1

Finite difference schemes for linear time-dependent problems

1.1 Basic concepts, definitions and notation

Consider a general initial value problem for linear partial differential equations:

$$\begin{aligned} u_t(x,t) &= \mathcal{P}\left(x,t,\frac{\partial}{\partial x}\right)u(x,t), \\ u(x,0) &= f(x), \end{aligned} \tag{1.1}$$

where x is a vector of s components: $x = (x_1, \cdots, x_s)$, u is a vector of p components: $u(x,t) = (u_1(x,t), \cdots, u_p(x,t))$ and $\mathcal{P}$ is a polynomial of $\frac{\partial}{\partial x}$. If the highest order time derivative in a linear partial differential equation is $\frac{\partial^m}{\partial t^m}u$, then we can always rewrite it in the form of (1.1) as a system for a unknown vector $[u, \frac{\partial}{\partial t}u, \cdots, \frac{\partial^{m-1}}{\partial t^{m-1}}u]^T$. For instance, the two way wave equation $u_{tt} = u_{xx}$ can be written as

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ u_t \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \frac{\partial^2}{\partial x^2} & 0 \end{pmatrix}\begin{pmatrix} u \\ u_t \end{pmatrix}. \tag{1.2}$$

Remark 1.1. *Let* $(v_1, v_2)^T$ *denote the unknown functions in* (1.2) *and take the Fourier transform, then we get*

$$\frac{\partial}{\partial t}\begin{pmatrix} \hat{v}_1(t) \\ \hat{v}_2(t) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}\begin{pmatrix} \hat{v}_1(t) \\ \hat{v}_2(t) \end{pmatrix}.$$

Notice that $\begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}$ *is diagonalizable with eigenvalues* $\pm \mathbb{i}\,\omega$*, thus* (1.2) *is the convection to two directions even though* $\frac{\partial^2}{\partial x^2}$ *is the only spatial differential operator in* (1.2)*.*

For computational convenience we will restrict the domain of the solution $u(x,t)$ of (1.1) to a bounded region, even though it might be defined for all x and all $t > 0$. On this bounded region we construct a grid of points, discretizing both space (in each of the space coordinates) and time. For this purpose, we specify the step sizes Δt and Δx_i for $i = 1, \cdots, s$ and define the grid points as points of the form $(x_1, \cdots, x_s, t_n)$, where:

$$t_n = n\Delta t$$

$$x_{j_i} = j_i \Delta x_i, i = 0, \cdots, N_i.$$

Although in many practical applications it is preferable to define suitable varying step sizes, we have chosen here constant ones for simplicity of notation. However, the concepts and properties discussed in this chapter can be readily generalized to the variable step-size case.

The main idea of any finite difference scheme attempting to approximate the values of $u(x,t)$ by computer methods is to construct a vector of p components for given integers n and $j_1, \cdots, j_s$ with $0 \leq j_i \leq N_i$, which we call $U^n_{j_1,\cdots,j_s}$ and which "approximates" the value of $u(x_{j_1}, \cdots, x_{j_s}, n\Delta t)$.

For fixed n, $U^n_{j_1,\cdots,j_s}$ is therefore a vector-valued function of the set of integers $\{j_i = 0, \cdots, N_i; 0 \leq i \leq s\}$. For such functions, we define the k-th shift operator E_k to be the operator that shifts the index j_k to its right $(j_k + 1)$, that is:

$$E_k U^n_{j_1,\cdots,j_k,\cdots,j_s} = U^n_{j_1,\cdots,j_k+1,\cdots,j_s}, \quad 1 \leq k \leq s.$$

Definition 1.1. *A finite difference scheme is a recursion formula of the form:*

$$B_0(E_1, \cdots, E_s)V^{n+1}_\alpha = B_1(E_1, \cdots, E_s))V^n_\alpha \tag{1.3}$$

where $\alpha = j_1, \cdots, j_s$ is a multi-index, and $B_0(E_1, \cdots, E_s)$ and $B_1(E_1, \cdots, E_s)$ are functions of the operators E_i, $1 \leq i \leq s$. If B_0 is the identity operator, we say the scheme is explicit. Otherwise it is called an implicit scheme.

We now give some examples to illustrate our notation.

Example 1.1. *Consider the two-dimensional problem:*

$$u_t = u_x + u_y,$$

where $u(x,y,t)$ is a real valued function. Let Δt, Δx and Δy be positive, fixed quantities. One possible finite difference scheme is given by:

$$\begin{aligned} U^{n+1}_{i,j} = &\frac{1}{4}\left(U^n_{i+1,j+1} + U^n_{i-1,j+1} + U^n_{i+1,j-1} + U^n_{i-1,j-1}\right) \\ &+ \frac{\Delta t}{2\Delta x}(U^n_{i+1,j} - U^n_{i-1,j}) + \frac{\Delta t}{2\Delta y}(U^n_{i,j+1} - U^n_{i,j-1}) \end{aligned}$$

which in terms of the shift operators E_1 and E_2, can be written in the form:

$$U_{i,j}^{n+1} = \left(\frac{1}{4}(E_1 + E_1^{-1})(E_2 + E_2^{-1}) + \frac{\Delta t}{2\Delta x}(E_1 - E_1^{-1}) + \frac{\Delta t}{2\Delta y}(E_2 - E_2^{-1})\right) U_{i,j}^n.$$

Thus $V^n = U^n$ in this case.

Example 1.2. *Conside the Leapfrog scheme for the one-way wave equation:*

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} = \frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x}$$

In order to write down this scheme in the form (1.3)*, we define the two dimensional vector:*

$$V_j^n = \begin{pmatrix} U_j^n \\ U_j^{n-1} \end{pmatrix},$$

and express $V_j^{n+1} = B_1(E)V_j^n$, where now $B_1(E)$ is a 2×2 matrix depending on the shift operator E. In fact, since:

$$\begin{pmatrix} U_j^{n+1} \\ U_j^n \end{pmatrix} = \begin{pmatrix} \frac{\Delta t}{\Delta x}(U_{j+1}^n - U_{j-1}^n) + U_j^{n-1} \\ U_j^n \end{pmatrix},$$

then

$$V_j^{n+1} = \begin{pmatrix} \frac{\Delta t}{\Delta x}(E - E^{-1}) & 1 \\ 1 & 0 \end{pmatrix} V_j^n$$

Example 1.3. *For the same equation $u_t = u_x$, consider the scheme:*

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(U_{j+1}^{n+1} - U_j^{n+1}).$$

This can be written in terms of the shift operator E in the following way:

$$(1 + \frac{\Delta t}{\Delta x} - \frac{\Delta t}{\Delta x}E)U_j^{n+1} = U_j^n.$$

For implicit schemes like the example above, to find U_j^{n+1}, we need to solve a globally coupled linear sytem. We shall assume that the operator B_0^{-1} exists and it is bounded, so the finite difference scheme (1.3) can always be written in matrix form as:

$$V^{n+1} = C(\Delta t, \Delta x, \bar{x}, t)V^n$$

where we are taking all the components $\{V_\alpha^n : \alpha = (j_1, ..., j_x); j_i = 0, ..., N_i\}$ ordered to form a vector (e.g., the $vec(X)$ operation in Chapter **??**); $\Delta x = (\Delta x_1, \cdots, \Delta x_s)$, and $\bar{x} = \{x_{j_i} : j_i = 0, ..., N_i, i = 1, ..., s\}$. We will assume that $\Delta x_i = h_i(\Delta t)$ for some functions h_i of the parameter Δt, for all space coordinates $i = 0, ..., s$. If the operator $\mathcal{P}$ in (1.1) does not depend on time,

then it is reasonable to limit our study to the case where C does not depend on time either. We will refer to this situation as the autonomous equation. If C does not depend on $\bar{x}$ either, we call the scheme a constant coefficient scheme, which we shall study in detail later. For the remainder of this chapter we shall simply write $C(\Delta t)$, keeping in mind that it may depend on t and $\bar{x}$ as well. We will therefore analyze the finite difference scheme (1.3) in its equivalent form:

$$V^{n+1} = C(\Delta t)V^n \tag{1.4}$$

where now $C(\Delta t)$ is an $N \times N$ matrix, with:

$$N = \prod_{i=1}^{s}(N_i + 1).$$

Recall that, N_i depends on Δx_i which is a function of Δt, thus the dimension of the matrix $C(\Delta t)$ depends on Δt.

Definition 1.2. *Let $\Delta x = (\Delta x_1, \cdots, \Delta x_{N_s})$, then for any fixed real number $t \geq 0$, we define the operator $Q_{\Delta x}$ by:*

$$Q_{\Delta x}u(x,t) = \{u(x_{j_1}, \cdots, x_{j_s}, t), \quad j = 0, \cdots, N_i, i = 0, \cdots, s.\}$$

So given a function $u(x,t)$, $Q_{\Delta x}u(x,t)$ is a vector with $N = \prod_{i=1}^{s}(N_i + 1)$ components, each of them representing a vector (recall that $u(x,t) = (u_1(x,t), ..., u_p(x,t))$ is a vector of p components).

At any fixed time t, $Q_{\Delta x}$ is an operator which "looks" at the values that $u(x,t)$ attains at the space grid points. In some cases it is more appropriate to specify projection operators which assign some values between the grid points. The space where we "project" the solution $u(x,t)$ via $Q_{\Delta x}$ is the same space where we are to construct the numerical solution, in accordance with (1.4). We are interested in studying the behavior of the collection of vectors in (1.4) for "small" values of Δt. We shall assume that:

$$\lim_{\Delta t \to 0} \Delta x_i = \lim_{\Delta t \to 0} h_i(\Delta t) = 0.$$

We now want to give a precise meaning to the statement *as Δt becomes smaller, the numeical solution gets closer to the analytical solution at any given time $t = n\Delta t$ held fixed.* Specifically, we want to compare the limit of V_i^n as $\Delta t \to 0$ and $n \to \infty$ such that $t = n\Delta t$ is constant, with the corresponding limit of $Q_{\Delta x}u$. This involves the concept of norms on the euclidean space $\mathbb{R}^N$ when the dimensjon N grows as $\Delta t \to 0$.

Definition 1.3. *For any vector $V = (V_1, ..., V_N)$, we define the norm $|V|_N$ by:*

$$|V|_N^2 = \frac{1}{N}\sum_{j=1}^{N}|V_j|^2$$

where, if each component V, is itself a vector, $|V_j|$ denotes the usual vector norm.

By our notations, (1.4) can denote a "one-step" method if $V^n = U^n$ where U^n approximates u at t_n, or a "k-step" method if

$$V^n = \begin{pmatrix} U^n \\ U^{n-1} \\ \vdots \\ U^{n-k} \end{pmatrix}. \tag{1.5}$$

1.2 Properties of Finite Difference Schemes

Throughout this section, we shall consider $u(x,t)$ to be the solution of a well posed initial value problem. That is, calling $S(t,t_0)$ the solution operator, the function u is specified by:

$$u(x,t) = S(t,t_0)u(x,t_0),$$

thus in particular:

$$u(x,(n+1)\Delta t) = S((n+1)\Delta t, n\Delta t)u(x,n\Delta t).$$

If the problem is *autonomous*,that is, the operator $\mathcal{P}$ in (1.1) is independent of time, then S is a function of the elapsed $(t-t_0)$ and we can simply write $S(t-t_0)$ and $S(\Delta t)$ in the above expressions.

Definition 1.4. *We say that the scheme $U^{n+1} = C(\Delta t)U^n$ is accurate of degree (or order) q_1 in space and q_2 in time, or more shortly, accurate (or consistent) of order (q_1,q_2) if for any fixed $t = n\Delta t$ and a very smooth solution $u(x,t)$:*

$$|[C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(t+\Delta t,t)]\,u(x,t)|_N \le K(t)\Delta t(|\Delta x|^{q_1} + \Delta t^{q_2}) \tag{1.6}$$

where

$$|\Delta x| = \sqrt{\sum_{i=1}^{s}(\Delta x_i)^2}.$$

If the system is autonomous, we can write (1.6) *in the form*

$$|[C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)]\,u(x,t)|_N \le K(t)\Delta t(|\Delta x|^{q_1} + \Delta t^{q_2}).$$

For a k-step method (1.4) *with* (1.5)*, accuracy of order (q_1,q_2) means*

$$\left| [C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)] \begin{pmatrix} u(x,t) \\ u(x,t-\Delta t) \\ \vdots \\ u(x,t-k\Delta t) \end{pmatrix} \right|_N \le K(t)\Delta t(|\Delta x|^{q_1} + \Delta t^{q_2}). \tag{1.7}$$

In most of the cases, it is desirable to have the same degree of accuracy in space and time, $q_1 = q_2 = q$, and when this happens, if no confusion arises, we will say that the scheme is accurate of degree q, or q-th order accurate. In some situations, however, we will work with accurate schemes for which $q_1 \neq q_2$.

This definition of accuracy (1.7) is simply an abstract description of the following local truncation error.

Definition 1.5. *Rewrite the scheme $V^{n+1} = C(\Delta t)V^n$ for solving 1.1 in the form approaching $u_t(x,t) - \mathcal{P}\left(x, t, \frac{\partial}{\partial x}\right) u(x,t) = 0$ as $\Delta t \to 0$ and $\Delta x \to 0$. The local truncation error is the residue of replacing numerical solutions by a smooth exact solution in the scheme of this form. The scheme is accurate of degree (or order) q_1 in space and q_2 in time if the local truncation error is equal to $\mathcal{O}(|\Delta x|^{q_1}) + \mathcal{O}(\Delta t^{q_2})$.*

We give now some examples of different schemes for the problem $u_t = u_x$.
Scheme 1:

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

This scheme is useless since it will never be stable as we have seen in Example **??**. Nonetheless let us consider its accuracy. We denote by u_j^n, the true solution at the grid points: $u_j^n = u(j\Delta x, n\Delta t)$. Then

$$[C(\Delta t)Q_{\Delta x}u(x, n\Delta t)]_j = u_j^n + \frac{\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n)$$

is the j-th component of the vector $C(\Delta t)Q_{\Delta x}u(x, n\Delta t)$. By definition,

$$S(\Delta t)u(x, n\Delta t) = u(x, (n+1)\Delta t),$$

and therefore:

$$[Q_{\Delta x}S(\Delta t)u(x, n\Delta t)]_j = u_j^{n+1}.$$

By the Taylor's expansion around (x_j, t_n), and the fact $u_t = u_x$, we get

$$\begin{aligned}
&|C(\Delta t)Q_{\Delta x}u(x, n\Delta t) - Q_{\Delta x}S(\Delta t)u(x, n\Delta t)|_j \\
=&|u_j^{n+1} - u_j^n - \frac{\Delta t}{2\Delta x}(u_{j+1}^n - u_{j-1}^n)| \\
=&|\Delta t(u_t)_j^n + \frac{1}{2}\Delta t^2(u_{tt})_j^n - \frac{\Delta t}{2\Delta x}(2\Delta x(u_x)_j^n + 2\frac{1}{6}\Delta x^3(u_{xxx})_j^n)| \\
=&|\frac{1}{2}\Delta t^2(u_{tt})_j^n - \frac{1}{6}\Delta t\Delta x^2(u_{xxx})_j^n| \\
\leq&\frac{1}{2}\max\left\{\max_x |u_{tt}(x, n\Delta t)|, \frac{1}{3}\max_x |u_{xxx}(x, n\Delta t)|\right\}\Delta t(\Delta t + \Delta x^2).
\end{aligned}$$

Assume there is a very smooth solution $u(x,t)$ s.t.

$$\max\left\{\max_x |u_{tt}(x, n\Delta t)|, \frac{1}{3}\max_x |u_{xxx}(x, n\Delta t)|\right\} \leq K,$$

then the scheme is accurate of order $(2,1)$. The scheme can be rewritten in the form approaching $u_t - u_x = 0$:

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} - \frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0$$

As an alternative way to check accuracy, we can compute the local truncation error as:

$$\tau_j^n = \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = u_t(x_j, t_n) + \mathcal{O}(\Delta t) - u_x(x_j, t_n) + \mathcal{O}(\Delta x^2)$$

and using now the equation $u_t = u_x$, satisfied by $u(x,t)$, we conclude that this scheme is accurate of second order in space and first order in time.
Scheme 2: Lax-Friedrich's Scheme:

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

This scheme is a first order accurate scheme, that is, $q_1 = q_2 = 1$.
Scheme 3: Upwind Scheme: Consider the one-sided difference for the spatial derivative:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(U_{j+1}^n - U_j^n).$$

This is a first order accurate scheme.
Scheme 4: Downwind Scheme: Consider the one-sided difference for the spatial derivative:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(U_j^n - U_{j-1}^n).$$

This is a first order accurate scheme. However this scheme is also useless. The domain of dependence (the exact solution is a wave travelling to the left thus $u(x_j, t_n + \Delta t) = u(x_j - \Delta t, t_n)$ thus U_j^{n+1} depends on values of U^n to the right of x_j) is not included in the scheme stencil (U_j^{n+1} is based on U_j^n and U_{j-1}^n) therefore such a scheme is unstable.
Scheme 5: Leapfrog Scheme: If we use the centered difference for both time and spatial derivatives, we get

$$U_j^{n+1} = U_j^{n-1} + \frac{\Delta t}{\Delta x}(U_{j+1}^n - U_{j-1}^n). \tag{1.8}$$

To find its accuracy, rewrite it as

$$\frac{U_j^{n+1} - U_j^{n-1}}{2\Delta t} - \frac{U_{j+1}^n - U_{j-1}^n}{2\Delta x} = 0,$$

and compute

$$\tau_j^n = \frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} - \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = u_t + \mathcal{O}(\Delta t^2) - u_x + \mathcal{O}(\Delta x^2).$$

So this is a second order accurate scheme. It should be noticed that in order to implement this scheme it is not enough to specify initial conditions U^0 since it involves two time steps. To obtain U^1 for initiate the computation, there are many different ways. For instance, we can use a one-step method to approximate U^1.

Scheme 5: Lax-Wendroff Scheme: This scheme was developed around 1960 - 1964 and it is very frequently used. It is based on the Taylor series expansion for $u(x,t)$ given by:

$$u(x,t+\Delta t) = u(x,t) + \Delta t u_t(x,t) + \frac{1}{2}\Delta t^2 u_{tt}(x,t) + \mathcal{O}(\Delta t^3),$$

which, using $u_t = u_x$, reduces to:

$$u(x,t+\Delta t) = u(x,t) + \Delta t u_x(x,t) + \frac{1}{2}\Delta t^2 u_{xx}(x,t) + \mathcal{O}(\Delta t^3).$$

Using the centered difference, we obtain a scheme with second order accuracy in both time and space:

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

Scheme 5: Cranck-Nicholson Scheme: This is a second order accurate implicit scheme

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

The mere fact that a scheme is accurate does not imply that it provides useful results. Therefore we would like to compare the behavior of the numerical solution with the true solution, and not only the discrepancies resulting from one step of the time iterations. This comparison is the underlying concept of *convergence*.

