11

Numerical Solution of Ordinary Differential Equations

In this chapter we deal with the numerical solutions of the Cauchy problem for ordinary differential equations (henceforth abbreviated by ODEs). After a brief review of basic notions about ODEs, we introduce the most widely used techniques for the numerical approximation of scalar equations. The concepts of consistency, convergence, zero-stability and absolute stability will be addressed. Then, we extend our analysis to systems of ODEs, with emphasis on *stiff* problems.

11.1 The Cauchy Problem

The Cauchy problem (also known as the initial-value problem) consists of finding the solution of an ODE, in the scalar or vector case, given suitable initial conditions. In particular, in the scalar case, denoting by I an interval of $\mathbb{R}$ containing the point t_0, the Cauchy problem associated with a first order ODE reads:

find a real-valued function $y \in C^1(I)$, such that

$$\begin{cases} y'(t) = f(t, y(t)), & t \in I, \\ y(t_0) = y_0, \end{cases} \tag{11.1}$$

where $f(t, y)$ is a given real-valued function in the strip $S = I \times (-\infty, +\infty)$, which is continuous with respect to both variables. Should f depend on t only through y, the differential equation is called *autonomous*.

Most of our analysis will be concerned with one single differential equation (scalar case). The extension to the case of systems of first-order ODEs will be addressed in Section 11.9.

If f is continuous with respect to t, then the solution to (11.1) satisfies

$$y(t) - y_0 = \int_{t_0}^{t} f(\tau, y(\tau))d\tau. \tag{11.2}$$

Conversely, if y is defined by (11.2), then it is continuous in I and $y(t_0) = y_0$. Moreover, since y is a primitive of the continuous function $f(\cdot, y(\cdot))$, $y \in C^1(I)$ and satisfies the differential equation $y'(t) = f(t, y(t))$.

Thus, if f is continuous the Cauchy problem (11.1) is equivalent to the integral equation (11.2). We shall see later on how to take advantage of this equivalence in the numerical methods.

Let us now recall two existence and uniqueness results for (11.1).

1. **Local existence and uniqueness.**

   Suppose that $f(t, y)$ is locally Lipschitz continuous at (t_0, y_0) with respect to y, that is, there exist two neighborhoods, $J \subseteq I$ of t_0 of width r_J, and Σ of y_0 of width r_Σ, and a constant $L > 0$, such that

   $$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2| \quad \forall t \in J, \ \forall y_1, y_2 \in \Sigma. \tag{11.3}$$

   Then, the Cauchy problem (11.1) admits a unique solution in a neighborhood of t_0 with radius r_0 with $0 < r_0 < \min(r_J, r_\Sigma/M, 1/L)$, where M is the maximum of $|f(t, y)|$ on $J \times \Sigma$. This solution is called the *local solution*.

   Notice that condition (11.3) is automatically satisfied if f has continuous derivative with respect to y: indeed, in such a case it suffices to choose L as the maximum of $|\partial f(t, y)/\partial y|$ in $\overline{J \times \Sigma}$.

2. **Global existence and uniqueness.** The problem admits a unique *global solution* if one can take $J = I$ and $\Sigma = \mathbb{R}$ in (11.3), that is, if f is *uniformly Lipschitz continuous* with respect to y.

In view of the stability analysis of the Cauchy problem, we consider the following problem

$$\begin{cases} z'(t) = f(t, z(t)) + \delta(t), & t \in I, \\ z(t_0) = y_0 + \delta_0, \end{cases} \tag{11.4}$$

where $\delta_0 \in \mathbb{R}$ and δ is a continuous function on I. Problem (11.4) is derived from (11.1) by perturbing both the initial datum y_0 and the source function f. Let us now characterize the sensitivity of the solution z to those perturbations.

