31 Numerical Solution of PDE

Although we have been discussing analytical methods to solve PDE, most problems are intractable by exact methods. In this section elementary numerical methods to solve PDE are outlined. We require a numerical scheme to be stable and consistent (i.e., converging to the original problem in the continuum limit). This is a section for ABC of numerical analysis.

Key words: discretization, consistency, stability, convergence, von Neumann condition, Courant-Friedrichs-Lewy condition

Summary:
(1) There are two major methods to discretize a continuum problem: the Galerkin method and sampling at space-time lattice points (**31.2**). There can be many unconventional discretization schemes (**31.4**).
(2) Any discrete scheme must recover the original problem in the continuum limit (consistency of the scheme). If the solution to a scheme is bounded, then the scheme is said to be stable. For linear problems Consistency and stability imply convergence of the scheme (i.e., the solution to the scheme converges to the true solution in the continuum limit) (**31.7**).
(3) Stability conditions for a scheme may be understood, roughly, by the condition that physical propagation speed of the signal does not outrun the numerical propagation speed (**31.9**, **31.11**).

31.1 Discretization. To use computers to solve a differential equation, unless we use symbolic manipulation, we must discretize everything and express quantities in a finite number of rational numbers. Thus the fundamental question of numerical computations of differential equations is how faithful this map to the discrete world is.
Numerical analysis is a discipline to analyze numerical algorithms and is as old as analysis itself. Already Newton discussed a series expansion method to solve ODE in his first calculus paper (1669). Euler introduced discretization methods in 1743.

Discussion.

Consider[389]

$$\frac{du}{dt} = f(u), \tag{31.1}$$

where f satisfies $f(0) = f(1) = 0$, $f(u) > 0$ for $u \in (0,1)$ and $f(u) < 0$ for $1 < u < \kappa$ for some positive $\kappa > 1$. Then, its Euler differencing result

$$u_{n+1} = u_n + \Delta t f(u_n) \tag{31.2}$$

exhibits chaos for $\Delta t > c_1$ for some positive c_1. Here 'exhibiting chaos' means that the solution has a 'natural' relation to random numbers (or the outcome of coin-tossing).[390]

(B) Consider the following *logistic equation*

$$\frac{du}{dt} = u(1-u). \tag{31.3}$$

(1) Solve this equation with the initial condition $u = u_0 \in (0,1)$ analytically.

(2) Get the following type of difference equation with the aid of the center differencing scheme:

$$u_{n+1} = v_n + \alpha u_n(1-u_n), \quad v_{n+1} = u_n, \tag{31.4}$$

where $\alpha = 2\Delta t$, $u_n = u(n\Delta t)$ and $v_n = u_{n-1}$.

(3) The equation (31.4) defines a map from $\boldsymbol{R}^2$ into itself. The map exhibits chaos irrespective of the size of Δt.[391] A more careful statement is as follows. Let time T be fixed and $N \equiv T/\Delta t$. If $\Delta t \to 0$, then up to N there is no pathological behavior. However, if Δt is fixed, then for sufficiently large N (consequently for large T), pathological behavior will show up.

(4) The equation (31.4) converges to (more generally, see **31.3** Discussion)

$$\frac{du}{dt} = v(1-v), \quad \frac{dv}{dt} = u(1-u). \tag{31.5}$$

This equation does not exhibit chaos, but is unstable near $u = v = 1$.

31.2 Two major methods of discretization. There are two major methods to map a continuous problem to a discrete problem. One is the sampling method (recall Green's approach →**1.8**), and the other is the Fourier expansion method.

The sampling method tries to represent a function $f(x)$ by a set of function values sampled at appropriately located sampling points, and is usually called "the discrete variable method." We have already used its primitive version in **1** (**1.15**, **1.18**, **1.20**).

The Fourier expansion method tries to describe a function $f(x)$ as a truncated generalized Fourier expansion $f_N(x)$ (→**20.14**). A typical method is the one called the *Galerkin method*: Put $f_N(x) = \sum_{n=1}^N a_n\phi_n(x)$,

[389] M. Yamaguti and H. Matano, Euler's finite difference scheme and chaos, Proc. Japan Acad. **55** Ser.A, 78-80 (1979).

