5.2 Accuracy and Stability for $u_t = c\,u_x$

This section begins a major topic in scientific computing: **Initial-value problems for partial differential equations**. Naturally we start with linear equations that involve only one space dimension x (and time t). The exact solution is $u(x,t)$ and its discrete approximation on a space-time grid has the form $U_{j,n} = U(j\Delta x, n\Delta t)$. We want to know if U is near u—how close they are and how stable U is.

Begin with the simplest wave equation (first-order, linear, constant coefficient):

$$\textbf{One-way wave equation} \qquad \boxed{\frac{\partial u}{\partial t} = c\,\frac{\partial u}{\partial x}\,.} \tag{1}$$

We are given $u(x,0)$ at time $t=0$. We want to find $u(x,t)$ for all $t>0$. For simplicity, these functions are defined on the whole line $-\infty < x < \infty$. There are no difficulties with boundaries (where waves could change direction and bounce back).

The solution $u(x,t)$ will have the typical feature of *hyperbolic equations*: *signals travel at finite speed*. Unlike the second-order wave equation $u_{tt} = c^2 u_{xx}$, this first-order equation $u_t = c\,u_x$ sends signals in one direction only.

Solution for $u(x,0) = e^{ikx}$

Throughout this chapter I will solve for a pure exponential $u(x,0) = e^{ikx}$. **At every time t, the solution remains a multiple Ge^{ikx}.** The growth factor G will depend on the frequency k and the time t, but different frequencies do not mix. Substituting $u = G(k,t)\,e^{ikx}$ into $u_t = c\,u_x$ yields a simple ordinary differential equation for G, because we can cancel e^{ikx}. The derivative of e^{ikx} produces the factor ik:

$$u_t = c\,u_x \quad \text{is} \quad \frac{dG}{dt}e^{ikx} = ikc\,Ge^{ikx} \quad \text{or} \quad \frac{dG}{dt} = ikcG\,. \tag{2}$$

The growth factor is $G(k,t) = e^{ikct}$. The initial value is $G = 1$.

$$\textbf{An exponential solution to } \frac{\partial u}{\partial t} = c\,\frac{\partial u}{\partial x} \ \textbf{ is } \ u(x,t) = e^{ikct}e^{ikx} = \boldsymbol{e^{ik(x+ct)}}\,. \tag{3}$$

Immediately we see two important features of this solution:

1. The growth factor $G = e^{ikct}$ has absolute value $|G| = 1$.
2. The initial function e^{ikx} moves to the left with fixed velocity c, to $e^{ik(x+ct)}$.

The initial value at the origin is $u(0,0) = e^{ik0} = 1$. This value $u = 1$ appears at all points on the line $x + ct = 0$. **The initial data propagates along the characteristic lines $x + ct = \textit{constant}$**, in Figure 5.3. Right now we know this fact for the special solutions $e^{ik(x+ct)}$. Soon we will know it for all solutions.

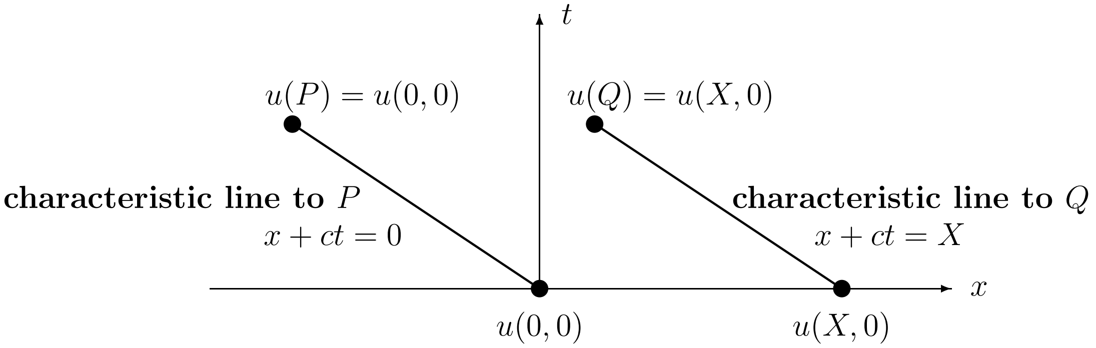


Figure 5.3: The solution $u(x,t)$ moves left with speed c, along characteristic lines.

Figure 5.3 shows the travel path of the solution in the x-t plane (we are introducing the characteristic lines). Figure 5.4 will graph the solution itself at times 0 and t. That step function combines exponentials $e^{ik(x+ct)}$ for different frequencies k. By linearity we can add those solutions.

