# Finite Difference Schemes for PDEs

Helen K. Lei

September 22, 2007

## 1 Definitions and Results

We consider a partial differential equation (PDE)

$$Pu = f$$

and a finite difference scheme (FDS)

$$P_{k,h}v = f.$$

**Definition 1.1.** (Consistency) We say the FDS is consistent with the PDE if

$$P\phi - P_{k,h}\phi \to 0 \quad \text{as} \quad k, h \to 0,$$

for any smooth function $\phi(t, x)$ and where the convergence is pointwise convergence at each grid point.

**Definition 1.2.** (Stability) We say a FDS $P_{k,h}v_m^n = 0$ for a first order PDE (first order in time derivative) is stable in some stability region $\Lambda$ if

$$\exists J \in \mathbb{N} \cup \{0\}$$

such that

$$\forall T > 0 \quad \exists C_T$$

such that

$$h\|v^n\|_{\ell^2(\mathbb{Z})} \leq C_T \sum_{j=0}^{J} \|v^j\|_{\ell^2(\mathbb{Z})}$$

$$\text{for } 0 \leq nk \leq T \text{ and } \forall (h, k) \in \Lambda.$$

Stability region $\Lambda$ is some subset of the first quadrant of $\mathbb{R}^2$ which contains $(0, 0)$ as an accumulation point.

**Definition 1.3.** (Well–Posedness) We say a PDE $Pu = 0$ is well–posed if

$$\forall T \geq 0 \quad \exists C_T$$

such that any solution $u(t,x)$ satisfies

$$\|u(t,\cdot)\|_{L^2(\mathbb{R})} \leq C_T \|u(0,\cdot)\|_{L^2(\mathbb{R})},.$$

$\forall 0 \leq t \leq T$.

**Remark 1.4.** (Duhamel's Principle) Note that the previous two definitions are for the homogeneous initial value problem, but we have Duhamel's principle, which says that solution to an inhomogeneous initial value problem can be regarded as the superposition of solutions to homogeneous initial value problems. Thus well–posedness and stability estimates for the former follows from those for the latter. It is for this reason that when we talk about showing a scheme is stable, we always talk about the homogeneous case.

**Theorem 1.5.** *(Lax–Richtmyer Equivalence Theorem) A consistent FDS for a well–posed initial value problem for a PDE is convergent if and only if it is stable.*

## 2 Fourier Analysis, Stability and Well–Posedness

The solution of a FDS $P_{k,h}v = 0$ at a fixed time $n$, $\{v_m^n\}_{m=-\infty}^{\infty}$ is a function defined on $\mathbb{Z}$. Whereas the usual perspective of Fourier series is to start with a $\hat{f} : [-\pi,\pi] \to \mathbb{R}$ and Fourier transform to obtain coefficients $v_m$ such that $\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{m=-\infty}^{\infty} e^{-im\xi} v_m$, for the analysis of FDSs, we start with the grid function $v^n(m) = v_m^n$ and write the Fourier inversion formula (in the space variable)

$$v_m^n = \frac{1}{\sqrt{2\pi}} \int_{-\pi}^{\pi} e^{im\xi} \hat{v}^n(\xi) d\xi.$$

We now describe how to perform von Neumann analysis. For simplicity suppose we have an explicit one–step scheme, which we write as

$$v_m^{n+1} = \sum_{l=-B}^{F} \alpha_l(k,h) v_{m+l}^n.$$

Using the Fourier representation in space and using that the Fourier transform is unique, we learn that (after cancelling a factor of $e^{imh\xi}$)

$$\hat{v}^{n+1}(\xi) = \hat{v}^n(\xi) \sum_{k=-B}^{F} \alpha_l(k,h) e^{ilh\xi}.$$

Letting $\theta = h\xi$, we can write this as

$$\hat{v}^{n+1} = g(\theta, k, h)\hat{v}^n.$$

Iterating this, we learn that

$$\hat{v}^n = g^n \hat{v}^0.$$

Suppose for simplicity that $g$ is independent of $k$ and $h$ and $|g| \leq 1$, then using Parseval's relation (the first line in the following display),

$$\begin{aligned}
\sum_{m=-\infty}^{\infty} |v_m^n|^2 &= \int_{-\pi}^{\pi} |\hat{v}^n|^2 d\xi \\
&= \int_{-\pi}^{\pi} |g|^{2n} |\hat{v}^0|^2 d\xi \\
&\leq \int_{-\pi}^{\pi} |\hat{v}^0|^2 d\xi \\
&= \sum_{m=-\infty}^{\infty} |v_m^0|^2,
\end{aligned}$$

and so the scheme is stable.

More generally, we have the following stability condition

**Theorem 2.1.** *A one–step FDS is stable in stability region $\Lambda$ if and only if there is a constant $K$*

$$|g(\theta, k, h)| \leq 1 + Kk$$

*$\forall\theta$ and $\forall(h, k) \in \Lambda$.*

Note that $O(k)$ growth in $g$ is allowed in the stability condition (as long as the growth can be uniformly bounded in $\xi$), since the constant in the definition of stability is allowed to depend on $T$ and so we have, for $nk \leq T$, the inequality $(1 + Kk)^n \leq e^{Kkn} \leq e^{KT}$. On the other hand, if it is the the case that $|g(\theta, k, h)| \geq 1 + Ck$ for some constant $C$ on some interval $[\theta_1, \theta_2]$, then we can take $\hat{v}^0$ to be a step function supported on $[\theta_1, \theta_2]$ and we see that $\|v^n\|^2 \geq C^* \|v^0\|^2$.

---

**Remark 2.2.** We list some cases where algebra can be simplified in performing von Neumann analysis.

- Since $O(k)$ growth is the only growth allowed in $g$, when $g$ is independent of $h$ and $k$, the restricted condition

  $$|g(\theta)| \leq 1$$

  may be used.

- We have that $g(\theta, k, h) = g(\theta, 0, 0) + O(h) + O(k)$. When $\lambda = k/h$ is a constant, as is often the case in hyperbolic systems ($u_t = -au_x + f$), terms which are $O(h)$ are also $O(k)$ (uniformly in $\theta$) and in this case, the stability condition becomes

$$|g(\theta, 0, 0)| \leq 1.$$

- Sometimes the algebra can be simplified by discarding lower order terms which only contribute $O(k)$ to $g$:
  - For the hyperbolic equation $u_t + au_x + bu = 0$, the stability is not affected by the $bu$ term, since in a consistent scheme, it can only contribute $O(k)$ to $g$.
  - For a parabolic equation $u_t = bu_{xx} - au_x + cu$, the same statement can be made about the $cu$ term. The $au_x$ term is more tricky as it is likely to contribute $O(k/h)$ to $g$ and in parabolic cases it is usually $\mu = k/h^2$ which is held constant, but sometimes in the expression for $|g|^2$ (e.g. when the $au_x$ term contributes a purely imaginary part to $g$), the contribution from the $au_x$ term simply becomes $(k/h)^2 = k\mu$, which can then be discarded.

---

These ideas can be easily adapted to accommodate multi–step schemes, but the algebra becomes much more involved. Basically one uses the Fourier representation to derive a recurrence relation for $\hat{v}^n$. The associated characteristic polynomial, denote $\Phi(g, \theta)$, is called the amplification polynomial. The roots of the amplification polynomial then determines the solution of the recurrence relation. From the form of the solution to a linear recurrence relation, the following theorem is not surprising:

**Theorem 2.3.** *If the amplification polynomial $\Phi(g, \theta)$ is independent of $h$ and $k$, then the necessary and sufficient condition for the FDS to be stable is that all root, $g_\nu(\theta)$, satisfy*

*a. $|g_\nu(\theta)| \leq 1$ and*

*b. If $|g_\nu(\theta)| = 1$, then $g_\nu(\theta)$ must be a simple root.*

---

For systems of equations, e.g. of the form

$$u_t + Au_x = 0,$$

where $u$ is a vector of functions of $d$ dimensions and $A$ is a $d \times d$ matrix, basically the same stability analysis can be carried out, except now the scalar quantities will now be replaced by matrices. Again for simplicity let's consider an explicit one–step scheme

$$v_m^{n+1} = \sum_{l=-B}^{F} \mathfrak{A}_l(k, h) v_{m+l}^n.$$

In this equation $v_j^\ell$'s are $d$–dimensional vectors and $\mathfrak{A}_l(k,h)$ is a matrix whose entries depend on $k$ and $h$. Using the Fourier representation on each entry, we obtain as before

$$\hat{v}_m^{n+1} = \hat{v}^n \sum_{l=-B}^{F} \mathfrak{A}_l(k,h) e^{il\theta}.$$

So we now have instead an amplification matrix

$$G(\theta,k,h) = \sum_{l=-B}^{F} \mathfrak{A}_l(k,h) e^{il\theta},$$

where

$$e^{il\theta} = (e^{il\theta_1}, \dots e^{il\theta_d})^T.$$

So we now have

$$\hat{v}^n = G^n \hat{v}^0.$$

To introduce the stability condition, we need matrix norms.

**Definition 2.4.** Let $A : (\mathbb{C}^N, |\cdot|) \to (\mathbb{C}^M, |\cdot|)$ be a $M \times N$ matrix, then the matrix norm is

$$\|A\| = \sup_{|v|=1} |Av|,$$

where to avoid clutter we have not explicitly specified the norm on $\mathbb{C}^N$ and $\mathbb{C}^M$.

We point out two properties of the matrix norm:

- From the definition it follows that for any $v \in \mathbb{C}^N$,

$$|Av| = |v| A\left(\frac{v}{|v|}\right) \leq \|A\||v|.$$

- In addition to satisfying the usual properties of being a norm, the matrix norm satisfies

$$\|AB\| \leq \|A\|\|B\|.$$

From this we see that

$$|\hat{v}^n| = |G^n \hat{v}^0| \leq \|G^n\||\hat{v}^0|.$$

So the stability condition is then

**Theorem 2.5.** *A one–step FDS for a system of equations with amplification matrix $G(\theta, h, k)$ is stable in stability region $\Lambda$ if any only if*

$$\forall T > 0, \exists C_T,$$

*such that*

$$\|G^n\| \leq C_T,$$

*for all $0 \leq nk \leq T$ and $\forall h, k \in \Lambda$.*

At the risk of sounding pedantic, we remark that since $\|G^n\| \leq \|G\|^n$ requiring that $\|G\| \leq 1 + Kk$ as in the scalar case would be "overkill".

---

We have the following result on finding matrix norms.

**Proposition 2.6.** *If $A : (\mathbb{C}^M, |\cdot|_p) \to (\mathbb{C}^N, |\cdot|_p)$, then*

$$\|A\| = \max_{1 \leq j \leq M} \sum_{i=1}^{N} |a_{ij}|, \quad \text{if } p = 1$$

*and*

$$\|A\| = \max_{1 \leq i \leq N} \sum_{j=1}^{M} |a_{ij}|, \quad \text{if } p = \infty,$$

*that is, in the case of $\ell^1$, the matrix norm is the maximum absolute value sum of column entries, in the case of $\ell^\infty$, the maximum absolute value sum of row entries. Also*

$$\|A\| = \rho(A^* A)^{1/2}, \quad \text{if } p = 2,$$

*where $A^*$ is the conjugate of $A$ and $\rho(A)$ is the spectral radius, which denotes the maximum magnitude of the eigenvalues of $A$.*

To verify the third statement in the above requires the use of Schur's Lemma, which says that any square matrix is equivalent to an upper triangular matrix via a unitary matrix, i.e. given a square matrix $A$, there is some unitary $U$ such that

$$U^* A U = T,$$

where $T$ is upper triangular. A unitary matrix $U$ leaves the $\ell^2$ norm invariant, i.e.

$$|Uv|_2 = |v|_2$$

(so in particular $\|U\|_2 = 1$) and satisfies

$$U^* = U^{-1}.$$

Next we list some useful properties of upper triangular matrices: Let $T$ be an upper triangular matrix.

- If $T$ has zero diagonal entries, then $T$ is nilpotent, i.e. $T^n \to 0$ as $n \to \infty$. This can be used to show that e.g. $\rho(A) < 1$ implies $\lim_{n\to\infty} \|A\|^n = 0$.
- The determinant of $T$ is equal to the product of its diagonal entries.
- The diagonal entries of $T^2$ are the squares of the diagonal entries of $T$.

We notice that in particular if $A$ is symmetric, then $A = A^*$, so that $\|A\| = \rho(A^2)^{1/2}$. By the Spectral Mapping Theorem (or just notice that if $\lambda$ is an eigenvalue of $A$, then certainly $\lambda^2$ is an eigenvalue of $A^2$; to complete the argument, use Schur's Lemma. In fact the same observations can be used to prove the Theorem itself), $\rho(A^2) = \rho(A)^2$, so we have

**Proposition 2.7.** *If $A : \mathbb{C}^N \to \mathbb{C}^N$ is symmetric, then*

$$\|A\|_2 = \rho(A).$$

In general, the relationship between $\rho(A)$ and $\|A\|$ is summarized by the following:

**Proposition 2.8.** *Let $A : \mathbb{C}^N \to \mathbb{C}^N$. Then*

- $\rho(A) \leq \|A\|$.
- $\lim_{n\to\infty} \|A^n\| = 0$ *if and only if* $\rho(A) < 1$.

