A comparison of finite difference methods for solving the one- and two-dimensional diffusion equation

Abstract

Introduction

Theory

Derivation of the heat equation

To derive the heat equation we consider some substance of volume V and with surface S. At any given time there is some heat flowing in or out of the substance. We define the heat flux density $\boldsymbol{q}$
as the amount of energy flow per unit area and per unit time. Fourier's law states that the heat flux density is proportional to the temperature gradient

$$\boldsymbol{q} = -k\boldsymbol{\nabla}T \tag{1}$$

The thermal conductivity k of the material is defined by Eq. (1). Integrating both sides of Eq. (1) over the surface S we obtain

$$\frac{\partial Q}{\partial t} = -\oiint_S d\boldsymbol{S}\cdot k\boldsymbol{\nabla}T \tag{2}$$

where dQ is the amount of heat that flows out of the substance during a time interval dt. Let U be the internal energy of the substance. Assuming no work is being done on the substance the first law of thermodynamics says that $dU = -dQ$. Using $dU = C_V dT$where C_V is the heat capacity of the material at constant volume we can rewrite the left side of Eq. (2) as

$$-C_V\frac{\partial T}{\partial t}$$

We can rewrite the right side of Eq. (2) using the divergence theorem as

$$-\iiint_V dV\, \boldsymbol{\nabla} \cdot (k \boldsymbol{\nabla} T)$$

Using $C_V = mc_V$ where m is the mass of the substance and c_V is the specific heat capacity of the material, Eq. (2) now reads

$$mc_V \frac{\partial T}{\partial t} = \iiint_V dV\, \boldsymbol{\nabla} \cdot (k \boldsymbol{\nabla} T) \tag{3}$$

By only considering an infinitesimal part of the substance with mass dm and volume dVwe can substitute $\iiint_V dV \to dV$ and $m \to dm = \rho dV$ in Eq. (3) where ρ is the density of the material. Dividing both sides by dV we obtain the heat equation

$$\rho c_V \frac{\partial T}{\partial t} = \boldsymbol{\nabla} \cdot (k \boldsymbol{\nabla} T) \tag{4}$$

A more stripped down version of Eq. (4) is the diffusion equation

$$\frac{\partial u}{\partial t} = \boldsymbol{\nabla}^2 u \tag{5}$$

where time and position are dimensionless. We will solve Eq. (5) numerically in one and two dimensions using various finite difference schemes, both explicit and implicit.

Analytical solution in one dimension

The one-dimensional diffusion is

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x} \tag{6}$$

which we will solve analytically on the interval $x \in [0,\ 1]$ for $t > 0$ with the boundary conditions $u(0,t) = 0$, $u(1,t) = 1$ and the initial condition $u(x,0) = 0$. We can think of $u(x,t)$ as the temperature distribution along a rod of length 1 with a constant heat source at one of the end points. From our daily life experience we know that the temperature distribution should reach a steady state after some amount of time, in which case we can set the left side of Eq. (6) to zero and obtain

$$u\big(x, t \to \infty\big) = ax + b$$

i.e the temperature distribution approaches some first degree polynomial. The coefficients a and $b$are determined by the boundary conditions. In our case the temperature distribution will converge to

$$u\big(x, t \to \infty\big) = x \tag{7}$$

Eq. (7) satisfies the diffusion equation for all times t since the time derivative and the second order position derivative are both zero, but it does not satisfy our choice of the initial condition. But since the diffusion equation is linear, a sum of two solutions is also a solution. We can therefore try a solution on the form

$$u(x,t) = x + f(x,t)$$

where $f(x,t)$ has to satisfy $f(0,t) = f(1,t) = 0$ and $f(x) \equiv f(x,0) = -x$. We can find $f(x,t)$by using the method of separation of variables and assume that it can be written on the form $f(x,t) = F(x)G(t)$. Inserting it into Eq. (6) we obtain

$$F\frac{dG}{dt} = G\frac{d^2F}{dx^2}$$

Or equivalently

$$\frac{1}{G}\frac{dG}{dt} = \frac{1}{F}\frac{d^2F}{dx^2} \tag{8}$$

Since the left side is a function of t alone and the right side is a function of x alone, both sides

of Eq. (8) must be a constant. It's convenient to call this constant $-k^2$ for reasons which will be apparent in a moment. We can then split Eq. (8) into two ordinary differential equations and obtain the respective general solutions

$$\frac{dG}{dt} = -k^2 G \rightarrow G(t) = e^{-k^2 t}$$

$$\frac{d^2 F}{dx^2} = -k^2 F \rightarrow F(x) = A\sin(kx) + B\cos(kx)$$

There is generally a constant in front of the exponential in $G(t)$, but it might as well be absorbed into $F(x)$. Choosing a constant on the form $-k^2$ ensures that there is no time dependence as $t \rightarrow \infty$. To find the solution $f(x,t) = F(x)G(t)$ which satisfies $f(0,t) = f(1,t) = 0$ it is convenient to set $t = 0$ so that $G = 1$ and $F(0) = F(1) = 0$. From the first boundary condition we get

