Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

CONVERGENCE AND STABILITY CONSTANT OF THE THETA-METHOD

István Faragó
Eötvös Loránd University, Institute of Mathematics and
HAS-ELTE Numerical Analysis and Large Networks Research Group
Pázmány P. s. 1/c, 1117 Budapest, Hungary
faragois@cs.elte.hu

Abstract

The Euler methods are the most popular, simplest and widely used methods for the solution of the Cauchy problem for the first order ODE. The simplest and usual generalization of these methods are the so called theta-methods (notated also as θ-methods), which are, in fact, the convex linear combination of the two basic variants of the Euler methods, namely of the explicit Euler method (EEM) and of the implicit Euler method (IEM). This family of the methods is well-known and it is introduced almost in any arbitrary textbook of the numerical analysis, and their consistency is given. However, in its qualitative investigation the convergence is proven for the EEM, only, almost everywhere. At the same time, for the rest of the methods it is usually missed (e.g., [1, 2, 7, 8]). While the consistency is investigated, the stability (and hence, the convergence) property is usually shown as a consequence of some more general theory. In this communication we will present an easy and elementary prove for the convergence of the general methods for the scalar ODE problem. This proof is direct and it is available for the non-specialists, too.

1. Motivation and basic of the theta-method

Many different problems (physical, chemical, etc.) can be described by the initial-value problem for first order ordinary differential equation (ODE) of the form

$$\frac{du}{dt} = f(t,u), \quad t \in (0,T), \tag{1}$$

$$u(0) = u_0. \tag{2}$$

We note that, using the semidiscretization, the time-dependent partial differential equations also lead to the problem (1)–(2). Hence, the solution of such problem plays a crucial role in mathematical modelling. (For simplicity, in sequel we consider only the scalar problem, i.e., when $f : \mathbb{R}^2 \to \mathbb{R}$.) We know that under the global Lipshitz condition, i.e., in case

$$|f(t,s_1) - f(t,s_2)| \le L|s_1 - s_2| \quad \text{for all } (t,s_1),(t,s_2) \in \mathrm{dom}(f) \tag{3}$$

with the Lipschitz constant $L > 0$, the problem (1)–(2) has unique solution on the entire domain $\mathrm{dom}(f)$.

Since we have no hope of solving the vast majority of differential equations in explicit, analytic form, the design of suitable numerical algorithms for accurately approximating solutions is essential. The ubiquity of differential equations throughout mathematics and its applications has driven the tremendous research effort devoted to numerical solution schemes, some dating back to the beginnings of the calculus. Therefore, we apply some numerical method. Hence, the numerical integration of the problem (1)–(2) – under the condition (3) – is one of the most typical tasks in the numerical modelling of real-life problems.

Our aim is to define some numerical solution at some fixed point $t^\star \in (0, T)$ to the Cauchy problem (1)–(2). Therefore, we construct the sequence of the uniform meshes with the mesh-size $h = t^\star/N$ of the form

$$\omega_h = \{t_n = n \cdot h, n = 0, 1, \ldots, N\},$$

and our aim is to define at the mesh-point $t^\star = t_N$ a suitable approximation y_N on each fixed mesh.

This requires to give the rule how to define the mesh-function $y_h : \omega_h \to \mathbb{R}$. The most popular, simplest and widely used method are the so-called single step (one-step) schemes, particularly, the theta-method, which is frequently notated as θ-method. Using the notation $y_h(t_n) = y_n$, the θ-method is defined as

$$\begin{aligned} &y_n = y_{n-1} + h\left(\theta f(t_n, y_n) + (1-\theta) f(t_{n-1}, y_{n-1})\right), \quad n = 1, \ldots, N, \\ &y_0 = u_0. \end{aligned} \tag{4}$$

Here $\theta \in [0, 1]$ is a fixed parameter, and, it is for $\theta = 0$ explicit, otherwise implicit method. The θ-method is considered here as basic method since it represents the most simple Runge-Kutta method (and also linear multistep method). For stiff systems the cases $\theta = 0.5$ trapezoidal rule and $\theta = 1$ implicit (backward) Euler are of practical interest, for non-stiff systems we can also consider $\theta = 0$ explicit (forward) Euler.