Definition 11.1 ([Hah67], [Ste71] or [PS91]). Let I be a bounded set. The Cauchy problem (11.1) is *stable in the sense of Liapunov* (or *stable*) on I if, for any perturbation $(\delta_0, \delta(t))$ satisfying

$$|\delta_0| < \varepsilon, \qquad |\delta(t)| < \varepsilon \quad \forall t \in I,$$

with $\varepsilon > 0$ sufficiently small to guarantee that the solution to the perturbed problem (11.4) does exist, then

$$\exists C > 0 \text{ independent of } \varepsilon \text{ such that } \quad |y(t) - z(t)| < C\varepsilon, \qquad \forall t \in I. \tag{11.5}$$

If I has no upperly bound we say that (11.1) is *asymptotically stable* if, as well as being Liapunov stable in any bounded interval I, the following limit also holds

$$|y(t) - z(t)| \to 0, \qquad \text{for } t \to +\infty. \tag{11.6}$$

■

The requirement that the Cauchy problem is stable is equivalent to requiring that it is well-posed in the sense stated in Chapter 2.

The uniform Lipschitz-continuity of f with respect to y suffices to ensure the stability of the Cauchy problem. Indeed, letting $w(t) = z(t) - y(t)$, we have

$$w'(t) = f(t, z(t)) - f(t, y(t)) + \delta(t).$$

Therefore,

$$w(t) = \delta_0 + \int_{t_0}^{t} [f(s, z(s)) - f(s, y(s))]\, ds + \int_{t_0}^{t} \delta(s) ds, \qquad \forall t \in I.$$

Thanks to previous assumptions, it follows that

$$|w(t)| \le (1 + |t - t_0|)\, \varepsilon + L \int_{t_0}^{t} |w(s)| ds.$$

Applying the Gronwall lemma (which we include below for the reader's ease) yields

$$|w(t)| \le (1 + |t - t_0|)\, \varepsilon e^{L|t - t_0|}, \qquad \forall t \in I$$

and, thus, (11.5) with $C = (1 + K_I)e^{LK_I}$ where $K_I = \max_{t \in I} |t - t_0|$.

Lemma 11.1 (Gronwall) *Let p be an integrable function nonnegative on the interval $(t_0, t_0 + T)$, and let g and φ be two continuous functions on $[t_0, t_0 + T]$, g being nondecreasing. If φ satisfies the inequality*

$$\varphi(t) \le g(t) + \int_{t_0}^{t} p(\tau)\varphi(\tau) d\tau, \qquad \forall t \in [t_0, t_0 + T],$$

then

$$\varphi(t) \le g(t) \exp\left(\int_{t_0}^{t} p(\tau) d\tau \right), \qquad \forall t \in [t_0, t_0 + T].$$

For the proof, see, for instance, [QV94], Lemma 1.4.1.

The constant C that appears in (11.5) could be very large and, in general, depends on the upper extreme of the interval I, as in the proof above. For that reason, the property of asymptotic stability is more suitable for describing the behavior of the *dynamical system* (11.1) as $t \to +\infty$ (see [Arn73]).

As is well-known, only a restricted number of nonlinear ODEs can be solved in closed form (see, for instance, [Arn73]). Moreover, even when this is possible, it is not always a straightforward task to find an explicit expression of the solution; for example, consider the (very simple) equation $y' = (y - t)/(y + t)$, whose solution is only implicitly defined by the relation $(1/2)\log(t^2 + y^2) + \tan^{-1}(y/t) = C$, where C is a constant depending on the initial condition.

For this reason we are interested in numerical methods, since these can be applied to any ODE under the sole condition that it admits a unique solution.

11.2 One-Step Numerical Methods

Let us address the numerical approximation of the Cauchy problem (11.1). Fix $0 < T < +\infty$ and let $I = (t_0, t_0 + T)$ be the integration interval and, correspondingly, for $h > 0$, let $t_n = t_0 + nh$, with $n = 0, 1, 2, \ldots, N_h$, be the sequence of discretization nodes of I into subintervals $I_n = [t_n, t_{n+1}]$. The width h of such subintervals is called the *discretization stepsize*. Notice that N_h is the maximum integer such that $t_{N_h} \le t_0 + T$. Let u_j be the approximation at node t_j of the exact solution $y(t_j)$; this solution will be henceforth shortly denoted by y_j. Similarly, f_j denotes the value $f(t_j, u_j)$. We obviously set $u_0 = y_0$.

