Investigation of the Basic Notions in Numerical Analysis

M.Sc. Thesis
by

Imre Fekete
Applied Mathematician M.Sc., Applied Analysis

Supervisor:
István Faragó
Professor and
Head of the
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University

Budapest
2012

Acknowledgement

I would like to express my gratitude to my supervisor István Faragó for his impressive lectures on Numerical Methods for ODE's, which piqued my interest on the topic and for those inspiring discussions on nonlinear numerical analysis.

I would like to express my thanks to Miklós Mincsovics for those valuable conversations.

Finally, I would like to express my deepest gratitude to my family for their support, understanding and endless patience.

Contents

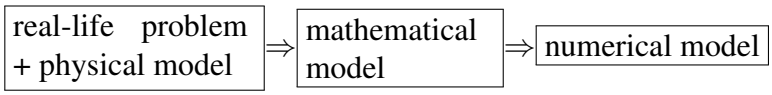
1 Preface

Many phenomena in nature can be described by mathematical models which consist of functions of a certain number of independent variables and parameters. In particular, these models often consist of equations, usually containing a large variety of derivatives with respect to the variables. Typically, we are not able to give the solution of the mathematical model in a closed (analytical) form, we construct some numerical and computer models that are useful for practical purposes. The ever-increasing advances in computer technology has enabled us to apply numerical methods to simulate plenty of physical and mechanical phenomena in science and engineering.

As a result, numerical methods do not usually give the exact solution to the given problem, they can only provide approximations, getting closer and closer to the solution with each computational step. Numerical methods are generally useful only when they are implemented on computer using a computer programming language. Using a computer, it is possible to gain quantitative (and also qualitative) information with detailed and realistic mathematical models and numerical methods for a multitude of phenomena and processes in physics and technology.

The application of computers and numerical methods has become ubiquitous. Computations are often cheaper than experiments; experiments can be expensive, dangerous or downright impossible. Real-life experiments can often be performed on a small scale only, and that makes their results less reliable.

The above modelling process of real-life phenomena can be illustrated as follows:



This means that the complete modelling process consists of three steps. In this thesis we will analyze the step when we transform the mathematical (usually continuous) model into numerical (usually discrete) models. Our aim is to guarantee that this step does not cause any significant loss of the information.

The discrete model usually yields a sequence of (discrete) tasks. During the construction of the numerical models the basic requirements are the following.

- Each discrete problem in the numerical model is a well-posed problem.
- In the numerical model we can efficiently compute the numerical solution.
- The sequence of the numerical solutions is convergent.
- The limit of this sequence is the solution of the original problem.

2 Mathematical background

When we model some real-life phenomenon with a mathematical model, it results in a – not necessarily linear – problem of the form

$$F(u) = 0\,, \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are normed spaces, $\mathcal{D} \subset \mathcal{X}$ and $F : \mathcal{D} \to \mathcal{Y}$ is a (nonlinear) operator. In the theory of numerical analysis it is usually *assumed* that there exists a unique solution, which will be denoted by $\bar{u}$.

On the other side, for any concrete applied problems *we must prove* the existence of $\bar{u} \in \mathcal{D}$. Even if it is possible to solve directly, the realization of the solving process is very difficult or even impossible. However, we need only a good approximation for the solution of problem (1). Therefore we construct numerical models by use of some discretization, which results in a sequence of simpler problems, i.e., a numerical method. With this approach we need to face the following difficulties:

- we need to compare the solution of the simpler problems with the solution of the original problem (1), which might be found in different spaces;
- this comparison seems to be impossible, since the solution of the original problem (1) is not known.

To get rid of the latter difficulty, the usual trick is to introduce the notions of consistency and stability, which are independent of the solution of the original problem (1) and are controllable. The convergence can be replaced with these two notions. Sometimes this popular "recipe" is summarized in the formula

$$\text{Consistency} + \text{Stability} = \text{Convergence}\,. \tag{2}$$

In the following we introduce and investigate these notions in an abstract framework, and we try to shed some light on the formula (2). Namely:

- how to define consistency and stability to ensure the formula (2);
- is it consistency or/and stability that is necessary for the convergence;
- how the nonlinear theory works in the linear case;

2.1 A demonstration example: the Cauchy problem

Definition 2.1. *Problem (1) can be given as a triplet $\mathscr{P} = (\mathcal{X}, \mathcal{Y}, F)$. We will refer to it as* problem $\mathscr{P}$.

Example 2.1. *Consider the following initial value problem:*

$$u'(t) = f(u(t)) \tag{3}$$

$$u(0) = u_0\,, \tag{4}$$

where $t \in [0,1],\ u_0 \in \mathbb{R}$ and $f \in C(\mathbb{R}, \mathbb{R})$ is a Lipschitz continuous function.

Then the operator F and the spaces $\mathcal{X}, \mathcal{Y}$ are defined as follows.

- $\mathcal{X} = C^1[0,1],\ \ \|u\|_{\mathcal{X}} = \max\limits_{t\in[0,1]} |u(t)|$
- $\mathcal{Y} = C[0,1] \times \mathbb{R},\ \ \left\| \begin{pmatrix} u \\ u_0 \end{pmatrix} \right\|_{\mathcal{Y}} = \max\limits_{t\in[0,1]} (|u(t)|) + |u_0|$
- $F(u) = \begin{pmatrix} u'(t) - f(u(t)) \\ u(0) - u_0 \end{pmatrix}.$

Definition 2.2. *We say that the sequence $\mathscr{N} = (\mathcal{X}_n, \mathcal{Y}_n, F_n)_{n\in\mathbb{N}}$ is a numerical method if it generates a sequence of problems*

$$F_n(u_n) = 0\,, \quad n = 1, 2, \dots\,, \tag{5}$$

where

- *$\mathcal{X}_n, \mathcal{Y}_n$ are normed spaces;*
- *$\mathcal{D}_n \subset \mathcal{X}_n$ and $F_n : \mathcal{D}_n \to \mathcal{Y}_n$.*

If there exists a unique solution of the (approximating) problems (5), it will be denoted by $\bar{u}_n$.

Example 2.2. *For $n \in \mathbb{N}$ we define the following sequence of triplets:*

- $\mathcal{X}_n = \mathbb{R}^{n+1},\ \ \mathbf{v}_n = (v_0, v_1, \dots, v_n) \in \mathcal{X}_n : \|\mathbf{v}_n\|_{\mathcal{X}_n} = \max\limits_{i=0,\dots,n} |v_i|$
- $\mathcal{Y}_n = \mathbb{R}^{n+1},\ \ \mathbf{y}_n = (y_0, y_1, \dots, y_n) \in \mathcal{Y}_n : \|\mathbf{y}_n\|_{\mathcal{Y}_n} = |y_0| + \max\limits_{i=1,\dots,n} |y_i|.$

- $F_n : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$, *and for any* $\mathbf{v}_n = (v_0, v_1, \ldots, v_n) \in \mathbb{R}^{n+1}$ *it acts as*

$$(F_n(\mathbf{v}_n))_i = \begin{cases} n\,(v_i - v_{i-1}) - f\,(v_{i-1})\,, & i = 1, \ldots, n, \\ \\ v_0 - u_0, & i = 0. \end{cases} \tag{6}$$

Definition 2.3. *We say that the sequence* $\mathscr{D} = (\varphi_n, \psi_n, \Phi_n)_{n\in\mathbb{N}}$ *is a* discretization *if*

- *the* φ_n*-s (respectively* ψ_n*-s) are restriction operators from* $\mathcal{X}$ *into* $\mathcal{X}_n$ *(respectively from* $\mathcal{Y}$ *into* $\mathcal{Y}_n$*), where* $\mathcal{X}, \mathcal{X}_n, \mathcal{Y}, \mathcal{Y}_n$ *are normed spaces;*
- $\Phi_n : \{F : \mathcal{D} \to \mathcal{Y} \,|\, \mathcal{D} \subset \mathcal{X}\} \to \{F_n : \mathcal{D}_n \to \mathcal{Y}_n \,|\, \mathcal{D}_n \subset \mathcal{X}_n\}$.

Example 2.3. *Based on Examples 2.1 and 2.2, in Definition 2.3 we define* $\mathcal{X} = C^1[0,1]$, $\mathcal{Y} = C[0,1] \times \mathbb{R}$, *and* $\mathcal{X}_n = \mathcal{Y}_n = \mathbb{R}^{n+1}$. $\mathbb{G}_n := \{t_i = \frac{i}{n},\ i = 0, \ldots, n\}$. *Then, we define the triplet of the operators as follows.*

- *For any* $u \in \mathcal{X}$ *we put* $(\varphi_n u)_i = u(t_i),\ \ i = 0, 1, \ldots, n,$
- *For any* $y \in \mathcal{Y}$ *we put*

$$(\psi_n y)_i = \begin{cases} y\,(t_{i-1})\,, & 1, \ldots, n, \\ \\ y(t_0), & i = 0. \end{cases}$$

- *In order to give* Φ_n*, we define the mapping* $\Phi_n : C^1[0,1] \to \mathbb{R}^{n+1}$ *in the following way:*

$$[(\Phi_n(F))\,u]_i = \begin{cases} n\,(u(t_i) - u(t_{i-1})) - f(u(t_{i-1})), & i = 1, \ldots, n, \\ \\ u(t_0) - u_0, & i = 0. \end{cases} \tag{7}$$

We note that the introduced notions of problem and numerical methods are independent of each other. However, for our purposes only those numerical methods $\mathscr{N}$ are interesting which are obtained when some discretization method $\mathscr{D}$ is applied to some certain problem $\mathscr{P}$.