Solution for Every $u(x,0)$

In almost all partial differential equations, the solution changes shape as it travels. Here the shape stays the same. All pure exponentials travel at the same velocity c, so *every* initial function moves with that velocity. We can write down the solution:

$$\textbf{General solution} \qquad \frac{\partial u}{\partial t} = c\,\frac{\partial u}{\partial x} \quad \text{is solved by} \quad \boldsymbol{u(x,t) = u(x+ct,0)}\,. \tag{4}$$

The solution is a function only of $x + ct$. That makes it constant along characteristic lines, where $x + ct$ is constant. This dependence on $x + ct$ also makes it satisfy the equation $u_t = c\,u_x$, by the chain rule. If we take $u = (x + ct)^n$ as an example, the extra factor c appears in $\partial u/\partial t$:

$$\frac{\partial u}{\partial x} = n\,(x+ct)^{n-1} \quad \text{and} \quad \frac{\partial u}{\partial t} = cn\,(x+ct)^{n-1} \quad \text{which is} \quad c\,\frac{\partial u}{\partial x}\,.$$

A Taylor series person would combine those powers (different n) to produce a large family of solutions. A Fourier series person combines exponentials (different k) to produce an even larger family. In fact *all* solutions are functions of $x + ct$ alone.

Here are two important initial functions—a light flashes or a dam breaks.

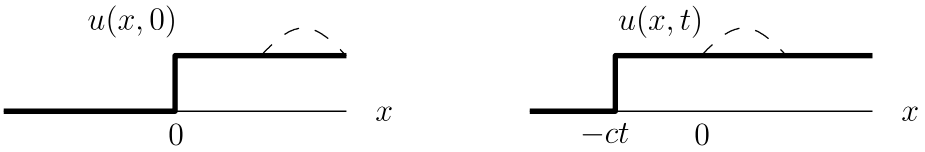
Example 1 $u(x,0) =$ delta function $\delta(x) =$ **flash of light** at $x = 0, t = 0$

By our formula (4), the solution is $u(x,t) = \delta(x + ct)$. The light flash reaches the point $x = -c$ at the time $t = 1$. It reaches $x = -2c$ at the time $t = 2$. The impulse is traveling to the left at speed $|dx/dt| = c$. In this example all frequencies k are present in equal amounts, because the Fourier transform of a delta function is a constant.

Notice that a point goes dark again as soon as the flash passes through. This is the Huygens principle in 1 and 3 dimensions. If we lived in two or four dimensions, the wave would not pass all at once and we wouldn't see clearly.

Example 2 $u(x,0) =$ step function $S(x) =$ **wall of water** at $x = 0, t = 0$

The solution $S(x+ct)$ is the moving step function in Figure 5.4. The wall of water travels to the left (*one-way wave*). At time t, the "tsunami" reaches the point $x = -ct$. The flash of light will get there first, because its speed c is greater than the tsunami speed. That is why a warning is possible for an approaching tsunami.



initial profile $S(x)$ at $t = 0$ later profile $S(x+ct)$ at time t

Figure 5.4: The wall travels left with velocity c (all waves e^{ikx} do too).

An actual tsunami is described by the nonlinear "shallow water equations" that come later. The feature of *finite speed* still holds.

Finite Difference Methods for $u_t = c\,u_x$

The one-way wave equation is a perfect example for creating and testing finite difference approximations. We can replace $\partial u/\partial t$ by a forward difference with step Δt. Here are four choices for the discrete form of $\partial u/\partial x$ at meshpoint $i\Delta x$:

1. **Forward** $= \dfrac{U_{i+1} - U_i}{\Delta x} =$ **upwind**: Low accuracy, conditionally stable for $c > 0$.

2. **Centered** $= \dfrac{U_{i+1} - U_{i-1}}{2\Delta x}$: Unstable after a few steps as we will prove!

3. **Lax-Friedrichs**: (20) has low accuracy, conditionally stable also for $c < 0$.

4. **Lax-Wendroff**: (14) has extra accuracy, conditionally stable also for $c < 0$.

The list doesn't end there. We have reached a central problem of scientific computing, to construct approximations that are *stable* and *accurate* and *fast*. That topic can't be developed on one page, especially when we move to nonlinear equations.

Conditionally stable means that the time step Δt is restricted. The need for this restriction was noticed by Courant, Friedrichs, and Lewy. When the space difference reaches no further than $x+\Delta x$, there is an automatic stability restriction:

$$\textbf{CFL requirement for stability} \qquad \boldsymbol{r = c\,\frac{\Delta t}{\Delta x} \le 1\,.} \tag{5}$$

That number $c\,\Delta t/\Delta x$ is often called the *Courant number*. (It was really Lewy who recognized that $r \le 1$ is necessary for stability and convergence.) The reasoning is straightforward, based on using the initial value that controls $u(x,t)$:

> The true solution at (x,t) equals the initial value $u(x+ct,0)$. Taking n discrete steps to reach $t = n\,\Delta t$ uses information on the initial values as far out as $x+n\,\Delta x$. If $x+ct$ is further than $x+n\,\Delta x$, the method can't work:

$$\textbf{CFL condition} \quad x + c\,t \le x + n\,\Delta x \quad \text{or} \quad c\,n\,\Delta t \le n\,\Delta x \quad \text{or} \quad r = c\,\frac{\Delta t}{\Delta x} \le 1\,. \tag{6}$$

If the difference equation uses $U(x+2\Delta x,t)$, then CFL relaxes to $r \le 2$.

