Global, rigorous and realistic bounds for the solution of dissipative differential equations. Part I: Theory

Arnold Neumaier

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974-0636
U.S.A.

Abstract. It is shown how interval analysis can be used to calculate rigorously valid enclosures of solutions to initial value problems for ordinary differential equations. In contrast to previously known methods, the enclosures obtained are valid over larger time intervals, and for uniformly dissipative systems even globally.

This paper discusses the underlying theory; main tools are logarithmic norms and differential inequalities. Numerical results will be given in a subsequent paper.

Zusammenfassung. Es wird gezeigt, wie man mit Hilfe von Intervall-Analysis rigorose Einschließungen von Lösungen von Anfangswertproblemen bei gewöhnlichen Differentialgleichungen berechnen kann. Im Gegensatz zu anderen Methoden sind die Einschließungen über größere Zeitintervalle, und für gleichmäßig dissipative Systeme sogar global gültig.

Diese Arbeit behandelt die zugrundeliegende Theorie; Hauptwerkzeuge sind logarithmische Normen und Differentialungleichungen. Numerische Ergebnisse werden in einer späteren Arbeit vorgestellt.

March 1993

Keywords: initial value problem, rigorous enclosure, Peano existence theorem, differential inequalities, interval arithmetic, logarithmic norms

1991 MSC Classification: primary 65L70, secondary 65G10, 65L07

1 Introduction

The solution of initial value problems for ordinary differential equations has proceeded to the stage where one can not only compute approximate solutions automatically, but also give (approximate) accuracy estimates based on local control of truncation error versus roundoff error. But due to the diversity of behaviour of dynamical systems, this local error control can be unreliable when a certain *global* accuracy need to be achieved.

There are methods for rigorous error control going back to MOORE [11] which are based on interval arithmetic (see [12] for a modern treatment of interval analysis). However, Moore observed that naive methods can lead to severe overestimation even on simple problems, due to so-called *wrapping* (cf. [13]). The current best rigorous code, due to LOHNER [9] takes measures against wrapping. It has no automatic step size control, but techniques of EIJGENRAAM [3] allow to control the step size adaptively. However, both Lohner's and Eijgenraam's methods use initial bounds related to explicit ODE methods like Euler's, and thus have severe step size restrictions for stiff systems.

In this paper, we

- relate local errors and global errors, using one-sided Lipschitz conditions (Theorem 2.8);
- survey the properties of logarithmic norms, needed for explicit work with the one-sided Lipschitz condition;
- prove a new existence theorem (Theorem 3.5) giving conditions under which an initial value problem has a solution which remains close (in a quantitatively specified sense) to a given approximation;
- give explicitly a set of sufficient conditions verifiable by computer (using interval arithmetic), and show that for uniformly dissipative problems, these conditions give global bounds for *all* times, with a global error of the approximation;
- indicate an adaptive strategy for the automatic enclosure of solutions of general initial value problems, with the property that no step size restrictions are expected for stiff problems.

2 Logarithmic norms

In this section we review and extend the known properties of logarithmic norms. Some of the results discussed here are not needed for the remaining sections, but are included for the sake of completeness.

Logarithmic norms were introduced by DAHLQUIST [2] and LOZINSKIJ [10]. They are extensively used in the book by COPPEL [1] (in particular, pp. 3, 41, 59) and the article by STRÖM [17], where further properties and references may be found.

Let V be a Banach space and define, for $u, v \in V$, $u \neq 0$,

$$\mu_h(u,v) := \frac{\|u+hv\| - \|u\|}{h\|u\|}. \tag{1}$$

2.1. Proposition. *For $h > 0$, $\mu_h(u,v)$ is monotone increasing in h and bounded from below by $-\|v\|/\|u\|$; hence the limit*

$$\mu(u,v) := \limsup_{h\to+0} \mu_h(u,v) = \lim_{h\to+0} \mu_h(u,v) = \inf_{h>0} \mu_h(u,v) \tag{2}$$

exists.

Proof. By the triangle inequality, $\Big|\|u+hv\| - \|u\|\Big| \leq \|hv\|$ and, for $h \geq k$,

$$\mu_h(u,v) - \mu_k(u,v) = \frac{\|h^{-1}u+v\| + (k^{-1}-h^{-1})\|u\| - \|k^{-1}u+v\|}{\|u\|} \geq 0. \quad \square$$

$\mu(u,v)$ is called the *logarithmic derivative* of the vector norm $\|\cdot\|$; cf. Remark 2.4 below. We first note some simple properties:

2.2. Proposition.

(i) We have

$$\|u+hv\| \geq \|u\|(1+h\mu(u,v)), \tag{3}$$

and, for $h \to 0$,

$$\|u+hv\| = \|u\|(1+h\mu(u,v)) + o(h).$$

(ii) We have

$$\mu(u,u) = 1 \quad \textit{for} \quad u \neq 0, \tag{4}$$

$$\mu(u,v) \leq \frac{\|v+su\|}{\|u\|} - s \quad \textit{for} \quad s \geq 0, \tag{5}$$

$$\mu(\alpha u, v) = \frac{1}{\alpha}\mu(u,v), \quad \mu(u,\alpha v) = \alpha\mu(u,v) \quad \textit{for} \quad \alpha > 0, \tag{6}$$

$$|\mu(u,v) - \mu(u,w)| \leq \frac{\|v-w\|}{\|u\|}. \tag{7}$$

Proof. All statements are straightforward consequences of (1) after taking limits; for (5) use $s = 1/h$. □

Central for the application of logarithmic norms is the following result, again a direct consequence of the definition.

2.3. Proposition. *The forward derivative*

$$\partial^+ f(t) := \lim_{h \to +0} \frac{f(t+h) - f(t)}{h} \tag{8}$$

of the norm of a differentiable vector function x is given by

$$\partial^+ \|x(t)\| = \mu(x(t), \dot{x}(t))\, \|x(t)\|. \tag{9}$$

In (9), the 2-sided derivative exists iff $\mu(x, -\dot{x}) = -\mu(x, \dot{x})$. □

2.4. Remark. If $N : V \to \mathbb{R}_+$ given by $N(u) := \|u\|$ is Frechet-differentiable at u, it follows that

$$\mu(u,v) = \frac{N'(u)v}{N(u)} = (\log N(u))'v.$$

2.5. Examples.

(i) For the norm $\|x\|_2 = \sqrt{\langle x|x\rangle}$ in a Hilbert space with inner product $\langle \cdot \mid \cdot \rangle$ we have

$$\mu_2(u,v) = \frac{\mathrm{Re}\langle u|v\rangle}{\|u\|^2},$$

and the norm $\|\cdot\|_2$ is smooth for $x \neq 0$.

(ii) For the norm $\|x\|_\infty$ in $V = \mathbb{R}^n$ we have

$$\mu_\infty(u,v) = \max\left\{(\operatorname{sgn} u_i)v_i \mid i \ \text{ with } \ |u_i| = \|u\|_\infty\right\},$$

and the norm $\|x\|_\infty$ is smooth if x has a unique absolutely largest component. □

The usefulness of logarithmic norms can be seen from the following *stability theorem.*

2.6. Theorem. *Assume the* **one-sided Lipschitz condition**

$$\mu(x-y, F(t,x)-F(t,y)) \le \mu_F(t) \quad \textit{for all} \quad x, y \in \mathbb{R}^n \tag{10}$$

and let

$$\kappa(s,t) := \int_s^t \mu_F(\tau)d\tau, \qquad [= (t-s)\mu_F \ \textit{in the autonomous case.}] \tag{11}$$

Then, for any two solutions x_1, x_2 of

$$\dot{x}(t) = F(t, x(t)), \tag{12}$$

the difference

$$r(s,t) := e^{-\kappa(s,t)}\|x_1(t) - x_2(t)\|$$

is monotone decreasing in t.