Example 2.4. *Let us define the numerical method* $\mathscr{N}$ *for the problem* $\mathscr{P}$ *from Example 2.1, and for the discretization* $\mathscr{D}$ *from Example 2.3. Then we solve the sequence of problems in the form* (5)*, where in the discretization for* g *and* c *we put* f *and* u_0 *from problem* (3)-(4)*, respectively. This yields that the mapping* $F_n : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ *is defined as follows: for the vector* $\mathbf{v}_n = (v_0, v_1, \ldots, v_n) \in \mathbb{R}^{n+1}$ *we have*

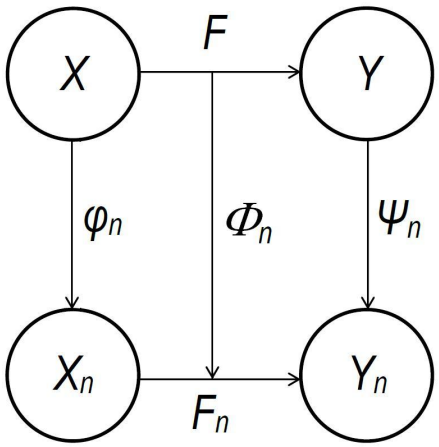


Figure 1: The general scheme of numerical methods.

$$\left(F_n(\mathbf{v})\right)_i = \begin{cases} n\,(v_i - v_{i-1}) - f(v_{i-1}), & i = 1, \ldots, n, \\ \\ v_0 - u_0, & i = 0. \end{cases} \tag{8}$$

Hence, the equation (5) *for* (8) *results in the task: we seek the vector* $\mathbf{v} = (v_0, v_1, \ldots, v_n) \in \mathbb{R}^{n+1}$ *such that*

$$\begin{cases} \dfrac{v_i - v_{i-1}}{1/n} = f(v_{i-1}), & i = 1, \ldots, n, \\ \\ v_0 = u_0, & i = 0. \end{cases} \tag{9}$$

Hence, the obtained numerical method is the well-known explicit Euler method on the mesh $\mathbb{G}_n$ *with step-size* $1/n$.

In sequel for the discretization $\mathscr{D} = (\varphi_n, \psi_n, \Phi_n)_{n\in\mathbb{N}}$ we assume the validity of the following assumption.

Assumption 2.1. *The discretization* $\mathscr{D}$ *possesses the property* $\psi_n(0) = 0$.

Obviously, when ψ_n are linear operators, then this condition is automatically satisfied. We also list two further natural assumptions about the discretization, which will be used later.

Assumption 2.2. *The discretization* $\mathscr{D}$ *generates a numerical method* $\mathscr{N}$ *which possesses the property* $\dim \mathcal{X}_n = \dim \mathcal{Y}_n < \infty$.

Theoretically, the normed spaces $\mathcal{X}$ and $\mathcal{Y}$ in the definitions of the problem and of the discretization might be different. However the application of the discretization to the problem is possible only when these normed spaces are the same. In the sequel this will be always assumed.

Assumption 2.3. *Let us apply the discretization $\mathscr{D}$ to the problem $\mathscr{P}$. We assume that F_n is continuous on the ball $B_R(\varphi_n(\bar{u}))$.*

3 Basic notions

In this part we analyze the general framework of a numerical method (according to Figure 1). We apply a discretization $\mathscr{D}$ for some problem $\mathscr{P}$, then it results in a numerical method $\mathscr{N}$, which generates the sequence of problems (5). Our aim is to guarantee the existence of the solutions $\bar{u}_n$ and the closeness of these to $\bar{u}$. To this aim we define the distance between these elements, which will be called global discretization error. (Since these elements belong to different spaces, this is not straightforward.) Independently of the form of the definition of the global error, it is hardly applicable in practice, because the knowledge of the exact solution $\bar{u}$ is assumed. Therefore, we introduce some further notions (consistency, stability), which help us in getting information about the behavior of the global discretization error.

3.1 Convergence

The usual approach for the characterization of the distance of the elements $\bar{u}$ and $\bar{u}_n$ is their comparison in $\mathcal{X}_n$ in the following way.

Definition 3.1. *The element $e_n = \varphi_n(\bar{u}) - \bar{u}_n \in \mathcal{X}_n$ is called global discretization error.*

Clearly, our aim is to guarantee that the global discretization error is arbitrary small, by increasing n. That is, we require the following property.

Definition 3.2. *The discretization $\mathscr{D}$ applied to the problem $\mathscr{P}$ is called* convergent *if*

$$\lim \|e_n\|_{\mathcal{X}_n} = 0 \tag{10}$$

holds. When

$$\|e_n\|_{\mathcal{X}_n} = \mathcal{O}(n^{-p})$$

we say that the order of the convergence *is p.*

Remark 3.1. *It is possible to define the distance between the elements $\bar{u}$ and $\bar{u}_n$ in the space $\mathcal{X}$, with the help of an operator $\bar{\varphi}_n : \mathcal{X}_n \to \mathcal{X}$, by the quantity $\|\bar{u} - \bar{\varphi}_n \bar{u}_n\|_{\mathcal{X}}$. For such an approach see Figure 2.*

Here we assume that $\lim(\varphi_n \circ \bar{\varphi}_n)v = v$ for any $v \in \mathcal{X}$. We note that this relation does not mean that $\bar{\varphi}_n$ is the inverse of φ_n, because φ_n is not invertible, typically it represents some interpolation. In this approach the convergence means that the numerical sequence $\|\bar{u} - \bar{\varphi}_n \bar{u}_n\|_{\mathcal{X}}$ tends to zero. Because this approach requires an additional interpolation, and the choice of the interpolation may influence the rate of the convergence, therefore this kind of convergence is less common.

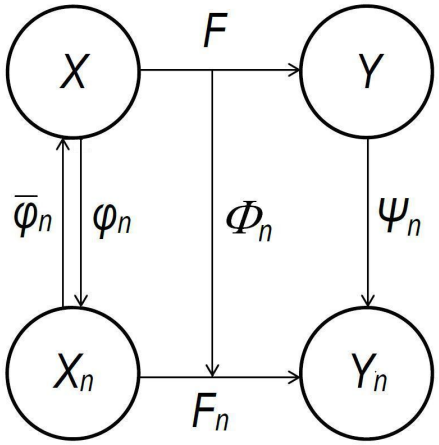


Figure 2: The general scheme of numerical methods with interpolation operator.

3.2 Consistency

Consistency is the notion which makes some connection between the problem $\mathscr{P}$ and the numerical method $\mathscr{N}$.

Definition 3.3. *The discretization $\mathscr{D}$ applied to problem $\mathscr{P}$ is called consistent at the element $v \in D$ if*

- *$\varphi_n(v) \in \mathcal{D}_n$ holds from some index,*
- *the relation*

$$\lim \|F_n(\varphi_n(v)) - \psi_n(F(v))\|_{\mathcal{Y}_n} = 0 \tag{11}$$

holds.

The element $l_n(v) = F_n(\varphi_n(v)) - \psi_n(F(v)) \in \mathcal{Y}_n$ in (11) plays an important role in the numerical analysis. When we fix some element $v \in \mathcal{D}$, we can transform it into the space in two different ways: $\mathcal{X} \to \mathcal{Y} \to \mathcal{Y}_n$ and $\mathcal{X} \to \mathcal{X}_n \to \mathcal{Y}_n$ (c.f. Figure 1). The magnitude $l_n(v)$ characterizes the difference of this two directions for the element v. Hence, the consistency at the element v yields that in limit the diagram of Figure 1 is commutative. A special role is played by the behavior of $l_n(v)$ on the solution of the problem (1), that is the elements $l_n(\bar{u})$. Later on we will use the following notions.

Definition 3.4. *The element $l_n(v) = F_n(\varphi_n(v)) - \psi_n(F(v)) \in \mathcal{Y}_n$ is called local discretization error at the element v. The element $l_n(\bar{u}) = F_n(\varphi_n(\bar{u})) - \psi_n(F(\bar{u})) = F_n(\varphi_n(\bar{u}))$ is called* local discretization error. *When*

$$\|l_n(v)\|_{\mathcal{X}_n} = \mathcal{O}(n^{-p}),$$

we say that the order of the consistency at v is p.

Remark 3.2. *For simplicity, we will use the notation l_n for $l_n(\bar{u})$. In the sequel, the consistency on $\bar{u}$ and its order will be called consistency and order of consistency.*

Example 3.1. *Consider the explicit Euler method. We apply it to the initial value problem of Example 2.1, i.e., to the problem (3)-(4). Then for the local discretization error we obtain*

$$l_n(\bar{u})(t_i) = \|[F_n(\varphi_n(\bar{u}))](t_i)\|_{\mathcal{Y}_n} = \left\| \frac{\bar{u}(t_i) - \bar{u}(t_{i-1})}{1/n} - \bar{u}'(t_{i-1}) \right\|_{\mathcal{Y}_n} =$$

$$= \|n[\bar{u}(t_i) - \bar{u}(t_{i-1})] - \bar{u}'(t_{i-1})\|_{\mathcal{Y}_n} = \max_{1\le i\le n} |\bar{u}'((i-1)/n) - n[\bar{u}(i/n) - \bar{u}((i-1)/n)]| =$$

$$= \max_{1\le i\le n} |n \int_{(i-1)/n}^{i/n} [\bar{u}'((i-1)/n) - \bar{u}'(s)]\mathrm{ds}| \le n \max_{1\le i\le n} \int_{t_{i-1}}^{t_i} |\bar{u}'(t_{i-1}) - \bar{u}'(s)|\mathrm{ds}.$$

Assume that $M_2(\bar{u}) := \sup_{t\in(0,1)} |\bar{u}''(t)| < \infty$, then we get

$$l_n(\bar{u}) \le nM_2(\bar{u}) \frac{1}{2n^2} = \frac{M_2(\bar{u})}{2n}.$$

Hence, for the class of problems (3)-(4) with Lipschitz continuous right-hand side f, the explicit Euler method is consistent, and the order of the consistency equals one.

One might ask whether consistency implies convergence. The following simple example shows that this is not true in general.

Example 3.2. *Let us consider the case $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\varphi_n = \psi_n = identity$. Our aim is to solve the scalar equation $F(x) = 0$, where we assume that it has a unique solution $\bar{x} = 0$. We define the numerical method $\mathcal{N}$ as $F_n(x) = (1 - x)/n$. Clearly, due to the linearity of φ_n and ψ_n, we have $l_n = F_n(0) - 0 = F_n(0)$. Since $F_n(0) \to 0$, therefore this discretization is consistent. However, it is not convergent, since the solution of each problem $F_n(x) = 0$ is $\bar{x}_n = 1$.*

Thus, convergence cannot be replaced by consistency in general.

4 Nonlinear stability

Generally, consistency in itself is not enough for convergence. In numerical analysis one of the most important task is to guarantee the convergence of the sequence of the numerical solutions. To guarantee this property we introduce the notion of stability.

Our main aim is to study how to define appropiately the notion of stability.

