Extensions to the basic analysis

5.1 Adding drift

Recall the advection-diffusion equation from 3.1,

$$\frac{\partial u}{\partial t} + \mu \frac{\partial u}{\partial x} = \frac{\sigma^2}{2} \frac{\partial^2 u}{\partial x^2}. \tag{5.1}$$

We focus our analysis on the case of constant μ and σ in the first instance. The solution is then the solution of the heat equation translated by an amount μt. Previously, we thought of time to be scaled such that the variance per unit time is one. We could achieve this here retrospectively by introducing a new timescale $\tau = \sigma^2 t$, as long as $\sigma \neq 0$, and write the equation in τ and x coordinates. However so as not to rule out the degenerate case $\sigma = 0$, resulting in the pure advection equation

$$\frac{\partial u}{\partial t} + \mu \frac{\partial u}{\partial x} = 0,$$

we keep σ in.

Various numerical schemes for (5.1) were introduced in 3.1. For simplicity, the analysis of the preceding sections was restricted to the heat equation though. The techniques developed there are equally applicable – with minor modifications – to the equation with drift. Indeed, we will see that

1. the order of the truncation error is additive, and therefore the analysis of local accuracy can be performed separately for each term;

2. stability is largely determined by the highest order term, here the second derivative and its discretisation. The character of the problem only changes if $\sigma = 0$, or in practice if it σ very small.

Consistency and stability lead to convergence by exactly the same mechanism studied earlier.

5.1.1 Accuracy of central differences vs upwinding

The main approximations used previously was a *central* difference, for which Taylor expansion shows

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

i.e. second order spatial accuracy. In contrast, one-sided approximations as in

$$\begin{aligned}\frac{u(x+\Delta x,t)-u(x,t)}{\Delta x} &= \frac{\partial u}{\partial x}(x,t)+O(\Delta x),\\ \frac{u(x,t)-u(x-\Delta x,t)}{\Delta x} &= \frac{\partial u}{\partial x}(x,t)+O(\Delta x),\end{aligned}$$

are only of first order accurate. These are right- and left-sided approximations respectively. Given the nature of the equation and its solution (see 3.1), a more meaningful distinction for a particular drift μ in (5.1) is whether the finite difference is taken *in the direction of or against the drift*. For $\mu > 0$, the left-sided difference is against the drift and is called an *upwind* difference, whereas for $\mu < 0$ it is taken in the direction of the drift and is called a *downwind* difference. The opposite is the case for right-sided differences.

This leads to the definitions

$$\delta_x^- u = \operatorname{sign}(\mu)\frac{u(x,t)-u(x-\operatorname{sign}(\mu)\Delta x,t)}{\Delta x} \quad \text{(upwind difference)}, \tag{5.2}$$

$$\delta_x^+ u = \operatorname{sign}(\mu)\frac{u(x+\operatorname{sign}(\mu)\Delta x,t)-u(x,t)}{\Delta x} \quad \text{(downwind difference)}. \tag{5.3}$$

In a spirit similar to that of the θ-timestepping method, one can combine the two to

$$\begin{aligned}\mu\,\delta_x u &= \eta|\mu|\frac{u(x,t)-u(x-\operatorname{sign}(\mu)\Delta x,t)}{\Delta x}+(1-\eta)|\mu|\frac{u(x+\operatorname{sign}(\mu)\Delta x,t)-u(x,t)}{\Delta x}\\ &= \mu\,\eta\,\delta_x^- u+\mu\,(1-\eta)\,\delta_x^+ u\end{aligned}$$

with $\eta \in [0,1]$. $\eta = 0$ is downwinding, whereas $\eta = 1$ is upwinding, and $\eta = 0.5$ is the central difference scheme.

With

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

the central second difference as previously,

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

a time difference, Taylor expansion shows for a solution $u(x,t)$ to (5.1)

$$\begin{aligned}\delta_t^- u(x,t) &= \theta\left[\frac{\sigma^2}{2}\delta_x^2 u(x,t)-\mu\delta_x u(x,t)\right]+(1-\theta)\left[\frac{\sigma^2}{2}\delta_x^2 u(x,t-\Delta t)-\mu\delta_x u(x,t-\Delta t)\right]\\ &\quad+\underbrace{(1-2\theta)O(\Delta t)+O(\Delta t^2)+(1-2\eta)O(\Delta x)+O(\Delta x^2)}_{=R(x,t)},\end{aligned}$$

where the remainder term $R(x,t)$ is derived by Taylor expansion.

