MATH2103: Lecture Note on Numerical Solution of Partial Differential Equations

Shixiao W. Jiang

Institute of Mathematical Sciences, ShanghaiTech University, Shanghai 201210, China

jiangshx@shanghaitech.edu.cn

2025 年 11 月 5 日

Contents

Chapter 1

Consistency, convergence and stability

1.1 Definition of the problems considered

In this chapter we shall gather together and formalise definitions that we have introduced in earlier chapters. This will enable us to state and prove the main part of the key **Lax Equivalence Theorem**. For simplicity we will not aim at full generality but our definitions and arguments will be consistent with those used in a more general treatment.

Thence we write the general form of the problems considered as

$$\begin{aligned} &\frac{\partial u}{\partial t} = \mathbf{L}(u) \quad \text{in } \Omega \times (0, t_F], \\ &\mathbf{g}(u) = g_0 \quad \text{on } \partial\Omega_1 \subset \partial\Omega, \\ &u = u^0 \quad \text{on } \Omega \text{ when } t = 0. \end{aligned} \tag{5.1}$$

We shall always assume that (5.1) defines a well-posed problem, in a sense which we shall define later; broadly speaking, it means that a solution always exists and depends continuously on the data.

1.2 The finite difference mesh and norms

Our finite difference approximation will be defined on a fixed mesh, with the time interval Δt constant both over the mesh and at successive time steps. The region Ω is covered by a mesh which for simplicity we shall normally assume has uniform spacing Δx, Δy, ... in Cartesian co-ordinates, or Δr, $\Delta\theta$, ... in polar co-ordinates. Individual values at mesh points will be denoted by U_j^n; in two or more space dimensions the subscript j will be used to indicate a multi-index, as a condensed notation for $U_{j,k}^n$, $U_{j,k,l}^n$, etc. We shall assume that a fixed, regular finite difference scheme is applied to a set of points where U_j^n is to be solved for and whose subscripts j lie in a set J_Ω, and **it is only these points which will be incorporated in the norms**. Usually this will be just the interior points of the mesh; and this means that where made necessary by curved boundaries, derivative boundary conditions etc., extrapolation to fictitious exterior points is used to extend the regular scheme to as many points as possible – see, e.g., Section 3.4 and the use of (3.35). There are other exceptional cases to consider too; when the regular finite difference operator is used at points on a symmetry boundary, as in Section 6.5 below, and also at points on boundaries where the boundary conditions are periodic, then these points are also included in J_Ω. The values of $\mathbf{U}$ at all such points on time level n will be denoted by $\mathbf{U}^n$:

$$\mathbf{U}^n := \{U_j^n, j \in J_\Omega\}. \tag{5.2}$$

To simplify the notation we will consider schemes which involve only two time levels: for one-step methods this means that each U_j^n, if a vector, has the same dimension as u. However, as we have seen with the leap-frog method in Section 4.9, we can include multi-step methods by extending the dimension of U_j^n compared with u. For example, if a scheme involves three time levels, so that U^{n+1} is given in terms of U^n and U^{n-1}, we can **define a new vector $\bar{\mathbf{U}}^n$ with twice the dimension, whose elements are those of** U^n and U^{n-1}.

To compare $\mathbf{U}$ with u we need to introduce norms which can be used on either, and in particular on their difference. Thus we first denote by u_j^n mesh values of the function $u(x,t)$ which will usually be the point values $u(x_j, t_n)$. We hope to show that the mesh values of $\mathbf{U}$ converge to these values of u. Then as for the mesh point values U_j^n above we define

$$\mathbf{u}^n := \{u_j^n, j \in J_\Omega\}. \tag{5.3}$$

We shall consider just two norms. Firstly, the **maximum norm** is given by

$$\|U^n\|_\infty := \max\{|U_j^n|, j \in J_\Omega\}. \tag{5.4}$$

If we evaluate the maximum norm of $\mathbf{u}^n$ the result will approximate the usual supremum norm $\|u\|_\infty$ with u considered as a function of x at fixed time t_n, but will not in general be equal to it. The norms will only be equal if the maximum value of the function $|u(x,t_n)|$ is attained at one of the mesh points.

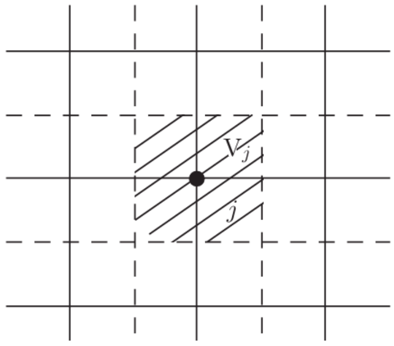

Fig. 5.1. Definition of control volume.

Figure 1.1: Definition of control volume.

