7 Introduction to PDEs

To those who do not know mathematics, it is difficult to get across a real feeling as to the beauty, the deepest beauty, of nature. If you want to learn about nature, to appreciate nature, it is necessary to understand the language that she speaks in.
—Richard Feynman

7.1 Introduction to PDEs

7.1.1 What are partial differential equations?

Partial differential equations (PDE) are differential equations in which there are two or more independent variables. Typically this means that we are dealing with a mathematical model that is concerned with rates of change in more than one direction (spatially and/or temporally). Consider the following simple example, which you have already seen in Multivariable calculus (I bet you didn't realize you were solving PDEs way back then).

Example 7.1.1.

$$\partial_x \partial_y u = 0.$$

As you already saw in multivariable calculus, this has a solution given by $u(x, y) = f(x) + g(y)$ for some arbitrary C^1 functions $f(x)$ and $g(y)$. This can be checked directly:

$$\partial_{xy} \left(f(x) + g(y) \right) = \partial_x \left(g'(y) \right) = 0.$$

Remark 7.1.2. The Principle of Verification applies to PDEs too. If you think you have found a solution of a PDE you can check that you have by verifying that it satisfies the PDE.

Remark 7.1.3. The solution provided above is called a *general solution* of the PDE. Just as a general solution of an ODE has unknown constants, the general solution of a PDE has unknown functions.

This simple example illustrates some of the fundamental differences between PDEs and ODEs.

(i) Whereas for an ODE we had an undetermined constant (determined by the initial conditions), for PDEs we now have undetermined functions. This is a bit more difficult to deal with, and specify precisely what these undetermined functions are (this is handled in Chapter 9).

(ii) We can no longer simply use the prime mark $\cdot'$ or dot symbol to denote differentiation. Instead we have to be more careful about notation. In particular we must take care when referring to the partial derivatives or total derivatives of a function.

For illustrative purposes and to clarify notation, we consider a scalar valued function of two variables (as in the previous example) $u(x, y)$. We denote partial derivatives using any of the following forms:

$$\frac{\partial u}{\partial x} = \partial_x u = u_x \qquad\qquad \frac{\partial u}{\partial y} = \partial_y u = u_y.$$

For C^2 functions, second-order derivatives are denoted similarly as

$$\frac{\partial^2 u}{\partial x^2} = \partial_{xx} u = u_{xx} \qquad\qquad \frac{\partial^2 u}{\partial y^2} = \partial_{yy} u = u_{yy},$$

and because $u(x, y)$ is C^2 the mixed second-order partial derivatives are the same:

$$\frac{\partial^2 u}{\partial x \partial y} = \partial_{xy} u = u_{xy} = u_{yx}.$$

Higher order derivatives denoted in a similar fashion.

Recall the chain rule for a function of two variables: for a C^1 function $u = u(x, y)$ if $x = x(t)$ and $y = y(t)$, then u is a function of t whose derivative is given by

$$\frac{du}{dt} = Du(x, y)\begin{bmatrix}\dot{x}(t)\\ \dot{y}(t)\end{bmatrix} = \frac{\partial u}{\partial x}\frac{dx}{dt} + \frac{\partial u}{\partial y}\frac{dy}{dt}.$$

When $u = u(x, y)$ is a C^1 function where $x = x(y)$ is C^1, then u is a C^1 function of y, and the ordinary derivative $\frac{du}{dy}$ and the partial derivative $\frac{\partial u}{\partial y}$ are related through the chain rule by

$$\frac{du}{dy} = Du(x, y)\begin{bmatrix}\frac{dx}{dy}\\ \frac{dy}{dy}\end{bmatrix} = \frac{\partial u}{\partial x}\frac{dx}{dy} + \frac{\partial u}{\partial y} = u_x x_y + u_y.$$

A similar relation holds between $\frac{du}{dx}$ and $\frac{\partial u}{\partial x}$ for $u = u(x, y)$ when $y = y(x)$.

As mentioned earlier, there are very few circumstances where solutions to ODEs are known exactly. This situation is even more pronounced for PDEs. Exact solutions are so rare that they are almost never encountered in the wild. Even when exact solutions can be found, their representative formulae are often so complicated as to be nearly useless.

Our primary focus here is to describe techniques that allow the interested mathematician to ask and answer the right questions about the behavior of the PDE without knowing the exact value of the solution. As with ODEs this is often more useful than an explicit solution. This is not to say that explicit solutions are not helpful. We do spend time computing exact solutions to certain types of PDEs, and these methods are essential when they are applicable. But exact solutions are a rarity and are rarely the primary goal.

Example 7.1.4. Consider Example 7.1.1 again. Here we demonstrate how you might obtain the explicit solution that was given there. Recall that we are trying to find a solution to the PDE

$$\partial_{xy} u = 0.$$

Integrating both sides of this equation in x and y, and remembering that any function of x alone is constant with respect to y, we obtain:

$$\begin{aligned} u(x,y) &= \int\int \partial_{xy} u(x,y)\,dy\,dx \\ &= \int \left(\int 0\,dy \right) dx \\ &= \int \varphi(x)\,dx \\ &= f(x) + g(y), \end{aligned}$$

where $f(x)$ is an antiderivative of the arbitrary C^1 function $\varphi(x)$ and $g(y)$ is an arbitrary C^1 function.

Example 7.1.5. Let a and b be constants and consider the PDE

$$au_x + bu_y = 0.$$

For any (differentiable) function $f(z)$ of a single variable, $u(x,y) = f(ay - bx)$ is a solution (more on this in Chapter 8). To see that this is a solution, let $z = ay - bx$, and note that

$$\frac{\partial f}{\partial x} = -bf_z \quad \frac{\partial f}{\partial y} = af_z$$
$$a\frac{\partial f}{\partial x} + b\frac{\partial f}{\partial y} = (-ab + ab)\frac{\partial f}{\partial z} = 0.$$

Remark 7.1.6. In the first part of this book we spent a significant amount of time talking about how to model with an ODE. The same principles apply for PDEs. The process of choosing a model, testing it, breaking it, then improving the model and checking if it is robust, all apply when modeling with PDEs as well. In addition, dimensional analysis plays a vital role in analyzing and simplifying systems with multiple parameters.

One of the primary differences between modeling with PDEs and ODEs is that most ODE systems can be understood decently well in the matter of a few hours or days of rigorous study or careful numerical simulation. This is not always the case for PDEs, and the modeling process can be stuck on the testing-improving steps for a significant amount of time. For this reason we don't emphasize modeling for PDEs as much here, not because it is not as important or relevant, but due to limits of time and scope.

7.1.2 General form of a PDE, linearity, and classification of PDEs in one dimension

As described in the first part of this book, ODE's are typically written as $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, and a linear, homogeneous ODE can be written as $\dot{\mathbf{x}} = A\mathbf{x}$ where A is a matrix (linear operator on $\mathbb{R}^n$). PDEs are typically written as $G(\mathbf{x}, t, u) = 0$, or more frequently as

$$F(\mathbf{x}, t, u) = f(\mathbf{x}, t),$$

where F and G are differential operators rather than just functions.

Example 7.1.7. We present two different PDEs, and then rewrite them in the general format introduced above. Note that there are certainly other ways to consider how these PDEs are written. We don't want to spend to much time on this, but it is important to be familiar with the standard notation.

$$\partial_x \partial_y u = 0 \Rightarrow F(x, y, u) = \partial_x \partial_y u \text{ and } f(x, y) = 0.$$
$$y \cos(x) + u_x \sin(u) = 0 \Rightarrow F(x, y, u) = u_x \sin(u) \text{ and } f(x, y) = -y \cos(x).$$

Remark 7.1.8. As you may have noticed by now, not all notation in mathematics is standardized. This is partially because mathematicians are free-spirited individuals that don't have a governing body that has enacted mathematical Laws and Rules to standardize notation, but it is primarily because most of mathematics (including the study of PDEs) has been developed in a rather organic way, that is constantly evolving and crosses cultural and societal norms. Hence, keep in mind whether this lack of uniformity in notation bothers you when you think of your own political opinion regarding government regulations.

This is all to say, that saying that 'PDEs are typically written as ...' really only applies to the background of one of the authors of this text. In reality there are many other ways in which PDEs are written, but you are being indoctrinated in just one approach.

Remark 7.1.9. In the following you will notice the choice of independent variables (x, y) and (x, t) for the function u in the PDEs. The variables x, y typically denote spatial variables and the variable t typically denotes a temporal (time) variable.

Definition 7.1.10. *For $u = u(x, y)$, the* order *of a PDE $F(x, y, u) = f(x, y)$ is the highest order of the partial derivatives that appears in $F(x, y, u)$, that is, the largest number of derivatives with respect to the independent variables present.*

Example 7.1.11.

- The PDE $u_t + u_{xx} = 0$ has order 2.
- The PDE $u^2 \cos(x) + u_t u_x^3 = 0$ has order 1.
- The PDE $u_{txx} + \cosh(u_{ttxx}) u_x = t^2 \cos(x)$ has order 4.

Linearity

Just as in the theory of ODEs, linear PDEs play an important and special role in the study of their more complicated nonlinear relatives. A linear PDE is usually written as

$$Lu(x, t) = f(x, t),$$

or simply $Lu = f$ where L is a linear differential operator, meaning that $L(u+v) = Lu+Lv$ for any functions u and v, and $L(cu) = cLu$ for any constant c.

Definition 7.1.12.

- *A PDE $Lu = f$ is linear if L is a linear operator.*
- *A linear PDE $Lu = f$ is called* homogenous *if $f = 0$ and* inhomogeneous *is $f \neq 0$.*
- *A PDE $Lu = f$ is* nonlinear *if L is not linear.*

Example 7.1.13. For $u = u(x,t)$ the PDE $u_{xx} + u_t = 0$ has the form $Lu = 0$ where

$$L = \partial_{xx} + \partial_t \qquad \text{and} \qquad Lu = u_{xx} + u_t.$$

For $u = u(x,y)$ the PDE $u_{xx} + u_{yy} = \exp(-x^2 - y^2)$ has the form $Lu = f$ where

$$L = \partial_{xx} + \partial_{yy} \qquad \text{and} \qquad Lu = u_{xx} + u_{yy} \qquad \text{and} \qquad f(x,y) = \exp(-x^2 - y^2).$$

For $u = (x,t)$ the PDE

$$a(x,t)u_{tt} + b(x,t)u_{xt} + c(x,t)u_{xx} + d(x,t)u_t + e(x,t)u_x + g(x,t)u = f(x,t)$$

has the form $Lu = f(x,t)$ where

$$L = a(x,t)\partial_{tt} + b(x,t)\partial_{xt} + c(x,t)\partial_{xx} + d(x,t)\partial_t + e(x,t)\partial_x + g(x,t)$$
$$Lu = a(x,t)u_{tt} + b(x,t)u_{xt} + c(x,t)u_{xx} + d(x,t)u_t + e(x,t)u_x + g(x,t)u.$$

The intuition that we carefully built up for linear ODEs still applies very nicely in the setting of PDEs.

Proposition 7.1.14. *If u_1 and u_2 are solutions of $Lu = 0$, then $c_1u_1 + c_2u_2$ is also a solution of $Lu = 0$ for any constants $c_1, c_2 \in \mathbb{R}$. In addition, if u is a solution of linear homogeneous problem and u_p is a particular solution of the same linear problem but now with an inhomogeneous term f, then $u+u_p$ is also a solution of the inhomogeneous problem $Lu = f$.*

Proof. The proof is a straightforward exercise. □

Remark 7.1.15. Up to this point we have not addressed the issue of the domain of the function $u(x,y)$ (or $u(x,t)$) and coefficient functions that appear in the PDE $F(x,y,u) = f(x,y)$ or $Lu = f(x,y)$. We specify the domain by a set U of $\mathbb{R}^2$ to which (x,y) belongs. The set U is typically open or compact with nonempty interior, and the space of functions to which $u(x,y)$ belongs is $C^k(U)$.

Example 7.1.16. The following PDE's are linear:

$$\cos(x)u_t + (x^3 - 1)u_{xx} = \sin(x) \qquad t > 0,\ x \in \mathbb{R}$$
$$u_{tx} + \tanh(x\cos(t))u_{xxt} - u = 0 \qquad t \in \mathbb{R},\ x \in (0,1).$$

What is the order of each of these examples?

Example 7.1.17. *Burger's equation* is a model commonly encountered in physics and models of traffic flow and is typically written as

$$u_t + uu_x = 0 \qquad t > 0,\ x \in \mathbb{R}.$$

This first-order PDE is nonlinear because the linear combination of two solutions is not always a solution: for solutions u_1 and u_2 we have

$$\partial_t u_1 + u_1 \partial_x u_1 = 0 \text{ and } \partial_t u_2 + u_2 \partial_x u_2 = 0,$$

so that

$$\begin{aligned}
\partial_t(u_1+u_2) + (u_1+u_2)\partial_x(u_1+u_2) &= \partial_t u_1 + \partial_t u_2 + (u_1+u_2)(\partial_x u_1 + \partial_x u_2) \\
&= \partial_t u_1 + \partial_t u_2 + u_1\partial_x u_1 + u_1\partial_x u_2 + u_2\partial_x u_1 + u_2\partial_x u_2 \\
&= \partial_t u_1 + u_1\partial_x u_1 + \partial_t u_2 + u_2\partial_x u_2 + u_1\partial_x u_2 + u_2\partial_x u_1 \\
&= 0 + 0 + +u_1\partial_x u_2 + u_2\partial_x u_1 \\
&= u_1\partial_x u_2 + u_2\partial_x u_1
\end{aligned}$$

which is not guaranteed to be zero.

Remark 7.1.18. As we show later, linear homogeneous PDEs are a lot like infinite-dimensional linear homogeneous ODEs. In this sense the linear operator L can be viewed as an infinite-dimensional matrix. This carries over directly to numerical approximations of PDEs wherein L is frequently approximated by a large (yet finite-dimensional) matrix whose exact formulation depends on the choice of numerical method.

7.1.3 Classification of second-order PDEs in dimension one

Consider the most general second-order linear PDE in one spatial dimension:

$$au_{tt} + bu_{xt} + cu_{xx} + du_t + eu_x + gu = f(x,t), \qquad (x,t) \in \Omega, \tag{7.1}$$

where $a, b, c, d, e, f, g \in C^1(\Omega)$.

- This is a linear system if none of the coefficients depends on u.
- If any of the coefficients do depend on u, but not on any derivatives of u then (7.1) is called *quasilinear*.

Such PDEs are classified in a manner analogous to real conic sections (parabola, hyperbola, and ellipse).

Definition 7.1.19. *Define the* discriminant *of* (7.1) *as*

$$D = b(x,t)^2 - 4a(x,t)c(x,t),$$

and classify (7.1) *in the following way:*

- *On the set of vales* (x,t) *where* $D > 0$ *then* (7.1) *is* hyperbolic *(wave-like).*
- *On the set of values* (x,t) *where* $D = 0$ *then* (7.1) *is* parabolic *(diffusive).*
- *On the set of values* (x,t) *where* $D < 0$*, then* (7.1) *is* elliptic *(equilibrium).*

There are versions of this type of classification for higher dimensions and even somewhat for higher-order PDEs, explained in more detail in later chapters. The reason for this classification is that typically the techniques developed for one class of PDEs do not apply to a different class of PDEs, although there are exceptions. We focus on the hyperbolic and parabolic equations, and only briefly discuss elliptic equations, which can often be seen as the steady state solution of parabolic equations.

Example 7.1.20. Some specific examples of each of these types of equations are:

- The heat equation in one dimension is $u_t = ku_{xx}$ and is parabolic. Some solutions of the heat equation are

$$u(x,t) = \sin(n\pi x)\exp(-kn^2\pi^2 t), \quad \text{for } n \in \mathbb{N},$$

because

$$u_t = -kn^2\pi^2 \sin(t\pi x)\exp(-kn^2\pi^2 t) \text{ and } u_{xx} = -n^2\pi^2 \sin(n\pi x)\exp(-kn^2\pi^2 t)$$

so that

$$u_t = ku_{xx} \text{ for all } \quad \text{for } n \in \mathbb{N}.$$

Any finite collection of these solutions is linearly independent, and a formal infinite series solution of the heat equation is

$$\sum_{n=1}^{\infty} a_n \sin(n\pi x)\exp(-kn^2\pi^2 t), \ a_n \in \mathbb{R}.$$

- The wave equation in one-dimension is $u_{tt} - c^2u_{xx} = 0$ and is hyperbolic. Some known solutions of this equation are

$$u(x,t) = \sin(n\pi x)\cos(cn\pi t), \quad \text{for } n \in \mathbb{N},$$

because

$$u_{tt} = -c^2n^2\pi^2 \sin(n\pi x)\cos(cn\pi t) \text{ and } u_{xx} = -n^2\pi^2 \sin(n\pi x)\cos(cn\pi t),$$

so that

$$u_{tt} - c^2u_{xx} = 0 \text{ for all } n \in \mathbb{N}.$$

Any finite collection of these solutions is linearly independent, and a formal infinite series solution of the wave equation is

$$\sum_{n=1}^{\infty} a_n \sin(n\pi x)\cos(cn\pi t), \ a_n \in \mathbb{R}.$$

- Laplace's equation is $u_{xx} + u_{yy} = 0$ and is elliptic. Some solutions of this equation are

$$u(x,y) = \sinh(n\pi x)\sin(n\pi y), \ n \in \mathbb{N},$$

because by verification

$$u_{xx} = n^2\pi^2 \sinh(n\pi x)\sin(n\pi y) \text{ and } u_{yy} = -n^2\pi^2 \sinh(n\pi x)\sin(n\pi t),$$

so that

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

Any finite collection of these solutions is linearly independent, and a formal infinite series solution of Laplace's equation is

$$\sum_{n=1}^{\infty} a_n \sinh(n\pi x)\sin(n\pi y), \ a_n \in \mathbb{R}.$$

Remark 7.1.21. In these examples of linear homogeneous PDEs $Lu = 0$ the dimension of the vector space of solutions is infinite dimensional.

7.2 Finite Differences, A Basic Introduction to Numerical Methods for PDEs

We noted previously that there is no hope in this text to have a comprehensive discussion of the numerical approximation of solutions of ODEs because there are so many different approaches. This is even more the case for PDEs where the problem of interest is so much more complicated. This section will give a very brief review of *some* numerical methods for a select few PDEs. Throughout the rest of this part of the text, additional sections will be directed at the numerical approximation of PDEs as the need arises. This scattershot approach is because some types of numerical methods are appropriate for one type of PDE, while a different type of method is more appropriate for a different type of PDE.

We start off by trying to approximate the solution of a relatively simple PDE, namely the heat equation in one spatial dimension.

$$u_t = u_{xx}. \tag{7.2}$$

In a later chapter we discuss what type of auxiliary conditions are necessary to make this a well posed problem (and even discuss what well-posedness means in this context), but for now we will forge onward as if the initial/boundary conditions will be supplied to us later.

7.2.1 Discretization

One of the first things we can do is very similar to how we derived the forward and backward Euler methods for ODEs. We consider what is called a finite difference approximation to the derivatives specified in (7.2). First, we approximate the time derivative using the forward temporal difference approximation of the partial derivative

$$u_t(x,t) \approx \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t}, \tag{7.3}$$

and similarly we can approximate the spatial derivatives as

$$u_x(x,t) \approx \frac{u(x+\Delta x,t) - u(x,t)}{\Delta x}, \tag{7.4}$$

or

$$u_x(x,t) \approx \frac{u(x,t) - u(x-\Delta x,t)}{\Delta x}. \tag{7.5}$$

While specifics are often debated between textbooks and different authors, generically (7.4) is referred to as a *forward finite difference approximation* to the derivative, and (7.5) is a *backward finite difference approximation* to the derivative. Yet another option is to use a centered finite difference which is really an average of (7.4) and (7.5):

$$u_x(x,t) \approx \frac{u(x+\Delta x,t) - u(x-\Delta x,t)}{2\Delta x}. \tag{7.6}$$

```python
import numpy as np

def forward_one(u,delX):
    #Assumes that the vector u is periodic on the interval
    #This function estimates the first derivative via
    #a first order forward finite difference
    u0 = u
    u1 = np.roll(u,-1)
    ux = (u1-u0)/delX
    return ux
```

Algorithm 7.1: An example Python implementation for calculating the first derivative of a given function, defined as a vector.

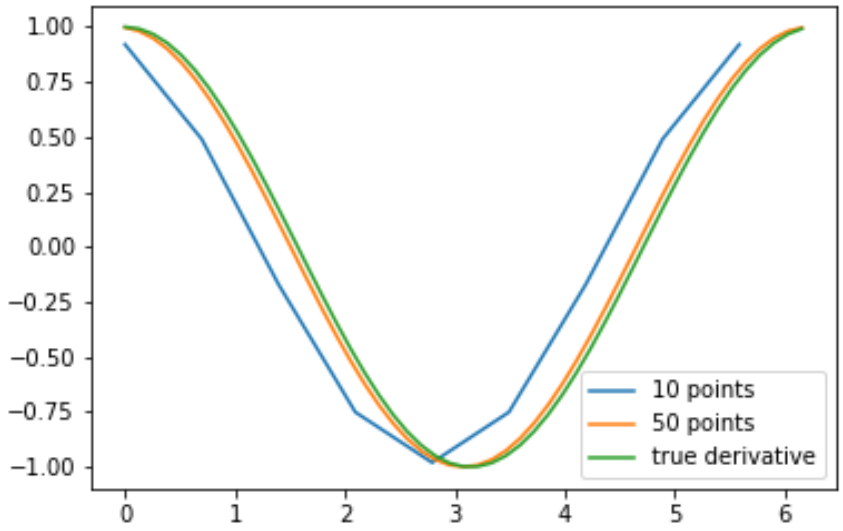


Figure 7.1: Approximations using the first-order forward differencing to the first derivative of $u(x) = \sin(x)$ following Algorithm 7.1, with either 10 or 50 grid points in the discretization. Note how quickly the approximation converges to the true value of the derivative ($u'(x) = \cos(x)$ in this case).

As an example implementation of the first-order, forward finite difference, consider Algorithm 7.1. To avoid worrying about boundaries, we have assumed that the domain in question is periodic, and that the grid is uniformly spaced with grid spacing Δx. This algorithm is used to estimate the derivative of the function $u(x) = \sin(x)$ for two different values of Δx which produce either 10 or 50 different grid points along the interval $[0, 2\pi)$ and the resultant estimated derivative is plotted in Figure 7.1. Note that the low resolution result is not only inaccurate, but appears to have some bias due to the forward differencing. When Δx is decreased sufficiently this bias is less noticeable although still present. Using a centered differencing approach would introduce a different type of error, but would do away with the bias seen here.

We can of course go through the same process for higher order derivatives. For example, we can combine the forward and backward spatial differences to come up with a finite difference approximation for the second derivative:

$$u_{xx}(x,t) \approx \frac{u_x(x+\Delta x,t) - u_x(x,t)}{\Delta x} \approx \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{(\Delta x)^2}, \qquad (7.7)$$

where we used a backward difference to approximate $u_x(x+\Delta x,t)$ and a forward and backward difference to approximate $u_x(x,t)$. This is of course not the only way to approximate $u_{xx}(x,t)$, but it is a reasonable one.

7.2.2 Truncation error and derivation of additional methods

We define the truncation error just as we did for ODEs.

Definition 7.2.1. *Let $\tilde{u}(x,t)$ be the finite difference approximation to some function $u(x,t)$. We define the truncation error $\tau(x,t)$ of $\tilde{u}(x,t)$ to be the difference between the approximate derivative that we seek, and the true value of the derivative of the function.*

Example 7.2.2. We find the truncation error for (7.4) and (7.7). You analyze the error for (7.5) and (7.6) in the exercises.

For (7.4) we note that the truncation error is defined as:

$$\begin{aligned}
\tau(x,t) &= u_x(x,t) - \frac{u(x+\Delta x,t) - u(x,t)}{\Delta x} \\
&= u_x(x,t) - \frac{1}{\Delta x}\left[u(x,t) + (\Delta x)u_x(x,t) + \frac{(\Delta x)^2}{2}u_{xx}(x,t) + O((\Delta x)^3) - u(x,t)\right] \\
&= u_x(x,t) - \frac{1}{\Delta x}\left[(\Delta x)u_x(x,t) + \frac{(\Delta x)^2}{2}u_{xx}(x,t) + O((\Delta x)^3)\right] \\
&= u_x(x,t) - u_x(x,t) - \frac{(\Delta x)}{2}u_{xx}(x,t) + O((\Delta x)^2) \\
&= -\frac{(\Delta x)}{2}u_{xx}(x,t) + O((\Delta x)^2).
\end{aligned}$$

Hence, the forward finite difference method is first order in Δx.

For (7.7) we proceed in the same fashion, but to save some space we first note that

$$u(x\pm\Delta x,t) \approx u(x,t) \pm \Delta x u_x(x,t) + \frac{(\Delta x)^2}{2}u_{xx}(x,t) \pm \frac{(\Delta x)^3}{6}u_{xxx}(x,t) + \frac{(\Delta x)^4}{24}u_{xxxx}(x,t).$$

Then the truncation error for the spatial finite difference for $u_{xx}(x,t)$ is

$$\begin{aligned}
\tau(x,t) &= u_{xx}(x,t) - \frac{1}{(\Delta x)^2}\left[u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)\right] \\
&= u_{xx}(x,t) - \frac{1}{(\Delta x)^2}\left[u(x,t) + (\Delta x)u_x(x,t) + \frac{(\Delta x)^2}{2}u_{xx}(x,t) + \frac{(\Delta x)^3}{3!}u_{xxx}(x,t)\right. \\
&\qquad\qquad + \frac{(\Delta x)^4}{4!}u_{xxxx}(x,t) - 2u(x,t) \\
&\qquad\qquad + u(x,t) - (\Delta x)u_x(x,t) + \frac{(\Delta x)^2}{2}u_{xx}(x,t) - \frac{(\Delta x)^3}{3!}u_{xxx}(x,t) \\
&\qquad\qquad \left. + \frac{(\Delta x)^4}{4!}u_{xxxx}(x,t) + O((\Delta x)^5)\right] \\
&= u_{xx}(x,t) - \frac{1}{(\Delta x)^2}\left[(\Delta x)^2 u_{xx}(x,t) + \frac{2(\Delta x)^4}{4!}u_{xxxx}(x,t) + O((\Delta x)^5)\right] \\
&= u_{xx}(x,t) - u_{xx}(x,t) - \frac{(\Delta x)^2}{12}u_{xxxx}(x,t) + O((\Delta x)^3) \\
&= -\frac{(\Delta x)^2}{12}u_{xxxx}(x,t) + O((\Delta x)^3).
\end{aligned}$$

Hence the centered finite difference approximation to the second derivative is actually second order.

Remark 7.2.3. This shouldn't be very surprising, and is actually almost the exact same calculation that we did when looking at time-stepping methods for ODEs. The only difference is that now we are dealing with potential error in both time t and space x. In fact we have only been dealing with the simplest case of one spatial dimension x when in reality we live in a three-dimensional world and hence should be concerned with PDEs (and hence approximations) in three dimensions as well. Hence we really are interested in errors that occur in all of the independent variables, and this can be problematic and complicated to compute.

With this definition of the truncation error in mind, we can discuss the general derivation of higher order methods (other than those that just pop up from the definition of the derivative). For finite differences, this can be done in a variety of different ways. We focus on a single approach that is meant to illustrate the general concept. We also restrict our attention to first-order derivatives, although the same approach easily generalizes to higher order derivatives.

We begin by supposing that we can approximate $u_x(x,t)$ by some linear combination of nearby points. For instance, we will consider the approximation

$$u_x(x,t) \approx \frac{1}{\Delta x}\sum_{k=-n}^{n} a_k u(x+k\Delta x,t), \tag{7.8}$$

for some fixed choice of n, and then choose the a_k to give us the desired order of accuracy. Because we are consider k from $-n$ to $+n$ then we refer to this as a centered finite difference scheme.

Example 7.2.4. The centered spatial finite difference

$$u_x(x,t) \approx \frac{u(x+\Delta x,t) - u(x-\Delta x,t)}{2\Delta x}$$

has this form where $n = 1$ and $a_{-1} = \frac{-1}{2}$, $a_0 = 0$ and $a_1 = \frac{1}{2}$.

A centered finite difference for $u_{xx}(x,t)$ is an expression of the form

$$u_{xx}(x,t) \approx \frac{1}{(\Delta x)^2} \sum_{k=-m}^{m} b_k u(x + k\Delta x, t)$$

for a fixed $m \in \mathbb{N}$ and constants b_k chosen to achieve a desired order of accuracy.

The centered spatial finite difference

$$u_{xx}(x,t) \approx \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{(\Delta x)^2}$$

has this form where $m = 1$ and $b_{-1} = 1$, $b_0 = -2$, and $b_1 = 1$.

Remark 7.2.5. Note that if we set $a_k = 0$ for $k < 0$ or for $k > 0$ then we have what is called a one-sided finite difference formula that gives us an approximation that depends only on information from the right or left of the current state. Sometimes this is desirable or necessary depending on the type of PDE we are interested in solving, and what auxiliary information is provided with the PDE.

The actual derivation of the finite difference methods is rather straightforward after what we have already done for time-stepping methods for ODEs, so the details are omitted here or left to the exercises. We do note that while this derivation is not overly complex, it is also not overly simple and is prone to algebraic errors and mistakes. This is often where symbolic computational packages become invaluable.

7.2.3 Returning to the PDE

Now we return to the original problem proposed in this section: we want to numerically approximate the solution to (7.2). The standard way to do this is to apply centered finite differences to the spatial derivatives, and then either a forward or backward one sided difference to the time derivative. This leads to the algebraic expression for the approximate solution:

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{(\Delta x)^2}, \tag{7.9}$$

which upon rearranging terms gives us the update formula:

$$u(x,t+\Delta t) = u(x,t) + \frac{\Delta t}{(\Delta x)^2}\left(u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)\right). \tag{7.10}$$

Example 7.2.6. To see how this update formula is actually used, suppose we know that

$$u(0,t) = T_1 \text{ for all } t \geq 0,\ u(1,t) = T_2 \text{ for all } t \geq 0 \text{ and } u(x,0) = f(x),\ 0 \leq x \leq 1.$$

where $f(0) = T_1$ and $f(1) = T_2$.

The first two are called boundary conditions and the last is called an initial condition.

Take $\Delta x = \frac{1}{4}$ and $\Delta t = 0.1$.

By the update formula

$$u(\frac{1}{4},0.1) \approx u(\frac{1}{4},0) + (0.1)\frac{u(\frac{1}{4}+\frac{1}{4},0) - 2u(\frac{1}{4},0) + u(\frac{1}{4}-\frac{1}{4},0)}{(\frac{1}{4})^2}$$
$$= f(\frac{1}{4}) + (0.1)\frac{f(\frac{1}{2}) - 2f(\frac{1}{4}) + T_1}{(\frac{1}{4})^2}.$$

What is $u(\frac{1}{2},0.1)$ and $u(\frac{3}{4},0.1)$?

[Draw the triangle diagram with four dots (three on the bottom and one above the middle dot on the bottom) of how the update formula uses values of u on the boundaries $x = 0$ and $x = 1$ and along the initial condition when $t = 0$ to compute the values of u at the points $(\frac{1}{4},0.1)$, $(\frac{1}{2},0.1)$, and $(\frac{3}{4},0.1)$.]

Now that we have these three values, how do we compute the approximations of u at the time $t = 0.2$, i.e., $u(\frac{1}{4},0.2)$, $u(\frac{1}{2},0.2)$, and $u(\frac{3}{4},0.2)$?

Remark 7.2.7. Before we run off and declare victory that we have found how to solve (7.2) numerically, we must reconsider what this actually means.

- From our work with solving ODEs numerically, we can see that this scheme is explicit, i.e. we only require knowledge of the current state (in time) to approximate the next state in the future.
- Updates on the spatial components of the solution are different though. Identifying the value of $u(x, t + \Delta t)$ requires knowledge of $u(x - \Delta x, t)$ and $u(x + \Delta x, t)$ as well as $u(x, t)$.
- Hence in order to approximate $u(x, t)$ at the next time, we need to also know what it's neighboring values (in space) are at the current time.
- This is related to what we will refer to as auxiliary conditions for PDEs, which are really an extension of the concept of an initial condition for an ODE. Although this is certainly a topic of immense importance and interest, we set it aside for now and will return to it later.

Another way of viewing this is to recast (7.2) as an ODE of the form $\dot{u} = f(u)$. with a messy right hand side $f(u) = u_{xx}$. Then reconsider the numerical approximation as having two steps:

(i) First use centered differences to approximate $f(u) = u_{xx}$.

(ii) Use forward Euler to approximate $u_t = f(u)$ where $f(u)$ is replaced by its finite difference approximation from the last step.

This illustrates that we actually have a lot of choices when it comes to approximating PDEs, particularly ones of this form. For instance, we could have used a higher order finite differencing for approximating u_{xx} or we could use a more sophisticated time-stepping method to approximate $u_t = f(u)$. Many of the same issues that we discussed previously for time-stepping methods arise, and particular issues for each PDE and its physical setting must be taken into account when selecting the numerical scheme.

Remark 7.2.8. One issue of utmost importance when dealing with this is considering the stability of the corresponding numerical scheme. Recall that the absolute stability region of the forward Euler scheme is given by the unit circle in the complex plane centered at $z = -1$. If we consider that our choice of approximation for u_{xx} is very similar to

$$f(u) = u_{xx} \approx \frac{1}{(\Delta x)^2}u(x,t), \tag{7.11}$$

where we have clearly neglected the fact that the differencing actually plays a role here, then we would expect this approach to be numerically stable so long as $\frac{\Delta t}{(\Delta x)^2} < 2$; that is, we have $z = \lambda \Delta t = \frac{\Delta t}{(\Delta x)^2}$. If we are trying to maintain a very fine mesh in x, this places an incredibly strict restriction on Δt as we must increase the temporal resolution twice as quickly as the spatial resolution.

This is why diffusive problems like this PDE are best treated via an implicit time-stepping method which can safely avoid most of these stability issues.

Some of you while reading this may be wondering why we don't go ahead and code up a solution to this PDE now (or at least ask you to do so in the exercises). This is because we have not properly set up the problem with initial and boundary conditions yet so we don't know how to initialize the solution $u(x,t)$ at previous time steps and/or spatial locations. Never fear though, we will get there soon enough and you will be able to simulate the heat equation.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

7.1. State whether the following PDEs are linear or nonlinear, and if linear, whether they are homogeneous or inhomogeneous. Also state the order of the PDE.

- $u_x(1+u_x^2)^{-\frac{1}{2}} + u_y(1+u_y)^{-\frac{1}{2}} = 0$.
- $u_{tt} - u_{xx} + x^2 = 0$.
- $u_t + u_{xxxx} + \sqrt{1+u} = 0$.
- $u_t - u_{xxt} + uu_x = 0$.

7.2. *Find the solution to the following in terms of arbitrary functions in x and y

$$u_{xx} + u = 3y \text{ where } u = u(x,y).$$

7.3. *Find the solution to the following in terms of arbitrary functions

$$tu_{xx} - 2u_x = 0 \text{ where } u = u(x,t).$$

7.4. *Find the solution to the following in terms of arbitrary functions

$$u_{xt} + \frac{1}{x}u_t = \frac{t}{x^2} \text{ where } u = u(x,t).$$

7.5. Find the general solution to the first-order linear equation

$$u_t + cu_x = 0$$

by invoking the change of variables $z = x - ct$.

7.6. Find all solutions of the heat equation $u_t = ku_{xx}$ that satisfy the ansatz $u(x,t) = f(z)$ where $z = \frac{x}{\sqrt{kt}}$. Hint: You may find it useful, at some point, to use the function $\text{erf}(y) = \frac{2}{\sqrt{pi}} \int_0^y e^{-s^2}\, ds$.

7.7. Are the following hyperbolic, parabolic, or elliptic?

(i) $u_{xx} - u_{xy} + 2u_y + u_{yy} - 3u_{yx} + 4u = 0$

(ii) $9u_{xx} + 6u_{xy} + u_{yy} + u_x = 0$.

7.8. Find regions in the xy-plane where the equation

$$(1+x)u_{xx} + 2xyu_{xy} - y^2 u_{yy} = 0$$

is elliptic, hyperbolic, or parabolic. Sketch these regions.

7.9. Show that (7.5) is first order.

7.10. Show that (7.6) is second order.

7.11. Derive a third-order one-sided finite difference approximation to $u_x(t,x)$. Make sure it is clear why your scheme is third order. How many terms did you have to retain in the approximation?

7.12. Use the approximation you derived in the previous problem to approximate the derivative of $\sin(x)^2$ on the interval $[0,\pi]$ for $\Delta x = 0.1,\ 0.01$. Compare against the true value of the derivative for this function. How do you handle the one-sided aspect of the differencing at the endpoints of the interval?

7.13. Is it possible to have a third-order centered finite difference approximation to $u_x(t,x)$? Why or why not? If it is possible, derive the coefficients for the scheme and write it out explicitly. If not, give an explanation why not.

Notes

8 Hyperbolic PDE

Today's scientists have substituted mathematics for experiments, and they wander off through equation after equation, and eventually build a structure which has no relation to reality.
—Nikola Tesla

8.1 Conservation Laws in dimension one

This section sketches the derivation of certain PDEs typically referred to as *conservation laws*. Conservation laws show up in a variety of physical circumstances. Typically a conservation law follows from one of the fundamental axioms of physics, i.e. Newton's first and second laws are often the basis for several PDEs. We proceed by the derivation of a simple example in a single dimension. After discussing how this example can be extended and generalized, in the next Section we briefly review of some concepts and notation from multivariable calculus, and then consider the multidimensional form of a generic conservation law.

Although this is a chapter on hyperbolic PDEs, many conservation laws, including several of those discussed below, do not result in hyperbolic PDEs. Nevertheless the bulk of all hyperbolic PDEs are derived via a conservation law of some sort. Hence, we'll discuss conservation laws in this chapter rather than later. Besides conservation laws are fun to work with.

8.1.1 Traffic flow in one dimension

Let $u = u(x,t)$ be the density of some quantity (for this example we will refer to the number of cars along a section of a one-way road with no entrances or exits). The total number of cars along the road from point x_1 to x_2 is given by

$$N(t) = \int_{x_1}^{x_2} u(x,t)dx.$$

Since there are no exits or entrances on this interval of the road, then this number of cars $N(t)$ is *conserved*, meaning that its rate of change (*time derivative*) depends only on the movement of cars into or out of the region at the two end points.

- The density $u(x,t)$ is an number bounded below by 0 for all (x,t) and it has dimesions $[u] = \mathsf{N}\mathsf{L}^{-1}$, where N is the dimension of number of cars.
- This situation also describes the number of red blood cells moving in an artery. Can you think of other situations that the function $u(x,t)$ describes?
- The objective of this section is to find an equation that the function $u(x,t)$ satisfies.

The number of vehicles on the road from the point x_1 of the road to the point x_2 of the road at the time t is the quantity

$$N(t) = \int_{x_1}^{x_2} u(x,t)\ dx.$$

We usually think of $x_1 < x_2$ so the $N(t)$ is bounded below by 0. Since dx has dimensions $[dx] = \mathsf{L}$ and the integral, like a sum, does not change dimension, the dimension of N is $[N] = \mathsf{N}\mathsf{L}^{-1}\mathsf{L} = \mathsf{N}$, as expected.

The quantity N changes in time according to

$$\begin{aligned}\frac{dN}{dt} &= \text{rate in at } x_1 - \text{ rate out at } x_2\\ &= u(x_1,t)V(x_1,t) - u(x_2,t)V(x_2,t).\end{aligned}$$

Let's check the units of these quantities to be sure these make sense:

$$\begin{aligned}\frac{dN}{dt} &= \frac{\text{cars}}{\text{time}},\\ u(x_1,t)V(x_1,t) &= \frac{\text{cars}}{\text{length}}\cdot\frac{\text{length}}{\text{time}} = \frac{\text{cars}}{\text{time}},\\ u(x_2,t)V(x_2,t) &= \frac{\text{cars}}{\text{length}}\cdot\frac{\text{length}}{\text{time}} = \frac{\text{cars}}{\text{time}}.\end{aligned}$$

We denote the quantity

$$J(x,t) = u(x,t)V(x,t)$$

as the flux of cars at position x at time t, that is, the cars per time entering or leaving point x at time t. Its units are $[J] = \mathsf{N}\mathsf{T}^{-1}$.

The Fundamental Theorem of Calculus gives

$$\int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(J(x,t)\big)\ dx = J(x_2,t) - J(x_1,t).$$

This implies that

$$\begin{aligned}u(x_1,t)V(x_1,t) - u(x_2,t)V(x_2,t) &= J(x_1,t) - J(x_2,t)\\ &= -\int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(J(x,t)\big)\ dx.\end{aligned}$$

Thus we obtain

$$\begin{aligned}\frac{dN}{dt} &= u(x_1,t)V(x_1,t) - u(x_2,t)V(x_2,t)\\ &= -\int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(J(x,t)\big)\ dx\\ &= -\int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big)\ dx.\end{aligned}$$

Since

$$N(t) = \int_{x_1}^{x_2} u(x,t)\ dx$$

we obtain the equation

$$\frac{d}{dt}\int_{x_1}^{x_2} u(x,t)\ dx = -\int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big)\ dx.$$

Assuming the partial derivative with respect to t and integration with respect to x can be interchanged (and they can if u is C^1), then we have

$$\int_{x_1}^{x_2} \frac{\partial}{\partial t}\big(u(x,t)\big)\ dx = -\int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big)\ dx.$$

Moving the two integrals to the left-hand side and combining them gives

$$\int_{x_1}^{x_2} \left\{ \frac{\partial}{\partial t}\big(u(x,t)\big) + \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big) \right\}\ dx = 0.$$

We have arrived at the integral form of a conservation law that the function $u(x,t)$ satisfies.

- Note that there was nothing special about our choice of the interval $[x_1, x_2]$ so this integral must be zero for any choices of x_1 and x_2.
- If the integrand is continuous and nonzero at some point $\widetilde{x}$ we can then choose an interval $[x_1, x_2]$ containing $\widetilde{x}$ on which the integrand is bounded away from 0 and hence the integral is not zero on $[x_1, x_2]$.
- This contradiction implies that the integrand must be identically equal to zero, giving the first-order PDE

$$\frac{\partial}{\partial t}\big(u(x,t)\big) + \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big) = 0.$$

- Rewriting this we have the differential form of the conservation law

$$u_t(x,t) + \partial_x\big(u(x,t)V(x,t)\big) = 0.$$

This gives a partial differential equation that the function $u(x,t)$ satisfies.

Now suppose that there are exits and/or entrances in this region represented by the local sink/source function $f(x,t)$ with units of cars/(length $\times$ time). Then it follows that

$$\frac{dN}{dt} = -\int_{x_1}^{x_2} \partial_x \left(u(x,t)V(x,t)\right) dx + \int_{x_1}^{x_2} f(x,t)dx$$

Then using

$$N(t) = \int_{x_1}^{x_2} u(x,t)\ dx$$

and interchanging the derivative and the partial integral, we obtain

$$\int_{x_1}^{x_2} \frac{\partial}{\partial t}\big(u(x,t)\big)\ dx + \int_{x_1}^{x_2} \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big)\ dx - \int_{x_1}^{x_2} f(x,t)\ dx = 0.$$

Combining the integrals gives

$$\int_{x_1}^{x_2} \left\{ \frac{\partial}{\partial t}\big(u(x,t)\big) + \frac{\partial}{\partial x}\big(u(x,t)V(x,t)\big) - f(x,t) \right\}\ dx = 0.$$

This gives an integral form of a conservation law in that the integral is always zero no matter the values of x_1 and x_2.

Arguing as before, the integrand must be zero, so that we arrive at the differential form of the conservation law

$$u_t(x,t) + \partial_x\big(u(x,t)V(x,t)\big) = f(x,t).$$

By suppressing the independent variables this has the form

$$u_t + \partial_x(uV) = f.$$

We have a first-order PDE that the function $u(x,t)$ satisfies.

Remark 8.1.1. The integral form of the conservation law is always valid, and can be used to motivate the numerical methods used to solve it (this is often done for what are called *finite volume* and *finite element* methods).

To finish this derivation, we need to have a functional relationship between $V(x,t)$ and $u(x,t)$. This is often a modeling choice, or in the case of physics, dictated by fundamental observations or classical laws of motion and mechanics. For the traffic flow problem, a reasonable velocity is

$$V(x,t) = V_\infty\left(1 - \frac{u(x,t)}{u_\infty}\right),$$

where V_∞ is the maximal velocity (a reasonable assumption may be 10 miles an hour over the speed limit) and $u = u_\infty$ is a maximal traffic density that indicates a traffic jam. With this choice of $V(x,t)$ we have

$$\partial_x(uV) = u_xV + uV_x = u_xV_\infty\left(1 - \frac{u}{u_\infty}\right) + uV_\infty\left(-\frac{u_x}{u_\infty}\right) = V_\infty\left(1 - \frac{2u}{u_\infty}\right)u_x.$$

We then arrive at the model

$$u_t + V_\infty\left(1 - \frac{2u}{u_\infty}\right)u_x = f.$$

With our choice of V we have introduced two parameters u_∞ and V_∞, and we should and do ask ourselves about the number of parameters this model depends on.

To find out we non-dimensionalize the model by setting

$$\widetilde{u} = \alpha u,\ \widetilde{x} = \beta x,\ \widetilde{t} = \gamma t.$$

The units of the parameters are

$$\alpha = \text{length per car},\ \beta = \text{ per length },\gamma = \text{ per time}.$$

With x held fixed we have

$$\frac{\partial\widetilde{u}}{\partial\widetilde{t}} = \frac{\partial\widetilde{u}}{\partial t}\frac{dt}{d\widetilde{t}} = \frac{\partial}{\partial t}[\alpha u]\frac{d}{d\widetilde{t}}\left[\frac{\widetilde{t}}{\gamma}\right] = \frac{\alpha}{\gamma}\frac{\partial u}{\partial t} \ \Rightarrow \frac{\partial u}{\partial t} = \frac{\gamma}{\alpha}\frac{\partial\widetilde{u}}{\partial\widetilde{t}},$$

and with t held fixed we have

$$\frac{\partial\widetilde{u}}{\partial\widetilde{x}} = \frac{\partial\widetilde{u}}{\partial x}\frac{dx}{d\widetilde{x}} = \frac{\partial}{\partial x}[\alpha u]\frac{d}{d\widetilde{x}}\left[\frac{\widetilde{x}}{\beta}\right] = \frac{\alpha}{\beta}\frac{\partial u}{\partial x} \ \Rightarrow \ \frac{\partial u}{\partial x} = \frac{\beta}{\alpha}\frac{\partial\widetilde{u}}{\partial\widetilde{x}}.$$

With these the PDE becomes

$$\frac{\gamma}{\alpha}\widetilde{u}_{\widetilde{t}} + V_\infty\left(1 - \frac{2\widetilde{u}}{\alpha u_\infty}\right)\frac{\beta}{\alpha}\widetilde{u}_{\widetilde{x}} = f(\widetilde{x}/\beta, \widetilde{t}/\gamma).$$

Multiplying through by $\frac{\alpha}{\gamma}$ gives

$$\widetilde{u}_{\widetilde{t}} + \frac{\beta}{\gamma} V_\infty \left(1 - \frac{2\widetilde{u}}{\alpha u_\infty}\right) \widetilde{u}_{\widetilde{x}} = \frac{\alpha}{\gamma} f(\widetilde{x}/\beta, \widetilde{t}/\gamma).$$

Choose

$$\alpha = \frac{1}{u_\infty} \text{ and } \frac{\beta}{\gamma} = \frac{1}{V_\infty}.$$

Recognize that the units for α are per length and the units for $\frac{\beta}{\gamma}$ are time per length.

Numerically we can assign $\beta = 1$ and $\gamma = V_\infty$ (where we are ignoring the units on V_∞ in assigning γ its value).

Then the non-dimensionalized PDE is

$$\widetilde{u}_{\widetilde{t}} + (1 - 2\widetilde{u})\widetilde{u}_{\widetilde{x}} = \frac{1}{u_\infty V_\infty} f(\widetilde{x}, \widetilde{t}/V_\infty).$$

Dropping the $\widetilde{}$ notation from the variables and reassigning the arbitrary right-hand side of the PDE to be $f(x,t)$ again (since it was arbitrary to begin with), we have

$$u_t + (1 - 2u)u_x = f.$$

The first-order PDE for the function $u(x,t)$ doesn't explicit depend on the two parameters V_∞ and u_∞, as these have been conveniently absorbed in the source/sink function f.

8.1.2 one-dimensional generic conservation laws

The traffic flow problem is just one example of a PDE that arises naturally as the result of the conservation of some key quantity. Conservation laws are typically written as

$$\begin{array}{l}\text{rate of change of}\\ \text{the quantity in some region}\end{array} = \begin{array}{l}\text{net rate of quantity}\\ \text{flowing into/out of the region}\end{array} + \text{influence of sources/sinks.}$$

In one dimension this can be written in integral form as

$$\frac{d}{dt}\int_a^b u(x,t)dx = J(a,t) - J(b,t) + \int_a^b f(x,t)dx, \tag{8.1}$$

or in PDE form as:

$$u_t + J_x = f, \tag{8.2}$$

assuming that all the derivatives exist and are nice and continuous (when might this break down?). The flux $J(x,t)$ is often not explicitly computable and so some necessary assumption or modeling must be made in order to proceed. For some physical systems this is not the case, and the flux is deterministically available, but this is not as likely for non-physics based models and situations.