Definition 1.6. *For a scheme $V^{n+1} = C(\Delta t)V^n$ in which U^n approximates u at $n\Delta t$, we say that the scheme converges if for arbitrary fixed $t > 0$ we have, for all n, Δt such that in $n\Delta t = t$:*

$$\lim_{\Delta t \to 0, \Delta x \to 0} |U^n - Q_{\Delta x} u(x, n\Delta t)|_N = 0.$$

Notice that the number N of points in the space grid becomes larger as $\Delta t \to 0$. If the initial condition of the original problem is $u(x,0) = f(x)$, then we can write:

$$u(x,t) = S(t)f(x) = S(t-t_1)S(t_1)f(x),$$

for any intermediate time $0 \leq t_1 \leq t$. In general we have:

$$u(x, n\Delta t) = S(\Delta t)^n f(x),$$

and analogously, a scheme in the form of $U^{n+1} = C(\Delta t)U^n$ can also be written as:

$$U^{n+1} = C(\Delta t)^n U^0$$

where $U^0 = Q_{\Delta x} f(x)$. In this notation, the convergence condition reads as follows:

$$\lim_{\Delta t \to 0, \Delta x \to 0} |(C(\Delta t)^n Q_{\Delta x} - Q_{\Delta x} S(\Delta t)^n) f(x)|_N = 0.$$

It should be clear now that convergence involves the difference between the values predicted by the numerical solution itself and those of the true solution "projected" at the grid points. On the other hand, to establish the accuracy of the scheme, we only need to check how the operator $C(\Delta t)$ changes the value of the true solution during only one time step, as compared to the true solution Δt units of time later. Convergence is the most important property of a numerical method. However, it cannot be established directly, since the true solution $u(x,t)$ is not known. We therefore look for ways to determine convergence indirectly, using only the partial differential equation and properties of the scheme that do not involve explicit knowledge of the function u that we want to approximate.

Definition 1.7. *We say that scheme $V^{n+1} = C(\Delta t)V^n$ is stable if for any fixed $t > 0$, there exist constants K and a such that:*

$$\|C(\Delta t)^n\| \leq K e^{\alpha n \Delta t},$$

for all n and Δt such that $n\Delta t = t$.. Here $\|C\|$ is the spectral norm for the matrix C.

Notice, first of all, that stability is indeed the discrete analog of well posedness. Recall from Chapter **??** that well posedness of the problem, in terms of the solution operator is equivalent to

$$|S(t,t_0)| \leq K e^{a(t-t_0)}, \quad \forall t \geq t_0,$$

where $|S|$ denotes the "operator norm". Therefore, for an autonomous system, where $S(n\Delta t) = S(\Delta t)^n$, we have:

$$|S(\Delta t)^n| \leq K e^{an\Delta t},$$

which is almost identical to the stability condition, except that here S is an operator acting on functions (generally belonging to some Hilbert space), and consequently the above norm refers to the corresponding operator norm, whereas in $|C(\Delta t)^n|$ is the matrix norm.

Theorem 1.1. Lax Equivalence Theorem. *Let $u(x,t)$ be a classical solution of the well posed linear problem* (1.1) *and let the finite difference scheme $V^{n+1} = C(\Delta t)V^n$ be accurate of order (q_1, q_2), i.e., the scheme satisfies* (1.6)*. If the scheme is stable, then for any T, there exists a bounded function $G(t)$ such that for all $t \in [0,T]$ and $n\Delta t = t$, the following holds*

$$|U^n - Q_{\Delta x}u(x, n\Delta t)|_N \leq G(t)(|\Delta x|^{q_1} + \Delta t^{q_2}).$$

Remark 1.2. *The theorem states not only the convergence but also the rate of convergence for a smooth solution of a wellposed initial value problem of any linear PDEs.*

Proof. For simplicity, we consider the one-step method $U^{n+1} = C(\Delta t)U^n$ and the extension to the k-step case is straightforward. Let

$$\delta^n = [C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)]u(x,t).$$

The actual error that we want to control to prove the convergence is

$$\begin{aligned}\varepsilon^{n+1} &= U^{n+1} - Q_{\Delta x}u(x,(n+1)\Delta t)\\ &= C(\Delta t)[U^n - Q_{\Delta x}u(x,n\Delta t)] + [C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)]u(x,t)\\ &= C(\Delta t)\varepsilon^n + \delta^n\end{aligned}$$

By solving $\varepsilon^{n+1} = C(\Delta t)\varepsilon^n + \delta^n$ and $\varepsilon^0 = 0$ (because we have $U^0 = Q_{\Delta x}u(x,0)$), we get

$$\varepsilon^n = \sum_{k=0}^{n-1} C(\Delta t)^{n-k-1}\delta^k,$$

thus

$$|\varepsilon^n|_N \leq \sum_{k=0}^{n-1} \|C(\Delta t)^{n-k-1}\||\delta^k|_N.$$

By stability, $\|C(\Delta t)^{n-k-1}\| \leq Ke^{a(n-k-1)\Delta t} \leq K_1$ for some constant K_1 and all k s.t. $0 \leq k \leq n-1$, and using accuracy on $|\delta^n|_N$:

$$|[C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)]\,u(x,t)|_N \leq K(t)\Delta t(|\Delta x|^{q_1} + \Delta t^{q_2}),$$

we get

$$|\varepsilon^n|_N \leq K_1 n\Delta t K(n\Delta t)(|\Delta x|^{q_1} + \Delta t^{q_2}),$$

which is the desired result, upon letting $G(t) = K_1 t K(t)$. □

Remark 1.3. *The Lax Equivalence Theorem applies to any linear scheme (for solving a linear PDE) in the form of $V^{n+1} = C(\Delta t)V^n$. For instance, in a finite element method in the form $U^{n+1} = C(\Delta t)U^n$ solving (1.1), U^n denotes the finite element basis coefficients (in contrast to point values in a finite difference method) and $Q_{\Delta x}u(x,t)$ denotes the projection of the exact solution onto the finite element space, then the same proof is still valid.*

In the Lax Equivalence Theorem, the assumption that $u(x,t)$ is a classical solution amounts to assume that the initial data $f(x)$ is a function with r continuous derivatives - where r is the degree of the polynomial $\mathcal{P}$ and with compact support, which we denote by $f(x) \in C_0^r$. Indeed, the assumption $f(x) \in C_0^r$ together with well posedness is equivalent to stating that $u(x,t)$ is a classical solution. However, the theorem can be generalized for the case where the initial function is not in C_0^r, provided that we can approximate this function in the L^2-sense by functions in C_0^r. To see this, assume that there exists a sequence $f_l(x) \in C_0^r, l = 1, 2, \cdots$ satisfying $\lim_{l\to+\infty} \|f - f_l\|^2 = 0$ where the norm is the L^2 norm (for example, $f(x)$ can be a step function multiplying a Gaussian, then $f(x)$ is not even continuous but can be approximated by a sequence of functions in C_0^r). Let $S(t - t_0)$ be the solution operator for the problem and define the sequence $u_l(x,t)$ as the corresponding solution with initial value $f_l(x)$, that is:

$$u_l(x,t) = S(t)f_l(x).$$

Since for any given $t \geq 0$, $S(t)$ is a bounded operator on L^2, it follows by convergence of f_l to f that the sequence of functions $u_l(x,t)$ for fixed t, converges in L^2 to some limit function $u(x,t)$ (which, however, may lack smoothness). Using the Lax Equivalence Theorem for each integer l we have:

$$|U_l^n - Q_{\Delta x}u_l(x, n\Delta t)|_N \leq G_l(t)(|\Delta x|^{q_1} + \Delta t^{q_2}), \tag{1.9}$$

where the U_l^n are defined for each l using scheme $U^{n+1} = C(\Delta t)U^n$ with initial value $U^0 = Q_{\Delta x}f_l(x)$. Therefore:

$$|U_l^n - U_m^n|_N \leq \|C(\Delta t)^n\||Q_{\Delta x}(f_l - f_m)|_N.$$

As a consequence of the L^2 convergence of f_l, it follows that for sufficiently large N, $|Q_{\Delta x}(f_l - f_m)|_N \to 0$ as $l, m \to \infty$, implying that the sequences $\{U_l^n : l \geq 1\}$ are Cauchy sequences for each n. This ensures the existence of the limiting vectors:

$$U^n = \lim_{l\to\infty} U_l^n \quad \text{for each} \quad n \geq 1.$$

Now to prove convergence we express the difference:

$$\begin{aligned} |Q_{\Delta x}u(x, n\Delta t) - U^n|_N &= |Q_{\Delta x}[u(x, n\Delta t) - u_l(x, n\Delta t)] + Q_{\Delta x}u_l(x, n\Delta t) - U_l^n + (U_l^n - U^n)|_N \\ &\leq |Q_{\Delta x}(u - u_l)|_N + |Q_{\Delta x}u_l - U_l^n|_N + |U_l^n - U^n|_N \end{aligned}$$

The first and third terms of this last inequality tend to zero as l increases, due to the definitions of u and U^n. The middle term satisfies (1.9), so all these facts together yield the convergence result for more general initial conditions. Notice that we loose information on the rate of convergence, since we do not know how the functions $G_l(t)$ in (1.9) behave with increasing l. Even if we know the rates for each l, the above inequality involves two limit processes.

Theorem 1.2. Kreiss Perturbation Theorem. *Suppose that the scheme:*

$$V^{n+1} = C(\Delta t)V^n$$

is stable. Then the perturbed scheme:

$$V^{n+1} = [C(\Delta t) + \Delta t D(\Delta t)]V^n$$

is stable, provided that $|D(\Delta t)| \leq H$, for some constant $H \geq 0$.

Before we give the proof of Kreiss perturbation theorein, we shall illustrate its usefulness.

Example 1.4. *Consider the partial differential equation:*

$$u_t = u_x - \beta u$$

and the scheme:

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

The Kreiss Perturbation Theorem states that it is enough to check stability for the leapfrog scheme (1.8)*, since $D(t) = -2\beta I$.*

Proof. We will just prove the "one-step" version, i.e., the case when $V^n = U^n$. The multi-step case is similar. Define the vectors W^n by the transformation:

$$W^n = e^{-n\Delta t\beta}U^n$$

where $\beta > 0$ is a constant to be determined later. The perturbed scheme becomes

$$W^{n+1} = e^{-(n+1)\Delta t\beta}[C(\Delta t)+\Delta t D(\Delta t)]e^{n\Delta\beta}W^n = e^{-\Delta t\beta}[C(\Delta t)+\Delta t D(\Delta t)]W^n,$$

so we get:

$$W^{n+1} = e^{-\Delta t\beta}C(\Delta t)W^n + \Delta t\bar{D}(\Delta t)W^n$$

where $\bar{D}(\Delta t) = e^{-\Delta t\beta}D(\Delta t)$. Let $\delta_n = \Delta t\bar{D}(\Delta t)W^n$, then the analog of the Duhamel principle for the finite difference equation:

$$W^{n+1} = e^{-\Delta t\beta}C(\Delta t)W^n + \delta_n$$

is given by

$$W^n = [e^{-\Delta t\beta}C(\Delta t)]^n W^0 + \sum_{k=0}^{n-1}[e^{-\Delta t\beta}C(\Delta t)]^{n-k-1}\delta_k$$

thus

$$W^n = [e^{-\Delta t\beta}C(\Delta t)]^n W^0 + \sum_{k=0}^{n-1}[e^{-\Delta t\beta}C(\Delta t)]^{n-k-1}\Delta t\bar{D}(\Delta t)W^k$$

By stability of $C(\Delta t)$ and boundedness of $D(\Delta t)$, there is a constant C_1 such that:

$$|C(\Delta t)^{n-k-1}||D(\Delta t)| \leq C_1$$

for all integers k with $0 \leq k \leq n-1$. Thus:

$$|W^n|_N \leq |C(\Delta t)|^n e^{-n\Delta t\beta}|W^0|_N + \left(C_1\Delta t\sum_{k=0}^{n-1} e^{-\Delta t\beta(n-k)}\right)\max_{0\leq k\leq n-1}|W^k|_N.$$

Let $z = e^{-\Delta t\beta}$, then $0 \leq z \leq 1$ and:

$$\sum_{k=0}^{n-1} e^{-\Delta t\beta(n-k)} = \sum_{k=0}^{n-1} e^{-\Delta t\beta k} = \sum_{k=0}^{n-1} z^k = \frac{1-z^n}{1-z} = \frac{1-e^{-n\Delta t\beta}}{1-e^{-\Delta t\beta}}$$

and since $1 - e^{-\Delta t\beta} \approx \beta\Delta t + \mathcal{O}(\Delta t^2)$, we may now pick β large enough so that the quantity:

$$C_1\Delta t\sum_{k=0}^{n-1} e^{-\Delta t\beta(n-k)} = C_1\frac{1-e^{-n\Delta t\beta}}{\beta + \mathcal{O}(\Delta t)}$$

is bounded by some constant $\gamma < 1/2$ for all integers n and all $\Delta t > 0$. Since $|C(\Delta t)|^n \leq C_2 e^{an\Delta t}$ for some constants C_2 and a, and using $U^0 = W^0$, it follows that:

$$|W^n|_N \leq C_2 e^{(a-\beta)n\Delta t}|U^0|_N + \gamma\max_{0\leq k\leq n-1}|W^k|_N.$$

We may assume that $a - \beta < 0$, for if this is not the case, we just increase the value of β and we will still have $\gamma < 1/2$. Then $e^{(a-\beta)\Delta t} \leq 1$ for all n, Δt, and:

$$|W^n|_N \leq C_2|U^0|_N + \gamma\max_{0\leq k\leq n-1} W|W^k|_N,$$

where C_2 does not depend on n. Now for any arbitrary large integer M, we take the maximum on both sides over $0 \leq n \leq M$:

$$\max_{0\leq n\leq M}|W^n|_N \leq C_2|U^0|_N + \max_{0\leq n\leq M}\gamma\max_{0\leq k\leq n-1}|W^k|_N,$$

thus

$$\max_{0\le n\le M}|W^n|_N \le C_2|U^0|_N + \gamma \max_{0\le n\le M}|W^k|_N.$$

Therefore,

$$(1-\gamma)|W^n|_N \le (1-\gamma)\max_{0\le n\le M}|W^n|_N \le C_2|U^0|_N,$$

for all $0 \le n \le M$. We finally get

$$|U^n|_N \le \frac{C_2}{1-\gamma}e^{\beta n\Delta t}|U^0|_N.$$

Stability follows from the fact that $U^n = C(\Delta t)^n U^0$. □

As can be deducted from the proof, it is extremely important that the perturbation be of the order Δt, more specifically, that it has the form $\Delta t D(\Delta t)$. In many practical situations one has to be careful in applying the result of this perturbation theorem, always checking first if the assumptions are indeed satisfied. The following is an example in which the perturbation has apparently the form $\Delta t D(\Delta t)$, yet it gives rise to an unstable scheme.

Example 1.5. *For the equation:*

$$u_t = u_{xx} + u_x$$

the term u_x is a lower order term. One possible scheme is the following:

$$U^{n+1}_j = U^n_j + \frac{\Delta t}{\Delta x^2}(U^n_{j+1} - 2U^n_j + U^n_{j-1}) + \frac{\Delta t}{2\Delta x}(U^n_{j+1} - U^n_{j-1}).$$

The last term corresponds to the "perturbation" and we want to know whether we can neglect this term by applying Kreiss Perturbation Theorem, in order to check stability. Although it appears that the perturbation is of order Δt, this may not be the case, for if we choose $\Delta t/\Delta x^2$ constant to achieve stability of the unperturbed scheme, then the perturbation is really of order $\sqrt{\Delta t}$ and the theorem is not applicable in this case.

1.3 Basic definitions and notations for stability

Next we will consider the stability for constant coefficient schemes. As already mentioned, stability of a finite difference scheme is the discrete analog of the concept of well posedness of a partial differential equation. We present the basic results and tools that allow us establish the stability of a constant coefficient finite difference scheme. Examples are inserted throughout the rest of this chapter in order to introduce and illustrate the concepts involved in the problem.

Example 1.6. *Consider*

$$u_t(x,t) = u_x(x,t), \quad x \in [0, 2\pi],$$

and we assume 2π-periodicity of the solution. We construct a grid of points with constant spacing Δx in space and Δt in time, such that:

$$\frac{\Delta t}{\Delta x} = \lambda \leq 1; \quad x_j = j\Delta x, j = 0, \cdots, N-1, \Delta x = \frac{2\pi}{N}.$$

Let us focus on the upwind scheme:

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

with the appropriate boundary conditions given by the periodicity requirement:

$$U_{-1}^n = U_{N-1}^n, U_N^n = U_0^n,$$

which holds for all $n > 0$. This upwind scheme can be written in the matrix form as:

$$U^{n+1} = C(\Delta t)U^n$$

where each U^n is a vector of N components:

$$\begin{pmatrix} U_0^{n+1} \\ U_1^{n+1} \\ \vdots \\ U_{N-2}^{n+1} \\ U_{N-1}^{n+1} \end{pmatrix} = \begin{pmatrix} 1-\lambda & \lambda & \cdots & 0 & 0 \\ 0 & 1-\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \vdots & 1-\lambda & \lambda \\ \lambda & 0 & \vdots & 0 & 1-\lambda \end{pmatrix} \begin{pmatrix} U_0^n \\ U_1^n \\ \vdots \\ U_{N-2}^n \\ U_{N-1}^n \end{pmatrix}.$$

The dimension of the matrix $C(\Delta t)$ depends on Δt itself, through the dependence of N on Δt. In general, this makes it hard to check directly the stability of the scheme, that is, to find constants K and α such that:

$$\|C(\Delta t)^n\| \leq Ke^{\alpha t},$$

for all n and Δt such that $t = n\Delta t$ is held fixed.

Before we analyze the stability, let us recall several matrix notations:

- A^T is the transpose of A. A^* is the complex conjugate transpose.
- Eigenvalues: $\text{eig}_i(S)$ denotes the eigenvalue of S with i-th largest magnitude.
- Jordan Normal Form: any matrix S can be decomposed as $S = P\Lambda P^{-1}$ where Λ is upper-triangular and the diagonal entries are eigenvalues of S.

- Singular Value: the i-th largest one is denoted as $\sigma_i(S) = \sqrt{\mathrm{eig}_i(SS^*)} = \sqrt{\mathrm{eig}_i(S^*S)}$.
- Normal Matrix: a matrix A is called normal if $AA^* = A^*A$.
- The following are equivalent:
  1. A is normal.
  2. A is diagonalizable by a unitary matrix, i.e., $A = P\Lambda P^*$, Λ is a diagonal matrix and $PP^* = I$.
  3. $\sigma_i(A) = |\,\mathrm{eig}_i(A)|$.