Proof. By (9) and (8), $\partial^+\|x_1(t)-x_2(t)\| \le \mu_F(t)\cdot\|x_1(t)-x_2(t)\|$,

$$\partial^+ r(s,t) = -\mu_F(t)r(s,t) + e^{-\kappa(s,t)}\,\partial^+\|x_1(t)-x_2(t)\| \le 0. \quad \square$$

2.7. Corollary. *For $t > s$, for any two solutions x_1, x_2 of* (11),

$$\|x_1(t) - x_2(t)\| \le e^{\kappa(s,t)} \|x_1(s) - x_2(s)\|. \tag{13}$$

As a consequence we can deduce the following result on local error propagation, which appears to be new.

2.8. Theorem. *Assume the one-sided Lipschitz condition* (10). *Let* $0 = t_0 < t_1 < \cdots$ *be a grid such that*

$$\int_{t_i}^{t_{i+1}} \mu_F(t)dt \le \kappa < 0 \quad \textit{for each } i. \tag{14}$$

For a solution $x(t)$ of (12), *consider an approximating grid function x_i, $i = 0, 1, \ldots$ whose* **local error (per step)** *satisfies*

$$\|x_i(t_{i+1}) - x_{i+1}\| \le r, \qquad i = 0, 1, \ldots \tag{15}$$

where $x_i(t)$ is the solution of (12) *with $x_i(t_i) = x_i$, then*

$$\|x(t_i) - x_i\| \le \max\left(\|x(0) - x_0\|, \frac{r}{1 - e^{\kappa}}\right). \tag{16}$$

Proof. Let $\epsilon_i := \|x(t_i) - x_i\|$, then, for $i = 0, 1, \ldots$

$$\epsilon_{i+1} \le \|x(t_{i+1}) - x_i(t_{i+1})\| + \|x_i(t_{i+1}) - x_{i+1}\| \le e^{\kappa} \cdot \epsilon_i + r \qquad \text{by (13)–(15).}$$

This implies, since $\kappa < 0$,

$$\epsilon_i \le \epsilon_0 e^{i\kappa} + r\frac{1 - e^{i\kappa}}{1 - e^{\kappa}} = \left(\epsilon_0 - \frac{r}{1 - e^{\kappa}}\right) e^{i\kappa} + \frac{r}{1 - e^{\kappa}} \le \max\left(\epsilon_0, \frac{r}{1 - e^{\kappa}}\right). \quad \square$$

2.9. Remarks. (i) If $\{x_i\}$ is generated by a one-step method

$$x_{i+1} := x_i + h_i F_{\text{num}}(h_i, t_i, x_i), \qquad h_i = t_{i+1} - t_i, \tag{17}$$

the local error bound r of (15) is a bound for $\rho_i + h_i\sigma_i$ where ρ_i is the local roundoff error (per step) and σ_i is the local discretization error.

(ii) In principle, one could use this for global error control by providing estimates for κ and r at each step.

(iii) Note that $\dfrac{1}{1 - e^{\kappa}} = -\dfrac{1}{\kappa} + \dfrac{1}{2} - \dfrac{\kappa}{12} + O(\kappa^2)$.

Working with $\mu(u,v)$ directly is sometimes cumbersome, and can be simplified using bounds in terms of logarithmic matrix norms. For a linear mapping $A \in \mathrm{Lin}(V)$ of V into itself (a $n\times n$ - matrix if $V = \mathbb{R}^n$), we define its *norm*

$$\|A\| := \sup_{u\neq 0} \|Au\|/\|u\| \tag{18}$$

and its *logarithmic norm*

$$\mu(A) := \limsup_{h\to+0} \left(\|I + hA\| - 1\right)/h. \tag{19}$$

Note that both $\|A\|$ and $\mu(A)$ may be infinite if $\dim V = \infty$, but we always have, from the triangle inequality,

$$\mu(A) \leq \|A\|, \qquad \mu(A+B) \leq \mu(A) + \mu(B). \tag{20}$$

Clearly, (18) implies

$$\|Au\| \leq \|A\|\,\|u\|, \tag{21}$$

and from (2), (19) we find the inequality

$$\mu(u, Au) \leq \mu(A), \tag{22}$$

and hence by (7) the important bound

2.10. Proposition.

$$\mu(u,v) \leq \mu(A) + \|v - Au\|/\|u\| \quad \textit{if} \quad u \neq 0. \tag{23}$$

With an appropriate choice of A, this formula yields computable bounds for $\mu(u,v)$ which are sufficiently good for the applications in Section 4.
In the finite-dimensional case (22) is sharp, i.e., we have

$$\mu(A) = \sup_{u\neq 0} \mu(u, Au). \tag{24}$$

The logarithmic norm is related to the *spectral abscissa*

$$\alpha(A) := \sup\{\mathrm{Re}\,\lambda \mid \lambda \in \mathrm{Spec}\,A\} = \lim_{h\to 0} \frac{\rho(I+hA) - 1}{h}, \tag{25}$$

which satisfies

$$\alpha(A) \le \mu(A) \le \|A\|. \tag{26}$$

In general,

$$\mu_2(A) = \alpha(A_{\text{sym}}) = \sup\{\lambda \mid \lambda \in \operatorname{Spec} A_{\text{sym}}\},$$

where

$$A_{\text{sym}} := \frac{1}{2}(A + A^*),$$

and for $n \times n$-matrices,

$$\mu_\infty(A) = \max\{\operatorname{Re} A_{ii} + \sum_{k \ne i} |A_{ik}| \mid i = 1, \dots n\}.$$

In particular,

$$\mu_2(A) \le \mu \iff \mu I - A_{\text{sym}} \text{ positive semidefinite.} \tag{27}$$

We have $\alpha(A) = \mu(A)$

- if $\|\cdot\|$ is monotone and A is diagonal, or
- if $\|\cdot\| = \|\cdot\|_\infty$ and A is quasimonotone, i.e. its off-diagonal entries are nonnegative, or
- if $\|\cdot\| = \|\cdot\|_2$ and A is normal (and in particular if A is self-adjoint).

Further properties of $\mu(A)$ are

$$\mu(\alpha A) = \alpha\mu(A) \quad \text{if} \quad \alpha > 0, \tag{28}$$

$$\mu(A + \alpha I) = \mu(A) + \operatorname{Re}\alpha. \tag{29}$$

2.11. Proposition. *The following inequalities hold:*

$$\|e^{At}\| \le e^{\mu(A)t} \tag{30}$$

$$\|(sI - A)^{-1}\| \le (\operatorname{Re} s - \mu)^{-1} \quad \textit{if} \quad \mu(A) \le \mu < \operatorname{Re} s \tag{31}$$

$$\|(I - A)^{-1}(I + A)\|_2 \le \begin{cases} 1 & \textit{if} \quad \mu(A) \le 0 \\ \frac{1+\mu}{1-\mu} & \textit{if} \quad \mu(A) \le \mu, \mu \in (0,1). \end{cases} \tag{32}$$

$$\mu_2(A) - \alpha(A) \le \sqrt{\tfrac{1}{2}(\operatorname{tr} A^*A - |\operatorname{tr} A^2|)} \tag{33}$$

Proof. For (30): For $\dot{x}(t) = Ax(t)$, we obtain $\mu_F = \mu(A)$ in (10). With $x_1(0) = x_0$, $x_2(0) = 0$, we have from (13), with $s = 0$,

$$\|x_1(t) - x_2(t)\| = \|e^{At}x_0\| \le e^{\mu(A)t}\|x_0\|$$

which implies (30).
For (31): Let $B = (sI - A)^{-1}$ so that $(sI - A)B = I$ giving

$$|1 + hs|\,\|B\| = \|B + hsB\| = \|(I + hA)B + hI\| \le \|I + hA\|\,\|B\| + h,$$
$$\frac{1}{\|B\|} + \frac{\|I + hA\| - 1}{h} \ge \frac{|1 + hs| - 1}{h}.$$

In the limit $h \to 0$, we find $\|B\|^{-1} + \mu(A) \ge \operatorname{Re} s$. With $\mu(A) \le \mu < \operatorname{Re} s$, we may conclude $\|B\| \le (\operatorname{Re} s - \mu)^{-1}$. For (32) and (33), see STRÖM [17] □

2.12. Remark. (32) does not hold for $\|\cdot\|_\infty$ in place of $\|\cdot\|_2$; e.g., if $A = \begin{pmatrix} -2 & 2 \\ 0 & 0 \end{pmatrix}$, then $\mu_\infty(A) = 0$, but $(I - A)^{-1}(I + A) = \frac{1}{3}\begin{pmatrix} -1 & 4 \\ 0 & 3 \end{pmatrix}$ has norm $\frac{5}{3} > 1$. Im particular, this implies that the following result of HAIRER ET AL. [6] does not generalize to $\|\cdot\|_\infty$.