Remark 8.1.2. There are two types of flux that appear in a large variety of applications.

(i) We have introduced the *advective*[28] flux in the derivation for traffic flow above, and this is one of the most common, i.e. $J(x,t) = u(x,t)v(x,t)$ for some velocity field $v(x,t)$ where the density of interest is $u(x,t)$.

(ii) The other most frequently encountered flux is a diffusive flux wherein $J(x,t) = -\kappa u_x(x,t)$. This is used to model the diffusion of the concentration from areas of high concentration to low concentration. This is the type of behavior that temperature in a rod will follow, and also accurately represents the diffusion of a dye in a fluid. Note that the constant must have dimension $[\kappa] = \mathsf{L}^2\mathsf{T}^{-1}$ in order for flux to have the correct dimensions of $[J] = \mathsf{N}\mathsf{T}^{-1}$.

[28] *Advection* is the transport of a quantity by the movement of a fluid. In the traffic example we treat the traffic like a fluid, so we still think of the flux in that example as advective.

These two fluxes appear in many different physical settings and can even be used in models of social interactions, finance, and many other applications. In cases where a quantity is diffuses and is carried by a fluid flow (for example dye in flowing water), the flux is the sum of both an advection term and a diffusion term. There are of course many other fluxes that are of interest, but these are so common and important that they're worth committing to memory.

Remark 8.1.3. Now of course the world is not one-dimensional, and we need to discuss how to handle/derive a conservation law in multiple dimensions. Before we can formulate such an approach in the next section, we will very briefly review some key results from multivariable calculus. The primary need for this review is to establish notation, but we also cover some standard results that are needed when converting a boundary flux term to a bulk integrated term.

8.2 Multi-dimensional conservation laws

One-dimensional conservation laws derived in the last section involved using the Fundamental Theorem of Calculus. In higher dimensions, the Fundamental Theorem of Calculus takes many different forms such as Green's Theorem, Stoke's Theorem, and Gauss' Divergence Theorem. These are some of the mathematical tools needed to derive conservation laws in higher dimensions.

8.2.1 Review of multivariable calculus

For $n \in \mathbb{N}$ let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be the standard Euclidean coordinates on $\mathbb{R}^n$. Let Ω be a *well-behaved* subset of $\mathbb{R}^n$, which is to say that its interior Ω^o is not empty and its boundary $\partial\Omega$ is piecewise smooth, a finite algebraic sum (the sign indicates the orientation of the surface) of surfaces

$$\partial\Omega = S_1 + S_2 + \cdots + S_\ell,$$

for some positive integer ℓ, where each component S_j of the boundary has a continuously varying outward unit normal vector $\mathbf{n}$. This happens precisely when each S_j is the graph of a C^1 function.

Example 8.2.1. The parallelepiped

$$\{(x_1, x_2, x_3) : 0 \le x_1 \le 1, 0 \le x_2 \le 1, 0 \le x_3 \le 1\}$$

in $\mathbb{R}^3$ is a well-behaved subset. This is a unit cube in $\mathbb{R}^3$. The boundary of the cube consists of 6 sides S_i for $i = 1, 2, 3, 4, 5, 6$, each of which is parallel to a coordinate plane.

A few other examples of well-behaved subsets Ω of $\mathbb{R}^n$ are:

- balls $B(\mathbf{x}_0, r) = \{\mathbf{x} \in \mathbb{R}^n | \|\mathbf{x} - \mathbf{x}_0\| < r\}$
- parallelepipeds (including rectangles).
- the portion of $\mathbb{R}^n$ between two parallel codimension 1 planes (sometimes called a *slab*).
- cylinders whose tops and bottoms are C^1 graphs.

Remark 8.2.2. The reason for considering well-behaved subsets Ω is so we can apply the Divergence Theorem and its many consequences to functions defined on those domains.

Remark 8.2.3. The spatial domain Ω of any scalar-valued time-dependent function $u : \Omega \times I \to \mathbb{R}$ or vector-valued time-dependent function $\mathbf{F} : \Omega \times I \to \mathbb{R}^n$ will always be assumed "well-behaved." This means we are essentially assuming that every function in the following is C^∞ unless stated otherwise.

- We think of $d\mathbf{x} = dx_1 dx_2 \cdots dx_n$ as the *volume form* on $\mathbb{R}^n$ so that for any well-behaved subset Ω we have
$$\int_\Omega d\mathbf{x} = \text{Vol}(\Omega).$$
Of course if $n > 3$, then by *volume* we mean the corresponding higher-dimensional analogue of volume. And if $n = 2$, then *volume* means area; and if $n = 1$, then *volume* means length.
- We think of dA as the surface area element on the boundary $\partial\Omega = S_1 + S_2 + \cdots + S_\ell$ of Ω so that the area of the boundary of Ω is
$$\int_{\partial\Omega} dA = \int_{\sum_{i=1}^{\ell} S_i} dA = \sum_{i=1}^{\ell} \text{area of } S_i.$$
Again, if $n > 3$, then *surface area* means the corresponding higher-dimensional analogue of surface area; if $n = 2$, then the boundary consists of curves and *surface area* means path length. If $n = 1$, then the boundary consists of isolated points, and *surface area* is just the number of those points (counted with appropriate sign, corresponding to orientation).
- Let $\mathbf{F}(\mathbf{x}, t)$ be a time-dependent vector field defined on the Cartesian product $\Omega \times I$ for a domain Ω in $\mathbb{R}^n$ and an open interval I in $\mathbb{R}$.

Let $\mathbf{n}$ be the unit normal to the boundary $\partial\Omega$. We can integrate the outward unit normal component of $\mathbf{F}$, that is, $\mathbf{F} \cdot \mathbf{n}$, on the boundary of Ω:

$$\int_{\partial\Omega} \mathbf{F} \cdot \mathbf{n} \, dA = \sum_{i=1}^{l} \int_{S_i} \mathbf{F}(\mathbf{x}, t) \cdot \mathbf{n} \, dA,$$

The spatial divergence operator is represented by the row vector

$$\nabla\cdot = \begin{bmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots & \frac{\partial}{\partial x_n} \end{bmatrix}.$$

This represents the flux of the vector field through the boundary surface.
The spatial divergence of a time-dependent vector field (column vector)

$$\mathbf{F}(\mathbf{x}, t) = \big(F_1(\mathbf{x}, t), F_2(\mathbf{x}, t), \cdots, F_n(\mathbf{x}, t)\big),$$

is the scalar-valued time-dependent function

$$\begin{aligned} \nabla \cdot \mathbf{F}(\mathbf{x}, t) &= \begin{bmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots & \frac{\partial}{\partial x_n} \end{bmatrix} \begin{bmatrix} F_1(\mathbf{x}, t) \\ F_2(\mathbf{x}, t) \\ \vdots \\ F_n(\mathbf{x}, t) \end{bmatrix} \\ &= \frac{\partial F_1}{\partial x_1} + \frac{\partial F_2}{\partial x_2} + \cdots + \frac{\partial F_n}{\partial x_n}. \end{aligned}$$

Let $w(\mathbf{x}, t)$ be a scalar-valued function whose domain is the Cartesian product $\Omega \times I$ for a domain Ω in $\mathbb{R}^n$ and an open subinterval I of $\mathbb{R}$.

The spatial gradient of $w(\mathbf{x},t)$ is the time-dependent vector field

$$\nabla w(\mathbf{x},t) = (D_{\mathbf{x}} w(\mathbf{x},t))^{\mathsf{T}} = \begin{bmatrix} \partial_{x_1} w(\mathbf{x},t) \\ \partial_{x_2} w(\mathbf{x},t) \\ \vdots \\ \partial_{x_n} w(\mathbf{x},t) \end{bmatrix}.$$

For $\mathbf{x} \in \partial\Omega$, the directional derivative of $w(\mathbf{x},t)$ in the direction $\mathbf{n}$ is known as the normal derivative of w and is

$$\frac{dw}{d\mathbf{n}} = \mathbf{n} \cdot \nabla w(\mathbf{x},t).$$

The surface integral of the normal derivative of $w(\mathbf{x},t)$ on the boundary of Ω is

$$\int_{\partial\Omega} \frac{dw}{d\mathbf{n}}\, dA = \sum_{i=1}^{l} \int_{S_i} \mathbf{n} \cdot \nabla w(\mathbf{x},t)\, dA.$$

The spatial Laplacian of a time-dependent scalar function $w(\mathbf{x},t)$ is the time-dependent scalar function

$$\begin{aligned} \Delta w = \nabla^2 w = \nabla \cdot \nabla w &= \begin{bmatrix} \dfrac{\partial}{\partial x_1} & \dfrac{\partial}{\partial x_2} & \cdots & \dfrac{\partial}{\partial x_n} \end{bmatrix} \begin{bmatrix} \partial_{x_1} w(\mathbf{x},t) \\ \partial_{x_2} w(\mathbf{x},t) \\ \vdots \\ \partial_{x_n} w(\mathbf{x},t) \end{bmatrix} \\ &= \frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2} + \cdots + \frac{\partial^2 w}{\partial x_n^2}. \end{aligned}$$

With these notations in place we can state the Divergence Theorem and some of its many corollaries that we will be using.

Theorem 8.2.4. *If the time-dependent vector field $\mathbf{F}(\mathbf{x},t)$ is in $C^1(\overline{\Omega \times I}) \cap C^2(\Omega \times I)$, then*

$$\int_{\Omega} \nabla \cdot \mathbf{F}\, d\mathbf{x} = \int_{\partial\Omega} \mathbf{F} \cdot \mathbf{n}\, dA.$$

Corollary 8.2.5. *If the time-dependent scalar-valued function $u(\mathbf{x},t)$ is $C^1(\overline{\Omega \times I}) \cap C^2(\Omega \times I)$, then*

$$\int_{\Omega} u_{x_k}\, d\mathbf{x} = \int_{\partial\Omega} u\, \mathrm{n}_k\, dA,\ k = 1, 2, \ldots, n.$$

Proof. Take $\mathbf{F}(\mathbf{x},t)$ to be the time-dependent vector field

$$\mathbf{F}(\mathbf{x},t) = (0, 0, \ldots, u(\mathbf{x},t), 0, \ldots, 0)$$

where the function $u(\mathbf{x},t)$ appears in the k^{th} slot. Then the spatial divergence of $\mathbf{F}$ is

$$\nabla \cdot \mathbf{F}(\mathbf{x}(t) = \partial_{x_k} u(\mathbf{x},t)$$

and the normal component of $\mathbf{F}$ is

$$\mathbf{F}(\mathbf{x},t) \cdot \mathbf{n} = u(\mathbf{x},t)\, \mathrm{n}_k.$$

By the Divergence Theorem we obtain

$$\int_{\Omega} \partial_{x_k} u(\mathbf{x},t)\, d\mathbf{x} = \int_{\Omega} \nabla \cdot \mathbf{F}\, d\mathbf{x} = \int_{\partial\Omega} \mathbf{F} \cdot \mathbf{n}\, dA = \int_{\partial\Omega} u(\mathbf{x},t)\, \mathrm{n}_k\, dA.$$

Since this holds for all $i = 1, 2, \ldots, n$, this gives the result. □

Remark 8.2.6. This corollary is equivalent to the Divergence Theorem. This can be seen by taking the scalar-valued function $u(\mathbf{x},t)$ to be the k^{th} component of the vector-valued function $\mathbf{F}(\mathbf{x},t)$, i.e., $u(\mathbf{x},t) = F_k(\mathbf{x},t)$, and applying the Corollary to the scalar components for each k

$$u_{x_k} = \frac{\partial F_k}{\partial x_k} \text{ and } u\,\mathrm{n}_k = F_k\,\mathrm{n}_k,$$

which gives

$$\int_\Omega \frac{\partial F_k}{\partial x_k}\,d\mathbf{x} = \int_\Omega u_{x_k}\,d\mathbf{x} = \int_{\partial\Omega} u\,\mathrm{n}_k\,dA = \int_{\partial\Omega} F_k\,\mathrm{n}_k\,dA.$$

This implies, since the divergence

$$\nabla\cdot\mathbf{F} = \sum_{k=1}^{n}\frac{\partial F_k}{\partial x_k},$$

and the normal component

$$\mathbf{F}\cdot\mathbf{n} = \sum_{k=1}^{n} F_k\,\mathrm{n}_k$$

are both sums, that

$$\begin{aligned}\int_\Omega \nabla\cdot\mathbf{F}\,d\mathbf{x} &= \int_\Omega \sum_{k=1}^{n}\frac{\partial F_k}{\partial x_k}\,d\mathbf{x}\\ &= \sum_{k=1}^{n}\int_\Omega \frac{\partial F_k}{\partial x_k}\,d\mathbf{x}\\ &= \sum_{k=1}^{n}\int_{\partial\Omega} F_k\,\mathrm{n}_k\,dA\\ &= \int_{\partial\Omega}\sum_{k=1}^{n} F_k\,\mathrm{n}_k\,dA\\ &= \int_{\partial\Omega}\mathbf{F}\cdot\mathbf{n}\,dA.\end{aligned}$$

This gives the Divergence Theorem.

Additional extensions of the Divergence Theorem are provided in the following Corollary

Corollary 8.2.7. *If time-dependent scalar-valued functions $u(\mathbf{x},t)$ and $w(\mathbf{x},t)$ and a time-dependent vector field $\mathbf{F}(\mathbf{x},t)$ are all $C^1(\overline{\Omega\times I})\cap C^2(\Omega\times I)$, then there holds*

(i) an integration by parts formula,

$$\int_\Omega w u_{x_k}\,d\mathbf{x} = -\int_\Omega u w_{x_k}\,d\mathbf{x} + \int_{\partial\Omega} uw\,\mathrm{n}_k\,dA,$$

(ii) Green's First Identity,

$$\int_\Omega \left(u\Delta w + \nabla u\cdot\nabla w\right)\,d\mathbf{x} = \int_{\partial\Omega} u\frac{dw}{d\mathbf{n}}\,dA,$$

(iii) and Green's Second Identity,

$$\int_\Omega u\Delta w\,d\mathbf{x} = \int_\Omega w\Delta u\,d\mathbf{x} + \int_{\partial\Omega}\left(u\frac{dw}{d\mathbf{n}} - w\frac{du}{d\mathbf{n}}\right)\,dA.$$

Proof. (i) The product scalar-valued function

$$v(\mathbf{x},t) = u(\mathbf{x},t)w(\mathbf{x},t)$$

is $C^1(\overline{\Omega \times I}) \cap C^2(\Omega \times I)$.

Thus we can apply the Corollary to it to obtain for each $k = 1, 2, \ldots, n$,

$$\int_\Omega v_{x_k}\ d\mathbf{x} = \int_{\partial\Omega} v\,\mathrm{n}_k\ dA.$$

Since

$$v_{x_k} = (uw)_{x_k} = u_{x_k}w + uw_{x_k}$$

we obtain

$$\int_\Omega v_{x_k}\ d\mathbf{x} = \int_\Omega \left(u_{x_k}w + uw_{x_k}\right)\ d\mathbf{x} = \int_\Omega u_{x_k}w\ d\mathbf{x} + \int_\Omega uw_{x_k}\ d\mathbf{x}.$$

Since $v = uw$ we obtain

$$\int_{\partial\Omega} v\,\mathrm{n}_k\ dA = \int_{\partial\Omega} uw\,\mathrm{n}_k\ dA.$$

Putting the pieces together gives

$$\int_\Omega u_{x_k}w\ d\mathbf{x} + \int_\Omega uw_{x_k}\ d\mathbf{x} = \int_\Omega v_{x_k}\ d\mathbf{x} = \int_{\partial\Omega} v\,\mathrm{n}_k\ dA = \int_{\partial\Omega} uw\,\mathrm{n}_k\ dA.$$

Commuting w and u_{x_k} in the integral with integrand $u_{x_k}w$ and moving the integral with integrand to the right-hand side gives

$$\int_\Omega wu_{x_k}\ d\mathbf{x} = -\int_\Omega uw_{x_k}\ d\mathbf{x} + \int_{\partial\Omega} uw\,\mathrm{n}_k\ dA.$$

Parts (ii) and (iii) are proved in the exercises. ∎

8.2.2 Conservation laws in higher dimensions

A conservation law is a fundamental law of nature that states that a quantity (say a chemical species in a chemical reaction) is balanced throughout a specified process (chemical reaction). The conservation law or balance for a chemical species is that the time rate of change of the total amount of the chemical species in some domain *must* equal the rate at which the chemical species flows into that domain minus the rate at which the chemical species flows out of that domain, plus the rate at which the chemical species is created (source) or consumed (sink) in that domain.

- In other words, the conservation law accounts for all of the chemical species by means of an equation.
- The conservation law or equation can be expressed mathematically using the Divergence Theorem to formulate everything nicely (in one dimension we relied on the Fundamental Theorem of Calculus).
- Suppose we have a quantity (think chemical species) distributed throughout a domain $\Omega \subset \mathbb{R}^n$.
- The density or concentration of this quantity is represented by a scalar-valued function $u : \Omega \times I \to \mathbb{R}$ for I an open interval of $\mathbb{R}$.
- For each open ball $B_r(\mathbf{x})$, $\mathbf{x} \in \Omega^o$, let $N_r(\mathbf{x})$ denote the amount of the quantity in the ball $B_r(\mathbf{x})$ at time t.

- Then the density or concentration of the quantity at the point $\mathbf{x}$ at time t is given by
$$u(\mathbf{x},t) = \lim_{r\to 0} \frac{N_r(\mathbf{x})}{\mathrm{Vol}(B_r(\mathbf{x}))}.$$
Note that the dimensions of $u(\mathbf{x},t)$ are quantity per volume: $[u] = \mathsf{NL}^{-1}$ for the three-dimensional case, and $[u] = \mathsf{NL}^{-n}$ for the n-dimensional case.

For a well-behaved subset E of Ω (in one dimension a well behaved subset would be any finite interval $[a,b]$), the time-dependent integral
$$N_E(t) = \int_E u(\mathbf{x},t)\ d\mathbf{x}$$
gives the amount of the quantity in the domain E at time t.

- Assuming additionally that $u(\mathbf{x},t)$ is C^1 then we have
$$\frac{dN_E}{dt} = \frac{d}{dt}\int_E u(\mathbf{x},t)\ d\mathbf{x} = \int_E u_t(\mathbf{x},t)\ d\mathbf{x}.$$
- Let $\mathbf{J} : \Omega \times I \to \mathbb{R}^n$ be the flux vector field of the quantity. The flux is a vector field that describes how the concentration of material moves into or out of a region. In $\mathbb{R}^3$ its dimensions are $[J] = \mathsf{NL}^{-2}\mathsf{T}^{-1}$ because it describes the rate at which the material flows through a small piece of surface (with dimensions L^2 at that point. in $\mathbb{R}^n$ its dimensions are $[J] = \mathsf{NL}^{-(n-1)}\mathsf{T}^{-1}$.
- For any well-behaved subset E of Ω, let $\mathbf{n}$ denote the outward-facing unit normal at each point of the boundary ∂E. Since $\mathbf{n}$ has been rescaled to have length 1, it is dimensionless. The normal component $J\cdot\mathbf{n}$ of the flux J represents the rate at which the quantity crosses ∂E. Its dimensions are the same as those of J, and represent the rate per unit area at which the chemical species crosses the boundary of E.

Since the normal vector is the outward normal, the net outward flux of the quantity across ∂E is the surface integral
$$\int_{\partial E} \mathbf{J}\cdot\mathbf{n}\ dA = \sum_{i=1}^{l}\int_{S_i} \mathbf{J}(\mathbf{x},t)\cdot\mathbf{n}\ dA,$$
where each of the S_i are the individual components of the boundary ∂E. Since dA has dimensions $[dA] = \mathsf{L}^{n-1}$, the dimensions of the integrand (and the resulting integral) are NT^{-1}, as expected. By the Divergence Theorem it follows that
$$\int_E \nabla\cdot\mathbf{J}\ d\mathbf{x} = \int_{\partial E} \mathbf{J}\cdot\mathbf{n}\ dA.$$
This integral accounts for that part of the quantity which leaves E; changing the sign gives the amount of the chemical species that enters E. All the terms in the divergence are first derivatives of J with respect to space, so they all change the dimensions of J by L^{-1}. The volume form $d\mathbf{x} = dx_1\dots dx_n$ has dimensions L^n, so again the integrand on the left and the resulting integral have the expected dimensions NT^{-1}.

We account for any sources and/or sinks for the quantity in Ω by means of a function $f : \Omega\times I \to \mathbb{R}$ in terms of the integral
$$\int_E f(\mathbf{x},t)\,d\mathbf{x}.$$
Putting all the pieces together we have for the concentration $u(\mathbf{x},t)$, the conservation law
$$\int_E u_t(\mathbf{x},t)\ d\mathbf{x} = -\int_E \nabla\cdot\mathbf{J}(\mathbf{x},t)\ d\mathbf{x} + \int_E f(\mathbf{x},t)\ d\mathbf{x}.$$
which holds for any well-behaved subset E of Ω.

Remark 8.2.8. The minus sign in front of the integral with the divergence of the flux vector field is because this integral represent that part of the quantity that leaves E across ∂E, i.e., the unit normal used is the outward unit normal.

Combining the integrals into one integral over E and suppressing the dependence of the functions on $\mathbf{x}$ and t gives the conservation law

$$\int_E \{u_t + \nabla \cdot \mathbf{J} - f\} \; d\mathbf{x} = 0. \tag{8.3}$$

Just in case you forgot, this holds for all well-behaved subsets E of Ω. Assuming the integrand is continuous (because we haven't made enough assumptions yet so we may as well start assuming away), if there were $\bar{x}$ where the integrand were not zero, then by continuity of the integrand there would be a compact subset E of Ω whose open interior E^o contains $\bar{x}$ such that $u_t + \nabla \cdot \mathbf{J} - f$ is strictly bounded away from zero on E, contradicting the conservation law. Hence because this integral form of the conservation law holds for all such E then we obtain the PDE version of the conservation law,

$$u_t + \nabla \cdot \mathbf{J} = f. \tag{8.4}$$

Remark 8.2.9. The final relationship that we need to define (we have carefully avoided this up to now) is a connection between the flux vector field $\mathbf{J}$ and the density/concentration $u(\mathbf{x}, t)$ of the quantity in question. This relationship is known as a constitutive relationship or an equation of state, and depends on assumptions about the physical properties of the medium the quantity is in, which in turn derive from empirical reasoning and experimentation. Whereas the conservation law is a fundamental law of nature that is expressed mathematically (and depends on some continuity and smoothness assumptions), the constitutive relation is an approximation whose origin lies almost entirely in empirical observation.

As in the one-dimensional case, there are two important types of flux that come up repeatedly, namely *advective* and *diffusive*.

- *Advective flux* describes flux of a quantity that is carried with fluid flow: $\mathbf{J}(t, \mathbf{x}) = u(t, \mathbf{x})\mathbf{v}(t, x)$, where $\mathbf{v}(t, \mathbf{x})$ is the velocity field of fluid. This is compatible with the dimensional analysis above $[u\mathbf{v}] = (\mathsf{NL}^{-n})(\mathsf{LT}^{-1} = \mathsf{NL}^{-(n-1)}\mathsf{T}^{-1} = [J]$.
- *Diffusive flux* describes the flux of a quantity, like heat in a solid, that diffuses over time: $\mathbf{J}(x, t) = -\kappa(\mathbf{x})\nabla u(t, \mathbf{x})$ for some nonnegative function $\kappa : \Omega \to \mathbb{R}$. The dimensions of ∇ are $[\nabla] = \mathsf{L}^{-1}$, so $[\nabla u] = \mathsf{NL}^{-(n+1)}$. Since the dimensions of flux are $[J] = \mathsf{NL}^{-(n-1)}\mathsf{T}^{-1}$, we must have $[\kappa] = \mathsf{L}^2 T^{-1}$.

These two fluxes need to be combined (added) for a quantity that both diffuses and is carried with the flow (like dye in flowing water).

Example 8.2.10. The diffusive flux

$$\mathbf{J}(\mathbf{x}, t) = -\kappa(\mathbf{x})\nabla u(t, \mathbf{x}),$$

is known as Fick's Law. Recall that the gradient points in the direction of steepest increase, so that this flux vector field points in the direction of steepest decrease (scaled by $\kappa(\mathbf{x})$). This agrees with the expectation that a quantity will move from regions of high density to regions of low density.

For many situations the function $\kappa(\mathbf{x})$ is independent of $\mathbf{x}$, and thus the PDE form of the conservation law is

$$u_t - \kappa \nabla \cdot \nabla u = f.$$

Recalling that $\nabla \cdot \nabla u = \Delta u$, we obtain the heat diffusion equation with sources/sinks,

$$u_t = \kappa \Delta u + f,$$

which is known as a *reaction-diffusion* equation (the reaction part arising from the source/sink term f and the diffusion part arising from the term $\kappa \Delta u$). In this case the constant κ is known as the thermal diffusivity and is a constant related to the material in which the heat is diffusing, i.e. it is called a material property (not dynamic) of the material in that different materials will have different values of κ.

8.3 Method of Characteristics

Before we dive more into different types of PDEs, we will first consider a particular method of solution that gives some insight into the nature of certain physical phenomena, and will be used below to motivate the physical intuition to a certain class of solutions and clarifies the effectiveness (or lack thereof) of certain numerical methods. The method of characteristics is typically applied to hyperbolic problems, but has its usefulness in other problems as well. With this in mind, we will consider the first-order, quasilinear, homogeneous PDE

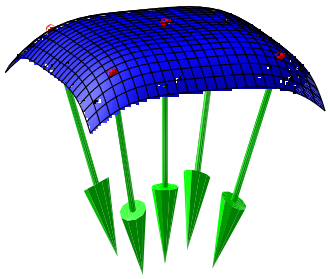
$$a(x, y, u)u_x + b(x, y, u)u_y = c(x, y, u). \tag{8.5}$$

The graph of $z = u(x, y)$ is a surface in xyz-space.

Remark 8.3.1. When u is a solution of (8.5) then the graph of u is called an integral surface for the PDE. Recall from multivariable calculus that the normal of the surface $z = u(x, y)$ at a point is the gradient of $w(x, y, z) = u(x, y) - z$, i.e.

$$\nabla w(x, y, z) = (u_x, u_y, -1).$$

We illustrate this graphically in the following where the graph of $u(x, y)$ is the blue surface and the five green arrows are a sampling of the vectors $(u_x, u_y, -1)$ located at the red points on the graph of $u(x, y)$.



On an integral surface $z = u(x, y)$ where $u(x, y)$ is a solution of (8.5), then at each point $(x, y, u(x, y))$ the normal $(u_x, u_y, -1)$ is perpendicular to the coefficient vector

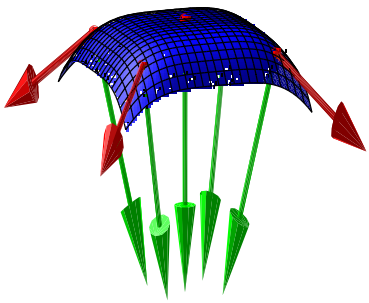
$$(a, b, c) = (a(x, y, u), b(x, y, u), c(x, y, u))$$

of the PDE, because

$$(u_x, u_y, -1) \cdot (a, b, c) = au_x + bu_y - c = 0.$$

- Recall that the tangent plane of a surface $z = u(x, y)$ at a point $(x, y, u(x, y))$ is the plane perpendicular to the normal vector $(u_x, u_y, -1)$ at that point, i.e. the tangent plane is defined as all those points whose dot product with the normal vector is zero.
- Thus at each point of the integral surface, the vector or *characteristic direction* (a, b, c) belongs to the tangent plane of the integral surface.
- Thus the characteristic direction (a, b, c) is a tangent vector on the integral surface $z = u(x, y)$.

This is illustrated in the following Figure, where the red arrows are tangent to the integral surface or graph of $u(x, y)$.



This characteristic direction or tangent vector to the integral surface is a geometric tool for solving a quasilinear first-order PDE.

Definition 8.3.2. *Making the the substitution $z = u(x, y)$ in a, b, and c, the characteristic direction (a, b, c) for (8.5) is a vector field on the integral surface. An integral curve of this vector field is a curve $(x(s), y(s), z(s))$ whose tangent vectors match the vector field at each point of the curve. That means the curve must satisfy the first-order system of ODEs*

$$\frac{dx}{ds} = a(x, y, z), \quad \frac{dy}{ds} = b(x, y, z), \quad \frac{dz}{ds} = c(x, y, z),$$

called the characteristic differential equations *of the PDE. We call such a curve a* characteristic curve *of the PDE.*

The graph of a characteristic curve $(x(s), y(s), z(s))$ lies on the integral surface $z = u(x, y)$ so that

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

Using the Chain Rule to take the derivative of this with respect to s gives

$$\frac{dz}{ds} = \frac{\partial u}{\partial x}\frac{dx}{ds} + \frac{\partial u}{\partial y}\frac{dy}{ds}. \tag{8.6}$$

Definition 8.3.3. *A characteristic projection is the projection of a characteristic curve $(x(s), y(s), z(s))$ to the independent variable space, i.e., the curve $(x(s), y(s))$.*

Remark 8.3.4. The variable s is a "dummy" variable that parameterizes the characteristic curves that lie on the integral surface. We can rescale the characteristic differential equations which reparameterize the characteristic curves. There is no physical meaning for the parameterizing variable s.

Example 8.3.5. The linear homogeneous PDE

$$a(x,y)u_x + b(x,y)u_y = 0 \tag{8.7}$$

is a quasilinear first-order PDE and the characteristic differential equations for it are

$$\frac{dx}{ds} = a(x,y), \ \frac{dy}{ds} = b(x,y), \ \frac{dz}{ds} = 0.$$

Since $\frac{dz}{ds} = 0$, equation (8.6) above becomes

$$0 = \frac{\partial u}{\partial x}\frac{dx}{ds} + \frac{\partial u}{\partial y}\frac{dy}{ds}.$$

Remark 8.3.6. To solve the first-order planar system

$$\frac{dx}{ds} = a(x,y), \ \frac{dy}{ds} = b(x,y),$$

we can eliminate the "dummy" variable s via the Chain Rule and convert it into a first-order equation:

$$\frac{dy}{dx} = \frac{dy}{ds}\frac{ds}{dx} = \frac{dy/ds}{dx/ds} = \frac{b(x,y)}{a(x,y)}.$$

Typically the solutions of this first-order equation are given implicitly by the level curves of some function $F(x,y)$; we assume this is the case for the sake of argument in all that follows.

Example 8.3.7 (Continuation of previous example). The level curves of the function $F(x,y)$ implicitly represent the characteristic projections onto the xy-plane because by implicit differentiation:

$$F(x,y) = \text{constant} \ \Rightarrow \ \frac{\partial F}{\partial x} + \frac{\partial F}{\partial y}\frac{dy}{dx} = 0 \ \Rightarrow \ -\frac{\partial F/\partial x}{\partial F/\partial y} = \frac{dy}{dx} = \frac{b(x,y)}{a(x,y)}.$$

The equation $\frac{dz}{ds} = 0$ says that z (recall that $z = u(x,y)$) is constant along each characteristic projection which is given by $F(x,y) = \text{constant}$.

Changing the characteristic projection $F(x,y) = \text{constant}$, i.e., changing the constant, means the value of the constant z may change too, i.e. the dependence of $z = u$ reduces to dependence on a single variable. In other words there is a function f such that

$$z = f(\text{constant}).$$

Putting the pieces together we obtain

$$u(x,y) = z = f(\text{constant}) = f(F(x,y)).$$

Assuming f and F are C^1 we verify that this form of u is indeed a solution of the PDE as follows:

$$\begin{aligned} a(x,y)u_x + b(x,y)u_y &= a(x,y)f'(F(x,y))\frac{\partial F}{\partial x} + b(x,y)f'(F(x,y))\frac{\partial F}{\partial y} \\ &= a(x,y)f'(F(x,y))\left[-\frac{b(x,y)}{a(x,y)}\frac{\partial F}{\partial y}\right] + b(x,y)f'(F(x,y))\frac{\partial F}{\partial y} \\ &= -b(x,y)f'(F(x,y))\frac{\partial F}{\partial y} + b(x,y)f'(F(x,y))\frac{\partial F}{\partial y} = 0. \end{aligned}$$

Remark 8.3.8. Note that in the derivation above, we have stealthily avoided considering issues that may arise if/when $a(x,y) = 0$ for any values of x and/or y. Such issues do arise in various settings, so our avoidance of the topic shouldn't be seen as justification that such circumstances don't occur. On the contrary, our avoidance is purely out of self-preservation, i.e. such assumptions make the derivation much simpler.

Example 8.3.9. Replacing y with t we consider the first-order linear homogeneous PDE

$$au_x + u_t = 0$$

where a is a constant. This PDE is known as the scalar advection equation and is written with the partial derivative with respect to t first:

$$u_t + au_x = 0.$$

The characteristic differential equations are

$$\frac{dx}{ds} = a, \quad \frac{dt}{ds} = 1, \quad \frac{dz}{ds} = 0.$$

The last equation says the value of $z = u(x,y)$ does not change along a characteristic curve.

By the Chain Rule the first two equations combine to give

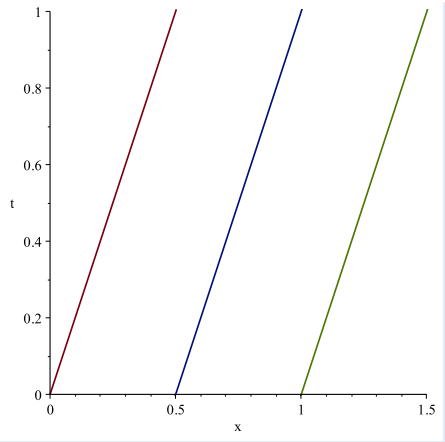
$$\frac{dx}{dt} = \frac{dx}{ds}\frac{ds}{dt} = a.$$

A characteristic projection (this is most frequently called a characteristic curve) in the xt-plane is

$$x = at + c_0 \text{ or } c_0 = x - at,$$

for an arbitrary constant c_0.

Here is a graph of some of the characteristic projections/curves in the xt-plane (for $a = \frac{1}{2}$).



Because this PDE is homogeneous, along these characteristic projections the value of u is constant, i.e., the value of u depends only on the quantity $c_0 = x - at$, and only by crossing the characteristic projections does the value of u change. Putting it another way, $u(x,t)$ maintains a fixed value along each one of these individual characteristic curves, i.e. the solution $u(x,t)$ at any point in time and space (x,t) is completely defined by which characteristic curve (x,t) lies on.

This means there is a function f such that

$$u(x,t) = f(c_0) = f(x - at).$$

Assuming f is C^1 we can verify that this is indeed a solution of the PDE:

$$u_t + au_x = -af'(x - at) + af'(x - at) = 0.$$

Since a solution $u(x,t)$ is constant along the characteristic projections $x - at =$ constant, numerically computing $u(x, t + \Delta t)$ can be done with no error, i.e. we just 'follow' the characteristics backward from $(x, t + \Delta t)$ to some previous point at time t where we know the solution value already! In this case, because the characteristics are straight lines, the algebra is relatively simple.

Suppose we know that value of $u(x,t)$ at time t for all values of x. For Δt and Δx satisfying

$$\frac{\Delta t}{\Delta x} = \frac{1}{a}$$

(the slope of the characteristic projection $x - at = c_0$ has slope $\frac{1}{a}$ in the xt-plane) the points

$$(x,t) \text{ and } (x + \Delta x, t + \Delta t)$$

lie on the same characteristic projection because their difference

$$(x + \Delta x, t + \Delta t) - (x,t) = (\Delta x, \Delta t)$$

has the ratio

$$\frac{\Delta t}{\Delta x} = \frac{1}{a}$$

and the point $(x+\Delta x, t+\Delta t)$ is on the characteristic projection $x-at=c_0$,

$$\begin{aligned}(x+\Delta t)-a(t+\Delta t) &= x+\Delta x-at-a\Delta t\\ &= x+\Delta x-at-a((\frac{1}{a})\Delta x)\\ &= x-at=c_0.\end{aligned}$$

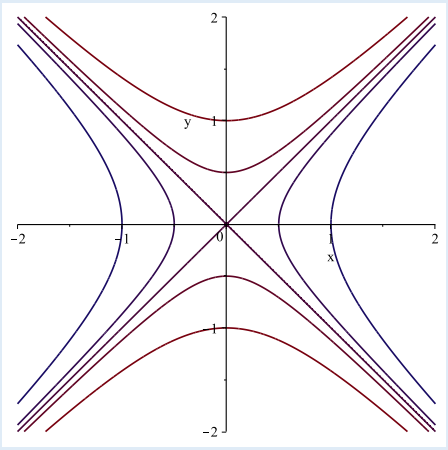
Example 8.3.10. Consider the first-order linear homogeneous PDE

$$yu_x+xu_y=0.$$

You will show that the characteristic (projections) are the level curves of

$$F(x,y)=x^2-y^2.$$

Here is a graph of some of the characteristic projections.



Since the linear PDE is homogeneous a solution u is constant along each characteristic projection.

For $c>0$ the characteristic projection $x^2-y^2=c$ is made up of two distinct curves

$$x=\sqrt{c+y^2} \text{ and } x=-\sqrt{c+y^2},\ y\in\mathbb{R}.$$

So when $c>0$ there are arbitrary C^1 functions f_1 and g_1 for which a solution is

$$u(x,y)=\begin{cases} f_1(x^2-y^2) & \text{if } x>0,\\ g_1(x^2-y^2) & \text{if } x<0.\end{cases}$$

Looking at the limits of these parts along the x-axis, i.e., at points of the form $(\pm x,0)$ we require for u to be continuous at $(0,0)$ that

$$f_1(0)=\lim_{x\to 0^+} f_1(x^2-0^2)=\lim_{x\to 0^-} g_1(x^2-0^2)=g_1(0).$$

For $c < 0$ the characteristic projection $x^2 - y^2 = c$ is made up of two distinct curves

$$y = \sqrt{-c + x^2} \text{ and } y = -\sqrt{-c + x^2},\ x \in \mathbb{R}.$$

So when $c < 0$ there are arbitrary C^1 functions f_2 and g_2 for which a solution is

$$u(x, y) = \begin{cases} f_2(x^2 - y^2) & \text{if } y > 0, \\ g_2(x^2 - y^2) & \text{if } y < 0. \end{cases}$$

Looking at the limits of each of these curves along the y-axis, i.e., at points of the form $(0, \pm y)$ we require for u to be continuous at $(0, 0)$ that

$$f_2(0) = \lim_{y \to 0^+} f_2(0^2 - y^2) = \lim_{y \to 0^-} g_2(0^2 - y^2) = g_2(0).$$

For all four parts of u to be continuous at the origin $(0, 0)$ we require that the function f_1, f_2, g_1, and g_2 satisfy

$$f_2(0) = f_1(0) = g_1(0) = g_2(0).$$

This common value will be the value of u on the characteristic projection $x^2 - y^2 = 0$ which contains the origin $(0, 0)$.

Remark 8.3.11. The method of characteristics can be applied to a quasilinear first-order PDE for a scalar-valued function u dependent on three or more variables. We define the characteristic differential equations for first-order linear homogeneous PDEs and illustrate this for a scalar-valued function u of three independent variables.

Definition 8.3.12. *Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be standard Euclidean coordinates on $\mathbb{R}^n$. For a first-order linear homogeneous PDE in the scalar-valued function $u(\mathbf{x})$ defined by*

$$a_1(\mathbf{x})u_{x_1} + a_2(\mathbf{x})u_{x_2} + \cdots + a_n(\mathbf{x})u_{x_n} = \sum_{k=1}^{n} a_k(\mathbf{x})u_{x_k} = 0,$$

the characteristic differential equations are:

$$\frac{dx_1}{ds} = a_1(\mathbf{x}), \frac{dx_2}{ds} = a_2(\mathbf{x}), \ldots, \frac{dx_n}{ds} = a_n(\mathbf{x}),\ \frac{dz}{ds} = 0,$$

where $z = u(\mathbf{x}) = u(x_1, x_2, \ldots, x_n)$.

Remark 8.3.13. The characteristic projections are curves in the n-dimensional space $(x_1, x_2, \ldots, x_n)$. In the event that none of the coefficient functions $a_k(\mathbf{x})$ vanish (equal zero) we can eliminate s from the first n equations for the characteristic projections. This is done by pairs: for distinct $i, j \in \{1, 2, \ldots, n\}$ the differential equations

$$\frac{dx_i}{ds} = a_i(\mathbf{x}) \text{ and } \frac{dx_j}{ds} = a_j(\mathbf{x})$$

imply by the Chain Rule that

$$\frac{dx_i}{dx_j} = \frac{dx_i}{ds}\frac{ds}{dx_j} = \frac{a_i(\mathbf{x})}{a_j(\mathbf{x})}.$$

From this we obtain

$$\frac{dx_i}{a_i(\mathbf{x})} = \frac{dx_j}{a_j(\mathbf{x})}.$$

Since this holds for distinct arbitrary $i, j \in \{1, 2, \ldots, n\}$ there holds (suppressing the dependence of a_i on $\mathbf{x}$)

$$\frac{dx_1}{a_1} = \frac{dx_2}{a_2} = \cdots = \frac{dx_n}{a_n}.$$

It is sufficient to solve $n-1$ of the pairings in the n variables implicitly to give

$$F_k(\mathbf{x}) = c_k, \ k = 1, 2, \ldots, n-1,$$

for constants $c_1, c_2, \ldots, c_{n-1}$.

Since the linear PDE is homogeneous, the solution is constant along the characteristic projections given implicitly by $F_k(\mathbf{x}) = c_k$, $k = 1, 2, \ldots, n-1$. Thus there exists an arbitrary function f such that

$$u(\mathbf{x}) = f(c_1, c_2, \ldots, c_{n-1}) = f(F_1(\mathbf{x}), F_2(\mathbf{x}), \cdots, F_{n-1}(\mathbf{x})).$$

Example 8.3.14. The characteristic projections of

$$u_{x_1} + 2u_{x_2} + 3u_{x_3} = 0$$

are given by

$$\frac{dx_1}{1} = \frac{dx_2}{2} = \frac{dx_3}{3}.$$

Choosing two pairings to solve, say

$$\frac{dx_1}{1} = \frac{dx_2}{2} \text{ and } \frac{dx_1}{1} = \frac{dx_3}{3},$$

gives

$$F_1(x_1, x_2, x_3) = 2x_1 - x_2 = c_1 \text{ and } F_2(x_1, x_2, x_3) = 3x_1 - x_3 = c_2.$$

Notice that for fixed c_1 and fixed c_2, the variables x_2 and x_3 are functions of x_1, i.e., we can rewrite $x_2 = 2x_1 - c_1$ and $x_3 = 3x_1 - c_2$.

Geometrically the two planes $F_1(x_1, x_2, x_3) = c_1$ and $F_2(x_1, x_2, x_3) = c_2$ intersect along the line

$$\{(x_1, x_2, x_3) = (x_1, 2x_1 - c_1, 3x_1 - c_2) : x_1 \in \mathbb{R}\}.$$

Since the linear PDE is homogeneous, a solution is constant along a characteristic projection, so there is an arbitrary function f of two variables such that

$$u(x_1, x_2, x_3) = f(c_1, c_2) = f(F_1(x_1, x_2, x_3), F_2(x_1, x_2, x_3)) = f(2x_1 - x_2, 3x_1 - x_3).$$

When f is C^1 we can verify this as follows:

$$u_{x_1} + 2u_{x_2} + 3u_{x_3} = 2\partial_{c_1} f + 3\partial_{c_2} f - 2(\partial_{c_1} f) - 3(\partial_{c_2} f) = 0.$$

Example 8.3.15 (Inviscid Burger's Equation). The quasilinear homogeneous first-order PDE

$$u_t + uu_x = 0 \tag{8.8}$$

is known as the *inviscid Burger's equation* (it doesn't have a u_{xx} term, i.e. a viscous term) in one-dimension.

Thinking of t as y and rewriting the equation as

$$uu_x + u_t = 0,$$

the characteristic differential equations are

$$\frac{dx}{ds} = u, \frac{dt}{ds} = 1, \frac{dz}{ds} = 0.$$

The last equation says that $z = u(x, y)$ is constant along the characteristic projections. We can combine the first two characteristic differential equations via the Chain Rule into one ODE:

$$\frac{dx}{dt} = \frac{dx}{ds}\frac{ds}{dt} = u.$$

Since u is constant along a characteristic projection this can be solved exactly to give

$$x - tu = c$$

for an unknown constant c. Solving this for u gives

$$u(x, t) = \frac{x - c}{t}$$

which is not defined when $t = 0$. This means that although we can come up with an explicit formula for $u(x, t)$ we can not use it over the entire interval that we are interested in.

Without writing out an explicit solution we can still identify that inviscid Burger's is going to have some problematic behavior. We return to the characteristic projection $x - tu = c$ and suppose for $x_1 < x_2$ that at time $t = 0$ we know

$$u(x_1, 0) = u_1 \text{ and } u(x_2, 0) = u_2.$$

The values of x_1 and x_2 determine the value of the arbitrary constant c at each starting point as

$$c_1 = x_1 - 0u_1 \text{ and } c_2 = x_2 - 0u_2.$$

Because the quasilinear PDE is homogeneous, the value of u remains constant along the characteristic projection, i.e. the solution along $x - tu = c_1$ is completely determined by the constant c_1.

The equations of the characteristic projections that start at the points $(x_1, 0)$ and $(x_2, 0)$ have the equations

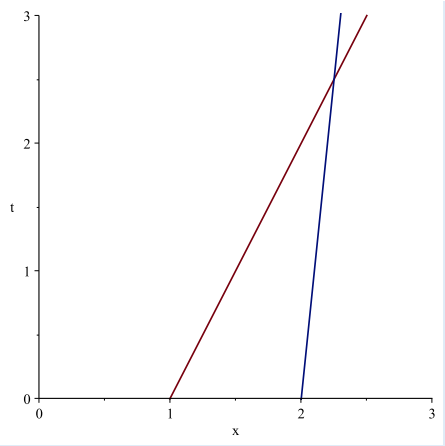
$$x - tu_1 = x_1 \text{ and } x - tu_2 = x_2.$$

Solving each of these for t as functions of x gives

$$t = \frac{x}{u_1} - \frac{x_1}{u_1} \text{ and } t = \frac{x}{u_2} - \frac{x_2}{u_2}.$$

What happens when $u_1 > u_2$?

Here is the graph of the two characteristic projections when $x_1 = 1$, $x_2 = 2$, $u_1 = 0.5$, and $u_2 = 0.1$.



At a *finite* time in the future the two characteristic projections intersect, at which time and position the solution u becomes multi-valued (meaning u is no longer a function). This is because the solution must remain constant along each of these distinct characteristics even though we have just assumed that the solution was uniquely defined on each one.

This illustrates the possibility of discontinuities or shocks appearing in solutions of nonlinear PDEs in finite time. This is only possible for nonlinear PDEs, and represents a specific class of phenomena that are fundamentally of great interest both physically and theoretically. Calculating the exact point of intersection of the two crossing characteristic projections is performed in the exercises.

Hopefully the last example illustrates for you just how useful the method of characteristics can be, even without formulating an exact solution to a given PDE. Without describing the explicit solution of Burger's equation, we were able to identify that unique solutions may not exist (depending on the initial conditions for the specific problem). This illustrates one of the reasons that characteristics are essential to understand, particularly for first-order PDEs where they are seen the most. As was already mentioned PDEs are not so simple that defining an initial condition will make things work out (we have stepped away from the world of ODEs), and the method of characteristics does a great job of illustrating what type of additional conditions are necessary to obtain a unique solution.

8.4 Auxiliary conditions and the method of characteristics

For first-order quasilinear PDEs solved using the method of characteristics it is sufficient to specify the solution along a curve that intersects every characteristic curve, exactly once. If the solution is prescribed at more than one point on the same characteristic there would be no solution because the solution can't satisfy 2 different conditions, and if there was a characteristic for which the solution was not specified anywhere than the solution would not be unique because there is nothing pinning the solution down in that case. More generally when a possibly finite union of parameterized manifolds of auxiliary conditions intersects each characteristic curve exactly once, then those characteristic curves uniquely define a solution along which the auxiliary conditions evolve.

We explore this concept via several well designed examples.

Example 8.4.1. We return to the one-dimensional advection equation

$$u_t + au_x = 0$$

for $a > 0$, whose characteristic differential equations are

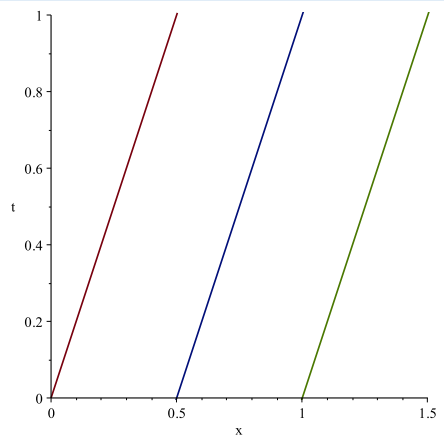
$$\frac{dx}{ds} = a, \ \frac{dt}{ds} = 1, \ \frac{dz}{ds} = 0.$$

The first two combine into one ODE $\frac{dx}{dt} = a$ whose solution

$$x = at + c \text{ or } x - at = c$$

for an arbitrary constant c, are the characteristic projections in the xt-plane. Recall since $\frac{dz}{ds} = 0$ that $u(x,t) = f(x-at)$ is the form of the solution of the PDE when f is an arbitrary C^1 function.

