## Maximum principles and monotone matrices

We begin the stability analysis by revisiting the fully explicit and implicit schemes, but now with boundaries.

**Example 5.2.2** (Explicit scheme). *The explict Euler scheme for interior nodes $0 < n < N$ can be written as*

$$u_n^m = A_n u_{n-1}^{m-1} + B_n u_n^{m-1} + C_n u_{n+1}^{m-1}. \tag{5.29}$$

*If*

$$|A_n| + |B_n| + |C_n| \leq 1, \tag{5.30}$$

*then for $0 < n < N$,*

$$|u_n^m| \leq \max_{0 \leq k \leq N} |u_k^{m-1}|. \tag{5.31}$$

*Including boundaries,*

$$\max_n |u_n^M| \leq \max\{\max_n |u_n^0|, \max_{m \leq M} |u_0^m|, \max_{m \leq M} |u_N^m|\}.$$

**Remark 5.2.3.** *In fact, for $A_n, B_n, C_n$ from (5.24) to (5.26),*

$$A_n + B_n + C_n = 1, \tag{5.32}$$

*and (5.32) is automatically given if*

$$A_n, B_n, C_n \geq 0. \tag{5.33}$$

*Conditions (5.33) provide a simple test for stability of the explicit scheme. Taking a step back and looking where (5.32) comes from, it says that the first and second finite differences of a constant are zero. This is necessary for consistency and therefore a generic property of finite difference schemes. NB: This gets lost in the presence of zero order terms.*

More generally, we see from this example that a scheme of the form $u^m = K u^{m-1}$ satisfies a discrete maximum principle, if

$$\|K\|_\infty \leq 1,$$

where

$$\|K\|_\infty = \sup_{|x|_\infty \leq 1} |Kx|_\infty = \max_i \sum_j |K_{ij}|$$

is the matrix (operator) norm associated with the maximum norm in $\mathbb{R}^{N-1}$.

**Example 5.2.4** (Implicit scheme)**.** *The implict Euler scheme can be written for interior nodes $0 < n < N$ as*

$$a_n u_{n-1}^m + b_n u_n^m + c_n u_{n+1}^m = u_n^{m-1}. \tag{5.34}$$

*Here want to conclude*

$$\begin{aligned} |u_n^m| &= \frac{1}{|b_n|} \left| u_n^{m-1} - a_n u_{n-1}^m - c_n u_{n+1}^m \right| \\ &\leq \frac{|1 - a_n - c_n|}{|b_n|} \max\left( |u_n^{m-1}|, |u_{n-1}^m|, |u_{n+1}^m| \right) \\ &\leq \max\left( |u_n^{m-1}|, |u_{n-1}^m|, |u_{n+1}^m| \right), \end{aligned}$$

*which is admissible if we assume that*

$$a_n, c_n \leq 0 \tag{5.35}$$

*and*

$$a_n + b_n + c_n \geq 1 \quad \Leftrightarrow \quad 0 \leq \frac{1 - a_n - c_n}{b_n} \leq 1. \tag{5.36}$$

*Then stability follows again inductively,*

$$\max_n |u_n^M| \leq \max\{\max_n |u_n^0|, \max_{m \leq M} |u_0^m|, \max_{m \leq M} |u_N^m|\}.$$

**Remark 5.2.5.** *Similar to the previous example,*

$$a_n + b_n + c_n = 1, \tag{5.37}$$

*implying (5.36) from (5.35). Again, this is generic for finite difference schemes of parabolic equations without zero-order term. The only conditions to check for maximum norm stability is then (5.35).*

The matrix $K_1$ which defines the implicit scheme via $K_1 u^m = u^{m-1}$, has the following property.

**Definition 5.2.6** (Diagonally dominant matrix)**.** A matrix is called *diagonally dominant* if

$$|K_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |K_{ij}| \quad \text{for} \quad 1 \leq i \leq n.$$

It is *strictly diagonally dominant* if the above inequality is strict.

Looking at the matrices in $K_1 = I + \Delta t A$, $I$ is strictly diagonally dominant, $A$ diagonally dominant but not strictly, and $K_1$ strictly diagonally dominant. A strictly diagonally dominant matrix is invertible and therefore $K_1$ is invertible, which we had tacidly assumed.

Another crucial ingredient to the derivation above was the "sign-condition" (5.35), which can easily be shown to guarantee monotonicity of the scheme. In particular, the solution $u^m$ to $K_1 u^m = u^{m-1}$ with non-negative right-hand-side is non-negative. This is equivalent to $K_1^{-1}$ being non-negative elementwise. This motivates the introduction of the following class of matrices, which we will return to later.

