Applications of Mathematics

Emil Vitásek

Approximate solution of an inhomogeneous abstract differential equation

Applications of Mathematics, Vol. 57 (2012), No. 1, 31–41

Persistent URL: http://dml.cz/dmlcz/141817

APPROXIMATE SOLUTION OF AN INHOMOGENEOUS ABSTRACT DIFFERENTIAL EQUATION*

Emil Vitásek, Praha

(Received November 19, 2009)

Abstract. Recently, we have developed the necessary and sufficient conditions under which a rational function $F(hA)$ approximates the semigroup of operators $\exp(tA)$ generated by an infinitesimal operator A. The present paper extends these results to an inhomogeneous equation $u'(t) = Au(t) + f(t)$.

Keywords: abstract differential equations, semigroups of operators, rational approximations, A-stability

MSC 2010: 34K30, 34G10, 35K90, 47D03

1. Preliminaries

Let $\mathcal{X}$ be a (complex) Banach space and let A be an infinitesimal generator of a continuous semigroup of operators $U(t)$, $t \in [0, T]$ (for the relevant literature see, for example, [1], [2], [4], [6]).

Let F be a rational function with poles in the right half-plane of the complex plane and let it be regular at infinity. Further, let the coefficients of the polynomials in the numerator and denominator of F be real and let F approximate the exponential function with order p, i.e., let

$$\exp(z) = F(z) + O(z^{p+1}) \quad \text{for } z \to 0, \tag{1.1}$$

where p is a positive integer.

The approximation of the given semigroup $U(t)$ will be meant in the following sense: Divide the interval $[0, T]$ into N subintervals $[t_j, t_{j+1}]$ of the length $h = T/N$

* The research was supported by the Academy of Sciences of the Czech Republic, Institutional Research plan No. AV0Z10190503.

by mesh points $0 = t_0 < t_1 < \ldots < t_N = T$ and define the sequence $\{u_j,\ j = 0, 1, \ldots, N\} \subset \mathcal{X}$ by the recurrence

$$u_{j+1} = F(hA)u_j, \quad j = 0, \ldots, N-1, \ \ u_0 = \eta. \tag{1.2}$$

Further, suppose that

$$\lim_{\substack{h \to 0 \\ jh \to t}} u_j = U(t)\eta$$

holds for any $\eta \in \mathcal{X}$ and any $t \in [0, T]$. Then we say that the rational function F approximates the semigroup $U(t)$ or, alternatively, that the method (1.2) is convergent on the class of abstract differential equations of the form

$$u'(t) = Au(t), \quad t \in [0, T], \tag{1.3}$$

with the initial condition

$$u(0) = \eta \in \mathcal{X}. \tag{1.4}$$

The following theorem was proved in [5].

Theorem 1.1. *A rational function F with its poles in the right half-plane, regular at infinity and satisfying* (1.1) *with some $p \geqslant 1$ generates the convergent method* (1.2) *if and only if there exists a constant $M = M(t)$ such that*

$$\|F^j(hA)\| \leqslant M \tag{1.5}$$

for any sufficiently small h and for any j satisfying $0 \leqslant jh \leqslant t$. Moreover, if $\eta \in \mathcal{D}(A^{p+1})$ then the convergence is of order h^p.

The aim of this paper is to generalize these results to the case of a nonhomogeneous equation.

2. Main result

Let us investigate the differential equation of the form

$$u'(t) = Au(t) + f(t), \quad t \in [0, T], \tag{2.1}$$

with the initial condition

$$u(0) = \eta \in \mathcal{X}. \tag{2.2}$$

Here, A is an infinitesimal generator of a continuous semigroup of operators $U(t)$ as in Section 1 and the function $f := [0, T] \to \mathcal{X}$ is continuous. It is well known that if we suppose, moreover, that the initial value η lies in $\mathcal{D}(A)$ then the classical solution of the problem (2.1)–(2.2) exists and is given by

$$u(t) = U(t)\eta + \int_0^t U(t-\tau) f(\tau) \,\mathrm{d}\tau. \tag{2.3}$$

However, (2.3) has sense for any $\eta \in \mathcal{X}$ even though the function $u(t)$ need not be differentiable in the general case. Nevertheless, we will suppose it to be the generalized solution of the problem (2.1)–(2.2).