With the θ-η notation from above, this leads to a θ-timestepping finite difference scheme

$$\begin{aligned}\frac{u_n^m-u_n^{m-1}}{\Delta t} &= \theta\frac{1}{2}\frac{u_{n+1}^m-2u_n^m+u_{n-1}^m}{\Delta x^2}+(1-\theta)\frac{u_{n+1}^{m-1}-2u_n^{m-1}+u_{n-1}^{m-1}}{\Delta x^2} \qquad (5.4)\\ &- \theta\mu\frac{(1-\eta)u_{n+1}^m+(2\eta-1)u_n^m-\eta u_{n-1}^m}{\Delta x}-(1-\theta)\mu\frac{(1-\eta)u_{n+1}^{m-1}+(2\eta-1)u_n^{m-1}-\eta u_{n-1}^{m-1}}{\Delta x}.\end{aligned}$$

Proposition 5.1.1. *The scheme is of second order accurate in Δt for $\theta = 1/2$ (Crank-Nicolson), otherwise of first order; of second order in Δx for $\eta = 1/2$ (central first difference), otherwise of first order.*

5.1.2 Stability

In lose analogy with timestepping methods, upwinding corresponds to a backward difference, whereas a central spatial difference is similar to a time central difference. The expectation is therefore that upwinding will have best stability properties.

Assume $\mu > 0$. The case $\mu < 0$ is analogous, and $\mu = 0$ is trivial.

Example 5.1.2. *The explicit Euler scheme with central spatial differences,*

$$\frac{u_n^m - u_n^{m-1}}{\Delta t} + \mu \frac{u_{n+1}^{m-1} - u_{n-1}^{m-1}}{2\Delta x} = \frac{\sigma^2}{2} \frac{u_{n+1}^{m-1} - 2u_n^{m-1} + u_{n-1}^{m-1}}{\Delta x^2},$$

can be written as

$$u_n^m = \frac{\lambda}{2}\left(\sigma^2 - \Delta x\mu\right) u_{n+1}^{m-1} + (1 - \sigma^2\lambda)u_n^{m-1} + \frac{\lambda}{2}\left(\sigma^2 + \Delta x\mu\right) u_{n-1}^{m-1} \tag{5.5}$$

with $\lambda = \Delta t/\Delta x^2$.

By checking the signs of the coefficients, we read from (5.5) that the scheme is monotone and hence stable in the maximum norm if

$$\sigma^2\lambda \leq 1, \tag{5.6}$$
$$\Delta x\mu \leq \sigma^2. \tag{5.7}$$

The first condition (5.6) is identical to the purely diffusive case, except that here the variance is not normalised to one and explicitly taken into account. For the central difference scheme, there is an additional stability condition (5.7). This is not going to be critical as long as $\sigma \neq 0$, *because we want to let* $\Delta x \to 0$ *(or pick* Δx *small in practice). For* $\sigma = 0$ *(or, in practice, very small) this* is *problematic and the scheme is unstable.*

Example 5.1.3. *For* $\sigma = 0$, *i.e. the pure drift case, the explicit Euler scheme with upwind difference,*

$$\frac{u_n^m - u_n^{m-1}}{\Delta t} + \mu \frac{u_n^{m-1} - u_{n-1}^{m-1}}{\Delta x} = 0,$$

can be written as

$$u_n^m = (1 - \mu\lambda)u_n^{m-1} + \mu\lambda u_{n-1}^{m-1} \tag{5.8}$$

with $\lambda = \Delta t/\Delta x$. *The scheme is stable if*

$$\mu\lambda \leq 1.$$

This is not a very prohibitive condition because the scheme is of first order accurate in both Δt *and* Δx, *so choosing* $\Delta t \sim \Delta x$ *is appropriate. The condition says that the process does not drift more than one grid cell per timestep.*

The full finite difference scheme (5.4) can be re-written as

$$au_{n-1}^m + bu_n^m + cu_{n+1}^m = Au_{n-1}^{m-1} + Bu_n^{m-1} + Cu_{n+1}^{m-1}$$

with

$$\begin{aligned} a &= -\theta\frac{\lambda}{2}\left(\sigma^2 + \eta\Delta x\mu\right) \\ b &= 1 + \theta\lambda\left(\sigma^2 + (2\eta-1)\Delta x\mu\right) \\ c &= -\theta\frac{\lambda}{2}\left(\sigma^2 - (1-\eta)\Delta x\mu\right) \end{aligned}$$

and

$$\begin{aligned} A &= (1-\theta)\frac{\lambda}{2}\left(\sigma^2 + \eta\Delta x\mu\right) \\ B &= 1 - (1-\theta)\lambda\left(\sigma^2 + (2\eta-1)\Delta x\mu\right) \\ C &= (1-\theta)\frac{\lambda}{2}\left(\sigma^2 - (1-\eta)\Delta x\mu\right) \end{aligned}$$

We leave a more detailed stability analysis to 5.2.2 where it is somewhat better placed, and only anticipate the result. The essential requirement is that $A, B, C \geq 0$, which leads to the following stability conditions.