Here is a graph of some of these characteristic projections for $a = \frac{1}{2}$.



Since $a \neq 0$ the characteristic projections crosses the x-axis exactly once. An *initial condition* $u(x,0) = f(x)$, $x \in \mathbb{R}$, obtained by setting $t = 0$ is an auxiliary condition along the x-axis (a curve) that intersects each characteristic projection exactly once. This is important because it means that the information from the initial condition at two different points in x will correspond to two unique characteristic projections that do not intersect and hence we don't have issues with multiple potential solutions such as was observed for the inviscid Burger's equation.

For a C^1 function f the solution of the PDE for the initial condition f is

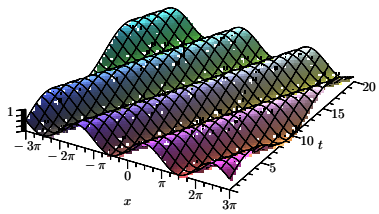
$$u(x,t) = f(x-at)$$

as can be checked by verification:

$$u_t + au_x = -af'(x-at) + af'(x-at) = 0.$$

Dynamically the profile given by $f(x)$ propagates to the right along the parallel characteristic projections at the speed a in time.

Here is the graph of $u(x,t) = \sin(x - 0.5t)$, the solution of $u_t + (\frac{1}{2})u_x = 0$, $u(x,0) = \sin(x)$.



This is all fine and dandy if we are allowing $x \in \mathbb{R}$, but what is the spatial domain if $x \in [0, L]$ for a finite $L > 0$?

We can continue to specify an initial condition

$$u(x, 0) = f(x), \ 0 \leq x \leq L$$

for a C^1 function $f(x)$, and this profile will propagate to the right along the parallel characteristic projections at the speed $a > 0$ in time, disappearing through the right boundary $x = L$ at time $t = \frac{L}{a}$.

The vanishing time $\frac{L}{a}$ of the information from the initial condition is obtained via the intersection of the characteristic projection $x - at = 0$ that emanates from $(0, 0)$, with the vertical line $x = L$. But what happens for $u(x, t)$ at those points (x, t) that lie above the characteristic projection

$$x - at = 0?$$

There is no "information" (yet) for the characteristic projections to propagate above this characteristic projection.

To remedy this situation requires a *boundary condition*, another auxiliary condition along the vertical axis (another curve),

$$u(0, t) = g(t), \ t \geq 0,$$

for a C^1 function g that satisfies the consistency condition,

$$g(0) = f(0).$$

Remark 8.4.2. We have illustrated in the previous Example the need for both initial and boundary conditions to form a well-posed solution of a PDE. In the above Example if the value of the speed a were negative, what would change in the required auxiliary conditions for a well-posed solution?

We next exhibit a scalar PDE for which some choices of an auxiliary condition can not result in a well-posed solution of the PDE.

Example 8.4.3. The characteristic differential equations for the first-order linear PDE

$$u_x + yu_y = u^2$$

are

$$\frac{dx}{ds} = 1, \ \frac{dy}{ds} = y, \ \frac{dz}{ds} = z^2.$$

Combining the first two by eliminating the "dummy" variable s by the Chain Rule gives

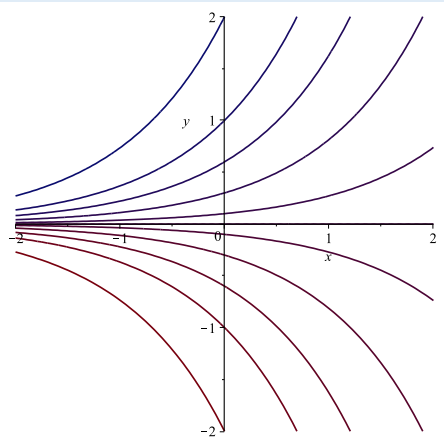
$$\frac{dy}{dx} = \frac{dy}{ds}\frac{ds}{dx} = y.$$

The general solution of this first-order ODE is

$$y = c_1 e^x \text{ or } c_1 = ye^{-x}$$

for an arbitrary constant c_1.

Here is a plot of some of these characteristic projections.



We combine $\frac{dx}{ds} = 1$ and $\frac{dz}{ds} = z^2$ by the chain rule to get

$$\frac{dz}{dx} = \frac{dz}{ds}\frac{ds}{dx} = z^2.$$

Solving

$$\frac{dz}{dx} = z^2 \ \Rightarrow \ \frac{dz}{z^2} = dx \ \Rightarrow \ -\frac{1}{z} = x + c_0$$

gives

$$z = -\frac{1}{x + c_0}$$

for a constant c_0. The constant c_0 is unique to each characteristic curve and so is connected with the constant c_1 from the characteristic projection by way of a C^1 function $g(c_1) = c_0$. Thus

$$u(x, y) = z = -\frac{1}{x + c_0} = -\frac{1}{x + g(c_1)} = -\frac{1}{x + g(y\exp(-x))}.$$

Notice that each characteristic projection crosses the vertical axis exactly once.

An auxiliary condition along the y-axis,

$$u(0, y) = f(y)$$

requires that

$$f(y) = -\frac{1}{0 + g(y\exp(-0))} = -\frac{1}{g(y)}.$$

This implies that

$$g(y) = -\frac{1}{f(y)}$$

and so the auxiliary condition $u(0, y) = f(y)$ gives the well-posed solution

$$u(x, y) = -\frac{1}{x - 1/f(y\exp(-x))}.$$

What if instead we try an auxiliary condition along the x-axis,

$$u(x, 0) = h(x)?$$

- The first issue is that the x-axis is a characteristic projection; so no characteristic projection crosses the x-axis exactly once.
- A second issue is that the value of u along the x-axis changes according to $\frac{dz}{dx} = z^2$, so when we seek to match

  $$u(x, 0) = -\frac{1}{x + g(0)}$$

  with $h(x)$, we encounter the problem that only certain (singular) choices of $h(x)$ are permitted (singular meaning there is a vertical asymptote at $x = -g(0)$).

Hence imposing the auxiliary condition $u(x, 0) = h(x)$ gives what is called an ill-posed problem.

Remark 8.4.4. The choice of auxiliary condition(s) to form a well-posed solution of PDE depends on a knowledge of the characteristic projections. Hopefully as we have demonstrated by these examples that the notion of the existence and uniqueness of solutions for PDEs with auxiliary conditions is a much more subtle problem than it was for ODEs with initial conditions.

We illustrate next the usefulness of characteristic projections when there are three independent variables. The characteristic projections are curves in three-dimensional space where it becomes more difficult to visualize their geometric nature, but they still yield valuable information about the possible auxiliary conditions that may lead to well-posed solutions. We will see that the possible auxiliary conditions for three independent variables are conditions on surfaces, the 'nicest' of these surfaces being planes.

Example 8.4.5. For a nonzero constant a, the characteristic differential equations for

$$u_t + xu_x - au_y = -u$$

are

$$\frac{dt}{ds} = 1, \ \frac{dx}{ds} = x, \ \frac{dy}{ds} = -a, \ \frac{dz}{ds} = -z.$$

By the Chain Rule we use $\frac{dt}{ds} = 1$ to reparameterize the remaining ODEs

$$\frac{dx}{dt} = x, \ \frac{dy}{dt} = -a, \ \frac{dz}{dt} = -z.$$

Solving the first two we obtain the characteristic projections

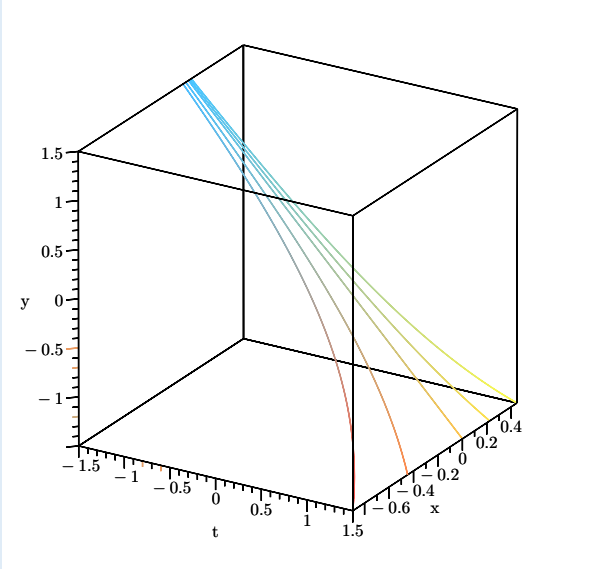
$$x = c_1 e^t, \ y = -at + c_0, \ t \in \mathbb{R}.$$

The characteristic projections are the curves

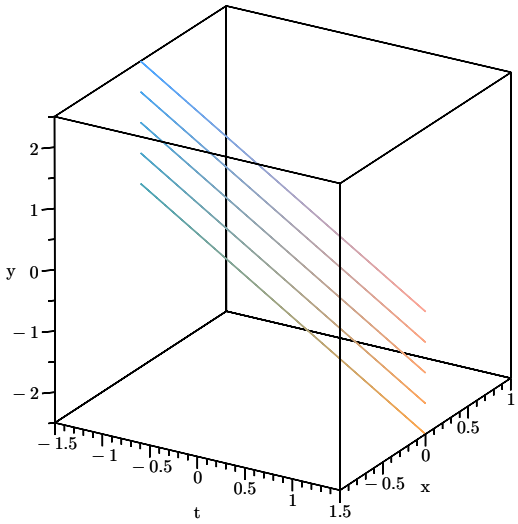
$$t \mapsto (t, c_1 e^t, -at + c_0)$$

in txy-space for each choice of the pair of constants (c_0, c_1).

Here is a graph of some of these characteristic curves when $c_0 = 0$ (and $a = 1$).



These characteristic curves lie in a plane $t + y = 0$ (for general $a \neq 0$ this would be the plane $at + y = 0$). Here is a graph of some of these characteristic curves when $c_1 = 0$ (and $a = 1$).



These characteristic curves lie in a vertical plane $x = 0$ (for general $a \neq 0$ this would still be the plane $x = 0$ with the slope of the characteristic lines changing).

What you will notice about the graphed characteristic projections is that each one crosses the vertical plane $t = 0$ exactly once. This should make you feel good inside, i.e. characteristic projections passing through a certain surface exactly once for each projection is a good thing. We can verify this behavior for the arbitrary characteristic projection $(t, c_1 e^t, -at + c_0)$ because its first component equals 0 if and only if $t = 0$, meaning it intersects the plane $t = 0$ exactly once, and its derivative

$$\frac{d}{dt}(t, c_1 e^t, -at + c_0)\Big|_{t=0} = (1, c_1 e^t, -a)\Big|_{t=0} = (1, c_1, a)$$

is not in the vertical plane $t = 0$, so each characteristic projection crosses the vertical plane $t = 0$.

Solving

$$\frac{dz}{dt} = -z$$

gives

$$z = c_2 e^{-t}$$

for an arbitrary constant c_2 is that unique to the characteristic curve. The value of c_2 is determined by an arbitrary C^1 function g of the two arbitrary constants c_0 and c_1 that determined the characteristic projection. Thus we obtain

$$u(t, x, y) = z = c_2 e^{-t} = g(c_0, c_1) e^{-t} = g(xe^{-t}, y + at) e^{-t}$$

where we have expressed

$$x = c_1 e^t \text{ and } y = -at + c_0$$

as

$$c_1 = xe^{-t} \text{ and } c_0 = y + at.$$

Can we use the initial condition

$$u(0, x, y) = f(x, y)$$

as an auxiliary condition? We can because the each characteristic projection crosses the vertical plane $t = 0$ exactly once.

Evaluation of the general form of u when $t = 0$ gives

$$u(0, x, y) = g(xe^{-0}, y + a(0)) = g(x, y),$$

so that we can take

$$f(x, y) = g(x, y).$$

Then a well-defined solution of the initial value problem is

$$u(t, x, y) = f(xe^{-t}, y + at)e^{-t}.$$

Summarizing

- The auxiliary condition was specified on a surface, in this case an initial condition on the $t = 0$ plane.
- Furthermore, this auxiliary condition gave a well-posed solution.

What auxiliary conditions would be needed if the spatial xy-domain was restricted to say the unit square $[0, 1] \times [0, 1]$ to get a well-posed solution? This would require a more detailed understanding of the nature of the characteristic projections in the slab

$$(t, x, y) \in \mathbb{R} \times [0, 1] \times [0, 1],$$

which can get rather complicated.

Remark 8.4.6. The method of characteristics is a technique that can be used to determine an appropriate set of auxiliary conditions to obtain a well-posed solution of a PDE.

The method of characteristics is a powerful approach to determining the correct type of boundary condition that makes sense. This is still an issue of direct relevance to current modeling efforts in the atmospheric sciences as discussed in the following example:

Example 8.4.7. Climatological and meteorological models are often designed to refine their spatial resolution in regions of particular interest. For instance, if an 'accurate' forecast is desired for the Hawaiian islands, it is not necessary to have a very refined grid throughout the entire Pacific which would be a substantial source for computational expense. Instead, a very rough grid is taken over the part of the Pacific away from Hawaii and then the grid is slowly refined to a finer grid as it gets closer to Hawaii. The model must then solve the same set of equations (sets of PDEs at the most fundamental level) on these refined grids.

The problem arose when the scientists prescribed boundary conditions or set values on the border of the refined grid (values which came from the coarser grid overlaying the refined grid). Without considering where the characteristics were pointing, boundary conditions were prescribed all around the boundary of the refined mesh, leading to an overly-prescribed set of auxiliary conditions and an ill-posed problem. Naturally the numerical calculations didn't make any sense and were highly inconsistent from one computation to another. (Need to get the reference from Joe Tribbia and Roger Temam)

Another, perhaps simpler example is if we return to the traffic flow problem that introduced this Chapter. If we are modeling the northbound flow of traffic on I-15 through Utah Valley, then does it make sense for us to have a boundary condition specifying the number of cars at the northernmost point in the valley? I hope not, because hopefully since we are on the northbound side of the freeway, traffic will only be flowing northward, and hence the information obtained by prescribing the traffic density at the northernmost point of our domain will have no consequence on the traffic density to the south (this is of course valid unless there is a traffic jam, but we won't worry about that yet). The characteristics of the traffic flow problem indicate that it only makes sense to prescribe the in-flow of traffic (on the northbound side) on the southern end of our domain.

The same principle applies for flow of a river. For example, if you wanted to predict the velocity of the Mississippi River on the Louisiana-Mississippi border on Wednesday, would it make the most sense to take measurements of the river at its head in the Gulf of Mexico on Tuesday, or to measure the velocity in St. Louis on Tuesday?

Remark 8.4.8. This most recent example is illustrative because one way of thinking of characteristics is that they dictate the flow or movement of information. For instance, a river only flows in one direction, and characteristics carry information (from initial or boundary conditions) in one direction. In higher dimensions this is of course more complicated, but the basic idea remains.

This discussion leads us to a definition that we will return to later on, but is extremely useful when dealing with mathematical problems that have realistic applications.

Definition 8.4.9. *A PDE with auxiliary conditions is well-posed if*

(i) a solution of the PDE exists that satisfies the auxiliary conditions,

(ii) that solution is unique, and

(iii) that solution depends continuously on the auxiliary conditions.

Otherwise the PDE with the auxiliary conditions is ill-posed.

Clearly whenever characteristics are available for a given PDE, they play a vital role in determining whether or not a problem is well-posed. As we will see, they are not the only method that is used to investigate well-posedness though.

8.5 Numerical methods for hyperbolic problems

We introduce some of the simplest approaches to discretizing PDEs using finite differences. With our understanding of the effects of characteristic projections and their influence on well-posedness of solutions, we can construct numerical methods that accurately capture the appropriate properties of the PDE.

8.5.1 Characteristics and numerical methods

We note at this point that there is a clever connection between the method of characteristics and stability/convergence of numerical methods that is directly related to the concept of auxiliary conditions. This may have been apparent from the use of Δt in Example 8.3.9. For instance, if we consider the scalar advection equation

$$u_t + au_x = 0,$$

with constant velocity a, and apply a first-order finite difference scheme in both space and time then we obtain the following discretization of the original PDE:

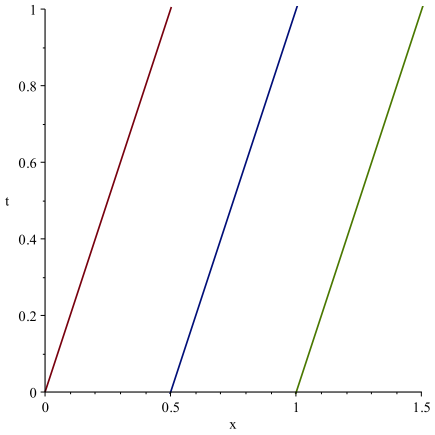
$$\frac{u(x,t+\Delta t)-u(x,t)}{\Delta t} + a\frac{u(x+\Delta x,t)-u(x,t)}{\Delta x} = 0, \tag{8.9}$$

which leads to the update equation

$$u(x,t+\Delta t) = u(x,t) - \frac{a\Delta t}{\Delta x}\left(u(x+\Delta x,t)-u(x,t)\right). \tag{8.10}$$

At first glance this seems like a plausible numerical scheme. We see below that this is problematic however as it has compatibility issues with the characteristic projections when $a > 0$.

Here is a graph of some of the characteristic projections when $a = \frac{1}{2}$.



Recall that the characteristic differential equations are

$$\frac{dx}{ds} = a,\ \frac{dt}{ds} = 1,\ \frac{dz}{ds} = 0.$$

Elimination of s via the Chain rule gives

$$\frac{dx}{dt} = a \text{ and } \frac{dz}{dt} = 0.$$

The first one integrates to give the characteristic projections

$$x = at + c \text{ or } x - at = c.$$

The characteristic projections are lines in the xt-plane with slope $\frac{1}{a}$.

- To stay on a characteristic projection by changing x to $x+\Delta x$ requires changing t to $t+\Delta t$ where $\Delta t = (\Delta x)/a$. Going backwards to stay on a characteristic projection from the point $(x, t+\Delta t)$ requires changing x to $x-\Delta x$ and $t+\Delta t$ to t where $(\Delta x)/a = \Delta t$ or $\Delta x = a\Delta t$, i.e.,
$$u(x, t+\Delta t) = u(x - a\Delta t, t).$$
- In the update formula given above, the value of $u(x, t+\Delta t)$ is determined by the value of $u(x,t)$ and $u(x+\Delta x, t)$. This actually isn't feasible because as just noted the true solution value of $u(x, t+\Delta t)$ is dependent on x values to the left of x at time t, i.e. the information at $(x+\Delta x, t)$ has no influence on the solution at $(x, t+\Delta t)$ if $a > 0$.
- Physically we can reason through this because if $u(x,t)$ is the concentration of a pollutant at position x in a pipe at time t and it is advecting or transporting to the right at speed $a > 0$, then to numerically compute its concentration at time $t+\Delta t$ we should have information on the solution for points to the left of x at time t.
- On the other hand if $a < 0$, then the characteristic projections go to the left, and the update formula for $u(x+\Delta t, t)$ in terms of $u(x,t)$ and $u(x+\Delta x, t)$ makes more sense; it looks in the "correction direction" consistent with the advection or transfer to the left.

How do we fix the update formula when $a > 0$? For $u_x(x,t)$ we use the finite difference
$$u_x(x,t) \approx \frac{u(x,t) - u(x-\Delta x, t)}{\Delta x}.$$
With this, the discretization of the PDE becomes
$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} + a\frac{u(x,t) - u(x-\Delta x,t)}{\Delta x} = 0,$$
which gives the update formula
$$u(x,t+\Delta t) = u(x,t) - \frac{a\Delta t}{\Delta x}\big(u(x,t) - u(x-\Delta x,t)\big). \tag{8.11}$$
This update formula looks in the "correct direction" with the advection or transfer to the right.

For the first-order advection PDE
$$u_t + a(x,t)u_x = 0$$
where $a(x,t)$ may change sign, the algorithm chooses the "correct direction" at each time step, either the update formula (8.10) when $a(x,t) < 0$ or the update formula (8.11) when $a(x,t) > 0$. This is known as the "upwind method."

Remark 8.5.1. All of the previous discussion was a qualitative introduction to the influence/effect that characteristics have on the derivation of numerical methods for hyperbolic problems. We formalize these ideas a bit further in the following example.

Example 8.5.2. For the one-dimensional advection PDE
$$u_t + au_x = 0$$
with $a > 0$ the characteristic projections are
$$x - at = c$$

and the upwind update formula for the numerical scheme is

$$u(x+\Delta x,t) = u(x,t) - \frac{a\Delta t}{\Delta x}\big(u(x,t) - u(x-\Delta x,t)\big).$$

Solutions of the PDE are constant along the characteristic projections

$$x = at + c_0.$$

The value of $u(x, t+\Delta t)$ is determined by the characteristic projection

$$x = a(t+\Delta t) + c_0.$$

Traveling back along this characteristic projection the point in space-time that dictates the solution at $u(x, t+\Delta t)$ at time t is the point

$$(x-\Delta x, t) \text{ where } \Delta x = a\Delta t,$$

that is this point satisfies

$$(x-\Delta x) - at = (x - a\Delta t) - at = x - a\Delta t - at = x - a(t+\Delta t) = c_0.$$

For the numerical scheme to have a chance to accurately approximate the value of $u(x, t+\Delta t)$ the spatial difference should "cover" the value of u at the point $(x - a\Delta t, t)$, i.e., the spatial difference should interpolate the value of u by the value of u at points on both sides of $(x - a\Delta t, t)$ for fixed t.

The upwind scheme uses the value of u at the two points (x, t) and $(x-\Delta x, t)$ to approximate the value of u at $(x, t+\Delta t)$. Using the observation above, we arrive at the inequalities for the three points,

$$x - \Delta x < x - a\Delta t < x.$$

Subtracting x from all three gives

$$-\Delta x < -a\Delta t < 0.$$

Dividing through by the negative $-\Delta x$ gives

$$0 < \frac{a\Delta t}{\Delta x} < 1.$$

This inequality is known as the Courant-Friedrichs-Lewy (CFL) condition; satisfying it leads to stable numerical schemes.

Remark 8.5.3. Violations of the CFL condition leads to unstable schemes and wild numerical outputs.

Remark 8.5.4. The value of $a > 0$ is not a choice we make; it is part of the PDE that typically arises through some empirical data. The values of Δx and Δt are choices we make when implementing the upwind numerical scheme.

- For fixed choices of Δx and Δt, if the CFL condition is satisfied, i.e., $\frac{\Delta t}{\Delta x} < \frac{1}{a}$, then there is a triangle formed by the three points $(x-\Delta x, t)$, (x, t) and $(x, t+\Delta t)$, where the last two points have the same first component so lie on the same vertical line, and the characteristic line passes through the bottom of the triangle (the line between the two points with same second component value) and exits through the vertex at the top.

- If the CFL condition is violated, i.e., $\Delta t/\Delta x \geq \frac{1}{a}$, then the characteristic line does not pass *through* the triangle formed by the three points $(x - \Delta x, t)$, (x, t), and $(x, t + \Delta t)$, it only intersects the triangle at the top vertex. In this case the numerical method is unstable

8.5.2 Actual implementation of these methods

To simplify matters, we will consider a periodic (in the spatial coordinate x) advection equation, i.e.

$$u_t + au_x = 0, \tag{8.12}$$

where a is a constant (at least for now), and $x \in [-L, L]$ where $u(-L) = u(L)$, and some initial condition is prescribed as $u(x, 0) = u_0(x)$. Application of the upwind method is pretty straightforward in this setting, with the primary difficulty dealing with indexing to capture the periodicity correctly. We demonstrate one such approach, but there are actually several different ways to enforce the periodic boundary conditions, some more efficient than others. An example implementation when $a > 0$ is shown in Algorithm 8.1.

There are several things worth noting about Algorithm 8.1 that we focus on right now:

- The initial condition and spatial discretization is a required input. This dictates the value of Δx and sets up the spatial domain (which is assumed to be $[-L, L]$).
- It is important to ensure that the initial condition $u0$ satisfies the periodic boundary conditions. What happens if it doesn't (try this out)?
- Rather than specifying the time step Δt this algorithm specifies a CFL number, and then the time step is adjusted to fit that value. There are of course other ways to handle the choice of time step/CFL number, but this is one frequent approach.
- Because this Algorithm assumes that $a > 0$ then the periodic boundary conditions only need to be handled at the left-most point (index $jj = 0$). How could you set this up if we don't assume that $a > 0$?
- This approach also introduces a *for* loop to handle the upwind scheme at each time-step. Is this necessary, or is it possible to handle this via array indexing or some other more clever approach?
- Is the use of the dummy variable v really necessary?

Of course, practical considerations must be taken into account for non-periodic boundaries, but even then we need to recognize the role that characteristics will play on the choice of boundary conditions. In addition, modifying this Algorithm to work for non-constant a is an important and certainly non-trivial task. In addition, we can modify this without to much effort to work for a nonlinear equation such as Burger's equation, but would we really expect it to work in that case?

Another modification that we won't get into here is extension of the upwind method to higher dimensions. We can also search for a higher order method by simply using higher order finite differences in either the approximation of the time derivative or the spatial derivative. A more practical approach is taken in Example 8.5.5 that considers the error from both types of approximations simultaneously.

To seek for higher order methods we take an approach that replaces partial derivatives of u with respect to t with partial derivatives of u with respect to x.

Example 8.5.5 (Lax-Wendroff). If a solution $u(x, t)$ of the one-dimensional advection equation

$$u_t + au_x = 0, \ a \neq 0,$$

```
import numpy as np
from matplotlib import pyplot as plt
from matplotlib import animation as ani

def advect_upwind(a,CFL,T,x=[],u0=[]):
    """solution of the 1D advection equation u_t+au_x=0.
    a>0 is assumed to be a positive constant
    CFL is the desired CFL number
    T is the final time that the simulation will be run
    x is the pre-selected spatial variable, and should be uniformly distributed
    u0 is the initial condition, and must be the same size as x

    Returns an array u which is the solution at all time steps up to T
    as dictated by the time-step chosen through the specified CFL number."""

    N = len(x)
    L = x[-1]
    deltaX = 2*L/(N-1)
    deltaT=CFL*deltaX/a
    u = []
    u.append(u0)
    v = u0.copy() #a dummy variable for the update process
    for nn in range(1,int(T/deltaT)):
        for jj in range(1,N):
            v[jj] = u[nn-1][jj] - CFL*(u[nn-1][jj] - u[nn-1][jj-1])
        #Now we handle the periodic BCs
        v[0] = u[nn-1][0] - CFL*(u[nn-1][0] - u[nn-1][-1])
        u.append(v.copy())

    return x,u
```

Algorithm 8.1: Python implementation of the one-dimensional upwind advection algorithm for $a > 0$.

is at least C^2, then taking the partial derivative of the PDE with respect to t gives

$$u_{tt} + au_{tx} = 0 \;\Rightarrow\; u_{tt} = -au_{tx},$$

and taking the partial derivative of the PDE with respect to x gives

$$u_{tx} + au_{xx} = 0.$$

Multiplying the latter through by a gives

$$au_{tx} + a^2 u_{xx} = 0 \;\Rightarrow -au_{tx} = a^2 u_{xx},$$

so that it combines with the former to give

$$u_{tt} = -au_{tx} = a^2 u_{xx}.$$

Note that if we assume that u is C^k we can obtain relations among the k^{th} order partial derivatives of u which as we see below would give us even higher order numerical methods for solving the advection equation.

Using the PDE $u_t + au_x = 0$, i.e., $u_t = -au_x$, the concomitant relation $u_{tt} = a^2 u_{xx}$, the centered finite differences

$$u_x(x,t) = \frac{u(x+\Delta x,t) - u(x-\Delta x,t)}{2\Delta x},$$

and

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

we Taylor expand our desired solution at $(x, t+\Delta t)$:

$$\begin{aligned} u(x,t+\Delta t) &= u(x,t) + (\Delta t)u_t(x,t) + \frac{(\Delta t)^2}{2}u_{tt}(x,t) + O((\Delta t)^3) \\ &= u(x,t) - a(\Delta t)u_x(x,t) + a^2\frac{(\Delta t)^2}{2}u_{xx}(x,t) + O((\Delta t)^3) \\ &= u(x,t) - \frac{a(\Delta t)}{2\Delta x}\big(u(x+\Delta x,t) - u(x-\Delta x,t)\big) \\ &\quad + \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\big(u(x+\Delta x) - 2u(x,t) + u(x-\Delta x,t)\big) + O((\Delta t)^3). \end{aligned}$$

Ignoring the $O((\Delta t)^2)$ and higher order terms this gives the update formula

$$u(x,t+\Delta t) = u(x,t) + \frac{a\Delta t}{2\Delta x}\big(u(x+\Delta x,t) - u(x-\Delta x,t)\big),$$

in which the spatial derivative u_x is approximated by a central finite difference.

Keeping the $O(\Delta t)^2$ terms and ignoring the higher order terms in the above Taylor expansion gives the update formula

$$\begin{aligned} u(x,t+\Delta t) = u(x,t) &- \frac{a(\Delta t)}{2\Delta x}\big(u(x+\Delta x,t) - u(x-\Delta x,t)\big) \\ &+ \frac{a^2(\Delta t)^2}{2(\Delta x)^2}\big(u(x+\Delta x) - 2u(x,t) + u(x-\Delta x,t)\big), \end{aligned}$$

which is best known as the Lax-Wendroff method for the advection equation.

The Lax-Wendroff and upwind methods are definitely not the only possibilities for hyperbolic problems, but they are the most frequently used. In fact most hyperbolic systems are well approximated using a mix of the upwind and Lax-Wendroff methods via a technique most commonly referred to as finite volume methods.

Remark 8.5.6. It is worth noting that this entire section is devoted to the advection equation, which is most certainly not the most interesting hyperbolic PDE. In fact, one may be justified in saying the advection equation is a rather boring example to spend so much time on. This is certainly true, particularly in one dimension. It turns out though that most hyperbolic PDEs can either be linearized to appear like the advection equation, or the method of characteristics can be used to effectively force the system to appear like a series of advection equations in different directions (based on eigen-directions of the Jacobian). We won't go into such details here as it is actually quite complicated how this is done in practice, but it is definitely possible to extend the ideas presented here for the advection equation to the multi-dimensional and nonlinear setting, leading to what are commonly called finite volume methods.

8.5.3 *Errors in numerical approximations and modified equations

If you let the upwind advection solver run for a sufficiently long time, what should happen to the solution? The initial state should be translated according to the velocity a, but the actual solution shouldn't change at all. Using the upwind scheme however, you will notice rather quickly that this isn't the case, and in fact the quality of the solution degrades the longer the simulation is run. To see why this is the case, we once again return to Taylor series applied to the upwind scheme (assuming that $a > 0$ for now). First, though we clarify that in this case we are considering a function $v(x,t)$ that satisfies the upwind scheme exactly, and then we compute a Taylor series (in both Δt and Δx) to see what differential equation that it most closely approximates.

$$v(x,t+\Delta t) = v(x,t) - \frac{a\Delta t}{\Delta x}\left(v(x,t) - v(x-\Delta x,t)\right) \tag{8.13}$$

$$\Rightarrow v_t(x,t) + \frac{\Delta t}{2}v_{tt}(x,t) + \frac{(\Delta t)^2}{6}v_{ttt}(x,t) + O((\Delta t)^3) \tag{8.14}$$

$$+ av_x(x,t) - \frac{a\Delta x}{2}v_{xx}(x,t) + \frac{a(\Delta x)^2}{6}v_{xxx}(x,t) + O((\Delta x)^3) = 0. \tag{8.15}$$

Retaining only those terms up to $O(\Delta t)$ we see that $v(x,t)$ satisfies:

$$v_t(x,t) + av_x(x,t) = \frac{1}{2}(a\Delta x v_{xx}(x,t) - \Delta t v_{tt}(x,t)), \tag{8.16}$$

from which we can also derive

$$v_{tt}(x,t) = -av_{xt}(x,t) + \frac{1}{2}\left(a\Delta x v_{xxt}(x,t) - \Delta t v_{ttt}(x,t)\right), \tag{8.17}$$

$$v_{tx} = -av_{xx}(x,t) + \frac{1}{2}\left(a\Delta x v_{xxx}(x,t) - \Delta t v_{ttx}(x,t)\right), \tag{8.18}$$

$$\Rightarrow v_{tt}(x,t) = a^2 v_{xx}(x,t) + O(\Delta t). \tag{8.19}$$

Inserting this back into the previously derived relation, and dropping the $O((\Delta t)^2)$ terms we arrive at

$$v_t(x,t) + av_x(x,t) = \frac{a\Delta x}{2}\left(1 - \frac{a\Delta t}{\Delta x}\right)v_{xx}(x,t). \tag{8.20}$$

This equation is referred to as the modified equation for the upwind method, and indicates what exact solutions to the upwind scheme more accurately approximate.

Hence, while the upwind scheme is a $O(\Delta t)$ accurate approximation to the advection equation it is actually an $O((\Delta t)^2)$ approximation to (8.20). As we see later on, this explains they type of error that we see from the upwind method. Numerically we can observe that the upwind method has what we refer to as a diffusive error, in that the solution is diffused for all $t > 0$, i.e. the general shape of the solution remains correct, but the amplitude of the initial state is diffused toward zero.

On the other hand, the Lax-Wendroff method has a very different type of error that is introduced. If we go through a similar process, we find that solutions of the Lax-Wendroff scheme more closely approximate

$$v_t(x,t) + av_x(x,t) + \frac{a(\Delta x)^2}{6}\left(1 - \left(\frac{a\Delta t}{\Delta x}\right)^2\right) v_{xxx}(x,t) = -\varepsilon v_{xxxx}(x,t), \tag{8.21}$$

where $\varepsilon > 0$ so long as Δx and Δt are chosen appropriately. As we see below, this is a dispersive equation at the highest order, which means that dispersive errors are introduced to the approximation of the advection equation. These errors appear in the form of additional waves in the solution that are not physically present, however the amplitude of the initial state is preserved.

8.6 Equation of motion for a vibrating string, i.e. the wave equation

Beyond the advection equation, one of the most fundamental hyperbolic PDEs is the wave equation which we introduce in this section. It is of fundamental importance both because it serves as a canonical example of the nature of hyperbolic PDEs, but also because it accurately represents physically interesting and relevant phenomena that occur everywhere around us. We start the introduction to this equation following a physical derivation, but emphasize that the basic principles espoused by this equation are applicable too many other phenomena than just playing the guitar (although honestly understanding how a guitar creates music should be enough motivation by itself).

Remark 8.6.1. In the spirit of using remarks as a form of venting, we will now comment on the value of modeling something just because you find it interesting. Breakthroughs in Applied mathematics are often motivated by problems of great scientific and practical interest. Sometimes though, such breakthroughs are the result of a curious mathematician wondering what the optimal rate of flipping a hamburger will best cook the meat for optimal flavor (without over-cooking). In either case it is important to recognize that the motivation for a specific problem may not be the most important part of the problem.

Recall that the second-order PDE $u_{tt} = c^2 u_{xx}$, or

$$u_{tt} - c^2 u_{xx} = 0$$

is hyperbolic or wave-like because its discriminant

$$D = 0^2 - 4(-c^2) = 4c^2 > 0.$$

Recall also that some solutions of this hyperbolic PDE are

$$u(x,t) = \sin(n\pi x)\cos(cn\pi t),\ n \in \mathbb{N},$$

whose graphs are bounded and are wave-like in the x direction and wave-like in the t direction (Amazing...wave-like solutions for the wave equation!)

8.6.1 Physically motivated derivation

Consider a homogeneous, flexible, elastic string that undergoes small, transverse vibrations. For example a violin or guitar string that is plucked perpendicular to its rest placement (a good example of this is at `https://www.youtube.com/watch?v=8YGQmV3NxMI`). Let $u(x,t)$ be the displacement of the string from equilibrium at time t and position x.

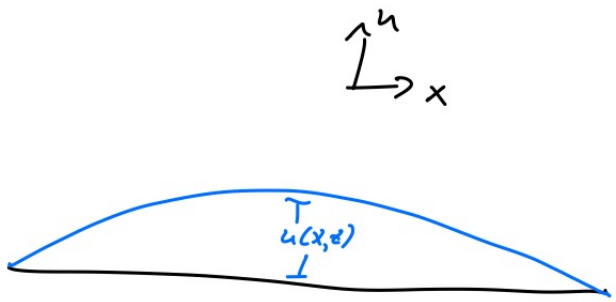


Figure 8.1: An illustration of a plucked string. The horizontal direction is x and the vertical is $u(x,t)$. The black line is where the string would be at rest (equilibrium), and the blue curve is after the string is plucked and released.

Remark 8.6.2. This description and the derivation that follows give you a lot of words to google: *homogeneous, flexible, elastic, and transverse.* These are all important terms, but unlike any appropriate math textbook, we will not dwell on their definition because they are physical adjectives for the string, and not actually mathematically defined objects.

We secure the infinitely long string into its resting position by pulling it (so it is taut) and securing its ends at pegs located at two points infinitely far apart.

We make several more (un)reasonable physical assumptions about this infinitely long string:

- the string is made of a perfectly homogeneous material,
- the string is infinitely flexible, and
- the string is perfectly elastic.
- We assume further that the string moves in a fixed vertical xu-plane with the x-axis as the resting position of the string.
- We let $u(x,t)$ denote the transverse or vertical displacement (in the direction u) of the string at position x at time t, with the assumption that the point x of the string moves only vertically (no horizontal displacement of the string).
- If the string is perfectly flexible then the tension $T(x,t)$ (the difference in the tension between two points will indicate the force over that interval) will be tangential to the string.
- Since the string is homogeneous (meaning there are no changes in its properties throughout the entire length of the one-dimensional object), the density ρ (mass per unit length) is constant.

The mass of the string between any two points $x_1 < x_2$ of the string is

$$m = \int_{x_1}^{x_2} \rho \, dx = \rho(x_2 - x_1).$$

Multiplying the constant density ρ by the acceleration u_{tt} gives ρu_{tt} whose units are

$$\text{mass per length} \times \text{length per time}^2 = \text{mass per time}^2.$$

Thus the integral

$$\int_{x_1}^{x_2} \rho u_{tt}\, dx \tag{8.22}$$

has units

$$\text{mass length per time}^2$$

which are the units of force.

Now assuming all of these assumptions, we want to use Newton's second law applied to any part of the string between two arbitrary points x_1 and x_2. This law states that force=mass×acceleration.

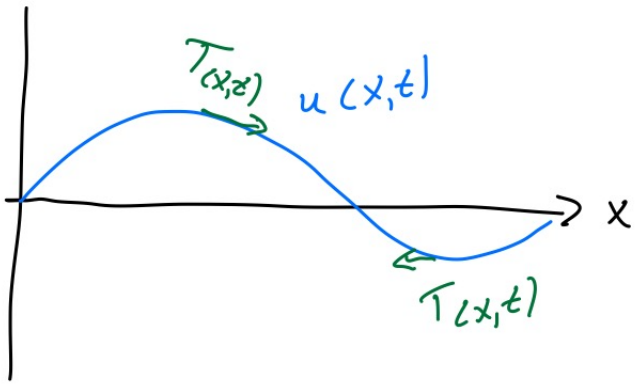


Figure 8.2: A sketch of a plucked string that includes the tension.

Before proceeding we note that in an infinitesimal setting we can refer to the slope of the string as $u_x(x,t)$, i.e. the x derivative of $u(x,t)$ can be used to determine the shape of an infinitesimal piece of the string. We need to consider Newton's second law applied in both the vertical and the horizontal direction, i.e. there must be a balance between the force and the acceleration in both components of the possible motion.

Again before we proceed, we note that because the slope of the tangent line to $u(x,t)$ is given by $u_x(x,t) = u_x(x,t)/1$ then this will describe a right triangle depicted in Figure 8.3 where the 'rise' is u_x and the 'run' is 1. Because the tension is tangent to the curve $u(x,t)$ then the tension $T(x,t)$ will be on the hypotenuse of a right triangle similar to this one (with the same angle θ).

To identify the sum of the forces, we need to break the tension up into horizontal and vertical components. Using the similar triangle depicted in Figure 8.3 we see that the horizontal component of the tension is

$$T\cos\theta = \frac{T}{\sqrt{1+u_x^2}},$$

and the vertical component of the tension will be

$$T\sin\theta = \frac{Tu_x}{\sqrt{1+u_x^2}}.$$

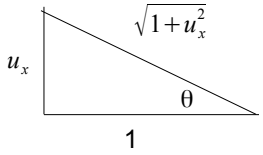


Figure 8.3: The similarity triangle used to describe the geometry of a vibrating string.TODO: this figure needs help

We are neglecting any longitudinal (horizontal) motion, which means from Newton's second law that the sum of the forces in the horizontal direction must be zero. The force in this direction is equal to the difference in the tension along an interval $[x_1, x_2]$ of the string. This indicates that

$$\left.\frac{T(x,t)}{\sqrt{1+u_x^2(x,t)}}\right|_{x_1}^{x_2} = 0. \tag{8.23}$$

We next need to apply Newton's second law to the vertical components of acceleration and force. To do this we first notice that the transverse (vertical) acceleration×mass on the interval $[x_1, x_2]$ is given by

$$\int_{x_1}^{x_2} \rho u_{tt} dx.$$

The density can also be used to compute the mass of the string in the interval $[x_1, x_2]$ as:

$$\int_{x_1}^{x_2} \rho dx.$$

It follows that Newton's second law applied to the transverse motion implies that

$$\left.\frac{T(x,t)u_x(x,t)}{\sqrt{1+u_x^2(x,t)}}\right|_{x_1}^{x_2} = \int_{x_1}^{x_2} \rho u_{tt} dx. \tag{8.24}$$

Equations (8.23) and (8.24) model the nonlinear evolution of the wave formed by a vibrating string. This is wonderful news! Note that these equations are highly nonlinear, and have a mix of integration and differentiation which is a combination of various types of nightmares all at once!

Remark 8.6.3. Note that in the derivation above, the interval $[x_1, x_2]$ was chosen arbitrarily. There is nothing special about either of these endpoints. Thus, just as in the derivation of conservation laws, these equations must hold for any values of x_1 and x_2 that lie in the domain of interest.

Now returning to the derivation, we assume that the displacement of the string in the vertical is small relative to the length of the entire string (after all any finite displacement is pretty small compared to a string tied down at $\pm\infty$), particularly over the interval $[x_1, x_2]$ in question. It follows that we are equivalently assuming that $|u_x|$ is 'small' on this interval, implying that $\sqrt{1+u_x^2} \approx 1 + \frac{1}{2}u_x^2 + \ldots$. Thus we see that the equations of motion considered above can be approximated as:

$$T(x,t)|_{x_1}^{x_2} = 0$$
$$T(x,t)u_x(x,t)|_{x_1}^{x_2} = \int_{x_1}^{x_2} \rho u_{tt} dx.$$

The first of these equations implies that $T(x,t)$ is constant between the arbitrarily chosen points x_1 and x_2, and thus it is constant throughout the entire domain of interest. Just as we did for conservation laws, we rewrite the second equation in integral form using the Fundamental Theorem of Calculus:

$$\int_{x_1}^{x_2} (Tu_x)_x dx = \int_{x_1}^{x_2} \rho u_{tt} dx,$$

and because the interval $[x_1, x_2]$ was chosen arbitrarily, then we can assume that

$$(Tu_x)_x = \rho u_{tt}$$

except on a set of measure zero, and thus assuming that T is independent of time as well as x, we can arrive at the celebrated one-dimensional wave equation

$$u_{tt} = c^2 u_{xx}, \tag{8.25}$$

where $c = \sqrt{\frac{T}{\rho}} > 0$.

8.6.2 D'Alembert's solution

We are interested in considering the characteristics of such a PDE as the wave equation, but our standard method of defining the characteristic curves does not apply here. Interestingly many of the same ideas still hold.

Theorem 8.6.4 (D'Alembert). *For the Cauchy problem (defined on the infinite domain),*

$$\begin{aligned} u_{tt} &= c^2 u_{xx}, \qquad x \in \mathbb{R}, \quad t > 0 \\ u(x,0) &= F(x) \quad \text{and} \quad u_t(x,0) = G(x), \quad x \in \mathbb{R}, \end{aligned}$$

with $F \in C^2(\mathbb{R})$ and $G \in C^1(\mathbb{R})$, the solution is given by

$$u(x,t) = \frac{1}{2}\left[F(x+ct) + F(x-ct)\right] + \frac{1}{2c}\int_{x-ct}^{x+ct} G(y)dy.$$

Proof. You show in the homework that the change of variables

$$\xi = x - ct, \ \eta = x + ct,$$

will transform the PDE $u_{tt} = c^2 u_{xx}$ into the PDE

$$u_{\xi\eta} = 0.$$

This implies that there are two arbitrary C^2 functions f and g such that

$$u(\xi, \eta) = f(\xi) + g(\eta).$$

Returning to the variables x and t we have

$$u(x,t) = f(x-ct) + g(x+ct).$$

Now we just need to identify how the arbitrary functions f and g are determined by the initial functions F and G.

Plugging the initial position $F(x)$ into the solution gives

$$F(x) = u(x,0) = f(x) + g(x).$$

Rearranging this gives

$$f(x) + g(x) = F(x). \tag{8.26}$$

Differentiating $u(x,t) = f(x-ct) + g(x+ct)$ with respect to t gives

$$u_t(x,t) = -cf'(x-ct) + cg'(x+ct).$$

Plugging in the initial velocity $G(x)$ gives

$$G(x) = u_t(x,0) = -cf'(x) + cg'(x).$$

This implies that

$$-f(x) + g(x) = \frac{1}{c}\int_{-\infty}^{x} G(y)\ dy. \tag{8.27}$$

Adding equations (8.26) and (8.27) gives

$$2g(x) = F(x) + \frac{1}{c}\int_{-\infty}^{x} G(y)\ dy.$$

Thus

$$g(x) = \frac{F(x)}{2} + \frac{1}{2c}\int_{-\infty}^{x} G(y)\ dy.$$

Subtracting equation (8.27) from equation (8.26) gives

$$2f(x) = F(x) - \frac{1}{c}\int_{-\infty}^{x} G(y)\ dy.$$

Thus

$$f(x) = \frac{F(x)}{2} - \frac{1}{2c}\int_{-\infty}^{x} G(y)\ dy.$$

With f and g thus related to F and G we see that

$$\begin{aligned} u(x,t) &= f(x-ct) + g(x+ct) \\ &= \frac{F(x-ct)}{2} - \frac{1}{2c}\int_{-\infty}^{x-ct} G(y)\ dy + \frac{F(x+ct)}{2} + \frac{1}{2c}\int_{-\infty}^{x+ct} G(y)\ dy \\ &= \frac{1}{2}\big[F(x+ct) + F(x-ct)\big] + \frac{1}{2c}\int_{x-ct}^{x+ct} G(y)\ dy. \end{aligned}$$

This gives the desired D'Alembert solution of the one-dimensional wave equation. □

Remark 8.6.5. The lines $x \pm ct =$ constant are technically characteristics for second-order quasilinear PDEs, but our previous discussion of "characteristic" curves is for first-order PDE, and the concept of "characteristic" for second-order PDEs is different. For the one-dimensional wave equation the lines $x \pm ct =$ constant come from the ODE

$$\frac{dt}{dx} = \frac{\beta \pm \sqrt{\beta^2 - \alpha\gamma}}{\alpha}$$

where the constants α, β, and γ are the coefficients in the PDE

$$\alpha u_{xx} + 2\beta u_{xt} + \gamma u_{tt} = 0.$$

For the one-dimensional wave equation $\alpha = c^2$, $\beta = 0$ and $\gamma = -1$, so that

$$\frac{dt}{dx} = \frac{0 \pm \sqrt{0 - c^2(-1)}}{c^2} = \pm\frac{1}{c} \ \Rightarrow x \pm ct = \text{constant}.$$

You will recognize the discriminant

$$\beta^2 - \alpha\gamma = 0 - c^2(-1) = c^2 > 0$$

which says that the one-dimensional wave equation is hyperbolic.

Remark 8.6.6. For nonzero nonconstant wave-like functions f and g, the form of the solution

$$u(x,t) = f(x - ct) + g(x + ct)$$

is that of two waves traveling in opposite directions along the string. The two waves will interact with each other either in a future time or did interact in a past time. We would like to think of the "characteristic" lines $x \pm ct =$ constant as transporting the wave. This is not quite correct. These "characteristic" lines always intersect because $c \neq 0$. Transporting the value of u along these intersecting lines would force u to be a constant, which it is not.

Remark 8.6.7. Although the "characteristic" lines $x \pm ct =$ constant do not transport the value of u, they play a role nonetheless in the propagation of the wave. This is what D'Alembert's solution of the Cauchy Problem for the one-dimensional wave equation says, which is demonstrated and explained below.

8.6.3 Domain of influence/ domain of dependence

The role of the "characteristic" lines $x \pm ct =$ constant appear in the domain of influence and in the domain of dependence of the solution in different times.