Definition 11.2 A numerical method for the approximation of problem (11.1) is called a *one-step* method if $\forall n \ge 0$, u_{n+1} depends only on u_n. Otherwise, the scheme is called a *multistep* method. ■

For now, we focus our attention on one-step methods. Here are some of them:

1. **forward Euler method**

$$u_{n+1} = u_n + hf_n; \tag{11.7}$$

2. **backward Euler method**

$$u_{n+1} = u_n + hf_{n+1}. \tag{11.8}$$

In both cases, y' is approximated through a finite difference: forward and backward differences are used in (11.7) and (11.8), respectively. Both finite differences are first-order approximations of the first derivative of y with respect to h (see Section 10.10.1).

3. **trapezoidal (or Crank-Nicolson) method**

$$u_{n+1} = u_n + \frac{h}{2}\left[f_n + f_{n+1}\right]. \tag{11.9}$$

This method stems from approximating the integral on the right side of (11.2) by the trapezoidal quadrature rule (9.11).

4. **Heun method**

$$u_{n+1} = u_n + \frac{h}{2}[f_n + f(t_{n+1}, u_n + hf_n)]. \tag{11.10}$$

This method can be derived from the trapezoidal method substituting $f(t_{n+1}, u_n + hf(t_n, u_n))$ for $f(t_{n+1}, u_{n+1})$ in (11.9) (i.e., using the forward Euler method to compute u_{n+1}).

In this last case, we notice that the aim is to transform an *implicit* method into an *explicit* one. Addressing this concern, we recall the following.

Definition 11.3 (explicit and implicit methods) A method is called *explicit* if u_{n+1} can be computed directly in terms of (some of) the previous values u_k, $k \leq n$. A method is said to be *implicit* if u_{n+1} depends implicitly on itself through f. ■

Methods (11.7) and (11.10) are explicit, while (11.8) and (11.9) are implicit. These latter require at each time step to solving a nonlinear problem if f depends nonlinearly on the second argument.

A remarkable example of one-step methods are the Runge-Kutta methods, which will be analyzed in Section 11.8.

11.3 Analysis of One-Step Methods

Any one-step explicit method for the approximation of (11.1) can be cast in the concise form

$$u_{n+1} = u_n + h\Phi(t_n, u_n, f_n; h), \quad 0 \leq n \leq N_h - 1, \quad u_0 = y_0, \tag{11.11}$$

where $\Phi(\cdot, \cdot, \cdot; \cdot)$ is called an *increment function*. Letting as usual $y_n = y(t_n)$, analogously to (11.11) we can write

$$y_{n+1} = y_n + h\Phi(t_n, y_n, f(t_n, y_n); h) + \varepsilon_{n+1}, \quad 0 \leq n \leq N_h - 1, \tag{11.12}$$

where ε_{n+1} is the residual arising at the point t_{n+1} when we pretend that the exact solution "satisfies" the numerical scheme. Let us write the residual as

$$\varepsilon_{n+1} = h\tau_{n+1}(h).$$

The quantity $\tau_{n+1}(h)$ is called the *local truncation error* (LTE) at the node t_{n+1}. We thus define the *global truncation error* to be the quantity

$$\tau(h) = \max_{0 \leq n \leq N_h - 1} |\tau_{n+1}(h)|$$

Notice that $\tau(h)$ depends on the solution y of the Cauchy problem (11.1).

The forward Euler's method is a special instance of (11.11), where

$$\Phi(t_n, u_n, f_n; h) = f_n,$$

while to recover Heun's method we must set

$$\Phi(t_n, u_n, f_n; h) = \frac{1}{2}\left[f_n + f(t_n + h, u_n + hf_n)\right].$$

A one-step explicit scheme is fully characterized by its increment function Φ. This function, in all the cases considered thus far, is such that

$$\lim_{h \to 0} \Phi(t_n, y_n, f(t_n, y_n); h) = f(t_n, y_n), \quad \forall t_n \geq t_0 \tag{11.13}$$

Property (11.13), together with the obvious relation $y_{n+1} - y_n = hy'(t_n) + \mathcal{O}(h^2)$, $\forall n \geq 0$, allows one to obtain from (11.12) that $\lim_{h \to 0} \tau_n(h) = 0$, $0 \leq n \leq N_h - 1$. In turn, this condition ensures that

$$\lim_{h \to 0} \tau(h) = 0$$

which expresses the *consistency* of the numerical method (11.11) with the Cauchy problem (11.1). In general, a method is said to be *consistent* if its LTE is infinitesimal with respect to h. Moreover, a scheme has *order* p if, $\forall t \in I$, the solution $y(t)$ of the Cauchy problem (11.1) fulfills the condition

$$\tau(h) = \mathcal{O}(h^p) \quad \text{for } h \to 0. \tag{11.14}$$

Using Taylor expansions, as was done in Section 11.2, it can be proved that the forward Euler method has order 1, while the Heun method has order 2 (see Exercises 1 and 2).

