Exercise 3.2. Suppose that central differencing on an equidistant mesh of width $h = 1/N$ is used to solve the problem

$$-\varepsilon u'' + au' + bu = 0 \ \text{ on } (0,1), \quad u(0) = 0,\ u(1) = 1,$$

where $a > 0$ and $b \geq 0$. Assume that $\min_x a(x) > 2\varepsilon/h$. By considering the signs of the coefficients in the difference scheme, show that the computed solution $\{u_i^N\}_{i=0}^N$ has $u_i^N u_{i+1}^N < 0$ for $i = 1, 2, \ldots, N-1$. This conclusion says that the computed solution oscillates around zero (the solution of the reduced problem) analogously to what we see in Figure 3.1. *Hint.* A similar argument can be found in [**MS12**].

Why does this standard method give us these oscillations? What has gone wrong? In the next section we will reveal the answers to these questions.

3.1. M-matrices, upwinding

A square matrix $A = (A_{ij})$ is said to be an *M-matrix*[1] if $A_{ij} \leq 0$ for all $i \neq j$ and A^{-1} exists with $(A^{-1})_{ij} \geq 0$ for all i, j. Difference schemes that employ M-matrices are common because they are desirable: they are generally stable and are more amenable to analysis.

Exercise 3.3. Let $A = (A_{ij})$ be an M-matrix. Prove that $A_{ii} > 0$ for all i.

Our central difference scheme above fails to satisfy the M-matrix sign condition on the off-diagonal entries since $B_{i,i+1} > 0$ when ε is small relative to h. If $h\|a\|_\infty \leq 2\varepsilon$, then the sign condition is satisfied, and it turns out that the difference method gives an acceptable computed solution, but to enforce this inequality when ε is small is impractical in many problems (especially in partial differential equations, where multiple dimensions are involved) since it can lead to an intolerable number of mesh points.

The second M-matrix requirement—that A^{-1} exists with $(A^{-1})_{ij} \geq 0$ for all i and j—does not seem easy to verify in practice. Fortunately there are more tractable alternatives, as stated in the next two lemmas.

A square matrix $A = (A_{ij})$ is said to be *strictly diagonally dominant* if $A_{ii} > \sum_{j \neq i} |A_{ij}|$ for all i.

Lemma 3.4. *Suppose that the square matrix $A = (A_{ij})$ satisfies $A_{ij} \leq 0$ for all $i \neq j$. Then A^{-1} exists and $(A^{-1})_{ij} \geq 0$ for all i, j if A is strictly diagonally dominant with $A_{ii} > 0$ for all i.*

Proof. See, e.g., [**QV94**, Lemma 2.1.1]. □

[1] *Historical Note.* The "M" in M-matrix refers to Hermann Minkowski, who studied some of their properties. M-matrices have been exhaustively analysed in the research literature.

Consider a vector $\mathbf{w} = (w_1, w_2, \ldots, w_n) \in \mathbb{R}^n$. By $\mathbf{w} > \mathbf{0}$ we mean that $w_i > 0$ for $i = 1, 2, \ldots, n$. Similarly, $\mathbf{w} \geq \mathbf{0}$ means $w_i \geq 0$ for all i. We set $|\mathbf{w}| = (|w_1|, |w_2|, \ldots, |w_n|)$. The discrete L^∞ norm $\|\cdot\|_{\infty,d}$ for vectors in $\mathbb{R}^n$ is defined by $\|\mathbf{w}\|_{\infty,d} = \max_i |w_i|$. The matrix norm $\|\cdot\|_{\infty,d}$ is the norm induced by the corresponding vector norm $\|\cdot\|_{\infty,d}$; for the $n \times n$ matrix $A = (A_{ij})$ it is the "maximum row sum" norm, viz., $\|A\|_{\infty,d} = \max_i \sum_j |A_{ij}|$. Matrix norms induced by vector norms are discussed in many basic numerical analysis books.

Lemma 3.5. *Suppose that the $n \times n$ matrix $A = (A_{ij})$ satisfies $A_{ij} \leq 0$ for all $i \neq j$. Then A^{-1} exists and $(A^{-1})_{ij} \geq 0$ for all i, j if and only if there exists a vector $\mathbf{w} > 0$ in $\mathbb{R}^n$ such that $A\mathbf{w} > 0$. Furthermore, we have*

$$\|A^{-1}\|_{\infty,d} \leq \frac{\|\mathbf{w}\|_{\infty,d}}{\min_k (A\mathbf{w})_k}. \tag{3.3}$$

Proof. See [**Boh81**] or [**AK90**]. □

One can often construct a vector $\mathbf{w}$ that satisfies the conditions of Lemma 3.5 by first finding a function $w(x)$ such that $w > 0$ and $Lw > 0$, then restricting w to the mesh to form $\mathbf{w}$.