2.13. Theorem.

(i) Suppose $R(z)$ is analytic in $\operatorname{Re} z < 0$, continuous on $\operatorname{Re} z = 0$. If

$$|Rz| \le 1 \quad \textit{for all} \quad z \in \mathbb{C} \quad \textit{with} \quad \operatorname{Re} z \le 0$$

then

$$\mu_2(A) \le 0 \Rightarrow \|R(A)\|_2 \le 1$$

(ii) Suppose $R(z)$ is analytic in $\operatorname{Re} z < \mu$, continuous on $\operatorname{Re} z = \mu$. Then

$$\|R(A)\|_2 \le \varphi_R(\mu_2(a)) \quad \textit{where} \quad \varphi_R(\mu) := \sup\{|R(z)| \mid \operatorname{Re} z \le \mu\}.$$

This theorem can be refined further; see SCHMITT [15].

For practical applications to rigorous enclosures, it is important to be able to calculate strict bounds for logarithmic norms using approximate arithmetic only.

Using a guess μ_0 for $\mu_2(A)$, one can compute a rigorous bound for $\mu_2(A)$ as follows.

Calculate an approximate modified Cholesky factorization

$$\mu_0 I - A_{\text{sym}} \approx LL^T - E \tag{34}$$

with diagonal $E \geq 0$ (using, e.g., the algorithm of SCHNABEL AND ESKOW [16]), and observe that (20) and (27) imply for arbitrary L

$$\mu_2(A) \leq \mu_0 + \|\mu_0 I - A_{\text{sym}} - LL^T\|_2 \,. \tag{35}$$

The norm term bounds rounding errors and truncation errors in the modified Cholesky factorization. In this special case where A_{sym} is nearly diagonal, sufficiently good bounds are already obtained by using (35) with $L = 0$ and $\mu_0 = \min A_{ii}$.

With the use of interval arithmetic and $\|B\|_2 \leq \|B\|_F = \sqrt{\operatorname{tr} BB^T}$, a rigorous upper bound for (35) can be found. If the initial guess was good, then $E \approx 0$ in (34) and the correction term in (35) will be small. If the correction term is large, one can repeat the process with an improved μ_0, obtained, e.g., by a few Lanczos iterations with A_{sym}. For a related technique to bound smallest singular values of matrices see RUMP [14].

3 A semilocal existence theorem

In this section we use differential inequalities and Peano's existence theorem for solutions of initial value problems to deduce verifiable conditions that the solution of an initial value problem exists and remains in a prescribed tube for some calculable time interval. We begin with an auxiliary result which establishes a sufficient condition that a function remains ≤ 0.

3.1. Lemma. *Let $f : [\underline{t}, \bar{t}\,] \to \mathbb{R}$ be a continuous function. If there are constants $\gamma, \delta \in \mathbb{R}$, $\delta > 0$, such that, for $t \in [\underline{t}, \bar{t}\,[$, the implication*

$$0 \leq f(t) \leq \delta \quad \Rightarrow \quad \lim_{h \to +0} \frac{f(t+h) - f(t)}{h} \leq \gamma f(t) \tag{1}$$

holds, then

$$f(\underline{t}) \le 0 \quad \Rightarrow \quad f(t) \le 0 \quad \textit{for all} \quad t \in [\,\underline{t}, \bar{t}\,]. \tag{2}$$

Proof. For given $\epsilon \in (0, \delta/(e^{\gamma \bar{t}} - e^{\gamma \underline{t}}))$, let T be the set of $t \in [\,\underline{t}, \bar{t}\,]$ where

$$f(t) \le \epsilon(e^{\gamma t} - e^{\gamma \underline{t}}). \tag{3}$$

Note that (3) implies $f(t) \le \delta$ for $t \in T$. We will show that $T = [\,\underline{t}, \bar{t}\,]$ if $f(\underline{t}) \le 0$, independently of ϵ; hence $\epsilon \to 0$ will yield (2).

a) Take $t \in T$, $t < \bar{t}$, and assume $f(t) \ge 0$. By (1), for every $\epsilon > 0$, there is a positive $\bar{h} \le \bar{t} - t$ such that

$$f(t+h) \le (1 + \gamma h) f(t) + \epsilon_0 h \quad \text{for} \quad h \in [0, \bar{h}].$$

We choose $\epsilon_0 = \epsilon \gamma e^{\gamma \underline{t}}$ and find, with $1 + \gamma h \le e^{\gamma h}$ and (3),

$$\begin{aligned} f(t+h) &\le e^{\gamma h} \epsilon (e^{\gamma t} - e^{\gamma \underline{t}}) + \epsilon \gamma e^{\gamma \underline{t}} h \\ &= \epsilon (e^{\gamma (t+h)} - e^{\gamma \underline{t}}) - \epsilon e^{\gamma \underline{t}} (e^{\gamma h} - 1 - \gamma h) \\ &\le \epsilon (e^{\gamma (t+h)} - e^{\gamma \underline{t}}) \end{aligned}$$

Thus $t + h \in T$ for $h \in [0, \bar{h}]$.

b) Now assume $f(t) < 0$ at $t < \bar{t}$, $t \in T$. By continuity, there is a positive $\bar{h} \le \bar{t} - t$ such that $f(t+h) \le 0$ and $t + h \in T$ for $h \in [0, \bar{h}]$.

For $f(\underline{t}) \le 0$, there is a maximal t^* such that $[\,\underline{t}, t^*] \le T$ since T is closed. By a) and b), $t^* = \bar{t}$. $\square$

3.2. Remarks. (i) Instead of constant γ, we may assume $\gamma = \dot{\beta}(t)$ where $\beta : [\,\underline{t}, \bar{t}\,] \to \mathbb{R}$ is continuously differentiable. The proof works, with γt replaced by $\beta(t)$ and similar changes. But this extended form seems not to be more useful.

(ii) One cannot put $\delta = 0$ in (1). A counterexample is: $\underline{t} = -1$, $\bar{t} = +1$, $f(t) = t^3$.

The following *comparison theorem* is a generalization of the well-known Gronwall inequality (see e.g. [7]).

3.3. Theorem. *For $s > 0$, let $u : [0, s] \to V$ (a Banach space), $\varphi : [0, s] \to \mathbb{R}_+ = \{x \in \mathbb{R} \mid x > 0\}$ be continuously differentiable functions. For fixed $\delta > 0$, let t_δ be the infimum of all $t \in [0, s]$ where the following two relations are simultaneously satisfied:*

$$\varphi(t) \le \|u(t)\| \le \varphi(t) + \delta \tag{4}$$

$$\dot{\varphi}(t) \le \mu\left(u(t), \dot{u}(t)\right)\, \varphi(t); \tag{5}$$

but if (4) and (5) are incompatible, let $t_\delta = s$. Then

$$\|u(0)\| \le \varphi(0) \quad \Rightarrow \quad \|u(t)\| \le \varphi(t) \quad \textit{for all} \quad t \in [0, t_\delta]. \tag{6}$$