11.3.1 The Zero-Stability

Let us formulate a requirement analogous to the one for Liapunov stability (11.5), specifically for the numerical scheme. If (11.5) is satisfied with a constant C independent of h, we shall say that the numerical problem is *zero-stable*. Precisely:

Definition 11.4 (zero-stability of one-step methods) The numerical method (11.11) for the approximation of problem (11.1) is *zero-stable* if

$$\exists h_0 > 0,\ \exists C > 0 :\ \forall h \in (0, h_0],\ |z_n^{(h)} - u_n^{(h)}| \leq C\varepsilon,\ 0 \leq n \leq N_h, \quad (11.15)$$

where $z_n^{(h)}$, $u_n^{(h)}$ are respectively the solutions of the problems

$$\begin{cases} z_{n+1}^{(h)} = z_n^{(h)} + h\left[\Phi(t_n, z_n^{(h)}, f(t_n, z_n^{(h)}); h) + \delta_{n+1}\right], \\ z_0 = y_0 + \delta_0, \end{cases} \quad (11.16)$$

$$\begin{cases} u_{n+1}^{(h)} = u_n^{(h)} + h\Phi(t_n, u_n^{(h)}, f(t_n, u_n^{(h)}); h), \\ u_0 = y_0, \end{cases} \quad (11.17)$$

for $0 \leq n \leq N_h - 1$, under the assumption that $|\delta_k| \leq \varepsilon$, $0 \leq k \leq N_h$. ■

Zero-stability thus requires that, in a bounded interval, (11.15) holds for any value $h \leq h_0$. This property deals, in particular, with the behavior of the numerical method in the limit case $h \to 0$ and this justifies the name of *zero*-stability. This latter is therefore a distinguishing property of the numerical method itself, not of the Cauchy problem (which, indeed, is stable due to the uniform Lipschitz continuity of f). Property (11.15) ensures that the numerical method has a weak sensitivity with respect to small changes in the data and is thus stable in the sense of the general definition given in Chapter 2.

Remark 11.1 The constant C in (11.15) is independent of h (and thus of N_h), but it can depend on the width T of the integration interval I. Actually, (11.15) does not exclude *a priori* the constant C from being an unbounded function of T. ■

The request that a numerical method be stable arises, before anything else, from the need of keeping under control the (unavoidable) errors introduced by the finite arithmetic of the computer. Indeed, if the numerical method were not zero-stable, the rounding errors made on y_0 as well as in the process of computing $f(t_n, u_n)$ would make the computed solution completely useless.

Theorem 11.1 (Zero-stability) *Consider the explicit one-step method* (11.11) *for the numerical solution of the Cauchy problem* (11.1). *Assume that the increment function Φ is Lipschitz continuous with respect to the second argument, with constant Λ independent of h and of the nodes $t_j \in [t_0, t_0 + T]$, that is*

$$\exists h_0 > 0,\ \exists \Lambda > 0 :\ \forall h \in (0, h_0]$$
$$|\Phi(t_n, u_n^{(h)}, f(t_n, u_n^{(h)}); h) - \Phi(t_n, z_n^{(h)}, f(t_n, z_n^{(h)}); h)| \leq \Lambda |u_n^{(h)} - z_n^{(h)}|,\ 0 \leq n \leq N_h. \quad (11.18)$$

Then, method (11.11) *is zero-stable.*

Proof. Setting $w_j^{(h)} = z_j^{(h)} - u_j^{(h)}$, by subtracting (11.17) from (11.16) we obtain, for $j = 0, \ldots, N_h - 1$,

$$w_{j+1}^{(h)} = w_j^{(h)} + h\left[\Phi(t_j, z_j^{(h)}, f(t_j, z_j^{(h)}); h) - \Phi(t_j, u_j^{(h)}, f(t_j, u_j^{(h)}); h)\right] + h\delta_{j+1}.$$