[390] Y. Oono, Period $\neq 2^n$ implies chaos, Prog. Theor. Phys. **59**, 1029-1030 (1978).

[391] S. Ushiki, Central differencing scheme and chaos, Physica D **4**, 407-424 (1982).

where ϕ_n denotes orthonormal functions (→**20.10**), into the original equation. Then, multiply $\phi_m(x)$ and integrate over x. This will give a set of equations for the Fourier coefficients. This is a finite set of algebraic equations, so there are many ways to solve it.[392]

31.3 Consistency, stability and convergence. If the discretization scheme recovers the original equation in the limit which recovers a function from its discretized version, we say the method is *consistent.* If the discretized solution is bounded in terms of the input data (initial condition, etc), we say the method is *stable.* Consistency and stability imply the convergence of the scheme. That is, if a numerical scheme is consistent and stable, then the scheme gives the solution which converges to the true solution of the original continuous problem in the limit recovering a function from its discretized version. There are consistent but unstable schemes.[393]

Discussion.

Probably the most famous example is the center differencing scheme:[394]
Since $dx/dt \simeq [x(t_{n+1}) - x(t_{n-1})]/2h$, where h is the time increment $t_{n+1} - t_n = h$ for all n, we might be able to rewrite $dx/dt = f(x)$ as

$$\frac{x(t_{n+1}) - x(t_{n-1})}{2h} = f(x(t_n)). \tag{31.6}$$

The scheme is called the *center differencing scheme.* It is known that this equation converges to the following simultaneous equation:

$$\frac{dx}{dt} = f(y), \quad \frac{dy}{dt} = f(x). \tag{31.7}$$

If $x = y$ is stable, then there is no problem, but often this is not the case. The method doubles the dimensionality of the phase space (= the space where the trajectories are). Hence, even a two dimensional ODE could produce chaos as artifact after center differencing discretization.

31.4 Discretization of PDE. The simplest method to discretize a PDE is to use a regular mesh on its domain and use the values of the functions sampled at the mesh points.[395] As explained in **31.2** we

[392] The Galerkin method is often used to solve PDE. In this case the resultant set of equations become a simultaneous set of ODEs. The method is also very important as a tool to prove the existence of the solutions to PDEs like the Navier-Stokes equation. See Ladyzhenskaya quoted in **1.21** Discussion.

[393] One might think that if a scheme is not consistent, then the scheme is useless. However, the situation is not this simple, because we do not take the $h \to 0$ limit in practice. Hence, even if the limit may be different from the original equation, still the numerical solution for a finite h may be a good solution.

[394] M. Mizutani, T. Niwa and T. Ohno, Chaos and bifurcation phenomena in limiting central difference scheme, J. Math. Kyoto Univ. **23**, 39-54 (1983).

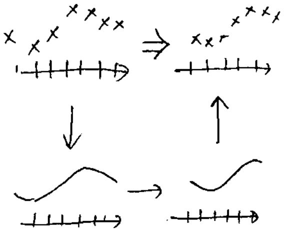
[395] A. Iserles, *A First Course in the Numerical Analysis of Differential Equations* (Cambridge UP, 1996) is an excellent introdution to the mathematical side of nu-

can also use the Galerkin method to discretize the PDE with the aid of generalized Fourier expansion (in terms of an appropriate complete set). Always the consistency and stability of the scheme are crucial. An important point recognized explicitly in recent years is that good modeling of physics on a discrete space can motivate a useful numerical solver for PDE.

Discussion.
A typical example is the numerical schemes for the simple equation

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \tag{31.8}$$

where c is a constant. We can solve this equation analytically easily (→**1.2B** Discussion(2), **2B.6**, **13A.4**), e.g., for the initial condition $u(x, 0) = 1$ for $x > 0$ and 0, otherwise. Ordinary discretization methods give miserable results (Try to solve this with the simple Euler scheme). However, we know the essence of the equation is the translational symmetry of space:



$$u(x, t + \delta t) = u(x - c\delta t, t) \tag{31.9}$$

for any δt (this is the equation for weak solutions, cf. **2B.3**). The problem is that if we discretize u, then we know only $u(x_i)$ at sampling points $\{x_i\}$. Therefore, it is very hard to describe the translational symmetry. The most natural idea is: (i) first reconstruct the continuous u from the discrete sampled values by interpolation, (ii) then translate the reconstructed continuous function according to (31.9), (iii) Finally sample the values of the shifted function at the grid points (see Figure). Actually, this reconstruction-resampling scheme is used in one of the best schemes for (31.8). Thus, the reader should keep in mind that there is still an ample room to devise unconventional numerical schemes for PDE.