Proof. The function $f : [0, s] \to \mathbb{R}$ defined by

$$f(t) := \|u(t)\| - \varphi(t) \tag{7}$$

is continuous. Hence the set $T := \{t \in [0, t_\delta] \mid 0 \le f(t) \le \delta\}$ is either empty (in which case there is nothing to prove) or compact. In this case,

$$\sup_{t\in T} \|\dot{u}(t)\| < \infty, \qquad \inf_{t\in T} \|u(t)\| \ge \inf_{t\in T} \varphi(t) > 0 \tag{8}$$

so that we can define (cf. (2.5))

$$\gamma := \sup_{t\in T} \mu(u(t), \dot{u}(t)) \le \sup_{t\in T} \|\dot{u}(t)\|/\|u(t)\| < \infty. \tag{9}$$

Take $t \in T$, $t < t_\delta$, so that, by the construction of t_δ, (5) cannot hold because (4) holds:

$$\dot{\varphi}(t) > \mu\left(u(t), \dot{u}(t)\right)\, \varphi(t). \tag{10}$$

For $h > 0$ and $t + h \in T$ we have

$$\begin{aligned}
\|u(t+h)\| &= \|u(t) + h\dot{u}(t)\| + o(h) \\
&= \left(1 + h\,\mu\left(u(t), \dot{u}(t)\right)\right) \|u(t)\| + o(h) \\
&= \left(1 + h\,\mu\left(u(t), \dot{u}(t)\right)\right) f(t) + \left(1 + h\,\mu\left(u(t), \dot{u}(t)\right)\right) \varphi(t) + o(h) \\
&\le (1 + h\gamma) f(t) + \varphi(t) + h\dot{\varphi}(t) + o(h)
\end{aligned}$$

by (2.3),(7), (9) and (10), so that

$$f(t+h) = \|u(t+h)\| - \varphi(t+h) \le (1 + h\gamma) f(t) + o(h)\,.$$

Hence (1) holds for $\underline{t} = 0$, $\bar{t} = t_\delta$ and (6) follows from (2) by the Lemma. □

3.4. Remark. Clearly, t_δ is a decreasing function of δ, hence the conclusion (6) is strongest for $\delta \to 0$. It would be interesting to show that $t_0 = \sup_{\delta>0} t_\delta$; then we could put $\delta = 0$ in (4). However, at present I cannot exclude the possibility that $t_0 > \sup_{\delta>0} t_\delta$.

We shall now apply the comparison theorem (Theorem 3.3) to give a constructive existence test for a solution of the initial value problem

$$F(t, x(t), \dot{x}(t)) = 0 \quad \text{with} \quad x(t_0) = x_0, \quad \dot{x}(t_0) = z_0. \tag{11}$$

Here F is a mapping from $\Omega \subseteq \mathbb{R} \times V \times V$ into V where $\Omega \supseteq D \times E$, $D \subseteq \mathbb{R} \times V$, $E \subseteq V$, and the initial values satisfy

$$F(t_0, x_0, z_0) = 0, \qquad (t_0, x_0) \in \operatorname{int} D, \qquad z_0 \in \operatorname{int} E. \tag{12}$$

Explicit ordinary differential equations are obtained as the special case

$$F(t, x, z) = F_0(t, x) - z\,; \tag{13}$$

however, it will be useful to consider the implicit form (11) since the solution of (11) for $\dot{x}$ may complicate the expression and lead to additional overestimations.

Actually, (11), a *differential-algebraic equation* (DAE), includes much more general situations than (13). We will consider only DAEs *of index zero:* For each triple (t_0, x_0, z_0) satisfying (12) there are neighborhoods $U \subseteq D$ of (t_0, x_0) and $U' \subseteq E$ of z_0 such that, for every $(t, x) \in U$, the equation $F(t, x, z) = 0$ has a unique solution $z \in U'$ and z depends continuously on (t, x).

By the local implicit function theorem, F has index zero in $D \times E$ if it is continuous in $D \times E$, continuously differentiable with respect to z, and if the partial derivative $F_z(t, x, z)$ has a bounded inverse for $(t, x, z) \in D \times E$. In particular, F has index zero if

$$F(t, x, z) = F_0(t, x) - G(t, x)z, \tag{14}$$

with continuous $F_0 : D_0 \to V$ and $G : D_0 \to \operatorname{Lin}(V)$, and if $G(t, x)$ has a bounded inverse for $(t, x) \in D_0$. Clearly, this covers the case (13) of explicit ODEs with continuous F_0.

The index zero property may be tested, either by symbolic computation or by numeric computation with intervals.

In the following, we aim to construct, for a solution $x(t)$ of (11), enclosures of the form

$$\|S^{-1}(x(t_0+h)-p(h))\| \le \varphi(h) \quad \text{for} \quad 0 \le h \le \bar{h}. \tag{15}$$

Here,

- $p(h)$ is a "known" approximation of an "unknown" solution $x(t_0+h)$ which a priori need not even be known to exist,
- S is an invertible linear mapping $\in \mathrm{Lin}(V)$ which, for $\|\cdot\| = \|\cdot\|_2$, defines the axes of an error ellipsoid.
- φ is a "simple" positive function (constant, linear, exponential) which bounds the error.

The comparison theorem may be used to prove the following sufficient conditions for a bound (15), with a time-dependent linear mapping $S(h)$.

3.5. Theorem. *Let $s > 0$, $D \subseteq \mathbb{R} \times V$ closed, $E \subseteq V$ compact, $D \times E \subseteq \Omega \subseteq \mathbb{R} \times V \times V$, and suppose that $F : \Omega \to V$ has index zero in $D \times E$.*

Let $p : [0,s] \to V$, $S : [0,s] \to \mathrm{Lin}(V)$, $\varphi : [0,s] \to \mathbb{R}_+$ be continuously differentiable and $S(h)$ invertible for $h \in [0,s]$. Let h^* be the infimum of all $h \in [0,s]$ for which there exist $u, v \in V$ such that

$$F\left(t_0+h,\, p(h)+S(h)u,\, \dot{p}(h)+\dot{S}(h)u+S(h)v\right) = 0, \tag{16}$$

$$\left(t_0+h,\, p(h)+S(h)u,\, \dot{p}(h)+\dot{S}(h)u+S(h)v\right) \in \partial(D \times E), \tag{17}$$

$$\|u\| \le \varphi(h). \tag{18}$$

For fixed $\delta > 0$, let h_δ be the infimum of all $h \in [0,h^*]$ for which there exist $u, v \in V$ such that the following relations are simultaneously satisfied: (16) *and*

$$(t_0+h,\, p(h)+S(h)u,\, \dot{p}(h)+\dot{S}(h)u+S(h)v) \in D \times E \tag{19}$$

$$\varphi(h) \le \|u\| \le \varphi(h)+\delta \tag{20}$$

$$\dot{\varphi}(h) \le \mu(u,v)\,\varphi(h). \tag{21}$$

If (12) *holds and*

$$\|S(0)^{-1}(x_0 - p(0))\| \le \varphi(0) \tag{22}$$

then any continuously differentiable solution $x : [t_0, t_0 + \bar{h}] \to V$ *of the initial value problem* (11) *with* $\bar{h} \in]0, h_\delta]$ *can be extended to a continuously differentiable solution* $x : [t_0, t_0 + h_\delta] \to V$ *which satisfies*

$$(t, x(t)) \in D, \qquad \dot{x}(t) \in E \quad for \quad t \in [t_0, t_0 + h_\delta] \tag{23}$$

$$\|S(h)^{-1}(x(t_0 + h) - p(h))\| \le \varphi(h) \quad for \quad h \in [0, h_\delta]. \tag{24}$$

3.6. Remarks. (i) $h \le h^*$, defined by (16)–(18), keeps the solution away from the boundary of $D \times E$ whereas $h \le h_\delta$, defined by (16) and (19)–(21) keeps the solution within (24).
(ii) If $D = [\,\underline{t}, \infty) \times V$, $E = V$, then $h^* = s$ since (17) is never satisfied.

