need. To allow generality, let us assume that we wish to solve

$$Lu = g,$$

where g is a known function, u is to be found, and L is a linear differential operator. Normally there are some constraints on u from boundary and/or initial conditions. Since the methods we shall be considering tend to be applied to elliptic problems (which are time independent), we shall assume that the constraints are boundary conditions. We remark that for time-dependent problems (such as parabolic and hyperbolic equations), the methods discussed here could be applied in the space variables to yields systems of ordinary differential equations.

Example 1.1. An example of a linear differential operator is

$$L = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \quad (\text{VII.1})$$

Then $Lu = 0$ is just Laplace's equation. ☒

The idea is to approximate u by u_N such that the residual $Lu_N - g$ is small in some sense.

Let U be a vector space of functions which contains u and which satisfy any boundary conditions. For the space U we can define an *inner product* $\langle \cdot, \cdot \rangle$, which has the following properties ($f, f_1, f_2 \in U$):

- (a) $\langle \alpha f_1, f_2 \rangle = \langle f_1, \alpha f_2 \rangle = \alpha \langle f_1, f_2 \rangle$ for all scalars α ;
- (b) $\langle f_1 + f_2, f \rangle = \langle f_1, f \rangle + \langle f_2, f \rangle$, $\langle f, f_1 + f_2 \rangle = \langle f, f_1 \rangle + \langle f, f_2 \rangle$;
- (c) $\langle f_1, f_2 \rangle = \langle f_2, f_1 \rangle$;
- (d) $\langle f, f \rangle \geq 0$ for all $f \in U$, and $\langle f, f \rangle = 0$ if and only if f is the zero function.

It may be shown that

$$\|f\| := \sqrt{\langle f, f \rangle}$$

has the properties of a norm. If U contains functions defined over some domain Ω , then an example of an inner product is

$$\langle f_1, f_2 \rangle = \int_{\Omega} f_1(\mathbf{x}) f_2(\mathbf{x}) \, d\mathbf{x}.$$

A u satisfying $Lu = g$ and the appropriate boundary conditions is known as the *classical solution*. A weak solution u is one which satisfies

$$\langle Lu - g, w \rangle = 0 \text{ for all } w \in U.$$

One may think of these w as weight functions. This last equation is the weak form of the differential equation. By using integration by parts (or higher dimensional analogues), it is possible to lower the differentiability requirements on the weak solution. For instance, if L is the differential operator given in (VII.1), then the classical solution u will need to have two partial derivatives in x and y . However, a weak solution needs to have only one derivative in x and y . To see this, suppose we wish to solve Laplace's equation on the unit square with boundary conditions $u(0, y) = u(1, y) = u(x, 0) = u(x, 1) = 0$. Then

$$\begin{aligned} \int_0^1 \int_0^1 [u_{xx} + u_{yy}]w \, dx \, dy &= \int_0^1 \left[u_x w \Big|_{x=0}^{x=1} - \int_0^1 u_x w_x \, dx \right] dy \\ &\quad + \int_0^1 \left[u_y w \Big|_{y=0}^{y=1} - \int_0^1 u_y w_y \, dy \right] dx \\ &= - \int_0^1 \int_0^1 u_x w_x \, dx \, dy - \int_0^1 \int_0^1 u_y w_y \, dy \, dx = 0, \end{aligned}$$

where we have used integration by parts in the first step and the boundary conditions in the second. Note that in the final form, we require u to have only first derivatives in x and y . So though the classical solution is always a solution of the weak form, a weak solution may not be a classical solution and may not even be differentiable!

§2 Weighted residual methods

Suppose we wish to solve $Lu = g$. Then $\langle Lu - g, w \rangle = 0$ for all $w \in U$. Let U_N be a N -dimensional subspace of U . In weighted residual methods, we find $u_N \in U_N$, such that

$$\langle Lu_N - g, v_N \rangle = 0 \quad \text{for all } v_N \in V_N,$$

where V_N is another approximating subspace of U . Let ψ_1, \dots, ψ_N be a basis for V_N . Writing

$$u_N = c_1\phi_1 + c_2\phi_2 + \dots + c_N\phi_N,$$

we obtain the equations

$$\sum_{j=1}^N \langle \psi_i, L\phi_j \rangle c_j = \langle g, \psi_i \rangle, \quad 1 \leq i \leq N.$$

(The functions ϕ_i are usually called *trial functions* and the ψ_i are usually called *test functions*.)

We wish to choose U_N and V_N such that the above equations have a solution so that u_N exists. Moreover, we want u_N to be a better approximation as N increases.

Different methods result from different choices of ψ_i :

- (a) If $\psi_i = \phi_i$, then we obtain the classical Galerkin (or Bubnov-Galerkin) method.
- (b) If $\psi_i \neq \phi_i$, then the resulting method is known as the Petrov-Galerkin method.
- (c) If $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ are N points in the domain of interest, another choice is to take $\psi_i(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}_i)$, where $\delta(\mathbf{x} - \mathbf{x}_i)$ is the Dirac delta function. Then the equations become

$$\sum_{j=1}^N (L\phi_j)(\mathbf{x}_i)c_j = g(\mathbf{x}_i), \quad 1 \leq i \leq n.$$

This method is more commonly known as the *collocation method*.