This is a consequence of following the "characteristic" lines forward and backwards, and the D'Alembert solution at a point (x_0, t_0) for $t_0 > 0$ which is

$$u(x_0,t_0) = \frac{1}{2}\big[F(x_0 + ct_0) + F(x_0 - ct_0)\big] + \frac{1}{2c}\int_{x_0-ct_0}^{x_0+ct_0} G(y)\ dy.$$

The "characteristic" lines

$$x + ct = x_0 + ct_0 \text{ and } x - ct = x_0 - ct_0$$

intersect at the point (x_0, t_0) on the horizontal line $t = t_0$ and intersect the horizontal line $t = 0$ at the points

$$x = x_0 - ct_0 \text{ and } x = x_0 + ct_0.$$

This says that the initial conditions $F(x) = u(x, 0)$ and $G(x) = u_t(x, 0)$ for x in the interval

$$[x_0 - ct_0, x_0 + ct_0]$$

uniquely determines the value of $u(x_0, t_0)$, i.e., this interval is the domain of dependence for $u(x_0, t_0)$ when $t = 0$.

For a value of $t_* \in (0, t_0)$ the "characteristic" lines intersect the horizontal line $t = t_*$ at the points

$$x = x_0 - c(t_0 - t_*) \text{ and } x = x_0 + c(t_0 - t_*).$$

This says that the "initial conditions" $F_*(x) = u(x, t_*)$ and $G_*(x) = u_t(x, t_*)$ for x in the interval

$$[x_0 - c(t_0 - t_*), x_0 + c(t_0 - t_*)]$$

uniquely determine the value of $u(x_0, t_0)$.

Conversely, the initial values $F(x_0)$ and $G(x_0)$, when $t = 0$, may influence the value of $u(x, t)$ for $t > 0$ whenever x belongs to the interval $[x_0 - ct, x_0 + ct]$. Since x_0 belongs to this interval, the integral

$$\int_{x_0-ct}^{x_0+ct} G(y)\ dy$$

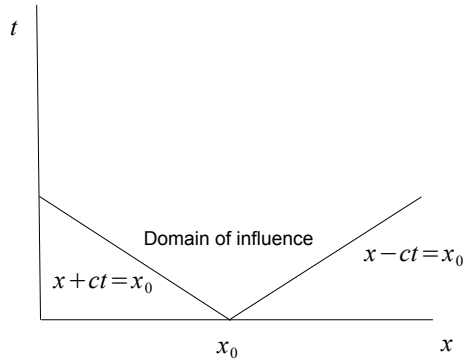


Figure 8.4: The domain of influence for the wave equation.

is influenced by the value of $G(x_0)$ (which G is assumed to be C^1). If x is an endpoint, then $x = x_0 - ct$ or $x = x_0 + ct$ so that $x_0 = x + ct$ or $x_0 = x - ct$ and $F(x_0) = F(x + ct)$ or $F(x_0) = F(x - ct)$, and $F(x_0)$ also influences the value of $u(x_0, t_0)$.

Instead if we consider the initial values $F_*(x) = u(x, t_*)$ and $G_*(x, t_*) = u_t(x, t_*)$ for some $0 < t_* < t_0$ and a point (x_*, t_*) such that the "characteristic" lines $x + ct = x_* + ct_*$ and $x - ct = x_* - ct_*$ emanating upward from the point (x_*, t_*) for which the point (x_0, t_0) is between the "characteristic" lines, then these "characteristic" lines intersect the horizontal line $t = t_0$ at the points

$$x + ct_0 = x_* + ct_* \text{ and } x - ct_0 = x_* - ct_*.$$

This gives the interval

$$[x_* - c(t_0 - t_*), x_* + c(t_0 - t_*)]$$

which contains x_0.

Thus the value of $u(x_0, t_0)$ is influenced by the value of $G_*(x_*)$ and possibly by the value of $F_*(x_*)$ (the latter if x_0 is an endpoint).

Definition 8.6.8. *The domain of influence at the point (x_0, t_0) is the region in the xt-plane for $t > t_0$ that is influenced by data at the point (x_0, t_0).*

The "characteristic" lines form the sides of the cone of influence emanating upward from the point (x_0, t_0) as shown in Figure 8.4

Definition 8.6.9. *The domain of dependence at the point (x_0, t_0) is the region in the xt-plane for $t < t_0$ that can affect the solution at the point (x_0, t_0).*

The "characteristic" lines form the sides of the cone emanating downward to the point (x_0, t_0) that intersects the horizontal line $t = t_*$ for $t_* < t_0$ in the interval of dependence. See Figure 8.5 for an illustration of these effects.

Numerical methods for the wave equation

Accounting for the domain of dependence and the domain of influence, it might be wise to use the centered finite difference approximation for both second-order derivatives in the wave equation.

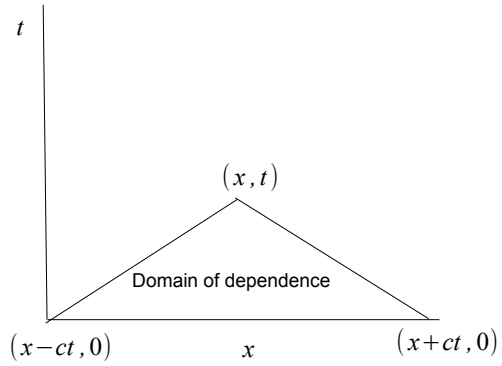


Figure 8.5: The domain of dependence for the wave equation.

This gives the discretization of the wave equation as,

$$\frac{u(x,t+\Delta t)-2u(x,t)+u(x,t-\Delta t)}{(\Delta t)^2} = c^2\frac{u(x+\Delta x)-2u(x,t)+u(x-\Delta x,t)}{(\Delta x)^2}.$$

This gives the update formula

$$\begin{aligned} u(x,t+\Delta) &= 2u(x,t)-u(x,t-\Delta t)+\frac{c^2(\Delta t)^2}{(\Delta x)^2}\big[u(x+\Delta x)-2u(x,t)+u(x-\Delta x,t)\big] \\ &= 2\left(1-\frac{c^2(\Delta t)^2}{(\Delta x)^2}\right)u(x,t)-u(x,t-\Delta t) \\ &\quad+\frac{c^2(\Delta t)^2}{(\Delta x)^2}\big[u(x+\Delta x)+u(x-\Delta x,t)\big]. \end{aligned}$$

You will notice in this update formula that to compute the value of $u(x,t+\Delta t)$ requires the values $u(x,t)$, $u(x\pm\Delta x,t)$, and $u(x,t-\Delta t)$.

Finite boundaries

What happens if the string is of finite length, say $x\in[0,L]$ for finite $L>0$? This means there are boundary conditions when $x=0$ and $x=L$.

The initial conditions $u(x,0)=F(x)$ and $u_t(x,0)=G(x)$ influence those points (x,t) contained in the triangle whose sides are the x-axis from $x=0$ to $x=L$, the "characteristic" line $x-ct=0$ emanating out of the point $(0,0)$, and the "characteristic" line $x+ct=L$ emanating out of point $(L,0)$.

For points outside this triangle, the boundary conditions influence the solution via those "characteristic" lines that intersect the boundaries.

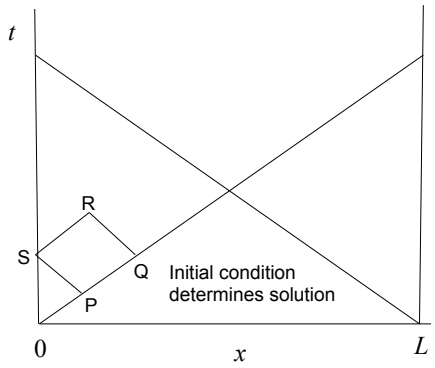


Figure 8.6: The influence of boundary conditions on the evolution of information for the wave equation.

8.7 Traveling and Standing Waves

Often in PDEs the full PDE is difficult to understand and solutions are intractable or not helpful when they are tractable, so we consider pieces of the PDE, or model it with simpler PDEs that are tractable. Traveling and standing waves are just two examples of how this is done for several classes of PDEs. To do this we make use of the ansatz method to identify these two different types of waves. The two types of ansatz each involve parameters that have physical meaning which we explain here. We illustrate each ansatz through examples.

8.7.1 Traveling waves

The discussion on traveling waves here is incorporated almost entirely by a single example. The ideas presented for this example apply generically to several other traveling waves for PDEs of varying complexity.

We have seen that function of the form $u(x,t) = U(x - ct)$, for a function U, is sometimes a solution of a PDE. The ansatz

$$u(x,t) = U(x - ct),$$

for a nonzero constant c and a smooth function U, is a traveling wave whose profile is the graph of the function U, which wave travels at speed c to the right if $c > 0$ or to the left if $c < 0$.

Remark 8.7.1. A short note about the following example is warranted before the unassuming reader gets lost in the weeds (details) of the derivation that follows. The precise details of the derivation given in the next example are *NOT* important. The key idea is the ansatz that the solution has the form $u(x,t) = U(x - ct)$. Inserting this into the PDE results in a messy (but calculable) ODE that can (theoretically) be solved exactly and that is the point of discussing traveling waves.

Example 8.7.2 (Burger's equation again). Consider Burger's equation in one dimension yet again:

$$u_t + uu_x = Du_{xx}, \quad D > 0.$$

Suppose that there is a traveling wave solution of the form $u(x,t) = U(x-ct)$ with wave speed c, yet to be determined. To help with the substitution of this ansatz into the PDE we set

$$s = x - ct.$$

Then the required partial derivatives of the ansatz are

$$\begin{aligned} u_t &= -cU'(s), \\ u_x &= U'(s), \\ u_{xx} &= U''(s). \end{aligned}$$

Substitution of the ansatz into the PDE gives

$$-cU'(s) + U(s)U'(s) = DU''(s).$$

Suppressing the variable s gives the second-order nonlinear ODE in U,

$$-cU' + UU' = DU''.$$

Recognizing by the Chain Rule that

$$\frac{d}{ds}\frac{1}{2}U^2 = UU',$$

the ODE becomes

$$-cU' + \frac{1}{2}(U^2)' = DU''.$$

Since the second derivative is the derivative of the first derivative the ODE becomes

$$-cU' + \frac{1}{2}(U^2)' = D(U')',$$

i.e. every term here has a derivative applied to it. Integration of the ODE then yields

$$-cU + \frac{1}{2}U^2 = DU' + B$$

for a constant B that is independent of s. Below we see that the constant B is actually positive, but for now we just note that B is independent of s.

We now have a separable first-order ODE

$$\begin{aligned} \frac{dU}{ds} &= U' \\ &= \frac{1}{D}\left[-cU + \frac{1}{2}U^2 - B\right] \\ &= \frac{1}{2D}\left[-2cU + U^2 - 2B\right] \\ &= \frac{1}{2D}\left[U^2 - 2cU - 2B\right]. \end{aligned}$$

Separating the variables gives

$$\frac{dU}{U^2 - 2cU - 2B} = \frac{ds}{2D}.$$

The roots of the quadratic in the denominator $U^2 - 2cU - 2B$ are

$$\frac{2c \pm \sqrt{4c^2 + 8B}}{2} = c \pm \sqrt{c^2 + 2B}.$$

Hence, we let

$$f_1 = c - \sqrt{c^2 + 2B},\ f_2 = c + \sqrt{c^2 + 2B}.$$

Note that

$$c = \frac{f_1 + f_2}{2}.$$

With the two roots f_1 and f_2 the quadratic denominator has the factorization

$$U^2 - 2cU - 2B = (U - f_1)(U - f_2).$$

Hence the separable form of the ODE becomes

$$\frac{dU}{(U - f_1)(U - f_2)} = \frac{ds}{2D}.$$

Assuming the constants B and c satisfy

$$c^2 > -2B,\ \text{i.e.,}\ c^2 + 2B > 0,$$

(the discriminant is positive) implies that the roots f_1 and f_2 are both real and satisfy

$$f_1 < f_2.$$

For the partial fraction decomposition we have

$$\frac{1}{(U - f_1)(U - f_2)} = \frac{\alpha}{U - f_1} + \frac{\beta}{U - f_2}$$

where

$$1 = \alpha(U - f_2) + \beta(U - f_1),$$

which implies (by taking $U = f_1$ and then by taking $U = f_2$) that

$$\alpha = \frac{1}{f_1 - f_2} \text{ and } \beta = \frac{1}{f_2 - f_1}.$$

Thus the partial fraction decomposition is

$$\frac{1}{(U - f_1)(U - f_2)} = \frac{1}{(f_1 - f_2)(U - f_1)} + \frac{1}{(f_2 - f_1)(U - f_2)}.$$

Hence the separated form of the ODE becomes

$$\frac{dU}{(f_1 - f_2)(U - f_1)} + \frac{dU}{(f_2 - f_1)(U - f_2)} = \frac{ds}{2D}.$$

Integrating this yields

$$\frac{1}{f_1 - f_2} \ln|U - f_1| + \frac{1}{f_2 - f_1} \ln|U - f_2| = \frac{s}{2D} + A,$$

where A is the arbitrary constant of integration.

We *assume* that $A = 0$ since we are seeking for a solution that is of the ansatz form (not necessarily all of them) and the choice of $A = 0$ will simplify the computation.

We combine the logarithms to obtain

$$\begin{aligned}\frac{1}{f_1 - f_2} \ln|U - f_1| + \frac{1}{f_2 - f_1} \ln|U - f_2| &= -\frac{1}{f_2 - f_1} \ln|U - f_1| + \frac{1}{f_2 - f_1} \ln|U - f_2| \\ &= \frac{1}{f_2 - f_1} \ln\left|\frac{1}{U - f_1}\right| + \frac{1}{f_2 - f_1} \ln|U - f_2| \\ &= \frac{1}{f_2 - f_1} \ln\left|\frac{U - f_2}{U - f_1}\right|.\end{aligned}$$

We *assume* that U satisfies

$$f_1 < U < f_2$$

so that

$$U - f_1 > 0 \text{ and } f_2 - U > 0.$$

Then

$$\frac{1}{f_2 - f_1} \ln\left|\frac{U - f_2}{U - f_1}\right| = \frac{1}{f_2 - f_1} \ln\left(\frac{f_2 - U}{U - f_1}\right).$$

So the relationship between U and s becomes

$$\frac{1}{f_2 - f_1} \ln\left(\frac{f_2 - U}{U - f_1}\right) = \frac{s}{2D}.$$

Now we solve this equation for U, which is a linear equation in U in disguise.

Isolating the logarithm gives

$$\ln\left(\frac{f_2 - U}{U - f_1}\right) = \frac{(f_2 - f_1)s}{2D}.$$

Exponentiating gives

$$\frac{f_2 - U}{U - f_1} = \exp\left(\frac{(f_2 - f_1)s}{2D}\right).$$

Multiplying through by $U - f_1$ gives

$$f_2 - U = (U - f_1) \exp\left(\frac{(f_2 - f_1)s}{2D}\right) = U \exp\left(\frac{(f_2 - f_1)s}{2D}\right) - f_1 \exp\left(\frac{(f_2 - f_1)s}{2D}\right).$$

Moving terms with U to the left-side and the remaining terms to the right-side

$$-U - U \exp\left(\frac{(f_2 - f_1)s}{2D}\right) = -f_2 - f_1 \exp\left(\frac{(f_2 - f_1)s}{2D}\right).$$

Multiplying through by -1 simplifies this to

$$U + U \exp\left(\frac{(f_2 - f_1)s}{2D}\right) = f_2 + f_1 \exp\left(\frac{(f_2 - f_1)s}{2D}\right).$$

Factoring out the U from the left-hand side gives

$$U\left(1+\exp\left(\frac{(f_2-f_1)s}{2D}\right)\right)=f_2+f_1\exp\left(\frac{(f_2-f_1)s}{2D}\right).$$

Solving for U we obtain

$$U=\frac{f_2+f_1\exp\left(\dfrac{(f_2-f_1)s}{2D}\right)}{1+\exp\left(\dfrac{(f_2-f_1)s}{2D}\right)}.$$

Setting

$$K=\frac{f_2-f_1}{2D}$$

simplifies the solution to

$$U=\frac{f_2+f_1\exp(Ks)}{1+\exp(Ks)}.$$

The constant K is positive because $D>0$ and $f_2-f_1>0$.

Because $K>0$, the function U has the limiting properties

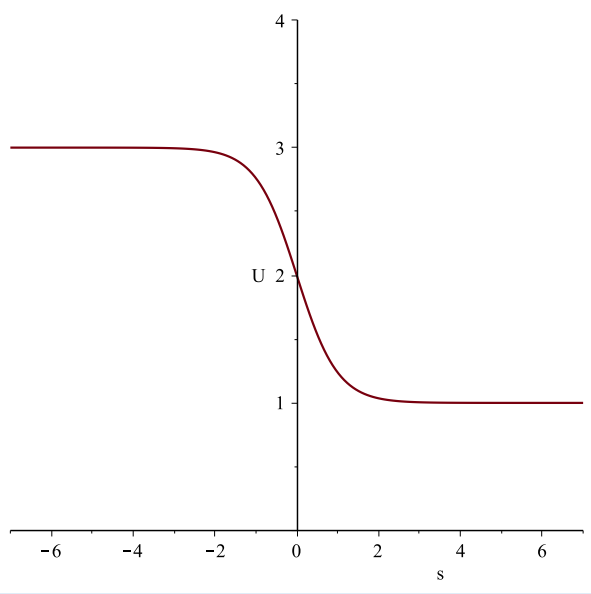
$$\lim_{s\to-\infty}U(s)=f_2 \text{ and } \lim_{s\to\infty}U(s)=f_1.$$

From the first-order ODE that U satisfies,

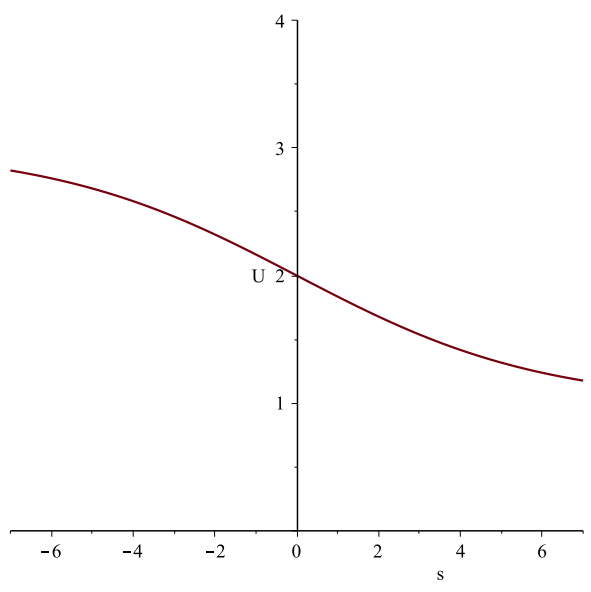
$$\frac{dU}{ds}=\frac{1}{2D}(U-f_1)(U-f_2)$$

and from the inequalities $f_1<U<f_2$, it follows that $\frac{dU}{ds}<0$, so that U is decreasing.

Here is the graph of $U(s)$ for $f_1=1$, $f_2=3$, $D=\frac{1}{2}$ (so $K=2$).



Here is the graph of $U(s)$ for $f_1=1$, $f_2=3$, and $D=6$ (so $K=\frac{1}{3}$).



You will notice that as we increase D the shape of U becomes more linear, or as D gets closer to 0 the shape of U becomes steeper near $s = 0$, i.e., $U'(0)$ becomes large.

Now that we have the profile function $U(s)$ we return to the variables (x, t) through $s = x - ct$ to complete the ansatz:

$$u(x,t) = U(s) = \frac{f_2 + f_1 \exp(K(x-ct))}{1 + \exp(K(x-ct))}.$$

With $c > 0$ the profile U travels with speed c to the right.

- The value of $u(x,t) = U(x-ct)$ at x in the past (as $t \to -\infty$) is f_2.
- The value of $u(x,t) = U(x-ct)$ at x in the future (as $t \to \infty$) is f_1.

Physically traveling waves occur in a variety of circumstances, including the formation of near-shocks ($D > 0$ is very small). This is a classical example of a traveling wave that appears in several other PDEs in addition to Burger's equation. The key in determining the nature of these traveling waves is to suppose there is a solution such as the ansatz $u(x,t) = U(x-ct)$ in the previous example, and then to work out what this means for the solution and its dependence on the auxiliary conditions of the problem. Traveling waves can appear even in nonlinear evolution equations such as Burger's equation, and as such can exhibit a very long life-time. Stability of these traveling waves is a subject of intense interest unto itself.

Remark 8.7.3. The key aspects of this example are not the details involved (in case you haven't noticed, there are *A LOT* of details here), but the ansatz that the solution looks like $u(x,t) = U(x-ct)$ where the speed c is to be determined.

As we have seen already in our study of the advection and wave equations, this type of solution is anticipated for hyperbolic PDEs. We may anticipate that part of the solution indeed does look like such a traveling wave, and this is the reason that there is such a focus on the topic. Just as we have emphasized previously, the exact calculations involved can be quite hideous, but the ansatz here is critically important. Often numerical methods are necessary to find such solutions and to study their stability and other properties, but the key is the form of the solution itself.

8.7.2 Standing/ plane waves

Another type of wave that appears frequently in nature, and one that we have already seen (unbeknown to the reader) is a standing or plane wave. Typically these waves are present only in linear PDEs, although several nonlinear PDEs admit plane waves as partial solutions.

Before we consider these standing waves, we note that we will make substantial use of Euler's formula

$$\exp(i\theta) = \cos\theta + i\sin(\theta)$$

to express periodic functions in complex notation. The ansatz we will use in this case has the complex variable form

$$u(x,t) = \exp(i(kx - \omega t))$$

for real constants k and ω. The partial derivatives of this ansatz are

$$u_t = -i\omega u$$
$$u_x = ik\omega u.$$

We will see that the two constants k and ω are not generally independent; there is a relationship between them that arises from the ansatz being a solution of a specific PDE.

Physical example and interpretation

To begin, let us consider yet again the one-dimensional advection equation:

$$u_t + au_x = 0,$$

which we know has solution $u(x,t) = f(x-at)$ for some function $f(z)$ of a single variable. Quite frequently we are interested in harmonic waves or, equivalently, in this case, periodic waves that can be expressed as $f(x-at) = A\cos(x-at)$.

Generically for a linear, constant coefficient PDE we can expect this type of a solution to appear, i.e. for the most general case we may consider solutions of the form $u(x,t) = e^{i(kx-\omega t)}$.

Example 8.7.4. Considering the advection equation in one-dimension:

$$u_t + au_x = 0,$$

we proceed with the ansatz that $u(x,t) = e^{i(kx-\omega t)}$. Inserting this ansatz into the PDE, we see that

$$-i\left(-\omega + ak\right)e^{i(kx-\omega t)} = 0 \Rightarrow \omega = ak,$$

so that such solutions can be written for each choice of k in the form

$$u(x,t) = e^{ik(x-at)} = \cos\left[k(x-at)\right] + i\sin\left[k(x-at)\right],$$

Because the PDE is linear homogeneous, we can use complex-valued linear combinations to obtain the real-valued solutions

$$\cos(k(x-at)) \text{ and } \sin(k(x-at)).$$

You may have noticed that we discovered the characteristic projections $x - at =$ constant with the ansatz.

The functional relationship $\omega = ak$ is referred to as the dispersion relation.

Definition 8.7.5. *For a linear (this often applies to nonlinear situations as well, but we won't enter that realm here) PDE with solutions satisfying the ansatz $u(x,t) = e^{i(kx-\omega t)}$ the relationship $\omega = \omega(k)$ is called the* dispersion relation.

Remark 8.7.6. As seen in the example above, the dispersion relation for a linear PDE can be computed simply by inserting the ansatz into the PDE and simplifying the result algebraically.

Definition 8.7.7. *The* wavelength *(or period) of any nonzero scalar multiple of the wave*

$$\cos(k(x-at))$$

is the quantity

$$\lambda = \frac{2\pi}{k}.$$

The quantity k is called the wavenumber *of the wave.*

- The units of k are per length, so the units of λ are length.
- The larger the wave-number the smaller the wave-length (or period), and the smaller the wave-number the larger the wave-length (or period).

In general plane waves are present if the PDE is linear, constant coefficient, and homogeneous, although as mentioned above, more complicated PDEs often admit plane waves if part of the PDE is linear with constant coefficients. In fact, the analysis of nonlinear PDEs can often be reduced to an accurate understanding of how the nonlinearity causes two or more plane waves (generated by the linear part of the PDE) to interact with each other.

For a linear PDE, if two different plane waves $u_1(x,t) = \tilde{u}_1 e^{i(k_1 x - \omega_1 t)}$ and $u_2(x,t) = \tilde{u}_2 e^{i(k_2 x - \omega_2 t)}$ are solutions then so is $u(x,t) = u_1(x,t) + u_2(x,t)$. In fact, as we will see later, all solutions to a linear, constant coefficient PDE are a superposition of plane waves. Thus, we want to understand how these waves behave and what they mean because they play a vital role in the determination of solutions to linear PDEs.

Phase Velocity

Suppose now that we want to determine the speed at which these waves propagate, that is how fast does the wave $\cos[(kx-\omega t)]$ propagate given that we have identified the dispersion relation $\omega = \omega(k)$?

- To evaluate this, we note that the temporal frequency of this solution is $f = \frac{\omega}{2\pi}$.
- Note that the frequency has units of inverse time.

What we are really asking then, is how fast the wave can travel one wave-length over one temporal period, i.e. we want to know the phase velocity $c_p = f\lambda = \frac{\omega}{k}$. If we had a linear plane wave in an infinite medium this would be the practical speed at which that wave would propagate.

Definition 8.7.8. *A plane wave with dispersion relation $\omega = \omega(k)$ has phase velocity $c_p = \omega(k)/k$.*

Example 8.7.9. Consider the following beam equation (representing vibrations of a beam)

$$u_{tt} + \gamma u_{xxxx} = 0, \quad \gamma > 0.$$

Applying the ansatz that solutions are of the form $u(x,t) = e^{i(kx-\omega t)}$ gives

$$-\omega^2 + \gamma k^4 = 0,$$

which implies that the dispersion relation is $\omega(k) = \pm\sqrt{\gamma}k^2$, so that waves propagate at speed $c_p = \frac{\omega}{k} = \pm\sqrt{\gamma}k$. Thus larger scale ($k$ is smaller) waves travel slower than smaller scale (k large) waves.

If we selected the positive root for the dispersion relation, i.e. $\omega = k^2\sqrt{\gamma}$ then the solution would look like

$$u(x,t) = \cos(kx - \omega t) = \cos(k(x - k\sqrt{\gamma}t)) = \cos(k(x - c_p t)).$$

Remark 8.7.10. One may legitimately ask why we are so focused on these topics. For the beam equation considered in the previous example, propagation of these waves is important to understand. As mentioned, the beam equation models the motion/vibration of a beam put under a certain type of pressure. Certain displacements or stresses of the beam will cause both large and small scale disturbances that will propagate throughout the length of the beam, but only the smallest scale disturbances will propagate rapidly. This means that the small vibrations will be felt first, followed (rather quickly according to our common time-scales) by larger scale oscillations that are often more devastating if the beam is holding up a ceiling or part of a bridge construction.

Group Velocity

Now let $u_1(x,t) = A(\cos(k_1 x - \omega_1 t)$ and $u_2(x,t) = A\cos(k_2 x - \omega_2 t)$ be two plane wave solutions to a linear PDE where $k_1 = k + \Delta k$, $k_2 = k - \Delta k$ and $\omega_1 = \omega + \Delta\omega$, $\omega_2 = \omega - \Delta\omega$. This means we are looking at two plane wave solutions that are only slightly perturbed relative to each other (both in their spatial and temporal dependence). Then

$$\begin{aligned} u(x,t) = u_1 + u_2 &= A[\cos((k+\Delta k)x - (\omega + \Delta\omega)t) + \cos((k-\Delta k)x - (\omega - \Delta\omega t))] \\ &= A[\cos(kx + \Delta kx - \omega t - \Delta\omega t) + \cos(kx - \Delta kx - \omega t + \Delta\omega t)] \\ &= A[\cos(kx - \omega t + \Delta kx - \Delta\omega t) + \cos(kx - \omega t - \Delta kx + \Delta\omega t)] \\ &= A[\cos(kx - \omega t + (\Delta kx - \Delta\omega t)) + \cos(kx - \omega t - (\Delta kx - \Delta\omega t))] \\ &= 2A\cos(kx - \omega t)\cos(\Delta kx - \Delta\omega t)) \end{aligned}$$

is also a solution with rapidly (the first cosine term) and slowly (the second cosine term which has very large scales both temporally and spatially) varying parts, where we used the trigonometric identity

$$2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta).$$

For an example, see Figure 8.7 which demonstrates how this works for $k = 10$ and $\Delta k = 0.5$ in this case.

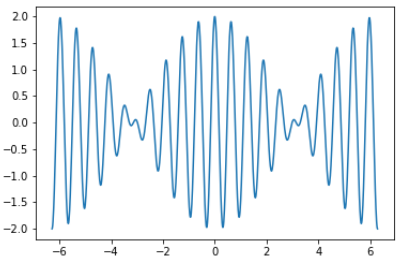


Figure 8.7: Demonstration of a wave packet. The smaller scales travel at a different speed than the larger wave packet.

- In this linear combination the coefficient $2A$ is the maximum amplitude, and the term
$$\cos(kx - \omega t) = \cos\left(k\left(x - \frac{\omega}{k}t\right)\right)$$
is a "rapidly" oscillating wave (with shorter wave-length $\frac{2\pi}{k}$),.
- The term
$$\cos(\Delta k x - \Delta\omega t) = \cos\left(\Delta k\left(x - \frac{\Delta\omega}{\Delta k}t\right)\right)$$
is a slowly oscillating wave (with longer wave-length $\frac{2\pi}{\Delta k}$).
- The shorter wave-length wave travels at the speed
$$\frac{\Delta\omega}{\Delta k}$$
that differs from the speed
$$\frac{\omega}{k}$$
at which longer wave-length waves travel.
- The larger wave-length (the one corresponding to the Δk) wave is called the wave packet.
- The ratio of $\Delta\omega/\Delta k$ approximates the derivative $\frac{d\omega}{dk}$ from the dispersion relation.

Definition 8.7.11. *For a set of plane waves that have the dispersion relation $\omega = \omega(k)$ the group velocity $c_g = \frac{d\omega}{dk}$ defines the velocity at which the wave packet travels.*

Example 8.7.12. This is the same PDE as the previous example, i.e.
$$u_{tt} + \gamma u_{xxxx} = 0,$$

wherein we found a dispersion relation of $\omega = \pm\sqrt{\gamma}k^2$ with phase velocity $c_p = \pm\sqrt{\gamma}k$. In this case the group velocity is easily computed as $c_g = \pm 2\sqrt{\gamma}k \neq c_p$. Thus, the wave packet (that is likely what is observed) is actually twice as fast as the speed of an individual wave.

Remark 8.7.13. We can also define a dispersion relation, group velocity and phase velocity in higher-dimensional settings. In higher dimensions we suppose that the solution has the form $u(\mathbf{x}, t) = e^{i(\mathbf{k}\cdot\mathbf{x}-\omega t)}$, where now $\mathbf{k}$ is a vector the same dimension as $\mathbf{x}$. The dispersion relation $\omega(\mathbf{k})$ follows just as in the scalar case, and the group velocity is defined as the gradient of $\omega(\mathbf{k})$ with respect to the wavenumber vector $\mathbf{k}$, i.e. $\mathbf{c}_g = \nabla_k\omega(\mathbf{k})$. The phase velocity is defined component-wise as the dispersion relation divided by the corresponding wavenumber in that dimension, i.e. the first component of the phase velocity is $\omega(\mathbf{k})/k_1$ etc.. This is typically written in the rather misleading form of $\mathbf{c}_p = \omega(\mathbf{k})/\mathbf{k}$.

Who cares about planar waves? (We do!)

Some numerical schemes may appear to have high accuracy, but if they compute an inaccurate dispersion relation, then the waves will travel at the wrong speeds, leading to a jumbled mess (to put things mildly). This means that certain waves that may not be important could dominate the solution, and the nonlinear interaction of waves traveling at different speeds will be completely wrong.

To see how this works, we once again consider the advection equation. We have already seen that the dispersion relation is $\omega(k) = ak$ in this case. Now we consider wave-like solutions to the upwind method, i.e. $u(x, t) = e^{i(kx-\omega t)}$ where we look at the spatial discretization only, but leave the time derivative in place:

$$u_t = -\frac{a}{\Delta x}(u(x, t) - u(x - \Delta x, t)).$$

This leads to the numerical dispersion relation for the upwind method:

$$-i\omega e^{i(kx-\omega t)} = -\frac{a}{\Delta x}\left(1 - e^{-ik\Delta x}\right)e^{i(kx-\omega t)} \tag{8.28}$$

$$\Rightarrow \omega = \frac{a}{i\Delta x}\left(1 - e^{-ik\Delta x}\right) \tag{8.29}$$

$$\Rightarrow \omega \approx \frac{a}{i\Delta x}\left(1 - 1 + ik\Delta x + \frac{k^2(\Delta x)^2}{2}\right) + O((\Delta x)^2), \tag{8.30}$$

where the final step was to expand the exponential in a Taylor series in Δx. Thus we see that the dispersion relation using the upwind method is actually pretty close to the true dispersion relation for the advection equation, i.e.

$$\omega_{upwind} = ak - i\frac{ak^2\Delta x}{2} + O((\Delta x)^2). \tag{8.31}$$

Inserting this back into the ansatz, we see why the upwind method results in decay of the amplitude of the solution:

$$u_{upwind}(x, t) = e^{i(kx-\omega t)} \approx e^{ik(x-at)}e^{-\frac{ak^2(\Delta x)^2}{2}t}. \tag{8.32}$$

For small values of k (very large scales), and sufficiently small values of Δx we may not notice this effect, but the very smallest scales (large k) will exponentially decay using the upwind method.

Remark 8.7.14. The opposite effect can be seen for the Lax-Wendroff method, where instead of the dominant error term producing decay in the solution (with time), we find that the dispersion relation for Lax-Wendroff introduces additional oscillatory effects (terms that are complex exponentials. This leads to numerical approximations that have additional non-physical oscillations.

The differences between the different types of numerical approximations, and their effects on the approximated solution is most apparent when we are trying to numerically capture a wave packet wherein it is critical to capture the correct group and phase velocity of the waves. As we have seen here, that is not an easy task, and we often have to balance the different type of errors that may be introduced from the numerical method of choice.

Remark 8.7.15. Clearly in the derivation of the dispersion relation for upwind, we neglected the effects of the temporal discretization without providing a valid reason for doing so. If we included the temporal discretization then we would end up with a more complicated version of the same dispersion relation, that upon expanding in terms of Δt would give us the same approximate relationship shown here.

Classification of linear PDEs

Another aspect for which planar waves is useful, is in the classification of linear PDEs. For a PDE of the general form $Lu = 0$ where L is a linear, constant coefficient differential operator with dispersion relation $\omega = \omega(k)$ then

- L is diffusive if $\omega(k)$ is complex valued.
- L is dispersive if $\omega(k)$ is real and $\omega(k)/k$ is independent of k.
- L is hyperbolic if $\omega(k)$ is real and $\omega(k)/k$ depends on k in which case $c_g \neq c_p$.

We have seen the second case with the advection equation $u_t + au_x = 0$ where $\omega(k) = ak$ and we saw the third case with the vibrating beam equation where $\omega(k) = \pm k^2\sqrt{\gamma}$.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

8.1. Come up with a different model of the velocity of the vehicle concentration for the traffic flow problem. Justify how this model works, and explain the physical relevance of it.

8.2. A homogeneous (constant ρ, C and K) metal rod has cross-sectional area $A(x)$, $0 < x < l$, and there is only a small variation of $A(x)$ with x, so that the assumption of constant temperature in any cross section remains valid. There are no sources and the flux is given by $-Ku_x(x,t)$. From a conservation law obtain a partial differential equation for the temperature $u(x,t)$ that reflects the area variation of the bar.

8.3. In the absence of sources, derive the diffusion equation for radial motion in the plane,

$$u_t = \frac{D}{r}(ru_r)_r,$$

from first principles. That is, take an arbitrary domain between circles $r = a$ and $r = b$ and apply a conservation law for the density $u = u(r,t)$, assuming that the flux is $J(r,t) = -Du_r$. Assume no sources.

8.4. The biomass density u (mass per unit volume) of zooplankton in a very deep lake varies as a function of depth x and time t. Zooplankton diffuse vertically with diffusion constant D, and buoyancy effects cause them to migrate toward the surface at a constant speed of ag, where g is the acceleration due to gravity and a is a positive constant. At any time, in any vertical tube of unit cross-sectional area, the total biomass of zooplankton is a constant U. From first principles derive a partial differential equation model for the biomass density of zooplankton. (Take $x = 0$ at the surface).

8.5. *Continuing from the last problem, find the steady-state biomass density as a function of depth. (Formulate any reasonable boundary conditions, or other auxiliary conditions, that are needed to solve the problem).

8.6. A fluid is called *incompressible* if for any region Ω the rate at which the fluid enters and leaves the region is 0, i.e. the fluid cannot be compressed. The prototypical incompressible fluid is water at room temperature. If an incompressible fluid has constant density, show that the velocity field of the fluid must be divergence free. Hint: Translating this statement into mathematics is the only difficulty for this problem.

8.7. *Show that the growth-diffusion (reaction-diffusion) equation $u_t = D\Delta u + ru$ can be transformed into a pure diffusion equation via the transformation $v = ue^{-rt}$.

8.8. Let $\Omega \subset \mathbb{R}^3$ be a region occupied by a fluid body (e.g. a lake), $\mathbf{v}(\mathbf{x},t)$ be the instantaneous velocity of the fluid at point $\mathbf{x} \in \Omega$ and time t, and $\rho(\mathbf{x},t)$ be the concentration of some pollutant at time t and position $\mathbf{x} \in \mathbb{R}^3$. In this case the flux has two parts—both an advection component and a diffusive component, which can be added together. Derive a PDE for ρ for a given velocity field $\mathbf{v}(\mathbf{x},t)$. The resulting PDE is called the *advection-diffusion* equation.

8.9. Prove Green's first identity (the third equality in Corollary 8.2.7). Hint: It may be useful to first prove the following version of the product rule: $\nabla \cdot (u\mathbf{v}) = \nabla u \cdot \mathbf{v} + u\nabla \cdot \mathbf{v}$, where u is scalar valued and $\mathbf{v}$ is a vector field.

8.10. Prove Green's second identity (the third equality in Corollary 8.2.7).

8.11. For the first-order PDE $au_x + bu_y = 0$ show that the characteristic curves given by

$$\frac{dx}{ds} = a \qquad \frac{dy}{ds} = b,$$

are level sets of solutions to the PDE. Hint: this is equivalent to showing that if $\gamma(s) = (x(s), y(s))$, $s \in I$, is a solution of $\frac{dx}{ds} = a(x,y)$ and $\frac{dy}{ds} = b(x,y)$, then $d(u \circ \gamma(s))/ds = 0$ for all $s \in I$

8.12. Find the characteristics of the first-order PDE $yu_x + xu_y = 0$.

8.13. For the scalar equation $u_t + \partial_x f(u) = 0$ show that the curves defined by $\frac{dt}{ds} = 1$ and $\frac{dx}{ds} = f'(u)$ are characteristics of the PDE, i.e. u is constant along these curves. ($f'(u)$ refers to the derivative of f with respect to u).

8.14. Consider the linear PDE

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

Derive a formula for the characteristics of this equation when $f \neq 0$. Hint: When $f = 0$ then u is constant along the characteristics, but when $f \neq 0$ then u is no longer constant along the characteristics.

8.15. For the inviscid Burger equation in one dimension with initial data $u(0,x_1) > u(0,x_2)$ where $x_1 < x_2$ find the point (in the $x-t$ plane) at which the uniqueness of the solution is no longer guaranteed classically. (This is the intersection of the two characteristic curves in the figure in Example 8.3.15).

8.16. *For the traffic flow problem,

$$u_t + (1-2u)u_x = 0, \quad u(x,0) = f(x),$$

use the method of characteristics with parameter s to determine the general solution. Hint: Use 'initial' conditions with respect to s as $x(s=0) = r$, $t(s=0) = 0$ and $u(s=0) = f(r)$. The solution will look something like $u(;r) = f(r)$, $x(t;r) = ...$

8.17. *Solve the problem above for initial data given by the initial condition below which is meant to imitate light density traffic entering heavy traffic:

$$f(x) = \begin{cases} \alpha, & x \leq 0 \\ \left(\frac{3}{4} - \alpha\right)x + \alpha, & 0 \leq x \leq 1 \\ \frac{3}{4}, & x \geq 1 \end{cases},$$

where $0 \leq \alpha \leq \frac{3}{4}$.
NOTE:The previous two problems highlight some of the potential usefulness of the method of characteristics, and also identifies the greatest draw back for this method, in that typically it is difficult to derive solutions of the more useful format $u(x,t)$ as a function of x and t.

8.18. For the advection equation

$$u_t + au_x = 0, \quad t > 0, \quad x \in [0,L],$$

where $a > 0$ and $u(x,t)$ represents the concentration of a pollutant in a pipe, which of the following are valid auxiliary conditions, and what could they mean physically?

(i) $u(x,0) = f(x)$ and $u(0,t) = \alpha$, and $u(L,t) = b$.

(ii) $u(x,0) = f(x)$ and $u(0,t) = \alpha$.

(iii) $u(x,0) = f(x)$ and $u(L,t) = b$.

(iv) $u(0,t) = \alpha$ and $u(L,t) = b$.

8.19. For the traffic flow problem

$$u_t + (1-2u)u_x = 0, \quad t > 0, \quad x \in [0,L],$$

which of the following are valid auxiliary conditions, and what could they mean physically if we are concerned with traffic along the northbound section of a freeway?

(i) $u(x,0) = f(x) > 0$ and $u(0,t) = a$, and $u(L,t) = b$ with $a \neq b$.

(ii) $u(x,0) = f(x) > 0$ and $u(0,t) = a$.

(iii) $u(x,0) = f(x) > 0$ and $u(L,t) = b$.

(iv) $u(0,t) = a$ and $u(L,t) = b$ with $a \neq b$.

8.20. Use the method of characteristics to solve the problem (Hint: you may want to review the derivatives of inverse trigonometric functions)

$$\begin{aligned}(1+x^2)u_x + u_y &= yu^2\\ u(x,0) &= g(x).\end{aligned}$$

8.21. Use the method of characteristics to solve the problem

$$\begin{aligned}u_t + xu_x &= \sin t\\ u(0,x) &= f(x).\end{aligned}$$

8.22. Would it make sense to solve the one-dimensional advection equation using a centered finite difference in the spatial variable and forward Euler in time? Why or why not?

8.23. Modify Algorithm 8.1 (still with $a > 0$ here) to work for the Lax-Wendroff scheme. Create an animation for both the upwind and Lax-Wendroff schemes for $a = 1$ on the periodic interval $[-\pi, \pi]$ with $u(x,0) = \sin(x)$. Run this over the time interval $[0, 10]$. Use the same CFL number (0.5) for both simulations. What do you see as the difference between the two solutions?

8.24. Modify Algorithm 8.1 to work for any constant value of a, i.e. both $a > 0$ and $a < 0$. Generate an animation that shows the evolution of the solution on the periodic interval $[-\pi, \pi]$ with initial condition $f(x) = \sin(x)$ up to time $T = 5$, for both $a = -1$ and $a = 1$. What happens to the solution as time goes on?

8.25. Modify your code from the previous problem to account for $a = a(x)$, a non-constant function. Generate an animation showing the evolution of the solution for the same domain, and initial condition as in the previous problem, but now with $a(x) = 2 + \cos(x)$. How do you select a time-step/CFL number in this case?

8.26. *Derive (8.21) from the Lax-Wendroff scheme directly.

8.27. *Let $u \in C^3(\mathbb{R}^2; \mathbb{R})$ be a solution to the wave equation

$$u_{tt} = u_{xx}. \tag{8.33}$$

Define the energy density as $e = \frac{1}{2}(u_t^2 + u_x^2)$ and the momentum density as $\rho = u_t u_x$.

(i) Prove that $e_t = \rho_x$ and $\rho_t = e_x$.

(ii) Show that both e and ρ satisfy (8.33).

8.28. *Let $u \in C^3(\mathbb{R}^2; \mathbb{R})$ be a solution to (8.33).

(i) Show that $u(t, x-y)$ is also a solution of (8.33) for each fixed $y \in \mathbb{R}$.

(ii) Show that u_x and u_t are also solutions of (8.33).

8.29. Let $u \in C^3(\mathbb{R}^2; \mathbb{R})$ be a solution to (8.33). Show that

$$v(t,x) = u(at, ax), \qquad a \in \mathbb{R} \setminus \{0\},$$

is also a solution of (8.33).

8.30. *Let $u \in C^3(\mathbb{R}^2;\mathbb{R})$ be a solution to (8.33). Prove for all $x,t,h,k \in \mathbb{R}$ that

$$u(t+h,x+k)+u(t-h,x-k)=u(t+k,x+h)+u(t-k,x-h).$$

8.31. Consider the damped wave equation

$$u_{tt}+\mu u_t = c^2 u_{xx}, \qquad \mu > 0. \tag{8.34}$$

If $u \in C^2(\mathbb{R}^2;\mathbb{R})$ and $u_x \to 0$ as $|x| \to \infty$ (in order to allow integration by parts), then prove that the total energy

$$E(t)=\int_{-\infty}^{\infty} \frac{1}{2}\left(u_t^2+c^2u_x^2\right)\,dx$$

is a decreasing function.

8.32. Derive an equation for the transverse motion of a string located in a medium that resists its movement, such as air or water. The resistance is expressed as a force opposite in direction and proportional in magnitude to the velocity of the string's displacement, i.e. there is an additional additive force given by $-ku_t$ for some positive constant k. Thus it does not affect the longitudinal (horizontal) direction of the string's evolution. Other than this condition, you may make the same assumptions we did in class in the absence of such resistance.

8.33. Show that the transformation

$$\xi = x-ct, \qquad \eta = x+ct$$

transforms the wave equation into the partial differential equation

$$\phi_{\xi\eta}=0, \qquad \phi(\xi,\eta)=u(x,t),$$

where $u(x,t)$ is a solution of the one-dimensional wave equation. As a consequence, show that the general solution of the wave equation is given by $u(x,t)=f(x+ct)+g(x-ct)$.

8.34. *Recall that for a 3-dimensional vector field $\mathbf{u}=(u(x,y,z),v(x,y,z),w(x,y,z))^T$, the curl is defined as

$$\nabla \times \mathbf{u} = (w_y-v_z,u_z-w_x,v_x-u_y)^T.$$

Prove that for a three-dimensional vector field $\mathbf{u}$, $\nabla\times(\nabla\times\mathbf{u})=\nabla(\nabla\cdot\mathbf{u})-\Delta\mathbf{u}$, where $\Delta\mathbf{u}$ is the component-wise Laplacian operator.

8.35. *Maxwell's equations for a nonconducting medium with permeability μ and permittivity ε are

$$\begin{aligned}\frac{\partial B}{\partial t}+\nabla\times E &= 0\\ \varepsilon\frac{\partial E}{\partial t} &= \frac{1}{\mu}\nabla\times B,\\ \nabla\cdot B &= 0\\ \nabla\cdot E &= 0,\end{aligned}$$

where B is the magnetic induction and E is the electric field (both vector fields). Show that the components of B and E satisfy the three-dimensional wave equation $u_{tt}-c^2(u_{xx}+u_{yy}+u_{zz})=0$ with propagation speed $c=(\varepsilon\mu)^{-\frac{1}{2}}$, where $u=u(t,x,y,z)$.

8.36. *Using D'Alembert's solution, solve

$$\begin{aligned} u_{tt} - u_{xx} = 0, \quad 0 < x < \pi, t > 0 \\ u(0,t) = u(\pi,t) = 0, \quad t > 0, \\ u(x,0) = 0, u_t(x,0) = 4\sin x, \quad 0 < x < \pi. \end{aligned}$$

8.37. Solve

$$\begin{aligned} u_{tt} =& u_{xx}, \quad x \in \mathbb{R}, t > 0, \\ u(0,x) =& \sin x, \\ u_t(0,x) =& 1, \end{aligned}$$

8.38. *Consider the non-homogeneous problem

$$\begin{aligned} u_{tt} - c^2 u_{xx} = f(x,t), \quad x \in \mathbb{R}, t > 0, \\ u(x,0) = 0, u_t(x,0) = 0, \quad x \in \mathbb{R}. \end{aligned}$$

By integrating over the characteristic triangle T bounded by characteristics emanating backward from the point (x_0, t_0) to the x axis, show that

$$u(x_0,t_0) = \frac{1}{2c} \int \int_T f(x,t) dx dt.$$

8.39. *Develop a solver for the one-dimensional wave equation based on (**??**) on the periodic spatial domain $x \in [-\pi, \pi]$ with initial conditions $u(x,0) = \sin(x)$ and $u_t(x,t) = \cos(x)$. For $c^2 = 1$ and $\Delta x = 0.01$, what is a value of Δt that gives reasonable results? Create an animation of the solution out to time $T = 2.0$.