Proposition 5.1.4. *For $\mu \geq 0$, the θ-η scheme satisfies a discrete maximum principle if*

$$\begin{aligned} \lambda(1-\theta)\left(\sigma^2 + \Delta x\mu(2\eta-1)\right) &\leq 1, \\ (1-\eta)\mu\Delta x &\leq \sigma^2. \end{aligned}$$

For $\mu \leq 0$, replace μ by $-\mu$ and $1-\eta$ by η in the second inequality.

For $\theta = 1$ (fully implicit timestepping) and $\eta = 1$ (full upwinding) the scheme is unconditionally stable, in all other cases there are constraints on the timestep and grid size respectively. As we let $\Delta x \to 0$, the second condition will always be satisfied eventually if $\sigma > 0$. Similarly, the first condition is essentially a restricion on $\Delta t/\Delta x^2$, and the lower order term Δx does not alter this much for fine grids.

The von Neumann analysis of this discretisation is left as an exercise (Exercise 2). It produces no substantially different insights and the solution of the central difference scheme in Fourier space is now

$$\widehat{u}^m(k) = R(\Delta x, \Delta t; k)\widehat{u}^{m-1}(k) = R^m(\Delta x, \Delta t; k)\widehat{u}^0(k), \tag{5.9}$$

where the symbol, here given for the special case of central differencing

$$R(\Delta x, \Delta t; k) = \frac{1 - (1-\theta)[2\sigma^2\Delta t/\Delta x^2 \sin^2(k/2) + i\mu\Delta t/\Delta x \sin(k)]}{1 + \theta[2\sigma^2\Delta t/\Delta x^2 \sin^2(k/2) + i\mu\Delta t/\Delta x \sin(k)]}, \tag{5.10}$$

has an imaginary part. An interpretation of this is that the drift introduces a "phase shift". All schemes with $\theta \geq 1/2$ are still unconditionally stable in l_2, as was seen earlier for the heat equation. The stability conditions for $\theta < 1/2$ are also essentially identical to those for the heat equation for $\sigma > 0$, but become degenerate for $\sigma = 0$.

Remark 5.1.5. *On a final note regarding the stability of upwinding, it is instructive to write*

$$\frac{u_n^m - u_{n-1}^m}{\Delta x} = \frac{u_{n+1}^m - u_{n-1}^m}{2\Delta x} - \frac{1}{2}\Delta x\frac{u_{n+1}^m - 2u_n^m + u_{n-1}^m}{\Delta x^2},$$

which shows that the upstream difference is a more accurate discretisation of a term

$$\frac{\partial u}{\partial x} - \frac{1}{2}\Delta x \frac{\partial^2 u}{\partial x^2}.$$

The "numerical diffusion" has a stabilising effect. This is best visible in numerical examples where the true solution is non-smooth, e.g. has discontinuities. Upwinding "smears out" out the solution.

5.2 Boundary value problems and the matrix analysis

In the previous examples and analysis, we considered random processes on infinite coordinate axes, resulting in PDEs on unbounded spatial domains, and an infinite number of coupled discretised equations in the case of implicit finite difference schemes. In practice, to make approximation schemes for this type of equations computationally tractable, we have to restrict the equations to a finite number. To "complete" the system, boundary conditions are required at the points where we truncate the coordinates. This is already present in the continuous system where an initial boundary value problem on an interval requires two boundary conditions, usually in the form of one at each end point. Since the solution is usually unknown (it is the objective of our computation!), we have to resort to approximations and the first question is how accurate these approximations are, and consequently what is the effect of these errors at the boundary on the solution in the interior. This is discussed in the next subsection.

A further question is if the introduction of boundary conditions has a noticable effect on the accuracy and stability of discretisation schemes. The answer is no and yes. Discrete maximum principles carry over easily to boundary value problems and are somewhat easier here because we have control over the boundary values via the imposed boundary conditions and do not have to consider e.g. asymptotic growth of the solution.