For the particular case in Example 1.6, we can find $\|C(\Delta t)^n\|$ since $C(\Delta t)$ is circulant thus can be diagonalized by the DFT matrix, which is a unitary matrix. By multiplyting $C(\Delta t)$ to the vector obtained by sampling $e^{\mathbb{i}\,nx}$ at $x = (0, \Delta x, 2\Delta x, \cdots, (N-1)\Delta x)^T$, we can get the eigenvalues as $\lambda_k = 1 - \lambda + \lambda e^{\mathbb{i}\,k\Delta x}$. Let T be the DFT matrix then $C(\Delta t) = T\Lambda T^*$ where the diagonal matrix Λ has diagonal entries λ_k $(k = 0, 1, \cdots, N-1)$. Since $C(\Delta t)^n = T\Lambda^n T^*$ (so $C(\Delta t)^n$ is a normal matrix), thus $C(\Delta t)^n[C(\Delta t)^n]^* = T[\Lambda\Lambda^*]^n T^*$, we get the singular values of $C(\Delta t)^n$ as

$$|\lambda_k|^n = [(1-\lambda)^2 + \lambda^2 + 2\cos(k\Delta x)\lambda(1-\lambda)]^{\frac{n}{2}}.$$

Next we use an easier alternative method instead of looking directly at the matrix $C(\Delta t)$. In Example 1.6, we can consider the discrete Fourier transform (**??**) and the inverse discrete Fourier transform (**??**) for U_j^n. Assume N is even, we use a normalized (also index shifted) version of (**??**) and (**??**):

$$\hat{U}_k^n = \sum_{j=0}^{N-1} e^{-\mathbb{i}\,kj\Delta x} U_j^n, \quad k = 0, \cdots, N-1.$$

$$U_j^n = \frac{1}{N}\sum_{k=0}^{N-1} e^{\mathbb{i}\,kj\Delta x}\hat{U}_k^n, \quad j = 0, \cdots, N-1,$$

We also have the Parseval's identity for the discrete Fourier transform above:

$$\sum_{j=0}^{N-1} |U_j^n|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |\hat{U}_k^n|^2.$$

Replace U^n by its inverse discrete Fourier transform in the upwind scheme, we get

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{\mathbb{i}\,kj\Delta x}\hat{U}_k^{n+1} = \frac{1}{N}\sum_{k=0}^{N-1} e^{\mathbb{i}\,kj\Delta x}\hat{U}_k^n + \lambda\Big(\frac{1}{N}\sum_{k=0}^{N-1} e^{\mathbb{i}\,k(j+1)\Delta x}\hat{U}_k^n - \frac{1}{N}\sum_{k=0}^{N-1} e^{\mathbb{i}\,kj\Delta x}\hat{U}_k^n\Big),$$

thus

$$\frac{1}{N}\sum_{k=0}^{N-1} e^{\mathbb{i}\,kj\Delta x}\left[\hat{U}_k^{n+1}-\hat{U}_k^n-\lambda(e^{\mathbb{i}\,k\Delta x}\hat{U}_k^n-\hat{U}_k^n)\right]=0.$$

which means that the inverse discrete Fourier transform of $\hat{U}_k^{n+1}-\hat{U}_k^n-\lambda(e^{\mathbb{i}\,k\Delta x}\hat{U}_k^n-\hat{U}_k^n)$ is equal to zero. Therefore, we get

$$\hat{U}_k^{n+1}-\hat{U}_k^n-\lambda(e^{\mathbb{i}\,k\Delta x}\hat{U}_k^n-\hat{U}_k^n)=0,$$

which can be written as

$$\hat{U}_k^{n+1}=g(k)\hat{U}_k^n,\quad g(k)=1+\lambda(e^{\mathbb{i}\,k\Delta x}-1). \tag{1.10}$$

Then we have

$$\begin{aligned}\sum_{j=0}^{N-1}|U_j^{n+1}|^2&=\frac{1}{N}\sum_{k=0}^{N-1}|\hat{U}_k^{n+1}|^2\\&=\frac{1}{N}\sum_{j=0}^{N-1}|g(k)|^2|\hat{U}_k^n|^2\\&\le\max_k|g(k)|^2\sum_{j=0}^{N-1}|U_j^n|^2.\end{aligned}$$

Thus if $\max_k|g(k)|^2$ is bounded for all possible values of $k\Delta x$, we have a bound for $\|C(\Delta t)\|$, which yields stability. The main idea is therefore to study the functions $g(k)$ instead of working with the matirx $C(\Delta t)$, even though these two methods are essentially equivalent. Recall that we have used the DFT matrix to diagonalize the circulant matrix and the DFT matrix represents precisely the discrete Fourier transform. Notice that $g(k)$ are exactly the eigenvalues of $C(\Delta t)$. Notheless, the second method is easier, because we can obtain (1.10) simply by asserting an ansatz $U_j^n=\hat{U}_k^n e^{\mathbb{i}\,kx_j}$ into the scheme.

Example 1.7. *We consider now a generalization of the previous example. Let A be a constant $p\times p$ matrix and and $u(x,t)=(u_1(x,t),\cdots,u_p(x,t))^T$ satisfying:*

$$u_t=Au_x,\quad x\in[0,2\pi],$$

and we also assume 2π-periodicity of $u(x,t)$. Consider a naive extension of the upwind scheme:

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

with boundary conditions

$$U_{-1}^n=U_{N-1}^n, U_N^n=U_0^n.$$

Now let us use the ansatz $U_j^n = \hat{U}_k^n e^{\mathbb{i}\, kj\Delta x}$, *which is equivalent to apply the discrete Fourier transform to the scheme or use the DFT matrix to "diagonalize"* $C(\Delta t)$. *We get*

$$\hat{U}_k^{n+1} e^{\mathbb{i}\, kj\Delta x} = \hat{U}_k^n e^{\mathbb{i}\, kj\Delta x} + \lambda A(\hat{U}_k^n e^{\mathbb{i}\, k(j+1)\Delta x} - \hat{U}_k^n e^{\mathbb{i}\, kj\Delta x}),$$

thus

$$\hat{U}_k^{n+1} = \mathcal{G}(k)\hat{U}_k^n, \quad \mathcal{G}(k) = I + \lambda A(e^{\mathbb{i}\, k\Delta x} - 1).$$

where $\mathcal{G}(k)$ *is a* $p \times p$ *matrix. Notice that* $C(\Delta t)$ *is a* $Np \times Np$ *matrix with* $N \to 0$ *while the size of* $\mathcal{G}(k)$ *is fixed.*

We now generalize the concepts introduced in the examples given above. As we recall from Chapter **??**, the general form of a partial differential equation with constant coefficients is given by

$$\begin{aligned} u_t &= \mathcal{P}(\partial\,/\,\partial\, x)u, \\ u(x,0) &= f(x), \end{aligned} \tag{1.11}$$

where $u(x,t) = (u_1(x,t), \cdots, u_n(x,t))^T$ is a function of $x = (x_1, \cdots, x_s)$ and time t. In an analogous way, we can define a finite difference scheme with constant coefficients in general form:

Definition 1.8. *Let* Δt *and* Δx_i *be any given step sizes, for* $i = 1, \cdots, s$, *and denote by* $X = \{x_{j_i} : j_i = 0, \cdots, N_i; i = 1, \cdots, s\}$ *the collection of all grid points in the space coordinates. A scheme of the form:*

$$V^{n+1} = C(\Delta t, X, t)V^n$$

is called a constant coefficient scheme if the matrix $C(\Delta t, X, t)$ *does not depend on* X *and* t, *so we can write:*

$$V^{n+1} = C(\Delta t)V^n.$$

Consider the constant coefficient scheme:

$$U^{n+1} = C(\Delta t, X, t)U^n.$$

For each multiindex $j = (j_1, \cdots, j_i)$ with $j_i = 0, \cdots, N_i$ and $i = 1, \cdots, s$, U_j^n is a vector of p components approximating the value of the true solution at the grid points $u(j_1\Delta x_1, \cdots, j_s\Delta x_s, n\Delta t)$. The discrete Fourier transform is now given by

$$\hat{U}_k^n = \sum_{j\in\mathcal{J}} e^{-\,\mathbb{i}\langle k, x_j\rangle} U_j^n,$$

$$U_j^n = \frac{1}{N} \sum_{k\in\mathcal{K}} e^{\mathbb{i}\langle k, x_j\rangle} \hat{U}_k^n.$$

where $N = \Pi_{i=1}^{s} N_i$, $k = (k_1, \cdots, k_s)$ is a multi-index, the sets $\mathcal{J}$ and $\mathcal{K}$ are the set of multiindex j and k so that $0 \leq k_i \leq N_i - 1$ and $0 \leq j_i \leq N_i - 1$. Using the discrete Fourier transform of the scheme yields the difference equations in the Fourier space:

$$\hat{U}_k^{n+1} = \mathcal{G}(\Delta t, k)\hat{U}_k^n.$$

We call the matrix $\mathcal{G}(\Delta t, k)$ the **amplification matrix**. If the problem is scalar (p = 1), then we write $g(\Delta t, k)$ or sometimes just $g(k)$, and usually call it the *amplification factor*.

1.4 von Neumann stability

Stability of the scheme $V^{n+1} = C(\Delta t)V^n$ can be written in terms of the amplification matrix as the following condition: given $t > 0$, there exist constants K and α such that for all multi-index k and all n such that $n\Delta t = t$,

$$\|\mathcal{G}(\Delta t, k)^n\| \leq Ke^{\alpha t}.$$

The condition must be satisfied for all multi-index k in order to establish stability of the scheme. This condition involves an infinite number of matrices being uniformly bounded, yet in practice it turns out to be remarkably easier to deal with the amplification matrices treating k as a parameter, than it is to study stability working directly with $C(\Delta t)$, whose dimension depends on the chosen discretization of space and time. Our first result presents a necessary although not sufficient condition for stability.

Theorem 1.3. The von Neumann Condition *The amplification matrix of a stable scheme satisfies the condition:*

$$\rho[\mathcal{G}(\Delta t, k)] \leq e^{\gamma \Delta t} = 1 + \mathcal{O}(\Delta t),$$

where $\rho[\mathcal{G}(\Delta t, k)]$ denotes the spectral radius (largest magnitude of eigenvalues) of the matrix $\mathcal{G}(\Delta t, k)$.

The von Neumann stability condition is necessary but not sufficient for stability. In most practical applications, turns out to be easily checked whether this condition holds or not, as we shall examplify later on. When determining stability of a scheme, our first step shall always be verifying whether this condition holds or not.

Proof. If the scheme is stable, then

$$\|\mathcal{G}^n\| \leq Ke^{\alpha t},$$

where $t = n\Delta t$. We need a fact for the spectral radius

$$\rho(A)^n \leq \|A^n\|.$$

To see why this is true, let v and λ be eigenvectors and eigenvalues of A, then

$$|\lambda|^k\|v\| = \|\lambda^k v\| = \|A^k v\| \leq \|A^k\| \cdot \|v\| \Rightarrow |\lambda|^k \leq \|A^k\|.$$

So we have

$$\rho[\mathcal{G}(\Delta t, k)] \leq \|\mathcal{G}^n\|^{\frac{1}{n}} \leq K^{\frac{1}{n}} e^{\alpha \Delta t}.$$

Since $t = n\Delta t$ is held fixed at a constant value, $K^{\frac{1}{n}} = K^{\frac{\Delta t}{t}}$. Let $\beta = \log K$, then

$$\rho(\mathcal{G}) \leq e^{\beta \Delta t/t} e^{\alpha \Delta t} = e^{(\beta/t+\alpha)\Delta t} = e^{\gamma \Delta t}$$

where $\gamma = \beta/t + \alpha$ is a positive constant for all n and Δt such that $t = n\Delta t$ is constant, yielding the von Neumann condition. □

Remark 1.4. *The von Neumann condition is also sufficient for stability in the following two cases:*

- *If $\mathcal{G}$ is a normal matrix (the scalar case $\mathcal{G} = g$ is a special case), then so is $\mathcal{G}^n$ thus $\|\mathcal{G}^n\| = \rho[\mathcal{G}^n]$.*
- *If $\mathcal{G}$ is diagonalizable $\mathcal{G}(\Delta t, k) = T\Lambda T^{-1}$ with $\|T\|\|T^{-1}\| \leq K$ for all Δt and k, then $\mathcal{G}^n = T\Lambda^n T^{-1}$ thus $\|\mathcal{G}\| \leq \|T\|\|\Lambda^n\|\|T^{-1}\| = \|T\|\rho[\mathcal{G}^n]\|T^{-1}\|$.*

1.5 The leapfrog scheme

1.5.1 The one way wave equation

In this section we first study in detail the leap frog scheme for the one dimensional scalar equation $u_t = u_x$ to understand the stability we have defined for finite difference schemes. The scheme is:

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

and we impose the periodic boundary conditions through the usual periodicity requirement:

$$U_{-1}^n = U_{N-1}^n, \quad U_0^n = U_N^n$$

Define the vector

$$V_j^n = \begin{pmatrix} U_j^n \\ U_j^{n-1} \end{pmatrix},$$

then we can rewrite the scheme into the form $V^{n+1} = C(\Delta t)V^n$. Let $\lambda = \frac{\Delta t}{\Delta x}$, then the scheme becomes

$$V_j^{n+1} = \begin{pmatrix} \lambda(E - E^{-1}) & 1 \\ 1 & 0 \end{pmatrix} V_j^n,$$

where E and E^{-1} are the shift operations, as introduced in Section 1.1. Therefore, although the original problem is a scalar one, here V is a 2-component vector: we have considered a "fictitious" component to be able to represent the scheme by $V^{n+1} = C(\Delta t)V^n$. Plugging in the ansatz $V_j^n = \hat{V}_k^n e^{\mathbb{i}\, kj\Delta x}$ (which is equivalent to plugging in the discrete Fourier transform of V^n), we get

$$\hat{V}_k^{n+1} e^{\mathbb{i}\, kj\Delta x} = \begin{pmatrix} \lambda(E - E^{-1}) & 1 \\ 1 & 0 \end{pmatrix} \hat{V}_k^n e^{\mathbb{i}\, kj\Delta x}.$$

Notice that the shift operators act only on the functions of $x_j = j\Delta x$, we have

$$E\hat{e}^{\mathbb{i}\, kj\Delta x} V_k^n = e^{\mathbb{i}\, k\Delta x} e^{\mathbb{i}\, kj\Delta x} \hat{V}_k^n,$$
$$E^{-1}\hat{e}^{\mathbb{i}\, kj\Delta x} V_k^n = e^{-\mathbb{i}\, k\Delta x} e^{\mathbb{i}\, kj\Delta x} \hat{V}_k^n,$$

thus

$$\hat{V}_k^{n+1} = e^{-\mathbb{i}\, kj\Delta x} \begin{pmatrix} \lambda(e^{\mathbb{i}\, k\Delta x} - e^{-\mathbb{i}\, k\Delta x}) & 1 \\ 1 & 0 \end{pmatrix} e^{\mathbb{i}\, kj\Delta x} \hat{V}_k^n = \begin{pmatrix} 2\,\mathbb{i}\,\lambda \sin(k\Delta x) & 1 \\ 1 & 0 \end{pmatrix} \hat{V}_k^n.$$

Therefore we have the explicit expression for the amplification matrix:

$$\mathcal{G}(\Delta x, k) = \begin{pmatrix} 2\,\mathbb{i}\,\lambda \sin(k\Delta x) & 1 \\ 1 & 0 \end{pmatrix}$$

The variable $k\Delta x$ appears in the expression of the amplilication matrix as the argument of a trigonometric function. This is in general true, and in order to analyze the amplification matrix in terms of its arguments, it is enough to consider the variable $\xi = k\Delta x$ restricted to $0 \leq \xi < 2\pi$. Throughout the rest of this text, we shall often write $\xi = k\Delta x$ without further mentioning that we actually consider ξ to be restricted to the interval $[0, 2\pi)$.

The eigenvalues of the amplification matrix $\mathcal{G}(\Delta x, k)$ can be calculated as:

$$\mu_1(\xi) = \mathbb{i}\,\lambda \sin\xi + \sqrt{1 - \lambda^2 \sin^2\xi},$$
$$\mu_1(\xi) = \mathbb{i}\,\lambda \sin\xi - \sqrt{1 - \lambda^2 \sin^2\xi},$$

We will check now the von Neumann condition as well as the conditions for stability of the leap frog scheme under study.

Case I: If $\lambda^2 > 1$, then for those values of k such that $\xi = k\Delta x = \frac{pi}{2}$ we have:

$$\mu_1(\pi/2) = \mathbb{i}(\lambda + \sqrt{\lambda^2 - 1}),$$

so $|\mu_1(\pi/2)| > 1$, yielding that the von Neumann stability condition is not satisfied by the amplification matrix. We conclude that the leap frog scheme is unstable when $\lambda > 1$.

Case II: If $\lambda^2 \leq 1$, then

$$|\mu_i(\xi)|^2 = \lambda^2 \sin^2 \xi + 1 - \lambda^2 \sin^2 \xi = 1,$$

for $i = 1, 2$ which holds for any value of ξ. Therefore $\rho[\mathcal{G}] = 1$ and the von Neumann condition is satisfied. Nonetheless, this does not imply that the scheme is stable for $\lambda \leq 1$. Indeed we will show that the scheme is actually unstable for $\lambda = 1$.

To see this, recall that stability requires that the family of matrices $\mathcal{G}(\Delta, k)$ be uniformly bounded by $Ke^{\alpha t}$ for all values of k. Consider now $\lambda = \Delta t/\Delta x = 1$, then for all n with $n\Delta t$ fixed, stability would certainly imply the uniform bound in $\|\mathcal{G}^n(\Delta, k)\|$ as $n \to \infty$ for all possible values of k. Notice that $\lambda = 1$ is also fixed. In order to prove our claim that this case is unstable, it suffices to show that for one particular value of $\xi = k\Delta x$, $\|\mathcal{G}^n(\Delta, k)\|$ is not bounded as $n \to \infty$. Let $\xi = \pi/2$ and k_0 denote the modes for which $k_0\Delta x = \frac{\pi}{2}$(modulo 2π), then

$$\mathcal{G}(\Delta t, k_0) = \begin{pmatrix} 2\,\mathbb{i} & 1 \\ 1 & 0 \end{pmatrix}.$$

Notice that $\mathcal{G}(\Delta t, k_0)$ has one repeated eigenvalue $\mu_1 = \mu_2 = \mathbb{i}$ and it is not diagonalizable (because the eigenspace is one-dimensional). Let v_1 the be one eigenvector and v_2 be one generalized eigenvector. Let $T = [v_1, v_2]$, then the Jordan form of this matrix can be written as

$$\mathcal{G}(\Delta t, k_0) = T \begin{pmatrix} \mathbb{i} & 1 \\ 0 & \mathbb{i} \end{pmatrix} T^{-1}.$$

Therefore,

$$\mathcal{G}^n(\Delta t, k_0) = T \begin{pmatrix} \mathbb{i} & 1 \\ 0 & \mathbb{i} \end{pmatrix}^n T^{-1} = T \begin{pmatrix} \mathbb{i}^n & n\,\mathbb{i}^{n-1} \\ 0 & \mathbb{i}^n \end{pmatrix} T^{-1}.$$

Obviously $\left\| \begin{pmatrix} \mathbb{i}^n & n\,\mathbb{i}^{n-1} \\ 0 & \mathbb{i}^n \end{pmatrix} \right\| \to \infty$ as $n \to \infty$, thus $\|\mathcal{G}^n(\Delta t, k_0)\| \to \infty$ as $n \to \infty$. Therefore, the leap frog scheme is unstable for $\lambda = 1$, although the von Neumann condition is satisfied.

