ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Botchev M.A.

Some topics in matrix analysis for time integration methods

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. КЕЛДЫША
Российской академии наук

M.A. Botchev

Some topics in matrix analysis for time integration methods

Москва — 2018

Михаил Александрович Бочев

Некоторые вопросы матричного анализа методов интегрирования по времени. Препринт Института прикладной математики им. М.В. Келдыша РАН, Москва, 2018.

Данный препринт содержит конспекты лекций, прочитанных в 2016 г. на Римско-Московской школе по матричным методам и прикладной линейной алгебре. Лекции посвящены некоторым задачам матричного анализа, возникающих при разработке и анализе схем интегрирования по времени систем обыкновенных дифференциальных уравнений и дифференциальных уравнений в частных производных. Материал лекций включает некоторые аспекты конечно-разностных пространственных аппроксимаций уравнений конвекции–диффузии (в контексте метода прямых), устойчивости систем обыкновенных дифференциальных уравнений, логарифмической матричной нормы и её применения, явно-неявных схем, методов расщепления, методов Розенброка в сочетании с приближёнными разложениями якобиана и схем с матричной экспонентой на основе подпространств Крылова. Препринт предназначен для аспирантов и студентов, а также для научных работников для ознакомления с указанной тематикой.

Ключевые слова: метод прямых, уравнение конвекции–диффузии, устойчивость дифференциальных уравнений и разностных схем, матричная экспонента, логарифмическая матричная норма, методы Розенброка, подпространства Крылова.

Mikhail A. Botchev

Some topics in matrix analysis for time integration methods. Preprint of Keldysh Institute of Applied Mathematics RAS, Moscow, 2018.

This report contains lecture notes used for the 2016 edition of the Rome-Moscow school of Matrix Methods and Applied Linear Algebra, held in Moscow and Rome (respectively, in August and September 2016). The notes deal with some matrix analysis problems which arise in construction and analysis of time integration methods for solving large systems of ordinary and partial differential equations (ODEs and PDEs). The material treated includes some aspects of finite-difference approximation of convection–diffusion operators (used, following the framework of the methods of lines, to reduce time-dependent convection–diffusion problems to ODE systems), stability of the ODE systems, the logarithmic matrix norm, stability of the implicit–explicit θ-method, splitting methods, Rosenbrock methods with approximate matrix factorizations and Krylov subspace exponential time integration.

Key words: method of lines, convection–diffusion equation, stability of differential equations and difference schemes, matrix exponent, logarithmic matrix norm, Rosenbrock methods, Krylov subspace.

1 Some facts from matrix analysis

Here we list some definitions and results, mostly without proofs, which will be used in our lectures. Marks $\Diamond$ and $\Box$ denote the end of an exercise and a proof, respectively.

By a vector $x \in \mathbb{C}^n$ we mean a *column* vector. Hence, the conventional inner product in $\mathbb{C}^n$ can be defined as

$$(x, y) = y^* x, \quad x, y \in \mathbb{C}^n,$$

where y^* denotes the conjugate of y. Let $A \in \mathbb{R}^{n\times n}$. The set of all the eigenvalues of A is called its spectrum. By $\rho(A)$ we denote the spectral radius of A, defined as

$$\rho(A) = \max\{|\lambda| \mid \lambda \in \text{spectrum of } A\}.$$

We now define some the following vector norms, called the 2-norm, the 1-norm and the max-norm, respectively,

$$\|x\|_2 = \sqrt{\sum_{k=1}^{n} |x_k|^2}, \quad \|x\|_1 = \sum_{k=1}^{n} |x_k|, \quad \|x\|_\infty = \max_{1\leqslant k\leqslant n} |x_k|. \tag{1.1}$$

The associated matrix norms, $\|A\| = \max_{x\neq 0}(\|Ax\|/\|x\|)$, are

$$\|A\|_2 = \sqrt{\rho(A^*A)}, \qquad \|A\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{ij}|, \qquad \|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|. \tag{1.2}$$

Note that $\|A\|_2$ can not be computed by an explicit formula. In fact, computing the 2-norm of a large matrix can be quite expensive. Thus, if we only need some norm of a matrix, we should avoid computing the 2-norm[1].

For any two matrix norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ the sharpest constant can be found such that for any matrix $A \in \mathbb{C}^{n\times n}$ holds $\|A\|_\alpha \leqslant C_{\alpha,\beta}\|A\|_\beta$. These constants are [21, Section 5.6]

$$\begin{aligned} C_{1,2} &= \sqrt{n}, \quad C_{1,\infty} = n, \\ C_{2,1} &= \sqrt{n}, \quad C_{2,\infty} = \sqrt{n}, \\ C_{\infty,1} &= n, \quad C_{\infty,2} = \sqrt{n}. \end{aligned} \tag{1.3}$$

A square matrix P is called a permutation matrix if its columns can permuted in such a way that the identity matrix is obtained.

A square matrix A is called *reducible or decomposable* if there exists a permutation matrix $\hat{P}$ such that

$$\hat{P}A\hat{P}^T = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix},$$

[1] Note that `norm(A)` in Matlab or Octave computes $\|A\|_2$. $\|A\|_1$ and $\|A\|_\infty$ can be computed as `norm(A,1)` and `norm(A,'inf')`, respectively.

where the matrices A_{11} and A_{22} are square. Otherwise, A is called *irreducible or nondecomposable* [28, 29]. Irreducible matrices are most easily characterized by their directed graphs. A directed graph of a matrix $A \in \mathbb{R}^{n\times n}$ is a graph of n vertices, where there is a directed link (an arrow) from vertex i to vertices j as soon as $a_{ij} \neq 0$. A matrix is irreducible if and only if its graph is (strongly) connected, i.e., there exists a directed path between any two vertices i and j [51, 28].

Theorem 1.1 *(Perron-Frobenius theorem) Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be an elementwise nonnegative ($a_{ij} \geqslant 0$) irreducible matrix. Then*
(1) A has a positive eigenvalue λ which equals the spectral radius of A: $\lambda = \rho(A)$;
(2) the eigenvector $x = (x_i)$ corresponding to λ can be chosen elementwise strictly positive: $x_i > 0$, $i = 1, \dots, N$;
(3) λ is a simple eigenvalue.
If the matrix A is only elementwise nonnegative then $\rho(A)$ is an eigenvalue of A and the corresponding eigenvector can be chosen elementwise nonnegative [21, Theorem 8.3.1].

A matrix $A \in \mathbb{R}^{n\times n}$ is called weakly diagonally dominant if

$$|a_{ii}| \geqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|, \qquad i = 1, \dots, n. \tag{1.4}$$

If the strict inequalities hold here for all i, the matrix is called strictly diagonally dominant. A matrix is called irreducibly diagonally dominant if it is irreducible, weakly diagonally dominant and at least for one i the diagonal dominance inequality holds strictly [28, 29].

Theorem 1.2 *[51, 28] A strictly or irreducibly diagonally dominant matrix is nonsingular and has nonzero diagonal entries.*

Exercise 1.1 Prove Theorem 1.2 for a strictly diagonally dominant matrix. $\Diamond$

A matrix A is called an M-matrix [47] if $A = sI - B$, with B being elementwise nonnegative ($B \geqslant 0$) and $s > \rho(B)$. If $s = \rho(B)$ then A is singular and is called a singular M-matrix [22, 33].

The following theorem can often be useful in establishing the property of being an M-matrix.

Theorem 1.3 *Let $A \in \mathbb{R}^{n\times n}$ be weakly diagonally dominant and let*

$$a_{ii} \geqslant 0, \ i = 1, \dots, n, \qquad a_{ij} \leqslant 0, \ i \neq j.$$

Then the eigenvalues of A have nonnegative real part and A is a possibly singular M-matrix.

Proof [22, Section 2.5] Define

$$s = \max_i a_{ii}, \qquad B = sI - A.$$

Note that B is elementwise nonnegative and it can be checked that

$$\|B\|_\infty \leqslant s. \tag{1.5}$$

Thus, from $\rho(B) \leqslant \|B\|_\infty$ follows $\rho(B) \leqslant s$ and A is a (possibly singular) M-matrix.
Furthermore, all the eigenvalues $\lambda(A)$ of A belong to the set $\{z \in \mathbb{C} \mid |z - s| \leqslant \rho(B)\}$. Hence, we have $\operatorname{Re}\lambda(A) \geqslant s - \rho(B)$. By the last part of Theorem 1.1, $\rho(B)$ is an eigenvalue of B and, thus, $s - \rho(B) \geqslant 0$ is an eigenvalue of A with the smallest real part. Hence, if it is known that A is nonsingular, for instance, by Theorem 1.2, then $s > \rho(B)$. □

Exercise 1.2 In the proof of Theorem 1.3, check that (1.5) holds. ♢

A regular splitting of $A \in \mathbb{R}^{n\times n}$ is a representation [47]

$$A = P - Q,$$

where P is nonsingular and P^{-1} and Q are elementwise nonnegative.

Theorem 1.4 *[47, 33] If A is a possibly singular M-matrix and $A = P - Q$ is its regular splitting then $\rho(P^{-1}Q) \leqslant 1$. The last inequality becomes strict as soon as A is nonsingular.*

A matrix $A = (a_{ij})$ is called Hermitian if A equals its conjugate transpose $A^* = (\bar{a}_{ji})$. If $A = A^T$ the matrix A is called symmetric. Real Hermitian matrices are symmetric. The eigenvalues of a Hermitian matrix are real. Skew-Hermitian matrices (i.e., matrices A for which holds $A^* = -A$) have purely imaginary eigenvalues. For every matrix $A \in \mathbb{R}^{n\times n}$, its symmetric and skew-symmetric parts are defined respectively as $\frac{1}{2}(A + A^T)$ and $\frac{1}{2}(A - A^T)$. A real square matrix is uniquely determined by its symmetric and skew-symmetric parts. Spectrum $\Lambda(A)$ of a real square matrix A is contained in a rectangular domain in the complex plane,

$$\Lambda(A) \subset [a, b] \times [-ic, ic] \subset \mathbb{C}, \qquad a, b, c \in \mathbb{R},$$

where a and b are respectively minimum and maximum eigenvalues of $\frac{1}{2}(A+A^T)$ and c is the spectral radius of $\frac{1}{2}(A - A^T)$.

2 The problem we solve. Examples

In these lectures we discuss some matrix problems arising when the following *initial value problem (IVP)* is solved numerically. For given $A \in \mathbb{R}^{n\times n}$, $g(t) : \mathbb{R} \to \mathbb{R}^n$ and $y^0 \in \mathbb{R}^n$, find a vector function $y(t) : \mathbb{R} \to \mathbb{R}^n$ such that

$$y'(t) = -Ay(t) + g(t), \qquad y(0) = y^0. \tag{2.1}$$

Such problems arise in many contexts, for instance, when we solve numerically parabolic or hyperbolic partial differential equations (PDEs). A possible way to solve a given time-dependent PDE is to first discretize it in space. If the PDE linear, the spatial discretization yields (2.1), which is then solved by a time integration method. This solution approach is called the method of lines (MOL) [23, Sect. 6.1].

In Section 5.5, we also briefly discuss solution of a more general, nonlinear autonomous ODE system $y'(t) = -Ay - R(y)$ known as advection–diffusion–reaction problem. Unless reported differently, we usually assume that A in (2.1) is such that its symmetric part $\frac{1}{2}(A + A^T)$ is positive definite.

2.1 Example: unsteady convection–diffusion problem

Assume $\Omega \subset \mathbb{R}^2$ is a domain with a smooth boundary $\partial\Omega$ and $L[u]$ is a linear differential operator acting on functions $u(x, y, t)$ from some functional space, with $t > 0$ and $(x, y) \in \Omega$. Consider the following problem:

$$\begin{aligned} &\text{(a)} \quad \frac{\partial u}{\partial t} + L[u] = \tilde{g}(x, y, t), \\ &\qquad\; u = u(x, y, t), \quad (x, y) \in \Omega, \quad t > 0, \\ &\text{(b)} \quad u(x, y, 0) = u^0(x, y), \\ &\text{(c)} \quad \text{conditions on } u(x, y, t)\big|_{(x,y)\in\partial\Omega} \text{ and its derivatives,} \end{aligned} \tag{2.2}$$

where the functions $\tilde{g}(x, y, t)$, $u^0(x, y)$ are given and $u(x, y, t)$ is unknown. This problem is an *initial-boundary-value problem* because relations (2.2)(b), (2.2)(c) provide respectively initial and boundary conditions on the unknown function $u(x, y, t)$.

