Lecture 12

Rational Approximations

In the previous lectures we have seen some examples for time discretisation methods, e.g., the explicit and implicit Euler methods, the Crank–Nicolson scheme, and the Radau II A method. We now turn to study numerical approximation schemes $F : [0, \infty) \to \mathscr{L}(X)$ (see Lecture 4) that are defined by means of a rational function r:

$$F(h) = r(hA).$$

Such were the previously mentioned time discretisation methods. In general, we first need to give meaning to the expression $r(hA)$, and—in view of the Lax equivalence theorem, Theorem 4.6—to study consistency and stability of these schemes. We start with the scalar case and make the following definition. Let $r : \mathbb{C} \to \mathbb{C}$ be a rational function, i.e., $r = \frac{P}{Q}$ with P, Q polynomials and $Q \neq 0$, and even if it is not stated we shall usually suppose that P and Q have no common zeros.

Definition 12.1. We call a rational function r a rational approximation of the exponential function of order p, **rational approximation of order** p for short, if there are constants $C, \delta > 0$ such that

$$|r(z) - \mathrm{e}^z| \leq C|z|^{p+1} \quad \text{for all } z \in \mathbb{C} \text{ with } |z| \leq \delta.$$

12.1 The scalar case

Recall the test equation, already seen in Section 1.1 and Appendix B, with the unknown function $u : [0, \infty) \to \mathbb{C}$:

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t} u(t) = \lambda u(t), & t > 0 \\ u(0) = u_0, \end{cases} \tag{12.1}$$

where the parameter $\lambda \in \mathbb{C}$ and the initial value $u_0 \in \mathbb{C}$ are given. Of course, the exact solution of problem (12.1) equals $u(t) = \mathrm{e}^{t\lambda} u_0$.

1. Consistency

That r is a rational approximation of order p means by definition that $F(h) := r(h\lambda)$ is a *finite difference scheme (method)* consistent of order p on $X = \mathbb{R}$ with the Cauchy problem (12.1), cf. Definition 4.11. By considering the power series expansion around $z = 0$ one sees immediately that p-order consistency in this case is equivalent to the conditions

$$r(0) = 1,\ r'(0) = 1,\ \ldots,\ r^{(p)}(0) = \exp^{(p)}(0).$$

This yields the next example.

Example 12.2. Consider

$$r(z) := 1 + z + \frac{z^2}{2!} + ... \cdots + \frac{z^s}{s!}.$$

Trivially, r is a rational approximation of order $p = s$. All **explicit s-stage Runge–Kutta methods** of order $p = s$ possess this stability function, see also Example 12.3 below.

If we apply the time discretisation method $F(h) = r(h\lambda)$ to obtain the numerical solution $u_{h,n}$ after n steps with time step h we are led to the recursion

$$u_{h,n} = r(h\lambda)u_{h,n-1}, \quad n \in \mathbb{N}.$$

Next we show that Runge–Kutta methods are of this form.

Example 12.3 (**Runge–Kutta methods**)**.** Recall the s-stage Runge–Kutta method from Appendix B given by the recursion (B.13), (B.14). For the special right-hand side of problem (12.1) we then have:

$$u_{h,n} = u_{h,n-1} + h\lambda \sum_{i=1}^{s} b_i k_i \tag{12.2}$$

with

$$k_i = u_{h,n-1} + h\lambda \sum_{j=1}^{s} a_{ij} k_j \tag{12.3}$$

with certain coefficients a_{ij}, b_i for $i, j = 1, ..., s$. We introduce the following vectors in $\mathbb{R}^s$:

$$\mathbf{k} = (k_1, ..., k_s)^\top, \quad \mathbf{1} = (1, ..., 1)^\top, \quad \mathbf{b} = (b_1, ..., b_s)^\top,$$

and the matrix $\mathbf{A} = (a_{ij})_{i,j=1,...,s} \in \mathbb{R}^{s\times s}$. Then formulae (12.2) and (12.3) can be written as

$$u_{h,n} = u_{h,n-1} + z\mathbf{b}^\top k$$

and

$$\mathbf{k} = (1 - z\mathbf{A})^{-1}\mathbf{1}u_{h,n-1} \tag{12.4}$$

with $z = h\lambda \in \mathbb{C}$. This implies for all $n \in \mathbb{N}$ that

$$u_{h,n} = u_{h,n-1} + z\mathbf{b}^\top(I - z\mathbf{A})^{-1}\mathbf{1}u_{h,n-1} = \big(1 + z\mathbf{b}^\top(I - z\mathbf{A})^{-1}\mathbf{1}\big)u_{h,n-1}, \tag{12.5}$$

that is, we obtain $u_{h,n} = r(z)u_{h,n-1}$ with $r(z) = 1 + z\mathbf{b}^\top(I - z\mathbf{A})^{-1}\mathbf{1}$ which is a rational function of $z \in \mathbb{C}$. To work out the details of the computations above is left as Exercise 1.