Proof. Consider the solution $x : [t_0, t_0 + \bar{h}\,] \to V$ of (11) of the Theorem, $\bar{h} \in [0, h_\delta]$.
(i) At first we show that (23), (24) hold for $t \in [t_0, t_0 + \bar{h}]$:
Let $h' \le \bar{h}$ be maximal such that (23), (24) hold for $t \in [t_0, t_0 + h']$, with $h' \ge 0$ by (12) and (22); suppose $h' < \bar{h}$. For $0 \le h \le \bar{h}$, let

$$u := u(h) := S(h)^{-1}\,(x(t_0 + h) - p(h))\,, \qquad v := v(h) = \dot{u}(h). \tag{25}$$

Then, for $t := t_0 + h$,

$$x(t) = p(h) + S(h)u, \qquad \dot{x}(t) = \dot{p}(h) + \dot{S}(h)u + S(h)v,$$

so that (16) holds.
If $(t, x(t), \dot{x}(t)) \in \partial(D \times E)$ for some $h \in [0, h']$, then (17) holds and (18) follows from (24). Thus $h \ge h^*$ contradicting $h^* \ge h_\delta \ge \bar{h} > h' \ge h$. Therefore

$$(t, x(t), \dot{x}(t)) \in \mathrm{Int}(D \times E) \quad \text{for all} \quad h \in [0, h'] \tag{26}$$

and (23) and hence (19) holds for $h \in [0, h'']$, $h'' \in (h', \bar{h}\,]$ sufficently close to h'.
Now we apply the comparison theorem (Theorem 3.3) with h in place of t: (20) and (21) correspond to (4)and (5) (cf. (25)), hence (6) asserts that (22)

implies (24) for $h \le \min(\bar{h}, h_\delta) = \bar{h}$. Thus h' has not been maximal and $h' = \bar{h}$.

(ii) Now we show that the solution may be extended to $t_0 + h_\delta$:

Suppose $\bar{h} < h_\delta$, then the argument above (26) yields $(\bar{t}, x(\bar{t}), \dot{x}(\bar{t})) \in \mathrm{Int}(D \times E)$ for $\bar{t} = t_0 + \bar{h}$. Since F has index 0 there are neighborhoods $U \subseteq D$ of $(\bar{t}, x(\bar{t}))$ and $U' \subseteq E$ of $\dot{x}(\bar{t})$ such that, for $(t, x) \in U$, the equation $F(t, x, z) = 0$ has a unique solution $z = z(t, x) \in U'$ depending continuously on (t, x).

By Peano's theorem, our original solution x of (11) in $[t_0, t_0 + \bar{h}]$ may be somewhat further extended by the solution of $\dot{x'}(t) = z(t, x(t))$ which also satisfies (11). Let t^* be the supremum of all $t' \le t_0 + h_\delta$ such that $x(t)$ can be extended to t^*, and assume $t^* < t_0 + h_\delta$.

Then we choose an increasing sequence $t_l \to t^*$ and extend each solution x_l in $[t_0, t_l]$ to a solution x_{l+1} in $[t_0, t_{l+1}]$. Thus $x(t) := x_l(t)$ for $t \in [\, t_l, t_{l+1})$, $l = 0, 1, \ldots$, is a solution in $[\, t_0, t^*)$. Since $\dot{x}(t) \in E$, $\beta := \sup \|\dot{x}(t)\|$ is finite and $x(t)$ is Lipschitz continuous. This implies that $x(t_l)$ is a Cauchy sequence whose limit, used as $x(t^*)$, extends x continuously to $[t_0, t^*]$.

Since E is compact, $\{\dot{x}(t_l)\}$ has a convergent subsequence with limit $z^* \in E$ which satisfies $F(t^*, x(t^*), z^*) = 0$, and $z^* = \lim_{t \to t^*} \dot{x}(t)$ since F has index 0. Thus $x(t)$ is a continuously differentiable solution of (11) in $[t_0, t^*]$ and can be further extended by the previous arguments. Hence $t^* = t_0 + h_\delta$. □

4 Bounds for initial value problems

In this section we show how logarithmic norms can be used to obtain global, rigorous and realistic enclosures for a class of ordinary differential equations containing those satisfying a uniform dissipation condition. This is done by rewriting Theorem 3.5 in a form more amenable to computer calculation. In particular, the global optimization problem for the determination of h_δ in Theorem 3.5 can be avoided if we do not insist on finding the optimal h_δ. Suboptimal lower bounds may be obtained by global linearization using *arithmetic on sets* (e.g., interval arithmetic). We use the following notation: Let $a, b \in V$, $A, B \in \mathrm{Lin}(V)$, $f : \mathrm{Lin}(V) \to \mathbb{R}$; $[A]$ denotes a "set of A's", $[a]$

a "set of a's" etc. Then

$$\begin{aligned} f([A]) &:= \{f(A) \mid A \in [A]\} \subset \mathbb{R} \\ [A] + [B] &:= \{A + B \mid A \in [A], B \in [B]\} \subseteq \mathrm{Lin}(V) \\ [A] \cdot [B] &:= \{A \cdot B \mid A \in [A], B \in [B]\} \subseteq \mathrm{Lin}(V) \\ [A]^I[B] &:= \{A^{-1}B \mid A \in [A], B \in [B]\} \subseteq \mathrm{Lin}(V) \\ [A]^I[b] &:= \{A^{-1}b \mid A \in [A], b \in [b]\} \subseteq V \\ [A] \cdot [b] &:= \{A \cdot b \mid A \in [A], b \in [b]\} \subseteq V \\ &\text{etc.} \end{aligned}$$

Such sets are introduced to control the rounding errors and the nonlinearities. Therefore, we may use "supersets" of the specified sets (i.e. sets including them) in an obvious fashion where necessary or convenient. In particular, we may use interval arithmetic (see e.g. Neumaier [12]) to calculate boxes containing these sets.

As in the previous section we consider the initial value problem

$$F(t, x(t), \dot{x}(t)) = 0 \quad \text{with} \quad x(t_0) = x_0, \quad \dot{x}(t_0) = z_0, \tag{1}$$

where the initial values satisfy

$$F(t_0, x_0, z_0) = 0, \qquad (t_0, x_0) \in \operatorname{int} D, \qquad z_0 \in \operatorname{int} E, \tag{2}$$

and we consider enclosures of the form

$$\|S(h)^{-1}(x(t_0 + h) - p(h))\| \le \varphi(h) \quad \text{for} \quad 0 \le h \le \bar{h}. \tag{3}$$

For the sake of simplicity, the following theorem is formulated only for the case where F is defined for all x, z, so that $h^\partial = s$ by Remark 3.6(ii). The method extends to the general case but leads to a very messy formulation.

The transformation of Theorem 3.5 to computable form is based on linearization of the problem function (2) in a neighborhood of the approximate solution. Instead of truncating the Taylor series we maintain rigor by using the mean value theorem for the linearization. Thus we get an exact linear expression for F – or rather a preconditioned form CF, cf. (5) –, however with coefficients which depend on unknown intermediate points. These coefficients can be enclosed rigorously by intervals, using interval arithmetic.

With this linear formulation, one can simplify the condition of Theorem 3.5 by using properties of the logarithmic norm (in particular, Proposition 2.10). This reduces computations to finding rigorous upper bounds for some interval expressions (namely (6) – (9) below) and a simple check on the closure condition.

We shall first give a general version of the linearization (Theorem 4.1) and then a constructively computable version (Proposition 4.3). Then we show (Corollary 4.5) that under suitable conditions, bounds can be obtained over arbitrarily long time intervals. An example how these results are applied is given in Example 4.11, after a discussion of natural choices for the various quantities occuring in the conditions guaranteeing the bounds.

4.1. Theorem. *Let* $\dim(V) < \infty$. *Let* $\omega, s > 0$, $D = [t_0, t_0 + s] \times V$, *and suppose that* $F : D \times V \to U$ *has index* 0 *in* $D \times V$. *Let* $p : [0, s] \to V$ *and* $S : [0, s] \to \mathrm{Lin}(V)$ *be continuously differentiable,* $S(h)$ *invertible for* $h \in [0, s]$.