4.1 First attempt: N-stability

The convergence yields that e_n tends to zero. Moreover, for the consistent methods we have information about the behaviour of the local discretization error, only. Intuitively, this means the following requirement. When $l_n(\bar{u}) = F_n(\varphi_n(\bar{u})) - F(\bar{u}_n)$ is small, then $e_n = \varphi(\bar{u}) - \bar{u}_n$ be small too. Because we don't know $\bar{u}$, in first approach we require this property for each pairs of the elements in $\mathcal{D}_n$.

This demand implies the requirement

$$\|z_n - w_n\|_{\mathcal{X}_n} \leq S \|F_n(z_n) - F_n(w_n)\|_{\mathcal{Y}_n} \tag{12}$$

holds for arbitrary $z_n,\ w_n \in \mathcal{D}_n$ and the stability constant is indepedent of the mesh-size parameter.

This idea leads to make the first attempt to define the nonlinear stability notion.

Definition 4.1. *The discretization $\mathscr{D}$ is called* N-stable *on the problem $\mathscr{P}$ if there exists positive stability constant S, such that for each $z_n, w_n \in \mathcal{D}_n$, the estimation (12) holds.*

Furthermore we will refer to this notion as the naiv stability (N-stability).

Example 4.1. *Consider the following periodic initial-value reaction-diffusion problem*

$$\partial_t u(x,t) = \partial_{xx} u(x,t) + f(s), \quad -\infty < x < \infty,\ 0 \leq t \leq T < \infty \tag{13}$$

$$u(x+1,t) = u(x,t), \quad -\infty < x < \infty,\ 0 \leq t \leq T < \infty \tag{14}$$

$$u(x,0) = u^0(x), \quad -\infty < x < \infty. \tag{15}$$

In (13), f is a smooth real function of the real variable s, $-\infty < s < \infty$. In (15), u^0 is a given real one-periodic function and it is assumed that f, T and u^0 are such that (13)-(15) possesses a unique smooth solution up to $t = T$. To set up the numerical scheme, choose a positive constant r and an integer $J > 2$. Set $h = 1/J$ and consider the grid points $x_j = jh$, where j an integer and the time levels $t_N = N\delta, \delta = rh^2, N = 0, \ldots, n = [T/\delta]$.

Then for $j = 1, \ldots, J$ and $N = 0, \ldots, n-1$

$$\frac{u_j^{N+1} - u_j^N}{\delta} - \frac{u_{j-1}^N - 2u_j^N + u_{j+1}^N}{h^2} - f(u_j^N) = 0, \tag{16}$$

where it is obviously understood that $u_0^N = u_J^N$ and $u_{J+1}^N = u_1^N$. Set

$$u_j^0 - u_0(x_j) = 0, \quad j = 1, \ldots, J. \tag{17}$$

Formulae (16)-(17) are cast in the format (5) as follows. Let $\mathcal{Z}_n$ denote the vector space of the grid functions $u = [u_1, \ldots, u_J]$ defined on $x_j : 1 \leq j \leq J$. For each N, all the numerical approximations U_j^N associated with the time level t_N form a vector u^N in $\mathcal{Z}_n$. Thus (16)-(17) may be rewritten

$$\frac{u^{N+1} - u^N}{\delta} - D^2 u^N - f(u^N) = 0, \quad N = 0, \ldots, n-1, \tag{18}$$

$$u^0 - u^0 = 0, \tag{19}$$

where $u^0 = [u^0(x_1), \ldots, u^0(x_J)]$ and D^2 is the standard matrix replacement of the second derivative operator with periodic boundary conditions, i.e.

$$D^2 = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 1 & -2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & -2 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & -2 \end{pmatrix}.$$

Choose $\mathcal{X}_n = \mathcal{D}_n = \mathcal{Y}_n$ equal to the product of $n+1$ copies $\mathcal{Z}_n \times \ldots \times \mathcal{Z}_n$. Thus $u_N := [u^0, \ldots, u^n]$ is a vector in $\mathcal{X}_n$ and (18)-(19) are clearly of the form (5) for a suitable choice of f_N. For $\varphi_n(\bar{u})$ the obvious choice is given by the vector of grid restrictions $[u_0, \ldots, u_n]$ of $\bar{u}$. In $\mathcal{Z}_n$ we use the maximum norm, in $\mathcal{X}_n$ we use the norm

$$\left\| [v^0, \ldots, v^n] \right\| = \max_n \left\| v_N \right\|, [v^0, \ldots, v^n] \in \mathcal{X}_n, \tag{20}$$

and in $\mathcal{Y}_n$ we use the norm

$$\left\| [\rho^0, \ldots, \rho^n] \right\| = \left\| \rho^0 \right\| + \sum_{N=1}^{n} \delta \left\| \rho^n \right\|, [\rho^0, \ldots, \rho^n] \in \mathcal{Y}_n. \tag{21}$$

We will prove that for globally Lipschitz function f (i.e. $|f(a) - f(b)| \leq L|a - b|$, for all $a, b \in \mathbb{R}$) the discretization is N-stable when $r \leq 1/2$.

Then, if $v_N = [v^0, \ldots, v^n], w_N = [w^0, \ldots, w^n]$, $f_n(v_N) = [\rho^0, \ldots, \rho^n]$ *and* $f_n(w_N) = [\sigma^0, \ldots, \sigma^n]$. *Then*

$$\frac{v^{N+1} - v^N}{\delta} - D^2 v^N - f(V^N) = \rho^{N+1}, N = 0, \ldots, n-1 \tag{22}$$

$$v^0 - u^0 = \rho^0 \tag{23}$$

$$\frac{w^{N+1} - w^N}{\delta} - D^2 w^N - f(w^N) = \sigma^{N+1}, N = 0, \ldots, n-1 \tag{24}$$

$$w^0 - u^0 = \sigma^0 \tag{25}$$

A method like (22) with $f \equiv 0$ *rewrite in the form*

$$v^{N+1} = C_N v^N + \delta \rho^{N+1}, \tag{26}$$

where $C_N = I + \delta D^2$ *is the transition matrix and* $\|C_N\| = 1$. *Subtract (24) from (22) and use (26), to obtain for* $N = 0, \ldots, n-1$,

$$v^{N+1} - w^{N+1} = C_N(v^N - w^N) + \delta[f(v^N) - f(w^N)] + \delta[\rho^{N+1} - \sigma^{N+1}]. \tag{27}$$

The globally Lipschitz property implies that (27) yields

$$\left\|v^{N+1} - w^{N+1}\right\| \leq (1 + \delta L) \left\|v^N - w^N\right\| + \delta \left\|\rho^{N+1} - \sigma^{N+1}\right\|.$$

A standard recursion leads to (12) with $S = e^{LT}$. *Thus, for* $r \leq 1/2$ *and* f *globally Lipschitz, the scheme is N-stable.*

4.1.1 A special case: linear stability

To study the nonlinear stability it is expedient to present the linear case, because no distinction on the linear or nonlinear character of (5). Let

$$F_n(u_n) = L_n(u_n) = 0\,, \quad n = 1, 2, \ldots\,, \tag{28}$$

where L_n is a linear oparator and $L_n : \mathcal{D}_n \to \mathcal{Y}_n$.

Consider the N-stability for linear case. It follows directly from (12).

Definition 4.2. *The discretization* $\mathscr{D}$ *is called* stable *on the problem* $\mathscr{P}$ *if there exist positive stability constant* S, *such that for each* $z_n \in \mathcal{D}_n$

$$\|z_n\|_{\mathcal{X}_n} \leq S \|L_n(z_n)\|_{\mathcal{Y}_n} \tag{29}$$

holds.

The bound (29) implies three basic properties:

i, In view of (29), we obtain the "basic theory of numerical analysis":

$$\text{Consistency + Stability = Convergence}$$

In fact, due to the linearity of L_n, by the choice $z_n = \varphi_n(\bar{u}) - \bar{u}_n$, we have

$$\|\varphi_n(\bar{u}) - \bar{u}_n\|_{\mathcal{X}_n} \leq S \, \|L_n(\varphi_n(\bar{u})) - L_n(\bar{u}_n)\|_{\mathcal{Y}_n} \tag{30}$$

which leads to the estimation

$$\|e_n\|_{\mathcal{X}_n} = \|\varphi_n(\bar{u}) - \bar{u}_n\|_{\mathcal{X}_n} \leq S \, \|L_n(\varphi_n(\bar{u}))\|_{\mathcal{Y}_n} = S \, \|l_n(\bar{u})\|_{\mathcal{Y}_n} \tag{31}$$

and for consistent methods, this implies the convergence.

ii, For any problems (28), the relation (29) shows that $L_n(z_n) = 0$ implies that $z_n = 0$, i.e., L_n^{-1}exists. Hence, the N-stability bound implies the existence and uniqueness of solutions of (28).

iii, Due to ii and (29), we have

$$\left\|L_n^{-1} w_n\right\|_{\mathcal{X}_n} \leq S \, \|w_n\|_{\mathcal{Y}_n}$$

for arbitrary $w_n \in \mathcal{Y}_n$. Therefore the uniform norm estimation holds for all n, i.e., $\|L_n^{-1}\|_{Lin(\mathcal{Y}_n, \mathcal{X}_n)} \leq S$.

Remark 4.1. *The "basic theory of numerical analysis" can be succesfully generalized for the nonhomogeneous linear equation $L(u) = f$.*
In this case, $F(u) \equiv L(u) - f$, hence $F(u) \equiv L(u) - f = 0$, and the discretization is

$$F_n(u_n) \equiv L_n(u_n) - f_n = 0,$$

where $f_n = \psi_n(f)$. Then,

$$l_n(\bar{u}) = F_n(\varphi_n(\bar{u})) - \psi_n(F(\bar{u})) = (L_n(\varphi_n(\bar{u})) - f_n) - \psi_n(L(\bar{u}) - f).$$

Assume that ψ_n is linear, hence Assumption 2.1 holds. Thus,

$$l_n(\bar{u}) = (L_n(\varphi_n(\bar{u})) - \psi_n(L(\bar{u}))) + (\psi_n(f) - f_n) = L_n(\varphi_n(\bar{u})) - \psi_n(L(\bar{u})).$$

Then, in view of (31), we get the statment.

The theory is more developed for linear problems, see [LR56, PS84a, PS84b, PS85]. Namely, there are two well-celebrated theorems: Lax-Richtmyer-Kantorovich equivalence theorem for linear initial-value problems in partial differential equations and Palencia and Sanz-Serna's extension theorem.