Secondly, we shall use a discrete ℓ_2 norm which will approximate the integral L_2 norm. To do so, we introduce a 'control volume' with measure V_j associated with each interior mesh point: these will be non-overlapping elements whose union approximates Ω. Usually, as shown in Fig. 5.1, a mesh point $\mathbf{x}_j$ will lie at the centre of the control volume – see also Section 4.7 on finite volume methods; but this need not be the case so long as there is a one-to-one correspondence between mesh points and control volumes. In three-dimensional Cartesian geometry, $V_j = \Delta x \Delta y \Delta z$; in three-dimensional cylindrical geometry, $V_j = r_j \Delta\theta \Delta r \Delta z$, and so on. Then, we define

$$\|U^n\|_2 := \left\{ \sum_{j \in J_\Omega} V_j |U_j^n|^2 \right\}^{1/2}. \tag{5.5}$$

For mesh points near the boundary the control volume may or may not be modified to lie wholly in Ω. In either case, the sum in (5.5) clearly approximates an integral so that $\|u^n\|_2$ approximates but does not in general equal the integral L_2 norm

$$\|u(\cdot, t_n)\|_2 := \left[\int_\Omega |u(x, t_n)|^2 dV \right]^{1/2} \tag{5.6}$$

at time t_n. However, if we define u_j^n as the root mean square value of $u(x, t_n)$ averaged over the jth control volume we clearly do have an exact match; we saw in Section 2.14 the value of making a similar interpretation when modelling heat conservation properties. For a single differential equation the notation $|U_j^n|$ is clear; if we are dealing with a system of differential equations, $\mathbf{U}_j^n$ is a vector and $|\mathbf{U}_j^n|$ denotes a norm of this vector. The choice of which vector norm to use is **immaterial** to the subsequent analysis, but of course it must be used consistently throughout.

Remark 1.2.1 ***We should perhaps note here some of the techniques in common practical use that are not included in this general framework. Many of them have to do with adaptivity: choosing the next time step on the basis of a current error estimate; choosing a backward or forward difference in a solution dependent manner as in the upwind scheme of (4.20); or locally refining the mesh, e.g., to follow a steep gradient. Some of these would require major changes in the analysis. On the other hand, to cover the case of a refinement path composed of nonuniform meshes would not be very difficult.***

1.3 Finite difference approximations

The general form of difference scheme we shall consider will be written

$$B_1 U^{n+1} = B_0 U^n + F^n. \tag{5.7}$$

As the notation implies, the difference operators B_0, B_1 are independent of n, corresponding to the assumption that $L(\cdot)$ does not depend explicitly on t; but, although based on fixed difference operators, they may depend on the point where they are applied. Thus at each point $j \in J_\Omega$, a linear difference operator B will be written in the form of a sum over near neighbours also in J_Ω:

$$(BU^n)_j = \sum_{k \in J_\Omega} b_{j,k} U_k^n \quad \forall j \in J_\Omega; \tag{5.8}$$

while for a nonlinear operator, nonlinear combinations of U_k^n would be involved. The notation $b_{j,k}$ denotes the fact that the coefficients may depend on j as well as k, for two reasons. Firstly, it enables us to cover the case when $L(\cdot)$ has spatially variable coefficients, while for a constant coefficient problem, $b_{j,k}$ would usually just depend on the difference $k - j$. Secondly, although the pattern of neighbours involved in (5.8) will be the same for all points well away from the boundary, at the boundary we are assuming that the numerical boundary conditions have already been incorporated in (5.8) so that values of U at all points outside J_Ω have been eliminated. **Thus the data term F^n in (5.7) includes not only data arising from inhomogeneous terms in the differential operator $L(u)$ but also inhomogeneous boundary data.**

We shall always assume that B_1 is linear, of the form (5.8), so that it can be represented by a square matrix. To extend the theory to nonlinear problems it would be necessary for B_0 to be nonlinear but not necessarily B_1; but to cover schemes like the box scheme by such an extension would require B_1 to be nonlinear too.

We shall furthermore assume that B_1 is invertible, i.e. its representing matrix is non-singular. Hence we can write (5.7) as

$$U^{n+1} = B_1^{-1}[B_0 U^n + F^n]. \tag{5.9}$$

We shall also assume that (5.7) is so scaled that formally it represents the differential equation in the limit and hence $B_1 = O(1/\Delta t)$. Thus

$$B_1 u^{n+1} - [B_0 u^n + F^n] \to \frac{\partial u}{\partial t} - L(u) \tag{5.10}$$

as the mesh intervals Δt, Δx, ... are refined in some manner which may depend on consistency conditions being satisfied. For example, in the θ-method (2.75) for the one-dimensional diffusion equation which we discussed in Chapter 2, away from the boundaries

$$B_1 = \frac{1}{\Delta t} - \theta\frac{\delta_x^2}{(\Delta x)^2}, \quad B_0 = \frac{1}{\Delta t} + (1-\theta)\frac{\delta_x^2}{(\Delta x)^2}. \tag{5.11}$$

Moreover, we shall combine these two conditions and assume that the matrix B_1 is uniformly well-conditioned in the sense that there is a constant K such that, in whichever norm is being used to carry out the analysis,

$$\|B_1^{-1}\| \leq K_1 \Delta t, \tag{5.12}$$

even though B_1^{-1} is represented by a matrix of ever-increasing dimension as the limit $\Delta t \to 0$ is approached.