Example 12.4. All the time discretisation methods introduced previously are Runge–Kutta methods. Therefore, the corresponding rational function can be obtained by the derivation in Example 12.3.

explicit Euler method: $r(z) = 1 + z$

implicit Euler method: $r(z) = \dfrac{1}{1 - z}$

Crank–Nicolson scheme: $r(z) = \dfrac{1 + \frac{z}{2}}{1 - \frac{z}{2}}$

Radau II A method: $r(z) = \dfrac{1 + \frac{2}{5}z + \frac{1}{10}\frac{z^2}{2}}{1 - \frac{3}{5}z + \frac{3}{10}\frac{z^2}{2} - \frac{1}{10}\frac{z^3}{6}}.$

Now let us return to general rational functions

$$r(z) = \frac{P(z)}{Q(z)},$$

with $k = \deg(P)$ and $l = \deg(Q)$, where we suppose that P and Q have no common zeros. If $k, l \in \mathbb{N}_0$ are fixed, the maximal order of approximation to the exponential function is $p = k + l$, see Exercise 5. Such rational approximations r are called **rational Padé approximations**. One can collect the corresponding functions r in the **Padé tableau**, see Table 12.1 for examples.

$l\backslash k$	0	1	2
0	$\frac{1}{1}$	$\frac{1+z}{1}$	$\frac{1+z+\frac{z^2}{2!}}{1}$
1	$\frac{1}{1-z}$	$\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$	$\frac{1+\frac{2}{3}z+\frac{2}{3}\frac{z^2}{2!}}{1-\frac{1}{3}z}$
2	$\frac{1}{1-z+\frac{z^2}{2!}}$	$\frac{1+\frac{1}{3}z}{1-\frac{1}{3}z+\frac{1}{3}\frac{z^2}{2!}}$	$\frac{1+\frac{1}{2}z+\frac{1}{6}\frac{z^2}{2!}}{1-\frac{1}{2}z+\frac{1}{6}\frac{z^2}{2!}}$
3	$\frac{1}{1-z+\frac{z^2}{2!}-\frac{z^3}{3!}}$	$\frac{1+\frac{1}{4}z}{1-\frac{1}{4}z+\frac{1}{2}\frac{z^2}{2!}-\frac{1}{4}\frac{z^3}{3!}}$	$\frac{1+\frac{2}{5}z+\frac{1}{10}\frac{z^2}{2!}}{1-\frac{3}{5}z+\frac{3}{10}\frac{z^2}{2!}-\frac{1}{10}\frac{z^3}{3!}}$

Table 12.1: Padé tableau.

2. Stability issues

As discussed in Lecture 4, stability is fundamental if one longs for convergence of the method for all initial values. More precisely, since $u_{h,n} = \big(r(h\lambda)\big)^n u_0$ is expected to be the approximation of the exact solution $u(t) = \mathrm{e}^{t\lambda}u_0$ at time $t = nh$, the recursion $u_{h,n} = r(h\lambda)u_{h,n-1}$ needs to be stable. This motivates the next definition. The set

$$S = S(r) = \big\{z \in \mathbb{C} : |r(z)| \le 1\big\}$$

is called the **stability region** of the corresponding rational approximation. Also note that, if one starts, say with some Runge–Kutta method as in Example 12.3, and derives a formula for the recursion, the appearing rational function r determines the stability of the method. Hence the rational function is also called **stability function**.

Example 12.5. Consider the following time discretisation methods, their stability functions, and stability regions.

1. For the explicit Euler method we have $r_1(z) = 1 + z$, which implies

$$S(r_1) = \{z \in \mathbb{C} : |1 + z| \le 1\}$$

the closed disc of radius 1, centred at the point -1.

2. The implicit Euler method has stability function $r_2(z) = \frac{1}{1-z}$, hence,

$$S(r_2) = \{z \in \mathbb{C} : |1 - z| \ge 1\},$$

which is the exterior of the circle with radius 1 and centre 1.

