

# The Logarithmic Norm. History and Modern Theory

GUSTAF SÖDERLIND

*Numerical Analysis*

*Centre for Mathematical Sciences*

*Lund University, Box 118*

*S-221 00 Lund, Sweden.*

*e-mail: Gustaf.Soderlind@na.lu.se*

In Memory of Germund Dahlquist

## Abstract.

In his 1958 thesis *Stability and Error Bounds* [3], Germund Dahlquist introduced the logarithmic norm in order to derive error bounds in initial value problems, using differential inequalities that distinguished between forward and reverse time integration. Originally defined for matrices, the logarithmic norm can be extended to bounded linear operators, but the extensions to nonlinear maps and unbounded operators have required a functional analytic redefinition of the concept.

This compact survey is intended as an elementary, but broad and largely self-contained, introduction to the versatile and powerful modern theory. Its wealth of applications range from the stability theory of IVPs and BVPs, to the solvability of algebraic, nonlinear, operator, and functional equations.

*AMS subject classifications:* 65L05.

*Key words:* Logarithmic norm, logarithmic Lipschitz constant, monotonicity, uniform monotonicity theorem, differential inequality, difference method, stability, error bound, Lax principle

## 1 Introduction

Let  $|\cdot|$  denote a vector norm on  $\mathbb{C}^d$  as well as its subordinate matrix (operator) norm on  $\mathbb{C}^{d \times d}$ . The classical definition of the *logarithmic norm* of a matrix  $A$ ,

$$(1.1) \quad \mu[A] = \lim_{h \rightarrow 0^+} \frac{|I + hA| - 1}{h},$$

was introduced in 1958 independently by Dahlquist [3] and Lozinskii [17]. The limit exists on account of every norm being left and right Gâteaux differentiable,

[7, p. 49]. The left and right limits generally do not coincide; they will turn out to be related to greatest lower bounds (left limit) and least upper bounds (right limit), respectively.

Here,  $\mu[A]$  is readily seen to be a *right G-differential*, [18, p. 66]. It fails to be a G-derivative as the logarithmic norm is a *sublinear* functional on  $\mathbb{C}^{d \times d}$ , i.e.,  $\mu[A + B] \leq \mu[A] + \mu[B]$ . In a few special cases  $\mu[\cdot]$  may nevertheless be linear. For example, if  $A = \lambda \in \mathbb{C}$  we have

$$(1.2) \quad |A| = |\lambda|; \quad \mu[A] = \operatorname{Re} \lambda.$$

Just as the operator norm measures “magnitude,” the logarithmic norm may be viewed as an extension of the notion of “real part.” For the spectrum of a general matrix  $A$  it holds that

$$(1.3) \quad \rho[A] \leq |A|; \quad \alpha[A] \leq \mu[A],$$

where  $\rho[A] = \max_i |\lambda_i|$  is the *spectral radius* and  $\alpha[A] = \max_i \operatorname{Re} \lambda_i$  is the *spectral abscissa*. For a given norm, equality is not often attained in (1.3), but there are important exceptions. For example, the Euclidean norms,  $|\cdot|_2$  and  $\mu_2[\cdot]$ , are sharp for the entire class of normal matrices.

Further, in analogy with (1.1), one notes that

$$(1.4) \quad \alpha[A] = \lim_{h \rightarrow 0^+} \frac{\rho[I + hA] - 1}{h}.$$

However,  $\rho[\cdot]$  is not submultiplicative and  $\alpha[\cdot]$  is not subadditive; this severely limits the use of these functionals in stability theory. The spectrum alone cannot characterize stability: while the spectrum is topologically invariant, *stability is a topological notion*. Nevertheless, the logarithmic norm has many links to spectral bounds, resolvents, Rayleigh quotients, and the numerical range of an operator, see e.g. [24, Ch. 17]. For instance,  $\mu[A]$  is the maximum real part of the numerical range, [21]. Several special fields in mathematics, such as semigroup theory, rely on notions that are directly related to the logarithmic norm.

There are only a few published surveys of the logarithmic norm and its applications, see e.g. [9, 20], although there are other sources that provide an excellent introduction, e.g. [8, pp 27–35]. These references focus on the original definition (1.1) and its applications to initial value problems for ODEs, emphasizing stability and perturbation theory. By reconsidering the definition of  $\mu[\cdot]$ , however, the logarithmic norm can be extended to nonlinear maps [21]; unbounded operators; matrix pencils [14]; and to maps on nonlinear manifolds [15]. Today’s theory is far richer than was originally expected, and the formalism can conveniently be used in an abundance of applications.

The original idea behind introducing the logarithmic norm was therefore to derive a topological (norm) condition on  $A$  that would guarantee that solutions to a linear dynamical system

$$(1.5) \quad \dot{x} = Ax + r$$

remain bounded whenever  $r$  is a bounded function of  $t$ . For  $t \geq 0$  the norm of  $x$  satisfies a differential inequality,

$$(1.6) \quad D_t^+ |x| \leq \mu[A] \cdot |x| + |r(t)|,$$

where  $D_t^+$  is the upper right *Dini derivative* with respect to time  $t$ . The upper left and right Dini derivatives are defined for a function  $\psi(t)$  by

$$(1.7) \quad (D_t^\pm \psi)(t) = \limsup_{h \rightarrow 0^\pm} \frac{\psi(t+h) - \psi(t)}{h}.$$

Using this notion, the following simple analysis of (1.5) led to (1.1):

$$\begin{aligned} (1.8a) \quad D_t^+ |x| &= \limsup_{h \rightarrow 0^+} \frac{|x(t+h)| - |x(t)|}{h} \\ (1.8b) &= \lim_{h \rightarrow 0^+} \frac{|x(t) + h\dot{x}(t)| - |x(t)|}{h} \\ (1.8c) &\leq \lim_{h \rightarrow 0^+} \frac{|(I + hA)x(t)| - |x(t)|}{h} + |r(t)| \\ (1.8d) &\leq \lim_{h \rightarrow 0^+} \frac{|I + hA| - 1}{h} |x(t)| + |r(t)| \\ (1.8e) &= \mu[A] \cdot |x(t)| + |r(t)|. \end{aligned}$$

The crucial step was to avoid using the triangle inequality  $|I + hA| \leq 1 + h|A|$  in (1.8d), as this would lead to the uninteresting differential inequality,

$$(1.9) \quad D_t^+ |x| \leq |A| \cdot |x(t)| + |r(t)|,$$

where forward and reverse time can no longer be distinguished. Consequently, every estimate of  $|x|$  would grow exponentially. By contrast, as Dahlquist remarked, “[the logarithmic norm] *can be a negative number*” (original emphasis; we also note that the term “logarithmic norm” is not used in the original reference), [3, p. 11]. Thus (1.6) does not only distinguish between forward and reverse time; it may also distinguish between stable and unstable systems.

The sensitivity to the sign of the operator is a key feature of the logarithmic norm. For example, describing irreversible processes, diffusion-type equations are well-posed in forward time but ill-posed in reverse. This will be reflected by the logarithmic norm, making it useful also in semi-group theory. Another example is in feedback control theory, where *negative* feedback is commonly used to stabilize a system.

In this paper we will first review the “classical” theory, which is concerned with matrix bounds and where the logarithmic norm is defined by (1.1). We will then develop a general theory, establishing the logarithmic norm within a framework of nonlinear functional analysis. To this end we shall replace the definition (1.1) and introduce a more general concept, extending the functional  $\mu[\cdot]$  to nonlinear maps as well as to unbounded operators. The basic relation (1.1) will however still hold for all bounded linear maps.

The extended theory will be shown to apply to initial and boundary value problems; to PDEs; to “algebraic” nonlinear equations; in discretization theory; and in questions of stability and error bounds. In particular, we will see how the logarithmic norm is related to, and can quantify, notions such as definiteness, ellipticity and monotonicity. A simple proof of the uniform monotonicity theorem will be given, and we will finally look at finite difference discretization theory and how the logarithmic norm links the Browder and Minty theorem to the Lax equivalence theorem.