For example, it is easy to deduce that in the case of (5.11) and the maximum norm, we have $K_1 = 1$: for the equation

$$B_1 U = F, \quad U = B_1^{-1} F \tag{5.13}$$

means, with $\mu = \Delta t/(\Delta x)^2$ and at a point away from the boundaries,

$$\begin{aligned} -\mu\theta U_{j-1} + (1+2\mu\theta)U_j - \mu\theta U_{j+1} &= \Delta t F_j, \\ \text{i.e.,} \quad (1+2\mu\theta)U_j &= \Delta t F_j + \mu\theta(U_{j-1} + U_{j+1}). \end{aligned} \tag{5.14}$$

With Dirichlet boundary conditions, $B_1 U = F$ at $j = 1$ and $J-1$ involves only two points of $J_\Omega \equiv \{j = 1, 2, \ldots, J-1\}$, giving

$$(1+2\mu\theta)U_1 - \mu\theta U_2 = \Delta t F_1 \quad \text{and}$$
$$-\mu\theta U_{J-2} + (1+2\mu\theta)U_{J-1} = \Delta t F_{J-1}$$

respectively. Thus for all values of $j \in J_\Omega$ we have

$$(1+2\mu\theta)|U_j| \leq \Delta t\|F\|_\infty + 2\mu\theta\|U\|_\infty$$

and hence

$$(1 + 2\mu\theta)\|U\|_\infty \leq \Delta t\|F\|_\infty + 2\mu\theta\|U\|_\infty \tag{5.15}$$

from which the result follows. Notice that the definition of $\|B_1^{-1}\|$ is defined as

$$\|B_1^{-1}\| = \max \frac{\|B_1^{-1}F\|}{\|F\|} = \max \frac{\|U\|}{\|F\|}$$

1.4 Consistency, order of accuracy and convergence

All limiting operations or asymptotic results refer (sometimes implicitly) to an underlying refinement path or set of refinement paths. That is, as in (2.47), a sequence of choices of the mesh parameters Δt, Δx, Δy, etc. is made such that they each tend to zero: and there may be inequality constraints between them. For brevity we shall characterise the whole of the spatial discretisation by a single parameter h: this may be just the largest of the mesh intervals Δx, $\Delta y, \ldots$, though this may need to be scaled by characteristic speeds in each of the co-ordinate directions; or h may be the diameter of the largest control volume around the mesh points. Then taking the limit along some designated refinement path we shall denote by '$\Delta t(h) \to 0$', or sometimes just $\Delta t \to 0$ or $h \to 0$: we shall always need Δt to tend to zero but stability or consistency may require that it does so at a rate determined by h, for example $\Delta t = O(h^2)$ being typical in parabolic problems and $\Delta t = O(h)$ in hyperbolic problems.

The **truncation error** is defined in terms of the exact solution u as

$$T^n := B_1 u^{n+1} - [B_0 u^n + F^n], \tag{1.1}$$

and **consistency** of the difference scheme (5.7) with the problem (5.1a)–(5.1c) as

$$T_j^n \to 0 \quad \text{as } \Delta t(h) \to 0 \quad \forall j \in J_\Omega \tag{1.2}$$

for all sufficiently smooth solutions u of (5.1a)–(5.1c). Note that this includes consistency of the boundary conditions through the elimination of the boundary values of U in the definition of B_0 and B_1.

If p and q are the largest integers for which

$$|T_j^n| \leq C[(\Delta t)^p + h^q] \quad \text{as } \Delta t(h) \to 0 \quad \forall j \in J_\Omega \tag{1.3}$$

for sufficiently smooth u, **the scheme is said to have order of accuracy p in Δt and q in h: or pth order of accuracy in Δt, and qth order of accuracy in h.**

Convergence on the other hand is defined in terms of all initial and other data for which (5.1a)–(5.1c) is well-posed, in a sense to be defined in the next section. Thus (5.7) is said to provide a **convergent approximation** to (5.1a)–(5.1c) in a norm $\|\cdot\|$ if

$$\|U^n - u^n\| \to 0 \quad \text{as } \Delta t(h) \to 0, n\Delta t \to t \in (0, t_F] \tag{1.4}$$

for every u^0 for which (5.1a)–(5.1c) is well-posed in the norm: here we mean either of the norms (5.4) or (5.5). From a practical viewpoint the advantage of this approach is that the effect of round-off errors can be immediately allowed for: if on the other hand convergence were established only for sufficiently smooth data, round-off errors would have to be accounted for in a separate analysis.