3. The stability function of the Crank–Nicolson scheme is $r_3(z) = \frac{1+\frac{z}{2}}{1-\frac{z}{2}}$, therefore,

$$S(r_3) = \{z \in \mathbb{C} : \operatorname{Re} z \le 0\},$$

i.e., the left half-plane.

The respective stability regions are shown in Figure 12.1.

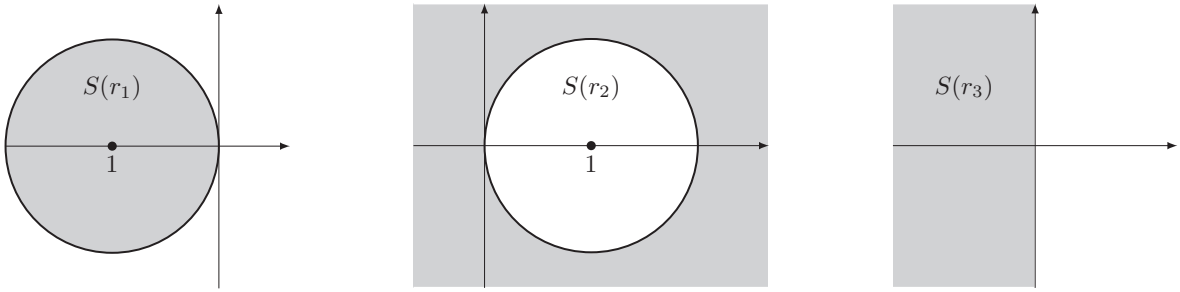

Figure 12.1: Stability regions: explicit Euler, implicit Euler, Crank–Nicolson methods.

From the examples above one can see that the stability of the recursion is not obvious. There is a restrictive condition on $z = h\lambda$. To achieve a stable recursion $u_{h,n} = r(z)u_{h,n-1}$, $n \in \mathbb{N}$, the complex number $z = h\lambda$ has to lie in the stability region S of the method. Since $\lambda \in \mathbb{C}$ is a given parameter in problem (12.6), this yields a condition on the step size h. Thus, if $\operatorname{Re}\lambda \le 0$, the explicit Euler method is not unconditionally stable in contrast to the the implicit Euler or Crank–Nicolson schemes (cf. Section B.1).

The rational approximations with stability region containing the entire left half-plane are called ***A*-stable**[1]. If the stability region contains a sector

$$\overline{\curlywedge}_\alpha = \{z \in \mathbb{C} : |\arg(-z)| \le \alpha\} = -\overline{\Sigma}_\alpha,$$

for some $\alpha \in [0, \frac{\pi}{2}]$, we speak about ***A*($\alpha$)-stability**. (For $\alpha = 0$ we set $\curlywedge_0 = (-\infty, 0)$ and $\Sigma_0 = (0, \infty)$.) Notice that A-stability is the same as $A(\frac{\pi}{2})$-stability. For example, the implicit Euler or Crank–Nicolson schemes are A-stable.

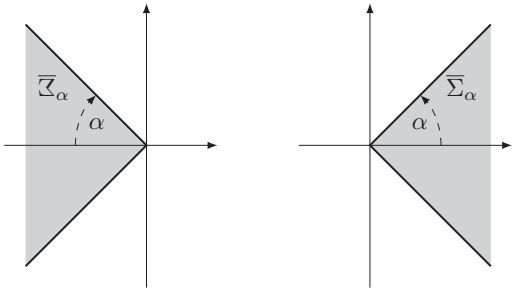

Figure 12.2: The sector Σ_α and its reflection.

[1]The terminology is due to G. Dahlquist. According to Hairer, Nørsett and Wanner he said: 'I didn't like all these "strong", "perfect", "absolute", "generalized", "super", "hyper", "complete" and so on in mathematical definitions, I wanted something neutral; and having been impressed by David Young's "property A", I chose the term "A-stable".'

Without proof we state an important stability property of Padé approximations first conjectured by Ehle[2], and then proved by Wanner, Hairer, and Nørsett.[3]

Theorem 12.6. *A rational Padé approximation $r = \frac{P}{Q}$ is A-stable if and only if*

$$\deg(Q) - 2 \leq \deg(P) \leq \deg(Q).$$

That is, only those Padé approximations are A-stable whose stability functions appear in the main diagonal or in the first or the second lower sub-diagonal of the Padé tableau.

Remark 12.7. Let $r = \frac{P}{Q}$ be an A-stable rational Padé approximation. If r is diagonal, i.e., $\deg(P) = \deg(Q)$, then $r(\infty) = 1$, otherwise $r(\infty) = 0$.