In the nonhomogeneous case, we will not construct the approximations of the solution of (2.1)–(2.2) directly from a rational function F approximating the exponential as was described in Section 1 but we will use the so-called *selfstarting block methods* as they were introduced in [3]. For the readers' convenience, the definition and basic properties of such methods will be summarized in Appendix.

Apply now the SB-method (3.5) to the problem (2.1)–(2.2). We obtain

$$\begin{pmatrix} u_{jk+1} \\ \vdots \\ u_{(j+1)k} \end{pmatrix} = \begin{pmatrix} u_{jk} \\ \vdots \\ u_{jk} \end{pmatrix} + hC \begin{pmatrix} Au_{jk+1} \\ \vdots \\ Au_{(j+1)k} \end{pmatrix} + hC \begin{pmatrix} f_{jk+1} \\ \vdots \\ f_{(j+1)k} \end{pmatrix} + h \begin{pmatrix} d_1 Au_{jk} \\ \vdots \\ d_k Au_{jk} \end{pmatrix} + h \begin{pmatrix} d_1 f_{jk} \\ \vdots \\ d_k f_{jk} \end{pmatrix}. \tag{2.4}$$

Let $G \otimes A$ be the tensor product of a matrix G (of order k) and the operator A, i.e. $G \otimes A := \mathcal{D}(A) \times \ldots \times \mathcal{D}(A) \to \mathcal{X} \times \ldots \times \mathcal{X}$ defined by

$$G \otimes A = \begin{pmatrix} g_{11}A & \ldots & g_{1k}A \\ \vdots & \ddots & \vdots \\ g_{k1}A & \ldots & g_{kk}A \end{pmatrix}. \tag{2.5}$$

This notation allows to rewrite (2.4) in the form

$$(I - hC \otimes A) \begin{pmatrix} u_{jk+1} \\ \vdots \\ u_{(j+1)k} \end{pmatrix} = (I + hD \otimes A) \begin{pmatrix} u_{jk} \\ \vdots \\ u_{jk} \end{pmatrix} + hC \begin{pmatrix} f_{jk+1} \\ \vdots \\ f_{(j+1)k} \end{pmatrix} + h \begin{pmatrix} d_1 f_{jk} \\ \vdots \\ d_k f_{jk} \end{pmatrix} \tag{2.6}$$

where D is the diagonal matrix with the components of the vector $\underline{d}$ on the diagonal.

The operator $I - hC \otimes A$ is generally unbounded. Thus, the question about the solvability of (2.4) should be answered first.

Before formulating the corresponding theorem, we recall that the resolvent $R(\lambda, A)$ of A has to satisfy the inequality

$$\|R(\lambda, A)^n\| \leqslant \frac{M}{(\Re(\lambda) - \omega)^n} \tag{2.7}$$

for $n = 1, 2, \ldots$ and for any λ for which $\Re(\lambda) > \omega$, and M, ω are real constants. It is so since A is the infinitesimal generator of a strongly continuous semigroup of operators, see, e.g., [1]. We also recall that the semigroup fulfills the inequality

$$\|U(t)\| \leqslant M \exp(\omega t). \tag{2.8}$$

Further, realize that any matrix of the form $I - zC$ is (at least for sufficiently small z) nonsingular so that it is possible to write its inverse in the form

$$(I - zC)^{-1} = \frac{1}{Q(z)} \begin{pmatrix} p_{11}(z) & \ldots & p_{1k}(z) \\ \vdots & \ddots & \vdots \\ p_{k1}(z) & \ldots & p_{kk}(z) \end{pmatrix}, \tag{2.9}$$

where

$$Q(z) = \det(I - zC) \tag{2.10}$$

and $p_{ij}(z)$ is the determinant of the matrix of order $k-1$ obtained from the matrix $(I - zC)$ by omitting the jth row and ith column and multiplying by $(-1)^{i+j}$. Note that any $p_{ij}(z)$ is a polynomial in z of degree at most $k-1$.