**Definition 5.2.7** (M-matrix)**.** An $n \times n$ matrix $K = (K_{ij})_{1 \le i,j \le n}$ is called an *M-matrix* if it satisfies the following two conditions:

1. $K_{ij} \le 0$ for $i \ne j$, $1 \le i, j \le n$
2. $K$ is invertible and $(K^{-1})_{ij} \ge 0$ for $1 \le i, j \le n$

An equivalent characterisation is given (see the remarkable Theorem 5.1.1 on page 129 in [Fiedler, 2008]) if instead of property 2. in Definition 5.2.7 one demands

$$\exists x \ge 0 : \quad Kx > 0. \tag{5.38}$$

This is sometimes easier to verify in practice, as one only has to find one particular $x$. In particular, by choosing $x = (1, \dots, 1)$ one finds that a strictly diagonally dominant matrix which has positive diagonal $K_{ii} > 0$, $1 \le i \le n$, and satisfies 1. in Definition 5.2.7, is an M-matrix. M-matrices are positive definite (and, by definition, invertible).

Consider now the matrix norm

$$\|K\|_\infty \equiv \max_{|x|_\infty = 1} |Kx|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |K_{ij}|$$

associated with the $\infty$-vector norm.

For an M-matrix it is then true that,

$$\|K^{-1}\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |(K^{-1})_{ij}|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} (K^{-1})_{ij}$$

and that

$$K(1, \dots, 1)' \ge (1, \dots, 1)' \qquad \Rightarrow \qquad K^{-1}(1, \dots, 1)' \le (1, \dots, 1)',$$

hence

$$\sum_{j=1}^{n} K_{ij} \ge 1 \text{ for all } i \qquad \Rightarrow \qquad \|K^{-1}\|_\infty \le 1.$$

It therefore follows that for an M-matrix with row sums greater than 1,

$$Ku^{m+1} = u^m \qquad \Rightarrow \qquad \|u^{m+1}\|_\infty \le \|u^m\|_\infty.$$

The $l_\infty$ norm often gives an overly pessimistic assessment of stability, and we therefore study the $l_2$ norm in comparison.

### 5.2.3 Eigenvalue analysis and mean-square convergence

Going back to (5.28), we concluded a numerical scheme is going to be stable, if the sequence $K^j$ is bounded in some sense. There are two complementary viewpoints. We can think of $\Delta t$ fixed and let the number $M$ of timesteps go to infinity, i.e. moving the final time ahead. Then $K$ is fixed and for a bounded solution it is necessary that

$$\|K^m\|^{1/m} \leq C^{1/m} \to 1, \quad m \to \infty.$$

The following notion is therefore useful here.

**Definition 5.2.8** (and Lemma, Spectral radius)**.** The *spectral radius* $\rho$ of an $n \times n$ matrix $K$, with eigenvalues (spectrum) $\lambda_1, \dots, \lambda_n$, is defined by either of the following equivalent definitions:

1. $\rho(K) = \max_{i=1}^{n} |\lambda_i|$,
2. $\rho(K) = \lim_{j\to\infty} \|K^j\|^{1/j}$,

where $\|\cdot\|$ is an arbitrary matrix norm.

For the solution to be bounded in perpetuity, we need the spectral radius to be smaller or equal to one. This is only sensible to ask for if the solution to the underlying PDE itself does not grow unbounded in time.

In practice, we are usually more interested in the situation where we consider a fixed time horizon $T$, and want to improve the accuracy of a numerical approximation by increasing the number $M$ of timesteps *within* this interval, hence letting $\Delta t = T/M \to 0$ for *fixed* final time $T$. In that case we want

$$\|K(\Delta t)^m\| \leq C \;\; \forall m \in \mathbb{N},$$

where $\Delta t$ is linked to $m$ via $\Delta t = t/m$. A *sufficient* condition for stability is then

$$\|K(\Delta t)\|^m \leq C \;\; \forall m \in \mathbb{N} \quad \Leftrightarrow \quad \|K\| \leq 1 + c\Delta t$$

for a constant $c$. The scheme is unconditionally stable if $c$ is independent of $\Delta x$.

The question arises if this condition is *necessary*. A necessary condition is clearly $\rho(K) \leq 1 + c\Delta t$, otherwise components parallel to the eigenvector with the largest eigenvalue will blow up. To illustrate this, assume for the sake of the argument that the eigenvectors of $K$, say $V_j$ with eigenvalues $\lambda_j$, form a basis of $\mathbb{R}^n$, then if $V = \sum_{j=1}^{k} \mu_j V_j$,

$$K^m V = \sum_{j=1}^{k} \lambda_j^m \mu_j V_j.$$

We have thus a necessary condition for stability expressed in the eigenvalues, and a sufficient one expressed in the norm.

