Applications of Mathematics

Miloslav Vlasák

Time discretizations for evolution problems

Applications of Mathematics, Vol. 62 (2017), No. 2, 135–169

Persistent URL: http://dml.cz/dmlcz/146700

62 (2017) APPLICATIONS OF MATHEMATICS No. 2, 135–169

TIME DISCRETIZATIONS FOR EVOLUTION PROBLEMS

Miloslav Vlasák, Praha

Received September 26, 2016. First published March 6, 2017.

Abstract. The aim of this work is to give an introductory survey on time discretizations for liner parabolic problems. The theory of stability for stiff ordinary differential equations is explained on this problem and applied to Runge-Kutta and multi-step discretizations. Moreover, a natural connection between Galerkin time discretizations and Runge-Kutta methods together with order reduction phenomenon is discussed.

Keywords: time discretizations; parabolic PDEs; stiff ODEs; Runge-Kutta methods; multi-step methods

MSC 2010: 65J10, 65L04, 65L20

1. Introduction

The aim of this work is to give an introductory survey on time discretizations for linear parabolic problems. We shall concentrate on a linear abstract parabolic problem discretized in space by an abstract finite element method (FEM). The resulting system of ordinary differential equations is usually stiff and therefore needs to be solved by sufficiently stable time integrator.

The most popular methods to deal with stiff problems are the famous BDF methods, since they are robust and cheap. In recent years, the development of computer power and algorithms of linear algebra enables the use of more expensive but also much more robust implicit Runge-Kutta methods. We shall focus on the basic principles of analysis of stiff problems and apply them to general Runge-Kutta methods and linear multi-step methods.

Some of the suitable Runge-Kutta methods can be interpreted as Galerkin discretizations in time similar to FEM. This brings a very nice unified point of view to

The research has been supported by the Grant No. P201/13/00522S of the Czech Science Foundation. The author is a junior researcher of the University Centre for Mathematical Modelling, Applied Analysis and Computational Mathematics (Math MAC).

DOI: 10.21136/AM.2017.0268-16

the whole discretization that can be then exploited in several ways, e.g. unified adaptivity, unified a posteriori error analysis, superconvergence analysis etc. Although the theory of implicit Runge-Kutta methods and their connections to collocation methods and Galerkin methods has been described in literature for some time, this knowledge is not widely spread among people in common practice.

Unfortunately, this paper cannot cover the entire topic and many aspects are not discussed in it. These aspects include adaptivity with respect to the step-size, adaptivity with respect to the order of the method, adaptive choice of the method itself, see e.g. [8], a posteriori estimates, the extension of the analysis to nonlinear problems, see e.g. [15] or [27], effective implementation of the methods discussed (especially implicit Runge-Kutta methods), see e.g. [33], exponential integrators, see e.g. [24] and many others.

1.1. Abstract parabolic problem. Let H be a Hilbert space with a scalar product $(\cdot,\cdot)$, the corresponding norm $\|\cdot\|$, and let X be a Hilbert space continuously embedded into a dense subspace of H. The duality on X we will express by a scalar product on H. We shall focus on an *abstract parabolic problem*: find $u \in L^2(0,T;X)$ with derivative $u' \in L^2(0,T;X^*)$ such that

$$\begin{aligned} u'(t) + \mathcal{A}u(t) &= f(t), \quad t \in (0,T), \\ u(0) &= u^0, \end{aligned} \tag{1.1}$$

where $\mathcal{A}\colon X \to X^*$ is bounded, linear, not necessarily self-adjoint operator, $f \in C([0,T];X^*)$, and $u^0 \in H$. Moreover, we assume ellipticity of $\mathcal{A}$: there exists a constant $c > 0$ such that

$$(\mathcal{A}u, u) \geqslant c\|u\|_X^2 \quad \forall\, u \in X. \tag{1.2}$$

The existence and uniqueness of the exact solution of the abstract parabolic problem follows from [30].

E x a m p l e 1.1. A typical example is the heat equation: Ω is a computational domain, $H = L^2(\Omega)$, $X = H_0^1(\Omega)$, and $\mathcal{A}$ is the weak Laplace operator, i.e. $(\mathcal{A}u, v) = (\nabla u, \nabla v)$.

1.2. Space discretization. In this section we will introduce the general conforming FEM discretization of abstract parabolic problem (1.1). Other discretizations (finite volume method, discontinuous Galerkin method, finite difference method, ...)

behave usually very similarly. Let X_h be a finite-dimensional subspace of X. Then we define the semi-discrete problem: find $u_h \in C^1(0,T,X_h)$ such that

$$(u_h'(t) + \mathcal{A}u_h(t), v_h) = (f(t), v_h) \quad \forall\, v_h \in X_h,\ t \in (0,T), \tag{1.3}$$
$$(u_h(0), v_h) = (u^0, v_h) \quad \forall\, v_h \in X_h.$$

The semi-discrete problem represents a system of ordinary differential equations (ODEs). The existence and uniqueness of the semi-discrete solution follows from the Picard-Lindelöf theorem, see e.g. [6]. The initial condition $u_h(0)$ is a simple H-orthogonal projection of the original initial condition u^0.

Using the natural isomorphism $X_h \cong X_h^*$ via the H-scalar product, we can define an operator $\mathcal{A}_h$ on X_h such that

$$u_h'(t) + \mathcal{A}_h u_h(t) = f_h(t) \quad \forall\, t \in (0,T), \tag{1.4}$$

where $f_h(t) \in X_h$ is the result of restriction $X^* \to X_h^* \cong X_h$.

Problem (1.4) is usually considered as *stiff*. The concept of stiffness is rather vague in literature. In our case the stiffness of the problem comes from the wide range of eigenvalues of $\mathcal{A}_h$.

We will demonstrate it by the following example (the heat equation on a rod):

E x a m p l e 1.2. Let $\Omega = (0,1)$, let $\mathcal{A}$ be the weak Laplace operator and $H = L^2(\Omega)$, $X = H_0^1(\Omega)$. Moreover, let us assume the FEM space $X_h \subset X$ to be the space of piece-wise linear functions on equidistant mesh with mesh-size h, $N = 1/h$. Then it is possible to determine the eigenvalues λ_i of $\mathcal{A}_h$ in this specific situation:

$$\lambda_i = \frac{12}{h^2} \frac{1 - \cos(\pi h i)}{4 + 2\cos(\pi h i)}, \quad i = 1, \dots, N-1, \tag{1.5}$$

see e.g. [3]. For small mesh-size h we get $\lambda_1 = \lambda_{\min} \approx \pi^2$ and $\lambda_{N-1} = \lambda_{\max} \approx 12/h^2$.

To simplify our next considerations about space discretization error we define a space $V \subset X$ of sufficiently regular functions and the Ritz projection $R_h \colon X \to X_h$ such that

$$(\mathcal{A}R_h u, v_h) = (\mathcal{A}u, v_h) \quad \forall\, v_h \in X_h. \tag{1.6}$$

The existence of the Ritz projection comes from the properties of the operator $\mathcal{A}$. Due to Cea's lemma, the approximation properties of R_h, i.e. $\|R_h u - u\|_X$, imitate the approximation properties of X_h in X, i.e. $\|R_h u - u\|_X \leqslant C \inf\limits_{v \in X_h} \|v - u\|_X$. Moreover, it is possible to see that $\|R_h u - u\| \leqslant C \|R_h u - u\|_X$, where the constant C comes

from the continuous embedding of X to H. We will denote the error caused by the general FEM space discretization by $\mathrm{err}(h)$. Note that $\mathrm{err}(h) \sim \|R_h u - u\|_X$ and $\mathrm{err}(h)$ depends on $\|u\|_V$ for $u \in V$.

To simplify the ideas in the analysis of the abstract parabolic problem it is advantageous to study the so-called *Dahlquist's equation*: find $y\colon (0,T) \to \mathbb{C}$ such that

$$\begin{aligned} y'(t) + \lambda y(t) = 0, \quad t \in (0,T), \\ y(0) = y^0. \end{aligned} \tag{1.7}$$

Dahlquist's equation naturally simplifies abstract parabolic problem (1.1) replacing the operator $\mathcal{A}_h$ by its eigenvalue $\lambda \in \mathbb{C}$.

We also simplify the forthcoming relations by not emphasizing the dependence on t, if it is not necessary.

2. One-step methods

In this section we shall start with a description of intuitive time discretizations via Euler methods and θ-scheme. On these methods we demonstrate the importance of stability for stiff problems. Then we extend these ideas to Runge-Kutta methods and derive some suitable stable higher order Runge-Kutta discretizations.

In order to discretize problems (1.4) and (1.7) we consider a time partition $0 = t_0 < t_1 < \ldots < t_r = T$ with time subintervals $I_m = (t_{m-1}, t_m)$, time steps $\tau_m = |I_m| = t_m - t_{m-1}$ and global step-size $\tau = \max_m \tau_m$. Moreover, we will use the usual notation for function values at the nodes, e.g. $u^m = u(t_m)$ or $y^m = y(t_m)$.

For all of the one-step methods mentioned below the initial condition used will be $U^0 = u_h(0)$.

2.1. Euler methods, θ-scheme. The simplest example of methods for ODEs are the well-known Euler methods. The derivation of these methods is based on the replacement of the time derivative in the original equation by the forward or backward difference

$$u_h'(t_{m-1}) \approx \frac{u_h^m - u_h^{m-1}}{\tau_m}, \qquad u_h'(t_m) \approx \frac{u_h^m - u_h^{m-1}}{\tau_m}. \tag{2.1}$$

Then the *forward* or *backward Euler method* discretization of (1.4) is a sequence $\{U^m\}_{m=0}^r \subset X_h$ satisfying respectively

$$U^m - U^{m-1} + \tau_m \mathcal{A}_h U^{m-1} = \tau_m f_h^{m-1}, \quad 1 \leqslant m \leqslant r, \tag{2.2}$$

or

$$U^m - U^{m-1} + \tau_m \mathcal{A}_h U^m = \tau_m f_h^m, \quad 1 \leqslant m \leqslant r. \tag{2.3}$$

Both these methods are step-marching schemes determining the new value U^m from U^{m-1}, where the starting value U^0 is determined from the initial condition. For computing U^m by the forward Euler method it is sufficient to evaluate the relation from the already known U^{m-1}, i.e. we get an *explicit* relation for U^m. On the contrary, to evaluate U^m by the backward Euler method we need to compute the solution of the system containing $\mathcal{A}_h$, i.e. the backward Euler method provides an *implicit* relation for U^m. For this reason the methods are also called the *explicit* or *implicit Euler* method, respectively.

The *θ-scheme* mimics Euler methods by replacing a convex combination of time derivatives by classical difference.

$$(1-\theta)u'(t_{m-1}) + \theta u'(t_m) \approx \frac{u^m - u^{m-1}}{\tau_m}, \quad \theta \in [0,1]. \tag{2.4}$$

Then the θ-scheme discretization of (1.4) is a sequence $\{U^m\}_{m=0}^r \subset X_h$ satisfying

$$\begin{aligned} U^m - U^{m-1} + \tau_m((1-\theta)\mathcal{A}_h U^{m-1} + \theta \mathcal{A}_h U^m) \\ = \tau_m((1-\theta) f_h^{m-1} + \theta f_h^m), \quad 1 \leqslant m \leqslant r. \end{aligned} \tag{2.5}$$

It is possible to see that the variant with $\theta = 0$ is the forward Euler method while $\theta = 1$ gives the backward Euler method. Other important possibility is $\theta = 1/2$, which is called the Crank-Nicolson method.