Summing over j gives, for $n = 1, \ldots, N_h$,

$$w_n^{(h)} = w_0^{(h)} + h\sum_{j=0}^{n-1}\delta_{j+1} + h\sum_{j=0}^{n-1}\left(\Phi(t_j, z_j^{(h)}, f(t_j, z_j^{(h)}); h) - \Phi(t_j, u_j^{(h)}, f(t_j, u_j^{(h)}); h)\right),$$

so that, by (11.18)

$$|w_n^{(h)}| \leq |w_0| + h\sum_{j=0}^{n-1}|\delta_{j+1}| + h\Lambda\sum_{j=0}^{n-1}|w_j^{(h)}|, \qquad 1 \leq n \leq N_h. \quad (11.19)$$

Applying the discrete Gronwall lemma, given below, we obtain

$$|w_n^{(h)}| \leq (1 + hn)\,\varepsilon e^{nh\Lambda}, \qquad 1 \leq n \leq N_h.$$

Then (11.15) follows from noticing that $hn \leq T$ and setting $C = (1 + T)\,e^{\Lambda T}$. ◇

Notice that zero-stability implies the boundedness of the solution when f is linear with respect to the second argument.

Lemma 11.2 (discrete Gronwall) *Let k_n be a nonnegative sequence and φ_n a sequence such that*

$$\begin{cases} \varphi_0 \leq g_0 \\ \varphi_n \leq g_0 + \sum_{s=0}^{n-1} p_s + \sum_{s=0}^{n-1} k_s \phi_s, \qquad n \geq 1. \end{cases}$$

If $g_0 \geq 0$ and $p_n \geq 0$ for any $n \geq 0$, then

$$\varphi_n \leq \left(g_0 + \sum_{s=0}^{n-1} p_s \right) \exp\left(\sum_{s=0}^{n-1} k_s \right), \qquad n \geq 1.$$

For the proof, see, for instance, [QV94], Lemma 1.4.2. In the specific case of the Euler method, checking the property of zero-stability can be done directly using the Lipschitz continuity of f (we refer the reader to the end of Section 11.3.2). In the case of multistep methods, the analysis will lead to the verification of a purely algebraic property, the so-called *root condition* (see Section 11.6.3).

11.3.2 Convergence Analysis

Definition 11.5 A method is said to be *convergent* if

$$\forall n = 0, \ldots, N_h, \qquad |u_n - y_n| \leq C(h)$$

where $C(h)$ is an infinitesimal with respect to h. In that case, it is said to be *convergent with order p* if $\exists C > 0$ such that $C(h) = Ch^p$. ■

We can prove the following theorem.

Theorem 11.2 (Convergence) *Under the same assumptions as in Theorem 11.1, we have*

$$|y_n - u_n| \leq (|y_0 - u_0| + nh\tau(h))\, e^{nh\Lambda}, \qquad 1 \leq n \leq N_h. \tag{11.20}$$

Therefore, if the consistency assumption (11.13) *holds and $|y_0 - u_0| \to 0$ as $h \to 0$, then the method is convergent. Moreover, if $|y_0 - u_0| = \mathcal{O}(h^p)$ and the method has order p, then it is also convergent with order p.*

Proof. Setting $w_j = y_j - u_j$, subtracting (11.11) from (11.12) and proceeding as in the proof of the previous theorem yields inequality (11.19), with the understanding that

$$w_0 = y_0 - u_0, \text{ and } \delta_{j+1} = \tau_{j+1}(h).$$

The estimate (11.20) is then obtained by applying again the discrete Gronwall lemma. From the fact that $nh \leq T$ and $\tau(h) = \mathcal{O}(h^p)$, we can conclude that $|y_n - u_n| \leq Ch^p$ with C depending on T and Λ but not on h. ◇

A consistent and zero-stable method is thus convergent. This property is known as the *Lax-Richtmyer theorem* or *equivalence theorem* (the converse: "a convergent method is zero-stable" being obviously true). This theorem, which is proven in [IK66], was already advocated in Section 2.2.1 and is a central result in the analysis of numerical methods for ODEs (see [Dah56] or

[Hen62] for linear multistep methods, [But66], [MNS74] for a wider classes of methods). It will be considered again in Section 11.5 for the analysis of multistep methods.