## 2 Classical theory

Let us assume that  $A$  is a constant matrix. By integration of (1.6), one finds the basic perturbation bound

$$(2.1) \quad |x(t)| \leq e^{t\mu[A]}|x(0)| + \int_0^t e^{(t-\tau)\mu[A]}|r(\tau)|d\tau$$

for  $t \geq 0$ . There are two cases of special interest. First, if  $r \equiv 0$ , an initial excitation  $x(0)$  from the zero equilibrium yields the solution  $x(t) = e^{tA}x(0)$ . By (2.1) we then have  $|e^{tA}x(0)| \leq e^{t\mu[A]}|x(0)|$  for every  $x(0)$  and  $t \geq 0$ , hence the following bound, [3, p. 14]:

PROPOSITION 2.1. *Let  $A \in \mathbb{C}^{d \times d}$ . The matrix exponential is bounded by*

$$(2.2) \quad |e^{tA}| \leq e^{t\mu[A]}; \quad t \geq 0.$$

Thus, if  $\mu[A] \leq 0$  the zero solution is stable, with exponential stability whenever  $\mu[A] < 0$ . In particular, the condition  $\mu[A] < 0$  implies that the matrix exponential is a *contraction (semi-) group*, i.e.,  $|e^{tA}| < 1$  for  $t > 0$ .

The second special case is obtained by taking  $x(0) = 0$ . A perturbation  $r \neq 0$  then leads to the following bound:

PROPOSITION 2.2. *Let  $A \in \mathbb{C}^{d \times d}$  and let  $x(0) = 0$  in (1.5). The solution  $x(t)$  is then bounded by*

$$(2.3) \quad |x(t)| \leq \frac{e^{t\mu[A]} - 1}{\mu[A]} \max_{0 \leq \tau \leq t} |r(\tau)|$$

on compact intervals  $[0, T]$ .

In the case  $\mu[A] < 0$ , the bound (2.3) also holds on infinite intervals. Define  $\|r\|_\infty = \sup_{t \geq 0} |r(t)|$ ; we then have the uniform bound

$$(2.4) \quad \|x\|_\infty \leq -\frac{\|r\|_\infty}{\mu[A]},$$

which demonstrates stability, now in the sense that  $x$  depends continuously on the perturbation  $r$ . In other words, if  $\|r\|_\infty \rightarrow 0$  then  $\|x\|_\infty \rightarrow 0$ .

Specializing further, consider  $\dot{x} = Ax + r$  with  $\mu[A] < 0$  and  $r \equiv \text{const}$ . As  $\mu[A] < 0$ , homogeneous solutions decay (see (2.1)); we can conclude that  $A$  is

nonsingular and that there exists a unique stable equilibrium  $x = -A^{-1}r$ . Take  $x(0) = -A^{-1}r$ . By (2.4) we then have  $|A^{-1}r| \leq -|r|/\mu[A]$  for all  $r$ . Hence we have the following unexpected, but important, algebraic property:

PROPOSITION 2.3. *Let  $A \in \mathbb{C}^{d \times d}$ . Then*

$$(2.5) \quad \mu[A] < 0 \quad \Rightarrow \quad |A^{-1}| \leq -\frac{1}{\mu[A]}.$$

The condition  $\mu[A] < 0$  generalizes the notion of a negative definite matrix, and Proposition 2.3 is a special case of the *uniform monotonicity theorem*, [18, p. 167], [8, p. 147]. The propositions above also apply to bounded linear operators. Moreover, in (1.6) the matrix  $A$  could also be time-dependent. As  $\mu[A(t)]$  is then time-dependent as well, a simple modification of (2.1) is needed.

In *Stability and Error Bounds* [3], Dahlquist considered (1.5) as a variational equation modelling the error propagation in strongly stable time-stepping methods. Without going into full detail,  $r(t)$  then represents a defect, while  $x(t)$  represents the accumulated global error at time  $t$ . Assuming that  $\|r\|_\infty = \kappa h^p$ , where  $h$  is the step size and  $\kappa$  a constant, the global error is bounded by (2.3), where  $\mu[A]$  captures the dynamics of the problem. The method is *convergent*, as

$$(2.6) \quad |x(T)| \leq C_M(T) \cdot \kappa h^p, \quad \text{with } C_M(T) = \frac{e^{T\mu[A]} - 1}{\mu[A]}$$

at a fixed time  $T$ ; the global error can be made arbitrarily small by reducing  $h$ .

Before the advent of logarithmic norms, “classical” convergence proofs (still seen in many texts) derived estimates using the Lipschitz constant,  $|A|$ , essentially starting from (1.9). Then

$$(2.7) \quad |x(T)| \leq C_L(T) \cdot \kappa h^p, \quad \text{with } C_L(T) = \frac{e^{T|A|} - 1}{|A|}.$$

Although it mathematically proves convergence, this bound is seriously flawed. In his lectures, Dahlquist suggested considering e.g.  $|A| = 10$  and  $T = 10$ ; then  $C_L(T) \approx 3 \cdot 10^{42}$ . “Such ‘constants’ do not belong in numerical analysis!”, he exclaimed with fervor; and one learned to look twice at every  $O(h^p)$  claim.

Yet this is a benign example, for in stiff problems the product  $T|A|$  may be huge. For example, consider a parabolic equation solved by the method of lines. Then one typically has  $\mu_2[A] < 0$  in (2.6), implying that *the error is uniformly bounded* for  $t \geq 0$ . The bound (2.7), however, is now “wrong in principle,” as Dahlquist put it, [6], as it estimates the error for the ill-posed, reverse time problem. Since  $|A| = |-A|$ , these cannot be distinguished, see (1.9).

Still, (2.7) is often called “sharp,” although (2.6) is sharper as a matter of course. This can be illustrated by considering  $A = \lambda \in \mathbb{C}$ ; then (1.2) shows that (2.6) is sharp for any  $\lambda \in \mathbb{C}$ , while (2.7) is merely sharp on a set of measure zero, i.e., when  $\lambda \in \mathbb{R}^+$ . For more general spaces, however, even (2.6) often significantly overestimates the error. Nevertheless, its ability to distinguish between forward and reverse time sets it apart both in theory and practice.

Both [3, 8] and the surveys [9, 20] provide tables for how to compute  $\mu[A]$  for the most common norms, as well as numerous basic properties of the logarithmic norm. We refer to [8, p. 35] for a summary; most of these properties will be found below in a more general functional analytic setting.

In spite of its virtues, the logarithmic norm needs critical examination. In particular, as the original definition (1.1) excludes applications that are well within reach, it appears to be unfortunately chosen. For example, noting that a one-step method with stability function  $R(z)$  approximates  $e^{hA}$  by the rational matrix function  $R(hA)$ , we have

$$(2.8) \quad \mu[A] = \lim_{h \rightarrow 0^+} \frac{|R(hA)| - 1}{h}.$$

If used as a definition, (2.8) might admit *unbounded linear operators*, e.g. for subdiagonal Padé approximations like  $R(z) = 1/(1-z)$ . Moreover, inner product norms offer further alternatives for unbounded operators. Considerations of this kind eventually led to redefinitions of the logarithmic norm in order to extend its applicability.

### 3 Lipschitz maps

Leaving the classical theory, we first turn to nonlinear maps. For a long time linearization was used, with the logarithmic norm applied to the Jacobian. The modern approach, however, is functional analytic.

Let  $f : D \subset X \rightarrow X$ . We define two functionals, the *least upper bound* (lub) and *greatest lower bound* (glb) *Lipschitz constants*, by

$$(3.1) \quad L[f] = \sup_{u \neq v} \frac{|f(u) - f(v)|}{|u - v|}; \quad l[f] = \inf_{u \neq v} \frac{|f(u) - f(v)|}{|u - v|},$$

for  $u, v \in D$ , where the domain is assumed to be path-connected, see [21]. With these definitions, it holds that

$$(3.2) \quad l[f] \cdot |u - v| \leq |f(u) - f(v)| \leq L[f] \cdot |u - v|.$$

If  $l[f] > 0$ , then  $f(u) \rightarrow f(v)$  implies that  $u \rightarrow v$ . Then  $f$  is an injection, with an inverse on  $f(D)$ , just as a matrix  $A$  is invertible if its glb is strictly positive. Moreover,  $L[f^{-1}] = l[f]^{-1}$ , where  $L[f^{-1}]$  is naturally defined over  $f(D)$ .

One easily shows that the lub Lipschitz constant is an *operator semi-norm*, generalizing the matrix norm: if  $f = A$  is a linear map then  $L[A] = |A|$ . Hence  $L[\cdot]$  has left and right G-differentials. For the class of Lipschitz maps, this allows us to define two more functionals on  $D$ , the *lub logarithmic Lipschitz constant* and the *glb logarithmic Lipschitz constant*, by

$$(3.3) \quad M[f] = \lim_{h \rightarrow 0^+} \frac{L[I + hf] - 1}{h}; \quad m[f] = \lim_{h \rightarrow 0^-} \frac{L[I + hf] - 1}{h},$$

where we note that the lub logarithmic Lipschitz constant generalizes the usual logarithmic norm; for every matrix  $A$  we have  $M[A] = \mu[A]$ . The glb logarithmic

Lipschitz constant satisfies  $m[f] = -M[-f]$ ; this is an analogy to the relations  $l[f] = L[f^{-1}]^{-1}$  and  $\text{glb}[A] = |A^{-1}|^{-1}$ .