To prove the first item, note that if $\lambda$ is an eigenvalue of $A$ with associated non–zero eigenvector $v$, then $|\lambda||v| = |Av| \leq \|A\||v|$. One direction of the second item follows from the relation $\rho(A^n) = \rho(A)^n$, which is a consequence of the Spectral Mapping Theorem, and the other direction follows from Schur's Lemma.

---

Returning to our discussion of stability, we make the following observations:

- The previous proposition implies that the condition

  $$|g_\nu| \leq 1 + Kk$$

  is necessary (but not sufficient) for stability. Indeed, if $\exists g_\nu$ such that $g_\nu > 1 + Ck$ for arbitrary $C$, then

  $$\|G^n\| \geq \rho(G^n) = \rho(G)^n \geq g_\nu^n > (1 + Ck)^n,$$

  and we see that $\|G^n\|$ cannot be bounded uniformly for $0 \leq nk \leq T$ since $C$ can be arbitrarily large.

- If $U$ is the matrix given by Schur's Lemma such that

$$U^*GU = T,$$

with $T$ triangular, and we can bound $T$, then we can also bound $G$, since $G^n = UT^nU^{-1}$,

$$\|G^n\| \leq \|U\|\|U^{-1}\|\|T^n\|.$$

Incidentally, given a non–singular matrix $A$, the quantity

$$K(A) = \|A\|\|A^{-1}\|$$

is called the condition number of the matrix $A$. Notice that $K(A) \geq 1$. It measures the accuracy of the approximate solution to $Ax = b$ given residual vector $r = b - A\tilde{x}$.

---

Diagonalizability merits some discussion. For a hyperbolic system, the matrix $A$ can always be diagonalized, and the system

$$u_t + Au_x = 0$$

decouples into $d$ equations. That is, $\exists P$ a change of basis matrix such that

$$A = P^{-1}DP,$$

where $D$ is diagonal. Note that the columns of $P$ are the eigenvectors of $A$, so $P$ sends the standard basis to the eigenbasis. Therefore multiplying by $P^{-1}$ will change from standard coordinates to eigencoordinates, so with $P^{-1}u = w$,

$$Pw_t + APw_x = 0, \quad \text{so} \quad w_t + Dw_x = 0,$$

and we obtain the $d$ scalar equations

$$w_t + a_i w_x = 0, \quad i = 1, \ldots, d,$$

where $a_i$ are now the eigenvalues of $A$.

One thing that one can do is to first decouple the equations and then apply scalar FDS to each equation, and analyze stability of each equation separately. The stability conditions will then involve the eigenvalues $a_i$. On the other hand, the procedure we described previously applies a FDS directly to the system, using the matrix $A$ to take the place of the scalar quantity $a$ in the scalar equation $u_t + au_x = 0$. If in this case the amplification matrix $G$ turns out to be a rational function of $A$, then the same $P$ which diagonalizes $A$ also diagonalizes $G$:

$$PA^nP^{-1} = (PAP^{-1})^n,$$

and

$$D^{-1} = PA^{-1}P^{-1}.$$

So if the FDS in question gives the amplification factor as $f(a)$, where $f$ is a rational function, then

$$P^{-1}GP = P^{-1}f(A)P = f(P^{-1}AP) = f(D),$$

and the (diagonal) entries of $f(D)$ are exactly $f(a_i)$, where $a_i$ are the eigenvalues of $A$. By the uniqueness of the diagonal representation, the two operations commute: whether we first decouple and then apply FDS or directly apply FDS to system and then diagonalize $G$ to analyze stability, the resulting diagonal form of $G$ will be the same. The same comments apply to parabolic systems: The same matrix which puts $B$ into upper triangular form also puts $G$ into upper triangular form if the scheme gives $G$ as a rational function of $B$.

For a higher–dimensional system of the form

$$u_t + Au_x + Bu_y = 0,$$

the same remarks apply, if we assume $A$ and $B$ are simultaneously diagonalizable.

---

We now describe how to use the Fourier transform to determine when an initial value problem is well–posed. Given a function $u : \mathbb{R} \to \mathbb{R}$, its Fourier transform is given by

$$\hat{u}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-i\xi x} u(x)\ dx.$$

The Fourier inversion formula is given by

$$u(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi x} \hat{u}(\xi)\ d\xi.$$

Suppose now that $u(t, x) : \mathbb{R}^+ \times \mathbb{R}$ and the Fourier transform has been taken in space, then differentiating the inversion formula, we learn that

$$\frac{\partial u}{\partial x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi x} (i\xi)\hat{u}(\xi)\ d\xi,$$

so

$$\left(\widehat{\frac{\partial u}{\partial x}}\right)(\xi) = (i\xi)\hat{u}(x).$$

We will also greatly take advantage of Parseval's relation

$$\|u\|_{L^2_x} = \|\hat{u}\|_{L^2_\xi}.$$

First let's consider a first–order equation (i.e. only one time derivative). After Fourier transform and using the derivative property, any first–order equation looks like

$$\hat{u}_t(t,\xi) = q(\xi)\hat{u}(t,\xi),$$

where $q$ is a (complex) polynomial. The solution is then

$$\hat{u}(t,\xi) = e^{q(\xi)t}\hat{u}_0(\xi).$$

By Parseval's relation, we then have the following condition for well–posedness:

**Theorem 2.9.** *A necessary and sufficient condition for a first–order equation to be well–posed is that there exists some $\overline{q}$ such that for all real values of $\xi$*

$$\Re e\ q(\xi) \leq \overline{q}.$$

The condition is necessary since if it does not hold then one may construct $\hat{u}_0(\xi)$ to be a step function supported on the interval where $\Re e\ q(\xi) > C$ for some constant $C$, which then quickly leads to the inequality $\|u(t,\cdot\| > C^*\|u_0\|$.

# 3 Symbols, Consistency and Order of Accuracy

For consistency the condition $P\phi - P_{k,h}\phi \to 0$ can usually be checked by Taylor expansions. Order of accuracy is a more delicate question. We begin with a definition.

**Definition 3.1.** A FDS $P_{k,h}v = R_{k,h}f$ that is consistent with $Pu = f$ is accurate of order $p$ in time and order $q$ in space if for any smooth function $\phi(t,x)$

$$P_{k,h}\phi - R_{k,h}P\phi = O(k^p) + O(k^q).$$

We then say the scheme is accurate of order $(p,q)$. The quantity $P_{k,h}\phi - R_{k,h}P\phi$ is called the truncation error of the scheme.

For schemes whose truncation error cannot be written as $O(k^p) + O(k^q)$, we modify the definition as follows: If we can write $k = \Lambda(h)$, where $\Lambda$ is smooth and $\Lambda(0) = 0$, then the scheme is accurate of order $r$ if

$$P_{k,h}\phi - R_{k,h}P\phi = O(h^r).$$

Note that

$$P_{k,h}\phi - R_{k,h}P\phi = (P_{k,h}\phi - P\phi) + (I - R_{k,h})P\phi,$$

where $I$ is the identity operator. So we see that consistency implies $R_{k,h}$ is an approximation to the identity operator.

Observe that we can think of the truncation error as a measurement of how much a solution of the PDE deviates from the FDS: Since $Pu = f$, $P_{k,h}u - R_{k,h}f = P_{k,h}u - R_{k,h}Pu$. However, $f$ does not explicitly appear in checking for accuracy: The only reason for the expression $R_{k,h}f$ is that it so happens that $f$ satisfies $Pu = f$. The only thing of interest is what the operator $R_{k,h}$ *does*, and therefore any smooth $\phi$ works just as well in checking accuracy, as long as we replace the placeholder $f$ by $P\phi$. Of course the same comment can already be made about the simpler definition of consistency.

---

The smooth function which is most useful in checking for accuracy is

$$\phi(t,x) = e^{st}e^{i\xi x}.$$

We remark that $\phi(t,x)$ is also the function that we integrate a function $\hat{u}(s,\xi)$ against if we wish to inverse Laplace transform in time and inverse Fourier transform in space. The key property we will use is that

$$\frac{\partial^{l+j}\phi}{\partial t^l \partial x^j} = s^l(i\xi)^j\phi.$$

This leads naturally to the following definition.

**Definition 3.2.** The symbol $p(s,\xi)$ of a differential operator $P$ is defined by

$$P(e^{st}e^{i\xi x}) = p(s,\xi)e^{st}e^{i\xi x}.$$

That is, the symbol is the complex polynomial in $s$ and $\xi$ in front of $e^{st}e^{i\xi x}$ in $P(e^{st}e^{i\xi x})$. For a difference operator $P_{k,h}$, we have a similar definition: We set the grid points to be $t = nk$, $x = mh$, and we obtain

$$P_{k,h}(e^{snk}e^{i\xi mh}) = p_{k,h}(s,\xi)e^{snk}e^{i\xi mh}.$$

So at first glance, $p_{k,h}(s,\xi)$ is a polynomial in $e^{sk}$ and $e^{i\xi h}$, with coefficients which are rational expressions of $h$ and $k$. But for the purposes of determining accuracy, we shall actually take the opposite perspective: $p_{k,h}(s,\xi)$ is a rational expression of $h$ and $k$ with coefficients which are polynomials in $e^{sk}$ and $e^{i\xi h}$.

Using symbols, we can more systematically check the order of accuracy of a scheme. If a scheme $P_{k,h}v = R_{k,h}f$ is accurate of order $(p,q)$, then taking $\phi(t,x) = e^{st}e^{i\xi x}$, we have

$$[p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi)]e^{snk}e^{i\xi mh} = P_{k,h}\phi - R_{k,h}P\phi = O(k^p) + O(h^q),$$

so that

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = O(k^p) + O(h^q),$$

since $e^{snk}e^{i\xi mh} = O(1)$ as $k, h \to 0$. Conversely, if we have a consistent scheme, then the Taylor expansion of $p_{k,h}(s,\xi)$ (expand the $e^{sk}$ and $e^{i\xi h}$ terms in $p_{k,h}(s,\xi)$) cannot contain negative powers of $h$ or $k$ due to the condition $P_{k,h}\phi - P\phi \to 0$ as $k, h \to 0$. So if we find

$$p_{k,h}(s,\xi) = \sum_{l,j\geq 0} A_{l,j}(k,h)s^l(i\xi)^j,$$

then

$$P_{k,h}(e^{st}e^{i\xi x}) = p_{k,h}(s,\xi)(e^{st}e^{i\xi x}) = (e^{st}e^{i\xi x})\sum_{l,j\geq 0} A_{l,j}(k,h)s^l(i\xi)^j = \left[\sum_{l,j\geq 0} A_{l,j}(k,h)\frac{\partial^{l+j}}{\partial t^l \partial x^j}\right](e^{st}e^{i\xi x}).$$

Linear independence of the set $\{s^l(i\xi)^j\}_{l,j\geq 0}$ implies that

$$P_{k,h} = \sum_{l,j\geq 0} A_{l,j}(k,h)\frac{\partial^{l+j}}{\partial t^l \partial x^j}.$$

Replacing $e^{st}e^{i\xi x}$ now by an arbitrary $\phi$, we learn the coefficient of $A_{l,j}(k,h)$ is zero in the expression $P_{k,h}\phi - R_{k,h}P\phi$ if and only if the coefficient is zero in the expression $p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi)$ (for all $s$ and $\xi$). The consistency condition can also be used to show that

$$r_{k,h}(s,\xi) = 1 + o(1),$$

and we have

**Theorem 3.3.** *A scheme $P_{k,h}v = R_{k,h}f$ consistent with $Pu = f$ is accurate of order $(p,q)$ if and only if for all $s$ and $\xi$,*

$$\frac{p_{k,h}(s,\xi)}{r_{k,h}(s,\xi)} - p(s,\xi) = O(k^p) + O(h^q).$$

In practice we use the condition

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = O(k^p) + O(h^q).$$

For schemes with $k = \Lambda(h)$, the above is modified to be

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = O(h^r).$$

**Remark 3.4.** We can use symbols to find the actual truncation error of the scheme as well. Suppose

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = \sum_{l,j\geq 0} A_{l,j}(k,h)s^l(i\xi)^j,$$

then by the same argument as above

$$[P_{k,h} - R_{k,h}P]\phi = \sum_{l,j\geq 0} A_{l,j}(k,h)\frac{\partial^{l+j}\phi}{\partial t^l \partial x^j}.$$

---

In the homogeneous case $Pu = 0$, order of accuracy can be determined without knowledge of $R_{k,h}$. This can be done in two ways:

- We may directly show that $P_{k,h}\phi = O(k^p) + O(h^q)$ for each *formal* solution $P\phi = 0$.
- We may consider this to be the case $Pu = f$ with $f = 0$. If we can find $R_{k,h}$ with symbol $r_{k,h}$ such that

  $$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = O(k^p) + O(h^q),$$

  then the scheme is certainly accurate of order $(p,q)$.

# 4 Hyperbolic

The typical homogeneous equation we consider is

$$u_t + au_x = 0,$$

with initial condition

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

The solution is found to be

$$u(t,x) = u_0(x - at).$$

We will also consider the inhomogeneous equation

$$u_t + au_x = f.$$

Systems of hyperbolic equations take the form

$$u_t + Au_x = 0,$$

where $u$ is a vector of functions of dimension $d$ and $A$ is a $d \times d$ matrix which is diagonalizable with real eigenvalues.

