Numerical Solution of Advection-Diffusion-Reaction Equations

Lecture notes, 2000, Thomas Stieltjes Institute

Willem Hundsdorfer
CWI, Amsterdam
(Willem.Hundsdorfer@cwi.nl)

Contents

0. Introduction

In these notes we shall discuss various numerical aspects for the solution of advection-diffusion-reaction equations. Problems of this type occur for instance in the description of transport-chemistry in the atmosphere and we shall consider the equations with this application as reference. Other examples for the occurrence of advection-diffusion-reaction equations can be found in the introduction of Morton (1996).

The advection-diffusion-reaction equations

The mathematical equations describing the evolution of chemical species can be derived from mass balances. Consider a concentration $u(x,t)$ of a certain chemical species, with space variable x and time t. Let $h > 0$ be a small number, and consider the average concentration $\bar{u}(x,t)$ in a cell $\Omega(x) = [x - \frac{1}{2}h, x + \frac{1}{2}h]$,

$$\bar{u}(x,t) = \frac{1}{h}\int_{x-h/2}^{x+h/2} u(x',t)dx' = u(x,t) + \frac{1}{24}h^2 u_{xx}(x,t) + \cdots .$$

If the species is carried along by a flowing medium with velocity $a(x,t)$ then the mass conservation law implies that the change of $\bar{u}(x,t)$ per unit of time is the net balance of inflow and outflow over the cell boundaries,

$$\frac{\partial}{\partial t}\bar{u}(x,t) = \frac{1}{h}\Big[a(x - \tfrac{1}{2}h, t)u(x - \tfrac{1}{2}h, t) - a(x + \tfrac{1}{2}h, t)u(x + \tfrac{1}{2}h, t)\Big].$$

Here $a(x \pm \frac{1}{2}h, t)u(x \pm \frac{1}{2}h, t)$ are the fluxes over the left and right cell boundaries. Now, if we let $h \to 0$, it follows that the concentration satisfies

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\Big(a(x,t)u(x,t)\Big) = 0.$$

This is called an advection equation (or convection equation). In a similar way we can consider the effect of diffusion. Then the change of $\bar{u}(x,t)$ is caused by gradients in the solution and the fluxes across the cell boundaries are $-d(x \pm \frac{1}{2}h, t)u_x(x \pm \frac{1}{2}h, t)$ with $d(x,t)$ the diffusion coefficient. The corresponding diffusion equation is

$$\frac{\partial}{\partial t}u(x,t) = \frac{\partial}{\partial x}\Big(d(x,t)\frac{\partial}{\partial x}u(x,t)\Big).$$

There may also be a change in $u(x,t)$ due to sources, sinks and chemical reactions, leading to

$$\frac{\partial}{\partial t}u(x,t) = f(u(x,t), x, t).$$

The overall change in concentration is described by combining these three effects, leading to the advection-diffusion-reaction equation

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\Big(a(x,t)u(x,t)\Big) = \frac{\partial}{\partial x}\Big(d(x,t)\frac{\partial}{\partial x}u(x,t)\Big) + f(x,t,u(x,t)) \ .$$

We shall consider the equation in a spatial interval $\Omega \subset \mathbb{R}$ with time $t \geq 0$. An initial profile $u(x,0)$ will be given and we also assume that suitable boundary conditions are provided.

More general, let $u_1(x,t), \cdots, u_s(x,t)$ be concentrations of s chemical species, with spatial variable $x \in \Omega \subset \mathbb{R}^d$ ($d = 2$ or 3), and time $t \geq 0$. Then the basic mathematical equations for transport and reaction are given by the following set of partial differential equations (PDEs)

$$\frac{\partial}{\partial t} u_j(x,t) + \sum_{k=1}^{d} \frac{\partial}{\partial x_k} \Big(a_k(x,t) u_j(x,t) \Big) =$$

$$\sum_{k=1}^{d} \frac{\partial}{\partial x_k} \Big(d_k(x,t) \frac{\partial}{\partial x_k} u_j(x,t) \Big) + f_j(u_1(x,t), ..., u_s(x,t), x, t) \quad , \quad j = 1, 2, ..., s$$

with suitable initial and boundary conditions. The quantities a_k that represent the velocities of the transport medium, such as water or air, are either given in a data archive or computed alongside with a meteorological or hydrodynamical code. (In such codes Navier-Stokes or shallow water equations are solved, where again advection-diffusion equations are of primary importance.) The diffusion coefficients d_k are constructed by the modellers and may include also parametrizations of turbulence. The final term $f_j(c,x,t)$, which gives a coupling between the various species, describes the nonlinear chemistry together with emissions (sources) and depositions (sinks). In actual models these equations are augmented with other suitable sub-grid parametrizations and coordinate transformations.

Numerical requirements in air pollution models

The data, such as the velocity field a, diffusion coefficients and reaction constants, are in general not very accurate. Therefore the accuracy requirements for the numerical solution are also low. On the other hand, with many models there are very many spatial grid points, for instance 10^4 for a domain covering Europe, 10^6 for global models. The number of species s may range typically from 10 to 100. So, the problems may be "very big" and we need

- fast, "cheap" numerical methods.

Often, one is interested in long term effects, so that the equations have to be integrated over long time intervals. Therefore, in spite of the low accuracy demands, the numerical solutions should be "qualitatively correct", and we need properties like

- mass conservation,
- positivity,
- small phase errors.

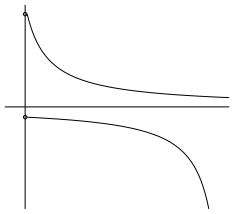
In most air pollution models the transport is advection dominated, and there can be strong, local sources. Hence we may expect steep spatial gradients in the solution and numerical schemes are needed with

- good resolution of steep gradients.

All these requirements together are already difficult to fulfill. In the next sections these various aspects will be addressed.

Further, it should also be noted that the reaction terms are usually very *stiff*, that is, some reactions take place on very small time scales compared to the overall time scale, due to large reaction constants. This implies that such terms have to be solved implicitly, which make them difficult and time consuming. Moreover, with large reaction constants *positivity* is often necessary to maintain *stability* of the model.

As a very simple example, consider $s = 1$ and $f(w) = -\kappa w^2$ with reaction constant $\kappa \gg 1$. Then solutions of $w'(t) = f(w(t))$ are only stable if we start with $w(0) \geq 0$. With $w(0) < 0$ there will be a "blow-up" of the solution.



Therefore, if such a reaction term occurs for a certain chemical component, the treatment of advection-diffusion should be such that negative values are avoided.

Relevant literature

Numerical problems arising with air pollution models are discussed in more detail in the review paper of McRea et al. (1982) and the book of Zlatev (1995). Standard text books that deal with advection-diffusion problems are Richtmyer & Morton (1967), Mitchell & Griffiths (1980), Hirsch (1988) and Strikwerda (1989).

In these notes we shall mainly consider finite-difference or finite-volume methods on simple grids. For spectral methods and finite elements we refer to Canuto et al. (1988) and Strang & Fix (1973), respectively. More recent material on finite element methods can be found in the monograph of Morton (1996).

Some implicit ODE methods that are suited for stiff chemistry calculations are listed in the appendix. A comprehensive treatment of such ODE methods is found in Hairer and Wanner (1991).

1. Some simple space discretizations and modified equations

1.1. Test equations

To introduce numerical schemes for the advection-diffusion-reaction equations we first consider some spatial discretizations for simple advection and diffusion equations with constant coefficients. We consider the following partial differential equations (PDEs)

$$u_t + au_x = 0 \qquad \text{for } x \in \mathbb{R},\ t \geq 0, \tag{1.1}$$

and

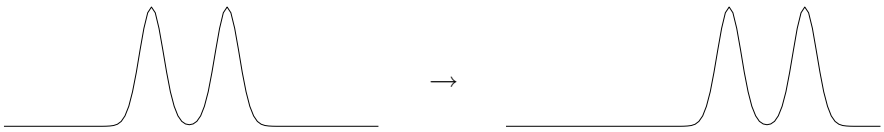
$$u_t = du_{xx} \qquad \text{for } x \in \mathbb{R},\ t \geq 0 \tag{1.2}$$

with given constants $a \in \mathbb{R}, d > 0$, given initial value $u(x,0)$ and the *periodicity condition*

$$u(x+1,t) = u(x,t).$$

The reason for considering periodicity conditions is mainly for the ease of presentation of the main concepts. Boundary conditions cause additional theoretical and numerical problems, as we shall see gradually in later sections. Note that with this periodicity condition we only have to compute the solution for $0 \leq x \leq 1$. In this section we shall look at some simple space discretizations and an attempt will be made to understand the qualitative behaviour of the discretizations. This will be done (in a heuristic way) by considering so-called modified equations. First we take a short look at the behaviour of the exact solutions.

Equation (1.1) is an *advection* (test-)problem. The solution simply is $u(x,t) = u(x-at,0)$. Initial profiles are shifted (carried along by the wind) with velocity a. The lines $x-at$ constant in the (x,t)-plane are the *characteristics* of this advection problem. Along these characteristics the solution $u(x,t)$ is constant.



Equation (1.2) is a *diffusion* (test-)problem. Insight in the behaviour of solutions can be obtained by *Fourier decompositions*. Consider

$$\varphi_k(x) = e^{2\pi ikx} \quad \text{for} \quad k \in \mathbb{Z}, \qquad (\varphi,\psi) = \int_0^1 \overline{\varphi(x)}\psi(x)dx.$$

The functions φ_k will be called Fourier modes, and (φ,ψ) is an inner product for the function space $L_2[0,1]$, consisting of all square integrable complex functions on $[0,1]$ with identification

of functions that differ only on sets of measure zero. The set $\{\varphi_k\}_{k\in\mathbb{Z}}$ is an orthonormal basis for this space. For any function $\psi \in L_2[0,1]$ we have

$$\psi(x) = \sum_{k\in\mathbb{Z}} \alpha_k \varphi_k(x) \qquad \text{with} \quad \alpha_k = (\varphi_k, \psi),$$

$$\|\psi\|_{L_2}^2 = \int_0^1 |\psi(x)|^2 dx = \sum_{k\in\mathbb{Z}} |\alpha_k|^2 \qquad \text{(Parseval's identity)}.$$

Formal proofs of these statements can be found in analysis text books where Fourier series are discussed, for example Pinkus & Zafrany (1997).

Now, consider (1.2) with initial profile $u(x,0) = \varphi_k(x)$ for some k. To find the solution we make the "Ansatz" (a motivated guess that will turn out right) by separation of variables

$$u(x,t) = \gamma(t)\varphi_k(x), \quad \gamma(0) = 1.$$

Inserting this into (1.2) leads to an equation for $\gamma(t)$,

$$\gamma'(t)\varphi_k(x) = -4\pi^2 k^2 d\,\gamma(t)\varphi_k(x),$$

$$\gamma(t) = e^{-4\pi^2 k^2 d\,t}.$$

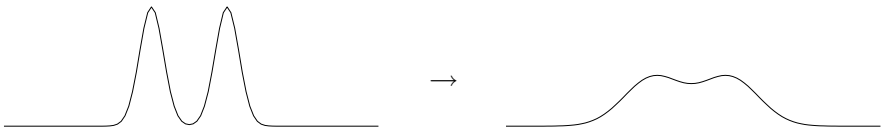
So we see that the Fourier modes are damped, the larger the frequency the stronger the damping. In general, if

$$u(x,0) = \sum_k \alpha_k \varphi_k(x),$$

then

$$u(x,t) = \sum_k \alpha_k e^{-4\pi^2 k^2 d\,t} \varphi_k(x).$$

Because the high frequencies are damped more rapidly than the low ones, the solution will become smoother. This is of course consistent with the physical interpretation of (1.2) as heat flow or diffusion caused by Brownian motion of particles.



For the advection-diffusion test problem

$$u_t + a u_x = d u_{xx} \tag{1.3}$$

with periodicity condition and $u(x,0) = \varphi_k(x)$ we get, in the same way as above,

$$u(x,t) = e^{(-2\pi ika - 4\pi^2 k^2 d)t}\varphi_k(x) = \underbrace{e^{-4\pi^2 k^2 dt}}_{damping}\ \underbrace{\varphi_k(x-at)}_{shift}$$

(superposition of previous cases). So, all Fourier modes are shifted with the same velocity and they are damped according to their frequency.

Remark. If d were negative, then the Fourier modes with high frequency would be strongly amplified and we would have instability in the L_2-norm (blow up). The sign of the velocity term a merely decides whether we have a shift to the left or to the right. $\diamond$

Remark. If $u(x,t)$ is a concentration then $\int_0^1 u(x,t)dx$ is the *mass* in [0,1] at time t. This is a conserved quantity:

$$\frac{d}{dt}\int_0^1 u(x,t)dx = \int_0^1 u_t(x,t)dx = \int_0^1 \Big(-au_x(x,t) + du_{xx}(x,t)\Big)dx =$$

$$= -a\Big(u(1,t) - u(0,t)\Big) + d\Big(u_x(1,t) - u_x(0,t)\Big) = 0,$$

due to the periodicity. $\diamond$

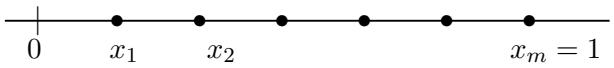
1.2. Finite difference discretizations

In this section we shall consider some simple space discretizations on a uniform grid $x_i = ih$ with mesh width $h = 1/m$. Approximations $w_i(t) \approx u(x_i,t)$, $i = 1,2,...,m$ are found by replacing the spatial derivatives by difference quotients. This gives a *finite difference discretization* in space. Setting $w(t) = (w_1(t),...,w_m(t))^T$, we then get a system of ordinary differential equations (ODEs)

$$w'(t) = F(t,w(t)), \tag{1.4}$$

with a given initial value $w(0)$. Often we shall deal with an F that is linear in w,

$$w'(t) = Aw(t) + g(t). \tag{1.5}$$



In later sections we shall also consider simple *finite volume discretizations* where the values $w_i(t)$ are interpreted as approximations to average values of $u(x,t)$ on the cells $[x_i - \frac{1}{2}h, x_i + \frac{1}{2}h]$. For the moment, with the above test equations, there is no difference between the two approaches.

Finite element and spectral discretizations are not considered here, but we note that also with such methods one arrives at ODE systems, $Bw'(t) = Aw(t) + Bg(t)$ with nonsingular mass matrix B, but the $w_i(t)$ will then refer to a weight of a basis function.

1.3. Discretizations for the advection operator

Consider the advection equation (1.1) with $a > 0$. The formula

$$\frac{1}{h}\Big(\psi(x-h) - \psi(x)\Big) = -\psi_x(x) + \mathcal{O}(h) \tag{1.6}$$

leads to the *1-st order upwind* discretization

$$w_i'(t) = \frac{a}{h}\Big(w_{i-1}(t) - w_i(t)\Big), \qquad i = 1, 2, ..., m, \tag{1.7}$$

with $w_0(t) = w_m(t)$ by periodicity. This is of the form (1.5) with $g = 0$ and

$$A = \frac{a}{h}\begin{pmatrix} -1 & & & & 1 \\ 1 & -1 & & & \\ & \ddots & \ddots & & \\ & & 1 & -1 & \\ & & & 1 & -1 \end{pmatrix}.$$

The formula

$$\frac{1}{2h}\Big(\psi(x-h) - \psi(x+h)\Big) = -\psi_x(x) + \mathcal{O}(h^2) \tag{1.8}$$

gives the *2-nd order central* discretization

$$w_i'(t) = \frac{a}{2h}\Big(w_{i-1}(t) - w_{i+1}(t)\Big), \qquad i = 1, 2, ..., m, \tag{1.9}$$

with $w_0(t) = w_m(t)$ and $w_{m+1}(t) = w_1(t)$. Here we have (1.5) with

$$A = \frac{a}{2h}\begin{pmatrix} 0 & -1 & & & 1 \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \\ -1 & & & 1 & 0 \end{pmatrix}.$$

For smooth profiles the 2-nd order scheme is better. However, consider $a = 1$ and initial profile $u(x, 0) = (\sin(\pi x))^{100}$. Solutions at $t = 1$ are given in the Figures 1.1 and 1.2 for $h = 1/50$, with dotted lines for the exact solution and solid lines for the numerical solution. The 1-st order scheme is not accurate, but the result of the 2-nd order scheme is also far from satisfactory: it gives oscillations, negative values and a significant phase error.

Figure 1.1. First order upwind advection

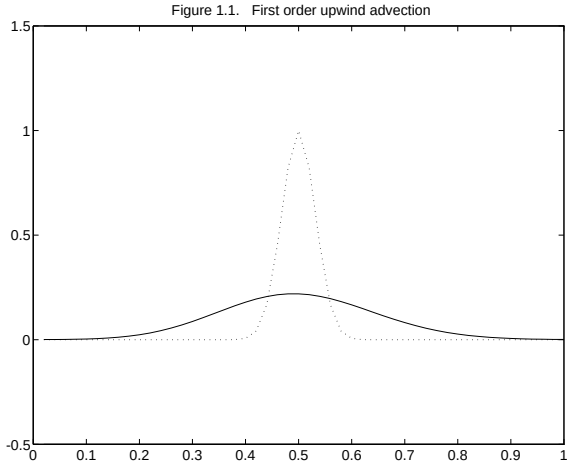
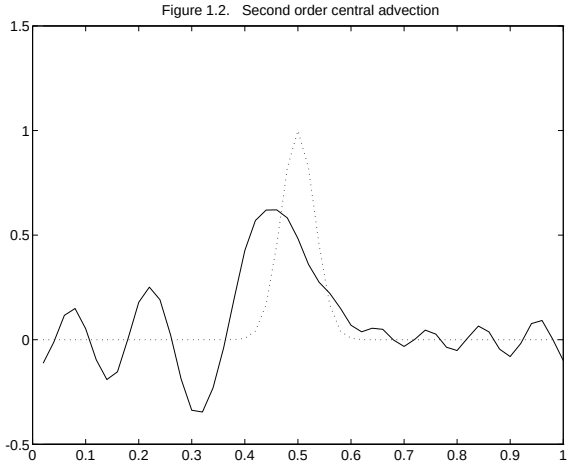


Figure 1.2. Second order central advection



The qualitative behaviour can be understood by considering the *modified equation* of the discretizations. Further expansion in formula (1.6) gives

$$\frac{1}{h}\Big(\psi(x-h)-\psi(x)\Big) = -\psi_x(x) + \frac{1}{2}h\psi_{xx}(x) + \mathcal{O}(h^2).$$

From this it can be seen (proof is given in the next section) that the upwind discretization (1.7) gives a first order approximation for $u_t + au_x = 0$, but it gives a second order approximation to the modified equation

$$\tilde{u}_t + a\tilde{u}_x = \frac{a}{2}h\tilde{u}_{xx}.$$

This explains the diffusive nature of the first order upwind discretization in Figure 1.1. Although we are seeking a solution to the advection problem, we are actually generating a solution to an advection-diffusion equation, with a numerical diffusion coefficient $\frac{1}{2}ah$.

Likewise, a further expansion in formula (1.8) gives

$$\frac{1}{2h}\Big(\psi(x-h)-\psi(x+h)\Big) = -\psi_x(x) - \frac{1}{6}h^2\psi_{xxx}(x) + \mathcal{O}(h^4),$$

from which it can be seen that the central discretization (1.9) gives a fourth order approximation to the modified equation

$$\tilde{u}_t + a\tilde{u}_x = -\frac{a}{6}h^2\tilde{u}_{xxx}$$

(again, arguments for convergence proof follow in the next section). The term $\tilde{u}_{xxx}$ gives rise to *dispersion*, that is, Fourier modes $\varphi_k(x)$ are shifted with with a velocity that depends on k. With initial value $\tilde{u}(x,0) = \varphi_k(x)$ the solution of this modified equation is

$$\tilde{u}(x,t) = e^{2\pi ik(x-a_kt)} = \varphi_k(x-a_kt), \qquad a_k = a(1-\frac{2}{3}\pi^2k^2h^2).$$

Hence Fourier modes with high frequencies move too slow.

If the initial profile is smooth, the coefficients in front of the high-frequency modes are very small. If the initial profile has large gradients then some high-frequency modes will be

significant, and then the dispersive effect will cause oscillations with the central discretization, see Figure 1.2.

Remark. If $\psi \in C^j(\mathbb{R})$ with period 1, $\psi(x) = \sum_k \alpha_k \varphi_k(x)$, then

$$|\alpha_k| \leq \frac{1}{(2\pi k)^j} \max_{0\leq x\leq 1} |\psi^{(j)}(x)|,$$

as can be seen by considering the inner product of $\psi^{(j)}$ with φ_k. (For a differentiable function we are allowed to differentiate its Fourier series, see Pinkus & Zafrany (1997, Sect.2.9).) Thus for smooth functions ψ the coefficients α_k are extremely small for large $|k|$.

If we consider a block-shaped function (say $\psi(x) = 1$ if $|x - \frac{1}{2}| \leq \frac{1}{4}$, $\psi(x) = 0$ otherwise) as an example of a function with a discontinuity, then it follows from direct calculation that $|\alpha_k| \sim k^{-1}$. Note that this function is piecewise C^1. $\diamond$

1.4. Discretization for the diffusion operator

Consider the diffusion equation (1.2). We have

$$\frac{1}{h^2}\Big(\psi(x-h) - 2\psi(x) + \psi(x+h)\Big) = \psi_{xx}(x) + \mathcal{O}(h^2). \tag{1.10}$$

This leads to the *second order central* discretization, for the diffusion equation,

$$w_i'(t) = \frac{d}{h^2}\Big(w_{i-1}(t) - 2w_i(t) + w_{i+1}(t)\Big), \quad i = 1, 2, ..., m, \tag{1.11}$$

with again $w_0 \equiv w_m$ and $w_{m+1} \equiv w_1$. This can be written as an ODE system with

$$A = \frac{d}{h^2}\begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{pmatrix}.$$

A further expansion in (1.10) gives

$$\frac{1}{h^2}\Big(\psi(x-h) - 2\psi(x) + \psi(x+h)\Big) = \psi_{xx}(x) + \frac{1}{12}h^2\psi_{xxxx}(x) + \mathcal{O}(h^4).$$

Therefore, the modified equation that is approximated with order 4 by this central discretization reads

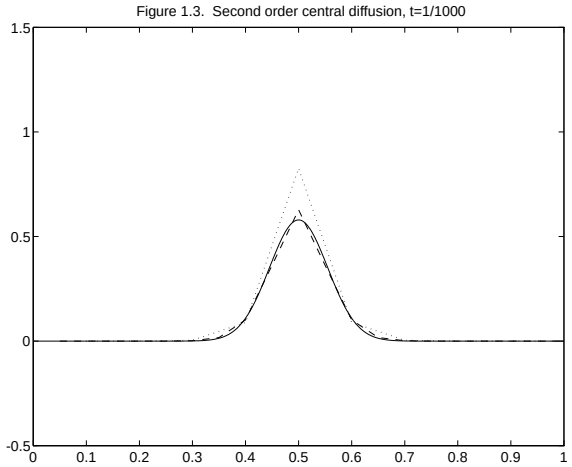
$$\tilde{u}_t = d\tilde{u}_{xx} + \frac{d}{12}h^2\tilde{u}_{xxxx}.$$

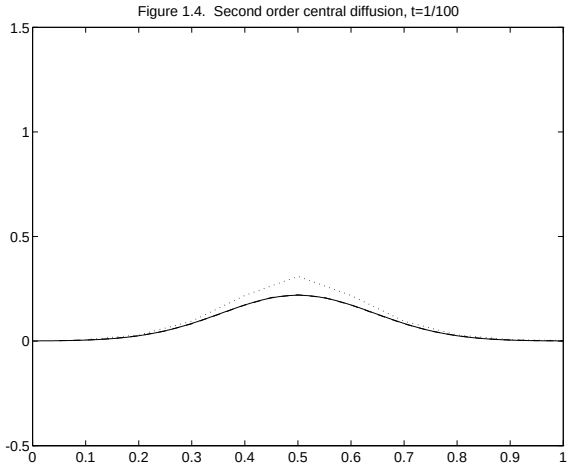
The heuristic analysis by this modified equation is not as easy as in the previous examples, due to the fact that this equation is not *well posed*: if $\tilde{u}(x,0) = \varphi_k(x)$ then $\tilde{u}(x,t) = \gamma(t)\varphi_k(x)$ with

$$\gamma(t) = e^{-4\pi^2k^2d(1-\frac{1}{3}\pi^2h^2k^2)t},$$

which gives exponential growth for $h^2k^2 > 3/\pi^2$. As we shall see below, it is only necessary to consider $|k| \leq \frac{1}{2}m$ since higher frequencies cannot be represented on the grid. This gives

$|hk| \leq \frac{1}{2}$ and thus all $\gamma(t)$ tend to zero. Under this restriction, the qualitative behaviour of the modified equation corresponds with that of the exact solution. In particular, there is no advection or dispersion, only damping. Indeed the qualitative behaviour is correct, see the Figures 1.3, 1.4. In these figures numerical solutions are plotted at time $t = 1/1000$ and $t = 1/100$, respectively, for the diffusion equation with $d = 1$ and initial profile $u(x, 0) = \sin(\pi x)^{100}$. The dotted line is the numerical solution with $h = 1/10$, the dashed line for $h = 1/20$ and solid line is the exact solution, also found numerically but with very small h (the numerical solution with $h = 1/40$ is already virtually the same). Note that even on the very coarse grid, with $h = 1/10$, the qualitative behaviour is correct (for a good quantitative behaviour we need a smaller h).





To see why we may restrict ourselves to $|k| \leq \frac{1}{2}m$ we consider *discrete Fourier decompositions.* Let

$$\phi_k = (\varphi_k(x_1), \varphi_k(x_2), ..., \varphi_k(x_m))^T \in \mathbb{C}^m \quad \text{for} \quad k \in \mathbb{Z},$$

and consider the inner product on $\mathbb{C}^m$

$$(v, w) = h \sum_{j=1}^{m} \overline{v_j} w_j.$$

We have

$$(\phi_k, \phi_l) = h \sum_{j=1}^{m} e^{2\pi i(l-k)x_j} = h \sum_{j=1}^{m} \rho^j, \qquad \rho = e^{2\pi i(l-k)h}.$$

If $k = l \bmod m$, then $\rho = 1$ and $(\phi_k, \phi_l) = 1$. Otherwise

$$(\phi_k, \phi_l) = h\rho \frac{1 - \rho^m}{1 - \rho} = 0 \quad \text{since} \quad \rho^m = e^{2\pi i(l-k)} = 1.$$

It follows that

$$\{\phi_{-k}, \phi_{-k+1}, ..., \phi_{m-k-1}\} \quad \text{is an orthonormal basis for } \mathbb{C}^m,$$

$$\phi_k = \phi_l \quad \text{if} \quad k = l \bmod m.$$

For the basis we can take $k = m/2$ if m is even, $k = (m-1)/2$ if m is odd. In conclusion, on our grid we can only represent Fourier modes with frequency $|k| \leq m/2$. To study a space discretization, and its modified equation, we thus may restrict ourselves to these modes.

We note that in above example for the diffusion test problem, one could also include higher order terms in the modified equation, leading to

$$\tilde{u}_t = d\tilde{u}_{xx} + \frac{d}{12}h^2\tilde{u}_{xxxx} + \frac{d}{460}h^4\tilde{u}_{xxxxxx}.$$

This equation is well posed, as can be seen by inserting Fourier modes. It is clear, however, that the modified equation approach, which easily gave insight in the advection discretizations, is rather cumbersome for the simple diffusion test equation. On the other hand, from a practical point of view, discretization of the diffusion equation poses much less problems than for the advection equation.

Note. Modified equations to study the behaviour of discretizations were introduced by Warming & Hyett (1974). We will consider such equations only in a heuristic fashion, with the aim of understanding the qualitative behaviour. A more general discussion on the subject can be found in Griffiths & Sanz-Serna (1986).

Remark. We have the following relation,

$$\frac{1}{h}\begin{pmatrix} -1 & & & 1 \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} = \frac{1}{2h}\begin{pmatrix} 0 & -1 & & 1 \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & -1 \\ -1 & & 1 & 0 \end{pmatrix} + \varepsilon\frac{1}{h^2}\begin{pmatrix} -2 & 1 & & 1 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & -2 \end{pmatrix},$$

with $\varepsilon = \frac{1}{2}h$. Thus the 1-st order upwind advection equals the 2-nd order central advection plus diffusion with numerical diffusion coefficient ε. Apparently, this numerical diffusion coefficient is too large, see Figure 1.1. So, an obvious question is whether better schemes can be constructed by a more careful addition of numerical diffusion. We shall return to this subject in connection with limited schemes. $\diamond$

2. Space discretizations: general considerations

2.1. Discretization and truncation errors

Consider a PDE solution $u(x,t)$ with $t \geq 0, x \in \Omega$ of an initial(-boundary) value problem. Discretization on a grid Ω_h, with $h > 0$ the mesh width (or maximal mesh width) yields an ODE system, the *semi-discrete* system,

$$w'(t) = F(t, w(t)), \quad w(0) \quad \text{given}, \tag{2.1}$$

with $w(t) \in \mathbb{R}^m$. The term semi-discrete is used to indicate that only space derivatives are discretized, the time is still continuous.

We want to compare $u(x,t)$ (function in x) with $w(t)$ (vector in $\mathbb{R}^m$). For this, let $w_h(t)$ be a suitable representation of the exact solution on $\mathbb{R}^m$. For example, for finite difference discretizations considered in the previous section the components of $w_h(t)$ will be function values $u(x_i,t)$ at the various grid points.

The *spatial (discretization) error* of the semi-discrete system is

$$w_h(t) - w(t).$$

In order to estimate this global quantity, we consider a suitable norm $\|\cdot\|$ on $\mathbb{R}^m$ (or $\mathbb{C}^m$), and we define the *space truncation error*

$$\sigma_h(t) = w_h'(t) - F(t, w_h(t)), \tag{2.2}$$

which is the residual obtained by substituting the exact PDE solution (or rather its representation on the grid) in the semi-discrete system. Assuming smoothness of the PDE solutions one obtains, by Taylor expansion, an estimate of the form

$$\|\sigma_h(t)\| = \mathcal{O}(h^q),$$

where $q \in \mathbb{N}$ is the order of the space discretization. We want, of course, a bound for the error $\|w_h(t) - w(t)\|$.

The analysis will be presented for linear systems,

$$F(t,v) = Av + g(t). \tag{2.3}$$

Some basic linear algebra concepts that will be used, with examples of vector and matrix norms, are listed in a next subsection. Here we shall use the following notations: for any $m \times m$ matrix B let

$$\|B\| = \max_{v \neq 0} \frac{\|Bv\|}{\|v\|}$$

stand for the induced matrix norm. The exponential function of a matrix is defined by the following series

$$e^B = I + B + \frac{1}{2}B^2 + \cdots + \frac{1}{k!}B^k + \cdots \quad ,$$

so that the solution of $w'(t) = Bw(t)$, $w(0) = v$ can be written as $w(t) = e^{tB}v$.

We can relate $\|w_h(t) - w(t)\|$ with $\|\sigma_h(t)\|$ if we make the following *stability assumption*

$$\|e^{tA}\| \leq Ke^{t\omega} \quad \text{for all} \quad t \geq 0 \tag{2.4}$$

with some "moderate" constants $K > 0$, $\omega \in \mathbb{R}$.

Theorem 2.1. Consider the linear system (2.1), (2.3) with stability assumption (2.4). Then

$$\|w_h(t) - w(t)\| \leq Ke^{\omega t}\|w_h(0) - w(0)\| + \frac{K}{\omega}(e^{\omega t} - 1) \max_{0 \leq s \leq t} \|\sigma_h(t)\|.$$

(Here we use the convention that $\frac{1}{\omega}(e^{\omega t} - 1) = t$ in case $\omega = 0$.)

Proof. The spatial error $\varepsilon(t) = w_h(t) - w(t)$ satisfies

$$\varepsilon'(t) = A\varepsilon(t) + \sigma_h(t), \qquad t \geq 0.$$

By the "variation of constants formula" we thus find

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

Hence

$$\|\varepsilon(t)\| \leq \|e^{tA}\|\, \|\varepsilon(0)\| + \int_0^t \|e^{(t-s)A}\|\, ds \max_{0 \leq s \leq t} \|\sigma_h(s)\|.$$

Using the stability assumption, the bound for the spatial error follows. □

Corollary 2.2. If (2.4) is valid and $w(0) = w_h(0)$, $\|\sigma_h(t)\| \leq Ch^q$ for $0 \leq t \leq T$, then

$$\|w_h(t) - w(t)\| \leq \frac{K}{\omega}(e^{\omega t} - 1)Ch^q \quad \text{for} \quad 0 \leq t \leq T.$$

□

In general, the term *stability* will be used to indicate that small perturbations give a small overall effect. This is just what we have in the above: the unperturbed system is $w'(t) = Aw(t) + g(t)$ with given $w(0)$, and w_h can be regarded as solution of the perturbed system $w_h'(t) = Aw_h(t) + g(t) + \sigma_h$, with perturbation σ_h and also a perturbation $w_h(0) - w(0)$ on the initial value.

The term "moderate" will be used in general to indicate something of order of magnitude 1, but this must be understood in an operational sense. For example, if we have perturbations with order of magnitude $\sim 10^{-6}$ and these perturbations are amplified with a factor $\sim 10^3$, then this factor might still be considered as "moderate enough" if one is only interested in 3 digits accuracy.

2.2. Example: 1-st order upwind discretization for advection operator

Consider the periodic advection equation $u_t + u_x = 0$ with first order upwind discretization and let $w_h(t) = (u(x_1,t),...,u(x_m,t))^T$. Then the i-th component of $\sigma_h(t)$ is

$$\sigma_{h,i}(t) = \frac{d}{dt}u(x_i,t) - \frac{1}{h}\Big(u(x_{i-1},t) - u(x_i,t)\Big) =$$

$$= -u_x(x_i,t) - \frac{1}{h}\Big(u(x_{i-1},t) - u(x_i,t)\Big) = -\frac{1}{2}hu_{xx}(x_i,t) + \mathcal{O}(h^2).$$

In the discrete L_2-norm $\|v\| = (h\sum_{i=1}^m |v_i|^2)^{1/2}$ we thus have

$$\|\sigma_h(t)\| \le \frac{1}{2}h \max_{0\le x\le 1}|u_{xx}(x,t)| + \mathcal{O}(h^2).$$

Note that if we consider the local truncation error $\tilde{\sigma}_h$ with respect to the modified equation $\tilde{u}_t + \tilde{u}_x = \frac{1}{2}h\tilde{u}_{xx}$, then we obtain $\|\tilde{\sigma}_h(t)\| = \mathcal{O}(h^2)$.

To apply Theorem 2.1 we have to verify the stability condition (2.4) for

$$A = \frac{1}{h}\begin{pmatrix} -1 & & & 1 \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}.$$

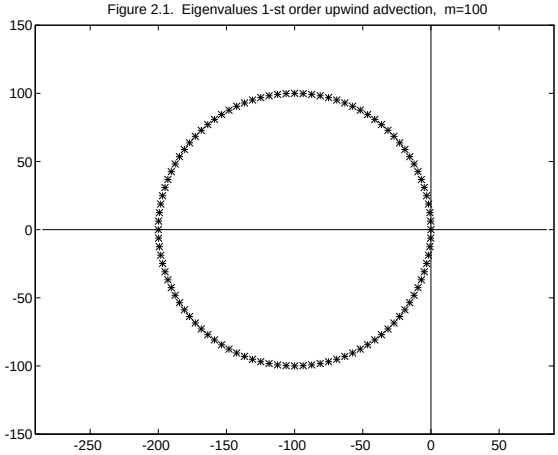
Consider the discrete Fourier modes $\phi_k = (e^{2\pi ikx_1},...,e^{2\pi ikx_m})^T$ for $1 \le k \le m$. We have for all components $j = 1,2,...,m$ (also for $j = 1$ due to periodicity)

$$(A\phi_k)_j = \frac{1}{h}\Big(e^{2\pi ikx_{j-1}} - e^{2\pi ikx_j}\Big) = \lambda_k e^{2\pi ikx_j} = \lambda_k(\phi_k)_j,$$

$$\lambda_k = \frac{1}{h}\Big(e^{-2\pi ikh} - 1\Big).$$

So, the discrete Fourier modes are the eigenvectors for A with eigenvalues λ_k in the left halve of the complex plane. In Figure 2.1 these are plotted for $m = 100$.

Figure 2.1. Eigenvalues 1-st order upwind advection, m=100



Further, any vector $v \in \mathbb{C}^m$ can be written as $v = \sum_{k=1}^m \alpha_k \phi_k \in \mathbb{C}^m$, and we have

$$\|v\|^2 = (v,v) = \sum_{k,l} \overline{\alpha_k} \alpha_l (\phi_k, \phi_l) = \sum_{k=1}^m |\alpha_k|^2$$

(discrete counterpart of Parseval's identity). So, consider $v'(t) = Av(t)$, $v(0) = \sum_{k=1}^m \alpha_k \phi_k$. Then

$$v(t) = \sum_{k=1}^m \alpha_k e^{t\lambda_k} \phi_k,$$

$$\|v(t)\|^2 = \sum_{k=1}^m |\alpha_k e^{\lambda_k t}|^2 \leq \sum_{k=1}^m |\alpha_k|^2 = \|v(0)\|^2.$$

For arbitrary $v(0) \in \mathbb{C}^m$ we thus have

$$\|v(t)\| = \|e^{tA} v(0)\| \leq \|v(0)\|,$$

which shows that $\|e^{tA}\| \leq 1$. Therefore we can apply Theorem 2.1 and Corollary 2.2 with $K = 1$, $\omega = 0$.

We note that in a somewhat more abstract setting the above can also be written as follows. Let $U = [\phi_1, \phi_2, ..., \phi_m] \in \mathbb{C}^{m \times m}$ and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_m)$. We have

$$A = U \Lambda U^{-1}$$

Therefore $e^{tA} = U e^{t\Lambda} U^{-1}$ and

$$\|e^{tA}\| \leq \|U\| \, \|e^{t\Lambda}\| \, \|U^{-1}\| = \max_{1 \leq k \leq m} |e^{t\lambda_k}| \, \mathrm{cond}(U) = 1,$$

since $\mathrm{Re}\lambda_k \leq 0$ with equality for $k = m$, and $\mathrm{cond}(U) = \|U\| \, \|U^{-1}\| = 1$, due to the fact that the discrete Fourier modes, which form the columns of U, are orthonormal. (Note that U iself is not unitary, but for $V = \sqrt{h} U$ we do have $V^* V = I$.)

So, in conclusion, we have shown that with the discrete L_2-norm:

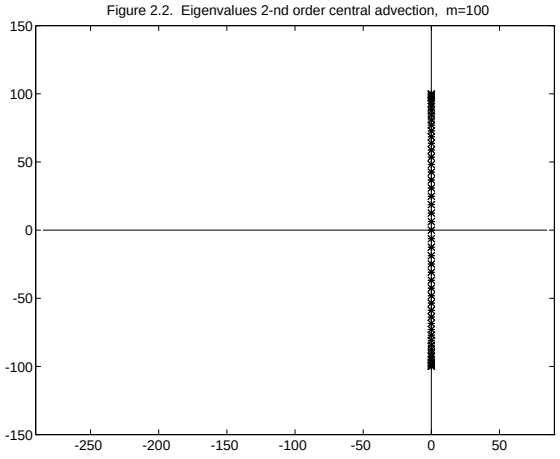
> The 1-st order upwind discretization (1.7) converges for $h \to 0$ with order 1 to the solution of $u_t + a u_x = 0$. Moreover, with respect to the modified equation $\tilde{u}_t + a \tilde{u}_x = \frac{1}{2} a h \tilde{u}_{xx}$ the order of convergence is 2.

2.3. Example: central discretizations for advection/diffusion operator

With the second order discretizations (1.9),(1.11) we can proceed similarly. Also for these instances the discrete Fourier modes are the eigenvectors for the discretized operator A, and all eigenvalues have nonpositive real part.

For (1.9) with $a = 1$, we get

$$A = \frac{1}{2h}\begin{pmatrix} 0 & -1 & & 1 \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & -1 \\ -1 & & 1 & 0 \end{pmatrix},$$



and by some calculations it is seen that $A\phi_k = \lambda_k\phi_k$ with the eigenvalues given by

$$\lambda_k = \frac{1}{2h}\Big(e^{-2\pi ikh} - e^{2\pi ikh}\Big) =$$

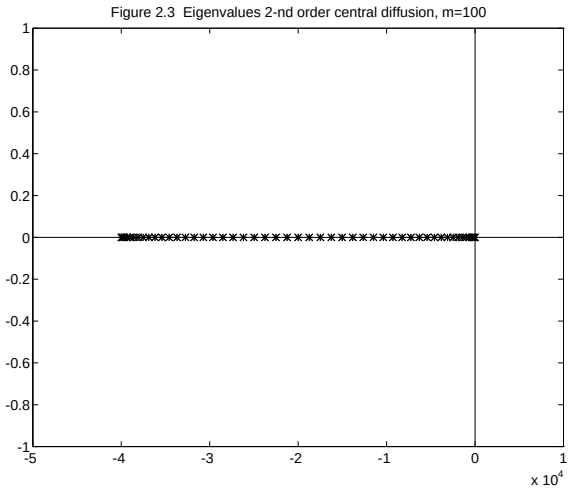
$$= -\frac{i}{h}\sin(2\pi kh),$$

see Figure 2.2. These are purely imaginary since A is skew-symmetric, that is, $A^T = -A$. The claims on convergence of this central advection discretization that were made in Section 1 can now be proven in the same way as for the 1-st order upwind discretization, by considering the space truncation error with respect to the advection equation $u_t + au_x = 0$ or the modified equation $\tilde{u}_t + a\tilde{u}_x = -\frac{1}{6}ah^2\tilde{u}_{xxx}$. We obtain convergence in the discrete L_2 norm with order 2 for the advection equation and order 4 for the modified equation.

In a similar way we can obtain stability and convergence results with the central discretization (1.11) for the diffusion equation.

Considering (1.11) with $d = 1$,

$$A = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & 1 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & -2 \end{pmatrix},$$



we have again $A\phi_k = \lambda_k\phi_k$, but now with real eigenvalues

$$\lambda_k = \frac{1}{h^2}\Big(e^{-2\pi ikh} - 2 + e^{2\pi ikh}\Big) =$$

$$= \frac{2}{h^2}\Big(\cos(2\pi kh) - 1\Big) = \frac{-4}{h^2}\sin^2(\pi kh),$$

see Figure 2.3.

Results for combined advection-diffusion $u_t + au_x = du_{xx}$ follow in the same way. With second order central differences the eigenvalues now become

$$\lambda_k = \frac{2d}{h^2}\Big(\cos(2\pi kh) - 1\Big) - \frac{ia}{h}\sin(2\pi kh).$$

These are on an ellipse in the left half plane. Note also that the eigenvalues are in a wedge

$\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \le \alpha\}$ with angle α such that $\tan\alpha \approx a/(2\pi d)$ (determined by $\lambda_1 \approx -2\pi^2 d - 2\pi i a$). With first order upwind advection discretization we obtain the same formula for the eigenvalues, except that then d should be replaced by $d + \frac{1}{2}|a|h$.

Remark. The above examples might give the impression that stability somehow follows automatically for consistent discretizations. This is not so. Consider, for instance, $u_t + u_x = 0$ with the "down-stream" discretization $w_i'(t) = \frac{1}{h}(w_i(t) - w_{i+1}(t))$. This has also a first order truncation error, but it has no practical significance due to the fact that it is not stable. The eigenvalues will be in the right half-plane, similar to Figure 2.1 but reflected around the imaginary axis, and therefore condition (2.4) only holds with $\omega = 2/h$. With this ω we have

$$\frac{1}{\omega}(e^{t\omega} - 1) \to \infty \quad \text{for} \quad h \to 0.$$

In fact, this discretization is only a reasonable one with respect to the truncation error. With $u_t + u_x = 0$ the time evolution at a point x_i is determined by what happens to the left of x_i. With the above down-stream discretization the evolution of $w_i(t)$ is determined by what happens to the right of x_i. So, the semi-discrete system gets its information from the wrong direction. We note that the instability of this system also follows from the classical paper of Courant, Friedrichs & Lewy (1928), the first paper where stability of difference schemes was discussed. ◇

2.4. Some linear algebra concepts

Here some basic properties are listed that are used throughout these notes. A good reference for linear algebra is the book of Horn & Johnson (1985). More advanced results can be found in Horn & Johnson (1991). The topic of numerical linear algebra will not be treated. The standard text book in this field is Golub & van Loan (1996).