2.2. Numerical example. Here we will briefly present the numerical behaviour of Euler methods. Let us consider the following problem partially described in Example 1.1 and Example 1.2:

$$\begin{gathered} u' - \Delta u = f, \quad t \in (0,1),\ x \in (0,1), \\ u(t,0) = u(t,1) = 0, \quad t \in (0,1), \\ u(0,x) = 0, \quad x \in (0,1), \end{gathered} \tag{2.6}$$

where $u' = \partial u / \partial t$ and the function f is such that the exact solution satisfies

$$u(t,x) = 4\frac{\mathrm{e}^{10t} - 1}{\mathrm{e}^{10} - 1} x(1-x). \tag{2.7}$$

Standard FEM piece-wise linear approximations on equidistant mesh with a mesh-size h ($N = 1/h$) is used for space discretization. The resulting semi-discrete system is solved by forward and backward Euler method with equidistant step-size τ.

N	$1/\tau$	forward Euler	backward Euler
100	5	3.09E+7	2.92E−1
100	20	1.68E+60	8.76E−2
100	40	4.02E+123	4.49E−2
100	80	—	2.27E−2
100	160	—	1.14E−2

Table 1. L^2-norm error at time level $t = 1$.

Table 1 shows *overkill* in discretization with respect to space, i.e., the error produced by the space discretization is negligible. Then the backward Euler method behaves perfectly as a first order accurate method and for larger time steps it still provides a reasonable error. On the other hand the forward Euler method in this situation completely fails to provide reasonable results and with decreasing step-size τ the situation is even worse.

N	$1/\tau$	forward Euler	backward Euler
20	2000	—	1.02E−3
20	2300	7.22E+35	1.01E−3
20	2400	1.82E−3	1.01E−3
40	9600	4.55E−4	2.52E−4
80	38400	1.14E−4	6.31E−5

Table 2. L^2-norm error at time level $t = 1$.

Table 2 shows the behaviour of the error for a very small step-size τ (in comparison with the mesh-size h). The backward Euler method in this situation produces reasonable results, but the behaviour of the error is much more influenced by the discretization in space. The results of the forward Euler method are more interesting. For the mesh-size $h = 1/20$ and step-sizes $\tau = 1/2000$ and $\tau = 1/2300$ the results are obviously very bad. But at mesh-size $\tau = 1/2400$ the results start to be comparable with the results obtained by the backward Euler method.

2.3. Backward Euler analysis. The analysis of the backward Euler method is straightforward. We decompose the error $e^m = U^m - u^m = \xi^m + \eta^m$ with $\xi^m = U^m - R_h u^m$ and $\eta^m = R_h u^m - u^m$, where η^m is the part of the error depending only on the quality of the space discretization. Our aim will be estimating ξ terms by η

terms. Subtracting (1.1) at time $t = t_m$ multiplied by τ_m from (2.3) and decomposing the error into ξ terms and η terms, we get the error equation

$$(\xi^m - \xi^{m-1}, v_h) + \tau_m(\mathcal{A}\xi^m, v_h) = (\tau_m u'(t_m) - u^m + u^{m-1}, v_h) - \tau_m(\mathcal{A}\eta^m, v_h) - (\eta^m - \eta^{m-1}, v_h) \quad \forall\, v_h \in X_h. \tag{2.8}$$

It is worth noticing that $(\mathcal{A}\eta^m, v_h) = 0$ for $v_h \in X_h$, see the definition of R_h. Moreover, under an assumption on the regularity of the exact solution, e.g. $u \in W^{2,\infty}(0,T,H) \cap W^{1,\infty}(0,T,V)$, it can be shown that

$$(\tau_m u'(t_m) - u^m + u^{m-1}, v_h) \leqslant C\tau_m^2 \|v_h\|, \qquad (\eta^m - \eta^{m-1}, v_h) \leqslant \tau_m \mathrm{err}(h)\|v_h\|. \tag{2.9}$$

Setting $v_h = \xi^m$, we get from (2.8) the left-hand side terms $(\xi^m - \xi^{m-1}, \xi^m) \geqslant \|\xi^m\|^2 - \|\xi^{m-1}\|\,\|\xi^m\|$ and $\tau_m(\mathcal{A}_h\xi^m, \xi^m) \geqslant 0$. Using these estimates together with (2.9) in the error equation (2.8), we get

$$(\|\xi^m\| - \|\xi^{m-1}\|)\|\xi^m\| \leqslant \tau_m(C\tau_m + \mathrm{err}(h))\|\xi^m\|. \tag{2.10}$$

Dividing by $\|\xi^m\|$ and summing over m, we obtain

$$\|\xi^m\| \leqslant \|\xi^0\| + t_m(C\tau + \mathrm{err}(h)), \tag{2.11}$$

i.e., ξ^m is estimated by the error in the initial condition and by the accumulated local errors. It is important to notice that the rate of accumulation of local errors is linear (additive) in this case.

2.4. Forward Euler analysis. The analysis of the forward Euler method is more complicated. For simplicity, let us assume the forward Euler discretizaton of Dahlquist's equation (1.7) with equidistant time step, i.e.

$$Y^m - Y^{m-1} + \tau\lambda Y^{m-1} = 0. \tag{2.12}$$

Then the equation for the error $e^m = Y^m - y^m$ comes from subtracting (1.7) at $t = t_{m-1}$ multiplied by τ from (2.12):

$$e^m - e^{m-1} + \tau\lambda e^{m-1} = \tau y'(t_{m-1}) - y^m + y^{m-1} = \mathrm{loc}^m. \tag{2.13}$$

The term $\mathrm{loc}^m = \tau y'(t_{m-1}) - y^m + y^{m-1}$ represents the local error and it is possible to show that it behaves as $O(\tau^2)$. Expressing the error e^m we get

$$e^m = (1-\tau\lambda)e^{m-1} + \mathrm{loc}^m = (1-\tau\lambda)^2 e^{m-2} + (1-\tau\lambda)\mathrm{loc}^{m-1} + \mathrm{loc}^m = \ldots = (1-\tau\lambda)^m e^0 + \sum_{s=1}^{m}(1-\tau\lambda)^{m-s}\mathrm{loc}^s. \tag{2.14}$$

This equation shows us how the initial error and local errors propagate to the global error e^m. We can see that this propagation has two completely different regimes. The former (stable) one is for $|1-\tau\lambda| \leqslant 1$. Then the powers of $1-\tau\lambda$ can be estimated from above by 1 and the accumulation of local errors is additive as in the backward Euler case. The latter regime is for $|1-\tau\lambda| > 1$ and is unstable. Then the error increases exponentially.

Coming back to Numerical example (Subsection 2.2) we can simulate the problem by Dahlquist's equation with $\lambda = \lambda_i$, where λ_i are the eigenvalues of $\mathcal{A}_h$. Then the condition $|1-\tau\lambda_i| \leqslant 1$ is equivalent to $\tau \leqslant 2/\lambda_i$. The worst case scenario is for $\lambda_i = \lambda_{\max} \approx 12/h^2$, see Example 1.2. Then the stability condition is equivalent to $\tau \leqslant h^2/6$ which perfectly fits the results obtained in the experiment. The stability condition $\tau \leqslant h^2/6$ is usually considered very restrictive, since for small mesh-size h which is needed for sufficient resolution of the discretization with respect to space the needed stable time step length must be extremely small. Therefore it is usually recommended for stiff problems to apply methods without step-size restrictions like the backward Euler method.

For completeness we present the idea of analysis of the forward Euler method applied to the original problem (1.4). We decompose the error into $\xi = U - R_h u$ and $\eta = R_h u - u$ similarly to the backward Euler case. Then after subtracting (1.1) at time $t = t_{m-1}$ multiplied by τ from (2.2) and decomposing the error into ξ terms and η terms we get the error equation

$$
\begin{aligned}
(\xi^m - \xi^{m-1}, v_h) + \tau(\mathcal{A}\xi^{m-1}, v_h) &= (\tau u'(t_{m-1}) - u^m + u^{m-1}, v_h) \\
&\quad - \tau(\mathcal{A}\eta^{m-1}, v_h) - (\eta^m - \eta^{m-1}, v_h) \quad \forall\, v_h \in X_h.
\end{aligned} \tag{2.15}
$$

This relation can be interpreted as

$$
\begin{aligned}
\xi^m &= (I - \tau\mathcal{A}_h)\xi^{m-1} + \mathrm{RHS}^m = \dots \\
&= (I - \tau\mathcal{A}_h)^m \xi^0 + \sum_{s=1}^{m} (1 - \tau\mathcal{A}_h)^{m-s} \mathrm{RHS}^s,
\end{aligned} \tag{2.16}
$$

where $\mathrm{RHS}^m \in X_h$ is the notation of the right-hand side of (2.15). Under assumptions on the regularity of the exact solution similar to the backward Euler case it is possible to show $\|\mathrm{RHS}^m\| \leqslant \tau(C\tau + \mathrm{err}(h))$. Once again, if $\|I - \tau\mathcal{A}_h\| \leqslant 1$ then the global error is accumulated from local errors contained in RHS^m additively. If $\|I - \tau\mathcal{A}_h\| > 1$, then unstable behaviour of the error appears and the global error increases exponentially.

2.5. General one-step method analysis for Dahlquist equation. Let us consider Dahlquist's equation (1.7) and its general one-step discretization, i.e. a pro-

cess $Y^{m-1} \xrightarrow{\tau,\lambda} Y^m$. For simplicity let us assume equidistant step-size. The exact solution of (1.7) behaves as

$$y^m = \mathrm{e}^{-\tau\lambda} y^{m-1}, \tag{2.17}$$

i.e. the exact solution develops from time level t_{m-1} to t_m as the multiple by factor $\mathrm{e}^{-\tau\lambda}$. Let us assume that a general one-step method behaves similarly and the discrete solution develops from time level t_{m-1} to t_m as the multiple by factor $R(\tau\lambda)$, i.e.

$$Y^m = R(\tau\lambda) Y^{m-1}. \tag{2.18}$$

The function $R\colon \mathbb{C} \to \mathbb{C}$ is called the *stability function*. It can be seen that the method is fully described by its stability function R.

We define the local error as the difference between the exact solution and the result of application of one step of the method starting from the exact solution, i.e.

$$\mathrm{loc}^m = R(\tau\lambda) y^{m-1} - y^m = (R(\tau\lambda) - \mathrm{e}^{-\tau\lambda}) y^{m-1}. \tag{2.19}$$

The (local) order of convergence simply results from the ability of the stability function $R(z)$ to approximate e^{-z}, e.g. the stability function of the forward Euler method is $R(z) = 1 - z = \mathrm{e}^{-z} + O(z^2)$ while that of the backward Euler method is $R(z) = 1/(1+z) = \mathrm{e}^{-z} + O(z^2)$.

Definition 2.1. Let us consider a general one-step method described by its stability function R. The set

$$S = \{z \in \mathbb{C}\colon\ |R(z)| \leqslant 1\} \tag{2.20}$$

is called the stability region. We say that the method is *stable* if $\tau\lambda \in S$. We call the method *unconditionally stable* if the method is stable for arbitrary step-size $\tau > 0$. Otherwise, we call the method conditionally stable. We call the method *A-stable*, if $\{z \in \mathbb{C}\colon\ \mathrm{Re}\, z \geqslant 0\} \subset S$.

A-stability is a key ingredient in the analysis of the abstract parabolic problem (1.1) satisfying (1.2), since it generalizes unconditional stability of Dahlquist's equation with $\mathrm{Re}\,\lambda \geqslant 0$.