4.2 T-stability and its application

In this subsection we introduce a new stability definition, which diverge of the previous ones. First we consider the Trenogin's stability definition.

Definition 4.3. *The discretization $\mathscr{D}$ is called* stable in Trenogin's sense *(T-stable) if there exists a continuous, strictly monotonically increasing function $\omega(s)$, defined for $s \geq 0$, such that $\omega(0) = 0$ and $\omega(\infty) = \infty$, and*

$$\omega\left(\left\|v_n^1 - v_n^2\right\|_{\mathcal{X}_n}\right) \leq \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{\mathcal{Y}_n}$$

holds for all $v_n^1, v_n^2 \in \mathcal{D}_n$.

Remark 4.2. *If we choose $\omega(s) = s/S$, then Definition 4.3 results in the notion of N-stability.*

In sense of Definition 4.3 theoretical results can be given.

Definition 4.4. *The sequence of $\|\cdot\|_{\mathcal{X}_n}$ norms is called consistent to the norm $\|\cdot\|_{\mathcal{X}}$, when for arbitrary $v \in \mathcal{X}$ the relation*

$$\lim \|\varphi_n(v)\|_{\mathcal{X}_n} = \|v\|_{\mathcal{X}} \tag{32}$$

holds.

Remark 4.3. *In most cases this condition is automatically satisfied. For Example 2.1 it is obviously true.*

Lemma 4.1. *When the sequence of $\|\cdot\|_{\mathcal{X}_n}$ norms is consistent, then the relation $\lim \|\varphi_n(v)\|_{\mathcal{X}_n} = 0$ implies that $v = 0$.*

Proof. We consider two cases.

i, If $v = 0$, then $\lim \|\varphi_n(v)\|_{\mathcal{X}_n} = \|v\|_{\mathcal{X}} = 0$;

ii, If $\lim \|\varphi_n(v)\|_{\mathcal{X}_n} = 0$, then $\|v\|_{\mathcal{X}} = 0$. Hence, $v = 0$.

□

Theorem 4.1. *Suppose that*

- *the sequence of $\|\cdot\|_{\mathcal{X}_n}$ norms is consistent,*
- *there exists the solution of the problem (1) and (5),*
- *the discretization $\mathscr{D}$ is consistent any solution $\bar{u}$ and it is T-stable.*

Then

i, $\bar{u}$ *is unique,*

ii, $\bar{u}_{n,\ n\in\mathbb{N}}$ *are unique,*

iii, the numerical method is convergent.

Proof. i, Let v_1, v_2 be solutions of (1) and for these elements, due to the consistency, the relations

$$\lim_{n\to\infty} \|F_n(\varphi_n(v_1))\|_{\mathcal{Y}_n} = 0; \lim_{n\to\infty} \|F_n(\varphi_n(v_2))\|_{\mathcal{Y}_n} = 0$$

hold. Then $\|\varphi_n(v_1 - v_2)\|_{\mathcal{X}_n} \leq \omega^{-1}(\|F_n(\varphi_n(v_1)) - F_n(\varphi_n(v_2))\|)_{\mathcal{Y}_n} \leq$

$\leq \omega^{-1}(\|F_n(\varphi_n(v_1))\|_{\mathcal{Y}_n} + \|F_n(\varphi_n(v_2))\|_{\mathcal{Y}_n}) \to 0$, for $n \to \infty$. Hence, we get

$$\lim_{n\to\infty} \|\varphi_n(v_1 - v_2)\|_{\mathcal{X}_n} = 0.$$

Due to Lemma 4.1 we gain $\|v_1 - v_2\|_{\mathcal{X}} = 0$. This relation implies, that the solution is unique.

ii, Assume that, v_1^n and v_2^n are solutions of (5). Then the relation in Definition 4.3 implies $0 \geq \omega(\|v_1^n - v_2^n\|_{\mathcal{X}_n})$. Since $\omega(\|v_1^n - v_2^n\|_{\mathcal{X}_n}) \geq 0$, so $\omega(\|v_1^n - v_2^n\|_{\mathcal{X}_n}) = 0$. Moreover, ω is strictly monotonically increasing function and $\omega(0) = 0$, therefore $\|v_1^n - v_2^n\|_{\mathcal{X}_n} = 0$. It implies $v_1^n = v_2^n$.

iii, From the Definition 4.3, from the continiuty of the function ω^{-1} at $t = 0$ and from (5) we gain

$$\|v_1^n - \varphi_n(v_1)\|_{\mathcal{X}_n} \leq \omega^{-1}(\|F_n(v_1^n) - F_n(\varphi_n(v_1))\|)_{\mathcal{Y}_n} = \omega^{-1}(\|F_n(\varphi_n(v_1))\|)_{\mathcal{Y}_n},$$

and the last term converges to 0 as $n \to \infty$.

□

Remark 4.4. *The property of Lemma 4.1 is the norm regularity and we say that* $\mathcal{X}_n,\ n \in \mathbb{N}$ *normed spaces are regularly normed.*

In the following part we revisited Definition 4.3 from the application point of view.

4.2.1 How to verify the T-stability?

To verify the T-stability of the problem (3)-(4) we consider the equation

$$F_n(x_n + z_n) - F_n(x_n) = y_n, \tag{33}$$

where x_n elements are parameters, z_n are unknows. If we can give an estimation in the form of

$$\|z_n\|_{\mathcal{X}_n} \leq \zeta(\|y_n\|)_{\mathcal{Y}_n}, \tag{34}$$

where the properties of $\zeta(s)$ correspond with the properties of $\omega(s)$, then by choice $\omega(s) := \zeta^{-1}(s)$ we prove the T-stability.

Let in (33) $x_n = x_1^n$, while $x_n + z_n = x_2^n$. Then $F_n(x_2^n) - F_n(x_1^n) = y_n$ and $x_2^n - x_1^n = z_n$. Arisen from the estimation (34)

$$\|x_2^n - x_1^n\|_{\mathcal{X}_n} \leq \zeta(\|F_n(x_2^n) - F_n(x_1^n)\|_{\mathcal{Y}_n}).$$

Because of the inverse of ζ exist and strictly monotonically increasing, we can write that

$$\zeta^{-1}(\|x_2^n - x_1^n\|_{\mathcal{X}_n}) \leq \|F_n(x_2^n) - F_n(x_1^n)\|_{\mathcal{Y}_n}.$$

This matches the stability estimation in Definition 4.3.

If the right-hand side of the equation is in the form $f(x,t)$, then we demand the condition $|f_x(x,t)| < L$, where L is the Lipschitz constant. To check the stability actually we can show that the estimation

$$\|z_n\|_{\mathcal{X}_n} \leq c\,\|y_n\|_{\mathcal{Y}_n} = c(\max_{1\leq i\leq n} |y_i| + |u_0|) \tag{35}$$

holds. Substituting F_n to (33), we gain

$$(y_n)_i = \begin{cases} n(z_i - z_{i-1}) + f(t_{i-1}, x_{i-1} + z_{i-1}) - f(t_{i-1}, x_{i-1}) = y_{i-1}, & i = 1, \ldots, n, \\ z_0 = u_0, & i = 0. \end{cases} \tag{36}$$

Then we can write the equation in the form:

$$z_i = (1 - hL_{i-1})z_{i-1} + hy_{i-1},\ i = 1, \ldots, n,$$

and $z_0 = y_0$, where

$$L_{i-1} = \int_0^1 f_x(t_{i-1}, x_{i-1} + \theta z_{i-1})d\theta.$$

Furthermore we know that $|L_{i-1}| \leq L,\ 1 \leq i \leq n$. Make these knowledge of use for z_i we get

$$|z_i| \leq (1 + Lh)|z_{i-1}| + h|y_{i-1}|, \quad 1 \leq i \leq n.$$

For all i index write out these, respectively apply recursively the above equation we gain, that

$$|z_1| \leq (1 + Lh)|u_0| + h\,\|y_n\|_\infty,$$

$$|z_2| \leq (1+Lh)^2|u_0| + [(1+Lh)+1]h\,\|y_n\|_\infty\,,$$
$$\vdots$$
$$|z_n| \leq (1+Lh)^n|u_0| + \sum_{k=1}^{n-1}(1+Lh)^k h\,\|y_n\|_\infty\,.$$

Estimating the following two terms:

$$(1+Lh)^k = \left(1+\frac{L}{n}\right)^k \leq \left(1+\frac{L}{n}\right)^n \leq e^L,$$

$$\sum_{k=0}^{n}\left(1+\frac{L}{n}\right)^k\frac{1}{n} = \frac{\left(1+\frac{L}{n}\right)^n-1}{\left(1+\frac{L}{n}\right)-1}\cdot\frac{1}{n} \leq \frac{e^L-1}{L}.$$

Then we get for the norm of z_n the following estimation:

$$\|z_n\|_{\mathcal{X}_n} \leq e^L|u_0| + \frac{e^L-1}{L}\,\|y_n\|_\infty\,.$$

Hence, with the choice $c = \max(e^L, \frac{e^L-1}{L})$ we can write an estimation in the form (35), so we prove the discretization is T-stable.

4.3 Local stability notions

In the paper of [LS88a] López-Marcos and Sans-Serna investigated the properties of the N-stability. We gain insight the following example to understand why have to introduce the so-called local stability notions.

Example 4.2. *Let $z \in \mathbb{R}^{n+1}$ be an arbitrary vector and $F_n : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is an operator. We define the following operator*

$$[F_n^\alpha(z)]_k = \begin{cases} \dfrac{z_k - z_{k-1}}{h} - z_{k-1}^2, & k = 1,2,\ldots,n \\ \\ z_0 - \alpha, & k = 0, \end{cases} \tag{37}$$

where h is the mesh-size parameter and $\alpha \in [0,1)$ is a fixed parameter. Taking the $\bar{v}^\alpha(t) = \alpha/[1-\alpha t]$ function, where $t \in [0,1]$ and applying the φ_n grid restriction to $\bar{v}^\alpha(t)$ we get

$$[\varphi_n(\bar{v}^\alpha)]_k \equiv (\bar{v}_n^\alpha)_k \equiv \bar{v}^\alpha(t_k) \equiv \frac{\alpha}{1-\alpha t_k}, \qquad k = 0,1,\ldots,n,$$

where t_k are the node points.