Solving (2.2) numerically by the *method of lines* approach, we first discretize the partial differential equation (PDE) given by (2.2)(a) in space and then arrive to a system of ordinary differential equations (ODEs):

$$\frac{\partial u}{\partial t} + L[u] = \tilde{g}(x, y, t) \quad \overset{\text{space discretization}}{\rightarrow} \quad y'(t) = -Ay(t) + g(t), \tag{2.3}$$

where the vector function $y(t)$, $y : \mathbb{R} \to \mathbb{R}^n$ approximates the unknown function $u(x, y, t)$ at n discrete points $(x_i, y_k) \in \Omega$, the matrix $A \in \mathbb{R}^{n\times n}$ approximates the operator $L[\cdot]$, $Aw \approx L[u]$, and $g(t)$ is a vector function, $g : \mathbb{R} \to \mathbb{R}^n$, whose coordinate functions $g_i(t)$ contain the values $\tilde{g}(x_i, y_k, t)$, plus possibly some contributions from boundary conditions (2.2)(c). The boundary conditions are also taken into account by the structure of the matrix A.

We now describe two simple finite difference space discretizations (2.3) of a nonstationary convection–diffusion problem. This problem is given by (2.2) with

$$L[u] = -(D_1 u_x)_x - (D_2 u_y)_y + v_1 u_x + v_2 u_y + Du, \tag{2.4}$$

where the given functions D_i, v_i and D satisfy

$$\begin{aligned} D_i = D_i(x,y) \geqslant 0, \quad v_i = v_i(x,y), \qquad i = 1,2, \\ D_1 + D_2 > 0, \quad (v_1)_x + (v_2)_y \equiv 0, \\ D = D(x,y) \geqslant 0, \qquad (x,y) \in \Omega, \end{aligned} \tag{2.5}$$

and the subindices $\cdot_x$ and $\cdot_y$ denote the derivatives with respect to x and y, respectively. D_i and v_i are called diffusion and convection coefficients, respectively. For simplicity, we assume that the domain Ω is convex and boundary conditions (2.2)(c) are homogeneous:

$$u\big|_{\partial\Omega} = 0, \qquad t > 0.$$

One of the finite-difference discretizations which we describe yields a matrix A such that $L[u] \approx Ay$ and

$$\begin{gathered} A_0 y = L_{\text{diff}}[u], \qquad A_1 y = L_{\text{conv}}[u], \\ L_{\text{diff}}[u] \equiv -(D_1 u_x)_x - (D_2 u_y)_y + Du, \qquad L_{\text{conv}}[u] \equiv v_1 u_x + v_2 u_y, \end{gathered} \tag{2.6}$$

where A_0 and A_1 are the symmetric and skew-symmetric parts of A, respectively. In other words, the symmetric part of A approximates the diffusion terms, and the skew-symmetric part the convection terms.

2.2 Finite difference relations. Central differences

We introduce a uniform Cartesian mesh covering Ω and consisting of n points $(x_i, y_k) \in \Omega$. The mesh has mesh sizes $h_1 > 0$ in the x-direction and $h_2 > 0$ in the y-direction, i.e., $x_{i+1} - x_i = h_1$, $y_{k+1} - y_k = h_2$ for all possible i and k. At each node (x_i, y_k) of the mesh, we approximate the terms in $L_{\text{diff}}[u]$ by finite differences as follows

$$\begin{aligned} (D_1 u_x)_x &\approx \frac{(D_1)_{i+1/2,k}(u_{i+1,k} - u_{i,k}) - (D_1)_{i-1/2,k}(u_{i,k} - u_{i-1,k})}{h_1^2}, \\ (D_2 u_y)_y &\approx \frac{(D_2)_{i,k+1/2}(u_{i,k+1} - u_{i,k}) - (D_2)_{i,k-1/2}(u_{i,k} - u_{i,k-1})}{h_2^2}, \\ Du &\approx D_{i,k} u_{i,k}, \end{aligned} \tag{2.7}$$

where the subindices $(\cdot)_{i,k}$ refer to the point (x_i, y_k), the subindices $(\cdot)_{i\pm1,k}$, $(\cdot)_{i,k\pm1}$, $(\cdot)_{i\pm1/2,k}$, $(\cdot)_{i,k\pm1/2}$ refer to the points shifted from (x_i, y_k) respectively by $\pm h_1$, $\pm h_2$, $\pm h_1/2$ or $\pm h_2/2$ in the x or y direction.

Before giving the finite difference relations for the convection terms, we rewrite them in the following form [24] (see relation (2.5)):

$$v_1 u_x + v_2 u_y = \frac{1}{2}(v_1 u_x + v_2 u_y) + \frac{1}{2}((v_1 u)_x + (v_2 u)_y). \tag{2.8}$$

The reason why we rewrite the convective terms in this way, will become clear a little bit later, in Exercise 2.2. Then, the finite difference approximations for these terms at the point (x_i, y_k) are given by

$$\begin{aligned}\frac{1}{2}(v_1u_x + (v_1u)_x) &\approx \frac{(v_1)_{i,k}(u_{i+1,k} - u_{i-1,k}) + ((v_1)_{i+1,k}u_{i+1,k} - (v_1)_{i-1,k}u_{i-1,k})}{4h_1},\\ \frac{1}{2}(v_2u_y + (v_2u)_y) &\approx \frac{(v_2)_{i,k}(u_{i,k+1} - u_{i,k-1}) + ((v_2)_{i,k+1}u_{i,k+1} - (v_2)_{i,k-1}u_{i,k-1})}{4h_2}.\end{aligned} \tag{2.9}$$

Combining relations (2.7) and (2.9), we can approximate the operator $L[u]$ at each mesh point (x_i, y_k) as follows:

$$L[u]\bigg|_{(x_i,y_k)} \approx W_{i,k}u_{i-1,k} + S_{i,k}u_{i,k-1} + C_{i,k}u_{i,k} + N_{i,k}u_{i,k+1} + E_{i,k}u_{i+1,k}, \tag{2.10}$$

with

$$\begin{aligned}W_{i,k} &= -\frac{(D_1)_{i-1/2,k}}{h_1^2} - \frac{(v_1)_{i,k} + (v_1)_{i-1,k}}{4h_1},\\ S_{i,k} &= -\frac{(D_2)_{i,k-1/2}}{h_2^2} - \frac{(v_2)_{i,k} + (v_2)_{i,k-1}}{4h_2},\\ C_{i,k} &= \frac{(D_1)_{i-1/2,k} + (D_1)_{i+1/2,k}}{h_1^2} + \frac{(D_2)_{i,k-1/2} + (D_2)_{i,k+1/2}}{h_2^2} + D_{i,k},\\ N_{i,k} &= -\frac{(D_2)_{i,k+1/2}}{h_2^2} + \frac{(v_2)_{i,k} + (v_2)_{i,k+1}}{4h_2},\\ E_{i,k} &= -\frac{(D_1)_{i+1/2,k}}{h_1^2} + \frac{(v_1)_{i,k} + (v_1)_{i+1,k}}{4h_1}.\end{aligned}$$

The notation in (2.10) corresponds to the positions of the nodes of the finite difference stencil $u_{i\pm1,k\pm1}$ with respect to the central node $u_{i,k}$ ($u_{i+1,k}$ lies to the East from $u_{i,k}$, hence we write $E_{i,k}u_{i+1,k}$; $u_{i,k+1}$ lies to the North, hence we denote $N_{i,k}u_{i,k+1}$, etc.)

2.3 Structure of the matrix

Written for every mesh point, the relations (2.10) can be combined into a matrix-vector product. Assume $\Omega = [0,1] \times [0,1]$ and the mesh is given by

$$x_i = ih_1, \quad y_k = kh_2, \quad \text{with} \quad h_j = 1/n_j, \quad j = 1, 2. \tag{2.11}$$

Writing the relations (2.10) successfully for the points
(x_1, y_1), (x_2, y_2), ..., (x_1, y_{n_2-1}),
(x_2, y_1), (x_2, y_2), ..., (x_2, y_{n_2-1}),
...

(x_{n_1-1}, y_1), (x_{n_1-1}, y_2), …, (x_{n_1-1}, y_{n_2-1}), we obtain:

$$\begin{aligned}
(x_1, y_1):&\quad W_{1,1}u_{0,1} + S_{1,1}u_{1,0} + C_{1,1}u_{1,1} + N_{1,1}u_{1,2} + E_{1,1}u_{2,1},\\
(x_1, y_2):&\quad W_{1,2}u_{0,2} + S_{1,2}u_{1,1} + C_{1,2}u_{1,2} + N_{1,2}u_{1,3} + E_{1,2}u_{2,2},\\
(x_1, y_3):&\quad W_{1,3}u_{0,3} + S_{1,3}u_{1,2} + C_{1,3}u_{1,3} + N_{1,3}u_{1,4} + E_{1,3}u_{2,3},\\
&\quad \dots\\
(x_1, y_{n_2-1}):&\quad W_{1,n_2-1}u_{0,n_2-1} + S_{1,n_2-1}u_{1,n_2-2} + C_{1,n_2-1}u_{1,n_2-1}\\
&\qquad + N_{1,n_2-1}u_{1,n_2} + E_{1,n_2-1}u_{2,n_2-1},\\
(x_2, y_1):&\quad W_{2,1}u_{1,1} + S_{2,1}u_{2,0} + C_{2,1}u_{2,1} + N_{2,1}u_{2,2} + E_{2,1}u_{3,1},\\
&\quad \dots\\
(x_{n_1-1}, y_{n_2-1}):&\quad W_{n_1-1,n_2-1}u_{n_1-2,n_2-1} + S_{n_1-1,n_2-1}u_{n_1-1,n_2-2} + C_{n_1-1,n_2-1}u_{n_1-1,n_2-1}\\
&\qquad + N_{n_1-1,n_2-1}u_{n_1-1,n_2} + E_{n_1-1,n_2-1}u_{n_1,n_2-1},\\
&\quad \dots
\end{aligned}$$

These relations, where $u_{i,k} = 0$ for $i = 0$, $i = n_1$, $k = 0$ and $k = n_2$ due to the homogeneous boundary conditions (2.2)(c), can be cast into the matrix-vector product form:

$$\begin{bmatrix}
C_{1,1} & N_{1,1} & \dots & \dots & E_{1,1} & & & \\
S_{1,2} & C_{1,2} & N_{1,2} & \dots & & E_{1,2} & & \\
 & \ddots & \ddots & & & & \ddots & \\
 & & S_{1,n_2-1} & C_{1,n_2-1} & 0 & & & \\
W_{2,1} & & & 0 & C_{2,1} & N_{2,1} & & \\
 & W_{2,2} & & & S_{2,2} & C_{2,2} & N_{2,2} & \\
 & & \ddots & & & \ddots & \ddots & \ddots
\end{bmatrix}
\begin{bmatrix}
u_{1,1}\\ u_{1,2}\\ \vdots\\ u_{1,n_2-1}\\ u_{2,1}\\ u_{2,2}\\ \vdots
\end{bmatrix} = Aw, \quad (2.12)$$

where we denote the matrix by A and the vector by w. The matrix A is five-diagonal with the main diagonal containing the coefficients $C_{i,k}$, the sub- and superdiagonals containing respectively $S_{i,k}$ and $N_{i,k}$ and two additional diagonals containing $W_{i,k}$ and $E_{i,k}$.

Note that each node in the finite difference mesh corresponds to a row in A and we could use any order of nodes, when forming A. Thus, the structure of A depends on the chosen node ordering, see e.g. [35] for more detail.

Note that the coordinates of the vector w in (2.12) are in fact functions of the time t, they are approximations of the unknown function $u(x, y, t)$ at the mesh points (x_i, y_k). Replacing in (2.2)(a) $\partial u/\partial t$ by a vector of the time derivatives of the coordinates of w and $L[u]$ by Aw, we obtain a system of ODEs (see (2.3)).

Exercise 2.1 How many zero entries appear in the first row of the matrix A between $N_{1,1}$ and $E_{1,1}$? Assume $n_1 = 20$, $n_2 = 10$. Which entry does A have in the position (9,10)? In the position (10,11)? Assume $n_2 = 5$. Write down the first five coordinates of the vector function $g(t)$ in (2.3). ♢

Exercise 2.2 Show that finite difference approximation (2.7), (2.9) possesses the property given in (2.6): the diffusion terms yield the Hermitian part of A, the convection terms the skew-Hermitian part. ◊

Let us now assume that the property $D_1 + D_2 > 0$ which holds for all $(x, y) \in \Omega$ (see (2.4)) extends to the mesh in such a way that

$$(D_1)_{i+1/2,k} + (D_2)_{i,k+1/2} > 0 \qquad \forall (x, y) \in \Omega.$$

It is not difficult to check that the graph of $\frac{1}{2}(A + A^T)$ is connected and that $\frac{1}{2}(A + A^T)$ is weakly diagonal dominant. In some rows, the diagonal dominance inequality holds strictly. Therefore, we conclude that the symmetric part of A is irreducibly diagonally dominant. Furthermore, based on the Theorems 1.2 and 1.3, it is easy to see that $\frac{1}{2}(A + A^T)$ is symmetric positive definite.