We carry out in detail the convergence analysis in the case of the forward Euler method, without resorting to the discrete Gronwall lemma. In the first part of the proof we assume that any operation is performed in exact arithmetic and that $u_0 = y_0$.

Denote by $e_{n+1} = y_{n+1} - u_{n+1}$ the error at node t_{n+1} with $n = 0, 1, \ldots$ and notice that

$$e_{n+1} = (y_{n+1} - u^*_{n+1}) + (u^*_{n+1} - u_{n+1}), \tag{11.21}$$

where $u^*_{n+1} = y_n + hf(t_n, y_n)$ is the solution obtained after one step of the forward Euler method starting from the initial datum y_n (see Figure 11.1). The first addendum in (11.21) accounts for the consistency error, the second one for the cumulation of these errors. Then

$$y_{n+1} - u^*_{n+1} = h\tau_{n+1}(h), \quad u^*_{n+1} - u_{n+1} = e_n + h\left[f(t_n, y_n) - f(t_n, u_n)\right].$$

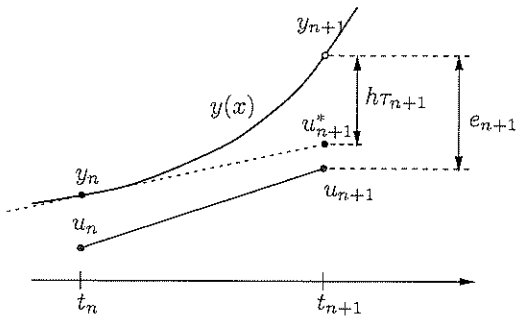

FIGURE 11.1. Geometrical interpretation of the local and global truncation errors at node t_{n+1} for the forward Euler method

As a consequence,

$$|e_{n+1}| \leq h|\tau_{n+1}(h)| + |e_n| + h|f(t_n, y_n) - f(t_n, u_n)| \leq h\tau(h) + (1 + hL)|e_n|,$$

L being the Lipschitz constant of f. By recursion on n, we find

$$\begin{aligned} |e_{n+1}| &\leq [1 + (1 + hL) + \ldots + (1 + hL)^n]\, h\tau(h) \\ &= \frac{(1 + hL)^{n+1} - 1}{L}\tau(h) \leq \frac{e^{L(t_{n+1} - t_0)} - 1}{L}\tau(h). \end{aligned}$$

The last inequality follows from noticing that $1+hL \leq e^{hL}$ and $(n+1)h = t_{n+1} - t_0$.

On the other hand, if $y \in C^2(I)$, the LTE for the forward Euler method is (see Section 10.10.1)

$$\tau_{n+1}(h) = \frac{h}{2}y''(\xi), \quad \xi \in (t_n, t_{n+1}),$$

and thus, $\tau(h) \leq (M/2)h$, where $M = \max_{\xi \in I} |y''(\xi)|$. In conclusion,

$$|e_{n+1}| \leq \frac{e^{L(t_{n+1}-t_0)} - 1}{L}\frac{M}{2}h, \quad \forall n \geq 0, \tag{11.22}$$

from which it follows that the global error tends to zero with the same order as the local truncation error.

If also the rounding errors are accounted for, we can assume that the solution $\bar{u}_{n+1}$, actually computed by the forward Euler method at time t_{n+1}, is such that

$$\bar{u}_0 = y_0 + \zeta_0, \quad \bar{u}_{n+1} = \bar{u}_n + hf(t_n, \bar{u}_n) + \zeta_{n+1}, \tag{11.23}$$

having denoted the rounding error by ζ_j, for $j \geq 0$.
Problem (11.23) is an instance of (11.16), provided that we identify ζ_{n+1} and $\bar{u}_n$ with $h\delta_{n+1}$ and $z_n^{(h)}$ in (11.16), respectively. Combining Theorems 11.1 and 11.2 we get, instead of (11.22), the following error estimate

$$|y_{n+1} - \bar{u}_{n+1}| \leq e^{L(t_{n+1}-t_0)}\left[|\zeta_0| + \frac{1}{L}\left(\frac{M}{2}h + \frac{\zeta}{h}\right)\right],$$

where $\zeta = \max_{1 \leq j \leq n+1} |\zeta_j|$. The presence of rounding errors does not allow, therefore, to conclude that as $h \to 0$, the error goes to zero. Actually, there exists an optimal (non null) value of h, h_{opt}, for which the error is minimized. For $h < h_{opt}$, the rounding error dominates the truncation error and the global error increases.