Lemma 1.1. *The leap frog scheme for $u_t = u_x$ is stable for $\lambda < 1$.*

Proof. Let $\lambda < 1$. Then the two eigenvalues $\mu_1(\xi)$ and $\mu_2(\xi)$ are distinct thus $\mathcal{G}$ is diagonalizable. Let T be the eigenvector matrix then we have

$$\mathcal{G} = T \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} T^{-1},$$

thus

$$\mathcal{G}^n = T \begin{pmatrix} \mu_1^n & 0 \\ 0 & \mu_2^n \end{pmatrix} T^{-1},$$

and

$$\|\mathcal{G}^n\| \leq= \|T\| \left\| \begin{pmatrix} \mu_1^n & 0 \\ 0 & \mu_2^n \end{pmatrix} \right\| \|T^{-1}\|$$

The spectral norm of $\begin{pmatrix} \mu_1^n & 0 \\ 0 & \mu_2^n \end{pmatrix}$ is equal to $\max_i |\mu_i|^n = 1$ (recall that $\|\mu_i\| = 1$) because we have (the singular values of A are square roots of eigenvalues of AA^*)

$$\begin{pmatrix} \mu_1^n & 0 \\ 0 & \mu_2^n \end{pmatrix} \begin{pmatrix} \bar{\mu}_1^n & 0 \\ 0 & \bar{\mu}_2^n \end{pmatrix} = \begin{pmatrix} |\mu_1|^{2n} & 0 \\ 0 & |\mu_2|^{2n} \end{pmatrix}.$$

Therefore $\|\mathcal{G}^n\| \leq \|T\|\|T^{-1}\|$. To conclude the uniform boundedness of $\|\mathcal{G}^n\|$ as $n \to \infty$, we still need to show $\|T\|\|T^{-1}\|$ are bounded as $n \to \infty$. This is true since T depends on only ξ and λ. The eigenvectors of $\mathcal{G}$ can be explicitly computed. For instance, we can take

$$T = \begin{pmatrix} \mu_1 & \mu_2 \\ 1 & 1 \end{pmatrix}, \quad T^{-1} = \frac{1}{\mu_1 - \mu_2} \begin{pmatrix} 1 & -\mu_2 \\ -1 & \mu_1 \end{pmatrix}.$$

We have

$$T^*T = \begin{pmatrix} 2 & \mu_1^*\mu_2 + 1 \\ \mu_1\mu_2^* + 1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & -(\mu_1 + \mu_2)\mu_2 \\ -(\mu_1 + \mu_2)\mu_1 & 2 \end{pmatrix}$$

whose eigenvalues are bounded at least by 4, yielding $\|T\| \leq 2$. Similarly,

$$(T^{-1})^*T^{-1} = \frac{1}{(\mu_1 - \mu_2)^2} \begin{pmatrix} 2 & -(\mu_1 + \mu_2) \\ 0 & \mu_1(\mu_1 - \mu_2) \end{pmatrix},$$

whose eigenvalues are also bounded. Indeed, since $\frac{1}{(\mu_1-\mu_2)^2} \leq \frac{1}{4(1-\lambda^2)}$, we have $\|T^{-1}\|^2 \leq \frac{C}{1-\lambda^2}$ for some constant C.

This implies that $\|\mathcal{G}^n(\Delta t, k)\|$ is bounded for all values of Δt, k and n such that $t = n\Delta t$ is held fixed, which proves the assertion. □

Example 1.8. *Now we use a numerical example to understand the stability condition $\lambda < 1$ that we have just derived. Consider using the leapfrog scheme to solve $u_t = u_x$ with periodic boundary conditions on the inverval $x \in [-1, 1]$ and initial condition $f(x) = \frac{1}{a}e^{-x^2/a^2}$ with $a = 0.02$.*

First we consider the exact initial conditions, i.e, suppose we take $U^0 = f(x)$ and $U^1 = f(x+\Delta t)$. See Figure 1.1 for three cases $\frac{\Delta t}{\Delta x} = \lambda = 0.9, \lambda = 1$, and $\lambda = 1.1$ at time $t = 0.8$. We can observe that:

1. *The numerical solution blows up for $\lambda = 1.1$, as expected.*

2. *The case $\lambda = 1$ gives the best solution. This is actually not a surprise because the numerical stencil happens to coincide with the characteristic lines of $u_t = u_x$. In other words, the numerical scheme produces the exact solution in this case. For instance, the exact solution at $(x_j, 2\Delta t)$ is $f(x_j + 2\Delta t)$, while the leapfrog scheme gives $U_j^2 = U_j^0 + (U_{j+1}^1 - U_{j-1}^1) = U_{j+1}^1 = f(x_{j+1} + \Delta t) = f(x_j + 2\Delta t)$, where we use facts that $U_j^0 = U_{j-1}^1$ and $U_{j+1}^1 = f(x_{j+1} + \Delta t)$ (both are due to exact initial conditions).*

3. *There are some oscillations in the case $\lambda = 0.9$. There is nothing contradictory to the stability $\|\mathcal{G}^n\| \leq Ke^{\alpha t}$ because this stability is 0-stability, similar to what we defined for ODE solvers in Chapter **??**. These oscillations imply the error at this specific grid is large. On the other hand, if we refine the mesh ($\Delta x \to 0$), these errors will go away in a second order rate since this is a smooth solution. In other words, the oscillations in Figure 1.1 (d) are accuracy issues rather than stability issues.*

It is counterintuitive that an unstable scheme $\lambda = 1$ can produce a very nice solution. Actually it produces the exact one, which cannot be better. However, we have used the exact initial conditions. Now let us see what will happen if using inexact initial conditions to initiate the leapfrog scheme. We consider the following consistent perturbed initial conditions:

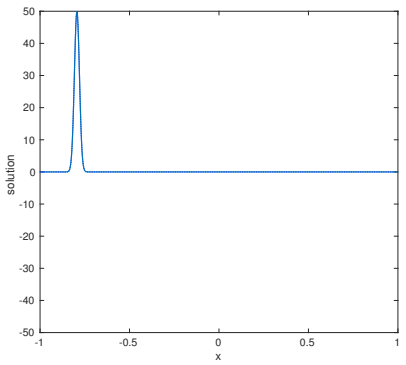
```
1  U_0=f(x);
2  U_1=f(x+dt)+dt*10*randn(size(x));
```

See Figure 1.2 for numerical solutions of $\lambda = 0.9$ and $\lambda = 1$ at a longer time $t = 2.8$. We can see that both stable and unstable schemes produce oscillations. However, the oscillations reduce when we refine the mesh in the stable scheme ($\lambda = 0.9$) while the oscillations increase when we refine the mesh in the unstable scheme ($\lambda = 1$). This is preciously what will happen for unstable schemes: we lose convergence (as $\Delta x \to 0, \Delta t \to 0$).

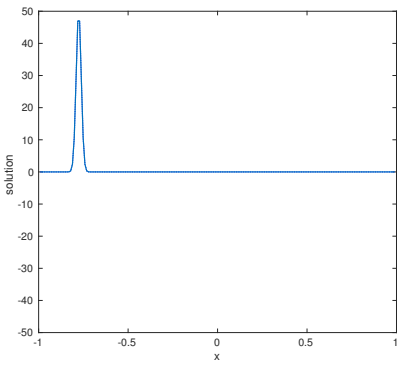
We have two interesting observations in this example:

- *An unstable scheme does not necessarily produce blow-ups. It is not enough to assert a scheme designed/implemented is stable if we only see the numerical solution on a coarse grid fits the reference solution well. It is necessary to validate the convergence by refining the mesh. For a linear problem, if there is no convergence (error stops to decrease when refining meshes), then there is no stability.*

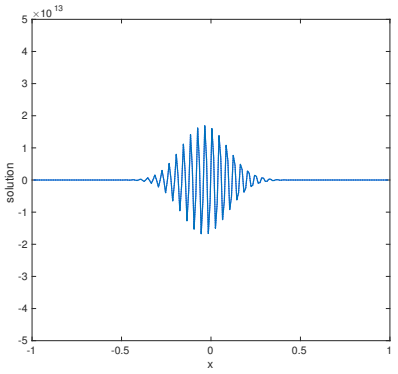
- *On some grid, an unstable scheme may produce better solutions, which does not imply any of its usefulness though.*



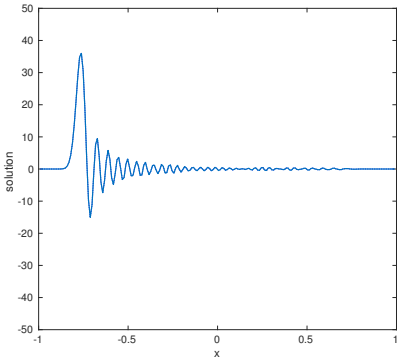
(a) Reference Solution.



(b) $\lambda = 1$ with exact initial conditions on 200 grid points.

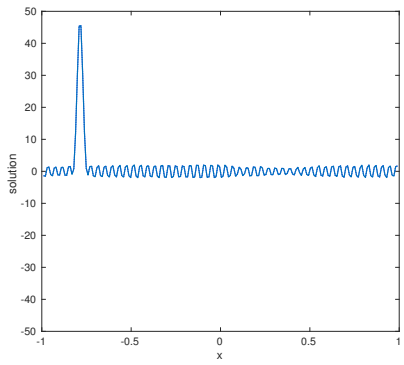


(c) $\lambda = 1.1$ with exact initial conditions on 200 grid points.

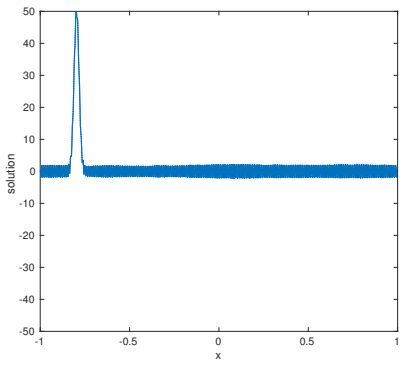


(d) $\lambda = 0.9$ with exact initial conditions on 200 grid points.

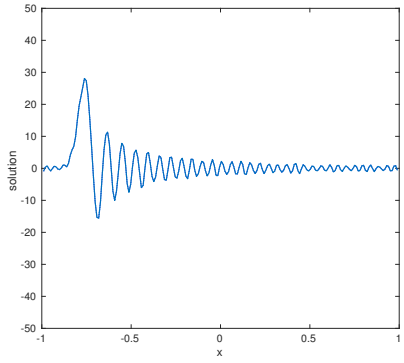
Figure 1.1: The leapfrog scheme for $u_t = u_x$ with exact initial conditions at time $t = 0.8$.



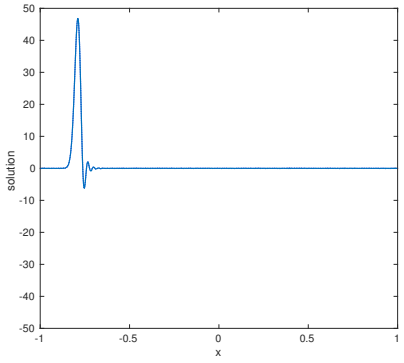
(a) $\lambda = 1$ on 200 grid points.



(b) $\lambda = 1$ on 800 grid points.

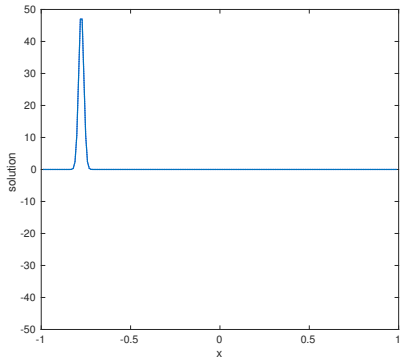


(c) $\lambda = 0.9$ on 200 grid points.

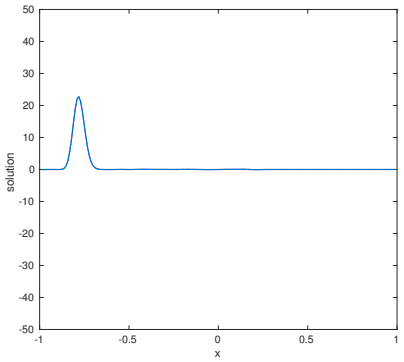


(d) $\lambda = 0.9$ on 800 grid points.

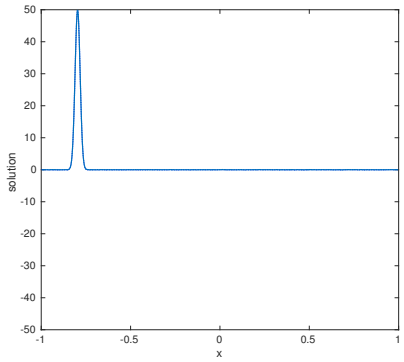
Figure 1.2: The leapfrog scheme for $u_t = u_x$ with consistent perturbed initial conditions at time $t = 2.8$. The oscillatory perturbation in the initial condition will vanish as $\Delta t \to 0$. However, the oscillations in the unstable scheme do not vanish as mesh refines.



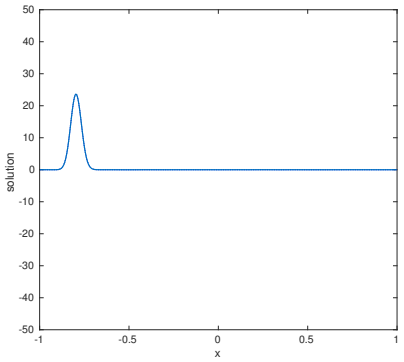
(a) $\lambda = 1$ on 200 grid points at time $t = 0.8$. Exact initial conditions.



(b) $\lambda = 0.9$ on 200 grid points at time $t = 0.8$. Exact initial conditions.



(c) $\lambda = 1$ on 800 grid points at time $t = 2.8$. Perturbed initial conditions.



(d) $\lambda = 0.9$ on 800 grid points at time $t = 2.8$. Perturbed initial conditions.

Figure 1.3: The upwind scheme for $u_t = u_x$.

Finally as a comparison, consider the first order accurate upwind scheme

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(U_{j+1}^n - U_j^n).$$

Plugging in the ansatz $U_j^{n+1} = \hat{U}_k^{n+1} e^{\mathbb{i}\, kj\Delta x}$, *we get*

$$\hat{U}_k^{n+1} e^{\mathbb{i}\, kj\Delta x} = \hat{U}_k^n e^{\mathbb{i}\, kj\Delta x} + \lambda(\hat{U}_k^n e^{\mathbb{i}\, k(j+1)\Delta x} - \hat{U}_k^n e^{\mathbb{i}\, kj\Delta x}),$$

thus the amplification factor is $g(k) = 1 - \lambda + \lambda e^{\mathbb{i}\, k\Delta x}$. *We have*

$$|g^n| = |g|^n = \left[(1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\cos\xi\right]^{\frac{n}{2}}.$$

For stability, we need $|g^n|$ *to be uniformly bounded as* $n \to \infty$, *which holds if and only if*

$$(1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\cos\xi \le 1,$$

i.e.,

$$2(1-\cos\xi)\lambda(\lambda - 1) \le 0.$$

So the upwind scheme is stable if and only if $\lambda \le 1$. *See Figure 1.3 for the performance of the upwind scheme with exact and similarly perturbed initial conditions. We can observe that*

- *The upwind scheme with* $\lambda = 1$ *also produces the exact solution with exact initial conditions, and it is stable.*

- *If we compare Figure 1.3 (b) with Figure 1.1 (d), then it may seem that the upwind scheme gives a better solution in some sense (less oscillatory), which is not contradictory to the fact that the leapfrog scheme is a more accurate scheme. Recall that we define the order of accuracy for* $\Delta x \to 0$ *for smooth solutions. In this example, the solution is smooth, but obviously it is underresovled on the 200-point mesh. In other words, comparison of accuracy of numerical schemes makes little sense (if there is any) on this mesh because even the sampling error (representing the initial data on 200 grid points) is huge. Recall that sampling in space is equivalent to periodization in frequency. Also see Shannon Sampling Theorem in Chapter* **??**.

Finally, let us try to understand the stability and "oscillations" in Figure 1.1 (d) from the perspective of stability region of ODE solvers. Recall that in Section **??** we defined the absolute stability for the linear multistep methods. In Example **??**, we found the stability region of the leagfrog method is the inverval $(-\mathbb{i}, \mathbb{i})$ on the imagnary axis. In particular, consider solving Example **??** by the leapfrog method. Namely, we solve the semidiscrete

scheme $\mathbf{U}'(t) = A\mathbf{U}$, with

$$A = \frac{1}{2\Delta x}\begin{pmatrix} 0 & 1 & & & & -1 \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 0 & 1 \\ 1 & & & & -1 & 0 \end{pmatrix},$$

by centered difference for the time derivative. Recall that A is circulant so DFT matrix diagonalizes it thus it is easy to find eigenvalues. The matrix A has purely imagnary eigenvalues because it is skew-symmetric. The eigenvalues are $\mathbb{i}\sin(k\Delta x)/\Delta x, k = 0, \cdots, N-1$. Since $\Delta x = \frac{2\pi}{N}$, the largest magnitude of the eigenvalues are $\mathbb{i}/\Delta x$ when $k\Delta x = \frac{\pi}{2}$. Thus to ensure the absolute stability, we need to take the time step to satisfy $\Delta t/\Delta x < 1$ (notice that $\lambda = 1$ will be on the outside of the stability region).

The "oscillations" in Figure 1.1 (d) do not "grow" in time. On the other hand, we need to see why we still have "oscillations" with absolute stability ensured. The absolute stability for a multistep method means the set of points z in complex plane so that the polynomial $\pi(\xi, z) = \rho(\xi) - z\sigma(\xi)$ satisfies the root condition. The root condition is derived from the initial value problem for the difference equation (for the leapfrog method solving $u' = au$),

$$U^{n+1} = U^{n-1} + 2\Delta t a U^n.$$

If $z = \Delta t a \in (-\mathbb{i}, \mathbb{i})$, then $\pi(\xi, z) = \rho(\xi) - z\sigma(\xi) = \xi^2 - 2z\xi - 1$ has two distinct roots ξ_1 and ξ_2 satisfying $|\xi_i| \leq 1$. The solution to this IVP can be written as

$$U^n = c_1\xi_1^n + c_2\xi_2^n.$$

Even though, $\|\xi_1^n\| \leq 1$ and $\|\xi_2^n\| \leq 1$, obviously we do not necessarily have $\|U^{n+1}\| \leq \|U^n\|$, which explains the "oscillations" in Figure 1.1 (d). However, the "energy" of U^n does not grow for fixed c_1 and c_2. In other words, the "oscillations" in Figure 1.1 (d) will not grow as time evolves.

1.5.2 The two way wave equation

The leapfrog method (second order centered difference for time and space derivatives) for the two-way wave equation $u_{tt} = u_{xx}$ is

$$\frac{U_j^{n+1} - 2U_j^n + U_j^{n-1}}{\Delta t^2} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2}. \tag{1.12}$$

The simplified 1D Maxwell's equations can be written as

$$\begin{cases} E_t &= H_x \\ H_t &= E_x \end{cases}, \tag{1.13}$$

which is equivalent to $E_{tt} = E_{xx}$ or $H_{tt} = H_{xx}$.

The FDTD method (second order centered difference for time and space derivatives) for (1.13) is defined on staggered grid for H:

$$\begin{cases} \frac{E_j^{n+1}-E_j^n}{\Delta t} & = \frac{H_{j+\frac{1}{2}}^{n+\frac{1}{2}}-H_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{\Delta x} \\ \frac{H_{j+\frac{1}{2}}^{n+\frac{1}{2}}-H_{j+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} & = \frac{E_{j+1}^n-E_j^n}{\Delta x} \end{cases} \tag{1.14}$$

It is a simple excersie to show that (1.14) is equivalent to (1.12) solving $E_{tt} = E_{xx}$ if we ignore the initial conditions.