The logarithmic Lipschitz constants have a number of important properties that are useful in nonlinear analysis. We state them without proof; all of them are easily proved. The left inequality of Proposition 3.1.1 corresponds to the uniform monotonicity theorem, see also (3.4) below.

**PROPOSITION 3.1.** *Let  $f$  and  $g$  be Lipschitz on a domain  $D$ . Then*

1.  $-l[f] \leq M[f] \leq L[f]$
2.  $M[f + zI] = M[f] + \text{Re } z$
3.  $M[\alpha f] = \alpha M[f]$ ,  $\alpha \geq 0$
4.  $M[f] + m[g] \leq M[f + g] \leq M[f] + M[g]$ .

With the definitions above, the classical theory is fully extended to nonlinear problems. This includes initial value problems of the form  $\dot{x} = f(x)$ , as well as algebraic equations of the general form  $f(x) = y$ , [21]. For example, (2.5) becomes

$$(3.4) \quad M[f] < 0 \quad \Rightarrow \quad L[f^{-1}] \leq -\frac{1}{M[f]}.$$

A map with  $M[f] < 0$  is called uniformly negative monotone; (3.4) says that  $f$  is then an injection, and has a Lipschitz continuous inverse defined on  $f(D)$ . As a result, the equation  $f(x) = y$  has a unique solution  $x \in D$  for any  $y \in f(D)$ .

For the problem  $\dot{x} = f(x)$ , error bounds like (2.3) and (2.6) carry over, with  $\mu[A]$  replaced by  $M[f]$ . Hence questions of stability, convergence and error bounds do not require linearization in this setting. As for stability, the difference between two solutions  $u$  and  $v$  to  $\dot{x} = f(x)$  satisfies the differential inequality

$$(3.5) \quad D_t^+ |u - v| \leq M[f] \cdot |u - v|.$$

We now assume that  $(I + hf)(D) \subset D$  for  $0 \leq h < \varepsilon_0$ , and denote the flow by  $e^{tf} : x(0) \mapsto x(t)$  (where, obviously, the flow has the group properties  $e^{0f} = I$ ;  $(e^{tf})^{-1} = e^{-tf}$ ; and  $e^{tf}e^{sf} = e^{(t+s)f}$ ). Then we have, by (3.5),

$$(3.6) \quad L[e^{tf}] \leq e^{tM[f]}; \quad t \geq 0.$$

For example, if  $M[f] < 0$ , then  $L[e^{tf}] < 1$  on  $D$  for  $t > 0$ , implying that solutions to the dynamical system  $\dot{x} = f(x)$  are exponentially stable. Thus, *to a monotone vector field  $f$  there corresponds a contractive flow  $e^{tf}$ .*

#### 4 Monotonicity and contractivity

A few numerical methods preserve this relation between monotonicity and contractivity, and approximate the flow by *discrete contractive maps*. As an example, the implicit Euler method applied to  $\dot{x} = f(x)$  yields the recursion

$$(4.1) \quad x_{n+1} = (I - hf)^{-1}(x_n).$$

Consider the map  $g := hf - I$ , and assume that  $g(D) = X$  (this allows the recursion above). Apply (3.4) under the assumption  $M[g] < 0$  and use Proposition 3.1.2–3 to obtain, for  $h$  positive,

$$(4.2) \quad hM[f] < 1 \quad \Rightarrow \quad L[(I - hf)^{-1}] \leq \frac{1}{1 - hM[f]}.$$

In particular, if  $M[f] < 0$ , then  $(I - hf)^{-1}$  is a *contraction* for all  $h > 0$ . Not only is the continuous flow contractive, but so is the discrete flow. Moreover, this holds irrespective of the size of  $hL[f]$ , which makes the property of interest in the analysis of stiff systems.

In 1963, Dahlquist introduced A-stability [4], as a method criterion to guarantee that every scalar problem  $\dot{x} = \lambda x$  with  $\operatorname{Re} \lambda \leq 0$  also has bounded numerical solutions, replicating the mathematical behaviour. As we saw in Section 2, if a one-step method is applied to a linear system, we obtain  $x_{n+1} = R(hA)x_n$ . The method is A-stable if  $R$  maps the left half-plane  $\mathbb{C}^-$  into the unit disk. Interestingly, by a minor modification (from polynomials to rational functions) of an operator theoretical result due to von Neumann [25], we have the following:

**PROPOSITION 4.1.** *Let  $R(z)$  be a rational function, and let  $\|\cdot\|_2$  and  $\mu_2[\cdot]$  be the Euclidean norm and logarithmic norm, respectively. Let  $A$  be a linear operator. If  $\operatorname{Re} z \leq 0 \Rightarrow |R(z)| \leq 1$ , then  $\forall h > 0$ ,  $\mu_2[hA] \leq 0 \Rightarrow |R(hA)|_2 \leq 1$ .*

Thus every A-stable one-step method applied to a linear, constant coefficient, monotone problem in Hilbert space preserves contractivity. This raises new and important questions: can this result be generalized to e.g. (i) arbitrary norms; (ii) nonlinear dynamics; and (iii) multistep methods?

As for general norms, the answer is no: the method's order cannot exceed  $p = 1$  for contractivity in the max norm, [22]. Thus the implicit Euler method is essentially the only method for which we have a result of the type (4.2).

The increasing interest in nonlinear stability concepts led in the 1970s to the study of dissipative problems, where inner products were used. Thus, if we let  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  denote an inner product on a (complex) Hilbert space  $\mathcal{H}$ , we have

$$(4.3) \quad M_{\mathcal{H}}[f] = \sup_{u \neq v} \frac{\operatorname{Re} \langle u - v, f(u) - f(v) \rangle_{\mathcal{H}}}{\langle u - v, u - v \rangle_{\mathcal{H}}},$$

which is equivalent to (3.3) if  $f$  is Lipschitz. The functional  $M_{\mathcal{H}}[\cdot]$  still has all the properties in Proposition 3.1.

Again,  $f$  is uniformly negative monotone if  $M_{\mathcal{H}}[f] < 0$ , in which case the continuous flow is contractive. Would A-stability be sufficient in order to preserve contractivity? Ideas related to this question can be found already in [4], and were successively refined when Dahlquist introduced the notion of *G-stability* for multistep methods. In turn, Butcher introduced *B-stability* for one-step methods [1] and derived the necessary conditions on the method coefficients for B-stability. An example is the *trapezoidal rule* and the *implicit midpoint method*. Both are A-stable and share the same stability function,

$$R(z) = \frac{1 + z/2}{1 - z/2},$$

but only the implicit midpoint method is B-stable (preserves contractivity). When applied to  $\dot{x} = f(x)$ , they produce two different recursions,

$$(4.4a) \quad x_{n+1} = (I - (h/2)f)^{-1} \circ (I + (h/2)f)(x_n) \quad (\text{TR})$$

$$(4.4b) \quad x_{n+1} = (I + (h/2)f) \circ (I - (h/2)f)^{-1}(x_n). \quad (\text{IM})$$

Due to non-commutativity,  $R(z)$  alone cannot fully characterize a method in the nonlinear case, implying that the methods have different nonlinear stability properties. Here only the second method above can be shown to preserve contractivity, on account of the *polarization identity* in Hilbert space, [21]. This differs from the case in Proposition 4.1, where linear operators commute.

For multistep methods, Dahlquist proved that G-stability is equivalent to A-stability in 1978, [5]. In view of (4.4) this claim needs qualification: it only applies to the class of “one-leg” methods. But each multistep method has a “one-leg twin” method, with identical linear stability properties. For example, the implicit midpoint method is the one-leg twin of the trapezoidal rule. The point was that if the one-leg method was contractive, then its multistep twin could trivially be proved *stable*, although not contractive.

The desire to derive error bounds under the sole assumption  $M_{\mathcal{H}}[f] < 0$ , i.e., irrespective of the size of the Lipschitz constant  $hL_{\mathcal{H}}[f]$ , led to the theory of *B-convergence*, [11, 16]. This departs from classical convergence theory, as the usual Lipschitz continuity (boundedness) assumption is dropped.