In mathematics and computational science, these methods are most basic method for numerical integration of ordinary differential equations and they are the simplest Runge-Kutta methods.

Let us define the local truncation error for the θ-method, under the assumption that f (and hence, the solution $u(t)$) is sufficiently smooth.

As it is known, the local truncation error $l_n(h)$ for the θ-method can be defined as

$$l_n(h) = u(t_n) - u(t_{n-1}) - h\theta f(t_n, u(t_n)) - h(1-\theta) f(t_{n-1}, u(t_{n-1})), \tag{5}$$

where $u(t)$ stands for the solution of the problem (1)–(2). Therefore, we have the relation

$$l_n(h) = u(t_n) - u(t_{n-1}) - h\theta u'(t_n) - h(1-\theta) u'(t_{n-1}). \tag{6}$$

Hence, by expanding $u(t_n) = u(t_{n-1} + h)$ and $u'(t_n) = u'(t_{n-1} + h)$ into the Taylor series around the point $t = t_{n-1}$, we get for the local approximation error the relation

$$l_n(h) = (1/2 - \theta) h^2 u''(t_{n-1}) + (1/6 - \theta/2) h^3 u'''(t_{n-1}) + \mathcal{O}(h^4). \tag{7}$$

The order of a numerical method is defined by the local truncation error: when $l_n(h) = \mathcal{O}(h^{p+1})$ then the method is called consistent of order p. This means that for both Euler methods ($\theta = 0$ and $\theta = 1$) the order of consistency is equal to one, while for the trapezoidal rule ($\theta = 0.5$) the order of consistency is equal to two.

However, as it is well-known, the consistency itself does not guarantee the convergence of a numerical method, the stability is also required.

Roughly speaking, the consistency is the characterization of the local (truncation) error of the method, which is the error committed by one step of the method. (That is, it is the difference between the result given by the method, assuming that no error was made in earlier steps and hence having the exact solution.) On the other hand, the stability guarantees that the numerical method produces a bounded solution whenever the solution of the exact differential equation is bounded, in other words, the local truncation errors are damped out. The convergence means that the numerical solution approximates the solution of the original problem, i.e., a numerical method is said to be convergent if the numerical solution converges to the exact solution as the step size of mesh h tends to zero.

Although the consistency analysis of the θ-method is introduced almost in any arbitrary textbook of the numerical analysis, typically the stability (and hence, the convergence) is shown directly for the explicit method, only.

Our aim is to give an easy and elementary prove for the convergence of the general θ-method, i.e., we consider the implicit methods. The proof is direct and it is available for the non-specialists, too. Moreover, we give the expression for the stability constant of the θ-method.

This paper extends the results of the paper [4] in two directions: we prove the convergence of any implicit θ-method, and we also give sharp estimate for the stability constant, improving the result obtained in paper [4].

The paper is organized as follows. In Section 2, for sake of completeness, we formulate the basic results for the explicit Euler method, proving its convergence and stability constant. Section 3 contains the simple and compact proof of the convergence of the θ-method, and we define the order of its convergence, too. Finally, we finish the paper with giving some remarks and conclusions.

2. Convergence and the stability constant of the explicit Euler method

In this section we use a sequence of meshes ω_h and we define the numerical solution at some fixed point $t^\star \in (0, T)$ to the Cauchy problem (1)–(2) for the θ-method with $\theta = 0$, i.e., by using the scheme

$$\begin{aligned} y_n &= y_{n-1} + hf(t_{n-1}, y_{n-1}), \quad n = 1, 2, \ldots, N, \\ y_0 &= u_0 \end{aligned} \tag{8}$$

with $Nh = t^\star$.

The following statement will be used several times within the paper.

Lemma 2.1 *Let $a \geq 1$, $b \geq 0$, and s_n be such numbers that the inequalities*

$$|s_n| \leq a|s_{n-1}| + b, \quad n = 1, 2, \ldots \tag{9}$$

hold. Then the estimate

$$|s_n| \leq a^n \left(|s_0| + n\frac{b}{a} \right), \quad n = 0, 1, 2, \ldots \tag{10}$$

is valid.