8.40. Find all of the traveling wave solutions to the reaction-diffusion equation

$$u_t + u_x = u(1-u).$$

8.41. *Complete the solution to the traveling wave ODE for the KdV equation by finding the solution to the ODE

$$(u')^2 = su^2 - 2u^3.$$

8.42. *Let $u_1 = A\cos(k_1 x - \omega_1 t)$ and $u_2 = A\cos(k_2 x - \omega_2 t)$. Using trigonometric identities, prove that $u_1 + u_2 = 2A\cos(kx - \omega t)\cos(\Delta k x - \Delta\omega t)$ where $k_1 = k + \Delta k,\ k_2 = k - \Delta k$ and $\omega_1 = \omega + \Delta\omega,\ \omega_2 = \omega - \Delta\omega$.

8.43. Consider the evolution of the potential vorticity linearized about a state of rest with rotation given by the β-plane (don't worry about all of this, it is just some geophysical mumbo-jumbo)

$$\frac{\partial}{\partial t}\left[\Delta\psi - \Gamma\psi\right] + \beta\frac{\partial \psi}{\partial x} = 0$$

where $\psi = \psi(x,y,t)$ is a scalar valued function of two spatial variables and one temporal one. In this context, the Laplace operator acts on the two spatial dimensions only, i.e. $\Delta\psi = \psi_{xx} + \psi_{yy}$. Allowing for plane wave solutions of the form $\psi(x,y,t) = e^{i(kx+ly-\omega t)}$ find the dispersion relation for this equation.

8.44. Find the group and phase velocities for the previous problem. Is this equation dispersive?

8.45. Although one usually thinks of waves in the context of water waves, it is in fact much more difficult to determine the exact nature of water waves. There are actually several different potential PDEs that describe the evolution of a water wave, dependent on the 'asymptotic' regime of interest. One example is when there are short wavelengths in deep water (meaning the depth of the water is greater than the wavelength of the wave) in which case the effect of gravity is dominant which yields the dispersion relation $\omega = \sqrt{gk}$ where g is the gravitational constant. If instead we have a very long wavelength wave (compared to the depth of the water) then the dispersion relation is $\omega = k\sqrt{gH}$ where H is the depth of the water (held constant in this case).
Suppose that an earthquake in the Pacific Ocean (with approximate depth of $4000m$) triggers a wave $2m$ in height with a wave-length of $20km$, what will the approximate phase and group velocities of the wave be? Does this represent a dispersive or nondispersive system?
NOTE: The reason such a wave is so dangerous is that most of its energy will remain intact over a very large distance, and as the water depth shortens near shore, all of this energy will be concentrated in a very narrow region, raising the amplitude of the wave.

8.46. *Use the Cole-Hopf map

$$u = -2\nu \frac{\phi_x}{\phi}$$

to transform the viscous Burgers' equation

$$\begin{aligned} u_t + uu_x &= \nu u_{xx} \\ u(0,x) &= u_0(x) \end{aligned}$$

into the heat equation

$$\begin{aligned} \phi_t &= \nu \phi_{xx} \\ \phi(0,x) &= \exp\left(\frac{-1}{2\nu}\int_0^x u_0(y)\,dy\right). \end{aligned}$$

8.47. Calculate the exact dispersion relation for the upwind scheme, including the temporal discretization. Then using a Taylor series in Δt find the approximate dispersion relation (8.31).

8.48. *Derive the dispersion relation for the Lax-Wendroff scheme, and show how it leads to spurious oscillations in the approximated solution.

Notes

9 Auxiliary Conditions, Well-Posedness, and Parabolic and Elliptic equations

Science is a differential equation. Religion is a boundary condition.
—Alan Turing

9.1 Boundary conditions

Now that we have already seen the effects of auxiliary conditions on several different PDEs, we will finally get around to formally discussing them, and determining the impact they have on 'solving' a system of PDEs.

For systems of ODEs we had to only specify initial conditions in order to get rid of undetermined constants. PDEs, as we have seen in the previous Chapters, have undetermined functions instead, so specifying what happens at a specific point in time and space will not be sufficient (as it was for ODEs). Instead, we need to give what we will refer to as auxiliary conditions on a surface in space-time.

Clearly this won't be as simple as it was for ODEs. We need to make certain we prescribe just enough information, too little would not give a unique solution, and to much information would not allow a solution to exist, i.e. conflicting auxiliary conditions would result such as when two characteristics that stem from different auxiliary conditions intersect.

As an example, a first-order ODE requires a single initial condition, whereas a second-order ODE would require two initial conditions (or an initial condition on two different parts of the solution). If we tried to assign two different initial conditions to a first-order ODE there would be no solution, and prescribing a single initial condition for a second-order ODE would not give a unique solution. As we noted above, this is far more complicated and specialized to each individual case for PDEs.

Usually auxiliary conditions come from some physical considerations that will guarantee the uniqueness of solutions. This is not always feasible however, as demonstrated in the next Section.

9.1.1 Typical boundary conditions

We start this Section off with two physical, specific examples, and then delve into the general case. First we return to the simple case of the wave equation in one dimension.

Example 9.1.1 (Boundary conditions for the wave equation). Recall that the amplitude of a vibrating string in one dimension will satisfy the PDE

$$u_{tt} = c^2 u_{xx}. \tag{9.1}$$

If we let $u(x,t)$ be defined on the interval $x \in [0, L]$ then we have the choice of the following three different types of boundary conditions, with corresponding physical interpretation. We will only specify these conditions at $x = 0$ here, but equivalent forms of the boundary conditions at $x = L$ are of course possible.

(i) The string is fixed or attached at the point $x = 0$ so that $u(x = 0, t)$ is fixed at a given value, typically $u(x = 0, t) = 0$. *This is called a Dirichlet boundary condition.*

Dirichlet boundary conditions at both ends would apply to modeling the taut string of a guitar. A "solution" of the wave equation satisfying Dirichlet boundary conditions at both end points is

$$u(x,t) = \sum_{n=1}^{\infty} \left\{ a_n \cos \frac{cn\pi t}{L} + b_n \sin \frac{cn\pi t}{L} \right\} \sin \frac{n\pi x}{L}$$

for sequences $\{a_n\}$ and $\{b_n\}$ of coefficients to be determined by the initial position and initial velocity of the string. [We are ignoring for the moment the issue of convergence of this infinite series of functions, and only viewing the infinite series formally.]

For each n the factor $\sin(n\pi \frac{x}{L})$ always equals 0 when $x = 0$ or $x = L$, thus satisfying the Dirichlet boundary conditions $u(0,t) = 0$ and $u(L,t) = 0$ for all $t \geq 0$. The time dependent linear combination inside the curly braces is a simple harmonic oscillator for each fixed $x \in (0, L)$.

(ii) The string is attached to a vertical (frictionless of course) track that allows the string to move freely up and down at $x = 0$. *This will lead to the Neumann boundary condition* $u_x(x = 0, t) = 0$.

A "solution" of the wave equation satisfying Neumann boundary conditions at both end points is

$$u(x,t) = \sum_{n=1}^{\infty} \left\{ a_n \cos \frac{cn\pi t}{L} + b_n \sin \frac{cn\pi t}{L} \right\} \cos \frac{n\pi x}{L}$$

for sequences $\{a_n\}$ and $\{b_n\}$ of coefficients to be determined by the initial position and initial velocity of the string. Formally differentiating the infinite series with respect to x gives

$$u_x(x,t) = -\sum_{n=1}^{\infty} \frac{n\pi}{L} \left\{ a_n \cos \frac{cn\pi t}{L} + b_n \sin \frac{cn\pi t}{L} \right\} \sin \frac{n\pi x}{L}$$

which satisfies the Neumann boundary conditions are each endpoint $x = 0$ and $x = L$.

(iii) The string is still attached to a vertical track at $x = 0$, but now we want to account for the effects of friction. *This leads to what is called a Robin boundary condition which interpolates somehow between the Dirichlet and Neumann conditions, i.e.* $u(x = 0, t) + \alpha u_x(x = 0, t) = 0$ *for some value of* $\alpha > 0$.

A similar Robin boundary condition can be specified at $x = L$ as well. As you may suspect, a "solution" of the wave equation with Robin boundary conditions at *both* endpoints will be akin to a linear combination of the solutions for Dirichlet and Neumann boundary conditions.

Remark 9.1.2. For the one-dimensional wave equation one end of the string can have one type of boundary condition while the other end can have another type of boundary condition. The situation might be that a Dirichlet boundary is warranted at $x = 0$ while a Robin boundary condition is warranted at $x = L$. Whatever the boundary conditions are, there is a well-known procedure for solving linear second-order PDEs of two variables called separation of variables that begins with the ansatz

$$u(x, t) = X(x)T(t),$$

which quite literally assumes the variables x and t can be separated from each other in the solution.

Did you notice that the formal series solutions presented were sums of products of functions of the form $X(x)T(t)$?

The next example is higher dimensional, and a bit more complicated, but the fundamental ideas remain the same. In almost all situations there are three possible types of boundary conditions that make sense physically and provide robust solutions to the underlying PDE. Note that it doesn't make physical sense for more than one of these types of conditions to be satisfied at the same location at the same time.

Example 9.1.3 (Rayleigh-Bénard Convection). Consider the motion of fluid between two horizontal plates (see Figure 9.1). If the velocity of the fluid is described as the vector valued (three-dimensional) function $\mathbf{u} = \mathbf{u}(x, y, z, t)$ then the temperature field $\theta = \theta(x, y, z, t)$ obeys an advection diffusion equation (which you know thanks to the homework!):

$$\theta_t + \mathbf{u} \cdot \nabla\theta = \kappa\Delta\theta,$$

where κ is the thermal diffusivity (typically a constant) of the fluid. See Figure 9.1 for an illustration of this problem. If the bottom plate is surrounded by a hot substance then the bottom boundary plate can be specified as $\theta(x, y, z = 0, t) = T_{hot}(x, y, t)$ (typically referred to as a Dirichlet boundary condition).

If the top plate is highly insulated at a specific value then we can instead suppose that the flux at this boundary is fixed, i.e.

$$\left.\frac{\partial\theta}{\partial z}\right|_{x,y,z=h,t} = F(x, y, t),$$

which is usually taken as a constant, i.e. $F = constant$ (typically referred to as a Neumann boundary condition).

If we are in an experimental setup where we are trying to maintain a constant temperature at the top by surrounding the top plate with a cold 'bath'. Materials do not instantly conduct this cold temperature from the bath to the fluid, so there is a mixed boundary condition (typically called a Robin boundary condition) wherein the temperature is not specified precisely but neither is the flux, i.e. at the top plate $z = h$

$$-\eta = \theta(x, y, z = h, t) + \eta\nabla\theta(x, y, z = h, t) \cdot \mathbf{n} = \theta(x, y, z = h, t) + \eta\theta_z(x, y, z = h, t)$$

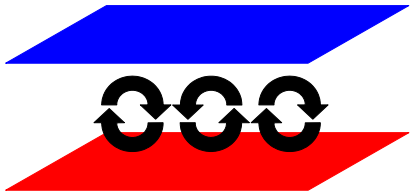


Figure 9.1: Rayleigh-Bénard convection is the motion of a fluid constrained between two plates, cold on the top and hot on the bottom. When the temperature difference is significant enough then the fluid begins to convect as illustrated and as the temperature difference increases then turbulence sets in and the motion becomes chaotic and unpredictable.

at $z = 1$ with Biot number η.

In general, for a PDE in a given domain Ω, there are typically three types of boundary conditions

(i) Dirichlet: $u(\mathbf{x}, t) = f(\mathbf{x}, t)$ for $\mathbf{x} \in \partial\Omega$.

(ii) Neumann: $\frac{du}{d\mathbf{n}}(\mathbf{x}, t) = \nabla u(\mathbf{x}, t) \cdot \mathbf{n} = g(\mathbf{x}, t)$ for $\mathbf{x} \in \partial\Omega$.

(iii) Robin: $\frac{du}{d\mathbf{n}}(\mathbf{x}, t) + au(\mathbf{x}, t) = h(\mathbf{x}, t)$ for $\mathbf{x} \in \partial\Omega$.

Generically speaking, for every independent variable in the PDE, you will need a boundary condition for each derivative in that variable of the PDE.

For example for the Rayleigh-Bénard convection example, there are two derivatives on the temperature field in each direction. In the setup we have described, we have given a boundary condition on the temperature at the top and bottom of the domain, i.e. two boundary conditions. As of yet we have not specified the horizontal boundaries for this setup. Some common examples are thermally insulated sidewalls (homogeneous Neumann boundaries), an infinite domain (see below) or even periodic boundary conditions.

For the wave equation considered earlier, there are two derivatives in the x direction, and so we would anticipate requiring two boundary conditions in x. This is typically by providing a condition at $x = 0$ and a different condition at $x = L$, the two endpoints of the string. For instance, a guitar string that is plucked will likely have the endpoints fixed so that $u(x = 0, t) = 0 = u(x = L, t)$.

Periodic boundary conditions are a unique set of conditions that are typically used to mathematically simplify the analysis of the PDE. In most cases periodicity is not a physically relevant assumption, but it reduces the complexity of the PDE in the mathematical sense and so it is often used in practice. Periodicity means that anything going out of the domain on the right will enter the domain on the left. This can be thought of as wrapping the domain around in a circular fashion. Physical domains where periodicity actually occurs are circles, the torus (donut), spheres, and cylinders.

- Although there are special circumstances that yield exceptions to this rule, in general it is *NOT* a good idea to prescribe more than one boundary condition at the same location.
- The number of parts of the boundary $\partial\Omega$ of the spatial domain Ω of the function u for the PDE may limit the number of boundary conditions to be imposed.

Remark 9.1.4. We have already seen in the last Chapter that characteristics dictate what auxiliary conditions are relevant and valid, so although the general rule of thumb is to prescribe a boundary condition for every spatial derivative in the PDE, we must realize that this isn't always going to work, and we need to consider the boundary conditions for every PDE separately.

In particular, we need to consider the actual boundary of the prescribed PDE on an individual basis. Most of the examples contained in these notes are very nice and simple (periodic boxes are simple to a mathematician), but reality is often much more messy and involves complicated geometries that must be considered in the context of the effect that boundary conditions will have on the solution.

Example 9.1.5. Consider the evolution of heat in a metal rod of length L. Neglecting the diameter and cross-sectional area of the rod and ignoring all potential sources, the evolution of the heat in the rod is described by

$$u_t = \kappa u_{xx}, \ x \in [0, L], \text{ and } t > 0.$$

Because this is a second-order (in x) PDE, we should expect to specify two boundary conditions on $u(x, t)$. This can be done in a variety of different ways.

(i) One standard approach is to say that the temperature at each end of the rod is known, i.e.

$$u(0, t) = f(t) \qquad u(L, t) = g(t),$$

where usually these functions are held fixed in time, i.e. the temperature is specified exactly at the endpoints of the rod. Typical Dirichlet boundary conditions at both ends of the rod are the constant temperature conditions

$$u(0, t) = T_0 \text{ and } u(L, t) = T_L \text{ for all } t \geq 0.$$

In this case, no matter the initial temperature across the rod, the temperature will limit to (as $t \to$)

$$u(x) = \frac{T_L - T_0}{L} \ x + T_0, \ 0 \leq x \leq L.$$

(ii) The flux of heat is known at each end of the rod, i.e.

$$u_x(0, t) = f(t) \qquad u_x(L, t) = g(t).$$

This will actually lead to problems concerning the uniqueness of the solution to the PDE as we will see later, but may seem like a natural set of boundary conditions to use. Without going any further, can you see why this may not yield a unique solution for $u(x, t)$? Also, why is this specifying what we refer to as the flux?

Typical Neumann boundary conditions at both ends are the perfect insulation conditions

$$u_x(0, t) = 0 \text{ and } u_x(L, t) = 0 \text{ for all } t \geq 0.$$

What would be the long term temperature in the rod with these boundary conditions?

(iii) A mixed set of boundary conditions may occur when we fix the temperature at one end of the rod, and the heat flux at the other end, i.e.

$$u(0,t) = f(t) \qquad u_x(L,t) = g(t).$$

This avoids the complications from specifying only Neumann boundary conditions, and allows for an easier measurement/estimate of the heat flux at $x = L$.

(iv) Specifying that the flux of heat at the endpoint $x = 0$ (equivalently at $x = L$) is determined by the temperature at that endpoint and a time dependent (prescribed) quantity will give the boundary conditions:

$$\alpha_1 u(0,t) + \beta_1 u_x(0,t) = f(t) \text{ for all } t \geq 0$$
$$\alpha_2 u(L,t) + \beta_2 u_x(L,t) = g(t) \text{ for all } t \geq 0$$

for nonzero constants α_1, α_2 and β_1, β_2. These are Robin boundary conditions. Typical Robin boundary conditions at both ends are

$$\alpha_1 u(0,t) + \beta_1 u_x(0,t) = 0 \text{ and } \alpha_2 u(L,t) + \beta_2 u_x(L,t) = 0 \text{ for all } t \geq 0.$$

(v) Suppose that the heat flux and temperature are known at one end, but the other end of the rod is unknown. The boundary conditions in this case would be

$$u(0,t) = f(t) \qquad u_x(0,t) = g(t).$$

The solutions are well-behaved in this case, but it is not very likely that instruments will really let you know the precise temperature and heat flux at the same point, so although this may be possible it is rather unlikely.

As this example illustrates, and as is demonstrated in our earlier discussion on characteristics, it is not enough to simply prescribe the correct number of boundary conditions. One must also be certain that these conditions do not over-prescribe the data, or under-prescribe it.

9.1.2 Infinite Domains

Suppose there is a problem that we are only interested in the local behavior, then is it feasible to completely neglect the boundary conditions?

As we can already see from our earlier discussion on ODEs, it is not a good idea to completely ignore boundary conditions, but such 'boundary conditions' may be rather ambiguous anyway.

Example 9.1.6. Suppose we are concerned with the migration of murder hornets, particularly as they get closer to home. In other words, if we live in North America, we likely don't care how many hornets there are in Patagonia. As far as the migration of hornets goes, we could consider Patagonia to be at ∞. We do know however that there are a finite number of hornets. Thus if $u(\mathbf{x}, t)$ is the concentration of the hornets at a given spatial and temporal location $(\mathbf{x}, t)$ then we would expect the total number of hornets to be constrained such that

$$\int_{\mathbb{R}^2} u(\mathbf{x}, t) d\mathbf{x} < \infty.$$

This implies in turn that $\lim_{|\mathbf{x}|\to\infty} u(\mathbf{x},t) = 0$.

Remark 9.1.7. The previous example may seem rather cooked up, but it is reasonable to approximate different problems by an infinite domain, particularly if the scales of interest are several orders of magnitude smaller than the entire physical domain. There are several instances where the infinite domain assumption is a valuable approximation.

9.2 Initial Conditions, well-posedness and the energy method

9.2.1 Initial conditions

In the study of ODEs we most often wanted to understand the evolution of the solution independent of the initial conditions. We were interested in the behavior of the entire phase space, or space of possible solutions starting at any initial condition. This is a much more difficult problem in PDEs, primarily because the phase space is actually infinite-dimensional. Instead, for most PDEs we instead solve each PDE individually for a specific initial condition. We are still very interested in the nature of solutions for any class of initial conditions, but this is a much more difficult problem, and one that we don't get into here.

If only one time derivative is present (Conservation Laws are a good example of this), a single initial condition is all that is necessary, for example $u(\mathbf{x}, t = 0) = f(\mathbf{x})$. If more than one time derivative is present (see the wave equation for example), then it is necessary to specify two initial conditions (just as in the ODEs case), for example $u(\mathbf{x}, t = 0) = f(\mathbf{x})$ and $u_t(\mathbf{x}, t = 0) = g(\mathbf{x})$.

Example 9.2.1. Suppose we are considering the heat equation from Example 9.1.5. We can specify an initial condition in this case because we may know the initial distribution of heat in the metallic rod. Some potential initial conditions for $x \in [0, L]$ are:

$$\begin{aligned} u(x, t = 0) &= \sin\left(2\pi \frac{x}{L}\right) \\ u(x, t = 0) &= \tanh(x) \\ u(x, t = 0) &= (x - L/2)^2 \\ u(x, t = 0) &= x(x - L). \end{aligned}$$

The trick here is to make sure that the initial condition matches with the corresponding boundary conditions. If they do not match, there will be an initially very rapid transition to ensure that the boundary condition is satisfied for $t > 0$. This can be very problematic (especially numerically), and physically doesn't make very much sense.

Remark 9.2.2. There are occasions where a PDE may be specified with what we may call an endpoint condition (we see more of this for ODEs in optimal control), that is the solution is specified at some point in time $t = T$ and the solution is sought for $t < T$. Remember that we brought this up for a snowball thrown at you, i.e. you know the snowball hit your face (final time T), but you want to find out where it came from.

Remark 9.2.3. A final remark is in order regarding PDEs where time is not an independent variable of interest, for example the elliptic equations referred to at the end of this Chapter. In this setting, initial conditions are not relevant and instead different types of boundary conditions are all that is required. The same principles apply even in these circumstances though, where now one of the spatial variables may play the role of the temporal one. In addition, there are instances of PDEs where time and spatial coordinates are no longer the relevant independent variables in which case one would need to proceed by analogy to the generic cases considered here.

As a final example, we return to the murder hornets that are potentially invading North America.

Example 9.2.4. Continuing on the murder hornets example, does it make sense to specify an initial condition $u(\mathbf{x}, t = 0) = \cos(\|\mathbf{x}\|)$, considering what the boundary conditions are (at infinity) in this case? What about $u(\mathbf{x}, t = 0) = \exp(-\|\mathbf{x}\|^2)$?

9.2.2 Well-posedness

From now on whenever we refer to a given PDE we will implicitly also be referring to the relevant auxiliary conditions. The question we really want to answer is what types of auxiliary conditions are appropriate for each PDE, but what does appropriate mean in this context? We recall the following definition which was introduced in the previous Chapter when working with characteristics TODO: get definition reference

Definition 9.2.5. *A well-posed problem is one in which:*

- *Solutions exist.*
- *The solution is unique.*
- *The solution does not change drastically due to slight changes in the auxiliary conditions (continuous dependence on auxiliary conditions).*

Unlike ODEs there is not a comprehensive theory for the well-posedness of PDEs. Instead each PDE must be considered on a case by case basis. In fact this is an entire field of mathematics, and typically requires several years of analysis courses to understand even the simplest arguments. For this reason, we do not go into very many details here, but do show one example where the uniqueness of the solution is guaranteed.

Before proceeding we emphasize that this is only one approach that will guarantee the uniqueness of solutions. There are multiple other approaches, just for establishing uniqueness of solutions let alone methods for constructing an existing solution, or establishing continuous dependence on the parameters of the problem. We will focus on this particular method because it is rather straightforward to work through, and generalizes to a wide class of parabolic equations.

Example 9.2.6. [Energy Method]
Consider the advection diffusion equation in one dimension,

$$\theta_t + u\theta_x = \kappa\theta_{xx},$$

where κ and u are constants. This is augmented with the auxiliary conditions:

$$\theta(x,0) = f(x), \qquad \theta(0,t) = 0, \qquad \text{and} \qquad \theta(L,t) = h(t),$$

where $0 \leq x \leq L$, i.e. Dirichlet boundary conditions with a specific initial condition (that we assume matches the boundary conditions for now). We desire to show that a solution to this PDE is unique.

Suppose to the contrary that there are two solutions $\theta_1(x,t)$ and $\theta_2(x,t)$ that satisfy this PDE including the auxiliary conditions. Then if we let $w(x,t) = \theta_1(x,t) - \theta_2(x,t)$ we see that $w(x,t)$ satisfies:

$$w_t = -uw_x + \kappa w_{xx}, \qquad w(x,0) = 0, \quad w(0,t) = 0 \quad \text{and} \quad w(L,t) = 0.$$

Defining the 'energy' of the system as

$$E(t) = \int_0^L w^2(x,t)dx.$$

It follows that the time derivative of the energy then satisfies (we are assuming that our solutions are sufficiently smooth so that we can commute derivatives, and carry them through integrals etc.):

$$\begin{aligned}\frac{dE(t)}{dt} &= 2\int_0^L ww_t dx \\ &= -2u\int_0^L ww_x dx + 2\kappa\int_0^L ww_{xx}dx,\end{aligned}$$

but integrating by parts we see that

$$\begin{aligned}\int_0^L ww_x dx &= w^2\big|_{x=0}^{x=L} - \int_0^L ww_x dx \\ &= -\int_0^L ww_x dx \\ \Rightarrow \int_0^L ww_x dx &= 0,\end{aligned}$$

where we used the boundary conditions at $x = 0$ and $x = L$. In addition we can integrate by parts on the final integral to see that

$$\begin{aligned}\int_0^L ww_{xx}dx = ww_x|_{x=0}^{x=L} - \int_0^L w_x^2 dx = -\int_0^L w_x^2 dx \\ \Rightarrow \frac{dE(t)}{dt} = -2\kappa\int_0^L |w_x|^2 dx \leq 0.\end{aligned}$$

Thus

(i) $E(t)$ is decreasing in time,

(ii) $E(t) \geq 0$ by definition, and

(iii) $E(0) = 0$.

Hence $E(t) = 0$ for all time t. Therefore $w(x,t) = 0$, and the solutions $\theta_1(x,t)$ and $\theta_2(x,t)$ are identical for all time, and solutions to this PDE (with these specific boundary conditions) are unique.

Remark 9.2.7. This energy method for proving uniqueness is of course relying on a few key assumptions such as the regularity (smoothness) of the solutions. In addition, we have assumed that at least one solution does exist. These are all assumptions that should be questioned for every PDE, and should be an issue of great concern if you are trying to simulate a complicated problem that doesn't seem to be giving you the results you would expect. It may be that the numerical discretization is in error, and it may also be a result of the ill-posedness of the original problem.

We don't spend an inordinate amount of time on justifying the well-posedness of a given PDE, but it is a very important concept, and one that should not be set aside and relegated to those hard core mathematical analysts who like to revisit such existential questions. In particular, the techniques that have arisen from the study of well-posedness of a given PDE have illustrated many different phenomenon that would not have been realized otherwise. For instance, generalizations of the energy method in the previous example are useful for a wide variety of other problems and circumstances as demonstrated in the exercises.

Example 9.2.8 (Hadamard's Example). Consider the PDE

$$u_{tt} + u_{xx} = 0, \quad t > 0, \quad x \in \mathbb{R},$$

with initial conditions

$$u(x,0) = 0, \quad u_t(x,0) = 0.$$

Note that this PDE is Laplace's equation in x and t. We can see readily enough that the solution to this problem is the zero solution $u(x,t) = 0$ for all x, t. Now, if we modify the initial conditions to

$$u(x,0) = 0, \quad u_t(x,0) = 10^{-m}\sin(10^m x),$$

which should be a very 'small' change in the auxiliary conditions for sufficiently large m, then the solution will be given by:

$$u(x,t) = 10^{-2m}\sin(10^m x)\sinh(10^m t).$$

As t increases, this is nothing like the zero solution, but instead experiences exponential growth in t with exponent 10^m which is very fast. This means that this problem is not well posed, i.e. Laplace's equation stated as an initial value problem is not well-posed because small changes to the initial data can lead to drastic changes in the solution.

Remark 9.2.9. The previous example is the classical example of an ill-posed problem, and is often referred to as the backward heat equation. If we think of what this would mean physically, we need to recall one method of deriving the heat equation from conservation laws, that is Fick's Law is used to determine the flux. Fick's Law states that a concentration of something (temperature in this case) will move towards regions of lower concentrations, meaning the density will diffuse throughout the domain. The backward heat equation is the opposite of this setting, the temperature (or other density variable) will move from areas of low concentration to areas of high concentration. One can see how physically this could create some problems. For instance, this would lead to an infinite spike in temperature at a single point rather than the temperature diffusing to be uniform throughout the prescribed domain. Thus, not only is Hadamard's example mathematically ill-posed, but it is physically unrealistic as well.

9.3 Parabolic equations: the canonical example of the heat equation

The heat equation which we have already seen a bit of is:

$$u_t - \kappa \Delta u = 0, \qquad \mathbf{x} \in \Omega \tag{9.2}$$

with possible boundary conditions

$$\begin{aligned} &u(\mathbf{x}, t) = 0, \quad \text{for } \mathbf{x} \in \partial\Omega \qquad \text{(Dirichlet BCs)} \\ \text{or } &\frac{du}{d\mathbf{n}}(\mathbf{x}, t) = \nabla u(\mathbf{x}, t) \cdot \mathbf{n} = 0, \quad \text{for } \mathbf{x} \in \partial\Omega \qquad \text{(Neumann)}, \end{aligned}$$

and initial conditions

$$u(\mathbf{x}, t = 0) = f(\mathbf{x}), \quad \text{for } \mathbf{x} \in \Omega.$$

9.3.1 Derivation of the Heat Equation

We can derive the heat equation as a conservation law, where the flux is given by Fick's law, i.e. $J(u) = \kappa \nabla u$, where κ is a constant that is typically referred to as the thermal diffusivity of the material in question. The conservation law in integral form becomes

$$\int_E u_t \, d\mathbf{x} = \int_{\partial E} \kappa \nabla u \cdot \mathbf{n} \, dA.$$

By the Divergence Theorem the conservation law becomes

$$\int_E u_t \, d\mathbf{x} = \int_E \kappa \, \mathrm{div}(\nabla u) \, d\mathbf{x}.$$

Since

$$\mathrm{div}(\nabla u) = \nabla \cdot \nabla u = \Delta u$$

we obtain

$$\int_E u_t \, d\mathbf{x} = \int_E \kappa \Delta u \, d\mathbf{x}.$$

Combining the two integrals into one gives

$$\int_E \left(u_t - \kappa \Delta u \right) \, d\mathbf{x} = 0.$$

Since this holds for arbitrary subdomains E of Ω we obtain the PDE

$$u_t - \kappa \Delta u = 0.$$

The heat equation is also referred to as the *diffusion equation*, and models the diffusion of the concentration of some pertinent quantity in a medium where Fick's law applies. As noted previously, Fick's law basically implies that the concentration $u(\mathbf{x}, t)$ will move from areas of high concentration to regions of low concentration, i.e. 'diffuse' everywhere throughout the domain.

9.3.2 Energy Decay

One of the key properties of a parabolic system is that the total energy of the system is decaying. To apply the Energy Method to the heat equation we need one of Green's Identities we saw in the previous Chapter.

Suppose scalar functions $u(\mathbf{x},t)$ and $w(\mathbf{x},t)$ are $C^1\left(\overline{\Omega \times I}\right) \cap C^2(\Omega \times I)$ where I is an interval. Then Green's First Identity is

$$\int_\Omega \left(u\Delta w + \nabla u \cdot \nabla w\right)\, d\mathbf{x} = \int_{\partial\Omega} u\frac{dw}{d\mathbf{n}}\, dA.$$

We can split the left-hand side and rearrange to obtain the identity

$$\int_\Omega u\Delta w\, d\mathbf{x} = -\int_\Omega \nabla u \cdot \nabla w\, d\mathbf{x} + \int_{\partial\Omega} u\frac{dw}{d\mathbf{n}}\, dA. \tag{9.3}$$

Definition 9.3.1. *For a solution $u(\mathbf{x},t)$ of the heat equation with either the Dirichlet boundary conditions $u(\mathbf{x},t) = 0$ for all $\mathbf{x} \in \partial\Omega$ and for all $t \geq 0$, or the Neumann boundary conditions $\nabla u(\mathbf{x},t) \cdot \mathbf{n} = 0$ for all $\mathbf{x} \in \partial\Omega$ and for all $t \geq 0$, the* total energy *of $u(\mathbf{x},t)$ at time t is*

$$E(t) = \frac{1}{2}\int_\Omega |u(\mathbf{x},t)|^2\, d\mathbf{x}.$$

In what follows we assume that $u(\mathbf{x},t)$ is smooth enough to justify all computations.

Theorem 9.3.2. *The total energy*

$$E(t) = \frac{1}{2}\int_\Omega |u|^2 d\mathbf{x},$$

defined for the solution $u(\mathbf{x},t)$ of (9.2) is never increasing in time for either the Dirichlet or Neumann boundary conditions provided.

Remark 9.3.3. The proof is remarkably similar to the energy method we used to prove the uniqueness of solutions to the advection diffusion equation. This indicates the real reason for the name 'energy method'.

Proof. Computing the time derivative of the energy,

$$\begin{aligned}
\frac{dE(t)}{dt} &= \int_\Omega u u_t d\mathbf{x} = \kappa\int_\Omega u\Delta u d\mathbf{x} \\
&= -\kappa\int_\Omega |\nabla u|^2 d\mathbf{x} + \kappa\int_{\partial\Omega} u\nabla u \cdot \mathbf{n} dA \\
&= -\kappa\int_\Omega |\nabla u|^2 d\mathbf{x} \\
&\leq 0,
\end{aligned}$$

where the boundary integral vanishes due to the prescribed boundary conditions. Thus the total integrated energy is not increasing in time. ∎

Remark 9.3.4. This is one of the key aspects of parabolic equations that sets them apart form hyperbolic ones. There is a clear diffusive behavior that proves invaluable when establishing uniqueness of solutions to parabolic equations. Such diffusion, or decay of energy is not present for the most part in hyperbolic systems. This decay of energy should feel eerily familiar to our introduction, discussion, and use of Lyapunov functions for ODEs.

9.3.3 The Maximum Principle

The maximum principle is one of the undergirding and fundamentally important concepts in all of PDEs. Although this may not be apparent when we go over this here, just trust me (or should this be 'us'?), it is extremely important. What is more, there are several different versions of the maximum principle with a variety of assumptions and levels of justification. We will not consider the most general case, nor we will we consider the most specific, but the reader should get the basic idea from the discussion presented here. In addition, we do not broach the maximum principle for the multi-dimensional form of the heat equation here, although the same ideas apply.

Before stating and proving the maximum principle that we will use here, we require the following lemma.

Lemma 9.3.5. *For $D = [a,b] \times [0,T]$ and $\kappa > 0$ suppose $v : D \to \mathbb{R}$ is a C^2 function. If*

$$v_t - \kappa v_{xx} < 0$$

for all (x,t) in $(a,b) \times (0,T]$, then $v(x,t)$ does not obtain a maximum on $(a,b) \times (0,T]$.

Proof. Suppose that v satisfies $v_t - \kappa v_{xx} < 0$ on $(a,b) \times (0,T]$. We will consider two cases for the domain and ensure that a maximum doesn't occur on either part of the domain:

(i) $(a,b) \times (0,T)$

(ii) $(a,b) \times \{T\}$.

(i) Suppose to the contrary that v obtains a maximum at a point $(x_0,t_0) \in (a,b) \times (0,T)$. Then v has a critical point at (x_0,t_0), which implies that

$$Dv(x_0,t_0) = \mathbf{0}^{\mathsf{T}} \;\Rightarrow\; v_x(x_0,t_0) = 0 \qquad \text{and} \qquad v_t(x_0,t_0) = 0.$$

Moreover, $D^2v(x_0,t_0)$ is negative semidefinite by the second-order necessary condition (see Volume 2, Theorem 12.1.12). This implies that $v_{xx}(x_0,t_0) = \mathbf{e}_1^{\mathsf{T}} D^2(x_0,t_0)\mathbf{e}_1 \leq 0$, where $\mathbf{e}_1 = (1,0)$.

Hence we have $v_t(x_0,t_0) = 0$ and $v_{xx}(x_0,t_0) \leq 0$ which imply

$$v_t(x_0,t_0) - \kappa v_{xx}(x_0,t_0) \geq 0,$$

but by assumption $v_t - \kappa v_{xx} < 0$ at all points of $(a,b) \times (0,T)$. This yields the contradiction, and hence the maximum cannot occur at any point of $(a,b) \times (0,T)$.

(ii) Now suppose that a maximum occurs on the boundary $(a,b) \times \{T\}$, so there is a point (x_0,T) with $x \in (a,b)$ at which $v(x_0,T)$ is maximum. The curve $s \mapsto v(x_0+s,T)$ lies on the boundary $t = T$ and for all sufficiently small s stays away from the points (a,T) and (b,T). By the same argument above we obtain $v_{xx}(x_0,T) \leq 0$.

To determine the sign of $v_t(x_0,T)$ use the one-sided curve $s \mapsto v(x_0,T+s)$ for $s \leq 0$ and sufficiently close to 0, i.e., we approach the boundary $t = T$ from the interior of D. The Taylor expansion of this curve gives

$$v(x_0,T+s) = v(x_0,T) + v_t(x_0,T)s + \cdots .$$

If $v_t(x_0,T) < 0$ then for $s < 0$ and s sufficiently close to 0, the value of v along the curve would get bigger, contradicting that the maximum is achieved on the boundary $(a,b) \times \{T\}$. Thus $v_t(x_0,T) \geq 0$.

Combining this with $v_t(x_0,T) \geq 0$ gives

$$v_t(x_0,T) - \kappa v_{xx}(x_0,T) \geq 0.$$

But this contradicts $v_t - \kappa v_{xx} < 0$ for all $(x,t) \in (a,b) \times (0,T]$.

Therefore the maximum value of v is realized when $x = a$, $x = b$, or when $t = 0$. ∎

Theorem 9.3.6 (Weak Maximum Principle). *For $D = [a, b] \times [0, T]$, a constant $\kappa > 0$, and $u : D \to \mathbb{R}$ a C^2 function, if u satisfies*

$$u_t - \kappa u_{xx} \leq 0 \text{ for all } (x, t) \in (a, b) \times (0, T],$$

then u attains its maximum value when $x = a$, $x = b$, or when $t = 0$.

Proof. For $\varepsilon > 0$ set

$$v(x, t) = u(x, t) + \varepsilon x^2.$$

Then

$$v_t = u_t \text{ and } v_{xx} = u_{xx} + 2\varepsilon.$$

Since $u_t - \kappa u_{xx} \leq 0$ for all $(x, t) \in (a, b) \times (0, T]$ and $\kappa > 0$ it follows that

$$v_t - \kappa v_{xx} = u_t - \kappa(u_{xx} + 2\varepsilon) = u_t - \kappa u_{xx} - 2\kappa\varepsilon < 0 \text{ for all } (x, t) \in (a, b) \times (0, T].$$

By Lemma 9.3.5 the maximum of $v(x, t)$ occurs when $x = a$, $x = b$, or $t = 0$.

Let Γ to be the part of the boundary of D where the maximum of v must occur:

$$\Gamma = \{(a, t) : t \in [0, T]\} \cup \{(x, 0) : x \in [a, b]\} \cup \{(b, t) : t \in [0, T]\}.$$

Since $v(x, t) = u(x, t) + \varepsilon x^2$ it follows that $u(x, t) \leq v(x, t)$ so that

$$\max_{(x,t)\in D} u(x, t) \leq \max_{(x,t)\in D} v(x, t) = \max_{(x,t)\in \Gamma} v(x, t).$$

The maximum of εx^2 on $[a, b] \times [0, T]$ is

$$\max_{(x,t)\in \Gamma} \varepsilon x^2 = \varepsilon \max\{a^2, b^2\}.$$

Since $v(x, t) = u(x, t) + \varepsilon x^2$ and Γ is compact it follows that

$$\max_{(x,t)\in \Gamma} v(x, t) \leq \max_{(x,t)\in \Gamma} u(x, t) + \max_{(x,t)\in \Gamma} \varepsilon x^2 = \max_{(x,t)\in \Gamma} u(x, t) + \varepsilon \max\{a^2, b^2\}.$$

Thus we obtain

$$\max_{(x,t)\in D} u(x, t) \leq \max_{(x,t)\in \Gamma} u(x, t) + \varepsilon \max\{a^2, b^2\}.$$

As this holds for all $\varepsilon > 0$ we obtain

$$\max_{(x,t)\in D} u(x, t) \leq \max_{(x,t)\in \Gamma} u(x, t).$$

Because $\Gamma \subset D$ the opposite inequality holds, i.e.,

$$\max_{(x,t)\in D} u(x, t) \geq \max_{(x,t)\in \Gamma} u(x, t)$$

Therefore we obtain

$$\max_{(x,t)\in D} u(x, t) = \max_{(x,t)\in \Gamma} u(x, t),$$

giving the result. ∎

Remark 9.3.7. The Weak Maximum Principle applies to the one-dimensional heat equation

$$u_t - \kappa u_{xx} = 0,\ 0 \le x \le L,\ t \ge 0.$$

For given initial conditions

$$u(x,0) = f(x),\ 0 \le x \le L,$$

and consistent Dirichlet boundary conditions

$$u(0,t) = g(t) \text{ and } u(L,t) = h(t) \text{ for } t \ge 0,$$

the maximum value of u on the domain $D = [a,b] \times [0,T]$ occurs when

(i) $x = 0$, i.e., $u(0,t) = g(t)$, $0 \le t \le T$,

(ii) $x = L$, i.e., $u(L,t) = h(t)$, $0 \le t \le T$, or

(iii) $t = 0$, i.e., $u(x,0) = f(x)$, $0 \le x \le L$.

Remark 9.3.8. The Weak Maximum Principle gives a uniform or L^∞ bound on the solution $u(x,t)$ of the one-dimensional heat equation. Suppose that $u(x,t)$ satisfies

$$\begin{aligned} &u_t - \kappa u_{xx} = 0,\ 0 \le x \le L,\ t \ge 0,\\ &u(x,0) = f(x) \ge 0,\ 0 \le x \le L,\\ &u(0,t) = 0 \text{ and } u(L,t) = 0 \text{ for } t \ge 0, \end{aligned}$$

where $f(x) > 0$ for some $x \in (0,L)$, and $f(0) = 0$ and $f(L) = 0$ for consistency.

By the Weak Maximum principle the sup-norm or max-norm of u on the domain $D = [0,L] \times [0,T]$ of the solution is

$$\|u\|_\infty = \sup_{(x,t)\in D} |u(x,t)| = \sup_{(x,t)\in \Gamma} |u(x,t)|$$

where

$$\Gamma = \{(0,t) : t \in [0,T]\} \cup \{(x,0) : x \in [0,L]\} \cup \{(L,t) : t \in [0,T]\}.$$

Since $u(0,t) = 0$ and $u(L,0) = 0$ for all $t \ge 0$, the sup-norm simplifies to

$$\|u\|_\infty = \sup_{x\in[0,L]} |u(x,0)| = \sup_{x\in[0,L]} |f(x)|.$$

Notice this is independent of $T > 0$ so that

$$\sup_{x\in[0,L]} |f(x)| = \sup_{(x,t)\in[0,L]\times[0,\infty)} |u(x,t)|.$$

If you look carefully at the proofs of Lemma 9.3.5 and the Weak Maximum Principle above, you will notice that it really didn't depend on the time interval being of the form $[0,T]$, i.e., that 0 was an endpoint. All that mattered was that time was taken in a compact interval. Hence we can use an appropriate adaptation of the Weak Maximum Principle and the arguments above to see that

$$\sup_{(x,t)\in[0,L]\times[t_*,T]} |u(x,t)| = \sup_{x\in[0,L]} |u(x,t_*)| = \sup_{(x,t)\in[0,L]\times[t_*,\infty)} |u(x,t)|.$$

The diffusive property of the heat equation is expressed in the L^∞ norm as: for $0 \le t_1 < t_2 < \infty$,

$$\sup_{(x,t)\in[0,L]\times[t_1,\infty)} |u(x,t)| \ge \sup_{(x,t)\in[0,L]\times[t_2,\infty)} |u(x,t)|$$

and ultimately

$$\lim_{t\to\infty} \sup_{(x,t)\in[0,L]\times[t,\infty)} |u(x,t)| = 0.$$

Contrast this with the Energy Method where

$$E(t) = \frac{1}{2}\int_0^L |u|^2\, dx = \frac{1}{2}\|u\|_2^2 \text{ and } \frac{dE}{dt} = -\kappa \int_0^L |u_{xx}|^2\, dx = -\kappa \|u_{xx}\|_2^2,$$

which relies on the L^2-norm of u and its derivatives.

On compact spatial domains, the L^∞-norm is stronger than the L^2-norm, i.e., every L^∞ function on a compact set is an L^2 function, but not every L^2 function on a compact set is an L^∞ function.

Remark 9.3.9. Just as the energy method was used to yield unique solutions of the heat equation in L^2, we can use the maximum principle to prove there are unique solutions of the heat equation in L^∞.

9.3.4 Equilibrium/Elliptic Equations

Elliptic equations are a vibrant area of active research and are fundamentally of significant interest in several different applications. That being said, we do not have the time nor space to devote to a sufficiently detailed investigation of these equations, so we will provide only a precursory overview.

To introduce the concepts of elliptic equations, consider the heat equation as discussed substantially above:

$$\begin{aligned} u_t &= \kappa\Delta u + f(\mathbf{x}), \qquad \mathbf{x}\in\Omega \\ u|_{\mathbf{x}\in\partial\Omega} &= G(\mathbf{x},t), \qquad u(\mathbf{x},t=0) = h(\mathbf{x}), \end{aligned}$$

where we will assume that $\lim_{t\to\infty} G(\mathbf{x},t) = g(\mathbf{x})$. It is reasonable to suppose that after a sufficiently long time, that $u(\mathbf{x},t)$ will reach a steady state independent of initial conditions, i.e.

$$\begin{aligned} -\kappa\Delta u &= f(\mathbf{x}), \qquad \mathbf{x}\in\Omega \\ u(\mathbf{x}) &= g(\mathbf{x}), \qquad \mathbf{x}\in\partial\Omega. \end{aligned}$$

This is called Poisson's equation. If $f(\mathbf{x}) = 0$ then this is Laplace's equation. Note that up to a change of sign, this is the same as the steady state for the wave equation as well. Thus, we often consider elliptic equations as those obtained from considering the steady-state approximation of parabolic and/or hyperbolic equations.

Laplace's equation arises in other contexts as well. If $f(z) = u(x,y) + iv(x,y)$ where $z = x + iy$ is holomorphic (meaning the complex valued derivative exists and is well-defined in a neighborhood of every point in the complex plane), then one can show (in the homework of course) that $\Delta u = 0$ and $\Delta v = 0$. Solutions such as these of Laplace's equation are called *harmonic functions*.

Harmonic functions have many nice properties, enough so that entire books may be written about them, but for our purposes, suffice it to say that they satisfy equilibrium solutions of the heat equation. As a teaser, we demonstrate just one surprising property that harmonic functions have:

For a harmonic function $u(x,y)$ defined on $\Omega\in\mathbb{R}^2$ the following Mean Value Property is satisfied:

$$u(x,y) = \frac{1}{2\pi}\int_0^{2\pi} u((x,y) + r(\cos\theta,\sin\theta))\, d\theta$$

whenever the curve

$$\theta \mapsto r(\cos\theta, \sin\theta),\ \theta \in [0, \pi],$$

belongs to Ω. This says that the value of $u(x, y)$ is determined by the average of the values of u on any circle in Ω centered at (x, y). This follows from taking the real or imaginary part of Gauss' Mean Value Theorem for holomoprhic functions (Volume 1, Corollary 11.4.3).

Remark 9.3.10. Clearly there is much more to be said about parabolic and elliptic PDEs, in fact there are entire textbooks dedicated to the subject. One need not suppose that the lack of material presented here is due to a lack of interest in the topic. Instead it is a matter of choosing the time and space wisely, and because we have already discussed many of the ideas relating to these equations previously, for example in the derivation of conservation laws in the previous chapter.

9.4 *Finite Difference approximations of the heat equation

We will consider numerical approximations of the one-dimensional heat equation only. Clearly there is much more interest in exploring higher-dimensional parabolic PDEs, and even considering more complicated examples than the heat equation, but once again finite time constraints force us to pick and choose. We assume one spatial dimension with the thin metal rod being of length $L = 1$, the thermal diffusivity $\kappa = 1$, and constant Dirichlet boundary conditions, namely the PDE

$$\begin{aligned} &u_t = u_{xx},\ 0 \le x \le 1,\ t \ge 0,\\ &u(x, 0) = f(x),\ 0 \le x \le 1,\\ &u(0, t) = a,\ u(1, t) = b,\ \text{for all } t \ge 0. \end{aligned}$$

where we are assuming the consistency conditions

$$f(0) = a \text{ and } f(1) = b.$$

We also restrict our attention here to finite difference methods even though there are several other methods that could be employed for this specific equation. Even with such restrictions, we will find that there are plenty of choices to be made, and there are plenty of opportunities to complicate the numerical approximation.

We first return to the forward Euler temporal, and centered spatial finite difference approximation to $u_t = u_{xx}$ given in (7.10), i.e.

$$\frac{u(x, t + \Delta t) - u(x, t)}{\Delta t} = \frac{u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)}{(\Delta x)^2}.$$

Solving this for $u(x, t + \Delta t)$ gives the update formula

$$u(x, t + \Delta t) = u(x, t) + \frac{\Delta t}{(\Delta x)^2}\big(u(x + \Delta x, t) - 2u(x, t) + u(x - \Delta x, t)\big).$$

Recall that the CFL number is the quantity

$$\text{CFL} = \frac{\Delta t}{(\Delta x)^2}$$

which should be smaller than $]\frac{1}{2}$ for stability of the numerical algorithm.