31.5 Discretization of Poisson's equation. Practically useful numerical schemes use simple discretization to solve a Poisson's equation:[396]

$$\Delta u = f \tag{31.10}$$

on a region D with the boundary condition $u = g$ on ∂D. Let us consider this in 2-space. To discretize this, we follow Euler: Let h be the lattice spacing of the sampling regular square lattice; the sampling points are (ih, jh), where i and j are integers. Let us denote the value of a function f at (ih, jh) as $f[i, j]$. The simplest scheme is

$$\Delta_h u[i, j] \equiv \frac{u[i+1, j] + u[i, j+1] + u[i-1, j] + u[i, j-1] - 4u[i, j]}{h^2} = f[i, j] \tag{31.11}$$

merical analysis. Although, as the author explicitly says, it is not for practitioners, still the comments in the end of each chapter contain updated information and are useful.

[396] If the domain is regular, say, a square, then, Fourier transform methods are practical.

with $u[i,j] = g[i,j]$ if (ih, jh) is on the discretized boundary. Let us denote the set of grid points in the domain by D_h and the discretized boundary by Γ_h.

31.6 Solvability of (31.11). (31.11) is a linear algebraic equation, so that if the matrix defined by Δ_h is non-singular, then we can solve it. The non-singularity of the matrix can be shown with the aid of the maximum principle (→**29.6**) which is still true after discretization, because the mean value theorem is correct as can be seen from the form of Δ_h (→**1.13**). More precisely, we can show easily that if

$$\Delta_h v \geq 0 \text{ on } D_h \text{ and } v \geq 0 \text{ on } \Gamma_h, \tag{31.12}$$

then $v \geq 0$ on $D_h \cup \Gamma_h$. This implies that if v and $-v$ both satisfy (31.12), then $v = 0$ on $D_h \cup \Gamma_h$. That is, if $\Delta_h v = 0$ on D_h and $v = 0$ on Γ_h, then its unique solution is $v = 0$ everywhere. Hence, the matrix defining the simultaneous linear equation (31.11) is regular, and (31.11) has a unique solution. The matrix is very sparse, so many sparse marix solvers can be used.

31.7 Consistency and stability ⇒ convergence. Is this discretization scheme consistent? That is, in the $h \to 0$ limit can we claim that the discretized version converges to the original equation? If u is C^3 on the domain, we can demonstrate

$$\max_{x\in D_h} |\Delta_h u - \Delta u| \leq \frac{h}{3} \max_{x\in D} \left\{ \left|\frac{\partial^3 u}{\partial x^3}\right|, \left|\frac{\partial^3 u}{\partial y^3}\right| \right\}. \tag{31.13}$$

Since we know the solution to Poisson's equation is very smooth (→**29.10**) this is enough to demonstrate that indeed our scheme is consistent.

Our scheme is also stable: the solution to (31.11) is bounded by the 'magnitudes' of f and g in the problem as

$$\max_{x\in D_h\cup\Gamma_h} |u(x)| \leq c(\max_{x\in D_h} |f| + \max_{x\in \Gamma_h} |g|), \tag{31.14}$$

where c is a positive constant independent of h, f and g.[397]
Now we have
Theorem. The solution u_h to (31.11) converges uniformly to the solution to the original problem. More precisely,

$$\max_{x\in D_h\cup\Gamma_h} |u_h(x) - u(x)| \leq ch \max_{x\in D} \left\{ \left|\frac{\partial^3 u}{\partial x^3}\right|, \left|\frac{\partial^3 u}{\partial y^3}\right|, \left|\frac{\partial u}{\partial x}\right|, \left|\frac{\partial u}{\partial y}\right| \right\}. \tag{31.15}$$

[397] In this case, we need not restrict the size of h, but usually the stability holds for h up to some upper bound as we will see in the case of diffusion equation (→**31.8**).