2.4 Upwind finite differences approximation

As an alternative to the central difference approximation (2.9) for the convection terms, we can also use the so-called upwind finite differences. As we will see in a moment, in this case it is not necessary to rewrite the convective terms $v_1 u_x + v_2 u_y$ as done in (2.8). For the upwind finite differences we get the familiar five point stencil approximation (2.10) with different coefficients

$$\begin{aligned}
W_{i,k} &= -\frac{(D_1)_{i-1/2,k}}{h_1^2} - \frac{(v_1)_{i,k} + |v_1|_{i,k}}{2h_1}, \\
S_{i,k} &= -\frac{(D_2)_{i,k-1/2}}{h_2^2} - \frac{(v_2)_{i,k} + |v_2|_{i,k}}{2h_2}, \\
C_{i,k} &= \frac{(D_1)_{i-1/2,k} + (D_1)_{i+1/2,k}}{h_1^2} + \frac{(D_2)_{i,k-1/2} + (D_2)_{i,k+1/2}}{h_2^2} + D_{i,k} + \\
&\quad \frac{|v_1|_{i,k}}{h_1} + \frac{|v_2|_{i,k}}{h_2} \\
N_{i,k} &= -\frac{(D_2)_{i,k+1/2}}{h_2^2} + \frac{(v_2)_{i,k} - |v_2|_{i,k}}{2h_2}, \\
E_{i,k} &= -\frac{(D_1)_{i+1/2,k}}{h_1^2} + \frac{(v_1)_{i,k} - |v_1|_{i,k}}{2h_1}.
\end{aligned} \tag{2.13}$$

As we see, now the convective terms do contribute to $C_{i,k}$ on the main diagonal of A. Hence, the contributions of the convective terms do not form a skew-symmetric matrix anymore (cf. Exercise 2.2).

Exercise 2.3 Show that the upwind finite difference approximation (2.10), (2.13) results in a matrix A which is an M-matrix. ◊

2.5 Two other examples

In fact, many other problems and applications involve solving an IVP of the form (2.1). We briefly discuss here another two examples where (2.1) has to be solved. Both examples are taken from [40]. The first example is a typical problem from the control theory, namely, find the state vector function $y(t)$ such that

$$y'(t) = -Ay(t) + Bu(t), \qquad y(0) = y^0, \tag{2.14}$$

where $A \in \mathbb{R}^{n\times n}$ is the state companion matrix, $u(t) : \mathbb{R} \to \mathbb{R}^m$ is the control function and $B \in \mathbb{R}^{n\times m}$.

The second example is a large group of problems where the concept of continuous time Markov chains is employed. As noted in [40], this *"successful and widely used way of modeling the behavior of many physical systems consists in enumerating the (mutually exclusive) states in which the system may be at a given time and then, describing the interaction between these states."* Under certain assumptions the physical process under consideration can then be described by the Chapman-Kolmogorov IVP:

$$y'(t) = -Ay(t), \qquad y(0) = y^0.$$

Its solution $y(t) = e^{-tA}y^0$ is the so-called transient probability distribution of the Markov chain and the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is called infinitesimal generator of order n, with n being the number of states in the Markov chain. Because of certain probability assumptions A is a singular M-matrix with zero column sums, i.e.,

$$a_{ij} \leqslant 0 \text{ for } i \neq j \quad \text{and} \quad a_{jj} = -\sum_{i\neq j} a_{ij} \geqslant 0.$$

Exercise 2.4 Based on the results of Section 1, show that A in the last example is indeed a singular M matrix. ◊

3 Well-posedness of the problem. Stability estimates

The material presented in this section follows closely [23, Sect. 2.3].

3.1 Matrix exponential. Variation of constants formula

To analyze the IVP (2.1) and numerical methods for its solution, we need the concept of the matrix exponential, defined, for a given matrix $A \in \mathbb{R}^{n\times n}$, by the power series

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k, \qquad A^0 = I. \tag{3.1}$$

This definition is one of many possible definitions of the matrix exponential, see e.g. [18].

Theorem 3.1 *Solution of homogeneous, i.e. with $g(t) \equiv 0$, IVP* (2.1) *is given by*

$$y(t) = e^{-tA}y^0. \tag{3.2}$$

Proof Writing down the power series (3.1) for e^{-tA}, we note that the terms in the series are bounded in norm by $\frac{1}{k!}t^k\|A\|^k$. Hence, the series converges and

$$\|e^{-tA}\| \leqslant e^{t\|A\|}. \tag{3.3}$$

The rest of the proof is left as an exercise. □

Exercise 3.1 Finish the proof of Theorem 3.1. ◊

Theorem 3.2 *Solution of IVP* (2.1) *is given by*

$$y(t) = e^{-tA}y^0 + \int_0^t e^{(s-t)A}g(s)\,ds. \tag{3.4}$$

The last relation is called the variation of constants formula.

Proof Multiplying the ODE system $y'(t) + Ay(t) = g(t)$ by the matrix e^{tA}, we get

$$e^{tA}y'(t) + e^{tA}Ay(t) = e^{tA}g(t) \quad \Leftrightarrow \quad \frac{d}{dt}\left(e^{tA}y(t)\right) = e^{tA}g(t).$$

The relation (3.4) can now be obtained by integrating the last equality:

$$\begin{aligned} \int_0^t \frac{d}{ds}\left(e^{sA}y(s)\right)\,ds &= \int_0^t e^{sA}g(s)\,ds, \\ e^{tA}y(t) - \underbrace{e^{0A}y(0)}_{y^0} &= \int_0^t e^{sA}g(s)\,ds. \end{aligned}$$

□

3.2 Stability estimates

The variation of constants formula (3.4) allows to obtain the so-called stability estimates for IVP (2.1), whose meaning is as follows. Consider, together with (2.1), a perturbed problem

$$\tilde{y}'(t) = -A\tilde{y}(t) + g(t) + \delta(t), \qquad \tilde{y}(0) = \tilde{y}^0, \tag{3.5}$$

where $\delta(t)$ and $\tilde{y}^0$ are given. Let $\varepsilon(t) = \tilde{y}(t) - y(t)$. We are interested in establishing stability estimates, i.e., estimates which show dependence of $\|\varepsilon(t)\|$ on $\|\varepsilon(0)\|$ and $\|\delta(t)\|$. Since $\varepsilon(t)$ satisfies

$$\varepsilon'(t) = -A\varepsilon(t) + \delta(t), \qquad \varepsilon(0) = \tilde{y}^0 - y^0,$$

we obtain, using the variation of constants formula,

$$\varepsilon(t) = e^{-tA}\varepsilon(0) + \int_0^t e^{(s-t)A}\delta(s)\,ds,$$

$$\|\varepsilon(t)\| \leqslant \|e^{-tA}\|\,\|\varepsilon(0)\| + \int_0^t \|e^{(s-t)A}\|\,ds\,\max_{s\in[0,t]}\|\delta(s)\|,$$

where we used the fact that for a continuous vector function $f : [a,b] \to \mathbb{R}^n$ holds $\|\int_a^b f(x)dx\| \leqslant \int_a^b \|f(x)\|dx$. If we assume that there exist constants K and ω such that

$$\|e^{-tA}\| \leqslant Ke^{-t\omega}, \qquad t \geqslant 0, \tag{3.6}$$

then

$$\|\varepsilon(t)\| \leqslant Ke^{-t\omega}\|\varepsilon(0)\| + K\frac{1-e^{-t\omega}}{\omega}\max_{s\in[0,t]}\|\delta(s)\|. \tag{3.7}$$

Exercise 3.2 Check that (3.7) is correct. Note that the factor $(1 - e^{-t\omega})/\omega$ is undefined for $\omega = 0$. However, we can formally assign a certain value to $(e^{-t\omega} - 1)/\omega$ for $\omega = 0$. Which value should it be? ♢

For the stability estimate (3.7) to be useful, the exponential estimate (3.6) should be sufficiently sharp. Such estimates can be obtained in various ways.

Exercise 3.3 Note that (3.3) also fits the form of (3.6). However, this estimate is not very useful. Explain why. Hint: consider the scalar case $n = 1$ and matrices $A = 1$, $A = -1$. ♢

Assume that A is diagonalizable as $A = VDV^{-1}$ (where D is a diagonal matrix with the eigenvalues λ_k of A being its entries). Then

$$\|e^{-tA}\| = \|Ve^{-tD}V^{-1}\| \leqslant \|V\|\|e^{-tD}\|\|V^{-1}\| = \kappa(V)\max_k|e^{-t\lambda_k}| = \kappa(V)e^{-t\min_k \operatorname{Re}\lambda_k}.$$

Here $\kappa(V) = \|V\|\|V^{-1}\|$ is the condition number of the eigenvector matrix V. We see that (3.6) holds with $K = \kappa(V)$ and $\omega = \min_k \operatorname{Re}\lambda_k$. If A is normal then $\kappa(V) = 1$ in the 2-norm. However, if A is far from normal, so that a large $\kappa(V)$ makes the estimate above useless, or if the information on spectrum is unavailble then we need a different sort technique which we now consider.

3.3 Logarithmic matrix norm

To obtain more sensible exponential estimates of the form (3.6), we introduce the so-called *logarithmic norm* of a matrix $A \in \mathbb{R}^{n\times n}$, defined as [23, Sect. 2.3]

$$\mu(A) = \lim_{\tau\to 0+}\frac{\|I+\tau A\| - 1}{\tau}, \tag{3.8}$$

where $\|\cdot\|$ is a matrix norm induced by some vector norm.

Exercise 3.4 Check that for $\tau > 0$

$$-\|A\| \leqslant \frac{\|I + \tau A\| - 1}{\tau} \leqslant \|A\|.$$

$\Diamond$

The fraction under the limit in (3.8) can be shown to be non-decreasing in $\tau > 0$. Indeed, for $0 < \sigma < 1$ we have

$$\|I + \sigma\tau A\| = \|I + \sigma\tau A + \sigma I - \sigma I\| \leqslant \sigma\|I + \tau A\| + 1 - \sigma,$$
$$\frac{\|I + \sigma\tau A\| - 1}{\sigma\tau} \leqslant \frac{\sigma\|I + \tau A\| - \sigma}{\sigma\tau} \leqslant \frac{\|I + \tau A\| - 1}{\tau}.$$

Hence, the limit in (3.8) exists and the convergence is monotone.

Exercise 3.5 Is the logarithmic norm a norm? $\Diamond$

The definition (3.8) of the logarithmic matrix norm shows that this special norm can be interpreted as a one-sided derivative of the mapping $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}$, evaluated in point $I \in \mathbb{R}^{n\times n}$ in the direction given by $A \in \mathbb{R}^{n\times n}$ [10, Section 1.5]. The name "logarithmic" becomes clear if we note that for any $A \in \mathbb{R}^{n\times n}$

$$\mu(A) = \lim_{\tau\to 0+} \frac{\ln \|e^{\tau A}\|}{\tau}.$$

Indeed, for sufficiently small $\tau > 0$, we have [10, Section 1.5]

$$\begin{aligned}\ln \|e^{\tau A}\| &= \ln\left(\|I + \tau A\| + O(\tau^2)\right) = \ln\left(1 + \left[\|I + \tau A\| - 1 + O(\tau^2)\right]\right)\\ &= \|I + \tau A\| - 1 + O(\tau^2).\end{aligned}$$

The importance of the logarithmic norm becomes clear from the following result [23].

Theorem 3.3 *Let $A \in \mathbb{R}^{n\times n}$ and $\omega \in \mathbb{R}$. We have*

$$\mu(-A) \leqslant -\omega \qquad \Leftrightarrow \qquad \|e^{-tA}\| \leqslant e^{-t\omega}\ \forall t \geqslant 0. \tag{3.9}$$

Proof Note that the last relation can be rewritten in an equivalent form

$$\mu(A) \leqslant \omega \qquad \Leftrightarrow \qquad \|e^{tA}\| \leqslant e^{t\omega}\ \forall t \geqslant 0. \tag{3.10}$$

We give a proof for this last relation. First, assume that $\mu(A) \leqslant \omega$. Then, for sufficiently small $\tau > 0$, by definition of $\mu(A)$,

$$\begin{gathered}\frac{\|I + \tau A\| - 1}{\tau} - \mu(A) = O(\tau),\\ \|I + \tau A\| - 1 - \tau\mu(A) = O(\tau^2),\\ \|I + \tau A\| = 1 + \tau\mu(A) + O(\tau^2),\\ \|I + \tau A\| \leqslant 1 + \tau\omega + O(\tau^2),\\ \|(I + \tau A)^k\| \leqslant (1 + \tau\omega + O(\tau^2))^k,\end{gathered}$$

where $t = k\tau$ is fixed. Taking a limit $\tau \to 0+$ in the both parts of the last inequality, we obtain

$$\|e^{tA}\| \leqslant e^{t\omega}.$$

Here, for $\tau \to 0+$ and fixed $t = k\tau$, $(I + \tau A)^k \to e^{tA}$ because $I + \tau A$ is the transfer matrix of the explicit Euler method applied to $y'(t) = Ay(t)$ (see (4.1) with $-A$ replaced by A).