11.3.3 The Absolute Stability

The property of *absolute stability* is in some way specular to zero-stability, as far as the roles played by h and I are concerned. Heuristically, we say that a numerical method is absolutely stable if, *for h fixed*, u_n remains bounded as $t_n \to +\infty$. This property, thus, deals with the asymptotic behavior of u_n, as opposed to a zero-stable method for which, for a fixed integration interval, u_n remains bounded as $h \to 0$.
For a precise definition, consider the linear Cauchy problem (that from now on, we shall refer to as the *test problem*)

$$\begin{cases} y'(t) = \lambda y(t), & t > 0, \\ y(0) = 1, \end{cases} \tag{11.24}$$

with $\lambda \in \mathbb{C}$, whose solution is $y(t) = e^{\lambda t}$. Notice that $\lim_{t \to +\infty} |y(t)| = 0$ if $\mathrm{Re}(\lambda) < 0$.

Definition 11.6 A numerical method for approximating (11.24) is *absolutely stable* if

$$|u_n| \longrightarrow 0 \quad \text{as} \quad t_n \longrightarrow +\infty. \tag{11.25}$$

Let h be the discretization stepsize. The numerical solution u_n of (11.24) obviously depends on h and λ. The *region of absolute stability* of the numerical method is the subset of the complex plane

$$\mathcal{A} = \{z = h\lambda \in \mathbb{C} : \text{(11.25) is satisfied}\}. \tag{11.26}$$

Thus, $\mathcal{A}$ is the set of the values of the product $h\lambda$ for which the numerical method furnishes solutions that decay to zero as t_n tends to infinity. ■

Let us check whether the one-step methods introduced previously are absolutely stable.

1. *Forward Euler method*: applying (11.7) to problem (11.24) yields $u_{n+1} = u_n + h\lambda u_n$ for $n \geq 0$, with $u_0 = 1$. Proceeding recursively on n we get

$$u_n = (1 + h\lambda)^n, \qquad n \geq 0.$$

Therefore, condition (11.25) is satisfied iff $|1 + h\lambda| < 1$, that is, if $h\lambda$ lies within the unit circle with center at $(-1, 0)$ (see Figure 11.3). This amounts to requiring that

$$h\lambda \in \mathbb{C}^- \quad \text{and} \quad 0 < h < -\frac{2\mathrm{Re}(\lambda)}{|\lambda|^2} \tag{11.27}$$

where

$$\mathbb{C}^- = \{z \in \mathbb{C} : \mathrm{Re}(z) < 0\}.$$

Example 11.1 For the Cauchy problem $y'(x) = -5y(x)$ for $x > 0$ and $y(0) = 1$, condition (11.27) implies $0 < h < 2/5$. Figure 11.2 (left) shows the behavior of the computed solution for two values of h which do not fulfill this condition, while on the right we show the solutions for two values of h that do. Notice that in this second case the oscillations, if present, damp out as t grows. •

2. *Backward Euler method*: proceeding as before, we get this time

$$u_n = \frac{1}{(1 - h\lambda)^n}, \qquad n \geq 0.$$

The absolute stability property (11.25) is satisfied *for any value of $h\lambda$* that does not belong to the unit circle of center $(1, 0)$ (see Figure 11.3, right).

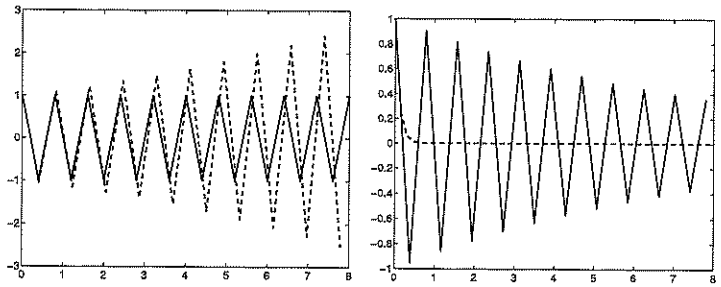

FIGURE 11.2. Left: computed solutions for $h = 0.41 > 2/5$ (dashed line) and $h = 2/5$ (solid line). Notice how, in the limiting case $h = 2/5$, the oscillations remain unmodified as t grows. Right: two solutions are reported for $h = 0.39$ (solid line) and $h = 0.15$ (dashed line)