Remark 4.5. *We mentioned that the discretization (37) is the application of the explicit Euler's rule to the problem*

$$\begin{cases} u'(t) = u^2(t), \quad t \in [0,1] \\ \\ u(0) = \alpha, \end{cases} \tag{38}$$

with the solution $u(t) = \alpha/[1-\alpha t]$.

Substituiting $\bar{v}_n^\alpha$ *into (37), we gain*

$$[F_n^\alpha(\bar{v}_n^\alpha)]_k = \begin{cases} \dfrac{\bar{v}^\alpha(t_k) - \bar{v}^\alpha(t_{k-1})}{h} - [\bar{v}^\alpha(t_{k-1})]^2, & k = 1,2,\ldots,n \\ \\ (\bar{v}_n^\alpha)_0 - \alpha, & k = 0. \end{cases} \tag{39}$$

Let $\bar{w}_n \in \mathbb{R}^{n+1}$ *be a vector with the components* w_k*, such that* $[F_n(\bar{w}_n)] = 0$*, where*

$$[F_n(\bar{w}_n)]_k = \begin{cases} \dfrac{w_k - w_{k-1}}{h} - w_{k-1}^2, & k = 1,2,\ldots,n \\ \\ w_0 - 1, & k = 0. \end{cases} \tag{40}$$

Introduce the following norms

$$\|x_k\|_{\mathcal{X}_n} = \max_{1 \le k \le n+1} |x_k|,$$

$$\|y_k\|_{\mathcal{Y}_n} = |y_0| + \sum_{k=1}^{n} h|y_k|.$$

We prove that there doesn't exist an independent stability constant S *in the estimation (12), thus the discretization won't be N-stable. We have to show that the estimation*

$$\|\bar{v}_n^\alpha - \bar{w}_n\|_{\mathcal{X}_n} \le S \, \|F_n^\alpha(\bar{v}_n^\alpha) - F_n(\bar{w}_n)\|_{\mathcal{Y}_n} \tag{41}$$

cannot be hold for all n*. Due to [SV86] the value* $(\bar{w}_n)$ *corresponding to the last grid point* $t = 1$ *in* $[0,1]$ *behaves like* $1/(h|\ln h|)$*. Thus,*

$$\lim_{n\to\infty} (\bar{w}_n)_n = \lim_{h\to 0} \frac{1}{h|\ln h|} = \infty.$$

Now, we will verify the estimation (41) cannot hold. Since $(\bar{v}_n^\alpha)_n \equiv \alpha/[1-\alpha]$ *and* $\alpha \in [0,1)$*, hence the value of* $(\bar{v}_n^\alpha)_n$ *is finite. So the left term of (41) converges to* ∞ *as* $n \to \infty$*, i.e.,*

$$\lim_{n\to\infty} \|\bar{v}_n^\alpha - \bar{w}_n\|_{\mathcal{X}_n} = \infty. \tag{42}$$

The first step to check the other term of (41) is that, we have to write the distinction of the convenient grid functions.

$$[F_n^\alpha(\bar{v}_n^\alpha) - F_n(\bar{w}_n)]_k = \begin{cases} \dfrac{\bar{v}^\alpha(t_k) - \bar{v}^\alpha(t_{k-1})}{h} - [\bar{v}^\alpha(t_{k-1})]^2, & k = 1, 2, \dots, n \\ \alpha - 1, & k = 0. \end{cases} \tag{43}$$

Using the introduced norm in $\mathcal{Y}_n$ to (43) and due to Example 3.1, we get

$$\|F_n^\alpha(\bar{v}_n^\alpha) - F_n(\bar{w}_n)\|_{\mathcal{Y}_n} = |\alpha - 1| + \sum_{k=1}^{n} h \cdot l_n(\bar{v}^\alpha(t_k)) \leq \frac{M_2(\bar{v}^\alpha)}{2}.$$

Thus,

$$\lim_{n\to\infty} \|F_n^\alpha(\bar{v}_n^\alpha) - F_n(\bar{w}_n)\|_{\mathcal{Y}_n} < \infty. \tag{44}$$

From (42) and (44) we can see the estimation (41) cannot hold. Thus, the discretization is not N-stable.

Remark 4.6. *The Example 4.2 also shows us, if $\bar{w}_n$ is far from $\bar{v}_n^\alpha$ (i.e., the perturbation $\bar{v}_n^\alpha$ is too large), we should not give an estimation it is results in the form (12). This consideration leads to motivate the notion of stability threshold.*

Among others Example 4.2 shows us the N-stability definition is too restrictive, because we require the condition (12) for any elements from $\mathcal{D}_n$. However, as we will see, it is enough to guarantee similar properties only for the elements from some smaller subdomain. Furthermore this introduces the important idea of local stability and stability threshold notions.

4.3.1 locT-, K- and S-stability notions

Definition 4.5. *The discretization $\mathscr{D}$ is called* locally T-stable *(locT-stable) if there exists a function $\omega(s)$ defined at $\mathcal{K}_{\tilde{R}}(0) \subset \mathcal{X}_n$ (some neighbourhood of zero), which is continuous and strictly monotonically increasing, such that $\omega(0) = 0$ and*

$$\omega\left(\left\|v_n^1 - v_n^2\right\|_{\mathcal{X}_n}\right) \leq \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{\mathcal{Y}_n} \tag{45}$$

holds for all $v_n^1, v_n^2 \in \mathcal{K}_{\tilde{R}}(0)$.

Remark 4.7. *The relation $\|v_n^1 - v_n^2\|_{\mathcal{X}_n} \leq \omega^{-1}\left(\|F_n(v_n^1) - F_n(v_n^2)\|_{\mathcal{Y}_n}\right)$ from the estimation (45) is obviously true.*

We are looking for the answer with this definition what type of theoretical result can be given.

Theorem 4.2. *Suppose that*

- *the sequence of $\|\cdot\|_{\mathcal{X}_n}$ norms is consistent,*
- *there exists the solution of the problem (1) and (5),*
- *the discretization $\mathscr{D}$ is consistent any solution $\bar{u}$ and it is locT-stable.*

Then the numerical method is convergent.

Proof. From the Definition 4.5, from the continiuty of the function ω^{-1} at $t = 0$ and from (5) we gain

$$\|v_1^n - \varphi_n(v_1)\|_{\mathcal{X}_n} \leq \omega^{-1}(\|F_n(v_1^n) - F_n(\varphi_n(v_1))\|)_{\mathcal{Y}_n} = \omega^{-1}(\|F_n(\varphi_n(v_1))\|)_{\mathcal{Y}_n},$$

and the last term converges to 0 as $n \to \infty$. □

Definition 4.6. *The discretization $\mathscr{D}$ is called* stable in Keller' sense *(K-stable) for problem $\mathscr{P}$ if there exist $S \in \mathbb{R}$, $R \in (0, \infty]$ such that*

- *$B_R(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ holds from some index;*
- *$\forall (v_n^1)_{n\in\mathbb{N}}, (v_n^2)_{n\in\mathbb{N}}$ which satisfy $v_n^i \in B_R(\varphi_n(\bar{u}))$ $(i = 1, 2)$, the estimate*

$$\left\|v_n^1 - v_n^2\right\|_{\mathcal{X}_n} \leq S \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{\mathcal{Y}_n} \tag{46}$$

holds.

Remark 4.8. *The constant S in Definition 4.6 may depend on $\bar{u}$.*

Remark 4.9. *Let $R > 0$ fixed. Then as we have seen in the Example 4.2 the condition $\bar{v}_n^\alpha, \bar{w}_n \in B_R(\bar{v}_n^\alpha)$ cannot be guaranteed. However, if we require the stability condition only for the elements from $B_R(\bar{v}_n^\alpha)$ (that is the stability notion in Definition 4.6), then the condition (46) is satisfied.*

Corollary 4.1. *If the discretization $\mathscr{D}$ is stable on problem $\mathscr{P}$ at the element $v \in \mathcal{X}$ with stability threshold R, then F_n is injective on $B_R(\varphi_n(v))$ from some index.*

We give less restrictive definition than the Definition 4.6.

Definition 4.7. *The discretization $\mathscr{D}$ is called* stable in Stetter's sense *(S-stable) for problem $\mathscr{P}$ if there exist $S \in \mathbb{R}$, $R \in (0, \infty]$ and $r \in (0, \infty]$ such that*

- *$B_R(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ holds from some index;*

- *for all* $(v_n^1)_{n\in\mathbb{N}}, (v_n^2)_{n\in\mathbb{N}}$ *from* $B_R(\varphi_n(\bar{u}))$*, such that* $F_n(v_n^i) \in B_r(F_n(\varphi_n(\bar{u})))$ $(i = 1, 2)$*, the estimate*

$$\left\|v_n^1 - v_n^2\right\|_{\mathcal{X}_n} \leq S \left\|F_n(v_n^1) - F_n(v_n^2)\right\|_{\mathcal{Y}_n}$$

holds.

Example 4.3. *In [FMF11] we showed that the explicit Euler method is S-stable on the problem (3)-(4) with* $S = e^L$ *and* $R = 1$*.*

Remark 4.10. *If we put* $r = \infty$ *in Definition 4.7, then we re-obtain the stability definition by Keller.*

Remark 4.11. *If we choose* $\omega(s) = s/S$ *in Definition 4.5,* $R = \tilde{R}$ *in Definition 4.6 and*

i, $B_{\tilde{R}}(\varphi_n(\bar{u})) \subset \mathcal{K}_{\tilde{R}}(0)$ *and* $\forall(v_n^i)_{n\in\mathbb{N}}$ *which satisfy* $v_n^i \in B_{\tilde{R}}(\varphi_n(\bar{u}))$ $(i = 1, 2)$*;*

ii, the previous two properties and $F_n(v_n^i) \in B_r(F_n(\varphi_n(\bar{u})))$ $(i = 1, 2)$*;*

then we re-obtain K- and S-stability.

4.3.2 Theoretical results

In this subsection we will see the advantages of the restricted Keller's local stability notion. The first step in this direction is done by introducing a simplified form of the notion of semistability in [LS88b].

Definition 4.8. *The discretization* $\mathscr{D}$ *is called* semistable *on the problem* $\mathscr{P}$ *if there exist* $S \in \mathbb{R}$*,* $R \in (0, \infty]$ *such that*

- $B_R(\varphi_n(\bar{u})) \subset \mathcal{D}_n$ *holds from some index;*
- $\forall(v_n)_{n\in\mathbb{N}}$ *which satisfy* $v_n \in B_R(\varphi_n(\bar{u}))$ *from that index, the relation*

$$\|\varphi_n(\bar{u}) - v_n\|_{\mathcal{X}_n} \leq S \|F_n(\varphi_n(\bar{u})) - F_n(v_n)\|_{\mathcal{Y}_n} \tag{47}$$

holds.