$$A\sin(0) + B\cos(0) = A \cdot 0 + B \cdot 1 = B = 0$$

And from $F(1) = 0$ we get

$$A\sin(k) = 0 \rightarrow k = n\pi$$

where n is some positive integer. n could be a negative integer too, but since $\sin(x)$ is an odd function, $\sin\big(-|n|\pi x\big) = -\sin\big(|n|\pi x\big)$ where the minus sign might as well be absorbed into A. A solution $f(x,t)$ which satisfies $f(0,t) = f(1,t) = 0$ is thus on the form

$$f(x,t) = A\sin(n\pi x)e^{-n^2\pi^2 t}$$

but it does not satisfy $f(x,0) = f(x) = -x$ for any n. Again we can exploit the linearity of the diffusion equation and write $f(x,t)$ as

$$f(x,t) = \sum_{n=1}^{\infty} A_n \sin(n\pi x)e^{-n^2\pi^2 t} \tag{9}$$

which is a Fourier series. The functions $\sin(n\pi x)$ are orthogonal with respect to the inner product

$$2\int_0^1 dx\,\sin(m\pi x)\sin(n\pi x) = \delta_{nm}$$

If we set $t = 0$ in Eq. (9) and "multiply" both sides by $2\int_0^1 dx\,\sin(m\pi x)$ we obtain

$$A_n = -2\int_0^1 dx\,\sin(n\pi x)x = (-1)^n\frac{2}{n\pi}$$

The analytical solution can thus be written as

$$u(x,t) = x + \sum_{n=1}^{\infty}(-1)^n\frac{2}{n\pi}\sin(n\pi x)e^{-n^2\pi^2 t}$$

Analytical solution in two dimensions

The two-dimensional diffusion equation is

$$\frac{\partial u(x,y,t)}{\partial t} = \frac{\partial^2 u(x,y,t)}{\partial x^2} + \frac{\partial^2 u(x,y,t)}{\partial y^2} \tag{10}$$

which for simplicity will be solved on the area $x, y \in [0,1]$ with the boundary conditions $u|_{x=0} = u|_{y=0} = u|_{x=1} = u|_{y=1} = 0$ and with initial condition $f(x,y) \equiv u(x,y,0)$ (it'll soon be apparent what's a convenient pick for the initial condition). Again we use the method of separation of variables and assume a solution on the form $u(x,y,t) = X(x)Y(y)T(t)$. Substituting this into Eq. (10) we obtain

$$XY\frac{dT}{dt} = YT\frac{d^2X}{dx^2} + XY\frac{d^2Y}{dy^2}$$

Dividing both sides by XYT we get

$$\frac{1}{T}\frac{dT}{dt} = \frac{1}{X}\frac{d^2X}{dx^2} + \frac{1}{Y}\frac{d^2Y}{dy^2} \tag{11}$$

where the left side is a function of t alone and the right side is a function of position alone. So both sides is equal to some constant $-k^2$. But we could move the t-term over to the right and either the x-or y-term over to the left, in which case the left side would be a function of either x or y alone. So all terms in Eq. (11) must be constants. Let the x-term be $-p^2$ and the y-term be $-q^2$ where $p^2 + q^2 = k^2$. Then we get the three ordinary differential equations

$$\frac{d^2X}{dx^2} = -p^2X \rightarrow X(x) = A\sin(px) + B\cos(px)$$

$$\frac{d^2Y}{dy^2} = -q^2Y \rightarrow Y(y) = C\sin(qy) + D\cos(qy)$$

$$\frac{dT}{dt} = -k^2T \rightarrow T(t) = e^{-k^2t}$$

where the coefficient in front of the exponential in $T(t)$ has been absorbed into $X(x)Y(y)$. Analogous to the one-dimensional case, the boundary conditions are satisfied by choosing $B = D = 0$ and $p = m\pi,\ q = n\pi$ for some positive integers m and n. Again by exploiting the linearity of the diffusion equation

$$u(x, y, t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} A_{m,n}\sin(m\pi x)\sin(n\pi y)e^{-\pi^2\left(m^2+n^2\right)t}$$

Setting $t = 0$ and "multiplying" both sides by $4\int_0^1 dx\int_0^1 dy\sin(m'\pi x)\sin(n'\pi y)$ we obtain

$$A_{m,n} = 4\int_0^1 dx\int_0^1 dy\ f(x, y)\sin(m\pi x)\sin(n\pi x)$$

In particular, if $f(x, y) = \sin(\pi x)\sin(\pi y)$ only $A_{1,1} = 1$ is nonzero. In this case the

analytical solution is

$$u(x, y, t) = \sin(\pi x)\sin(\pi y)e^{-2\pi^2 t}$$

where the factor of 2 in the exponential comes from $\left(m^2 + n^2\right) = 1^2 + 1^2 = 2$.