For mathematical (but not numerical) applications a theory for unbounded operators already existed. In nonlinear dissipative evolution equations, the contractivity and convergence of the implicit Euler method are used as standard proof techniques for existence and uniqueness, [2]. Nevertheless, in order to consider a numerical treatment of unbounded operators, such as partial differential operators and their discretizations, one must consider alternative definitions of the logarithmic norm. For evolutions in Hilbert space (4.3) has recently been used to improve convergence results for Runge–Kutta and multistep methods, [12, 13]. For Banach space applications, special techniques are needed.

## 5 Unbounded operators in Banach space

In Banach spaces one can emulate some of the properties in Hilbert space theory by the use of semi-inner products, [7, p. 123].

DEFINITION 5.1. *Let  $X$  be a real Banach space. For  $u, v \in X$ , the left  $(\cdot, \cdot)_-$  and right  $(\cdot, \cdot)_+$  semi-inner products are defined by*

$$(5.1) \quad (u, v)_{\pm} = \|u\| \lim_{h \rightarrow 0_{\pm}} \frac{\|u + hv\| - \|u\|}{h}.$$

As every norm possesses left and right G-differentials, the limits in (5.1) exist. From (5.1) it follows that the semi-inner product induces the norm in the usual way:  $(u, u)_{\pm} = \|u\|^2$ . Conversely, if the norm  $\|\cdot\|$  is induced by a true inner

product, then  $(u, v)_\pm = \langle u, v \rangle$ . One of the most important properties of the semi-inner products is that they satisfy the Cauchy–Schwarz inequalities

$$(5.2) \quad -\|u\| \cdot \|v\| \leq (u, v)_\pm \leq \|u\| \cdot \|v\|.$$

The following elementary properties of the semi-inner products are straightforward consequences of the vector norm properties:

PROPOSITION 5.1. *For  $u, v, w \in X$ ,*

1.  $(u, -v)_\pm = -(u, v)_\mp$
2.  $(u, \alpha v)_\pm = \alpha(u, v)_\pm; \quad \alpha \geq 0$
3.  $(u, v)_- + (u, w)_\pm \leq (u, v + w)_\pm \leq (u, v)_+ + (u, w)_\pm$ .

Each equality and inequality of Proposition 5.1 is two-fold: one combines the left and right semi-inner products by choosing either the upper or the lower subscripts. Proposition 5.1.3 shows that the two semi-inner products are almost equal. Equality holds, i.e.  $(u, v)_+ = (u, v)_-$ , if and only if the dual space  $X^*$  is strictly convex. That condition excludes  $\|\cdot\|_\infty$ , but the two semi-inner products are nevertheless equal a.e. in this topology, [7, p. 124].

Semi-inner products are directly related to the upper Dini derivatives through

$$(5.3) \quad D_t^\pm \|u\| = \frac{(u, \dot{u})_\pm}{\|u\|^2} \|u\|,$$

see [7, p. 124]. Differential inequalities generalizing (1.6) and (3.5) are then obtained by defining the logarithmic Lipschitz constants accordingly. Therefore, in view of (4.3) and (5.3), we make the following, fully general, definition.

DEFINITION 5.2. *Let  $f : X \rightarrow X$ . The least upper bound logarithmic Lipschitz constants with respect to the semi-inner products  $(\cdot, \cdot)_\pm$  are defined by*

$$(5.4) \quad M^\pm[f] = \sup_{u \neq v} \frac{(u - v, f(u) - f(v))_\pm}{\|u - v\|^2}.$$

*Similarly, we define the greatest lower bound logarithmic Lipschitz constants by*

$$(5.5) \quad m^\pm[f] = \inf_{u \neq v} \frac{(u - v, f(u) - f(v))_\pm}{\|u - v\|^2}.$$

The classical definition (1.1) is a G-differential of a supremum, the Lipschitz constant. The extended definition (5.4) is essentially the converse, the supremum of an expression containing a G-differential. Although [20] comes close to suggesting a similar idea, this approach was never explored in the classical theory, as dissipative evolution equations were not considered.

Comparing (5.4) to (3.3), we note that  $M^-[f] \leq M^+[f] \leq M[f]$  for any  $f$ . However,  $M^-[f] = M^+[f] = M[f]$  if the norm is induced by an inner product. Further, the basic properties of the logarithmic Lipschitz constant (5.4) show that Definition 5.2 is compatible with, and covers, the previous theory:

PROPOSITION 5.2.

1.  $-l[f] \leq M^\pm[f] \leq L[f]$
2.  $M^\pm[f + zI] = M^\pm[f] + \operatorname{Re} z$
3.  $M^\pm[\alpha f] = \alpha M^\pm[f], \quad \alpha \geq 0$
4.  $M^\pm[f + g] \leq M^\pm[f] + M^+[g]$ .

The inequalities of Proposition 5.2.4 combine the left and right semi-inner products (as in Proposition 5.1, one chooses either the plus or the minus signs). Since subadditivity is important when deriving stability conditions and perturbation bounds, one may prefer to use the right semi-inner product and  $M^+[f]$ , even if marginally stronger results might be obtained with the left. As already mentioned, the left inequality of Proposition 5.2.1 merits special attention:

PROPOSITION 5.3. (Uniform Monotonicity Theorem) *Let  $D \subset X$  be path-connected and assume that  $f : D \rightarrow f(D) \subset X$ . Then*

$$(5.6) \quad M^\pm[f] < 0 \quad \Rightarrow \quad L[f^{-1}] \leq -\frac{1}{M^\pm[f]}.$$

Using semi-inner products, the proof is completely elementary. By combining the *lower* Cauchy–Schwarz inequality (5.2) with (5.4) we have

$$-\|u - v\| \cdot \|f(u) - f(v)\| \leq (u - v, f(u) - f(v))_\pm \leq M^\pm[f] \cdot \|u - v\|^2.$$

Hence

$$\sup_{u \neq v} -\frac{\|f(u) - f(v)\|}{\|u - v\|} \leq M^\pm[f] < 0$$

and it follows that

$$-l[f] \leq M^\pm[f] < 0,$$

see (3.1). Therefore,  $l[f] \geq -M^\pm[f] > 0$  implies that  $f^{-1}$  exists on  $f(D)$ , and

$$L[f^{-1}] \leq -\frac{1}{M^\pm[f]},$$

where  $L[f^{-1}]$  is defined on  $f(D)$ . If  $D = X$  one needs to prove, in addition, that  $f^{-1}$  is defined on all of  $X$ ; this usually follows by coercivity.

## 6 Unbounded operators in Hilbert space

Although the definitions above are general, Hilbert space theory offers the most important examples of how to use the logarithmic norm in infinite dimensional spaces. As a simple illustration, we consider the 1D reaction–diffusion problem

$$(6.1) \quad u_t = u_{xx} + g(u)$$

with boundary data  $u(t, 0) = u(t, 1) = 0$ . We let  $\mathcal{H} = L^2[0, 1]$  and consider functions  $u, v \in H_0^1 \cap H^2 \subset L^2[0, 1]$ , with the usual inner product and norm,

$$(6.2) \quad \langle u, v \rangle_{\mathcal{H}} = \int_0^1 u(x)v(x) \, dx; \quad \|u\|_{\mathcal{H}}^2 = \langle u, u \rangle_{\mathcal{H}}.$$

Introduce the operator  $f = \partial^2/\partial x^2 + g$ , consisting of the unbounded diffusion operator  $\partial^2/\partial x^2$ , with a bounded (Lipschitz) nonlinearity  $g$  representing the reaction part. The problem (6.1) is then an ODE on a Hilbert space,

$$(6.3) \quad \dot{u} = f(u).$$

We are interested in the stability of  $u(t, \cdot)$  as  $t \rightarrow \infty$ , and whether there is a unique equilibrium solution, satisfying the two-point boundary value problem

$$(6.4) \quad u'' + g(u) = 0; \quad u(0) = u(1) = 0,$$

where  $'$  denotes  $d/dx$ . The logarithmic Lipschitz constant answers both questions, the first via Dini derivatives and differential inequalities, and the second via the uniform monotonicity theorem.