The analysis of the global error follows the idea of decomposition to the local error and to the error at previous time level multiplied by the stability function. From (2.19) we get

$$\begin{aligned} e^m = Y^m - y^m &= Y^m - R(\tau\lambda) y^{m-1} + R(\tau\lambda) y^{m-1} - y^m \\ &= R(\tau\lambda) e^{m-1} + \mathrm{loc}^m = \ldots = R(\tau\lambda)^m e^0 + \sum_{s=1}^{m} R(\tau\lambda)^{m-s} \mathrm{loc}^s. \end{aligned} \tag{2.21}$$

Relation (2.21) demonstrates the contribution of the local errors and the initial error to the global error. Assuming that the method (for given λ and with chosen step-size τ) is stable, i.e. $|R(\tau\lambda)| \leqslant 1$, we get the desired global error estimate

$$|e^m| \leqslant |e^0| + \sum_{s=1}^{m} |\mathrm{loc}^s| \quad \forall\, m \geqslant 1. \tag{2.22}$$

Assuming $\mathrm{loc}^s = O(\tau^{p+1})$ with $p \geqslant 1$ or $\sum\limits_{s=1}^{m} |\mathrm{loc}^s| \to 0$ for $\tau \to 0$ gives convergence.

The situation for non-equidistant step-size is more complicated. Following the same idea as in (2.21), we get

$$e^m = Y^m - y^m = \Big(\prod_{s=1}^{m} R(\tau_s\lambda)\Big)e^0 + \sum_{s=1}^{m}\Big(\prod_{j=s+1}^{m} R(\tau_j\lambda)\Big)\mathrm{loc}^s. \tag{2.23}$$

The stability in this situation comes from boundedness of the product terms $\prod\limits_{j=s+1}^{m} R(\tau_j\lambda)$ that is uniform with respect to $m \geqslant 1$ and $0 \leqslant s \leqslant m$. Assuming

$$\prod_{j=s+1}^{m} |R(\tau_j\lambda)| \leqslant C, \quad m \geqslant 1,\ 0 \leqslant s \leqslant m, \tag{2.24}$$

the global error can be estimated by $|e^m| \leqslant C\Big(|e^0| + \sum\limits_{s=1}^{m} |\mathrm{loc}^s|\Big)$. Comparing Definition 2.1 and (2.24), we can see that (2.24) is a natural generalization of condition $|R(\tau\lambda)| \leqslant 1$ designed for equidistant step-size.

2.6. Runge-Kutta methods, basic description. Let us consider the semidiscrete problem (1.4). The Runge-Kutta discretization can be expressed by

$$\begin{aligned} K_i + \mathcal{A}_h\Big(U^{m-1} + \tau_m \sum_{j=1}^{k} a_{i,j}K_j\Big) &= f_h(t_{m-1} + \tau_m c_i) \quad \forall\, i = 1, \dots, k, \\ U^m - U^{m-1} &= \tau_m \sum_{i=1}^{k} b_i K_i, \end{aligned} \tag{2.25}$$

where the initial condition U^0 is again the H-orthogonal projection of u^0 to X_h. The general Runge-Kutta method can be described by its coefficients $k \geqslant 1$, $a_{i,j}$, b_i and c_i. This description of the general Runge-Kutta method may seem very messy, but the individual terms and coefficients can be usually interpreted in a manner that

may bring some light into the method. To explain this, it is best to start from (1.4) which we want to solve. After integration over the subinterval I_m we get

$$u_h^m - u_h^{m-1} = \int_{I_m} u_h' \,\mathrm{d}t = \int_{I_m} f_h - \mathcal{A}_h u_h \,\mathrm{d}t. \tag{2.26}$$

The right-hand side of this equation is difficult to compute and therefore, it is approximated by some quadrature. Then the coefficients c_i (they are typically in $[0,1]$) are quadrature nodes on the reference interval $[0,1]$ and $t_{m-1} + \tau_m c_i$ are the corresponding quadrature nodes on I_m. The values $K_i \in X_h$ are called the inner stages and they represent the approximations of u_h' (or equivalently $f_h - \mathcal{A}_h u_h$) at quadrature nodes. Finally, b_i are quadrature weights on the reference interval $[0,1]$ and since the method manipulates with the nodal values only we approximate

$$\begin{aligned} u_h(t_{m-1} + \tau_m c_i) &= u_h^{m-1} + \int_{t_{m-1}}^{t_{m-1}+\tau_m c_i} u_h' \,\mathrm{d}t \\ &\approx u_h^{m-1} + \tau_m \sum_{j=1}^{k} a_{i,j} u_h'(t_{m-1} + \tau_m c_j) \\ &\approx u_h^{m-1} + \tau_m \sum_{j=1}^{k} a_{i,j} K_j \end{aligned} \tag{2.27}$$

with the aid of the coefficients $a_{i,j}$. The coefficients $a_{i,j}$ play the role of quadrature weights for the approximation of integrals $\int_{t_{m-1}}^{t_{m-1}+\tau_m c_i}$.

Sometimes, Runge-Kutta methods can be expressed by alternative (and equivalent) description using another inner stages g_i. These inner stages g_i represent approximations of u_h instead of u_h' at the quadrature nodes. Using this idea, we can reformulate (2.25) into

$$\begin{aligned} g_i - U^{m-1} + \tau_m \sum_{j=1}^{k} a_{i,j} \mathcal{A}_h g_j &= \tau_m \sum_{j=1}^{k} a_{i,j} f_h(t_{m-1} + \tau_m c_j), \quad i = 1, \dots, k, \\ U^m - U^{m-1} + \tau_m \sum_{i=1}^{k} b_i \mathcal{A}_h g_i &= \tau_m \sum_{i=1}^{k} b_i f_h(t_{m-1} + \tau_m c_i). \end{aligned} \tag{2.28}$$

It can be shown that

$$g_i = U^{m-1} + \tau_m \sum_{j=1}^{k} a_{i,j} K_j, \quad K_i = f_h(t_{m-1} + \tau_m c_i) - \mathcal{A}_h g_i. \tag{2.29}$$

The coefficients of Runge-Kutta methods are usually described by Butcher's array

$$\begin{array}{c|ccc} c_1 & a_{1,1} & \dots & a_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ c_k & a_{k,1} & \dots & a_{k,k} \\ \hline & b_1 & \dots & b_k \end{array} = \begin{array}{c|c} c & A \\ \hline & b^{\mathrm{T}} \end{array}. \tag{2.30}$$

Zero entries in the matrix A are often omitted in notation.

Runge-Kutta methods with the matrix A strictly lower triangular, i.e. $a_{i,j} = 0$ for $i \leqslant j$, are called explicit. Inner stages of explicit Runge-Kutta methods can be computed from the previous solution state U^{m-1} and from the other already computed inner stages. Otherwise, the methods are called implicit and the inner stages are linked in more complicated manner. A fully implicit Runge-Kutta method requires to solve a linear problem of the size $k \cdot \dim(X_h)$. A special case of implicit Runge-Kutta methods are those with matrix A lower triangular, i.e. $a_{i,j} = 0$ for $i+1 \leqslant j$. These methods are called diagonally implicit (DIRK) and require to solve an individual implicit relation for each inner stage separately. Even more specialized DIRK methods are called SDIRK with a single value on the diagonal of A, i.e. $a_{i,i} = \gamma$ for each i. The advantage of SDIRK is that they have the same matrix in the implicit relation for each inner stage. For the details about DIRK and SDIRK methods see [2].

The existence of the explicit Runge-Kutta solution follows from the explicit character of the method. The implicit case is more delicate. The existence of the implicit Runge-Kutta solution can be guaranteed by the Banach fixed-point theorem for sufficiently small step-size τ. This option is completely unsuitable for stiff problems, since then the step-size is severely restricted. The theory of the existence and uniqueness for stiff problems is studied in [7] or [12]. For all implicit Runge-Kutta methods applied to problem (1.4) or (1.7) with $\operatorname{Re}\lambda \geqslant 0$ discussed in this paper the existence and uniqueness of the solution is guaranteed for arbitrary step-size.

Some examples of explicit Runge-Kutta methods are the forward Euler method, the Runge method and the most popular Runge-Kutta method of order 4:

$$\begin{array}{c|c} 0 & \\ \hline & 1 \end{array}, \quad \begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 0 & 1 \end{array}, \quad \begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}. \tag{2.31}$$

The simplest examples of implicit Runge-Kutta methods are backward Euler method, midpoint rule, Crank-Nicolson method and θ-scheme, $0 \leqslant \theta \leqslant 1$:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}, \quad \begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}, \quad \begin{array}{c|cc} 0 & & \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}, \quad \begin{array}{c|cc} 0 & & \\ 1 & 1-\theta & \theta \\ \hline & 1-\theta & \theta \end{array}. \tag{2.32}$$

Some examples of implicit Runge-Kutta methods of higher order: Radau IIA method (of order 3) and Kunzmann-Butcher (or Gauss-Legendre) method (of order 4):

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}, \qquad \begin{array}{c|cc} 1/2-a & 1/4 & 1/4-a \\ 1/2+a & 1/4+a & 1/4 \\ \hline & 1/2 & 1/2 \end{array}, \tag{2.33}$$

where $a = \sqrt{6}/3$.

2.7. Stability functions of Runge-Kutta methods. Applying a general Runge-Kutta method with equidistant step-size to Dahlquist equation (1.7), we get

$$\begin{aligned} g_i - Y^{m-1} + \tau\lambda \sum_{j=1}^{k} a_{i,j} g_j = 0, \\ Y^m - Y^{m-1} + \tau\lambda \sum_{i=1}^{k} b_i g_i = 0. \end{aligned} \tag{2.34}$$

Using the notation $g = (g_1, \dots, g_k)^{\mathrm{T}}$ and $\mathbf{1} = (1, \dots, 1)^{\mathrm{T}}$ we can rewrite these relations in a more comfortable matrix-vector formulation

$$\begin{aligned} g - Y^{m-1}\mathbf{1} + \tau\lambda A g = 0, \\ Y^m - Y^{m-1} + \tau\lambda b^{\mathrm{T}} g = 0. \end{aligned} \tag{2.35}$$

Eliminating the vector g from these relations, we get

$$Y^m = (1 - \tau\lambda b^{\mathrm{T}}(I + \tau\lambda A)^{-1}\mathbf{1}) Y^{m-1} \tag{2.36}$$

and this immediately implies that the stability function is $R(z) = 1 - zb^{\mathrm{T}}(I + zA)^{-1}\mathbf{1}$. We are able to show another formula for the stability function of Runge-Kutta methods that would be more useful for our considerations. Rewriting (2.35) in one linear system

$$\left(\begin{array}{c|c} I + \tau\lambda A & 0 \\ \hline \tau\lambda b^{\mathrm{T}} & 1 \end{array}\right) \left(\begin{array}{c} g \\ \hline Y^m \end{array}\right) = Y^{m-1} \left(\begin{array}{c} \mathbf{1} \\ \hline 1 \end{array}\right), \tag{2.37}$$

we can apply Cramer's rule to compute Y^m. Then we get

$$R(z) = \frac{\det(I + zA - z\mathbf{1}b^{\mathrm{T}})}{\det(I + zA)}. \tag{2.38}$$

This formula immediately implies that $R(z)$ is a rational function in z with polynomial degrees in both the numerator and denominator $\leqslant k$. Moreover, for explicit methods the matrix A is strictly lower triangular and the determinant in the denominator equals 1, i.e. $R(z)$ is in fact a polynomial in z in this situation. Therefore, the stability regions of explicit Runge-Kutta methods are bounded and these methods cannot be unconditionally stable.