Numerical solution to tridiagonal matrix equations

The two implicit schemes Backward Euler and Crank-Nicolson requires us to solve a tridiagonal matrix equation, so we'll take a moment to go through how this can be done efficiently. Consider the matrix equation

$$\begin{bmatrix} 1 & & & & \\ a_1 & d_1 & b_1 & & \\ & & \ddots & & \\ & & a_{n-2} & d_{n-2} & b_{n-2} \\ & & & & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ \\ y_{n-1} \end{bmatrix}$$

The indices has been chosen to match the standard array indexing in the C/C++ and Python programming languages. The matrix is not completely tridiagonal; a 1 has been added to the top left and the bottom right elements to allow us to choose any boundary condition. Omitting these elements correspond to setting $y_0 = y_{n-1} = 0$, i.e Dirichlet boundary conditions. Multiplying out the matrix we get a set of linear equations where the first few are

$$x_0 = y_0 \tag{12}$$

$$a_1 x_0 + d_1 x_1 + b_1 x_2 = y_1 \tag{13}$$

$$a_2 x_1 + d_2 x_2 + b_2 x_3 = y_2$$

What we're going to do is to eliminate the a_i in every equation using the previous equation. Multiplying Eq. (12) by a_1 and subtracting the result from Eq. (13) we get

$$d_1 x_1 + b_1 x_2 = \widetilde{y}_1 \tag{14}$$

$$a_2 x_1 + d_2 x_2 + b_2 x_3 = y_2 \tag{15}$$

Where $\widetilde{y}_1 \equiv y_1 - a_1 y_0$. Multiplying Eq. (14) by $a_2 \,/\, d_1$ and subtracting the result from Eq. (15) we get

$$\left(d_2 - \frac{a_2}{d_1} b_1\right) x_2 + b_2 x_3 = y_2 - \frac{a_2}{d_1} \widetilde{y}_1$$

Defining $\widetilde{d}_2 \equiv d_2 - \frac{a_2}{d_1} b_1$ and $\widetilde{y}_2 \equiv y_2 - \frac{a_2}{d_1} \widetilde{y}_1$ the next few equations are

$$\widetilde{d}_2 x_2 + b_2 x_3 = \widetilde{y}_2 \tag{16}$$

$$a_3 x_2 + d_3 x_3 + b_3 x_4 = y_3 \tag{17}$$

Notice that Eq. (16) and (17) are on the same form as Eq. (14) and Eq. (15). Repeating this process using the recurrence relations

$$\widetilde{d}_i = d_i - \frac{a_i}{\widetilde{d}_{i-1}} b_{i-1}, \;\; \widetilde{d}_1 = d_1 \tag{18}$$

$$\widetilde{y}_i = y_i - \frac{a_i}{\widetilde{d}_{i-1}} \widetilde{y}_{i-1}, \;\; \widetilde{y}_1 = y_1 - a_0 y_0 \tag{19}$$

for $i = 2, \; 3, \;\; \ldots n-2$ we have effectively rewritten the matrix equation as

$$\begin{bmatrix} 1 & & & & \\ & \widetilde{d}_1 & b_1 & & \\ & & \ddots & & \\ & & & \widetilde{d}_{n-2} & b_{n-2} \\ & & & & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ \\ x_{n-1} \end{bmatrix} = \begin{bmatrix} y_0 \\ \widetilde{y}_1 \\ \vdots \\ \widetilde{y}_{n-2} \\ y_{n-1} \end{bmatrix}$$

Eq. (18) and (19) are a type of forward substitution algorithms. The last few equations are now

$$\begin{aligned} x_{n-3}\widetilde{d}_{n-3} + b_{n-3}x_{n-2} &= \widetilde{y}_{n-3} \\ x_{n-2}\widetilde{d}_{n-2} + b_{n-2}x_{n-1} &= \widetilde{y}_{n-2} \\ x_{n-1} &= y_{n-1} \end{aligned} \tag{20}$$

Solving Eq. (20) for x_{n-2} we get

$$x_{n-2} = \frac{\widetilde{y}_{n-2} - b_{n-2}x_{n-1}}{\widetilde{d}_{n-2}}$$

It's now easy to see that we can solve for all the x_i's using the backward substitution algorithm

$$x_i = \frac{\widetilde{y}_i - b_i x_{i+1}}{\widetilde{d}_i}, \;\; x_{n-1} = y_{n-1} \tag{21}$$

for $i = n-2,\; n-3,\; \ldots,\; 1$ (x_0 is already given by Eq. (12)).

The Forward Euler scheme

Consider a function $f(x)$ and its Taylor expansion around the point $x = a$.