Next, we consider the scheme (1.12) on the interval $x \in [0, 2\pi]$ with periodic boundary conditions. Let $\lambda = \frac{\Delta t}{\Delta x}$, then (1.12) can be written as

$$U_j^{n+1} = 2U_j^n + \lambda^2(E - 2 + E^{-1})U_j^n - U_j^{n-1},$$

where E is the shift operator. Define

$$V_j^n = \begin{pmatrix} U_j^n \\ U_j^{n-1} \end{pmatrix},$$

then we get

$$V_j^{n+1} = \begin{pmatrix} 2 + \lambda^2(E - 2 + E^{-1}) & -1 \\ 1 & 0 \end{pmatrix} V_j^n. \tag{1.15}$$

Plugging in the ansatz $V_j^n = \hat{V}_k^n e^{\mathbb{i}\, kj\Delta x}$, we get

$$\hat{V}_k^{n+1} = \begin{pmatrix} 2 + \lambda^2(e^{\mathbb{i}\, k\Delta x} - 2 + e^{-\mathbb{i}\, k\Delta x}) & -1 \\ 1 & 0 \end{pmatrix} \hat{V}_k^n.$$

Thus

$$\mathcal{G} = \begin{pmatrix} 2 + \lambda^2(2\cos\xi - 2) & -1 \\ 1 & 0 \end{pmatrix}.$$

The eigenvalues of $\mathcal{G}$ are $\mu_1 = a + \sqrt{a^2 - 1}, \mu_2 = a - \sqrt{a^2 - 1}$ with $a = 1 + \lambda^2(\cos\xi - 1)$. Notice that $-1 \le a \le 1$ if and only if $1 - \frac{2}{\lambda^2} \le \cos\xi \le 1$.

- If $\lambda > 1$, consider those ξ_0 such that $\cos\xi_0 < 1 - \frac{2}{\lambda^2}$. Then $a(\xi_0) < -1$ and $|\mu_2(\xi_0)| = |a - \sqrt{a^2 - 1}| > 1$. The von Neumann stability is violated thus not stable.

- If $\lambda \le 1$, then $a^2 - 1 \le 0$ thus $\mu_1 = a + \mathbb{i}\sqrt{1 - a^2}, \mu_2 = a - \mathbb{i}\sqrt{1 - a^2}$. So $|\mu_i| = 1$ and the von Neumann stability is satisfied. On the other hand, $\mathcal{G}$ is not a normal matrix and $\|\mathcal{G}\| > 1$. Nonetheless, $\mathcal{G}$ is diagonalizable

if $\mu_1 \neq \mu_2$, which is true if $\cos\xi \neq 1$. So $\mathcal{G} = T\begin{pmatrix}\mu_1 & 0\\ 0 & \mu_2\end{pmatrix}T^{-1}$ implies $\mathcal{G}^n = T\begin{pmatrix}\mu_1^n & 0\\ 0 & \mu_2^n\end{pmatrix}T^{-1}$, thus

$$\|\mathcal{G}^n\| \le \|T\|\left\|\begin{pmatrix}\mu_1^n & 0\\ 0 & \mu_2^n\end{pmatrix}\right\|\|T^{-1}\| = \|T\|\|T^{-1}\|\max_i|\mu_i^n| = \|T\|\|T^{-1}\|. \tag{1.16}$$

We still need to discuss $\|T\|$ and $\|T^{-1}\|$ and the case $\cos\xi = 1$ (or $\xi = 0$), see the discussion below for stability.

First we estimate $\|T\|$ and $\|T^{-1}\|$ for the case $\lambda \le 1$ and $\xi \neq 0$ (since $\xi = k\Delta x$, we consider $k = 1, 2, \cdots, N-1$). By using the fact $\mu_1\mu_2 = 1$, the eigenvectors can be chosen as

$$T = \begin{pmatrix}\mu_1 & \mu_2\\ 1 & 1\end{pmatrix}, \quad T^{-1} = \frac{1}{\mu_1 - \mu_2}\begin{pmatrix}1 & -\mu_2\\ -1 & \mu_1\end{pmatrix}.$$

Since $\mu_1^* = \mu_2$ and $\mu_2^* = \mu_1$, we have

$$T^*T = \begin{pmatrix}2 & \mu_1^*\mu_2 + 1\\ \mu_1\mu_2^* + 1 & 2\end{pmatrix} = \begin{pmatrix}2 & (\mu_1+\mu_2)\mu_2\\ (\mu_1+\mu_2)\mu_1 & 2\end{pmatrix}$$

whose eigenvalues are bounded at least by 4 (let x be an eigenvalue of TT^*, then $(x-2)^2 = (\mu_1+\mu_2)^2\mu_1\mu_2 = 4a^2$), yielding $\|T\| \le 2$. With $\mu_1\mu_2 = 1$, $\mu_1 + \mu_2 = 2a$ and $\mu_1 - \mu_2 = 2\,\mathrm{i}\,\sqrt{1-a^2}$, we have

$$(T^{-1})^*T^{-1} = \frac{1}{(\mu_1-\mu_2)(\mu_1^*-\mu_2^*)}\begin{pmatrix}2 & -(\mu_1+\mu_2)\\ -(\mu_1+\mu_2) & 2\mu_1\mu_2\end{pmatrix} = \frac{1}{2(1-a^2)}\begin{pmatrix}1 & -a\\ -a & 1\end{pmatrix}.$$

Let x_i be eigenvalues of $(T^{-1})^*T^{-1}$, then

$$x_1 = \frac{1}{2}\frac{1}{1-a}, \quad x_2 = \frac{1}{2}\frac{1}{1+a}.$$

Since $a = 1 + \lambda^2(\cos(k\Delta x) - 1)$, for fixed $\lambda \le 1$, by Taylor expansion on $\cos\Delta x$, we have

$$|x_i| \le \frac{1}{2\lambda^2}\frac{1}{1-\cos\Delta x} = \mathcal{O}(\Delta x^{-2}),$$

thus

$$\|T^{-1}\| \le C\Delta x^{-1}.$$

With (1.16), we have $\|\mathcal{G}^n\| \le C\Delta x^{-1}$, which means the scheme (1.12) is not stable according to the definition of stability.

Notice that we have used inequalities in (1.16), which might not be sharp. We can also compute $\mathcal{G}^n$ directly by

$$\mathcal{G}^n(k,\Delta t) = T\begin{pmatrix}\mu_1^n & 0\\ 0 & \mu_2^n\end{pmatrix}T^{-1} = \frac{1}{\mu_1-\mu_2}\begin{pmatrix}\mu_1^{n+1}-\mu_2^{n+1} & -\mu_1^n+\mu_2^n\\ \mu_1^n-\mu_2^n & -\mu_1^{n-1}+\mu_2^{n-1}\end{pmatrix}.$$

Since $|\mu_i|=1$, we can rewrite them as $\mu_1 = e^{\mathbb{i}\,\theta}$, $\mu_2 = e^{-\mathbb{i}\,\theta}$. So

$$\begin{aligned}\mathcal{G}^n(k,\Delta t) &= \frac{1}{e^{\mathbb{i}\,\theta}-e^{-\mathbb{i}\,\theta}}\begin{pmatrix}e^{\mathbb{i}(n+1)\theta}-e^{-\mathbb{i}(n+1)\theta} & -e^{\mathbb{i}\,n\theta}+e^{-\mathbb{i}\,n\theta}\\ e^{\mathbb{i}\,n\theta}-e^{-\mathbb{i}\,n\theta} & -e^{\mathbb{i}(n-1)\theta}+e^{-\mathbb{i}(n-1)\theta}\end{pmatrix}\\ &= \begin{pmatrix}\frac{\sin(n+1)\theta}{\sin\theta} & -\frac{\sin n\theta}{\sin\theta}\\ \frac{\sin n\theta}{\sin\theta} & -\frac{\sin(n-1)\theta}{\sin\theta}\end{pmatrix}\end{aligned}$$

As $k\longrightarrow 0$, $\theta\longrightarrow 0$ thus $\frac{\sin(n+1)\theta}{\sin\theta}\approx n+1$. So we have shown the following result

Lemma 1.2. *For the scheme* (1.12)*, for fixed* $\lambda\le 1$*, each entry of* $\mathcal{G}^n(k,\Delta t)$ *for* $k=1$ *is* $\mathcal{O}(n)$*.*

Next we look at what may happen when $\xi=0$ and $\lambda<1$ (similarly for the case $\lambda=1$ with $\xi=\pi$). Recall the discrete frequencies are $k=0,1,\cdots,N-1$ in the discrete Fourier transform that we used to derive the amplification matrix $\mathcal{G}(\Delta t,k)$. We have

$$\mathcal{G}(\Delta t,0) = \begin{pmatrix}2 & -1\\ 1 & 0\end{pmatrix}$$

with a repeated eigenvalue $\mu=1$. The Jordan form and the eigen-decomposition are

$$\mathcal{G}(\Delta t,0) = T\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}T^{-1},$$

thus

$$\mathcal{G}(\Delta t,0)^n = T\begin{pmatrix}1 & n\\ 0 & 1\end{pmatrix}T^{-1}.$$

Obviously we have $\|\mathcal{G}(\Delta t,0)^n\|\to\infty$ and $\|\mathcal{G}(\Delta t,N/2)^n\|\to\infty$ (assume N is even) as $n\to\infty$. So the scheme is unstalbe by the original definition. Now let us try to understand what it means that the stability is lost only when $k=0$ for $\lambda<1$ (and also $k=0,N/2$ for $\lambda=1$). The fact $\|\mathcal{G}(\Delta t,0)^n\|\to\infty$ implies that $\lim_{n\to\infty}|\hat{U}_0^n|=\infty$. In the discrete Fourier transform, the zero-th frequency corresponds to

$$\hat{U}_k^n = \sum_{j=0}^{N-1} e^{-\mathbb{i}\,kj\Delta x}U_j^n,\quad k=0,$$

thus

$$\hat{U}_0^n = \sum_{j=0}^{N-1} U_j^n.$$

Therefore this means that any perturbation in the total sum of the initial condition will not vanish as mesh refines. For instance, consider solving the IVP

$$u_{tt} = u_{xx}, u(x,0) = 0, u_t(x,0) = 0,$$

with periodic b.c. on an interval. Then the exact solution is constant zero. If we use the leapfrog scheme with the following initial conditions:

$$U^0 \equiv 0, U^1 \equiv \Delta t.$$

Plugging initial conditions into the scheme (1.12), we obtain $U^m = m\Delta t$. For any n satisfying $n\Delta t = t$, $U^n \equiv t$ thus we do not have convergence at all. On the other hand, if we use a second order accurate initial conditions:

$$U^0 \equiv 0, U^1 \equiv \Delta t^2,$$

then $U^n \equiv n\Delta t^2 = t\Delta t \to 0$ as $n \to \infty$.

Similar discussion holds for $k = N$ when $\lambda = 1$. The frequency $k = 0$ corresponds to the vector $[1 \quad 1 \quad \cdots \quad 1]$ while $k = N$ corresponds to the vector $v(N) = [1 \quad -1 \quad 1 \quad -1 \cdots -1]$ (if N is even). So any perturbation of the form $\Delta t v(N)$ in the initial condition will destroy convergence.

Therefore, at least for the case $\lambda < 1$, as long as we have an accurate initial condition so that the perturbation in the total sum is smaller than $\mathcal{O}(\Delta t)$ (a second order initial condition can be achieved by Taylor expansion $u(x, \Delta t) \approx u(x,0) + \Delta t u_t(x,0)$ since both $u(x,0)$ and $u_t(x,0)$ are given), it is still possible to have convergence.

1.5.3 Convergence for the two way wave equation

We can modify the proof of the Lax equivalence theorem to prove the convergence for the scheme (1.12). First, replace U_j^n by $u(x_j, t^n)$ in (1.12), the residue is the local truncation error

$$\tau^n = \mathcal{O}(\Delta t^2) + O(\Delta x^2).$$

Second, replace U_j^n by $u(x_j, t^n)$ in (1.15), the residue is

$$\Delta t^2 \tau^n = \Delta t^2 [\mathcal{O}(\Delta t^2) + O(\Delta x^2)].$$

Let $V^{n+1} = C(\Delta t)V^n$ denote the leapfrog scheme. Suppose

$$V^{n+1} = \begin{pmatrix} U_0^{n+1} \\ U_0^n \\ U_1^{n+1} \\ U_1^n \\ \vdots \\ U_{N-1}^{n+1} \\ U_{N-1}^n \end{pmatrix},$$

then define $Q_{\Delta x}$ as the sampling operator of any function at the spatial grid points and two time steps:

$$Q_{\Delta x}u(x,t) = \begin{pmatrix} u(x_0,t) \\ u(x_0,t-\Delta t) \\ u(x_1,t) \\ u(x_1,t-\Delta t) \\ \vdots \\ u(x_{N-1},t) \\ u(x_{N-1},t-\Delta t) \end{pmatrix}.$$

Define

$$\delta^n = [C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)]u(x,t),$$

where $S(\Delta t)u(x,t) = u(x,t+\Delta t)$ is the exact solution operator. Then

$$\delta^n = \Delta t^2 \tau^n.$$

The actual error that we want to control to prove the convergence is

$$\begin{aligned} \varepsilon^{n+1} &= V^{n+1} - Q_{\Delta x}u(x,(n+1)\Delta t) \\ &= C(\Delta t)[V^n - Q_{\Delta x}u(x,n\Delta t)] + [C(\Delta t)Q_{\Delta x} - Q_{\Delta x}S(\Delta t)]u(x,n\Delta t) \\ &= C(\Delta t)\varepsilon^n + \delta^n \end{aligned}$$

By solving $\varepsilon^{n+1} = C(\Delta t)\varepsilon^n + \delta^n$ (we no longer assume $\varepsilon^0 = 0$), we get

$$\varepsilon^n = C(\Delta t)^n \varepsilon^0 + \sum_{k=0}^{n-1} C(\Delta t)^{n-k-1}\delta^k.$$

Let F denote the $2N \times 2N$ matrix representing the linear transformation of taking the discrete Fourier transform for U^n and U^{n+1} respectively in V^{n+1}, i.e.,

$$FV^{n+1} = \begin{pmatrix} \hat{U}_0^{n+1} \\ \hat{U}_0^n \\ \hat{U}_1^{n+1} \\ \hat{U}_1^n \\ \vdots \\ \hat{U}_{N-1}^{n+1} \\ \hat{U}_{N-1}^n \end{pmatrix}.$$

Let $\hat{V}^n = FV^n$ and the amplification matrix be $\mathcal{G}(k)$ ($k = 0, \cdots, N-1$ is the discrete frequency). Then the scheme $V^{n+1} = C(\Delta t)V^n$ is equivalent to $\hat{V}^{n+1} = FC(\Delta t)F^{-1}\hat{V}^n$ and $FC(\Delta t)F^{-1}$ is a block diagonal matrix:

$$FC(\Delta t)F^{-1} = G = \begin{pmatrix} \mathcal{G}(0) & & & \\ & \mathcal{G}(1) & & \\ & & \ddots & \\ & & & \mathcal{G}(N-1) \end{pmatrix}.$$

In otherwords, the discrete Fourier transform that we have been using can block diagonalize the matrix $C(\Delta t)$.

Thus the error satisfies

$$\varepsilon^n = F^{-1}G^nF\varepsilon^0 + \sum_{k=0}^{n-1} F^{-1}G^{n-k-1}F\delta^k,$$

$$F\varepsilon^n = G^nF\varepsilon^0 + \sum_{k=0}^{n-1} G^{n-k-1}F\delta^k,$$

$$\hat{\varepsilon}^n = G^n\hat{\varepsilon}^0 + \sum_{k=0}^{n-1} G^{n-k-1}\hat{\delta}^k.$$

For $\lambda = \frac{\Delta t}{\Delta x} \leq 1$, in the previous subsection we have shown $\|\mathcal{G}^n(k)\| = \mathcal{O}(n) = \mathcal{O}(\Delta t^{-1})$ for any n and Δt satisfying $n\Delta t = t$ for fixed time t. Thus $\|G^n\| = \mathcal{O}(\Delta t^{-1})$

For the local truncation error part, since F is unitary, $\hat{\delta}^k = \Delta t^2\hat{\tau}^k = \Delta t^2[\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)]$. Thus

$$\left\|\sum_{k=0}^{n-1} G^{n-k-1}\hat{\delta}^k\right\| \leq \sum_{k=0}^{n-1} \|G^{n-k-1}\|\|\hat{\delta}^k\| = \sum_{k=0}^{n-1} \mathcal{O}(\Delta t)[\mathcal{O}(\Delta t^2)+\mathcal{O}(\Delta x^2)] = \mathcal{O}(\Delta t^2)+\mathcal{O}(\Delta x^2).$$

In other words, $\|G^n\| = \mathcal{O}(\Delta t^{-1})$ does not decrease the order of convergence by Δt^{-1} for the local truncation error part!

We only need to look at the numerical initial conditions. If the initial condition is second order accurate, i.e., $\varepsilon^0 = \mathcal{O}(\Delta t^2)$ thus $\hat{\varepsilon}^0 = \mathcal{O}(\Delta t^2)$. Then $G^n\hat{\varepsilon}^0 = \mathcal{O}(\Delta t)$, which is only first order. For instance, $\hat{\varepsilon}^0(0)$ denote the first two components in the vector $\hat{\varepsilon}^0$, then

$$\mathcal{G}(0)^n\hat{\varepsilon}^0(0) = T\begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}T^{-1}\begin{pmatrix} \mathcal{O}(\Delta t^2) \\ \mathcal{O}(\Delta t^2) \end{pmatrix} = T\begin{pmatrix} \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta t^2)n \\ \mathcal{O}(\Delta t^2) \end{pmatrix} = T\begin{pmatrix} \mathcal{O}(\Delta t) \\ \mathcal{O}(\Delta t^2) \end{pmatrix}.$$

So we still have the convergence $\lim_{\Delta t\to 0}\|\varepsilon^n\| = 0$.

To summarize, **the scheme** (1.12) **is not stable** by the stability definition even though we can still have convergence with more assumptions on initial conditions. On the other hand, the scheme (1.14) is stable if $\frac{\Delta t}{\Delta x} < 1$ thus (1.14) is convergent with consistent initial conditions.

Remark 1.5. *Recall that the 1D Maxwell equation* (1.13) *is equivalent to the equation* $E_{tt} = E_{xx}$*. However, the initial value problems for these two equations (even with periodic boundary conditions) are not necessarily equivalent. Consider the following initial value problems on the interval* $x \in [0, 2\pi]$ *with periodic boundary conditions:*

1. $E_{tt} = E_{xx}$ *with* $E(x,0)$ *and* $E_t(x,0)$ *given.*

2. *The system* (1.13) *with* $E(x,0)$ *and* $H(x,0)$ *given.*

For these two IVPs to be equivalent, $H(x,0) = \int E_t(x,0)\,dx$ *must hold. In other words, for generic periodic initial data* $E(x,0)$*,* $E_t(x,0)$ *and* $H(x,0)$*, these two IVPs are not equivalent.*

Problem 1.1. *Show the scheme* (1.14) *with periodic boundary conditions on* $x \in [0, 2\pi]$ *is stable for* $\frac{\Delta t}{\Delta x} < 1$*, by plugging in the ansatz* $E_j^n = \hat{E}_k^n e^{\mathbb{i}\, kj\Delta x}$ *and* $H_j^{n+\frac{1}{2}} = \hat{H}_k^{n+\frac{1}{2}} e^{\mathbb{i}\, k(j+\frac{1}{2})\Delta x}$*. The ansatz* $H_j^{n+\frac{1}{2}} = \hat{H}_k^{n+\frac{1}{2}} e^{\mathbb{i}\, k(j+\frac{1}{2})\Delta x}$ *is equivalent to using the transform* $H_j^{n+\frac{1}{2}} = \sum_{k=0}^{N-1} \hat{H}_k^{n+\frac{1}{2}} e^{\mathbb{i}\, k(j+\frac{1}{2})\Delta x}$ *(why?).*

Problem 1.2. *Recall that the scheme* (1.14) *is formally equivalent to* (1.12)*. However, these two schemes are obviously different since one is stable and the other one is not. To understand the difference or advantage on a staggered grid, consider the initial data* $E(x,0) = E_t(x,0) = H(x,0) \equiv 0$*, with which the two IVPs are equivalent. The scheme* (1.12) *is not convergent with the initial condition* $E^0 \equiv 0, E^1 \equiv \Delta t$*. Derive an initial condition* E^0 *and* $H^{\frac{1}{2}}$ *so that the solution to* (1.14) *is the same the solution to* (1.12) *with* $E^0 \equiv 0, E^1 \equiv \Delta t$*. What does this initial condition imply? Is there any contradiction to the fact that* (1.14) *is convergent with consistent initial conditions?*

1.6 Dissipative schemes

In practical applications, the spectral radius of the amplification matrix is often easy to evaluate. Looking for a sufficient condition, this time in terms of the spectral radius leads us to the concept of dissipation of a scheme, to which we now turn our attention.