Assume now that $\|e^{tA}\| \leqslant e^{t\omega}$ for all $t > 0$. Then

$$\|I + \tau A\| = \|e^{tA} + O(\tau^2)\| \leqslant \|e^{tA}\| + O(\tau^2) \leqslant e^{t\omega} + O(\tau^2) = 1 + \tau\omega + O(\tau^2),$$

from which $\mu(A) \leqslant \omega$ easily follows. □

The following result lists some more important properties of the logarithmic matrix norm.

Theorem 3.4 *Let $A \in \mathbb{R}^{n\times n}$ and let $\mu(A)$ be defined by* (3.8)*. We have*

$$\mu(sI + A) = s + \mu(A), \qquad \forall s \in \mathbb{R}, \tag{3.11}$$

$$\mu(tA) = |t|\mu(\text{sign}(t)A), \qquad \forall t \in \mathbb{R}, \tag{3.12}$$

$$\mu(A + B) \leqslant \mu(A) + \mu(B), \tag{3.13}$$

$$|\mu(A) - \mu(B)| \leqslant \|A - B\|, \tag{3.14}$$

$$\mu(A) \geqslant -\frac{\|Ax\|}{\|x\|}, \qquad \forall 0 \neq x \in \mathbb{C}^n, \tag{3.15}$$

where $\|\cdot\|$ is the norm which defines the logarithmic norm $\mu(\cdot)$ and sign *is the sign function.*

Proof To prove (3.11), we write, taking into account that $1 + \tau s \geqslant 0$ for small $\tau > 0$,

$$\begin{aligned}\mu(sI + A) &= \lim_{\tau\to 0+} \frac{\|I + \tau(sI + A)\| - 1}{\tau} = \lim_{\tau\to 0+} \frac{(1 + \tau s)\|I + \frac{\tau}{1+\tau s}A\| - 1}{\tau} \\ &= \lim_{\tau\to 0+} \frac{\|I + \frac{\tau}{1+\tau s}A\| - \frac{1}{1+\tau s}}{\frac{\tau}{1+\tau s}} = \lim_{\tau\to 0+} \frac{\|I + \frac{\tau}{1+\tau s}A\| - \frac{1+\tau s-\tau s}{1+\tau s}}{\frac{\tau}{1+\tau s}} = \mu(A) + s.\end{aligned}$$

Proof of (3.12) is left as an exercise.

Next, the property (3.13) can be proven, using the estimate [26]

$$\begin{aligned}\|I + \tau(A + B)\| - 1 &= \|\frac{1}{2}(I + 2\tau A) + \frac{1}{2}(I + 2\tau A)\| - 1 \\ &\leqslant \frac{1}{2}(\|I + 2\tau A\| - 1) + \frac{1}{2}(\|I + 2\tau B\| - 1).\end{aligned}$$

Furthermore, (3.14) can be shown based on the fact that $-\|A\| \leqslant \mu(A) \leqslant \|A\|$ (see Exercise 3.4).

Finally, to see that (3.15) holds, we write

$$\begin{aligned} \|x\| &= \|(I+\tau A)x - \tau Ax\| \leqslant \|(I+\tau A)x\| + \tau\|Ax\|, \\ -\|Ax\| &\leqslant \frac{\|(I+\tau A)x\| - \|x\|}{\tau} \leqslant \frac{\|I+\tau A\|-1}{\tau}\|x\|. \end{aligned}$$

□

Exercise 3.6 Finish all the details in the proof of Theorem 3.4. ♢

Since the logarithmic norm is introduced for any matrix norm, in practice we may want to use a norm best suitable for a particular situation.

Exercise 3.7 Check that for the most often used vector norms (1.1) and associated matrix norms (1.2), the corresponding logarithmic norms can be computed as

$$\begin{aligned} \mu_2(A) &= \max_{x\neq 0}\frac{\operatorname{Re}(Ax,x)}{(x,x)} = \max\{\lambda \mid \lambda \in \text{spectrum of } \frac{1}{2}(A+A^*)\}, \\ \mu_1(A) &= \max_j(\operatorname{Re} a_{jj} + \sum_{i\neq j}|a_{ij}|), \\ \mu_\infty(A) &= \max_i(\operatorname{Re} a_{ii} + \sum_{j\neq i}|a_{ij}|). \end{aligned} \tag{3.16}$$

♢

3.4 Examples

To see how the results of this section can be used, let us consider several examples taken from [23, Sect. 2.3]. Assume we solve (2.1) and it is known that the matrix $\frac{1}{2}(A+A^T)$ (the symmetric part of A) is positive semidefinite. As we see from (3.9) and (3.16), $\|e^{-tA}\|_2 \leqslant 1$ if and only if $\frac{1}{2}(A+A^T)$ is positive semidefinite. Thus, the stability estimate (3.7) holds in the 2-norm. If $\frac{1}{2}(A+A^T)$ is positive definite then we see that relation (3.6) holds with ω being the smallest eigenvalue of $\frac{1}{2}(A+A^T)$.

Furthermore, if it is known that the diagonal elements of A are positive and A is row-wise (weakly) diagonally dominant then

$$\mu_\infty(-A) = \max_i(-a_{ii} + \sum_{j\neq i}|a_{ij}|) = -\underbrace{\min_i(a_{ii} - \sum_{j\neq i}|a_{ij}|)}_{\text{denote by } \delta} \leqslant 0$$

and the stability estimate (3.7) can be established in the max-norm with $\omega = \delta$. Similarly, stability in the 1-norm can be obtained if A is column-wise (weakly) diagonally dominant with positive diagonal elements.

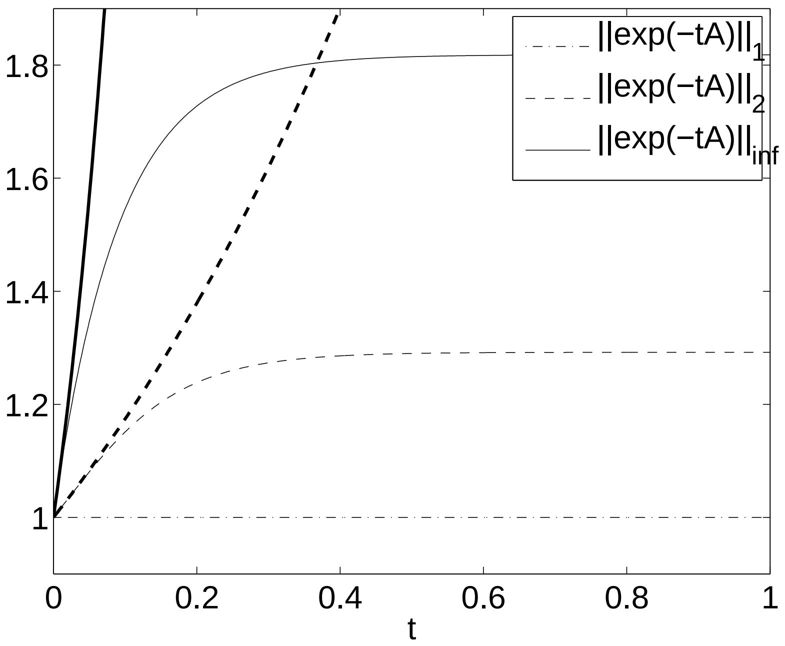

Figure 1: The time dependence of the norms $\|e^{-tA}\|_1$ (dash-dotted line), $\|e^{-tA}\|_2$ (dashed line) and $\|e^{-tA}\|_\infty$ (solid line). The upper bounds for the two last norms are plot in bold.

We now consider a more specific example, taken from [23, Sect. 2.3]. An IVP

$$y'(t) = -Ay(t), \qquad A = \begin{bmatrix} k_1 & -k_2 \\ -k_1 & k_2 \end{bmatrix}, \qquad y(0) = \begin{bmatrix} y_1^0 \\ y_2^0 \end{bmatrix}, \tag{3.17}$$

models the two-way chemical reaction $y_1 \xrightarrow{k_1} y_2 \xrightarrow{k_2} y_1$.

Exercise 3.8 Check, by diagonalization of A, that the exact solution of (3.17) is

$$\begin{aligned} y_1(t) &= ak_2 + be^{-(k_1+k_2)t}, \\ y_2(t) &= ak_1 - be^{-(k_1+k_2)t}, \qquad a = \frac{y_1^0 + y_2^0}{k_1 + k_2}, \quad b = \frac{k_1 y_1^0 - k_2 y_2^0}{k_1 + k_2}, \end{aligned}$$

where y_1^0 and y_2^0 are the given initial values. ◊

Let us consider solution of (3.17) for the $0 \leqslant t \leqslant T = 1$, $k_1 = 1$, $y_1^0 = 0.1$ and $y_2^0 = 0.9$. If $k_2 \gg k_1 = 1$ we have $\|A\| \gg 1$ and the stability estimate (3.7) with (3.3),(3.6) (i.e., $\omega = -\|A\|$) suggests an instability (an ill-posedness) of the problem. On the other hand, the logarithmic norms are

$$\mu_1(-A) = 0, \qquad \mu_2(-A) = -\hat{\omega} = -\frac{k_1 + k_2}{2} + \sqrt{\frac{k_1^2 + k_2^2}{2}} > 0, \qquad \mu_\infty(-A) = |k_2 - k_1|.$$

This, due to (3.9), implies

$$\begin{aligned} &\|e^{-tA}\|_1 \leqslant 1 \qquad \forall t \geqslant 0, \\ &\|e^{-tA}\|_2 \leqslant e^{-t\hat{\omega}} \qquad \forall t \geqslant 0, \\ &\|e^{-tA}\|_\infty \leqslant e^{t|k_2 - k_1|} \qquad \forall t \geqslant 0. \end{aligned}$$

As we see from Figure 1, the last two estimates are far from being sharp for large $t > 0$. To get sharper estimates, we note that for any $n \times n$ matrix B holds (see (1.3))

$$\|B\|_2 \leqslant \sqrt{n}\|B\|_1, \qquad \|B\|_\infty \leqslant n\|B\|_1.$$

This yields the estimates

$$\begin{aligned} \|e^{-tA}\|_2 &\leqslant \sqrt{2} \qquad \forall t \geqslant 0, \\ \|e^{-tA}\|_\infty &\leqslant 2 \qquad \forall t \geqslant 0, \end{aligned}$$

which show good stability properties of the problem.

4 Basic time integration schemes. Their stability

Let $\tau > 0$ be a time step size and denote by y_k a numerical solution of (2.1) at $t = k\tau$, $y^k \approx y(k\tau)$. Some standard numerical schemes for integrating (2.1) in time read

$$\frac{y^{k+1} - y^k}{\tau} = -Ay^k + g^k, \tag{4.1}$$

$$\frac{y^{k+1} - y^k}{\tau} = -Ay^{k+1} + g^{k+1}, \tag{4.2}$$

$$\frac{y^{k+1} - y^k}{\tau} = -(1-\theta)Ay^k - \theta Ay^{k+1} + (1-\theta)g^k + \theta g^{k+1}, \quad \theta \in [0,1], \tag{4.3}$$

called respectively forward (or explicit) Euler scheme, backward (or implicit) Euler scheme, and the implicit–explicit θ-method. For $\theta = \frac{1}{2}$ we get the implicit trapezoidal rule scheme, whereas the choices $\theta = 0$ and $\theta = 1$ yield forward Euler and backward Euler schemes, respectively. Note that the implicit trapezoidal rule scheme is also known as the Crank–Nicolson scheme, proposed in 1947 by Crank and Nicolson [9].

Many time integration schemes for (2.1) with $g \equiv 0$ can be written as

$$y^{k+1} = R(-\tau A)y^k,$$

where $R(z)$ is some rational function. For a particular method, this function can be easily derived by applying the method to a scalar test equation (the so-called Dahlquist test problem [23, Sect. I.2])

$$y'(t) = \lambda y(t), \ y(0) = y^0, \qquad \lambda \in \mathbb{C}. \tag{4.4}$$

$R(z)$ is called the stability function of the method.

Exercise 4.1 Check that the stability function of the θ-method reads

$$R(z) = \frac{1 + (1-\theta)z}{1 - \theta z}.$$

◊

The name "stability function" becomes clear if we consider the set

$$\mathcal{S} = \{z \in \mathbb{C} \mid |R(z)| \leqslant 1\}.$$

Since, to have stability for the problem (4.4), it is sufficient to require $\tau\lambda = z \in \mathcal{S}$, we call $\mathcal{S}$ the stability region of the method.

The following result, used later, is called the maximum modulus theorem.

Theorem 4.1 *[23, Sect. I.2] Let φ be a nonconstant complex function which is analytic on a set $\mathcal{D} \subset \mathbb{C}$ and continuous on its closure. Then the maximum of φ is attained on the boundary $\partial\mathcal{D}$ of $\mathcal{D}$. In particular, if φ is rational without poles in $\mathbb{C}^- = \{z \in \mathbb{C} \mid \operatorname{Re} z \leqslant 0\}$ then $\max_{z\in\mathbb{C}^-} |\varphi(z)| = \max_{y\in\mathbb{R}} |\varphi(iy)|$.*

A time integration method is called A-stable if its stability region contains the left complex half-plane $\mathbb{C}^- = \{z \in \mathbb{C} \mid \operatorname{Re} z \leqslant 0\}$. A-stability means that the method applied to (4.4) with $\lambda \in \mathbb{C}^-$ is unconditionally stable (i.e., stable for all $\tau > 0$).