1.5 Stability and the Lax Equivalence Theorem

None of the definitions (5.16)–(5.19) in the last section was limited to linear problems: they are quite general. In this section (and most of the rest of the chapter) however we are able to consider only linear problems. Suppose two solutions V^n and W^n of (5.7) or (5.9) have the same inhomogeneous terms F^n but start from different initial data V^0 and W^0: we say the scheme is **stable** in the norm $\|\cdot\|$ and for a given refinement path if there exists a constant K such that

$$\|V^n - W^n\| \leq K\|V^0 - W^0\|, \quad n\Delta t \leq t_F; \tag{5.20}$$

the constant K has to be independent of V^0, W^0 and of $\Delta t(h)$ on the refinement path, so giving a uniform bound. (In the nonlinear case, restrictions would normally have to be placed on the initial data considered.) The assumption that V^n and W^n have the same data F^n is a simplification that mainly stems from our decision to limit the consideration of boundary effects in the evolutionary problems treated in this and earlier chapters – this is why we have been able to set boundary conditions to zero when applying Fourier analysis to study stability. We shall study boundary effects much more in Chapter 6 on elliptic problems. Note too that we can also include the effect of interior data by application of Duhamel's principle, as was done in Section 2.11 of Chapter 2.

Since we are dealing with the linear case (5.20) can be written

$$\|(B_1^{-1}B_0)^n\| \leq K, \quad n\Delta t \leq t_F. \tag{5.21}$$

Notice that for implicit schemes the establishment of (5.12) is an important part of establishing (5.21); consider, for example, the box scheme for linear advection, and the effect of having boundary conditions on one side or the other.

It is now appropriate to formalise our definition of well-posedness. We shall say that the problem (5.1) is well-posed in a given norm $\|\cdot\|$ if, for all sufficiently small h, we can show that (i) a solution exists for all data u^0 for which $\|u^0\|$ is bounded independently of h, and (ii) there exists a constant K' such that for any pair of solutions v and w,

$$\|v^n - w^n\| \leq K'\|v^0 - w^0\|, \quad t_n \leq t_F. \tag{5.22}$$

This differs from the usual definition in that we are using discrete norms; but we have chosen each of these so that it is equivalent to the corresponding function norm as $h \to 0$, if this exists for u, and we define u_j^n appropriately. An important feature of either definition is the following: for u to be a classical solution of (5.1a) it must be sufficiently smooth for the derivatives to exist; but suppose we have a sequence of data sets for which smooth solutions exist and these data sets converge to arbitrary initial data u^0 in the $\|\cdot\|$ norm, uniformly in h; then we can define a **generalised solution** with this data as the limit at any time t_n of the solutions with the smooth data, because of (5.22). Thus the existence of solutions in establishing well-posedness has only to be proved for a dense set of smooth data (with the definition of denseness again being uniform in h).

There is clearly a very close relationship between the definition of well-posedness for the differential problem and that of stability given by (5.20) for the discrete problem. This definition of stability, first formulated by Lax in 1953, enabled him to deduce the following key theorem:

Theorem 5.1 (Lax Equivalence Theorem) *For a consistent difference approximation to a well-posed linear evolutionary problem, which is uniformly solvable in the sense of (5.12), the stability of the scheme is necessary and sufficient for convergence.*

Proof (of sufficiency). Subtracting (5.16) from (5.7) we have

$$B_1\left(U^{n+1}-u^{n+1}\right)=B_0\left(U^n-u^n\right)-T^n,$$

i.e.,

$$U^{n+1}-u^{n+1}=\left(B_1^{-1}B_0\right)\left(U^n-u^n\right)-B_1^{-1}T^n. \tag{5.23}$$

Assuming that we set $U^0=u^0$, it follows that

$$U^n-u^n=-[B_1^{-1}T^{n-1}+\left(B_1^{-1}B_0\right)B_1^{-1}T^{n-2}+\cdots+\left(B_1^{-1}B_0\right)^{n-1}B_1^{-1}T^0]. \tag{5.24}$$

Now in applying the theorem, (5.12) and (5.21) are to hold in the same norm, for which we shall also deduce (5.19); we can combine these two to obtain

$$\|\left(B_1^{-1}B_0\right)^m B_1^{-1}\|\le KK_1\Delta t \tag{5.25}$$

from which (5.24) gives

$$\|U^n-u^n\|\le KK_1\Delta t\sum_{m=0}^{n-1}\|T^m\|.$$

Thus convergence in the sense of (5.19) follows from the consistency of (5.17), if u is sufficiently smooth for the latter to hold. For less smooth solutions, convergence follows from the hypotheses of well-posedness and stability: general initial data can be approximated arbitrarily closely by data for smooth solutions and the growth of the discrepancy is bounded by the well-posedness of the differential problem and the stability (5.19) of the discrete problem.