The update formula determines the value $u(x, t + \Delta t)$ using the values of

$$u(x - \Delta x, t),\ u(x, t),\ u(x + \Delta, t).$$

To see how this is implemented in Python, we perform the following steps:

- Partition the spatial domain $[0, 1]$ into subintervals $[x_{i-1}, x_i]$ of equal length $\Delta x = \frac{1}{M}$ with endpoints
$$x_i = i\Delta x, i = 0, 1, \ldots, M.$$
- For $T > 0$ subdivide the time interval $[0, T]$ into subinterval $[t_{j-1}, t_j]$ of equal length $\Delta t = \frac{T}{K}$ and endpoints
$$t_j = j\Delta t,\ j = 0, 1, 2, \ldots, K,$$
where we assume that the integer K is chosen so that
$$\frac{T/K}{(\Delta x)^2} = \frac{\Delta t}{(\Delta x)^2} = \text{CFL} < 1.$$
- The points (x_i, t_j) for $(i, j) \in \{0, 1, 2, \ldots, M\} \times \{0, 1, 2, \ldots, K\}$ partition the rectangle $[0, 1] \times [0, T]$ into subrectangles.
- The update formula then takes the form
$$u(x_i, t_{j+1}) = u(x_i, t_j) + \frac{\Delta t}{(\Delta x)^2}\big(u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)\big).$$

This leads to the following update strategy:

- We start with the first time row $t = t_0$. The values of $u(x_i, t_0)$, $i = 0, 1, 2, \ldots, M$, are all know form either boundary conditions or initial conditions.
- For $i = 0$ the value $u(x_0, t_0) = a$, the boundary condition at $x = 0$ since $x_0 = 0$.
- For $i = 1, 2, \ldots, M - 1$, the value of $u(x_i, t_0)$ is that of $f(x_i)$, the initial condition.
- For $i = M$, the value $u(x_M, t_0) = b$, the boundary condition at $x = 1$ since $x_M = 1$.
- Next we compute the values of u on the time row $t = t_1$.
- The grid point (x_0, t_1) is on the left boundary where the value of u is a, so we set
$$u(x_0, t_1) = a.$$
- The grid point (x_M, t_1) is on the right boundary where the value of u is b, so we set
$$u(x_M, t_1) = b.$$
- At the grid point (x_1, t_1) the update formula gives
$$\begin{aligned} u(x_1, t_1) &= u(x_1, t_0) + \frac{\Delta t}{(\Delta x)^2}\big(u(x_2, t_0) - 2u(x_1, t_0) + u(x_0, t_0)\big) \\ &= f(x_1) + \frac{\Delta t}{(\Delta x)^2}\big(f(x_2) - 2f(x_1) + a\big). \end{aligned}$$
- At the grid point (x_2, t_2) the update formula gives
$$\begin{aligned} u(x_2, t_1) &= u(x_2, t_0) + \frac{\Delta t}{(\Delta x)^2}\big(u(x_3, t_0) - 2u(x_2, t_0) + u(x_1, t_0)\big) \\ &= f(x_2) + \frac{\Delta t}{(\Delta x)^2}\big(f(x_3) - 2f(x_2) + f(x_1)\big). \end{aligned}$$
- Going near to the end of the time row $t = t_1$, at the grid point (x_{M-1}, t_1) the update formula gives
$$\begin{aligned} u(x_{M-1}, t_1) &= u(x_{M-1}, t_0) + \frac{\Delta t}{(\Delta x)^2}\big(u(x_M, t_0) - 2u(x_{M-1}, t_0) + u(x_{M-2}, t_0)\big) \\ &= f(x_{M-1}) - \frac{\Delta t}{(\Delta x)^2}\big(b - 2f(x_{M-1}) + f(x_{M-2})\big). \end{aligned}$$

- Having established the values of u on the time row $t = t_1$ we can proceed to the time row $t = t_2$ and apply the update formula to compute the values of u when $t = t_2$ at the equally spaced x_i points.
- Continuing this for each time row gives a numerical approximation of the solution u on the grid of points in the compact domain $[0, 1] \times [0, T]$.

What happens for other boundary conditions, where the value of $u(x, t)$ is not fixed, i.e. non-Dirichlet boundaries?

```
import numpy as np

def forward_euler_1D_heat(a,b,CFL,T,x=[],u0=[]):
    """solution of the 1D heat equation u_t=u_xx.
    u(0) = a, and u(1) = b are the BCs
    CFL=deltaT/(deltaX^2)
    T is the final time that the simulation will be run
    x is the pre-selected spatial variable,
    and should be uniformly distributed
    u0 is the initial condition, and must be the same size as x

    Returns an array u which is the solution at all
    time steps up to T
    as dictated by the time-step chosen through the
    specified 'CFL' number."""
    N = len(x)
    deltaX = 1/(N-1)
    deltaT = CFL*(deltaX)**2
    u = []
    u.append(u0)
    v = u0.copy() #a dummy variable for the update process
    for nn in range(1,int(T/deltaT)):
        # double check that the BCs are maintained
        v[0] = a
        v[-1] = b
        for jj in range(1,N-1):
            v[jj] = u[nn-1][jj] +
            CFL*(u[nn-1][jj-1]-2*u[nn-1][jj]+u[nn-1][jj+1])
        u.append(v.copy())
    return x,u
```

Algorithm 9.1: Python implementation of the one-dimensional heat equation for forward Euler.

Example 9.4.1. As we already have seen, we don't want to revert to purely Neumann boundary conditions for the heat equation, but we can consider mixed boundary conditions, i.e.

$$u_x(0,t) = 0, \quad u(1,t) = 0, \tag{9.4}$$

for example. For this particular case, we don't need to treat the boundary at $x = 1$ any differently, but we do need to consider carefully what we do near $x = 0$. We need to somehow fix the derivative, without fixing the actual value of the function. There are a variety of ways to do this, but we will consider fixing an approximation of the derivative. For example, suppose that we approximate $u_x(0,t)$ by a centered finite difference as

$$0 = u_x(0,t) = \frac{u(\Delta x, t) - u(-\Delta x, t)}{2\Delta x}. \tag{9.5}$$

We can then derive an approximation for $u(-\Delta x, t)$ as

$$u(-\Delta x, t) \approx u(\Delta x, t), \tag{9.6}$$

and we can use this to update $u(0,t)$ in the update scheme outlined in Algorithm 9.1, i.e.

$$u(0, t+\Delta t) = u(0,t) + \frac{2\Delta t}{(\Delta x)^2}\left(u(\Delta x, t) - u(0,t)\right). \tag{9.7}$$

Remark 9.4.2. The previous example creates the function value at the fictitious point $(-\Delta x, t)$ in order to handle the Neumann boundary condition. This is a pretty standard approach and the fictitious point is referred to as a ghost point, or ghost boundary condition. If our boundary conditions had been more complicated (Neumann, but not homogeneous) then we would need to be a bit more careful with how we develop the scheme, but this particular condition worked out quite nicely so we don't even have to keep track of the ghost value explicitly. In higher dimensions this is certainly not the case anymore.

We now note that Algorithm 9.1 can be recast in terms of matrix vector multiplication. To obtain it we simplify the boundary conditions to homogeneous Dirichlet:

$$\begin{aligned} &u_t = u_{xx}, \ 0 \le x \le 1, \ t \ge 0, \\ &u(x,t) = f(x), \ 0 \le x \le 1, \\ &u(0,t) = 0, \ u(1,t) = 0, \ \text{for all } t \ge 0. \end{aligned}$$

For each time row j, the homogeneous Dirichlet boundary conditions fix the terms $u(x_0, t_j) = 0$ and $u(x_M, t_j) = 0$, while the remaining values of $u(x_i, t_{j+1})$, $i = 1, \ldots, M-1$, are completely determined by the values of $u(x_i, t_j)$, $i = 1, 2, \ldots, M-1$. On each time row there are $M-1$ points x_i where values of u are computed.

In the update formula

$$u(x_i, t_{j+1}) = u(x_i, t_j) + \frac{\Delta t}{(\Delta x)^2}\big(u(x_{i+1}, t_j) - 2u(x_i, t_j) + u(x_{i-1}, t_j)\big),$$

the terms $u(x_{i+1}, t_j)$ when $i = M-1$ and $u(x_{i-1}, t_j)$ when $i = 1$ are the boundary terms whose values have already been set to 0. We do not need to track the values of $u(x_{i+1}, t_j)$ when $i = M-1$, i.e., $u(x_M, t_j) = u(1, t_j)$, and $u(x_{i-1}, t_j)$ when $i = 1$, i.e., $u(x_0, t_j) = u(0, t_j)$. That is

$$u(x_1, t_{j+1}) = u(x_1, t_j) + \frac{\Delta t}{(\Delta x)^2}\big(u(x_{i+1}, t_j) - 2u(x_i, t_j) + 0\big)$$

and

$$u(x_{M-1}, t_{j+1}) = u(x_{M-1}, t_j) + \frac{\Delta t}{(\Delta x)^2}\big(0 - 2u(x_{M-1}, t_j) + u(x_{M-2}, t_j)\big).$$

To track the values of $u(x_i, t_j)$ for $i = 1, 2, \ldots, M-1$, we form a $M-1$ vector for the time row t_j by

$$\mathbf{u}(t_j) = \begin{bmatrix} u(x_1, t_j) \\ u(x_2, t_j) \\ \vdots \\ u(x_{M-2}, t_j) \\ u(x_{M-1}, t_j) \end{bmatrix}.$$

In matrix form the update formula for these values of u becomes

$$\mathbf{u}(t_{j+1}) = \mathbf{u}(t_j) + (\Delta t) A \mathbf{u}(t_j)$$

where

$$A = \frac{1}{(\Delta x)^2} \begin{bmatrix} -2 & 1 & 0 & 0 & \cdot & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \cdot & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \cdot & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 & -2 \end{bmatrix}.$$

The $(M-1) \times (M-1)$ matrix A is called a tridiagonal matrix, and for large spatial partitions the matrix A has a lot of zeros, known as a sparse matrix. From this matrix version of the algorithm we obtain

$$\mathbf{u}(t_{j+1}) = \big(I + (\Delta t)A\big)\mathbf{u}(t_j).$$

Defining

$$B = I + (\Delta t)A$$

the algorithm is

$$\mathbf{u}(t_{j+1}) = B\mathbf{u}(t_j).$$

Remark 9.4.3. We first note that the implementation in Algorithm 9.2 is not optimal. In fact, a very simple adjustment to using sparse matrices (arrays) will significantly speed things up and reduce the spatial complexity of the Algorithm as well. Another consideration is that if we only care what the final solution is, we could have just computed $B^n \, \mathrm{U}(t)$ rather than iterating through the loop as is done here. This matrix exponentiation can be performed in a manner similar to fast modular exponentiation (See Volume 2, Section 1.9.3) and is almost always faster than the iteration formed here. For certain sparse matrices it can be made much faster than the iteration.

Remark 9.4.4. Something that you will notice about using forward Euler for the time-stepping is that it actually requires ridiculously small Δt to ensure stability. This is because $\Delta t \sim (\Delta x)^2$ is required. Just imagine what would happen if we had a fourth order PDE, i.e. something like the hyper-diffusive case $u_t = -u_{xxxx}$. This is not something that you want to try to integrate explicitly, but an implicit method would work quite well.

```python
import numpy as np

def forward_euler_1D_heat_array(CFL,T,x=[],u0=[]):
    """solution of the 1D heat equation u_t=u_xx.
    u(0) = 0, and u(1) = 0 are the BCs
    CFL=deltaT/(deltaX^2)
    T is the final time that the simulation will be run
    x is the pre-selected spatial variable, and should be uniformly distributed
    u0 is the initial condition, and must be the same size as x

    Returns an array u which is the solution at all time steps up to T
    as dictated by the time-step chosen through the specified 'CFL' number."""
    N = len(x)
    deltaX = 1/(N-1)
    deltaT = CFL*(deltaX)**2
    u = []
    u.append(u0)
    v = u0.copy() #a dummy variable for the update process
    A = np.diag(np.ones(N-3),-1)-2*np.diag(np.ones(N-2),0)+np.diag(np.ones(N-3),1)
    B = np.eye(N-2)+CFL*A
    for nn in range(1,int(T/deltaT)):
        # double check that the BCs are maintained
        v[0] = 0
        v[-1] = 0
        v[1:-1] = np.matmul(B,u[nn-1][1:-1])
        u.append(v.copy())
    return x,u
```

Algorithm 9.2: Python implementation of the one-dimensional heat equation for forward Euler using a matrix vector representation as given in (**??**).

Remark 9.4.5. Yet another thing to comment on related to stability of the numerical method here is that we can notice that (ignoring the boundary terms which are not affected by anything) $\mathrm{U}(t + n\Delta t) = B^n \mathrm{U}(t)$. This means that another approach to analyzing stability is to ensure that B^n does *not* grow as $n \to \infty$ (this would not be good for our solution). If we assume that B is diagonalizable (it is) then $B^n = PD^nP^{-1}$ where D is a diagonal matrix of the eigenvalues of B. Thus to show that B^n doesn't grow with n we need only show that all of the eigenvalues of B have magnitude less than one. This gives us an equivalent condition that demonstrates the restrictions on the time step Δt relative to the spatial discretization Δx. In the interest of saving space (and time) we don't go into such exciting details here, but do want to point out that this is a very natural way to investigate stability of the numerical scheme.

The primary reason for introducing Algorithm 9.2 instead of Algorithm 9.1 isn't because we like matrix vector products (although that is oddly enough a very good reason), but because it immediately shows us what we would do for an implicit method such as backward Euler or the trapezoid method. In the exercises you carry out these two modifications and find that the stability restriction on the time step is drastically reduced. For implicit methods you end up needed to solve a matrix vector equation at each time step, rather than performing a matrix vector product. For sparse systems such as those we are dealing with here in one dimension, there are very fast solvers that work incredibly efficiently and make this quite fast.

Remark 9.4.6. In case you aren't sufficiently sick of remarks here is another one. You may have realized that we suddenly went from non-homogeneous boundary conditions in Algorithm 9.1 to homogeneous boundary conditions in Algorithm 9.2 with no explanation as to why (tricky, tricky). For the very astute and careful reader who noticed this, we now admit to our deception and clarify the reasons/justification for doing so.

Because the heat equation is linear, we can incorporate the boundary conditions additively by supposing that $u(x,t) = v(x,t)+w(x)$ where $w(x)$ satisfies Laplace's equation i.e. $w_{xx} = 0$ in one dimension, but now where $w(x)$ satisfies the non-homogeneous boundary conditions and $v(x,t)$ satisfies homogeneous conditions of the same type. Specifically, if we are trying to solve the heat equation where $u(0,t) = a$ and $u(1,t) = b$ then we would allow $w(x) = (b-a)x + a$ so that $v_t = v_{xx}$ with $v(0,t) = v(1,t) = 0$, and then use Algorithm 9.2 or its implicit siblings to find what $v(x,t)$ is for a specific initial condition. The same approach can be taken for Neumann or mixed boundary conditions that are non-homogeneous.

Finally we also note that the matrices A and B constructed above are only valid in one dimension. We can do similar things for the heat equation in dimension greater than one, but it takes quite a bit more work. The finite difference approximation to Δu requires a grid of points, and the adequate approximation takes some time to work out. The primary issue with higher dimensions though is how to label the grid points so that the matrix A remains sparse in a way that makes the calculations as efficient as possible. This is not a trivial task, and leads to some very interesting mathematics in its own right.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

9.1. Describe/explain why a unique solution in the entire xy plane could not be derived for the auxiliary condition $u(x,a) = h(x)$ where $a > 0$ prescribed for Example 8.4.3.

9.2. For the traffic flow problem, describe a physical situation in which a Dirichlet boundary condition would make the most sense.

9.3. For the traffic flow problem again, describe a physically relevant situation in which a Neumann boundary condition would make sense.

9.4. Consider a simple cube $[0, L]^3 \subset \mathbb{R}^3$ containing an incompressible fluid with velocity $\mathbf{u} = (u, v, w)$. If the cube has an impermeable boundary meaning the fluid cannot pass through the boundary, what is a valid boundary condition at the top and bottom of the box, i.e. at $z = 0, L$ (the spatial variable is $\mathbf{x} = (x, y, z)$)?

9.5. *Continuing on the previous problem, suppose that the walls of the cube are sufficiently rough so that the fluid cannot 'slip' along the boundary, i.e. the fluid is stuck to the boundary. What additional boundary condition would this impose on the velocity field $\mathbf{u}$ of the fluid? *This is called the 'no-slip' boundary condition in the fluid mechanics literature*

9.6. Suppose that $u(\mathbf{x}, t)$ represents the population density of the state of Utah at time t and geographical location $\mathbf{x}$. If $\mathbf{x}_0$ represents the point at which I-15 enters the state on the northern boundary, then what is a reasonable boundary condition at that point, i.e. $u(\mathbf{x}_0, t) =$?

9.7. For the initial conditions listed in Example 9.2.1, determine what type of boundary conditions are consistent with each. Formulate a specific set of boundary conditions for each case.

9.8. *A homogeneous body occupying the solid region Ω is completely insulated. Its initial temperature is $f(x)$. Find the steady-state temperature that it reaches after a long time. Hint: No heat is gained or lost.

9.9. Use the energy method to prove that a solution to the initial boundary value problem

$$\begin{gathered} u_t - k u_{xx} = 0, \quad 0 < x < l, \quad 0 < t < T, \\ u(x, 0) = f(x), \quad 0 < x < l, \\ u_x(0, t) = 0, u(l, t) = h(t), \quad 0 < t < T, \end{gathered}$$

must be unique.

9.10. Use the energy method to prove uniqueness for the problem

$$\begin{gathered} u_t = \Delta u, \quad \mathbf{x} \in \Omega, \quad t > 0, \\ u(\mathbf{x}, 0) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \\ u(\mathbf{x}, t) = g(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega, t > 0. \end{gathered}$$

9.11. Suppose that u satisfies $u_{tt} - c^2 u_{xx} = 0$ on the interval (a, b), for $t > 0$. Assuming that the solution is sufficiently smooth so that integration by parts holds and derivatives can pass through the integral, under what boundary conditions on u is energy defined as $E = \int_a^b (u_t^2 + c^2 |u_x|^2) dx$ conserved? Prove it.

9.12. The PDE of the previous problem represents the oscillations of a vibrating string. Give a physical interpretation of the boundary conditions you found in the previous problem in this case.

For the next 3 problems, we will consider the Korteweg-De Vries (KdV) equation. Solitary waves (often called solitons) are waves that can travel a great distance without changing their shape. Tsunamis are one example, but such waves occur naturally in other instances as well. The first scientific study of a soliton was initiated when Scott Russell followed such a wave in a channel on horseback in 1834. Russell was fascinated by it's rapid pace and unchanging shape. In 1895, Korteweg and De Vries showed that the evolution of the profile of a soliton is governed by the equation

$$u_t + 6uu_x + u_{xxx} = 0.$$

Suppose that $x \in \mathbb{R}$ and $t > 0$ and that u and all of its derivatives decay to 0 as $x \to \pm\infty$.

9.13. Let $p = \int_{-\infty}^{\infty} u(x,t)dx$. Show that p is constant in time (*physically p is the momentum of the wave.*

9.14. Let $E = \int_{-\infty}^{\infty} u(x,t)^2 dx$. Show that E is constant in time (*physically E is the energy of the wave.*

9.15. *Find an additional non-trivial conserved quantity that is not a linear combination of p and E for this PDE.

9.16. *Solve the Cauchy problem

$$\begin{aligned} u_t &= ku_{xx}, \qquad -\infty < x < \infty, t > 0, \\ u(0,x) &= e^{ax}, \qquad a > 0. \end{aligned}$$

9.17. *Solve the Cauchy problem

$$\begin{aligned} u_t + du &= ku_{xx}, \qquad -\infty < x < \infty, t > 0, \\ u(0,x) &= g(x), \end{aligned}$$

where $d \in \mathbb{R}$ is constant and g is integrable. Hint: Write

$$u(t,x) = e^{-dt}v(t,x).$$

9.18. For an open, bounded subset Ω of $\mathbb{R}^n$, and $u(\mathbf{x},t)$ satisfying the heat equation

$$u_t - \kappa\Delta u = 0$$

with Neumann zero boundary conditions on $\partial\Omega$, and $\kappa > 0$ as usual, show that the total heat $H(\mathbf{x},t) = \int_\Omega u(\mathbf{x},t)d\mathbf{x}$ is conserved. This is not true for Dirichlet zero boundary conditions, why?

9.19. Let $\Omega \subset \mathbb{R}^d$ be an open, bounded subset of $\mathbb{R}^d$ and $C = \Omega \times (0,T)$ be the open 'cylinder' that defines the space-time domain for a scalar valued function $u(\mathbf{x},t)$. Now suppose that $u_t - \kappa\Delta u \leq 0$ in C and u is continuous up to the boundary of C. Then prove that u will attain it's maximum on the sides or bottom of C, i.e. on the set $\Gamma = \{(\mathbf{x},t) | \mathbf{x} \in \partial\Omega \text{ or } t = 0\}$. Hint: If u attains a maximum at some point $(\mathbf{x}_0, t_0)$ then it attains a maximum in every spatial direction. Thus $u_{x_i x_i} \leq 0$, and consequently $\Delta u \leq 0$ at that point.

9.20. Prove that if the Dirichlet problem

$$\begin{aligned} -\Delta u &= \lambda u, \quad \mathbf{x} \in \Omega, \\ u &= 0, \quad \mathbf{x} \in \partial\Omega, \end{aligned}$$

has a nontrivial solution where $u(\mathbf{x})$ is smooth and where both $\mathbf{u}$ and is first derivatives are in $L^2(\Omega)$, then the constant λ must be positive.

9.21. Let $f(z) = u(x, y) + iv(x, y)$ be a holmorphic function where $z = x + iy$ implying that the Cauchy-Riemann equations

$$u_x = v_y \qquad u_y = -v_x$$

are satisfied. If both $f(z)$ and $f'(z)$ (meaning the complex derivative of $f(z)$) are holomorphic (something that is in fact equivalent), show that $\Delta u = 0$ and $\Delta v = 0$, that is u and v satisfy Laplace's equation in x and y.

9.22.* Modify Algorithm 9.1 for the mixed boundary condition described in Example 9.4.1. Create an animation that displays the solution of this problem for initial condition $u(x, 0) = x^2 \sin(2\pi x)$ with $u(1, t) = 0$ and $u_x(0, t) = 0$. How fine of a grid in x did you need to use, and what was the time step you were forced to use as a result to maintain stability?

9.23.* Modify Algorithm 9.2 for backward Euler time-stepping. For initial condition $u(x, 0) = \sin(2\pi x)$ and boundary conditions $u(0) = u(1) = 0$, create an animation that shows how the solution evolves using this method, now using time steps at least $10x$ larger than you did with forward Euler.

9.24.* Modify your algorithm from the previous problem to work for the trapezoid time-stepping method, create an animation of the solution, and compare it with your previous two solutions. *This is such an important discretization that it has two names attached to it: 'Crank–Nicolson'. The Crank–Nicolson scheme is the canonical approach to solving diffusive problems because it eliminates the severe time step restriction that appears for explicit methods and it is higher order than simply backward Euler.*

Notes

10 Eigenfunction Expansions

A little knowledge is a dangerous thing.
—Alexander Pope

So is a lot.
—Anon. (falsely attributed to Einstein)

Linear algebra is the driving force behind a very large portion of applied mathematics. It turns out that every time we think we have escaped from needing to use linear algebra, we have deceived ourselves. As we have already discussed, linearization is one of the greatest tools in an applied mathematician's toolbox. Even without the use of linearization, we can make use of the theory developed for linear operators to pin down the nonlinear behavior, much like we use Duhamel's principle to handle the nonlinear terms in the proof of the Stable Manifold Theorem.

Basically this just means that we really need to study linear PDEs very carefully if we want to be able to make progress on the more interesting nonlinear ones. This chapter and the next develop some of the critical theory for linear PDEs that forms the basis for the study of most nonlinear PDEs.

To motivate this chapter, let's return to a conservation law with no sources/sinks

$$u_t = -\nabla \cdot \mathbf{f}(u).$$

If we suppose this flux divergence is linear, then we can represent it as $-\nabla \cdot \mathbf{f}(u) = \mathscr{L}u$ where $\mathscr{L}$ is some linear differential operator (for the heat equation $\mathscr{L}u = \kappa\Delta u$). Now, from our study of ODEs we may suppose that we can write a solution as $u(\mathbf{x}, t) = e^{\mathscr{L}t}u_0(\mathbf{x})$ where $u(\mathbf{x}, t = 0) = u_0(\mathbf{x})$. It turns out this is an accurate statement/analogy, but what do we make of the exponential of a linear differential operator? There are a plethora of ways we could go about defining this, including the series definition used previously for ODEs. In that case though we would be left with many questions about the repeated application of a differential operator, and convergence of the series would be a rather ornerous task, something that we leave to the interested reader to investigate more on their own.

When we write $\mathscr{L}$ we are implicitly assuming some valid boundary conditions as well. This means that even if the differential operator is identical but the boundary conditions are different, then the exponential will be fundamentally very different. In fact the choice of the boundary conditions will fundamentally alter the exponential operator and the resulting solution.

Remark 10.0.1. Usually you don't see solutions of linear PDEs written as $u(\mathbf{x},t) = e^{\mathscr{L}t}u_0(\mathbf{x})$, but a lot of the ideas we had for computing e^{At} for an $n \times n$ matrix A apply here anyway. Specifically eigenvalues and eigenvectors for the finite-dimensional system become eigenvalues and eigenfuctions for the PDE version, i.e.

$\mathscr{L}u = \lambda u$ with "appropriate" BCs implicitly defined by the linear operator $\mathscr{L}$.

Once we know something about the eigenvalues and eigenfunctions of $\mathscr{L}$ then we can determine how to write $e^{\mathscr{L}t}u_0(\mathbf{x})$, and we will know what it really means. These eigenfunctions are usually found via a method called 'separation of variables' for most linear PDEs. We will demonstrate on some simple examples first, and then work our way up to some more complicated examples as well as computation of the full solution in later Sections in this Chapter.

10.0.1 Analogy to finite systems

To place things on a firmer footing before we continue in our analogy for linear PDEs, recall that the analysis of linear ODEs takes place in an inner product space $(X, \langle\cdot,\cdot\rangle)$ with the additional property that the resulting norm $\|\cdot\|$ is complete (which makes $(X, \|\cdot\|)$ a Banach space. Such a space is called a *Hilbert* space. The analysis of ODEs was done in a finite-dimensional Hilbert space (necessarily isomorphic to $\mathbb{R}^n$ or $\mathbb{C}^n$ with the usual inner product).

For a linear operator $\mathscr{L} : \mathbb{R}^n \to \mathbb{R}^n$, a linear differential equation is

$$\begin{aligned} \mathbf{x}_t &= \mathscr{L}(\mathbf{x}), \\ \mathbf{x}(0) &= \mathbf{x}_0 \in \mathbb{R}^n, \end{aligned}$$

where $\mathbf{x}_t$ means the derivative of $\mathbf{x}$ with respect to t. With a choice of coordinates on $\mathbb{R}^n$ the operator $\mathscr{L}$ is represented by a matrix A, but for now we forego a choice of coordinates.

The solution of the linear differential equation is

$$\mathbf{x}(t) = \exp(\mathscr{L}t)\mathbf{x}_0,$$

where the operator exponential is

$$\exp(\mathscr{L}t) = \sum_{k=0}^{\infty} \frac{t^k \mathscr{L}^k}{k!}.$$

If $\mathscr{L}$ is self-adjoint,

$$\langle \mathbf{x}, \mathscr{L}\mathbf{y}\rangle = \langle \mathscr{L}\mathbf{x}, \mathbf{y}\rangle \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n,$$

(if a matrix representation A of $\mathscr{L}$ satisfies $A^{\mathsf{T}} = A$), then $\mathscr{L}$ has n linearly orthonormal eigenvectors

$$\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$$

corresponding to the real eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ (listed with any repeats) of $\mathscr{L}$, i.e.,

$$\mathscr{L}(\mathbf{v}_k) = \lambda_k \mathbf{v}_k, \ k = 1, 2, \ldots, n.$$

For the initial condition $\mathbf{x}_0$ there are unique constants $c_1, c_2, \ldots, c_n$ such that

$$\mathbf{x}_0 = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \sum_{k=1}^{n} c_k\mathbf{v}_n, \ c_k = \langle \mathbf{x}_0, \mathbf{v}_k\rangle, \ k = 1, 2, \ldots, n.$$

Using spectral calculus we showed in Chapter 4 that

$$\exp(\mathscr{L}t)\mathbf{x}_0 = \sum_{\lambda \in \sigma(\mathscr{L})} \exp(t\lambda) \left(P_\lambda + \sum_{k=1}^{m_\lambda - 1} \frac{t^k D_\lambda^k}{k!} \right) \mathbf{x}_0.$$

Since we are assuming that $\mathscr{L}$ is self-adjoint, each $D_\lambda = 0$, so that

$$\begin{aligned}\exp(\mathscr{L}t)\mathbf{x}_0 &= \sum_{k=1}^{n} \exp(t\lambda_k) P_{\lambda_k}\left(\sum_{l=1}^{n} c_l \mathbf{v}_l\right) \\ &= \sum_{k=1}^{n} \exp(t\lambda_k) c_k \mathbf{v}_k\end{aligned}$$

because $P_{\lambda_k}(\mathbf{v}_l) = \mathbf{v}_k$ when $l = k$ and $P_{\lambda_k}(\mathbf{v}_l) = 0$ when $l \neq k$.

The expression

$$\mathbf{x}(t) = \sum_{k=1}^{n} \exp(t\lambda_k) c_k \mathbf{v}_k$$

is known as the *eigenvalue-eigenvector* expansion of the solution of the linear differential equation.

10.0.2 Infinite-dimensional setting

The analysis of linear PDEs takes place in an infinite-dimensional Hilbert space $\{V, \langle\cdot,\cdot\rangle\}$ of smooth real-valued functions on a domain Ω in $\mathbb{R}^n$.

For a linear operator $\mathscr{L} : V \to V$ consider the linear PDE

$$\begin{aligned}&u_t = \mathscr{L}(u),\ x \in \Omega,\ t \geq 0, \\ &u(\mathbf{x}, 0) = f(\mathbf{x}),\ x \in \Omega, \\ &\alpha u(\mathbf{x}, t) + \beta u_t(\mathbf{x}, t) = g(\mathbf{x}, t),\ x \in \partial\Omega,\ t \geq 0,\end{aligned}$$

for constants α and β, with the appropriate consistency condition $f(\mathbf{x}) = g(\mathbf{x}, 0)$ for all $\mathbf{x} \in \partial\Omega$.

Examples of linear operators $\mathscr{L}$ are

$$\begin{aligned}&\mathscr{L}(u) = -au_x \text{ if } x \in \mathbb{R}, \\ &\mathscr{L}(u) = u_{xx} \text{ if } x \in \mathbb{R}, \\ &\mathscr{L}(u) = -\kappa\Delta u = -\kappa(u_{xx} + u_{yy} + u_{zz}) \text{ if } \mathbf{x} = (x, y, z) \in \mathbb{R}^3.\end{aligned}$$

There are known conditions (like self-adjointness) that can be checked for the linear operator $\mathscr{L}$ subject to the boundary conditions that guarantee that $\mathscr{L}$ has a countable collection of real eigenvalues $\{\lambda_k\}$ and orthonormal eigenfunctions $\{\phi_k(\mathbf{x})\}$ (we show some of these properties later...'YAY'), i.e.,

$$\mathscr{L}(\phi_k(\mathbf{x})) = \lambda_k \phi_k(\mathbf{x}),\ k \in \mathbb{N},$$

such that the eigenfunctions form a basis for V.

The initial condition $f(\mathbf{x})$ expanded in the basis of eigenfunctions determines constants $\{c_k\}$ such that

$$f(\mathbf{x}) = \sum_{k=1}^{\infty} c_k \phi_k(\mathbf{x}),\ c_k = \langle f(\mathbf{x}), \phi_k(\mathbf{x})\rangle,\ k \in \mathbb{N}.$$

Using the infinite-dimensional version of spectral calculus the solution of the linear PDE has the form

$$u(\mathbf{x}, t) = \exp(\mathscr{L}t) f(\mathbf{x}) = \sum_{\lambda \in \sigma(\mathscr{L})} \exp(t\lambda) P_\lambda f(\mathbf{x}).$$

The solution of the linear PDE has the eigenvalue-eigenfunction expansion

$$u(\mathbf{x}, t) = \exp(\mathscr{L}t) f(\mathbf{x}) = \sum_{k=1}^{\infty} \exp(\lambda_k t) c_k \phi_k(\mathbf{x}).$$

Remark 10.0.2. One might correctly point out that there were a lot of suppositions and unjustified assumptions in the discussion provided above. Nevertheless for many types of linear differential operators these very ideas hold true, and in fact this is the most common fundamental approach to computing solutions of linear PDEs, even those where the time-dependence is different than that discussed above. This is of course reliant on specific properties of the linear operator $\mathscr{L}$ which we review briefly in the following sections, but remarkably even these specific properties are fulfilled in almost every linear PDE that arises in a non-artificial way just like finite-dimensional systems are generically represented by a linear system with a semi-simple matrix.

Remark 10.0.3. You may have noticed already that the infinite-sum form of the solution is a sum of scalar multiples of products of functions of t with functions of $\mathbf{x}$. In other words, the ansatz

$$u(\mathbf{x}, t) = X(\mathbf{x})T(t)$$

is a starting point for the solution technique known as the method of separation of variables, i.e. separation of variables (a topic which most frequently encapsulates an entire semester long course (or more) in many scientific departments will be briefly introduced here and justified in a matter of a few days.

Remark 10.0.4. Another aspect of this comparison is if we consider a finite difference discretization of the linear operator $\mathscr{L}$ with spatial discretization Δx. As shown in Algorithm 9.2 this discretized solution can be represented by a linear, finite-dimensional ordinary differential equation. This approximation converges to the linear PDE as $\Delta x \to 0$ wherein the matrix is then infinite-dimensional. Hence, the analogy that linear PDEs are infinite-dimensional ODEs is actually quite natural.

The computation of the eigenvalues and eigenfunctions for even the simplest of linear differential operators is painstaking, and as such we have a limited number of examples in these notes. We try to illustrate this going from the 'simplest' examples to the more complicated ones in the following sections. The painstaking nature of these examples should illustrate why many other courses take entire semesters dedicated to this topic...even a single example can take several days to work out.

10.1 Some Fundamental Examples

The first thing we will investigate is calculating the eigenfunctions and eigenvalues for a spatial, linear PDE operator (with corresponding boundary conditions). In later sections we will use these eigenfunctions to get an expansion solution for a related linear PDE (with potential temporal dependence).

Example 10.1.1. We first investigate the linear differential operator

$$\mathscr{L}u = u_{xx}, \tag{10.1}$$

with boundary conditions $u(0) = 0$ and $u_x(1) = 0$. As noted, once we have identified the eigenvalues and eigenfunctions of this linear operator, we will be able to compute a series solution to PDEs such as $u_t = u_{xx}$ and $u_{tt} = u_{xx}$ with the same boundary conditions.

A function u is an eigenvector of $\mathscr{L}$ with eigenvalue λ if

$$u_{xx} = \lambda u. \tag{10.2}$$

Solutions to (10.2) are exponential functions of the form $u(x) = e^{rx}$, where

$$r^2 = \lambda. \tag{10.3}$$

This leads to three different types of potential solutions:

(i) $\lambda = \mu^2 > 0$. In this case the solutions $u(x)$ can be written as a linear combination of either exponential functions or hyperbolic trigonometric functions (either are appropriate because $\sinh(\mu x) = \frac{1}{2}\left(e^{\mu x} - e^{-\mu x}\right)$ and $\cosh(\mu x) = \frac{1}{2}\left(e^{\mu x} + e^{-\mu x}\right)$), i.e.

$$u(x) = a_0 e^{\mu x} + a_1 e^{-\mu x},$$

or

$$u(x) = c_0 \cosh(\mu x) + c_1 \sinh(\mu x),$$

are appropriate general solutions. We use the hyperbolic trigonometric functions because it makes the algebra simpler. Applying the first boundary condition leads to

$$u(0) = c_0 \cosh(\mu 0) = c_0 = 0,$$

so the solution is really given by $u(x) = c_1 \sinh(\mu x)$. Using the second boundary condition, we obtain

$$u_x(1) = \mu c_1 \cosh(\mu) = 0, \tag{10.4}$$

however a quick review of our hyperbolic trigonometric functions tells us that this is impossible for any value of $\mu \neq 0$ unless $c_1 = 0$. Thus the only possible solution in this case is $u(x) = 0$ i.e. the trivial solution.

Now, just as when we were calculating eigenvectors in linear algebra, the zero eigenfunction is not really an eigenfunction at all (the zero vector isn't an eigenvector), and hence we have failed for this case. This means that we must have $\lambda \leq 0$.

(ii) $\lambda = 0$. In this case the solutions are given by

$$u(x) = c_0 + c_1 x,$$

(these are not the same c_i as the previous case, I just was going to run out of the English alphabet if I used something else, and most readers don't like use Greek and/or Hebrew letters for constants). Inserting the boundary condition at $x = 0$ leads to

$$u(0) = c_0 = 0,$$

and using the right boundary condition

$$u_x(1) = c_1 = 0,$$

once again yields the trivial solution, i.e. $\lambda = 0$ is NOT an eigenvalue. Hence $\lambda < 0$ is our only hope.

(iii) $\lambda = -\mu^2 < 0$ in which case the roots of the characteristic equation (10.3) are $r = \pm i\mu$. In this case the solution can more easily be written as

$$u(x) = c_0 \cos(\mu x) + c_1 \sin(\mu x).$$

Now if we insert the left boundary condition we arrive at

$$u(0) = c_0 = 0.$$

The right boundary condition leads to

$$u_x(1) = \mu c_1 \cos(\mu) = 0. \tag{10.5}$$

The only way that (10.5) will hold without resorting to the trivial case is if

$$\mu_n = \frac{\pi}{2} + n\pi, \quad n \in \mathbb{N}.$$

This means that we have found our eigenfunctions and eigenvalues simultaneously! Such excitement, what thrilling news! The eigenfunctions for this linear operator (and corresponding boundary conditions) are given by

$$\phi_n(x) = \sin\left[\left(\frac{\pi}{2} + n\pi\right)x\right], \quad n \in \mathbb{N}, \tag{10.6}$$

with corresponding eigenvalues

$$\lambda_n = -\mu_n^2 = -\left(\frac{\pi}{2} + n\pi\right)^2, \quad n \in \mathbb{N}. \tag{10.7}$$

We can verify that these are the eigenfunctions of $u_{xx} = \lambda u$ for the prescribed boundary conditions. Checking the first boundary condition we have

$$\phi_n(0) = \sin\left(\left(\frac{\pi}{2} + n\pi\right)(0)\right) = 0 \text{ for all } n \in \mathbb{N}.$$

Checking the second boundary condition we have

$$\left(\frac{\pi}{2} + n\pi\right)\cos\left(\left(\frac{\pi}{2} + n\pi\right)(1)\right) = 0 \text{ for all } n \in \mathbb{N}.$$

Computing the second derivative of $\phi_n(x)$ we have

$$\mathscr{L}(\phi_n(x)) = -\left(\frac{\pi}{2} + n\pi\right)^2 \sin\left(\left(\frac{\pi}{2} + n\pi\right)x\right) = \lambda_n \phi_n(x).$$

Remark 10.1.2. We will of course immediately jump into a more difficult example (this is a math textbook after all), but first let us make a few comments about the general setup. First we note that there were three potential cases that we had to go through but only one that led to a viable eigenfunction. This is actually a very common feature of eigenfunction calculations. Typically there are multiple different types of solutions (not necessarily three, but three does seem to happen a lot), but only a small subset of those possible solution types yield a non-trivial eigenfunction. We are only interested in the non-trivial eigenfunctions, so we can throw out the other options.

Second, note that you should probably review your trigonometric and hyperbolic trigonometric functions. Personally I think wikipedia is a great resource for this, but if you like memorizing such facts I will not hold it against you.

Third, this was a bit more complicated and time consuming than finding an eigenvector-eigenvalue pair for a finite-dimensional linear system. Every calculation for linear PDEs is quite the process, which is why there are entire textbooks and courses dedicated to this very subject.

Remark 10.1.3. You may also notice that we mysteriously switched from using $u(x)$ to $\phi_n(x)$ when we specified the eigenfunction. There was no specific reason for doing this...it just seemed appropriate to refer to the eigenfunctions by a special variable whereas $u(x)$ is reserved for generic statements.

Now we are ready to consider a slightly more complicated example, one in which the eigenvalues are actually not as easily computable.

Example 10.1.4. Consider now the linear differential operator given by

$$\mathscr{L}u = u_{xx} - u_x,$$

with boundary conditions $u(0) = 0$ and $u'(1) = 0$. We are looking at this in order to solve the problem $u_t = \mathscr{L}u$, i.e. we want these eigenfunctions and eigenvalues so we can solve the advection-diffusion equation $u_t + u_x = u_{xx}$. To find the eigenvalues and eigenfunctions for this operator, we first recognize that the solutions of

$$u_{xx} - u_x = \lambda u$$

will be given by exponential functions of the form $u(x) = e^{rx}$. This implies that in this setting

$$r^2 e^{rx} - re^{rx} = \lambda e^{rx},$$

so that r must satisfy the quadratic equation

$$r^2 - r - \lambda = 0,$$

implying that

$$r = \frac{1}{2} \pm \sqrt{\frac{1}{4} + \lambda}.$$

As in the previous example, this gives three possible cases:

(i) $\lambda = -\frac{1}{4}$. This means that the solution is given by

$$u(x) = c_0 e^{\frac{x}{2}} + c_1 x e^{\frac{x}{2}}.$$

Applying the first boundary condition implies that

$$u(0) = c_0 = 0,$$

and we note that

$$u'(1) = c_1 \left(e^{\frac{x}{2}} + \frac{x}{2} e^{\frac{x}{2}} \right) \Big|_{x=1} = c_1 \frac{3}{2} e^{\frac{1}{2}} = 0,$$

so that $c_1 = 0$. This means $\lambda = -\frac{1}{4}$ is not an eigenvalue.

(ii) Now suppose that $\lambda > -\frac{1}{4}$. In this case we will have two possible solutions $r_\pm = \frac{1}{2} \pm \sigma$ where $\sigma = \sqrt{\frac{1}{4} + \lambda}$ is real. As previously, we could view the solution in terms of the exponential only, but it is much more convenient to write

$$u(x) = e^{\frac{x}{2}} \left(c_0 \cosh(\sigma x) + c_1 \sinh(\sigma x) \right).$$

Applying the first boundary condition we get

$$u(0) = c_0 = 0.$$

For the second boundary condition we first note that this means that

$$u' = c_1 e^{\frac{x}{2}} \left(\frac{1}{2} \sinh(\sigma x) + \sigma \cosh(\sigma x) \right).$$

Evaluating this at $x = 1$ yields

$$c_1 e^{\frac{1}{2}} \left(\frac{1}{2} \sinh(\sigma) + \sigma \cosh(\sigma) \right) = 0.$$

This can only happen if

$$\frac{1}{2} \sinh(\sigma) + \sigma \cosh(\sigma) = 0,$$

however if $\sigma > 0$ (which it is in this case) then this isn't possible because this function is a positive function everywhere. Thus there is no solution for $\lambda > -\frac{1}{4}$.

(iii) Now we suppose that $\lambda < -\frac{1}{4}$. This means that the roots are given by $r = \frac{1}{2} \pm i\sigma$ where $\sigma = \sqrt{-\lambda - \frac{1}{4}}$ is real (and strictly positive). This means that solutions are given by:

$$u(x) = e^{\frac{x}{2}} \left(c_0 \cos(\sigma x) + c_1 \sin(\sigma x) \right).$$

The first boundary condition indicates that $u(0) = c_0 = 0$ so that $u(x) = c_1 e^{\frac{x}{2}} \sin(\sigma x)$, and

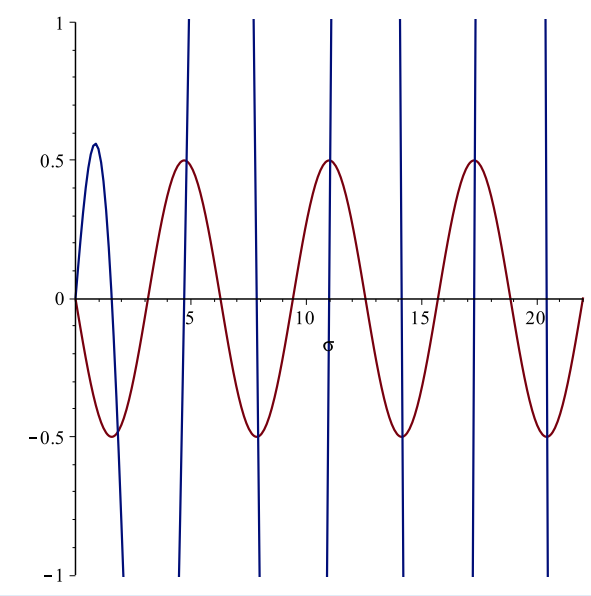
$$u' = c_1 e^{\frac{x}{2}} \left(\frac{1}{2} \sin(\sigma x) + \sigma \cos(\sigma x) \right) = 0.$$

This means that we are interested in all positive values of σ so that

$$\sigma \cos(\sigma) = -\frac{1}{2} \sin(\sigma).$$

This has a countable number of solutions σ_n which can be seen by plotting the function $f(\sigma) = \sigma \cos(\sigma) + \frac{1}{2} \sin(\sigma)$ and identifying the roots of this equation. Actual computation of these σ_n can be done via a root finding method, and asymptotic approximations are often used to approximate their actual values.

There is no algebraic way to solve this for σ, but it has infinitely many solutions, that are asymptotic to the extremizers of $-(\frac{1}{2})\sin(\sigma)$, i.e., to $\frac{\pi}{2} + n\pi$, (which is the red graph in the following, the blue graph is that of $\sigma \cos(\sigma)$).



With infinitely many values σ_n, $n \in \mathbb{N}$, we have infinitely many eigenvalues

$$\sigma = \sqrt{-\lambda - \frac{1}{4}} \;\Rightarrow\; \lambda_n = -\sigma_n^2 - \frac{1}{4}, \quad n \in \mathbb{N}.$$

The corresponding eigenfunctions are

$$\phi_n(x) = \exp\left(\frac{x}{2}\right)\sin(\sigma_n x), \quad n \in \mathbb{N}.$$

These two examples in this section hopefully illustrate the main ideas for how to calculate the eigenfunctions and eigenvalues of a linear operator. In addition, the second example illustrates that not all eigenvalues are computable in a nice closed form, but are frequently solutions to nonlinear, transcendental equations. Despite the inherent difficulties, these eigenfunction-eigenvalue calculations are quite doable once you get used to them.

10.2 Multiple dimensions, i.e. separation of variables

The examples from the previous Section were complicated in their own right, but were only concerned with a PDE in one spatial dimension only. Now we will consider what happens if we are looking at a spatial linear operator with 2 or more independent spatial variables. All of the examples we consider are in 2 dimensions, but the extension to higher dimensions is rather immediate (at least in theory, in practice they take much longer).

Example 10.2.1. Let

$$\mathscr{L}u = \Delta u = u_{xx} + u_{yy},$$

for (x, y) on the domain $\Omega = [0, \pi]^2$ with boundary conditions

$$u(x, 0) = u_y(x, \pi) = 0, \quad \text{and} \quad u(0, y) = u(\pi, y) = 0.$$

We want to determine the eigenfunctions of this differential operator, that is we want to identify functions $u(x, y)$ such that $\mathscr{L}u = -\lambda u$, where we have cleverly absorbed a minus sign into the eigenvalue λ.

Now, suppose that $u(x, y) = X(x)Y(y)$ where X has no dependence on y and Y has no dependence on x. Then the eigenvalue problem becomes

$$-X''(x)Y(y) - X(x)Y''(y) = \lambda X(x)Y(y) \Rightarrow -\frac{X''}{X} - \frac{Y''}{Y} = \lambda$$
$$\Rightarrow -\frac{X''}{X} = \left(\frac{Y''}{Y} + \lambda\right),$$

where we have dropped the independent variables x and y for simplicity. Now since the left hand side depends only on x and the right hand side only on y then both sides must equal a constant value which we call μ. In the following it is very important to keep in mind that μ is just a constant which we do not yet know, i.e. it is something that we 'choose' below and is related to the eigenvalue λ. This gives us two problems:

(i) $-X'' = \mu X$ $\qquad X(0) = X(\pi) = 0$

(ii) $-Y'' = (\lambda - \mu)Y$ $\qquad Y(0) = Y'(\pi) = 0,$

i.e. note that the boundary conditions are decoupled just as the function is itself.

For the first problem there are three distinct possibilities:

(a) $\mu = 0$. This implies that the solution is given by

$$X(x) = c_0 x + c_1.$$

The boundary condition $X(0) = 0$ implies that $c_1 = 0$, and the boundary condition $X(\pi) = 0 \Rightarrow c_0 = 0$, so this gives only a trivial solution, which by the way isn't what we are looking for (did we consider the zero vector a valid eigenvector for a matrix A?).

(b) $\mu = -\sigma^2 < 0$. Then

$$X(x) = c_0 \sinh(\sigma x) + c_1 \cosh(\sigma x).$$

Note that this is equivalent to letting $X(x) = a_0 e^{\sigma x} + a_1 e^{-\sigma x}$ where a_0 and a_1 are linear combinations of c_0 and c_1. We choose the hyperbolic trigonometric functions instead because it makes one of the boundary conditions easier to work with.