**Technical point: Norm vs eigenvalues**

Only in special situations is the norm of a matrix $K$ identical to the spectral radius. One important class of matrices for which this is true are *normal* matrices, defined by the property $KK' = K'K$, where the prime denotes the matrix transpose. Specifically this is true if $K$ is *symmetric*, $K = K'$. For normal (symmetric) matrices, there is a *unitary (orthogonal) transformation* $U^*KU = D$ to diagonal form, with a *unitary* (orthogonal) matrix $U$, i.e. one with $U^* = \overline{U}' = U^{-1}$. $D$ then contains the eigenvalues of $K$ and $\|K\|_2 = \rho(K)$.

However, analysing only the eigenvalues may be misleading. The eigenvectors of the matrix may not be independent, i.e. the matrix not diagonalisable. In this scenario, the spectral radius may give no indication of the "size" of the matrix. (For instance, it can be zero for non-zero matrices.) To add a further complication, if we are looking to unconditional stability across different grids, the dimension of the vectors grows when $\Delta x$ is refined along with $\Delta t$, so we are comparing solutions in different spaces. The basis transformation with the eigenvector matrix $V = (V_1, \dots, V_n)$ – when it exists – takes place in different spaces and generally the norm of $V$ will depend on the grid size $\Delta x$. Normal matrices are an exception because the unitary transformations have unit norm independent of the dimension.

The bad news is that the only normal matrices we will encounter are the disretisation matrices for the heat equation. These are in fact symmetric. The good news is that for many examples we will find, e.g., diffusion equations with drift and even with variable coefficients, the discretisation matrices are "almost" normal. In essence, the reason for this is that the leading order term comes from the highest derivative, the diffusion, and if the coefficients vary smoothly these can be seen as almost constant for fine grids.

Summing up, the eigenvalues of the discretisation matrix are going to be useful, if they describe the asymptotic behaviour of the scheme. In practice, they can often be estimated more easily than the $l_2$ norm itself. We will now investigate an example where the eigenvalues and eigenvectors are known explicitly.

**Examples and applications**

We start with a special case of the matrices in (5.27),

$$K = \mathrm{diag}(a, b, c, n) = \begin{pmatrix} b & c & & \dots & 0 \\ a & b & c & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & a & b & c \\ 0 & \dots & 0 & a & b \end{pmatrix}, \tag{5.39}$$

where the entries are constant along the diagonals. Matrices of this type are called *Töplitz* matrices. This particular structure of a tridiagonal Töplitz matrix arises for a finite difference discretisation of the differntial operator (5.15) with constant coefficients $\mu$ and $\sigma$.

**Example 5.2.9.** *The $n \times n$ matrix $K = \text{diag}(a, b, c, n)$ has eigenvalues*

$$\lambda_k = b + 2\sqrt{ac}\cos(k\xi) \tag{5.40}$$

*and corresponding eigenvectors*

$$U_k = \left(r\sin(k\xi), r^2\sin(2k\xi)\ldots, r^n\sin(k\xi n)\right) \tag{5.41}$$

*where $r = \sqrt{a/c}$ and $\xi = \pi/(n+1)$.*

Note that $K$ in (5.39) is *not* normal if $a \neq c$, i.e. if $K$ is not symmetric. To see this, it is sufficient to calculate, for $n = 2$,

$$(KK')_{11} = b^2 + c^2 \neq a^2 + b^2 = (K'K)_{11},$$

and similarly for the second diagonal entry. Exercise 5 in 5.4 shows that this is only due to the boundaries (the $2 \times 2$ system has only "boundaries") and the discretisation of the constant coefficient PDE as such is "normal". This is no longer the case for variable coefficients.

**Example 5.2.10.** *Consider $K_0$ and $K_1$ from (5.27) with constant coefficients $\sigma$ and $\mu$. We analyse the spectral radius of the iteration matrix $K = K_1^{-1}K_0$ and want to show for the spectrum $\mathcal{S}$, i.e. the set of all eigenvalues,*

$$\mathcal{S}(K) \subset (-1, 1] \quad \text{for} \quad \theta \geq 1/2. \tag{5.42}$$

*Assume $\sigma^2 < \Delta x|\mu|$ such that $a, c < 0$, then the eigenvalues of $K_1$ are*

$$\lambda_k = 1 + \theta\lambda\sigma^2(1 + \cos(k\xi)\sqrt{1 - \Delta x^2\mu^2/\sigma^4}) > 1. \tag{5.43}$$

*From (5.43), $\mathcal{S}(K_1) \subset (1, \infty)$, follows from the inverse eigenvalue theorem $\mathcal{S}(K_1^{-1}) \subset (0, 1]$, which proves the statement (5.42) for the fully implicit scheme. We now show that (5.42) is also true for $K$ from the $\theta$-scheme for $\theta \geq 1/2$. Define $A$ such $K_1 = I + \Delta t\theta A$, then*

$$\begin{aligned} K_0 &= I - (1-\theta)\Delta t A = \frac{1}{\theta}I - \frac{1-\theta}{\theta}K_1, \\ K &= K_1^{-1}K_0 = \frac{1}{\theta}K_1^{-1} - \frac{1-\theta}{\theta}I. \end{aligned}$$