Consider the vector spaces $\mathbb{R}^m$ and $\mathbb{C}^m$ and let $h = 1/m$. Some vector norms used in these notes are the discrete L_p-norms, with $p = 1, 2$ or ∞,

$$\|v\|_2 = \Big(h\sum_{j=1}^{m} |v_j|^2\Big)^{1/2}, \qquad \|v\|_1 = h\sum_{j=1}^{m} |v_j|, \qquad \|v\|_\infty = \max_{1\le j\le m} |v_j|, \tag{2.5}$$

for $v = (v_1, v_2, ..., v_m)^T$. The L_2-norm is generated by the inner product

$$(u, v)_2 = h\sum_j \bar{u}_j v_j. \tag{2.6}$$

Given a vector norm, the induced matrix norm for $m \times m$ matrices B is defined as

$$\|B\| = \max_{v\neq 0} \frac{\|Bv\|}{\|v\|}. \tag{2.7}$$

We have $\|AB\| \le \|A\|\|B\|$ for any two A and B in $\mathbb{R}^{m\times m}$ or $\mathbb{C}^{m\times m}$.

Further, if $B = (b_{jk})$ then $B^* = (\overline{b_{kj}})$ denotes the Hermitian adjoint. If B is real this is the same as the transpose B^T. The set of eigenvalues of B, denoted by $\sigma(B)$, is called the

spectrum of B. The spectral radius of B, $\max\{|\lambda| : \lambda \in \sigma(B)\}$, is denoted by $\rho(B)$, and we always have $\rho(B) \le \|B\|$. Some examples of induced matrix norms are

$$\|B\|_2 = \sqrt{\rho(B^*B)}, \qquad \|B\|_1 = \max_{1\le k\le m} \sum_{j=1}^{m} |b_{jk}|, \qquad \|B\|_\infty = \max_{1\le j\le m} \sum_{k=1}^{m} |b_{jk}|. \tag{2.8}$$

As for vectors, we also have for matrices the Hölder inequality $\|B\|_2 \le \sqrt{\|B\|_1 \|B\|_\infty}$. This follows from

$$\|B\|_2^2 = \rho(B^*B) \le \|B^*\|_\infty \|B\|_\infty = \|B\|_1 \|B\|_\infty.$$

A vector norm is called *monotone* if $\|u\| = \|v\|$ for any two vectors whose components have equal modulus, $|u_i| = |v_i|$ for all i. This is equivalent with the property

$$\|\Lambda\| = \max_j |\lambda_j| \qquad \text{for any diagonal matrix} \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_m).$$

The above L_p-norms are monotone. If we consider arbitrary inner products $(u, v) = u^*Gv$ with $G = H^*H$ and H nonsingular, that is G positive definite, then the corresponding norm $\|v\| = \sqrt{(v, v)}$ is only monotone if G is diagonal.

The matrix B is *unitary* if $B^*B = I$, that is $B^{-1} = B^*$. This implies that $\|Bv\|_2 = \|v\|_2 = \|B^{-1}v\|_2$ for any vector v, and in particular

$$\|B\|_2 = 1, \qquad \mathrm{cond}_2(B) = \|B\|_2 \|B^{-1}\|_2 = 1.$$

The matrix B is said to be *normal* if $BB^* = B^*B$. A normal matrix has a complete set of orthogonal eigenvectors (see Horn & Johnson (1985, Sect.2.5)), and it can be decomposed as

$$B = U\Lambda U^{-1}$$

with unitary U and diagonal $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_m)$. Note that the columns of U are the eigenvectors of B, that is, $Bu_j = \lambda_j u_j$ if $U = [u_1, u_2, ..., u_m]$.

Examples of normal matrices are the unitary (orthogonal) matrices $B^*B = I$, the Hermitian (symmetric) matrices $B^* = B$, and the skew-Hermitian (skew-symmetric) matrices $B^* = -B$. The eigenvalues of a unitary matrix are all on the unit circle, the eigenvalues of an Hermitian matrix are all real and those of a skew-Hermitian matrix are all purely complex.

If $P(z) = p_0 + p_1 z + \cdots + p_s z^s$ is a polynomial and A an $m \times m$ matrix, we define

$$P(A) = p_0 I + p_1 A + \cdots + p_s A^s.$$

The eigenvalues of $P(A)$ are $P(\lambda)$, with λ eigenvalue of A. For a rational function $R(z) = P(z)/Q(z)$ we define $R(A) = P(A)[Q(A)]^{-1}$, provided R is analytic on the spectrum of A. The exponential function of a matrix is defined by the series

$$e^A = I + A + \frac{1}{2}A^2 + \cdots + \frac{1}{k!}A^k + \cdots.$$

If $A = U\Lambda U^{-1}$ with $\Lambda = \mathrm{diag}(\lambda_j)$ and φ is a rational or exponential function, we have

$$\varphi(A) = U\varphi(\Lambda)U^{-1}, \qquad \varphi(\Lambda) = \mathrm{diag}(\varphi(\lambda_1), \varphi(\lambda_2), ..., \varphi(\lambda_m)). \tag{2.9}$$

In case A is normal it thus holds that

$$\|\varphi(A)\|_2 = \max_{1\le j\le m} |\varphi(\lambda_j)|. \tag{2.10}$$

The stability condition (2.4) was easy to verify in the examples of the previous subsections, due to the fact that we did consider problems that are linear with constant coefficients and without boundary conditions. This leads to a normal matrix A for which it easy to obtain results in the L_2-norm.

A useful concept for stability results with non-normal matrices is the *logarithmic norm*, defined as

$$\mu[A] = \lim_{\tau \downarrow 0} \frac{\|I + \tau A\| - 1}{\tau}. \tag{2.11}$$

For $\tau > 0$ the difference ratio on he right hand side side is easily seen to be in the interval $[-\|A\|, \|A\|]$, and it is monotonically nondecreasing in τ: if $0 < \theta < 1$ then

$$\frac{1}{\theta\tau}\Big(\|I + \theta\tau A\| - 1\Big) \leq \frac{1}{\theta\tau}\Big(\|\theta I + \theta\tau A\| + |1 - \theta| - 1\Big) = \frac{1}{\tau}\Big(\|I + \tau A\| - 1\Big).$$

Hence the limit in (2.11) exists. Note that the logarithmic norm is not a matrix norm; it can be negative. The importance of this logarithmic norm lies in the following result.

Theorem 2.3. Let $A \in \mathbb{R}^{m\times m}$. We have

$$\mu[A] \leq \omega \quad \text{iff} \quad \|e^{tA}\| \leq e^{t\omega} \ \text{ for all } t \geq 0.$$

Proof. Suppose that $\mu[A] \leq \omega$. Then

$$\|I + \tau A\| \leq 1 + \omega\tau + o(\tau), \quad \tau \downarrow 0,$$

and consequently

$$\|(I + \tau A)^n\| \leq (1 + \omega\tau + o(\tau))^n \to e^{t\omega} \quad \text{as } \tau \downarrow 0,\ t = n\tau \text{ fixed}.$$

Since

$$e^{tA} = \lim_{\tau\downarrow 0}(I + \tau A)^n \quad \text{as } \tau \downarrow 0,\ t = n\tau \text{ fixed},$$

it follows that $\|\exp(tA)\| \leq e^{t\omega}$.

On the other hand, suppose that $\|e^{tA}\| \leq e^{t\omega}$ for all $t > 0$. Since $I + \tau A = e^{\tau A} + \mathcal{O}(\tau^2)$ it follows that

$$\|I + \tau A\| \leq 1 + \tau\omega + \mathcal{O}(\tau^2) \quad \text{for } \tau \downarrow 0,$$

and hence $\mu[A] \leq \omega$. □

For the L_p vector norms, the corresponding logarithmic norm of a real matrix A is given by

$$\mu_2[A] = \max_{v\neq 0} \frac{(Av, v)_2}{(v, v)_2} = \max\{\lambda : \lambda \ \text{ eigenvalue of } \ \frac{1}{2}(A + A^T)\}, \tag{2.12}$$

$$\mu_1[A] = \max_j \Big(a_{jj} + \sum_{i\neq j} |a_{ij}|\Big), \qquad \mu_\infty[A] = \max_i \Big(a_{ii} + \sum_{j\neq i} |a_{ij}|\Big). \tag{2.13}$$

These expressions can be derived directly from the formulas for the matrix norms in the definition of $\mu[A]$. In particular, we have $\mu_2[A] \le 0$ iff $(v, Av)_2 \le 0$ for all $v \in \mathbb{R}^m$. If the diagonal elements of A are negative, we have $\mu_\infty[A] \le 0$ whenever A is row-wise diagonally dominant, and $\mu_1[A] \le 0$ when A is column-wise diagonally dominant.

Further properties of the logarithmic norms can be found in Coppel (1965) or Dekker & Verwer (1984). We note that the concept of logarithmic norms was introduced in 1958 by G. Dahlquist and S.M. Lozinskij, see loc.cit.

In the following some examples (formulated as exercises) are given where the logarithmic norm can be used to prove stability.

Example. Consider $u_t + (a(x)u)_x = 0$ for $0 \le x \le 1$ with periodicity in x and with smooth periodic velocity $a(x)$. Show that $\int_0^1 u(x,t)dx$ is constant in t.

Consider the space discretization, on uniform grid $x_j = jh$,

$$w_j' = \frac{1}{h}\Big(a_{j-1/2}w_{j-1/2} - a_{j+1/2}w_{j+1/2}\Big),$$

with $a_{j+1/2} = a(\frac{1}{2}x_j + \frac{1}{2}x_{j+1})$ and with $w_{j+1/2} = \frac{1}{2}w_j + \frac{1}{2}w_{j+1}$ (central). Let

$$\omega = \frac{1}{2h}\max_j(a_{j+1/2} - a_{j-1/2}) = \mathcal{O}(1).$$

Determine the truncation error of the discretization. Write the ODE system as $w'(t) = Aw(t)$ and show that $(Av, v)_2 \le \omega(v, v)_2$ for all vectors v (hint: A is a skew-symmetric matrix plus a diagonal matrix). Consequently, $\mu_2[A] \le \omega$.

Suppose that $a(x) \ge 0$ and consider $w_{j+1/2} = w_j$ (upwind). Show that now $\mu_1[A] \le 0$ and $\mu_\infty[A] \le 2\omega$. Using the Hölder inequality for matrices it now follows that $\|e^{tA}\|_2 \le e^{t\omega}$. $\diamond$

Example. Consider $u_t + a(x)u_x = 0$ with periodicity as above. Show that $u(\xi(t), t)$ is constant along the characteristics $(\xi(t), t)$ in the (x,t)-plane, defined by $\xi'(t) = a(\xi(t))$.

Consider here discretizations

$$w_j' = \frac{1}{h}a(x_j)\Big(w_{j-1/2} - w_{j+1/2}\Big),$$

with $w_{j+1/2}$ as above, either central or upwind. Show that we have consistency and stability in the L_2-norm, provided $\frac{1}{h}\max_j |a(x_j) - a(x_{j+1})| = \mathcal{O}(1)$. With upwind we now have $\mu_\infty[A] \le 0$. $\diamond$

Example. Consider $u_t = (d(x)u_x)_x + g(x,t)$ with $d(x) > 0$, again with periodicity as above. Determine consistency for

$$w_j' = \frac{1}{h^2}\Big(d_{j-1/2}(w_{j-1} - w_j) - d_{j+1/2}(w_j - w_{j+1})\Big) + g(x_j, t)$$

with $d_{j+1/2} = d(\frac{1}{2}x_j + \frac{1}{2}x_{j+1})$. Show that $\mu_1[A] \le 0$, $\mu_\infty[A] \le 0$. Using the Hölder inequality for matrices, applied to e^{tA}, it follows that also $\mu_2[A] \le 0$. $\diamond$

3. Time discretizations: MOL and von Neumann stability

Suppose our PDE, with solution $u(x,t)$, has been discretized in space, resulting in the semi-discrete system (of ODEs)

$$w'(t) = F(t, w(t))$$

with $w(t) = (w_i(t))_{i=1}^m \in \mathbb{R}^m$, m being proportional to the number of grid points in space. Fully discrete approximations $w_i^n \approx u(x_i, t_n)$ can now be obtained by applying some suitable ODE method with step size τ for the time levels $t_n = n\tau$. In the following we use $w_n = (w_i^n)_{i=1}^m$ to denote the vector (grid function) containing the discrete numerical solution.

The approach of considering space and time discretizations separately is called the *method of lines* (MOL). This is not a "method" in the numerical sense, it is a way to construct and analyze certain numerical methods. A typical MOL reasoning goes as follows: if we know that $\|w(t) - w_h(t)\| \le Ch^q$ for our space discretization and the ODE theory tells us that $\|w(t_n) - w_n\| \le C\tau^p$, then we have an error bound for the fully discrete approximations

$$\|w_h(t_n) - w_n\| \le C\tau^p + Ch^q.$$

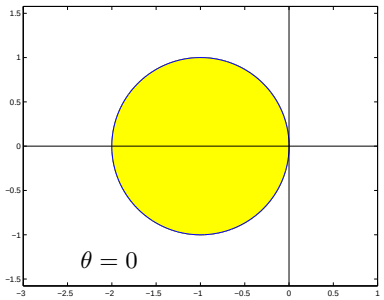
For the error bound in time we need of course to verify consistency and stability of our numerical ODE method. The stability considerations in the MOL literature are usually based on the *stability functions* and *stability regions* of the ODE methods. On the other hand, in the traditional PDE literature one usually sees stability considerations based on Fourier decomposition, the so-called *von Neumann analysis*.

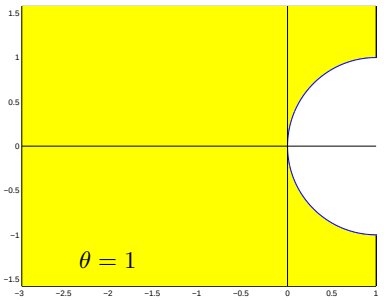
In this section we shall consider these concepts. For the ODE method we consider, as an example, the *θ-method*

$$w_{n+1} = w_n + \tau(1-\theta)F(t_n, w_n) + \tau\theta F(t_{n+1}, w_{n+1}) \tag{3.1}$$

with as special cases the explicit (forward) Euler method ($\theta = 0$), the trapezoidal rule ($\theta = \frac{1}{2}$) and the implicit (backward) Euler method ($\theta = 1$). As we shall see, the order is $p = 2$ if $\theta = \frac{1}{2}$, and $p = 1$ otherwise. Application of the method to the scalar, complex test equation $w'(t) = \lambda w(t)$ gives approximations

$$w_{n+1} = R(\tau\lambda)w_n, \qquad R(z) = \frac{1 + (1-\theta)z}{1 - \theta z}.$$





This R is the *stability function* of the method. Near $z = 0$ we have $R(z) = 1 + z + \theta z^2 + \mathcal{O}(z^3)$. The *stability region* is the set

$$\mathcal{S} = \{z \in \mathbb{C} : |R(z)| \leq 1\}$$

in the complex plane. An ODE method that has the property that $\mathcal{S}$ contains the left half-plane $\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re} z \leq 0\}$ is called *A-stable*. The θ-method is A-stable for $\theta \geq \frac{1}{2}$.

Examples of more general ODE methods, together with some basic properties, are given in the appendices. Here we review some of these ODE concepts by means of the θ-method.

3.1. Convergence of ODE methods

Inserting the ODE solution $w(t)$ into (3.1) gives

$$w(t_{n+1}) = w(t_n) + \tau(1-\theta)w'(t_n) + \tau\theta w'(t_{n+1}) + \tau\rho_n \tag{3.2}$$

with *truncation error* ρ_n. By Taylor expansion around $t = t_n$ it follows that

$$\rho_n = \frac{1}{2}(1-2\theta)\tau w''(t_n) + \frac{1}{6}(1-3\theta)\tau^2 w'''(t_n) + \mathcal{O}(\tau^4).$$

Thus, for $w(t)$ sufficiently smooth, we get $\|\rho_n\| = \mathcal{O}(\tau)$ if $\theta \neq \frac{1}{2}$, and $\|\rho_n\| = \mathcal{O}(\tau^2)$ if $\theta = \frac{1}{2}$.

For further analysis, assume as before that the problem is linear, $F(t, v) = Av + g(t)$. Let $\varepsilon_n = w(t_n) - w_n$, $n \geq 0$, stand for the *global discretization error*. We want to find an upper bound for $\|\varepsilon_n\|$. Subtraction of (3.1) from (3.2) leads to the recursion

$$\varepsilon_{n+1} = \varepsilon_n + (1-\theta)\tau A\varepsilon_n + \theta\tau A\varepsilon_{n+1} + \tau\rho_n$$

for $n \geq 0$. It follows that

$$\varepsilon_{n+1} = R(\tau A)\varepsilon_n + \delta_n \quad (n \geq 0), \qquad \varepsilon_0 = w(0) - w_0, \tag{3.3}$$

with

$$R(\tau A) = (I - \theta\tau A)^{-1}(I + (1-\theta)\tau A), \qquad \delta_n = (I - \theta\tau A)^{-1}\tau\rho_n.$$

We see from (3.3) that the matrix $R(\tau A)$ determines how an error already present at time level t_n is propagated to the next time level. On the other hand, during this time step also a new error δ_n is introduced. This δ_n is the *local discretization error*. The ODE method is said to be *consistent of order* p if $\|\delta_n\| = \mathcal{O}(\tau^{p+1})$ whenever the exact solution is sufficiently smooth. Note that we do have $\|\delta_n\| \leq C\tau\|\rho_n\|$ provided that

$$\|(I - \theta\tau A)^{-1}\| \leq C.$$

The existence of this inverse simply means that the implicit relation in the θ-method has a unique solution, and a bound on the inverse will hold if we can bound $\|R(\tau A)\|$, since $(I - \theta\tau A)^{-1} = \theta R(\tau A) + (1-\theta)I$. We then find $p = 2$ if $\theta = \frac{1}{2}$, and $p = 1$ for the other values of θ.

To relate the local discretization errors to the global errors we need stability. Assume

$$\|R(\tau A)^n\| \leq K \quad \text{for } n \geq 0,\ n\tau \leq T. \tag{3.4}$$

Theorem 3.1. The stability assumption (3.4) implies

$$\|w(t_n) - w_n\| \le K\|w(t_0) - w_0\| + K\sum_{j=0}^{n-1} \|\delta_j\| \qquad \text{for } n\tau \le T.$$

Proof. Elaboration of the error recursion (3.3) gives

$$\varepsilon_n = R(\tau A)^n \varepsilon_0 + R(\tau A)^{n-1}\delta_0 + \cdots + R(\tau A)\delta_{n-2} + \delta_{n-1},$$

from which the result directly follows. □

So, with this theorem, if $\|\delta_j\| \le C\tau^{p+1}$ for all j and $w_0 = w(0)$, we obtain the global error bound

$$\|w(t_n) - w_n\| \le C'\tau^p \quad \text{for } n\tau \le T,$$

with constant $C' = KTC$, and thus we have *convergence of order p*. Obviously, stability is the crucial point here.

Now, if we consider a *fixed* matrix A, then

$$R(\tau A) = I + \tau A + \mathcal{O}(\tau^2), \quad \tau \downarrow 0,$$

and hence

$$\|R(\tau A)^n\| \le \Big(1 + \tau\|A\| + \mathcal{O}(\tau^2)\Big)^n \le e^{2t_n\|A\|} \qquad \text{for } n\tau \le T,$$

provided $\tau > 0$ is sufficiently small (in fact, $\tau\|A\|$ should be sufficiently small). Thus for fixed, bounded A we will have stability.

However, if A results from spatial discretization of a PDE problem, it will contain negative powers of the mesh width h and its dimension will also grow with decreasing h. The stability assumption (3.4) then must be carefully examined, since we want estimates that hold *uniformly* in h. To do this, we can consider the homogeneous equation $w'(t) = Aw(t)$ and prove that $\|w_n\| \le K\|w_0\|$ for arbitrary w_0, with K independent of h. In the remainder of this section this will be worked out for some examples.

Note. Theorem 3.1 can be viewed as a time-discrete version of Theorem 2.1. Both results essentially state that

$$\text{consistency \& stability} \quad \Longrightarrow \quad \text{convergence.}$$

Within a certain technical framework, the reverse also holds. This "iff" result is known as the *Lax equivalence theorem*, see Richtmyer & Morton (1967).

3.2. Example: explicit Euler for the diffusion problem

Consider the diffusion test problem $u_t = u_{xx}$ with periodicity condition at $x = 0, 1$. The standard semi-discrete system is (see Section 1)

$$w_j'(t) = \frac{1}{h^2}\Big(w_{j-1}(t) - 2w_j(t) + w_{j+1}(t)\Big), \quad j = 1, 2, ..., m,$$

with $w_0 \equiv w_m$ and $w_{m+1} \equiv w_1$. Application of the explicit Euler method now gives the fully discrete scheme

$$w_j^{n+1} = w_j^n + \frac{\tau}{h^2}\Big(w_{j-1}^n - 2w_j^n + w_{j+1}^n\Big). \tag{3.5}$$

To study stability, we can proceed as in the previous sections by inserting discrete Fourier modes. Thus we put $w_j^0 = (\phi_k)_j = e^{2\pi i k x_j}$ and we make the "Ansatz" $w_j^{n+1} = r w_j^n$, that is $w_j^n = r^n e^{2\pi i k x_j}$ for $n \geq 0$. Insertion in (3.5) yields

$$r^{n+1} e^{2\pi i k x_j} = r^n e^{2\pi i k x_j}\Big(1 + \frac{\tau}{h^2}(e^{-2\pi i k h} - 2 + e^{2\pi i k h})\Big).$$

Thus we find the *amplification factor* for the k-th Fourier mode

$$r = r_k = 1 + \frac{\tau}{h^2}(e^{-2\pi i k h} - 2 + e^{2\pi i k h}) = 1 - \frac{4\tau}{h^2}\sin^2(\pi h k).$$

The von Neumann criterion for stability is

$$|r_k| \leq 1 \quad \text{for all} \quad k = 1, 2, ..., m,$$

which is fulfilled here if

$$\frac{\tau}{h^2} \leq \frac{1}{2}. \tag{3.6}$$

If this holds then the numerical solution will stay bounded, because if we consider an arbitrary starting vector $w_0 = \sum_k \alpha_k \phi_k \in \mathbb{C}^m$ then $w_n = \sum_k \alpha_k (r_k)^n \phi_k$ (superposition of results for individual Fourier modes), and thus

$$\|w_n\|^2 = \sum^k |\alpha_k|^2 |r_k|^{2n} \leq \sum^k |\alpha_k|^2 = \|w_0\|^2$$

in the discrete L_2-norm.

In general, stability means that perturbations are not amplified too much. For example, if we would start with a perturbed initial value $\tilde{w}_0$, we want the difference $\|\tilde{w}_n - w_n\|$ to be bounded by $C\|\tilde{w}_0 - w_0\|$, with a moderate constant $C > 0$. For linear problems we can simply look at boundedness of solutions v_n of the homogeneous equation (that is, without source terms), since the difference $\tilde{w}_n - w_n$ will satisfy the same recursion as v_n.

The MOL approach would lead to the equivalent, but conceptually different reasoning: our semi-discrete system can be written as $w'(t) = Aw(t)$ with A given by (1.11). We know from Section 2 that

$$A = U\Lambda U^{-1}, \quad \Lambda = \mathrm{diag}(\lambda_k), \quad \mathrm{cond}(U) = 1.$$

Application of the explicit Euler method to this system of ODEs gives $w_{n+1} = R(\tau A)w_n$. Hence

$$w_n = R(\tau A)^n w_0,$$

and thus to be sure that w_n stays bounded we need to know whether this holds for $R(\tau A)^n$. We have

$$R(\tau A)^n = U R(\tau\Lambda)^n U^{-1}, \quad R(\tau\Lambda)^n = \mathrm{diag}(R(\tau\lambda_k)^n).$$

It follows that, in the L_2-norm

$$\|R(\tau A)^n\| = \max_{1\le k\le m} |R(\tau\lambda_k)^n|,$$

and thus we will have stability provided that

$$\tau\lambda_k \in \mathcal{S} \quad \text{for all} \quad k.$$

In this example the above eigenvalue criterion is the same as the von Neumann criterion, since

$$r_k = R(\tau\lambda_k), \quad \lambda_k = -\frac{4}{h^2}\sin^2(\pi hk).$$

It is important to note that in the present example the eigenvalue criterion is sound because the matrix A is normal (orthogonal basis of eigenvectors, namely the discrete Fourier modes).

For the θ-method with $\theta > 0$ we can proceed in a similar way. If $\theta < 1/2$ the ratio τ/h^2 must be bounded to achieve stability. The precise bound is given in Table 3.1.

3.3. Step size restrictions for advection/diffusion

For the other examples considered thus far the matrix A was also normal, due to the fact that we consider problems with constant coefficients and no boundary conditions. So, with arbitrary stability function R we have, as for the Euler scheme,

$$R(\tau A)^n = UR(\tau\Lambda)^n U^{-1}, \quad R(\tau\Lambda)^n = \mathrm{diag}(R(\tau\lambda_k)^n),$$

and thus we have in the L_2-norm

$$\|R(\tau A)^n\| = \max_{1\le k\le m} |R(\tau\lambda_k)^n|.$$

So, to verify stability we look at the *eigenvalue criterion*

$$\tau\lambda_k \in \mathcal{S} \quad \text{for all} \quad k. \tag{3.7}$$

Direct insertion of the Fourier modes in the scheme would lead to growth factors $r_k = R(\tau\lambda_k)$, and thus the von Neumann analysis leads to the same result.

Combination of the pictures for the eigenvalues of A given in Section 2 with pictures of the stability regions of the θ-methods ($\mathcal{S} = \{z \in \mathbb{C} : |z + \alpha| \le \alpha\}$ with $\alpha = 1/(1-2\theta)$ for $\theta < 1/2$) directly leads to the following conditions on the step size that have to be satisfied for stability.

	$\theta < \frac{1}{2}$	$\theta \ge \frac{1}{2}$
upwind advection (1.7)	$a\tau/h \le 1/(1-2\theta)$	$\tau \le \infty$
central advection (1.9)	$a\tau/h \le 0$	$\tau \le \infty$
central diffusion (1.11)	$d\tau/h^2 \le 1/(2-4\theta)$	$\tau \le \infty$

Table 3.1. Von Neumann conditions for stability with θ-methods.

For the schemes with $\theta \geq \frac{1}{2}$ there are no step size restrictions (*unconditional stability*), due to the fact that these θ-methods are A-stable.

Although the von Neumann stability analysis can be applied, in strict mathematical sense, only to a very restricted class of problems (no boundary conditions, constant coefficients), in practice it often gives a good criterion for much more general problems. In this section we have only considered the θ-methods for time integration. Of course, many more methods are available, either of the Runge-Kutta type or linear multistep type. Although it is not the intention to go deeply into the choice of particular ODE methods here, a few comments are in order.

Explicit ODE methods always have a bounded stability domain. Application to an advection equation will lead to a stability condition of the form

$$\frac{a\tau}{h} \leq C,$$

a so-called CFL-restriction (after Courant-Friedrichs-Lewy), where C depends on the particular method and space discretization. If the space discretization is central then the eigenvalues will be on the imaginary axis, and the ODE method should be selected such that a portion of the imaginary axis is contained in the stability region.

Application of an explicit ODE method to a diffusion equation will give rise to a stability condition

$$\frac{d\tau}{h^2} \leq C,$$

with again C determined by the method and space discretization. Since solutions of diffusion problems often give rise to rather smooth solutions, this time step restriction makes explicit methods unattractive for such problems. With implicit methods we can avoid such restrictions.

As a rule, with exceptions, explicit methods are more efficient for advection dominated problems than implicit methods. For problems with significant diffusion the implicit methods are in general to be preferred. In the appendices some ODE methods and stability restrictions are listed.

Remark. If semi-discretization leads to an ODE system $w'(t) = Aw(t) + g(t)$ with A not normal, then straightforward application of the eigenvalue criterion might still seem to be possible, but in such a situation this may lead to *wrong* conclusions.

A notorious example is given by the 1-st order upwind discretization of the initial-boundary value problem

$$u_t + u_x = 0, \quad u(0,t) = 0,$$

leading to

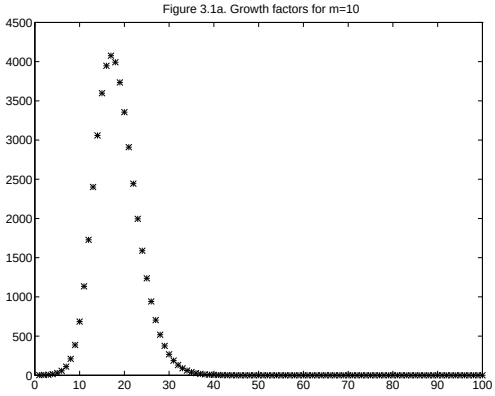
$$w_j'(t) = \frac{1}{h}\Big(w_{j-1}(t) - w_j(t)\Big), \quad j = 1, 2, ..., m,$$

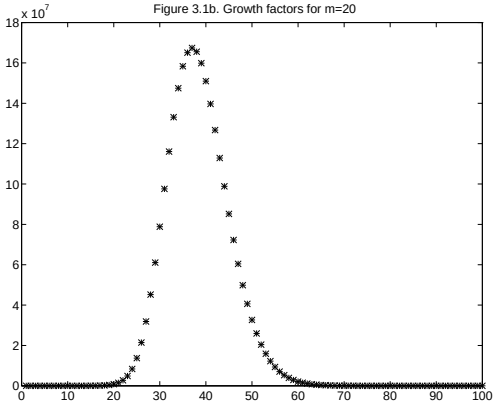
with $h = 1/m$ and $w_0(t) = 0$ (inflow boundary condition). In vector form we have $w'(t) = Aw(t)$ with

$$A = \frac{1}{h}\begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix}.$$

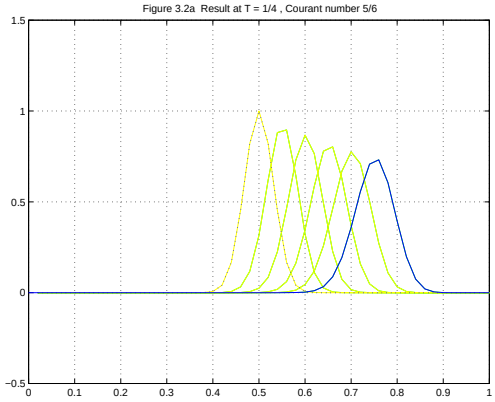
This matrix has only one eigenvalue, namely $\lambda = -1/h$. So, with the explicit Euler method we have $\tau\lambda \in \mathcal{S}$ iff $\tau/h \leq 2$. On the other hand, the von Neumann stability condition (ignoring boundaries) reads $\tau/h \leq 1$, and this is the *correct* condition.

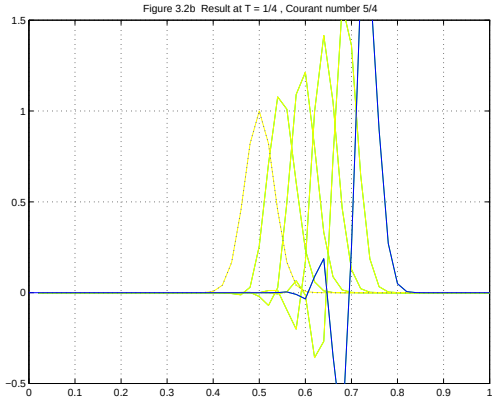
For example, in the Figures 3.1 the L_2-norm $\|R(\tau A)^n\|$ is plotted versus n, for $m = 10, 20$ with $\tau/h = 3/2$. Clearly, with this Courant number $\tau/h = 3/2$ the scheme is not stable (with moderate constants). Although we see that $\|R(\tau A)^n\| \to 0$ for $n \to \infty$ due to the fact that the eigenvalues are damped, before this happens $\|R(\tau A)^n\|$ can become very large, leading to an unacceptable error propagation.





A further illustration is given in the Figures 3.2 with time integration from $t = 0$ till $t = T = \frac{1}{4}$, initial condition $u(x, 0) = \sin(\pi x)^{100}$ and $h = 1/50$. In Figure 4.1a the numerical results are shown for the above scheme with $\tau = 1/60$ (15 time steps, Courant number 5/6) and in Figure 3.2b with $\tau = 1/40$ (10 time steps, Courant number 5/4).





We shall not pursue this matter here. The message simply is that the eigenvalue criterion (3.7) should be handled with great care if the matrix is not normal. For a thorough discussion

we refer to Morton (1980), Dorsselear et al. (1993). Further stability results can also be found in Strikwerda (1989), Thomée (1990). ◇

Remark. In some instances stability results in inner product norms are easy to obtain using the logarithmic norms. Let $\|u\| = \sqrt{(u,u)}$ with $(\cdot,\cdot)$ an inner product on $\mathbb{R}^m$, and let $D(r) = \{\zeta \in \mathbb{C} : |\zeta + r| \leq r\}$. Then the following holds,

$$\|A + \alpha I\| \leq \alpha, \quad D(r) \in \mathcal{S} \quad \Longrightarrow \quad \|R(\tau A)\| \leq 1 \ \text{ for } \ \tau\alpha \leq r.$$

In the limit α to ∞, this leads to

$$\mu[A] \leq 0, \quad \mathbb{C}^- \subset \mathcal{S} \quad \Longrightarrow \quad \|R(\tau A)\| \leq 1 \ \text{ for all } \ \tau > 0.$$

These results are based on a theorem of J. von Neumann from 1941, which states that $\|\varphi(B)\| \leq \max\{|\varphi(z)| : z \in \mathbb{C}, |z| \leq 1\}$ if $\|B\| \leq 1$ and φ is analytic on the unit disc. Indeed, this is the same von Neumann as in the "von Neumann analysis", but this analysis refers to Fourier decompositions, whereas in the above results non-normal matrices are allowed.

The last result tells us that with A-stable methods there will be unconditional stability provided that $\mu[A] \leq 0$. A very elegant proof of this statement, due to M. Crouzeix, can be found in Hairer & Wanner (1991). In that book a similar stability result can be found for multi-step methods, due to O. Nevanlinna, where R is the companion matrix. ◇

3.4. Simultaneous space-time discretizations

The MOL approach, where space and time discretizations are considered separately, is conceptually simple and flexible. However, sometimes it is better to consider space and time errors simultaneously: there may be cancellation of the various error terms.

Example. Consider once more the explicit Euler, 1-st order upwind discretization for the advection test equation $u_t + au_x = 0$, with $a > 0$, given initial profile and periodicity condition at $x = 0, 1$,

$$w_j^{n+1} = w_j^n + \frac{a\tau}{h}\Big(w_{j-1}^n - w_j^n\Big), \quad j = 1, 2, ..., m, \tag{3.8}$$

and with $a\tau/h \leq 1$ for stability. This scheme is also known as the Courant-Isaacson-Rees scheme. If we insert the exact PDE solution into this difference scheme we get

$$u(x_j, t_{n+1}) = u(x_j, t_n) + \frac{a\tau}{h}\Big(u(x_{j-1}, t_n) - u(x_j, t_n)\Big) + \tau\rho_j^n, \quad j = 1, 2, ..., m,$$

with a (residual) local truncation error

$$\rho_j^n = \Big[(u_t + \frac{1}{2}\tau u_{tt} + \cdots) + a(u_x - \frac{1}{2}hu_{xx} + \cdots)\Big](x_j, t_n) =$$

$$= -\frac{1}{2}ah\Big(1 - \frac{a\tau}{h}\Big)u_{xx}(x_j, t_n) + \mathcal{O}(h^2).$$

We have

$$w_h(t_{n+1}) - w_{n+1} = R(\tau A)(w_h(t_n) - w_n) + \tau\rho_n$$

where $\rho_n = (\rho_1^n, ..., \rho_m^n)^T$ is the space-time local truncation error. If $\|R(\tau A)\| \leq 1$, which can be shown in this example easily for the L_1, L_2 and L_∞-norms if $a\tau/h \leq 1$, then it follows in a standard fashion that

$$\|w_h(t_n) - w_n\| \leq \frac{1}{2} t_n a h \Big(1 - \frac{a\tau}{h}\Big) \max_{x,t} \|u_{xx}(x,t)\| + \mathcal{O}(h^2).$$

If we let $\tau \to 0$ with h fixed, we just reobtain the bound for the spatial error. We see, however, that the error for the above scheme will actually *decrease* for $\tau > 0$, and it will be less than the error of the semi-discrete system with exact time integration. Apparently the error of the explicit Euler time stepping counteracts the error of 1-st order upwind space discretization. $\diamond$

In the above example the discrete scheme still could be viewed within the MOL framework, only a more refined analysis is needed to obtain the true error behaviour. There are also some schemes which cannot be regarded as an ODE method applied to a certain space discretization. We conclude this section with a short description of two such schemes.

Example. One of the most popular schemes for advection equations is the *Lax-Wendroff scheme*. For the model equation $u_t + au_x = 0$ this scheme reads

$$w_j^{n+1} = w_j^n + \frac{a\tau}{2h}\Big(w_{j-1}^n - w_{j+1}^n\Big) + \frac{1}{2}(\frac{a\tau}{h})^2\Big(w_{j-1}^n - 2w_j^n + w_{j+1}^n\Big). \tag{3.9}$$

The scheme is stable, in the sense of von Neumann, under under the CFL condition $|a\tau/h| \leq 1$. This can be shown by inserting Fourier modes and computing the amplification factors. The local truncation error in space and time, defined as in the previous example, is

$$\rho_j^n = \frac{1}{6} a h^2 \Big(1 - (\frac{a\tau}{h})^2\Big) u_{xxx}(x_j, t_n) + \mathcal{O}(h^3).$$

For h fixed and $\tau \to 0$ we get the same bound as for the semi-discrete system (1.9) with central differences (not surprisingly, if we divide (3.9) by τ and then consider $\tau \to 0$). As in the previous example, the error becomes smaller for $\tau > 0$.

The Lax-Wendroff scheme can be interpreted in terms of the characteristics: we know that $u(x_j, t_{n+1}) = u(x_j - \tau a, t_n)$, and to find the value for $u(x_j - \tau a, t_n)$ one can apply quadratic interpolation using the values $u(x_{j-1}, t_n)$, $u(x_j, t_n)$ and $u(x_{j+1}, t_n)$, leading to (3.9). $\diamond$

Example. A special scheme for the diffusion equation $u_t = du_{xx}$ is the *DuFort-Frankel scheme*,

$$w_j^{n+1} = w_j^{n-1} + 2d\frac{\tau}{h^2}\Big(w_{j-1}^n - w_j^{n-1} - w_j^{n+1} + w_j^n\Big). \tag{3.10}$$

This is an explicit 2-step scheme. It is unconditionally stable (in the sense of von Neumann), which is of course very peculiar for an explicit scheme. The stability result is not as straightforward as in the other examples since this is a 2-step scheme, but it can be done by writing the 2-step recursion for w_n as a 1-step recursion for $(w_n, w_{n-1})^T$, see Richtmyer & Morton (1967). The local truncation error of this scheme equals

$$\rho_j^n = 2d\frac{\tau^2}{h^2} u_{tt}(x_j, t_n) + \mathcal{O}(\tau^2) + \mathcal{O}(h^2).$$

In spite of the unconditional stability, the time step cannot be chosen large, since this local truncation error is proportional to $(\tau/h)^2$.

Due to the fact that with standard explicit schemes one needs a bound on τ/h^2, the DuFort-Frankel scheme is still occasionally used for calculations where accuracy of the diffusion calculation is not so important, but it is not a method that can be recommended for general use. $\diamond$

4. Linear space discretizations and positivity

In this section we shall look at more general space discretizations than the first and second order examples treated thus far, and we consider the requirements for having positive solutions.

4.1. Linear advection discretizations

We consider again

$$u_t + au_x = 0 \quad \text{with} \quad a > 0,$$

for $0 \le x \le 1$ with periodicity condition and given initial profile $u(x,0)$. As we already saw in Section 1, the 1-st order upwind discretization is too diffusive, whereas the 2-nd order central discretization gives oscillations ("wiggles") and negative values. Therefore we consider the general spatial discretization formula

$$w_j'(t) = \frac{1}{h}\sum_{k=-s}^{r} \gamma_k w_{j+k}(t), \quad j = 1, 2, ..., m, \tag{4.1}$$

with $w_{i+m} \equiv w_i$. The spatial truncation error is

$$u_t(x,t) - \frac{1}{h}\sum_k \gamma_k u(x+kh,t) = -au_x - \frac{1}{h}\sum_k \gamma_k \Big(u + khu_x + \frac{1}{2}k^2h^2u_{xx} + \cdots\Big) \quad \Big|_{(x,t)} =$$

$$= -\frac{1}{h}\sum_k \gamma_k u - \Big(a + \sum_k k\gamma_k\Big)u_x - \frac{1}{2}h\sum_k k^2\gamma_k u_{xx} - \cdots \quad \Big|_{(x,t)}.$$

The conditions for order q are

$$\sum_k \gamma_k = 0, \ \sum_k k\gamma_k = -a, \ \sum_k k^2\gamma_k = 0, \cdots, \sum_k k^q\gamma_k = 0.$$

This can be satisfied $q \le r+s$. Schemes with order $q = r+s$ are called *optimal order schemes*. For each r and s there is precisely one such scheme. There is a fundamental result on the stability of these schemes:

> The optimal order schemes, with $q = r+s$, are stable for $r \le s \le r+2$ and unstable otherwise.

This result is due to Iserles & Strang (1983). A proof for the stability of the methods with $s = r$, $s = r+1$, $s = r+2$ can be found in Iserles & Nørsett (1991), pages 124,125. In that book also the instability of the other schemes is demonstrated, but this is complicated and relies on the theory of *order stars*. We note that the sufficiency for stability was proved already by Strang (1962) for fully discrete schemes.

Example: 3-th and 4-th order advection discretizations

For $s = 2$, $r = 1$ we obtain the 3-th order upwind biased discretization

$$w_j'(t) = \frac{a}{h}\Big(-\frac{1}{6}w_{j-2}(t) + w_{j-1}(t) - \frac{1}{2}w_j(t) - \frac{1}{3}w_{j+1}(t)\Big). \tag{4.2}$$

The modified equation for this discretization, which is approximated with order 4, reads $\tilde{u}_t + a\tilde{u}_x = -\frac{1}{12}ah^3\tilde{u}_{xxxx}$. The term $-\tilde{u}_{xxxx}$ is a higher order dissipation, giving damping of the high-frequency Fourier modes, but still giving some oscillations and over and under-shoot. (It should be noted that the equation $u_t = -u_{xxxx}$ does not satisfy the maximum principle. For instance, if $u(x,0) = 1 - \cos(2\pi x)$ then $u(0,0) = 0$ and $u_t(0,0) = -(2\pi)^4 < 0$.)

Figure 4.1 gives the numerical solution at $t = 1$ for $h = 1/50$ and $u(x,0) = (\sin(\pi x))^{100}$, the same as in the Figures 1.1 and 1.2. We see that this 3-th order discretization still gives some (rather small) oscillations, but the phase-speed is very good, which is in accordance with the modified equation.

If $a < 0$, the 3-th order upwind-biased discretization reads

$$w_j'(t) = \frac{a}{h}\Big(\frac{1}{3}w_{j-1}(t) + \frac{1}{2}w_j(t) - w_{j+1}(t) + \frac{1}{6}w_{j+2}(t)\Big),$$

which is a reflection of formula (4.2).

For $r = s = 2$ we get the 4-th order central discretization

$$w_j'(t) = \frac{a}{h}\Big(-\frac{1}{12}w_{j-2}(t) + \frac{2}{3}w_{j-1}(t) - \frac{2}{3}w_{j+1}(t) + \frac{1}{12}w_{j+2}(t)\Big). \tag{4.3}$$

The local truncation error will now only contain dispersion terms, no damping. For nonsmooth solutions this gives strong oscillations, see Figure 4.2 (with same initial profile and mesh width as before).