□

We thus know that u_h converges to the true solution, but actually this is shown only on the dense set that are limit points of the lattice points. Since we know from the general theory that the true solution is very smooth, this should be enough.

31.8 Discretizing diffusion equation: θ-method. Let us consider 1-space diffusion equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f \tag{31.16}$$

on $Q_T = \{(x,t); x \in (0,1), t \in (0,T)\}$. We impose the initial condition $u(x,0) = a(x)$ for $x \in (0,1)$. We must also specify a boundary condition at $x = 0$ and 1, but we will not explicitly write it down. The 1-space version of Δ_h is given by

$$\Delta_h u[i] = \frac{u[i+1] + u[i-1] - 2u[i]}{h^2}. \tag{31.17}$$

We must discretize the time axis with the spacing τ. We introduce the following notation

$$u_n[i] = u(ih, n\tau), \tag{31.18}$$

and

$$u_{n+\theta}[i] = \theta u_{n+1}[i] + (1-\theta)u_n[i]. \tag{31.19}$$

We introduce the following scheme called the *θ-method*:

$$\frac{u_{n+1}[i] - u_n[i]}{\tau} = \Delta_h u_{n+\theta}[i] + f_{n+\theta}[i]. \tag{31.20}$$

For $\theta = 0$ this is the standard Euler method; for $\theta = 1/2$ it is called the *Cranck-Nicholson method*; for $\theta = 1$ it is called the *backward Euler method*. The latter two methods are called implicit methods, because we cannot immediately read off the updated data.

31.9 Stability analysis. A standard method to analyze the stability of a scheme is to compute the so-called *amplification factorA*:

$$u_n[l] = A^n e^{ikl}. \tag{31.21}$$

The basic idea is that we prepare spatially bounded 'initial condition' (that is why e^{ikl}) and study its time evolution. If $|A| > 1$, we are in trouble.

31.10 Von Neumann's stability condition.[398] In our case the scheme is stable if $u_n[i]$ is bounded for all i and n by a number proportional to the 'magnitude' of the initial condition a. Let us measure the 'magnitude' with the following 'normalized ℓ_2-norm':

$$||v||_h \equiv \left\{\frac{1}{N}\sum_{i=0}^{N-1} v[i]^2\right\}^{1/2}. \tag{31.22}$$

The stability is defined by the inequality

$$||u_n||_h \leq c||a||_h \tag{31.23}$$

for all n with some positive constant c independent of a, h and $\tau(< 1)$.
Theorem [von Neumann]. A necessary and sufficient condition for the scheme (31.20) to be stable is that there is a nonnegative constant b such that for any k

$$\left|\frac{h^2 - 4(1-\theta)\sin^2\frac{\pi k}{4N}}{h^2 + 4\theta\tau\sin^2\frac{\pi k}{4N}}\right| < 1 + b\tau \tag{31.24}$$

for any $k \in \boldsymbol{Z}$. In particular, the scheme is stable for $\theta \in [1/2, 1]$ unconditionally and for $\theta \in [0, 1/2]$ under the condition

$$\frac{\tau}{h^2} \leq \frac{1}{2(1-2\theta)}, \tag{31.25}$$

which is called the stability condition.[399] □
Generally speaking, implicit schemes are more stable as seen here. However, implicit schemes are usually computationally more time consuming. The reader might think that exploiting the stability, we can choose a large τ to compensate the complexity. Sometimes, this indeed works, but stability does not mean that the obtained solution is accurate, so that choosing a large τ is not usually wise.

Discussion.
(A) In (31.20) put $\theta = 0$ and $f = 0$. Assume

$$u_{n,j} = \lambda(k)^n e^{ik\,(jh)}. \tag{31.26}$$

Then, this is a solution to (31.20), if

$$\lambda(k) = 1 - 4\frac{\tau}{h^2}\sin^2(kh/2). \tag{31.27}$$

[398] John von Neumann, 1903-1957.

[399] The stability condition may depend on the norm used. If we use the ℓ_∞-norm, then the RHS of (31.25) reads $1/2(1-\theta)$ for $\theta \in [0, 1]$.