Theorem 2.1. *Let A have its spectrum in the half-plane $\Re(\lambda) \leqslant \omega$ and let C have its eigenvalues in the half-plane $\Re(\lambda) > 0$. Then the operator $(I - hC \otimes A)$ has for sufficiently small h a bounded inverse, and*

$$(I - hC \otimes A)^{-1} = M \equiv \begin{pmatrix} m_{11} & \ldots & m_{1k} \\ \vdots & \ddots & \vdots \\ m_{k1} & \ldots & m_{kk} \end{pmatrix}, \tag{2.11}$$

holds, where

$$m_{ij} = p_{ij}(hA) Q^{-1}(hA) \tag{2.12}$$

and the polynomials $p_{ij}(z)$ and $Q(z)$ are given by (2.9) *and* (2.10), *respectively.*

P r o o f. In this proof we use some results from the theory of functions of unbounded operators, see again, e.g., [1]. The degree of the polynomial Q is exactly k, since C is regular. Moreover, there exists a constant $\lambda_C > 0$ such that all roots of Q lie in the half-plane $\Re(\lambda) \geqslant \lambda_C$, since the eigenvalues of C lie in the open right half-plane. Without loss of generality we can suppose that $\omega > 0$ and let us choose h_0 in such a way that $h_0 < \lambda_C/\omega$. Then the spectrum of the operator hA lies in the half-plane $\Re(\lambda) \leqslant h_0\omega < \lambda_C$ for any $0 < h \leqslant h_0$. Further, the degree of any of the polynomials $p_{ij}(z)$ is at most $k-1$ and the degree of the polynomial $Q(z)$ is exactly k as we have already said above. Consequently, the rational function $p_{ij}(z)Q^{-1}(z)$ is regular at the infinity, and it is also regular in the half-plane $\Re(\lambda) < \lambda_C$, as follows from the properties of the roots of the polynomial Q. Thus, the operators $p_{ij}(hA)Q^{-1}(hA)$ are correctly defined and they are bounded operators in $\mathcal{X}$.

The definition of the functions $p_i^{(j)}$ gives immediately that

$$\sum_{s=1}^{k}(\delta_i^{(s)} - zc_{is})p_{sj}(z)Q^{-1}(z) = \delta_i^{(j)}, \tag{2.13}$$

where $\delta_i^{(s)}$ is the Kronecker symbol. Note that formula (2.13) is nothing else than the commonly known Cramer's rule. Since $\delta_i^{(s)} - zc_{is}$ is a polynomial of degree 1 and since $p_{sj}(z)Q^{-1}(z)$ has a root at the infinity, it follows that

$$\sum_{s=1}^{k}(\delta_i^{(s)} I - c_{is}hA)p_{sj}(hA)Q^{-1}(hA)x = \delta_i^{(j)} x \tag{2.14}$$

for any $x \in \mathcal{X}$, and the operators behind the summation sign are well-defined bounded operators. But (2.14) gives immediately that

$$(I - hC \otimes A)M\underline{x} = \underline{x} \tag{2.15}$$

for any $\underline{x} \in \mathcal{X} \times \ldots \times \mathcal{X}$. In the similar way we prove that

$$M(I - hC \otimes A)\underline{x} = \underline{x} \tag{2.16}$$

for any $\underline{x} \in \mathcal{D}(A) \times \ldots \times \mathcal{D}(A)$. Now equations (2.15)–(2.16) complete the proof of theorem. □

Supposing that C satisfies the assumptions of Theorem 2.1, we can rewrite (2.6) in the form

$$u_{(j+1)k} = F(hA)u_{jk} + q_j, \quad j = 0, 1, \ldots, \tag{2.17}$$

where the rational function F is given by

$$F(z) = P(z)Q^{-1}(z), \tag{2.18}$$

the polynomial P by

$$P(z) = \sum_{s=1}^{k} p_{ks}(z)(1 + d_s z) \tag{2.19}$$

and

$$q_j = h \sum_{i=1}^{k} \sum_{s=1}^{k} c_{is} m_{ki} f_{jk+s} + h \sum_{i=1}^{k} d_i m_{ki} f_{jk}. \tag{2.20}$$

Hence, if the matrix C has its eigenvalues in the right half-plane the corresponding SB-method can be used even for the approximation of the generalized solution. Naturally, the method must be understood in the form (2.17). The convergence is controlled—as can be expected—by the behaviour of the powers of the operator $F(hA)$. Before formulating the corresponding convergence theorem we prove an auxiliary assertion.