Figure 4.1. Third order upwind-biased advection

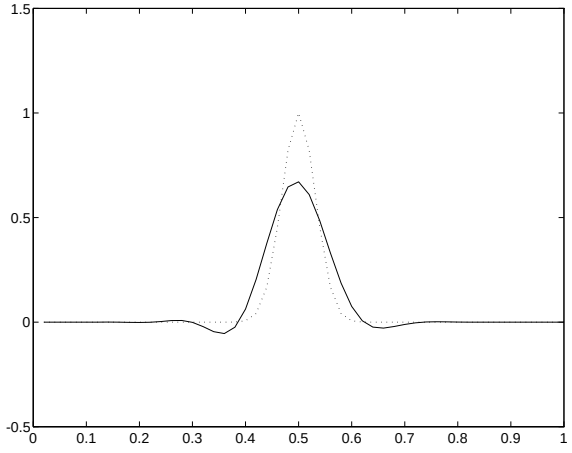
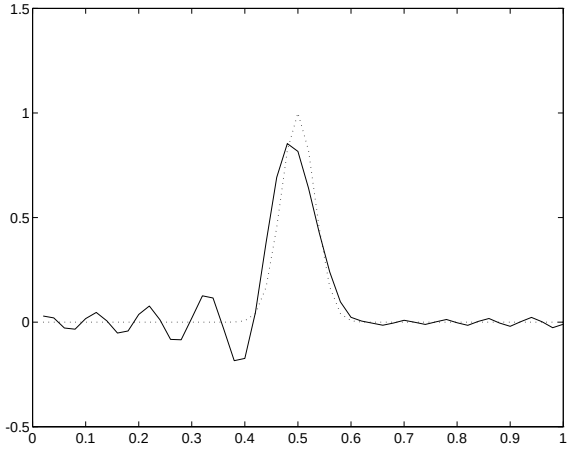


Figure 4.2. Fourth order central advection

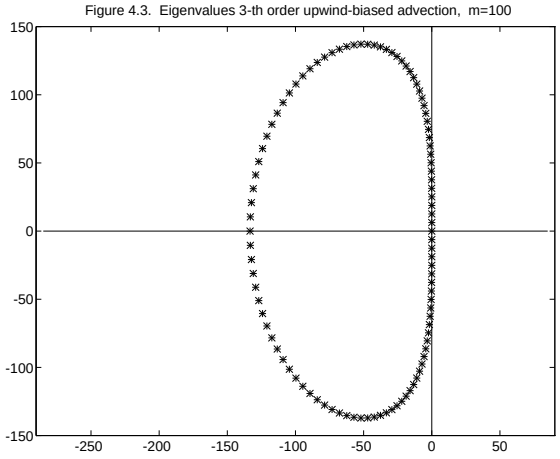


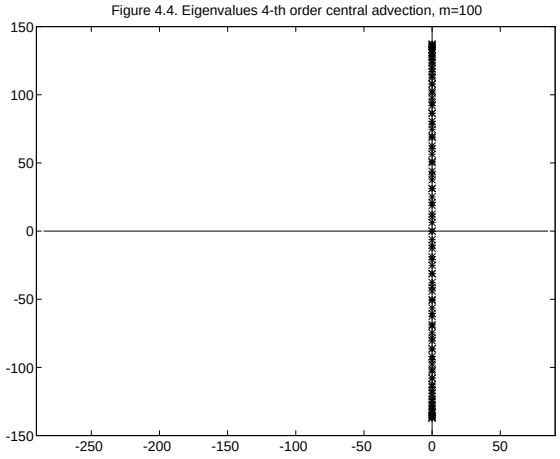
If we insert Fourier modes into the 3-th order discretization (4.2), in the same way as in Section 2, we obtain growth factors $e^{t\lambda_k}$, $k = 1, 2, ..., m$ with eigenvalues

$$\lambda_k = \frac{a}{h}\Big(-\frac{1}{6}e^{-4\pi ikh} + e^{-2\pi ikh} - \frac{1}{2} - \frac{1}{3}e^{2\pi ikh}\Big) =$$

$$= -\frac{4}{3}\frac{a}{h}\sin^4(\pi kh) - \frac{i}{3}\frac{a}{h}\sin(2\pi kh)\Big(4 - \cos(2\pi kh)\Big),$$

see Figure 4.3. Note that, although there is damping, many eigenvalues stay very close to the imaginary axis.





It can be shown that the explicit Euler method applied to this 3-th order discretization will be unstable for fixed Courant numbers $a\tau/h$ as $h \to 0$, due to the fact that many eigenvalues are almost purely imaginary. Higher order explicit Runge-Kutta methods are conditionally stable, see Appendix A.

The eigenvalues of the 4-th order central discretization are all on the imaginary axis, see Figure 4.4, similar to the 2-nd order central discretization, since this discretization is also skew-symmetric.

4.2. Positive space discretizations

For advection-diffusion equations we know, by physical interpretation, that

$$u(x,0) \geq 0 \quad \text{for all } x \quad \Longrightarrow \quad u(x,t) \geq 0 \quad \text{for all } x \text{ and } t > 0.$$

As we have seen, space discretizations may destroy this property. We would like to have a criterion that tells us when positivity is maintained.

Consider a semi-discrete ODE system in $\mathbb{R}^m$

$$w_i'(t) = F_i(t, w(t)), \quad i = 1, 2, ..., m. \tag{4.4}$$

This system will be called *positive* (short for "nonnegativity preserving") if

$$w_i(0) \geq 0 \quad \text{for all } i \quad \Longrightarrow \quad w_i(t) \geq 0 \quad \text{for all } i \text{ and } t > 0.$$

We want to have a criterion on F that tells us whether the system is positive.

Theorem 4.1. Suppose that $F(t, v)$ is continuous and satisfies a Lipschitz condition with respect to v. Then, system (4.4) is positive iff for any vector $v \in \mathbb{R}^m$ and all $i = 1, 2, ..., m$, $t \geq 0$,

$$v_i = 0, \quad v_j \geq 0 \quad \text{for all } j \neq i \quad \Longrightarrow \quad F_i(t, v) \geq 0.$$

Proof. Necessity of the above criterion easily follows. As for sufficiency, note that the criterion is equivalent with

$$w_i(t) = 0, \quad w_j(t) \geq 0 \quad \text{for all } j \neq i \quad \Longrightarrow \quad w_i'(t) \geq 0.$$

This is not enough to prove positivity, we also need the Lipschitz condition. (A counterexample, provided by Z. Horvath (1994, private communications), is $w(t) = (1-t)^3$ satisfying the scalar equation $w'(t) = -3(w(t))^{2/3}$. Note that in this example the right hand side $-3w^{2/3}$ does not satisfy a Lipschitz condition.)

It would be enough to have

$$w_i(t) = 0, \quad w_j(t) \geq 0 \quad \text{for all } j \neq i \quad \Longrightarrow \quad w_i'(t) > \varepsilon > 0,$$

since then $w(t)$ cannot cross the hyperplanes $\{w \in \mathbb{R}^m : w_i = 0 \text{ for some } i\}$. This will hold for the perturbed system with components $\tilde{F}_i(t,w) = F_i(t,w) + \varepsilon$. Using the Lipschitz condition, we can apply a standard stability argument for ODEs to show that the solution of the unperturbed system will be approximated with any precision by a solution of the perturbed system if we let $\varepsilon \to 0$, see for instance Coppel (1965). □

Corollary 4.2. A linear system $w'(t) = Aw(t)$ is positive iff

$$a_{ij} \geq 0 \quad \text{for all } j \neq i.$$

Proof. This is a consequence of Theorem 4.1. A more direct proof for linear systems follows from the relations

$$e^{\tau A} = I + \tau A + \mathcal{O}(\tau^2)$$

to show necessity, and

$$e^{t_n A} = \lim_{n \to 0} (I + \tau A)^n \quad \text{with } t_n = n\tau \text{ fixed}$$

to show sufficiency. The elaboration of this is left as exercise. □

Positivity may also imply a stronger property, namely a *maximum principle*. For linear PDEs without boundary conditions, the semi-discrete system will often satisfy the *affine invariance* property

$$F(t, \alpha v + \beta e) = \alpha F(t,v) \quad \text{for all } \alpha, \beta \in \mathbb{R} \text{ and } v \in \mathbb{R}^m,$$

with $e = (1,1,...,1)^T \in \mathbb{R}^m$. This means that if $w(t)$ is a solution of (4.4) and $v(0) = \alpha w(0) + \beta e$, then $v(t) = \alpha w(t) + \beta e$ is also a solution of (4.4). So, in particular, if $0 \leq w_i(0) \leq 1$ and $v_i(0) = 1 - w_i(0)$, for all components i, then positivity of $v(t)$ implies $w_i(t) \leq 1$. More general, if we have affine invariance, then positivity implies the maximum principle

$$\min_j w_j(0) \leq w_i(t) \leq \max_j w_j(0) \qquad \text{for all } t > 0, \tag{4.5}$$

and thus global overshoots and undershoots cannot arise.

4.3. Positivity for advection discretizations

Returning to our discretizations (4.1) for the advection equation, we see that the requirement for positivity is

$$\gamma_k \geq 0 \quad \text{for all } k \neq 0. \tag{4.6}$$

This is satisfied by the 1-st order upwind discretization, which is very inaccurate and very diffusive. Unfortunately, it is also "optimal" under the positive advection discretizations:

> For $q \geq 2$ we need $\sum_k k^2\gamma_k = 0$, and therefore
>
> $$(4.6) \Longrightarrow q \leq 1.$$
>
> Furthermore, if $q = 1$ then the leading term in the truncation error is proportional to $\sum_k k^2\gamma_k$. Since we have $\sum_k k\gamma_k = -a$, it follows that
>
> $$(4.6) \Longrightarrow \sum_k k^2\gamma_k \geq a,$$
>
> and the minimal error coefficient $\sum_k k^2\gamma_k = a$ is achieved by the 1-st order upwind discretization.

Consequently, if we want positivity and better accuracy than 1-st order upwind we have to consider *nonlinear discretizations*.

Note. Positivity for the advection equation is related to more general monotonicity and contractivity properties for nonlinear hyperbolic equations, as discussed in the monograph of LeVeque (1992). We note that the order barrier $q \leq 1$ for positive or monotone advection schemes is due to Godunov, 1959, see loc. cit.

4.4. Linear diffusion discretizations

In the same way as for the advection equation, we can consider linear discretizations for the diffusion equation

$$u_t = du_{xx}$$

with periodicity condition and given initial values. A general formula for the spatial discretization is

$$w_j'(t) = \frac{1}{h^2}\sum_{k=-s}^{r} \gamma_k w_{j+k}(t), \quad j = 1, 2, ..., m, \tag{4.7}$$

with $w_{i+m} \equiv w_i$. We assume that $s = r$ and $\gamma_{-k} = \gamma_k$, symmetry in space.

For the symmetric discretization the spatial truncation error is

$$u_t(x,t) - \frac{1}{h^2}\sum_k \gamma_k u(x + kh, t) =$$

$$= du_{xx} - \frac{1}{h^2}\sum_k \gamma_k \left(u + khu_x + \frac{1}{2}k^2h^2u_{xx} + \cdots\right) \quad \Big|_{(x,t)} =$$

$$= -\frac{1}{h^2}\sum_k \gamma_k u + \Big(d - \sum_k \frac{1}{2}k^2\gamma_k\Big)u_{xx} - \frac{1}{4!}h^2\sum_k k^4\gamma_k u_{xxxx} - \cdots \Big|_{(x,t)}.$$

So, the conditions for order q (q is even, due to symmetry) are

$$\sum_k \gamma_k = 0,\ \sum_k k^2\gamma_k = 2d,\ \sum_k k^4\gamma_k = 0, \cdots, \sum_k k^q\gamma_k = 0,$$

which is possible for $q \le 2r$.

Example. For $r = 2$ we obtain the 4-th order central diffusion discretization

$$w_j'(t) = \frac{1}{h^2}\Big(-\frac{1}{12}w_{j-2}(t) + \frac{4}{3}w_{j-1}(t) - \frac{5}{2}w_j(t) + \frac{4}{3}w_{j+1}(t) - \frac{1}{12}w_{j+2}(t)\Big).$$

The eigenvalues corresponding to this discretization are easily seen to be on the negative real axis, and thus the discretization is stable. ◇

The above 4-th order discretization fails to be positive, due to the $-\frac{1}{12}$ coefficients. Indeed, the requirement of positivity, $\gamma_k \ge 0$ for all $k \ne 0$, again leads to an order barrier:

> For $q > 2$ we need $\sum_k k^4\gamma_k = 0$, and therefore
>
> $$(4.6) \Longrightarrow q \le 2.$$
>
> Furthermore, if $q = 2$ then the leading term in the truncation error is proportional to $\sum_k k^4\gamma_k$. Since we have $\sum_k k^2\gamma_k = 2d$, it follows that
>
> $$(4.6) \Longrightarrow \sum_k k^4\gamma_k \ge 2d,$$
>
> The minimal error coefficient $\sum_k k^4\gamma_k = 2d$ is achieved by the standard 2-nd order central discretization with $r = 1$ and $\gamma_{-1} = \gamma_1 = d, \gamma_0 = -2d$.

Although this is again somewhat disappointing, the situation is not as bad as for the advection equation, since for many practical purposes this second order discretization is sufficiently accurate.

Remark. The restriction in the above to symmetric discretizations for the diffusion equation is reasonable, since, if $w'(t) = Aw(t)$ is a non-symmetrical semi-discrete system (4.7), then the symmetrical system $w'(t) = \frac{1}{2}(A + A^T)w(t)$ can be shown to be more accurate (no dispersion terms) and at least as stable. ◇

Remark. Consider the advection-diffusion equation $u_t + au_x = du_{xx}$. If we use 2-nd order central discretization for both advection and diffusion we get the semi-discrete system

$$w_j' = \frac{a}{2h}(w_{j-1} - w_{j+1}) + \frac{d}{h^2}(w_{j-1} - 2w_j + w_{j+1}).$$

This system will be positive if $|ah/d| \le 2$. The number $|ah/d|$ is called the *cell Péclet number*.

If we discretize the equation in space with first order upwind for the advection and second order central for the diffusion part, we get

$$w_j' = \frac{a^+}{h}(w_{j-1} - w_j) + \frac{a^-}{h}(w_j - w_{j+1}) + \frac{d}{h^2}(w_{j-1} - 2w_j + w_{j+1}),$$

with $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$. This semi-discrete system is always positive. It will also satisfy the translation invariance property mentioned in this section, and therefore we will have $\|w(t)\|_\infty \leq \|w(0)\|_\infty$. This implies stability for the linear semi-discrete system in the max-norm. From this, it can be easily be shown that the space discretization will convergence in the max-norm. The claim made in this section that solutions of the advection-diffusion equation with non-negative initial profile stay non-negative ("by physical interpretation") can be proven mathematically this way. $\diamond$

5. A nonlinear advection discretization by flux-limiting

Positive solutions can of course always be obtained by simply "cutting off" negative approximations. However, in this way we are adding mass, and we do not eliminate over/under shoot. So, the aim is to derive a space discretization for the advection equation that will give positive solutions, no over/under shoot and better accuracy than the 1-st order upwind scheme.

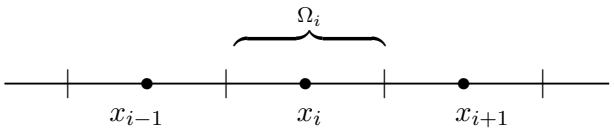
5.1. Flux forms

Mass conservation is guaranteed if we consider discretizations in the *flux form* (or *conservation form*)

$$w_i'(t) = \frac{1}{h}\Big(f_{i-\frac{1}{2}}(w(t)) - f_{i+\frac{1}{2}}(w(t))\Big). \tag{5.1}$$

Such a form is natural for *finite volume* schemes where $w_i(t)$ approximates the average value in the cell $\Omega_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$,

$$w_i(t) \approx \frac{1}{h}\int_{\Omega_i} u(x,t)dx.$$



Then $f_{i-\frac{1}{2}}, f_{i+\frac{1}{2}}$ are the fluxes at the cell boundaries. Note that the flux that "leaves" Ω_i is the flux that "enters" Ω_{i+1}, and therefore we will always have mass conservation regardless of the actual choice for the fluxes.

Examples. For the advection test problem $u_t + au_x = 0$, with $a > 0$, some flux forms are

$$f_{i+\frac{1}{2}}(w) = aw_i,$$

for the 1-st order upwind flux, and

$$f_{i+\frac{1}{2}}(w) = \frac{1}{2}a(w_i + w_{i+1}).$$

for the 2-nd order central fluxes. This last form can also be written as the 1-st order flux plus a correction ("anti-diffusion") $f_{i+\frac{1}{2}}(w) = aw_i + \frac{1}{2}a(w_{i+1} - w_i)$. For the 3-th order upwind biased formula we have the fluxes

$$f_{i+\frac{1}{2}}(w) = a(-\frac{1}{6}w_{i-1} + \frac{5}{6}w_i + \frac{1}{3}w_{i+1}).$$

Writing this as a correction to the 1-st order flux, we get

$$f_{i+\frac{1}{2}}(w) = a\Big[w_i + (\frac{1}{3} + \frac{1}{6}\theta_i)(w_{i+1} - w_i)\Big]$$

where

$$\theta_i = \frac{w_i - w_{i-1}}{w_{i+1} - w_i}.$$

$\diamond$

In the following we consider the more general form

$$f_{i+\frac{1}{2}}(w) = a\Big[w_i + \psi(\theta_i)(w_{i+1} - w_i)\Big], \tag{5.2}$$

with *limiter function* ψ, which is to be chosen such that we have better accuracy than 1-st order upwind but still positivity. For a smooth profile we have $\theta_i \approx 1$, except near extrema. Therefore we will take $\psi(\theta)$ equal to $\frac{1}{3} + \frac{1}{6}\theta$ in a region around $\theta = 1$, so that the accuracy of the third order scheme will be maintained away from extrema.

Note that (5.1),(5.2) are affine invariant. Hence, if we achieve positivity we will also avoid under/over shoot. Further it should be noted that for $a < 0$ we get, by reflection,

$$f_{i+\frac{1}{2}}(w) = a\Big[w_{i+1} + \psi(\frac{1}{\theta_{i+1}})(w_i - w_{i+1})\Big],$$

which is the same formula as (5.2), only seen from the "backside".

5.2. Choice of limiter function

The discretization (5.1),(5.2), written out in full, gives

$$w_i'(t) = \frac{a}{h}\Big[w_{i-1} + \psi(\theta_{i-1})(w_i - w_{i-1}) - w_i - \psi(\theta_i)(w_{i+1} - w_i)\Big] =$$

$$= \frac{a}{h}\Big(1 - \psi(\theta_{i-1}) + \frac{1}{\theta_i}\psi(\theta_i)\Big)(w_{i-1} - w_i),$$

with $w_i = w_i(t)$. In view of Theorem 4.1 we thus require

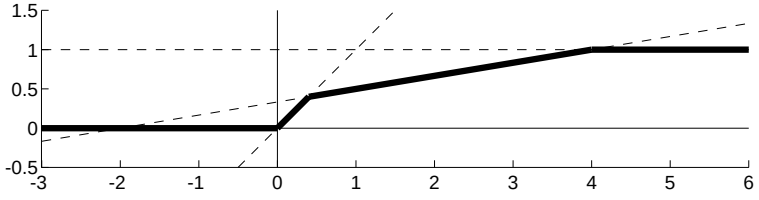
$$1 - \psi(\theta_{i-1}) + \frac{1}{\theta_i}\psi(\theta_i) \geq 0. \tag{5.3}$$

Here θ_{i-1} and θ_i can assume any value in $\mathbb{R}$, independent of each other. A sufficient condition on the limiter function is

$$0 \leq \psi(\theta) \leq 1, \quad 0 \leq \frac{1}{\theta}\psi(\theta) \leq \mu \quad \text{for all } \theta \in \mathbb{R}, \tag{5.4}$$

where μ is a positive parameter. The function that satisfies this condition and is as close as possible to $\frac{1}{3} + \frac{1}{6}\theta$ is given by

$$\psi(\theta) = \max\Big(0, \min\Big(1, \frac{1}{3} + \frac{1}{6}\theta, \mu\theta\Big)\Big). \tag{5.5}$$



The role of the parameter μ will become clear in the next section where we discuss time discretizations. For the moment, following Koren (1993), we take $\mu = 1$.

Remark. Nonlinear stability and accuracy results are lacking, but in practice these types of spatial discretizations perform well. Further we note that in actual implementations of limiters (5.2) one usually adds a small number ϵ to the denominator of the θ_i, to prevent division by 0. This may result in small negative values with order of magnitude ϵ. $\diamond$

Numerical results for the limited space discretization (5.5) are given in Figure 5.1 and 5.2 for the test equation $u_t + u_x = 0$ for $t \geq 0$, $0 \leq x \leq 1$ with periodicity. The plots are for $t = 1$ with $h = 1/50$ with initial profiles $u(x, 0) = (\sin(\pi x))^{100}$ and the block-function $u(x, 0) = 1$ for $0.3 \leq x \leq 0.7$, 0 otherwise. The exact solution is dotted ($\cdots$), the non-limited 3-th order discretization is dashed (- -) and the limited counterpart is indicated with solid lines (—).

Figure 5.1. Third order upwind biased advection with limiting

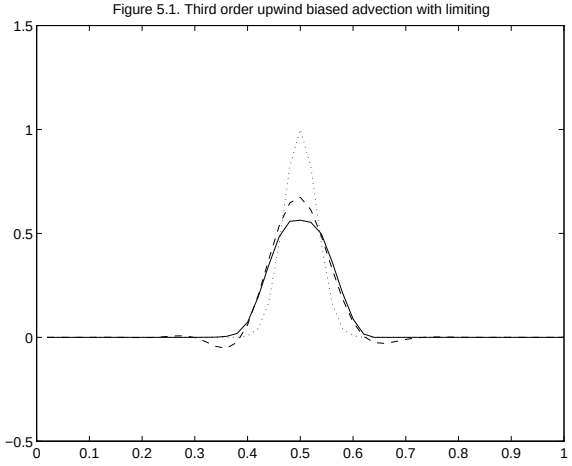
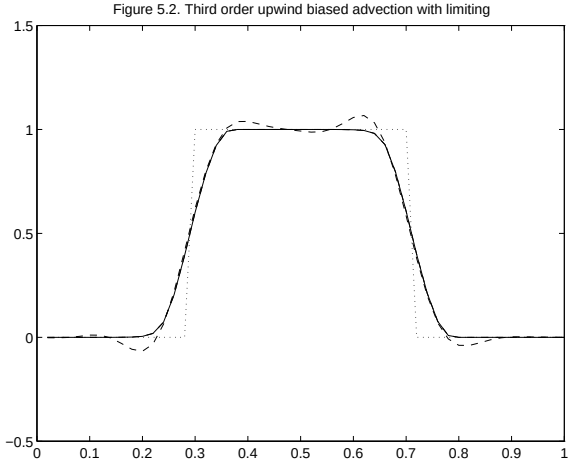


Figure 5.2. Third order upwind biased advection with limiting



The result for the $\sin^{100}$-function show that the limited discretization still has a very good phase speed, but the amplitude error has increased by the limiting procedure near the extremum. At the extremum we have $\theta_i \leq 0$ and thus the limiter will switch to $f_{i+\frac{1}{2}} = aw_i$, the 1-st order upwind flux. Note that the inaccuracy caused by this remains confined to a small region near the extremum. The result for the block-function shows that limiting can also have an overall favourable effect on the accuracy.

Formal statements on the accuracy near an extremum seem difficult to obtain, due the various switches in the discretization. In the following Table 5.3 the errors are given for a smooth function $u(x, 0) = \sin^2(\pi x)$ in the L_1-norm ($\|v\|_1 = h \sum |v_i|$), the L_2-norm ($\|v\|_2 = (h \sum |v_i|^2)^{1/2}$) and the L_∞-norm ($\|v\|_\infty = \max |v_i|$), together with the estimated order upon halving the mesh width $h = 1/m$. Also included are results for the limiter (5.5) with $\mu = 3$.

	h	L_1-error	L_2-error	L_∞-error
Non-limited	1/10	$0.37\ 10^{-1}$	$0.41\ 10^{-1}$	$0.57\ 10^{-1}$
	1/20	$0.50\ 10^{-2}$ (*2.87*)	$0.56\ 10^{-2}$ (*2.89*)	$0.78\ 10^{-2}$ (*2.86*)
	1/40	$0.64\ 10^{-3}$ (*2.98*)	$0.71\ 10^{-3}$ (*2.98*)	$0.10\ 10^{-2}$ (*2.97*)
	1/80	$0.80\ 10^{-4}$ (*3.00*)	$0.89\ 10^{-4}$ (*3.00*)	$0.12\ 10^{-3}$ (*2.99*)
	1/160	$0.10\ 10^{-4}$ (*3.00*)	$0.11\ 10^{-4}$ (*3.00*)	$0.15\ 10^{-4}$ (*3.00*)
Limiter $\mu = 1$	1/10	$0.70\ 10^{-1}$	$0.88\ 10^{-1}$	0.15
	1/20	$0.16\ 10^{-1}$ (*2.06*)	$0.22\ 10^{-1}$ (*2.00*)	$0.49\ 10^{-1}$ (*1.64*)
	1/40	$0.36\ 10^{-2}$ (*2.20*)	$0.58\ 10^{-2}$ (*1.92*)	$0.16\ 10^{-1}$ (*1.58*)
	1/80	$0.81\ 10^{-3}$ (*2.18*)	$0.25\ 10^{-2}$ (*1.92*)	$0.55\ 10^{-2}$ (*1.57*)
	1/160	$0.16\ 10^{-3}$ (*2.33*)	$0.39\ 10^{-3}$ (*1.97*)	$0.18\ 10^{-2}$ (*1.58*)
Limiter $\mu = 3$	1/10	$0.50\ 10^{-1}$	$0.62\ 10^{-1}$	0.11
	1/20	$0.74\ 10^{-2}$ (*2.76*)	$0.12\ 10^{-1}$ (*2.39*)	$0.31\ 10^{-1}$ (*1.84*)
	1/40	$0.15\ 10^{-2}$ (*2.25*)	$0.26\ 10^{-2}$ (*2.15*)	$0.94\ 10^{-2}$ (*1.75*)
	1/80	$0.32\ 10^{-3}$ (*2.27*)	$0.65\ 10^{-3}$ (*2.03*)	$0.29\ 10^{-2}$ (*1.68*)
	1/160	$0.64\ 10^{-4}$ (*2.32*)	$0.16\ 10^{-3}$ (*2.03*)	$0.93\ 10^{-3}$ (*1.66*)

TABLE 5.3. Errors and estimated orders for $u_t + u_x = 0$, $u(x,0) = \sin^2(\pi x)$.

Note. Limiters of the above type (5.2) were introduced for Lax-Wendroff type methods by Sweby (1984) based on previous work of Osher, van Leer and others. References and a more general discussion can be found in the monograph of LeVeque(1992).

The above limiter (5.5) with $\mu = 1$ was proposed by Koren (1993). This is just one of many possibilities, but it has been chosen here because it gives good results for linear advection. An example of a somewhat smoother limiter, due to van Leer (1974), is

$$\psi(\theta) = \frac{1}{2}\,\frac{\theta + |\theta|}{1 + |\theta|}. \tag{5.6}$$

The results for this limiter are slightly more diffusive than for (5.5). Many more examples and pictures can be found in the review paper of Zalesak (1987).

5.3. Numerical example: an adsorption test

As the test example we regard an adsorption-desorption model from soil mechanics. Consider a flow through a porous medium with a macroscopic velocity a and consider a chemical species that dissolves in the fluid but which can also be adsorbed by the solid medium. Let u be the dissolved concentration and v the adsorbed concentration. The conservation law for the total concentration $u + v$ then reads

$$(u + v)_t + (au)_x = 0.$$

The local balance between u and v is given by

$$v_t = -k\Big(v - \psi(u)\Big)$$

where $k > 0$ is the reaction rate and

$$\psi(u) = \frac{k_1 u}{1 + k_2 u}$$

describes the steady state ratio between u and v, with $k_1, k_2 > 0$. In soil mechanics ψ is known as a Langmuir isoterm. These equations can be written as a system of advection-reaction equations

$$\begin{aligned} u_t + (au)_x &= k\Big(v - \psi(u)\Big), \\ v_t &= -k\Big(v - \psi(u)\Big). \end{aligned} \tag{5.7}$$

Having $u, v \geq 0$ is necessary for the model to make physical sense. Moreover $\psi(u)$ has a singularity at $u = -1/k_2$. As a consequence, non-limited higher order advection discretizations cannot be used here if we have steep gradients near a state $u = 0$, since this will lead to negative values or even divergence by the singularity in ψ.

We shall take the values $k = 1000$, $k_1 = k_2 = 100$, and we solve the equations as a stiff advection-reaction system. The velocity a is spatially homogeneous and given by

$$a(t) = \begin{cases} 1 & \text{if } t \leq 1, \\ -1 & \text{if } t > 1, \end{cases}$$

the initial condition is $u, v \equiv 0$, and we have given boundary conditions $u(0,t) = 1$ for $t \leq 1$, $u(1,t) = 0$ for $t > 1$. The equation is considered on the time interval $0 \leq t \leq T = \frac{5}{4}$. An illustration of the solution is given in Figure 5.4, where the concentrations u, v and total concentration $u + v$ are plotted as function of x at time $t = 1$ and $t = T = \frac{5}{4}$.

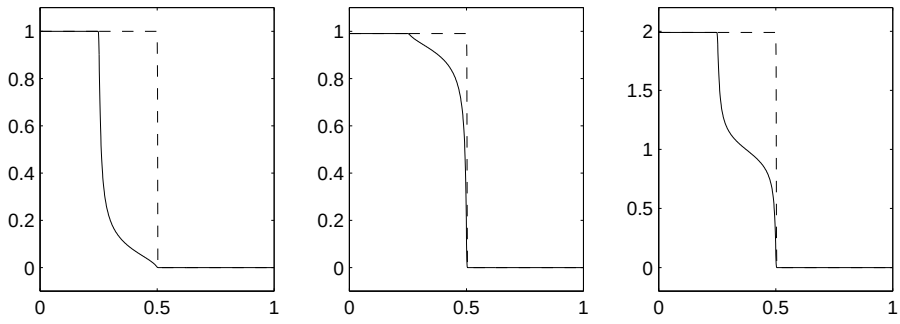


Figure 5.4. Absorption-desorption test (5.7). Plots of dissolved concentration u (left), adsorbed concentration v (middle) and total concentration $u + v$ (right) at time $t = 1$ (dashed) and $t = \frac{5}{4}$ (solid).

We see that for $0 \leq t \leq 1$ there is a shock front traveling to the left. The speed of the front is not equal to the advective velocity $a = 1$ but only approximately half of it, since the propagation is slowed down by adsorption. After $t = 1$ the advective velocity is reversed and then a rarefaction wave is formed due to advection of u to the left and dissolution of adsorbed concentration. Note that the top of this rarefaction wave (where u becomes 1) now travels with speed $a = -1$.

For the spatial discretization we consider the 1-st order upwind scheme and the limited discretization (5.2), (5.5) with $\mu = 1$. Time integration is performed with very small time steps so that no temporal errors are visible here. As said before non-limited discretizations cannot be used directly due to negative values. As an illustration for the need of mass conservation we have included in the experiment the non-limited 3-rd order upwind-biased discretization where in each time step the numerical approximation w_n is replaced by $\max(w_n, 0)$. We refer to this as "clipping". It amounts to adding mass at those regions where the advection discretization produces negative values. The results are given in the following Figure 5.5.

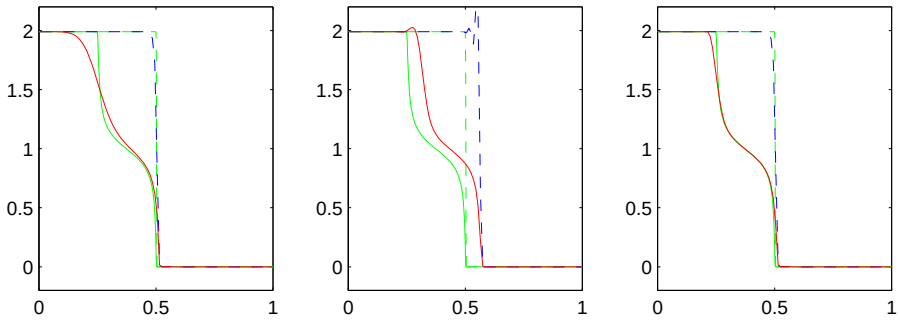


Figure 5.5. Numerical solutions for $u + v$ at $t = 1, \frac{5}{4}$ with $h = \frac{1}{100}$ and 1-st order upwind (left), 3-rd order scheme with "clipping" of negative values (middle) and limited scheme (right). The reference solutions is indicated by gray lines.

Comparing the 1-st order upwind and limited discretization, we see little difference up to $t = 1$. This can be expected since in the shock the limiter will also switch to 1-st order upwind and outside the shock the solution is constant, so there any consistent discretization gives the same result. (Enlargement of the plot would show a small difference; with 1-st order upwind the shock is a bit more smeared.) For $t > 1$ we clearly see that the first order upwind scheme gives larger errors due to numerical diffusion. Even on the short interval $t \in [1, \frac{5}{4}]$ the error has become significant. The limited scheme also adds some diffusion but much less. The difference between these two schemes would be more pronounced if the time interval were larger.

For the non-limited scheme with clipping of negative values we see that for $0 \leq t \leq 1$ the front speed is too large. This is caused by the fact that we are adding mass in the front. Therefore the adsorption will be quickly saturated and this speeds up the total solution. As a consequence the errors are very large.

Remark. Incorrect shock speeds for schemes without mass conservation are typical for non-linear hyperbolic equations. Such equations are outside the scope of these notes, but in connection to (5.6) we note the following. In the above experiments the reaction constant was given by $k = 1000$, and an increase of k hardly changes the solution. In the limit $k \to \infty$ we have $v = \psi(u)$, or

$$(u + \psi(u))_t + au_x = 0,$$

which can be formulated as a nonlinear conservation law for $\bar{u} = u + \psi(u)$,

$$\bar{u}_t + a\phi(\bar{u})_x = 0, \tag{5.8}$$

with ϕ implicitly defined by the relation

$$\bar{u} = u + \psi(u) \quad \Longrightarrow \quad u = \phi(\bar{u}).$$

We can discretize (5.8) in space by

$$\bar{w}_i'(t) = \frac{a}{h}\Big(\phi(\bar{w}_{i-\frac{1}{2}}(t)) - \phi(\bar{w}_{i+\frac{1}{2}}(t))\Big) \tag{5.9}$$

with $\bar{w}_{i+1/2} = f_{i+1/2}(\bar{w})$ computed as before, and this leads to a solution that is virtually the same as in Figure 5.4. For a good introduction to nonlinear conservation laws and the numerical solution of such equations we refer to LeVeque (1992). ◇

5.4. Formulas for non-constant coefficients and multi-dimensional problems

The advection equation

$$u_t + (a(x,t)u)_x = 0, \tag{5.10}$$

with variable velocity a, can be discretized in space as

$$w_i'(t) = \frac{1}{h}\Big(f_{i-\frac{1}{2}}(t, w(t)) - f_{i+\frac{1}{2}}(t, w(t))\Big), \tag{5.11}$$

with the fluxes given by

$$\begin{aligned} f_{i+\frac{1}{2}}(t,w) = a^+(x_{i+\frac{1}{2}},t)\Big[w_i + \psi(\theta_i)(w_{i+1} - w_i)\Big] + \\ + a^-(x_{i+\frac{1}{2}},t)\Big[w_{i+1} + \psi(\tfrac{1}{\theta_{i+1}})(w_i - w_{i+1})\Big], \end{aligned} \tag{5.12}$$

where $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$. We can take ψ as in (5.5) with $\mu = 1$. The semi-discrete system is then positive for arbitrary velocities a. If $\psi \equiv 0$ we reobtain the 1-st order upwind discretization, and $\psi \equiv \frac{1}{2}$ gives 2-nd order central (for central schemes the a^+, a^- formulation is unnecessary, of course).

For the diffusion equation

$$u_t = (d(x,t)u_x)_x, \tag{5.13}$$

with $d(x,t) > 0$, we consider the 2-nd order central discretization

$$w_i'(t) = \frac{1}{h^2}\Big(d(x_{i-\frac{1}{2}},t)(w_{i-1}(t) - w_i(t)) - d(x_{i+\frac{1}{2}},t)(w_i(t) - w_{i+1}(t))\Big). \tag{5.14}$$

Also this system is always positive, as can be verified by Corollary 4.2. It has the same form as (5.11) with diffusion fluxes $h^{-1}d(x_{i+\frac{1}{2}},t)(w_i(t) - w_{i+1}(t))$.

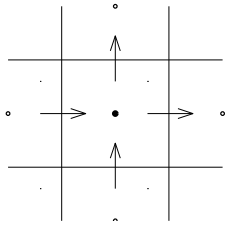
The right-hand sides of (5.11) with (5.12) and of (5.14) can be regarded as finite difference approximations to $(a(x,t)u)_x$ and $(d(x,t)u_x)_x$, respectively. Superposition of these finite

differences gives space discretizations for the general multi-dimensional advection-diffusion equation

$$\frac{\partial}{\partial t}u + \sum_{k=1}^{d} \frac{\partial}{\partial x_k}\Big(a_k(x,t)u\Big) = \sum_{k=1}^{d} \frac{\partial}{\partial x_k}\Big(d_k(x,t)\frac{\partial}{\partial x_k}u\Big) \tag{5.15}$$

with $x = (x_1, x_2, ..., x_d)^T \in \mathbb{R}^d$. (Here x_k denotes the k-th direction.) Working on regular Cartesian grids, we can simply plug in our 1-dimensional discretizations for the individual terms $(a_k u)_{x_k}$ and $(d_k u_{x_k})_{x_k}$ in the various directions. Reaction terms are also easily included. It is this possibility of superposition that makes the method of lines approach popular. Methods with combined space-time discretizations, such as the Lax-Wendroff method, are much harder to formulate for multi-dimensional advection-diffusion-reaction problems.

Remark. Deriving the multi-dimensional discretizations can also be done entirely within the finite-volume framework by considering in- and outflows over cell boundaries, say



$$w'_{ij} = \frac{1}{\Delta x}(f_{i-\frac{1}{2},j} - f_{i+\frac{1}{2},j}) + \frac{1}{\Delta y}(f_{i,j-\frac{1}{2}} - f_{i,j+\frac{1}{2}}),$$

in 2D with x and y coordinates.

Irrespective of the spatial dimension, the difference between a cell-average value and the value in a cell-center is $\mathcal{O}(h^2)$, and therefore the order of a discretization may depend on the interpretation, either as finite differences or as finite volumes. This will happen only if the order is larger than 2.

Indeed, by a Taylor expansion (and tedious calculations) it can be seen that the 1D scheme (5.11),(5.12), with $\psi(\theta) = \frac{1}{3} + \frac{1}{6}\theta$, has a third order truncation error as a finite volume scheme and only second order for the finite difference approximations. However, in 2 dimensions the order becomes also 2 for the finite volume interpretation, due to the fact that the fluxes over the cell boundaries are only evaluated in the middle of the edges (midpoint approximation to an integral). A third order finite difference approximation could be obtained by applying the discretization (4.2) directly to ac, instead of c, but in connection with limiting the form (5.11),(5.12) seems more appropriate. ◇

Remark. Equation (5.10) is called the *conservative form* of the advection equation. The *advective form* is given by $\tilde{u}_t + a(x,t)\tilde{u}_x = 0$. More general, for multi-dimensional problems we have the forms

$$u_t + \sum_k \frac{\partial}{\partial x_k}(a_k u) \;=\; u_t + \mathrm{div}(\underline{a}u) = 0$$

and

$$\tilde{u}_t + \sum_k a_k \frac{\partial \tilde{u}}{\partial x_k} \;=\; \tilde{u}_t + \underline{a} \cdot \mathrm{grad}(\tilde{u}) = 0,$$

respectively. Both forms have physical meaning. For a chemical species carried along by some fluid medium (for example, wind or water) with velocity a, the concentration u will satisfy the conservative form, reflecting the fact that mass is conserved. On the other hand, the mixing ratio, defined as concentration divided by the density ρ (the sum of all concentrations), will

satisfy the advective form. These mixing ratios are constant along the characteristics $(\xi(t), t)$ given by $\xi'(t) = \underline{a}(\xi(t), t)$. The two forms are equivalent if the velocity field is divergence free, that is $\sum(\partial a_k/\partial x_k) = 0$, which means that the fluid is incompressible and that the density is constant. Even if this holds, a numerical discretization of the advective form will in general give rise to a scheme that does not conserve mass. In air pollution modelling one usually encounters the conservative form, also due to the fact that chemical reactions are most frequently defined in terms of concentrations. $\diamond$

6. Positive time discretizations

Consider a linear semi-discrete system $w'(t) = Aw(t)$, where A satisfies

$$a_{ij} \geq 0 \quad \text{for } i \neq j, \qquad a_{ii} \geq -\alpha \quad \text{for all } i, \tag{6.1}$$

with $\alpha > 0$. As we saw in Section 4, this guarantees positivity of the system, irrespective of the value of α. Of course, we want to maintain positivity when time discretization is performed. As introduction, we first consider the explicit (forward) and implicit (backward) Euler time discretizations.

Application of the forward Euler method to the linear system gives $w_{n+1} = w_n + \tau A w_n$, that is,

$$w_{n+1} = (I + \tau A) w_n.$$

It is easily seen that $I + \tau A \geq 0$ (inequality componentwise) provided that $1 + \tau a_{ii} \geq 0$ for all i. This will hold if the step size is restricted such that $\alpha\tau \leq 1$.

The backward Euler method gives $w_{n+1} = w_n + \tau A w_{n+1}$, and this can be written as

$$w_{n+1} = (I - \tau A)^{-1} w_n.$$

Suppose that

$$A \text{ has no eigenvalues on the positive real axis.} \tag{6.2}$$

Then $I - \tau A$ is invertible for all $\tau > 0$, and so the Backward Euler relation has a unique solution. In fact, this solution will also be positive. The conditions (6.1),(6.2) imply

$$(I - \tau A)^{-1} \geq 0 \quad \text{for all } \tau > 0.$$

The proof of this statement will be given in Lemma 6.3 for a nonlinear case.

With these results for the forward and backward Euler method it is also possible to derive positivity results for certain other simple schemes. For example, the trapezoidal rule

$$w_{n+1} = w_n + \frac{1}{2}\tau A w_n + \frac{1}{2}\tau A w_{n+1}$$

can be viewed as a combination of two half steps with the forward and backward Euler method, respectively, and thus positivity is guaranteed for $\tau\alpha \leq 2$. We shall return to this when discussing non-linear systems. For linear systems there exists a nice, complete theory, the results of which will be presented next.

In the results we shall use rational functions with matrix arguments, see Section 2.

6.1. Positivity results of Bolley & Crouzeix

Consider a one-step method of order p and stability function R. Application to the linear semi-discrete system $w'(t) = Aw(t)$ will give

$$w_{n+1} = R(\tau A) w_n.$$

The rational function R is said to be *absolutely monotonic* on an interval $[-\gamma, 0]$ if R and all its derivatives are nonnegative on this interval. Let γ_R be the largest γ for which this holds. If there is no $\gamma > 0$ such that R is absolutely monotonic on $[-\gamma, 0]$, we set $\gamma_R = 0$.

We consider in the following the class $\mathcal{M}_\alpha$ consisting of all matrices satisfying (6.1),(6.2) with fixed $\alpha > 0$. The following theorem is due to Bolley & Crouzeix (1978). We elaborate the proof somewhat because it gives insight in the occurrence of the derivatives of R.

Theorem 6.1. $R(\tau A) \geq 0$ for all $A \in \mathcal{M}_\alpha$ iff $\alpha\tau \leq \gamma_R$.

Proof. Let $\mu = \alpha\tau$ and write τA as $-\mu I + N$ with $N = \mu I + \tau A$. We have the following expansion

$$R(\tau A) = \sum_{j\geq 0} \frac{1}{j!} R^{(j)}(-\mu)\, N^j.$$

The validity of this series expansion for $\mu \leq \gamma_R$ can be demonstrated in the following way:

(i) by the absolute monotonicity it can be shown that the power series (for scalar, complex arguments) has a radius of convergence larger than μ,

(ii) by using the Perron-Frobenius theorem for nonnegative matrices it can be shown that $\rho(N) \leq \mu$.