Example 11.2 The numerical solution given by the backward Euler method in the case of Example 11.1 does not exhibit any oscillation for any value of h. On the other hand, the same method, if applied to the problem $y'(t) = 5y(t)$ for $t > 0$ and with $y(0) = 1$, computes a solution that decays *anyway* to zero as $t \to \infty$ if $h > 2/5$, despite the fact that the exact solution of the Cauchy problem tends to infinity. •

3. *Trapezoidal (or Crank-Nicolson) method*: we get

$$u_n = \left[\left(1 + \frac{1}{2}\lambda h\right) / \left(1 - \frac{1}{2}\lambda h\right)\right]^n, \qquad n \geq 0,$$

hence (11.25) is fulfilled for any $h\lambda \in \mathbb{C}^-$.

4. *Heun's method*: applying (11.10) to problem (11.24) and proceeding by recursion on n, we obtain

$$u_n = \left[1 + h\lambda + \frac{(h\lambda)^2}{2}\right]^n, \qquad n \geq 0.$$

As shown in Figure 11.3 the region of absolute stability of Heun's method is larger than the corresponding one of Euler's method. However, its restriction to the real axis is the same.
We say that a method is *A-stable* if $\mathcal{A} \cap \mathbb{C}^- = \mathbb{C}^-$, i.e., if for $\mathrm{Re}(\lambda) < 0$, condition (11.25) is satisfied for all values of h.

The backward Euler and Crank-Nicolson methods are A-stable, while the forward Euler and Heun methods are conditionally stable.

Remark 11.2 Notice that the implicit one-step methods examined so far are *unconditionally absolutely stable*, while explicit schemes are *conditionally absolutely stable*. This is, however, not a general rule: in fact, there exist implicit unstable or only conditionally stable schemes. On the contrary, there are no explicit unconditionally absolutely stable schemes [Wid67]. ■

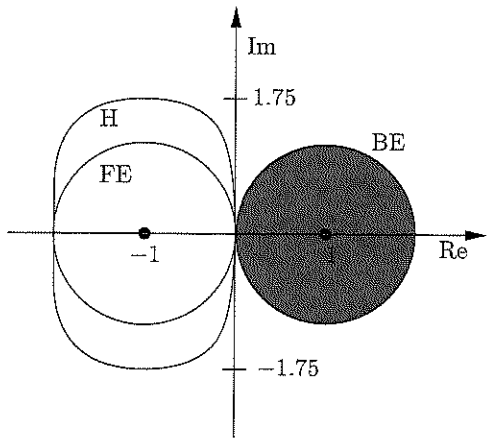

FIGURE 11.3. Regions of absolute stability for the forward (FE) and backward Euler (BE) methods and for Heun's method (H). Notice that the region of absolute stability of the BE method lies outside the unit circle of center $(1, 0)$ (shaded area)

11.4 Difference Equations

For any integer $k \geq 1$, an equation of the form

$$u_{n+k} + \alpha_{k-1}u_{n+k-1} + \ldots + \alpha_0 u_n = \varphi_{n+k}, \quad n = 0, 1, \ldots \tag{11.28}$$

is called a *linear difference equation* of order k. The coefficients $\alpha_0 \neq 0$, $\alpha_1, \ldots, \alpha_{k-1}$ may or may not depend on n. If, for any n, the right side φ_{n+k} is equal to zero, the equation is said *homogeneous*, while if the α_j's are independent of n it is called *linear difference equation with constant coefficients*.
Difference equations arise for instance in the discretization of ordinary differential equations. Regarding this, we notice that all the numerical methods examined so far end up with equations like (11.28). More generally, equations like (11.28) are encountered when quantities are defined through linear recursive relations. Another relevant application is concerned with the discretization of boundary value problems (see Chapter 12). For further details on the subject, we refer to Chapters 2 and 5 of [BO78] and to Chapter 6 of [Gau97].