Applying the first boundary condition, we see that

$$X(0) = 0 \Rightarrow c_1 = 0,$$

and hence applying the second boundary condition at $x = \pi$,

$$X(\pi) = c_0 \sinh(\sigma\pi) = 0 \Rightarrow c_0 = 0,$$

so $X = 0$ which gives only the trivial solution, where we have relied on the fact that $\cosh(x) \neq 0$ for any values of x.

(c) $\mu = \sigma^2 > 0$, so that:

$$X(x) = c_0 \sin(\sigma x) + c_1 \cos(\sigma x).$$

Applying the first boundary condition we see that $X(0) = 0 \Rightarrow c_1 = 0$. The second boundary condition indicates that $X(\pi) = c_0 \sin(\sigma\pi) = 0 \Rightarrow \sigma = n$ for any integer n, and thus $\mu_n = n^2$ corresponding to

$$X_n(x) = \sin(nx), \quad \mu_n = n^2, \quad n \in \mathbb{N}.$$

Hence we have found the solution for $X(x)$, but still need to do something with $Y(y)$. For this, we are now looking at the problem

$$-Y'' = (\lambda - n^2)Y,$$

with

$$Y(0) = Y'(\pi) = 0.$$

Letting $\nu = \lambda - n^2$ we again have three cases:

(a) $\nu = 0$. This implies that

$$Y(y) = c_0 y + c_1.$$

Just as before, we apply the boundary conditions to see that $Y(0) = c_1 = 0$ and $Y'(\pi) = c_0 = 0 \Rightarrow c_0 = 0$, so that we once again get the trivial case so that $\nu = 0$ can be discarded.

(b) $\nu = -\sigma^2 < 0$. In this case,

$$Y(y) = c_0 \sinh(\sigma y) + c_1 \cosh(\sigma y).$$

Applying the boundary conditions once again leads to $Y(0) = c_1 = 0$ and $Y'(\pi) = \sigma c_0 \cosh(\sigma\pi) = 0 \Rightarrow c_0 = 0$,, i.e. the trivial solution yet again.

(c) $\nu = \sigma^2 > 0$. This leads to the solution

$$Y(y) = c_0 \sin(\sigma y) + c_1 \cos(\sigma y).$$

The boundary conditions yield $Y(0) = c_1 = 0$ and $Y'(\pi) = c_0\sigma \cos(\sigma\pi) = 0 \rightarrow \sigma = \frac{2k-1}{2}$ for $k \in \mathbb{Z}^+$ in order to obtain a nontrivial solution.

It follows that $\nu_k = \frac{(2k-1)^2}{4}$ and $-\lambda_{k,n} = -n^2 - \frac{(2k-1)^2}{4}$ are the eigenvalues of this problem with corresponding eigenfunctions

$$\phi_{k,n}(x, y) = \sin(nx) \sin\left(\frac{2k-1}{2} y\right). \tag{10.8}$$

Now recall that our motivation for this was to find a way to compute $e^{\mathscr{L}} u$ where $\mathscr{L}u = \Delta u$ (this is to identify a solution of $u_t = \mathscr{L}u$). We can use the resolvent for this, which is related to the development of Green's functions in the next Chapter, but for now, we note that for these BC's, the operator $\mathscr{L}$ is simple, so there are no repeated eigenvalues and hence we can write out $e^{\mathscr{L}t} u_0(x)$ as a (countably in this case) infinite sum of the inner product of the eigenfunctions with $u_0(\mathbf{x})$.

For instance, in analogy with a finite-dimensional system where $\dot{\mathbf{x}} = A\mathbf{x}$ where A is simple, we can write

$$\begin{aligned} e^{At}\mathbf{v} &= \sum_{\lambda\in\sigma(A)} e^{\lambda t} P_\lambda \mathbf{v} \\ &= \sum_{\lambda\in\sigma(A)} e^{\lambda t}\langle \mathbf{x}_\lambda, \mathbf{v}\rangle \mathbf{x}_\lambda \end{aligned}$$

where $\mathbf{x}_\lambda$ is the eigenvector corresponding to eigenvalue λ and $\langle \mathbf{x}_\lambda, \mathbf{v}\rangle$ is the inner product given by the dot product. The extension of this to the linear differential operator considered here is:

$$u(x,y,t) = e^{\mathscr{L}t}u_0(x,y) = \sum_{k,n>0} e^{\lambda_{k,n}t} c_{k,n}\phi_{k,n}(x,y),$$

where, as we found above

$$-\lambda_{k,n} = \frac{(2k-1)^2}{4} + n^2 \text{ and } u_{k,n}(x,y) = \sin(nx)\sin\left(\frac{2k-1}{2}y\right),$$

and

$$c_{k,n} = \langle u_0(x,y), \phi_{k,n}(x,y)\rangle\,,$$

but what inner product can we use?

What is the only inner product space on functions that we know of? $L^2(\Omega)$ (appropriately weighted to ensure the selected eigenfunctions are normalized).

$$\langle g_1(x,y), g_2(x,y)\rangle_{L_2} = \frac{4}{\pi^2}\int_0^\pi \int_0^\pi g_1(x,y)g_2(x,y)dxdy.$$

Now we note that the eigenfunctions we constructed form an orthonormal set under this inner product, i.e.

$$\langle \phi_{k,n}, \phi_{k',n'}\rangle = \delta_{k-k',n-n'},$$

where

$$\delta_{k-k',n-n'} = \begin{cases} 1 & \text{if } k=k' \;\&\; n=n' \\ 0 & \text{otherwise} \end{cases}$$

is the Dirac delta operator.

Without justification, we note that as long as $f(x,y) = u_0(x,y) \in L^2(\Omega)$ then

$$f(x,y) = \sum_{k,n} \langle f(x,y), \phi_{k,n}(x,y)\rangle\, \phi_{k,n}(x,y),$$

which allows for the computation of the solution as given above.

In this example, miraculously not only did we find that the eigenfunctions were 'nice' functions, but that they actually form an orthogonal basis of $L^2(\Omega)$. It turns out this isn't quite as miraculous as it may seem, as this linear operator fits into a class of operators that yield such nice behavior. As already mentioned it seems that classifying linear operators in this way is restrictive, but it turns out that 'most' (think of this in a measure theoretic sense) linear differential operators fall under this classification. This is discussed in a little detail in the next section.

Actually making all of the steps in the previous example completely rigorous are beyond the scope of this course, but it suffices to note that such results are actually quite commonplace and can be shown in a graduate course on functional analysis.

Remark 10.2.2. It is noteworthy to point out that there are an infinite (at least countable) number of eigenvalues and eigenfunctions for the linear operator of the previous example. The exponential of this linear operator is precisely why we can think of linear PDEs as an infinite-dimensional system of ODEs. This eigen-decomposition demonstrates the differences between this infinite-dimensional setting, and that achieved in the finite-dimensional case for ODEs.

Remark 10.2.3. Note that all of the examples we have seen so far have been for homogenous boundary conditions. We treat non-homogeneous boundary conditions much like we did numerically for the heat equation, that is we construct a solution that satisfies these non-homogeneous boundary conditions, and subtract it from the full solution. The remaining term then satisfies a linear operator with homogeneous boundary conditions.

10.3 Sturm-Liouville Problems

In order to see how the examples from the previous two sections generalize, we define a particular class of second-order linear PDEs that retain all the nice properties that were observed in the examples provided above. The full theory developed for these systems is beyond the scope of this text, but the application and utility of this theory is for the most part a rather straight forward extension of linear algebra. Basically keep in mind that we are really extending linear algebra to infinite-dimensional Hilbert spaces.

Definition 10.3.1. *For $\mathbf{x} \in \Omega \subset \mathbb{R}^n$ (typically $n = 1, 2, 3$) define*

$$\mathscr{L}u = -\nabla \cdot (p\nabla u) + qu,$$

where $p : \overline{\Omega} \to \mathbb{R}$ and $q : \overline{\Omega} \to \mathbb{R}$ are C^2 functions with $p(\mathbf{x}) > 0$. The regular Sturm-Louville (SL in all that follows) problem is defined as

$$\mathscr{L}u = \lambda r u, \tag{10.9}$$

where $r : \overline{\Omega} \to \mathbb{R}$ and $r(\mathbf{x}) > 0$,with either Dirichlet, Neumann, or Robin boundary conditions. If $r(\mathbf{x}) = 0$ or $p(\mathbf{x}) = 0$ for some value of $\mathbf{x}$ then the SL is singular (or if the domain Ω is unbounded).

Remark 10.3.2. Expanding the operator $\mathscr{L}$ gives:

$$\begin{aligned}\mathscr{L}u &= -\nabla \cdot (p\nabla u) + qu \\ &= -\nabla p \cdot \nabla u - p\Delta u + qu.\end{aligned}$$

Example 10.3.3. If $p = 1$, $q = 0$ and $r = 1$, then $\mathscr{L}$ becomes $\mathscr{L}u = -\Delta u$. So the (negative) Laplace operator is a special case of SL.

Definition 10.3.4. *For this particular form of SL, define the inner product*

$$\langle f, g\rangle_{L^2_r(\Omega)} = \int_\Omega r(\mathbf{x}) f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}.$$

Remark 10.3.5. We often shorten notation so that $\langle \cdot, \cdot \rangle_{L^2_r(\Omega)} = \langle \cdot, \cdot \rangle_r$.

For complex-valued functions the inner product for the Hilbert space $L^2_r(\Omega)$ is defined as

$$\langle f, g \rangle_{L^2_r(\Omega)} = \int_\Omega r(\mathbf{x}) f(\mathbf{x}) \overline{g}(\mathbf{x})\, d\mathbf{x}$$

Remark 10.3.6. Higher order spatial derivative operators can be treated as higher order SL as well, but the analysis is significantly more complicated in that case. In addition, with few exceptions most physical applications appear in the form of at most a second-order operator. For these reasons we focus only on second-order SL operators.

Lemma 10.3.7 (Lagrange's Identity). *For a regular SL problem, with $u : \overline{\Omega} \to \mathbb{R}$ and $v : \overline{\Omega} \to \mathbb{R}$ both being C^2 then*

$$\int_\Omega (u\mathscr{L}v - v\mathscr{L}u) d\mathbf{x} = \int_{\partial\Omega} [-u(\mathbf{x})p(\mathbf{x})\nabla v(\mathbf{x}) + v(\mathbf{x})p(\mathbf{x})\nabla u(\mathbf{x})] \cdot \mathbf{n} dA.$$

Proof. Using integration by parts (the product rule for $\nabla\cdot$) and the Divergence Theorem gives

$$\begin{aligned}\int_\Omega [\overline{u}\mathscr{L}v]\, d\mathbf{x} &= -\int_\Omega [u\nabla \cdot (p\nabla v)]\, d\mathbf{x} + \int_\Omega [quv] d\mathbf{x} \\ &= \int_\Omega [(\nabla u) \cdot (p\nabla v)]\, d\mathbf{x} - \int_{\partial\Omega} [up\nabla v] \cdot \mathbf{n} dA + \int_\Omega [quv] d\mathbf{x}.\end{aligned}$$

Noting that the final volume integrals above are symmetric in u and v, the result follows immediately. ∎

The proof above relies on the following form of the Divergence Theorem (integration by parts):

$$\int_\Omega f\nabla \cdot \mathbf{G} d\mathbf{x} = -\int_\Omega (\nabla f) \cdot \mathbf{G} d\mathbf{x} + \int_{\partial\Omega} [f\mathbf{G}] \cdot \mathbf{n} dA,$$

where $f : \overline{\Omega} \to \mathbb{R}$ is C^2 and $\mathbf{G} : \overline{\Omega} \to \Omega$ is C^2 as well.

Remark 10.3.8. From the exercises we notice that the boundary integral in Lagrange's identity vanishes for appropriately chosen boundary conditions (Dirichlet, Neumann or Robin). Thus for these regular BC's the operator $\mathscr{L}$ is self-adjoint (meaning that the inner product between $\mathscr{L}u$ and v is the same as the inner product between $\mathscr{L}v$ and u) in $L^2(\Omega)$. It turns out that this is true even if we are considering the inner product space $L^2_r(\Omega)$, that is $\langle u, \mathscr{L}v \rangle_{L^2_r(\Omega)} = \langle u, \mathscr{L}v \rangle_{L^2_r(\Omega)}$.

Remark 10.3.9. The regular SL problem $\mathscr{L}(u) = \lambda r u$ doesn't quite look like an eigenvalue/eigenfunction equation because of the positive function r on the right-hand side. But it is equivalent to an eigenvalue/eigenfunction equation since with $r > 0$ we have

$$\mathscr{L}(u) = \lambda r(u) \;\Rightarrow\; \frac{1}{r}\mathscr{L}(u) = \lambda u.$$

Recall that

$$\mathscr{L} = -\nabla p \cdot \nabla u - p\Delta u + qu,$$

and so

$$\frac{1}{r}\mathscr{L}(u) = -\frac{1}{r}\nabla p \cdot \nabla u - \frac{p}{r}\Delta u + \frac{q}{r}u.$$

So the eigenvalue/eigenfunction equation associated to a regular SL problem is

$$-\frac{1}{r}\nabla p\cdot\nabla u-\frac{p}{r}\Delta u+\frac{q}{r}u=\lambda u.$$

When $\lambda\in\mathbb{C}$ the eigenfunction u may be complex-valued.

Self-adjoint linear operators have a lot of nice properties, just like Hermitian or symmetric matrices have in finite dimensions. One of these properties is shown in the following Theorem, and is a critical component of the theory for SL problems.

Theorem 10.3.10. *Eigenvalues of a regular SL are real.*

Proof. Note that if $\mathscr{L}u=r\lambda u$ then $\mathscr{L}\bar{u}=r\bar{\lambda}\bar{u}$ (this is because $r(\mathbf{x})>0$ is real). Thus

$$\begin{aligned}0&=\langle u,\mathscr{L}u\rangle-\langle\mathscr{L}u,u\rangle\\&=\langle u,r\lambda u\rangle-\langle r\lambda u,u\rangle\\&=\bar{\lambda}\langle u,ru\rangle-\lambda\langle ru,u\rangle\\&=\bar{\lambda}\int_\Omega ur\bar{u}d\mathbf{x}-\lambda\int_\Omega ru\bar{u}d\mathbf{x}\\&=(\bar{\lambda}-\lambda)\int_\Omega r(\mathbf{x})|u(\mathbf{x})|^2d\mathbf{x}\end{aligned}$$

so that $\bar{\lambda}=\lambda$ is the only feasible option (recall that $r(\mathbf{x})>0$). ∎

Remark 10.3.11. As shown in the homework, if λ_j is an eigenvalue of a regular SL then $\lambda_j>0$. Also shown in the homework, if $\lambda_j\neq\lambda_k$ are two distinct eigenvalues with eigenfunctions ϕ_j and ϕ_k then ϕ_j and ϕ_k are orthogonal, i.e. $\langle\phi_j,\phi_k\rangle_r=0$.

This shows why SL problems are so important: they have orthogonal eigenfunctions with real, positive eigenvalues. In an infinite-dimensional space such as $L^2(\Omega)$ it is necessary to not only recognize that the eigenfunctions are orthogonal, we also need to determine that they span the entire space. In the finite-dimensional setting, this could amount to counting the elements of the basis and comparing to the dimension of the vector space in question, but in an infinite-dimensional Hilbert space this is not the same thing, although it turns out that we do have enough information.

We will rely on the following theorem, although it is not proved here:

Theorem 10.3.12. *For a regular SL problem the eigenvalues can be ordered*

$$0<\lambda_1<\lambda_2<\cdots,\quad \textit{and}\quad \lim_{n\to\infty}\lambda_n=+\infty.$$

In addition the orthogonal eigenfunctions of a regular SL form a basis for the space of integrable functions $L^2(\Omega)$. A rigorous proof can be found in a good course on functional analysis (see, for example [?]).

Remark 10.3.13. Although we do not present the full proof here, it is worth noting that the rigorous justification of this theorem relies heavily on the use of the resolvent for the linear operator $\mathscr{L}$. The resolvent in this case is defined exactly as it was in the finite-dimensional setting, and the eigenvalues and eigenfunctions of the resolvent can be investigated rigorously because the resolvent is a compact operator whereas $\mathscr{L}$ is not.

To highlight some of the usefulness of the formal designation of SL, we present an informal proof of the *Fredholm alternative*, which is similar to the infamous Invertible Matrix Theorem in finite dimensions.

Theorem 10.3.14 (Fredholm Alternative). *Let $\mathscr{L}$ be a regular SL operator with appropriate BCs and assume that $f \in L^2(\Omega)$ is given. Consider the equation*

$$\mathscr{L}u = \mu u + f(\mathbf{x}). \tag{10.10}$$

The associated homogeneous problem *is*

$$\mathscr{L}u = \mu u,$$

corresponding to setting $f = 0$ in (10.10)*. The equation* (10.10) *has*

(a) a unique solution if μ is not *an eigenvalue of the associated homogeneous problem.*

(b) no solution or infinitely many, depending on $f(\mathbf{x})$, if μ is an eigenvalue of the associated homogeneous problem.

Proof. (Sketch of proof) Because f and u are in L^2 they can be represented by Fourier series of the eigenfunctions of $\mathscr{L}\phi_\lambda = \lambda\phi_\lambda$, that is,

$$f(\mathbf{x}) = \sum_{\lambda\in\sigma(\mathscr{L})} f_\lambda\phi_\lambda, \qquad\qquad u(\mathbf{x}) = \sum_{\lambda\in\sigma(\mathscr{L})} u_\lambda\phi_\lambda.$$

The expression $\mathscr{L}u = \mu u + f$ becomes

$$\sum_{\lambda\in\sigma(\mathscr{L})} \lambda u_\lambda\phi_\lambda = \mu \sum_{\lambda\in\sigma(\mathscr{L})} u_\lambda\phi_\lambda + \sum_{\lambda\in\sigma(\mathscr{L})} f_\lambda\phi_\lambda.$$

Taking the inner product of both sides of this equation with $\phi_{\tilde\lambda}$ implies, via orthogonality of the ϕ_λ, that

$$\tilde\lambda u_{\tilde\lambda} = \mu u_{\tilde\lambda} + f_{\tilde\lambda},$$

which if $\mu \neq \tilde\lambda$ for all λ will have solution $u_\lambda = \frac{f_\lambda}{\lambda-\mu}$. If $\mu = \lambda$ for some eigenvalue λ then this will give either infinitely many solutions (when $f_\lambda = 0$ meaning $f(\mathbf{x})$ is in the null space of $\mathscr{L}$) or none (when $f_\lambda \neq 0$). □

The Fredholm alternative may seem like a simple extension of the theory we have developed so far, and in a sense it is, but it is also key to many arguments in the theory of linear parabolic PDEs. Even if the immediate specific application of the Fredholm alternative is not immediately obvious, the formal proof above does indicate how the theory of SL can be applied to different situations.

10.4 Applications of the theory

To see how to use the theory developed above, and the construction of the relevant eigenvalues and eigenfunctions, we return to Example 10.2.1.

Example 10.4.1. Consider the PDE

$$u_t = \Delta u,$$

with boundary conditions

$$u(x,0,t) = u_y(x,\pi,t) = 0 \qquad u(0,y,t) = u(\pi,y,t) = 0.$$

Because $-\Delta$ is an SL operator with weight $r = 1$ (see Example 10.3.3), the operator $\mathscr{L}u = \Delta\mathbf{u}$ with these boundary conditions has distinct negative eigenvalues $\lambda_{n,k}$ and eigenfunctions $\phi_{n,k}(x, y)$ which are orthogonal:

$$\phi_{n,k}(x, y) = \sin(nx) \sin\left(\frac{2k-1}{2}y\right) \qquad \lambda_{n,k} = -n^2 - \left(\frac{2k-1}{2}\right)^2.$$

These are orthonormal with respect to the rescaled inner product

$$\langle g_1, g_2\rangle = \frac{4}{\pi^2}\int_0^\pi \int_0^\pi g_1(x, y)g_2(x, y)dxdy = \langle g_1, g_2\rangle. \tag{10.11}$$

The factor of $\frac{4}{\pi^2}$ is chosen to make the eigenfunctions have norm 1, but otherwise this is just the standard inner product (note that $\langle\cdot,\cdot\rangle_r = \langle\cdot,\cdot\rangle$ because $r = 1$). Under this inner product, these eigenfunctions form an orthonormal basis on the Hilbert space of square integrable functions satisfying the specific boundary conditions introduced above on the domain $\Omega = [0, \pi] \times [0, \pi]$.

To solve the initial value problem with $u(x, y, t = 0) = f(x, y)$ we rewrite the solution as:

$$u(x, y, t) = \sum_{n,k} c_{n,k}(t)\phi_{n,k}(x, y).$$

Inserting this into the PDE implies that

$$\sum_{n,k} c'_{n,k}(t)\phi_{n,k}(x, y) = \sum_{n,k} \lambda_{n,k}c_{n,k}(t)\phi_{n,k}(x, y)$$

where

$$u(x, y, t = 0) = \sum_{n,k} c_{n,k}(0)\phi_{n,k}(x, y) = \sum_{n,k}\langle f, \phi_{n,k}\rangle\phi_{n,k}(x, y),$$

and $c'_{n,k}(t)$ refers to the time derivative of the coefficients $c_{n,k}(t)$.

This is an infinite series, and although we may be tempted to look at the solution term by term, we all know that we need to be more careful. Instead we consider a single eigenfunction $\phi_{l,m}(x, y)$ and take the inner product of both sides of the equalities above with respect to $\phi_{l,m}(x, y)$, recalling that

This leads to:

$$\sum_{n,k} c'_{n,k}(t)\langle\phi_{n,k}, \phi_{l,m}\rangle = \sum_{n,k} \lambda_{n,k}c_{n,k}(t)\langle\phi_{n,k}, \phi_{l,m}\rangle,$$
$$\sum_{n,k} c_{n,k}(0)\langle\phi_{n,k}, \phi_{l,m}\rangle = \sum_{n,k}\langle f, \phi_{n,k}\rangle\langle\phi_{n,k}, \phi_{l,m}\rangle.$$

Now, recalling that the $\phi_{n,k}$ form an orthonormal set, we see that

$$\langle\phi_{n,k}, \phi_{l,m}\rangle = \begin{cases} 1 & \text{if } l = n,\ m = k, \\ 0 & \text{otherwise} \end{cases}$$

Thus, the evolution of the $c_{n,k}$ above for each of the indices n and k is reduced to:

$$c'_{l,m}(t) = \lambda_{l,m}c_{l,m}(t), \quad c_{l,m}(0) = f_{l,m} = \langle f, \phi_{l,m}\rangle,$$

which has the explicit solution

$$c_{l,m}(t) = f_{l,m} e^{\lambda_{l,m} t}.$$

For the problem at hand, suppose that the initial conditions are given by

$$u(x, y, 0) = f(x, y) = xy(\pi - x)(2\pi - y),$$

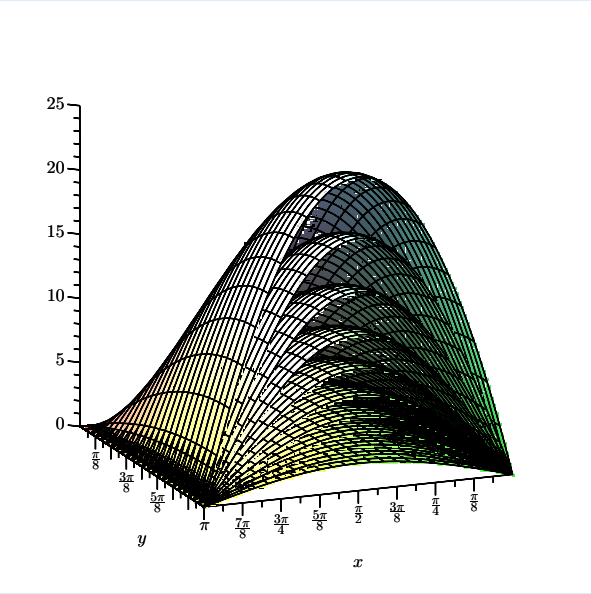
which was carefully chosen to satisfy the boundary conditions exactly. It follows that the coefficients $f_{n,k}$ are computed as:

$$\begin{aligned} f_{n,k} &= \frac{4}{\pi^2} \int_0^\pi \int_0^\pi xy(\pi - x)(2\pi - y) \sin(nx) \sin\left(\frac{2k-1}{2} y\right) dydx \\ &= \frac{4}{\pi^2} \frac{32(1 - (-1)^n)}{n^3(2k-1)^3}, \end{aligned}$$

where this integral was computed via Wolfram Alpha. The full solution is then given by:

$$u(x, y, t) = \sum_{n,k} \frac{128(1 - (-1)^n)}{\pi^2 n^3 (2k-1)^3} \sin(nx) \sin\left(\frac{2k-1}{2} y\right) e^{-\left[n^2 + \left(\frac{2k-1}{2}\right)^2\right] t}.$$

Here is a depiction showing the graphs of $u(x, y, t)$ at time intervals starting at $t = 0$ and ending at $t = 2.0$, increasing by 0.2.



Notice the homogeneous Dirichlet condition on the left (the other two homogeneous Dirichlet boundary conditions are on the back and the right).

The outline provided for this example can be extended to similar situations even for problems in which the temporal dependence is more complicated, or has higher order temporal derivatives (think of the wave equation). The key is to first determine the eigenfunctions and eigenvalues of the spatial differential operator (including the relevant boundary conditions). As long as this operator is a regular SL then the eigenfunctions can be used to expand the solution as well as any other potentially hazardous terms (including non-homogeneous parts of the PDE) in terms of the orthogonal eigenfunctions. Then using orthogonality, the temporal dependence can be determined as the coefficients of each eigenfunction expansion are matched up, and the resultant ODE is solved.

Remark 10.4.2. One may speculate as to the usefulness of these solutions to linear PDEs. For instance, we have written down an explicit solution in the previous example for the heat equation in the domain $[0, \pi] \times [0, \pi]$, but is this solution really all that useful? An infinite series in terms of trigonometric functions doesn't seem to be the ideal way to describe any phenomenon I recall, and evaluating this infinite series is also not a very practical exercise. In practice however, only the first handful of terms in the series are really necessary.

To see why this is the case, consider the time-dependent coefficients in front of each eigenfunction in the final form of the solution above. For n, k near zero, these coefficients are relatively large at least for short times, but either as $t \to \infty$ or more importantly for this discussion, $n, k \to \infty$ then these coefficients tend to zero exponentially. Thus we see that after only a few terms in the series, the remaining terms can justifiably be neglected as they are proportionally much smaller in magnitude. For the graph in the example above we used $n, k \in \{1, 2, 3, 4, 5\}$.

Remark 10.4.3. You have seen Fourier series and the Fourier transform before. We now note that the derivation presented in this Chapter yields a general approach for creating what is called a 'generalized Fourier series'. Every regular SL problem produces a set of orthogonal eigenfunctions that can be used to describe such a series.

It is also worth noting at this juncture that all of our results regarding these solutions are in $L^2(\Omega)$ or more generally $L^2_r(\Omega)$ when the weighting function $r(\mathbf{x})$ appears. What does this mean for the properties of the solution to the time dependent PDE? For instance, what does convergence of the generalized Fourier series mean in this space, and what does it mean for solutions of a differential operator to be defined in a space where discontinuous solutions are allowed?

What we get in the end for a generalized Fourier series solution, is one that converges to the true solution in L^2 but may diverge at a point. This is a phenomenon referred to commonly as Gibb's ringing, and is seen in many places where this method is applied without considering the ramifications first. In addition a generalized Fourier series is allowing for solutions of the PDE that are not continuous let alone differentiable. This should also raise a red flag. What we can say for certainty is that if the solution exists and is nice and differentiable to whatever order we need it to be, then the generalized Fourier series solution will converge to that solution except on a set of measure zero, i.e. it will converge in L^2. Thus our representation of the solution may not look very good at a few points, but the overall characteristics of the solution will be captured well.

Remark 10.4.4. One more remark and then we will have one last simpler example. Note that construction of the eigenfunctions of the SL rely not only on the form of the operator and on the boundary conditions, but also on the shape of the domain itself. In all of the examples considered in this chapter, nice square domains (intervals in 1D) are considered, but in general that is not what we find. For domains of different shapes, different eigenfunctions will be the result. For instance, cylindrical geometry yields Bessel functions, and spherical geometry produces Legendre polynomials and/or spherical Bessel functions (often called spherical harmonics), and in each of these cases we get a set of orthogonal eigenfunctions. Letting that sink in for a few moments we can see that the SL approach to these problems is powerful indeed. If orthogonal sets of functions are so nice (and yes they really are) then every SL on every unique domain with every different boundary condition, will give us another set of orthogonal eigenfunctions which we can use for a variety of purposes, not just solving time dependent linear PDEs.

As a final example, we return to Example 10.1.4, only looking at the solution of the advection diffusion equation in one dimension.

Example 10.4.5. Consider solutions of

$$u_t + u_x = u_{xx}, \quad u(0,t) = 0, \quad u_x(1,t) = 0,$$

with initial condition

$$u(x,0) = f(x) = x\sin(\pi x).$$

From Example 10.1.4 we already know the eigenfunctions and eigenvalues of the corresponding spatial operator here, and just as the previous example, we know that the solution will look like

$$u(x,t) = \sum_k c_k f_k e^{\lambda_k t} e^{x/2} \sin(\sigma_k x),$$

where σ_k satisfy $\sigma_k \cos(\sigma_k) = -\frac{1}{2}\sin(\sigma_k)$ and $\lambda_k = -\sigma_k^2 - \frac{1}{4}$, $k \in \mathbb{N}$ (note also that as $k \to \infty$ then $\sigma_k \sim \frac{\pi}{2} + k\pi$). We also specify the computable normalization constant

$$c_k = \sqrt{\langle \phi_k, \phi_k \rangle} = \sqrt{\int_0^1 e^x \sin^2(\sigma_k x) dx},$$

to make the presentation simpler, which yields the orthonormal eigenfunctions

$$\frac{\phi_k(x)}{c_k} = \frac{1}{c_k} e^{x/2} \sin(\sigma_k x).$$

The final remaining step is to identify the coefficients f_k which are found as

$$f_k = c_k \int_0^1 x\sin(\pi x) e^{x/2} \sin(\sigma_k x) dx,$$

which while a rather uncomfortable looking integral, is computable for all k.

Remark 10.4.6. This second example demonstrates how the eigenfunction expansion technique can quickly get out of hand for rather mild looking PDEs (looks can be deceiving). We remind the reader though, that even though this eigenfunction expansion to the advection diffusion equation is quite the mess, the solution itself is actually quite well behaved, i.e. this is still a linear, dissipative PDE whose solutions remain unique.

One may legitimately ask if the eigenfunction expansion technique produces such a headache of a computation, then what good is it really? The answer of course isn't immediate. There is one possibility which we mention below, but the other is that a numerical discretization (finite differences etc.) would have been more efficient than working through the numerical approximation of the integrals in the previous example. But even when other methods are easier to use computationally, the eigenfunction expansion is extremely useful as a theoretical tool.

10.5 Spectral and pseudo-spectral methods

These eigenfunction expansions lead to another numerical method that is particularly powerful for certain domains of interest. We focus primarily on periodic boundary conditions in a single dimension, but we note that similar methods are available for a host of other situations including multidimensional domains.

10.5.1 Linear PDEs and spectral methods

The basic idea of a spectral method is to approximate the solution of a PDE as a truncated series of known, simple functions such as a certain class of polynomials or trigonometric functions. If we use functions that are eigenfunctions of a linear differential operator on the domain of interest (with the specific boundary conditions we desire) then this can significantly improve our representation of the solution. In particular the boundary conditions are guaranteed to be satisfied for such an expansion.

Example 10.5.1. Consider the PDE given by:

$$u_t + u_{xx} = -u_{xxxx}, \quad u(0,t) = u(1,t) = 0.$$

If we recall that on this domain, and with these boundary conditions, $\phi_k(x) = \sqrt{2}\sin(k\pi x)$ are the eigenfunctions of the linear operator $\mathscr{L}u = -u_{xx}$ with eigenvalues $\lambda_k = k^2\pi^2$, then we will suppose that our solution is given by:

$$u(x,t) = \sum_k c_k(t)\phi_k(x).$$

Inserting this into the original PDE, and appropriately using the orthonomality of the eigenfunctions $\phi_k(x)$, we arrive at the decoupled infinite set of ODEs

$$c_k'(t) - k^2\pi^2 c_k(t) = -k^4\pi^4 c_k(t).$$

This can of course be readily solved explicitly, or integrated using a time-stepping method for the ODE.

The numerical approximation comes when we truncate this infinite series at a finite value, i.e. we suppose that

$$u(x,t) = \sum_{k=1}^{N} c_k(t)\phi_k(x),$$

with the computation of the initial data

$$c_k(0) = \langle f(x), \phi_k(x)\rangle,$$

where $u(x, t = 0) = f(x)$ is the initial condition.

There is nothing particularly remarkable about the previous example, especially when you consider it in the context of this entire chapter. What is remarkable is that we could use any orthonormal set of functions defined on the domain of interest with the correctly selected boundary conditions. This means that we could pick whatever set of basis functions $\{\phi_k(x)\}$ make the computation of $c_k(0)$ the 'simplest' or most efficient computationally. For periodic boundary conditions, $\sin(kx)$ and $\cos(kx)$ or e^{ikx} are a clear choice.

Example 10.5.2. Consider the periodic advection-diffusion equation in one dimension

$$u_t + au_x = u_{xx}, \quad u(0) = u(2\pi).$$

A convenient choice of basis functions would be $\sin(kx)$ and $\cos(kx)$, but these can be more compactly represented via the traditional Fourier series using e^{ikx}. Thus, we allow

$$u(x, t) = \sum_k \hat{u}_k(t) e^{ikx},$$

with the inner product

$$\langle g_1(x), g_2(x) \rangle = \frac{1}{2\pi} \int_0^{2\pi} g_1(x) \overline{g_2(x)}\, dx,$$

for complex valued functions $g_1, g_2 : [0, 2\pi] \to \mathbb{C}$. With respect to this inner product the eigenfunctions e^{ikx} form an orthonormal basis for $L^2([0, 2\pi])$, the space of square-integrable complex valued functions g with $g(0) = g(2\pi)$.

The real benefit of this is that if we truncate this series then we can use the FFT to compute the coefficients $\hat{u}_k(0)$. In addition, we could make use of the fact that $u(x, t)$ is a real valued function, and hence $\hat{u}_k(t) = \overline{\hat{u}_{-k}(t)}$. We can also determine the initial values $\hat{u}_k(0)$ from the initial condition:

$$f(x) = u(x, 0) = \sum_k \hat{u}_k(0) e^{ikx},$$

so that

$$\hat{u}_k(0) = \langle f(x), e^{ikx} \rangle, \ k \in \mathbb{Z}.$$

If we insert this expansion into the original PDE, then we can see that each coefficient $\hat{u}_k$ will satisfy the simple ODE:

$$\hat{u}_k'(t) = (-aik - k^2)\hat{u}_k(t).$$

This gives us a direct approach for a spectral solver to the advection diffusion equation as shown in Algorithm 10.1. The application of spectral methods is actually incredibly simple as demonstrated here, but the devil is in the indexing that *NumPy* uses to calculate the FFT. Once you have those details in place, the rest of the algorithm is quite straightforward.

```python
import numpy as np

def advec_diffuse(a,dt,T,x=[],u0=[]):
    """Use spectral decomposition and forward Euler to solve
    the 1D advection diffusion equation u_t + au_x = u_xx.
    Assume the domain is [0,2pi] or else the wavenumbers would need
    to be weighted differently.
    a is the advection speed
    dt is the time-step
    T is the final time we run to
    x is the spatial discretization
    u0 is the initial condition

    Because this uses the fft, it is best to have len(x) = 2^K for some integer K
    """

    N = len(x)
    #The most difficult part is figuring out how numpy indexes the wavenumbers
    k = np.array([1j*y for y in range(0,int(N/2))] + [0] + [1j*y for y in range(int(-N/2)+1,0)])
    k2 = k**2;
    u = []
    u.append(u0)
    uhat = np.fft.fft(u0)
    totalStep = int(T/dt)
    for nn in range(totalStep):
        # forward Euler time-stepping
        new_uhat = uhat + dt*(-a*k + k2)*uhat
        uhat = new_uhat.copy()
        # convert back to time domain
        u.append(np.real(np.fft.ifft(uhat)))
    return x,u
```

Algorithm 10.1: Python implementation of a spectral solution of the periodic one-dimensional advection diffusion equation considered in Example 10.5.2. Note that there are certainly better ways to implement this in Python. For instance the real FFT will simplify the calculation significantly.

Remark 10.5.3. One of the first things we can notice about the spectral approach is that the primary difficulty we have is figuring out how the FFT is implemented. Every package/programming language has a different approach to doing this, and the indexing can get quite complicated to figure out. In fact figuring out this indexing is far more difficult than anything else you do to analyze the PDE.

Just to add some complexity for the sake of computational speed up, we could also modify Algorithm 10.1 to use a real FFT, i.e. one which recognizes that we only need to track half of the Fourier coefficients because the others are complex conjugates of the same. This makes tracking the wavenumbers k even more complex, but it cuts the storage space for the FFT in half, and also speeds up (the already speedy) FFT algorithm itself. For settings where we are dealing with a very large vector u this can yield a significant gain.

If we wanted to consider the absolute numerical stability of this algorithm it is actually quite a bit simpler than what we saw for our previous discretizations of advection and/or diffusion. In this case, we have reduced the PDE to a series of ODEs that look just like

$$u' = \lambda u,$$

where λ changes with each wavenumber k and is given by $\lambda_k = -aik - k^2$. If you recall from TODO: chapter reference, we know that forward Euler is stable so long as $z = \lambda \Delta t$ is in the unit circle in the complex plane centered at $z = -1$. Hence we need to pick Δt so that $\Delta t(-aik - k^2)$ is in this region for all values of the wavenumber k. Once we know the absolute stability region for other time-stepping methods the approach is the same.

This is incredibly informative about the nature of the solution. In fact, we can see that these spectral methods are clearly very closely connected to the concept of planar waves that we saw previously.

10.5.2 Quasi-linear PDEs, eigenfunction expansions, and pseudo-spectral methods

We first consider the effects of an eigenfunction expansion on a quasilinear PDE, and then show how to modify spectral methods to apply to the quasilinear setting as well.

As an illustration of why we may want to use eigenfunction expansions or generalized Fourier Series, consider the quasilinear PDE:

$$u_t + uu_x + = u_{xx},$$

with some appropriate boundary conditions (we will use homogeneous Dirichlet ones here to match the exercises, but you can imagine how this can be extended to the periodic case as well) on the domain $[0, 1]$. This is viscous Burger's equation. The dissipative (second derivative) term eliminates the previous issues we had with characteristics crossing, and hence the breakdown of uniqueness of solutions. It also introduces a potential new approach. E-function expansions critically rely on the linearity of the PDE, so we can't directly use such an expansion here, but we can take advantage of the fact that part of the PDE is linear.

The point is that the linear part of this operator still yields a complete set of orthonormal eigenfunctions that we can use as a basis expansion of $u(x, t)$. When we insert this expansion into the equation, the orthonormality of the eigenfunctions no longer guarantees that each term in the infinite series decouples, but it still leads us to a system of ODEs (though infinite) that we can analyze.

Example 10.5.4. We specifically consider the viscous Burgers equation for homogeneous Dirichlet boundary conditions, i.e. $u(0, t) = u(1, t) = 0$. As shown in the exercises, the linear operator $\mathscr{L}u = -u_{xx}$ with these boundary conditions yields orthonormal eigenfunctions

$$\phi_k(x) = \sqrt{2}\sin(k\pi x),$$

with eigenvalues $\lambda_k = k^2\pi^2$ and the standard (real-valued) inner product

$$\langle g_1, g_2 \rangle = \int_0^1 g_1(x) g_2(x) dx.$$

Using these eigenfunctions, we suppose that the solution to the viscous Burgers equation is given by:

$$u(x,t) = \sum_k c_k(t)\phi_k(x) = \sqrt{2}\sum_k c_k(t)\sin(k\pi x).$$

If we plug this into the viscous Burgers equation then we get:

$$\sqrt{2}\sum_k c_k'(t)\sin(k\pi x) + 2\left(\sum_k c_k(t)\sin(k\pi x)\right)\left(\sum_n n\pi c_n(t)\cos(n\pi x)\right) = -\sqrt{2}\pi^2\sum_k k^2 c_k(t)\sin(k\pi x).$$

Multiplying by $\sqrt{2}\sin(m\pi x)$ and integrating over all of $[0,1]$ then this becomes:

$$c_m'(t) + 2\sqrt{2}\pi\sum_{k,n} c_k(t)c_n(t)n\int_0^1 \sin(k\pi x)\cos(n\pi x)\sin(m\pi x)dx = -m^2\pi^2 c_m(t).$$

The double summation above looks bad, but it turns out that the integral vanishes for most values of n and k, leaving a still infinite sum of terms that influence $c_m(t)$, but not the full double summation shown above. You can identify which terms are nonzero by working extensively with trigonometric identities if you like, but we will let it suffice to recognize that several of these terms disappear.

It may not be obvious from this context what the benefit of looking at the PDE in this sense is, but it helps to identify the effects of the diffusive term and boundary conditions on the nonlinearity and vice versa.

The use of an eigenfunction expansion of the linear part for a quasilinear operator is a standard approach to identifying the movement of energy between spatial length scales. The nonzero terms from the nonlinearity demonstrate how the energy is transferred. Actual approximations to the PDE are usually done by truncating the summation, which gives a reasonable numerical approximation to the full solution.

Nonlinear PDEs and pseudo-spectral methods

It is rather straightforward to modify Algorithm 10.1 for other linear PDEs, and once we have worked with higher order time-stepping methods we can also implement those quite easily. As demonstrated in Example 10.5.4 it is clear that these type of methods will not be as simple for nonlinear PDEs. One approach is to write out the resultant coupled system of ODEs for the truncated system, and to apply a time-stepping algorithm to this setup. This can be incredibly time-consuming to code up, particularly when you keep in mind how much fun we are already having with the indices of the FFT. Instead, we typically take a pseudo-spectral approach where the nonlinear term is treated by converting it back into real space, calculating the nonlinearity, and then applying the FFT to the resultant product.

The Fourier Transform is the bijective linear transformation

$$\hat{u}(k,t) = \mathscr{F}(u) = \int_0^{2\pi} u(x,t)\exp(-ikx)\ dx,\ \ k \in \mathbb{Z},$$

from $L^2([0,1])$ to $l^2(\mathbb{Z})$ (the inner product space of square summable bi-infinite sequences).

The Fourier Transform has the properties

$$\hat{u}_t = \frac{\partial}{\partial t}\mathscr{F}(u) = \frac{\partial}{\partial t}\int_0^{2\pi} u(x,t)\exp(-ikx)\ dx = \int_0^{2\pi} u_t(x,t)\exp(-ikx)\ dx$$

and

$$\mathscr{F}(u_x) = ik\hat{u}(k,t).$$

The second property follows because by integration by parts we have

$$\begin{aligned}\int_0^1 u_x(x,t)\exp(-ikx)\ dx &= u(x,t)\exp(-ikx)\Big|_{x=0}^{x=1} - \int_0^{2\pi}(-ik)u(x,t)\exp(-ikx)\ dx \\ &= 0 + ik\hat{u}(x,t),\end{aligned}$$

where we have used the boundary conditions $u(0,t) = 0$ and $u(2\pi,0) = 0$ for all $t \geq 0$.

The second property implies that

$$\mathscr{F}(u_{xx}) = (i)^2k^2\hat{u}(k,t) = -k^2\hat{u}(k,t).$$

Applying the Fourier Transform to the viscous Berger's equation gives

$$\mathscr{F}(u_t + uu_x) = \mathscr{F}(u_{xx}).$$

By linearity and the properties above we obtain

$$\hat{u}_t + \mathscr{F}(uu_x) = -k^2\hat{u}.$$

The Fourier Transform of the product can be written as

$$\mathscr{F}(uu_x) = \mathscr{F}\big(\mathscr{F}^{-1}(\hat{u})\mathscr{F}^{-1}(ik\hat{u})\big)$$

because

$$\mathscr{F}(u_x) = ik\hat{u} \ \Leftrightarrow\ u_x = \mathscr{F}^{-1}(ik\hat{u}).$$

Thus we obtain

$$\hat{u}_t + \mathscr{F}\big(\mathscr{F}^{-1}(\hat{u})\mathscr{F}^{-1}(ik\hat{u})\big) = -k^2\hat{u}.$$

This is implemented in Algorithm 10.2.

Remark 10.5.5. Of course in practice, periodic boundary conditions, while the clear choice for most mathematicians, are an incredibly naive choice for a serious modeler. If you want to be taken seriously outside of mathematics it would be best to not use periodic boundary conditions everywhere (unless of course you are interested in determining the spread of heat through a donut). This presents a problem because we have just shown that these spectral methods are remarkably efficient if we use the traditional Fourier series. The answer is of course, in the details. Using some very clever transformations, we can compute the coefficients of a Chebyshev representation using the FFT as well, and this is often the choice for non-periodic boundary conditions. For specific geometries, similar transformations exist in which case other bases are used such as Legendre or Laguerre polynomials.

Remark 10.5.6. We have said nothing about aliasing errors in these spectral methods due to a lack of space in this text, but that is not to mean they aren't important. In fact, aliasing errors in spectral and pseudo-spectral methods are a fundamentally concerning issue. If you google 'dealiasing' for spectral methods you will see that this is definitely an issue of fundamental concern.

```python
import numpy as np

def diffusive_burgers(dt,T,x=[],u0=[]):
    """Uses a spectral decomposition and forward Euler to solve
    the 1D diffusive Burgers equation u_t + uu_x = u_xx.
    Assumes the domain is [0,2pi] or else the wavenumbers would need
    to be weighted differently.
    dt is the time-step
    T is the final time we run to
    x is the spatial discretization
    u0 is the initial condition

    Because this uses the fft, it is best to have len(x) = 2^K for some integer K"""

    N = len(x)
    #The most difficult part is figuring out how numpy indexes the wavenumbers
    k = np.array([1j*y for y in range(0,int(N/2))] + [0] + [1j*y for y in range(int(-N/2)+1,0)])
    k2 = k**2;
    u = []
    u.append(u0)
    uhat = np.fft.fft(u0)
    totalStep = int(T/dt)
    for nn in range(totalStep):
        # forward Euler time-stepping
        new_uhat = uhat + dt*(np.fft.fft(np.fft.ifft(k*uhat)*np.fft.ifft(uhat)) + k2*uhat)
        uhat = new_uhat.copy()
        # revert back to real values (not necessary unless we want
        # to save our solution in real space)
        u.append(np.real(np.fft.ifft(uhat)))
    return x,u
```

Algorithm 10.2: Python implementation of a pseudo-spectral solution of the periodic one-dimensional diffusive Burgers equation.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

10.1. Find the eigenvalues λ and corresponding eigenfunctions for the boundary value problem

$$\begin{aligned} -u_{xx} - 2u_x &= \lambda u, \quad 0 < x < 1, \\ u(0) = 0, &\quad u(1) = 0. \end{aligned}$$

10.2. Find the eigenvalues and eigenfunctions of the boundary value problem

$$\begin{aligned} -x^2 u_{xx} - x u_x &= \lambda u, \quad 1 < x < e^\pi \\ u(1) = 0, &\quad u(e^\pi) = 0. \end{aligned}$$

Hint: Google Cauchy-Euler problems to see how to solve this one.

10.3. Show that $\phi_k(x) = \sqrt{2}\sin(\pi x)$ are the eigenfunctions of the diffusion operator defined by:

$$\begin{aligned} -u_{xx} &= \lambda u, \quad 0 < x < 1, \\ u(0) = u(1) &= 0. \end{aligned}$$

10.4. What are the eigenfunctions and eigenvalues for the same linear operator as in Example 10.2.1 with the same boundary conditions but now on the domain given by $x \in [0, L]$ and $y \in [0, 2\pi]$.

10.5. Find the eigenfunctions and eigenvalues for the 2D Laplacian Δu with boundary conditions $u_x(0, y) = 0 = u(\pi, y)$ and $u(x, 0) = u_y(x, \pi) = 0$.

10.6. Show that the general second-order boundary value problem in one dimension,

$$-P(x)u'' - Q(x)u' + R(x)u = \lambda u$$

can be rewritten as a one-dimensional Sturm-Liouville problem

$$\mathscr{L}y = (-py')' + qy = \lambda r y.$$

Hint: Use an integrating factor of the form $\mu = \frac{e^{\int \frac{Q}{P}\,dx}}{P}$. There is an analog for this in higher dimensions, can you find it?

10.7. Using the method you derived in the previous problem, convert the equation $-y'' + x^4 y' = \lambda y$ into Sturm-Liouville form.

10.8. Show that for Dirichlet and Neumann boundary conditions on $u(\mathbf{x})$ and $v(\mathbf{x})$ that the boundary integral on the right hand side of Lagrange's identity (Lemma 10.3.7) is zero.