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2 + \mathcal{O}\left(x^3\right) \tag{22}$$

We want to find an approximation to the first and second derivative of $f(x)$ at an arbitrary point x. If h is a small number we can accomplish this by substituting $a \to x$ and $x - a \to h$ in Eq. (22)to obtain

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)}{2}h^2 + \mathcal{O}(h^3) \tag{23}$$

Solving for $f'(x)$ we get

$$f'(x) = \frac{f(x+h) - f(x)}{h} + \mathcal{O}(-h) \tag{24}$$

Truncating all the terms containing powers of h larger than or equal to 1 we get an approximation to the first derivative which is first-order accurate, because the truncation error $\mathcal{O}(-h)$ is approximately proportional to the first power of h. To obtain a similar expression to Eq. (24) for the second derivative we can substitute $a \to x$ and $x - a \to -h$ in Eq. (22).

$$f(x-h) = f(x) - f'(x)h + \frac{f''(x)}{2}h^2 + \mathcal{O}(-h^3) \tag{25}$$

Adding Eq. (23) and (25) together we get

$$f(x+h) + f(x-h) = 2f(x) + f''(x)h^2 + \mathcal{O}(h^4)$$

Solving for $f''(x)$:

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + \mathcal{O}(-h^2) \tag{26}$$

where the approximation to the second derivative is second-order accurate. We can now write Eq. (6) as

$$\frac{u(x,t+\Delta t)-u(x,t)}{\Delta t}+\mathcal{O}(\Delta t)=\frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{\Delta x^2}+\mathcal{O}\left(\Delta x^2\right) \tag{27}$$

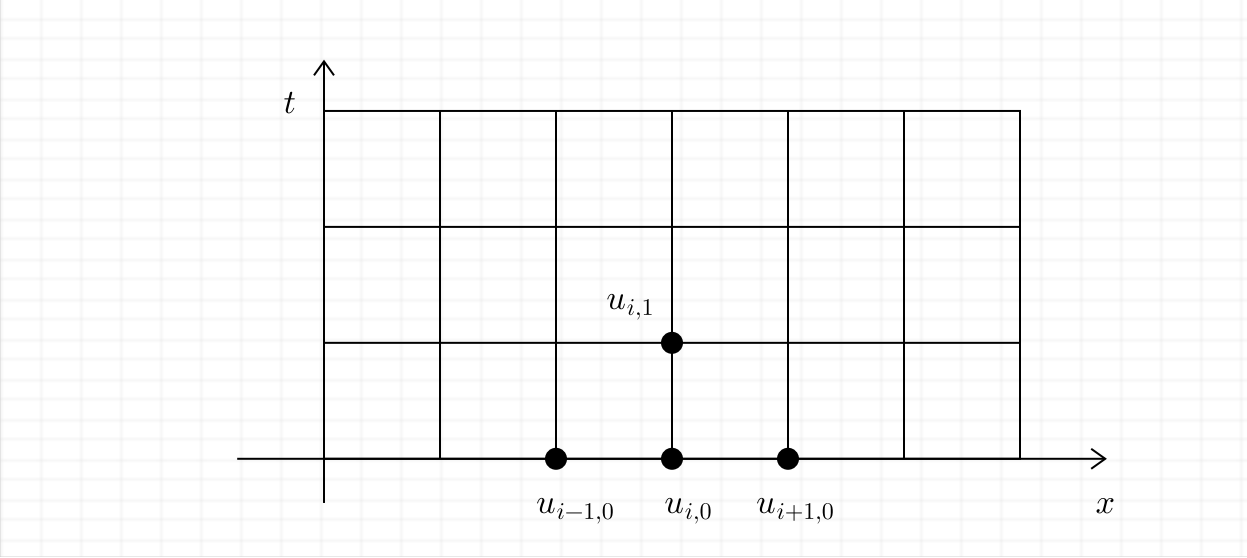
We'll now discretize the domain $x \in [0,1]$ into points $x_i = i\Delta x$ and times $t \in [0,\ T]$ into $t_i = j\Delta t$ for $i = 0,\ 1,\ \ldots,\ N-1$ and $j = 0,\ 1,\ \ldots, n-1$ where $x_{N-1} = 1$ and $t_{n-1} = T$. The step sizes are then given by

$$\Delta x = \frac{1}{N-1},\ \Delta t = \frac{T}{n-1}$$

Substituting $u_{i,j} \equiv u(x_i, t_j)$ into Eq. (27) and sloppily throwing away the truncation errors we can write

$$u_{i,j+1} = \alpha u_{i+1,j} + (1-2\alpha)u_{i,j} + \alpha u_{i-1,j} \tag{28}$$

where $\alpha \equiv \Delta t\,/\,\Delta x^2$. Eq. (28) is the Forward Euler scheme. It's an explicit scheme, meaning that there is only a single unknown quantity explicitly given in terms of known quantities.



This is illustrated in the figure above which shows the numerical solution $u_{i,j}$ as a function of x and t. All quantities $u_{i,\,0}$ for $i = 0,\ \ldots,\ N-1$ on the first row are known because they come from the initial condition and the boundary conditions. Thus every single quantity $u_{i,1}$ for $i = 1,\ \ldots,\ N-2$ can be calculated. After that every single quantity on the third row can be calculated, and so on. This shows that when we implement Eq. (28) in our program the

inner loop has to loop over position and the outer loop has to loop over time.