We present here examples of stability functions of Runge-Kutta methods mentioned at the end of Section 2.6. Stability functions of the forward Euler method, the Runge method and the 4th-order Runge-Kutta method are

$$R(z) = 1 - z, \quad R(z) = 1 - z + \frac{z^2}{2}, \quad R(z) = 1 - z + \frac{z^2}{2} - \frac{z^3}{6} + \frac{z^4}{24}, \tag{2.39}$$

respectively.

Figure 1 shows the boundedness of stability regions of explicit Runge-Kutta methods of maximal order.

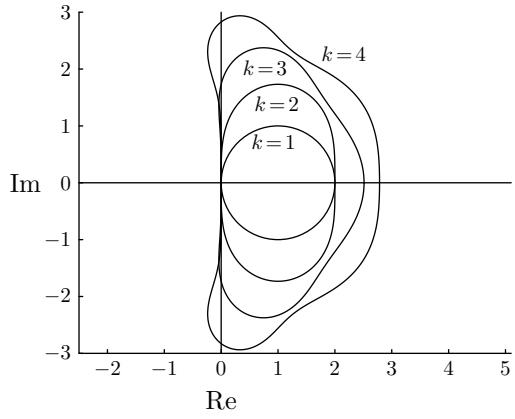

Figure 1. Stability regions of explicit Runge-Kutta methods for $k = 1, 2, 3, 4$ of maximal order k.

The stability function of the backward Euler method, the midpoint rule (the Crank-Nicolson method has the same stability function as the midpoint rule) and the θ-scheme are

$$R(z) = \frac{1}{1+z}, \quad R(z) = \frac{1 - z/2}{1 + z/2}, \quad R(z) = \frac{1 - (1-\theta)z}{1 + \theta z}, \tag{2.40}$$

respectively. Figure 2 shows the evolution of stability regions of the θ-scheme. For $\theta = 0$, i.e. the forward Euler method, the stability region is a unit disc centred at $z = 1$. As θ increases to $1/2$, the centre of the stability region travels to $z = \infty$. This means that the stability region of the Crank-Nicolson method, i.e. $\theta = 1/2$,

is the whole right half-plane. As θ increases from 1/2 to 1, the boundary of the stability region is again a circle with its centre traveling from $z = \infty$ to $z = -1$ (backward Euler method). Then the corresponding stability region is the exterior of the boundary circle in this case.

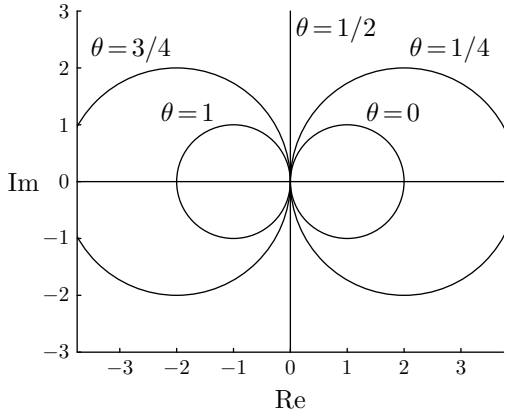

Figure 2. Stability regions of θ-scheme $\theta = 0, 1/4, 1/2, 3/4, 1$.

The stability function of the Radau IIA method (order 3) and the Kunzmann-Butcher method (order 4) are

$$R(z) = \frac{1 - z/3}{1 + 2z/3 + z^2/6}, \quad R(z) = \frac{1 - z/2 + z^2/12}{1 + z/2 + z^2/12}, \tag{2.41}$$

respectively.

Figure 3 enables us to observe A-stability of Radau IIA methods ($k = 1, 2, 3$). As we will see later, the stability region of all Kuntzmann-Butcher methods is exactly the right half-plane.

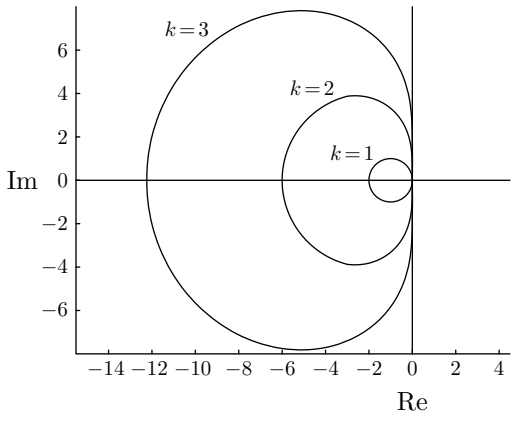

Figure 3. Stability regions of Radau IIA methods $k = 1, 2, 3$.

From the examples mentioned above we can see that many stability functions are *Padé approximations* to e^{-z}.

2.8. Padé approximations. Padé approximations are rational approximations to e^{z} that are optimal with respect to polynomial degrees in the numerator and denominator, see [28]. In our setting we will need to approximate e^{-z} instead of e^{z}. The resulting relations for e^{z} and e^{-z} are completely the same up to the substitution $z \to -z$.

We call the rational function

$$R_{i,j}(z) = \frac{P_{i,j}(z)}{Q_{i,j}(z)} \tag{2.42}$$

with the numerator and denominator

$$P_{i,j}(z) = \sum_{s=0}^{i} (-1)^s \frac{i!\,(i+j-s)!}{s!\,(i-s)!\,(i+j)!} z^s, \tag{2.43}$$

$$Q_{i,j}(z) = \sum_{s=0}^{j} \frac{j!\,(i+j-s)!}{s!\,(j-s)!\,(i+j)!} z^s$$

the Padé approximation of e^{-z}.

The lowest order examples are

(2.44)

	$i=0$	$i=1$	$i=2$
$j=0$	$\dfrac{1}{1}$	$\dfrac{1-z}{1}$	$\dfrac{1-z+z^2/2}{1}$
$j=1$	$\dfrac{1}{1+z}$	$\dfrac{1-z/2}{1+z/2}$	$\dfrac{1-2z/3+z^2/6}{1+z/3}$
$j=2$	$\dfrac{1}{1+z+z^2/2}$	$\dfrac{1-z/3}{1+2z/3+z^2/6}$	$\dfrac{1-z/2+z^2/12}{1+z/2+z^2/12}$

Theorem 2.1. *Let $R_{i,j}$ be the Padé approximation of e^{-z}. Then*

$$R_{i,j}(z) - \mathrm{e}^{-z} = O(z^{i+j+1}). \tag{2.45}$$

P r o o f. The proof can be found in [28]. □

Let us point out that $R_{i,j}$ is the only rational function with degree i in the numerator and degree j in the denominator such that $R_{i,j}(z) - \mathrm{e}^{-z} = O(z^{i+j+1})$ and all other rational functions with the same degrees provide approximations of lower order.

There is a natural question which Padé approximations lead to A-stable methods.

Theorem 2.2. *The Padé approximation $R_{i,j}(z)$ of e^{-z} is A-stable, that is $|R_{i,j}(z)| \leqslant 1$ for any z with $\operatorname{Re} z \geqslant 0$ if and only if $i \leqslant j \leqslant i+2$.*

P r o o f. The proof can be found in [22]. □

According to Theorem 2.2, we shall focus on methods with diagonal Padé approximations, because of their maximal achievable order of convergence $2k$ and A-stability, and methods with first subdiagonal Padé approximations, since these methods are even more robust (have larger stability regions) than the methods based on the diagonal approximations they are of still very attractive order $2k-1$.

2.9. Collocation methods and Galerkin methods. Let us consider (1.4). Let $U^0 = u_h(0)$ and c_i, $i = 1, \dots, k$ be distinct coefficients, typically $c_i \in [0,1]$. At each time interval I_m we define a polynomial $p \in P_k(I_m, X_h)$, i.e. a polynomial of degree k in time, such that

$$p(t_{m-1}) = U^{m-1}, \tag{2.46}$$
$$p'(t_{m-1} + \tau_m c_i) + \mathcal{A}_h p(t_{m-1} + \tau_m c_i) = f_h(t_{m-1} + \tau_m c_i), \quad i = 1, \dots, k.$$

The polynomial p is called the *collocation polynomial* and we define one step of the *collocation method* as $U^m = p(t_m)$.

Now, we will show that collocation methods are a special class of implicit Runge-Kutta methods.

Lemma 2.1. *Let c_i, $i = 1, \dots, k$ be the coefficients of a collocation method. Then this collocation method is equivalent to the Runge-Kutta method using the same coefficients c_i and*

$$a_{i,j} = \int_0^{c_i} l_j \,\mathrm{d}t, \quad b_i = \int_0^1 l_i \,\mathrm{d}t, \tag{2.47}$$

where l_j are the Lagrange interpolation basis polynomials of degree $k-1$ satisfying $l_j(c_i) = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker symbol.

P r o o f. The original proof can be found in [20] or [37]. Let p be the collocation polynomial corresponding to the coefficients c_i. Since p' is a polynomial of degree $k-1$, we can write

$$p'(t_{m-1} + \tau_m t) = \sum_{j=1}^{k} K_j l_j(t), \tag{2.48}$$

where $K_j = p'(t_{m-1} + \tau_m c_j)$. Since the collocation polynomial satisfies the collocation conditions (2.46), we get

$$K_i = p'(t_{m-1} + \tau_m c_i) = f_h(t_{m-1} + \tau_m c_i) - \mathcal{A}_h p(t_{m-1} + \tau_m c_i). \tag{2.49}$$

Using

$$p(t_{m-1}+\tau_m c_i) = p(t_{m-1}) + \tau_m \int_0^{c_i} p'(t_{m-1}+\tau_m t) \tag{2.50}$$
$$= U^{m-1} + \tau_m \sum_{j=1}^{k} a_{i,j} K_j,$$

we get relation (2.25) for inner stages of Runge-Kutta methods. Finally, the relation for the solution update can be obtained from

$$U^m = p(t_m) = p(t_{m-1}) + \tau_m \int_0^1 p'(t_{m-1}+\tau_m t) = U^{m-1} + \tau_m \sum_{j=1}^{k} b_i K_j. \tag{2.51}$$

□

It should be noted that the proof shows not only that the solutions of the two methods coincide, but also that the values at the collocation points coincide with the inner stages of the corresponding Runge-Kutta method.

To define Galerkin discretizations we introduce spaces

$$Y_h^\tau = \{v \in C([0,T], X_h)\colon\ v(0) = u_h(0), v|_{I_m} \in P_k(I_m, X_h)\}, \tag{2.52}$$
$$Z_h^\tau = \{v \in L^2(0,T,X_h)\colon\ v|_{I_m} \in P_{k-1}(I_m, X_h)\}.$$

We should mention that functions from Z_h^τ are not continuous in general at discretization nodes $\{t_m\}_{m\geqslant 0}$. Therefore, we set the notation for one-sided limits $v_\pm^m = v(t_m\pm) = \lim\limits_{t\to t_m\pm} v(t)$ and jumps $[v]_m = v_+^m - v_-^m$. The design of the spaces Y_h^τ and Z_h^τ is motivated by an alternative weak formulation to problem (1.1): find $u \in Y$ such that

$$\int_0^{\mathrm{T}} (u' + \mathcal{A}u, v)\,\mathrm{d}t = \int_0^{\mathrm{T}} (f, v)\,\mathrm{d}t, \quad v \in Z, \tag{2.53}$$

where

$$Z = L^2(0,T,X), \quad Y = \{v \in Z\colon\ v' \in L^2(0,T,X^*), v(0) = u^0\}. \tag{2.54}$$

Then Z_h^τ is a natural finite-dimensional subspace of Z and Y_h^τ is a finite-dimensional space that is a subspace of Y except for satisfying the initial condition, since $u_h(0) \neq u^0$ in general.