For M-matrices we have the following discrete analogues of Lemma 1.8 and Corollary 1.12.

Lemma 3.6 (Discrete maximum principle)**.** *Let A be an M-matrix. If $\mathbf{w}$ is a vector with $A\mathbf{w} \geq \mathbf{0}$, then $\mathbf{w} \geq \mathbf{0}$.*

Proof. $\mathbf{w} = (A^{-1})(A\mathbf{w}) \geq \mathbf{0}$, because $A^{-1} \geq 0$ and $A\mathbf{w} \geq \mathbf{0}$. □

Lemma 3.7 (Discrete barrier function)**.** *Let A be an M-matrix. If $\mathbf{w}, \mathbf{z}$ are vectors such that $|A\mathbf{w}| \leq A\mathbf{z}$, then $|\mathbf{w}| \leq \mathbf{z}$.*

Proof. Now $A(\mathbf{z} - \mathbf{w}) \geq \mathbf{0}$, so $\mathbf{w} - \mathbf{z} \geq \mathbf{0}$ by Lemma 3.6. Similarly, one has $\mathbf{z} + \mathbf{w} \geq \mathbf{0}$, and the result follows. □

When we take A to be the matrix arising from a discretisation of a boundary value problem, at first sight the boundary data requirement of Corollary 1.12 (the continuous analogue of Lemma 3.7) seems to be missing from Lemma 3.7, but this is deceptive. The first and last rows of A will include this information—see the construction of our matrix B above.

Returning to our difference scheme and its failure to generate an M-matrix, we see that the "incorrect" sign of $B_{i,i+1}$ comes from the central difference approximation $u'(x_i) \approx (u_{i+1}^N - u_{i-1}^N)/(2h)$. This approximation is generally recommended in basic courses in numerical methods because it gives an $\mathcal{O}\left(h^2\right)$ consistency error, but this consistency property is useless when the method is (as we saw) unstable. To cure the instability, for

convection-diffusion problems one can approximate $u'(x_i)$ by the *simple upwinding* formula $(u_i^N - u_{i-1}^N)/h$. Although the consistency error is now only $\mathcal{O}(h)$, the ith row of the scheme is

$$\left(0 \cdots 0 \quad -\frac{\varepsilon}{h^2} - \frac{a_i}{h} \qquad \frac{2\varepsilon}{h^2} + \frac{a_i}{h} + b_i \qquad -\frac{\varepsilon}{h^2} \qquad 0 \cdots 0\right),$$

which has the correct sign pattern. Hence, writing B for the associated $(N+1) \times (N+1)$ matrix that incorporates the boundary conditions, one has $B_{ij} \leq 0$ for $i \neq j$, as desired.

Lemma 3.8. *The coefficient matrix B for the simple upwind scheme is an M-matrix, and the scheme is uniformly stable with respect to the perturbation parameter*

$$\|u_h\|_{\infty,d} \leq C \|B u_h\|_{\infty,d},$$

with a stability constant C that is independent of ε and h.

Proof. Clearly, $B_{ij} \leq 0$ for $i \neq j$. We construct a suitable majorizing vector. Choose $w(x) := 1 + x$, so $Lw(x) \geq \alpha$. Let $\mathbf{v}$ be the restriction of w to the mesh. A quick computation yields $B\mathbf{v} \geq \min\{1, \alpha\}\mathbf{1}$, where $\mathbf{1} = (1, 1, \ldots, 1)^T$. Thus by Lemma 3.5 the matrix B is an M-matrix, and one gets the desired stability bound with stability constant $C = 1/\min\{1, \alpha\}$. □

Simple upwinding for (2.14) uses the one-sided difference $(u_i^N - u_{i-1}^N)/h$ to approximate $u'(x_i)$, but the alternative one-sided difference $(u_{i+1}^N - u_i^N)/h$ would not give the correct sign pattern in the matrix. Upwinding (of which there are many variants) means taking a nonsymmetric finite difference approximation that is *weighted away from the layer*. With simple upwinding, for $\varepsilon \ll h^2$ the scheme almost decouples the boundary condition at $x = 1$ from the values at the interior nodes. This is exactly what is needed to avoid the anomaly described in Remark 3.1.

Remark 3.9. In its various forms, upwinding uses discretisations of the convection term that are suitable for solving the reduced problem (2.18). This is more evident when dealing with the reduced problem (4.3) for convection-diffusion problems posed on domains in two dimensions. The construction and analysis of numerical methods for such "first-order hyperbolic" equations and their nonlinear generalisations has been the subject of much research.

Figure 3.2, where $N = 10$ so the mesh points are $0, 0.1, 0.2, \ldots, 1$, illustrates the difference between the central difference and upwind approximations of $u'(x_{N-1})$ in the typical case when $N \ll 1/\varepsilon$. Clearly, the central difference approximation (the slope of the dashed line through the