- (d) If $\psi_i = L\phi_i$, then the resulting method is the least squares method.
- (e) Subdomain methods arise when the domain is divided into N subdomains Ω_i , and the ψ_i are given by

$$\psi_i = \begin{cases} 1 & \text{inside } \Omega_i, \\ 0 & \text{outside } \Omega_i. \end{cases}$$

§3 An introduction to spectral methods and orthogonal polynomials

In finite element methods, the trial functions are local smooth functions (typically polynomials of fixed degree which are nonzero only on certain subdomains of Ω). By contrast, in spectral methods, the trial functions are taken to be global smooth functions. Typically these include Fourier series (for periodic problems) and orthogonal polynomials (for non-periodic problems).

Commonly used orthogonal polynomials are the Legendre polynomials and the Chebyshev polynomials of the first kind. The Legendre polynomials are given by

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_2(x) = \frac{1}{2}(3x^2 - 1),$$

and

$$(m+1)P_{m+1}(x) = (2m+1)xP_m(x) - mP_{m-1}(x), \quad x \in [-1, 1].$$

They are orthogonal on $[-1, 1]$ with respect to the weight function 1. We have

$$\int_{-1}^1 P_m(x)P_n(x) dx = \begin{cases} 0, & m \neq n; \\ \frac{2}{2m+1}, & m = n. \end{cases}$$

These Legendre polynomials satisfy the Sturm-Liouville equation

$$\frac{d}{dx} \left((1-x^2) \frac{dP_m}{dx} \right) + m(m+1)P_m(x) = 0.$$

The Chebyshev polynomials of the first kind $T_m(x) = \cos(m \cos^{-1}(x))$, $m = 0, 1, \dots$, are orthogonal on $[-1, 1]$ with respect to the weight function $1/\sqrt{1-x^2}$. We have

$$\int_{-1}^1 \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}} dx = \begin{cases} 0, & m \neq n; \\ \pi, & m = n = 0; \\ \pi/2, & m = n > 0. \end{cases}$$

We see that $T_0(x) = 1$ and $T_1(x) = x$. These polynomials satisfy the recurrence relation

$$T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x), \quad m \geq 1.$$

With $\theta = \cos^{-1}(x)$, this follows from:

$$T_{m+1}(x) + T_{m-1}(x) = \cos((m+1)\theta) + \cos((m-1)\theta) = 2\cos(\theta)\cos(m\theta) = 2xT_m(x).$$

Hence $T_2(x) = 2x^2 - 1$ etc.

Moreover, these Chebyshev polynomials satisfy the singular Sturm-Liouville equation

$$\frac{d}{dx} \left(\sqrt{1-x^2} \frac{dT_m}{dx} \right) + \frac{m^2}{\sqrt{1-x^2}} T_m(x) = 0.$$

More properties of these Chebyshev polynomials are given in the next theorem.

Theorem VII.1. *The Chebyshev polynomials T_m have the following properties:*

- (a) T_m is a polynomial of degree m with leading coefficient 2^{m-1} and is an even function when m is even, and an odd function when m is odd;
- (b) the roots of T_m are given by

$$x_i = \cos \left(\frac{2i+1}{2m} \pi \right), \quad 0 \leq i \leq m-1;$$

- (c) $T_m(\pm 1) = (\pm 1)^m$ and hence $\max_{x \in [-1, 1]} |T_m(x)| = 1$.

Proof. Part (a) follows by induction, while for part (b) we have

$$T_m(x_i) = \cos(m \cos^{-1}(x_i)) = \cos \left(m \frac{2i+1}{2m} \pi \right) = \cos \left(\frac{2i+1}{2} \pi \right) = 0.$$

Now for $x = 1$, we have $T_m(1) = \cos(m \cos^{-1}(1)) = \cos(0) = 1$ while for $x = -1$, we have $T_m(-1) = \cos(m \cos^{-1}(-1)) = \cos(m\pi) = (-1)^m$. Hence $\max_{x \in [-1, 1]} |T_m(x)| = 1$. Thus part

(c) is proved. \square

In some texts there is a distinction between Galerkin methods in which the trial functions individually satisfy the boundary conditions and tau-methods in which most of the test functions are the same as the trial functions, but the trial functions do not satisfy the boundary conditions. Other texts define tau-methods to be methods in which the test functions are the Chebyshev polynomials and the inner product is taken to be the one usually associated with Chebyshev polynomials.

Most spectral methods are classified as *interpolating* or *non-interpolating*. The former corresponds to a collocation method in which the approximation satisfies the PDE at certain points. They are sometimes known as *pseudo-spectral* methods, but be aware that some texts use "pseudo-spectral" in a different sense. The other type correspond to Galerkin-type methods. These latter methods are harder to implement because of the integrals that need to be evaluated.