To facilitate stability analysis we will let $\lambda = k/h$ and $\theta = h\xi$.

### 4.1 Some Basic Schemes

These schemes are of the form $P_{k,h}v = 0$ and are derived by replacing the derivative operators by various finite difference approximations. The consistency can all be shown by Taylor expanding around $(t_n, x_m)$.

- Forward–time forward space (explicit, one–step, order (1, 1), stable if $-1 \le a\lambda \le 0$):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_m^n}{h} = 0$$

We rewrite the scheme as

$$\begin{aligned} v_m^{n+1} &= v_m^n - a\lambda(v_{m+1}^n - v_m^n)) \\ &= (1 + a\lambda)v_m^n - a\lambda v_{m+1}^n. \end{aligned}$$

The amplification factor is

$$\begin{aligned} g(\theta) &= 1 - a\lambda(e^{i\theta} - 1) \\ &= 1 - a\lambda(\cos\theta - 1) + ia\lambda\sin\theta. \end{aligned}$$

$\cos\theta = 1 - 2\sin^2\frac{\theta}{2}$, so $\cos\theta - 1 = -2\sin^2\frac{\theta}{2}$. So

$$g(\theta) = 1 + a\lambda\sin^2\frac{\theta}{2} + ia\lambda\sin\theta$$

and

$$\begin{aligned} |g(\theta)|^2 &= (1 + a\lambda\sin^2\frac{\theta}{2})^2 + (a\lambda\sin\theta)^2 \\ &= 1 + 4a\lambda\sin^2\frac{\theta}{2} + a^2\lambda^2 sin^4\frac{\theta}{2} + 4a^2\lambda^2\sin^2\frac{\theta}{2}\cos^2\frac{\theta}{2}, \\ &= 1 + 4a\lambda\sin^2\frac{\theta}{2}(1 + \cos^2\frac{\theta}{2} + \frac{1}{4}\sin^2\frac{\theta}{2}) \end{aligned}$$

and we see that $|g(\theta)| \le 1$ necessitates that $-1 \le a\lambda \le 0$.

---

In this case it is actually easier to use the definition of stability directly. Let $\alpha = 1 + a\lambda$ and $\beta = a\lambda$. Then we have

$$v_m^{n+1} = \alpha v_m^n - \beta v_{m+1}^n.$$

Now let's look at $\|v_m^{n+1}\|_{\ell^2(\mathbb{Z})}$.

$$\begin{aligned}\sum_m |v_m^{n+1}|^2 &= \sum_m |\alpha v_m^n - \beta v_{m+1}^n|^2 \\ &\leq (|\alpha| + |\beta|)^2 \sum_m |v_m^n|^2,\end{aligned}$$

where we have used $2xy \leq x^2 + y^2$ and a minor rearrangement of terms (all terms are positive). So for stability we need

$$|\alpha| + |\beta| \leq 1.$$

Hence this scheme is stable if $|1 + a\lambda| + |a\lambda| \leq 1$, i.e. $-1 \leq a\lambda \leq 0$.

- Forward–time backward–space (explicit, one–step, order (1, 1), stable if $0 \leq a \leq 1$):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = 0$$

We rewrite the scheme as

$$\begin{aligned}v_m^{n+1} &= v_m^n - a\lambda(v_m^n - v_{m-1}^n) \\ &= (1 - a\lambda)v_m^n + a\lambda v_{m-1}^n.\end{aligned}$$

The amplification factor is then

$$\begin{aligned}g(\theta) &= 1 - a\lambda + a\lambda e^{-i\theta} \\ &= 1 + a\lambda(\cos\theta - 1) - ia\lambda\sin\theta,\end{aligned}$$

which is exactly the same as that found for the Forward–time forward–space scheme modulo a sign change of $a$. Again here it is better to use the definition of stability directly. Exactly the same analysis as in the previous item yields that the scheme is stable if $0 \leq a \leq 1$.

- Forward–time central–space (explicit, one–step, order (1, 2), unstable):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$$

Rewrite the scheme as

$$v_m^{n+1} = v_m^n - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n).$$

The amplification factor is then

$$\begin{aligned}g(\theta) &= 1 - \frac{a\lambda}{2}(e^{i\theta} - e^{-i\theta}) \\ &= 1 - ia\lambda\sin\theta,\end{aligned}$$

so $|g(\theta)| > 1$ unless $\theta = 0, \pi$, so this scheme is unstable.

- Lax–Friedrichs (explicit, one–step, order $O(k^{-1}h^2)+O(k)+O(h^2) = O(h)$ for $k = \lambda h$, stable if $|a\lambda| \leq 1$):

$$\frac{v_m^{n+1} - \frac{1}{2}(v_{m+1}^n + v_{m-1}^n)}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$$

First we demonstrate the scheme is consistent. For the spatial term, we have that (expanding around $(t_n, x_m)$),

$$\phi_{m+1}^n = \phi_m^n + h\phi_x + \frac{1}{2}h^2\phi_{xx} + O(h^3),$$

and

$$\phi_{m-1}^n = \phi_m^n - h\phi_x + \frac{1}{2}h^2\phi_{xx} + O(h^3),$$

so

$$\frac{\phi_{m+1}^n - \phi_{m-1}^n}{2h} = \phi_x + O(h^2).$$

For the time term, note that we also have

$$\phi_{m+1}^n + \phi_{m-1}^n = 2\phi_m^n + h^2\phi_{xx},$$

so that

$$\frac{1}{2}(\phi_{m+1}^n + \phi_{m-1}^n) = \phi_m^n + \frac{1}{2}h^2\phi_{xx} = \phi_m^n + O(h^2).$$

Substituting these expressions in, we learn that

$$P_{k,h}\phi - P\phi = O(k^{-1}h^2) + O(k) + O(h^2),$$

so the Lax–Friedrichs scheme is consistent as long as $k^{-1}h^2 \to 0$.

---

For stability, let's rewrite the scheme as

$$\begin{aligned} v_m^{n+1} &= \frac{1}{2}(v_{m+1}^n + v_{m-1}^n) - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n) \\ &= \frac{1}{2}[(1 - a\lambda)v_{m+1}^n + (1 + a\lambda)v_{m-1}^n]. \end{aligned}$$

The amplification factor is then

$$\begin{aligned} g(\theta) &= \frac{1}{2}(1 - a\lambda)e^{i\theta} + (1 + a\lambda)e^{-i\theta} \\ &= \cos\theta - ia\lambda\sin\theta. \end{aligned}$$

The condition $|g(\theta)| \leq 1$ then implies that we need $|a\lambda| \leq 1$.

Having considered a few explicit schemes, we should mention that we in fact have the following result:

**Theorem 4.1.** *For an explicit scheme for the hyperbolic equation of the form* $v_m^{n+1} = \alpha v_{m-1}^n + \beta v_m^n + \gamma v_{m+1}^n$ *with* $\lambda$ *held constant, a necessary condition for stability is the Courant–Friedrichs–Lewy (CFL) condition,*

$$|a\lambda| \leq 1.$$

In addition,

**Theorem 4.2.** *There are no explicit and consistent finite difference scheme for hyperbolic systems of partial differential equations which is stable for all values of* $\lambda$.

This theorem certainly does not apply to implicit schemes, as the next 3 schemes show.

- Backward–time forward–space (implicit, one–step, order (1, 1), stable if $a \leq 0$):

$$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_m^n}{h} = 0$$

Rewrite the scheme as

$$v_m^n = v_m^{n-1} - a\lambda(v_{m+1}^n - v_m^n).$$

So

$$(1 - a\lambda)v_m^n + a\lambda v_{m+1}^n = v_m^{n-1}.$$

The amplification factor is then

$$\begin{aligned} g(\theta) &= [(1 - a\lambda) + a\lambda e^{i\theta}]^{-1} \\ &= \frac{1}{1 + a\lambda(\cos\theta - 1) + ia\lambda\sin\theta}. \end{aligned}$$

We need $|g(\theta)| \geq 1$, i.e.

$$\begin{aligned} ((1 - a\lambda) + a\lambda\cos\theta)^2 + (a\lambda\sin\theta)^2 &= (1 - a\lambda)^2 + 2a\lambda(1 - a\lambda)\cos\theta + (a\lambda\cos\theta)^2 + (a\lambda\sin\theta)^2 \\ &= (1 - a\lambda)^2 + (a\lambda)^2 + 2a\lambda(1 - a\lambda)\cos\theta \\ &\geq 1. \end{aligned}$$

The above can be minimized at $\theta = 0, \pi$. For $\theta = 0$, we get $(1-a\lambda)^2+(a\lambda)^2+2a\lambda(1-a\lambda) = (1 - a\lambda + a\lambda)^2 = 1$. For $\theta = \pi$, we get $(1 - 2a\lambda)^2$ which is great than or equal to 1 only when $a \leq 0$. So the scheme is stable if $a \leq 0$.

- Backward–time backward–space (implicit, one–step, order (1, 1), stable if $a \geq 0$):

$$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = 0$$

Rewrite the scheme as

$$v_m^n = v_m^{n-1} - a\lambda(v_m^n - v_{m-1}^n).$$

So

$$(1 + a\lambda)v_m^n - a\lambda v_{m-1}^n = v_m^{n-1}.$$

The analysis of this is the same as in the previous item with a change of sign in $a$, so the stability condition is $a \geq 0$.

- Backward–time central–space (implicit, one–step, order (1, 2), unconditionally stable):

$$\frac{v_m^n - v_m^{n-1}}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$$

Rewrite the scheme as

$$v_m^n = v_m^{n-1} - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n).$$

So

$$v_m^n + \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n) = v_m^{n-1}.$$

The amplification factor is then

$$\begin{aligned} g(\theta) &= [1 + \frac{a\lambda}{2}(e^{i\theta} - e^{-i\theta})]^{-1} \\ &= \frac{1}{1 + ia\lambda \sin\theta}. \end{aligned}$$

So $|g(\theta)| \leq 1$ for all $a\lambda$, so the scheme is unconditionally stable.

- Leapfrog (explicit, multi–step, order (2, 2), stable if $|a\lambda| < 1$):

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = 0$$

Rewrite the scheme as

$$v_m^{n+1} = v_m^{n-1} - a\lambda(v_{m+1}^n - v_{m-1}^n).$$

Using the Fourier representation, we get

$$\hat{v}^{n+1} = \hat{v}^{n-1} - 2a\lambda i \sin\theta \hat{v}^n.$$

The amplification polynomial is then

$$\Phi(g, h, k) = g^2 + (2a\lambda i \sin\theta)g - 1.$$

The roots of this equation are given by the quadratic formula to be

$$g_\pm = -ia\lambda \sin\theta \pm \sqrt{1 - (a\lambda \sin\theta)^2}.$$

The cases $g_+ = g_-$ and $g_+ \neq g_-$ must be considered separately.

If $g_+ \neq g_-$, then the solution to the recurrence relation is given by

$$\hat{v}^n(\xi) = A(\xi)g_+^n + B(\xi)g_-^n,$$

where $A(\xi)$ and $B(\xi)$ are determined by initial conditions to be

$$A = \frac{\hat{v}^0 g_- - \hat{v}^1}{g_- - g_+}, \quad B = \frac{\hat{v}^1 - \hat{v}^0 g_+}{g_- - g_+}.$$

We may choose initial conditions to make either $A = 0$ or $B = 0$, which implies that a necessary condition for the stability of this scheme is that $g_+$ and $g_-$ must separately satisfy the stability condition, i.e. , we need

$$|g_\pm| \leq 1,$$

where we have taken $\lambda$ to be constant and have employed the restricted condition. Of course this condition is also sufficient, since if $|g_\pm| \leq 1$, then $|Ag_+^n + Bg_-^n| \leq 2M$, where $M = \max\{|A|, |B|\}$. For $a\lambda \leq 1$, the discriminant is positive and we have

$$|g_\pm|^2 = (a\lambda \sin\theta)^2 + 1 - (a\lambda \sin\theta)^2 = 1,$$

whereas for $a\lambda > 1$, at $\theta = \pi/2$, we have $|g_-| > 1$. So in the case that $g_+ \neq g_-$, the stability condition is $a\lambda \leq 1$.

We must now consider the case where $g_- = g_+$. We see that this can only happen when $(a\lambda \sin\theta)^2 = 1$, which implies $a\lambda = 1$ and $\theta = \pm\pi/2$. In this case $g_\pm = -i$ and the solution to the recurrence relation is (recall $\xi = \theta/h$)

$$\hat{v}^n(\pm(\pi/2h)) = A(\pm(\pi/2h))(\pm i)^n + B(\pm(\pi/2h))n(\pm i)^{n-1}.$$

We see that $\hat{v}^n$ behaves very badly around $\theta = \pi/2$, i.e. exhibits linear growth in $n$, so the scheme is unstable if $a\lambda = 1$. We conclude that the necessary and sufficient condition for the scheme to be stable is $|a\lambda| < 1$.

### 4.2 Order (2, 2) Schemes: Lax–Wendroff and Crank–Nicolson

In this section we study three higher order schemes of the form $P_{k,h}v = R_{k,h}f$ for the hyperbolic equation. The inhomogenous term $f(t,x)$ is incorporated into these schemes.