A particular finite difference equation might require a tighter restriction on Δt for stability. It might even be unstable for all ratios r (we hope not). The only route to *unconditional stability* for all Δt is an *implicit method*, which computes x-differences at the new time $t+\Delta t$. This will be useful later for diffusion terms like u_{xx}. For advection terms (first derivatives), explicit methods with a CFL limitation are usually accepted because a much larger Δt would lose accuracy as well as stability.

To repeat, if $r > 1$ then the finite difference solution at x,t does not use initial value information near the correct point $x^* = x + ct$. Hopeless.

Accuracy of the Upwind Difference Equation

Linear problems with constant coefficients are the ones to understand first. Exactly as for differential equations, we can follow each pure exponential e^{ikx}. After a single time step, there will be a growth factor in $U(x,\Delta t) = Ge^{ikx}$. That growth factor $G(k,\Delta t,\Delta x)$ may have magnitude $|G| < 1$ or $|G| > 1$. This will control stability or instability. The order of accuracy (if we compute in the k-ω domain) comes from comparing G with the true factor $e^{ikc\Delta t}$ from the differential equation.

We now determine that the order of accuracy is $p = 1$ for the *upwind method*.

$$\textbf{Forward differences} \qquad \frac{U(x,t+\Delta t) - U(x,t)}{\Delta t} = c\,\frac{U(x+\Delta x,t) - U(x,t)}{\Delta x}. \tag{7}$$

We will test the accuracy in the x-t domain and then the k-ω domain. Either way we use Taylor series to check the leading terms. Substituting the true solution $u(x,t)$ in place of $U(x,t)$, its forward differences are

$$\textbf{Time} \qquad \frac{1}{\Delta t}\left[u(x,t+\Delta t) - u(x,t)\right] \;=\; u_t + \frac{1}{2}\Delta t\,u_{tt} + \cdots \tag{8}$$

$$\textbf{Space} \qquad \frac{c}{\Delta x}\left[u(x+\Delta x,t) - u(x,t)\right] \;=\; c\,u_x + \frac{1}{2}c\,\Delta x\,u_{xx} + \cdots \tag{9}$$

On the right side, $u_t = c\,u_x$ is good. One more derivative gives $u_{tt} = c\,u_{xt} = c^2\,u_{xx}$. Notice c^2. Then $\Delta t\,u_{tt}$ matches $c\,\Delta x\,u_{xx}$ *only in the special case* $c\Delta t = \Delta x$:

$$\frac{1}{2}\Delta t\, c^2\, u_{xx} \quad \text{equals} \quad \frac{1}{2}c\,\Delta x\, u_{xx} \quad \text{only if} \quad r = \frac{c\Delta t}{\Delta x} = 1.$$

For any ratio $r \neq 1$, the difference between (8) and (9) has a ***first-order error***. Let me show this also in the k-ω Fourier picture and then improve to second-order.

Fix the ratio $r = c\Delta t/\Delta x$ as $\Delta x \to 0$ and $\Delta t \to 0$. In the difference equation (7), write each new value at time $t + \Delta t$ as a combination of two old values of U:

Difference equation $$U(x, t + \Delta t) = (\mathbf{1} - \boldsymbol{r})\, U(x, t) + \boldsymbol{r}\, U(x + \Delta x, t)\,. \qquad (10)$$

Starting from $U(x, 0) = e^{ikx}$ we quickly find the growth factor $\boldsymbol{G}$ at time Δt:

After 1 step $$(1-r)e^{ikx} + r\, e^{ik(x+\Delta x)} = \left[\mathbf{1} - \boldsymbol{r} + \boldsymbol{r}\, \boldsymbol{e^{ik\Delta x}}\right] e^{ikx} = \boldsymbol{G}\, e^{ikx}\,. \qquad (11)$$

To test the accuracy, compare this $G = G_{\textbf{approx}}$ to the exact growth factor $e^{ick\Delta t}$. Use the power series $1 + x + x^2/2! + \cdots$ for any e^x:

Accuracy $$G_{\textbf{approx}} = 1 - r + r\, e^{ik\Delta x} = (1-r) + r + r(ik\Delta x) + \frac{1}{2}r\,(ik\Delta x)^2 + \cdots$$

$$G_{\textbf{exact}} = e^{ick\Delta t} = e^{irk\Delta x} = 1 + irk\Delta x + \frac{1}{2}(irk\Delta x)^2 + \cdots \qquad (12)$$

The first terms agree as expected. Forward differences replaced derivatives, and the method is ***consistent***. We saw $u_t = c\, u_x$ in comparing (8) with (9). The next terms do *not* agree unless $r = r^2$:

Compare $\frac{1}{2}r(ik\Delta x)^2$ with $\frac{1}{2}r^2(ik\Delta x)^2$. **Single-step error of order $(\boldsymbol{k\Delta t})^{\mathbf{2}}$.** (13)

After $1/\Delta t$ steps, those errors of order $k^2(\Delta t)^2$ give a final error $O(k^2\Delta t)$. Forward differences are only first order accurate, and so is the whole method.