We first need to find  $M_{\mathcal{H}}[d^2/dx^2]$  on  $H_0^1 \cap H^2[0, 1]$ . Integrating by parts, making use of the homogeneous boundary conditions, we find

$$\langle u, u'' \rangle_{\mathcal{H}} = -\langle u', u' \rangle_{\mathcal{H}} = -\int_0^1 |u'(x)|^2 \, dx \leq -\pi^2 \int_0^1 |u(x)|^2 \, dx = -\pi^2 \langle u, u \rangle_{\mathcal{H}},$$

where the standard Sobolev-type inequality at the center is easily proved by Fourier analysis. Hence  $\langle u, u'' \rangle_{\mathcal{H}} \leq -\pi^2 \langle u, u \rangle_{\mathcal{H}}$  for all  $u \in H_0^1 \cap H^2[0, 1]$ , with equality for the function  $u(x) = \sin \pi x$ . We therefore have:

PROPOSITION 6.1. *Let  $M_{\mathcal{H}}[\cdot]$  be defined by (4.3). Then, on  $H_0^1 \cap H^2[0, 1]$ ,*

$$(6.5) \quad M_{\mathcal{H}}[d^2/dx^2] = -\pi^2.$$

Equivalently,  $m_{\mathcal{H}}[-d^2/dx^2] = \pi^2$ . This quantifies the *ellipticity* of  $-d^2/dx^2$ . We now investigate the stability of (6.1). As  $M_{\mathcal{H}}[\cdot]$  is subadditive, it holds that

$$(6.6) \quad M_{\mathcal{H}}[f] = M_{\mathcal{H}}[\partial^2/\partial x^2 + g] \leq M_{\mathcal{H}}[\partial^2/\partial x^2] + M_{\mathcal{H}}[g] = -\pi^2 + M_{\mathcal{H}}[g].$$

Let  $u$  and  $v$  be two different solutions of (6.1). Then, using Dini derivatives and the differential inequality (3.5), we have

$$(6.7) \quad \|u(t, \cdot) - v(t, \cdot)\|_{\mathcal{H}} \leq e^{(M_{\mathcal{H}}[g] - \pi^2)t} \|u(0, \cdot) - v(0, \cdot)\|_{\mathcal{H}}.$$

Hence if the reaction term satisfies  $M_{\mathcal{H}}[g] < \pi^2$  (this allows “stiff” reactions but limits their instability) the solution  $u(t, \cdot)$  of the system is exponentially stable.

Moreover, whenever  $M_{\mathcal{H}}[g] < \pi^2$ , (6.4) is readily seen to have a unique solution. By (6.6), the map  $f = d^2/dx^2 + g$  is then uniformly negative monotone,

with a Lipschitz continuous inverse on  $L^2[0, 1]$  according to the uniform monotonicity theorem 5.3. This analysis applies to perturbations of the type

$$(6.8) \quad v'' + g(v) = p(x),$$

with  $p \in L^2[0, 1]$ , using the original, unperturbed boundary conditions for  $v$ . As  $L_{\mathcal{H}}[f^{-1}] \leq -1/M_{\mathcal{H}}[f]$ , one then obtains

$$(6.9) \quad \|u - v\|_{\mathcal{H}} \leq \frac{\|p\|_{L^2}}{\pi^2 - M_{\mathcal{H}}[g]}.$$

Hence  $p \rightarrow 0$  implies  $u \rightarrow v$ , and the solution therefore depends continuously on the data, cf. (2.4). For clarity, the norm of the “data space,”  $\|\cdot\|_{L^2}$ , is here distinguished from the norm on the “solution space”  $H_0^1 \cap H^2 \subset L^2$  of continuously differentiable functions. The conditions above are sufficient conditions only, and solutions may both exist and be unique under weaker conditions.

**Discretization.** The simplest finite difference method for (6.1) on an equidistant grid with  $\Delta x = 1/(N + 1)$  approximates  $u_{xx}$  by central differences

$$(6.10) \quad u_{xx} \approx \frac{v_{i-1} - 2v_i + v_{i+1}}{\Delta x^2},$$

with boundary conditions  $v_0 = 0$  and  $v_{N+1} = 0$ . The diffusion equation  $u_t = u_{xx}$  is therefore approximated by a method of lines ODE system,

$$(6.11) \quad \dot{v} = T_{\Delta x} v,$$

where  $v_i(t) \approx u(t, i \cdot \Delta x)$ , and the  $N \times N$  Toeplitz matrix  $T_{\Delta x}$ , approximating  $\partial^2/\partial x^2$  on the grid, is given by

$$(6.12) \quad T_{\Delta x} = \frac{1}{\Delta x^2} \text{tridiag}(1 \ -2 \ 1).$$

The eigenvalues of  $T_{\Delta x}$  are known analytically, and are given by

$$(6.13) \quad \lambda_k(T_{\Delta x}) = -4(N + 1)^2 \sin^2 \frac{k\pi}{2(N + 1)}; \quad k = 1, \dots, N.$$

For  $k \ll N$ , these are in very good agreement with the eigenvalues of  $\partial^2/\partial x^2$ , which are  $\lambda_k(\partial^2/\partial x^2) = -(k\pi)^2$  for  $k \in \mathbb{Z}^+$ . Moreover, as  $T_{\Delta x}$  is normal, its Euclidean logarithmic norm equals its spectral abscissa,  $\lambda_1(T_{\Delta x})$ . Hence

$$(6.14) \quad \mu_2[T_{\Delta x}] = -\frac{4}{\Delta x^2} \sin^2 \frac{\pi \Delta x}{2} = -\pi^2 + \Delta x^2 \pi^4/12 + O(\Delta x^4).$$

The logarithmic norm of the discretization is therefore very close to that of the continuous problem,  $M_{\mathcal{H}}[\partial^2/\partial x^2] = -\pi^2$ . The method of lines ODE (6.11) has exponentially stable solutions, as well as a unique equilibrium (which for the homogeneous diffusion problem is the trivial solution 0).

Can one compare  $\mu_2[\cdot]$  on  $\mathbb{R}^N$  to  $M_{\mathcal{H}}[\cdot]$  on  $H_0^1 \cap H^2$ ? The answer is yes: for the continuous problem the norm is the standard  $L^2$  norm, and

$$\|u\|_{\mathcal{H}}^2 = \int_0^1 |u(x)|^2 dx \approx \frac{1}{N+1} \sum_1^N |v_i|^2 = \Delta x \cdot |v|_2^2 =: \|v\|_{\Delta x}^2,$$

where the integral is approximated by the trapezoidal rule, and  $u(x_i)$  by  $v_i$ . The norm  $\|v\|_{\Delta x}$  is the discrete  $L^2$  norm, with the corresponding operator norm

$$(6.15) \quad \|A\|_{\Delta x}^2 = \sup_{v \neq 0} \frac{\|Av\|_{\Delta x}^2}{\|v\|_{\Delta x}^2} = \sup_{v \neq 0} \frac{\Delta x \cdot |Av|_2^2}{\Delta x \cdot |v|_2^2} \equiv |A|_2^2.$$

Likewise,  $\mu_{\Delta x}[A] \equiv \mu_2[A]$ . Therefore the Euclidean logarithmic norm  $\mu_2[T_{\Delta x}]$  is the discrete counterpart to the  $L^2$  logarithmic Lipschitz constant  $M_{\mathcal{H}}[\partial^2/\partial x^2]$ .

**Stability, CFL conditions, and stiffness.** Due to the stiffness of the system (6.11), explicit methods suffer severe step size restrictions caused by stability conditions. Consider the explicit Euler method with step size  $\Delta t$ ,

$$(6.16) \quad v_{j+1} = (I + \Delta t \cdot T_{\Delta x})v_j.$$

As the Euclidean norm is sharp for normal matrices in general and symmetric matrices such as  $T_{\Delta x}$  in particular, stability can in this particular case be studied in terms of eigenvalues. As  $I + \Delta t \cdot T_{\Delta x}$  is symmetric as well, the solution remains bounded if *and only if*

$$(6.17) \quad |I + \Delta t \cdot T_{\Delta x}|_2 \leq 1.$$

Here  $|I + \Delta t \cdot T_{\Delta x}|_2 = \rho[I + \Delta t \cdot T_{\Delta x}]$ . Therefore, by (6.13),

$$(6.18) \quad |I + \Delta t \cdot T_{\Delta x}|_2 = |1 + \Delta t \cdot \lambda_N(T_{\Delta x})| = |1 - 4 \frac{\Delta t}{\Delta x^2} (1 - O(\Delta x^2))|.$$

We may therefore approximate the stability condition (6.17) by

$$(6.19) \quad |1 - 4 \frac{\Delta t}{\Delta x^2}| \leq 1,$$

which leads to the classical CFL condition  $\Delta t/\Delta x^2 \leq 1/2$ .