The necessity of stability for convergence follows from the principle of uniform boundedness in functional analysis, working in the framework of a single Banach space for the continuous and discrete problems; this is where our simplified approach based on discrete norms has its disadvantages, and therefore a consideration of this principle is beyond the scope of this book – interested readers may find a proof tailored to this application in Richtmyer and Morton (1967), pp. 34–36, 46. □

Thus for any scheme where consistency is readily established, we need only be concerned with establishing the conditions for stability; that is, we need only work with the discrete equations. As we have seen, consistency will usually hold for any sequence $\Delta t\to 0, h\to 0$; but there are a few cases where one has to be careful. For example, the Dufort-Frankel scheme for the one-dimensional heat equation,

$$\frac{U_j^{n+1}-U_j^{n-1}}{2\Delta t}=\frac{U_{j+1}^n-U_j^{n+1}-U_j^{n-1}+U_{j-1}^n}{(\Delta x)^2}, \tag{5.26}$$

has the advantage of being explicit and yet is unconditionally stable, something contrary to our experience so far. However, one finds for the truncation error

$$T=(u_t-u_{xx})+(\Delta t/\Delta x)^2u_{tt}+O((\Delta t)^2+(\Delta x)^2+((\Delta t)^2/\Delta x)^2). \tag{5.27}$$

Thus this is consistent with the heat equation only if $\Delta t=o(\Delta x)$ and is first order accurate only if $\Delta t=O((\Delta x)^2)$. As a result, it is the consistency condition rather than the stability condition that determines the refinement paths that can be used to obtain convergence.

This serves to emphasise the fact that, in the Lax Equivalence Theorem, there is implicit not only a choice of norm for defining stability and convergence but also a choice of refinement path.

1.6 Calculating stability conditions

As we are dealing with linear problems, if in (5.20) V^n and W^n are solutions of the difference equations (5.7), then the difference $V^n - W^n$ is a solution of the homogeneous difference equations with homogeneous boundary data. That is, establishing stability is equivalent to establishing the following:

$$B_1 U^{n+1} = B_0 U^n \text{ and } n\Delta t \leq t_F \Rightarrow \|U^n\| \leq K \|U^0\|, \tag{5.28}$$

which is what is meant by (5.21). The constant K will generally depend on the time interval t_F and allows for the sort of exponential growth that might occur with $u_t = u_x + u$, for example. **For simple problems one will often find: either $K = 1$, there is no growth and the scheme is stable; or $U^n \sim \lambda^n U^0$, with $|\lambda| > 1$ even as $\Delta t \to 0$ for some mode, so the scheme is unstable.**

Thus when we established a maximum principle in Section 2.6 and elsewhere we were also establishing stability in the maximum norm: strictly speaking, we had also to establish a minimum principle so as to be able to say not only

$$U_j^{n+1} \leq \max_k U_k^n \leq \|U^n\|_\infty \tag{1.5}$$

but also

$$U_j^{n+1} \geq \min_k U_k^n \geq -\|U^n\|_\infty \tag{1.6}$$

and could then deduce

$$\|U^{n+1}\|_\infty \leq \|U^n\|_\infty. \tag{1.7}$$

For parabolic problems proving stability by this means is very natural because a maximum (or minimum) principle is a very attractive and appropriate attribute for a difference scheme to possess over and above stability. Also, as we have seen in examples, quite general linear problems with variable coefficients and mixed boundary conditions can be dealt with: in each case we were able to deduce simple algebraic conditions on Δt for which a scheme could be shown to have a maximum principle and hence to be stable in the maximum norm. In many cases Fourier analysis could then show that, for the corresponding problem with constant coefficients and periodic boundary conditions, failing to satisfy these conditions leads to instability which, as we shall see below, is then in the ℓ_2 norm. However, in other cases as with the θ-method, there was a gap between the two conditions given by the two methods. In an influential paper in 1952,[1] Fritz John showed that, for a wider class of parabolic problems and corresponding difference approximations, schemes that satisfy a von Neumann condition obtained from a locally applied Fourier analysis are also stable in the maximum norm. **Thus, some schemes which do not satisfy a maximum principle are actually stable in the maximum norm; although see below in Section 5.7 for more comment on these cases.**

对于抛物型问题，通过这种方式证明稳定性是非常自然的，因为极值原理（最大值或最小值原理）是差分格式除稳定性之外应具备的一个非常吸引人且合适的特性。此外，正如我们在示例中所见，该方法能够处理相当一般的具有变系数和混合边界条件的线性问题：在每种情况下，我们都能推导出关于Δt 的简单代数条件，使得格式能被证明满足极值原理，从而在最大范数下稳定。在许多情况下，傅里叶分析随后可以表明，对于具有常系数和周期性边界条件的相应问题，若不满足这些条件会导致不稳定，而这种不稳定性，正如我们将在下文看到的，是ℓ_2 范数意义上的。然而，在其他情况下，例如θ 方法，两种方法给出的稳定性条件之间存在差距（基于极值原理的稳定性分析方法和基于傅里叶分析（冯·诺依曼条件）的稳定性分析方

[1] John, F. (1952) On the integration of parabolic equations by difference methods, *Comm. Pure Appl. Math.* **5**, 155.