Since both spaces Y_h^τ and Z_h^τ have the same dimension $r \cdot k \cdot \dim X_h$, we can formulate the natural *continuous (conforming) (Petrov-)Galerkin* discretization: find $U \in Y_h^\tau$ such that

$$\int_0^{\mathrm{T}} (U' + \mathcal{A}U, v_h^\tau)\,\mathrm{d}t = \int_0^{\mathrm{T}} (f, v_h^\tau)\,\mathrm{d}t, \quad v \in Z_h^\tau. \tag{2.55}$$

A similar approach can be used for the *discontinuous (nonconforming) Galerkin* discretization, where we seek $U \in Z_h^\tau$, although $Z_h^\tau \not\subset Y$. This fact needs to be compensated by additional (penalization) terms: find $U \in Z_h^\tau$ such that

$$\int_0^{\mathrm{T}} (U' + \mathcal{A}U, v_h^\tau)\,\mathrm{d}t + \sum_{m=1}^{r} ([U]_{m-1}, v_{h+}^{\tau\,m-1}) = \int_0^{\mathrm{T}} (f, v_h^\tau)\,\mathrm{d}t, \quad v \in Z_h^\tau. \tag{2.56}$$

The penalization term plays the role of the mediator between the solutions at I_{m-1} and I_m. The usual interpretation for this choice comes from the use of upwind numerical flux, see e.g. [32].

Continuous and discontinuous Galerkin methods have integrals in their formulations. These integrals could be naturally approximated by suitable quadrature formulae. We denote these quadratures by

$$\int_{I_m} F(t)\,\mathrm{d}t \approx Q_m(F) = \tau_m \sum_{i=1}^{k} \omega_i F(t_{m-1} + \tau_m \psi_i). \tag{2.57}$$

For continuous Galerkin methods we employ the classical Gauss quadrature on k quadrature nodes. The resulting quadrature has degree $2k - 1$. This implies that the integrals containing linear terms are evaluated exactly. For the discontinuous Galerkin variant we employ the right (Gauss)-Radau quadrature on k quadrature nodes, i.e. one quadrature node lies at t_m and the remaining quadrature nodes are distributed in such a way that the maximal degree for this situation is achieved. The resulting quadrature has degree $2k - 2$ and once again the integrals containing linear terms are evaluated exactly. Both versions of Galerkin methods with these quadratures will be called *quadrature variants*.

The following lemmas show that quadrature variants of both Galerkin methods are equivalent to the collocation methods based on Gauss or Radau quadrature nodes, respectively.

Lemma 2.2. *Let U be the solution of the quadrature variant of the continuous Galerkin method. Then $U(t) = p(t)$, $t \in I_m$, where p is the collocation polynomial obtained from the collocation method using Gauss quadrature nodes as collocation points.*

Lemma 2.3. *Let $r \in P_k$ be the Radau polynomial satisfying $r(0) = 1$, $r(1) = 0$, and $r \perp P_{k-2}$. Let U be the solution of the quadrature variant of the discontinuous Galerkin method. Then*

$$U(t) - [U]_{m-1} r\Big(\frac{t - t_{m-1}}{\tau_m}\Big) = p(t), \quad t \in I_m, \tag{2.58}$$

where p is the collocation polynomial obtained from the collocation method using right Radau quadrature nodes as collocation points.

An alternative definition of Radau polynomial is that the zeros of r lie at (reference) right Radau quadrature nodes in $(0, 1]$ and $r(0) = 1$. Using this property, we can interpret (2.58). Although $U \neq p$, they are equal at quadrature points $t_{m-1} + \tau_m c_i$. Since $c_k = 1$ it means that $U_-^m = p(t_m) = U^m$.

The proofs of both the lemmas are essentially the same. For this reason we will prove only the latter (more difficult) one. The original proof of Lemma 2.2 can be found in [25].

P r o o f. Setting the test function

$$v_h^\tau(t, x) = l_i\Big(\frac{t - t_{m-1}}{\tau_m}\Big) w_h(x), \tag{2.59}$$

with $w_h \in X_h$ arbitrary, we find that

$$\begin{aligned} \tau_m \omega_i U'(t_{m-1} + \tau_m c_i) + \tau_m \omega_i \mathcal{A}_h U(t_{m-1} + \tau_m c_i) + [U]_{m-1} l_i(0) \\ = \tau_m \omega_i f_h(t_{m-1} + \tau_m c_i). \end{aligned} \tag{2.60}$$

Comparing this relation with (2.46), we can see that this is almost what we want. Now, it is only necessary to show that $l_i(0) = -\omega_i r'(c_i)$. Using the exactness of the Radau quadrature for polynomials of degree $2k - 2$, per-partes and properties of the Radau polynomial, we get

$$\begin{aligned} -\omega_i r'(c_i) = -\omega_i r'(c_i) l_i(c_i) = -\int_0^1 r'(t) l_i(t) \,\mathrm{d}t \\ = \int_0^1 r(t) l_i'(t) \,\mathrm{d}t - r(1) l_i(1) + r(0) l_i(0) = l_i(0). \end{aligned} \tag{2.61}$$

□

Runge-Kutta methods equivalent to the quadrature version of continuous Galerkin methods are known as Kuntzmann-Butcher (or Gauss-Legendre) methods, first introduced in [26] and [5]. Runge-Kutta methods equivalent to the quadrature version of discontinuous Galerkin methods are known as Radau IIA methods, first introduced in [14].

The next theorem comes from the famous paper [5]:

Theorem 2.3. *Let the coefficients of the Runge-Kutta method satisfy (simplifying order conditions)*

$$\sum_{i=0}^{k} b_i c_i^{q-1} = \frac{1}{q}, \quad q = 1, \dots, p, \tag{2.62}$$

$$\sum_{j=1}^{k} a_{i,j} c_j^{q-1} = \frac{c_i^q}{q}, \quad i = 1, \dots, k, \ q = 1, \dots, p_1,$$

$$\sum_{i=1}^{k} a_{i,j} b_i c_i^{q-1} = \frac{b_j}{q}(1 - c_j^q), \quad j = 1, \dots, k, \ q = 1, \dots, p_2,$$

where $p \leqslant p_1 + p_2 + 1$ and $p \leqslant 2p_1 + 2$. Then the method is of order p.

In the next theorem we present the properties of Kuntzmann-Butcher methods and Radau IIA methods.

Theorem 2.4. *Kutzmann-Butcher methods (and equivalently collocation methods on Gauss quadrature nodes and the quadrature variant of continuous Galerkin methods) have as their stability function the diagonal Padé approximations, order $2k$ and are A-stable. Radau IIA Runge-Kutta methods (and right Radau collocation methods and the quadrature variant of discontinuous Galerkin methods) have as their stability function the first subdiagonal Padé approximations, order $2k - 1$ and are A-stable.*

P r o o f. The proof is very similar for the two methods. For this reason we will present only the Radau IIA variant of the proof. The coefficients $a_{i,j}$ and b_i satisfy the relations (2.47), where c_i are right Radau quadrature nodes in $(0, 1]$. Then it is obvious that the first relation of (2.62) is satisfied up to $p = 2k - 1$, since the right Radau quadrature is accurate up to degree $2k - 2$. For similar reasons we can see that the second relation is satisfied up to $p_1 = k$. Moreover, satisfying the first relation up to $p = 2k - 1$ and the second up to $p_1 = k$ implies that the third is satisfied up to $p_2 = k - 1$, see e.g. [22]. Then following Theorem 2.3 we can confirm that the order of Radau IIA methods is $2k - 1$. Since $c_k = 1$, we have $a_{k,i} = b_i$ and the degree of the nominator of the stability function is $\leqslant k - 1$. Comparing this fact with Theorem 2.1 we conclude that the stability functions must be the Padé approximations $R_{k-1,k}$. The A-stability then follows from Theorem 2.2. □

2.10. Order reduction. We have shown that Galerkin methods have (formally) order $2k$ (continuous version) or $2k - 1$ (discontinuous version). In the context of

Galerkin methods these high orders are usually called *superconvergence*, see e.g. [1]. But lower order of convergence is often observed for stiff problems. This, the so-called *order reduction* phenomenon, was studied first in [29]. We will show this phenomenon on the numerical example motivated by [31].

E x a m p l e 2.1. Let us consider the equation

$$y'(t) + \lambda y(t) = f(t), \quad t \in (0, 2), \tag{2.63}$$

with parameter $\lambda = 1$, $\lambda = 100$ or $\lambda = 10^5$ and prescribed exact solution $y(t) = \sin(2\pi t)$. We discretize this problem by 2-stages Radau IIA Runge-Kutta method with equidistant time steps.

	$\lambda = 1$		$\lambda = 10^2$		$\lambda = 10^5$	
$1/\tau$	error	rate	error	rate	error	rate
10	9.02E−3	–	7.78E−3	–	9.27E−6	–
20	1.09E−3	3.05	1.71E−3	2.19	2.64E−6	1.81
40	1.34E−4	3.02	3.12E−4	2.45	6.81E−7	1.96
80	1.67E−5	3.01	5.05E−5	2.63	1.71E−7	1.99
160	2.09E−6	3.00	7.40E−6	2.77	4.28E−8	2.00
320	2.61E−7	3.00	1.01E−6	2.87	1.07E−8	2.01
640	3.26E−8	3.00	1.33E−7	2.93	2.64E−9	2.01
1280	4.07E−9	3.00	1.70E−8	2.96	6.48E−10	2.03
2560	5.10E−10	3.00	2.16E−9	2.98	1.57E−10	2.05
5120	6.24E−11	3.03	2.70E−10	3.00	3.53E−11	2.15

Table 3. Errors and convergence rates at time $t = 2$.

Table 3 shows different behaviour of the order of the same method for different $\tau\lambda$. The first column ($\lambda = 1$) represents the non-stiff case, where $\tau\lambda$ is mainly influenced by τ. In this case the experimental order of the Radau IIA method is $2k - 1 = 3$. On the other hand, the last column ($\lambda = 10^5$) represents a stiff problem. Here $\tau\lambda$ remains large even for quite small τ and the experimental order achieved is $k = 2$. The interesting situation appears in the middle column ($\lambda = 100$), since this column represents a transition state between the stiff and non-stiff problem. For τ sufficiently large $\tau\lambda$ remains large enough and the problem behaves as stiff. Decreasing τ makes $\tau\lambda$ smaller and the problem is fluently turned into non-stiff.

Here we shall explain the strange behaviour of orders in Example 2.1. We apply a Runge-Kutta method with equidistant time step to problem (2.63):

$$\begin{aligned} g - \mathbf{1}Y^{m-1} + \tau\lambda A g &= \tau A F, \\ Y^m - Y^{m-1} + \tau\lambda b^{\mathrm{T}} g &= \tau b^{\mathrm{T}} F, \end{aligned} \tag{2.64}$$

where the elements of the vector F are $f(t_{m-1}+\tau c_i)$. Similar relations hold also for the exact solution:

$$y(t_{m-1}+\tau c_i) - y^m = \tau \sum_{j=1}^{k} a_{i,j} y'(t_{m-1}+\tau c_j) + \Delta_i, \quad i = 1, \dots, k, \tag{2.65}$$

$$y^m - y^{m-1} = \tau \sum_{i=1}^{k} b_i y'(t_{m-1}+\tau c_i) + \Delta_0.$$

The defects are $\Delta_0 = O(\tau^{p+1})$, $\Delta_i = O(\tau^{p_1+1})$, where p and p_1 are coefficients from (2.62) and typically $p_1 \leqslant p$, e.g. Radau IIA methods have $p = 2k-1$ and $p_1 = k$. It is important to notice that these orders are not influenced by the stiffness of the problem, since λ is not present in (2.65). Setting $\Delta = (\Delta_1, \dots, \Delta_k)^{\mathrm{T}}$, we get the local error equation

$$e^m = R(\tau\lambda) e^{m-1} + \tau\lambda b^{\mathrm{T}} (I + \tau\lambda A)^{-1} \Delta - \Delta_0, \tag{2.66}$$

where the term $\tau\lambda b^{\mathrm{T}}(I+\tau\lambda A)^{-1}\Delta - \Delta_0$ plays the role of the local error. If order conditions of Theorem 2.3 are satisfied then the local error can be estimated from above by $C(\tau\lambda)\tau^p$. For stiff problems the constant $C(\tau\lambda)$ can be very high and such a local error estimate is unrealistic. In general, it is possible to estimate the elements of the vector $zb^{\mathrm{T}}(I+zA)^{-1}$ by $O(1)$. This estimate leads to suboptimal local order (and afterwards global order) of convergence, namely $\min(p, p_1)$. This reduced order is sometimes called the *stiff* or *effective* order. The typical order reduction from (non-stiff) order p to the stiff order p_1 can be decreased, if additional assumptions are satisfied.