The CFL condition can be linked to *stiffness*. In order to do this, we note that in the case above,  $|I + \Delta t \cdot T_{\Delta x}|_2 = \Delta t \cdot |T_{\Delta x}|_2 - 1$ . Hence the exact CFL condition

$$(6.20) \quad \Delta t \leq \frac{2}{|T_{\Delta x}|_2}.$$

This condition is identical to the classical view on stiffness around 1970: an explicit method requires  $\Delta t \cdot L[f]$  to be small, therefore (some) “problems with large Lipschitz constants” would be stiff.

But that characterization fails to distinguish forward and reverse time. Instead, we express (6.17) in terms of the glb logarithmic Lipschitz constant (cf. Section 3 and (5.5)), once again using the sharpness of the Euclidean norms:

$$(6.21) \quad |I + \Delta t \cdot T_{\Delta x}|_2 = -m_2[I + \Delta t \cdot T_{\Delta x}] = -1 - m_2[\Delta t \cdot T_{\Delta x}] \leq 1.$$

Hence the exact CFL condition can now be written

$$(6.22) \quad \Delta t \leq -\frac{2}{m_2[T_{\Delta x}]}, \quad \text{or equivalently} \quad \Delta t \leq \frac{2}{\mu_2[-T_{\Delta x}]}.$$

Just as  $L[e^{\Delta t \cdot f}] \leq e^{\Delta t \cdot M[f]}$  measures maximal perturbation growth over an interval  $\Delta t$  in forward time,  $L[e^{-\Delta t \cdot f}] \leq e^{-\Delta t \cdot m[f]}$  measures perturbation sensitivity in reverse time. This suggests that *stiffness can be measured by the product*  $|\Delta t \cdot m[f]|$ . Stiffness thus depends *in part on the problem*, as quantified by  $m[f]$ , and *in part on the method and accuracy criterion*, which together determine  $\Delta t$ .

The CFL condition (6.22) is interpreted as expressing the fact that the flow of (6.11) is “nearly” a semigroup. Therefore, the reverse time problem is numerically ill conditioned, as can be expected from time discretizations of parabolic problems. But this is a general property of stiff dissipative problems. Although not formulated mathematically, similar thoughts have occurred in the literature: “A way we prefer for describing the latter condition is that  $[v(t)]$  is very unstable in the opposite direction [of time]”, see [19, p. 4].

Turning now to implicit methods, we apply the implicit Euler method to (6.11):

$$(6.23) \quad v_{j+1} = (I - \Delta t \cdot T_{\Delta x})^{-1} v_j.$$

As a symmetric matrix has a symmetric inverse, (6.23) is stable if and only if

$$(6.24) \quad |(I - \Delta t \cdot T_{\Delta x})^{-1}|_2 \leq 1.$$

Now, as  $\mu_2[\Delta t \cdot T_{\Delta x}] \approx -\Delta t \cdot \pi^2 < 1$  we can apply (4.2) and obtain

$$(6.25) \quad |(I - \Delta t \cdot T_{\Delta x})^{-1}|_2 \leq \frac{1}{1 - \mu_2[\Delta t \cdot T_{\Delta x}]} \approx \frac{1}{1 + \Delta t \cdot \pi^2} < 1.$$

The method is *unconditionally stable* for  $\Delta t > 0$ ; there is no CFL condition.

Finally, if we look at the trapezoidal rule (Crank–Nicolson method) we have

$$(6.26) \quad v_{j+1} = \left(I - \frac{\Delta t}{2} T_{\Delta x}\right)^{-1} \left(I + \frac{\Delta t}{2} T_{\Delta x}\right) v_j,$$

and the method is stable if

$$(6.27) \quad \left| \left(I - \frac{\Delta t}{2} T_{\Delta x}\right)^{-1} \left(I + \frac{\Delta t}{2} T_{\Delta x}\right) \right|_2 \leq 1.$$

Here  $T_{\Delta x}$  is *linear*, allowing us to apply von Neumann’s Proposition 4.1. Hence, as  $\mu_2[T_{\Delta x}] < 0$ , the Crank–Nicolson method is contractive for every  $\Delta t > 0$ , showing its unconditional stability.

Above only the linear diffusion part has been treated. For the fully nonlinear reaction-diffusion problem one needs estimates based on the subadditivity of the logarithmic Lipschitz constant, see (6.6), but they may no longer be sharp.

**Error analysis, stability and convergence.** Logarithmic norms can also be used to analyze the stationary problem (6.4), which after discretization becomes

$$(6.28) \quad T_{\Delta x}v + G(v) = 0,$$

where  $v_i \approx u(x_i)$  and the function  $G$  is defined by  $G_i(v) = g(v_i)$ . The numerical solution satisfies the boundary conditions  $v_0 = v_{N+1} = 0$ .

To represent the exact solution of (6.4) we introduce the vector  $U$  with components  $U_i = u(x_i)$  and boundary conditions  $U_0 = U_{N+1} = 0$ . Similarly, let  $U''$  have components  $U''_i = u''(x_i)$ . Then  $u'' + g(u) = 0$  may be written

$$(6.29) \quad U'' + G(U) = 0$$

on the grid  $\{x_i\}$ . Inserting the exact, global, solution  $U$  into (6.28) we obtain

$$(6.30) \quad T_{\Delta x}U + G(U) = -\delta,$$

where the residual  $\delta = U'' - T_{\Delta x}U$  is the “local error.” (One hastens to add that “global residual” would technically be a more appropriate term.) By Taylor series expansion of  $u(x)$ , we find

$$(6.31) \quad \delta = U'' - T_{\Delta x}U = \gamma_1 \Delta x^2 + \gamma_2 \Delta x^4 + \dots$$

provided that  $u$  is sufficiently differentiable. Therefore, as  $\Delta x \rightarrow 0$ , we have

$$(6.32) \quad T_{\Delta x}U + G(U) \rightarrow U'' + G(U).$$

In other words, (6.28) is consistent with (6.4). In order to prove that the finite difference approximation is also convergent, we need to prove that the global error  $\varepsilon = v - U \rightarrow 0$  as  $\Delta x \rightarrow 0$ . In a formal error analysis we find the global error by solving (6.28) and (6.30). To this end we introduce the map  $F_{\Delta x} := T_{\Delta x} + G$ . If this map is invertible we have

$$(6.33a) \quad v = F_{\Delta x}^{-1}(0)$$

$$(6.33b) \quad U = F_{\Delta x}^{-1}(-\delta),$$

and hence the global error bound

$$(6.34) \quad \|v - U\|_{\Delta x} \leq L_2[F_{\Delta x}^{-1}] \cdot \|\delta\|_{\Delta x},$$

provided that the *stability condition*  $L_2[F_{\Delta x}^{-1}] < \infty$  holds. This will follow from the uniform monotonicity theorem, but as we are working with different spaces as  $\Delta x \rightarrow 0$ , we need to establish that  $F_{\Delta x}^{-1}$  exists and is *Lipschitz equicontinuous* with respect to  $\Delta x \rightarrow 0$ . This requires a single constant  $0 < C_h < \infty$  such that

$L_2[F_{\Delta x}^{-1}] \leq C_h$  for the *entire family* of maps  $\{F_{\Delta x}^{-1}\}$  with  $0 < \Delta x \leq h$ . Then convergence will follow from consistency in accordance with the Lax principle.

For the map  $F_{\Delta x}$  we have  $M_2[F_{\Delta x}] \leq \mu_2[T_{\Delta x}] + M_2[G]$ . Therefore, by (6.14), it holds that  $M_2[F_{\Delta x}] < 0$  if

$$(6.35) \quad M_2[G] < \pi^2 - \Delta x^2 \pi^4 / 12 + O(\Delta x^4).$$

However, for the continuous problem we saw that  $M_{\mathcal{H}}[g] < \pi^2$  guarantees a unique solution, cf. (6.9), and we need to show that  $M_2[G] \approx M_{\mathcal{H}}[g]$ .