Proof. By using induction, we can readily verify the statement. Indeed, for $n = 0$ (10) is clearly valid. Now, under the assumption that (10) holds for $n - 1$, from (9) we have

$$\begin{aligned} |s_n| &\leq a \left[a^{n-1} \left(|s_0| + (n-1)\frac{b}{a} \right) \right] + b \\ &= a^n \left(|s_0| + n\frac{b}{a} \right) \underbrace{-a^{n-1}b + b}_{\leq 0} \leq a^n \left(|s_0| + n\frac{b}{a} \right), \end{aligned} \tag{11}$$

which yields the statement. □

For the EEM the local truncation error at the mesh-point $t = t_n$ can be written as

$$l_n(h) = u(t_n) - u(t_{n-1}) - hu'(t_{n-1}) = \frac{h^2}{2} u''(\vartheta_n^{EEM}), \tag{12}$$

where $\vartheta_n^{EEM} \in (t_{n-1}, t_n)$ is a given value. Hence, setting $M_2 = \max_{[0,t^\star]} |u''|$, we get

$$l_n(h) \leq l(h) := M_2 \frac{h^2}{2}. \tag{13}$$

Let us consider the EEM defined by the one-step recursion (8). Due to (12), we have

$$u(t_n) = u(t_{n-1}) + hf(t_{n-1}, u(t_{n-1})) + l_n(h). \tag{14}$$

Hence, for the global error $e_n = u(t_n) - y_n$ at the mesh-point $t = t_n$ we get the recursion in the form

$$e_n = e_{n-1} + h\left(f(t_{n-1}, u(t_{n-1})) - f(t_{n-1}, y_{n-1}) \right) + l_n(h). \tag{15}$$

Hence, using the Lipschitz property (3) and (13), we obtain

$$|e_n| \leq |e_{n-1}| + hL|e_{n-1}| + l(h) = (1 + Lh)|e_{n-1}| + l(h), \tag{16}$$

for any $n = 1, 2, \ldots, N$. Then, by choosing $a = 1 + Lh$ and $b = l(h)$, and using the inequality $1 + x \leq \exp(x)$ for $x \geq 0$, Lemma 2.1 implies the estimate

$$|e_n| \leq [\exp(hL)]^n \left[|e_0| + \frac{nl(h)}{1 + Lh} \right] \leq [\exp(hL)]^n \left[|e_0| + nl(h) \right]. \tag{17}$$

Since $nh = t_n \le t^\star$, the following relations obviously hold for any $n = 1, 2, \ldots, N$:

$$[\exp(hL)]^n = \exp(Lhn) = \exp(Lt_n) \le \exp(Lt^\star),$$
$$nl(h) = nM_2\frac{h^2}{2} = \frac{M_2 t_n}{2}h \le \frac{M_2 t^\star}{2}h.$$

Because $e_0 = 0$, the relation (17) results in the estimate

$$|e_n| \le C_{EEM} \cdot h, \tag{18}$$

for all $n = 1, 2, \ldots, N$ with $C_{EEM} = \exp(Lt^\star)\dfrac{M_2 t^\star}{2}$. Putting $n = N$ into (18), we get

$$|e_N| \le C_{EEM} \cdot h. \tag{19}$$

This proves the first order convergence of the EEM with the stability constant C_{EEM}.

3. Convergence of the implicit theta methods

The convergence of the implicit θ-method (i.e., for $\theta \in (0, 1]$) cannot be proven directly as it was done previously. The main reason is that from the corresponding error recursion the inequality (9) cannot be obtained directly, due to the implicitness with respect to e_n. The usual way of proving the convergence of the θ-method is to show the zero-stability, by using its first characteristic polynomial. (The proof is complicated, and it can be found in [6, 10].)

In the sequel, using Lemma 2.1, we give an elementary proof of the convergence.