Semistability is a purely theoretical notion, which, similarly as the consistency, cannot be checked directly, due to the fact, that $\bar{u}$ is unknown. However, the following statement clearly shows the relation of the three important notions.

Lemma 4.2. *We assume that the discretization $\mathscr{D}$*

- *is consistent at $\bar{u}$ and semistable with stability threshold R on the problem $\mathscr{P}$;*
- *generates a numerical method $\mathscr{N}$ that Equation (5) has a solution in $B_R(\varphi_n(\bar{u}))$ from some index.*

Then the sequence of these solutions of Equation (5) converges to the solution of problem $\mathscr{P}$, and the order of convergence is not less than the order of consistency.

Proof. Having the relation $F_n(\bar{u}_n) = \psi_n(F(\bar{u})) = 0$, we get

$$\|\varphi_n(\bar{u}) - \bar{u}_n\|_{\mathcal{X}_n} \leq S \, \|F_n(\varphi_n(\bar{u})) - F_n(\bar{u}_n)\|_{\mathcal{Y}_n} = S \, \|F_n(\varphi_n(\bar{u})) - \psi_n(F(\bar{u}))\|_{\mathcal{Y}_n} \ .$$

This yields that $\|e_n\|_{\mathcal{X}_n} \leq S\|l_n\|_{\mathcal{Y}_n}$, which proves the statement. □

This lemma has some drawbacks. First, we cannot verify its conditions because this requires the knowledge of the solution. Secondly, we have no guarantee that equation (5) has a (possibly unique) solution in $B_R(\varphi_n(\bar{u}))$ from some index. The introduced K-stability notion gets rid of the second problem.

Remark 4.12. *Obviously, the stability on the solution of problem* (1) *(i.e., at the element $\bar{u} \in \mathcal{X}$) implies the semistability.*

The following statements demonstrate the usefulness of the stability notion, given in Definition 4.6.

Lemma 4.3. *We assume that*

- *$\mathcal{V}$, $\mathcal{W}$ are normed spaces with the property $\dim \mathcal{V} = \dim \mathcal{W} < \infty$;*
- *$G : B_R(v) \to \mathcal{W}$ is continuous, where $B_R(v) \subset \mathcal{V}$ is a ball for some $v \in \mathcal{V}$ and $R \in (0, \infty]$;*
- *for all v^1, v^2 which satisfy $v^i \in B_R(v)$, the stability estimate*

$$\left\|v^1 - v^2\right\|_{\mathcal{V}} \leq S \left\|G(v^1) - G(v^2)\right\|_{\mathcal{W}} \tag{48}$$

holds.

Then

- *G is invertible, and $G^{-1} : B_{R/S}(G(v)) \to B_R(v)$;*
- *G^{-1} is Lipschitz continuous with the constant S.*

Proof. It is enough to show that $B_{R/S}(G(v)) \subset G(B_R(v))$, due to Corollary 4.1. We assume indirectly that there exists $w \in B_{R/S}(G(v))$ such that $w \notin G(B_R(v))$. We define the line $w(\lambda) = (1-\lambda)G(v) + \lambda w$ for $\lambda \geq 0$, and introduce the number $\hat{\lambda}$ as follows:

$$\hat{\lambda} := \begin{cases} \sup\{\lambda' > 0 \,|\, w(\lambda) \in G(B_R(v)) \, \forall \lambda \in [0, \lambda')\} \ , \text{ if it exists,} \\ 0 \, , \text{ else.} \end{cases}$$

Then clearly the inequality $\hat{\lambda} \leq 1$ holds. We will show that $\hat{w} =: w(\hat{\lambda}) \in G(B_R(v))$.

For $\hat{\lambda} = 0$ this trivially holds. For $\hat{\lambda} > 0$ we observe that G is invertible on $w(\hat{\lambda} - \varepsilon)$, (i.e., the operators $G^{-1}(w(\hat{\lambda} - \varepsilon)) \in B_R(v)$ exist) for all $\varepsilon : \ \hat{\lambda} \geq \varepsilon > 0$. Thus, we can use the stability estimate (48)

$$\begin{aligned} \left\| G^{-1}(w(\hat{\lambda} - \varepsilon)) - v \right\|_{\mathcal{V}} \leq S \left\| w(\hat{\lambda} - \varepsilon) - G(v) \right\|_{\mathcal{W}} = \\ S(\hat{\lambda} - \varepsilon) \underbrace{\|w - G(v)\|_{\mathcal{W}}}_{= \frac{R}{S} - \frac{\delta}{S}} < \hat{\lambda}(R - \delta) \leq R - \delta \, , \end{aligned}$$

for some $\delta > 0$, and using again the stability estimate we can conclude that the function $h(\varepsilon) = G^{-1}(w(\hat{\lambda} - \varepsilon))$ is uniformly continuous at $\varepsilon \in (0, \hat{\lambda}]$. Thus, there exists $\lim_{\varepsilon \searrow 0} h(\varepsilon) =: z \in B_R(v)$. Using the continuity of G, we get $G(z) = \hat{w}$.

Now we can choose a closed ball $\bar{B}_r(z) \subset B_R(v)$, $(r > 0)$ whose image $G(\bar{B}_r(z))$ contains a neighborhood of $\hat{w}$, due to the Brouwer's invariance domain theorem. This results in a contradiction.

Finally, the Lipschitz continuity with the constant S is a simple consequence of (48). □

Lemma 4.4. *For the discretization $\mathscr{D}$ we assume that*

- *it is consistent and K-stable at $\bar{u}$ with stability threshold R and constant S on problem $\mathscr{P}$;*
- *Assumptions 2.2 and 2.3 are satisfied.*

Then the discretization $\mathscr{D}$ generates a numerical method $\mathscr{N}$ such that equation (5) has a unique solution in $B_R(\varphi_n(\bar{u}))$, from some index.

Proof. Due to Lemma 4.3, F_n is invertible, and $F_n^{-1} : B_{R/S}(F_n(\varphi_n(\bar{u}))) \to B_R(\varphi_n(\bar{u}))$. Note that $F_n(\varphi_n(\bar{u})) = l_n \to 0$, due to the consistency. This means that $0 \in B_{\frac{R}{S}}(F_n(\varphi_n(\bar{u})))$, from some index. This proves the statement. □

Theorem 4.3. *We assume that*

- *the discretization $\mathscr{D}$ is consistent and K-stable at $\bar{u}$ with stability threshold R and constant S on problem $\mathscr{P}$;*
- *Assumptions 2.2 and 2.3 are true.*

Then the discretization $\mathscr{D}$ is convergent on problem $\mathscr{P}$, and the order of the convergence is not less than the order of consistency.

Proof. The statement is the consequence of Lemmas 4.4 and 4.2. □

5 Basic Notions – Revisited from the Application Point of View

Theorem 4.3 is not yet suitable for our purposes: the condition requires to check the stability and the consistency on the unknown element $\bar{u}$. Therefore, this statement is not applicable for real problems. Since we are able to verify the above properties on *some set of points* (sometimes on the entire $\mathcal{D}$), we extend the previously given pointwise (local) definitions to the set (global) ones.

Definition 5.1. *The discretization $\mathscr{D}$ is called* consistent *on problem $\mathscr{P}$ if there exists a set $\mathcal{D}_0 \subset \mathcal{D}$ whose image $F(\mathcal{D}_0)$ is dense in some neighborhood of the point $0 \in \mathcal{Y}$, and it is consistent at each element $v \in \mathcal{D}_0$.*

The order of the consistency in $\mathcal{D}_0$ is defined as $\inf\{p_v : v \in \mathcal{D}_0\}$, where p_v denotes the order of consistency at the point v.

Example 5.1. *Let us consider the explicit Euler method, given in Examples 2.2, 2.3 and 2.4. We apply it to the Cauchy problem of Example 2.1, i.e., to the problem (3)-(4). We verify the consistency and its order on the set $\mathcal{D}_0 \subset \mathcal{D}$, where $\mathcal{D} := C^1[0,1]$ and $\mathcal{D}_0 := C^2[0,1]$. Then for the local discretization error we obtain*

$$[F_n(\varphi_n(v)) - \psi_n(F(v))](t_i) = \begin{cases} \frac{1}{2n}v''(\theta_i) & i = 1, \dots, n, \\ 0, & i = 0, \end{cases} \tag{49}$$

where $\theta_i \in (t_{i-1}, t_i)$ are given numbers. Then $\|l_n(v)\|_{\mathcal{X}_n} = \mathcal{O}(n^{-1})$ from Definition 3.4.

Hence, for the class of problems (3)-(4) with Lipschitz continuous right-hand side f, the explicit Euler method is consistent, and the order of the consistency equals one.

In Section 3 (c.f. Example 3.2) we have shown that the pointwise consistency at the solution in itself is not enough for the convergence. One may think that the stronger notion of consistency, given by Definition 5.1, already ensures convergence. The following example shows that this is not true.

Example 5.2. *Let us choose the normed spaces as $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\varphi_n, = \psi_n = identity$. Our aim is to solve the scalar equation $F(x) = 0$, where the function $F \in C(\mathbb{R}, \mathbb{R})$ is given as follows*

$$F(x) = \begin{cases} |x|\ , \text{ if } x \in (-1,1)\,, \\ 1\ , \text{ if } x \in (-\infty,-1] \cup [1,\infty)\,. \end{cases}$$

Clearly this problem has a unique solution $\bar{x} = 0$. We define the numerical method $\mathscr{N}$ as

$$F_n(x) = \begin{cases} \frac{1}{n}\,, \text{ if } x \in \left[-\frac{1}{n}, \frac{1}{n}\right], \\ x\,, \text{ if } x \in \left(\frac{1}{n}, 1\right), \\ 1\,, \text{ if } x \in (-\infty, -1] \cup [1, n) \cup [n+2, \infty), \\ -x\,, \text{ if } x \in \left(-1, -\frac{1}{n}\right), \\ |x - (n+1)|\,, \text{ if } x \in [n, n+2). \end{cases}$$

For the given problem this discretization is consistent on the entire $\mathbb{R}$, however it is not convergent, since the solutions of the discrete problems $F_n(x) = 0$ are $\bar{x}_n = n + 1$ and therefore $\bar{x}_n \nrightarrow \bar{x}$.