In order to find  $M_2[G]$ , we need to compute, for suitably defined  $V$  and  $W$ ,

$$(6.36) \quad M_2[G] = \sup_{V \neq W} \frac{(V - W)^T (G(V) - G(W))}{|V - W|_2^2}.$$

Let  $v, w \in H_0^1 \cap H^2$  and define  $V_i = v(x_i)$  and  $W_i = w(x_i)$ . Then, by invoking the trapezoidal rule for numerical integration,

$$\begin{aligned} \langle v - w, g(v) - g(w) \rangle_{\mathcal{H}} &= \int_0^1 (v - w)(g(v) - g(w)) \, dx \\ &\approx \Delta x \sum_{i=1}^N (V_i - W_i)(g(V_i) - g(W_i)) \\ &= \Delta x (V - W)^T (G(V) - G(W)), \end{aligned}$$

and likewise  $\|v - w\|_{\mathcal{H}}^2 = \int_0^1 |v - w|^2 \, dx \approx \Delta x |V - W|_2^2$ . Both approximations hold with an error bounded by  $C\Delta x^2$ . Hence, by (6.36), it holds that

$$(6.37) \quad M_2[G] = M_{\mathcal{H}}[g] + O(\Delta x^2).$$

As a consequence, (6.35) is merely a discrete counterpart to the assumption  $M_{\mathcal{H}}[g] < \pi^2$ . Let us therefore suppose that  $M_2[F_{\Delta x}] < 0$ . By the uniform monotonicity theorem, it then follows that  $F_{\Delta x}^{-1}$  is indeed Lipschitz equicontinuous. Moreover, for all  $\Delta x \rightarrow 0$ , we have the uniform bound

$$(6.38) \quad L_2[F_{\Delta x}^{-1}] \leq - \frac{1}{\mu_2[T_{\Delta x}] + M_2[G]}.$$

Now, by combining (6.31) and (6.34) it follows that

$$(6.39) \quad \|v - U\|_{\Delta x} = \beta_1 \Delta x^2 + \beta_2 \Delta x^4 + \dots$$

Therefore the discretization method is convergent of second order.

## 7 Stability and error bounds: a functional analytic setting

A formal setting reveals that the analysis above is of general interest in numerical analysis. Consider the operator equation

$$(7.1) \quad \mathcal{F}(x) = y,$$

where  $\mathcal{F}$  is a bijection of the form  $\mathcal{F} : \text{Dom}(\mathcal{F}) \subset \xi + X \rightarrow \text{Im}(\mathcal{F}) \subset Y$ . Here  $X$  and  $Y$  are Banach spaces equipped with norms  $\|\cdot\|_X$  and  $\|\cdot\|_Y$ , respectively, and  $\xi + X$  denotes an affine space so that initial or boundary conditions can be accounted for. The solution of (7.1) is written  $x = \mathcal{F}^{-1}(y)$ .

If we introduce perturbed data  $\tilde{y} = y + \delta y$ , we obtain a perturbed solution  $\tilde{x} = x + \delta x = \mathcal{F}^{-1}(y + \delta y)$ . Stability is now a matter of bounding the deviation  $\delta x$  in terms of the data perturbation  $\delta y$ . For the two neighboring solutions

$$\begin{aligned} x + \delta x &= \mathcal{F}^{-1}(y + \delta y) \\ x &= \mathcal{F}^{-1}(y) \end{aligned}$$

the difference  $\delta x$  can be bounded by

$$(7.2) \quad \|\delta x\|_X \leq L_{XY}[\mathcal{F}^{-1}] \|\delta y\|_Y,$$

where the Lipschitz constant  $L_{XY}[\mathcal{F}^{-1}]$  is defined as

$$(7.3) \quad L_{XY}[\mathcal{F}^{-1}] = \sup_{y_1 \neq y_2} \frac{\|\mathcal{F}^{-1}(y_2) - \mathcal{F}^{-1}(y_1)\|_X}{\|y_2 - y_1\|_Y}.$$

In general, stability means that such a Lipschitz constant exists and is finite, i.e.,  $L_{XY}[\mathcal{F}^{-1}] \leq C$ . Stability is therefore equivalent to a continuous data dependence, referred to as “well-posedness” in mathematics. In numerical analysis, however, stability is not only a qualitative property. It is often quantified, and measured in terms of the size of  $L_{XY}[\mathcal{F}^{-1}]$ . Depending on the context, this Lipschitz constant is referred to as the “stability constant” or the (absolute) “condition number.” If small, the problem is considered “well-conditioned.”

**EXAMPLE 7.1** In the analysis of initial value problems  $\dot{x} = Ax$  with data  $x(0) = x_0$ , *Lyapunov stability* of the zero solution is characterized as a continuity condition:  $x = 0$  is *stable* if

$$(7.4) \quad \forall \varepsilon > 0 \quad \exists \delta > 0 : |x_0| < \delta \Rightarrow \|x\|_\infty < \varepsilon.$$

In terms of the flow of the ODE,  $e^{tA} : x_0 \mapsto x(t)$ , we have

$$(7.5) \quad |x(t)| \leq |e^{tA}| \cdot |x_0|.$$

Therefore the zero solution is stable if  $|e^{tA}| \leq C$  for all  $t \geq 0$ , in which case we can take  $\delta = \varepsilon/C$ . Thus Lyapunov stability follows if the one-parameter family (group) of maps  $\{e^{tA}\}$  is *equicontinuous* (with respect to matrix–vector multiplication) on  $t \in [0, \infty)$ . (Recall that a bounded linear operator is continuous.) In other words, the stability constant  $C$  must hold uniformly for all maps in the family. According to Proposition 2.1, the logarithmic norm condition  $\mu[A] \leq 0$  is sufficient for stability, and implies that we can take  $C = 1$ .

**EXAMPLE 7.2** In numerical analysis the *Lax Principle* plays an important role. The operator equation is often a parameterized family of *numerical problems*

$$(7.6) \quad \mathcal{G}_\theta(x_\theta) = y; \quad \theta \rightarrow 0+$$

that approximate a limiting *mathematical problem*  $\mathcal{F}(x) = y$ . The parameter  $\theta$  can e.g. be a discretization parameter or the accuracy of finite precision arithmetic; it is of interest to study stability as  $\theta \rightarrow 0+$ . We define the *error* by

$$(7.7) \quad \delta x_\theta = x_\theta - x.$$

The *residual* is found by inserting the exact solution  $x$  into the numerical problem, and it is defined by

$$(7.8) \quad \delta y_\theta = y - \mathcal{G}_\theta(x).$$

Assuming that  $\mathcal{F}$  and the family  $\{\mathcal{G}_\theta\}$  are invertible, we then formally have

$$\begin{aligned} \|x_\theta - x\|_X &= \|(\mathcal{G}_\theta^{-1} - \mathcal{F}^{-1})(y)\|_X \\ &= \|(\mathcal{G}_\theta^{-1} - \mathcal{G}_\theta^{-1}\mathcal{G}_\theta\mathcal{F}^{-1})(y)\|_X \\ &\leq L_{XY}[\mathcal{G}_\theta^{-1}] \cdot \|y - \mathcal{G}_\theta(x)\|_Y. \end{aligned}$$

We therefore have an error bound in terms of the residual,

$$(7.9) \quad \|\delta x_\theta\|_X \leq L_{XY}[\mathcal{G}_\theta^{-1}] \cdot \|\delta y_\theta\|_Y,$$

which illustrates the Lax Principle: *consistency and stability imply convergence*. These notions have the following precise interpretations:

- *Consistency*: the *residual* satisfies  $\|\delta y_\theta\|_Y \rightarrow 0$  as  $\theta \rightarrow 0+$ . This is equivalent to  $\mathcal{G}_\theta \rightarrow \mathcal{F}$  (limit in the sense of strong convergence).
- *Numerical stability*:  $L_{XY}[\mathcal{G}_\theta^{-1}] \leq C$  as  $\theta \rightarrow 0+$ . This is equivalent to *Lipschitz equicontinuity* of the family of inverse maps  $\{\mathcal{G}_\theta^{-1}\}$ .
- *Convergence*: the *error* satisfies  $\|\delta x_\theta\|_X \rightarrow 0$  as  $\theta \rightarrow 0+$ . This is equivalent to  $\mathcal{G}_\theta^{-1} \rightarrow \mathcal{F}^{-1}$  (limit in the sense of strong convergence).

The Lax Principle is therefore merely a matter of continuity: if the approximating family of numerical problems has equicontinuous inverses, then the numerical error can be made arbitrarily small by making the residual sufficiently small.

While it is usually straightforward to construct methods such that the residual is forced to zero (consistency), it is considerably harder to prove that the method is also numerically stable. As we have seen in the previous sections, the uniform monotonicity theorem offers a path around this difficulty: if  $M[\mathcal{G}_\theta] < 0$ , then  $L[\mathcal{G}_\theta^{-1}] \leq -1/M[\mathcal{G}_\theta]$ . Hence the logarithmic Lipschitz constant elegantly bridges the gap between the uniform monotonicity theorem and the Lax Principle: assuming that the numerical family of maps is negative monotone not only guarantees stability but immediately gives us the stability constant as well.