法）。在1952 年的一篇有影响力的论文中，弗里茨·约翰证明了，对于更广泛类别的抛物型问题及相应的差分近似，满足通过局部应用傅里叶分析得到的冯·诺依曼条件的格式，在最大范数下也是稳定的。因此，一些不满足极值原理的格式实际上在最大范数下是稳定的；尽管关于这些情况的更多评论，请参见下文第5.7 节。

Furthermore, a maximum principle is seldom available or even appropriate for hyperbolic problems. **As we have noted, the first order scheme (4.20) satisfies a maximum principle whenever $0 \le \nu \le 1$ so that it is then stable in the maximum norm: but we can show that this can never be true of a second order scheme.** For example, consider the Lax-Wendroff method written in the form (4.36). If it were to satisfy a maximum principle, then for any set of non-positive values for U^n one should never have $U^{n+1} > 0$: yet if $0 < \nu < 1$, setting $U^n_{j-1} = U^n_j = 0$ and $U^n_{j+1} = -1$ gives a positive value for U^{n+1}_j. **This does not of course demonstrate that the scheme is actually unstable in the maximum norm, merely that we cannot prove such stability by this means.**

For this reason, and also because hyperbolic differential equations are much more commonly well-posed in the L_2 norm than in the supremum norm, for hyperbolic problems we have to adopt the more modest target of proving stability in the ℓ_2 norm (5.5). This gives weaker results because we have, recalling that V_j is the measure of the jth control volume,

$$\left[\min_{j\in J_\Omega} V_j\right]^{1/2} \|U\|_\infty \le \|U\|_2 \le \left[\sum_{j\in J_\Omega} V_j\right]^{1/2} \|U\|_\infty; \tag{5.32}$$

in the bounded region we are working with, the coefficient on the right is a finite constant while that on the left tends to zero as the mesh is refined. It is clear that we would prefer to derive a maximum norm error bound from a stability analysis but, if we have only ℓ_2 stability and so obtain a bound for the ℓ_2 norm of the error $\|E^n\|_2$, then (5.32) gives a poor result for $\|E^n\|_\infty$.

However, it is the ℓ_2 norm which is appropriate for Fourier analysis because of Parseval's relation. Suppose we can assume periodicity on a normalised region $[-\pi, \pi]^d$ which is covered by a uniform (Cartesian) mesh of size $\Delta x_1 = \Delta x_2 = \ldots = \Delta x_d = \pi/J$. Then the Fourier modes that can be distinguished on the mesh correspond to wave numbers, which we denote by the vector $\mathbf{k}$, having components given by

$$k = 0, \pm 1, \pm 2, \ldots, \pm J, \tag{5.33}$$

where the last two with $k\Delta x = \pm\pi$ are actually indistinguishable. Hence we can expand any periodic function on the mesh as

$$U(\mathbf{x}_j) = \frac{1}{(2\pi)^{d/2}} \sideset{}{'}\sum_{(\mathbf{k})} \hat{U}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}_j} \tag{5.34}$$

where the prime on the summation sign means that any term with $k_s = \pm J$ has its weight halved, and we have also used a vector notation $\mathbf{x}_j$ for mesh points. This discrete Fourier expansion has an inverse which is the discrete Fourier transform

$$\hat{U}(\mathbf{k}) = \frac{1}{(2\pi)^{d/2}} \sum_{(\mathbf{j})} (\Delta x)^d U(\mathbf{x}_j) e^{-i\mathbf{k}\cdot\mathbf{x}_j}, \tag{5.35}$$

where each component of $\mathbf{j}$ runs from $-J$ to J with the mesh points on the periodic boundaries again having their weights halved so that all the weights are equal to the V_j introduced in (5.5).

Lemma 5.1 The Fourier modes $(2\pi)^{-d/2}e^{i\mathbf{k}\cdot\mathbf{x}_j}$ with components given by (5.33) are orthonormal with respect to the ℓ_2 inner product used in (5.35), namely

$$\langle U, W\rangle_2 := (\Delta x)^d \sum_{(\mathbf{j})}{}' U_j \overline{W}_j. \tag{5.36}$$

Hence we have, with V_j the control volume measure,

$$\begin{aligned} \|U\|_2^2 = \sum_{j\in J_\Omega} V_j |U_j|^2 &\equiv \sum_{(\mathbf{j})}{}' (\Delta x)^d |U(\mathbf{x}_j)|^2 \\ &= \left(\frac{\Delta x}{2\pi}\right)^d \sum_{(\mathbf{k})} \left|\hat{U}(\mathbf{k})\right|^2 \left(\frac{2\pi}{\Delta x}\right)^d, \end{aligned} \tag{5.40}$$

i.e.,

$$\|\hat{U}\|_2^2 := \sum_{(\mathbf{k})} \left|\hat{U}(\mathbf{k})\right|^2 = \|U\|_2^2, \tag{5.41}$$

which is the appropriate form of Parseval's relation.