10.9. Show that if λ_j is an eigenvalue of a Sturm-Liouville problem:

$$-\nabla \cdot (p\nabla u) + qu = \lambda r u$$

for $\mathbf{x} \in \Omega$ with $r(\mathbf{x}) > 0$, $p(\mathbf{x}) > 0$ and $q(\mathbf{x}) > 0$ as well as Robin boundary conditions $\frac{du}{dn} + a(\mathbf{x})u = 0$ for $\mathbf{x} \in \partial\Omega$ where $a(\mathbf{x}) > 0$, then $\lambda_j > 0$. Hint: Compute $\langle u, \mathscr{L}u\rangle_{L^2} = \langle u, r\lambda u\rangle_{L^2}$ and solve for λ. This results in a formula for the eigenvalue λ that is often referred to as Rayleigh's Quotient and is occasionally used to motivate a variational formula (a concept that you will see in the next Part of this book) for computing eigenvalues.

10.10. Suppose that $\lambda_j \neq \lambda_k$ are two distinct eigenvalues of the Sturm-Liouville problem stated in the previous problem. Show that the corresponding eigenfunctions ϕ_j and ϕ_k are orthogonal in the weighted L^2_r space, i.e. $\langle \phi_j, \phi_k\rangle_r = 0$.

10.11. Use the eigenfunction expansion method to solve the following problem

$$\begin{gathered} u_{tt} = c^2 u_{xx}, \quad 0 < x < 1, t > 0, \\ u(0,t) = u(1,t) = 0, \quad t > 0, \\ u(x,0) = x(1-x), u_t(x,0) = 0, \quad 0 < x < 1. \end{gathered}$$

10.12. Transform the problem

$$\begin{gathered} u_t = u_{xx}, \quad 0 < x < \pi, t > 0, \\ u(0,t) = 3, u(\pi,t) = 1, \\ u(x,0) = f(x), \end{gathered}$$

into one with homogeneous boundary conditions. *(HINT: let $u(x,t) = v(x,t) + A(x)$, where A is chosen to satisfy the boundary conditions.)*

10.13. Solve the problem given in the last problem for $f(x) = 1 + 2\sin\left(\frac{x}{2}\right)$.

10.14. *Using the results of the previous two problems, solve

$$\begin{gathered} u_t = u_{xx} + \sin\left(\frac{x}{2}\right), \quad 0 < x < \pi, t > 0, \\ u(0,t) = u(\pi,t) = 0, \\ u(x,0) = 1. \end{gathered}$$

10.15. Modify Algorithm 10.1 to solve the advection diffusion equation using a trapezoid time-stepping method. For $N = 128$ grid points on the periodic interval $[0, 2\pi]$, create an animation starting with initial condition $u_0(x) = \sin(x^2) + \cos(5x)$ and running up to time $T = 50$, with advection speed $a = 4$. What time-step did you need to use to ensure your solution was stable?

10.16. Modifying Algorithm 10.2, code up a forward Euler discretization to the Kuramato-Sivashinsky (KS) equation on $[0, 2\pi]$:

$$u_t + uu_x + u_{xx} + u_{xxxx} = 0.$$

NOTE: (KS) is often used as the simplest mathematical model of chaotic behavior in a PDE, but it was originally derived for a variety of different physical situations, including the evolution of a flame (think of a candle). The actual derivation is a bit of a mess, but the dynamics it models are incredibly interesting on their own.

10.17. Create an animation for the solution of (KS) with $N = 128$ grid points, and starting from initial condition $u_0(x) = \sin(x^2) + \cos(5x)$, again running up to time $T = 0.5$. How did you choose a time-step that ensured the solution was numerically stable?

10.18. Modify your code for (KS) for the 2-stage, second-order Runge-Kutta method, and repeat the same simulation and animation as in the previous problem. Can you notice any differences? How much more restrictive is the time-step for RK2 than for forward Euler?

Notes

11 Green's Functions

The formulation of a problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skills.
—Leopold Infeld

Green's functions aren't necessarily unique to PDEs, but they are an extremely useful tool when identifying solutions of PDEs. The best way to think of a Green's function is that it defines the kernel for the inverse of a differential operator (defined with appropriate boundary conditions). In this sense, this chapter is dedicated to defining, and identifying the inverse operator to the solution operator $e^{\mathscr{L}}$ where $\mathscr{L}$ is some differential operator. This is difficult because this solution operator is defined in infinite dimensions, we can't simply apply $e^{-\mathscr{L}}$ and expect to have reached our goal. Something much more sinister (and fun!) is going on.

Returning to the proof of the Fredholm Alternative, consider a linear PDE (with prescribed BCs)

$$\mathscr{L}u = f.$$

If we could compute the inverse of $\mathscr{L}$ then we would calculate $u = \mathscr{L}^{-1}f$. Allow $\mathscr{L}$ to have eigenfunctions $\phi_n(\mathbf{x})$ and eigenvalues λ_n, and suppose that $f \in L^2(\Omega)$ and $f(\mathbf{x})$ is continuously differentiable. Then

$$f(\mathbf{x}) = \sum_{n=1}^{\infty} f_n \phi_n(\mathbf{x}), \quad \text{where} \quad f_n = \langle f(\mathbf{x}), \phi_n(\mathbf{x}) \rangle = \int_\Omega f(\mathbf{x}) \phi_n(\mathbf{x}) d\mathbf{x}.$$

The solution of this PDE is written as

$$u(\mathbf{x}) = \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}),$$

where (so long as $\lambda_n \neq 0$ for all n) $c_n = \frac{f_n}{\lambda_n}$. Thus

$$\begin{aligned} u(\mathbf{x}) &= \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}) \\ &= \sum_{n=1}^{\infty} \frac{f_n}{\lambda_n} \phi_n(\mathbf{x}) \\ &= \sum_{n=1}^{\infty} \frac{\int_\Omega f(\xi)\phi_n(\xi)d\xi}{\lambda_n} \phi_n(\mathbf{x}) \\ &= \int_\Omega \left[\sum_{n=1}^{\infty} \frac{\phi_n(\xi)\phi_n(\mathbf{x})}{\lambda_n} \right] f(\xi)d\xi \\ &= \mathscr{L}^{-1} f \\ &= \int_\Omega G(\mathbf{x},\xi) f(\xi) d\xi, \end{aligned}$$

where $G : \Omega \times \Omega \to \mathbb{R}$ is defined as

$$G(\mathbf{x},\xi) = \sum_{n=1}^{\infty} \frac{\phi_n(\xi)\phi_n(\mathbf{x})}{\lambda_n}$$

and is called the *Green's function* for $\mathscr{L}$. This is referred to as the bilinear expansion of the Green's function. We note that the solution of the problem $\mathscr{L}u = f$ is satisfied by integrating the source term $f(\mathbf{x})$ against the Green's function. This is what we meant by referring to a *pseudoinverse* operator.

Example 11.0.1. For $\mathscr{L}u = -\Delta u = -u_{xx} - u_{yy}$ with $u(x,0) = u_y(x,\pi) = 0$ and $u(0,y) = u(\pi,y) = 0$ the eigenvalues and eigenvectors are given by:

$$\lambda_{n,k} = n^2 + \left(k - \frac{1}{2}\right)^2 \quad \text{and} \quad \phi_{n,k} = \frac{2}{\pi} \sin(nx) \sin\left(\frac{2k-1}{2} y\right),$$

for $k, n \in \mathbb{N}$ so that

$$G(\mathbf{x},\xi) = \frac{4}{\pi^2} \sum_{n,k} \frac{\sin(nx)\sin(\frac{2k-1}{2}y)\sin(n\xi_1)\sin(\frac{2k-1}{2}\xi_2)}{n^2 + \frac{(2k-1)^2}{4}}.$$

The construction here is indicative of the nature of the Green's function, i.e. the inverse of a linear differential operator. The Green's function is the function which convolved against, will invert $\mathscr{L}$, i.e. it defines the inverse operator which naturally is an integral operator. Green's functions are quite convenient to use because once they are found for the linear operator $\mathscr{L}$ then it is quite easy to solve $\mathscr{L}u = f$ for any $f \in L^2(\Omega)$.

Another useful application of Green's functions is that they give a formula for employing the Spectral Decomposition Theorem because the resolvent $R(z) = (z - \mathscr{L})^{-1}$ can be explicitly computed, where it usually makes the most sense to refer to the action of the resolvent on the function $u(\mathbf{x}, t)$ as $R(z)u = (z - \mathscr{L})^{-1}u$ and from here the spectral decomposition theorem can be applied. This is most useful for proving various properties of the linear operator in question rather than for practical computation. Essentially this means that all of the analysis we did on ODEs in a Banach space applies here as well to the infinite-dimensional setting. The key restriction (which we will not address at any length) is that the linear operator $\mathscr{L}^{-1}$ be compact with no essential spectrum (meaning the eigenvalues are discrete), in which case most of the nice theory we developed for finite-dimensional operators carries over.

Remark 11.0.2. Now that we have introduced the idea of a Green's function as the inverse of a linear differential operator, we need to return to methods for computing it, and then actually using it for computational purposes. The eigenfunction expansion method is not the most practical, and is actually unnecessarily lengthy.

The issue is that computation of the Green's function requires some deeper analysis then we have dealt with up to this point. We need to go back and revisit our concept of solutions to a PDE before we can really understand what is going on here.

11.1 Distributions

11.1.1 General definitions

Before we can give a more practical derivation and construction of Green's functions, we need to delve back into some definition based analysis, i.e. we are going back to the etheral life of a mathematician that is far more interested in defining the correct spaces than dealing with the humdrum of physical reality. To begin with, we need to generalize our definition of a function so we can start discussing solutions of PDEs that are not only not differentiable, but aren't even defined as functions. Hence, we seek to extend the ideas that we have come to love for functions, to potentially messier objects that can be used to interpret solutions of PDEs in a more general setting.

Before doing any of this, we first need to refer to a specific class of very nice functions.

Definition 11.1.1. *For an open domain Ω in $\mathbb{R}^n$ let $\mathscr{D}(\Omega) = C_0^\infty(\Omega; \mathbb{R})$ be the collection of functions $f : \Omega \to \mathbb{R}$ that have continuous derivatives of all orders (are infinitely continuously differentiable) and have compact support, i.e., the closure of*

$$\{x \in \Omega : f(x) \neq 0\}$$

is a compact subset of Ω.

Remark 11.1.2. The collection $\mathscr{D}(\Omega)$ is a metric vector space where the metric is a bit complicated to describe. In words, two functions ϕ and ψ in $\mathscr{D}(\Omega)$ are close if $\|\phi - \psi\|$ is small, $\|D\phi - D\psi\|$ is small, $\|D^2\phi - D^2\psi\|$ is small, etc. The point is that the metric vector space $\mathscr{D}(\Omega)$ is not a normed vector space. We call the functions $\phi \in \mathscr{D}(\Omega)$ *test functions* on Ω.

Remark 11.1.3. The fact that each test function $f \in \mathscr{D}(\Omega)$ has compact support inside the open set Ω means that at any point $\mathbf{x}$ of the boundary of Ω there is an open set $B(\mathbf{x}, \varepsilon)$ on which f and its derivative Df both vanish. In particular, f and Df must vanish at each point of the boundary of Ω.

Now that we have these test functions in our back pocket, we are prepared to generalize what we think of as a function.

Definition 11.1.4. *A* distribution *on $\mathscr{D}(\Omega)$ is a continuous linear functional v on $\mathscr{D}(\Omega)$, that is to say, v is a continuous linear transformation from $\mathscr{D}(\Omega)$ to $\mathbb{R}$. We denote the result of evaluating v on a test function $\phi \in \mathscr{D}(\Omega)$ by $v[\phi]$. The set of all distributions of $\mathscr{D}(\Omega)$ is denoted $\mathscr{D}^*(\Omega)$.*

Example 11.1.5. A function $u : \Omega \to \mathbb{R}$ is locally integrable if for every compact subset K of Ω then

$$\int_K |u(\mathbf{x})| \, d\mathbf{x} < \infty.$$

We denote the collection of such functions by $L^1_{\text{loc}}(\Omega)$.

A locally integrable function $u \in L^1_{\text{loc}}(\Omega)$ defines a functional $\mathscr{D}(\Omega)$ by the standard inner product:

$$\phi \mapsto \langle u(\mathbf{x}), \phi(\mathbf{x}) \rangle = \int_\Omega u(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x}.$$

We denote the action of this functional on ϕ by $u[\phi]$. This functional is linear:

$$u[a\phi + b\psi] = \int_\Omega u(a\phi + b\psi) \, d\mathbf{x} = a \int_\Omega u\phi \, d\mathbf{x} + b \int_\Omega u\psi \, d\mathbf{x} = au[\phi] + bu[\psi].$$

It is also continuous, but the proof of continuity requires some functional analysis we don't have (in particular the metric on $\mathscr{D}(\Omega)$). Because the functional is both linear and continuous it is a distribution in $\mathscr{D}^*(\Omega)$. We usually abuse notation and just write u for this distribution.

Typically we express a distribution as though it were a function even when there isn't a $u \in L^1_{\text{loc}}(\Omega)$ that represents the functional through integration. Notationally we typically write $u \in \mathscr{D}^*(\Omega)$ instead of $(u, \phi) \in \mathscr{D}^*(\Omega)$.

Here is a list of properties of distributions that follow as a consequence of the integral representation.

- For $u(\mathbf{x}) \in \mathscr{D}^*(\Omega)$ and $a(\mathbf{x}) \in C^\infty(\Omega)$ there holds $(au, \phi) = (u, a\phi)$ because

$$(au, \phi) = \int_\Omega a(\mathbf{x})u(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} = \int_\Omega u(\mathbf{x})a(\mathbf{x})\phi(\mathbf{x}) \, d\mathbf{x} = (u, a\phi)$$

  and the product of two C^∞ functions is a C^∞ function.

- We define the "weak derivative" of a distribution $u(\mathbf{x}) \in \mathscr{D}^*(\Omega)$ by

$$\left(\frac{\partial u}{\partial x_k}, \phi \right) = - \left(u, \frac{\partial \phi}{\partial x_k} \right)$$

because, by integration by parts, there formally holds

$$\begin{aligned}\left(\frac{\partial u}{\partial x_k},\phi\right) &= \int_\Omega \frac{\partial u}{\partial x_k}(\mathbf{x})\phi(\mathbf{x})\ d\mathbf{x}\\ &= u(\mathbf{x})\phi(\mathbf{x})\Big|_{\partial\Omega} - \int_\Omega u(\mathbf{x})\frac{\partial \phi}{\partial x_k}(\mathbf{x})\ d\mathbf{x}\\ &= -\int_\Omega u(\mathbf{x})\frac{\partial \phi}{\partial x_k}(\mathbf{x})\ d\mathbf{x}\\ &= -\left(u,\frac{\partial \phi}{\partial x_k}\right)\end{aligned}$$

where we have used the vanishing of the test function ϕ on the boundary of Ω.

- Higher order derivatives of $u(\mathbf{x})$ are defined similarly.
- Two distributions $u(\mathbf{x})$ and $v(\mathbf{x})$ in $\mathscr{D}^*(\Omega)$ are the same if $(u,\phi) = (v,\phi)$ for all $\phi \in \mathscr{D}(\Omega)$.

Example 11.1.6 (Heaviside and Dirac Delta 'Functions'). For $\Omega = \mathbb{R}$, consider the distribution $H(x)$ defined by

$$\int_{-\infty}^{\infty} H(x)\phi(x)\ dx = -\Phi(0)$$

where $\Phi(x)$ and $\phi(x)$ are related by

$$\frac{d}{dx}\Phi(x) = \phi(x).$$

There is no ambiguity in the choice of the antiderivative Φ of ϕ because of the vanishing of Φ outside of compact sets that fixes the arbitrary constant of integration.

There is a "function" form for the distribution $H(x)$, namely

$$H(x) = \begin{cases} 0 & \text{if } x < 0,\\ \frac{1}{2} & \text{if } x = 0,\\ 1 & \text{if } x > 0.\end{cases}$$

This can be checked:

$$\int_{-\infty}^{\infty} H(x)\phi(x)\ dx = \int_0^{\infty} \phi(x)\ dx = \lim_{A\to\infty}\Phi(A) - \Phi(0) = -\Phi(0).$$

The limit goes to 0 because $\Phi(x)$ belongs to $\mathscr{D}(\Omega)$, i.e., the test function vanishes outside a compact set. The value of $\frac{1}{2}$ for H at $x = 0$ does not affect the value of the integral; the value of $H(0)$ can be any number you want.

The weak derivative of $H(x)$ is another distribution denoted by $\delta(x)$. Using integration by parts we have

$$\begin{aligned}\int_{-\infty}^{\infty} H(x)\phi'(x)\ dx &= H(x)\phi(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} H'(x)\phi(x)\ dx\\ &= -\int_{-\infty}^{\infty} H'(x)\phi(x)\ dx,\end{aligned}$$

where the evaluations are 0 because $\phi(x) \in \mathscr{D}(\Omega)$, i.e., the test function vanishes outside a compact set. Thus the weak derivative of $H(x)$ is defined by

$$\int_{-\infty}^{\infty} \delta(x)\phi(x)\ dx = \int_{-\infty}^{\infty} H'(x)\phi(x)\ dx = -\int_{-\infty}^{\infty} H(x)\phi'(x)\ dx.$$

Since the distribution $H(x)$ gives -1 times the value of the antiderivative of the test function at $x = 0$, we obtain that the distribution $\delta(x)$ gives the value of the antiderivative of $\phi'(x)$ at $x = 0$, namely $\phi(0)$.

Thus for all $\phi \in \mathscr{D}(\Omega)$ there holds

$$\int_{-\infty}^{\infty} \delta(x)\phi(x)\ dx = \phi(0).$$

The distribution $H(x)$ is called the Heaviside distribution and the distribution $\delta(x)$ is called the *Dirac delta* distribution.

The Dirac delta 'function' is one of the most important distributions, particularly for use in the computation of the Green's function as we will see below. To reiterate the punch line from the previous example, $\delta(x-\xi) = \delta_\xi(x)$ satisfies $(\delta(x-\xi), \phi) = \phi(\xi)$ for all test functions $\phi \in D(\Omega)$. Similar definitions and results hold for dimension $d > 1$.

Remark 11.1.7. It is important to note that the Dirac delta 'function' isn't actually a function at all, but the 'function' appellation in this case is a historical adjective that has been attached to this particular distribution for far too long to correct it now. This distribution is not only critical to the derivation of Green's functions as we see below, but it is also key to many other areas of mathematics, and appears more frequently than many students would ever care to realize.

11.1.2 Distributions and PDE's

We will introduce a handful of ways in which distributions are useful in the study of PDEs. For instance in what follows we will generalize solutions of PDEs to distributions, and not just differentiable functions.

Definition 11.1.8. *The adjoint of a linear operator $\mathscr{L}$ is the linear operator $\mathscr{L}^*$ that satisfies*

$$(\mathscr{L}(u), \phi) = (u, \mathscr{L}^*(\phi)) \text{ for all } \phi \in \mathscr{D}(\Omega).$$

We will sidestep the issue of existence of the adjoint; in simple cases of linear differential operators, an integration by parts argument will give the existence of the adjoint.

Definition 11.1.9. *For a linear PDE $\mathscr{L}(u) = f$ on a domain Ω, with appropriate boundary conditions, and a distribution $f \in \mathscr{D}^*(\Omega)$, we say that $u \in \mathscr{D}^*(\Omega)$ is a distribution solution of $\mathscr{L}(u) = f$ if u satisfies*

$$(u, \mathscr{L}^*(\phi)) = (f, \phi) \text{ for all } \phi \in \mathscr{D}(\Omega).$$

In the integral presentation this is

$$\int_\Omega u(\mathbf{x})\mathscr{L}^*(\phi(\mathbf{x}))\ d\mathbf{x} = \int_\Omega f(\mathbf{x})\phi(\mathbf{x})\ d\mathbf{x} \text{ for all } \phi \in \mathscr{D}(\Omega).$$

Definition 11.1.10. *The* fundamental solution *for a linear operator $\mathscr{L}$ is the distribution solution $u(\mathbf{x}, \xi)$ of $\mathscr{L}(u(\mathbf{x}, \xi)) = \delta(\mathbf{x} - \xi)$.*

Remark 11.1.11. We show that the fundamental solution $u(\mathbf{x}, \xi)$ is the Green's function $G(\mathbf{x}, \xi)$ for $\mathscr{L}$.

To see that this fundamental solution $u(\mathbf{x}, \xi)$ is indeed the Green's function we informally derived early, we use the notation

$$\delta_{\mathbf{x}}(\xi) = \delta(\mathbf{x} - \xi).$$

For $f \in \mathscr{D}(\Omega)$ there holds $(u(\mathbf{x}, \xi), \mathscr{L}^*(f(\xi)) = (\mathscr{L}(u(\mathbf{x}, \xi)), f(\xi)) = (\delta_{\mathbf{x}}(\xi), f(\xi))$, and so

$$(\mathscr{L}(u(\mathbf{x}, \xi)), f(\xi)) = (\delta_{\mathbf{x}}(\xi), f(\xi)) = f(\mathbf{x}) = \mathscr{L}(u(\mathbf{x}))$$

where the third equality follows because the Dirac Delta distribution returns the value of $f(\xi)$ when $\xi = \mathbf{x}$, i.e., $\mathbf{x} - \xi = 0$.

Since the pairing is represented by an integral we obtain

$$\int_\Omega \mathscr{L}(u(\mathbf{x}, \xi)) f(\xi)\ d\xi = \mathscr{L}(u(\mathbf{x})).$$

The linear operator $\mathscr{L}$ acts on $\mathbf{x}$, not on ξ (so $f(\xi)$ is a scalar for $\mathscr{L}$), and so we obtain

$$\mathscr{L}\left(\int_\Omega u(\mathbf{x}, \xi) f(\xi)\ d\xi\right) = \mathscr{L}(u(\mathbf{x})).$$

Assuming the linear operator $\mathscr{L}$ has trivial kernel (it does not have 0 as an eigenvalue, which we assumed when deriving the Green's function), we obtain

$$\int_\Omega u(\mathbf{x}, \xi) f(\xi)\ d\xi = u(\mathbf{x}).$$

Thus we see that the fundamental solution has exactly the same property as the Green's function, and so the Green's function is the fundamental solution, $\mathscr{L}(G(\mathbf{x}, \xi)) = \delta(\mathbf{x} - \xi)$.

It is the mathematical properties of distributions that enable this other approach to computing the Green's function.

It turns out that computation of the Green's function isn't the only reason why we brought up the definition of a distribution (although it is the primary motivation here), but such general ideas are beneficial when interpreting solutions to PDEs. Before we can do this, we recall the following definition.

Definition 11.1.12. *Let $\mathscr{L}$ be a linear operator on the domain Ω (with appropriate boundary conditions) and suppose $f \in L^1_{\text{loc}}(\Omega)$. A weak solution of $\mathscr{L}(u) = f$ is a function $u \in L^1_{\text{loc}}(\Omega)$ such that*

$$(u, \mathscr{L}^*(\phi)) = (f, \phi) \text{ for all } \phi \in \mathscr{D}(\Omega).$$

In terms of the integral presentation of the pairing this condition is

$$\int_\Omega u(\mathbf{x}) \mathscr{L}^*(\phi(\mathbf{x}))\ d\mathbf{x} = \int_\Omega f(\mathbf{x}) \phi(\mathbf{x})\ d\mathbf{x}.$$

Remark 11.1.13. Note that weak solutions of a PDE need not be differentiable. In fact, the definition given here allows for weak solutions that only need to be integrable. This is because the differential operator in the definition of the weak solution never acts on $u(\mathbf{x})$ itself, but its adjoint acts on $\phi(\mathbf{x})$ which is infinitely continuously differentiable. Clearly this expands the concept of a solution to the PDE well beyond our traditional one.

In particular, weak solutions allow the presence of shocks, and discontinuities that don't allow evaluation of the actual differential operator, but must be evaluated via the adjoint as illustrated here. This generalization of solutions is essential to a thorough understanding of several physical phenomena that are not everywhere infinitely continuously differentiable, including the presence of shocks.

This definition of a weak derivative also demonstrates the weak form of a PDE. We have already seen integral forms of conservation laws, and weak forms of a PDE are similar, but now we are taking the adjoint of the PDE operating on the set of test functions instead.

Example 11.1.14. The weak form of the linear PDE $-u_{xx} = f$ for $f \in L^1_{\text{loc}}(\mathbb{R})$ is

$$(u, \phi_{xx}) = (f, \phi) \text{ for all } \phi \in D(\mathbb{R}).$$

In the integral presentation of the pairing this is

$$\int_{-\infty}^{\infty} u(x)\phi_{xx}(x)\ dx = \int_{-\infty}^{\infty} f(x)\phi(x)\ dx.$$

The adjoint of $\mathscr{L}(u) = -u_{xx}$ is $\mathscr{L}^*(\phi) = -\phi_{xx}$ (it is self-adjoint) because

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathscr{L}(u(x))\phi(x)\ dx &= -\int_{-\infty}^{\infty} u_{xx}(x)\phi(x)\ dx \\
&= -u_{xx}\phi_x\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} u_x(x)\phi_x(x)\ dx \\
&= u_x(x)\phi(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} u(x)\phi_{xx}\ dx \\
&= -\int_{-\infty}^{\infty} u(x)\phi_{xx}(x)\ dx \\
&= \int_{-\infty}^{\infty} u(x)\mathscr{L}^*(\phi(x))\ dx.
\end{aligned}$$

Definition 11.1.15. *A solution to $\mathscr{L}u = f$ that is $C^2(\Omega)$ (we are assuming that $\mathscr{L}$ is at most second order) is called a classical or strong solution.*

Thus there are three types of solutions:

(1) Distribution solutions: distributions, may not have a functional form at all.

(2) Weak Solutions: may have a functional form, but derivatives are 'weak' or calculated in the distributional sense.

(3) Strong/classical: standard $C^2(\Omega)$ (for second-order PDE).

In the theory of PDEs, existence of weak solutions is often an easier task than existence of strong or classical solutions.Then, sometimes, using what is called "boot-strapping" argument, the weak solution can be shown to be a strong solution. The space $L^1_{\text{loc}}(\Omega)$ contains the candidates for the strong solutions, but the existence techniques for weak solutions in $L^1_{\text{loc}}(\Omega)$ don't guarantee the solution is C^n.

There are several canonical nonlinear PDEs for which weak solutions are the only guaranteed solutions i.e. it is unknown if strong solutions exist, but weak solutions do exist. In some cases this may seem to be a negative aspect of the PDE itself, but if we are trying to model certain discontinuous behavior such as shocks then such a feature of the PDE is desirable and advantageous.

Remark 11.1.16. For our purposes, we have introduced distributions and weak solutions so we can appropriately define and compute Green's functions. This is one important aspect of the theory of distributions but is most certainly not the only one. In fact the analysis of nonlinear PDEs is littered with the use of distributions and weak solutions.

Example 11.1.17. One common place where weak solutions of a PDE are particularly important are in the development of finite element numerical methods. There are a lot of reasons for considering finite elements, but the real motivation is not apparent until you move to greater than two dimensions in spatial coordinates. One motivation for using finite elements is irregular geometry or narrow/small regions in the spatial domain where important aspects of the solution will occur. A finite element mesh is then created that covers these regions as desired.

Once the mesh is created then a set of test functions is generated based on this mesh. A standard set of test functions are those that are 1 at a given mesh point and then decay rapidly to 0 at nearby mesh points. These test functions $\phi_k(\mathbf{x})$ can then be used to form the weak form of the PDE. The solution is represented as a sum of these test functions with appropriate weights, and the weights are solved for via the weak form of the PDE. The solution reduces to a set of linear equations, which can be attached to any given time-stepping method for a time dependent PDE, or just a single solution can be used for a time-independent PDE.

This is all horribly confusing, so we will consider something a little more explicit. For now we restrict our attention to the one-dimensional problem of solving $-u_{xx} = f(x)$ on the interval $[0, 1]$. If we select a set of grid points (mesh in 1D) $\{x_k\}$ then we will define our test functions as

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i} & \text{if } x \in [x_{i-1}, x_i] \\ \frac{x_{i+1} - x}{h_{i+1}} & \text{if } x \in [x_i, x_{i+1}] \\ 0 & \text{otherwise,} \end{cases}$$

where $h_i = x_i - x_{i-1}$. These hat or tent functions are commonly used for the most basic test functions for the development of finite elements. Because each of these $\phi_i(x)$ vanish at the endpoints $x = 0$ and $x = 1$ then we can rewrite our PDE $-u_{xx} = f(x)$ in weak form as:

$$\int_0^1 \phi_i'(x) u_x(x) dx = \int_0^1 f(x)\phi_i(x) dx,$$

which must be satisfied for every $\phi_i(x)$.

We use the ansatz

$$u(x) = \sum_i c_i \phi_i(x)$$

which after substitution into the weak form of the PDE (the first and third integrals) leads to a linear system of equations in the coefficients c_i.

Note that technically the test functions ϕ_i do not belong to $D((0, 1))$ but they have C^∞ approximations that do.

11.2 Computation of the Green's Function in one dimension

We have employed the Dirac Delta distribution $\delta(\mathbf{x} - \xi)$ to define the Green's function $G(\mathbf{x}, \xi)$ for a linear operator $\mathscr{L}$ as the solution $\mathscr{L}(G(\mathbf{x}, \xi)) = \delta(\mathbf{x} - \xi)$, i.e.,

$$\int_\Omega G(\mathbf{x}, \xi) \mathscr{L}^*(\phi(\xi))\, d\xi = \int_\Omega \delta(\mathbf{x} - \xi)\phi(\xi)\, d\xi = \phi(\mathbf{x}) \text{ for all } \phi \in \mathscr{D}(\Omega).$$

Generally speaking, finding $G(\mathbf{x},\xi)$ by solving the distributional equation $\mathscr{L}(G(\mathbf{x},\xi)) = \delta(\mathbf{x}-\xi)$ isn't much easier than finding the eigenvalues and eigenfunctions of $\mathscr{L}$ and computing $G(\mathbf{x},\xi)$ as series.

But however we seek to find the Green's function there are properties it always has. Recall that when λ_n are the eigenvalues and ϕ_n are the eigenfunctions of a linear operator $\mathscr{L}$, then the Green's function for $\mathscr{L}$ is

$$G(\mathbf{x},\xi) = \sum_{n=1}^{\infty} \frac{\phi_n(\mathbf{x})\phi_n(\xi)}{\lambda_n}$$

provided 0 is not an eigenvalue of $\mathscr{L}$.

A consequence of this series representation of the Green's function is that it is symmetric in $\mathbf{x}$ and ξ, i.e.,

$$G(\mathbf{x},\xi) = G(\xi,\mathbf{x}) \text{ for all } \mathbf{x},\xi \in \overline{\Omega}.$$

Since we are expecting the Green's function to give the solution of the PDE $\mathscr{L}(u) = f$, i.e.,

$$u(\mathbf{x}) = \int_{\Omega} G(\mathbf{x},\xi) f(\xi)\, d\xi,$$

we expect the Green's function, when multiplied by f, is integrable so that the integral defining $u(\mathbf{x})$ is defined.

You may want the Green's function to be at least continuous to guarantee this (and we will shortly), but continuity of the Green's function cannot always be guaranteed. It is known that the Green's function for the two-dimensional Laplace operator is not continuous, but is such that $G(\mathbf{x},\xi)f(\xi)$ is integrable.

When $\mathscr{L}$ is a regular SL problem then there is an algorithmic approach to computing the Green's function for $\mathscr{L}$. We focus on some of the simplest types of regular SL problems in dimension one.

Example 11.2.1. Consider as a first example the 1D problem given by

$$\mathscr{L}w = w'' - k^2 w, \quad \text{with} \quad w(0) = w(1) = 0.$$

We want to find the solution to

$$\mathscr{L}w = \delta(x-\xi),$$

with these BC's. For $x \neq \xi$ this has solutions ($w'' = k^2 w$) of the form

$$w(x) = c_0 \sinh(kx) + c_1 \cosh(kx).$$

We can not simultaneously satisfy both boundary conditions with this solution, but this is expected because of the impulse at $x = \xi$, i.e. we are supposing that $\delta(x-\xi)$ represents an infinite pulse of input energy at $x = \xi$.

Instead of forcing both boundary conditions to simultaneously be satisfied, we define a piecewise form of the solution. Thus for $x < \xi$ we force the solution to satisfy the boundary condition at $x = 0$ and for $x > \xi$ the solution satisfies the boundary condition at $x = 1$. This leads to the piecewise defined solution for the Green's function:

$$G(x;\xi) = \begin{cases} c_0 \sinh(kx) & x < \xi \\ c_1(\cosh(kx) - \coth(k)\sinh(kx)) & x > \xi \end{cases}$$

This leaves us with the two constants $c_0,\ c_1$ (which are actually functions of ξ) to be determined from two additional considerations.

For the first requirement, we enforce the condition that $G(x;\xi)$ is continuous at $x = \xi$, i.e.

$$\begin{aligned} c_0 \sinh(k\xi) &= c_1(\cosh(k\xi) - \coth(k)\sinh(k\xi)) \\ &\Rightarrow c_0 = c_1(\coth(k\xi) - \coth(k)) \\ &\Rightarrow G(x;\xi) = c_1 \begin{cases} (\coth(k\xi) - \coth(k))\sinh(kx) & x < \xi \\ \cosh(kx) - \coth(k)\ \sinh(kx) & x > \xi \end{cases} \end{aligned}$$

To find out what c_1 is, we need to return to the PDE and use the properties we derived earlier of the delta function to determine exactly how the Green's function behaves near the point $x = \xi$. For the Green's function candidate to be the Green's function, it should satisfy $\mathscr{L}(G(x,\xi)) = \delta(x-\xi)$, i.e.,

$$\frac{d^2}{dx^2}G(x,\xi) - k^2 G(x,\xi) = \delta(x-\xi).$$

We apply integration to this ODE on the interval $[\xi-\varepsilon, \xi+\varepsilon]$, i.e., on a small interval near $x = \xi$.

This gives

$$\int_{\xi-\varepsilon}^{\xi+\varepsilon} \frac{d^2}{dx^2}G(x,\xi)\ dx - k^2 \int_{\xi-\varepsilon}^{\xi+\varepsilon} G(x,\xi)\ dx = \int_{\xi-\varepsilon}^{\xi+\varepsilon} \delta(x-\xi)\ dx.$$

To evaluate the integral involving the Dirac Delta distribution we notice that $\delta(x-\xi)$ is being multiplied by any function $\phi \in \mathscr{D}(\Omega)$ that is equal to 1 on the interval $[\xi-\varepsilon, \xi+\varepsilon]$ and is zero outside some larger compact interval, so that

$$\int_{\xi-\varepsilon}^{\xi+\varepsilon} \delta(x-\xi)\ dx = \int_{\xi-\varepsilon}^{\xi+\varepsilon} \delta(x-\xi)\phi(x)\ dx = \phi(\xi) = 1.$$

To the integral with the second-order derivative we apply the (measure-theoretic version of the) FTC to get

$$\int_{\xi-\varepsilon}^{\xi+\varepsilon} \frac{d^2}{dx^2}G(x,\xi)\ dx = \frac{\partial}{\partial x}G(x,\xi)\bigg|_{x=\xi+\varepsilon} - \frac{\partial}{\partial x}G(x,\xi)\bigg|_{x=\xi-\varepsilon}.$$

Putting these pieces together gives

$$\frac{\partial}{\partial x}G(x,\xi)\bigg|_{x=\xi+\varepsilon} - \frac{\partial}{\partial x}G(x,\xi)\bigg|_{x=\xi-\varepsilon} - k^2 \int_{\xi-\varepsilon}^{\xi+\varepsilon} G(x,\xi)\ dx = 1.$$

Taking the limit in this as $\varepsilon \to 0$ gives the jump condition

$$\frac{\partial}{\partial x}G(x,\xi)\bigg|_{x=\xi^+} - \frac{\partial}{\partial x}G(x,\xi)\bigg|_{x=\xi^-} = 1,$$

where the integral with G goes to 0 by the continuity of G when $x = \xi$.

For this specific problem, this becomes:

$$-c_1 k[\coth(k\xi) - \coth(k)]\cosh(k\xi) + c_1[k\sinh(k\xi) - k\coth(k)\cosh(k\xi)] = 1$$
$$\Rightarrow c_1 = \frac{1}{k\sinh(k\xi) - k\coth(k\xi)\cosh(k\xi)}.$$

Therefore we obtain the Green's function

$$G(x,\xi) = \begin{cases} \dfrac{(\coth(k\xi) - \coth(k))\sinh(kx)}{k\sinh(k\xi) - k\coth(k\xi)\cosh(k\xi)} & \text{if } x < \xi, \\ \dfrac{\cosh(kx) - \coth(k)\sinh(kx)}{k\sinh(k\xi) - k\coth(k\xi)\cosh(k\xi)} & \text{if } x > \xi. \end{cases}$$

This may not look symmetric, $G(x,\xi) = G(\xi,x)$, but it actually is if you go through everything very carefully.

With this Green's function we now have the integral representation of the solution

$$w(x) = \int_0^1 G(x,\xi) f(\xi)\, d\xi$$

for the PDE $\mathscr{L}(w) = f$ where $\mathscr{L}(w) = w'' - k^2 w$, subject to the boundary conditions $w(0) = 0$ and $w(1) = 0$.

For the general derivation of the Green's function, we consider a boundary value problem $\mathscr{L}u = \delta(x-\xi)$ with homogeneous boundary conditions

$$B_1 u(a) = \alpha_1 u(a) + \alpha_2 u(a) = 0$$
$$B_2 u(b) = \beta_1 u(b) + \beta_2 u(b) = 0,$$

(typically $\mathscr{L}$ is actually a Sturm-Liouville operator, but this is not always necessary) where the two algebraic operators B_i define the type of prescribed boundary condtion.

(i) First construct the Green's function as piecewise defined solutions to this problem satisfying the left and right boundary conditions separately.

(ii) Enforce continuity at $x = \xi$ to remove one of the remaining unknown constants.

(iii) Returning to the ODE, remove the final constant via a 'jump condition' by integrating over a small interval about $x = \xi$. Sturm-Liouville operators will result in the same type of jump condition shown in the previous example.

Remark 11.2.2. Some final important properties of the Green's function for the Sturm Liouville operator $\mathscr{L}u = -(pu')' + qu$ with boundary conditions as described previously, are:

(i) $\mathscr{L}G(x;\xi) = \delta(x-\xi)$ in the weak (distributional as well) sense.

(ii) $G(x;\xi)$ satisfies the specified boundary conditions.

(iii) $G(x;\xi)$ is continuous in the chosen domain (the interval $[a,b]$ for example).

(iv) $G(x;\xi)$ is differentiable for $x \neq \xi$ and

$$\frac{\partial G}{\partial x}(\xi^+;\xi) - \frac{\partial G}{\partial x}(\xi^-;\xi) = -\frac{1}{p(\xi)},$$

which is the jump condition at $x = \xi$.

(v) The Green's function is symmetric, i.e. $G(x;\xi) = G(\xi;x)$.

To cement these ideas even further, we consider the canonical example of the negative Laplacian for a particular set of boundary conditions.

Example 11.2.3. We will construct the Green's function for $\mathscr{L}u = -u''$ where $u(0) = 0 = u'(1)$. For $x \neq \xi$ the solution looks like $G(x;\xi) = c_0 + c_1 x$. This means that the solution is given by:

$$G(x;\xi) = \begin{cases} c_1 x & x < \xi \\ c_0 & x > \xi, \end{cases}$$

where c_0 and c_1 are functions of ξ. Applying the continuity condition at $x = \xi$ leads to

$$c_1 \xi = c_0,$$

so that

$$G(x;\xi) = c_1 \begin{cases} x & x < \xi \\ \xi & x > \xi. \end{cases}$$

The jump condition for the Green's function is

$$-\int_{\xi-\varepsilon}^{\xi+\varepsilon} \frac{d^2}{dx^2} G(x,\xi)\, dx = \int_{\xi-\varepsilon}^{\xi+\varepsilon} \delta(x-\xi)\, dx = 1.$$

Carrying out the integral on the left we obtain

$$-\left[\frac{\partial}{\partial x} G(x,\xi)\Big|_{x=\xi+\varepsilon} - \frac{\partial}{\partial x} G(x,\xi)\Big|_{x=\xi-\varepsilon} \right] = 1.$$

Computing the partial derivatives, carrying out the evaluations, and taking the limit as $\varepsilon \to 0$ gives

$$c_1 = 1,$$

so that

$$G(x;\xi) = \begin{cases} x & x < \xi \\ \xi & x > \xi. \end{cases}$$

Suppose that we wanted to use this information to find the solution to the boundary value problem $-u'' = \cos(x)$ where $u(0) = u'(1) = 0$. Then we simply find that

$$\begin{aligned} u(x) &= \int_0^1 G(x;\xi)\cos(\xi) d\xi \\ &= \int_0^x \xi \cos(\xi) d\xi + \int_x^1 x \cos(\xi) d\xi, \end{aligned}$$

which while non-trivial to calculate is really not a bad integral.

There are approaches to define and compute the Green's function for higher order differential operators, and it proves to be a powerful tool in the analysis of linear PDEs. One should keep in mind that the Green's function is specific not only to the linear differential operator and the corresponding boundary conditions, but is very specific to the domain in question. This combined with the linearity of the operators under consideration provides an avenue for constructing solutions to linear PDEs in complicated domains by linear combinations of solutions from much simpler domains. This technique is often referred to as the method of images and plays a significant role in the development of several numerical methods as well as the analysis of solutions.

11.3 Green's function in the plane

Of course reality is rarely one-dimensional, even though that would make for an interesting existential question to ask the philosophers. Instead, we must consider what happens for higher-dimensional Green's functions. We will see what happens in the planar case, and you will get to investigate higher-dimensional issues in the exercises (so exciting!). We focus here only on the full plane, i.e. all of $\mathbb{R}^2$ but we do comment on what happens for other domains. In addition we will focus on the Poisson equation

$$\Delta u = f,$$

although the same derivation can be imitated for other second-order PDEs as well.

Example 11.3.1. Consider the PDE $\Delta u = f(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^2$ on the whole plane. The Green's function $G(\mathbf{x};\boldsymbol{\xi})$ will then satisfy

$$\Delta G = \delta(\mathbf{x} - \boldsymbol{\xi}).$$

In particular, that means $\Delta G(\mathbf{x}, \boldsymbol{\xi}) = 0$ when $\mathbf{x} \neq \boldsymbol{\xi}$. If we restrict attention to the region Ω with boundary $\partial\Omega$ then we also enforce the Dirichlet condition $u(\mathbf{x}) = g(\mathbf{x})$ for $\mathbf{x} \in \partial\Omega$, but, as shown below, we want $G(\mathbf{x};\boldsymbol{\xi}) = 0$ for $\mathbf{x} \in \partial\Omega$.

Recall one of Green's identities in this setting (although the Green's function is continuous but not differentiable, these identities still hold so long as u is a test function on $\mathbb{R}^2$):

$$\int_\Omega (u\Delta G - G\Delta u)\, d\mathbf{x} = \int_{\partial\Omega} (u\nabla G - G\nabla u) \cdot \mathbf{n} dA.$$

Assume that u satisfies the PDE $\Delta u = f$ and the boundary condition. Combining this with the PDE for G and the properties of the delta function gives

$$u(\mathbf{x}) = \int_\Omega G(\mathbf{x};\boldsymbol{\xi}) f(\boldsymbol{\xi}) d\boldsymbol{\xi} + \int_{\partial\Omega} g(\mathbf{x}) \nabla G(\mathbf{x};\boldsymbol{\xi}) \cdot \mathbf{n} dA$$

Thus if we find $G(\mathbf{x};\boldsymbol{\xi})$ that satisfies the previously stated conditions then we have the solution $u(\mathbf{x})$.

To derive $G(\mathbf{x};\boldsymbol{\xi})$ we first resort to cylindrical coordinates, i.e. if we let r be the distance from the point $\mathbf{x}$, and note that this problem is axisymmetric i.e. there is no dependence on the angle θ. Then we find that the Poisson equation for G becomes:

$$\frac{\partial^2 G}{\partial r^2} + \frac{1}{r}\frac{\partial G}{\partial r} = \delta(r),$$

with $G(\mathbf{x};\boldsymbol{\xi})$ vanishing on $\partial\Omega$. For $r > 0$, we can integrate this ODE to find that $G(r) = a\log(r) + b$ where b is a constant that we can specify/prescribe later for specific domains/boundary conditions. To find what a is, and to incorporate the point source at $r = 0$ we integrate the entire PDE in a small ball D_ε of radius ε about $r = 0$. This yields

$$1 = \int_{D_\varepsilon} \delta(r)d\mathbf{x} = \int_{D_\varepsilon} \Delta G d\mathbf{x} = \int_{\partial D_\varepsilon} \nabla G \cdot \mathbf{n} dA = \int_0^{2\pi} \varepsilon \frac{a}{\varepsilon} d\theta = 2\pi a.$$

Thus we refer to the Fundamental Solution or Green's function for the Poisson equation in the plane as $G(r) = \frac{1}{2\pi}\log(r)$, where $r = \|\mathbf{x} - \boldsymbol{\xi}\|$.

For a specific domain which is a subset of the plane, then we would specify the Green's function as

$$G(\mathbf{x};\boldsymbol{\xi}) = \frac{1}{2\pi}\log(r) + h(\mathbf{x};\boldsymbol{\xi}),$$

where

$$\Delta h = 0, \quad \boldsymbol{\xi} \in \Omega,$$
$$h(\mathbf{x};\boldsymbol{\xi}) = -\frac{1}{2\pi}\log(r), \quad \boldsymbol{\xi} \in \partial\Omega.$$

The boundary condition here forces the full Green's function $G(\mathbf{x};\boldsymbol{\xi})$ to satisfy the proper type of boundary condition.

This seems like a pretty special case, and it sort of is, but we can modify this Green's function to work for any reasonable domain in $\mathbb{R}^2$ so long as we keep track of everything carefully. Most frequently, a half plane or particular region of the plane is used. You can also see what would happen from the derivation in this example if Neumann boundary conditions were used instead...you just need to watch the correct use of Green's identities/integration by parts.

Exercises

Note to the student: Each section of this chapter has several corresponding exercises, all collected here at the end of the chapter. The exercises between the first and second line are for Section 1, the exercises between the second and third lines are for Section 2, and so forth.

You should **work every exercise** (your instructor may choose to let you skip some of the advanced exercises marked with *). We have carefully selected them, and each is important for your ability to understand subsequent material. Many of the examples and results proved in the exercises are used again later in the text. Exercises marked with ⚠ are especially important and are likely to be used later in this book and beyond. Those marked with † are harder than average, but should still be done.

Although they are gathered together at the end of the chapter, we strongly recommend you do the exercises for each section as soon as you have completed the section, rather than saving them until you have finished the entire chapter.

11.1. Prove that $\alpha(x)\delta'(x) = -\alpha'(0)\delta(x) + \alpha(0)\delta'(x)$ where the prime here refers to the distributional derivative of the delta function, and $\alpha \in C^\infty(\mathbb{R})$.

11.2. Compute the distributional derivative of $H(x)\cos(x)$, where H is the Heaviside function. Does the derivative exist in the weak sense?

11.3. Find the distribution of $\left(\frac{d}{dx} - \lambda\right)\left(H(x)e^{\lambda x}\right)$ where $H(x)$ is the Heaviside function. Hint: The expression $\left(\frac{d}{dx} - \lambda\right)\left(H(x)e^{\lambda x}\right)$ is an operator $\left(\frac{d}{dx} - \lambda\right)$ acting on a distribution $\left(H(x)e^{\lambda x}\right)$. Your assignment is to 1. figure out how to write the result as a distribution and 2. prove that your answer is correct.

11.4. Assume that $\mathscr{L}$ is a linear operator with adjoint $\mathscr{L}^*$. Given a PDE of the form $\mathscr{L}u = g$, the weak form of the PDE is the equation $(u, \mathscr{L}^* f) = (g, f)$ for any $f \in \mathscr{D}$. If u is a function that satisfies the equation $\mathscr{L}u = g$, then it is called a *strong* solution of the PDE. If the distribution defined by $(u, f) = \int_\Omega uf \, d\mathbf{x}$ satisfies the weak form of the PDE $(u, \mathscr{L}^* f) = (g, f)$ for any $f \in \mathscr{D}$, then u is called a *weak solution* of the PDE. Prove that any strong solution defines a weak solution.

11.5. Find the weak form of each of the following PDEs:

- $u_{xx} - u_{xy} + \alpha u = 0$.
- $-u_{xx} + au_x = 0$.
- $-u_{xx} - u_{yy} - u_{zz} = 0$.

11.6. Derive the Green's function for the linear operator

$$\mathscr{L}u = -u''$$

with boundary conditions $u(0) = u(1) = 0$.

11.7. Try to derive the Green's function for the linear operator

$$\mathscr{L}u = -u''$$

with boundary conditions $u'(0) = u'(1) = 0$. Can you derive the Green's function here? Why or why not?

11.8. Derive the Green's function for the linear operator

$$\mathscr{L}u = u'' - k^2 u$$

with boundary conditions $u'(0) = u(1) = 0$.

11.9. Find the fundamental solution for Poisson's equation in $\mathbb{R}^3$. This means you will look for axisymmetric solutions in spherical coordinates.

11.10. Find the Green's function in the upper half plane for Dirichlet boundary conditions on the horizontal axis.

Notes