Suppose that there are sets $[a], [b] \subset V$, $[A], [B] \subset \mathrm{Lin}(V)$ *such that, for* $u, v \in V$ *with*

$$\|u\| < \omega \tag{4}$$

and $h \in [0, s]$

$$C(h)\cdot F\left(t_0 + h, p(h) + S(h)u, \dot{p}(h) + \dot{S}(h)u + S(h)v\right) = a + bh + Bu - Av \tag{5}$$

for suitable $a \in [a]$, $b \in [b]$, $A \in [A]$, $B \in [B]$, *and* $C : [0, s] \to \mathrm{Lin}(V)$.

Suppose further that all $A \in [A]$ *are invertible, and define real constants* μ, α, β, γ *such that*

$$\mu\left([A]^I[B]\right) \le \mu \tag{6}$$

$$\|S(0)^{-1}(x_0 - p(0))\| \le \alpha \tag{7}$$

$$\|[A]^I[a]\| + \alpha\mu \le \beta \tag{8}$$

$$\|[A]^I[b]\| + \beta\mu \le \gamma \tag{9}$$

and the function $\varphi : [0, s] \to \mathbb{R}_+$ *with*

$$\varphi(h) := \alpha + \beta h + \gamma h^2 \exp_2(\mu h) \tag{10}$$

where

$$\exp_2(\tau) := \begin{cases} (e^\tau - 1 - \tau)/\tau^2 & \text{for } \tau \neq 0, \\ \frac{1}{2} & \text{for } \tau = 0. \end{cases} \tag{11}$$

If (2) *and the* **closure condition**

$$\varphi(h) < \omega \quad \text{for} \quad h \in [0, \bar{h}], \qquad 0 < \bar{h} \leq s, \tag{12}$$

hold, then there exists a continuously differentiable solution $x : [t_0, t_0 + \bar{h}] \to V$ *of* (1) *which satisfies* (3).

4.2. Remark. ω of (4) is an *a priori estimate* of the (transformed) error of $p(h)$. C is a *preconditioner* for the implicit formulation (1) of the differential system. The regularity assumption for $[A]$ is a strengthening of the index 0 hypothesis. In (6), (8), (9), it would suffice to use $[A^{-1}B]$, $[A^{-1}a]$ and $[A^{-1}b]$ respectively.

Proof. Without loss of generality, we restrict ourselves to the compact set $D = [t_0, t_0 + s] \times D_0$ where D_0 is a compact set containing all $p(h) + S(h)u$ for $h \in [0, s]$, $\|u\| \leq \omega$ in its interior. Since F has index 0, $F(t, x, z) = 0$ has, for each $(t, x) \in D$ one solution z depending continuously on (t, x); therefore the range of $z(t, x)$ is in the interior of a bounded set E (compact because $\dim(V) < \infty$). Thus D and E have the properties required in Theorem 3.5. For sufficiently small $\epsilon > 0$, let

$$\varphi_\epsilon(h) := \alpha + (\beta + \epsilon)h + (\gamma + \epsilon\mu)h^2 \exp_2(\mu h) \tag{13}$$

and take $\delta = \delta(\epsilon) > 0$ so small that

$$\varphi_\epsilon(h) + \delta \leq \omega \quad \text{for} \quad h \in [0, \bar{h}]; \tag{14}$$

this is possible because of (12). After some computation we find that

$$\dot{\varphi}_\epsilon(h) = \mu\varphi_\epsilon(h) + (\beta - \alpha\mu) + (\gamma - \beta\mu)h + \epsilon. \tag{15}$$

We wish to apply the semilocal existence theorem (Theorem 3.5) with φ_ϵ in place of φ and $\bar{h} \leq h_\delta$ (and $h^* = s$, see above). Assume $\bar{h} > h_\delta$, i.e. there exists an $h < \bar{h}$ such that (3.16) and (3.19–21) are simultaneously satisfied for some $u, v \in V$.

By (3.20) and (14), (4) is satisfied; hence by (3.16) and (5), we have $0 = a + bh + Bu - Av$, i.e.

$$v = A^{-1}(a + bh) + A^{-1}Bu,$$

with suitable $a \in [a]$, $b \in [b]$, $A \in [A]$, $B \in [B]$. By Proposition 2.10, this implies

$$\mu(u, v) \leq \mu(A^{-1}B) + \|A^{-1}(a + bh)\|/\|u\|$$

and by (3.20) and (3.21), with φ replaced by φ_ϵ, we get

$$\dot{\varphi}_\epsilon(h) \leq \mu(u, v)\varphi_\epsilon(h) \leq \mu(A^{-1}B)\varphi_\epsilon(h) + \|A^{-1}(a + bh)\|.$$

By (6), (8), (9), this implies

$$\dot{\varphi}_\epsilon(h) \leq \mu\varphi_\epsilon(h) + (\beta - \alpha\mu) + (\gamma - \beta\mu)h$$

which is a contradiction to (15). Hence $\bar{h} \leq h_\delta$.

Since (3.22) is a consequence of (7) and (13), the assumptions of Theorem 3.5 are valid, and there exists a solution $x(t)$ of (1) satisfying (3), with φ_ϵ in place of φ, for $h \in [0, \bar{h}]$. With $\epsilon \to 0$, the conclusion of the theorem is obtained. □

Theorem 4.1 can be applied constructively once it is known how to find reasonable enclosures for a, b, A, B in (5). We now consider this problem for the most important special case

$$F(t, x, z) = F_0(t, x) - Gz, \tag{16}$$

$G \in \mathrm{Lin}(V)$ invertible with bounded inverse. (More general situations with $\|F(t, x, z) - F(t, x, 0)\| \geq \gamma(t, x)\|z\|$ can be treated in a similar but messier way.)

For simplicity we shall force $b = 0$; this simplifies the formulae a little without degrading the enclosure much.

4.3. Proposition. *Suppose*

$$[t] = [t_0, t_0 + s], \tag{17}$$

$$[x] \supseteq \{p(h) + S(h)u \mid h \in [0, s], \|u\| \leq \omega\}, \tag{18}$$

$$[H] \supseteq \text{ closed convex hull of } \left\{ C \cdot \frac{\partial F_0}{\partial x}(t,x) \;\middle|\; t \in [t], x \in [x] \right\}, \tag{19}$$

$$[a] \supseteq \{CF_0(t_0 + h, p(h)) - (CG)\dot{p}(h) \mid h \in [0,s]\} \tag{20}$$

$$[A] \supseteq \{(CG)S(h) \mid h \in [0,s]\} \tag{21}$$

$$[B] \supseteq \left\{ H \cdot S(h) - (CG)\dot{S}(h) \;\middle|\; h \in [0,s], H \in [H] \right\}; \tag{22}$$

then (5) *holds with* $a \in [a]$, $b = 0$, $A \in [A]$, $B \in [B]$. *Moreover,* (9) *is satisfied with* $\gamma = \beta\mu$, *and* (10) *simplifies to*

$$\varphi(h) = \alpha + \beta h \exp_1(\mu h) \tag{23}$$

where

$$\exp_1(\tau) := \begin{cases} (e^\tau - 1)/\tau & \text{for } \tau \neq 0, \\ 1 & \text{for } \tau = 0. \end{cases} \tag{24}$$

Proof. By the mean value theorem,

$$C \cdot F_0\left(t_0 + h, p(h) + S(h)u\right) = C \cdot F_0\left(t_0 + h, p(h)\right) + H \cdot S(h)u, \tag{25}$$

where

$$H = C \cdot \int_0^1 \frac{\partial F_0}{\partial x}\left(t_0 + h, p(h) + \tau S(h)u\right) d\tau \in [H].$$

Thus,

$$\begin{aligned} & C \cdot F(t_0 + h, p(h) + S(h)u, \dot{p}(h) + \dot{S}(h)u + S(h)v) \\ &= C \cdot F_0(t_0 + h, p(h) + H \cdot S(h)u - (CG)(\dot{p}(h) + \dot{S}(h)u + S(h)v) \\ &= C \cdot (F_0(t_0 + h, p(h)) - G\dot{p}(h)) + (H \cdot S(h) - (CG)\dot{S}(h))u - (CG)S(h)v, \end{aligned}$$

which is (5) with the asserted enclosures. (23) is straightforward. □

4.4. Remarks. (i) A sharper enclosure of the form (25) can be obtained by using slopes (Krawczyk & Neumaier [8]); this saves some computational effort and reduces the radius of $[H]$ by roughly a factor of 2.