3. Convergence

The first motivating result shows that $A(\alpha)$-stable rational approximations converge in the scalar case. However, this result will be fundamental later, when we pass to rational approximation of analytic semigroups.

Proposition 12.8. *Let $\alpha \in (0, \frac{\pi}{2}]$, and let r be an $A(\alpha)$-stable rational approximation of order $p \in \mathbb{N}$. Then for all $\theta \in (0, \alpha)$ and $\varepsilon_0 > 0$ there exist constants $C, c \geq 0$ such that*

$$|r(z)^n - \mathrm{e}^{nz}| \leq Cn|z|^{p+1}\mathrm{e}^{-nc|z|} \quad \textit{holds for all } z \in \overline{\Sigma}_\theta \textit{ with } |z| \leq h_0, \textit{ and for all } n \in \mathbb{N}.$$

In particular, we have for all $\lambda \in \Sigma_\alpha$ a constant $K > 0$ such that

$$|r(h\lambda)^n - \mathrm{e}^{t\lambda}| \leq Kn|h|^{p+1}\mathrm{e}^{-nhc} = K\mathrm{e}^{-tc}\frac{t^p}{n^p} \quad (t = nh)$$

for all $h \geq 0$ and $n \in \mathbb{N}$, i.e., the method is convergent of order p.

Proof. First of all, let us fix $C' \geq 0$ so that

$$|r(z) - \mathrm{e}^z| \leq C'|z|^{p+1} \quad \text{for all } z \in \overline{\Sigma}_\alpha \text{ with } |z| \leq h_0.$$

This is possible by the assumption about the approximation order. Note that

$$|\mathrm{e}^z| = \mathrm{e}^{\operatorname{Re} z} \leq \mathrm{e}^{-|z|\cos(\theta)} \quad \text{holds for all } z \in \overline{\Sigma}_\theta.$$

We next claim that for some $c' > 0$ the inequality

$$|r(z)| \leq \mathrm{e}^{-c'|z|} \quad \text{holds for all } z \in \overline{\Sigma}_\theta \text{ with } |z| \leq h_0.$$

We argue by contradiction and assume the contrary, i.e., that for all $n \in \mathbb{N}$ there is $z_n \in \overline{\Sigma}_\theta$ with $|z_n| \leq h_0$ such that

$$|r(z_n)| > \mathrm{e}^{-\frac{|z_n|}{n}}.$$

By passing to a subsequence we may assume that $(z_n) \subseteq \overline{\Sigma}_\theta \cap \overline{\mathrm{B}}(0, h_0)$ is convergent to a limit $z \in \overline{\Sigma}_\theta \cap \overline{\mathrm{B}}(0, h_0)$. Then we obtain $|r(z)| \geq 1$ and the $A(\alpha)$-stability yields $|r(z)| = 1$. By the

[2] B. L. Ehle, "A-stable methods and Padé approximations to the exponential," SIAM J. Math. Anal. **4** (1973), 671–680.

[3] G. Wanner, E. Hairer and S. P. Nørsett, "Order stars and stability theorems," BIT Num. Math. **18** (1978), 475–489.

maximum principle for the modulus of holomorphic functions (applied to r on $-\mathfrak{T}_\alpha$) we obtain that $z = 0$. Thus we conclude

$$\mathrm{e}^{-\frac{|z_n|}{n}} \leq |r(z_n)| \leq |r(z_n) - \mathrm{e}^{z_n}| + |\mathrm{e}^{z_n}| \leq C'|z_n|^{p+1} + \mathrm{e}^{-|z_n|\cos(\theta)}.$$

Therefore

$$\tfrac{1}{|z_n|}\Big(\mathrm{e}^{|z_n|(\cos(\theta)-\frac{1}{n})} - 1\Big) \leq C'|z_n|^p \mathrm{e}^{|z_n|\cos(\theta)} \to 0$$

as $n \to \infty$. This yields, however, a contradiction.