For this technical (but interesting) part of the proof we refer to Bolley & Crouzeix (1978). Here we simply assume that the expansion is valid. Then, sufficiency of the condition $\alpha\tau \leq \gamma_R$ follows directly from the fact that $N \geq 0$.

To prove necessity, consider the first order upwind discretization for $u_t + u_x = 0$, $u(0,t) = 0$, giving the semi-discrete system $w'(t) = Aw(t)$ with

$$A = \frac{1}{h}\begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} = \frac{1}{h}(-I + E) \in \mathbb{R}^{m\times m},$$

where E denotes the backward shift operator on $\mathbb{R}^m$. Taking $\mu = \tau/h$, we have $\tau A = -\mu I + \mu E$ and therefore

$$R(\tau A) = R(-\mu)I + \mu R'(-\mu)E + \frac{1}{2}\mu^2 R''(-\mu)E^2 + \cdots + \frac{1}{(m-1)!}\mu^{m-1} R^{(m-1)}(-\mu)E^{m-1}.$$

This is a lower triangular matrix with elements $R^{(j)}(-\mu)$ for $j = 0, 1, ..., m-1$. Thus we see that in order to have $R(\tau A) \geq 0$ for arbitrarily large m, it is necessary to have $R^{(j)}(-\mu) \geq 0$ for all $j \geq 0$. □

Of course we would like to have large γ_R, preferably $\gamma_R = \infty$ in which case we will have unconditional positivity. This can only hold for implicit methods, where R is not a polynomial. We already mentioned that it holds for the backward Euler method, see also Lemma 6.3. One might hope to find more accurate methods with this property, but Bolley & Crouzeix (1978) showed that

$$\gamma_R = \infty \quad \Longrightarrow \quad p \leq 1,$$

and therefore the backward Euler method is the only well-known method with $\gamma_R = \infty$. (A proof of this last result is based on a characterisation already given by Bernstein in 1928, see also Hairer & Wanner (1991. p. 188).)

It is easy to see that, for $0 \le \theta \le 1$,

$$R(z) = (1 + (1-\theta)z)/(1-\theta z) \qquad \Longrightarrow \qquad \gamma_R = 1/(1-\theta).$$

This is relevant to the θ-methods, considered in Section 3. As a further result we have

$$R(z) = 1 + z + \frac{1}{2}z^2 + \cdots + \frac{1}{p!}z^p \qquad \Longrightarrow \qquad \gamma_R = 1.$$

For the calculation of this bound one can use, in a repeated fashion, the fact that if $0 < \gamma \le 1$ and P is a polynomial with $P(0) = 1$, $0 \le P'(z) \le 1$ for $z \in [-\gamma, 0)$, then also $0 \le P(z) \le 1$ for $z \in [-\gamma, 0)$. This is relevant to well-known Runge-Kutta methods up to order $p = 4$. A table of values of γ_R for Padé approximations can be found in Hairer & Wanner (1991, p. 188), mainly based on previous work of Kraaijevanger (see loc. cit.).

Note. The threshold factors γ_R also occur in the study of contractivity $\|R(\tau A)\| \le 1$ in the max-norm or the sum-norm, for problems satisfying $\|e^{tA}\| \le 1$ for all $t \ge 0$, see Spijker (1983). As an example we mention that if $\|A + \alpha I\|_\infty \le \alpha$, then $\|R(\tau A)\|_\infty \le 1$ provided that $\tau\alpha \le \gamma_R$.

As a generalization, we now consider the linear system with source term

$$w'(t) = Aw(t) + g(t), \tag{6.3}$$

with $A \in \mathcal{M}_\alpha$ and $g(t) \ge 0$ for all $t \ge 0$. Application of a one-step method (say, Runge-Kutta or Rosenbrock type) will then lead to a recursion

$$w_{n+1} = R(\tau A)w_n + \sum_{j=1}^{s} Q_j(\tau A)\tau g(t_n + c_j\tau), \tag{6.4}$$

Therefore positivity is ensured if

$$R(\tau A) \ge 0 \quad \text{and} \quad Q_j(\tau A) \ge 0, \quad j = 1, 2, ..., s.$$

Example. As an example, we consider the *explicit trapezoidal rule*, which consists of the implicit trapezoidal rule with Euler predictor,

$$\begin{aligned} \bar{w}_{n+1} &= w_n + \tau F(t_n, w_n), \\ w_{n+1} &= w_n + \tfrac{1}{2}\tau F(t_n, w_n) + \tfrac{1}{2}\tau F(t_{n+1}, \bar{w}_{n+1}), \end{aligned} \tag{6.5}$$

and the related method, consisting of the implicit midpoint rule with Euler predictor,

$$\begin{aligned} \bar{w}_{n+1/2} &= w_n + \tfrac{1}{2}\tau F(t_n, w_n), \\ w_{n+1} &= w_n + \tau F(t_{n+1/2}, \bar{w}_{n+1/2}). \end{aligned} \tag{6.6}$$

Both methods are of order 2 and they have the same stability function $R(z) = 1 + z + \frac{1}{2}z^2$. Application of the explicit trapezoidal rule (6.5) to the inhomogeneous system (6.3) gives

$$w_{n+1} = R(\tau A)w_n + \frac{1}{2}(I + \tau A)\tau g(t_n) + \frac{1}{2}\tau g(t_{n+1}),$$

and thus positivity is ensured for $\alpha\tau \le 1$. For (6.6) we get

$$w_{n+1} = R(\tau A)w_n + \frac{1}{2}\tau^2 A g(t_n) + \tau g(t_{n+1/2}),$$

and for this method we get the step size restriction $\alpha\tau \le 0$, that is $\tau = 0$, if we insist on positivity with arbitrary $g(t) \ge 0$. Under the mild, extra condition $2g(t_{n+1/2}) - g(t_n) \ge 0$ we will again have positivity for $\alpha\tau \le 1$.

Note that for the implicit trapezoidal rule and implicit midpoint rule the requirement for positivity is $\alpha\tau \le 2$. Although this is better than what we get for the explicit counterparts, it is not comparable to the difference in stability requirements between the implicit and explicit schemes. ◇

We note that Bolley and Crouzeix also derived positivity results for linear multi-step methods with *arbitrary* nonnegative starting values. The conditions for this are very restrictive. For example, with the BDF_2 method

$$w_{n+2} = \frac{4}{3}w_{n+1} - \frac{1}{3}w_n + \frac{2}{3}F(t_{n+2}, w_{n+2})$$

one never has $w_2 \ge 0$ for arbitrary $w_0, w_1 \ge 0$, due to the presence of the factor $-\frac{1}{3}$ in the formula. It seems more reasonable to consider this method with a starting procedure, say $w_1 = w_0 + \tau F(t_1, w_1)$. It was shown by M. van Loon (1996, private communications) that the resulting scheme is then positive for linear systems (6.3) under the restriction $\tau\alpha \le 1/2$. Van Loon did prove this by considering the recursion

$$(I - \frac{2}{3}\tau A)(2w_{n+2} - w_{n+1}) = \frac{2}{3}(2w_{n+1} - w_n) + \frac{1}{3}(w_{n+1} + 2\tau A w_{n+1}) + \frac{4}{3}\tau g(t_{n+2}),$$

$$(I - \tau A)(2w_1 - w_0) = w_0 + \tau A w_0 + 2\tau g(t_1).$$

By induction it can be shown that $2w_{n+2} \ge w_{n+1} \ge 0$. Results of this kind for general multistep methods seem unknown.

6.2. Nonlinear positivity

Consider a general, nonlinear ODE system

$$w'(t) = F(t, w(t)).$$

The counterpart of condition (6.1) is: there is an $\alpha > 0$ such that

$$v + \tau F(t, v) \ge 0 \quad \text{for all } t \ge 0, v \ge 0 \text{ and } \alpha\tau \le 1. \tag{6.7}$$

This guarantees of course positivity for the forward Euler method. Further we assumed for linear systems that A has no eigenvalues on the positive real axis, so that the implicit relations for backward Euler have a unique solution. As nonlinear counterpart we will now assume

$$\begin{array}{l}\text{for any } v \ge 0,\ t \ge 0,\ \tau > 0 \text{ the equation } u = v + \tau F(t, u)\\ \text{has a unique solution that depends continuously on } \tau,\ v.\end{array} \tag{6.8}$$

This means that the backward Euler method is well defined. It also implies unconditional positivity.

Lemma 6.3. Conditions (6.7),(6.8) imply positivity for the backward Euler scheme for any step size $\tau > 0$.

Proof. For given t, v and with τ variable, we consider $u = v + \tau F(t, u)$ and we call this solution $u(\tau)$. We have to show that $v \geq 0$ implies $u(\tau) \geq 0$ for all positive τ. By continuity it is sufficient to show that $v > 0$ implies $u(\tau) \geq 0$. This is true (even $u(\tau) > 0$), for, if we assume that $u(\tau) > 0$ for $\tau \leq \tau_0$ but $u_i(\tau_0) = 0$, then

$$0 = u_i(\tau_0) = v_i + \tau_0 F_i(t, u(\tau_0)).$$

According to (6.7) we have $F_i(t, u(\tau_0)) \geq 0$ and thus $v_i + \tau_0 F_i(t, u(\tau_0)) > 0$, which is a contradiction. □

Note. A sufficient condition for (6.8) is that F is continuously differentiable and

$$\|(I - \tau F'(t, v))^{-1}\| \leq C \quad \text{for any } v \in \mathbb{R}^M, t \geq 0, \tau > 0,$$

with C some positive constant and $F'(t, v)$ the Jacobi matrix $(\partial F_i(t, v)/\partial v_j)$. Existence and uniqueness of the solution then follows from Hadamard's theorem, and by the implicit function theorem this solution depends continuously on τ, t and v, see for instance Ortega & Rheinboldt (1970), p.128 and p.137.

A theory for general Runge-Kutta methods is lacking at present. However, following an idea of Shu & Osher (1988) for explicit methods, it is easy to derive results for a class of diagonally implicit methods. We consider methods of the Runge-Kutta type, with internal vectors $w_{1n}, w_{2n}, ..., w_{s+1,n}$. To compute w_{n+1} from w_n, we set $w_{1n} = w_n$,

$$w_{in} = \sum_{j=1}^{i-1} \Big[p_{ij} w_{jn} + \tau q_{ij} F(t_n + c_j \tau, w_{jn}) \Big] + \tau r_i F(t_n + c_i \tau, w_{in}), \quad i = 2, 3..., s+1, \qquad (6.9)$$

giving the next approximation $w_{n+1} = w_{s+1,n}$. Here the parameters p_{ij}, q_{ij}, r_i and c_j define the method, with $\sum_{j=1}^{i-1} p_{ij} = 1$.

Theorem 6.4. If all parameters p_{ij}, q_{ij}, r_i with $1 \leq j < i \leq s+1$ are nonnegative, then method (6.9) will be positive for any F satisfying (6.7),(6.8) under the step size restriction

$$\alpha \tau \leq \min_{i,j} \frac{p_{ij}}{q_{ij}}$$

(convention: $p_{ij}/0 = +\infty$ for $p_{ij} \geq 0$). For explicit methods, with all $r_i = 0$, we only have to assume that F satisfies (6.7). For implicit methods we also need (6.8).

Proof. The proof follows by induction with respect to i, from the above results for the explicit and implicit Euler method. □

Example. For the explicit trapezoidal rule (6.5) we can write the second stage as

$$w_{n+1} = \frac{1}{2}w_n + \frac{1}{2}\bar{w}_{n+1} + \frac{1}{2}\tau F(t_{n+1/2}, \bar{w}_{n+1}).$$

Therefore, we have nonlinear positivity for $\alpha\tau \le 1$, the same condition as for linear systems.

With the midpoint discretization (6.6) we can write the second stage as

$$w_{n+1} = (1-\theta)w_n - \frac{1}{2}\theta\tau F(t_n, w_n) + \theta\bar{w}_{n+1/2} + \tau F(t_{n+1/2}, \bar{w}_{n+1/2})$$

with arbitrary $\theta \in \mathbb{R}$, but we cannot achieve a form (6.9) with all $p_{ij}, q_{ij} \ge 0$. In fact we already saw for the linear inhomogeneous equations that we cannot have positivity if only (6.7) and (6.8) are assumed. ◇

Example. The classical 4-th order Runge-Kutta method reads

$$\left.\begin{aligned} w_{1n} = w_n, \qquad w_{in} = w_n + c_i\tau F(t_n + c_{i-1}\tau, w_{i-1,n}) \quad (i = 2,3,4), \\ w_{n+1} = w_{1n} + \textstyle\sum_{i=1}^{4} b_i\tau F(t_n + c_i\tau, w_{in}), \end{aligned}\right\} \tag{6.10}$$

with $c_1 = 0, c_2 = c_3 = \frac{1}{2}, c_4 = 1$ and $b_1 = b_4 = \frac{1}{6}, b_2 = b_3 = \frac{1}{3}$. The stability function of this method is

$$R(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{6}z^3 + \frac{1}{24}z^4,$$

and thus we have for linear homogeneous systems,

$$w'(t) = Aw(t), \quad A \in \mathcal{M}_\alpha \quad \Longrightarrow \quad \text{positivity for } \alpha\tau \le 1.$$

Further the rational functions Q_j in (6.4) are found to be

$$Q_1(z) = \frac{1}{6}(1 + z + \frac{1}{2}z^2 + \frac{1}{4}z^3), \quad Q_2(z) = \frac{1}{6}(2 + z + \frac{1}{2}z^2), \quad Q_3(z) = \frac{1}{6}(2+z), \quad Q_4(z) = \frac{1}{6}.$$

(Actually, Q_2 and Q_3 should be taken together since $c_2 = c_3$.) It follows that we have for linear inhomogeneous equations,

$$w'(t) = Aw(t) + g(t), \quad A \in \mathcal{M}_\alpha, \quad g(t) \ge 0 \quad \Longrightarrow \quad \text{positivity for } \alpha\tau \le 2/3.$$

This bound is determined by Q_1''.

This Runge-Kutta method cannot be written in the form (6.9) with nonnegative coefficients p_{ij}, q_{ij}, and therefore we get no result for nonlinear positivity. (A proof for the nonexistence of nonnegative p_{ij}, q_{ij} can be obtained, in a roundabout way, from the contractivity results of Kraaijevanger (1991)). ◇

Note. More general results on nonlinear positivity of Runge-Kutta methods can be found in Horvath (1998). These results are closely related to Kraaijevanger (1991) where contractivity in arbitrary norms is considered.

As an illustration we consider the parabolic initial-boundary value problem

$$u_t = u_{xx}, \quad u(0,t) = u(1,t) = 0, \quad u(x,0) = \begin{cases} 0 & \text{for } 0 < x < \frac{1}{2}, \\ 1 & \text{for } \frac{1}{2} \le x < 1, \end{cases}$$

with discontinuities at $x = \frac{1}{2}$ and 1 for $t = 0$. Space discretization with 2-nd order central differences gives approximations $w_i(t) \approx u(x_i, t)$ by

$$w'(t) = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix} w(t), \qquad w_i(0) = \begin{cases} 0 & \text{for } 1 \le i < \frac{1}{2}m, \\ 1 & \text{for } \frac{1}{2}m \le i \le m, \end{cases}$$

with $x_i = ih$ and $h = 1/(m+1)$. Application of backward Euler and the trapezoidal rule (Crank-Nicolson) with $\tau = h = 1/50$ gives the following Figure 6.1. Note that a sufficient condition for positivity of the Crank-Nicolson scheme is $\tau/h^2 \le 1$, which is clearly not satisfied here.

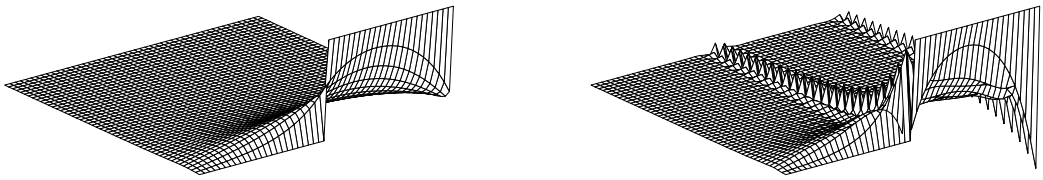


Figure 6.1. Discontinuous diffusion solutions with backward Euler (left) and Crank-Nicolson (right).

In practice, problems with positivity are not very often encountered with parabolic equations. The solutions for such problems are in general rather smooth and then accuracy takes care of negative values. Also in the discontinuous example presented here negative values can be avoided by starting the Crank-Nicolson scheme with small τ and then gradually increasing the time step.

Note. The condition $\tau/h^2 \le 1$ for positivity in the above example is obtained from Theorem 6.1, which was formulated for arbitrary matrices $A \in \mathcal{M}_\alpha$. For the above matrix $A = h^{-2}\text{tridiag}(1, -2, 1)$ this condition is a bit too strict. MATLAB experiments show that for this particular matrix the actual upper bound on τ/h^2 is approximately 1.17, so this is pretty close to the general bound from Theorem 6.1. For the Crank-Nicolson scheme with periodicity conditions it can be proven that the scheme will be positive provided that $\tau/h^2 \le 1.5$, see Dautray & Lions (1993, p. 50). We also note that Bolley & Crouzeix (1978) showed the condition of Theorem 6.1 to be necessary for matrices of the type $A = -\mu I + \epsilon\,\text{tridiag}(1, -2, 1)$ with $\mu > 0$ and with $\epsilon > 0$ sufficiently small.

For the advection equation $u_t + au_x = 0$ with $a > 0$, the discretization (5.1),(5.2) with limiter (5.5) leads to a semi-discrete system of the form

$$w_i'(t) = \alpha_i(w(t))\Big(w_{i-1}(t) - w_i(t)\Big)$$

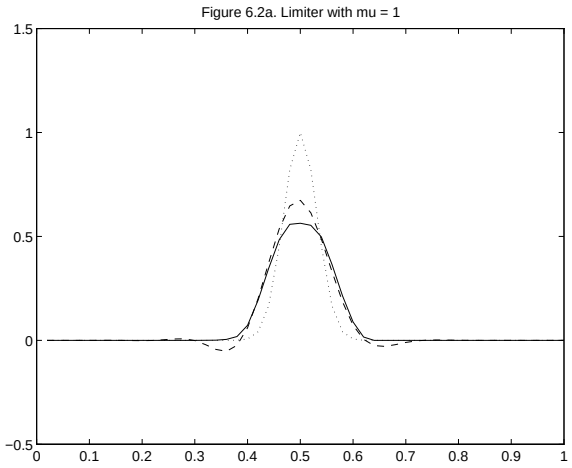
with

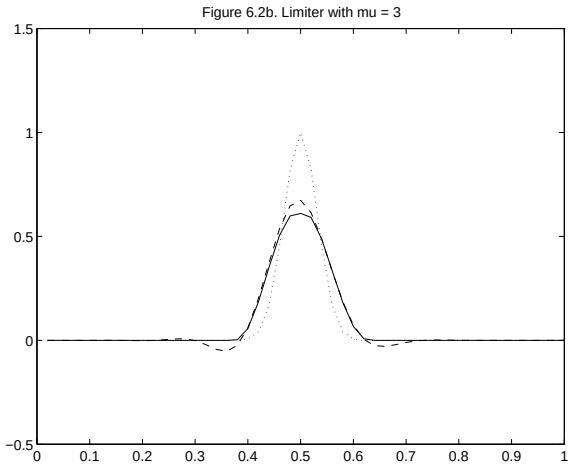
$$0 \le \alpha_i(w) \le \frac{a}{h}(1+\mu),$$

see Section 5. Condition (6.7) is satisfied with constant $\alpha = (a/h)(1+\mu)$. Therefore, with the explicit trapezoidal rule (6.5) positivity is known to hold if the Courant number $\nu = a\tau/h$ satisfies

$$(1+\mu)\nu \le 1.$$

Taking μ large gives a slightly better accuracy, especially near peaks, see Table 5.3 and the Figures 6.2 (left picture the same as Figure 5.1). This is also to be expected since taking large μ means that the limiter is less often applied. However, the above theoretical result predicts that we then need smaller time steps, that is, more computer time. Therefore the choice $\mu = 1$ seems a good compromise.





The dependence on μ of the maximal allowable Courant number was confirmed in numerical experiments (Hundsdorfer et al. (1995)). These numerical experiments were further not very conclusive. For example with $\mu = 1$ and the classical Runge-Kutta method positive values were obtained in 1 space dimension up to $\nu = 1.37$, whereas in 2 dimensions this method always gave negative results. Furthermore, little difference was found between the methods (6.5) and (6.6). With $\mu = 1$ both these methods did give positive results up to $\nu = 1$ in 1D, and $\nu \approx \frac{2}{3}$ in 2D. The theoretical bound for positivity with the explicit trapezoidal rule (6.5) is $\nu = \frac{1}{2}$.

So, the general theory seems to have some relevance for this specific advection discretization, but it is not able to give very accurate bounds. Based on theoretical and experimental observations, the explicit trapezoidal rule (6.5) seems a good choice for the time integration of advection with limiters.

Remark. Application of the forward Euler method to the flux-limited advection discretization gives very peculiar results. The scheme does satisfy a maximum principle if $(1+\mu)\nu \leq 1$, but smooth initial profiles are turned into blocks or staircases.

The reason for this is the instability of forward Euler for the underlying 3-th order discretization, see Figure 4.3 and the remark thereafter. With limiting we get the interesting *nonlinear* phenomenon: instability combined with maximum principle. In particular this shows that for nonlinear systems boundedness is not sufficient for having stability. $\diamond$

7. Boundary conditions and spatial accuracy

Consider the advection-diffusion equation

$$u_t + au_x = du_{xx} \quad \text{for } t \geq 0,\ 0 \leq x \leq 1 \tag{7.1}$$

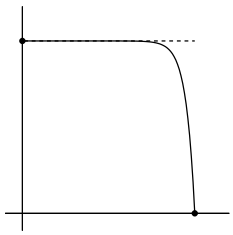
with given initial profile $u(x,0)$. If $d > 0$ we need boundary conditions at $x = 0$ and $x = 1$. Periodicity conditions do not often occur in practice. It is more common to impose *Dirichlet* conditions, where the values at the boundaries are prescribed,

$$u(0,t) = \gamma_0, \quad u(1,t) = \gamma_1, \tag{7.2}$$

or, more general, with time dependent boundary values $\gamma_0(t)$ and $\gamma_1(t)$.

On the other hand, for the pure advection problem with $d = 0$ we need only conditions at the *inflow* boundary, that is, at $x = 0$ if $a > 0$ and at $x = 1$ if $a < 0$. If $d > 0$ but $d \approx 0$ (more precisely, if the Péclet number $|a/d|$ is large), then the Dirichlet condition at the outflow boundary will give rise to a *boundary layer*.

Example. Let $u(0,t) \equiv 1$ and $a > 0$. Then the advection equation $u_t + au_x = 0$ gives the stationary solution $u(x,t) \equiv 1$.
On the other hand, if we consider the advection-diffusion equation $u_t + au_x = du_{xx}$ with $u(1,t) \equiv 0$ we get the stationary solution $u(x,t) = (e^{a/d} - e^{ax/d})/(e^{a/d} - 1)$.



A boundary layer of this type will be absent if the *Neumann* condition $u_x = 0$ is imposed at the outflow boundary. If $a > 0$ we then have

$$u(0,t) = \gamma_0, \quad u_x(1,t) = 0. \tag{7.3}$$

With this condition rapid changes may still occur in the spatial derivatives of u, but u itself will not show the nearly discontinuous behaviour that arises with Dirichlet conditions.

In practice, finding correct boundary conditions is a difficult task for the modellers, and much physical insight is needed for systems of equations. Boundary conditions also give rise to several numerical difficulties, some of which will be shortly addressed in this section. First, our numerical scheme may "require" more boundary conditions than the physical model. Secondly, the von Neumann stability analysis is no longer applicable (even with constant coefficients and L_2-norm). Thirdly, even if there are no boundary layers, a more refined error analysis may be needed to predict the correct order of convergence of the numerical scheme.

The issue of stability is extensively treated in Richtmyer & Morton (1967, Chapter 6) and Strikwerda (1989, Chapter 11). Here we shall confine ourselves to a few remarks on this subject. By means of some examples the issue of accuracy for smooth solutions will be discussed. Further a brief description will be given of a local grid refinement procedure that can be used to resolve boundary layers and other sharp gradients.

Note. Another major numerical difficulty with boundary conditions occurs for multi-dimensional problems when the spatial domain is not aligned with the grid, especially if this domain has a complicated shape. This is outside the scope of these notes. Also special fitted schemes that are suited to deal with boundary layers are not considered here. For these topics we refer to the monograph of Morton (1996).

7.1. Spatial accuracy

Consider the linear semi-discrete system

$$w'(t) \;=\; F(t,w(t)) \;=\; Aw(t)+g(t),$$

and let $w_h(t)$ be the exact PDE solution restricted to the space grid, and $\sigma_h(t) = w_h'(t) - F(t,w_h(t))$ the spatial truncation error. We shall use the stability assumption

$$\|e^{tA}\| \le K \quad \text{for all } t\in[0,T]\;, \tag{7.4}$$

with moderate $K>0$, independent of the mesh width h. As we saw in Section 2, the estimate $\|\sigma_h(t)\| = \mathcal{O}(h^q)$ leads to a bound $\|w_h(t)-w(t)\| = \mathcal{O}(h^q)$ for the spatial error. Sometimes this can be improved.

Lemma 7.1. Suppose the stability assumption (7.4) holds, $w(0)=w_h(0)$ and we have the decomposition

$$\sigma_h(t) = A\xi(t)+\eta(t) \quad \text{with} \quad \|\xi(t)\|,\|\xi'(t)\|,\|\eta(t)\| \le Ch^r$$

for $0\le t\le T$. Then $\|w_h(t)-w(t)\|\le C'h^r$ for $0\le t\le T$, with C' depending on C,K and T.

Proof. Let $\varepsilon(t)=w_h(t)-w(t)$. Then $\varepsilon(0)=0$ and

$$\varepsilon'(t) = A\varepsilon(t)+\sigma_h(t) = A\Big(\varepsilon(t)+\xi(t)\Big)+\eta(t).$$

Hence $\hat\varepsilon(t)=\varepsilon(t)+\xi(t)$ satisfies

$$\hat\varepsilon'(t) = A\hat\varepsilon(t)+\xi'(t)+\eta(t), \quad \hat\varepsilon(0)=\xi(0).$$

In the same way as in section 2, it follows that

$$\|\hat\varepsilon(t)\| \le K\|\xi(0)\| + Kt\max_{0\le s\le t}\|\xi'(s)+\eta(s)\|,$$

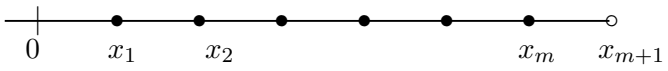
and since $\|\varepsilon(t)\|\le\|\hat\varepsilon(t)\|+\|\xi(t)\|$, the estimate for $\|\varepsilon(t)\|$ also follows. □

Due to negative powers of h contained in A, the assumption of the lemma does not imply $\|\sigma_h(t)\| = \mathcal{O}(h^r)$. So, the above result tells us that we can have convergence of order r while the truncation error has a lower order. We shall consider some simple examples where this can occur at the boundaries.

Note. Similar phenomena also occur when *nonuniform grids* are considered. For (technical) results on this, the interested reader is referred to the paper of Manteuffel & White (1986).

Example : outflow with central advection discretization

Consider the advection equation $u_t + u_x = 0$, for $0 \le x \le 1$, with given inflow condition $u(0,t) = \gamma_0(t)$ and initial profile $u(x,0)$. Let $h = 1/m$ and $x_j = jh$ for $j = 0, 1, ..., m$.



Second order central discretization gives

$$w_j'(t) = \frac{1}{2h}\Big(w_{j-1}(t) - w_{j+1}(t)\Big), \quad j = 1, 2, ..., m,$$

with $w_0(t) = \gamma_0(t)$. Here $w_{m+1}(t)$ represents the value at the virtual point $x_{m+1} = 1 + h$. This value can be found by extrapolation, for example,

$$w_{m+1}(t) = \theta w_m(t) + (1 - \theta) w_{m-1}(t).$$

We consider $\theta = 1$ (constant extrapolation) and $\theta = 2$ (linear extrapolation). This last choice seems more natural; in fact we then apply the 1-st order upwind discretization at the outflow point.

For the spatial truncation error $\sigma_h(t) = (\sigma_{h,1}(t), ..., \sigma_{h,m}(t))^T$ we find $\sigma_{h,j}(t) = \mathcal{O}(h^2)$ for $j < m$, whereas at the outflow point

$$\sigma_{h,m}(t) = \frac{d}{dt}u(t, x_m) - \frac{1}{2h}\Big(\theta u(t, x_{m-1}) - \theta u(t, x_m)\Big) =$$

$$= -\frac{1}{2}(2 - \theta)u_x - \frac{1}{4}\theta h u_{xx} + \cdots \quad \Big|_{(x_m,t)}.$$

So, for the space truncation error we have the bounds

$$\|\sigma_h\|_\infty = \mathcal{O}(h^s), \quad \|\sigma_h\|_2 = \mathcal{O}(h^{s+\frac{1}{2}}), \quad \|\sigma_h\|_1 = \mathcal{O}(h^{s+1})$$

in the L_∞, L_2 and L_1 norms, with $s = 0$ if $\theta = 1$ and $s = 1$ if $\theta = 2$.

Numerical experiments, however, show that $\|w_h(t) - w(t)\| = \mathcal{O}(h^{s+1})$ for all three norms. This is in accordance with Lemma 7.1. We have

$$A = \frac{1}{2h}\begin{pmatrix} 0 & -1 & & & \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \\ & & & \theta & -\theta \end{pmatrix}, \qquad \sigma_h = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} Ch^s + \mathcal{O}(h^{s+1}),$$

with $C = -\frac{1}{2}u_x(1,t)$ if $\theta = 1$, and $C = -\frac{1}{2}u_{xx}(1,t)$ if $\theta = 2$. If we ignore the higher order terms in σ_h, then $A\xi = \sigma_h$ gives

$$\xi_{j-1} - \xi_{j+1} = 0 \quad (j = 1, 2, ..., m \quad \text{with} \quad \xi_0 = 0),$$

$$\theta\xi_{m-1} - \theta\xi_m = 2Ch^{s+1}.$$

Hence $\xi = (\xi_1, 0, \xi_1, 0, ...)^T$ with $\xi_1 = \pm 2\theta^{-1}Ch^{s+1}$, the sign depending on the parity of m, and thus we find $\|\xi\| = \mathcal{O}(h^{s+1})$ in the L_1, L_2 and L_∞ norms.

For a complete convergence proof we need stability results. In the present example this is easy in the L_2-norm. Consider the inner product on $\mathbb{R}^m$,

$$(u, v) = h\sum_{j=1}^{m} \delta_j u_j v_j, \quad \text{with } \delta_j = 1 \text{ for } j < m \text{ and } \delta_m = 1/\theta,$$

and corresponding norm $\|v\| = (v, v)^{1/2}$. We have, for any $v \in \mathbb{R}^m$,

$$(v, Av) = -\frac{1}{2}v_m^2 \le 0.$$

Hence, if $u'(t) = Au(t)$ then

$$\frac{d}{dt}\|u(t)\|^2 = \frac{d}{dt}(u(t), u(t)) = 2(u(t), u'(t)) = 2(u(t), Au(t)) \le 0,$$

showing that $\|u(t)\|$ is nonincreasing. Consequently $\|e^{tA}\| \le 1$ for $t \ge 0$.

If $\theta = 1$ the norm $\|\cdot\|$ is the L_2-norm. For $\theta = 2$ it is equivalent to the L_2-norm,

$$\|v\|_2^2 \ \ge \ \|v\|^2 = \|v\|_2^2 - \frac{1}{2}hv_m^2 \ \ge \ \frac{1}{2}\|v\|_2^2,$$

and so in this case (7.4) holds with $K = \sqrt{2}$, in the L_2-norm.

Note. The L_2-convergence result in the above example is basically due to Gustafsson (1975). The results in that paper are more general (hyperbolic systems with multi-step time integration), but the derivations are also much more complicated.

Example : Neumann boundary condition for diffusion

Consider the diffusion test problem $u_t = u_{xx}$ with $u(0, t) = \gamma_0(t)$ and $u_x(1, t) = 0$, with the same grid as above. Second order central differences now give

$$w_j'(t) = \frac{1}{h^2}\Big(w_{j-1}(t) - 2w_j(t) + w_{j+1}(t)\Big), \quad j = 1, 2, ..., m,$$

with $w_0(t) = \gamma_0(t)$. The Neumann condition at $x = 1$ can be discretized as $h^{-1}(w_{m+1}(t) - w_m(t)) = 0$ or as $(2h)^{-1}(w_{m+1}(t) - w_{m-1}(t)) = 0$. Thus we set, with parameter $\theta = 0$ or 1,

$$w_{m+1}(t) = \theta w_m(t) + (1 - \theta)w_{m-1}(t).$$

For a smooth solution it can be assumed that both the differential equation and the Neumann condition are valid at $x_m = 1$. This implies that $u_x(1, t) = u_{xxx}(1, t) = \cdots = 0$. Inserting the exact solution in the difference scheme, we find a 2-nd order truncation error, except at x_m where

$$\sigma_{h,m}(t) = \frac{1}{2}\theta u_{xx}(1, t) + \mathcal{O}(h^2).$$

So, if $\theta = 0$ we have an $\mathcal{O}(h^2)$ truncation error. If $\theta = 1$ we have an inconsistency at $x_m = 1$, but still we can prove first order convergence: ignoring the $\mathcal{O}(h^2)$ terms we have $A\xi = \sigma_h$ if

$$\xi_{j-1} - 2\xi_j + \xi_{j+1} = 0 \ (j = 1, 2, ..., m-1; \xi_0 = 0), \quad \xi_{m-1} - \xi_m = \frac{1}{2-\theta} C h^2,$$

with $C = \frac{1}{2}\theta u_{xx}(1,t)$, giving $\xi_j = -j(2-\theta)^{-1} C h^2$ for $1 \le j \le m$. Hence $\|\xi\| = \mathcal{O}(h)$ in the L_1, L_2 and L_∞ norms.

Stability in the L_2-norm can be proven here just as in the previous example. In the present example we have $\|e^{tA}\|_\infty \le 1$, due to diagonal dominance in the rows, as can be seen from the formula for the logarithmic norm (2.13).

Remark. The choice $w_{m+1} \equiv w_{m-1}$ in this example presents itself in a natural way if we consider, instead of $u(x,t)$ for $0 \le x \le 1$, the function $\bar{u}(x,t)$ for $0 \le x \le 2$, defined by

$$\bar{u}(x,t) = u(x,t) \quad \text{for } 0 \le x \le 1,$$

$$\bar{u}(x,t) = u(1-x,t) \quad \text{for } 1 \le x \le 2.$$

For $\bar{u}$ we then have $\bar{u}_t = \bar{u}_{xx} (0 \le x \le 2)$, $\bar{u}(0,t) = \bar{u}(2,t) = \gamma_0(t)$, and the Neumann condition at $x = 1$ is automatically fulfilled due to symmetry around the point $x = 1$. Discretizing this extended problem with central differences will give the same symmetry in the semi-discrete system, so that $\bar{w}_{m+j}(t) = \bar{w}_{m-j}(t)$.

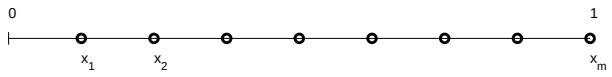
Numerical example: diffusion on cell/vertex centered grids

In the previous two examples the grid was chosen such that the boundaries did coincide with grid points. If fluxes are prescribed on the boundaries it would be more natural to let the boundaries coincide with cell vertices. As an example we consider

$$u_t = u_{xx}, \quad u(0,t) = \gamma_0(t), \quad u_x(1,t) = \gamma_1(t)$$

for $0 \le x \le 1$, $t \ge 0$ with given initial profile, and we shall present numerical results on different grids.

Vertex centered grid:

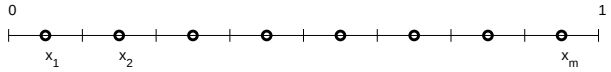


Let $x_i = ih$ with $h = 1/m$. Using $x_{m+1} = 1 + h$ as virtual point with value w_{m+1} such that $\frac{1}{2h}(w_{m+1} - w_{m-1}) = \gamma_1$, we obtain the system

$$\begin{cases} w_1' = \frac{1}{h^2}(-2w_1 + w_2) + \frac{1}{h^2}\gamma_0, \\ w_j' = \frac{1}{h^2}(w_{j-1} - 2w_j + w_{j+1}), \quad j = 2, ..., m-1, \\ w_m' = \frac{1}{h^2}(2w_{m-1} - 2w_m) + \frac{2}{h}\gamma_1. \end{cases}$$

The truncation error is $\mathcal{O}(h^2)$ at all points except at the right boundary where we find $\sigma_m = \frac{1}{3}h u_{xxx}(1,t) + \mathcal{O}(h^2)$.

Cell centered grid:



Let $x_i = (i - \frac{1}{2})h$ with $h = 1/m$. Now the right boundary condition fits in a natural way but we have to think about how to implement the Dirichlet condition at the left. Using the virtual point $x_0 = -\frac{1}{2}h$ with virtual value w_0 such that $\frac{1}{2}(w_0 + w_1) = \gamma_0$, we now obtain the semi-discrete system

$$\begin{cases} w_1' = \frac{1}{h^2}(-3w_1 + w_2) + \frac{2}{h^2}\gamma_0, \\ w_j' = \frac{1}{h^2}(w_{j-1} - 2w_j + w_{j+1}), \quad j = 2, ..., m-1, \\ w_m' = \frac{1}{h^2}(w_{m-1} - w_m) + \frac{1}{h}\gamma_1. \end{cases}$$

In this case we even have an inconsistency at the left, $\sigma_1 = \frac{1}{4}u_{xx}(0,t) + \mathcal{O}(h^2)$, at the right boundary we get $\sigma_m = \frac{1}{24}hu_{xxx}(1,t) + \mathcal{O}(h^2)$, and elsewhere the truncation errors are $\mathcal{O}(h^2)$.

The names vertex/cell centered are used here because of the relationship with finite volume schemes, see Morton (1996, p. 216). We can also combine the grids by taking $h = 1/(m + \frac{1}{2})$ and $x_j = jh$, so that now on both sides the boundary conditions fit naturally. On the left we get the vertex centered discretization and on the right the cell centered. We shall refer to this as "hybrid".

In the following table the errors on the three grids are given in the max-norm and L_2-norm for the solution $u(x,t) = 1 + e^{-\frac{1}{4}\pi^2 t}\cos(\frac{1}{2}\pi x)$ at output time $T = \frac{1}{4}$. Obviously there is no reduction in accuracy. Even with the cell centered case, where the truncation error is only $\mathcal{O}(h^0)$ at the left boundary, we find second order convergence in the max-norm. This can be explained just as in the previous examples (left as exercise). Further it should be noted that although the rates of convergence seem the same in the three cases, the error constants are smallest in the hybrid case.

#points m	vertex centered		cell centered		hybrid	
	L_2-error	L_∞-error	L_2-error	L_∞-error	L_2-error	L_∞-error
10	$.11\ 10^{-2}$	$.21\ 10^{-2}$	$.12\ 10^{-2}$	$.17\ 10^{-2}$	$.11\ 10^{-3}$	$.19\ 10^{-3}$
20	$.26\ 10^{-3}$	$.52\ 10^{-3}$	$.32\ 10^{-3}$	$.42\ 10^{-3}$	$.29\ 10^{-4}$	$.56\ 10^{-4}$
40	$.63\ 10^{-4}$	$.13\ 10^{-3}$	$.79\ 10^{-4}$	$.10\ 10^{-3}$	$.76\ 10^{-5}$	$.15\ 10^{-4}$
80	$.16\ 10^{-4}$	$.33\ 10^{-4}$	$.20\ 10^{-4}$	$.26\ 10^{-4}$	$.19\ 10^{-5}$	$.39\ 10^{-5}$

TABLE 7.1. Spatial errors with vertex/cell centered grids.

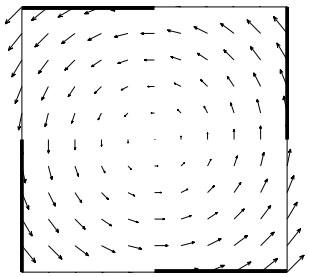
7.2. Local grid refinements

As we saw, boundary conditions may give rise to boundary layers in advection-diffusion problems with large Péclet numbers. To maintain accuracy in such a situation a refinements of the grid near that boundary will be needed to be able to represent the solution properly on the grid.

Strong spatial gradients may also be caused by local source terms, non-smooth initial profiles or nonlinear reaction terms. In case the location of these gradients is not known in advance, a local refinement of the grid should adapt itself to the solution. A relatively simple procedure of this kind has been derived by Trompert & Verwer (1991). Basically their method works as follows: for a time step $t_n \mapsto t_{n+1}$ one first performs a time step on a coarse grid. In those regions were one is not satisfied with the solution (for instance if an estimate for $|u_{xx}(x, t_{n+1})|$ is too large) the grid is is refined by bisection, and there the step $t_n \mapsto t_{n+1}$ is redone. For this step on a part of the grid, starting values may already be present, otherwise they are found by interpolation, and likewise for boundary values. For a detailed description of this process we refer to the paper of Trompert & Verwer (1991). References on related approaches can also be found in that paper.

As an illustration of grid refinement, we consider the so-called Molenkamp-Crowley test, which consists of the 2D advection equation

$$u_t + (au)_x + (bu)_y = 0$$



with $t \geq 0$ and $0 \leq x, y \leq 1$ and with given velocities

$$a(x, y) = -2\pi(y - \frac{1}{2}), \quad b(x, y) = 2\pi(x - \frac{1}{2}).$$

With this velocity field any initial profile is rotated around the center of the domain. At time $t = 1$ one rotation will be completed. Dirichlet conditions are prescribed at the inflow boundaries. The initial profile is a cylinder with height 1, radius 0.1 and center $(\frac{1}{2}, \frac{3}{4})$.

Figure 7.1 gives the numerical results after one rotation on a uniform grid with $h = 1/80$ and with locally refined grids using $h = 1/10, 1/20, 1/40, 1/80$. Spatial discretization is done as in Section 5 (limiter with $\mu = 1$). At the boundaries quadratic extrapolation is used to find missing values outside the domain. In the interior, at the mesh interfaces, linear extrapolation is used. Time integration is done with the classical 4-th order Runge-Kutta method with sufficiently small time steps, so that the temporal errors are not visible. The solution on the locally refined grid has the same quality as the solution on the overall fine grid.

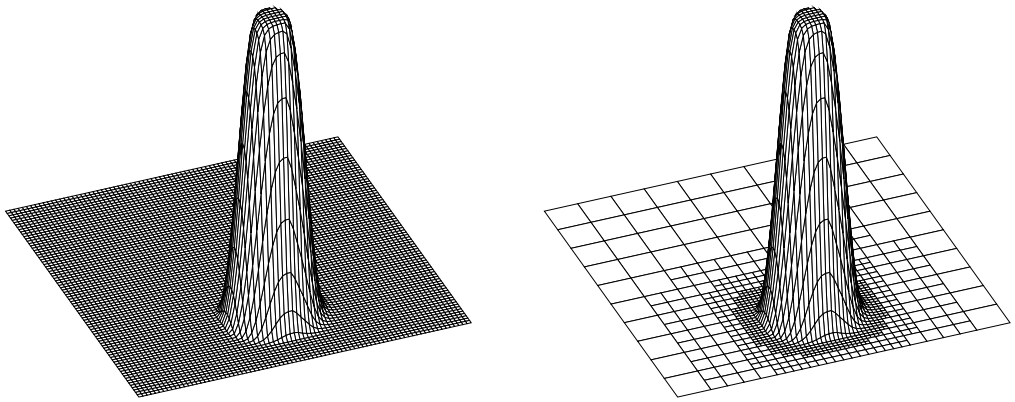


Figure 7.1. Rotating cylinder in Molenkamp test.

Application of local grid refinements to a smog prediction model is discussed in the thesis of van Loon (1996). The model is used at RIVM to give a smog forecast for several days, using meteo data from a whether prediction model. The domain covers a large part of Europe, whereas the region of interest is the Netherlands. The domain is taken so large to avoid influence of boundary conditions, which are not well known. Local grid refinements are used to improve accuracy in the region of interest, without introducing too many grid points.