Lemma 2.1. *Let an SB-method of order $p \geqslant 1$ be given and let the corresponding matrix C have its eigenvalues in the right half-plane of the complex plane. Further, let* (1.5) *be satisfied. Finally, let $f(t)$ be continuous in $[0, T]$ and define $\Delta(h)$ by*

$$\Delta(h) = \sup_{\substack{nh \leqslant t \\ 0 \leqslant \tau \leqslant t}} \|[F^n(hA) - U^n(h)]f(\tau)\|. \tag{2.21}$$

Then

$$\lim_{h \to 0} \Delta(h) = 0. \tag{2.22}$$

P r o o f. Suppose that (2.22) is not true. Then there exist $\varepsilon_0 > 0$ and sequences $\{h_k\}$, $\{n_k\}$, and $\{\tau_k\}$ satisfying

$$h_k \to 0, \quad n_k h_k \leqslant t, \quad \tau_k \leqslant t \tag{2.23}$$

such that

$$\|[F^{n_k}(h_k A) - U^{n_k}(h_k)]f(\tau_k)\| \geqslant \varepsilon_0 \tag{2.24}$$

holds for $k = 1, 2, \ldots$. Passing if necessary to subsequences, we can assume here that $n_k h_k \to t_0$, $\tau_k \to t_1$ for $k \to \infty$. Under these assumptions we have

$$\|[F^{n_k}(h_k A) - U^{n_k}(h_k)]f(t_1)\| \geqslant \frac{1}{2}\varepsilon_0, \quad k = 1, 2. \ldots. \tag{2.25}$$

This estimate follows from the definition of supremum, since $f(t)$ is continuous and $\|F^n(hA) - U^n(h)\|$ is bounded (see (2.7) and (1.5)). On the other hand,

$$\begin{aligned}\|[F^{n_k}(h_k A) - U^{n_k}(h_k)]f(t_1)\| \\ \leqslant \|[F^{n_k}(h_k A) - U(t)]f(t_1)\| + \|[U(t) - U(n_k h_k)]f(t_1)\| \to 0\end{aligned} \tag{2.26}$$

in virtue of Theorem 2.1 and the continuity of $U(t)$. Thus, we have a contradiction proving the lemma. □

Now we have all ready to prove the convergence theorem for the approximation of problem (2.1)–(2.2).

Theorem 2.2. *Let an SB-method of order $p \geqslant 1$ be given and let the corresponding matrix C have its eigenvalues in the right half-plane of the complex plane. Further, let* (1.5) *be satisfied. Finally, let u_{jk} be the approximate solution of the problem* (2.1)–(2.2), *where $f(t)$ is continuous, $f(t) \in \mathcal{D}(A)$ for $t \in [0, T]$ and $Af(t)$ is also continuous (cf.* (2.17)–(2.20)*). Then*

$$\lim_{\substack{h \to 0 \\ jh \to t}} u_{jk} = u(t). \tag{2.27}$$

P r o o f. If we take into account (2.17) we can write the approximation u_{jk} in the form

$$u_{jk} = F^j(hA) + \sum_{\nu=0}^{j-1} F^{j-1-\nu}(hA)q_\nu, \tag{2.28}$$

where q_j are given by (2.20). From Theorem 1.1 we know that $F^j(hA)\eta \to U(t)\eta$ for $h \to 0$ and $jh \to t$. So it remains to prove that

$$\sum_{\nu=0}^{j-1} F^{j-1-\nu}(hA)q_\nu \to \int_0^t U(t-\tau)f(\tau)\,\mathrm{d}\tau. \tag{2.29}$$

To achieve this let us investigate the operators q_j. Begin with the obvious identity $\sum\limits_{r=1}^{k} m_{kr}(\delta_r^{(i)} I - c_{ri} hA) = \delta_k^{(i)} I$ (see (2.16)) and rewrite it in the form