Theorem 2.5. *Let us consider the equation* (1.1) *and a time discretization of this problem via the A-stable Runge-Kutta method. Let p and q be the non-stiff order and the stiff order of the Runge-Kutta method, respectively. Let the exact solution satisfy additional regularity assumptions*

$$u^{(s)} \in L^\infty(0, T, \mathrm{Dom}(\mathcal{A}^{p+1-s})), \quad s = q, \dots, p+1. \tag{2.67}$$

Then the method has order p and the error can be estimated by

$$\|U^m - u^m\| \leqslant C\tau^p \max_{q+1 \leqslant s \leqslant p+1} \|\mathcal{A}^{p+1-s} u^{(s)}\|, \tag{2.68}$$

where the constant C is independent of $\mathcal{A}$.

P r o o f. The proof can be found in [4]. □

Strange regularity conditions (2.67) in Theorem 2.5 are (up to a mild generalization) necessary and sufficient conditions for achieving order p in stiff case, i.e. the constant in the error analysis is independent of $\mathcal{A}$ or λ, see [4]. On the other hand, it is possible to design some weaker version of these conditions to achieve order of convergence between q and p, e.g. order $q+1$ is discussed in [15]. It should be also mentioned that regularity conditions (2.67) are usually considered unnatural, especially for higher orders. For example, in classical parabolic PDE case this means that the time derivatives of the exact solution must satisfy some regularity with respect to space and some higher order boundary conditions. Finally, assuming homogeneity, i.e. $f = 0$, we can obtain by repeated differentiation of (1.1) that (2.67) is equivalent to the traditional condition of boundedness of $u^{(p+1)}$.

2.11. Error analysis, parabolic problem. Now, we will summarize the results obtained for one-step methods and apply them to the error analysis of time discretization of (1.4). Let us consider the A-stable Runge-Kutta discretization of stiff order $q = \min(p, p_1)$, where p and p_1 are constants from (2.62). We shall assume the exact solution to be sufficiently smooth, namely

$$u \in W^{q+1,\infty}(0,T,H) \cap W^{1,\infty}(0,T,V). \tag{2.69}$$

Then it is possible to prove

$$\begin{aligned} &\|R_h u(t_{m-1}+\tau_m c_i) - u(t_{m-1}+\tau_m c_i) - R_h u^{m-1} + u^{m-1}\| \leqslant \tau_m \mathrm{err}(h), \\ &\left\| u(t_{m-1}+\tau_m c_i) - u^{m-1} - \tau_m \sum_{j=1}^{k} a_{i,j} u'(t_{m-1}+\tau_m c_j) \right\| \leqslant C\tau_m^{q+1}, \\ &\left\| u^m - u^{m-1} - \tau_m \sum_{i=1}^{k} b_i u'(t_{m-1}+\tau_m c_i) \right\| \leqslant C\tau_m^{q+1}. \end{aligned} \tag{2.70}$$

The first relation comes from the approximation property of R_h. The second and third relations are consequences of (2.62) and the constant C depends on $\|u\|_{W^{q+1,\infty}(0,T,H)}$. Similarly to the analysis of Euler methods we decompose the error into ξ and η. Let us denote the error at the inner stages $e_{g,i} = g_i - u(t_{m-1}+\tau_m c_i)$ and its corresponding parts $\xi_{g,i} = g_i - R_h u(t_{m-1}+\tau_m c_i)$, $\eta_{g,i} = R_h u(t_{m-1}+\tau_m c_i) - u(t_{m-1}+\tau_m c_i)$. Once again, the η terms can be directly estimated by the properties of the space discretization, i.e. $\|\eta^m\| \leqslant \mathrm{err}(h)$. Our aim

is providing the error estimate for ξ^m. From (2.28) and from

$$(u(t_{m-1}+\tau_m c_i)-u^{m-1},v_h)+\tau_m\sum_{j=1}^{k}a_{i,j}(\mathcal{A}u(t_{m-1}+\tau_m c_j),v_h) \tag{2.71}$$
$$=\tau_m\sum_{j=1}^{k}a_{i,j}(f(t_{m-1}+\tau_m c_j),v_h)+(u(t_{m-1}+\tau_m c_i)-u^{m-1},v_h)$$
$$-\tau_m\sum_{j=1}^{k}a_{i,j}(u'(t_{m-1}+\tau_m c_j),v_h)$$

we can derive error equations for inner stages

$$\xi_{g,i}-\xi^{m-1}+\tau_m\sum_{j=1}^{k}a_{i,j}\mathcal{A}_h\xi_{g,j}=\Delta_i^m,\quad i=1,\dots,k, \tag{2.72}$$

where $\|\Delta_i^m\|\leqslant C\tau_m^{q+1}+\tau_m\mathrm{err}(h)$. Similarly we can derive the error equation for the solution update

$$\xi^m-\xi^{m-1}+\tau_m\sum_{i=1}^{k}b_i\mathcal{A}_h\xi_{g,i}=\Delta_0^m, \tag{2.73}$$

where $\|\Delta_0^m\|\leqslant C\tau_m^{q+1}+\tau_m\mathrm{err}(h)$. Eliminating the inner stage errors $\xi_{g,i}$ we obtain the error equation

$$\xi^m=R(\tau_m\mathcal{A}_h)\xi^{m-1}+\mathrm{loc}^m=R(\tau_m\mathcal{A}_h)^m\xi^0+\sum_{s=1}^{m}R(\tau_m\mathcal{A}_h)\,\mathrm{loc}^s \tag{2.74}$$

with

$$R(\tau_m\mathcal{A}_h)=\mathcal{I}-(b^{\mathrm{T}}\otimes(\tau_m\mathcal{A}_h))(I\otimes\mathcal{I}+A\otimes(\tau_m\mathcal{A}_h))^{-1}(\mathbf{1}\otimes\mathcal{I}), \tag{2.75}$$
$$\mathrm{loc}^m=-(b^{\mathrm{T}}\otimes(\tau_m\mathcal{A}_h)(I\otimes\mathcal{I}+A\otimes(\tau_m\mathcal{A}_h))^{-1}\Delta^m+\Delta_0^m),$$

where $\Delta^m=(\Delta_1^m,\dots,\Delta_k^m)^{\mathrm{T}}$, I is the identity on $\mathbb{R}^k$, $\mathcal{I}$ is the identity on X_h and $\otimes$ is the tensor product. Similarly to Section 2.10 it is possible to prove

$$\|\mathrm{loc}^m\|\leqslant C\max_{i=0,\dots,k}\|\Delta_i^m\|\leqslant C\tau_m(\tau^q+\mathrm{err}(h)), \tag{2.76}$$

unless other regularity assumptions are satisfied, see Theorem 2.5. At last it is necessary to estimate $\|R(\tau_m\mathcal{A}_h)^m\|$.

Theorem 2.6. *Let a rational function R be bounded for $\mathrm{Re}\,z\geqslant 0$. Let an operator $\mathcal{B}\colon X\to X^*$ satisfy $(\mathcal{B}u,u)\geqslant 0$ for all $u\in X$. Then*

$$\|R(\mathcal{B})\|\leqslant\sup_{\mathrm{Re}\,z\geqslant 0}|R(z)|. \tag{2.77}$$

P r o o f. The proof can be found in [35]. □

Since the method is A-stable and since $\mathcal{A}_h$ satisfies (1.2), applying Theorem 2.6 to (2.74), we immediately obtain

$$\|\xi^m\| \leqslant \|\xi^0\| + Ct_m(\tau^q + \mathrm{err}(h)). \tag{2.78}$$

3. Linear multi-step methods

Higher order methods discussed so far have achieved their order by using complicated relations among the inner stages. Another very popular approach for achieving higher order is using the already computed solutions at the previous time levels instead of inner stages.

3.1. Multi-step methods, basic description. For simplicity, we will only discuss equidistant time steps τ. Variable step-size multi-step methods and the related analysis can be found in e.g. [8], [17], [18], [21].

Let coefficients α_v and β_v, $v = 0, \ldots, k$, satisfy $\alpha_k > 0$ and $|\alpha_0| + |\beta_0| > 0$. Then we define the *linear k-step method* applied to (1.4) by

$$\sum_{v=0}^{k} \alpha_v U^{m+v} + \tau \sum_{v=0}^{k} \beta_v \mathcal{A}_h U^{m+v} = \tau \sum_{v=0}^{k} \beta_v f_h^{m+v}. \tag{3.1}$$

To be able to compute U^{m+k} it is necessary to know U^{m+v}, $v = 0, \ldots, k-1$. Here a problem arises at the beginning of the process when typically only U^0 can be directly obtained from the initial condition and the remaining U^v, $v = 1, \ldots, k-1$, are unknown. In practice, these values are usually computed by some one-step method. In our next consideration we will ignore the source of these values and on the contrary provide an error analysis that takes into consideration the error at these initial values.

Assuming U^{m+v}, $v = 0, \ldots, k-1$, are known the question of solvability of system (3.1) is equivalent to that of the nonsingularity of the matrix $\alpha_k I + \beta_k \tau \mathcal{A}_h$. The nonsingularity can be guaranteed with $\beta_k \geqslant 0$. For this reason we restrict our next considerations to the case $\beta_k \geqslant 0$.

The method is called explicit if $\beta_k = 0$, otherwise the method is called implicit.

A strange situation occurs when the generating polynomials

$$\alpha(\zeta) = \sum_{v=0}^{k} \alpha_v \zeta^v \qquad \text{and} \qquad \beta(\zeta) = \sum_{v=0^k} \beta_v \zeta^v \tag{3.2}$$

have a common divisor. Dividing these polynomials by their common divisor results in reduced polynomials $\alpha^*(\zeta)$ and $\beta^*(\zeta)$ with coefficients α_v^* and β_v^*, respectively. These polynomials generate another k^*-step method with $k^* < k$. Assuming that both the methods are started from the same initial values $Y^0, \ldots, Y^{k^*-1}$ and the rest of the initial values of the former method are computed with the aid of the new reduced method, both the methods produce exactly the same solution. For this reason it is possible to show that both the methods have similar properties, e.g. the same local order of convergence, and the former method is not of much use. Therefore, we will always assume in the following text that the multi-step method is *irreducible*.