Definition 1.9. *A finite difference scheme* $V^{n+1} = C(\Delta t)V^n$ *is called dissipative of order* $2r$ *if the amplification matrix satisfies:*

$$\rho[\mathcal{G}(\Delta t, k)] \leq 1 - \delta|\xi|^{2r},$$

where $\xi = k\Delta x$ *for all* $\Delta t, k$ *and* $\delta > 0$ *is independent of* k *and* Δt*.*

This condition means that the eigenvalues of the amplification matrix are bounded away from one in a way proportional to the parameter ξ. As mentioned earlier, it is in general true that even stable schemes have eigenvalue 1 for the mode $\xi = 0$. We shall let return to this fact in examples to come. Dissipation allows this case to happen, but all other eigenvalues are strictly inside the unit circle.

When a scheme is dissipative, it is very likely to be stable, even in the variable coefficient case, a fact that makes dissipation an important property of the schemes. We present some examples to illustrate the concept of dissipation and its relation to stability and "growth" of the numerical solution.

Example 1.9. *Consider the Lax-Wendroff scheme for $u_t = u_x$ with periodic boundary conditions:*

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

This scheme is second order accurate, both in space and time. By the ansatz $U_j^n = e^{\mathbb{i}\, kj\Delta t}\hat{U}_k^n$, we get the corresponding amplification factor

$$g(\xi) = 1 + \mathbb{i}\,\lambda \sin\xi + \lambda^2(\cos\xi - 1).$$

For convenience, let $\eta = \sin(\xi/2)$ then $\sin\xi = 2\sin(\xi/2)\cos(\xi/2) = 2\eta\sqrt{1-\eta^2}$, thus

$$g(\xi) = 1 - 2\lambda^2\eta^2 + 2\,\mathbb{i}\,\lambda\eta\sqrt{1-\eta^2},$$

and

$$|g(\xi)|^2 = 1 - 4\lambda^2(1-\lambda^2)\eta^4.$$

If $\lambda \leq 1$, we get $|g(\xi)| \leq 1$ for all ξ thus we have stability. If $\lambda < 1$, the scheme is dissipative of order 4. To see why this is true, we have

$$|g(\xi)|^2 = 1 - 4\lambda^2(1-\lambda^2)\left(\frac{\sin^4(\xi/2)}{(\xi/2)^4}\right)\left(\frac{\xi}{2}\right)^4 = 1 - 4\lambda^2(1-\lambda^2)\gamma\left(\frac{\xi}{2}\right)^4.$$

We also have

$$\frac{\sin\theta}{\theta} = \frac{\sin|\theta|}{|\theta|} \geq \frac{2}{\pi},$$

thus

$$\gamma = \frac{\sin^4(\xi/2)}{(\xi/2)^4} \geq \left(\frac{2}{\pi}\right)^4.$$

So we get

$$|g(\xi)|^2 \leq 1 - 4\lambda^2(1-\lambda^2)\left(\frac{2}{\pi}\right)^4\left(\frac{\xi}{2}\right)^4 = 1 - \frac{4}{\pi^4}\lambda^2(1-\lambda^2)\xi^4 \leq 1.$$

Notice, however, that if $\lambda = 1$, then the scheme is not dissipative.

Dissipation of a scheme may be desirable in some problems, as is the case of highly fluctuating initial data (or "noisy information"), and it ensures stability. But in other cases, if the effect of dissipation is too strong, we might lose our solution by an exaggerated smoothing mechanism, which is very likely to occur if we want to perform a large number of time iterations. Therefore, whether we should choose a dissipative scheme or not strongly depends on the particular problem we want to solve.

Example 1.10. *Consider again the problem $u_x = u_x$ with periodicity conditions, and the scheme:*

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

This scheme is both accurate and stable when the CFL condition $\lambda \leq 1$. holds. The amplification factor is given by:

$$g(\xi) = \cos\xi + \mathbb{i}\sin\xi$$

and therefore:

$$|g(\xi)| = \cos^2\xi + \lambda^2\sin^2\xi = 1 - (1-\lambda^2)\sin^2\xi.$$

By a similar argument to the one given in the previous example, it can be shown that when $\lambda < 1$, the scheme is dissipative of order 2. At this point we should notice that, as seen directly from expression of $|g(\xi)|$, the values $\xi = -\pi, 0, \pi$ yield $|g(\xi)| = 1$. Although the definition of dissipation does not hold exactly in the way stated, the inequality fails only for a finite number of values of ξ. We in general consider these schemes as dissipative ones. Again we have that the scheme reproduces the exact solution at the grid points when $\lambda = 1$, so in that case there is no dissipation.

Example 1.11. *Consider now the leap frog scheme for approximating the solution of $u_x = u_x$ with periodic boundary conditions. The amplification matrix is*

$$\mathcal{G}(\xi) = \begin{pmatrix} 2\,\mathbb{i}\,\lambda\sin\xi & 1 \\ 1 & 0 \end{pmatrix}.$$

If $\lambda < 1$, then the scheme is stable as discussed before, and the eigenvalues satisfy

$$\mu_1(\xi) = \mathbb{i}\,\lambda\sin\xi + \sqrt{1-\lambda^2\sin^2\xi},$$

$$\mu_2(\xi) = \mathbb{i}\,\lambda\sin\xi - \sqrt{1-\lambda^2\sin^2\xi},$$

$$|\mu_i(\xi)| = 1.$$

Thus $\rho(\mathcal{G}) = 1$ for all values of k and Δt, which implies the leapfrog scheme is not dissipative.

The following example illustrates how a non-dissipative scheme may give rise to a very a bad approximation of a system for which energy is being dissipated.

Example 1.12. *Let $u(x,t)$ be the solution of:*

$$u_x = u_x - \beta u,$$

where $\beta > 0$, and assume periodicity conditions. The usual energy estimates for this system can be evaluated multiplying by it and integrating by parts, yielding:

$$\frac{d}{dt}\|u(x,t)\|^2 = \frac{d}{dt}\int_0^{2\pi} |u(x,t)|^2\,dx = \int_0^{2\pi} \frac{d}{dx}u^2(x,t)\,dx - 2\beta\int_0^{2\pi} u^2(x,t)\,dx$$

$$= -2\beta\|u(x,t)\|^2,$$

where we have used $u(0,t) = u(2\pi,t)$ for all $t > 0$. Integrating with respect to time we obtain:

$$\|u(x,t)\|^2 = 2^{-2\beta t}\|u(x,0)\|^2$$

so the solution decreases in time, that is, the system is dissipating energy.

Consider now the leap frog scheme for this problem, given by:

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

with periodicity conditions:

$$U_{-1}^n = U_{N-1}^n, U_0^n = U_N^n.$$

It is easy to check that this scheme is second order accurate. If $\lambda = \frac{\Delta t}{\Delta x} < 1$, then $U_j^{n+1} = U_j^{n-1} + \frac{\Delta t}{\Delta x}(U_{j+1}^n - U_{j-1}^n)$ is stable. Thus by the Kreiss Perturbation Theorem 1.2, the scheme in this example is also stable.

The amplification matrix is

$$\mathcal{G}(\xi) = \begin{pmatrix} 2\,\mathrm{i}\,\lambda\sin\xi - 2\beta\Delta t & 1 \\ 1 & 0 \end{pmatrix}.$$

In particular, for the mode corresponding to $\xi = \pi$

$$\mathcal{G}(\pi) = \begin{pmatrix} -2\beta\Delta t & 1 \\ 1 & 0 \end{pmatrix}.$$

whose eigenvalues are

$$\mu_1(\pi) = -\beta\Delta t - \sqrt{1 + \beta^2\Delta t^2},$$

$$\mu_2(\pi) = -\beta\Delta t + \sqrt{1 + \beta^2\Delta t^2}.$$

If Δt is small but positive, we have:

$$\mu_1(\pi) \approx -\beta\Delta t - (1 + \frac{1}{2}\beta^2\Delta t^2) \approx -1 - \beta\Delta t$$

and thus, upon calling $T = n\Delta t$,

$$\mu_1(\pi)^n \approx (-1)^n(1 + \beta\Delta t)^n = (1 + \beta T/n)^n(-1)^n.$$

We know the scheme is accurate and stable, so by Lax equivalence theorem, it converges. Nonetheless, the concept of convergence involves taking limits of the approximations as $\Delta t \to 0$ with $T = n\Delta t$ fixed. In practice we deal with a fixed positive $\Delta t > 0$ and compute n time steps. As the number of time steps increases, the eigenvalue grows exponentially as:

$$\mu_1(\pi)^n \approx (-1)^n e^{\beta T}.$$

As discussed before, the energy of the true solution decreases exponentially with time, for any initial condition, whereas the scheme might give rise to increasing numerical solution in practice. To verify this statement, it is enough to consider a particular case for the initialization of the scheme and show that the corresponding numerical solution U^n grows in time. Consider the initial condition:

$$U_j^0 = u(x_j, 0) = (-1)^j.$$

In order to implement the scheme, we need to specify also the first time step U^1, which we give as

$$U_j^1 = \mu_1(\pi)(-1)^j.$$

Then the numerical solution is

$$U_j^n = (-1)^j q^n$$

where q satisfies:

$$q^{n+1} = q^{n-1} - 2\beta\Delta q^n.$$

This equation is equivalent to the quadratic equation:

$$q^2 + 2\beta\Delta t q - 1 = 0,$$

whose roots are preciously $\mu_1(\pi)$ and $\mu_2(\pi)$. Thus the general solution for q is of the form:

$$q = \alpha_1\mu_1(\pi) + \alpha_2\mu_2(\pi).$$

Since U_j^1 must coincide with the initialization given above, it follows that $q = \mu_1(\pi)$ thus $U_j^n = (-1)^j\mu_1^n(\pi)$, which grows exponentially with the number of iterations performed, keeping $\Delta t > 0$ fixed. In other words, if using a fixed time step Δt, for computing longer and longer time T, the energy of the numerical solution grows exponentially in T.

1.6.1 0-stability V.S. absolute stability

For a finite difference scheme $V^{n+} = C(\Delta t)V^n$, the stability that we defined in this chapter is to require $\|C(\Delta t)^n\| \leq Ke^{\alpha t}$ for any n and Δt satisfying $n\Delta t = t$, which is also called **Lax-Richtmyer stability**, which is very similar to the 0-stability as defined in Chapter **??**. On the other hand, we did not define the absolute stability for the scheme $V^{n+} = C(\Delta t)V^n$. Nonetheless, sometimes we achieved the absolute stability by requiring the Lax-Richtmyer stability. For instance, the amplification factor for the upwind scheme solving $u_t = u_x$ is $g(k) = 1 - \lambda + \lambda e^{\mathrm{i}\, k\Delta x}$, and

$$|g^n| = |g|^n = \left[(1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\cos\xi\right]^{\frac{n}{2}}.$$

For the Lax-Richitmyer stability, we need $|g^n|$ to be uniformly bounded as $n \to \infty$, which holds if and only if

$$(1-\lambda)^2 + \lambda^2 + 2\lambda(1-\lambda)\cos\xi \leq 1,$$

i.e.,

$$2(1-\cos\xi)\lambda(\lambda - 1) \leq 0.$$

Therefore, the Lax-Richitmyer stability holds if and only if $|g| \leq 1$, which is very similar to the absolute stability defined in Chapter **??**. In other words, we actually have the "absolute stability" for the schemes which perform well numerically, e.g., upwind and leagfrog schemes for $u_t = u_x$.

However, the "absolute stability" is less general than the Lax-Richtmyer stability, which is one of the reasons that we did not introduce or define the "absolute stability". For instance, the leapfrog scheme has Lax-Richtmyer stability and the "absolute stability" for the equation $u_t = u_x$. For the perturbed equation $u_t = u_x - \beta u$ with any $\beta > 0$ in Example 1.12, the leapfrog scheme also has Lax-Richtmyer stability due to the Kreiss Perturbation Theorem (Theorem 1.2), but the "absolute stability" is lost.

1.7 Difference schemes for hyperbolic systems in one dimension

It is often the case that the dimension of the space variable may change dramatically the properties of the numerical schemes, here we shall focus on problems in one dimension. Throughout this section, x will denote a scalar, and $u(x,t) = (u_1(x,t), \cdots, u_p(x,t))^T$ will denote the solution of a system of hyperbolic partial differential equations. We will be interested in approximating the solution $u(x,t)$ of the general, nonlinear equation of the form:

$$u_t(x,t) = \frac{\partial}{\partial x} F(u(x,t)), \tag{1.17}$$

where $F(u)$ is a function $F(u_1, \cdots, u_p) = (F_1(u_1, \cdots, u_p), \cdots, F_p(u_1, \cdots, u_p))^T$, e.g., the Euler equations discussed in Chapter **??**. Therefore we have:

$$\frac{\partial}{\partial x} F(u(x,t)) = \frac{\partial}{\partial u} F(u) \frac{\partial}{\partial x} u(x,t),$$

where $\frac{\partial}{\partial u} F(u)$ denotes the gradient matrix $A(u)$ with components $a_{ij}(u) = \frac{\partial F_i(u)}{\partial u_j}$ so that the nonlinear system can be written in the form:

$$u_t = A(u) u_x. \tag{1.18}$$

We now generalize the definitions given previously in Chapter **??** for hyperbolic partial differential equations in the nonlinear case.

Definition 1.10. *The nonlinear equation* (1.18) *is called weakly, strongly, symmetric or strictly hyperbolic if for every u_0 fixed, the corresponding linearized system:*

$$u_t = A(u_0) u_x$$

is weakly, strongly, symmetric or strictly hyperbolic, respectively.

As already mentioned before, the Lax equivalence theorem states basically that an accurate scheme is stable if and only if it converges, provided that the problem is strongly well posed. Weak well posedness may give rise to instabilities. Therefore, we shall consider only problems which are strongly, symmetric or strictly hyperbolic, yielding strong well posedness. We study separately the schemes which are accurate of order $(1,1)$, or first order schemes, and schemes which are accurate of order $(2,2)$, or second order schemes.

1.7.1 First order schemes

We shall consider two schemes: Friedrich's scheme and the upwind schemes. We will assume that the problem (1.18) is strongly well posed. The accuracy of the schemes can be checked directly in the nonlinear form (1.18), but in order to establish stability, as done for well posedness, we look at the linearized scheme substituting $A(u)$ by a constant matrix of the form $A(u_0)$, for which the problem is strongly well posed, as our previous assumption implies. Consider Friedrich's scheme:

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

where $F_{j+1}^n = F(U_{j+1}^n)$. This scheme is based on a first order approximation of the derivatives using Taylor expansion, and it can be easily shown that this scheme is first order accurate, and details are left to the reader. Linearizing the function $F(u)$ around some arbitrary value to u_0 we replace $A(u)$ by

a constant matrix A, so that the linearized problem is equivalent to the original problem with $F(u) = Au$. Substituting in the Friedrich's scheme, we get the linearized form of the scheme as:

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

The corresponding amplification matrix is given by:

$$\mathcal{G}(\xi) = I\cos\xi + \mathbb{i}\,\lambda A \sin\xi,$$

where $\xi = k\Delta x$, and I is the $p \times p$ identity matrix. If the original problem is strongly or strictly hyperbolic, then it follows that the matrix $A = A(u_0)$ is diagonalizable, i.e. there exists a matrix T such that

$$T^{-1}AT = \begin{pmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_p \end{pmatrix}$$

where $a_1, \cdots, a_p$ are the real eigenvalues of A. Therefore:

$$T^{-1}\mathcal{G}(\xi)T = I\cos\xi + \mathbb{i}\,\lambda \begin{pmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_p \end{pmatrix} \sin\xi$$

which is also diagonal with entries (eigenvalues) given explicitly by:

$$\mu_k(\xi) = cos\xi + \mathbb{i}\,\lambda a_k \sin\xi,$$

which implies that:

$$|\mu_k(\xi)|^2 = \cos^2\xi + \lambda^2 a_k^2 \sin^2\xi = 1 - (1 - \lambda^2 a_k^2)\sin^2\xi.$$

Therefore, if $\rho(A) = \max_k |a_k|$ satisfies the inequality

$$\frac{\Delta t}{\Delta x}\rho(A) \leq 1,$$

then von Neumann stability condition will hold and $|\mu_k(\xi)| \leq 1$ for all k and ξ. Furthermore, if $\frac{\Delta t}{\Delta x}\rho(A) < 1$, then $|\mu_k(\xi)|$ will be bounded away from 1 for all $0 \leq \xi < 2\pi$ except for $\xi = 0, \pi$. It is left as an exercise to prove that under strict inequality of von Neumann condition, the scheme is dissipative of order 2. Since $\mathcal{G}$ is diagonalized by a constant matrix T, the scheme for the linearized system is stable when $\frac{\Delta t}{\Delta x}\rho(A) \leq 1$. In practice, if the scheme for the linearized system is stable under the CFL condition $\frac{\Delta t}{\Delta x}\rho(A) \leq 1$, then the scheme for the nonlinear system is usually "stable" under the CFL condition $\frac{\Delta t}{\Delta x}\max_u \rho(A(u)) \leq 1$ for solving smooth solutions.

We now turn to the study of upwind schemes. These schemes are motivated by the scalar equation:

$$u_t = au_x$$

, when p = 1. If $a > 0$ the characteristics are straight lines moving to the left, and the scheme constructed in order to "follow" the physical characteristics is:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x} a(U_{j+1}^n - U_j^n), \quad a > 0, \tag{1.19}$$

and, as discussed before, the scheme is accurate and stable for $0 < a\lambda \leq 1$, for $\lambda = \Delta t/\Delta x$. On the other hand, if $a < 0$, then the characteristics point to the right and it is more reasonable to use the information carried by U_j^n and U_{j-1}^n, in order to evaluate U_j^{n+1} through the scheme:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x} a(U_j^n - U_{j-1}^n), \quad a < 0, \tag{1.20}$$

In this case, stability follows from the condition $-1 \leq \lambda a < 0$.

In order to extend the concept of upwind schemes to systems of hyperbolic equations some care must be taken. We shall gradually construct the general recursion formula.