The Backward Euler scheme

The Backward Euler scheme looks just a tiny bit different from the Forward Euler scheme, but the way we actually obtain the numerical solution $u_{i,j}$ is very different. We still use Eq. (26) to approximate the second-order position derivative in Eq. (6), but the new approximation to the time-derivative is obtained by using Eq. (25)

$$f'(x) = \frac{f(x) - f(x-h)}{h} + \mathcal{O}(h) \tag{29}$$

We can use Eq. (26) and Eq. (29) to write Eq. (6) as

$$\frac{u(x,t) - u(x,t-\Delta t)}{\Delta t} + \mathcal{O}(\Delta t) = \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2} + \mathcal{O}\left(\Delta x^2\right) \tag{30}$$

We discretize the domain $x \in [0,1]$ and the times $t \in [0,\ T]$ similarly to what we did in the previous section. Again substituting $u_{i,j} \equiv u(x_i, t_j)$ and introducing $\alpha = \Delta t\,/\,\Delta x^2$ we can put all u_j's and u_{j-1}'s on separate sides of the equality sign and write Eq. (30) as

$$u_{i,j-1} = -\alpha u_{i+1,j} + (1+2\alpha)u_{i,j} - \alpha u_{i-1,j} \tag{31}$$

If we set $j = 1$ we see that $u_{i,j-1}$ is the only one known of all the $u_{i,j}$'s in Eq. (31). So a single equation is not going to cut it. What we can do is to write Eq. (31) as a system of linear equations

$$u_{0,j} = u_{0,j-1}$$

$$-\alpha u_{0,j} + (1+2\alpha)u_{1,j} - \alpha u_{2,j} = u_{1,j-1}$$

$$\vdots$$

$$-\alpha u_{N-3,j} + (1+2\alpha)u_{N-2,j} - \alpha u_{N-1,j} = u_{N-2,j-1}$$

$$u_{N-1,j} = u_{N-1,j-1}$$

which we can write as the matrix equation

$$A\boldsymbol{u}_j = \boldsymbol{u}_{j-1}$$

$$\begin{bmatrix} 1 & & & & \\ -\alpha & 1+2\alpha & -\alpha & & \\ & & \ddots & & \\ & & -\alpha & 1+2\alpha & -\alpha \\ & & & & 1 \end{bmatrix} \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ \vdots \\ \\ u_{N-1,j} \end{bmatrix} = \begin{bmatrix} u_{0,j-1} \\ u_{1,j-1} \\ \vdots \\ \\ u_{N-1,j-1} \end{bmatrix}$$

and solve using the recurrence relations in Eq. (18), (19) and (21). The fact that we have to solve a system of linear equations to obtain the numerical solution $u_{i,j}$ makes the Backward Euler scheme an implicit scheme.

The Crank-Nicolson scheme

We can derive the Crank-Nicolson scheme in a simple way using the Forward and Backward Euler schemes, but first we have to rewrite the Backward Euler scheme slightly, shifting it up one time step:

$$\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} = \frac{u(x+\Delta x, t+\Delta t) - 2u(x,t+\Delta t) + u(x-\Delta x, t+\Delta t)}{\Delta x^2}$$

What we're going to do now is to approximate the first time derivative of and the second position derivative of $u(x,t)$ at the point $(x, t+\Delta t\,/\,2)$ by taking the average of the Forward and Backward Euler schemes. The approximation to the time derivative is

$$\frac{1}{2}\left[\overbrace{\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t}}^{\text{Forward Euler}} + \overbrace{\frac{u(x,t+\Delta t) - u(x,t)}{\Delta t}}^{\text{Rewritten Backward Euler}}\right] = \frac{u(x,t+\Delta t) - u(x,t)}{\Delta t} \quad (32)$$

and the approximation to the second order position derivative is

$$\frac{1}{2}\Big[\overbrace{\frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{\Delta x^2}}^{\text{Forward Euler}} + \underbrace{\frac{u(x+\Delta x,t+\Delta t)-2u(x,t+\Delta t)+u(x-\Delta x,t+\Delta t)}{\Delta x^2}}_{\text{Rewritten Backward Euler}} \tag{33}$$

Since the approximation to the time derivative is centered at $t+\Delta t\,/\,2$ and evaluated at t and $t+\Delta t$ it corresponds to

$$f'(t) = \frac{f(t+\Delta t)-f(t-\Delta t)}{\Delta t} + \mathcal{O}\left(\Delta t^2\right) \tag{34}$$

which is second-order accurate in Δt. The approximation to the second-order position derivative is however centered at x and evaluated at x and $x \pm \Delta x$, so it is still "only" second-order accurate in Δx^2 as with the Forward and Backward Euler schemes.