Figure 7.2 gives a numerical prediction (5 day period, ending at 24-7-1989, 14:00 MET) for the ozone concentrations over Europe in $\mu g\ m^{-3}$. The coarse grid solution, with 52×55 cells, is given in the left picture. For the picture on the right 4 levels of refinement were used in the central region, giving higher ozone concentrations over the Netherlands and southern England. The locally refined solution corresponds better with actual observations of that date, see van Loon (1996).

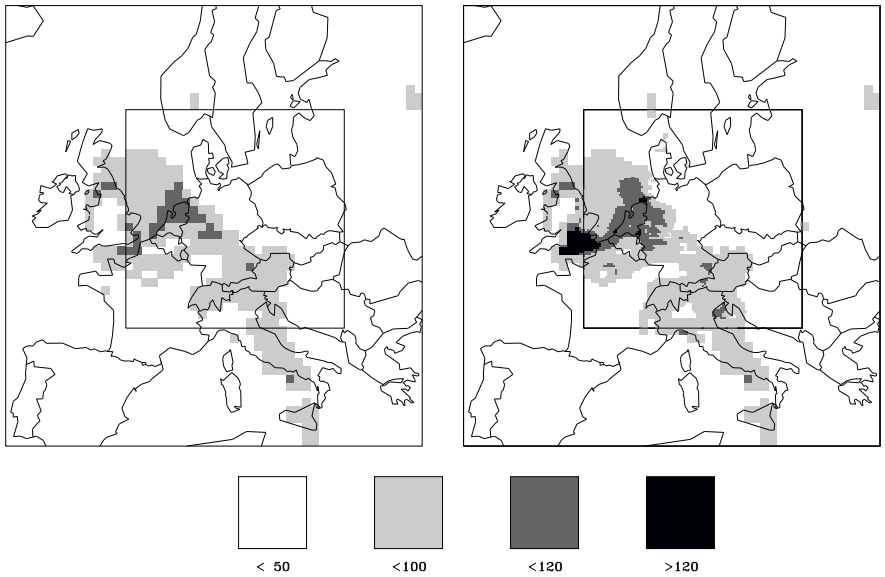


Figure 7.2. Computed O_3 distribution. Coarse grid (left) and with 4 levels refinement (right).

Note. A code based on this refinement procedure for general parabolic problems in 2D or 3D has been written by J. Blom. The code is easy to use and it can be obtained at the URL "`www.cwi.nl/ gollum`" under the headings VLUGR2 and VLUGR3.

8. Boundary conditions and temporal accuracy

Surprisingly, boundary conditions may also have a negative impact on the *temporal* accuracy. For standard one-step methods, of the Runge-Kutta type, this phenomenon will usually only manifest itself for high accuracies. As an illustration we consider

$$u_t + u_x = u^2 \quad \text{for } 0 \le x \le 1, \;\; 0 \le t \le 1/2, \tag{8.1}$$

with solution $u(x,t) = \sin^2(\pi(x-t))/(1 - t\sin^2(\pi(x-t)))$. We assume that $u(x,0)$ is given, together with either the inflow Dirichlet condition

$$u(0,t) = \sin^2(\pi t)/(1 - t\sin^2(\pi t)), \tag{8.2}$$

or the periodicity condition

$$u(x,t) = u(x \pm 1, t). \tag{8.3}$$

For a numerical experiment, we consider space discretization for u_x with 4-th order central differences, see Section 4. With the Dirichlet conditions (8.2) we use 3-th order differences (4 point stencil) at points near the boundaries. In view of the results of the previous section, we then still expect a 4-th order spatial error. Time discretization is done with the classical 4-th order Runge-Kutta method. We consider $\tau, h \to 0$ with fixed Courant number $\tau/h = 2$.

Bound. cond.	h	L_2-error	L_∞-error
Dirichlet (8.2)	h = 1/ 40	0.18 10^{-3}	0.30 10^{-3}
	h = 1/ 80	0.13 10^{-4} (*3.80*)	0.24 10^{-4} (*3.57*)
	h = 1/160	0.86 10^{-6} (*3.90*)	0.19 10^{-5} (*3.75*)
	h = 1/320	0.56 10^{-7} (*3.96*)	0.12 10^{-6} (*3.91*)
	h = 1/640	0.35 10^{-8} (*3.98*)	0.79 10^{-8} (*3.96*)
Periodic (8.3)	h = 1/ 40	0.17 10^{-3}	0.21 10^{-3}
	h = 1/ 80	0.11 10^{-4} (*3.98*)	0.14 10^{-4} (*3.98*)
	h = 1/160	0.67 10^{-6} (*3.99*)	0.85 10^{-6} (*3.99*)
	h = 1/320	0.42 10^{-7} (*4.00*)	0.53 10^{-7} (*4.00*)
	h = 1/640	0.26 10^{-8} (*4.00*)	0.33 10^{-8} (*4.00*)

Table 8.1. Spatial errors (and estimated orders) with Dirichlet and periodicity conditions.

Numerical results are given in Table 8.1 for the spatial errors at time $t = 1/2$ in L_2 and L_∞ norms, together with the estimated orders, showing 4-th order convergence for both cases (8.2) and (8.3). Table 8.2 gives the errors if we use the Runge-Kutta method. Now, with Dirichlet conditions there is a clear *order reduction*, we get approximately order 2.5 in the L_2-norm and order 2 in the L_∞-norm.

Bound. cond.	$\tau = 2h$	L_2-error	L_∞-error
Dirichlet (8.2)	$\tau = 1/\ 20$	$0.76\ 10^{-3}$	$0.13\ 10^{-2}$
	$\tau = 1/\ 40$	$0.68\ 10^{-4}$ (*3.48*)	$0.16\ 10^{-3}$ (*2.96*
	$\tau = 1/\ 80$	$0.95\ 10^{-5}$ (*2.84*)	$0.46\ 10^{-4}$ (*1.83*)
	$\tau = 1/160$	$0.17\ 10^{-5}$ (*2.52*)	$0.12\ 10^{-4}$ (*1.98*)
	$\tau = 1/320$	$0.30\ 10^{-6}$ (*2.48*)	$0.29\ 10^{-5}$ (*1.99*)
Periodic (8.3)	$\tau = 1/\ 20$	$0.75\ 10^{-3}$	$0.11\ 10^{-2}$
	$\tau = 1/\ 40$	$0.55\ 10^{-4}$ (*3.76*)	$0.87\ 10^{-4}$ (*3.72*)
	$\tau = 1/\ 80$	$0.37\ 10^{-5}$ (*3.90*)	$0.59\ 10^{-5}$ (*3.88*)
	$\tau = 1/160$	$0.24\ 10^{-6}$ (*3.95*)	$0.38\ 10^{-6}$ (*3.95*)
	$\tau = 1/320$	$0.15\ 10^{-7}$ (*3.98*)	$0.24\ 10^{-7}$ (*3.97*)

TABLE 8.2. Errors (and estimated orders) for RK4, $\tau = 2h$, with Dirichlet and periodicity conditions.

It should be noted that the results of Table 8.1 were also found numerically with the 4-th order Runge-Kutta method, but there a very small time step was chosen, and it was experimentally verified that temporal errors were negligible.

In this section we want to explain the result of Table 8.2. This will be done in a general framework, also including implicit methods.

Throughout the section a given norm $\|\cdot\|$ on $\mathbb{R}^m$ is considered and the following notation will be used: for $v \in \mathbb{R}^m$ depending on τ, h, we write $v = \mathcal{O}(\tau^\alpha h^\beta)$ if it holds that $\|v\| \le C\tau^\alpha h^\beta$ with $C > 0$ *independent* of τ and h. So, in particular $v = \mathcal{O}(\tau^\alpha)$ means that no negative powers of h are hidden in the bound. The same notation is also used for matrices. If it is necessary to specify the norm we write $\|v\| = \mathcal{O}(\tau^\alpha h^\beta)$.

8.1. Local error analysis

Consider a linear semi-discrete system in $\mathbb{R}^m$,

$$w'(t) = Aw(t) + g(t), \quad w(0) = w_0, \tag{8.4}$$

with a smooth solution so that $w^{(k)}(t) = \mathcal{O}(1)$ for all derivatives arising in the analysis. If the underlying PDE problem has non-homogeneous boundary conditions, these boundary data are incorporated in $g(t)$, together with genuine source terms. We assume for the moment that $w(t) = w_h(t)$, that is, spatial errors are neglected. Application of a one-step method, say Runge-Kutta type, will lead to a recursion

$$w_{n+1} = R(\tau A)w_n + \sum_{j=1}^{s} Q_j(\tau A)\tau g(t_n + c_j\tau), \tag{8.5}$$

with stability function R and rational functions Q_j determined by the method, see also Section 6. If we insert the exact solution into (8.4), we get a local error δ_n,

$$w(t_{n+1}) = R(\tau A)w(t_n) + \sum_{j=1}^{s} Q_j(\tau A)\tau g(t_n + c_j\tau) + \delta_n. \tag{8.6}$$

Note that this δ_n is the error which is made in one single step, that is, if we have $w_n = w(t_n)$ then $\delta_n = w(t_{n+1}) - w_{n+1}$.

Using $g(t) = w'(t) - Aw(t)$ we can express δ_n as a Taylor series in terms of the exact solution w and its derivatives,

$$\delta_n = \sum_{k\geq 0} \frac{1}{k!} H_k(\tau A)\tau^k w^{(k)}(t_n) \tag{8.7}$$

with rational functions

$$H_0(z) = 1 - R(z) + z\sum_{j=1}^{s} Q_j(z), \qquad H_k(z) = 1 + \sum_{j=1}^{s}(zc_j^k - kc_j^{k-1})Q_j(z) \quad \text{for } k \geq 1.$$

The Taylor expansion can, of course, be truncated at any level τ^k with a remainder term proportional to τ^{k+1}, involving derivatives $w_i^{(k+1)}$ of the components $i = 1, ..., m$ at intermediate points in $[t_n, t_{n+1}]$.

We assume in the following that the integers p and q are such that

$$H_0(z) = H_1(z) = \cdots = H_q(z) \equiv 0 \tag{8.8}$$

and

$$\text{the method has order } p. \tag{8.9}$$

The first condition means that the method is exact if $w(t)$ is a polynomial of degree q or less. In the second condition the order refers to the standard concept for ODEs, and so this means that $\delta_n = \mathcal{O}(\tau^{p+1})$ provided that $A = \mathcal{O}(1)$, the non-stiff case. Note that for semi-discrete PDEs we will have $A \sim h^{-k}$ with $k = 1$ for advection and $k = 2$ for diffusion, and thus the non-stiff estimate is not applicable. We do have $q \leq p$ and

$$H_k(z) = \mathcal{O}(z^{p+1-k}),\ z \to 0 \quad \text{for } q+1 \leq k \leq p. \tag{8.10}$$

(This can be seen by considering the scalar equation with $A = \lambda$, $|\lambda| = 1$, $w(t) = \frac{1}{k!}t^k$ and $n = 0$.)

In general, we can formulate assumptions such that $H_k(\tau A) = \mathcal{O}(1)$, but this only gives the estimate $\delta_n = \mathcal{O}(\tau^{q+1})$. For local error bounds applicable to semi-discrete PDEs property (8.10) does not necesserily lead to higher order estimates.

Example. For the classical Runge-Kutta method, see Section 6, we have $p = 4$, $q = 1$ and

$$H_2(z) = \frac{1}{48}z^3.$$

Therefore, the leading error term in (8.7) is given by

$$\delta_n^* = \frac{\tau^2}{96}[\tau A]^3 w''(t_n).$$

For stability with this explicit method we have to impose a step size restriction such that $\tau A = \mathcal{O}(1)$, which leads to the local error bound $\delta_n^* = \mathcal{O}(\tau^2)$. If we know, in addition, that $Aw''(t) = \mathcal{O}(1)$ then we get the bound

$$\delta_n^* = \frac{1}{96}\tau^3[\tau A]^2[Aw''(t_n)] = \mathcal{O}(\tau^3).$$

Likewise, if $A^2 w''(t) = \mathcal{O}(1)$, then $\delta_n^* = \mathcal{O}(\tau^4)$, and so on. However, whether $A^k w''(t) = \mathcal{O}(1)$ is true or not will depend on the *boundary conditions*. $\diamond$

Example. Condider the familiar example, arising from 1-st order upwind advection with inflow Dirichlet condition,

$$A = \frac{1}{h} \begin{pmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

and consider a vector $v = (v_j) \in \mathbb{R}^m$ with $v_j = \psi(x_j), x_j = jh$, for some fixed, smooth function ψ, for instance $\psi = u_{tt}$. Then

$$Av = -\frac{1}{h}\psi(0) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \psi_x(x_1) \\ \psi_x(x_2) \\ \vdots \\ \psi_x(x_m) \end{pmatrix} + \cdots,$$

and therefore $Av = \mathcal{O}(1)$ in L_2 and L_∞ norms iff $\psi(0) = 0$. Otherwise we will have $\|Av\|_2 \sim h^{-1/2}$ and $\|Av\|_\infty \sim h^{-1}$.

In case that $\psi(0) = 0$, we have

$$A^2 v = -\frac{1}{h}\psi_x(0) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} \psi_{xx}(x_1) \\ \psi_{xx}(x_2) \\ \vdots \\ \psi_{xx}(x_m) \end{pmatrix} + \cdots,$$

and thus we see that for having $A^2 v = \mathcal{O}(1)$ in L_2 and L_∞ norms we need $\psi(0) = \psi_x(0) = 0$. Likewise for higher powers of A.

On the other hand, for

$$B = \frac{1}{h} \begin{pmatrix} -1 & & & 1 \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

and ψ a smooth periodic function we get simply

$$Bv = (\psi_x(x_j)) + \mathcal{O}(h), \qquad B^2 v = (\psi_{xx}(x_j)) + \mathcal{O}(h), \qquad \cdots .$$

Higher order discretizations of advection and diffusion, with Dirichlet or Neumann boundary conditions, or with periodicity conditions, can be considered in a similar way. $\diamond$

In view of the above, the result of Table 8.2 for the periodic case (8.3) does not come as a surprise. With periodicity conditions we get local errors of $\mathcal{O}(\tau^{p+1})$ and thus global errors of $\mathcal{O}(\tau^p)$.

Also the fact that with Dirichlet conditions a lower order of convergence was observed is no longer surprising. Note however that the results for the Dirichlet case (8.2) are still not well explained. Since (8.2) gives an inhomogeneous Dirichlet condition, the above suggests that we will have $\|\delta_n\|_\infty = \mathcal{O}(\tau^2)$ and $\|\delta_n\|_2 = \mathcal{O}(\tau^{2.5})$ for the *local* errors. In Table 8.2, however, the accumulated *global* errors are given.

8.2. Global error analysis

Consider a recursion

$$\varepsilon_{n+1} = S\varepsilon_n + \delta_n \quad (n = 0, 1, ..., N), \qquad \varepsilon_0 = 0,$$

with stability assumption $\|S^n\| \le K$ for all $n = 0, 1, ..., N$. Here δ_n and ε_n will stand for the local and global errors, respectively.

Lemma 8.1. Suppose that

$$\delta_n = (I - S)\xi_n + \eta_n$$

with $\|\xi_n\| \le C\tau^r$, $\|\eta_n\| \le C\tau^{r+1}$ and $\|\xi_{n+1} - \xi_n\| \le C\tau^{r+1}$ for all n. Then there is a $C' > 0$, depending on C, K and $T = N\tau$, such that $\|\varepsilon_n\| \le C'\tau^r$ for all $0 \le n \le N$.

Proof. We have

$$\varepsilon_{n+1} = S\varepsilon_n + (I - S)\xi_n + \eta_n, \quad \varepsilon_0 = 0.$$

Introducing $\hat{\varepsilon}_n = \varepsilon_n - \xi_n$, we get

$$\hat{\varepsilon}_{n+1} = S\hat{\varepsilon}_n + \eta_n - (\xi_{n+1} - \xi_n), \quad \hat{\varepsilon}_0 = \xi_0.$$

This gives in the standard way $\hat{\varepsilon}_n = \mathcal{O}(\tau^r)$, and thus also $\varepsilon_n = \mathcal{O}(\tau^r)$, with constants determined by C, K and T. □

We note that the decomposition of the local error δ_n, as used in this lemma, can also be shown to be necessary for having $\varepsilon_n = \mathcal{O}(\tau^r)$ in case the δ_n are constant, see Hundsdorfer (1992).

The above lemma will be applied with $\varepsilon_n = w(t_n) - w_n$. By subtracting (8.5) from (8.6), we see that these global errors satisfy

$$\varepsilon_{n+1} = R(\tau A)\varepsilon_n + \delta_n, \quad \varepsilon_0 = 0, \tag{8.11}$$

with local errors δ_n given by (8.6).

To understand the behaviour of the errors it sufficient to consider the leading error term in δ_n. The contribution of the other terms to the global error is found in a similar way. So, we consider here

$$\delta_n = \frac{1}{(q+1)!} H_{q+1}(\tau A)\tau^{q+1} w^{(q+1)}(t_n). \tag{8.12}$$

Define, for $\alpha \ge 0$,

$$\varphi_\alpha(z) = (1 - R(z))^{-1} H_{q+1}(z) z^{-\alpha}.$$

Theorem 8.2. Consider the recursion (8.11),(8.12). Assume $\|R(\tau A)^n\| \le K$ for all n, and

$$\|\varphi_\alpha(\tau A)\| = \mathcal{O}(1), \qquad A^\alpha w^{(q+j)}(t) = \mathcal{O}(1) \quad \text{for } j = 1, 2, \tag{8.13}$$

uniformly in $t \in [0,T]$. Then $\varepsilon_n = \mathcal{O}(\tau^{q+1+\alpha})$ for $n\tau \le T$.

Proof. This is a consequence of Lemma 8.1. Take $\eta_n = 0$ and

$$\xi_n = \frac{1}{(q+1)!}\tau^{q+1+\alpha}\varphi_\alpha(\tau A)\Big(A^\alpha w^{(q+1)}(t_n)\Big).$$

□

To study the assumptions of the theorem, consider the discrete L_2-norm. Let $\mathcal{S}$ be the stability region of the method, and let $\mathcal{D} \subset \mathcal{S}$. We assume that A is diagonalizable, $A = V\Lambda V^{-1}$, with $\mathrm{cond}(V) = K = \mathcal{O}(1)$ and $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_m)$ such that

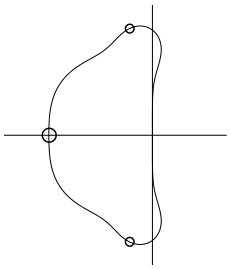
$$\tau\lambda_j \in \mathcal{D} \subset \mathcal{S}. \tag{8.14}$$

Then we have stability, $\|R(\tau A)^n\| \le K$ for all n. If we assume in addition that

$$|\varphi_\alpha(z)| \le C \quad \text{for all } z \in \mathcal{D}, \tag{8.15}$$

then $\|\varphi_\alpha(\tau A)\| = \mathcal{O}(1)$.

To apply this result we have to chose some suitable region $\mathcal{D} \subset \mathcal{S}$. We want the point 0 to be on its boundary, so that the result can be applied for a step size interval $(0, \tau_0]$. For this we need boundedness of $\varphi_\alpha(z)$ near $z = 0$. This holds for $\alpha \le p - q - 1$, due to (8.10). Further we can take $\mathcal{D}$ arbitrary in $\mathcal{S}$, except for points $z \ne 0$ on the boundary of $\mathcal{S}$ where $R(z) = 1$.



Example. For the classical Runge-Kutta method we have

$$\varphi_\alpha(z) = -\frac{1}{48}\,\frac{z^{2-\alpha}}{1 + \frac{1}{2}z + \frac{1}{6}z^2 + \frac{1}{24}z^3},$$

which is bounded near 0 if $\alpha \le 2$. The L_2-order 2.5 result of Table 8.2 follows if we can show that $A^\alpha w''(t) = \mathcal{O}(1)$ for α up to 1/2. This seems difficult to proof. An alternative is to take $\alpha = 1$, write the local error as

$$\delta_n = \Big(I - R(\tau A)\Big)\ \varphi_1(\tau A)\ \tau^{2.5}\Big(\tau^{0.5} A w''(t_n)\Big),$$

and then use the fact that for τ/h constant we will have $\|\tau^{0.5} A w''(t_n)\|_2 = \mathcal{O}(1)$. We also note that the order 2 convergence in the max-norm in Table 8.2 indicates that $\|\varphi_0(\tau A)\|_\infty = \mathcal{O}(1)$. ◇

Example. The implicit midpoint rule,

$$w_{n+1} = w_n + \tau F(t_{n+1/2}, \frac{1}{2}w_n + \frac{1}{2}w_{n+1}), \tag{8.16}$$

gives the form (8.5) with $s = 1, c_1 = \frac{1}{2}$ and

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, \qquad Q_1(z) = \frac{1}{1 - \frac{1}{2}z}.$$

We have $p = 2$, $q = 1$ and $H_2(z) = -\frac{1}{4}z/(1 - \frac{1}{2}z)$. The local error

$$\delta_n = -\frac{1}{8}(I - \frac{1}{2}\tau A)^{-1}\tau A\, \tau^2 w''(t_n) + \mathcal{O}(\tau^3)$$

will give rise to local order reduction unless $Aw''(t) = \mathcal{O}(1)$. For example, for the parabolic initial-boundary value problem $u_t = u_{xx} + f(t)$ with time-dependent Dirichlet conditions one can observe $\|\delta_n\|_2 = \mathcal{O}(\tau^{2.25})$ and $\|\delta_n\|_\infty = \mathcal{O}(\tau^2)$, see Verwer (1986). However, by noting that

$$\delta_n = -\frac{1}{8}\Big(I - R(\tau A)\Big)\, \tau^2 w''(t_n) + \mathcal{O}(\tau^3),$$

we see that the global error will show nicely an $\mathcal{O}(\tau^2)$ behaviour if we have stability, even in case $Aw''(t) \neq \mathcal{O}(1)$. ◇

Example. For the trapezoidal rule,

$$w_{n+1} = w_n + \frac{1}{2}\tau F(t_n, w_n) + \frac{1}{2}\tau F(t_{n+1}, w_{n+1}), \tag{8.17}$$

we get $s = 2$, $c_1 = 0$, $c_2 = 1$ and

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z}, \qquad Q_1(z) = Q_2(z) = \frac{\frac{1}{2}}{1 - \frac{1}{2}z}.$$

Here we have $p = q = 2$, so no order reduction will take place. ◇

8.3. The total space-time error

For simplicity, the spatial errors were neglected in this section. These errors can be simply included in the analysis. If we insert $w_h(t)$, the restriction of the PDE solution to the spatial grid Ω_h, into (8.5), we obtain

$$w_h(t_{n+1}) = R(\tau A)w_h(t_n) + \sum_{j=1}^{s} Q_j(\tau A)\tau g(t_n + c_j\tau) + \rho_n,$$

with ρ_n the local error in space and time. Using $w_h' = Aw_h + g + \sigma_h$ to eliminate the terms $g(t_n + c_j\tau)$, we get

$$\rho_n = \underbrace{\sum_{k\geq q+1} \frac{1}{k!} H_k(\tau A)\tau^k w_h^{(k)}(t_n)}_{\delta_n} + \underbrace{\sum_{j=1}^{s} Q_j(\tau A)\tau\sigma_h(t_n + c_j\tau)}_{\gamma_n}.$$

The contribution of γ_n to the global error can be studied in the same way as for δ_n. Suppose

$$\sigma_h(t) = A\xi(t) + \eta(t) \quad \text{with} \quad \xi(t), \xi'(t), \eta(t) = \mathcal{O}(h^\beta),$$

see Lemma 7.1. Then

$$\gamma_n = \sum_{j=1}^{s} \tau A Q_j(\tau A)\xi(t_n) + \mathcal{O}(\tau h^\beta)$$

assuming boundedness of the rational expressions $Q_j(\tau A)$ and $\tau A Q_j(\tau A)$. Since $\sum_{j=1}^{s} z Q_j(z) = -(1 - R(z))$, we get

$$\gamma_n = -(I - R(\tau A))\xi(t_n) + \mathcal{O}(\tau h^\beta).$$

Application of Lemma 8.1 with $r = 1$ shows that the local errors γ_n will give an $\mathcal{O}(h^\beta)$ contribution to the global errors $w_h(t_n) - w_n$.

Notes. The first analysis on order reduction is due to Crouzeix (1975, thesis) for implicit Runge-Kutta methods applied to parabolic problems, see also Brenner, Crouzeix & Thomée (1982). For the presentation of the results on the local error in this section the latter paper was closely followed. Similar results for nonlinear equations can be found in Lubich & Ostermann (1993). More recently, it was shown by Lubich & Ostermann (1995) that for parabolic problems and strongly A-stable methods, the classical order of convergence p will still be valid in the interior of the spatial domain.

The occurrence of order reduction for explicit methods and hyperbolic equations was first discussed in Sanz-Serna et al. (1987). In that paper also some examples are given where this temporal order reduction is avoided by a transformation of the problem. For example, a problem with inhomogeneous, time dependent Dirichlet conditions can be transformed to a problem with homogeneous conditions, and this will increase the temporal accuracy. A more general discussion on this subject, with numerical examples, can be found in Pathria (1997).

As we saw, there is no order reduction for the trapezoidal rule. In fact, order reduction never occurs for linear multistep methods, due to the fact that in such methods no lower order intermediate vectors are used. This is no longer true if such methods are used in a usual predictor-corrector fashion.

9. Time splitting methods

If we consider advection, diffusion combined with chemistry,

$$\frac{\partial}{\partial t}u + \sum_{k=1}^{d} \frac{\partial}{\partial x_k}\Big(a_k u\Big) = \sum_{k=1}^{d} \frac{\partial}{\partial x_k}\Big(d_k \frac{\partial}{\partial x_k} u\Big) + f(u,x,t),$$

for a vector $u(x,t) = (u_1(x,t), u_2(x,t), ..., u_s(x,t))^T$ containing concentration values of s chemical species, one might want to apply different time stepping methods to the different parts of the equations. For example, the chemistry can be very stiff, which calls for an implicit ODE method. On the other hand, if the advection is discretized in space with a flux-limiter, then explicit methods seem much more suitable for that part of the equation. Moreover, use of an implicit method to the full equation will lead to a huge algebraic system, with coupling over the species as well as over the space.

In this section we shall discuss some methods were the equation is split into several parts, which are all solved independently on the time intervals $[t_n, t_{n+1}]$. Such methods are usually called *(time) splitting* methods or *fractional step* methods. In case the splitting is such that different physical processes are separated, the term "operator splitting" is also used. If a multi-dimensional problem is split into 1-dimensional sub-problems, this is often called "dimensional splitting".

9.1. First order splitting

Consider an ODE system, linear for simplicity,

$$w'(t) = Aw(t),$$

with $A = A_1 + A_2$, arising for example from a linear PDE with homogeneous boundary conditions or periodicity conditions. We have

$$w(t_{n+1}) = e^{\tau A} w(t_n). \tag{9.1}$$

If we are only able, or willing, to solve the "sub-problems" $w'(t) = A_1 w(t)$ and $w'(t) = A_2 w(t)$, then (9.1) can be approximated by

$$w_{n+1} = e^{\tau A_2} e^{\tau A_1} w_n, \tag{9.2}$$

which is the simplest splitting method. In actual computations the terms $e^{\tau A_k}$ will, of course, be approximated by some suitable ODE method.

Replacing (9.1) by (9.2) will introduce an error, the so-called *splitting error* for this particular splitting. Inserting the exact solution into (9.2) gives $w(t_{n+1}) = e^{\tau A_2} e^{\tau A_1} w(t_n) + \tau \rho_n$ with local truncation error ρ_n. Note that $\tau \rho_n$ is the error introduced per step. We have

$$e^{\tau A} = \Big(I + \tau(A_1 + A_2) + \frac{1}{2}\tau^2 (A_1 + A_2)^2 + \cdots\Big),$$

$$e^{\tau A_2} e^{\tau A_1} = \Big(I + \tau(A_1 + A_2) + \frac{1}{2}\tau^2 (A_1^2 + 2A_2 A_1 + A_2^2) + \cdots\Big).$$

Hence the local truncation error equals

$$\frac{1}{\tau}\Big(e^{\tau A} - e^{\tau A_2}e^{\tau A_1}\Big)w(t_n) = \frac{1}{2}\tau[A_1, A_2]w(t_n) + \mathcal{O}(\tau^2), \tag{9.3}$$

with $[A_1, A_2] = A_1A_2 - A_2A_1$ the commutator of A_1 and A_2. We see that (9.2) will be a 1-st order process, unless A_1 and A_2 commute. Note that we assume here tacitly that terms like $A_1A_2w(t)$ are $\mathcal{O}(1)$, which seems reasonable only if there are no boundary conditions or the PDE solution satisfies certain compatibility conditions, see Section 8.

For general nonlinear ODE systems

$$w'(t) = F_1(t, w(t)) + F_2(t, w(t)),$$

we can apply (9.2) if the terms e^{tA_k} are interpreted as *solution operators*. Written out, we solve subsequently

$$\begin{aligned}
&\tfrac{d}{dt}w^*(t) = F_1(t, w^*(t)) \qquad \text{for } t_n \le t \le t_{n+1} \text{ with } w^*(t_n) = w_n,\\
&\tfrac{d}{dt}w^{**}(t) = F_2(t, w^{**}(t)) \qquad \text{for } t_n \le t \le t_{n+1} \text{ with } w^{**}(t_n) = w^*(t_{n+1}),
\end{aligned}$$

giving $w_{n+1} = w^{**}(t_{n+1})$ as the next approximation. If $w_n = w(t_n)$ we now get the local truncation error

$$\frac{1}{2}\tau\Big[\frac{\partial F_1}{\partial w}F_2 - \frac{\partial F_2}{\partial w}F_1\Big](t_n, w(t_n)) + \mathcal{O}(\tau^2),$$

similar to (9.3). This formula can be derived by Taylor expansions of $w^{**}(t_{n+1})$ and $w^*(t_{n+1})$ around $t = t_n$.

Note. The structure of the global error of (9.2) becomes transparent by using the Baker-Campbell-Hausdorff formula,

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

with

$$\begin{aligned}
\tilde{A} = A + \tfrac{1}{2}\tau[A_2, A_1] + \tfrac{1}{12}\tau^2\Big([A_2, [A_2, A_1]] + [A_1, [A_1, A_2]]\Big) +\\
+ \tfrac{1}{24}\tau^3[A_2, [A_1, [A_1, A_2]]] + \mathcal{O}(\tau^4).
\end{aligned} \tag{9.4}$$

This formula can be derived by power series developments of $e^{\tau A_1}e^{\tau A_2}$ and $e^{\tau \tilde{A}}$ and comparing the terms with the same powers of τ. The calculation of the terms in $\tilde{A}$ quickly become cumbersome if done in a straightforward fashion, but it can also be done in a recursive way, see Sanz-Serna & Calvo (1994) and the references given there. Using Lie formalism, a similar formula can also be obtained for nonlinear autonomous equations.

From formula (9.4) we can reobtain the truncation error (9.3), but we can also apply it in a global fashion, using $(e^{\tau A_2}e^{\tau A_1})^n = e^{t_n\tilde{A}}$. Hence, when applied with constant step size τ, the splitting process (9.1) will solve the modified equation $w'(t) = \tilde{A}w(t)$, rather than the original problem.

In (9.2) one starts in all steps with A_1. Interchanging the order of A_1 and A_2 after each step will lead to more symmetry and better accuracy. Carrying out two half steps with reversed sequence gives the following splitting, due to Strang (1968),

$$w_{n+1} = \Big(e^{\frac{1}{2}\tau A_2} e^{\frac{1}{2}\tau A_1}\Big)\Big(e^{\frac{1}{2}\tau A_1} e^{\frac{1}{2}\tau A_2}\Big) w_n = e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\frac{1}{2}\tau A_2} w_n. \tag{9.5}$$

By a series expansion and some tedious calculations it follows that the local truncation error is given by

$$\frac{1}{24}\tau^2\Big([A_2,[A_2,A_1]] - 2[A_1,[A_1,A_2]]\Big) w(t_n) + \mathcal{O}(\tau^4). \tag{9.6}$$

This can also be found by repeated application of formula (9.4). Due to symmetry, the truncation error will only contain even order terms.

If we work with constant step sizes, then (9.5) will require almost the same amount of computational work as (9.2), since for constant τ we can write the total process (9.5) as

$$w_n = e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\tau A_2} \cdots e^{\tau A_1} e^{\frac{1}{2}\tau A_2} w_0.$$

In general, with variable step sizes it will be more expensive, of course.

Generalization to nonlinear systems is straightforward, we get

$$\begin{aligned}
&\tfrac{d}{dt} w^*(t) = F_2(t, w^*(t)) \qquad \text{for } t_n \le t \le t_{n+1/2} \text{ with } w^*(t_n) = w_n,\\
&\tfrac{d}{dt} w^{**}(t) = F_1(t, w^{**}(t)) \qquad \text{for } t_n \le t \le t_{n+1} \text{ with } w^{**}(t_n) = w^*(t_{n+1/2}),\\
&\tfrac{d}{dt} w^{***}(t) = F_2(t, w^{***}(t)) \qquad \text{for } t_{n+1/2} \le t \le t_{n+1} \text{ with } w^{***}(t_{n+1/2}) = w^{**}(t_{n+1}),
\end{aligned}$$

giving $w_{n+1} = w^{***}(t_{n+1})$ as the approximation on the new time level. The local truncation error now contains many terms. If we assume that the equation is autonomous, then Taylor expansion leads to the following expression for this truncation error, see LeVeque (1982),

$$\begin{aligned}
&\frac{1}{6}\tau^2\Big[\ \frac{1}{4}\frac{\partial}{\partial w}\Big(\frac{\partial F_2}{\partial w}F_2\Big)F_1 - \frac{1}{2}\frac{\partial}{\partial w}\Big(\frac{\partial F_2}{\partial w}F_1\Big)F_2 + \frac{1}{4}\frac{\partial}{\partial w}\Big(\frac{\partial F_1}{\partial w}F_2\Big)F_2 -\\
&-\frac{1}{2}\frac{\partial}{\partial w}\Big(\frac{\partial F_1}{\partial w}F_1\Big)F_2 + \frac{\partial}{\partial w}\Big(\frac{\partial F_1}{\partial w}F_2\Big)F_1 - \frac{1}{2}\frac{\partial}{\partial w}\Big(\frac{\partial F_2}{\partial w}F_1\Big)F_1\ \Big](w(t_n)) + \mathcal{O}(\tau^4).
\end{aligned}$$

An other 2-nd order splitting, also due to Strang (1963), is given by

$$w_{n+1} = \frac{1}{2}\Big(e^{\tau A_1} e^{\tau A_2} + e^{\tau A_2} e^{\tau A_1}\Big) w_n. \tag{9.7}$$

The truncation error for this splitting is given by

$$-\frac{1}{12}\tau^2\Big([A_1,[A_1,A_2]] + [A_2,[A_2,A_1]]\Big) w(t_n) + \mathcal{O}(\tau^3). \tag{9.8}$$

The splitting (9.7) is, however, more expensive than (9.2) and will also require more computer memory. On the other hand, the factors $e^{\tau A_1} e^{\tau A_2}$ and $e^{\tau A_2} e^{\tau A_1}$ can be computed in parallel. For nonlinear problems the same considerations hold.

With regards to stability of the splittings, assume that we have $\|e^{tA_k}\| \le 1$ for $t \ge 0$ and $k = 1, 2$. Then it follows trivially that we have $\|w_{n+1}\| \le \|w_n\|$ in the splitting processes (9.2), (9.5) and (9.7). In the same way we get for these splittings

$$\|e^{tA_k}\| \le e^{t\omega_k} \ (k = 1, 2, \ t > 0) \quad \Longrightarrow \quad \|w_{n+1}\| \le e^{\tau\omega}\|w_n\| \ \text{ with } \omega = \omega_1 + \omega_2.$$

General stability results for these splittings, under the assumption that $\|e^{tA_k}\| \le K$ for $t \ge 0$, seem unknown. On the other hand, in practice the first order splitting and Strang splittings appear to be very stable. In general it is more the accuracy that needs improvement.

The above splitting methods fit in the more general form

$$w_{n+1} = \sum_{i=1}^{s} \alpha_i \Big(\prod_{j=1}^{r} e^{\tau\beta_{ij}A_1} e^{\tau\gamma_{ij}A_2} \Big) w_n \tag{9.9}$$

with $\alpha_1 + \cdots + \alpha_s = 1$. If we assume again that $\|e^{tA_k}\| \le 1$ for $t \ge 0, k = 1, 2$ and if all coefficients $\alpha_i, \beta_{ij}, \gamma_{ij} \ge 0$ we obtain as above the stability estimate $\|w_{n+1}\| \le \|w_n\|$. One could try to find suitable parameter choices that give higher order processes, but it was shown by Sheng (1989) that for having order $p > 2$ some of the coefficients must be negative. More recently, the result of Sheng was refined by Goldman & Kaper (1996), who showed that if $p > 2$ and all $\alpha_i > 0$ then $\min \beta_{ij} < 0$ and $\min \gamma_{ij} < 0$, and thus a step with negative time is necessary for both A_1 and A_2. The proof of these results are long and technical. Therefore we only discuss here briefly two examples of higher order splittings.

Examples. Let

$$S_\tau = e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\frac{1}{2}\tau A_2}$$

be the 2-nd order Strang splitting operator. By using local Richardson extrapolation, one obtains the 4-th order splitting

$$w_{n+1} = \Big(\frac{4}{3}(S_{\frac{1}{2}\tau})^2 - \frac{1}{3} S_\tau \Big) w_n,$$

with a negative weight $-1/3$. Because of this negative weight the simple stability considerations considered above no longer hold. In fact, it seems unknown for what kind of problems this scheme will be stable or unstable. (Using global extrapolation at the end point $t = T$ might be beneficial for stability, but then constant step sizes have to be used.)

Another 4-th order splitting, derived by Yoshida (1990) and Suzuki (1990), reads

$$w_{n+1} = S_{\theta\tau}\, S_{(1-2\theta)\tau}\, S_{\theta\tau}\, w_n,$$

with $\theta = (2 - \sqrt[3]{2})^{-1} \approx 1.35$. Here we have $1 - 2\theta < 0$, so that a step with negative time has to be taken.

For partial differential equations with boundary conditions such splittings with negative time steps seem of limited value. We note, however that they are frequently used for time reversible problems, which arise for instance with certain mechanical problems, see Sanz-Serna & Calvo (1994). $\diamond$

If $A = A_1 + A_2 + A_3$ then the first order splitting (9.2) can be generalized to

$$w_{n+1} = e^{\tau A_3} e^{\tau A_2} e^{\tau A_1} w_n.$$

Likewise, the Strang splitting (9.5) leads to the 2-nd order formula

$$w_{n+1} = e^{\frac{1}{2}\tau A_3} e^{\frac{1}{2}\tau A_2} e^{\tau A_1} e^{\frac{1}{2}\tau A_2} e^{\frac{1}{2}\tau A_3} w_n.$$

Note that this is just a repeated application of (9.5): first approximate $e^{\tau A}$ by $e^{\frac{1}{2}\tau A_3} e^{\tau(A_1+A_2)} e^{\frac{1}{2}\tau A_3}$, and then approximate $e^{\tau(A_1+A_2)}$ in the same fashion.

Application to more components and nonlinear systems carries over in the same way.

Remark. Repeated application of (9.7) leads to rather complicated formulas. For linear equations, with $A = A_1 + A_2 + A_3$, the formula

$$w_{n+1} = \frac{1}{2}\Big(e^{\tau A_1} e^{\tau A_2} e^{\tau A_3} + e^{\tau A_3} e^{\tau A_2} e^{\tau A_1}\Big) w_n,$$

gives also a 2-nd order truncation error. Probably (not verified!) this generalization remains of 2-nd order for nonlinear equations (this has to be verified separately since this is not a repeated application of a 2-nd order splitting). Many more variants are possible, of course.

We proceed with a brief description of some concrete examples for splittings, and their advantages. Obviously, for combined problems with more than two components, the advantages can also be combined. The disadvantage of splitting is, of course, the introduction of a splitting error on top of the errors that will be made when solving the sub-problems.

Example ("operator splitting"). For the advection-reaction equation

$$u_t + \sum_{k=1}^{d} (a_k u)_{x_k} = f(u, x, t) \in \mathbb{R}^s,$$

splitting of the advection and reaction terms has obvious computational advantages. We can then use an explicit method for the advective terms and an implicit method for the reaction. Further, in the advection sub-step there will be only coupling in space, whereas in the reaction sub-step we will have only coupling between the chemical species at the same place, and so there much parallelism.

The truncation error of the 1-st order splitting is

$$\frac{1}{2}\tau\Big[\sum_k (a_k f_{x_k}) + \sum_k (a_k)_{x_k}(f - f_u u)\Big] + \mathcal{O}(\tau^2).$$

There will be no splitting error if the velocity field is divergence-free, $\sum_k (\partial a_k/\partial x_k) = 0$, and the reaction term does not depend explicitly on the space variable, $(\partial f/\partial x_k) = 0$ $(k = 1, ..., d)$. If this does not hold, we can use Strang splitting to obtain 2-nd order accuracy. $\diamond$

Example ("dimension splitting"). Solving the 2D advection equation

$$u_t + (au)_x + (bu)_y = 0$$

with finite differences in space and explicit time stepping, will lead to a CFL condition for stability of the type

$$\frac{\tau}{\Delta x}|a| + \frac{\tau}{\Delta y}|b| \leq C_0,$$

with C_0 determined by the method. If we split the equation into an x-part and a y-part, while using the same discretizations, we get a stability restriction

$$\max\Big(\frac{\tau}{\Delta x}|a|, \frac{\tau}{\Delta y}|b|\Big) \leq C_0,$$

which allows larger time steps. Moreover, this splitting also allows the use of tailored 1-D schemes of the Lax-Wendroff type, for which good multi-dimensional extensions are difficult to derive.

The leading term in the truncation error of the 1-st order splitting now becomes

$$\frac{1}{2}\tau\Big((a(bu)_y)_x - (b(au)_x)_y\Big) = \frac{1}{2}\tau\Big(-(a_y bu)_x + (ab_x u)_y\Big),$$

and this will vanish if $a_y = b_x = 0$. If it does not, one should use Strang splitting. ◇

Example ("dimension splitting"). Consider the diffusion equation

$$u_t = (du_x)_x + (eu_y)_y,$$

with 2-nd order central differences to discretize the diffusion operators and implicit time stepping. Here splitting of the x-part and y-part makes the implicit relations much easier to solve. For example, setting $e = 0$ in the first sub-step, leads to a number of uncoupled 1D tri-diagonal systems.

With 1-st order splitting the leading term in the truncation error now reads

$$\frac{1}{2}\tau\Big((d(eu_y)_{xy})_x - (e(du_x)_{xy})_y\Big),$$

which is zero in case $d_y = e_x = 0$. ◇

9.4. Solving the fractional steps

To solve the sub-steps, one may select a method such as Euler or Trapezoidal Rule. If these are applied with the same step size τ that is used for the splitting itself, a specific splitting method arises. Numerous examples can be found in Yanenko (1971), Gourlay & Mitchell (1972), Mitchell & Griffiths (1980), Marchuk (1990).

For instance, first order splitting combined with backward Euler gives the first order method

$$\begin{aligned} w^*_{n+1} &= w_n + \tau F_1(t_{n+1}, w^*_{n+1}), \\ w_{n+1} &= w^*_{n+1} + \tau F_2(t_{n+1}, w_{n+1}). \end{aligned} \tag{9.10}$$

If F_1 and F_2 contain discretized space derivatives in x and y direction, respectively, this method is called the 1-st order LOD method (locally one dimensional) of Yanenko. It is obvious that we can generalize this method for $F = F_1 + F_2 + ... + F_s$.