For a rectangular region of general dimensions a simple scaling will reduce the situation to the above case. However, note that not only is Δx then changed but we will also generally have $\Delta k \neq 1$ and that such a coefficient will be needed in the definition of $\|\hat{U}\|_2$ for (5.41) to hold. It is also worth noting that when for example we have a problem on $[0, 1]$ with $u(0) = u(1) = 0$ we extend this to a periodic problem on $[-1, 1]$ by imposing antisymmetry at $x = 0$ and using a sine series. This is why we have taken $[-\pi, \pi]$ as our standard case above.

To establish (5.28) then, for a constant coefficient problem with periodic boundary conditions, we expand arbitrary initial data in the form (5.34) and, from the discrete Fourier transform of (5.28), obtain the same form at successive time levels with the coefficients given by

$$\hat{B}_1(\mathbf{k})\hat{U}^{n+1}(\mathbf{k}) = \hat{B}_0(\mathbf{k})\hat{U}^n(\mathbf{k}), \tag{5.42}$$

where, if the U^n are p-dimensional vectors, $\hat{B}_0$ and $\hat{B}_1$ are $p \times p$ matrices. The matrix

$$G(\mathbf{k}) = \hat{B}_1^{-1}(\mathbf{k})\hat{B}_0(\mathbf{k}) \tag{5.43}$$

is **called the amplification matrix as it describes the amplification of each mode by the difference scheme.** Because we have assumed that $\hat{B}_0$ and $\hat{B}_1$ are independent of t we can write

$$\hat{U}^n = [G(\mathbf{k})]^n \hat{U}^0 \tag{1.8}$$

and using (5.41) have

$$\begin{aligned} \sup_{U^0} \frac{\|U^n\|_2}{\|U^0\|_2} &= \sup_{\hat{U}^0} \frac{\left|\sum_{(\mathbf{k})}' |\hat{U}^n(\mathbf{k})|^2\right|^{1/2}}{\left|\sum_{(\mathbf{k})}' |\hat{U}^0(\mathbf{k})|^2\right|^{1/2}} \\ &= \sup_{\mathbf{k}} \sup_{\hat{U}^0(\mathbf{k})} \frac{|\hat{U}^n(\mathbf{k})|}{|\hat{U}^0(\mathbf{k})|} = \sup_{\mathbf{k}} \|G(\mathbf{k})^n\|. \end{aligned} \tag{5.44}$$

Thus stability in the ℓ_2 norm is equivalent to showing that

$$\|[G(\mathbf{k})]^n\| \leq K \quad \forall \mathbf{k}, \quad n\Delta t \leq t_F. \tag{5.46}$$

Here $\|G^n\|$ means the $p \times p$ matrix norm subordinate to the vector norm used for U_j^n and $\hat{U}(\mathbf{k})$.

Remark 1.6.1 *We start from the very beginning,*

$$\begin{aligned} B_1U^{n+1} &= B_0U^n, \\ B_1E^{-1}EU^{n+1} &= B_0E^{-1}EU^{,} \\ \hat{B}_1\hat{U}^{n+1} &= \hat{B}_0\hat{U}^n, \\ G &= \hat{B}_1^{-1}\hat{B}_0 = EB_1^{-1}B_0E^{-1} = \Lambda, \end{aligned}$$

where in fact ***we want to find the diagonalization of*** $B_1^{-1}B_0$ ***using discrete Fourier transform. Here,*** E ***is the unitary Fourier matrix.***

Then clearly we have the following result.

Theorem 5.2 (von Neumann Condition) A necessary condition for stability is that there exist a constant K' such that

$$|\lambda(\mathbf{k})| \leq 1 + K'\Delta t \quad \forall \mathbf{k}, \quad n\Delta t \leq t_F, \tag{5.47}$$

for every eigenvalue $\lambda(\mathbf{k})$ of the amplification matrix $G(\mathbf{k})$.

Proof. By taking any eigenvector of $G(\mathbf{k})$ as $\hat{U}(\mathbf{k})$ it is obviously necessary that there be a constant K such that $|\lambda^n| \leq K$: then by taking $n\Delta t = t_F$ we have

$$|\lambda| \leq K^{\Delta t/t_F} \leq 1 + (K-1)\Delta t/t_F \text{ for } \Delta t \leq t_F, \tag{1.9}$$

the last inequality following from the fact that K^s is a convex function of s. □

1.7 Practical (strict or strong) stability

Clearly the von Neumann condition is very important both practically and theoretically. Even for variable coefficient problems it can be applied locally (with local values of the coefficients) and because instability is a local phenomenon, due to the high frequency modes being the most unstable, it gives necessary stability conditions which can often be shown to be sufficient. However, for some problems the presence of the arbitrary constant in (5.47) is too generous for practical purposes, though being adequate for eventual convergence.