$$m_{ki} = hA \sum_{r=1}^{k} c_{ri} m_{kr} + \delta_k^{(i)} I. \tag{2.30}$$

After substituting (2.30) into (2.20), we obtain

$$
\begin{aligned}
(2.31)\qquad q_\nu &= h^2 A \sum_{i=1}^{k}\sum_{s=1}^{k} c_{is} \sum_{r=1}^{k} c_{ri} m_{kr} f_{\nu k+s} + h \sum_{i=1}^{k}\sum_{s=1}^{k} c_{is}\delta_k^{(i)} f_{\nu k+s} \\
&\quad + h^2 A \sum_{i=1}^{k} d_i \sum_{r=1}^{k} c_{ri} m_{kr} f_{\nu k} + h \sum_{i=1}^{k} d_i \delta_k^{(i)} f_{\nu k} \\
&= h^2 A \sum_{i=1}^{k}\sum_{r=1}^{k} c_{ri} m_{kr} \left(\sum_{s=1}^{k} c_{is} f_{\nu k+s} + d_i f_{\nu k} \right) \\
&\quad + h \left(\sum_{s=1}^{k} c_{ks} f_{\nu k+s} + d_k f_{\nu k} \right).
\end{aligned}
$$

The continuity of f implies that

$$
(2.32)\qquad f_{\nu k+s} = f_{\nu k} + \varphi_s,
$$

where

$$
(2.33)\qquad \|\varphi_s\| = o(1) \quad \text{for } h \to 0.
$$

Observing now that the norms of the operators m_{kr} are uniformly bounded (note that $m_{kr} = p_{kr}(hA)Q^{-1}(hA)$ and the degree of p_{kr} is less than the degree of $Q(z)$) and that the function $\|Af(t)\|$ is continuous and, therefore, also bounded, we obtain from (2.31)–(2.33) that

$$
(2.34)\qquad q_\nu = h\left(\sum_{s=1}^{k} c_{ks} + d_k \right) f_{\nu k} + \psi_\nu,
$$

where

$$
(2.35)\qquad \|\psi\| = o(h).
$$

But $\sum_{s=1}^{k} c_{ks} + d_k = 1$ since the order of the method used is at least 1 and, hence, (2.34) implies that

$$
(2.36)\qquad q_\nu = h f_{\nu k} + \psi_\nu.
$$

The substitution of (2.35) in the left-hand part of (2.29) gives

$$
(2.37)\qquad \sum_{\nu=0}^{j-1} F^{j-1-\nu}(hA) q_\nu = h \sum_{\nu=0}^{j-1} F^{j-1-\nu}(hA) f_{\nu k} + o(1).
$$

Obviously,

$$h\sum_{\nu=0}^{j-1} F^{j-1-\nu}(hA)f_{\nu k} - \int_0^{t_{jk}} U(t_{jk}-\tau)f(\tau)\,\mathrm{d}\tau = h\sum_{\nu=0}^{j-1} F^{j-1-\nu}(hA)f_{\nu k} - \sum_{\nu=0}^{j-1}\int_{t_{\nu k}}^{t_{(\nu+1)k}} U(t_{jk}-\tau)f(\tau)\,\mathrm{d}\tau. \tag{2.38}$$

The integral in the last sum of the right-hand term of (2.38) can be estimated as

$$\int_{t_{\nu k}}^{t_{(\nu+1)k}} U(t_{jk}-\tau)f(\tau)\,\mathrm{d}\tau = hU(t_{jk}-t_{\nu k})f_{\nu k} + o(h), \tag{2.39}$$

since the function $U(t_{jk}-\tau)f(\tau)$ is continuous. Thus, it remains to investigate the behaviour of the expression $h\sum_{\nu=0}^{j-1}(F^{j-1-\nu}(hA) - U(t_{jk}-t_{\nu k}))f_{\nu k}$. But

$$\begin{aligned} h\sum_{\nu=0}^{j-1}&(F^{j-1-\nu}(hA) - U(t_{jk}-t_{\nu k}))f_{\nu k} \\ &= h\sum_{\nu=0}^{j-1}(F^{j-1-\nu}(hA) - U^{j-\nu}(h))f_{\nu k} \\ &= h\sum_{\nu=0}^{j-1}(F^{j-1-\nu}(hA) - U^{j-1-\nu}(h))f_{\nu k} + h\sum_{\nu=0}^{j-1} U^{j-1-\nu}(h)(I-U(h))f_{\nu k}. \end{aligned} \tag{2.40}$$

If we use now Lemma 2.1 and the estimate (2.8) the assertion of the theorem follows immediately. □

3. Appendix

Let an ordinary differential equation

$$u'(t) = f(t,u), \quad t\in[0,T], \tag{3.1}$$

with the initial condition

$$u(0) = \eta \in \mathbb{R} \tag{3.2}$$

be given. The right-hand term of (3.1) is supposed to be defined, continuous, and Lipschitzian with respect to u in the strip $0 \leqslant t \leqslant T$, $-\infty < y < \infty$ so that the existence and uniqueness of (3.1), (3.2) is guaranteed in the whole interval $[0,T]$.