The question is whether this stability considerations for the scalar test problem (4.4) can be extended to the problem (2.1) for a nonnormal matrix[2] The following theorem answers this question.

Theorem 4.2 *[17, Sect. IV.11] (Theorem of John von Neumann) Let a rational function $R(z)$ be bounded in $\mathbb{C}^- = \{z \in \mathbb{C} \mid \operatorname{Re} z \leqslant 0\}$ and let $A \in \mathbb{C}^{n\times n}$ be such that*

$$\operatorname{Re}(y, Ay) \geqslant 0, \qquad \forall y \in \mathbb{C}^n.$$

Then in the matrix norm corresponding to the scalar product we have

$$\|R(-\tau A)\| \leqslant \max_{z\in\mathbb{C}^-} |R(z)|. \tag{4.5}$$

Proof [17, Sect. IV.11] To simplify the notation, the proof is given for $-\tau A$ replaced by A. Then for this new A we have to prove that $\|R(A)\| \leqslant \max_{z\in\mathbb{C}^-} |R(z)|$. It holds

$$\operatorname{Re}(y, Ay) \leqslant 0, \qquad \forall y \in \mathbb{C}^n.$$

Assume that A is nonnormal (otherwise the proof is left as Exercise 4.2). Introduce, for $\alpha \in \mathbb{C}$,

$$A(\alpha) = \frac{\alpha}{2}(A + A^*) + \frac{1}{2}(A - A^*).$$

Note that $A(1) = A$. It is not difficult to see that

$$(v, A(\alpha)v) = \bar{\alpha}\operatorname{Re}(v, Av) + i\operatorname{Im}(v, Av). \tag{4.6}$$

This shows that

$$\operatorname{Re}(y, A(\alpha)y) \leqslant 0, \qquad \forall y \in \mathbb{C}^n,$$

[2] If A is close to normal or if some information on the (pseudo)spectrum of A is available, then estimates for $\|R(-\tau A)\|$ as in the end of Section 3.2 can be used.

holds as long as $\operatorname{Re}\alpha \geqslant 0$. Hence, for $\operatorname{Re}\alpha \geqslant 0$ the eigenvalues of $A(\alpha)$ also have nonpositive real part. Therefore, the rational function

$$\varphi(\alpha) = \|R(A(\alpha))v\|^2,$$

where v is fixed, has no poles in $\operatorname{Re}\alpha \geqslant 0$. From Theorem 4.1 it follows then

$$\begin{aligned}\|R(A)v\|^2 = \varphi(1) &\leqslant \max_{y\in\mathbb{R}} \varphi(iy) = \max_{y\in\mathbb{R}} \|R(A(iy))v\|^2 \\ &\leqslant \max_{y\in\mathbb{R}} \|R(A(iy))\|^2 \|v\|^2.\end{aligned}$$

It can be checked that the matrix $A(iy)$ is normal. Since the proof holds for normal matrices (see Exercise 4.2), we have

$$\|R(A(iy))\| \leqslant \max_{z\in\mathbb{C}^-} |R(z)| \qquad \forall y \in \mathbb{R},$$

which leads to

$$\|R(A)v\|^2 \leqslant \left(\max_{z\in\mathbb{C}^-} |R(z)|\right)^2 \|v\|^2.$$

□

Exercise 4.2 Prove Theorem 4.2 for a normal matrix A. ♢

Exercise 4.3 Check the relation (4.6) and normality of the matrix $A(iy)$. ♢

To appreciate the strength of Theorem 4.2 and the elegance of its proof, let us now try to prove a similar result for a specific time integration scheme, namely, for the θ-method. It is not difficult to see that the θ-method applied to the problem (2.1) with $g \equiv 0$ reads

$$y^{k+1} = R(-\tau A)y^k, \qquad R(Z) = (I - \theta Z)^{-1}(I + (1-\theta)Z),$$

where the matrix $Z = -\tau A$ is introduced.

Exercise 4.4 Check that for any square matrix Z we have

$$(I - \theta Z)^{-1}(I + (1-\theta)Z) = (I + (1-\theta)Z)(I - \theta Z)^{-1}.$$

♢

For any vector norm corresponding to the chosen inner product and the induced operator norm induced by this vector norm, we then can write

$$\frac{\|y^{k+1}\|^2}{\|y^k\|^2} = \frac{\|R(Z)y^k\|^2}{\|y^k\|^2} = \frac{\|(I + (1-\theta)Z)(I - \theta Z)^{-1}y^k\|^2}{\|y^k\|^2} = \frac{\|(I + (1-\theta)Z)u\|^2}{\|(I - \theta Z)u\|^2},$$

where $u = (I - \theta Z)^{-1}y^k$. The last relation can be rewritten as

$$\frac{\|y^{k+1}\|^2}{\|y^k\|^2} = \frac{1 + 2(1-\theta)\operatorname{Re}(v, Zv) + (1-\theta)^2\|Zv\|^2}{1 - 2\theta\operatorname{Re}(v, Zv) + \theta^2\|Zv\|^2} = |R(\zeta)|^2, \tag{4.7}$$

with $v = u/\|u\|$ and $\zeta = \operatorname{Re}(v, Zv) + i\sqrt{\|Zv\|^2 - (\operatorname{Re}(v, Zv))^2}$.

Exercise 4.5 Show that (4.7) with the given ζ is correct. ◊

We thus obtain

$$\|R(-\tau A)\| = |R(\zeta)|,$$

where ζ is just defined. We would now like to get a bound on $\|R(-\tau A)\|$ by localizing ζ. It is natural to assume (cf. (3.9)) that

$$\mu_2(-A) \leqslant -\omega, \tag{4.8}$$

where $\mu_2(\cdot)$ is defined with respect to the same inner product operator norm.

Exercise 4.6 Propose a condition in terms of ω, θ and τ which guarantees that all the eigenvalues of the matrix $I + \theta\tau A$ have a positive real part. Under this condition the matrix $I + \theta\tau A$ is nonsingular. ◊

Using (4.8), we have in the estimate above

$$\operatorname{Re}\zeta = \operatorname{Re}(v, Zv) = \tau \operatorname{Re}(v, -Av) \leqslant \tau\mu_2(-A) \leqslant -\tau\omega,$$

so that

$$\|R(-\tau A)\| \leqslant \max_{\operatorname{Re}\zeta\leqslant-\tau\omega} |R(\zeta)| = \max\bigl\{|R(-\tau\omega)|, \underbrace{\lim_{z\to\infty} |R(z)|}_{1-1/\theta}\bigr\}.$$

The last equality is due to the maximum modulus theorem. We proved the following result.

Theorem 4.3 *[23, Sect. I.2] Let $\|\cdot\|$ denote a vector or induced matrix norm corresponding to a inner product and let $A \in \mathbb{C}^{n\times n}$ be such that $\mu_2(-A) \leqslant -\omega$ with ω satisfying the condition derived in Exercise 4.6. Furthermore let*

$$R(z) = \frac{1 + (1-\theta)z}{1-\theta z}$$

be the stability function of the θ-method (4.3)*. We have*

$$\|R(-\tau A)\| \leqslant \max_{\operatorname{Re} z\leqslant-\tau\omega} |R(z)| = \max\bigl\{|R(-\tau\omega)|, 1 - \frac{1}{\theta}\bigr\}.$$

Exercise 4.7 Show that the bound (4.8) on the logarithmic norm holds if and only if

$$\operatorname{Re}(v, Av) \geqslant \omega\|v\|^2, \qquad \forall v \in \mathbb{C}^n.$$

◊

Note that the stability results for the θ-method in this section are proven for an inner product vector norm (and the induced operator matrix norm). Obtaining stability results for the θ-method in other norms is often difficult unless $\theta = 1$ [23, Sect. I.2]. For the implicit Euler method ($\theta = 1$) stability results are not only easier to obtain but also some stability results hold exclusively for $\theta = 1$. For instance, the requirement for the θ-method

$$\|R(-\tau A)\|_* \leqslant 1, \qquad \text{with } * = 1 \text{ or } * = \infty,$$

necessarily implies $\theta = 1$ [17, Sect. IV.11].

5 Operator splitting

5.1 Introducing splitting methods

The material of this subsection closely follows the lines of [23, Sect. IV.1]. One of the very useful concepts in numerical time integration methods is the so-called operator splitting. To understand this concept, assume we solve (2.1) with no source term ($g \equiv 0$) and[3]

$$A = A_1 + A_2.$$

If ODE systems $y' = -A_1 y(t)$ and $y' = -A_1 y(t)$ are easier to solve than $y' = -Ay(t)$, then numerical solution of (2.1) after one time step, namely,

$$y^1 \approx y(\tau) = e^{-\tau A} y^0, \tag{5.1}$$

can be approximated by first solving the ODE system with A_1 and then with A_2. More precisely, we successfully solve two IVPs

$$\begin{aligned} \tilde{y}' &= -A_1 \tilde{y}(t), \quad \text{for } t \in [0,\tau] \quad \text{with } \tilde{y}(0) = y^0, \\ \hat{y}' &= -A_2 \hat{y}(t), \quad \text{for } t \in [0,\tau] \quad \text{with } \hat{y}(0) = \tilde{y}(\tau), \end{aligned} \tag{5.2}$$

where the output $\tilde{y}(\tau)$ of the first subproblem is the input $\hat{y}(0)$ of the second subproblem. The splitting procedure in (5.2) is repeated at all subsequent time steps $k = 2, 3, \dots$. It forms a simplest splitting method called sequential splitting.

Assume now that y^1 in (5.1) and subproblem solutions (5.2) are computed exactly, which means

$$y^1 = e^{-\tau A} y^0, \qquad y^1_{\text{split}} = e^{-\tau A_2} e^{-\tau A_1} y^0.$$

Comparing the exact solution y^1 with the splitting solution y^1_{split} we see

$$\begin{aligned} e^{-\tau A} &= I + \tau(-A_1 - A_2) + \frac{\tau^2}{2}(A_1 + A_2)^2 + \dots, \\ e^{-\tau A_2} e^{-\tau A_1} &= I + \tau(-A_1 - A_2) + \frac{\tau^2}{2}(A_1^2 + 2A_2 A_1 + A_2^2) + \dots. \end{aligned}$$

Hence, the splitting introduces an additional error at every time step. The error committed at one time step started at the exact solution is called local error.

[3] Here A_1 should not be confused with the skew-symmetric part of A.

As we see, in the sequential splitting method the local error is

$$\left(e^{-\tau A} - e^{-\tau A_2}e^{-\tau A_1}\right) y^0 = \frac{\tau^2}{2}(A_1A_2 - A_2A_1) + O(\tau^3).$$

Since the local error is $O(\tau^2)$, the global error, i.e., the error after all time steps are done, is $O(\tau)$. Thus, we see that the sequential splitting (5.2) is first order accurate. We call

$$[A_1, A_2] = A_1A_2 - A_2A_1$$

the commutator of A_1 and A_2.

Exercise 5.1 Prove that if A_1 and A_2 are diagonalizable and commute then

$$e^{-\tau A_2}e^{-\tau A_1} = e^{-\tau A_2 - \tau A_1} = e^{-\tau A}. \tag{5.3}$$

Thus, the sequential splitting is exact in this case. If A_1 and A_2 commute but are not necessarily diagonalizable then (5.3) still holds, which can be seen from the power expansion of the matrix exponential. ◊

For two noncommuting matrices it can sometimes be very useful to express a product of their matrix exponentials as a single matrix exponential, i.e., to find, for given $A_{1,2}$, such a matrix $\tilde{A}$ that

$$e^{\tau A_2}e^{\tau A_1} = e^{\tau \tilde{A}}.$$

The matrix $\tilde{A}$ is then given by the Baker-Campbell-Hausdorff formula:

$$\begin{aligned}\tilde{A} =&(A_1 + A_2) + \frac{\tau}{2}[A_2, A_1] + \frac{\tau^2}{12}\left([A_2, [A_2, A_1]] + [A_1, [A_1, A_2]]\right)\\ &+ \frac{\tau^3}{24}[A_2, [A_1, [A_1, A_2]]] + O(\tau^4).\end{aligned}$$

The higher order terms in this formula are quite cumbersome but can be computed recursively [39].