3.2. Error analysis. Here we will focus on the analysis of the general linear multi-step method (3.1) applied to Dahlquist's equation (1.7). The resulting scheme simplifies into

$$\sum_{v=0}^{k} \alpha_v Y^{m+v} + \tau\lambda \sum_{v=0}^{k} \beta_v Y^{m+v} = \sum_{v=0}^{k} (\alpha_v + \tau\lambda\beta_v) Y^{m+v} = 0. \tag{3.3}$$

We define the local error

$$\mathrm{loc}^m = \sum_{v=0}^{k} (\alpha_v + \tau\lambda\beta_v) y^{m+v} = \sum_{v=0}^{k} \alpha_v y^{m+v} - \tau \sum_{v=0}^{k} \beta_v y'(t_{m+v}). \tag{3.4}$$

This definition of local error is independent of λ and therefore, it is independent of stiffness. Expanding y^{m+v} and $y'(t_{m+v})$ into Taylor series gives the following order conditions:

Lemma 3.1. *Let the coefficient of the multi-step method* (3.1) *satisfy*

$$\begin{gathered} \sum_{v=0}^{k} \alpha_v = 0, \\ \sum_{v=0}^{k} \alpha_v v^s = s \sum_{v=0}^{k} \beta_v v^{s-1}, \quad s = 1, \ldots, p. \end{gathered} \tag{3.5}$$

Then the method has local order p, i.e. $\mathrm{loc}^m = O(\tau^{p+1})$.

To derive global error estimates of $e^{n+k} = Y^{n+k} - y^{n+k}$ we will use a technique presented in [11] and [23]. Let us simplify the notation of the coefficients: $a_v = a_v(\tau\lambda) = \alpha_v + \tau\lambda\beta_v$. Then

$$a_k e^{m+k} + a_{k-1} e^{m+k-1} + a_{k-2} e^{m+k-2} + \ldots + a_0 e^m = -\mathrm{loc}^m. \tag{3.6}$$

Multiplying this relations by γ_{n-m} and summing over $m = 1, \dots, n$, we arrive at

$$\begin{aligned}(3.7)\qquad & \gamma_0 a_k e^{n+k} + (\gamma_0 a_{k-1} + \gamma_1 a_k) e^{n+k-1} + \dots \\ & + \left(\sum_{v=0}^{k} \gamma_v a_v \right) e^n + \dots + \left(\sum_{v=0}^{k} \gamma_{n-k+v} a_v \right) e^k + \dots \\ & + (\gamma_{n-1} a_0 + \gamma_n a_1) e^1 + \gamma_n a_0 e^0 = - \sum_{m=1}^{n} \gamma_{n-m} \mathrm{loc}^m.\end{aligned}$$

Suitable coefficients γ_j can be defined by starting values $\gamma_j = 0$ for $j \leqslant -1$, $\gamma_0 = 1/a_k$ and by the difference equation

$$(3.8)\qquad \sum_{v=0}^{k} \gamma_{j+v} a_v = 0.$$

Using such coefficients γ_j simplifies (3.7) into

$$(3.9)\qquad e^{n+k} + \left(\sum_{v=0}^{k-1} \gamma_{n+1-k+v} a_v \right) e^{k-1} + \dots + \gamma_n a_0 e^0 = \sum_{m=1}^{n} \gamma_{n-m} \, \mathrm{loc}^m.$$

Motivated by (3.9) we define the stability of multi-step methods as the uniform boundedness of $\{\gamma_j\}_{j \geqslant 0}$. Let

$$(3.10)\qquad \varrho(\zeta, z) = \sum_{v=0}^{k} (\alpha_v + z\beta_v) \zeta^v = \sum_{v=0}^{k} a_v(z) \zeta^v = \alpha(\zeta) + z\beta(\zeta)$$

be the *characteristic polynomial* of the difference equation (3.8) or (3.3), respectively. Let us denote by $\zeta_i = \zeta_i(z)$, $i = 1, \dots, l$, the zeros of $\varrho(\zeta, z)$ with multiplicities m_i. Then the solution of (3.8) can be expressed by

$$(3.11)\qquad \gamma_j = \gamma_j(z) = \sum_{i=1}^{l} p_i(j) \zeta_i^j,$$

where p_i are polynomials of degree $m_i - 1$. This implies that the sequence $\{\gamma_j\}_{j \geqslant 0}$ is bounded if and only if the zeros ζ_i are bounded by 1 in modulus and those with modulus equal to 1 have multiplicity 1.

Definition 3.1. Let us consider a general multi-step method described by a characteristic polynomial $\varrho(\zeta, z)$. Let us denote the zeros of $\varrho(\zeta, z)$ as $\zeta_i(z)$ and their multiplicities as m_i. Then the set

$$(3.12)\qquad S = \{z \in \mathbb{C} \colon \ |\zeta_i(z)| \leqslant 1, \ |\zeta_i(z)| = 1 \Rightarrow m_i = 1 \ \forall\, i\}$$

is called the *stability region*. We say that the method is stable if $\tau\lambda \in S$. We call the method unconditionally stable if the method is stable for arbitrary step-size $\tau > 0$. Otherwise, we call the method conditionally stable. We call the method *A-stable*, if $\{z \in \mathbb{C}\colon \operatorname{Re} z \geqslant 0\} \subset S$, and we call the method *D-stable*, if $0 \in S$. We call the method *strictly stable* at z, if $z \in S$ and $|\zeta_i(z)| = 1$ implies $\zeta_i(z) = 1$.

The restriction for the local order of D-stable methods is described by the First Dahlquist's Barrier:

Theorem 3.1. *The order p of a D-stable k-step method satisfies*

$$\begin{aligned} p &\leqslant k+2 \quad \text{if } k \text{ is even,} \\ p &\leqslant k+1 \quad \text{if } k \text{ is odd,} \\ p &\leqslant k \quad \text{if } \frac{\beta_k}{\alpha_k} \leqslant 0 \text{ (in particular if the method is explicit).} \end{aligned} \tag{3.13}$$

P r o o f. The proof can be found in [11] or in [21]. □

Moreover, from Definition 3.1 we can deduce that explicit multi-step methods always have bounded stability regions. We can prove it by applying Vieta's formulae

$$\sum_{1 \leqslant i_1 < i_2 < \ldots < i_v \leqslant k} (-1)^v \zeta_{i_1}(z)\zeta_{i_2}(z)\ldots\zeta_{i_v}(z) = \frac{a_{k-v}(z)}{a_k(z)}, \quad v = 1, \ldots, k. \tag{3.14}$$

Assuming $z \in S$, we can see that the left-hand side is always bounded by a constant depending on k, but independent of z. On the other hand, there exists $\beta_{k-v} \neq 0$ for some v and therefore for this v the right-hand side tends to infinity as $z \to \infty$.

In Chapter 2, we have been mainly interested in A-stable methods. In the context of multi-step methods this condition is very strict as can be seen from the famous Second Dahlquist's Barrier:

Theorem 3.2. *An A-stable multi-step method must be of order $p \leqslant 2$.*

P r o o f. There are several proofs in the literature, see e.g. [19] or [36]. □

According to Theorem 3.2, it is suitable to study some milder stability properties for higher order multi-step methods. One of the options used most frequently is the so-called *$A(\alpha)$-stability* assuming the method is unconditionally stable for arbitrary λ lying in the sector $\{z \in \mathbb{C}\colon |\arg z| \leqslant \alpha\}$.

Lemma 3.2. *Let the stability region of the multi-step method S be a closed subset of $\overline{\mathbb{C}}$. Then there exists a constant $C_1 > 0$ independent of z such that the sequence*

$\{\gamma_j\}_{j\geqslant 0}$ defined by (3.8) *and by initial conditions* $\gamma_0(z) = 1/a_k(z)$, $\gamma_j(z) = 0$, $j \leqslant -1$, *is bounded by*

$$|\gamma_j(z)| \leqslant C \quad \forall\, z \in S,\ |z| \leqslant C_1,\ j \geqslant 0, \tag{3.15}$$

$$|\gamma_j(z)| \leqslant \frac{C}{1+|z|} \quad \forall\, z \in S,\ |z| \geqslant C_1,\ j \geqslant 0. \tag{3.16}$$

The constant C is independent of j and z.

P r o o f. The proof follows the ideas presented in [9]. Since the complete proof is quite long and technical we skip it. □

Now, we are going back to the global error estimate. Let us assume the multi-step method of order p is stable for given τ and λ, i.e. $\tau\lambda \in S$. Then the global error estimate is a direct consequence of (3.9) and Lemma 3.2, since all the terms in (3.9) containing γ_j or $\gamma_j a_v$ can be estimated uniformly by some constant C:

$$|e^{n+k}| \leqslant C\Big(\sum_{s=0}^{k-1}|e^s| + \sum_{s=1}^{n}|\mathrm{loc}^s|\Big) \leqslant C\Big(\sum_{s=0}^{k-1}|e^s| + \tau^p\Big). \tag{3.17}$$

Using the results in [9] and [27] we can generalize the ideas from Lemma 3.2 which are suited for Dahlquist's equation (1.7) to the original parabolic problem (1.1).

Theorem 3.3. *Let the regularity of the exact solution of* (1.1) *be*

$$u \in W^{p+1,\infty}(0,T,H) \cap W^{1,\infty}(0,T,V). \tag{3.18}$$

Let the multi-step method be $A(\alpha)$-stable and strictly stable at zero and at infinity. Let the multi-step method be of order $p \geqslant 1$. Let the eigenvalues λ_i of $\mathcal{A}_h$ lie in the sector $\{z \in \mathbb{C}\colon\ |\arg z| \leqslant \alpha'\}$, $\alpha' < \alpha$. Then there exists a constant $C > 0$ such that

$$\|U^m - u^m\| \leqslant C\Big(\mathrm{err}(h) + \tau^p + \sum_{s=0}^{k-1}\|U^s - u^s\|\Big), \tag{3.19}$$

where the constant C depends on T and α', but is independent of m and $\mathcal{A}$.

3.3. Examples of multi-step methods. In this section we present examples of multi-step methods, namely Adams methods and BDF.

3.3.1. Adams methods. Adams methods can be divided into explicit Adams-Bashforth methods and implicit Adams-Moulton methods. Adams methods in general are multi-step methods characterized by the choice $\alpha_k = 1$, $\alpha_{k-1} = -1$, and the remaining $\alpha_v = 0$. The usual interpretation of this choice is

$$u_h^{m+k} - u_h^{m+k-1} = \int_{I_{m+k}} u_h' = \int_{I_{m+k}} f_h - \mathcal{A}_h u_h \approx \tau \sum_{v=0}^{k} \beta_v (f_h^{m+v} - \mathcal{A}_h u_h^{m+v}), \tag{3.20}$$

where the last integral on the right-hand side is difficult to evaluate exactly and therefore, it is evaluated by an interpolation quadrature based on the values of u_h at t_{m+v}, $v = 0, \ldots, k-1$ (explicit variant) and $v = 0, \ldots, k$ (implicit variant).

For $k = 1$ Adams methods become the familiar forward Euler method and the Crank-Nicolson method. The coefficients for $k = 1, 2, 3$ are displayed in Table 4.

	Adams-Bashforth			Adams-Moulton		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
β_0	1	$-\frac{1}{2}$	$\frac{5}{12}$	$\frac{1}{2}$	$-\frac{1}{12}$	$\frac{1}{24}$
β_1	0	$\frac{3}{2}$	$-\frac{4}{3}$	$\frac{1}{2}$	$\frac{2}{3}$	$-\frac{5}{24}$
β_2	–	0	$\frac{23}{12}$	–	$\frac{5}{12}$	$\frac{19}{24}$
β_3	–	–	0	–	–	$\frac{3}{8}$

Table 4. Coefficients of Adams-Bashforth and Adams-Moulton methods $k = 1, 2, 3$.

The local order of convergence is directly connected with the order of interpolation that directly depends on the number of interpolation nodes. Using this idea, it is possible to determine that the order of k-step Adams-Bashforth methods is k and the order of k-step Adams-Moulton methods is $k + 1$.