- Lax–Wendroff (Explicit, one–step, order (2, 2), stable if $|a\lambda| \leq 1$ ):

  $$\frac{v_m^{n+1}-v_m^n}{k}+a\frac{v_{m+1}^n-v_{m-1}^n}{2h}-\frac{a^2k}{2}\frac{(v_{m+1}^n-2v_m^n+v_{m-1}^n)}{h^2}=\frac{1}{2}(f_m^{n+1}+f_m^n)-\frac{ak}{4h}(f_{m+1}^n-f_{m-1}^n)$$

  To derive the Lax–Wendroff scheme, we start as usual with the Taylor expansion in $t$ to obtain

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

  We will then replace the time derivatives on the right hand side by space derivatives according to the differential equation. We have

  $$u_t=-au_x+f$$

  and

  $$u_{tt}=-a(u_x)_t+f_t=-a(u_t)_x+f_t.$$

  Now differentiation the $u_t$ term with respect to $x$ and using the differential equation again, we get

  $$u_{tt}=a^2u_{xx}-af_x+f_t.$$

  Plugging this back into the Taylor expansion and replacing the $x$ derivatives by second–order accurate difference and $f_t$ by a forward difference, we have

  $$\begin{aligned}u(t+k,x)&=u(t,x)-aku_x(t,x)+kf(t,x)+\frac{a^2k^2}{2}u_{xx}(t,x)-\frac{ak^2}{2}f_x(t,x)+\frac{k^2}{2}f_t(t,x)+O(k^3)\\&=u(t,x)-aku_x(t,x)+\frac{a^2k^2}{2}u_{xx}(t,x)-\frac{ak^2}{2}f_x(t,x)+kf(t,x)+\frac{k^2}{2}f_t(t,x)+O(k^3).\end{aligned}$$

  The $x$ derivatives we will replace with second–order accurate differences. For the last two terms we note that

  $$\begin{aligned}kf(t,x)+\frac{k^2}{2}f_t(t,x)&=kf(t,x)+\frac{k^2}{2}\frac{[f(t+k,x)-f(t,x)]}{k}+O(k^3)\\&=\frac{k}{2}[f(t+k,x)+f(t,x)]+O(k^3).\end{aligned}$$

  Making all these replacements, we get

  $$\begin{aligned}u(t+k,x)=\;&u(t,x)-ak\frac{u(t,x+h)-u(t,x-h)}{2h}+\frac{a^2k^2}{2}\frac{u(t,x+h)-2u(t,x)+u(t,x-h)}{h^2}\\&-\frac{ak^2}{2}\frac{[f(t,x+h)-f(t,x-h)]}{2h}+\frac{k}{2}[f(t+k,x)+f(t,x)]\\&+O(kh^2)+O(k^3).\end{aligned}$$

---

To perform stability analysis, we consider the homogeneous case $u_t + au_x = 0$. Then scheme is

$$\begin{aligned} v_m^{n+1} &= v_m^n - \frac{a\lambda}{2}(v_{m+1}^n - v_{m-1}^n) + \frac{a^2\lambda^2}{2}(v_{m+1}^n - 2v_m^n + v_{m-1}^n) \\ &= v_m^n(1 - a^2\lambda^2) + \frac{a\lambda}{2}(a\lambda - 1)v_{m+1}^n + \frac{a\lambda}{2}(a\lambda + 1)v_{m-1}^n. \end{aligned}$$

So the amplification factor is

$$\begin{aligned} g(\theta) &= (1 - a^2\lambda^2) + \frac{a\lambda}{2}(a\lambda - 1)e^{i\theta} + \frac{a\lambda}{2}(a\lambda + 1)e^{-i\theta} \\ &= (1 - a^2\lambda^2) + a^2\lambda^2\cos\theta - ia\lambda\sin\theta \\ &= 1 - 2a^2\lambda^2\sin^2\frac{1}{2}\theta - i2a\lambda\sin\frac{1}{2}\theta\cos\frac{1}{2}\theta. \end{aligned}$$

So

$$\begin{aligned} |g(\theta)|^2 &= (1 - 2a^2\lambda^2\sin^2\frac{1}{2}\theta)^2 + 4a^2\lambda^2\sin^2\frac{1}{2}\theta\cos^2\frac{1}{2}\theta \\ &= 1 - 4a^2\lambda^2(1 - \cos^2\frac{1}{2}\theta) + 4a^2\lambda^4\sin^4\frac{1}{2}\theta \\ &= 1 - 4a^2\lambda^2(1 - a^2\lambda^2)\sin^2\frac{1}{2}\theta. \end{aligned}$$

So we see that the scheme is stable if and only if $|a\lambda| \le 1$.

---

Next we show that the scheme is accurate of order $(2, 2)$ using symbols. We have

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{e^{sk} - 1}{k} + \frac{a}{2h}(e^{i\xi h} - e^{-i\xi h}) - \frac{a^2 k}{2h^2}(e^{i\xi h} - 2 + e^{-i\xi h}) \\ r_{k,h}(s,\xi) &= \frac{1}{2}(e^{sk} + 1) - \frac{ak}{4h}(e^{i\xi h} - e^{-i\xi h}) \\ p(s,\xi) &= s + ai\xi \end{aligned}$$

First,

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{k}(e^{sk} - 1) + \frac{a}{h}i\sin(h\xi) - \frac{a^2 k}{h^2}(\cos(h\xi) - 1) \\ &= (s + \frac{s^2 k}{2} + O(k^2)) + ai(\xi - \frac{h^2\xi^3}{3!} + O(h^4)) - a^2 k(-\xi^2 + \frac{h^2\xi^4}{4!} + O(h^4)) \\ &= s + \frac{s^2 k}{2} + ai\xi - \frac{ai}{6}h^2\xi^3 + a^2\xi^2 k - \frac{a^2}{24}\xi^4 k h^2 + O(k^2) + O(h^4) \\ &= s + ai\xi + \frac{s^2 k}{2} + a^2\xi^2 k + O(k^2) + O(h^2). \end{aligned}$$

Next,

$$\begin{aligned} r_{k,h}(s,\xi) &= 1 + \frac{sk}{2} + \frac{(sk)^2}{4} + O(k^3) - \frac{ak}{2h} i \sin(h\xi) \\ &= 1 + \frac{sk}{2} + \frac{(sk)^2}{4} - \frac{ak}{2} i(\xi - \frac{h^2\xi^3}{6}) + O(k^3) + O(h^4) \\ &= 1 + \frac{sk}{2} - \frac{aik\xi}{2} + O(k^2) + O(kh^2). \end{aligned}$$

So

$$r_{k,h}(s,\xi)p(s,\xi) = s + ai\xi + \frac{s^2k}{2} + \frac{ai\xi sk}{2} - \frac{aik\xi s}{2} + \frac{a^2\xi^2 k}{2} + O(k^2) + O(kh^2).$$

We see that

$$p_{k,h}(s,\xi) - r_{k,h}(s,\xi)p(s,\xi) = O(k^2) + O(h^2).$$

- Crank–Nicolson (Implicit, one–step, order (2, 2), unconditionally stable):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n}{4h} = \frac{f_m^{n+1} + f_m^n}{2}$$

The Crank–Nicolson scheme is derived by using approximations for the derivatives about $(t + \frac{1}{2}k, x)$ instead of $(t, x)$ as is usually done. We use a forward difference for $u_t$, but for the $u_x$ term, we start with the Taylor expansion

$$u_x(t + \frac{1}{2}k, x) = \frac{u_x(t + k, x) + u_x(t, x)}{2} + O(k^2).$$

We then replace $u_x(t+k, x)$ and $u_x(t, x)$ by second order accurate differences to obtain

$$\begin{aligned} u_x(t + \frac{1}{2}k, x) = \frac{1}{2}\left[\frac{u(t + k, x + h) - u(t + k, x - h)}{2h} + \frac{u(t, x + h) - u(t, x - h)}{2h}\right] \\ + O(k^2) + O(h^2). \end{aligned}$$

---

To perform stability analysis, consider the homogeneous case and write the scheme as

$$v_m^{n+1} = v_m^n - \frac{a\lambda}{4}(v_{m+1}^{n+1} - v_{m-1}^{n+1} + v_{m+1}^n - v_{m-1}^n).$$

So

$$v_m^{n+1} + \frac{a\lambda}{4}(v_{m+1}^{n+1} - v_{m-1}^{n+1}) = v_m^n - \frac{a\lambda}{4}(v_{m+1}^n - v_{m-1}^n).$$

The amplification factor then satisfies

$$g(\theta)[1+\frac{a\lambda}{4}(e^{i\theta}-e^{-i\theta})]=1-\frac{a\lambda}{4}(e^{i\theta}-e^{-i\theta}).$$

So

$$g(\theta)=\frac{1-i\frac{1\lambda}{2}\sin\theta}{1+i\frac{a\lambda}{2}\sin\theta}$$

and $|g(\theta)|\leq 1$ unconditionally since $g(\theta)$ is the ratio of conjugate complex numbers. We conclude the scheme is unconditionally stable.

---

Now we show that the scheme is accurate of order $(2,2)$ using symbols. We have

$$\begin{aligned}
p_{k,h}(s,\xi)&=\frac{1}{k}(e^{sk}-1)+\frac{a}{4h}(e^{sk}e^{i\xi h}-e^{sk}e^{-i\xi h}+e^{i\xi h}-e^{-i\xi h})\\
r_{k,h}(s,\xi)&=\frac{1}{2}(e^{sk}+1)\\
p(s,\xi)&=s+ai\xi
\end{aligned}.$$

First,

$$\begin{aligned}
p_{k,h}(s,\xi)&=s+\frac{s^2k}{2}+O(k^2)+\frac{ai}{2h}(e^{sk}\sin(h\xi)+\sin(h\xi))\\
&=s+\frac{s^2k}{2}+\frac{ai}{2h}\sin(h\xi)(e^{sk}+1)+O(k^2)\\
&=s+\frac{s^2k}{2}+\frac{ai}{2h}(h\xi-\frac{h^3\xi^3}{6})(2+sk+O(k^2))+O(k^2)\\
&=s+\frac{s^2k}{2}+\frac{ai\xi}{2}(2+sk)+O(k^2)+O(h^2).
\end{aligned}$$

Next,

$$r_{k,h}(s,\xi)=1+\frac{sk}{2}+O(k^2).$$

So

$$r_{k,h}(s,\xi)p(s,\xi)=s+ai\xi+\frac{s^2k}{2}+\frac{ai\xi sk}{2}+O(k^2).$$

We see that

$$p_{k,h}(s,\xi)-r_{k,h}(s,\xi)p(s,\xi)=O(h^2)+O(k^2).$$

# 5 Parabolic

The typical homogeneous equation we consider is

$$u_t = bu_{xx},$$

with $b > 0$ and initial condition

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

This is also known as the heat equation. By Fourier transform, the solution is found to be

$$\begin{aligned} u(t, x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{i\xi x} e^{-b\xi^2 t} \hat{u}_0(\xi)\ d\xi \\ &= \frac{1}{\sqrt{4\pi bt}} \int_{-\infty}^{\infty} e^{-(x-y)^2/4bt} u_0(y)\ dy. \end{aligned}$$

We also consider the inhomogeneous equation

$$u_t = bu_{xx} + f.$$

Systems of parabolic equations take the form

$$u_t = Bu_{xx},$$

where $u$ is a vector of functions of dimension $d$ and $B$ is a $d \times d$ matrix such that all eigenvalues of $B$ have positive real part.

To facilitate stability analysis we let $\mu = k/h^2$ and $\theta = h\xi$.

## 5.1 Some Basic Schemes

These schemes are of the form $P_{k,h}v = 0$.

- Forward–time central–space (Explicit, one–step, order (1, 2), stable in both $\|\cdot\|_2$ and $\|\cdot\|_\infty$ if $b\mu \le 1/2$):

$$\frac{v_m^{n+1} - v_m^n}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

  Rewrite the scheme as

$$v_m^{n+1} = v_m^n + b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n).$$

The amplification factor is then

$$\begin{aligned} g(\theta) &= 1 + b\mu(e^{i\theta} - 2 + e^{-i\theta}) \\ &= 1 + 2b\mu(\cos\theta - 1) \\ &= 1 - 4b\mu \sin^2 \frac{\theta}{2}. \end{aligned}$$

We see then that $|g(\theta)| \leq 1$ for all $\theta$ if and only if $4b\mu \leq 2$ if and only if $b\mu \leq 1/2$. We note that even though the scheme is only first order accurate in time, the requirement $b\mu \leq 1/2$ forces it to be second order accurate in time.

---

Note that if we write the scheme as

$$v_m^{n+1} = (1 - 2b\mu)v_m^n + b\mu(v_{m+1}^n + v_{m-1}^n),$$

then we see that if $b\mu \leq 1/2$, then coefficients can be pulled out and the triangle inequality applied, so that we also have the maximum–norm inequality

$$\|v^{n+1}\|_\infty \leq \|v^n\|_\infty.$$

Iterating, we see that the scheme is also stable under the maximum norm. Notice that the actual solution $u(t, x)$ also satisfies

$$\|u(t, \cdot)\|_\infty \leq \|u(0)\|_\infty.$$

- Backward–time central–space (implicit, one–step, order (1, 2), unconditionally stable in $\|\cdot\|_2$ and $\|\cdot\|_\infty$):

$$\frac{v_m^n - v_m^{n-1}}{k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

Rewrite the scheme as

$$v_m^n = v_m^{n-1} + b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n).$$

The amplification factor is then

$$\begin{aligned} g(\theta) &= [1 - b\mu(e^{i\theta} - 2 + e^{-i\theta})]^{-1} \\ &= \frac{1}{1 + 4b\mu \sin^2 \frac{\theta}{2}}. \end{aligned}$$

So we see that $|g(\theta)| \leq 1$ and the scheme is unconditionally stable.