Let an integer k, $k \geqslant 1$, real numbers $\mu_1, \dots, \nu_{k-1}$, a square matrix C of order k, a k-dimensional vector $\underline{d}$ and a positive real number h be given. Putting

$$t_{rk} = rh, \quad r = 0, 1, \dots \tag{3.3}$$

(these points will be called the basic points),

$$t_{rk+i} = t_{rk} + \mu_i h, \dots, k-1 \tag{3.4}$$

(the intermediate points), and denoting the approximate solution at the point t_j by u_j, the selfstarting block method (SB-method briefly) is defined by the formula

$$\begin{pmatrix} u_{rk+1} \\ \vdots \\ u_{(r+1)k} \end{pmatrix} = u_{rk}\underline{i} + hC \begin{pmatrix} f_{rk+1} \\ \vdots \\ f_{(r+1)k} \end{pmatrix} + hf_{rk}\underline{d}, \quad r = 0, \dots, \tag{3.5}$$

where $\underline{i} = (1, \dots, 1)^\top$ and $f_j = f(t_j, u_j)$. One step of the SB-method consists therefore in computing k values of the approximate solution simultaneously from the generally nonlinear system of equations and the next step is started with the last of them. Note that the Lipschitz property of f guarantees that the system (3.5) has—at least for any sufficiently small h—exactly one solution.

The local truncation error of an SB-method is defined in the usual way, i.e. by

$$\underline{L}(u(t); h) = \begin{pmatrix} u(t+\mu_1 h) \\ \vdots \\ u(t+\mu_k h) \end{pmatrix} - u(t)\underline{i} - hC \begin{pmatrix} u'(t+\mu_1 h) \\ \vdots \\ u'(t+\mu_k h) \end{pmatrix} - hu'(t)\underline{d} \tag{3.6}$$

where $\mu_k = 1$. Using this definition, the given SB-method will be said to have the order p (p positive integer), if

$$L_i(u(t); h) = O(h^{p+1}) \quad \text{for } i = 1, \dots, k, \tag{3.7}$$

where L_i is the ith component of the vector $\underline{L}$.

The order of the method depends only on the parameters of the method and does not depend on the particular function u. For example, the assertion that the order of the method is at least 1 is equivalent to k algebraic equalities

$$\sum_{s=1}^{k} c_{is} + d_i = \mu_i, \quad i = 1, \dots, k. \tag{3.8}$$

The following two theorems can be proved very simply (see [3]).

Theorem 3.1. *The selfstarting method of order at least* 1 *is convergent.*

Theorem 3.2. *Let the solution of* (3.1), (3.2) *have* $p+1$ *continuous derivatives in* $[0,T]$. *Then the error of an SB-method of order* p *is of the order* $O(h^p)$.

The subclass of overimplicit block methods formed by such SB-methods for which $\mu_i = i/k$ and the order of which is at least k is not empty, as was shown also in [3]. We denote them as SBK-methods. Note that these methods play an important role in the study of methods for solving stiff differential equations.

References

[1] *N. Dunford, J. T. Schwartz*: Linear Operators. 1. General Theory. Interscience Publishers, New York-London, 1958.
[2] *T. Kato*: Perturbation Theory for Linear Operators. Springer, Berlin-Heidelberg-New York, 1966.
[3] *M. Práger, J. Taufer, E. Vitásek*: Overimplicit multistep methods. Apl. Math. *18* (1973), 399–421.
[4] *A. E. Taylor*: Introduction to Functional Analysis. John Wiley & Sons, New York, 1958.
[5] *E. Vitásek*: Approximate solutions of abstract differential equations. Appl. Math. *52* (2007), 171–183.
[6] *K. Yosida*: Functional Analysis. Springer, Berlin-Heidelberg-New York, 1971.

Author's address: *E. Vitásek*, Institute of Mathematics of the Academy of Sciences of the Czech Republic, Žitná 25, CZ-115 67 Prague 1, Czech Republic, e-mail: `vitas@math.cas.cz`.