5.2 Second order splittings

The accuracy in the sequential splitting (5.2) can be improved if we repeat the splitting steps in the opposite order:
(1) a step for the subproblem with A_1;
(2) a step for the subproblem with A_2;
(3) a step for the subproblem with A_2;
(4) a step for the subproblem with A_1.
More precisely, assuming again that the splitting subproblems can be solved exactly, we have at the first time step $k = 1$

$$y^1_{\text{split}} = e^{-\frac{\tau}{2}A_1}e^{-\frac{\tau}{2}A_2}e^{-\frac{\tau}{2}A_2}e^{-\frac{\tau}{2}A_1}y^0 = e^{-\frac{\tau}{2}A_1}e^{-\tau A_2}e^{-\frac{\tau}{2}A_1}y^0. \tag{5.4}$$

It can be shown, after quite a few manipulations, that

$$\left(e^{-\tau A}-e^{-\frac{\tau}{2}A_1}e^{-\tau A_2}e^{-\frac{\tau}{2}A_1}\right)y^0=\frac{\tau^3}{24}\left([A_1,[A_1,A_2]]+2[A_2,[A_1,A_2]]\right)y(\tau/2)+O(\tau^5),$$

which shows that this symmetric splitting is second order accurate. This splitting method was proposed in 1968 independently by Marchuk [27] and by Strang [43]. We call (5.4) Marchuk–Strang splitting.

Another second order splitting, proposed in 1963 by Strang [42], reads at the first time step $k=1$

$$y^1_{\text{split}}=\frac{1}{2}\left(e^{-\tau A_1}e^{-\tau A_2}+e^{-\tau A_2}e^{-\tau A_1}\right)y^0. \tag{5.5}$$

Here, for simplicity of the presentation we again assumed that the time steps for subproblems are carried out exactly. Note that the substeps $e^{-\tau A_1}e^{-\tau A_2}y^0$ and $e^{-\tau A_2}e^{-\tau A_1}y^0$ can be computed in parallel. For this reason splitting (5.5) is called parallel or symmetrically weighted splitting.

5.3 Examples of splitting

In the previous section we used the matrix exponentials e^{-tA_j}, $j=1,2$, only to describe different splitting methods in a compact form. In practice, each of the splitting substeps can be carried out by any suitable time integration method. Of course, the splitting can be applied (and, indeed, is widely applied) to any system of ODEs

$$y'(t)=f(t,y(t)),\qquad f:\mathbb{R}\times\mathbb{R}^n\to\mathbb{R}^n,$$

as soon as we have a splitting $f(t,y)=f_1(t,y)+f_2(t,y)$. Simplicity of splitting methods makes them very popular. In complex mathematical models, they allow to treat different processes independently, in separate modules of the software.

To emphasize the versatility of the operator splitting concept, we now name just a few possible splitting methods applicable to the time dependent advection–diffusion problem (2.3).

1. The θ-method. See Exercise 5.2.
2. Directional splitting methods where A_1 contain all the contributions of the partial derivatives with respect to x (and $A_2=A-A_1$). Famous splitting methods of this type are ADI (alternating direction implicit) (see e.g. [32]) and LOD (locally one-dimensional) schemes [50].
3. Splitting methods based on physical process, where, for instance, A_1 corresponds to diffusion and A_2 to advection.
4. Special splitting schemes where for the splitting steps carried out by an implicit scheme A_j are chosen such that the linear system with $I+\tau A_j$ is easy to solve.

Exercise 5.2 Show that the θ-method can seen as a sequential splitting method where the part with $A_1 = (1-\theta)A$ is carried out by the explicit Euler scheme and the part with $A_2 = \theta A$ by the implicit Euler scheme. $\diamondsuit$

5.4 Splitting with M-matrices

Let $A \in \mathbb{R}^{n\times n}$ be weakly diagonally dominant with nonnegative diagonal and nonpositive off-diagonal entries. As Theorem 1.3 states, A is then an (possibly singular) M-matrix. In some situations considered below, a splitting

$$A = M + N \tag{5.6}$$

can be useful, where M and N have the same properties as A, i.e., they are weakly diagonally dominant matrices with nonnegative diagonal entries and nonpositive off-diagonal entries. Such splittings are sometimes called replicative [3]. Consider the following implicit–explicit (IMEX) time integration scheme for solving (2.1):

$$\frac{y^{k+1}-y^k}{\tau} = -My^{k+1} - Ny^k + g^{k+1/2}. \tag{5.7}$$

The scheme can be rewritten as

$$y^{k+1} = (I+\tau M)^{-1}(I-\tau N)y^k + \tau(I+\tau M)^{-1}g^{k+1/2}. \tag{5.8}$$

We now give a stability result for this scheme.

Theorem 5.1 *[3] Let matrices M and N be weakly diagonally dominant with nonnegative diagonal entries and nonpositive off-diagonal entries. If $N \neq 0$ then the time integration scheme* (5.8),(5.6) *with*

$$\tau \leqslant \frac{1}{\max_i n_{ii}} \tag{5.9}$$

is stable, i.e.,

$$\|(I+\tau M)^{-1}(I-\tau N)\|_\infty \leqslant \|(I+\tau M)^{-1}\|\|(I-\tau N)\|_\infty \leqslant 1, \tag{5.10}$$

$$\|y^{k+1}\|_\infty \leqslant \|y_k\|_\infty + \tau\|g^{k+1/2}\|_\infty, \tag{5.11}$$

and monotone, i.e.,

$$g^{k+1/2} \geqslant 0,\ l \leqslant m \quad \Rightarrow \quad y^k \geqslant 0,$$

where the inequalities understood elementwise. Furthermore if $A = M + N$ is nonsingular then

$$\rho((I+\tau M)^{-1}(I-\tau N)) < 1.$$

If $N = 0$ then all the stability and monotonicity estimates of this theorem hold for any $\tau > 0$.

Proof For brevity, in the proof we omit the subscript $\cdot_\infty$ for the norms. By Theorem 1.3, M is a possibly singular M-matrix and we can write $M = sI - B$ where $s = \max_i m_{ii} \geqslant \|B\| \geqslant \rho(B)$ and B is elementwise nonnegative (see the proof of Theorem 1.3). Then

$$\begin{aligned}
\|(I+\tau M)y\| &= \|(I+\tau sI - \tau B)y\| \geqslant |\|(1+\tau s)y\| - \|\tau By\|| \geqslant \\
&\geqslant \|(1+\tau s)y\| - \|\tau By\| = (1+\tau s)\|y\| - \tau\|By\| \geqslant \\
&\geqslant \|y\| + \tau s\|y\| - \tau\|B\|\|y\| = \|y\| + \tau(s - \|B\|)\|y\|, \\
\|(I+\tau M)^{-1}\| &= \max_{x\neq 0} \frac{\|(I+\tau M)^{-1}x\|}{\|x\|} = \max_{y\neq 0}\frac{\|y\|}{\|(I+\tau M)y\|} \leqslant \\
&\leqslant \max_{y\neq 0}\frac{\|y\|}{\|y\| + \tau(s-\|B\|)\|y\|} \leqslant 1.
\end{aligned}$$

It is not difficult to check that $\|I - \tau N\| \leqslant 1$ provided (5.9) holds. Thus (5.10) holds and (5.11) follows. The monotonicity estimate results from the fact that $(I+\tau M)^{-1}$ and $I - \tau N$ are elementwise nonnegative. Finally, the bound on the spectral radius follows from the observation that $P - Q$, with $P = I + \tau M$ and $Q = I - \tau N$, is a regular splitting of the M-matrix τA (see Theorem 1.4). □

Exercise 5.3 (a) Check that bound (5.9) implies $\|(I - \tau N)\|_\infty \leqslant 1$.

(b) Check if $g(t) \equiv 0$ then from (5.8), (5.6), (5.9) follows

$$\|y^{k+1}\|_\infty \leqslant \frac{1 - \tau \min_i \sum_j n_{ij}}{1 + \tau(s - \|B\|_\infty)}\|y^k\|_\infty.$$

◊

5.5 Reducing the splitting error: Rosenbrock methods

In some applications the splitting errors can be rather harmful. This is often the case if the eigenvalues of the matrices A_1 and A_2 significantly differ, as for example in the stiff problems, where A_1 and A_2 may have eigenvalues of a different order of magnitude [41, 48]. In this case a nice alternative to splitting methods are the so-called Rosenbrock methods [49, 2]. Applied to a nonlinear autonomous IVP

$$y'(t) = f(y), \qquad y(0) = y^0, \tag{5.12}$$

with $f : \mathbb{R}^n \to \mathbb{R}^n$ and $y^0 \in \mathbb{R}^n$ given, a two stage Rosenbrock scheme called ROS2 reads

$$\begin{aligned}
y^{k+1} &= y^k + \frac{3}{2}k_1 + \frac{1}{2}k_2, \\
(I + \gamma\tau\hat{A})k_1 &= \tau f(y^k), \\
(I + \gamma\tau\hat{A})k_2 &= \tau f(y^k + k_1) - 2k_1,
\end{aligned} \tag{5.13}$$

where $-\hat{A}$ is supposed to be an approximation to the Jacobian $f'(y^k)$, $\tau > 0$ is the time step size and $\gamma > 0$ is a parameter defined below. The Rosenbrock schemes in general and the scheme (5.13) in particular have a remarkable property: their consistency order does not depend on how well $-\hat{A}$ approximates the Jacobian $f'(y^k)$. By applying the ROS2 to the Dahlquist test problem, it is not difficult to derive the ROS2 stability function. Then, for $-\hat{A} = f'(y^k)$, it can be shown that the ROS2 scheme is A-stable for $\gamma \geqslant \frac{1}{4}$. Furthermore, one can show that the ROS2 scheme has a second order consistency for any $\hat{A}$ [49], [23, Sect. IV.5.2]. For $\hat{A}$ approximating the Jacobian as $-\hat{A} = f'(y^k) + O(\tau)$ this two stage scheme can be modified such that it has a third order consistency for a specific value of γ [23, Sect. IV.5.2], [25]. We note that Rosenbrock methods allowing arbitrary approximations to the Jacobian matrices are also called W-methods, see e.g. [17, Sect. IV.7].

As already noted, an important attractive property of Rosenbrock schemes such as ROS2 is that they can be used as an alternative to the splitting methods. Indeed, since the matrix $\hat{A}$ can be chosen arbitrarily, we can take $\hat{A}$ such that

$$I + \gamma\tau\hat{A} = (I + \gamma\tau A_1)(I + \gamma\tau A_2), \qquad \text{with} \quad A_1 + A_2 = -f'(y^k). \tag{5.14}$$

For small τ this implies an approximation

$$I + \gamma\tau\hat{A} = I + \gamma\tau(A_1 + A_2) + (\gamma\tau)^2 A_1 A_2 = I - \gamma\tau f'(y^k) + O(\tau^2), \tag{5.15}$$

and, hence, we refer to (5.14) as *approximate matrix factorization* (AMF). The concept of AMF can be traced back to papers [14, 1] and to classic work on alternating direction implicit (ADI) methods [32]. A combination of the ROS2 scheme with AMF (called ROS2-AMF) can then be employed instead of a splitting method with A_1 and A_2 [49, 45]. Here, in fact, the splitting is put from time integration to the linear algebra level. This approach has been successfully applied in time integration of advection–diffusion–reaction problems [49, 16, 48, 2].

If A_1 and A_2 in (5.14) do not commute, a strict stability analysis of Rosenbrock methods with AMF is a difficult task, with many open problems (see e.g. [49, 30, 7]). For instance, consider an advection–diffusion–reaction problem (5.12) with

$$f(y) = -A_{\text{adv}}y - A_{\text{diff}}y - R(y), \tag{5.16}$$

where A_{adv} and A_{diff} are respectively discretized advection and diffusion operators (cf. (2.6)) and R represents the reaction operator. A usual assumption is that the chemical species under consideration only react among each other if they belong to the same mesh position (the same mesh cell). Under this assumption the Jacobian $R'(y)$ is a block diagonal matrix, where the number of blocks equals the number of the mesh points. When applying (5.13) with AMF (5.14) to solve (5.12),(5.16), it is reasonable not to include the advection terms to the matrix $\hat{A}$ because these terms are usually not large (one can call them "non-stiff"). Hence, they do not lead to a severe restriction on the step size of explicit time integration schemes. Not including some terms into $\hat{A}$ effectively means that these terms are treated explicitly by the ROS2 scheme. Explicit treatment of advection is also advocated because it often yields a better accuracy than implicit treatment.

In opposite, diffusion and reaction terms are usually stiff and should be treated implicitly. Hence, we adjust the AMF (5.14) as follows:

$$\begin{gathered} I+\gamma\tau\hat{A} = (I+\gamma\tau A_1)(I+\gamma\tau A_2), \qquad \text{with} \\ A_1 = A_{\text{diff}},\ A_2 = R'(y^k),\ A_1 + A_2 = -f'(y^k) - A_{\text{adv}}. \end{gathered} \tag{5.17}$$

To study stability of the ROS2-AMF scheme (5.13),(5.17) for this problem, one might consider a test problem similar to the Dahlquist test problem (4.4), i.e.,

$$y'(t) = \lambda_{\text{adv}} y + \lambda_{\text{diff}} y + \lambda_{\text{react}} y.$$

It is not difficult to see that the original problem (5.12),(5.16) can be reduced to this last problem under a strict and unrealistic assumption that the reaction terms are linear and all the three operators involved (advection, diffusion, reaction) commute.