*Now from the inverse eigenvalue theorem,*

$$\mathcal{S}(K_1^{-1}) \subset (0, 1] \;\Rightarrow\; \mathcal{S}(K) = \frac{1}{\theta}\mathcal{S}(K_1^{-1} - (1-\theta)I) \subset (-\frac{1-\theta}{\theta}, 1] \subset [-1, 1].$$

**Theorem 5.2.11.** *Consider the IBVP (5.11) to (5.14) with constant coefficients $\sigma > 0$ and $\mu$, such that the coefficients in the $\theta$-central difference scheme (5.16) to (5.19), (5.27) with entries (5.21) to (5.26), are independent of $n$ and $m$. Then the scheme is*

1. *unconditionally stable (convergent) in the $|\cdot|_2$ norm for $\theta \in [\frac{1}{2}, 1]$, which includes the Crank-Nicolson scheme $\theta = 1/2$;*
2. *conditionally stable (convergent) for $\theta \in [0, \frac{1}{2})$.*

*Proof.* We only cover in detail the case $\theta \in [\frac{1}{2}, 1]$ and leave the other case as an exercise. We have already shown in Example 5.2.10 that the eigenvalues of the matrix $K$ are contained in $(-1, 1)$. We now deduce that $\|K(\Delta t)^M\|$ is bounded.

Assume $\Delta x$ sufficiently small, such that $c < a < 0$. Denote by $\mathcal{S}(K)$ again the spectrum of a matrix $K$. We first observe that the eigenvectors of $K_1$ are identical to those of $K_0$, $K_1^{-1}$ and consequently to those of $K = K_1^{-1}K_0$. This means that $K$ is diagonalised by the matrix of eigenvectors $U_j$ from Example 5.2.9 with $U_{ij} = (r^i \sin(ij\xi))$, $K = UDU^{-1}$. Writing $U = RQ$ with $R = \mathrm{diag}(r^i, n)$ and $Q$ orthogonal,

$$V^m = K^m V^0 = UD^m U^{-1} V^0 = RQD^m Q' R^{-1} V^0. \tag{5.44}$$

From

$$r = \sqrt{\frac{1 + \Delta x \mu / \sigma^2}{1 - \Delta x \mu / \sigma^2}}, \tag{5.45}$$

Taylor expansion shows $1 - 3\Delta x|\mu|/\sigma^2 \leq r \leq 1 + 3\Delta x|\mu|/\sigma^2$ and $\exp(1 - 4|\mu|/\sigma^2) \leq r^n \leq \exp(1 + 4|\mu|/\sigma^2)$ for sufficiently small $\Delta x$, so

$$\|R\|, \|R^{-1}\| \leq \exp(1 + 4|\mu|/\sigma^2) \leq C.$$

The middle symmetric part $QD^mQ'$ of the matrix product in (5.44) has norm $\max_k(|D_{kk}|) = \rho(K) \leq 1$, so

$$|V^m|_2 \leq C|V^0|_2.$$

□

**Remark 5.2.12.** *The technique adopted here, i.e. transformation of the original problem to a simpler (e.g. symmetric one) with a well-behaved matrix, is a useful tool to keep in mind. The reason the transformation is stable in the above example is that the matrix is almost symmetric, in the sense that the ratio $r$ of the off-diagonals (5.45) is close to one. Similarly, for smoothly varying coefficients of the PDE, neighbouring diagonal entries are close. For an application to other non-symmetric problems, including ones with variable coefficients, see Exercises 6 and 7 in 5.4.*

In most practically relevant cases, the eigenvalues of the discretisation matrices are not known, because usually the coefficients vary with the coordinates and with time. The following result is very useful for estimating the size of the eigenvalues of a matrix in such cases.

**Lemma 5.2.13** (Gershgoring circle theorem)**.** *Let $K \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$), and define*

$$R_i = \sum_{j\neq i} |K_{ij}|.$$

1. *Each eigenvalue $\lambda_j$ lies in at least one Gershgorin disk*

$$D_i = D(K_{ii}; R_i)$$

   *in the complex plane, with centre $K_{ii}$ and radius $R_i$.*

2. *If a union of $i$ discs is disjoint from the other disks, then this union contains exactly $i$ eigenvalues.*

**Remark 5.2.14.** *It is probably clear by now that the finite-difference stability analysis of general initial-boundary value problems with variable coefficients is technically more involved than the simple von Neumann analysis. As a rule of thumb, the von Neumann stability analysis of a constant-coefficient problem of the same type normally gives a very good indication for stability of a scheme, and in practice suffices as a test.*