The special case $r = 1$ means $c\Delta t = \Delta x$. The difference equation is *exactly correct*. The true and approximate solutions at $(x, \Delta t)$ are both $u(x + \Delta x, 0)$. We are *on the characteristic line* in Figure 5.3. This is an interesting special case (the golden Δt, but hard to repeat in scientific computing when c varies).

Conclusion Except when $r = r^2$, ***the upwind method is first-order accurate***.

Higher Accuracy for Lax-Wendroff

To upgrade the accuracy, we need to match the $\frac{1}{2}\Delta t\, u_{tt}$ error term in the forward time difference by an additional space difference that gives $\frac{1}{2}\Delta t\, c^2 u_{xx}$. This is achieved by the **Lax-Wendroff method**:

$$\frac{U(x, t+\Delta t) - U(x,t)}{\Delta t} = c\,\frac{U(x+\Delta x, t) - U(x - \Delta x, t)}{2\Delta x} + \frac{\Delta t}{2}c^2\left(\frac{U(x+\Delta x, t) - 2U(x,t) + U(x-\Delta x, t)}{(\Delta x)^2}\right). \qquad (14)$$

Substituting the true solution, that second difference produces $\frac{1}{2}c^2\Delta t\, u_{xx}$ plus higher order terms. This cancels the $\frac{1}{2}\Delta t\, u_{tt}$ error term in the time difference, computed in equation (8). (Remember $u_{tt} = cu_{xt} = c^2 u_{xx}$. The centered difference has no Δx term.) Thus Lax-Wendroff has **second-order accuracy**.

To see this in the k-ω frequency domain, rewrite the LW difference equation (14):

$$U(x,t+\Delta t) = (1-r^2)U(x,t) + \frac{1}{2}(r^2+r)U(x+\Delta x,t) + \frac{1}{2}(r^2-r)U(x-\Delta x,t)\,. \quad (15)$$

Substitute $U(x,t) = e^{ikx}$ to find the one-step growth factor G at time $t+\Delta t$:

Growth factor for LW $$G = (\mathbf{1}-\boldsymbol{r}^{\mathbf{2}}) + \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{r}^{\mathbf{2}}+\boldsymbol{r})e^{ik\Delta x} + \frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{r}^{\mathbf{2}}-\boldsymbol{r})e^{-ik\Delta x}.$$

Expanding $e^{ik\Delta x}$ and $e^{-ik\Delta x}$ in powers of $ik\,\Delta x$, this becomes

$$G = 1 + r(ik\Delta x) + \frac{1}{2}r^2(ik\Delta x)^2 + O(k\Delta x)^3\,. \quad (16)$$

Comparing with $G_{\textbf{exact}} = e^{irk\Delta x}$ in equation (12), three terms agree. So the one-step error is of order $(k\Delta x)^3$. After $1/\Delta t$ steps the second-order accuracy of Lax-Wendroff is confirmed.

Figure 5.5 shows by actual computation the improvement in accuracy. For a first-order method, the "wall of water" is smeared out. High frequencies have growth factors $|G(k)|$ much smaller than 1. There is too much dissipation. For the first-order Lax-Friedrichs method, the dissipation is even worse (Problem 2). The second-order Lax-Wendroff method stays much closer to the discontinuity. But it's not perfect—those oscillations are not good.

For an ideal difference equation, we want to add enough dissipation very *close to the shock*, to avoid that oscillation (the Gibbs phenomenon). A lot of thought has gone into high resolution methods, to capture shock waves cleanly.

Greater accuracy is achievable by including more terms in the difference equation. If we go from the three terms in Lax-Wendroff to five terms, we can reach fourth-order accuracy. If we use *all* values $U(j\Delta x, n\Delta t)$ at every time step, which requires more work, we can achieve **spectral accuracy**. Then the error decreases faster than any power of Δx, provided $u(x,t)$ is smooth enough to allow derivatives of all orders. Section ____ gives a separate discussion of this **spectral method**.

Stability of the Four Finite Difference Methods

Now we turn from accuracy to stability. Accuracy requires G to stay close to the true $e^{ick\Delta t}$. Stability requires G to stay *inside the unit circle*. We need $|G| \le 1$ for all frequencies k or the finite difference approximation $G^n e^{ikx}$ will blow up.