The spectral analysis has more differences between finite and infinite grids. On an infinite lattice, we saw that Fourier modes with arbitrary wave length retain their shape through the timesteps of a constant coefficient finite difference scheme, where they may be dampened or magnified. In other words, the infinite grid vector is the eigenvector of a linear operator (an infinite "matrix") defined by the discretisation, and the symbol is the corresponding eigenvalue. Fourier theory gives us a clear translation between grid space and frequency space. For a (finite) matrix system, there is a finite set of eigenvectors and a discrete spectrum of eigenvalues. In simple cases, we can still find these analytically, more generally we have to use approximations. More fundamentally, the question of what the eigenvalues tell us about the stability of the scheme are less clear-cut and we have to dig deeper into matrix analysis

5.2.1 Problem formulation and discretisation

An initial-boundary value problem (IBVP)

We consider the IBVP

$$\mathcal{L}u(x,t) = 0, \qquad x \in (\underline{x}, \overline{x}), t \in (0,T) \tag{5.11}$$

$$u(\underline{x}, t) = \underline{f}(t) \qquad t \in (0,T) \tag{5.12}$$

$$u(\overline{x}, t) = \overline{f}(t) \qquad t \in (0,T) \tag{5.13}$$

$$u(x,0) = g(x) \qquad x \in [\underline{x}, \overline{x}] \tag{5.14}$$

on an interval $(\underline{x}, \overline{x})$, with a (parabolic) differential operator

$$\mathcal{L}u = \frac{\partial u}{\partial t} + \mu\frac{\partial u}{\partial x} - \frac{1}{2}\sigma^2\frac{\partial u}{\partial x}. \tag{5.15}$$

We do not necessarily assume that the coefficients are constant, but can have $\mu = \mu(x,t)$ and $\sigma = \sigma(x,t)$.

The boundary conditions are assumed of *Dirichlet* type, i.e. the value of the solution (as opposed to the derivative or other quantities) is given at the boundary points. As the solution is not known, we have to come up with approximate values. Often, an asymptotic approximation for large values of $-\underline{x}$ and $\overline{x}$ can be derived. For instance, if the initial condition g is localised (i.e. zero outside an interval), $u(x,t) \to 0$ for $x \to \pm\infty$ for all t and one can easily derive crude bounds on $u(\pm L, t)$ for large L. This allows us to choose L large enough such that $|u(\pm L, t)| \leq \epsilon$ for a desired accuracy ϵ.

The following stability result ensures that the error made by approximating the problem on the whole real line by one on a (large) finite interval is not larger than the error introduced at the boundaries.

Proposition 5.2.1. *If there are two solutions u and v, with boundary $\underline{f}_u$, $\overline{f}_u$, g_u and $\underline{f}_v$, $\overline{f}_v$, g_v respectively, then*

$$\max_{x,t}|u(x,t) - v(x,t)| \leq \max\left(\max_t|\underline{f}_u(t) - \underline{f}_v(t)|, \max_t|\overline{f}_u(t) - \overline{f}_v(t)|, \max_x|g_u(x) - g_v(x)|\right).$$

We defer the discussion of other boundary conditions to later.

Discretisation

Introducing a grid of N intervals of length $\Delta x = |\overline{x} - \underline{x}|/N$, M timesteps of length $\Delta t = T/M$, gives a discritised version of (5.11)–(5.14) as

$$Lu_n^m = 0, \qquad n \in \{1, \ldots, N-1\}, m \in \{0, \ldots, M-1\} \tag{5.16}$$
$$u_0^m = \underline{f}(t_m) \qquad m \in \{0, \ldots, M\} \tag{5.17}$$
$$u_N^m = \overline{f}(t_m) \qquad m \in \{0, \ldots, M\} \tag{5.18}$$
$$u_n^0 = g(x_n) \qquad n \in \{0, \ldots, N\} \tag{5.19}$$

where the finite difference operator at time $t_m = m\Delta t$ and grid point $x_n = n\Delta x$ reads

$$Lu_n^m = \delta_t^+ u_n^m - \left(\frac{1}{2}\sigma^2(x_n, t_{m+\theta})\delta_x^2 - \mu(x_n, t_{m+\theta})\delta_x\right)(\theta u_n^{m+1} + (1-\theta)u_n^m). \tag{5.20}$$

In the spirit of the θ-scheme, we have chosen to evaluate the time-dependent coefficients at time $t_{m+\theta} = (m+\theta)\Delta t = \theta t_{m+1} + (1-\theta)t_m$.