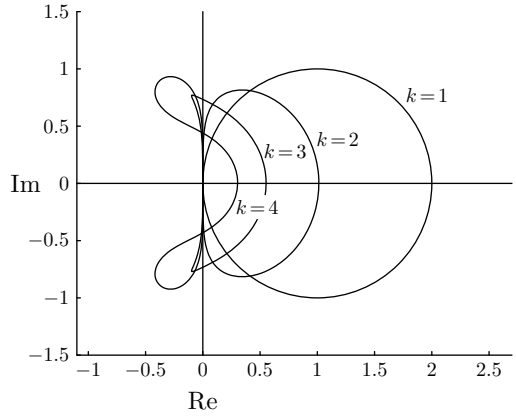

Figure 4. Stability regions of Adams-Bashforth methods $k = 1, 2, 3, 4$.

From Figure 4 and Figure 5 it is possible to see that the Adams methods with the exception of the Crank-Nicolson method (Adams-Moulton, $k = 1$) have bounded stability regions and therefore, these methods are not very useful for stiff problems.

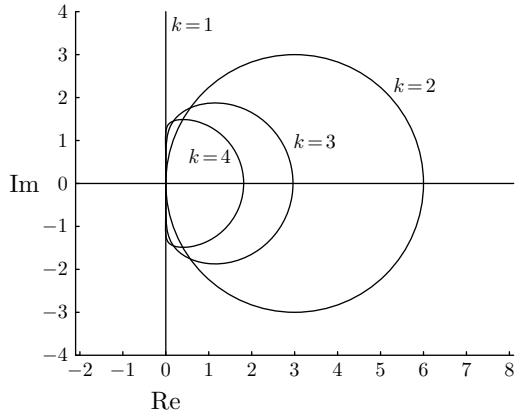

Figure 5. Stability regions of Adams-Moulton methods $k = 1, 2, 3, 4$.

3.3.2. BDF is the often used abbreviation for *Backward Differentiation Formulae*. These methods were introduced in [10] and since the work [16], they were often considered as the *method of the first choice* for solving stiff ordinary differential equations. In the past, the main reason for this was that even more robust implicit Runge-Kutta methods were too expensive for practical use. In recent years the situation is changing and the preference of BDF methods is not so obvious. For a comparison of BDF methods and discontinuous Galerkin methods, see e.g. [13].

BDF methods are characterized by the choice of coefficients $\beta_k = 1$ and the remaining coefficients $\beta_v = 0$. The coefficients α_v can be interpreted as the coefficients of the higher order backward difference

$$\sum_{v=0}^{k} \alpha_v u_h^{m+v} \approx \tau u_h'(t_{m+k}) = \tau f_h^{m+k} - \tau \mathcal{A}_h u_h^{m+k}. \tag{3.21}$$

It is possible to derive an explicit formula for the BDF coefficients

$$\begin{aligned} \alpha_k &= \sum_{v=0}^{k} \frac{1}{v}, \\ \alpha_v &= (-1)^{k-v} \binom{k}{v} \frac{1}{k-v}, \quad v = 0, \dots, k-1, \end{aligned} \tag{3.22}$$

see [34]. The BDF method for $k = 1$ is in fact the backward Euler method. Substituting the coefficients (3.22) into the order conditions (3.5), it can be shown that the k-step BDF method has local order k.

Theorem 3.4. *BDF methods are D-stable if and only if* $k \leqslant 6$.

P r o o f. The proof can be found in [19] or [21]. □

According to Theorem 3.4, BDF methods are applicable only up to the order 6. Figure 6 shows that BDF methods are A-stable for $k = 1, 2$, which is in a good agreement with the Second Dahlquist's Barrier. The remaining BDF methods ($k = 3, 4, 5, 6$) are A(α)-stable only. The corresponding angle of stability is in Table 5.

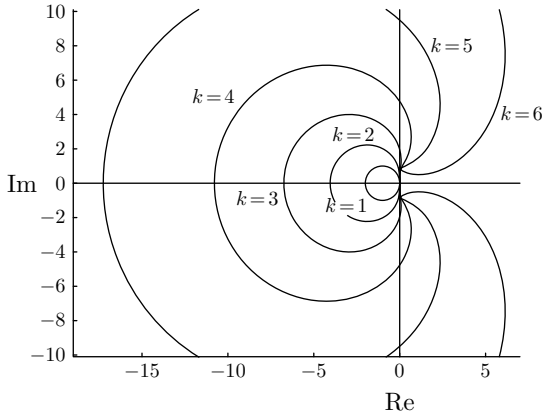

Figure 6. Stability regions of BDF $k = 1, 2, 3, 4, 5, 6$.

k	1	2	3	4	5	6
α	90°	90°	86.03°	73.35°	51.84°	17.84°

Table 5. The angles of stability of BDF methods.

References

[1] *G. Akrivis, C. Makridakis, R. H. Nochetto*: Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. Numer. Math. *118* (2011), 429–456. zbl MR doi

[2] *R. Alexander*: Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s. SIAM J. Numer. Anal. *14* (1977), 1006–1021. zbl MR doi

[3] *D. Boffi*: Finite element approximation of eigenvalue problems. Acta Numer. *19* (2010), 1–120. zbl MR doi

[4] *P. Brenner, M. Crouzeix, V. Thomée*: Single-step methods for inhomogeneous linear differential equations in Banach space. RAIRO, Anal. Numér. *16* (1982), 5–26. zbl MR doi

[5] *J. C. Butcher*: Implicit Runge-Kutta processes. Math. Comput. *18* (1964), 50–64. zbl MR doi

[6] *E. A. Coddington, N. Levinson*: Theory of Ordinary Differential Equations. McGraw-Hill Book Company, New York, 1955. zbl MR

[7] *M. Crouzeix, W. H. Hundsdorfer, M. N. Spijker*: On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods. BIT *23* (1983), 84–91. zbl MR doi

[8] *M. Crouzeix, F. J. Lisbona*: The convergence of variable-stepsize, variable-formula, multistep methods. SIAM J. Numer. Anal. *21* (1984), 512–534. zbl MR doi

[9] *M. Crouzeix, P.-A. Raviart*: Approximation des équations d'évolution linéaires par des méthodes à pas multiples. C. R. Acad. Sci., Paris, Sér. A *283* (1976), 367–370. zbl MR
[10] *C. F. Curtiss, J. O. Hirschfelder*: Integration of stiff equations. Proc. Natl. Acad. Sci. USA *38* (1952), 235–243. zbl MR doi
[11] *G. Dahlquist*: Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. *4* (1956), 33–53. zbl MR doi
[12] *K. Dekker*: On the iteration error in algebraically stable Runge-Kutta methods. Report NW 138/82, Math. Centrum, Amsterdam, 1982.
[13] *V. Dolejší, M. Feistauer*: Discontinuous Galerkin Method. Analysis and Applications to Compressible Flow. Springer Series in Computational Mathematics 48, Springer, Cham, 2015. zbl MR doi
[14] *B. L. Ehle*: On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems. Thesis (Ph.D)-University of Waterloo, Ontario, 1969. MR
[15] *R. Frank, J. Schneid, C. W. Ueberhuber*: Order results for implicit Runge-Kutta methods applied to stiff systems. SIAM J. Numer. Anal. *22* (1985), 515–534. zbl MR doi
[16] *C. W. Gear*: Numerical Initial Value Problems in Ordinary Differential Equations. Prentice-Hall, Englewood Cliffs, 1971. zbl MR
[17] *C. W. Gear, K. W. Tu*: The effect of variable mesh size on the stability of multistep methods. SIAM J. Numer. Anal. *11* (1974), 1025–1043. zbl MR doi
[18] *R. D. Grigorieff*: Stability of multistep-methods on variable grids. Numer. Math. *42* (1983), 359–377. zbl MR doi
[19] *R. D. Grigorieff, H. J. Pfeiffer*: Numerik gewöhnlicher Differentialgleichungen. Band 2: Mehrschrittverfahren. Teubner Studienbücher: Mathematik. B. G. Teubner, Stuttgart, 1977. zbl MR doi
[20] *A. Guillou, J. L. Soulé*: La résolution numérique des problèmes différentiels aux conditions initiales par des méthodes de collocation. Rev. Franç. Inform. Rech. Opér. *3* (1969), 17–44. zbl MR doi
[21] *E. Hairer, S. P. Nørsett, G. Wanner*: Solving Ordinary Differential Equations. I: Nonstiff Problems. Springer Series in Computational Mathematics 8, Springer, Berlin, 1993. zbl MR doi
[22] *E. Hairer, G. Wanner*: Solving Ordinary Differential Equations. II: Stiff and Differential-Algebraic Problems. Springer Series in Computational Mathematics 14, Springer, Berlin, 1996. zbl MR doi
[23] *P. Henrici*: Discrete Variable Methods in Ordinary Differential Equations. John Wiley and Sons, New York, 1962. zbl MR
[24] *M. Hochbruck, A. Ostermann*: Exponential integrators. Acta Numerica *19* (2010), 209–286. zbl MR doi
[25] *B. L. Hulme*: One-step piecewise polynomial Galerkin methods for initial value problems. Math. Comput. *26* (1972), 415–426. zbl MR doi
[26] *J. Kuntzmann*: Neuere Entwicklungen der Methode von Runge und Kutta. Z. Angew. Math. Mech. *41* (1961), T28–T31. zbl doi
[27] *C. Lubich*: On the convergence of multistep methods for nonlinear stiff differential equations. Numer. Math. *58* (1991), 839–853. zbl MR doi
[28] *H. Padé*: Sur la représentation approchée d'une fonction par des fractions rationnelles. Ann. Sci. Éc. Norm. Supér. (3) *9* (1892), 3–93. zbl MR
[29] *A. Prothero, A. Robinson*: On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. Math. Comput. *28* (1974), 145–162. zbl MR doi
[30] *K. Rektorys*: The Method of Discretization in Time and Partial Differential Equations. Mathematics and Its Applications (East European Series) 4, D. Reidel Publishing, Dordrecht; SNTL-Publishers of Technical Literature, Praha, 1982. zbl MR

[31] *F. Roskovec*: Numerical solution of nonlinear convection-diffusion problems by adaptive methods. Master Thesis, 2014. (In Czech.)
[32] *E. F. Toro*: Riemann Solvers and Numerical Methods for Fluid Dynamics. A Practical Introduction, Springer, Berlin. 1999. zbl MR doi
[33] *M. Vlasák, V. Dolejší, J. Hájek*: A priori error estimates of an extrapolated space-time discontinuous Galerkin method for nonlinear convection-diffusion problems. Numer. Methods Partial Differ. Equations *27* (2011), 1456–1482. zbl MR doi
[34] *M. Vlasák, Z. Vlasáková*: Derivation of BDF coefficients for equidistant time step. Programs and Algorithms of Numerical Mathematics 15. Proc. Seminar, Dolní Maxov, Academy of Sciences of the Czech Republic, Institute of Mathematics, Praha, 2010, pp. 221–226. zbl MR
[35] *J. von Neumann*: Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. Math. Nachr. *4* (1951), 258–281. zbl MR doi
[36] *O. B. Widlund*: A note on unconditionally stable linear multistep methods. BIT, Nord. Tidskr. Inf.-behandl. *7* (1967), 65–70. zbl MR doi
[37] *K. Wright*: Some relationships between implicit Runge-Kutta, collocation Lanczos τ methods, and their stability properties. BIT, Nord. Tidskr. Inf.-behandl. *10* (1970), 217–227. zbl MR doi

Author's address: *Miloslav Vlasák*, Charles University, Faculty of Mathematics and Physics, Department of Mathematical Analysis, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: `vlasak@karlin.mff.cuni.cz`.