We now check whether or not $|G| \le 1$, in the four methods.

1. Forward differences in space and time: $\boldsymbol{\Delta U/\Delta t = c\,\Delta U/\Delta x}$ (upwind).

Recall from equation (11) that $G = 1 - r + re^{ik\Delta x}$. If the Courant number r is between 0 and 1, the triangle inequality gives $|G| \le 1$:

$$\textbf{Stability for } \boldsymbol{0 \le r \le 1} \qquad |G| \le |1-r| + |re^{ik\Delta x}| = 1 - r + r = 1\,. \tag{17}$$

This sufficient condition $0 \le c\,\Delta t/\Delta x \le 1$ is exactly the same as the Courant-Friedrichs-Lewy necessary condition! They reasoned that $U(x, n\Delta t)$ depends on the initial values between x and $x + n\Delta x$. That **domain of dependence** must include the point $x + cn\Delta t$. (Otherwise, changing the initial value at the point $x + cn\Delta t$ would change the true solution u but not the approximation U.) Then $cn\Delta t$ must lie between 0 and $n\Delta x$, which means that $0 \le r \le 1$.

Figure 5.5 shows G in the stable case $r = \frac{2}{3}$ and the unstable case $r = \frac{4}{3}$ (when Δt is too large). As k varies, and $e^{ik\Delta x}$ goes around a unit circle, the complex number $G = 1 - r + re^{ik\Delta x}$ goes in a circle of radius r. The center is $1 - r$. Always $G = 1$ at zero frequency (constant solution, no growth).

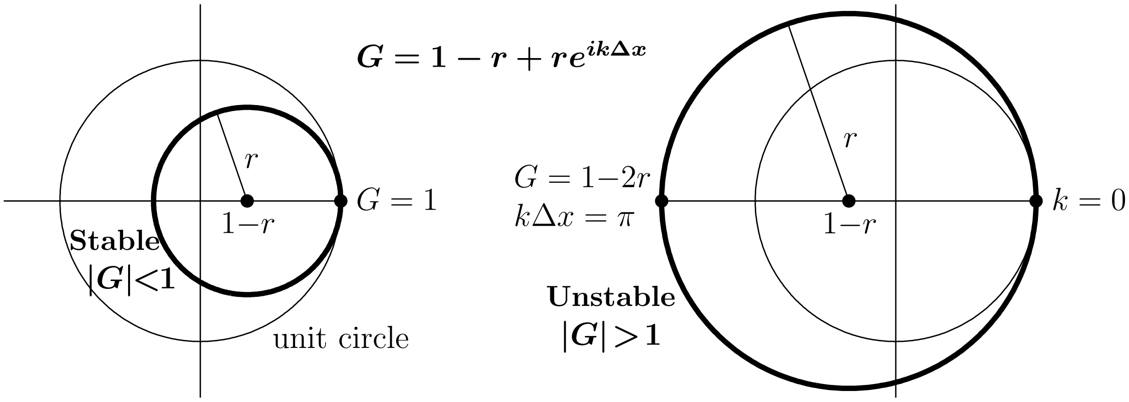


Figure 5.5: Stable (upwind) and unstable (centered): CFL numbers $r = \frac{2}{3}$ and $r = \frac{4}{3}$.

2. Forward difference in time, centered difference in space.

This combination is never stable! The shorthand $U_{j,n}$ will stand for $U(j\Delta x, n\Delta t)$:

$$\frac{U_{j,n+1} - U_{j,n}}{\Delta t} = c\,\frac{U_{j+1,n} - U_{j-1,n}}{2\Delta x} \quad \text{or} \quad U_{j,n+1} = U_{j,n} + \frac{r}{2}\left(U_{j+1,n} - U_{j-1,n}\right). \tag{18}$$

Those coefficients 1 and $r/2$ and $-r/2$ go into the growth factor G, when the solution is a pure exponential and e^{ikx} is factored out:

$$\textbf{Unstable: } \boldsymbol{|G| > 1} \qquad G = 1 + \frac{r}{2}e^{ik\Delta x} - \frac{r}{2}e^{-ik\Delta x} = \boldsymbol{1 + ir\,\textbf{sin}\,k\Delta x}\,. \tag{19}$$

The real part is 1. The magnitude is $|G| \ge 1$. Its graph is on the left side of Figure 5.6.