We obtain therefore the existence of a $c' > 0$ asserted above, and we set $c := \min\{c', \cos(\theta)\}$. By the standard telescopic identity we conclude

$$\begin{aligned}\big|r(z)^n - \mathrm{e}^{nz}\big| &\leq \big|r(z) - \mathrm{e}^z\big| \sum_{j=0}^{n-1} \big|r(z)\big|^j \big|\mathrm{e}^{(n-j-1)z}\big| \leq \big|r(z) - \mathrm{e}^z\big| \sum_{j=0}^{n-1} \mathrm{e}^{-cj|z|}\mathrm{e}^{-c(n-j-1)|z|} \\ &\leq C'|z|^{p+1}\mathrm{e}^{-c(n-1)|z|} \leq C'\mathrm{e}^c|z|^{p+1}\mathrm{e}^{-nc|z|} = C|z|^{p+1}\mathrm{e}^{-nc|z|}\end{aligned}$$

for all $z \in \overline{\mathfrak{T}}_\theta$ with $|z| \leq h_0$. □

12.2 Rational functions of operators

Let A be a linear operator on a Banach space X with nonempty resolvent set. Given a rational function $r = \frac{P}{Q}$ we would like to define $r(A)$. First of all, we recall the case when $r = P$ is a polynomial. Suppose $P(z) = z^k$. In this case, as we have already seen, we set $D(A^0) = X$ and $A^0 = I$, and for $k \in \mathbb{N}$ we define

$$\begin{aligned}D(A^k) &:= \big\{f \in D(A^{k-1}) : A^{k-1}f \in D(A)\big\}, \\ A^k f &= AA^{k-1}f \quad \text{for } f \in D(A^k)\end{aligned}$$

by recursion. Then A^k is a closed operator for every $k \in \mathbb{N}_0$, cf. Exercise 4.1. For a general polynomial $P \neq 0$

$$P(z) = a_0 + a_1 z + a_z^2 + \ldots + a_k z^k$$

with $a_k \neq 0$ we set $D(P(A)) := D(A^k)$ and

$$P(A) := a_0 I + a_1 A + a_2 A^2 + \ldots + a_k A^k,$$

which is again a closed operator, see Exercise 3.

Of course, we can write

$$P(z) = a_k(z - z_1)^{m_1}(z - z_2)^{m_2} \cdots (z - z_n)^{m_n}$$

where $z_j \in \mathbb{C}$ are pairwise different. Thus we have the identity

$$P(A) = a_k(A - z_1)^{m_1}(A - z_2)^{m_2} \cdots (A - z_n)^{m_n}.$$

If $P \neq 0$ and the zeros of P all lie in $\rho(A)$ then $(A - z_j)$ are in particular all injective, and we obtain

$$P(A)^{-1} = \frac{1}{a_k}(A - z_1)^{-m_1} \cdots (A - z_n)^{-m_n} = \frac{(-1)^k}{a_k} R(z_1, A)^{m_1} \cdots R(z_n, A)^{m_n} \in \mathscr{L}(X).$$

It is easy to see that $\operatorname{ran}(P(A)^{-1}) = D(A^{\deg(P)})$.

Next we define

$$\mathcal{R}_A = \big\{ r = \tfrac{P}{Q} : P, Q \text{ are polynomials with } Q \text{ having all zeros in } \rho(A) \big\}$$

and for $r \in \mathcal{R}_A$, $r \neq 0$ we set

$$r(A) := P(A)Q(A)^{-1},$$

with

$$D(r(A)) = \big\{ f \in X : Q(A)^{-1} f \in D(P(A)) \big\}.$$

Then $r(A)$ is a closed operator by Exercise 7.1, and we have

$$D(r(A)) = \begin{cases} D\big(A^{\deg(P)-\deg(Q)}\big) & \text{if } \deg(P) \geq \deg(Q) \\ X & \text{otherwise.} \end{cases}$$

Note that $r(A)$ is well-defined, i.e., if $r = \frac{P_1}{Q_1} = \frac{P_2}{Q_2}$ with Q_1, Q_2 having zeros in $\rho(A)$ then

$$P_1(A)Q_1(A)^{-1} = P_2(A)Q_2(A)^{-1}.$$

Finally, let us recall the **partial fraction decomposition** of a rational function. If r is a rational function with poles z_i of order $\nu_i \in \mathbb{N}$, then there is a unique polynomial P_0 and coefficients $c_{ij} \in \mathbb{C}$ such that

$$r(z) = P_0(z) + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} \frac{c_{ij}}{(z - z_i)^j}.$$

This provides yet another evaluation of $r(A)$ for $r \in \mathcal{R}_A$:

$$r(A) = P_0(A) + \sum_{i=1}^{\nu} \sum_{j=1}^{\nu_i} (-1)^j c_{ij} R(z_i, A)^j.$$

We can now at last define $r(hA)$.

Definition 12.9. A finite difference scheme F such that $F(h) = r(hA)$ holds for $h \in [0, h_0]$ with some $h_0 > 0$ is called a **rational approximation scheme**.