Discretizing Eq. (32) and (33) similarly to what we did with the Forward and Backward Euler schemes and substituting $u_{i,j} \equiv u(x_i, t_j)$ we can bring it all together.

$$\frac{u_{i,j+1}-u_{i,j}}{\Delta t} = \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}+u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{2\Delta x^2}$$

Again introducing $\alpha \equiv \Delta t\,/\,\Delta x^2$ and moving all u_j's and u_{j+1}'s on separate sides of the equality sign we obtain

$$-\alpha u_{i-1,j} + 2(1+\alpha)u_{i,j} - \alpha u_{i+1,j} = \alpha u_{i-1,j-1} + 2(1-\alpha)u_{i,j-1} + \alpha u_{i+1,j-1} \tag{35}$$

Setting $j=1$ we see that there are three unknowns $u_{i-1,1}$, $u_{i,1}$ and $u_{i+1,1}$ in Eq. (35), so we need to write it as a system of linear equations.

$$u_{0,j} = u_{0,j-1}$$

$$-\alpha u_{0,j} + 2(1+\alpha)u_{1,j} - \alpha u_{2,j} = \alpha u_{0,j-1} + 2(1-\alpha)u_{1,j-1} + \alpha u_{2,j-1}$$

$$\vdots$$

$$-\alpha u_{N-3,j} + 2(1+\alpha)u_{N-2,j} - \alpha u_{N-1,j} = \alpha u_{N-3,j-1} + 2(1-\alpha)u_{N-2,j-1} + \alpha u_{N-1,j-1}$$

$$u_{N-1,j} = u_{N-1,j-1}$$

This can be written as the matrix equation

$$(2I + \alpha B)\boldsymbol{u}_j = (2I - \alpha B)\boldsymbol{u}_{j-1} \tag{36}$$

$$\begin{bmatrix} 1 & & & & & \\ \alpha & 2(1-\alpha) & \alpha & & & \\ & & \ddots & & & \\ & & \alpha & 2(1-\alpha) & \alpha & \\ & & & & 1 & \end{bmatrix} \begin{bmatrix} u_{0,j-1} \\ u_{1,j-1} \\ \vdots \\ u_{n-1,j-1} \\ u_{N-1,j-1} \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ -\alpha & 2(1+\alpha) & -\alpha & & & \\ & & \ddots & & & \\ & & -\alpha & 2(1+\alpha) & -\alpha & \\ & & & & 1 & \end{bmatrix} \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ \vdots \\ u_{n-1,j} \\ u_{N-1,j} \end{bmatrix}$$

where I is the identity matrix and

$$B = \begin{bmatrix} 0 & & & & \\ 1 & -2 & 1 & & \\ & & \ddots & & \\ & & 1 & -2 & 1 \\ & & & & 0 \end{bmatrix}$$

Defining $\widetilde{\boldsymbol{u}}_{j-1} \equiv (2I - \alpha B)\boldsymbol{u}_{j-1}$ we can write Eq. (36) as

$$(2I + \alpha B)\boldsymbol{u}_j = \widetilde{\boldsymbol{u}}_{j-1}$$

Since $2I + \alpha B$ is on the same form as the matrix discussed in section SECTION HERE XXXXXXXXXX we can use the recurrence relations in Eq. (18), (19) and (21) to solve for $\boldsymbol{u}_j$.

The two-dimensional Forward Euler scheme

To solve the two-dimensional diffusion equation (Eq. (10)) we approximate the two second order position derivatives by Eq. (26) while the time derivative is approximated by Eq. (24).

Here the index i represents x, j represents y and n represents t.

$$\frac{u_{i,j}^{n+1} - u_{i,j}^{n}}{\Delta t} = \frac{u_{i+1,j}^{n} - 2u_{i,j}^{n} + u_{i-1,j}^{n}}{\Delta x^2} + \frac{u_{i,j+1}^{n} - 2u_{i,j}^{n} + u_{i,j-1}^{n}}{\Delta y^2}$$

We'll use the same step sizes in the x-and y-coordinates $h \equiv \Delta x = \Delta y$. Defining $\alpha \equiv \Delta t \,/\, h^2$ and solving for $u_{i,j}^{n+1}$ we get

$$u_{i,j}^{n+1} = u_{i,j}^{n} + \alpha\left(u_{i+1,j}^{n} + u_{i-1,j}^{n} + u_{i,j+1}^{n} + u_{i,j-1}^{n} - 4u_{i,j}^{n}\right) \tag{37}$$

Setting $n = 0$ makes it clear that $u_{i,j}^{n+1}$ is the only unknown in Eq. (37) since all the $u^{n=0}$ are known from the initial condition. The two-dimensional Forward Euler scheme is thus also an explicit scheme.

Von Neumann stability analysis

In this section we will use Von Neumann stability analysis to determine for what values of Δx and
Δt the various schemes discussed in the previous sections are stable, but first we need to go though some definitions.

A finite difference scheme is said to be *convergent* if the numerical solution of the scheme converges to the exact analytical solution of the partial differential equation as Δx and Δt approaches zero.

A finite difference scheme is said to be *consistent* if it is at least first order accurate in Δx and Δt. The truncation errors are $\mathcal{O}\left(\Delta x^{\beta}\right)$ and $\mathcal{O}\left(\Delta t^{\gamma}\right)$ where $\beta, \gamma \geq 1$.

A finite difference scheme is said to be *unconditionally stable* if the error, that is the difference between the numerical solution and the exact solution, do not tend to infinity for any choice of $\Delta x$and Δt.

A finite difference scheme is said to be *conditionally stable* if the error does not tend to infinity for some, but not all choices of Δx and Δt.

These definitions are brought together by the Lax-Equivalence theorem which states that a finite difference scheme is convergent if and only if it is consistent and stable. All the finite

difference schemes discussed in the previous sections are consistent, so to show that they are convergent we'll use Von Neumann stability analysis to figure out if and when they are stable.