3. Lax-Friedrichs Method (upwind-downwind).

We can recover stability for centered differences by changing the time difference. Replace $U_{j,n}$ by the average $\frac{1}{2}(U_{j+1,n} + U_{j-1,n})$ of its neighbors:

$$\textbf{Lax-Friedrichs} \qquad \frac{U_{j,n+1} - \frac{1}{2}(U_{j+1,n} + U_{j-1,n})}{\Delta t} = c\,\frac{U_{j+1,n} - U_{j-1,n}}{2\Delta x}\,. \tag{20}$$

Two old values $U_{j+1,n}$ and $U_{j-1,n}$ produce each new value $U_{j,n+1}$. Moving terms to the right-hand side, the coefficients are $\frac{1}{2}(1+r)$ and $\frac{1}{2}(1-r)$. The growth factor is

$$G = \frac{1+r}{2}\,e^{ik\Delta x} + \frac{1-r}{2}\,e^{-ik\Delta x} = \mathbf{cos}\,\boldsymbol{k\Delta x} + \boldsymbol{ir}\,\mathbf{sin}\,\boldsymbol{k\Delta x}\,. \tag{21}$$

The absolute value is $|G|^2 = (\cos k\Delta x)^2 + r^2(\sin k\Delta x)^2$. In Figure 5.6, $|G| \le 1$ when $r^2 \le 1$. This stability condition agrees again with the CFL condition.

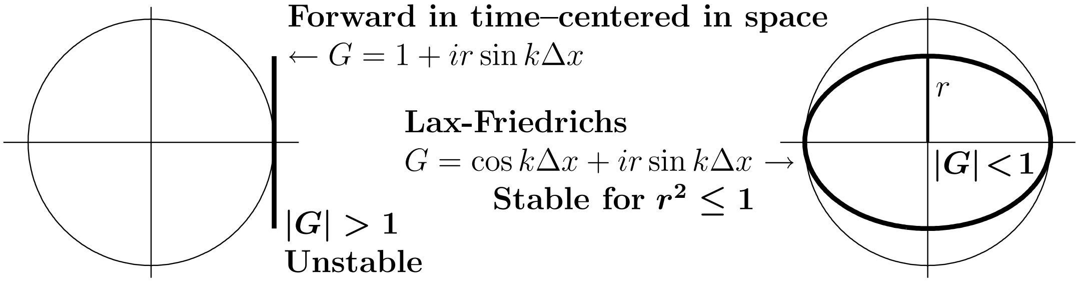

Forward in time–centered in space
$\leftarrow G = 1 + ir\sin k\Delta x$
Lax-Friedrichs
$G = \cos k\Delta x + ir\sin k\Delta x \rightarrow$
Stable for $\boldsymbol{r^2 \le 1}$
$\boldsymbol{|G| > 1}$
Unstable
r
$\boldsymbol{|G| < 1}$

Figure 5.6: Equation (18) is unstable for all r. Equation (20) is stable for $r^2 \le 1$.

Notice that $\boldsymbol{c}$ and $\boldsymbol{r}$ can be negative. The wave can go either way! This will be useful for the two-way wave equation, but the accuracy is still first-order. The Lax-Friedrichs G matches the next term in the exact growth factor only if $r^2 = 1$:

$$G = \cos k\Delta x + ir\sin k\Delta x = 1 + irk\Delta x - \frac{1}{2}(k\Delta x)^2 + \cdots \tag{22}$$

$$G_{\textbf{exact}} = e^{ikr\Delta x} = 1 + irk\,\Delta x + \frac{1}{2}i^2r^2(k\,\Delta x)^2 + \cdots$$

In the exceptional cases $r = 1$ and $r = -1$, G agrees with G_{exact}. Staying exactly on the characteristic line, $U_{j,n+1}$ matches the true $u(j\Delta x, t+\Delta t)$. For $r^2 < 1$, Lax-Friedrichs has an important advantage and disadvantage:

Good Each new $U_{j,n+1}$ is a **positive** combination of old values.

Not good The accuracy is only **first-order**.

Problem 6 will show that second-order is impossible with positive coefficients.

4. Lax-Wendroff Method (second-order accurate).

The LW difference equation (14) combines $U_{j,n}$ and $U_{j-1,n}$ and $U_{j+1,n}$ to compute the new value $U_{j,n+1}$. The coefficients of these three old values go into G:

$$\textbf{Lax-Wendroff} \qquad G = (1-r^2) + \frac{1}{2}(r^2+r)e^{ik\Delta x} + \frac{1}{2}(r^2-r)e^{-ik\Delta x}. \qquad (23)$$

This is $G = 1 - r^2 + r^2 \cos k\Delta x + ir \sin k\Delta x$. At the dangerous frequency $k\Delta x = \pi$, the growth factor is $1 - 2r^2$. That stays above -1 if $r^2 \le 1$.

Problem 5 shows that $|G| \le 1$ for every $k\Delta x$. **Lax-Wendroff is stable whenever the CFL condition $r^2 \le 1$ is satisfied**. Again the wave can go either way (or both ways) since c and r can be negative. This is the most accurate of the five methods in Figure 5.7.