Alternatively, to study stability, we can consider a simpler, lower order scheme called ROS1 (one stage Rosenbrock)

$$y^{k+1} = y^k + k_1, \qquad (I+\tau\hat{A})k_1 = \tau f(y^k), \tag{5.18}$$

where $\tau > 0$ is the time step size. One can easily check that that for $f(y) = -Ay$ the ROS1 scheme can be written as

$$B\frac{y^{k+1}-y^k}{\tau} + Ay^k = 0, \qquad B = I + \tau\hat{A}. \tag{5.19}$$

This formula is known in Russian numerical literature as a canonical form of a two-level difference scheme, see e.g. [36, 37, 38]. If A is symmetric positive definite, and $\|y\|_A$ and $\|S\|_A$ are vector and matrix norms associated with A,

$$\|y\|_A = \sqrt{(Ay,y)}, \quad \|S\|_A^2 = \inf\left\{M \mid (ASy,Sy) \leqslant M(Ay,y)\ \forall y \in \mathbb{R}^n\right\},$$

the following stability result (due to Samarskii) can be established.

Theorem 5.2 *[36, 37, 38] Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix and let $B \in \mathbb{R}^{n\times n}$ be a matrix such that $B + B^T$ is positive definite. The time integration scheme* (5.19) *applied to the ODE system $y'(t) = -Ay$ is stable in the sense that*

$$\|S\|_A \leqslant 1, \qquad S = B^{-1}(I - \tau A),$$

if and only if

$$(Bx,x) \geqslant \frac{\tau}{2}(Ax,x), \quad \forall x \in \mathbb{R}^n.$$

It is instructive to consider several examples to illustrate this result. First of all, for $B = I$ the ROS1 turns into the Euler forward scheme. According to Theorem 5.2, the scheme is stable if and only if

$$(x, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n,$$

which can be checked to be equivalent to

$$\tau \leqslant \frac{2}{\|A\|_2}.$$

Furthermore, ROS1 with $B = I + \tau A$ yields the Euler backward scheme for which the stability condition

$$((I + \tau A)x, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n,$$

trivially holds for all $\tau > 0$. Finally, the θ-method (4.3) corresponds to the choice $B = I + \tau\theta A$ and has the stability condition

$$((I + \tau\theta A)x, x) \geqslant \frac{\tau}{2}(Ax, x), \quad \forall x \in \mathbb{R}^n.$$

This can be rewritten as

$$1 + (\theta - \frac{1}{2})\tau\frac{(Ax, x)}{(x, x)} \geqslant 0, \quad \forall x \in \mathbb{R}^n,$$

which holds for all $\tau > 0$ provided $\theta \geqslant \frac{1}{2}$ (this can be easily confirmed based on results from Section 4). For $\theta < \frac{1}{2}$ we can recast the stability condition into the form

$$\tau \geqslant \frac{2}{(1 - 2\theta)\|A\|_2}.$$

5.6 An enhanced matrix factorization AMF+

The ROS2-AMF method usually leads to a better accuracy for advection–diffusion–problems than splitting schemes [48, 2]. However, in some cases the error of ROS2-AMF can be significant. This is by no means a surprise if we take a closer look at (5.15). Denoting the diffusion–reaction Jacobian $A_1 + A_2$ in (5.17) by A, we have

$$\text{AMF error} = I + \gamma\tau\hat{A} - (I + \gamma\tau A) = (\gamma\tau)^2 A_1 A_2.$$

We see that the error in the AMF is small *asymptotically* for $\tau \to 0$. In a global air pollution model TM5 [44], for typically used step sizes τ the eigenvalues of τA_1 (diffusion terms) range in absolute value from 10^{-5} to 10, whereas the eigenvalues of τA_2 (reaction terms) range in absolute value from 10^{-5} to 10^6 [7]. Thus, even though AMF works in this case, we can not expect that it provides a reasonable approximation to the true diffusion–reaction Jacobian.

As shown in [7], the approximation of the AMF can be improved provided that the diffusion matrix A_1 is columnwise weakly diagonally dominant (i.e., the entries of A_1^T satisfy (1.4)) and permits an LU factorization of $I + \gamma\tau A_1$ without pivoting. Indeed, let $I + \gamma\tau A_1 = LU$ be such an LU factorization and, furthermore, let the diagonal entries in the lower triangular matrix L be all ones (why is it possible?). Consider the following improved AMF which we call AMF+:

$$I + \gamma\tau\hat{A} = L(U + \gamma\tau A_2), \tag{5.20}$$

where $A_2 = R'(y^k)$ is the reaction Jacobian. For AMF+ we have

$$\text{AMF+ error} = I + \gamma\tau\hat{A} - (I + \gamma\tau A) = \gamma\tau(L - I)A_2.$$

By applying the result of Exercise 5.4 (given below) to the LU factorization of the matrix $I + \gamma\tau A_1$, we see that off-diagonal entries in L are $O(\tau)$. Hence, the error of the AMF+ is $O(\tau^2)$. This result is again asymptotic and does not guarantee that the AMF+ error is bounded for realistic values of τ. As we will show now, L inherits columnwise diagonal dominance of $I + \gamma\tau A_1$ and, hence,

$$\|L - I\|_1 \leqslant 1.$$

Thus, we have

$$\|\text{AMF+ error}\|_1 \leqslant \gamma\tau\|A_2\|_1.$$

This makes the difference between AMF and AMF+ visible: the error of AMF+ is $O(\tau^2)$ for small τ but is bounded by a constant times τ for any $\tau > 0$.

Exercise 5.4 Assume that the matrix $A \in \mathbb{R}^{n\times n}$ is such that the LU factorization without partial pivoting of $I + \tau A$ exists: $I + \tau A = LU$, where L and U are respectively lower and upper triangular matrices. Furthermore, let $l_{ii} = 1$, $i = 1, \dots, n$. Show that for all off-diagonal entries of L holds $l_{ij} = O(\tau)$. Hint: use mathematical induction on the matrix size n. ♢

We now prove the following theorem.

Theorem 5.3 *If $A \in \mathbb{R}^{n\times n}$ is columnwise diagonally dominant and $LU = A$ is its LU factorization with $l_{ii} = 1$, $i = 1, \dots, n$, then L is also columnwise diagonally dominant:*

$$\sum_{i=j+1}^{n} |l_{ij}| \leqslant |l_{jj}| = 1.$$

Proof The proof is by induction on the matrix dimension n. It is not difficult to check the result holds for $n = 2$. Assuming it holds for $n - 1$, we partition $A \in \mathbb{R}^{n\times n}$ as

$$A = \begin{bmatrix} a_{11} & a_U^T \\ a_L & A_{n-1} \end{bmatrix}, \qquad a_L, a_U \in \mathbb{R}^{n-1},\ A_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Then we can partition the LU factors of A in a similar way:

$$L = \begin{bmatrix} 1 & 0 \\ l & L_{n-1} \end{bmatrix}, \qquad U = \begin{bmatrix} a_{11} & a_U^T \\ 0 & U_{n-1} \end{bmatrix}.$$

The entries in the first column of L are $l_{i1} = a_{i1}/a_{11}$, $i = 2, \dots, n$, and hence

$$\sum_{i=2}^{n} |l_{i1}| = \frac{1}{|a_{11}|} \sum_{i=2}^{n} |a_{i1}| \leqslant \frac{1}{|a_{11}|} |a_{11}| = 1.$$

Thus, we have diagonal dominance for the first column in L. We have

$$L_{n-1} U_{n-1} = A_{n-1} - l a_U^T,$$

and therefore, due to the induction assumption, L_{n-1} will be columnwise diagonally dominant provided $A_{n-1} - l a_U^T$ is columnwise diagonally dominant. A check that, indeed, $A_{n-1} - l a_U^T$ possesses this property is left as an exercise. □

Exercise 5.5 Finish the proof of Theorem 5.3. ♢

The assumption of this section that A_1 is columnwise weakly diagonally dominant is made because this property holds for the TM5 model [2]. Note that if the matrix A_1 is *row-wise* weakly diagonally dominant then we can adjust the definition (5.20) as

$$I + \gamma\tau \hat{A} = (L + \gamma\tau A_2) U,$$

where $LU = I + \gamma\tau A_1$ is the LU factorization with $u_{kk} = 1$, $k = 1, \dots, n$. For this modified AMF+ we have $\|U - I\|_\infty \leqslant 1$ and $\|\text{AMF+ error}\|_\infty \leqslant \gamma\tau \|A_2\|_\infty$.

6 Krylov subspace methods for matrix exponential actions

6.1 Krylov subspace and matrix polynomials

For zero source term g, the explicit solution of (2.1) is given by

$$y(t) = e^{-tA} y^0. \tag{6.1}$$

An approximate action of the matrix exponential operator on the vector y_0 can be computed using the Krylov subspace framework [31, 12, 13, 8, 19] as follows. Using the so-called modified Gram-Schmidt process it is easy to compute the matrices $V_{k+1} \in \mathbb{R}^{n \times (k+1)}$ and upper-Hessenberg[4] $H_{k+1,k} \in \mathbb{R}^{(k+1) \times k}$ such that (see e.g. [35, 46])

$$V_{k+1} = \begin{bmatrix} v_1 & \dots & v_{k+1} \end{bmatrix}, \quad V_{k+1}^T V_{k+1} = I \in \mathbb{R}^{(k+1)\times(k+1)},$$
$$\operatorname{colspan}(V_{k+1}) = \operatorname{span}(y^0, Ay^0, \dots, A^k y^0)$$

and

$$A V_k = V_{k+1} H_{k+1,k} = V_k H_{k,k} + h_{k+1,k} v_{k+1} e_k^T, \tag{6.2}$$

[4] A matrix $H = (h_{ij})$ is called upper-Hessenberg if $h_{ij} = 0$ for $i > j + 1$.

where v_i is the ith column of V_k, $v_1 = y^0/\|y^0\|_2$, $H_{k,k}$ is $H_{k+1,k}$ with the last row omitted and $e_k = (0, \dots, 0, 1)^T \in \mathbb{R}^k$. The subspace spanned by the columns of V_k is called the Krylov subspace and denoted by $\mathcal{K}_k(A, y^0)$:

$$\mathcal{K}_k(A, y^0) = \operatorname{span}(y^0, Ay^0, \dots, A^{k-1}y^0).$$

Using the just constructed V_k and $H_{k,k}$ we can compute an approximation to (6.1) as

$$y(t) = e^{-tA}y^0 = \beta e^{-tA}V_k e_1 \approx \beta V_k e^{-tH_{k,k}} e_1, \tag{6.3}$$

where $e_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^k$ and $\beta = \|y^0\|_2$. Why can it be a good approximation? To understand this, we follow [34] and give a number of arguments.

First, the following result holds.

Lemma 6.1 *[34] Let $V_k \in \mathbb{R}^{n\times k}$ and upper-Hessenberg $H_{k,k} \in \mathbb{R}^{k\times k}$ be the matrices as defined above. Then for any polynomial p_j of degree $j \leqslant k-1$ we have*

$$p_j(A)v_1 = V_k p_j(H_k)e_1,$$

where the notation is as defined above.

Proof [34] Denote $\pi_k = V_k V_k^T$. Using induction, let us check that $A^j v_1 = V_k H_k^j e_1$, $j = 0, 1, \dots, k-1$. For $j = 0$ we have $v_1 = V_k e_1$ and, thus, the relation holds for $j = 0$. Assuming that it holds for a certain $j \leqslant k-2$, consider the case $j+1$. Note that $A^{j+1}v_1, A^j v_1 \in \mathcal{K}_k(A, y_0)$. Then we obtain

$$\begin{aligned} A^{j+1}v_1 = \pi_k A^{j+1}v_1 = \pi_k A A^j v_1 = \pi_k A \pi_k A^j v_1 = V_k H_k V_k^T A^j v_1 = \\ = V_k H_k V_k^T V_k H_k^j e_1 = V_k H_k^{j+1} e_1. \end{aligned}$$

□

Second, a well known fact is that if ν is the degree of the minimal polynomial of A then any power of A is a polynomial in A of degree not exceeding $\nu - 1$.

Third, the following fundamental result holds (see [15] for a proof).

Theorem 6.1 *[15] Let $A \in \mathbb{R}^{n\times n}$ have the minimal polynomial of degree ν. Then for any function f analytic in an open set containing the spectrum $\Lambda(A)$ of A holds*

$$f(A) = p_{\nu-1}(A),$$

where $p_{\nu-1}$ interpolates f on $\Lambda(A)$ in the Hermite sense with the eigenvalues repeated according their multiplicities[5].