The 2-nd order LOD method is obtained by combining Strang splitting with the trapezoidal rule (or, likewise, the implicit midpoint rule),

$$\begin{aligned} w^*_{n+1} &= w_n + \tfrac{1}{2}\tau\Big(F_1(t_n, w_n) + F_1(t_n + (\tfrac{1}{2}+c)\tau, w^*_{n+1})\Big), \\ w_{n+1} &= w^*_{n+1} + \tfrac{1}{2}\tau\Big(F_2(t_n + (\tfrac{1}{2}-c)\tau, w^*_{n+1}) + F_2(t_{n+1}, w_{n+1}))\Big), \\ w^*_{n+2} &= w_{n+2} + \tfrac{1}{2}\tau\Big(F_2(t_{n+1}, w_{n+1}) + F_2(t_{n+1} + (\tfrac{1}{2}+c)\tau, w^*_{n+2})\Big), \\ w_{n+2} &= w^*_{n+2} + \tfrac{1}{2}\tau\Big(F_1(t_{n+1} + (\tfrac{1}{2}-c)\tau, w^*_{n+2}) + F_1(t_{n+1}, w_{n+2}))\Big). \end{aligned} \tag{9.11}$$

Note that here Strang splitting is applied on the interval $[t_n, t_{n+2}]$. For c we can take for example $c = 0$ or $c = \frac{1}{2}$. What is best will depend on the problem, and there is no choice that seems preferable a priori. This is due to the fact that the intermediate vectors w^*_{n+j} are not a consistent approximation to the full problem at some given time level. Again, generalization to more F-components is straightforward.

An other familiar splitting method is the second order Peaceman-Rachford ADI method (alternating direction implicit)

$$\begin{aligned} w^*_{n+1/2} &= w_n + \tfrac{1}{2}\tau F_1(t_n, w_n) + \tfrac{1}{2}\tau F_2(t_{n+1/2}, w^*_{n+1/2}), \\ w_{n+1} &= w^*_{n+1/2} + \tau F_1(t_{n+1/2}, w_{n+1}) + \tfrac{1}{2}\tau F_2(t_{n+1}, w^*_{n+1/2}). \end{aligned} \tag{9.12}$$

This could be viewed as a Strang splitting with alternative use of forward and backward Euler, in a symmetrical fashion to obtain second order, but it seems more natural to consider this ADI method as a method of its own. Note that the intermediate value $w^*_{n+1/2}$ is consistent with the whole equation, unlike with the LOD methods. On the other hand, this ADI method does not have a natural extension for more than two components F_j. In the next section a related ADI method is discussed in detail that does allow more components.

With the above splitting methods all sub-problems are treated in the same fashion and with the same time step. In general, it seems better to solve the fractional steps with a method that is suited for that particular sub-step, possibly with a sub-time step $\bar{\tau} \le \tau$. Here one may chose, for example, an implicit or explicit Runge-Kutta method, depending whether the sub-problem $w'(t) = F_j(t, w(t))$ is stiff or non-stiff, with an appropriate $\bar{\tau}$.

9.5. Boundary corrections

The major difficulties with splitting methods occur for problems were the boundary conditions are important. If we consider a PDE problem with boundary conditions, then these are physical conditions for the whole process and boundary conditions for the sub-steps (which may have little physical meaning) are missing.

Therefore one may have to reconstruct boundary conditions for the specific splitting under consideration. For example, consider a linear semi-discrete problem $w'(t) = Aw(t) + g(t)$, were $g(t)$ contains the given boundary conditions. Suppose that

$$Av + g(t) = \Big(A_1 v + g_1(t)\Big) + \Big(A_2 v + g_2(t)\Big),$$

with $g_k(t)$ containing the boundary conditions relevant to A_k. The exact solution satisfies

$$w(t_{n+1}) = e^{\tau A} w(t_n) + \int_0^\tau e^{(\tau-s)A} g(t_n+s) ds.$$

If we consider 1-st order splitting, with inhomogeneous terms $\tilde{g}_1, \tilde{g}_2$, then

$$w_{n+1} = e^{\tau A_2} e^{\tau A_1} w_n + e^{\tau A_2} \int_0^\tau e^{(\tau-s)A_1} \tilde{g}_1(t_n+s) ds + \int_0^\tau e^{(\tau-s)A_2} \tilde{g}_2(t_n+s) ds.$$

Even with commuting matrices, $A_1 A_2 = A_2 A_1$, and constant boundary terms we will get a splitting error if we take $\tilde{g}_k = g_k$. An exact formula for this case is obtained by choosing

$$\tilde{g}_1(t_n+s) = e^{-sA_2} g_1(t_n+s), \quad \tilde{g}_2(t_n+s) = e^{(\tau-s)A_1} g_2(t_n+s).$$

Note that this correction for g_1 requires a *backward* time integration with A_2, and this may not be feasible with an implicit ODE method, due to the fact that the implicit algebraic relations need no longer be well defined with negative step size. One might replace e^{-sA_2} by some explicit polynomial approximation $P(-sA_2)$, but the effect of this on stability and accuracy is unclear.

As a rule of thumb, it can be said that the treatment of the boundaries should coincide as much as possible with the scheme in the interior of the domain. Examples for specific LOD or ADI methods can be found in Mitchell & Griffiths (1980, Chapter 2). A general analysis of boundary conditions for splitting methods is, at present, still lacking. Therefore we conclude this subject with an example.

Example. Consider the model advection-reaction equation, already used in Section 8,

$$u_t + u_x = u^2, \qquad 0 \le x \le 1, \quad 0 \le t \le 1/2$$

with given initial value at $t = 0$ and Dirichlet condition at $x = 0$, derived from the exact solution

$$u(t,x) = \frac{\sin(\pi(x-t))^2}{1 - t\sin(\pi(x-t))^2}.$$

As before, spatial discretization is performed with 4-th order central differences in the interior and 3-rd order one-sided approximations at the boundaries. The advection step is solved with the classical Runge-Kutta method at Courant number $\tau/h = 2$, and the "reaction" $u_t = u^2$ is solved exactly. Since the nonlinear term is nonstiff, splitting is not really necessary in this example, but for comparison it is instructive to consider the same example as in Section 8.

We consider :
(i) simple splitting (with reaction followed by advection) where in the advection step the given boundary values are used;
(ii) Strang splitting where after each time step the order of the fractional steps is reversed, also with the given boundary conditions;
(iii) the same splitting as in (i) but with corrected boundary conditions

$$u^{**}(t,0) = \frac{u(t,0)}{1 - (t_{n+1} - t)u(t,0)} \quad \text{for } t \in [t_n, t_{n+1}].$$

The errors in the L_2-norm, together with the estimated orders of convergence, are given in the following table.

	Simple splitting	Strang splitting	Corrected boundary
$\tau = 1/\ 20$	$0.26\ 10^{-1}$	$0.14\ 10^{-1}$	$0.88\ 10^{-3}$
$\tau = 1/\ 40$	$0.14\ 10^{-1}$ (*0.94*)	$0.48\ 10^{-2}$ (*1.58*)	$0.91\ 10^{-4}$ (*3.27*)
$\tau = 1/\ 80$	$0.72\ 10^{-2}$ (*0.96*)	$0.17\ 10^{-2}$ (*1.54*)	$0.13\ 10^{-4}$ (*2.80*)
$\tau = 1/160$	$0.36\ 10^{-2}$ (*0.98*)	$0.58\ 10^{-3}$ (*1.52*)	$0.22\ 10^{-5}$ (*2.57*)

TABLE 9.1. Relative L_2-errors (and estimated orders) for (4.1) at $t = 1/2$ with $\tau = 2h$.

Note that the simple splitting with boundary corrections is more accurate than its Strang type counterpart. With this correction we reobtain an accuracy comparable to that of Table 8.2.

The convergence rate of the scheme with boundary corrections is less than 4, but this is due to order reduction of the Runge-Kutta method, it is not caused by the splitting procedure. A similar order reduction can be observed with Strang splitting: in the absence of boundary conditions it has (at least) order 2, but in the above table an order 1.5 behaviour can be observed. ◇

10. IMEX, ADI and AF methods

With time splitting by the fractional step approach we have to solve sub-problems that are not consistent with the full model. As we saw this creates difficulties with boundary conditions, and similar problems arise with interface conditions. Also, stationary solutions of the problem are not stationary solutions of the fractional step methods. Moreover in the time splitting approach multi-step schemes cannot be used in a natural fashion.

In this section some alternatives to time splitting will be briefly reviewed. The methods considered here are still subject to ongoing research, and we shall refer to recent papers for proofs of the technical results.

10.1. The θ-IMEX method

Suppose that the semi-discrete system is of the form

$$w'(t) = F(t, w(t)) = F_0(t, w(t)) + F_1(t, w(t)) \tag{10.1}$$

where F_0 is a term suitable for explicit time integration, for instance discretized advection, and F_1 requires an implicit treatment, say discretized diffusion or stiff reactions.

We consider the following simple method

$$w_{n+1} = w_n + \tau F_0(t_n, w_n) + (1-\theta)\tau F_1(t_n, w_n) + \theta\tau F_1(t_{n+1}, w_{n+1}), \tag{10.2}$$

with parameter $\theta \geq \frac{1}{2}$. Here the explicit Euler method is combined with the implicit θ-method. Such mixtures of implicit and explicit methods are called IMEX schemes. Note that in contrast to the time splitting methods there are no intermediate results which are inconsistent with the full equation.

Insertion of the exact solution in the scheme gives the truncation error

$$\frac{1}{\tau}\Big(w(t_{n+1} - w(t_n)\Big) - (1-\theta)F(t_n, w(t_n)) - \theta F(t_{n+1}, w(t_{n+1}))-$$

$$-\theta\Big(F_0(t_{n+1}, w(t_{n+1})) - F_0(t_n, w(t_n))\Big) = (\frac{1}{2} - \theta)\tau w''(t_n) + \theta\tau\varphi'(t_n) + \mathcal{O}(\tau^2)$$

where $\varphi(t) = F_0(t, w(t))$. If F_0 denotes discretized advection and nonstiff terms, smoothness of w will also imply smoothness of φ, independent of boundary conditions or small mesh widths h. Therefore the structure of the truncation error is much more favourable than with the time splitting methods considered in the preceding section. For example, with a stationary solution $w(t) = w(0)$ we now have a zero truncation error. However, with methods of this IMEX type it is stability that needs a careful examination.

Let us consider the scalar, complex test equation

$$w'(t) = \lambda_0 w(t) + \lambda_1 w(t), \tag{10.3}$$

and let $z_j = \tau\lambda_j$, $j = 0, 1$. In applications to PDEs these λ_j will represent eigenvalues of the two components F_0 and F_1, found by inserting Fourier modes. One would hope that having $|1 + z_0| \leq 1$ (stability of the explicit method) and $\operatorname{Re} z_1 \leq 0$ (stability of the implicit method) would be sufficient to guarantee stability of the IMEX scheme, but this is not so in general.

Application of the IMEX scheme to this test equation yields $w_{n+1} = Rw_n$ where $R = R(z_0, z_1)$ is given by

$$R = \frac{1 + z_0 + (1-\theta)z_1}{1 - \theta z_1} \,. \tag{10.4}$$

Stability for the test equation thus requires $|R| \leq 1$.

First, consider the set

$$\mathcal{D}_0 = \{z_0 : \text{the IMEX scheme is stable for any } z_1 \in \mathbb{C}^-\}. \tag{10.5}$$

So, here we insist on A-stability with respect to the implicit part. Using the maximum principle, it follows by some straightforward calculations that $z_0 = x_0 + iy_0$ belongs to this set iff

$$\theta^2 y_0^2 + (2\theta - 1)(1 + x_0)^2 \leq 2\theta - 1.$$

Plots are given in Figure 10.1. If $\theta = 1$ we reobtain the stability region of the explicit Euler method, but for smaller values of θ the set start to shrink and for $\theta = \frac{1}{2}$ it reduces to the line segment $[-2, 0]$ on the negative axis.

Alternatively, one can insist on using the full stability region of the explicit method $\mathcal{S}_0 = \{z_0 : |1 + z_0| \leq 1\}$, but then z_1 has to be restricted to the set

$$\mathcal{D}_1 = \{z_1 : \text{the IMEX scheme is stable for any } z_0 \in \mathcal{S}_0\}. \tag{10.6}$$

It easily follows that $z_1 \in \mathcal{D}_1$ iff

$$1 + (1-\theta)|z_1| \leq |1 - \theta z_1|,$$

see the right plot in Figure 10.1. Again it is only for $\theta = 1$ that we get the stability region of the implicit θ-method. If $\theta = \frac{1}{2}$ the set $\mathcal{D}_1$ equals the negative real line $\mathbb{R}^-$.

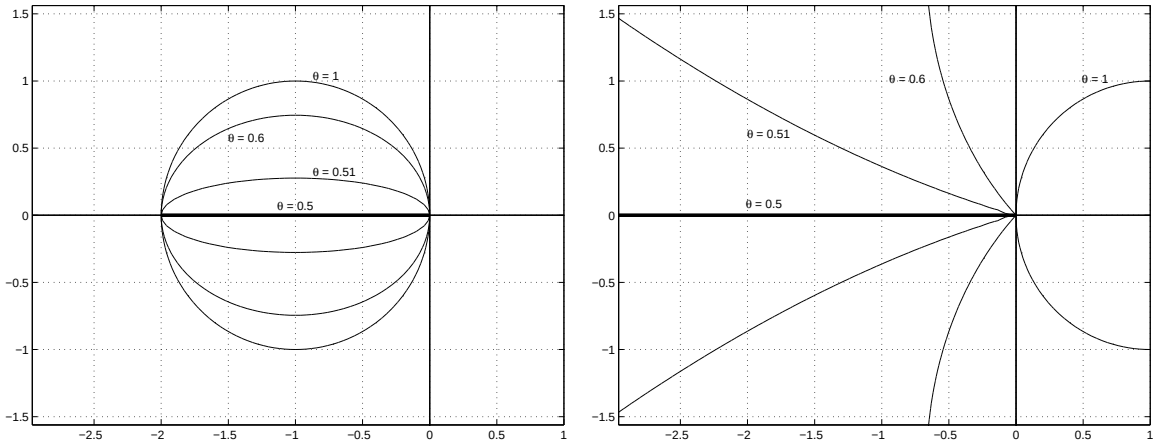


Figure 10.1. Boundaries of regions $\mathcal{D}_0$ (left) and $\mathcal{D}_1$ (right) for the θ-IMEX method (10.2) with $\theta = 0.5, 0.51, 0.6$ and 1.

Note that the implicit θ-method with $\theta > \frac{1}{2}$ is *strongly* A-stable (that is, A-stable with damping at ∞) whereas the trapezoidal rule, $\theta = \frac{1}{2}$, is "just" A-stable. Apparently, using a strongly implicit method gives better stability properties within an IMEX formula.

On the other hand, the above criteria are rather strict. For instance, if we take z_0 such that $|\rho + z_0| \le \rho$ with $\rho < 1$, then the method with $\theta = \frac{1}{2}$ will be stable if $z_1 = x_1 + iy_1 \in \mathbb{C}^-$ is within the hyperbole $\rho^2 y_1^2 + 4\rho^2(1-\rho) \le 4(1-\rho)(\rho - x_1)^2$ (exercise). Therefore, the IMEX method with $\theta = \frac{1}{2}$ should not be discarded, only extra care should be given to stability when applying this method.

In the above the values of λ_0 and λ_1 have been considered as independent, which is a reasonable assumption if F_0 and F_1 act in different directions, for instance if $F_0 \approx a(\partial/\partial x)$ (horizontal coupling) and $F_1 \approx d(\partial^2/\partial z^2)$ (vertical coupling) or F_1 a reaction term (coupling over chemical species).

Different results are obtained if there is a dependence between λ_0 and λ_1. Then the implicit treatment of λ_1 can stabilize the process so that we do not even need $z_0 \in \mathcal{S}_0$. Consider for example the 1D advection-diffusion equation $u_t + au_x = du_{xx}$ with periodicity in space and with second order spatial discretization. If advection is treated explicitly and diffusion implicitly, then

$$z_0 = i\nu \sin 2\phi, \qquad z_1 = -4\mu \sin^2 \phi \tag{10.7}$$

with $\nu = a\tau/h$, $\mu = d\tau/h^2$ and $0 \le \phi \le \pi$, see Section 2. A straightforward calculation shows that $|R| \le 1$ iff

$$1 - 8(1-\theta)\mu s + 16(1-\theta)^2\mu^2 s^2 + 4\nu^2 s(1-s) \le 1 + 8\theta\mu s + 16\theta^2\mu^2 s^2$$

where $s = \sin^2 \phi$. This holds for all $s \in [0,1]$ iff

$$\nu^2 \le 2\mu \quad \text{and} \quad 2(1-2\theta)\mu \le 1. \tag{10.8}$$

So for any $\theta \ge \frac{1}{2}$ we now just have the condition $\nu^2 \le 2\mu$, that is $a^2\tau \le 2d$.

Finally we note that the above IMEX method with $\theta = 1$ could be viewed as a time splitting method where we first solve $v'(t) = F_0(t, v(t))$ on $[t_n, t_{n+1}]$ with forward Euler and then $v'(t) = F_1(t, v(t))$ with backward Euler. This explains the favourable stability results with this method. However, the structure of the truncation error is very different from the time splitting methods. This is due to interference of the first order splitting error with the first order Euler errors.

In the following subsections we shall consider several generalizations of (10.2). Such generalizations are necessary for practical problems since the explicit Euler method is not well suited for advection, and also first order accuracy is often not sufficient. Moreover, we may want additional splittings of the implicit terms to resolve the implicit relations more efficiently.

10.2. IMEX multi-step methods

As mentioned already, in the time splitting approach multi-step schemes cannot be used in a natural fashion. Straightforward use of a multi-step scheme with step size τ to solve the sub-problems $v'(t) = F_j(t, v(t))$, $t_n \le t \le t_{n+1}$ leads to inconsistencies since the available past values $w_{n-1}, w_{n-2}, \cdots$ are approximations to the whole problem, not to the particular sub-problem at hand. Here we shall consider an other approach to combine implicit and explicit multi-step methods.

One of the most popular implicit methods is the second order BDF2 method

$$\frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} = \tau F(t_{n+1}, w_{n+1})$$

where the left hand side is the 2-step backward differentiation formula, hence the name BDF. Along with w_0, the starting value w_1 should be known. It can be computed by a one-step method, for instance Euler. The popularity of this implicit BDF method is due to its stability and damping properties. These are very useful properties for diffusion equations.

Convection equations are often treated more efficiently by an explicit method, such as

$$\frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} = 2\tau F(t_n, w_n) - \tau F(t_{n-1}, w_{n-1}),$$

to which we shall refer as the explicit BDF2 method. The stability region of this explicit method is plotted in Figure 10.2.

With advection-diffusion-reaction problems,

$$u_t + \nabla \cdot (au) = \nabla(d\nabla u) + f(u),$$

explicit advection and implicit diffusion-reaction can then be combined through the IMEX formula

$$\frac{3}{2}w_{n+1} - 2w_n + \frac{1}{2}w_{n-1} = 2\tau F_0(t_n, w_n) + \tau F_0(t_{n-1}, w_{n-1}) + \tau F_1(t_{n+1}, w_{n+1}),$$

where F_0 contains convective terms only and F_1 denotes discretized diffusion together with reaction.

The above can be generalized as follows: consider a fully implicit multistep method

$$\sum_{j=0}^{k} \alpha_j w_{n+1-j} = \tau \sum_{j=0}^{k} \beta_j \Big(F_0(t_{n+1-j}, w_{n+1-j}) + F_1(t_{n+1-j}, w_{n+1-j}) \Big), \tag{10.9}$$

with implicit treatment of advection and diffusion-reaction. We can handle the advection explicitly by applying an extrapolation formula

$$\varphi(t_{n+1}) = \sum_{j=1}^{k} \gamma_j \varphi(t_{n+1-j}) + \mathcal{O}(\tau^q) \tag{10.10}$$

with $\varphi(t) = F_0(t, w(t))$. This leads to the method

$$\sum_{j=0}^{k} \alpha_j w_{n+1-j} = \tau \sum_{j=1}^{k} \beta_j^* F_0(t_{n+1-j}, w_{n+1-j}) + \tau \sum_{j=0}^{k} \beta_j F_1(t_{n+1-j}, w_{n+1-j}), \tag{10.11}$$

with new coefficients $\beta_j^* = \beta_j + \beta_0 \gamma_j$. Methods of this implicit-explicit multistep type were introduced by Crouzeix (1980) and Varah (1980).

Accuracy of the IMEX multistep methods is easy to establish.

Theorem 10.1. Assume the implicit multistep method has order p and the extrapolation procedure has order q. Then the IMEX method has order $r = \min(p, q)$.

Proof. With $\varphi(t) = F_0(t, w(t))$, the local truncation error can be written as

$$\frac{1}{\tau}\sum_{j=0}^{k}\Big(\alpha_j w(t_{n+1-j}) - \tau\beta_j w'(t_{n+1-j})\Big) + \beta_0\Big(\varphi(t_{n+1}) - \sum_{j=1}^{k}\gamma_j\varphi(t_{n+1-j})\Big)$$

$$= C\tau^p w^{(p+1)}(t_n) + \mathcal{O}(\tau^{p+1}) + \beta_0 C'\tau^q\varphi^{(q)}(t_n) + \mathcal{O}(\tau^{q+1}),$$

with constants C, C' determined by the coefficients of the multistep method and the extrapolation procedure. □

Note that in this truncation error only total derivatives arise, and therefore the error is not influenced by large Lipschitz constants (negative powers of the mesh width) in F_0 or F_1.

Stability results for the IMEX multistep methods are quite complicated, even for the simple test problem (10.3). We consider here two classes of 2-step IMEX methods. Let $\mathcal{S}_0, \mathcal{S}_1$ be the stability regions of the explicit and implicit method, respectively.

The first class is based on the BDF2 method,

$$\begin{aligned}\tfrac{3}{2}w_{n+1} - 2w_n + \tfrac{1}{2}w_{n-1} = 2\tau F_0(t_n, w_n) - \tau F_0(t_{n-1}, w_{n-1}) + \\ + \theta\tau F_1(t_{n+1}, w_{n+1}) + 2(1-\theta)\tau F_1(t_n, w_n) - (1-\theta)\tau F_1(t_{n-1}, w_{n-1})\end{aligned} \tag{10.12}$$

with parameter $\theta \geq 0$. The order is 2 and the implicit method is A-stable for $\theta \geq \frac{3}{4}$. With $\theta = 1$, $F_0 = 0$ we reobtain the fully implicit BDF2 method. If $\theta = \frac{3}{4}$ the implicit method is "just" A-stable (equivalent with the trapezoidal rule).

We also consider the following class of IMEX methods, based on the two step Adams formulas,

$$\begin{aligned}w_{n+1} - w_n = \tfrac{3}{2}\tau F_0(t_n, w_n) - \tfrac{1}{2}\tau F_0(t_{n-1}, w_{n-1}) + \\ + \theta\tau F_1(t_{n+1}, w_{n+1}) + (\tfrac{3}{2} - 2\theta)\tau F_1(t_n, w_n) + (\theta - \tfrac{1}{2})\tau F_1(t_{n-1}, w_{n-1}),\end{aligned} \tag{10.13}$$

again with order 2. Here the implicit method is A-stable if $\theta \geq \frac{1}{2}$. If $\theta = \frac{1}{2}$ the implicit method reduces to the trapezoidal rule.

In the Figure 10.2 the stability regions $\mathcal{S}_0$ of the explicit methods are plotted together with the regions $\mathcal{D}_0$, defined as in (10.5). We see from the figure that here $\mathcal{D}_0$ is really smaller than $\mathcal{S}_0$ and if the implicit method is just A-stable, the region $\mathcal{D}_0$ reduces to a line. Formulas for the boundary of $\mathcal{D}_0$ can be found in Frank et al. (1997). In that paper also results on the set $\mathcal{D}_1$, see (10.6), are presented. It seems that, as a rule, if $z_0 \in \mathcal{S}_0$ and $z_1 < 0$, then the IMEX scheme is stable. Moreover, if the implicit method is strongly A-stable then the IMEX scheme is stable for z_1 in a wedge $\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \leq \alpha\}$, with positive angle α. These results were not proven for arbitrary IMEX schemes, only for some specific schemes in the above BDF2 and Adams2 class.

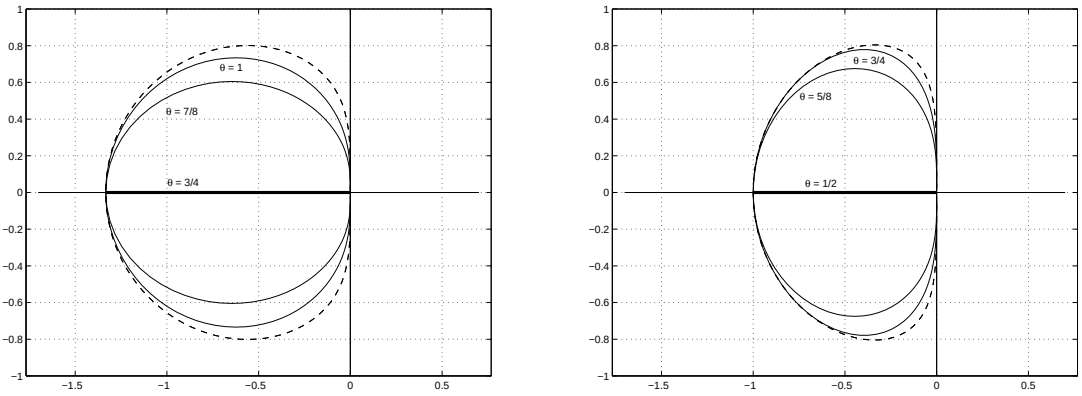


Figure 10.2. Explicit stability regions $\mathcal{S}_0$ (dashed) and regions $\mathcal{D}_0$ for the IMEX BDF2 methods (left) and ADAMS2 methods (right).

With these regions $\mathcal{D}_0$, z_0 and z_1 are considered as independent. As said before, this holds for example if F_0 represents horizontal advection and F_1 stands for vertical diffusion plus reaction (for air pollution problems these are the most relevant terms, the other processes, such as horizontal diffusion, are small and they can be lumped into F_0). Results for 1D advection-diffusion equations can be found in Varah (1980) and Ascher et al. (1995). More general stability results of this type, valid for noncommuting operators, are given in Crouzeix (1980).

10.3. Douglas ADI methods

Suppose we have a decomposition

$$F(t,v) = F_0(t,v) + F_1(t,v) + \cdots + F_s(t,v). \tag{10.14}$$

It will be assumed here that the term F_0 is nonstiff, or mildly stiff, so that this term can be treated explicitly. The other terms will be treated implicitly, in a sequential fashion.

The θ-IMEX method regarded at the beginning of this section can be generalized as follows,

$$\left.\begin{array}{rll} v_0 & = w_n + \tau F(t_n, w_n), & \\ v_j & = v_{j-1} + \theta\tau\Big(F_j(t_{n+1}, v_j) - F_j(t_n, w_n)\Big) & (j = 1, 2, \cdots, s), \\ w_{n+1} & = v_s, & \end{array}\right\} \tag{10.15}$$

with internal vectors v_j. In case $F_0 = 0$ this is the first order Douglas-Rachford ADI method if $\theta = 1$, and the second-order Brian-Douglas ADI method if $\theta = \frac{1}{2}$, see Douglas & Gunn (1964) and Mitchell & Griffiths (1980). This method is also known as the method of Stabilizing Corrections, see Marchuk (1990). Note that all internal vectors v_j are consistent with $w(t_{n+1})$ and therefore the accuracy for problems where the boundary conditions are influential is often better than with the time splitting schemes considered in the previous section. In particular, stationary solutions $\bar{w}$ of $w'(t) = F(w(t))$, that is $F(\bar{w}) = 0$, are also stationary solutions of the ADI method, as can be seen by considering consecutive v_j.

Observe that in this ADI method the implicit terms also allow a splitting, which is not the case with the IMEX multistep methods. However, as with the IMEX methods, stability of the method should be carefully examined. The most simple test problem is

$$w'(t) = \lambda_0 w(t) + \lambda_1 w(t) + \cdots + \lambda_s w(t). \tag{10.16}$$

Let $z_j = \tau\lambda_j$, $j = 0, 1, ..., s$. Then the ADI method yields a recursion $w_{n+1} = Rw_n$ with $R = R(z_0, z_1, ..., z_s)$ given by

$$R = 1 + \Big(\prod_{j=1}^{s}(1 - \theta z_j)\Big)^{-1} \sum_{j=0}^{s} z_j. \tag{10.17}$$

Obviously, stability for the test problem requires $|R| \le 1$.

Consider the wedge $\mathcal{W}_\alpha = \{\zeta \in \mathbb{C} : |\arg(-\zeta)| \le \alpha\}$ in the left half-plane. We consider here stability under the condition that $z_j \in \mathcal{W}_\alpha$, $j \ge 1$. If F_j is a discretized advection-diffusion operator and λ_j an eigenvalue in the Fourier decomposition, then $\alpha < \frac{1}{2}\pi$ means that advection is not allowed to dominate too much (see Section 2.3). For pure diffusion we have $z_j = \tau\lambda_j \in \mathcal{W}_0$, the line of non-positive real numbers. As before, $z_0, z_1, \ldots, z_s$ are assumed to be independent of each other.

Theorem 10.2. Suppose $z_0 = 0$ and $s \ge 2$, $1 \le r \le s - 1$. For any $\theta \ge \frac{1}{2}$ we have

$$|R| \le 1 \ \text{ for all } z_i \in \mathcal{W}_\alpha,\ 1 \le i \le s \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-1}\,\frac{\pi}{2}\,, \tag{10.18}$$

$$\left.\begin{array}{l} |R| \le 1 \ \text{ for all } z_1, \ldots, z_{s-r} \in \mathcal{W}_\alpha \\ \qquad \text{and } z_{s-r+1}, \ldots, z_s \le 0 \end{array}\right\} \quad \Longleftrightarrow \quad \alpha \le \frac{1}{s-r}\,\frac{\pi}{2}\,. \tag{10.19}$$

Proof. Necessity in (10.18) is easy to show: if we take all $z_j = -te^{i\alpha}$, $j \ge 1$, then for $t \to \infty$ we get

$$R = 1 - \frac{ste^{i\alpha}}{\theta^s t^s e^{is\alpha} + \mathcal{O}(t^{s+1})} = 1 - \frac{s}{\theta^s} t^{1-s} e^{i\alpha(1-s)}(1 + \mathcal{O}(t^{-1})),$$

and consequently $\mathrm{Re}(R) > 1$ if t is sufficiently large and $\alpha(1 - s) > \frac{1}{2}\pi$.

To illustrate necessity in (10.19), consider $s = 3$ and $z_3 \le 0$. Since R is fractional linear in z_3, it follows that we have $|R| \le 1$ for all $z_3 \le 0$ iff this holds with z_3 equal to 0 or ∞. This amounts to verification of the inequalities

$$\Big|\, 1 + \frac{z_1 + z_2}{(1 - \theta z_1)(1 - \theta z_2)} \,\Big| \le 1, \qquad \Big|\, 1 - \frac{1}{\theta(1 - \theta z_1)(1 - \theta z_2)} \,\Big| \le 1.$$

For the first inequality we know from (10.18) that $\alpha \le \frac{1}{2}\pi$ is necessary and sufficient, but for the second inequality it can be shown as above that we need $\alpha \le \frac{1}{4}\pi$. The proof of the other results is technical; these can be found in Hundsdorfer (1998,1999). □

Note that in (10.19), with $r = 1$ we get the same angles α as for $r = 0$. Moreover, it is somewhat surprising that there is no difference between $\theta = \frac{1}{2}$ and $\theta = 1$. In Hundsdorfer (1999) also results are given for $|1 + z_0| \le 1$, and then the having $\theta = \frac{1}{2}$ or $\theta = 1$ makes a

difference. If $\theta = 1$ the above statements remain the same. If $\theta = \frac{1}{2}$ we now need $\alpha = 0$, as we saw already with the θ-IMEX method.

In the following figure the boundary of the stability region $|R| \leq 1$ is plotted for two special choices, namely $z_0 = 0$, $z_j = z$ $(1 \leq j \leq s)$ and $z_0 = 0$, $z_j = z$ $(1 \leq j \leq s-1)$, $z_s = \infty$. Plots for the method with $\theta = \frac{1}{2}$ look very similar. Also drawn, as dotted curved lines, are contour lines of $|R|$ at 0.1, 0.2,,...,0.9. From this it is seen that we have little damping in general. If there are two z_j with large values then $|R|$ will be close to 1. The same holds if we are outside the region of stability, where we may have $|R| > 1$ but very close to 1. Consequently, there may be a very slow instability.

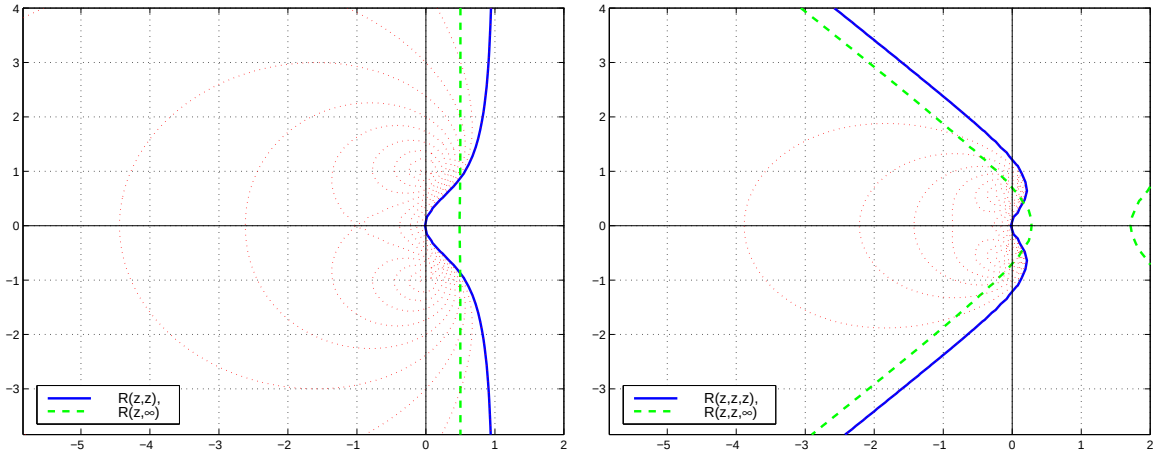


Figure 10.3. Regions of stability $|R| \leq 1$ for $\theta = 1$, $z_0 = 0$, with equal $z_j = z$ or $z_s = \infty$. Left picture $s = 2$, right picture $s = 3$.

In conclusion, if we consider $\alpha = \frac{1}{2}\pi$, then the essential condition for stability is $z_1 \in \mathcal{W}_{\pi/2}$ and $z_2, \ldots, z_s \leq 0$, so only one of the implicit term should have eigenvalues that are large in modulus and not near the negative real axis. If this is violated, instability can be expected. This instability will be quite slow and therefore difficult to detect before it is too late.

Example. To illustrate the slow onset of instability, we consider the following advection equation with a simple linear reaction term,

$$u_t = a u_x + b u_y + G\, u, \quad (x, y) \in [0, 1]^2,\ 0 \leq t. \tag{10.20}$$

The velocities are given by $a(x, y, t) = 2\pi(y - \frac{1}{2})$, $b(x, y, t)) = 2\pi(\frac{1}{2} - x)$. Further,

$$u = u(x, y, t) = \begin{pmatrix} u_1(x, y, t) \\ u_2(x, y, t) \end{pmatrix}, \qquad G = \begin{pmatrix} -k_1 & k_2 \\ k_1 & -k_2 \end{pmatrix}.$$

We take $k_1 = 1$. The second reaction constant k_2 can be used to vary the stiffness of the reaction term, and is taken here as 2000. Note that the matrix G has eigenvalues 0 and $-(k_1 + k_2)$, and we have a chemical equilibrium if $u_1/u_2 = k_2/k_1$.

The initial condition is chosen as

$$u_1(x, y, 0) = c, \quad u_2(x, y, 0) = (1 - c) + 100\, k_2^{-1} \exp(-80((x - \tfrac{1}{2})^2 - 80(y - \tfrac{3}{4})^2),$$

with $c = k_2/(k_1 + k_2)$. After the short transient phase, where most of the Gaussian pulse is transferred from u_2 to u_1, this is purely an advection problem, and the velocity field gives a rotation around the center of the domain. At $t = 1$ one rotation is completed. The exact solution is easily found by superimposing the solution of the reaction term onto the rotation caused by the advection terms.

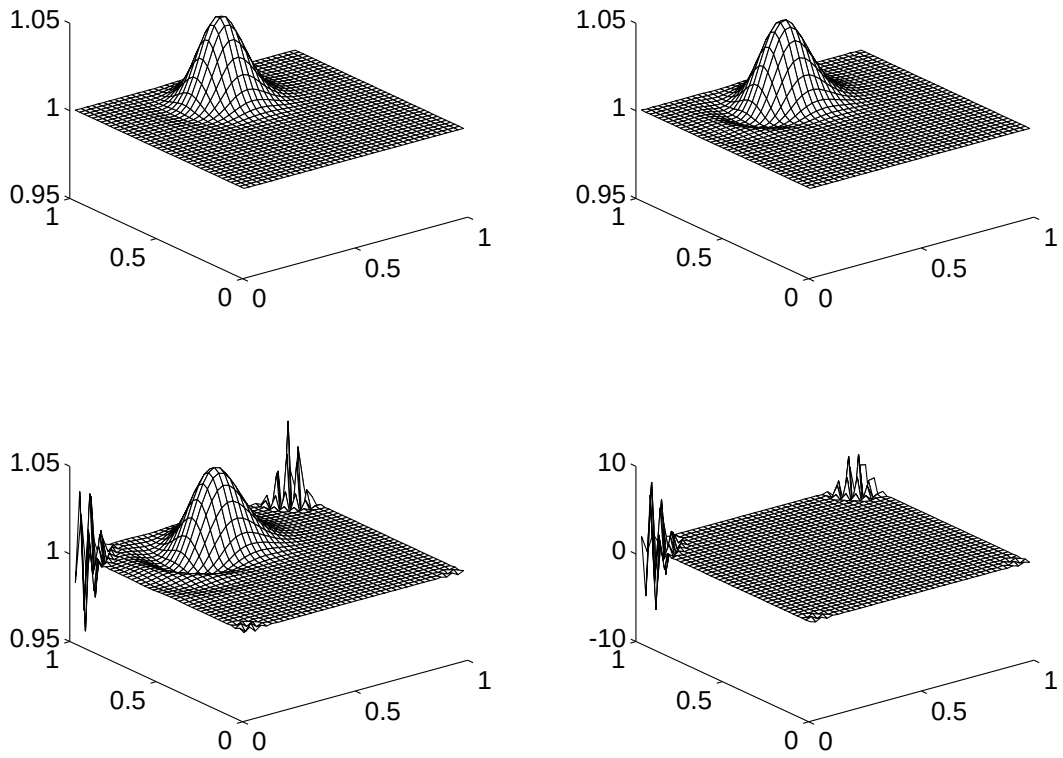


Figure 10.4. Numerical solutions advection-reaction problem (10.20) at $t = 1, 2, 3, 4$.

Dirichlet conditions are prescribed at the inflow boundaries. At the outflow boundaries we use standard upwind discretization, in the interior second order central differences are used. We consider splitting with F_1, F_2 the finite difference operators for advection in the x and y direction, respectively, and with F_3 defined by the linear reaction term. All three terms are treated implicitly. The corresponding eigenvalues λ_1, λ_2 will be close to the imaginary axis whereas $\lambda_3 = 0$ or $-(k_1 + k_2)$. The test has been performed on a fixed 80×80 grid, and with $\tau = 1/160$.

The numerical solution of the first component u_1 for the scheme with $\theta = \frac{1}{2}$ is given in in Figure 10.4 at time $t = 1$ (top left), $t = 2$ (top right), $t = 3$ (bottom left) and $t = 4$ (bottom right; different scale). There are some smooth oscillations in the wake of the Gaussian pulse, but these are caused by the spatial discretization with central differences. The instabilities occur near the corners where both advection speeds, in x and y direction, are large. The build

up of the instabilities is very slow, and therefore it will be difficult to detect this with error estimators. To some extend the slowness of the instability can be attributed to the fact that they occur near an outflow boundary, but related tests have shown that it is mainly caused by the fact that we have amplification factors only slightly larger than 1 in modulus.

Finally it should be noted that the advection treatment here, implicit with central differences, is only justified for problems with smooth solutions. If steep gradients may arise some upwinding or flux limiting is to be preferred. The experiment here merely serves as an illustration of the theoretical results on the stability of the ADI method with $s = 3$. $\diamond$

10.4. Error analysis for the Douglas ADI method

In this subsection an error analysis si predented for the Douglas ADI method with $s \leq 3$. As usual, we restrict ourselves to linear problems, that is

$$F_j(t, w) = A_j w + g_j(t)$$

It is assumed that the problem represents a semi-discrete PDE, so the dimension depends on the mesh width in space h and some of the matrices A_j will contain negative powers of h. For nonhomogeneous boundary conditions, the terms g_j will contain the boundary values relevant to A_j, which will also lead to negative powers of h.

It is assumed that the problem is such that the scheme is stable and $\|(I-\theta\tau A_j)^{-1}\| \leq C$ for arbitrary $\theta, \tau > 0$. The error bounds derived here will not be adversely affected by the mesh width h in the spatial discretization. In particular, we shal write $\mathcal{O}(\tau^p)$ to denote vectors or matrices whose norm can be bounded by $C\tau^p$ with $C > 0$ independent of h. Further, we shall use throughout the paper the notation $Z_j = \tau A_j$, $Z = Z_0 + Z_1 + \cdots + Z_s$ and $Q_j = I - \theta Z_j$.

As the starting point, we consider along with (10.15) the perturbed scheme

$$\left.\begin{aligned} &\tilde{v}_0 = \tilde{w}_n + \tau F(t_n, \tilde{w}_n) + \rho_0, \\ &\tilde{v}_j = \tilde{v}_{j-1} + \theta\tau\Big(F_j(t_{n+1}, \tilde{v}_j) - F_j(t_n, \tilde{w}_n)\Big) + \rho_j \qquad (j = 1, 2, \cdots, s), \\ &\tilde{u}_{n+1} = \tilde{v}_s. \end{aligned}\right\} \tag{10.21}$$

The perturbations ρ_j may stand for round-off or errors introduced in the solution of the implicit systems, for instance. We shall use them to derive an expression for the local discretization errors.

Let $e_n = \tilde{w}_n - w_n$, $\varepsilon_j = \tilde{v}_j - v_j$. Subtraction of (10.15) from (10.21) gives the relations

$$\begin{aligned} &\varepsilon_0 = e_n + Ze_n + \rho_0, \\ &\varepsilon_j = \varepsilon_{j-1} + \theta Z_j(\varepsilon_j - e_n) + \rho_j \qquad (j = 1, 2, \cdots, s), \\ &e_{n+1} = \varepsilon_s. \end{aligned}$$

We can eliminate the internal quantities ε_j by using $\varepsilon_j - e_n = Q_j^{-1}(\varepsilon_{j-1} - e_n + \rho_j)$, leading to

$$e_{n+1} = Re_n + d_n \tag{10.22}$$

with stability matrix

$$R = I + Q_s^{-1} \cdots Q_2^{-1} Q_1^{-1} Z$$

and with d_n containing the internal perturbations,

$$d_n = Q_s^{-1} \cdots Q_1^{-1}(\rho_0 + \rho_1) + Q_s^{-1} \cdots Q_2^{-1}\rho_2 + \cdots Q_s^{-1}\rho_s. \tag{10.23}$$

So, the matrix R determines how an error already present at time t_n will be propagated to t_{n+1}, whereas d_n stands for the local error introduced during the step.

Let $\tilde{w}_n = w(t_n)$ so that $e_n = w(t_n) - w_n$ is the global discretization error. To derive an expression for the local discretization error d_n we are free to chose the $\tilde{v}_j$; it is only the global relation (10.22) that matters. Simple expressions for the residuals ρ_j are obtained by taking $\tilde{v}_j = w(t_{n+1})$ for $j = 0, 1, ..., s$. Then

$$\rho_0 = \frac{1}{2}\tau^2 w''(t_n) + \tfrac{1}{6}\tau^3 w'''(t_n) + \cdots,$$

$$\rho_j = -\theta\tau\Big(\varphi_j(t_{n+1}) - \varphi_j(t_n)\Big) = -\theta\tau^2\varphi_j'(t_n) - \frac{1}{2}\theta\tau^3\varphi_j''(t_n) - \cdots, \quad j = 1, ..., s.$$

Inserting these residuals into (10.23) yields the local discretization error

$$d_n = Q_s^{-1} \cdots Q_1^{-1} \cdot \frac{1}{2}\tau^2 w''(t_n) - \sum_{j=1}^{s} Q_s^{-1} \cdots Q_j^{-1} \cdot \theta\tau^2\varphi_j'(t_n) + \mathcal{O}(\tau^3). \tag{10.24}$$

Note that boundedness of the Q_j^{-1} factors implies that $d_n = \mathcal{O}(\tau^2)$ uniformly in the mesh width h, and by the stability assumption we obtain at least first-order convergence of the global errors e_n independent of h.