This $\lambda(k)$ is the amplification factor for the mode k. From this we conclude that

$$\frac{\tau}{h^2} < \frac{1}{2} \tag{31.28}$$

is required for the scheme to be stable. The condition can be rewritten as

$$D < \frac{h^2}{2\tau}. \tag{31.29}$$

This may be interpreted as a condition for the numerical diffusion constant to be larger than the physical diffusion constant.

If $\tau/h^2 = 1/2$, the scheme may violate the maximum principle.

(B) In **31.8** try the same and derive the formula for the amplification factor for the θ method:

$$\lambda(k) = \frac{1 - 4(1-\theta)(\tau/h^2)\sin^2(kh/2)}{1 + 4\theta(\tau/h^2)\sin^2(kh/2)}. \tag{31.30}$$

From this the stability condition is given by (the von Neumann stability condition **31.9**)

$$\frac{\tau}{h^2}(1 - 2\theta) < \frac{1}{2}. \tag{31.31}$$

For $\theta = 1/2$, the method is called the *Cranck-Nicolson scheme.* In this case, if $\tau/h^2 = 1$, the scheme is stable, but does not satisfy the maximum principle (the number of peaks may increase).

(C) Consider the following diffusion-advection equation:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - b\frac{\partial u}{\partial x}, \tag{31.32}$$

where b is a continuous function of x and t with boundedness: $|b| < B$. Apply a discretization scheme (not complicated one, please) and study its stability.

31.11 Consistency and convergence of θ-method. If u is smooth enough,[400] then we can show that the θ-method is consistent. Under the stability condition discussed in **31.10**, the solution u_h to (31.20) converges to the solution to the original PDE in the $h \to 0$ limit. More precisely,

$$\max_n ||u_{hn} - u_n||_h \le T\left\{\left|\theta - \frac{1}{2}\right|\tau \max\left|\frac{\partial^2 u}{\partial t^2}\right| + \frac{\tau^2}{12}\max\left|\frac{\partial^3 u}{\partial t^3}\right| + \frac{h^2}{12}\max\left|\frac{\partial^4 u}{\partial t^4}\right|\right\}. \tag{31.33}$$

31.12 Courant-Friedrichs-Lewy condition. Let us return to the simple advection problem (31.8). Consider the following simple Euler scheme

$$\frac{u_n[i] - u_{n-1}[i]}{\tau} + c\frac{u_n[i] - u_n[i-1]}{h} = 0. \tag{31.34}$$

[400] C^4 in space and C^3 in time, for example.

This is called the *upstream approximation*, because if c is interpreted as the stream velocity, the scheme uses the upstream information only. The scheme satisfies the stability condition, if

$$\tau \leq \frac{h}{c}. \tag{31.35}$$

The condition is called the *Courant-Friedrichs-Lewy condition*[401] (CFL condition). This implies that the numerical propagation speed h/τ must not be smaller than the physical propagation speed c. In other words, if physics outruns computation, the scheme becomes unstable. A similar interpretation may be possible for **31.10**.

Exercise.
(1) Compute the amplification factor for (13.28) and derive the Courant-Friedrichs-Lewy condition.
(2) Show that the down stream scheme, which replaces $u_n[i] - u_n[i-1]$ in the upstream scheme with $u_n[i+1] - u_n[i]$ is always unstable.

31.13 Wave equation. A standard differencing practice to solve 1-space wave equation $u_{tt} - c^2 u_{xx} = 0$ is the simple Euler scheme:

$$\begin{aligned} u_{n+1}(i) &= 2u_n(i) - u_{n-1}(i) \\ &\quad + \left(\frac{c\Delta t}{\Delta x}\right)^2 \{u_n(i+1) + u_n(i-1) - 2u_n(i)\}. \end{aligned} \tag{31.36}$$

It is easy to generalize this to d-space (The stability limit due to the CFL condition is $c\Delta t/\Delta x \leq 1/\sqrt{d}$). This is a very stable and simple scheme, and is widely used. However, it suffers from the dispersion error (The scheme conserves energy very well, but distorts initial conditions with steep wave fronts.)

Exercise.
Study the stability condition of this simple scheme and demonstrate that we indeed need the Courant-Friedrichs-Lewy condition (the numerical propagation speed must be faster than the physical speed).

[401] Richard Courant, 1888-1972; Kurt Otto Friedrichs, 1901-1983.