To this aim, we first give a uniform estimate for the local approximation error, which, by (6), has the form

$$\begin{aligned} l_n(h) &= u(t_n) - u(t_{n-1}) - (1-\theta)hu'(t_{n-1}) - \theta u'(t_n) \\ &= \theta\left(u(t_n) - u(t_{n-1}) - hu'(t_n)\right) + (1-\theta)\left(u(t_n) - u(t_{n-1}) - hu'(t_{n-1})\right). \end{aligned} \tag{20}$$

The Taylor polynomial with Lagrange remainder gives

$$\begin{aligned} u(t_{n-1}) &= u(t_n) - hu'(t_n) + \frac{h^2}{2}u''(t_n) - \frac{h^3}{6}u'''(\vartheta_n^1), \\ u(t_n) &= u(t_{n-1}) + hu'(t_{n-1}) + \frac{h^2}{2}u''(t_{n-1}) + \frac{h^3}{6}u'''(\vartheta_n^2). \end{aligned} \tag{21}$$

Using the relation $u''(t_n) = u''(t_{n-1}) + hu'''(\vartheta_n^3)$ (where $\vartheta_n^i \in (t_{n-1}, t_n)$ for $i = 1, 2, 3$), substitution (21) into (20) results in the equality

$$l_n(h) = \frac{h^2}{2}(1-2\theta)u''(t_{n-1}) + \frac{h^3}{6}\left(-3\theta u'''(\vartheta_n^3) + \theta u'''(\vartheta_n^1) + (1-\theta)u'''(\vartheta_n^2)\right). \tag{22}$$

Hence, using the notation $M_3 = \max_{[0,t^\star]}|u'''|$, we obtain

$$|l_n(h)| \le l(h) = C_2^\theta h^2 + C_3^\theta h^3, \tag{23}$$

where

$$C_2^\theta = \frac{|1-2\theta|}{2} M_2, \quad C_3^\theta = \frac{1+3\theta}{6} M_3. \tag{24}$$

We consider the θ-method, which means that the values y_n at the mesh-points ω_h are defined by the one-step recursion (4). Rearranging the local truncation error for θ-method of the form (5), and using the formula (4), for global error e_n we get the recursion

$$\begin{aligned} e_n = e_{n-1} &+ h\theta \left(f(t_n, u(t_n)) - f(t_n, y_n) \right) \\ &+ h(1-\theta) \left(f(t_{n-1}, u(t_{n-1})) - f(t_{n-1}, y_{n-1}) \right) + l_n(h), \quad n = 1, \ldots, N, \end{aligned} \tag{25}$$

with $e_0 = 0$. This equality, by using the Lipschitz continuity, implies the relation

$$|e_n| \le |e_{n-1}| + \theta L h |e_n| + (1-\theta) L h |e_{n-1}| + |l_n(h)|, \quad n = 1, \ldots, N. \tag{26}$$

Using the uniform estimate (23), (26) yields that with the choice

$$a = \frac{1 + (1-\theta)Lh}{1 - \theta L h}, \quad b = \frac{l(h)}{1 - \theta L h} \tag{27}$$

the recursion

$$|e_n| \le a|e_{n-1}| + b, \quad n = 1, 2, \ldots, N\,, \tag{28}$$

holds for the values

$$0 < h < \frac{1}{\theta L}. \tag{29}$$

Due to the obvious relations

$$a = 1 + \frac{Lh}{1 - \theta L h} \ge 1, \quad b \ge 0, \tag{30}$$

Lemma 2.1 is applicable to the recursion (28), which results in the validity of the estimate

$$|e_n| \le a^n \left(|e_0| + n\frac{b}{a} \right) = a^n \left(|e_0| + t_n \frac{l(h)}{h} \frac{1}{1 + (1-\theta)Lh} \right) \le a^n \left(t_n \frac{l(h)}{h} \right), \tag{31}$$

for any $n = 0, 1, 2, \ldots, N$ and h, satisfying (29).

We give an estimate for a^n. According to (30), we have

$$a = 1 + \frac{Lh}{1 - \theta L h} = 1 + \frac{1}{\theta} \cdot \frac{\theta L h}{1 - \theta L h}. \tag{32}$$

Let $\varepsilon > 0$ be arbitrary fixed number. Then for any $x \in (0, \varepsilon/(1+\varepsilon))$ the inequality $x^2/(1-x) \le \varepsilon x$ holds. Therefore, owning to the identity

$$\frac{x}{1-x} = x + \frac{x^2}{1-x},$$

we have the estimate

$$\frac{x}{1-x} \le (1+\varepsilon)x, \quad \text{for any } x \in \left(0, \frac{\varepsilon}{1+\varepsilon}\right). \tag{33}$$