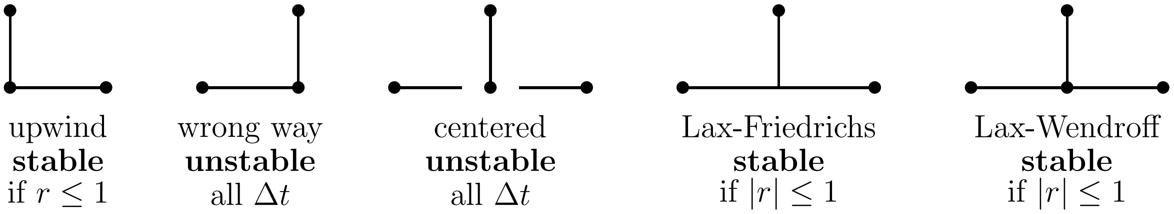


Figure 5.7: Difference methods for the one-way wave equation $u_t = cu_x$.

Equivalence of Stability and Convergence

Does the discrete solution U approach the true solution u as $\Delta t \to 0$? The expected answer is yes. But there are two requirements for convergence, and one of them—*stability*—is by no means automatic. The other requirement is *consistency*—the discrete problem must approximate the correct continuous problem. The fact that these two properties are sufficient for convergence, and also *necessary* for convergence, is the ***fundamental theorem of numerical analysis***:

Lax equivalence theorem Stability is equivalent to convergence, for a consistent approximation to a well-posed linear problem.

Lax proved the equivalence theorem for initial-value problems. The rate of convergence is given in (26). The theorem is equally true for boundary-value problems, and for the approximation of functions, and for the approximation of integrals. It applies to every discretization, when the given problem $Lu = f$ is replaced by $L_h U_h = f_h$. Assuming the inputs f and f_h are close, we will prove that u and U_h are close—provided L_h is stable. The key points of the proof take only a few lines when the equation is linear, and you will see the essence of this fundamental theorem.

Suppose f is changed to f_h and L is replaced by L_h. The requirements are

Consistency: $f_h \to f$ and $L_h u \to Lu$ for smooth solutions u.
Well-posed: The inverse of L is bounded: $\|u\| = \|L^{-1}f\| \le C\|f\|$.
Stability: The inverses L_h^{-1} remain uniformly bounded: $\|L_h^{-1}f_h\| \le C\|f_h\|$.

Under those conditions, the approximation $U_h = L_h^{-1}f_h$ will approach u as h goes to zero. We subtract and add $L_h^{-1}Lu = L_h^{-1}f$ when u is smooth:

Convergence $$u - U_h = L_h^{-1}(L_h u - Lu) + L_h^{-1}(f - f_h) \to 0\,. \qquad (24)$$

Consistency controls the quantities in parentheses (they go to zero). Stability controls the operators L_h^{-1} that act on them. Well-posedness controls the approximation of all solutions by smooth solutions. Then always U_h converges to u.

If stability fails, there will be an input for which the approximations $U_h = L_h^{-1}f$ are not bounded. The *uniform boundedness theorem produces this bad* f, from the inputs f_h on which instability gives $\|L_h^{-1}f_h\| \to \infty$. Convergence fails for this f.

A perfect equivalence theorem goes a little further, after careful definitions:

Consistency + Stability $\Longleftrightarrow$ Well-posedness + Convergence.

Our effort will now concentrate on initial-value problems, to estimate the error (the convergence rate) in $u - U_h$. The parameter h becomes Δt. We take n steps.

The Rate of Convergence

Consistency means that the error at each time step goes to zero as the mesh is refined. Our Taylor series estimates have done more: **The order of accuracy gives the rate** that this one-step error goes to zero. The problem is to extend this local rate to a global rate of convergence, accumulating the errors over n time steps.

Let me write S for a single finite difference step, so $\boldsymbol{U(t+\Delta t) = S\,U(t)}$. The corresponding step for the differential equation will be $\boldsymbol{u(t+\Delta t) = R\,u(t)}$. Then consistency means that Su is close to Ru, and the order of accuracy p tells how close:

Accuracy of discretization $\|Su - Ru\| \le C_1(\Delta t)^{p+1}$ for smooth solutions u.
Well-posed problem $\|R^n u\| \le C_2\|u\|$ for $n\,\Delta t \le T$.
Stable approximations $\|S^n U\| \le C_3\|U\|$ for $n\,\Delta t \le T$.