We now have a simple **functional calculus** for $r \in \mathcal{R}_A$. For the study of its various algebraic and analytic properties we refer to the Appendix A.6 of the monograph[4] by M. Haase. However to obtain the convergence of the rational approximation method obtained from $r \in \mathcal{R}_A$ further structural properties of r and A are needed. This will be the subject of forthcoming lectures. In the present one we shall give some illustration of results that can be expected, in a situation that is quite near to the scalar case.

[4] M. Haase: The Functional Calculus for Sectorial Operators, vol. 169 of Operator Theory: Advances and Applications, Birkhäuser Basel, 2006.

12.3 Multiplication operators

In this section we consider multiplication operators and start by briefly recalling some results from Exercises 1.4, 7.2 and Examples 7.2, 9.6. Let $(m_n) \subseteq \mathbb{C}$ be a sequence, and consider the multiplication operator $A = M_m$ with maximal domain

$$D(M_m) := \big\{(x_k) \in \ell^2 : (m_k x_k) \in \ell^2\big\}.$$

The norm of M_m is

$$\|M_m\| = \sup_{k\in\mathbb{N}} |m_k|,$$

provided the latter expression is finite. Our standing assumption will be that

$$\big\{m_n : n \in \mathbb{N}\big\} \subseteq \overline{\mathcal{Z}}_\delta$$

holds for some $\delta \in [0, \frac{\pi}{2})$. Or in other words:

Assumption 12.10. Let $A = M_m$ be a multiplication operator with spectrum

$$\sigma(M_m) = \overline{\big\{m_n : n \in \mathbb{N}\big\}} \subseteq \overline{\mathcal{Z}}_\delta$$

for some $\delta \in [0, \frac{\pi}{2})$.

Under this condition $A = M_m$ generates an analytic contraction semigroup T given by

$$T(t) = \mathrm{e}^{tA} = M_{\mathrm{e}^{tm}}.$$

For $\beta \geq 0$ the fractional power $(-A)^\beta$ of $-A = -M_m$ is given by

$$(-A)^\beta = M_{(-m)^\beta} \quad \text{with maximal domain} \quad D(M_{(-m)^\beta}) = \big\{(x_n) \in \ell^2 : ((-m_k)^\beta x_k) \in \ell^2\big\}.$$

Consider now the abstract Cauchy problem

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}u(t) = Au(t), & t > 0 \\ u(0) = u_0 \end{cases} \tag{12.6}$$

on the Banach space $X = \ell^2$ with $u_0 \in D(A)$. We use some rational approximation schemes to obtain the numerical solution $u_{h,n}$ at time t, i.e., after n steps with time step $h = \frac{t}{n}$.

The first illustrating result states the stability of some suitable schemes and explains why A-stability may be relevant in the general situation of rational approximation schemes.

Proposition 12.11 (Stability theorem). *Suppose $A = M_m$ is as above, and let r be an $A(\delta)$-stable rational approximation. Then the estimate*

$$\big\|\big(r(hA)\big)^n\big\| \leq 1$$

holds for all $h > 0$ and $n \in \mathbb{N}$.

Proof. Note that for $z \in \overline{\mathcal{Z}}_\delta$ and $h \geq 0$ we have $zh \in \overline{\mathcal{Z}}_\delta$. Then by the preparatory remarks we have

$$\big\|\big(r(hA)\big)^n\big\| \leq \sup_{k\in\mathbb{N}} \big\|\big(r(hm_k)\big)^n f\big\| \leq \sup_{z\in\overline{\mathcal{Z}}_\delta} |r(z)| \leq 1. \qquad \square$$

$\square$

We can extend the convergence result from the scalar case as follows. Our inspiration is the paper by M.-N. Le Roux[5].

Theorem 12.12 (**Convergence theorem I.**). *Suppose $A = M_m$ is as above. Let r be the stability function of an $A(\alpha)$-stable rational approximation of order p with $|r(\infty)| < 1$ and $\alpha \in (\delta, \frac{\pi}{2}]$. Then there is a constant $K > 0$ such that*

$$\big\|r(hA)^n - \mathrm{e}^{tA}\big\| \le K\frac{h^p}{t^p} = \frac{K}{n^p} \qquad (t = nh)$$

holds for all $n \in \mathbb{N}$, $t \ge 0$, i.e., one has the convergence of the rational approximation method in the operator norm.