(ii) Care must be taken to get a realistic enclosure of the preconditioned residual (20) since this generally involves substantial cancellation. It is important to use a centered form or a boundary value form (cf. Neumaier

[12]) for the full expression in (20), perhaps together with some splitting of the interval over which h ranges.
(iii) In the enclosure of $[a]$, $[A]$, and $[B]$, the products CG and $CF_0'([t],[x])$ (cf.(19)) should be explicitly computed (enclosed to cover round-off) to reduce overestimation. (It is difficult to exploit any sparsity structure present since C is generally dense.)

We finally show that the quality of the attained bounds must be quite good for dissipative systems since we can deduce from Theorem 4.1 the following.

4.5. Corollary. *Let $F : [0,\bar{t}] \times \mathbb{R}^n \to \mathbb{R}^n$, let $S \in \mathbb{R}^{n\times n}$ be invertible, and let $p : [0,\bar{t}\,] \to \mathbb{R}^n$ be an approximate solution of the initial value problem*

$$\dot{x}(t) = F(t,x(t)), \qquad x(0) = x_0, \tag{26}$$

in the sense that

$$\|S^{-1}(x_0 - p(0))\| \le \delta \tag{27}$$
$$\|S^{-1}(F(t,p(t)) - \dot{p}(t))\| \le \epsilon \qquad \textit{for} \quad t \in [0,\bar{t}\,]. \tag{28}$$

If

$$\mu\left(S^{-1}\frac{\partial F}{\partial x}(t,x)S\right) \le \mu \quad \textit{for} \quad t \in [0,\bar{t}\,], \quad x \in \mathbb{R}^n, \tag{29}$$

then (26) has a solution $x : [0,\bar{t}\,] \to \mathbb{R}^n$ satisfying

$$\|S^{-1}(x(t) - p(t))\| \le \delta e^{\mu t} + \epsilon t \exp_1(\mu t) \quad \textit{for} \quad t \in [0,\bar{t}\,]. \tag{30}$$

Proof. In Proposition 4.3 we put $t_0 = 0$, $S(h) = S$, $C = S^{-1}$, $G = I$, $[a] := \{r \in \mathbb{R}^n \mid \|r\| \le \epsilon\}$, $[H] = \left\{S^{-1}\frac{\partial F}{\partial x}(t,x) \mid t \in [0,\bar{t}\,], x \in \mathbb{R}^n\right\}$, $[A] := I$, $[B] = \{HS \mid H \in [H]\}$; from (7), (8) we obtain $\alpha = \delta$ and $\beta = \epsilon + \alpha\mu$. We choose s and ω so large that (12) holds for any specified $\bar{h}$. Then, by (23),

$$\varphi(t) = \alpha + \beta t \exp_1(\mu t) = \alpha e^{\mu t} + \epsilon t \exp_1(\mu t)$$

and the result follows from Theorem 4.1. □

In particular, if we can globally bound $S^{-1}\frac{\partial F}{\partial x}(t,x)$ then we may obtain a global bound on the error of an approximate solution for all times, and this bound (30) is proportional to the residual error multiplied by an exponential term. Moreover, this term decays when the differential equation (26) satisfies the *uniform dissipation condition*

$$\sup_{t\in]0,\bar{t}],\, x\in\mathbb{R}^n} \mu\left(S^{-1}\frac{\partial F}{\partial x}(t,x)S\right) < 0. \tag{31}$$

Together with the freedom of choosing the approximate solution to high accuracy, this allows the construction of rigorous and realistic error bounds for uniformly dissipative systems.

Selection of parameters

Now that we have computable expressions for all quantities required in Theorem 4.1 we discuss the selection of the various quantities which we can choose freely.

4.6. Choice of. ω and s:

ω must be chosen such that (12) can be satisfied for large $\bar{h}$. In view of the form (23) of φ, we certainly need $\omega > \alpha$, and this is already sufficient to guarantee a positive step. For negative μ, (23) implies $\varphi(h) \le \alpha + \beta/|\mu|$; since β from (8) will typically be small (note that $[a]$ and α are residuals), $\omega = \alpha + \min(\alpha, \alpha_0)$ with a small $\alpha_0 > 0$ seems to be a good choice.

s should be chosen such that $\bar{h} \approx s$. The special form (23) of φ allows the explicit determination of the smallest zero h_0 of $\varphi(h) - \omega$:

$$h_0 := \begin{cases} +\infty & \text{for } \beta \le 0 \\ (\omega-\alpha)/\beta & \text{for } \beta > 0, \mu = 0 \\ \frac{1}{\mu}\log(1+\mu(\omega-\alpha)/\beta) & \text{for } \beta > 0, \mu \neq 0. \end{cases} \tag{32}$$

(On the computer, $+\infty$ must be replaced by a large machine number.) If $h_0 \ll s$ or $h_0 \gg s$ then s should be replaced by $\sqrt{h_0 s}$ and the calculation repeated. If h_0 is close to s (say within a factor 2), we accept s and put $\bar{h} = \min(h_0, s)$. In this way, a good step is obtained.

4.7. Choice of. C**:**
The preconditioning matrix C mainly serves to reduce $[A]$ to a diagonally dominant matrix; thus the choice $C \approx (GS(0))^{-1}$ is natural. It is sufficient to compute an approximate inverse. With a diagonally dominant $[A]$, the enclosure of the expressions $[A]^I[\cdots]$ presents no problems, e.g. with interval Gauss elimination.

4.8. Choice of. $S(h)$**:**
With our crude enclosure $[H]$ of the partial derivative of F_0, there is no point in keeping S variable. (A linear S might be useful if $[H]$ is split into $[H_0] + [H_1]h$.) Thus we take $S(h) = S$ constant. Then (22) amounts to $[B] \supseteq HS$, and since $[A] \supseteq (CG)S$ we find $[A]^I[B] \supseteq S^{-1}(CG)^{-1}HS$. This matrix determines μ in (6) and hence the magnitude of the exponential part in (23).

Since μ occurs in the exponent of the bound (23), it is essential to get a good and preferable negative bound for $\mu([A]^I[B])$; the best choice of S would diagonalize the matrix $(CG)^{-1}H$. Thus we approximately solve the linear eigenvalue problem (cf. (19))

$$H_0 x = \lambda G x \quad \text{for} \quad H_0 := \frac{\partial F_0}{\partial x}\,(t_0, p(0)) \tag{33}$$

and choose for the columns of S the real and imaginary parts of a full set of eigenvectors. If $G^{-1}H_0$ is nearly defective, one should instead choose linearly independent basis vectors from low dimensional subspaces. It is essential that S is well-conditioned (i.e. that the invariant subspaces used are "sufficiently disjoint") since otherwise the initial error α in (7) gets magnified too much. Then, with $C \approx (GS)^{-1}$, $[A] \supseteq (CG)S$, $[B] \supseteq HS$ (cf. (ii) above and (21), (22)) one forms $[A]^I[B] =: [M]$ by Krawczyk's method or interval Gauss elimination. With a "thin" $[H]$ (with zero radii) and exact calculation, one would have a thin block-diagonal matrix $[M]$ with diagonal blocks (λ) or $\begin{pmatrix} \operatorname{Re}\lambda & \operatorname{Im}\lambda \\ -\operatorname{Im}\lambda & \operatorname{Re}\lambda \end{pmatrix}$; so, in practice, $[M]_{\mathrm{sym}}$ will be nearly diagonal $\mu_2([M]) = \mu_2([M]_{\mathrm{sym}})$ can be found by (2.35) from an approximate Cholesky decomposition of $\mu I - \check{M}$ where $\check{M} := \operatorname{mid}[M]$ or, simpler, from (2.35) with $L = 0$ and $\mu_0 = \min\left(\bigcup_{i=1}^n [M]_{ii}\right)$.