Example 1.13. *Consider first the following example:*

$$\begin{pmatrix} u \\ v \end{pmatrix}_t = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_x$$

where $c > 0$. This system is equivalent to:

$$(u+v)_t = c(u+v)_x$$
$$(u-v)_t = -c(u-v)_x$$

Therefore an upwind schemecan be constructed naturally for $U + V$ and $U - V$, which yields:

$$U_j^{n+1} + V_j^{n+1} = U_j^n + V_j^n + \frac{\Delta t}{\Delta x} c(U_{j+1}^n + V_{j+1}^n - U_j^n - V_j^n)$$
$$U_j^{n+1} + V_j^{n+1} = U_j^n - V_j^n - \frac{\Delta t}{\Delta x} c(U_j^n + V_j^n - U_{j-1}^n - V_{j-1}^n).$$

Adding and subtracting these two equations, we get the following equivalent scheme:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x} c(V_{j+1}^n - V_{j-1}^n) + \frac{\Delta t}{2\Delta x} c(U_{j+1}^n - 2U_j^n + U_{j-1}^n)$$
$$V_j^{n+1} = V_j^n + \frac{\Delta t}{2\Delta x} c(U_{j+1}^n - U_{j-1}^n) + \frac{\Delta t}{2\Delta x} c(V_{j+1}^n - 2V_j^n + V_{j-1}^n). \tag{1.21}$$

Looking at these equations it is clear that even though we started with upwind schemes, (1.21) *are centered expressions: there is no longer any explicit direction for the characteristics. Also, it should be noticed that besides an approximation of a first order derivative, the above equations also contain an approximation to a second order derivative, which we did not have when we started the construction of the schemes. When generalizing the concept of an upwind scheme, we must allow for centered expressions that at first sight may not seem to carry information along characteristics, keeping in mind this simple example. The scheme* (1.21) *is a first order accurate (both in time and space) scheme for the first order convection equation, and a second order accurate (in space) scheme for the convection diffusion equation:*

$$u_t = cv_x + c\Delta x u_{xx}$$
$$v_t = cu_x + c\Delta x v_{xx},$$

which is called the modified equation for the scheme (1.21)*. Thus at least intuitively we expect the scheme* (1.21) *produces a smoother numerical solution than the exact solution to the original first order convection equation.*

We now write an equivalent expression for (1.19) *and* (1.20) *by adding and subtracting the appropriate terms*

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x}a(U_{j+1}^n - U_{j-1}^n) + \frac{\Delta t}{2\Delta x}a(U_{j+1}^n - 2U_j^n + U_{j-1}^n), \quad a > 0 \tag{1.22}$$

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x}a(U_{j+1}^n - U_{j-1}^n) - \frac{\Delta t}{2\Delta x}a(U_{j+1}^n - 2U_j^n + U_{j-1}^n), \quad a < 0 \tag{1.23}$$

from where it is clear now that the general form of upwind Scheme for the scalar case $p = 1$ *is given by:*

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x}a(U_{j+1}^n - U_{j-1}^n) + \frac{\Delta t}{2\Delta x}|a|(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \tag{1.24}$$

While schemes (1.22) *and* (1.23) *are hard to generalize to the variable coefficient case in the form they are usually written, it is straightforward to implement* (1.24) *in the case* $a = a(x)$ *using the values of* $a(x)$ *and* $|a(x)|$*.*

As can be verified from (1.25)*, there is indeed a term that approximates a second derivative within the upwind schemes. Furthermore, this term has a positive coefficient,* $|a| > 0$*, which in turn introduces a dissipative mechanism for the scheme.*

In order to generalize (1.24) for systems of hyperbolic equations we first have to define the "absolute" value of a matrix that will play the role of $|a|$ in the scalar case. Consider again the linearized, strongly hyperbolic system, so that the matrix A is a constant, diagonalizable matrix:

$$u_t = Au_x.$$

Definition 1.11. *Let A be diagonalizable by T, so that:*

$$\Lambda = T^{-1}AT = \begin{pmatrix} a_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_p \end{pmatrix}.$$

The absolute value of Λ is defined by:

$$|\Lambda| = \begin{pmatrix} |a_1| & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & |a_p| \end{pmatrix}$$

and the absolute value of the matrix A is defined to be $|A| = T|\Lambda|T^{-1}$, so that $|A|$ is also a $p \times p$ matrix which T itself diagonalizes.

Example 1.14. *For the matrix $A = \begin{pmatrix} 0 & c \\ c & 0 \end{pmatrix}$, we have:*

$$A = T \begin{pmatrix} c & 0 \\ 0 & -c \end{pmatrix} T^{-1}, \quad T = T^{-1} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Thus

$$|A| = T \begin{pmatrix} c & 0 \\ 0 & c \end{pmatrix} T^{-1} = cI.$$

The generalization of scheme (1.24) *to the system $u_t = Au_x$ is given by the scheme:*

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x} A(U_{j+1}^n - U_{j-1}^n) + \frac{\Delta t}{2\Delta x}|A|(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \quad (1.25)$$

It is straightforward to verify that (1.21) *satisfy* (1.25).

Definition 1.12. *Let $f(x)$ be a real valued function of the variable x. We define the positive part f^+ of f as the function $f^+(x) = \max\{0, f(x)\}$ or equivalently:*

$$f^+(x) = \begin{cases} f(x), & \text{if} \quad f(x) > 0 \\ 0, & \text{if} \quad f(x) \le 0 \end{cases}$$

and analogously, the negative part $f^-(x)$ of f is defined by $f^-(x) = -\max\{0, -f(x)\}$. Using these definitions, it follows that:

$$f = f^+ + f^-, \quad |f| = f^+ - f^-.$$

Substituting in (1.24) the values of a and $|a|$ in terms of the positive and negative parts, we obtain an alternative expression for the upwind scheme in the scalar case, namely,

$$U_j^{n+1} = U_j^n + (a^+)\frac{\Delta t}{\Delta x}(U_{j+1}^n - U_j^n) + (a^-)\frac{\Delta t}{\Delta x}(U_j^n - U_{j-1}^n). \quad (1.26)$$

This representation of the upwind scheme has the advantage that it shows explicitly the directions of the characteristics that the scheme "picks' according to the sign of a, which becomes more useful when a is a variable coefficient $a(x)$. Following the natural extension, we can now define the positive and negative part of a diagonalizable matrix in terms of the absolute value.

Definition 1.13. *Let A be a diagonalizable matrix. We define the positive (negative) part of A by:*

$$A^+ = \frac{A + |A|}{2}, \quad A^- = \frac{A - |A|}{2}.$$

Scheme (1.25) can now be written in a more compact form using the positive and negative parts of the matrix A, yielding:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x} A^+ (U_{j+1}^n - U_j^n) + \frac{\Delta t}{\Delta x} A^- (U_j^n - U_{j-1}^n).$$

This expression gives the general form of the upwind scheme for approximating the solution of symmetric, strongly or strictly hyperbolic systems with constant coefficients. To generalize the scheme to the nonlinear case, where $A = A(u)$, we need some material on nonlinear equations in a more general scope. The foundamental method of this type is called Godunov's scheme, which we will not introduce in this chapter. We summarize now the concepts related to the upwind scheme in the linear case:

$$u_t = A u_x$$

where A is diagonalizable, so that the problem is strongly well posed. Accuracy of order (1,1) and stability of the upwind scheme follow straightforward assuming:

$$\rho(A) \frac{\Delta t}{\Delta x} \leq 1.$$

Indeed, one can decouple the system using the transformation $w = Tu$, which yields

$$w_t = \Lambda w_x.$$

The corresponding scheme defined by $W_j^n = TU_j^n$ has p components that satisfy schemes (1.24) or, equivalently, (1.26) with a, $|a|$, and a^+, a^- replaced in terms of the eigenvalues a_k of A. For each component the scheme for W_j^n is accurate of first order and stable, as an approximation of the solution of $w_t = \Lambda w_x$. Applying the bounded, linear transformation T to W_j^n, the result for the original problem is established.

1.7.2 Second order schemes

Roughly speaking, we can divide second order schemes into the dissipative and the non-dissipative ones. As before, we will assume strong well posedness of the problem $u_t = A(u)u_x$. Accuracy of the schemes can be evaluated directly for the schemes in general form, but in order to establish stability, we shall consider the linearized versions, as we did in the previous section. A representative of the class of dissipative schemes of second order accuracy is the Lax-Wendroff scheme, which we shall study first.

Definition 1.14. *A scheme for approximating the solution of $u_t = A(u)u_x$ is called a Lax-Wendroff scheme if under the assumption $A(u) = A$ (or $F(u) = Au$ is linear), the scheme reduces to:*

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x} A(U_{j+1}^n - U_{j-1}^n) + \frac{1}{2}\left(\frac{\Delta t}{\Delta x} A\right)^2 (U_{j+1}^n - 2U_j^n + U_{j-1}^n). \quad (1.27)$$

It may be shown that the scheme (1.27) is actually **the only second order scheme** for the linear problem that uses U_{j-1}^n, U_j^n, and U_{j+1}^n, to evaluate U_j^{n+1}.

Lax-Wendroff schemes arise from the idea of replacing time derivatives by space derivatives, using the equation $u_t = F(u)$, and approximating the latter by finite differences. Using a Taylor expansion for u, we have:

$$u(x, t + \Delta t) = u(x, t) + \Delta t u_t(x, t) + \frac{\Delta^2}{2} u_{tt}(x, t) + \mathcal{O}(\Delta t^3).$$

Since $u_t(x, t) = F(u(x, t))$, in the linear case where $F(u) = Au$ we get:

$$u_t(x, t) = Au_x(x, t), \quad u_{tt}(x, t) = A^2 u_{xx}(x, t).$$

Using now finite difference approximations for u_x and $u_x x$, it follows that the linear form of the scheme (1.27) is accurate of order (2,2). The amplification matrix of the linear form of the Lax-Wendroff scheme is:

$$\mathcal{G}(\xi) = I + \mathbb{i}\, \lambda A \sin \xi + \lambda^2 A^2 (\cos \xi - 1),$$

where, as usual, $\xi = k\Delta t$ and $\lambda = \Delta t/\Delta x$. Calling $\eta = sin(\xi/2)$ we can write:

$$\mathcal{G}(\xi) = I + 2\,\mathbb{i}\, \lambda A \eta \sqrt{1 - \eta^2} - 2\lambda^2 A^2 \eta^2.$$

Therefore any eigenvalue $\mu(\eta)$ of the amplification matrix will be of the form:

$$\mu(\eta) = 1 + 2\,\mathbb{i}\, \lambda \mu(A) \eta \sqrt{1 - \eta^2} - 2\lambda^2 \mu(A)^2 \eta^2,$$

which follows from the fact that A is diagonalizable. From the expression of the eigenvalues $\mu(\eta)$ of the amplification matrix we have:

$$|\mu(\eta)|^2 = 1 - \lambda^2 \mu(A)^2 \eta^4 (1 - \lambda^2 \mu(A)^2)$$

which holds for every eigenvalue of $\mathcal{G}(\xi)$. Recall that the spectral radius of $\mathcal{G}(\xi)$ is defined as the maximum value of $\mu(\eta)$. Therefore, upon letting μ_* be the eigenvalue of A which maximizes the above expression for $|\mu(\eta)|$ we get:

$$|\rho(\mathcal{G})|^2 = 1 - \lambda^2\mu_*^2\eta^4(1 - \lambda^2\mu_*^2)$$

Clearly, von Neumann condition will be satisfied if

$$\lambda\rho(A) \leq 1$$

which implies $\lambda\mu(A) \leq 1$ for all eigenvalues of A. Furthermore, if $\lambda\mu_* < 1$, then the scheme given by (1.27) is dissipative of order 4. Here the dissipation can be controlled through the parameter λ, or, equivalently, through the choice of Δt. In the nonlinear case we can construct different schemes which fall within the class of LaxWendroff schemes, depending on the way we approximate the derivatives.

For the nonlinear case, we have

$$u_{tt} = [F(u)]_{xt} = [F(u)_t]_x = [A(u)u_t]_x = [A(u)F(u)_x]_x.$$

Substituting $u_t = F(u)_x$ and the above expression in the Taylor expansion, we get:

$$u(x, t + \Delta t) = u(x, t) + \Delta t F(u)_x + \frac{\Delta t^2}{2}[A(u)F(u)_x]_x + \mathcal{O}(\Delta t^3).$$

The scheme originally proposed by Lax and Wendroff is based on approximating the space derivatives in the expansion above up to order $O(\Delta x^2)$ and is given by:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x}(F_{j+1}^n - F_{j-1}^n) + \frac{1}{2}\left(\frac{\Delta t}{\Delta x}\right)^2 (A_{j+\frac{1}{2}}^n(F_{j+1}^n - F_j^n) - A_{j-\frac{1}{2}}^n(F_j^n - F_{j-1}^n)). \tag{1.28}$$

where:

$$F_j^n = F(U_j^n), \quad A_{j+\frac{1}{2}}^n = A(\frac{U_{j+1}^n + U_j^n}{2}).$$

Scheme (1.28) becomes rather inefficient in practical applications due to the many computations involved at each time step iteration in order to evaluate A, and F. A modification of this scheme which is very popular considers approximating derivatives at "half stages" of the iteration, using:

$$u(x, t + \Delta t) = u(x, t) + \Delta t u_t(x, t + \frac{1}{2}\Delta t) + \mathcal{O}(\Delta t^2),$$

and it is known as the *MacCormack* scheme. Each iteration has two steps corresponding to first order approximations of the solution at half steps.

The Scheme is given by:

$$U_j^* = U_j^n + \frac{\Delta t}{\Delta x}(F_{j+1}^n - F_j^n)$$
$$U_j^{n+1} = \frac{1}{2}\left(U_j^n + U_j^* + \frac{\Delta t}{\Delta x}(F_j^* - F_{j-1}^*)\right)$$

where

$$F_j^n = F(U_j^N), \quad F_j^* = F(U_j^*).$$

This scheme is a two-stage scheme which evaluates a "predictor" U_j^* and a "corrector" $U_j^{**} = U_j^* + \frac{\Delta t}{\Delta x}(F_j^* - F_{j-1}^*)$, and then forms U_j^{n+1} as the average $(U_j^* + U_j^{**})/2$.

It is clear that in order to evaluate U_j^{n+1} the scheme uses the same points in the grid at time in as the Lax-Wendroff scheme. Notice, however, that here we go from right to left at the middle stage $*$, and then from left to right. The "efficiency" of a scheme is often related to the cost in computer time of each iteration. In these terms, one can compare different schemes. For the Lax-Wendroff scheme, we need to evaluate F_{j-1}^n, F_j^n, F_{j+1}^n, $A_{j+\frac{1}{2}}^n$ and $A_{j-\frac{1}{2}}^n$ and perform matrix multiplications in each iteration, whereas MacCormack Scheme requires only the evaluation of F_j^n, F_{j+1}^n, F_j^* and F_{j-1}^*.

It only remains to prove the order of accuracy of MacCormack scheme. The fact that it belongs to the class of Lax-Wendroff schemes follows straightforward replacing $F(u)$ by Au with A a constant matrix.

The local truncation error of the MacCormack scheme is $\mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) + \mathcal{O}(\Delta t \Delta x)$, in which we assume $\Delta t = \mathcal{O}(\Delta x)$. Thus it is a second order accurate in space and time.

Among the class of second order non-dissipative schemes is the leap frog scheme. For the general non-linear equation, the scheme is given by:

$$U_j^{n+1} = U_j^{n-1} + \frac{\Delta t}{\Delta x}(F_{j+1}^n - F_{j-1}^n). \tag{1.29}$$

We analyzed this scheme in detail for the linear case, and found out that it is not dissipative and it is stable, provided that $\frac{\Delta t}{\Delta x}\rho(A) < 1$. The fact that (1.29) is accurate of second order follows a straightforward calculation. This scheme is generally more efficient than Lax-Wendroff schemes, although it needs roughly twice as much memory due to the dependence on two previous time stages to evaluate U^{n+1}, therefore in practice, we usually face the trade-off between efficiency and storage requirements. Since this is a non-dissipative scheme, it will not give good approximations for nonlinear equations. We now proceed to describe a way to introduce a dissipative term in (1.29) to deal with this problem. When adding a dissipative term in the form of a small perturbation, care must be taken so that the resulting linear scheme retains stability. Recall that in the linear case $F(u) = Au$,

the amplification matrix $\mathcal{G}(\xi)$ is a $2p \times 2p$ matrix (A itself is a $p \times p$ matrix) given by:

$$\mathcal{G}(\xi) = \begin{pmatrix} 2\,\mathbb{i}\,\lambda A \sin\xi & I \\ I & 0 \end{pmatrix}.$$

where now each of the entries is itself a $p \times p$ matrix. In order to express the eigenvalues $\mu(\xi)$ of $\mathcal{G}$ in terms of those of A, we use the fact that if A is diagonalizable by a matrix T, then $\hat{\mathcal{G}}$ possesses the same eigenvalues of $\mathcal{G}$, for:

$$\hat{\mathcal{G}}(\xi) = \begin{pmatrix} T^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} 2\,\mathbb{i}\,\lambda A \sin\xi & I \\ I & 0 \end{pmatrix} \begin{pmatrix} T & 0 \\ 0 & I \end{pmatrix} = \begin{pmatrix} 2\,\mathbb{i}\,\lambda T^{-1}AT \sin\xi & I \\ I & 0 \end{pmatrix}.$$

Recall that $T^{-1}AT$ is a diagonal matrix with diagonal entries $a_1, \cdots, a_p$. From this expression (by rearranging rows/columns, $\hat{\mathcal{G}}(\xi)$ is similar to a block diagonal matrix with 2×2 diagonal blocks $\begin{pmatrix} 2\,\mathbb{i}\,\lambda a_j \sin\xi & 1 \\ 1 & 0 \end{pmatrix}$), it follows that any eigenvalue $\mu(\xi)$ of the amplification matrix satisfies:

$$\mu^2(\xi) = 1 + 2\,\mathbb{i}\,\lambda a_j \sin\xi \mu(\xi),$$

for some $j = 1, \cdots, p$.

We will show that, if we add a dissipative term to the leap frog scheme at time level n, this will give rise to instabilities. By a dissipative term we mean an approximation to a second derivative, as would be a term of the form:

$$\varepsilon(U^n_{j+1} - 2U^n_j + U^n_{j-1}), \tag{1.30}$$

added to the scheme (1.29), where ε is a "small" perturbation. Notice that any modification at time level n will affect the first block in the amplification matrix. If the term added is (1.30), then the modified amplification matrix will be of the form:

$$\mathcal{G}(\xi) = \begin{pmatrix} 2\,\mathbb{i}\,\lambda A \sin\xi + \varepsilon \sin^2(\xi/2) I & I \\ I & 0 \end{pmatrix}$$

and therefore the eigenvalues will now satisfy:

$$\mu^2(\xi) = 1 + (2\,\mathbb{i}\,\lambda a_j \sin\xi + \varepsilon \sin^2(\xi/2))\mu(\xi).$$

In general, if E denotes the shift operator $EU^n_j = U^n_{j+1}$, adding a dissipative term at time level n amounts to modifying (1.29) yielding the scheme:

$$U^{n+1}_j = U^{n-1}_j + \frac{\Delta t}{\Delta x} A(U^n_{j+1} - U^n_{j-1}) + \varepsilon P(E) U^n_j, \tag{1.31}$$

where $P(E)$ is a function of the shift operator (in particular, the term in (1.30) corresponds to $P(E) = E - 2I + E^{-1}$). Since $P(E)$ approximates a

second order derivative, its Fourier transform $\hat{P}(\xi)$ will be a real function of ξ. It is this function $\hat{P}(\xi)$ which will appear now added in the first block of the amplification matrix and thus the modified eigenvalues will in general satisfy:

$$\mu^2(\xi) = 1 + (2\,\mathbb{i}\,\lambda a_j \sin\xi + \varepsilon\hat{P}(\xi))\mu(\xi).$$

for some eigenvalue at of A. The fact that (1.31) is an unstable scheme follows now from the following lemma, applied to the eigenvalues $\mu(\xi)$.