If $F_0 = 0$ and $\theta = \frac{1}{2}$ this estimate can be improved, but then we need to take a closer look on the error propagation. We shall elaborate this for $s \le 3$, where we have

$$d_n = Q_3^{-1}Q_2^{-1}Q_1^{-1}\Big(\tfrac{1}{4}\tau^2 Z_1\varphi_2'(t_n) + \tfrac{1}{4}\tau^2(Z_1 + Z_2 - \frac{1}{2}Z_1Z_2)\varphi_3'(t_n)\Big) + \mathcal{O}(\tau^3). \tag{10.25}$$

In case $s = 2$ we can use this formula with $Z_3 = 0$, $\varphi_3 = 0$.

According to the general condition formulated in Section 8, we have second-order convergence if the local error can be decomposed as

$$d_n = (I - R)\xi_n + \eta_n \quad \text{with } \xi_n = \mathcal{O}(\tau^2),\ \eta_n = \mathcal{O}(\tau^3) \text{ and } \xi_{n+1} - \xi_n = \mathcal{O}(\tau^3).$$

Using this framework, convergence results are now easily obtained.

Theorem 10.3. Let $\theta = \frac{1}{2}$, $F_0 = 0$. Consider scheme (10.15) with $s = 2$ and assume that $A^{-1}A_1\varphi_2^{(k)}(t) = \mathcal{O}(1)$ for $k = 1, 2$ and $t \in [0, T]$. Then $e_n = \mathcal{O}(\tau^2)$ for $t_n \in [0, T]$.

Proof. If $s = 2$ we have

$$d_n = Q_2^{-1}Q_1^{-1}\tfrac{1}{4}\tau^2 Z_1\varphi_2'(t_n) + \mathcal{O}(\tau^3) = (R - I)\tfrac{1}{4}\tau^2 Z^{-1}Z_1\varphi_2'(t_n) + \mathcal{O}(\tau^3).$$

Thus we can take $\xi_n = Z^{-1}Z_1\varphi_2'(t_n) = A^{-1}A_1\varphi_2'(t_n)$ and η_n containing the remaining $\mathcal{O}(\tau^3)$ terms. □

For many splittings with standard advection-diffusion problems we will have $\|A^{-1}A_1\| \leq 1$, and hence the assumption $A^{-1}A_1\psi_2 = \mathcal{O}(1)$, $\psi_2 = \varphi_2', \varphi_2''$, in this theorem is natural. Furthermore we note that if A is singular, the above can be easily generalized: what we need to prove second-order convergence is the existence of a vector $v = \mathcal{O}(1)$ such that $Av = A_1\psi_2$. In all of the following such generalizations can be made.

Theorem 10.4. Let $\theta = \frac{1}{2}$, $F_0 = 0$. Consider scheme (10.15) with $s = 3$ and assume that $A^{-1}A_1\varphi_j^{(k)}(t) = \mathcal{O}(1)$ (j=2, 3), $A^{-1}A_2\varphi_3^{(k)}(t) = \mathcal{O}(1)$ and $A^{-1}A_1A_2\varphi_3^{(k)}(t) = \mathcal{O}(\tau^{-1})$ for $k = 1, 2$ and $t \in [0, T]$. Then $e_n = \mathcal{O}(\tau^2)$ for $t_n \in [0, T]$.

Proof. Since $R = I + Q_3^{-1}Q_2^{-1}Q_1^{-1}Z$, the local discretization error can be written as

$$d_n = (R - I)Z^{-1}\Big(\tfrac{1}{4}\tau^2 Z_1\varphi_2'(t_n) + \tfrac{1}{4}\tau^2(Z_1 + Z_2 - \frac{1}{2}Z_1Z_2)\varphi_3'(t_n)\Big) + \mathcal{O}(\tau^3).$$

Note that $A^{-1}A_1A_2\varphi_3^{(k)} = \mathcal{O}(\tau^{-1})$ implies $Z^{-1}Z_1Z_2\varphi_3^{(k)} = \mathcal{O}(1)$. Thus we can proceed in the same way as before with ξ_n containing the $\mathcal{O}(\tau^2)$ terms. □

Compared to the situation for $s = 2$, Theorem 10.3, the essential new condition here with $s = 3$ is $A^{-1}A_1A_2\psi_3 = \mathcal{O}(\tau^{-1})$, $\psi_3 = \varphi_3', \varphi_3''$, that is,

$$Z^{-1}Z_1Z_2\psi_3 = \mathcal{O}(1).$$

This may hold also if $Z^{-1}Z_1Z_2 \neq \mathcal{O}(1)$. As an example, consider A_j to be the standard second-order difference operator for $\partial^2/\partial {x_j}^2$, $j = 1, 2, 3$ with nonhomogeneous Dirichlet conditions at the boundaries. Then the matrices A_j commute and $\|A^{-1}A_j\| \leq 1$ for all h. Further it holds that $A_1^\gamma A_2^\gamma\psi_3 = \mathcal{O}(1)$ for any $\gamma < \frac{1}{4}$ (with $\gamma = \frac{1}{4}$ we have $A_1^\gamma A_2^\gamma\psi_3 = \mathcal{O}(\log(h))$, see Hundsdorfer & Verwer (1989). So we can write

$$Z^{-1}Z_1Z_2\psi_3 = \tau^{2\gamma}Z^{-1}Z_1^{1-\gamma}Z_2^{1-\gamma}[A_1^\gamma A_2^\gamma\psi_3].$$

Taking $\tau \sim h^{1+\epsilon}$ with $\epsilon = 1 - 4\gamma > 0$, it follows that

$$\|Z^{-1}Z_1Z_2\psi_3\| \sim \tau^{2\gamma}\Big(\frac{\tau}{h^2}\Big)^{1-2\gamma} = \mathcal{O}(1).$$

Thus the conditions in Theorem 10.4 are fulfilled under a step size restriction $\tau \sim h^{1+\epsilon}$ with $\epsilon > 0$ arbitrarily small. In a similar way it can also be shown that if $\tau \sim h$ then the global errors e_n can be bounded by $\tau^2\log(\tau)$, convergence with order practically equal to 2. Further we note that if φ_3', φ_3'' satisfy homogeneous boundary conditions on the boundaries relevant to A_1 and A_2 then no condition on the ratio τ/h is necessary, since then $A_1A_2\psi_3 = \mathcal{O}(1)$.

In conclusion, Theorem 10.4 indicates that also with $s = 3$ we will often have second-order convergence, although a mild restriction on the step size might be necessary in this case.

For larger values of s a similar analysis could be performed, but verification of the accuracy conditions becomes increasingly technical. For example, if $s = 4$ we get, in addition to conditions as in Theorem 10.4, the requirement $Z^{-1}Z_1Z_2Z_3\psi_4(t) = \mathcal{O}(1)$, that is, $A^{-1}A_1A_2A_3\psi_4(t) = \mathcal{O}(\tau^{-2})$. Although this may be fulfilled in many special cases, in general an order of convergence between 1 and 2 must now be expected.

Note. The above derivation is taken from Hundsdorfer (2000). A similar analysis was obtained in Hundsdorfer & Verwer (1989) for the Peaceman-Rachford ADI method and in Hundsdorfer (1992) for LOD methods. With the LOD method based on Strang splitting very low orders of convergence may occur (for example 1/2 in the L_2-norm), and then boundary corrections are necessary. Such corrections are given in Mitchell & Griffiths (1980) for ADI and LOD methods. With the above ADI method boundary corrections give in general somewhat more accurate results, but it also gives more complicated programs, and as we saw the order of convergence is in general already 1 or 2 (if $F_0 = 0$, $\theta = \frac{1}{2}$), which is the same as for fixed (nonstiff) ODEs.

10.5. Rosenbrock methods with approximate factorization

With the above ADI method we still still are dealing with the explicit Euler method for F_0. To allow a second order explicit method we first consider a linearization of this ADI method. In the following only autonomous equations are considered.

As starting point we consider the linearized θ-method

$$w_{n+1} = w_n + (I - \theta\tau A)^{-1}\tau F(w_n) \tag{10.26}$$

where A approximates the Jacobian matrix $F'(w_n)$. This is a so-called Rosenbrock method. It has order 1 if $\theta \neq \frac{1}{2}$ and order 2 if $\theta = \frac{1}{2}$ and $A - F'(w_n) = \mathcal{O}(\tau)$.

We consider the form where in the Jacobian approximation the nonstiff term is omitted and the rest is factorized in approximate fashion, that is

$$w_{n+1} = w_n + (I - \theta\tau A_s)^{-1} \cdots (I - \theta\tau A_2)^{-1}(I - \theta\tau A_1)^{-1}\tau F(w_n) \tag{10.27}$$

with $A_j \approx F_j'(w_n)$. The order of this approximate factorization method is 1 in general. For second order we need $\theta = \frac{1}{2}$ and $F_0 = 0$. If the problem is linear this approximate factorization method is identical to the Douglas ADI method. Hence the linear stability properties are the same. Approximate factorization methods of the above type were introduced by Beam & Warming (1976).

A 2-stage generalization of the above approximate factorization method is given by

$$\begin{aligned} w_{n+1} &= w_n + \tfrac{3}{2}k_1 + \tfrac{1}{2}k_2, \\ Mk_1 = \tau F(t_n, w_n), &\qquad Mk_2 = \tau F(t_n + c\tau, w_n + k_1) - 2k_1, \end{aligned} \tag{10.28}$$

where $M = \prod_{j=1}^{s}(I - \theta\tau A_j)$, $A_j \approx F_j'(w_n)$ and θ is a free parameter. The order of this method is 2 (in the classical ODE sense). If $F_0 = 0$ and $F_1 = F$ this is a well-known Rosenbrock method that has the special property that the order is not influenced by the Jacobian approximation. This Rosenbrock method is A-stable for $\theta \geq \frac{1}{4}$. On the other hand, if $F = F_0$ we now get a second order explicit Runge-Kutta method.

The above method has been proposed in Verwer et al. (1999), and in that paper the scheme was applied successfully on some 3D atmospheric transport-chemistry. problems. There operator splitting was used with F_0 advection, F_1 diffusion and F_2 reaction, and the free parameter was taken as $\theta = 1 + \frac{1}{2}\sqrt{2}$ to have optimal damping (L-stability). The eigenvalues of F_1 and F_2 were close to the negative real axis, and therefore stability problems were not expected, and indeed did not occur.

It is for such problems, where the structure of the eigenvalues can be well predicted in advance, that these approximate factorization methods seem suited. For general applications values θ in the range $[\frac{1}{2}, 1]$ seem more suitable than $\theta = 1 + \frac{1}{2}\sqrt{2}$, because the latter value gives relatively large error constants.

The above Rosenbrock methods are formulated here for autonomous problems. A nonautonomous problem $w'(t) = F(t, w(t))$ can be written as $v'(t) = G(v(t))$ with $v = (t, w)^T$ and $G(v) = (1, F(t, w))^T$, and so the methods can be applied to this artificial autonomous problem. Then t is formally also considered as an unknown, but it is easily seen that the approximation t_n found with this method still equals $n\tau$. When reformulated on the original level, in terms of w_n, the methods will now also involve approximations to the derivatives $F_t(t, w)$. For example, with $A_j \approx \partial_w F_j(t_{n+\theta}, w_n) \in \mathbb{R}^{m\times m}$, $b_j \approx \partial_t F_j(t_{n+\theta}, w_n) \in \mathbb{R}^m$ and

$$B_j = \begin{pmatrix} 1 & 0 \cdots 0 \\ (I - \theta\tau A_s)^{-1}\theta\tau b_j & (I - \theta\tau A_s)^{-1} \end{pmatrix} \in \mathbb{R}^{(m+1)\times(m+1)},$$

the factorized Rosenbrock scheme (10.26) then reads

$$\begin{pmatrix} t_{n+1} \\ w_{n+1} \end{pmatrix} = \begin{pmatrix} t_n \\ w_n \end{pmatrix} + B_s \cdots B_2 B_1 \begin{pmatrix} \tau \\ \tau F(t_n, w_n) \end{pmatrix}.$$

We will have $t_{n+1} = t_n + \tau$, as it should be, and the computation of w_{n+1} can be written in the more transparent recursive form

$$f_0 = F(t_n, w_n), \quad f_j = (I - \theta\tau A_j)^{-1}(\theta b_j + f_{j-1}) \ \ (1 \le j \le s), \quad w_{n+1} = w_n + \tau f_s.$$

Note. It is also possible to linearize a multistep method and then use approximate factorization. Such methods can be found in Warming & Beam (1979). Runge-Kutta methods of the IMEX type have been studied in Ascher et al. (1997); if such methods are applied in a linearized form, they are similar to the above factorized Rosenbrock methods with $s = 1$.

Remark. Instead of the above techniques, one could also use a well-known fully implicit method and then try to modify the Newton process such that the computational ease is comparable to the IMEX or approximate factorization methods. The advantage is that if the iteration converges, then the theoretical properties of the fully implicit method are valid.

Consider a generic implicit relation

$$w_{n+1} = W_n + \theta\tau F(w_{n+1}), \tag{10.29}$$

where W_n contains the information up to t_n. This may be for instance Backward Euler ($\theta = 1$, $W_n = w_n$), the Trapezoidal Rule ($\theta = \frac{1}{2}$, $W_n = w_n + \frac{1}{2}\tau F(t_n, w_n)$) or the BDF2 method ($\theta = \frac{2}{3}$, $W_n = \frac{4}{3}w_n - \frac{1}{3}w_{n-1}$). Then the Newton iteration to solve the implicit relation will look like

$$u_{i+1} = u_i - M^{-1}\Big(u_i - \theta\tau F(u_i) - W_n\Big), \quad i = 0, 1, 2, ... \tag{10.30}$$

with initial guess u_0. Standard modified Newton would be $M = I - \theta\tau A$ with $A \approx F'(v_0)$. For systems of multi-dimensional PDEs this leads to a very big linear algebra problem that has to be solved by a preconditioned conjugate gradient or multigrid method for example.

As an alternative one can consider approximate factorization inside the Newton process,

$$M = \prod_{j=1}^{s} (I - \theta\tau A_j) \tag{10.31}$$

with $A_j \approx F_j'(v_0)$, but now we have to look at convergence of the iteration.

When applied to the scalar test equation this iteration process has a convergence factor

$$S = 1 - \Big(\prod_{j=1}^{s}(1 - \theta z_j)\Big)^{-1}\Big(1 - \theta\sum_{j=0}^{s} z_j\Big) \tag{10.32}$$

and for the iteration to converge we need $|S| < 1$. This looks very similar to the stability factor with the Douglas ADI method. Indeed, the statements given previously for $|R| \le 1$ with the z_j in wedges are also valid for the convergence factor, see Hundsdorfer (1999).

In the next figure the boundaries of the convergence region are plotted for special choices of z_j with $z_0 = 0$, similar to Figure 10.3. The dotted curved lines are the contour lines for $|S|$ with all z_j equal. If the z_j assume large negative values, then $|S|$ is close to 1 and thus the convergence will be very slow. Moreover divergence may occur if $s \ge 3$ and two or more of the z_j are close to the imaginary axis.

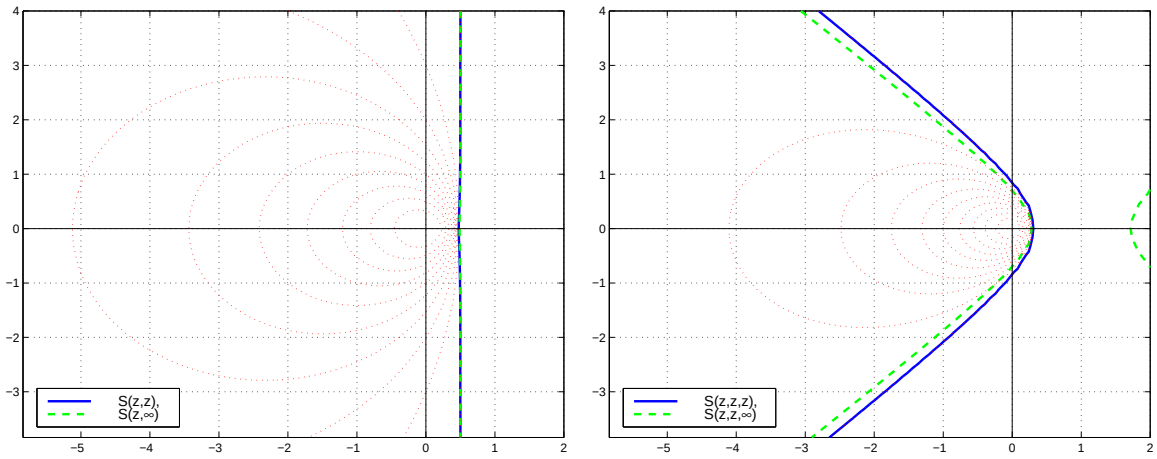


Figure 10.5. Regions of convergence $|S| < 1$ for $\theta = 1$ with equal $z_j = z$ or $z_s = \infty$. Left picture $s = 2$, right picture $s = 3$.

In conclusion it can be said that the convergence of such a modified Newton iteration with approximate factorizations is often very poor, so it is not an approach that is recommended for general equations. Of course, there are special cases, especially with smooth solutions (no high Fourier harmonics), where this approach may work well. However the class of problems where the iteration does not diverge seems close to the class where the Rosenbrock schemes

with approximate factorizations are be stable, see Figures 10.3 and 10.5. In those cases the simpler Rosenbrock schemes with approximate factorizations will be more efficient, and with such Rosenbrock schemes smoothness of the solution is not required.

10.6. Numerical illustration

In this section some numerical illustrations are given for the schemes applied to 2D advection-diffusion-reaction equations. We shall refer to the 1-stage scheme (10.26) as ROS1 and to the 2-stage scheme (10.28) as ROS2, and for both schemes parameter values $\theta = \frac{1}{2}$ and 1 are considered.

We consider here the following 2D equation, on spatial domain $\Omega = [0,1]^2$ and $t \in [0,1]$,

$$u_t + \alpha(u_x + u_y) = \epsilon(u_{xx} + u_{yy}) + \gamma u^2(1-u), \tag{10.33}$$

with traveling wave solution

$$u(x,y,t) = \Big(1 + \exp(a(x+y-bt)+c)\Big)^{-1}. \tag{10.34}$$

Here $a = \sqrt{\gamma/4\epsilon}$ determines the smoothness of the solution, $b = 2\alpha + \sqrt{\gamma\epsilon}$ is the velocity of the wave and $c = a(b-1)$ a shift parameter. Initial and Dirichlet boundary conditions are prescribed so as to correspond with this solution. Due to the time-dependent boundary conditions, the semi-discrete problem is non-autonomous and the Rosenbrock methods are applied to the extended autonomous form mentioned in Section 10.5.

For this scalar test example splitting is not really necessary, but the structure of the equations is similar to many real-life problems where splitting cannot be avoided with present day computer (memory) capacities. In Verwer et al. (1999) application the ROS2 method can be found for a large scale 3D problem from atmospheric dispersion.

Reaction-diffusion test. First we consider the above test equation with $\alpha = 0$. To give an illustration of the convergence behaviour of the various methods we take $\gamma = 1/\epsilon = 10$, which gives a relatively smooth solution.

For this smooth problem the spatial derivatives are discretized with standard second order finite differences. Let $D^{(x)}(t,u) = A^{(x)}u + g^{(x)}(t)$ stand for the finite difference approximation of ϵu_{xx} with the associated time-dependent boundary conditions for $x = 0$ and $x = 1$. Likewise $D^{(y)}(t,u)$ approximates ϵu_{yy} with boundary conditions at $y = 0$, $y = 1$. Further, $G(t,u)$ represents the reaction term $\gamma u^2(1-u)$ on the spatial grid. We consider the following two splittings with $s = 3$ and $F_0 = 0$,

$$(A) \quad \cdots \qquad F_1 = D^{(x)}, \quad F_2 = D^{(y)}, \quad F_3 = G,$$

and

$$(B) \quad \cdots \qquad F_1 = G, \quad F_2 = D^{(x)}, \quad F_3 = D^{(y)}.$$

Since the reaction term in (10.33) with $\gamma = 10$ is not stiff, we also consider here the case where this term is taken explicitly,

$$(C) \quad \cdots \qquad F_0 = G, \quad F_1 = D^{(x)}, \quad F_2 = D^{(y)}.$$

The spatial grid is uniform with mesh width h in both directions. The errors in the L_2-norm are calculated at output time $T = 1$ with $\tau = h = 1/N$, $N = 10, 20, 40, 80$. In the Figure 10.6 these errors are plotted versus τ on a logarithmic scale. The results for the ROS1 scheme are indicated by dashed lines with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$. Likewise, the results for the ROS2 scheme are indicated by solid lines with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$.

For comparison, results of the well-known fractional step (LOD) method of Yanenko (1971) are included, indicated by dotted lines with stars. With this method fractional steps are taken with the implicit trapezoidal rule $v_j = v_{j-1} + \frac{1}{2}\tau F_j(t_n, v_{j-1}) + \frac{1}{2}\tau F_j(t_{n+1}, v_j)$, with $v_0 = w_n$. After each step the order of the F_j is interchanged to achieve symmetry and second order (in the classical ODE sense), see formula (9.11) with $c = \frac{1}{2}$. If an explicit term F_0 is present, the implicit trapezoidal rule is replaced by its explicit counterpart for the fractional step with F_0.

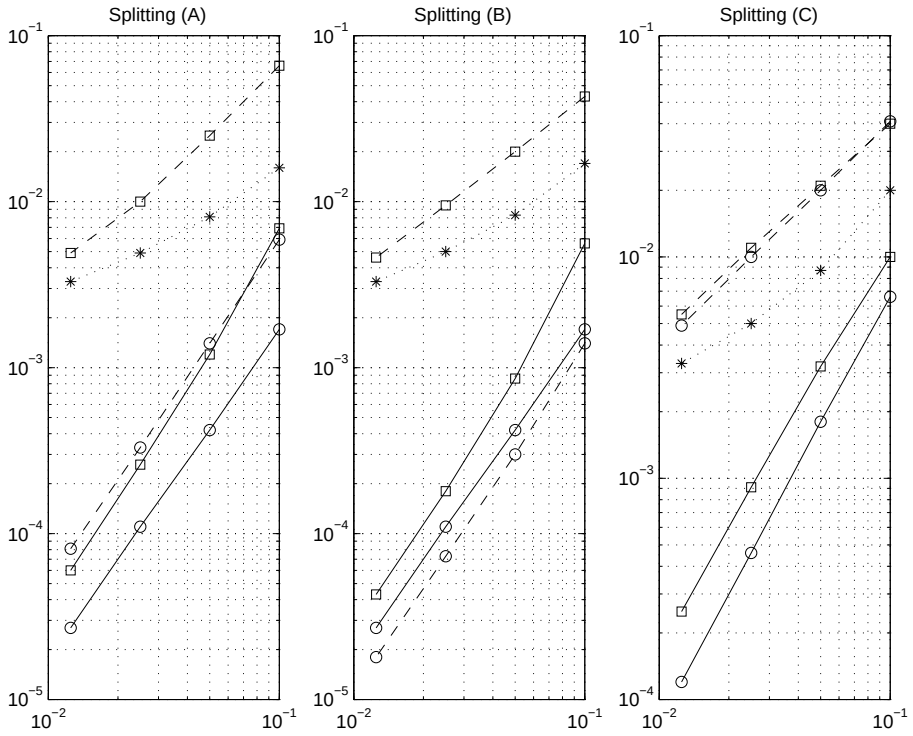


FIGURE 10.6. L_2-errors versus $\tau = h$ for the splittings (A), (B) and (C). Methods ROS1 (dashed lines) and ROS2 (solid lines) with $\theta = \frac{1}{2}$ (circles) and $\theta = 1$ (squares). Results for Yanenko's method are indicated with stars.

It is known that Yanenko's method needs boundary corrections to obtain second-order convergence for initial-boundary value problems, otherwise order of convergence can be lower, see Hundsdorfer (1992). In the present test we get convergence with order $\frac{1}{2}$ approximately. The test was repeated with boundary corrections, but still the results were less accurate than with the second-order ROS schemes. Finally we note that boundary corrections were also attempted on the scheme (1.3), similar to formula (101) in Mitchell & Griffiths (1980). In the above test this did lead to smaller errors, reduction with a factor ranging between 1.2 and 2, but

the convergence behaviour did not change fundamentally. Since boundary corrections have to be derived for each individual problem, it is a favourable property of the stabilizing correction schemes that such corrections are not necessary to get a genuine second-order behaviour.

Advection-diffusion-reaction test. To illustrate the improved stability behaviour of the 2-stage scheme ROS2 over ROS1 if a substantial explicit term is present, we now consider the test equation with a advection term with $\alpha = -1$ that will be taken explicitly. Further we choose $\gamma = 100$ and $\epsilon = 0.01, 0.001$ which gives solutions that have a steep gradient, relative to the mesh widths used here.

The splitting is such that F_0 contains the convective terms, F_1, F_2 diffusion in x and y direction, respectively, and F_3 the nonlinear reaction term. The convective terms are discretized with third-order upwind-biased differences (4-point stencil). For the diffusion terms standard second-order differences are used as before.

The results with $\epsilon = 0.01$ are given in the Figures 10.7, 10.8. In the plots of Figure 10.7 the solutions $h = 1/40$ and $\tau = 1/80$ are found, represented as contour lines at the levels 0.1, 0.2,...,0.9, with solid lines for the numerical solution and dotted lines for the exact solution. Quantitative results are given in Figure 10.8, where the L_2-errors are plotted as function of the time step for a 40×40 and 80×80 grid with $\tau = h, \frac{1}{2}h$ and so on. As in Figure 10.6 results for ROS1 are indicated with dashed lines, for ROS2 with solid lines, and with squares if $\theta = 1$ and circles if $\theta = \frac{1}{2}$.

It is obvious that the 2-stage schemes ROS2 give much better results than the corresponding 1-stage schemes ROS1. To achieve a level of accuracy comparable to the ROS2 schemes we need much smaller time steps with the ROS1 schemes, see Figure 10.8. This is primarily due to the more stable treatment of the explicit advection term with the ROS2 schemes. The explicit 2-stage Runge-Kutta method underlying ROS2 is stable for third-order advection discretization up to Courant number 0.87 (experimental bound). On the other hand, some of the eigenvalues associated with this discretization are always outside the stability region of the explicit Euler scheme. In this test it is the (implicit) diffusion part that provides a stabilization for the smaller step sizes. (In fact, for $\epsilon = 0.01$ similar results were obtained with second order central advection discretization, but not anymore with $\epsilon = 0.001$). Further we note that instabilities do not lead to overflow since the solutions are pushed back to the range [0,1] by the reaction term, but the resulting numerical solutions are qualitatively wrong.

Decreasing the value of the diffusion coefficient ϵ gives a clearer distinction between the methods. Results with $\epsilon = 0.001$ are given in the Figures 10.9 and 10.10. The grids chosen are 80×80 and 160×160, since the 40×40 grid gives quite large spatial errors with this small ϵ. The results are essentially the same as above: the 1-stage schemes ROS1 need much smaller time steps than the ROS2 schemes to obtain reasonable solutions.

For more realistic problems, with stiff reaction terms, nonlinear advection discretizations with flux limiters are recommended to avoid oscillations, and this fits easily into the present framework, see Verwer et al. (1999) for instance.

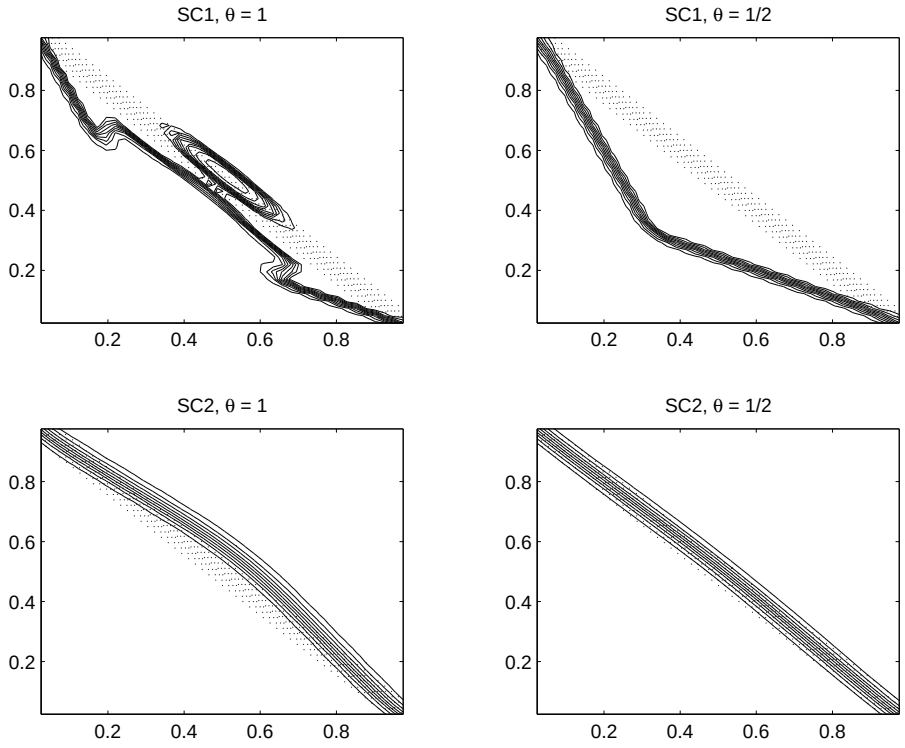


FIGURE 10.7 Contour lines for $\epsilon = 0.01$ with $h = 1/40$, $\tau = 1/80$.

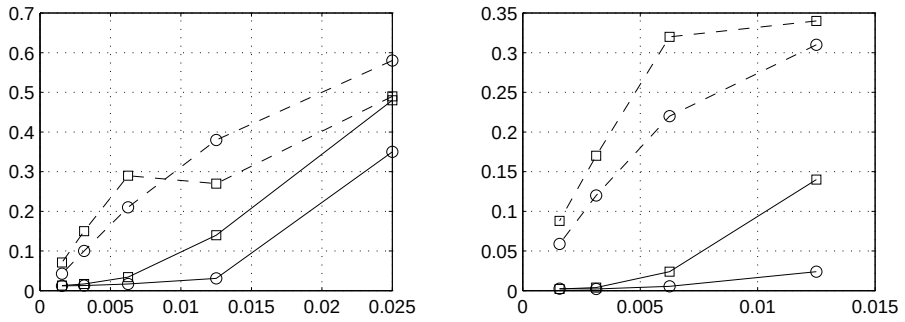


FIGURE 10.8. L_2-errors versus time step τ on 40×40 grid (left) and 80×80 grid (right) for $\epsilon = 0.01$. Various methods indicated as in Figure 10.6.

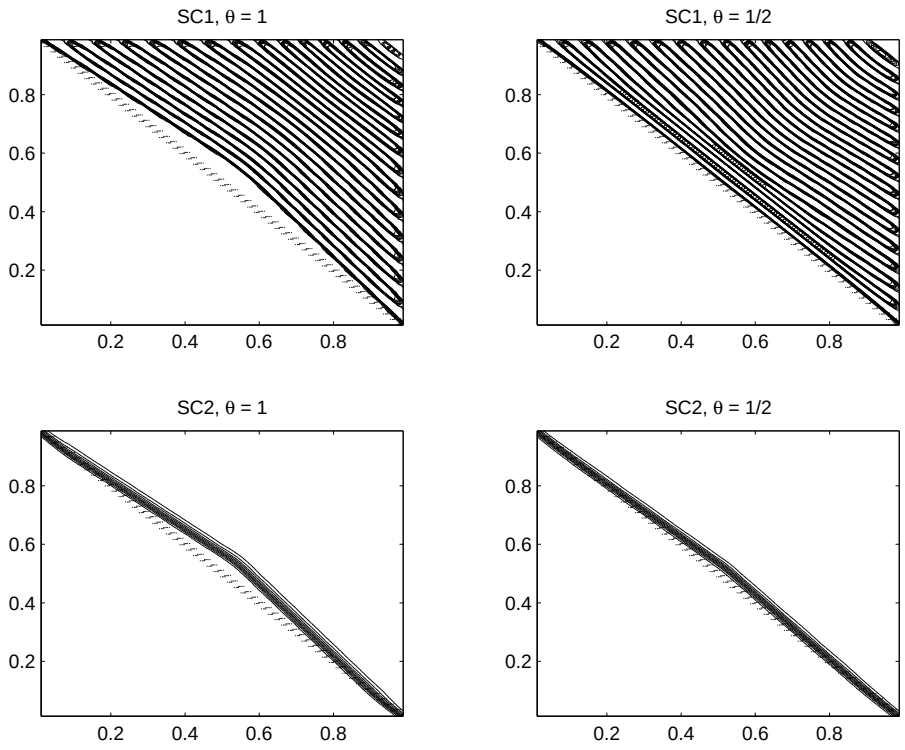


FIGURE 10.9 Contour lines for $\epsilon = 0.001$ with $h = 1/80$, $\tau = 1/160$.

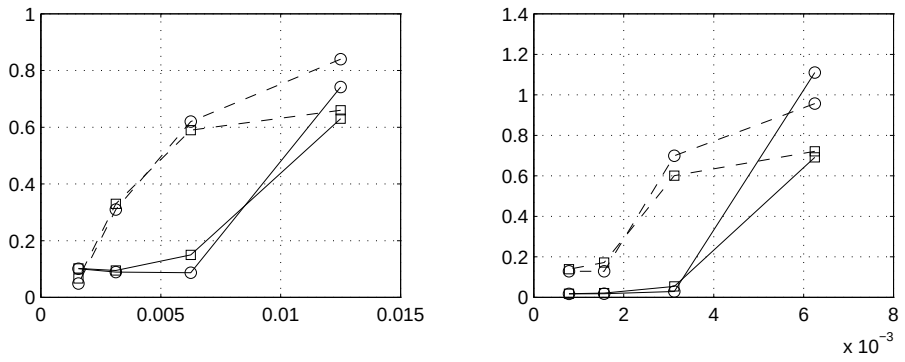


FIGURE 10.10. L_2-errors versus time step τ on 80×80 grid (left) and 160×160 grid (right) for $\epsilon = 0.001$. Various methods indicated as in Figure 5.1.

11. Appendices on ODE methods

For the solution of initial value problems for systems of ODEs there are many sophisticated and efficient computer codes, usually based on Runge-Kutta methods or linear multi-step methods. Here we give some examples of such methods, together with a few properties. As a rule of thumb: for problems on sufficiently large time intervals where the step sizes need not be changed too drastically, linear multi-step methods seem more efficient, whereas if we either have short integration intervals (for instance, in a splitting method) or if the step sizes need frequent and big adjustments, then the Runge-Kutta methods seem to be preferable. Good general references on ODE methods are provided by the books of Hairer, Nørsett & Wanner (1987), Hairer & Wanner (1991) and Lambert (1991).

For convenience we only consider methods with fixed step sizes, but it should be emphasized that in many applications variable step sizes are crucial to obtain an efficient code. The exact solution of the ODE problem

$$w'(t) = F(t, w(t)), \qquad w(0) = w_0$$

will be approximated in the points $t_n = n\tau$, $n = 0, 1, 2, \cdots$, with $\tau > 0$ being the step size. The numerical approximations are $w_n \approx w(t_n)$.

11.1. Appendix A : Runge-Kutta methods

When solving the differential equation $w'(t) = F(t, w(t))$ with a Runge-Kutta method, one obtains a new approximation w_{n+1} by first computing intermediate approximations $w_{ni} \approx w(t_n + c_i\tau)$, $i = 1, 2, \ldots, s$, where the integer s is called the number of stages used in the method. The general form of a Runge-Kutta method is

$$w_{n+1} = w_n + \tau \sum_{i=1}^{s} b_i F(t_n + c_i\tau, w_{ni}) \tag{A.1a}$$

$$w_{ni} = w_n + \tau \sum_{j=1}^{s} a_{ij} F(t_n + c_j\tau, w_{nj}), \qquad i = 1, \ldots, s, \tag{A.1b}$$

with $n = 0, 1, 2, \cdots$. Here a_{ij} and b_i are coefficients defining the particular method and $c_i = \sum_{j=1}^{s} a_{ij}$. The method is explicit if $a_{ij} = 0$ for $j \geq i$, since then the internal vectors $w_{n1}, w_{n2}, \ldots, w_{ns}$ can be computed one after another from an explicit relation. A Runge-Kutta method can be represented in a compact way by the array

$$\begin{array}{c|c} c & A \\ \hline & b^T \end{array} \quad = \quad \begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}$$

The method is said to have *order* p if $w(t_1) - w_1 = O(\tau^{p+1})$ whenever $w(0) = w_0$ and F is sufficiently smooth. This means that the *local* discretization error, that is, the error introduced

in one step of the time-integration method, is of $O(\tau^{p+1})$. The *global* discretization-error $w(t_n) - w_n$ is formed by n such local errors and will be of $O(\tau^p)$ if $w_0 = w(0)$, the function F is sufficiently smooth and $t_n \leq T$. Thus a method of order p is convergent of order p when applied to a *fixed*, smooth ODE problem.

The order of a Runge-Kutta method is, of course, determined by its coefficients a_{ij}, b_i, c_i. By making Taylor developments of $w(t_1)$ and w_1 in powers of τ, and requiring that these developments are identical up to $O(\tau^p)$ one obtains the order conditions for the coefficients. The conditions for $p = 1, 2, 3, 4$ are summarized in the following table, with $C = \text{diag}(c_i)$ and $e = (1, 1, \ldots, 1)^T$.

order p	order conditions	
1	$b^T e = 1$	
2	$b^T c = 1/2$	
3	$b^T c^2 = 1/3$	$b^T Ac = 1/6$
4	$b^T c^3 = 1/4$	$b^T CAc = 1/8$
	$b^T Ac^2 = 1/12$	$b^T A^2 c = 1/24$

The derivation of higher order methods is quite complicated and involve many order conditions. A systematic approach consists of the use of *Butcher trees*, see Butcher (1987) or Hairer et al. (1987).

The *stage order* q is the minimal order over all internal stages, that is, q is such that $w(c_i\tau) - w_{0i} = O(\tau^{q+1})$ for $i = 1, \ldots, s$ whenever $w(0) = w_0$ and F is sufficiently smooth. Although we are not interested in accuracy of the intermediate vectors, this stage order has some relevance for the accuracy of the approximations w_n for semi-discrete systems arising from PDEs with boundary conditions. For any reasonable method it holds that $q \leq p$, and for many methods q is substantially smaller than p.

Example A.1. The most simple explicit method is the forward Euler method. Two well-known second order explicit Runge-Kutta methods are given by the arrays

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}$$

The first method is called the explicit trapezoidal rule. A typical example of an explicit method with a higher order is the following method with $p = s = 4$

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

In fact, this method used to be called *the* method of Runge-Kutta. We shall refer to it as the classical 4-th order method. Any explicit Runge-Kutta method has stage order $q = 1$, since the second stage is the forward Euler method with step size τa_{21} (the first stage is trivial, $w_{n1} = w_n$). ◇

Example A.2. Some simple implicit methods are :

(i) the backward Euler method ($p = q = 1$)

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \text{,i.e.,} \qquad w_{n+1} = w_n + \tau F(t_{n+1}, w_{n+1}),$$

(ii) the implicit midpoint rule ($p = 2, q = 1$)

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array} \qquad \text{,i.e.,} \qquad w_{n+1} = w_n + \tau F(t_n + \frac{1}{2}\tau, \frac{1}{2}w_n + \frac{1}{2}w_{n+1}),$$

(iii) the trapezoidal rule ($p = q = 2$)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \text{,i.e.,} \qquad w_{n+1} = w_n + \frac{1}{2}F(t_n, w_n) + \frac{1}{2}F(t_{n+1}, w_{n+1}).$$

Generalizations of the above methods are the *collocation methods*, see Hairer et al. (1987), which have high order and good stability properties, but a full matrix A. ◇

With implicit methods the internal vectors have to be solved from a system of algebraic equations, usually by a Newton type iteration. If A is a full matrix the dimension of this system is ms, where m is the dimension of the differential equation. A compromise is found in the *diagonally implicit* methods where A is lower triangular so that we can first solve w_{1n}, then w_{2n}, and so on.

Example A.3. Two classes of diagonally implicit methods, with a parameter $\theta > 0$, are

$$\begin{array}{c|cc} \theta & \theta & \\ 1-\theta & 1-2\theta & \theta \\ \hline & 1/2 & 1/2 \end{array} \quad , \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 2\theta & \theta & \theta & \\ 1 & b_1 & b_2 & \theta \\ \hline & b_1 & b_2 & \theta \end{array} \qquad \begin{array}{l} b_1 = \frac{3}{2} - \theta - \frac{1}{4\theta} \\ b_2 = -\frac{1}{2} + \frac{1}{4\theta} \end{array}$$

Both methods have order $p = 3$ if $\theta = \frac{1}{2} \pm \frac{1}{6}\sqrt{3}$, and $p = 2$ for other θ values. The first method has stage order $q = 1$ since the first stage consists of a backward Euler step, whereas the second method has $q = 2$ (its first nontrivial stage is a trapezoidal rule step). ◇

Implicit methods are more expensive per step than explicit ones. Yet, implicit methods are often used, for instance for parabolic problems and stiff chemistry problems, because of their superior stability properties.

The stability function

The stability properties of ODE methods are, to a large extend, determined by the behaviour of the methods on the scalar, complex test equation

$$w'(t) = \lambda w(t).$$

Let $z = \tau\lambda$. Application of a Runge-Kutta to the test equation gives

$$w_{n+1} = R(z)w_n,$$

with a rational function R, the so-called *stability function*. For the general Runge-Kutta method (A.1) this function can be found to be

$$R(z) = 1 + zb^T(I - zA)^{-1}e \tag{A.2}$$

where $e = (1, 1, \ldots, 1)^T$. By considering $(I - zA)^{-1}$ in terms of determinants it follows that for explicit methods $R(z)$ is a polynomial of degree $\leq s$. For implicit methods it is a rational function with degree of both denominator and numerator $\leq s$.

If the Runge-Kutta method has order p , then

$$R(z) = e^z + O(z^{p+1}), \quad z \to 0.$$

This can be seen by considering he scalar test equation with $w_0 = w(0)$ and $|\lambda| = 1$, since we then know that $w(t_1) - w_1 = O(\tau^{p+1})$ but also $w(t_1) - w_1 = e^\tau - R(\tau)$.

The *stability region* of the method is defined as the set

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

If $\mathcal{S}$ encloses the whole left-half plane $\mathbb{C}^-$, then the method is said to be *A-stable*. Explicit methods cannot be A-stable.

The exact solution of the test equation satisfies $w(t_{n+1}) = e^{\tau\lambda}w(t_n)$, so the solution does not grow in modulus if $\mathrm{Re}\lambda \leq 0$. For an A-stable method, the numerical approximations have the same property no matter how large the step size is chosen. According to the maximum modulus principle, A-stability is equivalent to saying that R has no poles in $\mathbb{C}^-$ and $|R(it)| \leq 1$ for all real t.

The stability function of an explicit method with $p = s$ (possible for $s \leq 4$) equals

$$R(z) = 1 + z + \frac{1}{2}z^2 + \cdots + \frac{1}{s!}z^s.$$

For $s = 1, 2, 4$, respectively, this gives the stability functions of the forward Euler method, the second order methods of Example A.1 and the classical 4-th order Runge-Kutta method. Pictures of the stability regions for the above stability functions with $s = 1, 2, 3, 4$ are given in Figure A.1. Pictures for some higher order methods can be found in Hairer & Wanner (1991)

The trapezoidal rule has the stability function

$$R(z) = \frac{1 + \frac{1}{2}z}{1 - \frac{1}{2}z},$$

which is the same as for the implicit midpoint rule. The stability region for this R is precisely the left-half plane. Note that $|R(\infty)| = 1$ so there is no damping at infinity. The stability function of the backward Euler method is

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

and this method is A-stable with $|R(\infty)| = 0$.