---

To derive a max–norm inequality, rewrite the scheme as

$$(1+2b\mu)v_m^n - b\mu(v_{m+1}^n + v_{m-1}^n) = v_m^{n-1}.$$

Let $v_{m^+}^n$ be the largest of $v_\cdot^n$. Then

$$v_{m^+}^n \le (1+2b\mu)v_{m^+}^n - b\mu(v_{m^++1}^n + v_{m^+-1}^n) = v_{m^+}^{n-1},$$

so

$$v_{m^+}^n \le \max_m v_m^n.$$

Now let $v_{m^-}^n$ be the smallest of $v_\cdot^n$, then

$$v_{m^-}^n \le (1+2b\mu)v_{m^-}^n - b\mu(v_{m^-+1}^n + v_{m^--1}^n) = v_{m^-}^{n-1},$$

so

$$v_{m^-}^n \ge \min_m v_m^n.$$

Altogether we can now conclude

$$\|v^n\|_\infty \le \|v^{n-1}\|_\infty.$$

## 5.2 Unconditionally Stable Schemes: Crank–Nicolson and Du Fort–Frankel

- Crank–Nicolson (implicit, one–step, order (2, 2), unconditionally stable in $\|\cdot\|_2$ and stable in $\|\cdot\|_\infty$ if $b\mu \le 1$):

$$\frac{v_m^{n+1} - v_m^n}{k} = \frac{1}{2}b\frac{v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}}{h^2} + \frac{1}{2}b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2} + \frac{1}{2}(f_m^{n+1} + f_m^n).$$

To derive the Crank–Nicolson scheme, as in the hyperbolic case, we use Taylor expansion around $(t+\frac{1}{2}k, x)$. We have

$$u_{xx}(t+\frac{1}{2}k, x) = \frac{u_{xx}(t+k,x) + u_{xx}(t,x)}{2} + O(k^2).$$

Now replace $u_{xx}(t+k,x)$ and $u_{xx}(t,x)$ by centered second order approximations.

---

For stability, rewrite the scheme as

$$v_m^{n+1} - \frac{b\mu}{2}(v_{m+1}^{n+1} - 2v_m^{n+1} + v_{m-1}^{n+1}) = v_m^n + \frac{b\mu}{2}(v_{m+1}^n - 2v_m^n + v_{m-1}^n).$$

So the amplification factor is

$$\begin{aligned} g(\theta) &= \frac{1 + \frac{b\mu}{2}(e^{i\theta} - 2 + e^{-i\theta})}{1 - \frac{b\mu}{2}(e^{i\theta} - 2 + e^{-i\theta})} \\ &= \frac{1 + b\mu(\cos\theta - 1)}{1 - b\mu(\cos\theta - 1)} \\ &= \frac{1 - 2b\mu \sin^2\frac{\theta}{2}}{1 + b\mu \sin^2\frac{\theta}{2}}. \end{aligned}$$

So we see that $|g(\theta)| \leq 1$ for all $\theta$ and the scheme is unconditionally stable.

---

To derive a max–norm inequality, first rewrite the scheme as

$$(1 + b\mu)v_m^{n+1} - \frac{b\mu}{2}(v_{m+1}^{n+1} + v_{m-1}^{n+1}) = (1 - b\mu)v_m^n + \frac{b\mu}{2}(v_{m+1}^n + v_{m-1}^n).$$

Now note that if $b\mu \leq 1$, and $v_{m^+}^{n+1}$ is the largest of $v_\cdot^{n+1}$, then

$$\begin{aligned} v_{m^+}^{n+1} &\leq (1 + b\mu)v_{m^+}^{n+1} - \frac{b\mu}{2}(v_{m^+ +1}^{n+1} + v_{m^+ -1}^{n+1}) \\ &= (1 - b\mu)v_{m^+}^n + \frac{b\mu}{2}(v_{m^+ +1}^n + v_{m^+ -1}^n). \end{aligned}$$

Since $b\mu \leq 1$, we may pull out coefficients and get

$$\begin{aligned} (1 - b\mu)v_{m^+}^n + \frac{b\mu}{2}(v_{m^+ +1}^n + v_{m^+ -1}^n) &\leq (1 - b\mu)\max_m v_m^n + \frac{b\mu}{2} 2 \max_m v_m^n \\ &= \max_m v_m^n. \end{aligned}$$

So $v_{m^+}^{n+1} \leq \max_m v_m^n$. Now let $v_{m^-}^{n+1}$ be the smallest of $v_\cdot^{n+1}$, then a similar argument shows

$$v_{m^-}^{n+1} \geq \min_m v_m^n,$$

so altogether we learn

$$\|v^{n+1}\|_\infty \leq \|v^n\|_\infty.$$

- Du Fort–Frankel (explicit, multi–step, order $O(h^2) + O(k^2) + O(k^2h^{-2}) = O(h^2)$, for $k = h^2$, unconditionally stable):

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - (v_m^{n+1} + v_m^{n-1}) + v_{m-1}^n}{h^2} + f_m^n$$

For stability we rewrite the scheme as

$$(1+2b\mu)v_m^{n+1} - 2b\mu(v_{m+1}^n + v_{m-1}^n) - (1-2b\mu)v_m^{n-1} = 0.$$

So the amplification polynomial is

$$(1+2b\mu)g^2 - 4b\mu\cos\theta g - (1-2b\mu) = 0.$$

By the quadratic formula, the roots are found to be

$$\begin{aligned} g_\pm(\theta) &= \frac{2b\mu\cos\theta \pm \sqrt{4(b\mu)^2(\cos^2\theta - 1) + 1}}{1+2b\mu} \\ &= \frac{2b\mu\cos\theta \pm \sqrt{1 - 4(b\mu\sin\theta)^2}}{1+2b\mu}. \end{aligned}$$

We consider three cases, depending on the discriminant. If the discriminant is positive, then

$$|g_\pm(\theta)| \le \frac{2b\mu+1}{1+2b\mu} = 1.$$

If the discriminant is negative, then

$$\begin{aligned} |g_\pm(\theta)|^2 &= \frac{(2b\mu\cos\theta)^2 - 1 + (2b\mu\sin\theta)^2}{(1+2b\mu)^2} \\ &= \frac{(2b\mu)^2}{1+2b\mu)^2} \\ &\le 1. \end{aligned}$$

If the discriminant is zero, then there is a single repeated root $g(\theta)$, but with $|g(\theta)| < 1$ unconditionally. In all three cases, no condition is required for stability, so the scheme is unconditionally stable.

---

Next we check the order of accuracy of the scheme using symbols. We have

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{2k}(e^{sk} - e^{-sk}) - \frac{b}{h^2}(e^{i\xi h} - (e^{sk} + e^{-sk}) + e^{i\xi h}) \\ r_{k,h}(s,\xi) &= 1 \\ p(s,\xi) &= s + b\xi^2 \end{aligned}.$$

So

$$\begin{aligned} p_{k,h}(s,\xi) &= \frac{1}{2k}(2sk + 2\frac{(sk)^3}{3!} + O(k^5)) - \frac{b}{h^2}(2\cos(\xi h) - (2 + (sk)^2 + O(k^4)) \\ &= s + \frac{s^3k^2}{3!} + O(k^4) - b(-\xi^2 + \frac{2\xi^4h^2}{4!} + O(h^4) - (\frac{sk}{h})^2 + O(k^4)) \\ &= s + b\xi^2 + O(k^2) + O(h^2) + O(k^2/h^2). \end{aligned}$$

Since $r_{k,h}(s,\xi)p(s,\xi) = s + b\xi^2$, we conclude that the scheme is accurate of order $O(k^2) + O(h^2) + O(k^2/h^2)$. So although this scheme is explicit and unconditionally stable, it is only consistent if $k/h \to 0$ as $k, h \to 0$.

The analogue of the CFL condition for hyperbolic systems is the following theorem:

**Theorem 5.1.** *An explicit, consistent scheme for the parabolic system is convergent only if $k/h \to 0$ as $k, h \to 0$.*

### 5.3 The Convection–Diffusion Equation

The convection–diffusion equation is

$$u_t + au_x = bu_{xx}.$$

The solution to this equation is the solution to the wave equation with a "traveling" coordinate system, i.e. change of variable $y = x - at$.

- Forward–time central–space (explicit, one–step, order (2, 2), stable in $\|\cdot\|_2$ if $b\mu \le 1/2$, stable in $\|\cdot\|_\infty$ if $h \le \frac{2b}{a}$):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

  For stability, rewrite the scheme as

$$v_m^{n+1} = v_m^n - \frac{ak}{2h}(v_{m+1}^n - v_{m-1}^n) + b\mu(v_{m+1}^n - 2v_m^n + v_{m-1}^n).$$

  Then the amplification factor is (with $\lambda = k/h$ and $\mu$ constant)

$$g(\theta, \lambda) = 1 - ia\lambda \sin\theta + 2b\mu(\cos\theta - 1).$$

  Since the term with $a\lambda$ is purely imaginary, it contributes $a^2\lambda^2 \sin^2\theta$ to $|g|^2$, but since $\lambda^2 = k\mu$, it contributes $O(k)$ and hence can be dropped from the analysis. Thus, the stability requirement is still $b\mu \le 1/2$.

---

  Next we are interested in a max–norm inequality. Rewrite the scheme as

$$v_m^{n+1} = (1 - 2b\mu)v_m^n + (b\mu - \frac{ak}{2h})v_{m+1}^n + (b\mu + \frac{ak}{2h})v_{m-1}^n.$$

If all coefficients are positive, then taking absolute value and using the triangle inequality and finally taking a maximum we will learn that

$$\|v^{n+1}\|_\infty \leq \|v^n\|_\infty.$$

To ensure coefficients are positive, we need (in both the case $a > 0$ and $a < 0$)

$$|a|\lambda \leq 2b\mu.$$

Rewriting this in terms of $k$ and $h$, we learn that we need

$$h \leq \frac{2b}{|a|}.$$

This, unlike a stability condition (which has to do with the limit as $k, h \to 0$ and usually restricts the value of $\lambda$ or $\mu$), is simply a restriction on the step size. Or, we can think of this as a restriction on the region of stability in the maximum norm.

We may improve the condition $h \leq \frac{2b}{|a|}$ by upwind differencing of the convection $(au_x)$ term. We have two cases, $a > 0$ and $a < 0$.

- Upwind differencing $(a > 0)$ (Order (1, 2), stable in $\|\cdot\|_2$ and $\|\cdot\|_\infty$ if $a\lambda + 2b\mu \leq 1$):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_m^n - v_{m-1}^n}{h} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

The amplification factor is (holding $\mu$ constant)

$$\begin{aligned} g(\theta, \lambda) &= 1 - a\lambda(1 - e^{-i\theta}) + 2b\mu(\cos\theta - 1) \\ &= (1 - a\lambda - 2b\mu) + (a\lambda + 2b\mu)\cos\theta - ia\lambda\sin\theta. \end{aligned}$$

As before, $\lambda^2 = k\mu$ so we can drop the $ia\lambda\sin\theta$ term. The stability condition is then

$$|(1 - (a\lambda + 2b\mu)) + (a\lambda + 2b\mu)\cos\theta| \leq 1.$$

So since $a > 0$, we need

$$a\lambda + 2b\mu \leq 1.$$

---

For a max–norm inequality, rewrite the scheme as

$$v_m^{n+1} = (1 - (a\lambda + 2b\mu))v_m^n + b\mu v_{m+1}^n + (a\lambda + b\mu)v_{m-1}^n.$$

So since $a > 0$, to ensure all coefficients are positive and derive a max–norm inequality, all we need is

$$a\lambda + 2b\mu \leq 1.$$

- Upwind differencing ($a < 0$) (Order (1, 2), stable in $\|\cdot\|_2$ and $\|\cdot\|_\infty$ if $|a|\lambda + 2b\mu \leq 1$):

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_m^n}{h} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}$$

For stability, rewrite the scheme as

$$v_m^{n+1} = (1 + a\lambda - 2b\mu)v_m^n + (b\mu - a\lambda)v_{m+1}^n + b\mu v_{m-1}^n.$$

So the amplification factor is (with $\mu$ constant)

$$\begin{aligned} g(\theta, \lambda) &= (1 + a\lambda - 2b\mu) + 2b\mu\cos\theta - a\lambda(\cos\theta + i\sin\theta) \\ &= (1 - (2b\mu + |a|\lambda)) + (2b\mu + |a|\lambda)\cos\theta + |a|\lambda i\sin\theta. \end{aligned}$$

As before the $|a|\lambda i\sin\theta$ term can be discarded. Then we see that to ensure $|g| \leq 1$, we need

$$|a|\lambda + 2b\mu \leq 1.$$

---

For a max–norm inequality, rewrite the scheme as

$$v_m^{n+1} = (1 - (2b\mu + |a|\lambda))v_m^n + (b\mu + |a|\lambda)v_{m+1}^n + b\mu v_{m-1}^n$$

and we see that the condition is the same as required for stability

$$|a|\lambda + 2b\mu \leq 1.$$

Note that in both cases the condition is $|a|\lambda + 2b\mu \leq 1$, but the schemes used are slightly different. The condition can be written as

$$|a|\frac{k}{h} + 2b\frac{k}{h^2} \leq 1.$$

Compared to

$$h \leq \frac{2b}{|a|},$$

using upwind differencing gives a less restrictive condition than the original scheme, especially when $a$ is large, at the cost of being only first order accurate in space.