Proof. We have to estimate

$$\big\|r(hA)^n - \mathrm{e}^{tA}\big\| = \sup_{k\in\mathbb{N}}\big|r(hm_k)^n - \mathrm{e}^{tm_k}\big| = \sup_{k\in\mathbb{N}}\big|r(hm_k)^n - \mathrm{e}^{nhm_k}\big|.$$

Since $|r(\infty)| < 1$ we can choose $h_0 > 0$ so large that

$$\sup\big\{z \in \overline{\complement}_\delta : |z| \ge h_0\big\} =: r_0 < 1.$$

Suppose first $|hm_k| \le h_0$. Then by Proposition 12.8 we obtain that

$$\big|r(hm_k)^n - \mathrm{e}^{tm_k}\big| \le Cn|hm_k|^{p+1}\mathrm{e}^{-nhc|m_k|} = \frac{C}{n^p}|tm_k|^{p+1}\mathrm{e}^{-tc|m_k|} \le \frac{C'}{n^p}$$

for some constants $C', c > 0$. On the other hand, if $|hm_k| > h_0$, then

$$|r(hm_k)| \le r_0 < 1.$$

Therefore with some appropriate constant $C'' > 0$ we have

$$|r(hm_k)|^n \le r_0^n \le \frac{C''}{n^p}.$$

We also have

$$|\mathrm{e}^{nhm_k}| = \mathrm{e}^{nh\,\mathrm{Re}\,m_k} \le \mathrm{e}^{-nh\cos(\alpha)|m_k|} = \mathrm{e}^{-nh_0\cos(\alpha)} \le \frac{C'''}{n^p}.$$

Hence in case $|hm_k| \ge h_0$ we obtain

$$|r(hm_k)^n - \mathrm{e}^{tm_k}| \le \frac{C''' + C''}{n^p}.$$

This and the estimate in the first case finish the proof. □

The drawback of this result is that it tells nothing about the diagonal Padé approximations, e.g., about the Crank–Nicolson scheme. For "smooth" initial data u_0, however, we can recover convergence without the assumption $|r(\infty)| < 1$, hence the next result applies also to the missing case of diagonal Padé approximations.

[5]M.-N. Le Roux, "Semidiscretization in time for parabolic problems," Math. Comp. **147** (1979), 919–931.

Theorem 12.13 (**Convergence theorem II.**). *Suppose $A = M_m$ is as above. Let r be the stability function of an $A(\alpha)$-stable rational approximation of order p with $\alpha \in (\delta, \frac{\pi}{2}]$. Then for all $\beta \in (0, p]$ there is a constant $K > 0$ such that*

$$\|u_{h,n} - u(t)\| = \|r(hA)^n u_0 - \mathrm{e}^{tA} u_0\| \le K h^\beta \|(-A)^\beta u_0\| \qquad (t = nh)$$

holds for all $n \in \mathbb{N}$, $t \ge 0$ and $u_0 \in D\big((-A)^\beta\big)$.

Proof. We first estimate the term

$$\sup_{\substack{k\in\mathbb{N}\\ m_k\neq 0}} \big| r(hm_k)^n (-m_k)^{-\beta} - \mathrm{e}^{tm_k}(-m_k)^{-\beta}\big|.$$

As before we choose $h_0 > 0$ with

$$\sup\big\{ z \in \overline{\sphericalangle}_\delta : |z| \ge h_0 \big\} =: r_0 < 1.$$

If $0 < |hm_k| \le h_0$, we obtain by Proposition 12.8 that

$$\begin{aligned} |r(hm_k)^n(-m_k)^{-\beta} - \mathrm{e}^{tm_k}(-m_k)^{-\beta}| &\le Cnh^\beta |hm_k|^{p+1-\beta} \mathrm{e}^{-nhc|m_k|} \\ &= \frac{Ch^\beta}{n^{p-\beta}} |tm_k|^{p+1} \mathrm{e}^{-tc|m_k|} \le \frac{C' h^\beta}{n^{p-\beta}} = \frac{C' h^p}{t^{p-\beta}}. \end{aligned}$$

On the other hand, suppose $|hm_k| > h_0$. Then by the $A(\alpha)$-stability we obtain

$$\big| r(hm_k)(-m_k)^{-\beta} - \mathrm{e}^{tm_k}(-m_k)^{-\beta}\big| \le \frac{2}{|m_k|^\beta} \le \frac{2h^\beta}{h_0^\beta}.$$

Therefore, for all $k \in \mathbb{N}$ with $m_k \neq 0$ we obtain

$$\sup_{\substack{k\in\mathbb{N}\\ m_k\neq 0}} \big| r(hm_k)^n (-m_k)^{-\beta} - \mathrm{e}^{tm_k}(-m_k)^{-\beta}\big| \le C'' h^\beta.$$