We now assume that all subdiagonal entries in $H_{k,k}$ are nonzero, i.e., $h_{j+1,j} \neq 0$, $j = 1, \dots, k-1$ (otherwise, if, for some k, $h_{k+1,k} = 0$ then the columns of V_k span an invariant subspace of A). Hence, the geometric multiplicity of all

[5] We say that a polynomial p interpolates a function f in the Hermite sense at given x repeated l times if $f^{(j)}(x) = p^{(j)}(x)$, $j = 0, \dots, l-1$.

the eigenvalues of $H_{k,k}$ is one and its minimal polynomial coincides with its characteristic polynomial. Therefore,

$$e^{H_{k,k}} = p_{k-1}(H_{k,k}), \tag{6.4}$$

with p_{k-1} being the unique polynomial of degree $k-1$ which interpolates the exponential function on $\Lambda(H_{k,k})$ in the Hermite sense with the eigenvalues repeated according their multiplicities.

All this brings us to the following result.

Theorem 6.2 *[34] For approximation* (6.3) *holds*

$$\beta V_k e^{-tH_{k,k}} e_1 = p_{k-1}(-tA)y^0,$$

with p_{k-1} defined by (6.4).

Proof

$$\beta V_k e^{-tH_{k,k}} e_1 = \beta V_k p_{k-1}(-tH_{k,k})e_1 \overset{\text{(Lemma 6.1)}}{=} \beta p_{k-1}(-tA)v_1 = p_{k-1}(-tA)y^0.$$

□

We note that the eigenvalues of $H_{k,k}$ are called the Ritz values of A and there is a lot of literature explaining why and how well the eigenvalues of A are approximated by the Ritz values for increasing k (see e.g. [35, 46]).

6.2 An alternative derivation of the approximation

Following [8, 11, 5], we finish this section by giving another derivation of the approximation (6.3). Assume we solve (2.1) with zero source term g approximately by projecting it in the Galerkin sense onto the Krylov subspace colspanV_k. This means that we look for an approximate solution $y_k(t) \approx y(t)$ such that

$$y_k(t) = V_k u(t) \quad \text{and} \quad r_k(t) \perp \text{colspan} V_k, \tag{6.5}$$

where $r_k(t)$ is the residual of $y_k(t)$ defined as [8, 11, 5]

$$r_k(t) = -y_k'(t) - Ay_k(t).$$

Substituting $y_k(t) = V_k u(t)$ into $y' = -Ay(t)$ and noticing that $V_k^T V_k$ is the identity, we arrive at the projected IVP for the function $u(t)$:

$$u'(t) = -\underbrace{V_k A V_k}_{H_{k,k}} u(t), \qquad u(0) = \beta e_1, \tag{6.6}$$

where all the notation is as defined in the beginning of the section. Note that $u(t) = \beta e^{-tH_{k,k}} e_1$. Moreover, using (6.2), we can obtain an expression for the residual $r_k(t)$ which allows us to control the quality of the approximate solution

$y_k(t)$. Indeed [8, 11],

$$\begin{aligned} r_k(t) &= -y_k'(t) - Ay_k(t) = -V_k u'(t) - AV_k u(t) = (V_k H_{k,k} - AV_k)u(t) \\ &= (V_k H_{k,k} - V_{k+1} H_{k+1,k})u(t) = -h_{k+1,k} v_{k+1} e_k^T u(t) = -h_{k+1,k} v_{k+1} e_k^T e^{-tH_{k,k}} u(0) = \\ &= -h_{k+1,k} v_{k+1} e_k^T e^{-tH_{k,k}} \beta e_1 = \underbrace{-h_{k+1,k} e_k^T e^{-tH_{k,k}} \beta e_1}_{\text{a scalar function of } t} v_{k+1} \perp \mathrm{colspan} V_k. \end{aligned}$$

This residual can be useful for different purposes, see e.g. [5, 6].

For the inhomogeneous problems (2.1), i.e., with nonzero source term $g(t)$, the Krylov subspace approximations to the matrix exponential can be employed in the framework of the so-called exponential integrators, see e.g. [20]. For these problems one can also use a projection on a single *block* Krylov subspace [4], similarly to (6.5),(6.6).

7 Acknowledgements

I would like to thank the organizers of the Rome–Moscow school for inviting me to teach in the school, the students of the school for their interest in the lectures and for pointing me errors in the draft versions of these notes (all remaining errors are mine). I am thankful to Nikolai Zamarashkin for his useful comments. Special thanks go to my former colleagues Jaap van der Vegt and Gerrit Zwier (University of Twente) for facilitating my participation in the school.

References

[1] R. M. Beam and R. F. Warming. An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. *J. Comput. Phys.*, 22:87–110, 1976.

[2] P. J. F. Berkvens, M. A. Botchev, M. C. Krol, W. Peters, and J. G. Verwer. Solving vertical transport and chemistry in air pollution models. In D. Chock and G. Carmichael, editors, *Atmospheric Modeling*, volume 130 of *IMA Volumes in Mathematics and its Applications*, pages 1–20. Springer, 2002.

[3] M. A. Bochev (Botchev). On the stability of nonselfadjoint difference schemes with M-matrices for evolution boundary value problems with an elliptic operator with respect to space. *Izv. Vyssh. Uchebn. Zaved. Mat.*, 9:15–22, 1995.

[4] M. A. Botchev. A block Krylov subspace time-exact solution method for linear ordinary differential equation systems. *Numer. Linear Algebra Appl.*, 20(4):557–574, 2013.

[5] M. A. Botchev, V. Grimm, and M. Hochbruck. Residual, restarting and Richardson iteration for the matrix exponential. *SIAM J. Sci. Comput.*, 35(3):A1376–A1397, 2013. `http://dx.doi.org/10.1137/110820191`.

[6] M. A. Botchev, I. V. Oseledets, and E. E. Tyrtyshnikov. Iterative across-time solution of linear differential equations: Krylov subspace versus waveform relaxation. *Computers & Mathematics with Applications*, 67(12):2088–2098, 2014. `http://dx.doi.org/10.1016/j.camwa.2014.03.002`.

[7] M. A. Botchev and J. G. Verwer. A new approximate matrix factorization for implicit time integration in air pollution modeling. *J. Comp. Appl. Math.*, 157:309–327, 2003. `http://dx.doi.org/10.1016/S0377-0427(03)00414-X`.

[8] E. Celledoni and I. Moret. A Krylov projection method for systems of ODEs. *Appl. Numer. Math.*, 24(2-3):365–378, 1997.

[9] J. Crank and P. Nicolson. A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. *Proc. Camb. Philos. Soc.*, 43:50–67, 1947.

[10] K. Dekker and J. G. Verwer. *Stability of Runge–Kutta methods for stiff nonlinear differential equations.* North-Holland Elsevier Science Publishers, 1984. Russian translation: К. Деккер, Я. Вервер. Устойчивость методов Рунге–Кутты для жёстких нелинейных дифференциальных уравнений.— М.: Мир, 1988.

[11] V. L. Druskin, A. Greenbaum, and L. A. Knizhnerman. Using nonorthogonal Lanczos vectors in the computation of matrix functions. *SIAM J. Sci. Comput.*, 19(1):38–54, 1998.

[12] V. L. Druskin and L. A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *U.S.S.R. Comput. Maths. Math. Phys.*, 29(6):112–121, 1989.

[13] V. L. Druskin and L. A. Knizhnerman. Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic. *Numer. Lin. Alg. Appl.*, 2:205–217, 1995.

[14] E. G. D'yakonov. Difference systems of second order accuracy with a divided operator for parabolic equations without mixed derivatives. *USSR Comput. Math. Math. Phys.*, 4(5):206–216, 1964.

[15] F. R. Gantmacher. *The Theory of Matrices. Vol. 1.* AMS Chelsea Publishing, Providence, RI, 1998. Translated from the Russian by K. A. Hirsch, Reprint of the 1959 translation.

[16] A. Gerisch and J. G. Verwer. Operator splitting and approximate factorization for taxis-diffusion-reaction models. *Appl. Numer. Math.*, 42:159–176, 2002.

[17] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential–Algebraic Problems.* Springer Series in Computational Mathematics 14. Springer–Verlag, 2 edition, 1996.

[18] N. J. Higham. *Functions of Matrices: Theory and Computation.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2008.

[19] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34(5):1911–1925, Oct. 1997.

[20] M. Hochbruck and A. Ostermann. Exponential integrators. *Acta Numer.*, 19:209–286, 2010.

[21] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1986. Russian translation: Р. Хорн, Ч. Джонсон. Матричный анализ.—М.: Мир, 1989.

[22] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.

[23] W. Hundsdorfer and J. G. Verwer. *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations.* Springer Verlag, 2003.

[24] L. A. Krukier. Implicit difference schemes and an iterative method for solving them for a certain class of systems of quasi-linear equations. *Sov. Math.*, 23(7):43–55, 1979. Translation from Izv. Vyssh. Uchebn. Zaved., Mat. 1979, No. 7(206), 41–52 (1979).

[25] B. Lastdrager, B. Koren, and J. G. Verwer. Solution of time-dependent advection-diffusion problems with the sparse-grid combination technique and a Rosenbrock solver. *Comput. Methods Appl. Math.*, 1(1):86–99, 2001.

[26] S. M. Lozinskiĭ. Error estimate for numerical integration of ordinary differential equations. I. *Izv. Vysš. Učebn. Zaved. Matematika 1958, no. 5 (6), 52-90; Translated as: Izvestija Vysših Učebnyh Zavedeniĭ Matematika*, 1959(5 (12)):222, 1959.

[27] G. I. Marčuk. Some application of splitting-up methods to the solution of mathematical physics problems. *Apl. Mat.*, 13:103–132, 1968.

[28] J. M. Ortega. *Matrix theory. A second course.* The University Series in Mathematics. Plenum Press, New York, 1987.

[29] J. M. Ortega. *Introduction to Parallel and Vector Solution of Linear Systems.* Plenum Press, 1988. Russian translation: Дж. Ортега. Введение в параллельные и векторные методы решения линейных систем.—М.: Мир, 1991.

[30] A. Ostermann. Stability of W-methods with applications to operator splitting and to geometric theory. *Appl. Numer. Math.*, 42(1–3):353–366, 2002. `http://dx.doi.org/10.1016/S0168-9274(01)00160-X`.

[31] T. J. Park and J. C. Light. Unitary quantum time evolution by iterative Lanczos reduction. *J. Chem. Phys.*, 85:5870–5876, 1986.

[32] D. W. Peaceman and H. H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3:28–41, 1955.

[33] D. J. Rose. Convergent regular splittings for singular M-matrices. *SIAM J. Algebraic Discrete Methods*, 5(1):133–144, 1984.

[34] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29(1):209–228, 1992.

[35] Y. Saad. *Iterative Methods for Sparse Linear Systems.* SIAM, 2d edition, 2003. Available from `http://www-users.cs.umn.edu/~saad/books.html`.

[36] A. A. Samarskii. Regularization of difference schemes. *USSR Comput. Math. and Math. Phys.*, 7:62–93, 1967.

[37] A. A. Samarskii. *Theorie der Differenzenverfahren.* Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig, 1984. Translated from the Russian by Gisbert Stoyan.

[38] A. A. Samarskii and E. S. Nikolaev. *Numerical methods for grid equations. Vol. I&II.* Birkhäuser Verlag, Basel, 1989.

[39] J. M. Sanz-Serna and M. P. Calvo. *Numerical Hamiltonian Problems.* Chapman & Hall, 1994.

[40] R. B. Sidje. Expokit. A software package for computing matrix exponentials. *ACM Trans. Math. Softw.*, 24(1):130–156, 1998. `www.maths.uq.edu.au/expokit/`.

[41] B. Sportisse. An analysis of operator splitting techniques in the stiff case. *J. Comput. Phys.*, 161(1):140–168, 2000.

[42] G. Strang. Accurate partial difference methods I: linear Cauchy problems. *Archive for Rational Mechanics and Analysis*, 12:392–402, 1963.

[43] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5(3):506–517, 1968.

[44] TM5: global chemistry transport model. Wageningen University, the Netherlands, 2016. `http://tm5.sourceforge.net/`.

[45] P. J. van der Houwen and B. P. Sommeijer. Approximate factorization for time-dependent partial differential equations. *J. Comput. Appl. Math.*, 128(1-2):447–466, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.

[46] H. A. van der Vorst. *Iterative Krylov methods for large linear systems.* Cambridge University Press, 2003.

[47] R. S. Varga. *Matrix Iterative Analysis.* Prentice-Hall, 1962.

[48] J. G. Verwer, W. Hundsdorfer, and J. G. Blom. Numerical time integration for air pollution models. *Surveys for Mathematics in Industry*, 10:107–174, 2002.

[49] J. G. Verwer, E. J. Spee, J. G. Blom, and W. Hundsdorfer. A second order Rosenbrock method applied to photochemical dispersion problems. *SIAM J. Sci. Comput.*, 20:456–480, 1999.

[50] N. N. Yanenko. *The method of fractional steps. The solution of problems of mathematical physics in several variables.* Springer-Verlag, New York, 1971.

[51] D. M. Young. *Iterative Solution of Large Linear Systems.* Academic Press, 1971.

Index

Contents