## 6 ADI: Peaceman–Rachford

We consider hyperbolic or parabolic equations in two spatial dimensions. THe equation

$$u_t = b_{11}u_{xx} + 2b_{12}u_{xy} + b_{22}u_{yy}$$

is parabolic if

$$b_{11}, b_{22} > 0 \quad \text{and} \quad b_{12}^2 < b_{11}b_{22}.$$

ADI stands for alternating direction implicit and refers to a method which reduces a higher–dimensional problem to one–dimensional implicit problems by factoring the scheme. Consider the equation

$$u_t = A_1 u + A_2 u,$$

where $A_1$ and $A_2$ are linear (differential) operators. Note that applying say the Crank–Nicolson scheme directly would require inverting matrices at each time step that are not tridiagonal. The ADI method factors the scheme so that at each step the matrices that need to be inverted are again tridiagonal. We describe one such method.

- Peaceman–Rachford (Order $(1, 2)$, unconditionally stable):

$$\left(I - \frac{k}{2}A_{1h}\right)\tilde{v}^{n+1/2} = \left(I + \frac{k}{2}A_{2h}\right)v^n$$
$$\left(I - \frac{k}{2}A_{2h}\right)v^{n+1} = \left(I + \frac{k}{2}A_{1h}\right)\tilde{v}^{n+1/2}$$

  To derive the scheme, write Taylor series expansions around $t = (n + \frac{1}{2})k$

$$\frac{u(t+k,x) - u(t,x)}{k} = u(t + \frac{k}{2}, x)_t + O(k^2).$$

  Also, from the differential equation

$$\begin{aligned} u(t + \frac{k}{2}, x) &= (A_1 + A_2)u(t + \frac{k}{2}, x) \\ &= \frac{(A_1 + A_2)[u(t+k,x) + u(t,x)]}{2} + O(k^2). \end{aligned}$$

  This gives the scheme

$$\frac{u^{n+1} - u^n}{k} = \frac{1}{2}(A_1 u^{n+1} + A_2 u^{n+1}) + \frac{1}{2}(A_1 u^n + A_2 u^n).$$

Multiplying by $k$ and rearranging, we get

$$\left(I - \frac{k}{2}A_1 - \frac{k}{2}A_2\right) u^{n+1} = \left(I - \frac{k}{2}A_1 - \frac{k}{2}A_2\right) u^n + O(k^3).$$

We want to factor the scheme, so add $k^2 A_1 A_2 u^{n+1}/4$ to both sides, to obtain

$$\begin{aligned}\left(I - \frac{k}{2}A_1\right)\left(I - \frac{k}{2}A_2\right) u^{n+1} &= \left(I + \frac{k}{2}A_1\right)\left(I + \frac{k}{2}A_2\right) u^n + \frac{k^2}{4}A_1 A_2(u^{n+1} - u^n) + O(k^3) \\ &= \left(I + \frac{k}{2}A_1\right)\left(I + \frac{k}{2}A_2\right) u^n + O(k^3)\end{aligned},$$

since $u^{n+1} = u^n + O(k)$. Note that we have suppressed the spatial indices $m_x, m_y$ to avoid clutter, but it should be understood that these indices are present (i.e. $u^n$ really should be $u^n_{m_x,m_y}$ and $A_{1h}$ will work with $m_x$ whereas $A_{2h}$ works with $m_y$). We can now plug in second order discretizations $A_{1h}$ and $A_{2h}$ for $A_1$ and $A_2$, then the total error would be of $O(kh^2) + O(k^3)$. Now that the scheme is factored we can introduce the intermediate variable $\tilde{v}^{n+1/2}$, and we have derived the Peaceman–Rachford algorithm.

---

We apply the method to

$$u_t = b(u_{xx} + u_{yy}),$$

with

$$A_{1h} = b\delta_x^2 v_{m_x} = b\frac{v_{m_x+1} - 2v_{m_x} + v_{m_x-1}}{h_x^2}$$

and

$$A_{2h} = b\delta_y^2 v_{m_y} = b\frac{v_{m_y+1} - 2v_{m_y} + v_{m_y-1}}{h_y^2},$$

Then using the Fourier representation, we have

$$\left(1 - \frac{bk}{2h_x^2}(e^{i\theta_x} - 2 + e^{-i\theta_x})\right)\hat{v}^{n+1/2} = \left(1 + \frac{bk}{2h_y^2}(e^{i\theta_y} - 2 + e^{-i\theta_y})\right)\hat{v}^n$$

and

$$\left(1 - \frac{bk}{2h_y^2}(e^{i\theta_y} - 2 + e^{-i\theta_y})\right)\hat{v}^{n+1} = \left(1 + \frac{bk}{2h_x^2}(e^{i\theta_x} - 2 + e^{-i\theta_x})\right)\hat{\tilde{v}}^{n+1/2}.$$

So with $\mu_x = k/h_x^2$ and $\mu_y = k/h_y^2$ constant,

$$g_1 = \frac{1 - b\mu_y(1 - \cos\theta_y)}{1 + b\mu_x(1 - \cos\theta_x)} \quad \text{and} \quad g_2 = \frac{1 - b\mu_x(1 - \cos\theta_x)}{1 + b\mu_y(1 - \cos\theta_y)}.$$

Concatenating, we get

$$\hat{v}^{n+1} = g_2 g_1 \hat{v}^n.$$

The amplification factor is then $g = g_2 g_1$ and we see that $|g| \leq 1$ and the scheme is unconditionally stable. Note that if the equation was $u_t = b_1 u_{xx} + b_2 u_{yy}$ with $b_1 \neq b_2$, it is still true that $|g| \leq 1$ unconditionally.

---

To apply the method to

$$u_t = b_{11} u_{xx} + 2b_{12} u_{xy} + b_{22} u_{yy},$$

we will need to add a discretization for the mixed derivative term. We use e.g.

$$\left(1 - \frac{k}{2} b_{11} \delta_x^2\right) v^{n+1/2} = \left((1 + \frac{k}{2} b_{22} \delta_y^2) + k b_{12} \delta_{0x} \delta_{0y}\right) v^n$$

$$\left(1 - \frac{k}{2} b_{22} \delta_y^2\right) v^{n+1} = \left((1 + \frac{k}{2} b_{11} \delta_x^2) + k b_{12} \delta_{0x} \delta_{0y}\right) v^{n+1/2},$$

where

$$\delta_0 v_m = \frac{v_{m+1} - v_{m-1}}{2h}$$

is a second order approximation to the first derivative, so

$$\delta_{0x} \delta_{0y} v_m = (v_m)_{xy} + O(h_x^2) + O(h_y^2),$$

and the scheme is accurate of order (1, 2).

# 7 Laplace Transform and Initial Boundary Value Problems

Let $f : [0, \infty) \to \mathbb{R}$. The Laplace transform of $f$ is defined as

$$\mathcal{L} f(s) = \int_0^\infty f(t) e^{-st} \, dt,$$

for any $s \in \mathbb{C}$ such that the integral exists. The Laplace transform is clearly linear. Now let's find the Laplace transfrom of $e^{at}$, $a \in \mathbb{C}$.

$$\mathcal{L}[e^{at}](s) = \lim_{N \to \infty} \int_0^N e^{(a-s)t} dt = \lim_{N \to \infty} \frac{e^{(a-s)N} - 1}{a - s},$$

which, if $\Re e(a) < \Re e(s)$, is equal to $\frac{1}{s-a}$. From this calculation we see that whenver $\exists M, \alpha \geq 0$ such that

$$|f(t)| \leq M e^{\alpha t}, \quad \forall t \geq 0,$$

then the Laplace transform exists for any complex $s$ with $\Re e(s) > \alpha$.

One usefulness of the Laplace transform can be seen by investigating $\mathcal{L}[f']$:

$$\begin{aligned}\mathcal{L}[f'](s) &= \int_0^\infty e^{-st} f'(t)\, dt \\ &= e^{-st} f(t)\Big|_0^\infty + \int_0^\infty s e^{-st} f(t) \\ &= s\mathcal{L}f(s) - f(0),\end{aligned}$$

where we have used the integration by parts and the fact that $e^{-st} f(t) \to 0$ as $t \to \infty$ (otherwise the integral is not even defined). The formula

$$\mathcal{L}[f'](s) = s\mathcal{L}f(s) - f(0)$$

can be iterated. Iterating once, we obtain

$$\mathcal{L}[f''](s) = s^2\mathcal{L}f(s) - sf(0) - f(0).$$

This makes the Laplace transform especially useful for solving initial value problems.

---

To derive additional properties of the Laplace transform, we make connection with the Fourier transform. Extend the domain of $f$ to $(-\infty, \infty)$ by setting $f(t) = 0$ for $t < 0$, then the Fourier transform of $f$ is

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty f(t) e^{-i\xi t}\, dt = \frac{1}{\sqrt{2\pi}} \int_0^\infty f(t) e^{-i\xi t}\, dt.$$

We are a change of variable away from the Laplace transform:

$$\mathcal{L}f(s) = \sqrt{2\pi}\hat{f}(-is),$$

so

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \mathcal{L}f(i\xi).$$

From the Fourier inversion formula then, we have

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \hat{f}(\xi)\, d\xi = \frac{1}{2\pi} \int_{-\infty}^\infty \mathcal{L}f(i\xi)\, d\xi.$$

With the change of variable $s = i\xi$, we have

$$f(t) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} \mathcal{L}f(s)\, ds.$$

Note that this inversion formula assumes $\mathcal{L}f(s)$ is defined for $s$ with $\Re e(s) = 0$. For functions which only have Laplace transforms defined for $\Re e(s)$ sufficiently large, we will need to make a change of variable to shift the line of integration to the right. Suppose $|f(t)| \leq Me^{\alpha t}$, then for any $a > \alpha$, $\mathcal{L}[e^{-at}f](s)$ is defined on $(-i\infty, i\infty)$ and so

$$\begin{aligned} e^{-at}f(t) &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \mathcal{L}[e^{-at}f](s)e^{st}\, ds \\ &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \left[\int_0^\infty f(t')e^{-(a+s)t'}\, dt'\right] e^{st}\, ds \\ &= \frac{1}{2\pi i}\int_{-i\infty}^{i\infty} \mathcal{L}f(a+s)e^{st} ds. \end{aligned}$$

Notice that we have learned

$$\mathcal{L}[e^{-at}f](s) = \mathcal{L}f(s+a),$$

that is, multiplying $f(t)$ by $e^{-at}$ has the effect of translation by $a$ on the dual variable. After multiplying both sides by $e^{at}$ and a linear change of variable $s \mapsto s + a$, we finally have

$$f(t) = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} \mathcal{L}f(s)e^{st}\, ds.$$

This is the so–called Bromwich integral and can usually be done by complexifying the integrand and using residue theory. The integral is usually approximated by a half–circle contour with the vertical segment on the line $x = a$. To ensure that the integral over the circular part goes to zero, we close the contour on the left when $t \geq 0$ (so that $\Re e(e^{zt}) < 0$) and on the right when $t < 0$. Since $f(t) = 0$ for $t < 0$, it is the case that all poles lie to the left of the line $x = a$.

A few things are worth pointing out at this point. Let's write $s = a + i\tau$, $a, \tau \in \mathbb{R}$.

- Notice that if we differentiate the Laplace inversion formula under the integral sign, we will learn that $\mathcal{L}f'(s) = s\mathcal{L}f(s)$, without the $-f(0)$ term. There is some subtlety here regarding what happens at $t = 0$ which we do not wish to go into here. A satisfactory resolution of this would require generalized functions.
- The information that we are interested in $t > 0$ is encoded in the fact that we take $a > 0$, so that $e^{-st} \to 0$ as $t \to \infty$.
- From another perspective, since $e^{-st} = e^{-at}e^{-i\tau t}$, we can basically think of the Laplace transform as a generalization of the Fourier transform, where we may multiply the function $f(t)$ by a decaying factor $e^{-at}$ before transforming.
- From yet another perspective, we can view the Laplace transform as a one-parameter family of Fourier transforms, parametrized by $a$.

We now want some form of Parseval's relation. Since $s$ is complex, we need to explain what we mean by the "$L^2$–norm" of $\mathcal{L}[f]$. Fix the real part of $s$, then we may integrate over $\tau$ (again using $f(t) = 0$ for $t < 0$) and use Parseval's relation for Fourier transform to obtain

$$\begin{aligned}\int_{-\infty}^{\infty} |\mathcal{L}f(a+i\tau)|^2 \, d\tau &= \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} [e^{-at} f(t)] e^{-i\tau t} \, dt \right|^2 \, d\tau \\ &= \int_{-\infty}^{\infty} \left| \widehat{(e^{-at} f)}(\tau) \right|^2 \, d\tau \\ &= \int_{-\infty}^{\infty} |e^{-at} f(t)|^2 \, dt.\end{aligned}$$

We can write this as

$$\|\mathcal{L}f\|^2_{L^2(a,i\mathbb{R})} = \|f\|^2_{L^2(e^{-a\mathbb{R}})},$$

and we introduce the notation

$$\|f\|_a^2$$

for this family of norms.