Lemma 1.3. *Let x_1 and x_2 be the solutions of the equation $x^2 - \alpha x - 1 = 0$. If both $|x_1| \leq 1$ and $|x_2| \leq 1$, then necessarily the coefficient α is purely imaginary.*

Proof. Let $x_1 = re^{\mathbb{i}\,\theta}$, then $x_1 x_2 = -1$ implies $x_2 = \frac{1}{r}e^{-\mathbb{i}\,\theta}$. Both $|x_1| \leq 1$ and $x_2 \leq 1$ imply $r = 1$. Thus $\alpha = x_1 + x_2 = 2\,\mathbb{i}\sin\theta$. □

Remark 1.6. *Using exactly the same analysis, we may conclude in general that the leap frog scheme gives rise to instabilities when it is used to approximate parabolic equations. For the heat equation, this can also be explained by the stability region of the leapfrog method, which is only on the imaginary axis, while the centered finite difference used in approximating the second order derivatives will give real eigenvalues, as discussed in Example* **??**.

In order to introduce the correct amount of dissipation, we must add the dissipation term at time level $n-1$. We shall use the following operator $E^{\frac{1}{2}}$ which is defined as $E^{\frac{1}{2}}U_j^n = U_{j+\frac{1}{2}}^n$. Using this notation, the leap frog scheme (1.29) can be rewritten in the form:

$$U_j^{n+1} = U_j^{n-1} + \frac{\Delta t}{\Delta x}(E^{\frac{1}{2}} - E^{-\frac{1}{2}})(E^{\frac{1}{2}} + E^{-\frac{1}{2}})F_j^n$$

We shall show now that the modification of the scheme that is dissipative is given in general form by the expression:

$$U_j^{n+1} = U_j^{n-1} + \frac{\Delta t}{\Delta x}(E^{\frac{1}{2}} - E^{-\frac{1}{2}})(E^{\frac{1}{2}} + E^{-\frac{1}{2}})F_j^n - \frac{\varepsilon}{16}(E^{\frac{1}{2}} - E^{-\frac{1}{2}})^4 U_j^{n-1}. \tag{1.32}$$

Let $\eta = \sin(\xi/2)$, the amplification matrix of the linearized scheme (1.32) is:

$$\mathcal{G}(\xi) = \begin{pmatrix} 2\,\mathbb{i}\,\lambda A\sin\xi & (1-\varepsilon\eta^4)I \\ I & 0 \end{pmatrix}$$

and the eigenvalues hold the relations:

$$\mu^2(\xi) = 1 - \eta^4 + 2\,\mathbb{i}\,\lambda\mu(A)\sin\xi\sin\xi\mu(\xi),$$

for some eigenvalue $\mu(A)$ of A. Therefore:

$$\mu(\xi) = \mathbb{i}\,\lambda\mu(A)\sin\xi \pm \sqrt{1 - |\mu(A)|^2\sin^2\xi - \varepsilon\eta^4}$$

and we have $|\mu(\xi)|^2 = 1 - \varepsilon\eta^4$, provided that

$$1 - |\lambda\mu(A)|^2 \sin^2 \xi - \varepsilon\eta^4 > 0, \tag{1.33}$$

for all eigenvalues of A and all ξ. Under this condition, the modified scheme (1.32) is stable and dissipative. Remark, though, that in order for (1.33) to hold, whenever we add dissipation ($\varepsilon > 0$), we must also decrease the value of $\lambda = \Delta t/\Delta x$. This means that for a fixed space grid, a larger number of time steps must be evaluated to get the approximation of the solution at some given time t.

2

A brief introduction to nonlinear conservation laws

Preliminaries

- **Model problem**: Scalar conservation law

$$u_t + f_x(u) = 0. \tag{2.1}$$

  Given initial: $u(x, 0)$. Note, the subscript in (2.1) denotes derivative, for instance $u_t = \partial_t u$ and $f_x = \partial_x f$.

- **Weak solution**: Multiply test function $\phi \in C_0^1(\mathbb{R} \times \mathbb{R}^+)$ on (2.1) and integrate over space and time

$$\int_0^{+\infty} \int_{-\infty}^{+\infty} (u_t + f_x(u))\, \phi \, \mathrm{d}x\mathrm{d}t = 0. \tag{2.2}$$

  Here, C_0^1 is the space of function that are continuous differentiable with compact support. Integrate by part, yield the weak solution

$$\int_0^{+\infty} \int_{-\infty}^{+\infty} (\phi_t u + \phi_x f(u))\ \mathrm{d}x\mathrm{d}t = -\int_{-\infty}^{+\infty} \phi(x, 0) u(x, 0)\, \mathrm{d}x. \tag{2.3}$$

- **Numerical scheme**

$$U_j^{n+1} = U_j^n - \frac{k}{h} \left(F(U^n; j) - F(U^n; j-1) \right) \tag{2.4}$$

  Here, $F(U^n; j)$ is a flux function which is allow to depend on any finite number of elements of the vector U^n, "*centered*" about the j^{th} point.

$$F(U^n; j) = F(U_{j-p}^n, U_{j-p+1}^n, \cdots, U_{j+q}^n). \tag{2.5}$$

- **Consistency**: The numerical flux function F reduces to the true flux f for the case of constant flow, namely for all $\bar{u} \in \mathbb{R}$,

$$F(\bar{u}; j) = f(\bar{u}) \tag{2.6}$$

  Recall the concept of *Lipschitz continuous*: $|F(U^n; j) - f(\bar{u})| \leq K \max_{-p \leq i \leq q} |U_{j+i} - \bar{u}|$.

- **Discrete conservation**: The numerical flux on cell interface is single-valued. Therefore, we have

$$\sum_j U_j^{n+1} = \sum_j U_j^n \quad \text{for all } n. \tag{2.7}$$

- **Example**: Lax–Friedrichs method, see book page 125, equation (12.15) and (12.16).

Lax–Wendroff Theorem

Theorem 2.1 (Lax–Wendroff)**.** *Consider a sequence of grids indexed by* $\ell = 1, 2, \cdots$*, with mesh parameters* $k_\ell, h_\ell \to 0$ *as* $\ell \to \infty$*, let* $U_\ell(x,t)$ *denote the numerical approximation computed with a* ***consistent*** *and* ***conservative*** *method on the* ℓ^{th} *grid. Suppose* U_ℓ *converges to a function* u *as* $\ell \to \infty$*, in the sense that:*

1. *The* 1*-norm convergences:* $\|U_\ell - u\|_{1,\Omega} \to 0$ *as* $\ell \to \infty$*, where* $\Omega = [a,b] \times [0,T]$*.*

2. *Total variation bounded: there exists an* $R > 0$*, such that* $\mathrm{TV}(U_\ell(\cdot,t)) < R$*, for all* $0 \le t \le T$*,* $\ell = 1, 2, \cdots$*.*

Then, u *is a weak solution of the conservation law.*

Remark 2.1. *The Lax–Wendroff theorem does not guarantee that we do converge.*

Remark 2.2. *Even we have convergence, the Lax–Wendroff theorem does not guarantee that the weak solution obtained satisfy the entropy condition.*

Remark 2.3. *In case of a subsequence from a scheme convergences to physically correct solution (satisfies entropy condition), then the limit of this subsequence is a weak solution.*

Outline proof of Lax–Wendroff Theorem

- **Starting point**: The numerical scheme (2.4) for discussion.
- **Goal**: Show the limit function u satisfies (2.3).
- "**Roadmap**" and **motivation**:
  - Step 1. Analog the argument of deriving weak solution, multiply "test function" $\phi(x_j, t_n)$ on both side of the numerical scheme (2.4).

$$\text{integrate}: \int_0^{+\infty} \int_{-\infty}^{+\infty} \quad \to \quad \text{sum}: \sum_{n=0}^{+\infty} \sum_{j=-\infty}^{+\infty}$$

  - Step 2. Analog the argument of deriving weak solution, "integra-

tion by part" becomes "summation by part". Here use formulae

$$\sum_{n=0}^{m} a_n(b_{n+1}-b_n) = a_m b_{m+1} - a_0 b_0 + \sum_{n=0}^{m-1}(a_n - a_{n+1})b_{n+1}, \tag{2.8}$$

$$\sum_{j=-m}^{m} a_j(b_j - b_{j-1}) = a_m b_m - a_{-m} b_{-m-1} - \sum_{j=-m}^{m-1}(a_{j+1} - a_j)b_j. \tag{2.9}$$

Idea: original sum involves the product of a_j with differences of b's. Rewrite: final sum involves the product of b_j with differences of a's

- Step 3. Figuring out suitable conditions for our goal.
  - We will see how the conditions 1 and 2 are applied in the proof.
  - Review the concept of *the* 1*-norm convergences* and *total variation*, see book page 131.

- **Something to keep in mind**: the support of test function is compact, namely $\phi(x_j, t_n) = 0$ for $|j|$ or n sufficiently large.

More details of the proof for Lax–Wendroff Theorem

- Apply Step 1, we get:

$$\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\phi(x_j,t_n)(U_j^{n+1}-U_j^n)=-\frac{k}{h}\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\phi(x_j,t_n)\left(F(U^n;j)-F(U^n;j-1)\right). \tag{2.10}$$

- Apply Step 2 (note the support of ϕ is compact), we get:
  - Apply (2.8) to the left-hand side in (2.10) for "index n",

$$\begin{aligned}\text{LHS}&=\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\phi(x_j,t_n)(U_j^{n+1}-U_j^n)\\&=-\sum_{j=-\infty}^{+\infty}\phi(x_j,t_0)U_j^0-\sum_{j=-\infty}^{+\infty}\sum_{n=1}^{+\infty}\left(\phi(x_j,t_n)-\phi(x_j,t_{n-1})\right)U_j^n.\end{aligned} \tag{2.11}$$

  - Apply (2.9) to the right-hand side in (2.10) for "index j",

$$\begin{aligned}\text{RHS}&=-\frac{k}{h}\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\phi(x_j,t_n)\left(F(U^n;j)-F(U^n;j-1)\right)\\&=\frac{k}{h}\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\left(\phi(x_{j+1},t_n)-\phi(x_j,t_n)\right)F(U^n;j).\end{aligned} \tag{2.12}$$

  Therefore, substitute (2.11) and (2.12) into (2.10), we obtain

$$\begin{aligned}&-\sum_{j=-\infty}^{+\infty}\phi(x_j,t_0)U_j^0-\sum_{j=-\infty}^{+\infty}\sum_{n=1}^{+\infty}\left(\phi(x_j,t_n)-\phi(x_j,t_{n-1})\right)U_j^n\\&\qquad-\frac{k}{h}\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\left(\phi(x_{j+1},t_n)-\phi(x_j,t_n)\right)F(U^n;j)=0.\end{aligned} \tag{2.13}$$

  Multiply h on both side above and move the last two terms to the right side, we get

$$\underbrace{hk\sum_{n=1}^{+\infty}\sum_{j=-\infty}^{+\infty}\frac{\phi(x_j,t_n)-\phi(x_j,t_{n-1})}{k}U_j^n}_{=\mathrm{T}_1}$$
$$\underbrace{+hk\sum_{n=0}^{+\infty}\sum_{j=-\infty}^{+\infty}\frac{\phi(x_{j+1},t_n)-\phi(x_j,t_n)}{h}F(U^n;j)}_{=\mathrm{T}_2}=\underbrace{-\sum_{j=-\infty}^{+\infty}\phi(x_j,t_0)U_j^0}_{=\mathrm{T}_3}. \tag{2.14}$$

- Apply Step 3, take limit $\ell \to \infty$ $(k_\ell, h_\ell \to 0)$
  - The term T_1 and T_3 are handled by using condition 1.
  - The term T_2 is handled by using condition 2.

How to take limit?

- Review our **goal**: we want to obtain the following convergence, as $\ell \to \infty$ ($k_\ell, h_\ell \to 0$).

$$\mathrm{T}_1 \to \int_0^{+\infty} \int_{-\infty}^{+\infty} \phi_t u \,\mathrm{d}x\mathrm{d}t, \tag{2.15}$$

$$\mathrm{T}_2 \to \int_0^{+\infty} \int_{-\infty}^{+\infty} \phi_x f(u) \,\mathrm{d}x\mathrm{d}t, \tag{2.16}$$

$$\mathrm{T}_3 \to \int_{-\infty}^{+\infty} \phi(x,0)u(x,0)\mathrm{d}x. \tag{2.17}$$

- Notice, ϕ has compact support. For each ℓ, only finitely many terms in the sum of terms T_1, T_2, T_3 are non-zero. thus the sums are well-defined.

- Employ the notation $U_\ell(x_j, t_n)$ for piecewise constant function defined by U_j^n for a grid ℓ on $[x_{j-1/2}, x_{j+1/2}) \times [t_n, t_{n+1})$. The (2.18) is a Riemann sum of step functions, which can be written as

$$\underbrace{\int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi_\ell(x,t) - \phi_\ell(x,t-k)}{k} U_\ell(x,t) \,\mathrm{d}x\mathrm{d}t}_{=\mathrm{T}_1}$$
$$+\underbrace{\int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi_\ell(x+h,t) - \phi_\ell(x,t)}{h} F(U_\ell(x-ph,t), \cdots, U_\ell(x+qh,t)) \,\mathrm{d}x\mathrm{d}t}_{=\mathrm{T}_2}$$
$$= -\underbrace{\int_{-\infty}^{+\infty} \phi_\ell(x,0) U_\ell(x,0) \,\mathrm{d}x}_{=\mathrm{T}_3}. \tag{2.18}$$

- For the term T_1, in order to obtain (2.15), we employ condition "*the 1-norm convergences*".

  - Recall our goal is to show (2.15). Insert term "$u - u$" after U_ℓ, we get

$$\begin{aligned}\mathrm{T}_1 = &\int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi_\ell(x,t) - \phi_\ell(x,t-k)}{k} u(x,t) \,\mathrm{d}x\mathrm{d}t \\ &+ \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\phi_\ell(x,t) - \phi_\ell(x,t-k)}{k} \Big(U_\ell(x,t) - u(x,t)\Big) \,\mathrm{d}x\mathrm{d}t.\end{aligned} \tag{2.19}$$

  - By mean value theorem, $\exists \xi_{\ell,t} \in [t-k, t]$, such that

$$\phi_\ell(x,t) - \phi_\ell(x,t-k) = k\phi_t(x, \xi_{\ell,t}) \tag{2.20}$$

    Recall $\phi \in C_0^1(\mathbb{R} \times \mathbb{R}^+)$ with compact support and the definition of ϕ_ℓ.

– Use the condition "*the 1-norm convergences*", consider the support of ϕ_t is compact and $\phi_t \le \|\phi_t\|_{L^\infty}$. Take limit of the following expression

$$\begin{aligned} \mathrm{T}_1 = &\int_0^{+\infty}\int_{-\infty}^{+\infty} \phi_t(x,\xi_{\ell,t})\, u(x,t)\,\mathrm{d}x\mathrm{d}t \\ &+ \underbrace{\int_0^{+\infty}\int_{-\infty}^{+\infty} \phi_t(x,\xi_{\ell,t})\,\Big(U_\ell(x,t)-u(x,t)\Big)\,\mathrm{d}x\mathrm{d}t}_{\le\ \|\phi_t\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\,\|U_\ell-u\|_{1,\Omega}}. \end{aligned} \tag{2.21}$$

- For the term T_2, insert term "$f(U_\ell(x,t)) - f(U_\ell(x,t))$" after F, we have

$$\begin{aligned} \mathrm{T}_2 = &\int_0^{+\infty}\int_{-\infty}^{+\infty} \frac{\phi_\ell(x+h,t)-\phi_\ell(x,t)}{h}\, f(U_\ell(x,t))\,\mathrm{d}x\mathrm{d}t \\ +&\int_0^{+\infty}\int_{-\infty}^{+\infty} \frac{\phi_\ell(x+h,t)-\phi_\ell(x,t)}{h}\,\Big(F(U_\ell(x-ph,t),\cdots,U_\ell(x+qh,t))-f(U_\ell(x,t))\Big)\,\mathrm{d}x\mathrm{d}t \\ &= \mathrm{S}_1+\mathrm{S}_2 \end{aligned} \tag{2.22}$$

Let us show $\mathrm{S}_2 \to 0$ as $\ell \to \infty$.

– By mean value theorem, $\exists \xi_{\ell,x} \in [x, x+h]$, such that

$$\phi_\ell(x+h,t)-\phi_\ell(x,t) = h\phi_x(\xi_{\ell,x},t). \tag{2.23}$$

Recall $\phi \in C_0^1(\mathbb{R}\times\mathbb{R}^+)$ with compact support and the definition of ϕ_ℓ.

– Rewrite the term S_2 into the summation with respect to n and j. Note ϕ has compact support, we can assume $\phi = 0$ for all $t > T$.

$$\begin{aligned} |S_2| &\le \|\phi_x\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\int_0^T\int_{-\infty}^{+\infty} |F(U_\ell(x-ph,t),\cdots,U_\ell(x+qh,t))-f(U_\ell(x,t))|\,\mathrm{d}x\mathrm{d}t \\ &= \|\phi_x\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\; hk\sum_{n=0}^{T/k}\sum_{j=-\infty}^{+\infty} |F(U_\ell(x_{j-p},t_n),\cdots,U_\ell(x_{j+q},t_n))-f(U_\ell(x_j,t))| \end{aligned} \tag{2.24}$$

– Flux F is Lipschitz continuous, see LeVeque's book page 126 equation (12.19).

$$\begin{aligned} &|F(U_\ell(x_{j-p},t),\cdots,U_\ell(x_{j+q},t))-f(U_\ell(x_j,t))| \\ \le &K\max_{-p\le i\le q}|U_\ell(x_{j+i},t)-U_\ell(x_j,t)|. \end{aligned} \tag{2.25}$$

Substitute (2.25) into (2.24), notice the width of stencil is finite $p+q+1$, the term S_2 can be bounded by (trick: telescoping

summation)

$$|S_2| \le \|\phi_x\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\ hk \sum_{n=0}^{T/k} \sum_{j=-\infty}^{+\infty} K \max_{-p\le i\le q} |U_\ell(x_{j+i},t_n) - U_\ell(x_j,t_n)|$$
$$\le K\|\phi_x\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\ h \sum_{n=0}^{T/k} k\Big((p+q+1) \sum_{j=-\infty}^{+\infty} |U_\ell(x_j,t_n) - U_\ell(x_{j-1},t_n)|\Big). \tag{2.26}$$

– Recall U_ℓ is bounded total variation, see LeVeque's book page 131 equation (12.39 and 12.40). The term S_2 can be bounded by

$$|S_2| \le K(p+q+1)\|\phi_x\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\ h\Big(k \sum_{n=0}^{T/k} \mathrm{TV}(U_\ell(\cdot,t_n))\Big)$$
$$\le KRT(p+q+1)\|\phi_x\|_{L^\infty(\mathbb{R}^+\times\mathbb{R})}\ h \tag{2.27}$$

Therefore, $S_2 \to 0$ as $h \to 0$ ($\ell \to \infty$).

– *Example*: Lax–Friedrichs flux (stencil width 2).

- The rest terms S_1 and T_3 can be processed similarly.

Bibliography