In the sequel, besides the Assumptions 2.2, 2.3, which we have already made, we assume the validity of the following new assumptions.

Assumption 5.1. *For the problem $\mathscr{P}$ we assume that F^{-1} is continuous at the point $0 \in \mathcal{Y}$.*

Assumption 5.2. *Let us apply the discretization $\mathscr{D}$ to problem $\mathscr{P}$. We assume that discretization $\mathscr{D}$ possesses the property: there exists $K_1 > 0$ such that for all $v \in \mathcal{D}$ the relation*

$$\|\varphi_n(\bar{u}) - \varphi_n(v)\|_{\mathcal{X}_n} \leq K_1 \|\bar{u} - v\|_{\mathcal{X}}$$

holds for all $n \in \mathbb{N}$.

Assumption 5.3. *We assume that discretization $\mathscr{D}$ possesses the property: there exists $K_2 > 0$ such that for all $y \in \mathcal{Y}$ the relation*

$$\|\psi_n(y) - \psi_n(0)\|_{\mathcal{Y}_n} \leq K_2 \|y - 0\|_{\mathcal{Y}}$$

holds for all $n \in \mathbb{N}$.

For the simplicity of the formulation, the collection of the Assumptions 2.1–2.3 and 5.1–5.3 will be called Assumption $A^\star$.

Lemma 5.1. *Besides Assumption $A^\star$ we assume that*

- *the discretization $\mathscr{D}$ on problem $\mathscr{P}$ is consistent,*
- *the discretization $\mathscr{D}$ on problem $\mathscr{P}$ at the element $\bar{u}$ is stable with stability threshold R and constant S.*

Then F_n is invertible at the point $\psi_n(0)$, i.e., there exists $F_n^{-1}(\psi_n(0))$ for sufficiently large indices n.

Proof. We can choose a sequence $(y^k)_{k\in\mathbb{N}}$ such that $y^k \to 0 \in \mathcal{Y}$ and $F^{-1}\left(y^k\right) =: u^k \to \bar{u}$, due to the continuity of F^{-1}. Then the discretization $\mathscr{D}$ on problem $\mathscr{P}$ at the element u^k is stable with stability threshold $R/2$ and constant S, for some sufficiently large indices k. Moreover, F_n is continuous on $B_{R/2}(\varphi_n(u^k))$. Thus, for these indices k and also for sufficiently large n there exists F_n^{-1} : $B_{R/2S}(F_n(\varphi_n(u^k))) \to B_{R/2}(\varphi_n(u^k))$ moreover, it is Lipschitz continuous with constant S, according to Lemma 4.3. Let us write a trivial upper estimate:

$$\left\|F_n(\varphi_n(u^k))\right\|_{\mathcal{Y}_n} \leq \left\|F_n(\varphi_n(u^k)) - \psi_n(F(u^k))\right\|_{\mathcal{Y}_n} + \left\|\psi_n(F(u^k))\right\|_{\mathcal{Y}_n} .$$

Here the first term tends to 0 as $n \to \infty$, due to the consistency. For the second term, based on (5.3) we have the estimate $\left\|\psi_n(y^k)\right\|_{\mathcal{Y}_n} \leq K_2 \left\|y^k\right\|_{\mathcal{X}_n}$. Since the right-hand side tends to zero as $k \to \infty$, this means that the centre of the ball $B_{R/2}(F_n(\varphi_n(u^k)))$ tends to $0 \in \mathcal{Y}_n$, which proves the statement. □

Corollary 5.1. *Under the conditions of Lemma 5.1, for sufficiently large indices k and n, the following results are true.*

- *There exists $F_n^{-1}(\psi_n(y^k))$, since $\psi_n(y^k) \in B_{R/2S}(F_n(\varphi_n(u^k)))$.*
- *$F_n^{-1}(\psi_n(y^k)), \varphi_n(F^{-1}(y^k)) \in B_{R/2}(\varphi_n(\bar{u}))$.*

Analogously to the consistency, the stability can also be defined on a set of points. (This makes it possible to avoid the direct knowledge of the usually unknown $\bar{u}$.)

Definition 5.2. *The discretization $\mathscr{D}$ is called* stable *on problem $\mathscr{P}$ if there exist $S \in \mathbb{R}$, $R \in (0, \infty]$ and a set $\mathcal{D}_1 \subset \mathcal{D}$ such that $\bar{u} \in \mathcal{D}_1$ and it is stable at each point $v \in \mathcal{D}_1$ with stability threshold R and constant S.*

Now we are in the position to formulate our basic result, in which the notion of convergence is ensured by the notions of consistency and stability on a set, which can usually be verified directly, without knowing the exact solution of problem $\mathcal{P}$.

Theorem 5.1. *Besides the Assumption $A^\star$ we suppose that the discretization $\mathscr{D}$ on problem $\mathscr{P}$ is*

- *consistent;*
- *stable with stability threshold R and constant S.*

Then the discretization $\mathscr{D}$ is convergent on problem $\mathscr{P}$, and the order of the convergence can be estimated from below by the order of consistency on the corresponding set $\mathcal{D}_0$.

Proof. By use of the triangle inequality, we have

$$\begin{aligned}
&\|\varphi_n(\bar{u}) - \bar{u}_n\|_{\mathcal{X}_n} = \left\|\varphi_n(F^{-1}(0)) - F_n^{-1}(\psi_n(0))\right\|_{\mathcal{X}_n} \leq \\
&\underbrace{\left\|\varphi_n(F^{-1}(0)) - \varphi_n(F^{-1}(y^k))\right\|_{\mathcal{X}_n}}_{I.} + \\
&\underbrace{\left\|\varphi_n(F^{-1}(y^k)) - F_n^{-1}(\psi_n(y^k))\right\|_{\mathcal{X}_n}}_{II.} + \\
&\underbrace{\left\|F_n^{-1}(\psi_n(y^k)) - F_n^{-1}(\psi_n(0))\right\|_{\mathcal{X}_n}}_{III.},
\end{aligned} \tag{50}$$

where the elements $y^k \in \mathcal{Y}$ are defined in the proof of Lemma 5.1.

In the next step we estimate the different terms on the left-hand side of (50).

I. For the first term, based on Assumption 5.2, we have the estimate

$$\left\|\varphi_n(F^{-1}(0)) - \varphi_n(F^{-1}(y^k))\right\|_{\mathcal{X}_n} \leq K_1 \left\|F^{-1}(0) - F^{-1}(y^k)\right\|_{\mathcal{X}}.$$

Since $y^k \to 0$ as $k \to \infty$, and F^{-1} is continuous at the point $0 \in \mathcal{Y}$, therefore this term tends to zero, independently of n.

II. This term can be written as $\left\|F_n^{-1}(F_n(\varphi_n(F^{-1}(y^k)))) - F_n^{-1}(\psi_n(y^k))\right\|_{\mathcal{X}_n}$. Due to Corollary 5.1, we can use the stability estimate, therefore for this term we have the estimate

$$\begin{aligned}
&\left\|\varphi_n(F^{-1}(y^k)) - F_n^{-1}(\psi_n(y^k))\right\|_{\mathcal{X}_n} \leq \\
&S\left\|F_n(\varphi_n(F^{-1}(y^k))) - \psi_n(y^k)\right\|_{\mathcal{Y}_n} = S\left\|F_n(\varphi_n(u^k)) - \psi_n(F(u^k))\right\|_{\mathcal{Y}_n}.
\end{aligned}$$

In this estimate the term on the right-hand side tends to zero because of the consistency at u^k.

III. For the estimation of the third term we can use the Lipschitz continuity of F_n^{-1}, due to Lemma 5.1 and Corollary 5.1. Hence, by using the Assumption 5.3, we have

$$\left\|F_n^{-1}(\psi_n(y^k)) - F_n^{-1}(\psi_n(0))\right\|_{\mathcal{X}_n} \leq S\left\|\psi_n(y^k) - \psi_n(0)\right\|_{\mathcal{Y}_n} \leq SK_2\left\|y^k\right\|_{\mathcal{Y}}.$$

The right-hand side of the above estimate tends to zero, independently of the index n.

These estimations complete the proof. □

Example 5.3. *Let us analyze the stability property of the explicit Euler method, given in Example 2.4.*

Let $\mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathcal{X}_n = \mathbb{R}^{n+1}$ *be two arbitrary vectors, and we use the notation* $\epsilon = \mathbf{v}^{(1)} - \mathbf{v}^{(2)} \in \mathbb{R}^{n+1}$. *We define the vector* $\delta = F_n\left(\mathbf{v}^{(1)}\right) - F_n\left(\mathbf{v}^{(2)}\right) \in \mathbb{R}^{n+1}$, *where* F_n *is defined in (6). (In the notation, for simplicity, we omit the use of the subscript* n *for the vectors. We recall that the coordinates of the vectors are numbered from* $i = 0$ *until* $i = n$.*)*

For the coordinates of the vector δ *we have the following relations.*

- *For the first coordinate (*$i = 0$*) we obtain:*

$$\delta_0 = \left(F_n\left(\mathbf{v}^{(1)}\right)\right)_0 - \left(F_n\left(\mathbf{v}^{(2)}\right)\right)_0 = \left(v_0^{(1)} - u_0\right) - \left(v_0^{(2)} - u_0\right) = \epsilon_0.$$

- *For the other coordinates* $i = 1, \ldots, n$ *we have*

$$\begin{gathered}\delta_i = v_i^{(1)} - v_i^{(2)} = \\ n(v_i^{(1)} - v_{i-1}^{(1)}) - f(v_{i-1}^{(1)}) - n(v_i^{(2)} - v_{i-1}^{(2)}) + f(v_{i-1}^{(2)}) = \\ n(v_i^{(1)} - v_i^{(2)}) - n(v_{i-1}^{(1)} - v_{i-1}^{(2)}) - (f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})) = \\ n\epsilon_i - n\epsilon_{i-1} - (f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})).\end{gathered}$$

We can express ϵ_i *from this relation as follows:*

$$\epsilon_i = \epsilon_{i-1} + \frac{1}{n}(f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})) + \frac{1}{n}\delta_i.$$