To do this it is convenient to write the general solution of the diffusion equation on the more compact form

$$f(x,t) = \sum_{k=-\infty}^{\infty} A_k e^{ikx} e^{-k^2 t}$$

But it suffices to use only one of the functions in the linear combination

$$f_k(x,t) = A_k e^{ikx} e^{-k^2 t}$$

Using discrete points $x_j = j\Delta x$ and times $t_n = n\Delta t$ we can write

$$f_k(x_j, t_n) = A_k e^{ikj\Delta x} e^{-k^2 n\Delta t}$$

$$= A_k e^{-k^2 n\Delta t} e^{ikj\Delta x} = \left(B_k e^{-k^2\Delta t}\right)^n e^{ikj\Delta x}$$

$$= \xi_a(k)^n e^{ikj\Delta x} \tag{38}$$

where $A_k = B_k^n$ and $\xi_a(k) \equiv B_k e^{-k^2\Delta t}$ is the analytical amplification factor. The various finite difference schemes may be thought of as difference equations where Eq. (38) is a solution. If $u_{j,n}$ is the numerical solution and $\epsilon_{j,n}$ is the error we can write

$$f_k(x_j, t_n) = u_{j,n} - \epsilon_{j,n} \tag{39}$$

where $\epsilon_{j,n} \equiv u_{j,n} - f_k(x_j, t_n)$. We can put $f_{j,n} \equiv f_k(x_j, t_n)$ into the Forward Euler scheme

$$f_{j,n+1} = \alpha f_{j+1,n} + (1-2\alpha) f_{j,n} + \alpha f_{j-1,n}$$

$$(u_{j,n+1} - \epsilon_{j,n+1}) = \alpha(u_{j+1,n} - \epsilon_{j+1,n}) + (1-2\alpha)(u_{j,n} - \epsilon_{j,n}) + \alpha(u_{j-1,n} - \epsilon_{j-1,n})$$

But the $u_{j,n}$'s are given exactly by the Forward Euler scheme (Eq. (28)), so we have the difference equation for the errors

$$\epsilon_{j,n+1} = \alpha\epsilon_{j+1,n} + (1-2\alpha)\epsilon_{j,n} + \alpha\epsilon_{j-1,n} \tag{40}$$

where $\epsilon_{j,n} = \xi(k)^n e^{ikj\Delta x}$ is a solution. $\xi(k)$ is called the amplification factor and we see that the errors tend to infinity with each time step if $\big|\xi(k)\big| > 1$. The stability criterion for a finite difference scheme is thus $\big|\xi(k)\big| \leq 1$ for all k. To solve for the amplification factor we can put $\epsilon_{j,n} = \xi(k)^n e^{ikj\Delta x}$ into Eq. (40)

$$\xi(k)^{n+1} e^{ikj\Delta x} = \alpha\xi(k)^n e^{ik(j+1)\Delta x} + (1-2\alpha)\xi(k)^n e^{ikj\Delta x} + \alpha\xi(k)^n e^{ik(j-1)\Delta x}$$

Dividing both sides by $\xi(k)^n e^{ikj\Delta x}$

$$\xi(k) = \alpha e^{ik\Delta x} + (1-2\alpha) + \alpha e^{-ik\Delta x}$$

$$= 1 - 2\alpha + 2\alpha\cos(k\Delta x) = 1 - 2\alpha\big[1 - \cos(k\Delta x)\big]$$

$$= 1 - 4\alpha\sin^2\left(\frac{k\Delta x}{2}\right)$$

By varying k we see that this expression can at most be 1 when the sine is zero and at least be $1-4\alpha$ when the sine is ± 1. Using the stability criterion $\big|\xi(k)\big| \leq 1$ we have

$$(1-4\alpha)^2 \leq 1$$

$$\alpha \leq \frac{1}{2}$$

The Forward Euler scheme is thus conditionally stable with stability criterion $\Delta t \,/\, \Delta x^2 \leq 1 \,/\, 2$. We can repeat this process for the other finite difference schemes. For the Backward Euler scheme we have (Eq. (31))

$$\epsilon_{j,n-1} = -\alpha\epsilon_{j+1,n} + (1+2\alpha)\epsilon_{j,n} - \alpha\epsilon_{j-1,n}$$

$$\xi(k)^{n-1} e^{ikj\Delta x} = -\alpha\xi(k)^n e^{ik(j+1)\Delta x} + (1+2\alpha)\xi(k)^n e^{ikj\Delta x} - \alpha\xi(k)^n e^{ik(j-1)\Delta x}$$

Dividing both sides by $\xi(k)^n e^{ikj\Delta x}$ we get

$$\begin{aligned}\xi(k)^{-1} &= -\alpha e^{ik\Delta x} + (1+2\alpha) - \alpha e^{-ik\Delta x} \\ &= 1 + 2\alpha - 2\alpha\cos(k\Delta x) \\ &= 1 + 2\alpha\big[1 - \cos(k\Delta x)\big] \\ &= 1 + 4\alpha\sin^2\left(\frac{k\Delta x}{2}\right)\end{aligned}$$