Applying (33) to the second term on right-hand side (32), we obtain

$$a < 1 + \frac{1}{\theta} \cdot (1+\varepsilon)\theta L h = 1 + (1+\varepsilon)Lh \quad \text{for any } h \in (0, h_0), \tag{34}$$

where

$$h_0 = h_0(\varepsilon) = \frac{\varepsilon}{(1+\varepsilon)\theta L}. \tag{35}$$

Hence, using again the estimation $1 + s < \exp(s)$ for $s > 0$, we get

$$a^n < \exp\left(L(1+\varepsilon)t_n\right), \quad h \in (0, h_0). \tag{36}$$

Since for $\varepsilon > 0$ the inequality $\varepsilon/(1+\varepsilon) < 1$ holds, therefore under the condition $h \in (0, h_0)$ the requirement (29) is satisfied, too. Hence, based on relations (31), (23) and (36), we can formulate our results in the following statements.

Theorem 3.1 *Let $\varepsilon > 0$ be any fixed number and ω_h a mesh with mesh-size $h < h_0$, where h_0 is given in* (35)*. Then for the global error e_n of the θ-method with $\theta \in (0, 1]$ the estimate*

$$|e_n| \le t_n \left(C_2^\theta h + C_3^\theta h^2\right) \exp\left(L(1+\varepsilon)t_n\right) \tag{37}$$

holds for any $n = 1, 2, \ldots, N$, with the constants C_2^θ and C_3^θ defined in (24)*.*

Let us apply Theorem 3.1 for the value $n = N$. Then we have the following statement.

Corollary 3.2 *Under the assumptions and notations of the Theorem 3.1, for the global error e_N the estimate*

$$|e_N| \le t^\star \left(C_2^\theta h + C_3^\theta h^2\right) \exp\left(L(1+\varepsilon)t^\star\right) \tag{38}$$

holds.

The formula (38) gives an estimate for the global error at the mesh-point $t^\star = t_N = Nh$ of the θ-method with $\theta \in (0, 1]$ for any fixed $h \in (0, h_0)$. Moreover, ε depends on h_0, and, due to (35), ε also tends to zero as $h_0 \to 0$. Therefore, letting $h_0 \to 0$ on both sides of (38), we get the following statement.

Theorem 3.3 *The θ-method with any fixed $\theta \in (0, 1]$ is convergent at any fixed point $t^\star \in (0, T)$. Moreover, it is of the first order for $\theta \ne 0.5$, and of the second order for $\theta = 0.5$, with the stability constants $C_2^\theta t^\star \exp(Lt^\star)$ and $C_3^\theta t^\star \exp(Lt^\star)$, respectively.*

Since for the explicit Euler method we have $\theta = 0$ and $C_2^0 = C_{EEM}$ (c.f. formulas (15) and (24)), we can summarize our results in the following statement.

Theorem 3.4 *For the Cauchy problem* (1)–(2) *under the Lipschitz condition* (3) *the θ-method with any fixed $\theta \in [0,1]$ is convergent at any fixed point $t^\star \in (0,T)$. The rate of convergence of the method is equal to two for $\theta = 0.5$, otherwise it is of the first order. The stability constant C^θ of the method is defined as*

$$C^\theta = \begin{cases} \dfrac{1+3\theta}{6} M_3 t^\star \exp(Lt^\star) & \text{for } \theta = 0.5, \\ \dfrac{|1-2\theta|}{2} M_2 t^\star \exp(Lt^\star) & \text{for } \theta \neq 0.5, \end{cases} \tag{39}$$

respectively.

4. Concluding remarks

Finally, we give some comments.

♢ The convergence on the interval $[0, t^\star]$ yields the relation

$$\lim_{h \to 0} \max_{n=1,2,\ldots,N} |e_n| = 0.$$

As one can easily see, based on the relations (15) (for the EEM) and (37) (for the θ-method) the global error $|e_n|$ at any mesh-point can be bounded by the expression $C_{EEM} \cdot h$ (for the EEM) and by term, standing on the right-hand side of (38) (for the IEM). This means that both methods are convergent on the interval $[0, t^\star]$ with the same order.