For $0 < n < N$, the discrete equations can be written as before,

$$a_n u_{n-1}^{m+1} + b_n u_n^{m+1} + c_n u_n^{m+1} = A_n u_{n-1}^m + B_n u_n^m + C_n u_n^m$$

where a_n, b_n, etc are given by

$$a_n = -\theta\frac{\lambda}{2}\left(\sigma^2(x_n, t_{m+\theta}) + \Delta x\mu(x_n, t_{m+\theta})\right) \tag{5.21}$$
$$b_n = 1 + \theta\lambda\sigma^2(x_n, t_{m+\theta}) \tag{5.22}$$
$$c_n = -\theta\frac{\lambda}{2}\left(\sigma^2(x_n, t_{m+\theta}) - \Delta x\mu(x_n, t_{m+\theta})\right) \tag{5.23}$$

and

$$A_n = (1-\theta)\frac{\lambda}{2}\left(\sigma^2(x_n, t_{m+\theta}) + \Delta x \mu(x_n, t_{m+\theta})\right) \tag{5.24}$$

$$B_n = 1 - (1-\theta)\lambda\sigma^2(x_n, t_{m+\theta}) \tag{5.25}$$

$$C_n = (1-\theta)\frac{\lambda}{2}\left(\sigma^2(x_n, t_{m+\theta}) - \Delta x \mu(x_n, t_{m+\theta})\right) \tag{5.26}$$

We suppress the time-dependence of the factors to keep the notation simple.

Initial and boundary conditions are evaluated pointwise. The boundary values of Dirichlet type can be eliminated from the system to obtain

$$\underbrace{\begin{pmatrix} b_1 & c_1 & 0 & \dots & 0 \\ a_2 & b_2 & c_2 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & a_{N-2} & b_{N-2} & c_{N-2} \\ 0 & \dots & 0 & a_{N-1} & b_{N-1} \end{pmatrix}}_{:=K_1} \begin{pmatrix} u_1^{m+1} \\ u_2^{m+1} \\ \vdots \\ u_{N-2}^{m+1} \\ u_{N-1}^{m+1} \end{pmatrix} = \underbrace{\begin{pmatrix} B_1 & C_1 & 0 & \dots & 0 \\ A_2 & B_2 & C_2 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & A_{N-2} & B_{N-2} & C_{N-2} \\ 0 & \dots & 0 & A_{N-1} & B_{N-1} \end{pmatrix}}_{:=K_0} \begin{pmatrix} u_1^m \\ u_2^m \\ \vdots \\ u_{N-2}^m \\ u_{N-1}^m \end{pmatrix} + d^m, \tag{5.27}$$

where

$$d^m = \begin{pmatrix} A_0 u_0^m - a_0 u_0^{m+1} \\ 0 \\ \vdots \\ 0 \\ A_N u_N^m - a_n u_N^{m+1} \end{pmatrix}.$$

This is an $(N-1) \times (N-1)$ linear system to be solved at every timestep. We are less concerned with the practicalities of this at the moment than the error analysis.

Recap: Error propagation and convergence

Solving "symbolically",

$$u^{m+1} = K_1^{-1} K_0 u^m + K_1^{-1} d^m = K u^m + K_1^{-1} d^m,$$

where $K = K_1^{-1} K_0$. Recalling the definition of the truncation error as the remainder term when inserting the exact solution u in the finite difference scheme,

$$u(\cdot, t_{m+1}) = K u(\cdot, t_m) + K_1^{-1} d^m + \Delta t T^m,$$

one gets a recursion for the error

$$e^{m+1} = u(\cdot, t_{m+1}) - u^{m+1} = K e^m + \Delta t T^m.$$

We mention explicitly that because u_0^m and u_N^m were set equal to the Dirichlet boundary data, these terms drop out. The error at final time $T = M\Delta t$ is therefore

$$e^M = \frac{T}{M} \sum_{m=0}^{M-1} K^{M-m-1} T^m,$$

such that

$$|e^M| \leq T \max_{0\leq m\leq M} \|K^m\| \max_{0\leq m\leq M} |T^m|.$$

This is basically a rerun of the analysis in 4.1, but now u^m lives on a finite grid (i.e. in a finite-dimensional space) and we can think of K more concretely as a matrix.

Consistency (viz the truncation error) as a local quality is analysed precisely as before.

Stability can be expressed in terms of the matrix K as

$$\max_{0\leq m\leq M} \|K^m\| \leq C. \tag{5.28}$$

So far, we used matrices only notationally and as data-structures for the implementation of numerical schemes. In the next section, and especially 5.2.3, we will employ results from matrix analysis for the stability analysis of difference schemes.