---

We are interested in initial boundary value problems for PDEs of the form

$$\begin{aligned} Pu &= f \\ u(0,x) &= u_0(x) \\ Bu &= \beta \quad \text{on } \partial\Omega, \end{aligned}$$

on some domain $\Omega \subset \mathbb{R}^n$. We want to determine if boundary condtions are well–posed, i.e. we want to obtain estimates bounding the norm of the solution in terms of the norms of $u_0$, $f$, and $\beta$. The general estimate takes the form

$$\|u\|_a^2 + |u|_a^2 \leq C(a)(|\beta|_a^2 + \|f\|_a^2 + \|u_0\|^2),$$

where e.g. $|u|_a^2$ denotes the corresponding norm over the boundary.

We will make a few simplifications, which we do not fully justify here.

- We extend the time interval to be $(-\infty, \infty)$.
- We will separate out the boundary analysis. We will assume that suitable extensions can be found so that the initial value problem has some solution $w$ defined on all of $\mathbb{R}^n$ (so that $Pw = f$ and $w(0,x) = u_0(x)$). We then write $u = w + v$ and look for a solution $v$ to just the boundary value problem, that is, $v$ now satisfies

  $$Pv = 0, v(0,x) \equiv 0, Bv = \beta \text{ on } \partial\Omega.$$

  This turns out to be sufficient.

- Well–posedness is local: To check well–posedness, it is enough to consider the frozen coefficient problem, that is, if the problem with $t$ and $x$ fixed at $t_0$ and $x_0$ is well–posed for all possible values of $t_0$ and $x_0$, then the original problem is well–posed.

---

As an illustration, suppose we have a problem of the form

$$u_t = bu_{xx}, u(0,x) \equiv 0, \text{ and } u_x(t,0) = \beta(t)$$

on the region $\{(t,x) : t \in \mathbb{R}, x \in \mathbb{R}^+\}$. The idea is basically the same as when trying to show well–posedness of an initial value problem, but whereas there the variable of interest was $t$, here the variable of interest is $x$, so instead of Fourier transforming in space to "algebraitize" the space variables so that we are left with an ordinary differential equation in time, here we will Laplace transform in time: We get the resolvent equation ($\tilde{u}_t = s\tilde{u}(s,x) - u(0,x)$, but $u(0,x) \equiv 0$)

$$s\tilde{u} = b\tilde{u}_{xx},$$

where now $\tilde{u} = \mathcal{L}u$. This is a second order ordinary differential equation in $x$, with solution

$$\tilde{u}(s,x) = \tilde{u}(s,0)e^{-\kappa x},$$

where $\kappa = \sqrt{s/b}$, $\Re e(\kappa) > 0$. $\hat{u}(s,0)$ is determined by the boundary condition: Laplace transform the boundary condition to obtain

$$\tilde{u}_x(s,0) = \tilde{\beta}(s).$$

Differentiating the solution once in $x$ and plugging into the last expression, we obtain

$$-\kappa\tilde{u}(s,0) = \tilde{\beta}(s).$$

Given $s$, we view this as a linear equation to solve for the unkonwn $\tilde{u}(s,0)$. The equation has a unique solution if and only if

$$\kappa\tilde{u}(s,0) = 0$$

has no non–trivial solution if and only if $\kappa = \sqrt{s/b} \neq 0$ if and only if $s = 0$. We conclude the boundary condition is well–posed.

---

Basically the general procedure for e.g. a problem with boundary at $x = 0$ is to Laplace transform in time and Fourier transform in all space variables except $x$ to form the resolvent equation

$$[s - Q(\partial_x, i\xi)]\hat{u} = 0,$$

which yields an ordinary differential equation in $x$ for $\hat{u}$. The boundary condition for the differential equation is then given by the original boundary condition and we will obtain one linear equation for each boundary condition. The analysis of uniqueness of solution then reduces to considering the homogeneous boundary condition

$$B\hat{u} = 0.$$

More precisely we have the following definition and theorem.

**Definition 7.1.** An admissible solution to the resolvent equation is a solution that is in $L^2(\mathbb{R}_+)$ as a function of $x$ when $\Re e(s) > 0$ and when $\Re e(s) = 0$, it is the limit of admissible solutions with $\Re e(s)$ positive.

**Theorem 7.2.** *The initial–boundary value problem for a PDE of the form*

$$u_t = P(\partial_x, \partial_y)u + f(t, x, y)$$

*for $x \in \mathbb{R}^+$ and $y \in \mathbb{R}^d$ with boundary conditions*

$$Bu = \beta(t, y)$$

*is well–posed in the strong sense if and only if there are no nontrivial admissible solutions to the resolvent equation that satisfy the homogeneous boundary condition.*

By well–posed in the strong sense we mean not only does there exist a unique solution, but there are estimates of the solution in terms of the boundary data.

---

The discrete analogue of the Laplace transform is the $z$–transform. Given a discrete sequence $\{a_n\}_{-\infty}^{\infty}$, the $z$–transform is defined as

$$A(z) = \sum_{n=-\infty}^{\infty} a_n z^{-n},$$

where the series converges. The part of the series with positive index converges inside some disk, whereas the part with nonpositive index converges outside some disk. Therefore the entire series converges inside the annulus

$$\limsup \sqrt[n]{|a_n|} < |z| < (\limsup \sqrt[n]{|a_{-n}|})^{-1}.$$

By complex analysis (observe that $\oint_{|z|=1} z^n = 0$ unless $n = 1$ in which case it is equal to $2\pi i$), we then have the inversion formula

$$a_n = \frac{1}{2\pi i} \oint_\Gamma A(z) z^{n-1}\, dz,$$

where $\Gamma$ is any positively oriented simple curve inside the annulus of convergence. We collect a few basic properties of the $z$–transform:

- The $z$–transform is linear.
- Shifting index in $\{a_n\}$ corresponds to multiplying $A(z)$ by $z$: The $z$–transform of $\{b_n\}$ is $zA(z)$ where $b_n = a_{n+1}$.
- Observe that $\frac{d}{dz}\left(\sum_{-\infty}^{\infty} a_n z^{-n}\right) = -\frac{1}{z}\sum_{-\infty}^{\infty} n a_n z^{-n}$ and the annulus of convergence does not change, so the $z$–transform of $\{na_n\}_{-\infty}^{\infty}$ is $-zA'(z)$.

What is perhaps more reminiscent of the Laplace transform is the causal $z$–transform:

$$A^+(z) = \sum_{n=0}^{\infty} a_n z^{-n}.$$

This is one–sided, so we can no longer shift for free but instead must "pay" at the boundary: If $b_n = a_{n+1}$, then the $z$–transform of $\{b_n\}$ is

$$z[A^+(z) - a(0)].$$

It is this equation and iterates thereof that makes the $z$–transform very useful for solving recurrence relations: One takes the $z$–transform of the sequence of interest and then use the above formula to transform the recurrence equations into expressions involving $A(z)$ and $z$. Then some combination of closed form formula (which usually involves geometric series manipulations), Taylor expansion, or contour integration can then be used to extract the coefficients $a_n$. In this connection the $z$–transform is often called generating functions.

---

It is not difficult to make the analogy with the Laplace (and hence Fourier) transform exact. As before, we need a few changes of variables. Let $k > 0$ be what will become the time step size and choose $a$ such that the circle of radius $e^{ka}$ lies inside the annulus of convergence of $A(z)$. Let $s = a + i\tau$ and $z = e^{sk}$. Then

$$A(z) = \sum_{-\infty}^{\infty} a_n e^{-skn}.$$

The inversion formula now becomes ($z = e^{ak}e^{ik\tau}$, so $dz = ikz\,d\tau = ike^{sk}$)

$$\begin{aligned} a_n &= \frac{1}{2\pi i}\oint_{|z|=e^{ka}} A(z) z^{n-1}\, dz \\ &= \frac{1}{2\pi i}\int_{-\pi/k}^{\pi/k} A(e^{ak}e^{ik\tau}) ike^{skn}\, d\tau \\ &= \frac{1}{2\pi}\int_{-\pi/k}^{\pi/k} kA(z)e^{skn}\, d\tau. \end{aligned}$$

Note the role the parameter $a$ plays in both the Laplace transform and the $z$–transform: In both cases it parametrizes the domain of definition, a family of lines in the former case and a family of circles in the latter case.

Also notice that we have introduced a factor of $k$ in from of $A(z)$ in the inversion formula. If we return to the definition of the $z$–transform and defined instead

$$A(z) = k \sum_{-\infty}^{\infty} a_n z^{-n},$$

then the inversion formula will no longer have this factor of $k$. Also, writing everything in terms of $s$ and making a dummy change of index $n \rightsquigarrow kn$, we get

$$A(z) = \sum_{-\infty}^{\infty} a_{kn} e^{-skn} k$$

which we recognize as basically a Riemann sum with $\Delta t = k$, $t \rightsquigarrow nk$, which in the $k \to 0$ limit appears to be the Laplace transform, except maybe the two–sidedness, but that can of course be easily cured by setting $a_n = 0$, $\forall n < 0$. Having said that, however, we shall return to our original definition ($A(z) = \sum_{-\infty}^{\infty} a_n z^{-n}$) to avoid clutter.

We will be interested in analyzing boundary conditions for finite difference schemes. So we will have some discrete function $v_m^n$, where $m$ indexes space and $n$ indexes time and the $z$–transform will be taken in time. To emphasize the connection with the Laplace transform, the $z$–transform of $v_m^n$ will be denoted by $\tilde{v}_m(z)$. After a change of variables to remove $k$, the inversion formula becomes

$$\begin{aligned} v_m^n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{v}_m(z) e^{sn} \, d\tau \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{v}_m(z) e^{an} e^{i\tau n} \, d\tau, \end{aligned}$$

where $i\xi = s$. Also, $\tilde{v}_m(z) = \sum_{-\infty}^{\infty} v_m^n e^{-an} e^{-i\tau n}$, which we recognize as the Fourier transform of $w_m^n = v_m^n e^{-an}$, so

$$e^{-an} v_m^n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{w}_m(\tau) e^{i\tau n} \, d\tau.$$

Parseval's relation then immediately follows

$$\sum_{-\infty}^{\infty} e^{-2an} |v_m^n|^2 = \int_{-\pi}^{\pi} |\tilde{v}_m(z)|^2 \, d\tau.$$

We use the notation

$$\|v_m^n\|_a^2$$

to denote the above quantity.

---

We now use the $z$–transform to analyze stability of boundary conditions for FDS. We will consider one–dimensional hyperbolic equations on $\mathbb{R}^+$, with the boundary at 0. Suppose we have a scheme $P_{k,h}v_m^n = 0$ (as already discussed it is enough to consider the homogeneous case) which is consistent and stable for the initial value problem. For simplicity also assume the restrictive stability condition $|g| \leq 1$ holds. The boundary conditions we write as

$$B_{k,h}v_0^n = \beta(t_n).$$

Stability of boundary condition means that we have the following estimate:

$$a\|v^n\|_a^2 + |v^n|_a^2 \leq C|\beta|_a^2.$$

---

The starting point is very similar to von Neumann analysis: $z$–transform the difference scheme to obtain the resolvent equation

$$\tilde{P}_{k,h}(z)\tilde{v}_m(z) = 0.$$

This is a a recurrence relation for the quantity $\tilde{v}_m(z)$. We may write down the characteristic polynomial as usual and if say the degree of the polynomial is 2, then without knowing anything, the general solution takes the form $A(z)\kappa_+^m + B(z)\kappa_-^m$ when $\kappa_+$ and $\kappa_-$ are distinct roots of $P(\kappa)$ and $A(z)\kappa_0^m + B(z)m\kappa_0^{m-1}$ when $\kappa_+(z) = \kappa_-(z) = \kappa_0(z)$, and the coefficients $A(z)$ and $B(z)$ are determined by the boundary condition.