♢ In our paper we did not consider roundoff error, which is always present in computer calculations. At the present time there is no universally accepted method to analyze roundoff error after a large number of time steps. The three main methods for analyzing roundoff accumulation are the analytical method, the probabilistic method and the interval arithmetic method, each of which has both advantages and disadvantages.

♢ In the implicit θ-method in each step we must solve a -usually non-linear- equations, namely, the root of the equation. This can be done by using some iterative method such as direct (function) iteration, Newton method and modified Newton method.

♢ In this paper we have been concerned with the stability and accuracy properties of the Euler methods in the asymptotic limit of $h \to 0$ and $N \to \infty$ while $N \cdot h$ is fixed. However, it is of practical significance to investigate the performance of methods in the case of fixed $h > 0$ and $n \to \infty$. Specifically, we would like to ensure that when applied to an initial value problem whose solution decays

to zero as $t \to \infty$, the Euler methods exhibit a similar behavior, for fixed $h > 0$ and $t_n \to \infty$. This problem is investigated on the famous Dahlquist scalar test equation, and it requires the so called A-stability property [3]. As it is known (e.g. in [8]), the θ-method is A-stable ("absolute stable") for the values $\theta \in [0.5, 1]$, otherwise the θ-method is bounded only under some strict condition for h. The latter makes these methods (including the EEM, too) unusable for several classes of the problem, like stiff problems.

♢ Why consider the θ-method, i.e., analyze the method with any θ in $[0, 1]$, not just 0, 0.5 and 1? We can list several reasons.

- The concept of order is based on assumption that error is concentrated on the leading order of Taylor series expansion (on real computers, h is small, but finite). As formula (7) shows, the case $\theta = 1/3$ gets rid of $\mathcal{O}(h^3)$ while retaining $\mathcal{O}(h^2)$. Hence, for different types of $f(t, u)$ one can tune θ to control whether $\mathcal{O}(h^3)$ and higher order terms or $\mathcal{O}(h^2)$ and higher order terms contribute to the overall error when h is finite.
- It may be possible to choose a θ that generates a close-to-optimal or smaller error. E.g., in [9] it is shown that the optimality criterion

$$\min_{\theta} \max_{-\infty<z<0} |\exp(z) - R(z)|$$

  leads to the value $\theta \approx 0.878$.
- θ-method is an example of a general approach to designing algorithms in which geometric intuition is replaced by Taylor series expansion. Invariably the implicit function theorem is also used in the design and analysis of this scheme.
- The implicit Euler method (the case $\theta = 1$) is very practical: it is a simple yet robust method for solving stiff ODE's.
- In some applications, a value such as $\theta = 0.55$ is used as trade-off between extended stability and second order accuracy.

♢ The qualitative analysis of the θ-method is investigated in several works, mainly, by its use to the numerical solution of some semidiscretized linear parabolic problems, (e.g. [5, 11]).

Acknowledgements

This work has been supported by the Hungarian Research Grant OTKA K 67819.

References

[1] Ascher, U. M. and Petzold, L. R.: *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM, Philadelphia, 1998.

[2] Bachvalov, N. S.: *Numerical methods*. Nauka, Moscow, 1975. (in Russian)

[3] Dahlquist, G.: Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. **4** (1956), 33–53.

[4] Faragó, I.: Note on the convergence of the implicit Euler method. (submitted)

[5] Faragó, I., Horváth, R.: Continuous and discrete parabolic operators and their qualitative properties. IMA J. Numer. Anal. **29** (2009), 606–631.

[6] Isaacson, E. and Keller, H. B.: *Analysis of numerical methods.* Wiley, New York, 1966.

[7] LeVeque, R.: *Finite difference mehods for ordinary and partial differential equations.* SIAM, Philadelphia, 2007.

[8] Lambert, J. D.: *Numerical methods for ordinary differential systems: The initial value problem.* John Wiley and Sons, Chicester, 1991.

[9] Liniger, W.: Global accuracy and A-stability of one- and two-step integration formulae for stiff ordinary differential equations. Lecture Notes in Mathematics **109** (1969), 188–193.

[10] Suli, E.: *Numerical solution of ordinary differential equations.* Oxford, 2010.

[11] Szabó, T.: On the discretization time-step in the finite element theta-method of the discrete heat equation. Lect. Notes Comp. Sci. **5434** (2009), 564–571.