The difference between $U = S^n u(0)$ and the true $u = R^n u(\quad)$ is $(S^n - R^n)u(\quad)$. The key idea is a "telescoping identity" that involves n single-step differences $S - R$:

$$S^n - R^n = S^{n-1}(S-R) + S^{n-2}(S-R)R + \cdots + (S-R)R^{n-1}\,. \qquad (25)$$

Each of those n terms has a clear meaning. First, a power R^k carries $u(0)$ to the true solution $u(k\,\Delta t)$. Then $(S-R)u(k\,\Delta t)$ gives the error at step k of order $(\Delta t)^{p+1}$. Then powers of S carry that one-step error forward to time $n\Delta t$. By stability, this amplifies the error by no more than C_3. There are $n \le T/\Delta t$ steps. **The final rate of convergence for smooth solutions is $(\boldsymbol{\Delta t})^{\boldsymbol{p}}$:**

$$\|U(n\,\Delta t) - u(n\,\Delta t)\| = \|(S^n - R^n)u(0)\| \le C_1C_2C_3\,\frac{T}{\Delta t}(\Delta t)^{p+1} = C_1C_2C_3\,T(\Delta t)^p\,. \tag{26}$$

Notice how smoothness was needed in the Taylor series (8) and (9), when Δt and Δx multiplied u_{tt} and u_{xx}. That first-order accuracy would not apply if u or u_t or u_x had a jump. Still the order of accuracy p gives a practical estimate of the overall approximation error $u - U$. The problem of scientific computing is to get beyond $p = 1$ while maintaining stability and speed.

Problem Set 5.2

1 Integrate $u_t = c\,u_x$ from $-\infty$ to ∞ to prove that mass is conserved: $dM/dt = 0$. Multiply by u and integrate $uu_t = c\,uu_x$ to prove that energy is also conserved:

$$M(t) = \int_{-\infty}^{\infty} u(x,t)\,dx \quad \text{and} \quad E(t) = \tfrac{1}{2}\int_{-\infty}^{\infty} (u(x,t))^2\,dx \quad \text{stay constant in time.}$$

2 Substitute the true $u(x,t)$ into the Lax-Friedrichs method (21) and use $u_t = cu_x$ and $u_{tt} = c^2u_{xx}$ to find the coefficient of the *numerical dissipation* u_{xx}.

3 The difference equation $U_{j,n+1} = \sum a_mU_{j+m,n}$ has growth factor $G = \sum a_me^{imk\Delta x}$. Show *consistency* with $e^{ick\Delta t}$ (first-order accuracy at least) when $\sum a_m = 1$ and $\sum ma_m = c\Delta t/\Delta x = r$.

4 The condition for second-order accuracy is $\sum m^2a_m = r^2$, from the Taylor series. Check this for Lax-Wendroff with $a_0 = 1-r^2, a_1 = \frac{1}{2}(r^2+r), a_{-1} = \frac{1}{2}(r^2-r)$. With *nonnegative coefficients* $a_m \ge 0$, the Schwarz inequality $(\sum m\sqrt{a_m}\sqrt{a_m})^2 \le (\sum m^2a_m)(\sum a_m)$ becomes an equality $r^2 = r^2$. This equality only happens if $m\sqrt{a_m} = (\text{constant})\sqrt{a_m}$. *Second-order is impossible with* $a_m \ge 0$, unless the difference equation has only one term $U_{j,n+1} = U_{j+m,n}$.

5 The Lax-Wendroff method has $G = 1 - r^2 + r^2\cos k\Delta x + ir\sin kx$. Square the real and imaginary parts to get (eventually!) $|G|^2 = 1-(r^2-r^4)(1-\cos k\Delta x)^2$. Prove stability, that $|G|^2 \le 1$ if $r^2 \le 1$.

6 Suppose the coefficients in a linear differential equation change as t changes. The one-step solution operators become S_k and R_k, for the step from $k\,\Delta t$ to $(k+1)\Delta t$. After n steps, products replace the powers S^n and R^n in U and u:

$$U(n\,\Delta t) = S_{n-1}S_{n-2}\ldots S_1S_0\,u(0) \quad \text{and} \quad u(n\,\Delta t) = R_{n-1}R_{n-2}\ldots R_1R_0\,u(0)\,.$$

Change the telescoping formula (25) *to produce this* $U - u$. Which parts are controlled by stability? Which parts by well-posedness (= stability of the differential equation)? Consistency still controls $S_k - R_k$.

7 Even an unstable method will converge to the true solution $u = e^{ick\Delta t}e^{ikx}$ for each separate frequency k. Consistency assures that the single-step growth factor G is $1 + ick\,\Delta t + O(\Delta t)^2$. Then for $t = n\,\Delta t$,

$$G^n = \left(1 + \frac{ickt}{n} + O(\frac{1}{n^2})\right)^n \longrightarrow e^{ickt} \quad \text{which is convergence.}$$

How can we have convergence for each $u(0) = e^{ikx}$ and still prove divergence for a combination of frequencies $u(0) = \sum_{-\infty}^{\infty} c_k\, e^{ikx}$?

8 The upwind method with $r > 1$ is unstable because the CFL condition fails. By Problem 3, it does converge to $e^{ik(x+ct)}$ based on values of $u(x,0) = e^{ikx}$ *that do not reach as far as* $x + ct$. The method must be finding a "correct" extrapolation of e^{ikx}. So propose an initial $u(x,0)$ for which convergence fails.