This approach, however, has its own difficulties. For example, one may know *a priori* that  $\mathcal{F}$  is monotone (see e.g. Proposition 6.1, or more generally, consider the Laplacian with Dirichlet boundaries), implying that  $M[\mathcal{F}] < 0$ . The crucial

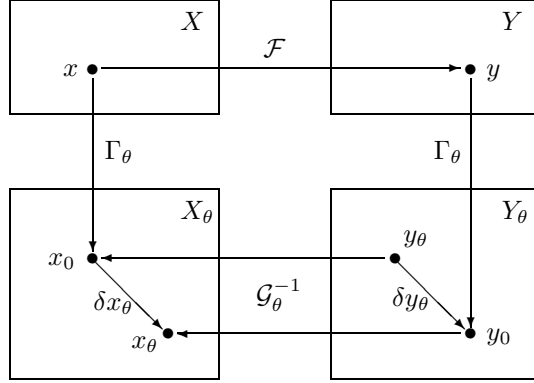


Figure 7.1: *Discretization of operator equation  $\mathcal{F}(x) = y$ .* Here  $\theta$  is a discretization parameter and  $\Gamma_\theta$  represents a restriction (grid map) from the space  $X$  to  $X_\theta$  and from  $Y$  to  $Y_\theta$ . Convergence requires  $\|\delta x_\theta\| \rightarrow 0$  as  $\theta \rightarrow 0+$ ; this will follow from consistency,  $\|\delta y_\theta\| \rightarrow 0$ , provided that we have numerical stability, i.e., if  $\mathcal{G}_\theta^{-1}$  is Lipschitz equicontinuous. Diagram adapted from Stetter [23].

question is then whether the discrete operator,  $\mathcal{G}_\theta$ , “inherits” that monotonicity, i.e., whether

$$(7.10) \quad M[\mathcal{F}] < 0 \quad \Rightarrow \quad M[\mathcal{G}_\theta] < 0.$$

Such a requirement typically restricts the choice of discretization methods; this is e.g. the central issue in the theory of  $B$ - and  $G$ -stability, see Section 4 above. Similarly, Section 6 above deals with the corresponding questions in a simple initial-boundary value problem, where we need to infer numerical stability from mathematical stability.

Example 7.2 above is simplified. Figure 7.1 offers a more complete illustration, in which the original mathematical problem,  $\mathcal{F}(x) = y$ , is set in a different space than the family of numerical problems,  $\mathcal{G}_\theta(x_\theta) = y_\theta$ . Here  $y_0 = \Gamma_\theta y$  is a restriction to a finite dimensional, discrete space. The analysis takes the form

$$\begin{aligned} \|\delta x_\theta\|_{X_\theta} &= \|(\mathcal{G}_\theta^{-1}\Gamma_\theta - \Gamma_\theta\mathcal{F}^{-1})(y)\|_{X_\theta} \\ &= \|(\mathcal{G}_\theta^{-1}\Gamma_\theta - \mathcal{G}_\theta^{-1}\mathcal{G}_\theta\Gamma_\theta\mathcal{F}^{-1})(y)\|_{X_\theta} \\ &\leq L_{X_\theta Y_\theta}[\mathcal{G}_\theta^{-1}] \cdot \|(\Gamma_\theta\mathcal{F} - \mathcal{G}_\theta\Gamma_\theta)(x)\|_{Y_\theta} \\ &= L_{X_\theta Y_\theta}[\mathcal{G}_\theta^{-1}] \cdot \|y_0 - \mathcal{G}_\theta(x_0)\|_{Y_\theta} = L_{X_\theta Y_\theta}[\mathcal{G}_\theta^{-1}] \cdot \|\delta y_\theta\|_{Y_\theta}. \end{aligned}$$

This is a *forward error analysis*. The role of the restriction operator  $\Gamma_\theta$  is of particular importance. In a *backward error analysis*, one would try to derive an error bound based on the assumption that the numerical solution can be identified to be the exact solution of a perturbed mathematical problem. The error bound would then be expressed in terms of the Lipschitz constant  $L_{XY}[\mathcal{F}^{-1}]$ .

Such a backward error bound is however impossible, *for structural reasons*: above  $\Gamma_\theta$  appears to the *left* of  $\mathcal{F}^{-1}$ , preventing the derivation of a rigorous backward error bound by extracting the desired Lipschitz constant. On the other hand,  $\Gamma_\theta$  appears to the *right* of  $\mathcal{G}_\theta^{-1}$ , and presents no difficulties there. Thus one can only derive a rigorous forward error bound. Consequently, *numerical stability is necessary*. Although this is known to every numerical analyst, it is of some interest that this is a structural (algebraic) property, due to non-commutativity, and independent of application or the type of discretization method.

## 8 Acknowledgments.

This paper is based on a series of lectures I gave at a stability workshop a few years ago. My aim was to develop the modern formalism of the logarithmic norm and give simple examples from a variety of applications, especially outside elementary matrix theory. I gave a first draft to Germund and suggested writing a joint paper. He liked the idea and suggested a division of work: “You write and I read!” For various reasons the project was never completed, but I hope that the present, condensed paper will be a worthwhile testimony to Germund’s scientific legacy, and to the scientific inspiration I owe to him.

Naturally, my interest in and views on the logarithmic norm were originally shaped by Germund’s teachings. Over the years, however, there were several other people whose contributions, interest and comments were very significant. In particular I would like to thank Marc Spijker, John Butcher, Des Higham, Nick Trefethen, Hans Stetter, Eitan Tadmor, Inmaculada Higuera, Jesper Opestrup, Eskin Hansen and Elza Farhi for many interesting discussions.

## REFERENCES

1. J.C. Butcher. A stability property of implicit Runge–Kutta methods, BIT 15 (1975), 358–361.
2. M.G. Crandall and T. Liggett. Generation of semigroups of nonlinear transformation on general Banach spaces. Amer. J. Math. 93 (1971), 265–298.
3. G. Dahlquist. Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations. Almqvist & Wiksells, Uppsala 1958; Transactions of the Royal Institute of Technology, Stockholm 1959.
4. G. Dahlquist. A special stability problem for linear multistep methods. BIT 3 (1963), 27–43.
5. G. Dahlquist. G-stability is equivalent to A-stability. BIT 18 (1978), 384–401.
6. G. Dahlquist, personal communication, c. 1985.
7. K. Deimling. Nonlinear Functional Analysis. Berlin: Springer-Verlag 1985.
8. K. Dekker and J.G. Verwer. Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations. New York: North Holland 1984.

9. C. Desoer and H. Haneda. The measure of a matrix as a tool to analyze computer algorithms for circuit analysis. *IEEE Trans. Circuit Theory* 19 (1972), 480–486.
10. A. Dontchev and F. Lempio. Difference methods for differential inclusions: a survey, *SIAM Review* 34 (1992), 263–294.
11. R. Frank, J. Schneid and C.W. Ueberhuber. The concept of B-convergence. *SIAM J. Num. Anal.* 18 (1981), 753–780.
12. E. Hansen. Convergence of multistep time discretizations of nonlinear dissipative evolution equations, To appear in *SIAM J. Numer. Anal.*
13. E. Hansen. Runge-Kutta time discretizations of nonlinear dissipative evolution equations, To appear in *Math. Comp.*
14. I. Higuera and B. García-Celayeta. Logarithmic norms of matrix pencils. *SIAM J. Matrix Anal.* (1997), 646–666.
15. I. Higuera and G. Söderlind. Logarithmic norms and nonlinear DAE stability. *BIT* 42 (2002), 823–841.
16. J.F.B.M. Kraaijevanger. B-convergence of the implicit midpoint rule and the trapezoidal rule. *BIT* 25 (1985), 652–666.
17. S.M. Lozinskii. Error Estimates for the Numerical Integration of Ordinary Differential Equations, Part I. *Izv. Vyss. Uceb. Zaved Matematika* 6 (1958), 52–90. (In Russian)
18. J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press 1970.
19. L.F. Shampine. What is stiffness? In *Stiff Computation*, ed. R.C. Aiken. New York: Oxford 1985.
20. T. Ström. On logarithmic norms. *SIAM J. Num. Anal.* 2 (1975), 741–753.
21. G. Söderlind. Bounds on Nonlinear Operators in Finite-dimensional Banach Spaces. *Num. Math.* 50 (1986), 27–44.
22. M. Spijker. Contractivity in the numerical solution of initial value problems. *Num. Math.* 42 (1983), 271–290.
23. H.-J. Stetter. *Analysis of discretization methods for ordinary differential equations*. Berlin–Heidelberg–New York: Springer-Verlag 1973.
24. L.N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton: Princeton University Press 2005.
25. J. von Neumann. Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes. *Math. Nachrichten* 4 (1951), 258–281.