Consider the following problem which is a mixture of our simple one-dimensional diffusion and advection problems:

$$u_t + au_x = \epsilon u_{xx}, \quad \epsilon > 0. \tag{5.48}$$

Let us approximate it by central differences in space and a forward difference in time:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{\Delta_{0x}U_j^n}{\Delta x} = \epsilon\frac{\delta_x^2 U_j^n}{(\Delta x)^2}. \tag{5.49}$$

We now have two mesh ratios

$$\nu := a\Delta t/\Delta x, \quad \mu := \epsilon\Delta t/(\Delta x)^2 \tag{5.50}$$

and a Fourier analysis gives the amplification factor

$$\lambda(k) = 1 - i\nu\sin k\Delta x - 4\mu\sin^2\frac{1}{2}k\Delta x, \tag{5.51a}$$

$$|\lambda|^2 = (1-4\mu s^2)^2 + 4\nu^2 s^2(1-s^2), \tag{5.51b}$$

where as usual $s = \sin \frac{1}{2} k \Delta x$. Putting $s^2 = 1$ shows that $\mu \leq \frac{1}{2}$ is necessary for stability; then

$$\nu^2 = (a\Delta t/\Delta x)^2 = (a^2/\epsilon)\mu\Delta t \tag{1.10}$$

and this implies

$$|\lambda|^2 \leq 1 + \frac{1}{2}(a^2/\epsilon)\Delta t \tag{5.52}$$

so that the von Neumann condition is satisfied. As this is a scalar pure initial-value problem, and a and ϵ are constants this is sufficient for stability. However, if $\nu = 1$, $\mu = \frac{1}{4}$ and $s^2 = \frac{1}{2}$ we have $|\lambda|^2 = \frac{5}{4}$ giving very rapid growth; by contrast, the differential problem damps all Fourier modes. （Δt不够小，由于ϵ太小了）

In practice, then, for finite values of Δx and Δt, **the von Neumann condition is too weak** when the exponential growth that it allows is inappropriate to the problem. We therefore introduce the following stricter definition:

Definition 5.1 A scheme is said to be practically (or strictly or strongly) stable if, when Fourier mode solutions of the differential problem satisfy

$$|\hat{u}(k, t + \Delta t)| \leq e^{\alpha \Delta t} |\hat{u}(k, t)| \quad \forall k \tag{5.53}$$

for some $\alpha \geq 0$, then the corresponding amplification factors for the difference scheme satisfy

$$|\lambda(k)| \leq e^{\alpha \Delta t} \tag{5.54}$$

for all k that correspond to discrete modes.

For the above example, we have $\alpha = 0$ and so require $|\lambda| \leq 1$. From (5.51b) we have

$$|\lambda|^2 = 1 - 4(2\mu - \nu^2)s^2 + 4(4\mu^2 - \nu^2)s^4. \tag{5.55}$$

By considering this expression as a positive quadratic function of s^2 on the interval $[0, 1]$, we obtain the conditions

$$|\lambda|^2 \leq 1 \quad \forall k \quad \text{iff} \quad \nu^2 \leq 2\mu \leq 1. \tag{5.56}$$

These are very well known and often important restrictions because they can be very severe if ϵ is small: apart from the expected condition $\mu \leq \frac{1}{2}$, we can write the first inequality as

$$\frac{a^2 \Delta t}{\epsilon} \equiv \frac{\nu^2}{\mu} \leq 2. \tag{1.11}$$

So it can be interpreted as placing a limit of 2 on a mesh Péclet number, in whose definition a is a velocity, $a\Delta t$ is interpreted as a mesh-length and ϵ is a diffusion or viscosity coefficient – see also the discussion in Section 2.15 where such restrictions arise from considering the application of a maximum principle.

Indeed, such is the practical importance of this criterion and resulting mesh restriction, the definition embodied in (5.53) and (5.54) is often used in the engineering field as the main definition of stability, the term being used without qualification. Note also that in Section 5.9 we refer to the property as *strong stability* when applied to problems which have no solution growth, corresponding to the concept of absolute stability in the discretisation of ODEs.

Using the same analysis as that given above, similar practical stability criteria can be found by combining any of the explicit schemes we have analysed for the diffusion equation with ones we have used for linear advection. Generally speaking the resultant condition will be a more severe combination of the corresponding results from the two parts. For example, if the upwind method (4.13) is used in (5.49) instead of the central difference we get

$$|\lambda|^2 \leq 1 \quad \forall k \quad \text{iff} \quad \nu^2 \leq \nu + 2\mu \leq 1 \tag{5.58}$$

compared with the conditions $0 \leq \nu \leq 1$ and $0 \leq \mu \leq \frac{1}{2}$, which are needed for the separate equations.