The two diagonally implicit methods of Example A.3 have the same stability function

$$R(z) = \frac{1 + (1-2\theta)z + (\frac{1}{2} - 2\theta + \theta^2)z^2}{(1-\theta z)^2},$$

and the methods are A-stable iff $\theta \geq \frac{1}{4}$. Thus for the two θ values leading to order 3 only $\theta = \frac{1}{2} + \frac{1}{6}\sqrt{3}$ gives A-stability. Further, $R(\infty) = 0$ if $\theta = 1 \pm \frac{1}{2}\sqrt{2}$.

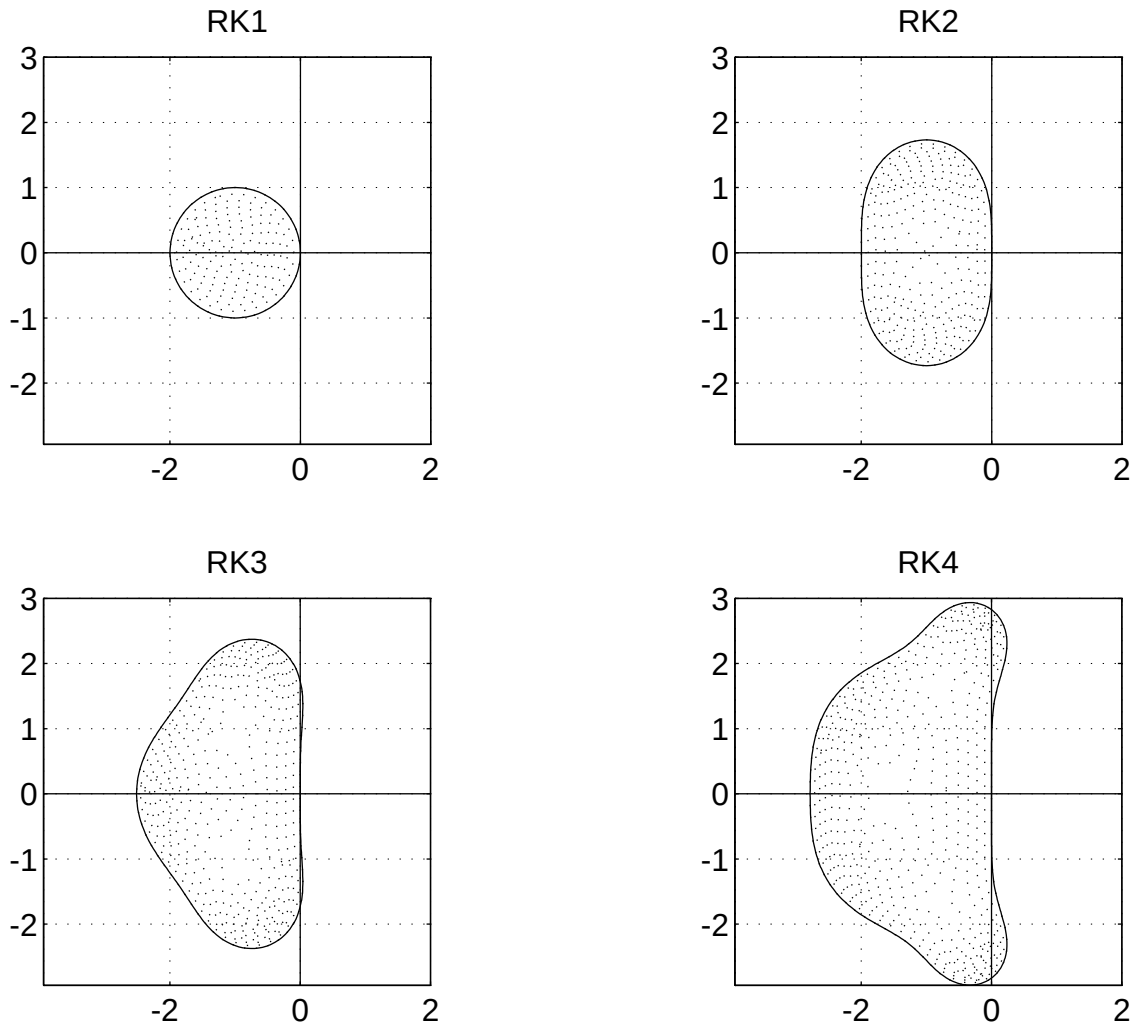


FIGURE A.1. STABILITY REGIONS FOR EXPLICIT RK METHODS.

Rosenbrock methods

With the implicit Runge-Kutta methods one has to solve at each step a system of nonlinear algebraic equations. Usually this is done with a modified Newton iteration, where the Jacobian is held fixed during the iteration. Test results in Hairer & Wanner (1991) show that for low accuracy requirements, good results can also be obtained with Rosenbrock methods. The simplest methods of this type can be viewed as linearizations of implicit Runge-Kutta methods. Here we give two examples of Rosenbrock methods, see Dekker & Verwer (1984), Hairer & Wanner (1991) for more general formulas. In the following, let $J(t,v)$ stand for the Jacobian matrix $(\partial F_i(t,v)/\partial v_j)$.

Example A.4. Application of one Newton iteration to the implicit Runge-Kutta method $w_{n+1} = w_n + \tau F(t_{n+\theta}, (1-\theta)w_n + \theta w_{n+1})$ leads to the scheme

$$w_{n+1} = w_n + (I - \theta\tau J_n)^{-1}\tau F(t_{n+\theta}, w_n),$$

where $J_n \approx J(t_{n+\theta}, w_n)$. We can regard this as a method in its own, rather than a special implementation of the Runge-Kutta method. By a Taylor series expansion it is seen that the order of this method is 2 if $\theta = \frac{1}{2}$ and $J_n - J(t_{n+\theta}, w_n) = \mathcal{O}(\tau)$. Otherwise the order is 1. The stability function of this method is given by

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

and hence the method is A-stable for any $\theta \geq \frac{1}{2}$. ◇

Example A.5. Consider the method

$$w_{n+1} = w_n + \Big(b_1 f_1 + b_2 f_2\Big),$$

$$(I - \theta\tau J_n)f_1 = \tau F(t_n, w_n), \qquad (I - \theta\tau J_n)f_2 = \tau F(t_n + c\tau, w_n + cf_1) - (2\theta c)\tau J_n f_1,$$

where $b_1 + b_2 = 1$, $b_2 = 1/(2c)$, and θ, c are free parameters. The stability function of this method is the same as with the diagonally implicit Runge-Kutta methods of Example A.3,

$$R(z) = \frac{1 + (1-2\theta)z + (\frac{1}{2} - 2\theta + \theta^2)z^2}{(1-\theta z)^2},$$

independent of c. An interesting feature of this Rosenbrock method is the fact that its order is 2 regardless of J_n, see Dekker & Verwer (1984). Of course, to obtain good stability the J_n should be related to the exact Jacobian.

For implementation the above method can be written in the slightly more efficient form

$$w_{n+1} = w_n + \Big((b_1 + 2b_2c)g_1 + b_2 g_2\Big),$$

$$(I - \theta\tau J_n)g_1 = \tau F(t_n, w_n), \qquad (I - \theta\tau J_n)g_2 = \tau F(t_n + c\tau, w_n + cg_1) - 2cg_1,$$

with the correspondence $g_1 = f_1$, $g_2 = f_2 - 2cf_1$. In this form one matrix-vector multiplication is saved.

◇

CFL restrictions

Stability restrictions for the advection equation $c_t + c_x = 0$ are called CFL restrictions (after Courant, Friedrichs and Lewy). The relevant eigenvalues in a von Neumann analysis for the standard advection discretizations of order 1,2,3 and 4 are

$$\lambda_{a,1} = \nu\Big(e^{-i\phi} - 1\Big), \tag{A.3a}$$

$$\lambda_{a,2} = \frac{\nu}{2}\Big(e^{-i\phi} - e^{i\phi}\Big), \tag{A.3b}$$

$$\lambda_{a,3} = \frac{\nu}{6}\Big(-e^{-2i\phi} + 6e^{-i\phi} - 3 - 2e^{i\phi}\Big), \tag{A.3c}$$

$$\lambda_{a,4} = \frac{\nu}{12}\Big(-e^{-2i\phi} + 8e^{-i\phi} - 8e^{i\phi} + e^{2i\phi}\Big), \tag{A.3d}$$

with $\phi \in [0, 2\pi]$ and $\nu = \tau/\Delta x$ the Courant number. The CFL restriction on ν is such that these eigenvalues are in the stability region. These restrictions are given in the following table for the standard explicit Runge-Kutta methods up to order 4. The values have been obtained experimentally.

	RK1	RK2	RK3	RK4
$\lambda_{a,1}$	1	1	1.25	1.39
$\lambda_{a,2}$	0	0	1.73	2.82
$\lambda_{a,3}$	0	0.87	1.62	1.74
$\lambda_{a,4}$	0	0	1.26	2.05

TABLE A.2. Stability restrictions on $\nu = \tau/\Delta x$ for advection.

Stability restrictions for diffusion equation $c_t = c_{xx}$ are obtained in a similar way. The relevant eigenvalues for the standard diffusion discretizations of order 2 and 4 are

$$\lambda_{d,2} = \mu\Big(e^{-i\phi} - 2 + e^{i\phi}\Big), \tag{A.4a}$$

$$\lambda_{d,4} = \frac{\mu}{12}\Big(-e^{-2i\phi} + 16e^{-i\phi} - 30 + 16e^{i\phi} - e^{2i\phi}\Big), \tag{A.4b}$$

where now $\mu = \tau/(\Delta x)^2$. The corresponding stability restrictions are given in the next table. Since a restriction on $\tau/(\Delta x)^2$ leads to a very small time step, diffusion equations are usually solved with implicit methods (or very special explicit methods with large interval $[-\beta, 0] \in \mathcal{S}$, see van der Houwen & Sommeijer (1980)).

	RK1	RK2	RK3	RK4
$\lambda_{d,2}$	0.5	0.5	0.62	0.69
$\lambda_{d,4}$	0.37	0.37	0.47	0.52

TABLE A.3. Stability restrictions on $\mu = \tau/(\Delta x)^2$ for diffusion.

Remark. All sorts of combinations are possible, of course. For $c_t + ac_x = dc_{xx}$ with second order central differences we get eigenvalues $2\mu(\cos\phi - 1) + i\nu\sin\phi$ with $\mu = d\tau/(\Delta x)^2$ and $\nu = a\tau/\Delta x$. The forward Euler method can be shown to be stable under the condition

$$\nu^2 \le 2\mu \le 1.$$

To solve an ODE $w'(t) = F(t, w(t))$ it is, on the one hand, quite natural to consider *one-step methods* where for the computation of $w_{n+1} \approx w(t_{n+1})$ only the previous approximation w_n is needed; after all, also the exact value $w(t_{n+1})$ is completely determined by $w(t_n)$. On the other hand, it seems wasteful not to use any past information, since available values $w_n, w_{n-1}, \ldots$ could be used with little cost to obtain already a reasonable approximation w_{n+1}, by extrapolation for example.

In this section we consider the important class of *linear multistep methods*

$$\sum_{j=0}^{k} \alpha_j w_{n+j} = \tau \sum_{j=0}^{k} \beta_j F(t_{n+j}, w_{n+j}) \tag{B.1}$$

for $n = 0, 1, \ldots$, yielding $w_{n+k} \approx w(t_{n+k})$ from already computed $w_{n+k-1}, \ldots, w_n$. We shall refer to (B.1) as a linear k-step method. The method is explicit if $\beta_k = 0$ and implicit otherwise. Formula (B.1) can be scaled, since multiplication of all coefficients α_j, β_j with a same factor will leave the computational scheme unchanged. Usually, scaling is used to set $\alpha_k = 1$ or $\beta_0 + \beta_1 + \cdots + \beta_k = 1$. We shall assume in the following that $\alpha_k > 0$.

A linear k-step method needs k *starting values* $w_0, w_1, \ldots, w_{k-1}$ to perform the first step in (B.1). Only the initial value $w_0 = w(0)$ is given. The other starting values can be computed with a Runge-Kutta method. An other possibility is to use a linear 1-step method to compute w_1, then a linear 2-step method for w_2, and so on, until all necessary starting values for (B.1) have been found.

If we insert the exact solution in (B.1), we obtain

$$\sum_{j=0}^{k} \alpha_j w(t_{n+j}) = \tau \sum_{j=0}^{k} \beta_j w'(t_{n+j}) + \tau r_{n+k}$$

with residual term τr_{n+k}. Usually, r_{n+k} is called the truncation error. The linear multistep method is said to have *order* p if $r_{n+k} = O(\tau^p)$ for all sufficiently smooth w. By a Taylor series expansion around $t = t_n$ it follows that

$$\tau r_{n+k} = C_0 w(t_n) + \tau C_1 w'(t_n) + \tau^2 C_2 w''(t_n) + \cdots$$

with

$$C_0 = \sum_{j=0}^{k} \alpha_j, \qquad C_i = \frac{1}{i!}\Big(\sum_{j=0}^{k} \alpha_j j^i - i \sum_{j=0}^{k} \beta_j i^{i-1}\Big) \quad \text{for} \quad i \geq 1.$$

Thus the method has order p if the order conditions

$$\sum_{j=0}^{k} \alpha_j = 0, \qquad \sum_{j=0}^{k} \alpha_j j^i = i \sum_{j=0}^{k} \beta_j i^{i-1} \quad \text{for} \quad i = 1, 2, \ldots, p \tag{B.2}$$

are satisfied.

We give a few examples of well-known multistep methods. More examples can be found in Hairer et al. (1987) and Lambert (1991).

Example B.1. The 2-step method

$$w_{n+2} - w_n = 2\tau F(t_{n+1}, w_{n+1})$$

is called the *explicit midpoint rule*. Its order is 2 and the method is often used for special classes of problems arising from hyperbolic PDEs. We shall see that it has rather poor stability properties for more general problems. ◇

Example B.2. *Adams methods* are characterized by

$$\alpha_k = 1, \qquad \alpha_{k-1} = -1, \qquad \alpha_j = 0 \quad (0 \le j \le k-2)$$

and with β_j chosen such that the order is optimal.

Explicit Adams methods, also called *Adams-Bashforth* methods, have order k.The method with $k = 1$ is simply the forward Euler method. The 2 and 3-step methods read

$$w_{n+2} - w_{n+1} = \frac{3}{2}\tau F_{n+1} - \frac{1}{2}F_n,$$
$$w_{n+3} - w_{n+2} = \frac{23}{12}\tau F_{n+2} - \frac{16}{12}\tau F_{n+1} + \frac{5}{12}\tau F_n$$

where F_j stands for $F(t_j, w_j)$.

The implicit Adams methods are also known as *Adams-Moulton* methods. The order is $k + 1$. The method with $k = 1$ is the trapezoidal rule, and for $k = 2, 3$ we get

$$w_{n+2} - w_{n+1} = \frac{5}{12}\tau F_{n+2} + \frac{8}{12}\tau F_{n+1} - \frac{1}{12}\tau F_n,$$
$$w_{n+3} - w_{n+2} = \frac{9}{24}\tau F_{n+3} + \frac{19}{24}\tau F_{n+2} - \frac{5}{24}\tau F_{n+1} + \frac{1}{24}\tau F_n.$$

The Adams methods are usually applied in a predictor-corrector fashion, that is, first we compute a predictor $\bar{w}_{n+k}$ from the explicit k-step method and this value is inserted in the right hand side of the implicit k-step method. The method thus obtained is explicit and has order $k + 1$, but it is no longer a genuine linear k-step method. It falls in the wider class of so-called multistep Runge-Kutta methods, with the prediction $\bar{w}_{n+k}$ playing the role of an internal vector. For $k = 1$ this procedure gives a 2-stage Runge-Kutta method, the so-called explicit trapezoidal rule, see Example A.1. ◇

Example B.3. *Backward differentiation formulas*, usually called *BDFs* or *BDF methods*, have

$$\beta_k = 1, \qquad \beta_j = 0 \quad (0 \le j \le k-1)$$

and the α_j are chosen such that the order is optimal, namely order k. The 1-step BDF method is the Backward Euler method. For $k = 2, 3$ the BDF methods read

$$\frac{3}{2}w_{n+2} - 2w_{n+1} + \frac{1}{2}w_n = \tau F_{n+2},$$
$$\frac{11}{6}w_{n+3} - 3w_{n+2} + \frac{3}{2}w_{n+1} + \frac{1}{3}w_n = \tau F_{n+3}.$$

Due to their favourable stability properties the BDF methods are well suited to solve parabolic problems with smooth solutions. The BDF methods were introduced by Curtiss and Hirschfelder in 1952 and their popularity can be attributed to a large extent to Gear (1971). ◇

Stability properties

When studying solutions of linear recursions of the type $\sum_{j=0}^{k} \gamma_k w_{n+j} = 0$ it is convenient to consider the characteristic polynomial $\pi(\zeta) = \sum_{j=0}^{k} \gamma_k \zeta^k$. Let $\zeta_1, \zeta_2, \ldots, \zeta_k$ be the zeros of this polynomial, with multiple zeros repeated. The general solution of the linear recursion can then be written as

$$w_n = c_1 n^{\nu_1} \zeta_1^n + c_2 n^{\nu_2} \zeta_2^n + \cdots + c_k n^{\nu_k} \zeta_k^n$$

with constants c_i determined by the starting values, and with $\nu_i = 0$ if ζ_i is a simple zero and $\nu_i = 0, \nu_{i+1} = 1, \ldots, \nu_{i+l} = l$ if $\zeta_i = \cdots = \zeta_{i+l}$ is a root of multiplicity $l+1$. The characteristic polynomial is said to satisfy the *root condition* if

$$|\zeta_i| \leq 1 \quad \text{for all } i \text{ , and} \quad |\zeta_i| < 1 \quad \text{if } \zeta_i \text{ is not simple.}$$

It is easily seen from the formula for the general solution that this condition is equivalent with boundedness of the sequence $\{w_n\}$ for arbitrary starting values.

Now, consider a linear multistep method (B.1) applied to the test equation

$$w'(t) = \lambda w(t)$$

and let $z = \tau\lambda$. Then we obtain the recursion

$$\sum_{j=0}^{k} (\alpha_j - z\beta_j) w_{n+j} \tag{B.3}$$

with characteristic polynomial $\pi_z(\zeta) = \sum_{j=0}^{k} (\alpha_j - z\beta_j)\zeta^j$. Defining

$$\rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j$$

we have $\pi_z(\zeta) = \rho(\zeta) - z\sigma(\zeta)$.

The *stability region* $\mathcal{S} \subset \mathbb{C}$ of the method is defined as the set consisting off all z such that $\{w_n\}$ is bounded for any choice of starting values $w_0, \ldots, w_{k-1}$. We have

$$z \in \mathcal{S} \quad \Leftrightarrow \quad \pi_z \quad \text{satisfies the root condition}$$

The method is called *zero-stable* if $0 \in \mathcal{S}$. This is equivalent to saying that $\rho(\zeta)$ satisfies the root condition. It is clear that methods which fail to be zero-stable are not suited as numerical methods for solving differential equations since such a method will not even integrate the trivial equation $w'(t) = 0$ properly. Zero-stability reduces the attainable order of linear multistep methods to $p = k$ for explicit methods and $p = 2\lfloor (k+2)/2 \rfloor$ for the implicit ones (the *1-st Dahlquist barrier*, see Dahlquist (1956) or Hairer et al. (1987)). For example, consider the class of explicit methods

$$w_{n+2} - (1+\alpha_0) w_{n+1} + \alpha_0 w_n = \frac{1}{2}\tau(3-\alpha_0) F_{n+1} - \frac{1}{2}\tau(1+\alpha_0) F_n.$$

If $\alpha_0 = 0$ this gives the explicit Adams method with $p = 2$. Taking $\alpha_0 = -5$ we obtain a method of order 3, but this method is not zero-stable. Nice numerical illustrations for the

unstable behaviour of this 3-th order method can be found for instance in Hairer et al. (1987) and Lambert (1991).

For the computation of the stability region of a linear multistep method, observe that on the boundary $\partial\mathcal{S}$ one of the roots of the characteristic polynomial must have modulus 1. Since $\pi_z(\zeta) = 0$ iff $z = \rho(\zeta)/\sigma(\zeta)$, it follows that any point on $\partial\mathcal{S}$ is of the form

$$\rho(e^{i\theta})/\sigma(e^{i\theta}) \quad \text{with} \quad 0 \le \theta \le 2\pi.$$

Example B.4. For the explicit midpoint rule we find that

$$\rho(e^{i\theta})/\sigma(e^{i\theta}) = (e^{2i\theta} - 1)/2e^{i\theta} = \frac{1}{2}(e^{i\theta} - e^{-i\theta}) = i\sin(\theta).$$

By considering the characteristic polynomial $\pi_z(\zeta) = \zeta^2 - 2z\zeta - 1$ and the roots

$$\zeta_{1,2} = z \pm \sqrt{1 + z^2},$$

it easily follows that the stability region is

$$\mathcal{S} = \{z \in \mathbb{C} : \mathrm{Re}z = 0, |z| < 1\},$$

so this is merely a line segment on the imaginary axis. ◇

The form of this stability region is a bit unusual since no r exists such that the disc $\mathcal{D}_r = \{z \in \mathbb{C} : |z + r| \le r\}$ is contained in the stability region. Pictures of the stability regions of several Adams and BDF methods can be found in Gear (1971) and Hairer & Wanner (1991), and for these methods $\mathcal{D}_r \in \mathcal{S}$ for r sufficiently small.

To define stability concepts stronger than zero-stability it is useful to include the point $z = \infty$ in our considerations. We shall say that $\infty \in \mathcal{S}$ if $\sigma(\zeta)$ satisfies the root condition. Observe that the roots of π_z tend to the roots of the polynomial σ for $z \to \infty$ (this is easily seen by dividing $\pi_z(\zeta)$ by z).

A linear multistep method is called *A-stable* if its stability domain contains $\{z \in \bar{\mathbb{C}} : \mathrm{Re}z \le 0 \quad \text{or} \quad z = \infty\}$. In contrast to the Runge-Kutta methods, there are not many linear multistep methods that are A-stable (the order of such methods is at most 2 , the *2-nd Dahlquist barrier*, see Dahlquist (1963) or Hairer & Wanner (1991)). Therefore we look at less demanding properties, allowing high order, which are still useful for semi-discrete PDEs.

A linear multistep method is said to be *$A(\alpha)$-stable* if its stability domain contains the infinite wedge $\{z \in \bar{\mathbb{C}} : z = 0, \infty \quad \text{or} \quad |\arg(-z)| \le \alpha\}$.

Example B.5. The BDF methods are $A(\alpha)$-stable for $k \le 6$ with angle α depending on k :

k	1	2	3	4	5	6	7
α	90°	90°	88°	73°	51°	18°	-

For $k \ge 7$ the methods are no longer zero-stable. Since the angle α for the 6-step method is rather small, the BDF methods are in general only used with $k \le 5$. ◇

The Adams methods of example B.2 all have bounded stability domains and thus these methods are not $A(\alpha)$-stable. The stability domains for the Adams-Bashforth (AB) methods

with $k = 2, 3$ are given in the top pictures of Figure B.1. These are rather small. The Adams-Moulton (AM) methods are implicit but still have a bounded stability regions. For this reason the Adams methods are usually implemented in a predictor-corrector fashion, where the explicit formula is inserted into the right hand side of the implicit formula. The stability regions of these methods with $k = 2, 3$ are given in the bottom pictures of Figure B.1, with fat lines for the predictor-corrector methods (ABM) and thin lines for the implicit ones. Pictures for higher order Adams methods and other multi-step methods can be found in Hairer & Wanner (1991).

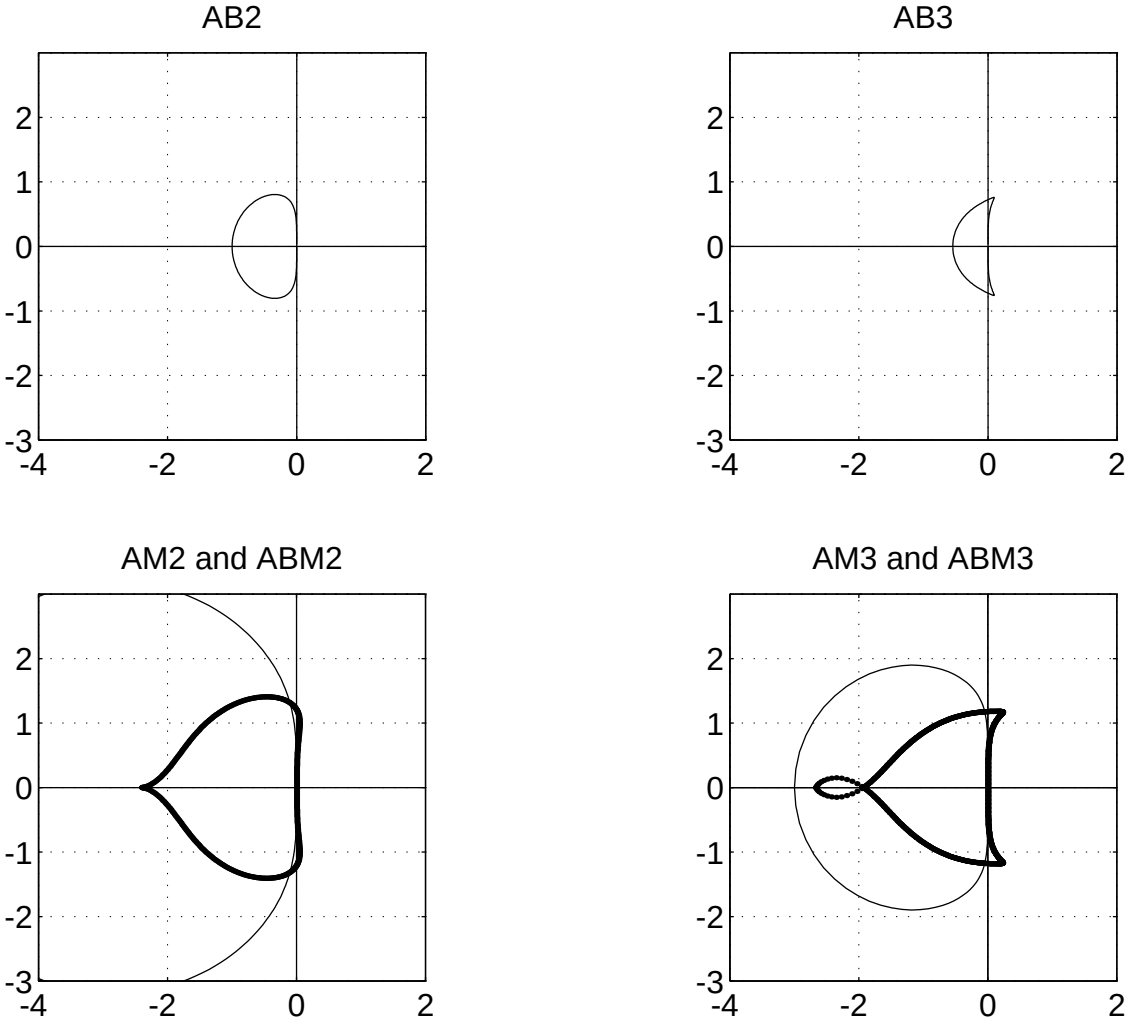


FIGURE B.1. STABILITY REGIONS FOR ADAMS METHODS.

Remark. It is often convenient for the analysis to write recursion (B.3) in a one-step form. First, observe that (B.3) is equivalent to

$$w_{n+k} = -\sum_{j=0}^{k} \frac{\alpha_j - z\beta_j}{\alpha_k - z\beta_k} w_{n+j}.$$

We can formulate this as a one-step recursion in a higher dimensional space by introducing

$$W_n = (w_{n+k-1}, \ldots, w_n)^T.$$

Then (B.3) can be written as

$$W_{n+1} = R(z)W_n \tag{B.4}$$

where

$$R(z) = \begin{pmatrix} r_1(z) & r_2(z) & \cdots & r_k(z) \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}, \qquad r_i(z) = -\frac{\alpha_{k-i} - z\beta_{k-i}}{\alpha_k - z\beta_k}. \tag{B.5}$$

This matrix is called the companion matrix of the muti-step method. From the equivalence of these recursions it is clear that $z \in \mathcal{S}$ iff the matrix $R(z)$ is power bounded.

For linear m-dimensional systems $w'(t) = Lw(t)$ we obtain in the same way $W_{n+1} = R(Z)W_n$ with $Z = \tau L$ and

$$R(Z) = \begin{pmatrix} r_1(Z) & r_2(Z) & \cdots & r_k(Z) \\ I & O & & \\ & \ddots & \ddots & \\ & & I & O \end{pmatrix}.$$

CFL restrictions

Below stability restrictions are given for the advection and diffusion discretizations that were considered in the previous subsection for Runge-Kutta methods. The multi-step methods considered are the 2 and 3-step Adams-Bashforth (AB) schemes and the Adams-Moulton schemes using Adams-Bashforth as predictor (ABM). The layout of the tables is the same as in the Tables A.2 and A.3 with Runge-Kutta methods.

	AB2	ABM2	AB3	ABM3
$\lambda_{a,1}$	0.5	0.98	0.27	0.79
$\lambda_{a,2}$	0	1.20	0.72	1.17
$\lambda_{a,3}$	0.58	1.02	0.39	0.80
$\lambda_{a,4}$	0	0.87	0.52	0.85

TABLE B.2. Stability restrictions on $\nu = \tau/\Delta x$ for advection.

	AB2	ABM2	AB3	ABM3
$\lambda_{d,2}$	0.25	0.6	0.13	0.48
$\lambda_{d,4}$	0.18	0.44	0.10	0.36

TABLE B.3. Stability restrictions on $\mu = \tau/(\Delta x)^2$ for diffusion.

U.M. Ascher, S.J. Ruuth, B. Wetton (1995), *Implicit-explicit methods for time-dependent PDE's*. SIAM J. Numer. Anal. 32 , pp. 797-823.

U.M. Ascher, S.J. Ruuth, R.J. Spiteri (1997), *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*. Appl. Num. Anal. 25, pp. 151-167.

R.M. Beam, R.F. Warming (1976), *An implicit finite-difference algorithm for hyperbolic systems in conservation-law form*. J. Comp. Phys. 22, pp. 87-110.

C. Bolley, M. Crouzeix (1978), *Conservation de la positivité lors de la discrétisation des problèmes d' évolution paraboliques*. RAIRO Anal. Numer. 12, pp. 237-245.

P. Brenner, M. Crouzeix, V. Thomée (1982), *Single step methods for inhomogeneous linear differential equations*. RAIRO Anal. Numer. 16, pp. 5-26.

J.C. Butcher (1987), *The Numerical Analysis of Ordinary Differential Equations*. John Wiley, New York.

W.A. Coppel (1965), *Stability and Asymptotic Behaviour of Differential Equations*. Heath Mathematical Monographs, D.C. Heath & Co., Boston.

C. Canuto, M.Y. Hussaini, A. Quarteroni, T.A. Zang (1988), *Spectral Methods in Fluid Dynamics*. Springer Series in Computational Physics, Springer-Verlag, Berlin.

R. Courant, K.O. Friedrichs, H. Lewy (1928), *Über die partiellen Diffrenzengleichungen der mathematischen Physik*. Math. Anal. 100, pp. 32-74.

M. Crouzeix (1975), *Sur l'approximation des équations différentielles opérationelles par des méthodes de Runge-Kutta*. Thèse, Univ. Paris VI.

M. Crouzeix (1980), *Une méthode multipas implicite-explicite pour l' approximation des équations d' évolution paraboliques*. Numer. Math. 35 , pp. 257-276.

G. Dahlquist (1956), *Convergence and stability in the numerical integration of ordinary differential equations*. Math. Scand. 4, pp. 33-53.

G. Dahlquist (1963), *A special stability problem for linear multistep methods*. BIT 3, pp. 27-43.

G. Dahlquist (1975), *Error analysis for a class of methods for stiff nonlinear initial value problems*. Numerical Analysis, Dundee 1975. Springer Lecture Notes in Mathematics 506, pp. 60-74.

R. Dautray, J.-L. Lions (1993), *Mathematical Analysis and Numerical Methods for Science and Technology 6 - evolution problems II*. Springer Verlag, Berlin.

K. Dekker, J.G. Verwer (1984), *Stability of Runge-Kutta Methods for Stiff nonlinear Differential Equations*. CWI Monograph 2, North-Holland, Amsterdam.

J.L.M. Dorsselear, J.F.B.M. Kraaijevanger, M.N. Spijker (1993), *Linear stability analysis in the numerical solution of initial value problems*. Acta Numerica 1993, pp. 199-237.

J. Douglas, J.E. Gunn (1964), *A general formulation of alternating direction methods.* Numer. Math. 6, pp. 428-453.

J. Frank, W. Hundsdorfer, J.G. Verwer (1997), *On the stability of implicit-explicit linear multistep methods.* Appl. Num. Math. 25 , pp. 193-205.

A. Friedman (1970), *Foundations of Modern Analysis.* Holt, Rinehart & Winston, Inc., New York.

C.W. Gear (1971), *Numerical Initial Value Problems in Ordinary Differential Equations.* Prentice Hall.

D. Goldman, T.J. Kaper (1996), Nth-order operator splitting schemes and nonreversible systems. SIAM J. Numer. Anal. 33, pp. 349-367.

G.H. Golub , C.F. van Loan (1996), *Matrix Computations*, third edition. John Hopkins Univ. Press, Baltimore.

A.R. Gourlay, A.R. Mitchell (1972), *On the structure of alternating direction implicit (A.D.I.) and locally one dimensional (L.O.D.) difference methods.* J. Inst. Maths. Applics. 9, pp. 80-90.

D.F. Griffiths, J.M. Sanz-Serna (1986), *On the scope of the method of modified equations.* SIAM J. Sci. Comput. 7, pp. 994-1008.

B. Gustafsson (1975), *The convergence rate for difference approximations to mixed initial boundary value problems.* Math. Comp. 29, pp. 396-406.

E. Hairer, S.P. Nørsett, G. Wanner (1987), *Solving Ordinary Differential Equations I – nonstiff problems*, Springer Series in Computational Mathematics 8, Springer Verlag, Berlin.

E. Hairer, G. Wanner (1991), *Solving Ordinary Differential Equations II – stiff and differential-algebraic problems.* Springer Series in Computational Mathematics 14, Springer Verlag, Berlin.

C. Hirsch (1988), *Numerical Computation of Internal and External Flows 1: fundamentals and numerical discretization.* John Wiley & Sons, Chichester.

R.A. Horn, C.R. Johnson (1985), *Matrix Analysis.* (1991), *Topics in Matrix Analysis.* Cambridge University Press.

Z. Horváth (1998), *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods.* To appear in Appl. Num. Math.

P.J. van der Houwen & B.P. Sommeijer (1980), *On the internal stability of explicit, m-stage Runge-Kutta methods for large m-values.* Z. Angew. Math. Mech. 60, pp. 479-485.

W. Hundsdorfer (1992), *Unconditional convergence of some Crank-Nicolson LOD methods for initial-boundary value problems.* Math. Comp. 53, pp. 81-101.

W. Hundsdorfer (1998), *A note on stability of the Douglas splitting method.* Math. Comp. 67, pp. 183-190.

W. Hundsdorfer (1999), *Stability of approximate factorizations with θ-methods.* BIT 39, pp. 473-483.

W. Hundsdorfer (2000), *Accuracy and stability of splitting with stabilizing corrections.* Report MAS-R9935, CWI, Amsterdam.

W. Hundsdorfer, B. Koren, M. van Loon, J.G. Verwer (1995), *A positive finite-difference advection scheme applied on locally refined grids.* J. Comp. Phys. 117, pp. 35-46.

W. Hundsdorfer, J.G. Verwer, *Stability and convergence of the Peaceman-Rachford ADI method for initial-boundary value problems.* Math. Comp. 53, pp. 81-101.

A. Iserles, G. Strang (1983), *The optimal accuracy of difference schemes.* Trans. Amer. Math. Soc. 277, pp.779-803.

A. Iserles, S.P. Nørsett (1991), *Order Stars*, Applied Mathematics and Mathematical Computation 2, Chapman & Hall, London.

B. Koren (1993), *A robust upwind discretization for advection, diffusion and source terms.* In : Numerical Methods for Advection-Diffusion Problems (C.B. Vreugdenhil and B. Koren, eds.), Notes on Numerical Fluid Mechanics 45, Vieweg, Braunschweig.

J.F.B.M. Kraaijevanger (1991), *Contractivity of Runge-Kutta methods.* BIT 31, pp. 482-528.

J.F.B.M. Kraaijevanger, H.W.J Lenferink, M.N. Spijker (1987), *Stepsize restrictions for stability in the numerical solution of ordinary differential equations.* J. Comp. Appl. Math. 20, pp.67-81.

J.D. Lambert (1991), *Numerical Methods for Ordinary Differential Equations, the initial value problem.* John Wiley & Sons, Chichester.

P.D. Lax, B. Wendroff (1960), *Systems of conservation laws.* Comm. Pure Appl. Math. 13, pp. 217-237.

B. van Leer (1974), *Towards the ultimate conservative difference scheme III. Monotonicity and conservation combined in a second order scheme.* J. Comput. Phys. 14, pp.361-370.

R.J. LeVeque (1982), *Time-split methods for partial differential equations.* PhD Thesis, Dept. Comp. Sc., Stanford Univ.

R.J. LeVeque (1992), *Numerical Methods for Conservation Laws.* Lecture Notes in Mathematics, ETH Zürich, Birkhäuser Verlag, Basel.

M. van Loon (1996), *Numerical methods in smog prediction.* Thesis, University of Amsterdam.

Ch. Lubich, A. Ostermann (1993), *Runge-Kutta methods for parabolic equations and convolution quadrature.* Math. Comp. 60, pp. 105-131.

Ch. Lubich, A. Ostermann (1995), *Interior estimates for time discretization of parabolic equations.* Appl. Num. Math. 18, pp. 241-251.

T.A. Manteuffel, A.B. White (1986), *The numerical solution of second-order boundary-value problems on nonuniform meshes.* Math. Comp. 47, pp. 511-535.

G.I. Marchuk (1990), *Splitting and alternating direction methods.* In: Handbook of Numerical Analysis I (P.G. Ciarlet and J.L. Lions, eds.), North-Holland, Amsterdam, pp. 197-462.

G.J. McRea, W.R. Goodin, J.H. Seinfeld (1982), *Numerical solution of atmospheric diffusion for chemically reacting flows.* J. Comput. Phys. 77 , pp. 1-42.

A.R. Mitchell, D.F. Griffiths (1980), *The Finite Difference Method in Partial Differential Equations*. John Wiley & Sons, Chichester.

K.W. Morton (1980), *Stability of finite difference approximations to a diffusion-Convection equation*. Int. J. Num. Meth. Eng. 15, pp. 677-683.

K.W. Morton (1996), *Numerical Solution of Convection-diffusion Problems*. Applied Mathematics and Mathematical Computation 12, Chapman & Hall.

J.M. Ortega, W.C. Rheinboldt (1970), *Iterative Solution of Nonlinear Equations in Several Variables*. Computer Science and Applied Mathematics, Academic Press, New York.

D. Pathria (1997), *The correct formulation of intermediate boundary conditions for Runge-Kutta time integration of initial boundary value problems*. SIAM J. Sci. Comput. 18, pp. 1255-1266.

A. Pinkus, S. Zafrany (1997), *Fourier Series and Intergral Transforms*. Cambrige University Press, Cambridge.

R.D. Richtmyer, K.W. Morton (1967), *Difference Methods for Initial-Value Problems*, 2-nd ed., Interscience Publishers, John Wiley & Sons, New York.

W. Rudin (1964), *Principles of Mathematical Analysis*, 2-nd ed., McGraw-Hill, New York.

J.M. Sanz-Serna, M.P. Calvo (1994), *Numerical Hamiltonian Problems*. Applied Mathematics and mathematical Computation 7, Chapman & Hall, London.

J.M. Sanz-Serna, J.G. Verwer, W. Hundsdorfer (1987), *Convergence and order reduction of Runge-Kutta schemes applied to evolutionary problems in partial differential equations*. Numer. Math. 50, pp. 405-418.

Q. Sheng (1989), *Solving partial differential equations by exponential splittings*. IMA J. Numer. Math. 9, pp. 199-212.

C.-W. Shu, S. Osher (1988), *Efficient implementation of essentially non-oscillatory shock-capturing schemes*. J. Comp. Phys. 77, pp. 439-471.

M.N. Spijker (1983), *Contractivity in the numerical solution of initial value problems*. Numer. Math 42, pp. 271-290.

M.N. Spijker (1996), *Numerical Stability Theory – resolvent conditions and stability estimates in the numerical solution of initial value problems*. Lecture Notes, Univ. of Leiden.

G. Strang (1962), *Trigonometric polynomials and difference methods of maximal accuracy*. J. Math. and Phys. 41, pp. 147-154.

G. Strang (1963), *Accurate partial difference methods I: linear Cauchy problems*. Arch. Rat. Mech. Anal. 12, pp. 392-402.

G. Strang (1968), *On the construction and comparison of difference schemes*. SIAM J. Numer. Anal. 5, pp. 506-517.

G. Strang, G.J. Fix (1973), *An Analysis of the Finite Element Method*. Prentice-Hall, New York.

C.J. Strikwerda (1989), *Finite Difference Schemes and Partial Differential Equations*, Chapman & Hall, New-York.

M. Suzuki (1990), *Fractal decomposition of exponential operators with applications to many-body theories and Monte Carlo simulations.* Phys. Lett. A 146, pp. 319-323.

P.K. Sweby (1984), *High resolution schemes using flux-limiters for hyperbolic conservation laws.* SIAM J. Numer. Anal. 21 , pp. 995-1011.

V. Thomée (1990), *Finite difference methods for linear parabolic equations.* In: Handbook of Numerical Analysis I (P.G. Ciarlet and J.L. Lions, eds.), North-Holland, Amsterdam, pp. 5-196.

R.A. Trompert, J.G. Verwer (1991), *A static regridding method for two-dimensional parabolic partial differential equations.* Appl. Numer. Math. 8 , pp. 65-90.

J.M. Varah (1980), *Stability restrictions on second order, three-level finite-difference schemes for parabolic equations.* SIAM J. Numer. Anal. 17 , pp. 300-309.

J.G. Verwer (1986), *Convergence and order reduction of diagonally implicit Runge-Kutta schemes in the method of lines.* In: Numerical Analysis (D.F. Griffiths and G.A. Watson, eds.), Pitman Research Notes in Mathematics 140, pp. 220-237.

J.G. Verwer, J.G. Blom, W. Hundsdorfer (1996), *An implicit-explicit approach for atmospheric transport-chemistry problems.* Appl. Num. Math. 20 , pp. 191-209.

J.G. Verwer, J.M. Sanz-Serna (1984), *Convergence of method of lines approximations to partial differential equations.* Computing 33, pp. 297-313.

J.G. Verwer, E. Spee, J.G. Blom, W. Hundsdorfer (1999), *A second order Rosenbrock method applied to photochemical dispersion problems*, SIAM J. Sci. Comput. 20, pp. 1456-1480.

G. Wanner, E. Hairer, S.P. Nørsett (1978), *Order stars and stability theorems.* BIT 18, pp. 475-489.

R.F. Warming, R.M. Beam (1979), *An extension of A-stability to alternating direction methods.* BIT 19, pp. 395-417.

R.F. Warming, B.J. Hyett (1974), *The modified equation approach to the stability and accuracy analysis of finite difference methods.* J. Comp. Phys. 14, pp.159-179.

N.N. Yanenko, *The Method of Fractional Steps.* Springer Verlag, Berlin, 1971.

H. Yoshida (1990), *Construction of higher order symplectic integrators.* Phys. Lett. A 150, pp. 262-268.

S.T. Zalesak (1987), *A preliminary comparison of modern shock-capturing schemes: linear advection.* In: Advances in Computer Methods for Partial Differential Equations, R. Vichnevetsky and R.S. Stapelman (eds.), IMACS Proceedings 1987, pp. 15-22.

Z. Zlatev (1995), *Computer Treatment of Large Air Pollution Models.* Kluwer, Dordrecht.