However, we have some information on the roots $\kappa_+$ and $\kappa_-$. First notice that our $z$–transform is one–sided ($v_m^n$ defined only for $n \geq 0$) and therefore recalling the definition, we see that we are interested in solutions of the resolvent equation for all $|z| > 1$ (since $z = e^s k = e^{a+i\tau}k$, this is the same as saying $a > 0$ or $\Re e(s) > 0$, exactly the same set of $s$ we were interested in for the analysis of well–posedness). Next we try to utilize the fact that the scheme is stable. Notice if we now plug the particular solution $\tilde{v}_m(z) = A(z)\kappa^m$ into the resolvent equation, the form of the resolvent equation then tells us that

$$\tilde{P}_{k,h}(z)[A(z)\kappa^m] = \tilde{p}(z,\kappa)[A(z)\kappa^m] = 0,$$

where $\tilde{p}(z,\kappa)$ is the symbol $p_{k,h}(s,\xi)$ with a change of variable:

$$\tilde{p}(e^{sk}, e^{ih\xi}) = p_{k,h}(s,\xi).$$

Now recall that the amplification polynomial $\Phi(g,\theta)$ ($\theta = h\xi$) is obtained by multiplying the scheme by $k$ and setting $v_m^n = g^n e^{im\theta}$. Therefore $\Phi(e^{sk},\theta) = kp_{k,h}(s,\xi)$, and so

$$k\tilde{p}(g, e^{i\theta}) = \Phi(g,\theta).$$

For a nontrivial solution of the form $A(z)\kappa^m$ to exist (with $|z| > 1$), we see from the resolvent equation that it must be the case that $\tilde{p}(z,\kappa)A(z) = 0$, that is

$$\tilde{p}(z,\kappa) = 0$$

If there is $\kappa$ such that $\tilde{p}(z,\kappa) = 0$ and $|\kappa| = 1$, then writing $\kappa = e^{i\theta}$, from the relation we just derived, we learn that

$$k\tilde{p}(z,\kappa) = \Phi(z,\theta) = 0,$$

so that $z$ is a root of the amplification polynomial. But since the scheme is stable, the (restricted) stability condition must hold, which implies that $|z| \leq 1$, a contradiction, since we have assumed that $|z| > 1$. We conclude that given a fixed value of $|z| > 1$, the roots of the equation $\tilde{p}(z,k) = 0$ separates into two groups

$$\mathfrak{K}_+(z) = \{k_-(z) : \tilde{p}(z,k_-(z)) = 0 \text{ and } |k_-(z)| < 1\}$$

and

$$\mathfrak{K}_+(z) = \{k_+(z) : \tilde{p}(z,k_+(z)) = 0 \text{ and } |k_+(z)| > 1\}.$$

Let $K_\pm(z) = \#(\mathfrak{K}_\pm(z))$ and let $N$ be the degree of $\tilde{p}(z,\kappa)$ viewed as a polynomial of $\kappa$ (e.g. $N = 2$ in the Lax–Wendroff scheme). Then $K_+(z) + K_-(z) = N$, a constant independent of $z$. For a fixed value of $z$, $\tilde{p}(z,\kappa)$ is an analytic function of $\kappa$, so by the Argument Principle (apply the Residue Theorem to $\oint_C \frac{f'(z)}{f(z)}dz = i\gamma(f(z))$, where $C$ denotes a closed curve and note that singularities in the integrand occur when $f$ has a zero or a pole), we have that

$$K_-(z) = \frac{1}{2\pi i}\oint_{|z|=1} \frac{\tilde{p}'(z,\kappa)}{\tilde{p}(z,\kappa)}\, d\kappa.$$

However, the integral is continuous with respect to $z$ and therefore $K_-(z)$ is an integer–valued continuous function of $z$ and hence must be a constant independent of $z$, from which we also conclude $K_+(z)$ is a constant independent of $z$. We have established the following

**Lemma 7.3.** *If the FDS is stable, then there are integers $K_-$ and $K_-$, independent of $z$, such that for $|z| > 1$ the roots $\kappa(z)$ of the equation $\tilde{p}(z,\kappa) = 0$ separate into two groups, one with $K_-$ roots all with $|\kappa_{-,\nu}(z)| < 1$, $\nu = 1,\ldots,K_-$ and one with $K_+$ roots, all with $|\kappa_{+,\nu}(z)| > 1$, $\nu = 1,\ldots,K_+$.*

In the case of a system of equations, $\tilde{p}(z,\kappa) = 0$ will be replaced by $\det\tilde{p}(z,\kappa) = 0$, and then $A(z)$ will have to be an eigenvector of $\tilde{p}(z,\kappa)$ with associated eigenvalue 0. The rest of the analysis does not change and the lemma continues to hold. It is of course the roots in $\mathfrak{K}_-(z)$ which provide good solutions (i.e. solutions which don't blow up as $m \to \infty$) to the

resolvent equation, so in the case where all roots in $\mathfrak{K}_-$ are distinct, the general solution looks like

$$\tilde{v}_m = \sum_{\nu=1}^{K_-} \alpha_\nu(z) A_\nu(z) k_{-,\nu}^m,$$

where $A_\nu$ are particular eigenvectors and $\alpha_n u$ are constants to be determined by the boundary condition.

**Definition 7.4.** An admissible solution to the resolvent equation is a solution which is in $L^2(h\mathbb{Z}_+)$ when $|z| > 1$, and when $|z| = 1$, it is the limit of admissible solutions with $|z| > 1$ (i.e. $v(z) = \lim_{\epsilon \to 0^+} v(z(1+\epsilon))$).

The set of admissible solutions is a vector space of dimension $K_-$. We have $K_-$ coefficients to solve for and therefore we see that we need $K_-$ boundary conditions. Suppose we have some boundary condition

$$B_{k,h} v_0^n = \beta^n.$$

Taking Laplace transform, we obtain

$$\tilde{B}\tilde{v}_0(z) = \tilde{\beta}(z),$$

which should give $K_-$ equations and so if we substitute in

$$\tilde{v}_0(z) = \sum_{\nu=1}^{K_-} \gamma_\nu(z) A_\nu(z)$$

we have a $K_- \times K_-$ linear system, which has a unique solution if and only if the homogeneous system

$$\tilde{B}\tilde{v}_0(z) = 0$$

has no non–trivial solution.

In general, the theorem for deciding stability of boundary conditions for hyperbolic systems is as follows:

**Theorem 7.5.** *The initial–boundary value problem for a scheme which is stable for the initial value problem is stable with a given boundary condition if and only if there are no nontrivial admissible solutions of the resolvent equation that satisfy the corresponding homogeneous boundary conditions.*

We need to be careful about admissible solutions when $|z| = 1$. Let's return to the expression $\tilde{p}(z, \kappa(z)) = 0$. This implicitly defines, as least locally, the functions $\kappa_{-,\nu}(z)$ as a function of $z$ (for details see [1] Chapter IV, Section 7, in particular Theorem 7.3). In

particular, we may define $\kappa_{-,\nu}(z)$ for $|z| = 1$ by (analytic) continuation. The $L^2$ requirement in the previous definition means that we must be careful about which solutions are admissible wen $|z| = 1$, as we must distinguish between limits of functions $\kappa_{-,\nu}$ versus $\kappa_{+,\nu}$ (limits of these are *not* admissible).

---

As an illustration of the method consider the Lax–Wendroff scheme for $u_t + au_x = 0$ for $t \geq 0$, $x \geq 0$:

$$\frac{v_m^{n+1} - v_m^n}{k} + a\frac{v_{m+1}^n - v_{m-1}^n}{2h} - \frac{a^2 k}{2}\frac{(v_{m+1}^n - 2v_m^n + v_{m-1}^n)}{h^2} = 0.$$

The exact solution to the PDE is $u(t,x) = u_0(x - at)$. If $a > 0$, the wave is moving to the right and the PDE needs a boundary condition at $x = 0$. On the other hand, if $a < 0$, the wave is coming into 0 from $\infty$ and so the PDE needs no boundary condition, but the scheme requires one numerical boundary condition, since $v_0^n$ cannot be obtained from the scheme, which would require the value at $v_{-1}^n$.

We $z$–transform the difference scheme in time to obtain a recurrence relation for $\tilde{v}_m$, which in this case is again called a resolvent equation. We have

$$\frac{z\tilde{v}_m(z) - \tilde{v}_m(z)}{k} + a\frac{\tilde{v}_{m+1}(z) - \tilde{v}_{m-1}(z)}{2h} - \frac{a^2 k}{2}\frac{(\tilde{v}_{m+1}(z) - 2\tilde{v}_m(z) + \tilde{v}_{m-1}(z))}{h^2} = 0.$$

Rearranging, we have

$$(\frac{a}{2h} - \frac{a^2 k}{2h^2})\tilde{v}_{m+1} + (\frac{1}{k}(z-1) + \frac{a^2 k}{h^2})\tilde{v}_m - (\frac{a^2 k}{2h^2} + \frac{a}{2h})\tilde{v}_{m-1} = 0,$$

so

$$k\tilde{p}(z,\kappa) = (\frac{a\lambda}{2} - \frac{(a\lambda)^2}{2})\kappa + ((z-1) + a^2\lambda^2) - (\frac{a\lambda}{2} + \frac{(a\lambda)^2}{2})\kappa^{-1},$$

where $\lambda = k/h$. Multiply the above by $2\kappa$, then we obtain a quadratic equation for $\kappa$

$$(\alpha - \alpha^2)\kappa^2 + 2((z-1) + \alpha^2)\kappa - (\alpha + \alpha^2) = 0$$

with roots

$$\kappa_\pm(z) = \frac{-((z-1) + \alpha^2)}{(\alpha - \alpha^2)} \pm \frac{\sqrt{((z-1) + \alpha^2)^2 + (\alpha + \alpha^2)(\alpha - \alpha^2)]}}{(\alpha - \alpha^2)},$$

where $\alpha = a\lambda$. The von Neumann stability condition is $|a\lambda| \leq 1$, so $|\alpha| \leq 1$. We are interested in solving the resolvent equation for $|z| > 1$ and since $z - 1$ appears in the expression, the easiest thing to do to determine $K_-$ and $K_+$ (recall they are constants) is perhaps to use $z = 1 + \epsilon$, so that $z - 1 = \epsilon$. Plugging this into the expression, we learn that

$$\kappa_\pm(1+\epsilon) = \frac{-(\epsilon + \alpha^2)}{(\alpha - \alpha^2)} \pm \frac{\sqrt{\epsilon^2 + (1+2\epsilon)\alpha^2}}{(\alpha - \alpha^2)}.$$

Let us see what happens at $\epsilon = 0$. We have

$$\kappa_+(1) = 1$$

and

$$\begin{cases} |\kappa_-(1)| < 1 & \text{for } a < 0 \ (\alpha < 0) \\ |\kappa_-(1)| > 1 & \text{for } a > 0 \ (\alpha > 0). \end{cases}$$

By continuity (in $\epsilon$) of the quadratic formula determining $\kappa_\pm$ we conclude it must be the case that $|\kappa_-(z)| < 1, \forall |z| > 1$, for $a < 0$ and $|\kappa_-(z)| > 1, \forall |z| > 1$ for $a > 0$: E.g. for $a < 0$, for small values of $\epsilon$, $|\kappa_-(1+\epsilon)| < 1$, and this implies $\kappa_-(z)$ (viewed as an (analytic) function of $z$) must satisfy $|\kappa_-(z)| < 1$ for all $\Re e(z) > 1$, since any other possibility would entail the root crossing the unit circle, a forbidden scenario due to the von Neumann stability. Varying $\epsilon$ (allowing it to be complex if necessary), we may cover the entire region $|z| > 1$. On the other hand, examing the equation for large values of $\epsilon > 0$, we see that for $|z| > 1$

$$\begin{cases} |\kappa_+(z)| > 1 & \text{for } a < 0 \ (\alpha < 0) \\ |\kappa_+(z)| < 1 & \text{for } a > 0 \ (\alpha > 0). \end{cases}$$

So for $|z| > 1$, one root is always inside the unit disk and the other is outside and so the solution looks like

$$\tilde{v}_m(z) = \gamma(z)\kappa_*^m(z),$$

where $\kappa_*(z)$ denotes the (analytical) expression for the root inside the unit disk, and is equal to $\kappa_-(z)$ for $a < 0$ and $\kappa_+(z)$ for $a > 0$.

Consider now the boundary condition

$$v_0^{n+1} = v_1^{n+1}.$$

Laplace transforming and plugging in the form of the solution, we get

$$\gamma(z)(1 - \kappa_*(z)) = 0.$$

According to the theorem, we need to check to see if there are any non–trivial solutions to this equation for $\gamma(z)$. Equivalently, we need to check whether $\kappa_*(z) = 1$ for some $|z| \geq 1$. For $|z| > 1$, we know that both roots are strictly inside or outside the unit circle. If we set $\kappa = 1$, we obtain the equation

$$(\alpha - \alpha^2) + 2((z-1) + \alpha^2) - (\alpha + \alpha^2) = 0$$

or

$$2(z - 1) = 0.$$

From which we conclude $z = 1$. Since $\kappa_+(1) = 1$ but $\kappa_*(z) = \kappa_-(z)$ for $a < 0$, whereas $\kappa_*(z) = \kappa_+(z)$ for $a > 0$, we have a non–trivial admissible solution only in the case that $a > 0$.

We conclude for the Lax–Wendroff scheme applied to $u_t + a u_x = 0$, $x \geq 0$, the boundary condition $v_0^{n+1} = v_1^{n+1}$ is

$$\begin{cases} \text{stable} & \text{if } a < 0, \\ \text{unstable} & \text{if } a > 0. \end{cases}$$

However, in the case $a > 0$, as already mentioned, the PDE itself requires a boundary condition and that is the boundary condition that should be implemented by the scheme.

## References


[1] Evgrafov, M. A. *Analytic Functions.* Dover Books on Advanced Mathematics, Dover Publications, Inc., New York, 1966. ISBN: 0-486-63648-8.

[2] E. B. Saff and A. D. Snider. *Fundamentals of Complex Analysis with Applications to Engineering and Science. Third Edition.* Pearson Education, Inc., Upper Saddle River, New Jersey, 2003. ISBN: 0-13-907874-6.

[3] John C. Strikwerda *Finite Difference Schemes and Partial Differential Equations.* Wadsworth, Inc., Belmont, California, 1989. ISBN: 0-534-09984-X.