This expression is at least 1 when the sine is zero, meaning that $|\xi(k)| \leq 1$ for all k. So the Backward Euler scheme is unconditionally stable. For the Crank Nicolson scheme we have (Eq. (35))

$$-\alpha\epsilon_{j-1,n} + 2(1+\alpha)\epsilon_{j,n} - \alpha\epsilon_{j+1,n} = \alpha\epsilon_{j-1,n-1} + 2(1-\alpha)\epsilon_{j,n-1} + \alpha\epsilon_{j+1,n-1}$$

$$\begin{aligned}&-\alpha\xi(k)^n e^{ik(j-1)\Delta x} + 2(1+\alpha)\xi(k)^n e^{ikj\Delta x} - \alpha\xi(k)^n e^{ik(j+1)\Delta x} \\ &= \alpha\xi(k)^{n-1} e^{ik(j-1)\Delta x} + 2(1-\alpha)\xi(k)^{n-1} e^{ikj\Delta x} + \alpha\xi(k)^{n-1} e^{ik(j+1)\Delta x}\end{aligned}$$

Again dividing both sides by $\xi(k)^n e^{ikj\Delta x}$ we get

$$\begin{aligned}&-\alpha e^{-ik\Delta x} + 2(1+\alpha) - \alpha e^{ik\Delta x} \\ &= \alpha\xi(k)^{-1} e^{-ik\Delta x} + 2(1-\alpha)\xi(k)^{-1} + \alpha\xi(k)^{-1} e^{ik\Delta x} \\ &2(1+\alpha) - 2\alpha\cos(k\Delta x) = \xi(k)^{-1}\big[2(1-\alpha) + 2\alpha\cos(k\Delta x)\big] \\ &1 + \alpha\big[1 - \cos(k\Delta x)\big] = \xi(k)^{-1}\Big\{1 - \alpha\big[1 - \cos(k\Delta x)\big]\Big\}\end{aligned}$$

$$1 + 2\alpha \sin^2\left(\frac{k\Delta x}{2}\right) = \xi(k)^{-1}\left[1 - 2\alpha \sin^2\left(\frac{k\Delta x}{2}\right)\right]$$

$$\xi(k) = \frac{1 - 2\alpha \sin^2(k\Delta x\,/\,2)}{1 + 2\alpha \sin^2(k\Delta x\,/\,2)}$$

This expression is 1 when the sines are zero. Otherwise the numerator is less than one and the denominator is larger than one, so it's easy to see that $|\xi(k)| \leq 1$ for all k. The Crank-Nicolson scheme is thus also unconditionally stable. Finally we have the two-dimensional Euler forward scheme (Eq. (37)) where we assume that the solution is on the form $\epsilon_{j,l}^n = \xi(k)^n e^{ikjh} e^{iklh}$. Here the index l represent y while the indices j and n still represent x and t respectively

$$\epsilon_{j,l}^{n+1} = \epsilon_{j,l}^n + \alpha\left(\epsilon_{j+1,l}^n + \epsilon_{j-1,l}^n + \epsilon_{j,l+1}^n + \epsilon_{j,l-1}^n - 4\epsilon_{j,l}^n\right)$$

$$\xi(k)^{n+1} e^{ikjh} e^{iklh} = \xi(k)^n e^{ikjh} e^{iklh}$$

$$+\alpha\xi(k)^n\left(e^{ik(j+1)h} e^{iklh} + e^{ik(j-1)h} e^{iklh} + e^{ikjh} e^{ik(l+1)h} + e^{ikjh} e^{ik(l-1)h} - 4e^{ikjh} e^{iklh}\right)$$

Dividing both sides by $\xi(k)^n e^{ikjh} e^{iklh}$ we get

$$\xi(k) = 1 + \alpha\left(e^{ikh} + e^{-ikh} + e^{ikh} + e^{-ikh} - 4\right)$$

$$= 1 + \alpha\left(4\cos(kh) - 4\right)$$

$$= 1 - 4\alpha\left[1 - \cos(kh)\right]$$

$$= 1 - 8\alpha \sin^2\left(\frac{kh}{2}\right)$$

This expression is at most 1 when the sine is zero and at least $1 - 8\alpha$ when the sine is one. The stability criterion gives

$$(1 - 8\alpha)^2 \leq 1$$

$$\alpha \leq \frac{1}{4}$$

So the two-dimensional Forward Euler scheme is conditionally stable with stability criterion $\Delta t \,/\, h^2 \leq 4$. The following table summarizes the results of this section.

Numerical scheme	**Stability criterion**
Forward Euler	$\Delta t \,/\, \Delta x^2 \leq 1 \,/\, 2$
Backward Euler	Stable for all $\Delta t,\ \Delta x$
Crank-Nicolson	Stable for all $\Delta t,\ \Delta x$
Two-dimensional Forward Euler	$\Delta t \,/\, h^2 \leq 1 \,/\, 4$

Procedure

Results

Conclusion