Under our assumption $f \in C(\mathbb{R}, \mathbb{R})$ *is a Lipschitz continuous function, therefore we have the estimation* $|f(v_{i-1}^{(1)}) - f(v_{i-1}^{(2)})| \leq L|v_{i-1}^{(1)} - v_{i-1}^{(2)}|$. *Hence, we get*

$$|\epsilon_i| \leq |\epsilon_{i-1}| + \frac{1}{n}L|v_{i-1}^{(1)} - v_{i-1}^{(2)}| + \frac{1}{n}|\delta_i| = |\epsilon_{i-1}|\left(1 + \frac{L}{n}\right) + \frac{1}{n}|\delta_i|.$$

If we apply this estimate consecutively to $|\epsilon_{i-1}|$, $|\epsilon_{i-2}|$, *etc., we obtain:*

$$\begin{gathered}|\epsilon_i| \leq |\epsilon_{i-2}|\left(1 + \frac{L}{n}\right)^2 + \frac{1}{n}|\delta_i| + \left(1 + \frac{L}{n}\right)\frac{1}{n}|\delta_{i-1}| \leq \ldots \\ |\epsilon_0|\left(1 + \frac{L}{n}\right)^n + \frac{1}{n}\sum_{i=1}^{n}|\delta_i|\left(1 + \frac{L}{n}\right)^{n-i}. \qquad (51)\end{gathered}$$

Since $\delta_0 = \epsilon_0$ *and* $\left\|\mathbf{v}^{(1)} - \mathbf{v}^{(2)}\right\|_{\mathcal{X}_n} = \max_{i=0,\ldots,n}|\epsilon_i|$, *hence we can write our estimation in the form*

$$\left\|\mathbf{v}^{(1)} - \mathbf{v}^{(2)}\right\|_{\mathcal{X}_n} \leq |\delta_0| \left(1 + \frac{L}{n}\right)^n + \frac{1}{n} \sum_{i=1}^{n} |\delta_i| \left(1 + \frac{L}{n}\right)^{n-i} \tag{52}$$

$$< e^L(\delta_0 + \max_{i=1,\ldots,n} |\delta_i|) = e^L \|\delta\|_{\mathcal{Y}_n} = e^L \left\|F_n\left(\mathbf{v}^{(1)}\right) - F_n\left(\mathbf{v}^{(2)}\right)\right\|_{\mathcal{Y}_n} . \tag{53}$$

This shows us that the discretization (8), i.e., the explicit Euler method is stable on the whole set $\mathcal{X} = C^1[0,1]$ with $S = e^L$ and $R = \infty$.

Hence, based on Theorem 5.1, the results of this example and Example 5.1, we can conclude that the explicit Euler method is convergent, and the order of its convergence is one.

6 Relation between consistency, stability and convergence

Theorem 5.1 shows that, under the Assumption $A^\star$, the consistency and stability of discretization $\mathscr{D}$ on problem $\mathscr{P}$ result in the convergence, i.e., consistency and stability together are a sufficient condition for convergence. (Roughly speaking, this implication is shown in (2).) However, from this observation we cannot get an answer to the question of the necessity of these conditions.

In the sequel, we raise a more general question: What is the general relation between the above listed three basic notions? Since each of them can be true (T) or false (F), we have to consider eight different cases, listed in Table 1.

	consistency	stability	convergence
1	T	T	T
2	T	T	F
3	T	F	T
4	T	F	F
5	F	T	T
6	F	T	F
7	F	F	T
8	F	F	F

Table 1: The list of the different cases (T: true, F: false).

Before giving the answer, we consider some examples. In each examples $\mathcal{X} = \mathcal{X}_n = \mathcal{Y} = \mathcal{Y}_n = \mathbb{R}$, $\mathcal{D} = \mathcal{D}_n = [0, \infty)$, $\varphi_n = \psi_n = identity$. Our aim is to solve the scalar equation

$$F(x) \equiv x^2 = 0\,, \tag{54}$$

which has the unique solution $\bar{x} = 0$.

Example 6.1. *For solving equation* (54) *we choose now the numerical method* $F_n(x) = 1 - nx$. *The roots of the discrete equations* $F_n(x) = 0$ *are* $\bar{x}_n = 1/n$, *therefore* $\bar{x}_n \to \bar{x} = 0$ *as* $n \to \infty$. *This means that the numerical method is convergent. We observe that* $\varphi_n(F_n(0)) = \varphi_n(1) = 1$, *and* $\psi_n(F(0)) = \psi_n(0) = 0$. *Hence, for the local discretization error we have* $|l_n| = 1$, *for any index* n. *This means that the numerical method is not consistent. One can easily check that* F_n *is invertible, and* $F_n^{-1}(x) = -x/n + 1/n$. *Hence the derivative of the inverse operators are uniformly bounded on* $[0, \infty)$ *by* 1 *for any* n. *Therefore the numerical method is stable.*

Example 6.2. *For solving equation* (54) *we choose the numerical method defined by the n-th Lagrangian interpolation, i.e., $F_n(x)$ is the Lagrangian interpolation polynomial of order n. Since the Lagrange interpolation is exact for $n \geq 2$, therefore $F_n(x) = x^2$ holds for all $n \geq 2$. Hence, clearly the numerical method is consistent and convergent. The operator F_n^{-1} can be defined easily, and it is $F_n^{-1}(x) = \sqrt{x}$. Hence its derivative is not bounded around the point $\bar{x} = 0$, therefore the numerical method is not stable.*

Example 6.3. *For solving equation* (54) *we choose the following numerical method: $F_n(x) = 1 - nx^2$. Then $\bar{x}_n = 1/\sqrt{n}$, and hence $\bar{x}_n \to \bar{x} = 0$ as $n \to \infty$. This means that the numerical method is convergent. Due to the relations $\varphi_n(F_n(0)) = \varphi_n(1) = 1$ and $\psi_n(F(0)) = \psi_n(0) = 0$, this method is not consistent. Since for this numerical method $F_n^{-1}(x) = \sqrt{(1-x)/n}$, therefore the derivatives are not bounded. Therefore the numerical method is not stable.*

Now, we are in the position to answer the question, posed at beginning of this section. Using the numeration of the different cases in Table 1, the answers are included in Table 2. (We note that two cases (case 6 and 8 in Table 1) are uninteresting from a practical point of view, therefore we have neglected their investigation.) The results particularly show that neither consistency, nor stability is a necessary condition for the convergence.

number of the case	answer	reason
1	always true	Theorem 5.1
2	always false	Theorem 5.1
3	possible	Example 6.2
4	possible	Examples 3.2 and 5.2
5	possible	Example 6.1
6	n.a.	n.a.
7	possible	Example 6.3
8	n.a.	n.a.

Table 2: The possibility of the different cases.

7 Summary

We have considered the numerical solution of non-linear equations in an abstract (Banach space) setting. The main aim was to guarantee the convergence of the numerical process. It was shown that, similarly to the linear case, this notion can be guaranteed by two notions: the consistency and the stability. We investigated how to define appropriately the notion of stability in nonlinear problems that helps us to claim that convergence can be replaced by these two notions. This turns out to be useful from the applicational point of view. Thanks to this investigation (through an example) we understood of the primary importance of the notion of the stability threshold and the so-called local stability notions.

Thus, the consistency and the stability together ensure the convergence. In the linear case this result is well known as the Lax (or sometimes Lax-Richtmyer-Kantorovich) theory. From the formulation of the main theorem it turns out that these two, directly checkable conditions (i.e., the consistency and stability) serve together as a sufficient condition of the convergence.

However, even in the linear theory, the necessity of these conditions is less investigated. By giving suitable examples we have shown that neither consistency, nor stability is necessary for the convergence, in general. As an example for the theory, we have investigated the numerical solution of a Cauchy problem for ordinary differential equations by means of the explicit Euler method. We have shown the first order consistency and the stability of this method, which, based on the basic theorem, yield first order convergence. (We note that, as opposed to the usual direct proof of the convergence of the explicit Euler method, the convergence in this example yields the convergence on the whole space-time domain, and not only at some fixed time level $t = t^*$.)

References

[FMF11] Faragó, I., Mincsovics, M. E., Fekete, I.: Notes on the basic notions in nonlinear numerical analysis.
E. J. Qualitative Theory of Diff. Equ., Proc. 9'th Coll. Qualitative Theory of Diff. Equ., No. 6, 1–22 (2011)

[K75] Keller, H. B.: Approximation Methods for Nonlinear Problems with Application to Two-Point Boundary Value Problems.
Math. Comput., 130, 464–474 (1975)

[LR56] Lax, P. D. and Richtmyer, R. D.: Survey of Stability of Linear Finite Difference Equations.
Comm. Pure Appl. Math., 9, 267–293 (1956)

[LS88a] López-Marcos, J. C. and Sanz-Serna, J. M.: A definition of stability for nonlinear problems.
Numerical Treatment of Differential Equations, Teubner-Texte zur Mathematik, Band 104, 216–226 (1988)

[LS88b] López-Marcos, J. C. and Sanz-Serna, J. M.: Stability and Convergence in Numerical Analysis III: Linear Investigation of Nonlinear Stability.
IMA J. Numer. Anal., 8, 71–84 (1988)

[PS84a] Palencia, C. and Sanz-Serna, J. M.: An Extension of the Lax-Richtmyer Theory.
Numer. Math., 44, 279–283 (1984)

[PS84b] Palencia, C. and Sanz-Serna, J. M.: Equivalence Theorems for Incomplete Spaces: an Appraisal.
IMA J. Numer. Anal., 4, 109–115 (1984)

[PS85] Palencia, C. and Sanz-Serna, J. M.: A General Equivalence Theorem in the Theory of Discretization Methods.
Math. of Comp., 45/171, 143–152 (1985)

[SV86] Sanz-Serna, J. M. and Verwer, J. G.: A Study of the Recursion $y_{n+1} = y_n + \tau y_n^m$.
J. Math. Anal. Appl., 116, 456–463 (1986)

[S91] Sanz-Serna, J. M.: Two topics in nonlinear stability.
Advances in Numerical Analysis, Will Light ed., Clarendon Press, Oxford, Vol. 1, 147–174 (1991)

[S73] Stetter, H. J.: Analysis of Discretization Methods for Ordinary Differential Equations.
Springer, Berlin, (1973)

[T80] Trenogin, V. A.: Functional Analysis.
Nauka, Moscow, (1980) (in Russian)