We now can write

$$\begin{aligned} \|r(hA)^n u_0 - \mathrm{e}^{tA} u_0\|_2^2 &= \sum_{\substack{k\in\mathbb{N}\\ m_k\neq 0}} \big| r(hm_k)^n u_0(k) - \mathrm{e}^{tm_k} u_0(k)\big|^2 \\ &\le \sup_{\substack{k\in\mathbb{N}\\ m_k\neq 0}} \big| r(hm_k)^n - \mathrm{e}^{tm_k} u_0(k)\big|^2 \cdot \|(-A)^\beta u_0\|_2^2 \le C''^2 h^{2\beta} \cdot \|(-A)^\beta u_0\|_2^2. \end{aligned}$$

The assertion is proved. □

Remark 12.14. 1. Of course, it was only for the sake of convenience that we stated the result above on $X = \ell^2$ for suitable multiplication operators $A = M_m$. Essentially the same proofs work for the general setting: Let $(\Omega, \mathscr{A}, \mu)$ be a σ-finite measure space (Ω a nonempty set, $\mathscr{A}$ a σ-algebra, μ a measure), and consider the Banach space $X = \mathrm{L}^2(\Omega, \mathscr{A}, \mu) = \mathrm{L}^2(\Omega)$. Suppose $m : \Omega \to \mathbb{C}$ is a measurable function such that the essential range of m is contained in the sector $\overline{\sphericalangle}_\alpha$ for some $\alpha \in [0, \frac{\pi}{2})$. Here the **essential range** is defined by

$$\operatorname{essran}(m) := \big\{ z \in \mathbb{C} : m^{-1}(\mathrm{B}(z, \varepsilon)) \text{ has positive } \mu\text{-measure for all } \varepsilon > 0 \big\}.$$

The spectrum of M_m is precisely essran(m), hence $\Sigma_{\pi-\alpha} \subseteq \rho(A)$. The multiplication operator $A = M_m$ with maximal domain

$$D(M_m) := \big\{ f \in \mathrm{L}^2(\Omega) : mf \in \mathrm{L}^2(\Omega) \big\}$$

generates an analytic semigroup T given by

$$T(t) = M_{\mathrm{e}^{tm}}.$$

For $\beta \geq 0$ the fractional power $(-A)^\beta$ of $-A = M_{-m}$ is given by

$$(-A)^\beta = M_{(-m)^\beta} \quad \text{with maximal domain} \quad D(M_{(-m)^\beta}) = \big\{ f \in \mathrm{L}^2(\Omega) : (-m)^\beta f \in \mathrm{L}^2(\Omega) \big\}.$$

Now the analogues of the results from above can be stated and proved with just a bit more work.

2. By part 1. and the spectral theorem for self-adjoint operators, we see that the results in this section remain valid for non-positive self-adjoint operators A on arbitrary Hilbert spaces.

Exercises

1. Consider an s-stage Runge–Kutta method applied for the problem (12.6) and defined by formulea (12.2) and (12.3)

a) Derive the formulae (12.2), (12.3).

b) Derive the recursion (12.5).

c) Show that the stability function

$$r(z) = \big(1 + z\mathbf{b}^\top (I - z\mathbf{A})^{-1}\mathbf{1}\big)$$

from formula (12.5) is a rational function. That is, $r(z) = \frac{P(z)}{Q(z)}$ with

$$P(z) = \det(I - z\mathbf{A} + z\mathbf{1}\mathbf{b}^\top) \qquad \text{and} \qquad Q(z) = \det(I - z\mathbf{A}).$$

d) Show that if the Runge–Kutta method is of order p, its stability function has the form

$$r(z) = 1 + z + \frac{z^2}{2!} + ... + \frac{z^p}{p!} + \mathcal{O}(z^{p+1}).$$

2. Prove directly that the Crank–Nicolson approximation is A-stable, cf. Example 12.5.

3. Work out the details of Section 12.2.

4. Prove the existence of a partial fraction decomposition for a rational function. *Hint: use complex analysis.*

5. Let $r(z) = \frac{P(z)}{Q(z)}$ with $k = \deg(P)$ and $l = \deg(Q)$, where we suppose that P and Q have no common zeros. Show that if $k, l \in \mathbb{N}_0$ are fixed, the maximal order of approximation to the exponential function is $p = k + l$.

6. Convince yourself about the details of Remark 12.14.