4.9. Choice of. $p(h)$:
A piecewise polynomial approximation of the solution is available from a Nordsieck method, or constructible from Runge-Kutta information. Alternatively, one may interpolate the discrete approximate solution obtained by any good numerical method. Rational interpolation is advisable.

4.10. Choice of. norm:
The 2-norm is useful since it allows an elegant computation of the logarithmic norm and takes account of imaginary parts automatically. However, it leads to an overestimation factor of $\approx \sqrt{n}$ in the enclosure of an ellipsoid by a box needed to compute the bounds (18) and (19). A way out would be the use of ellipsoid arithmetic (GUDERLEY & KELLER [5], NEUMAIER [13]).

A better alternative may be the use of a mixed $(2,\infty)$-norm: If $x = (x_1,\ldots,x_k)^T$ is the partition π into blocks defined by the separable real invariant subspaces of (33), we can define

$$\|x\|_\pi := \max_{i=1(1)k} \|x_i\|_2 \tag{34}$$

and find

$$\mu_\pi(M) \le \max_{i=1(1)k} \left(\mu_2(M_{ii}) + \sum_{j\neq i} \|M_{ij}\|_2 \right) \tag{35}$$

where M is analogously partitioned into submatrices M_{ij}. Now, the ellipsoid – box transformation is needed on small blocks only, typically of size ≤ 2.

If the resulting *bound* (3) is *not good enough*, one may try to reduce $s, \bar{h}$, or to improve the approximation $p(h)$ (if (20) is large) by defect correction. Since one has an approximate eigensystem, one can do the defect correction with an explicit method on the transformed variables $y(h) := S^{-1}\,(x(t_0+h) - p(h))$.

4.11. Example. Consider the simple second order initial value problem

$$\begin{aligned} m\ddot{q} + c\dot{q} + kq &= f(t), \\ q(0) = q_0, \quad \dot{q}(0) &= \dot{q}_0, \end{aligned}$$

with $m, c, k > 0$. For $m << k$ and $c^2 << mk$ there are rapid, slowly decaying oscillations typical for a singularly perturbed problem, which forces the established enclosure methods to take tiny steps only.

Suppose that we have an approximate solution $Q(t)$ with residual

$$r(t) := f(t) - m\ddot{Q}(t) - c\dot{Q}(t) - kQ(t)$$

bounded by

$$\begin{aligned} |r(t)| &\le \varepsilon \qquad \text{for} \quad t \in [0,T]\,, \\ |q_0 - Q(0)| &\le \varepsilon_0, \quad |\dot{q}_0 - \dot{Q}(0)| \le \varepsilon_1\,. \end{aligned}$$

If we introduce

$$x = \begin{pmatrix} q \\ \dot{q} \end{pmatrix}, \quad p(t) = \begin{pmatrix} Q(t) \\ \dot{Q}(t) \end{pmatrix}, \quad p = \begin{pmatrix} q_0 \\ \dot{q}_0 \end{pmatrix}$$

we can write the system as

$$m\dot{x} = \begin{pmatrix} 0 & m \\ -k & -c \end{pmatrix} x + \begin{pmatrix} 0 \\ f(t) \end{pmatrix} =: F_0(t,x)$$

which is of the form (16) with $G = mI$. The matrix (33) becomes

$$H_0 = \begin{pmatrix} 0 & m \\ -k & -c \end{pmatrix},$$

and has eigenvectors $(1,\lambda)^T$ corresponding to the eigenvalues $\lambda = \mu \pm i\omega$, solutions of $m\lambda^2 + c\lambda + k = 0$. Here

$$\mu = -\frac{c}{2m}\,, \quad \omega = \frac{1}{m}\sqrt{mk - c^2/4}\,,$$

$$c = -2m\mu, \quad k = m(\omega^2 + \mu^2).$$

The choice recommended in 4.8 and 4.7 gives

$$S = \begin{pmatrix} 1 & 0 \\ \mu & \omega \end{pmatrix}, \quad C = (GS)^{-1} = \frac{1}{m\omega}\begin{pmatrix} \omega & 0 \\ -\mu & 1 \end{pmatrix}.$$

After some calculation, the parameters in Proposition 4.3 (for the Euclidean norm) are seen to be

$$H = CH_0 = \tfrac{1}{\omega}\begin{pmatrix} 0 & \omega \\ -\mu^2 - \omega^2 & \mu \end{pmatrix}, \quad A = I\,,$$

$$B = HS = \begin{pmatrix} \mu & \omega \\ -\omega & \mu \end{pmatrix}, \quad B_{\rm sym} = \begin{pmatrix} \mu & 0 \\ 0 & \mu \end{pmatrix},$$
$$\mu(A^I B) = \mu(B) = \lambda_{\max}(B_{\rm sym}) = \mu$$
$$CF_0 - CG\dot{p} = C(F_0 - G\dot{p}) = C\begin{pmatrix} 0 \\ r(t) \end{pmatrix} = \tfrac{1}{m\omega}\begin{pmatrix} 0 \\ r(t) \end{pmatrix},$$
$$[a] = (0, \tfrac{\varepsilon}{m\omega}[-1,1])^T.$$

The box for x is irrelevant since the problem is here linear, and therefore (unlike in nonlinear problems), the bound ω in (4) and (12) – not to be confused with the frequency ω in the present example – can be chosen arbitrary large. Thus the closure condition (12) becomes trivial, and the constraints in (7) and (8) can be satisfied with

$$\begin{aligned} \|S^{-1}(x_0 - p(0))\| &= \left\| \frac{1}{\omega} \begin{pmatrix} \omega & 0 \\ -\mu & 1 \end{pmatrix} \begin{pmatrix} q_0 - Q(0) \\ \dot{q}_0 - \dot{Q}(0) \end{pmatrix} \right\| \\ &\le \sqrt{\varepsilon_0^2 + \left(\frac{\varepsilon_1 + |\mu|\varepsilon_0}{\omega} \right)^2} =: \alpha, \\ \beta &= \|[a]\| + \alpha\mu = \frac{\varepsilon}{m\omega} + \alpha\mu, \end{aligned}$$

and we find for all $t \in [0,T]$ the bound

$$\|S^{-1}(x(t) - p(t))\| \le \varphi(t) = \alpha + \beta t \exp_1(\mu t) = \alpha e^{\mu t} + \frac{2\varepsilon}{\omega c}(1 - e^{\mu t}).$$

Since $\mu < 0$, this gives a realistic long time error of asymptotically $\frac{2\varepsilon}{\omega c}$, independent of the time interval used for the error estimation. (Of course, this illustrative linear example does not tell the full story: For sufficiently large problems, "mixing" due to strongly changing eigensystems may cause much overestimation, which is hard to avoid.)

5 Adaptive enclosure

For dissipative systems Corollary 4.5 shows that everything goes well. For non-dissipative systems one may have to apply the enclosure repeatedly over shorter time steps. In the course of several steps, one may try to *keep parameters constant* (e.g. C, S, $\check{M}$) and to adapt others in a simpler way

(e.g., $s = 1.5h_{\text{old}}$) to save work. It is important to realize that the composition of several enclosures must be done with care in order to avoid an excessive *wrapping effect* (see e.g. NEUMAIER [13] and references there). However, the theorem may be considered as a first step towards the construction of *large step methods* for the rigorous enclosure of solutions to arbitrary systems of ordinary differential equations. This would remove difficulties of the methods of EIJGENRAAM [3] and LOHNER [9], which, especially for stiff systems, are often forced to take very small step sizes.

An adaptive algorithm would roughly consist of the following steps (initially, $i = 0, \xi_0 = x_0$):

(i) Use a spectral factorization of $F_x(t_i, \xi_i)$ to find a transformation matrix S, and wrap the ellipsoid enclosing $x(t_i)$ by one of the form (7) with $x(t_i)$ in place of x_0.

(ii) Approximate the solution in $[t_i, t_i + \overline{h}]$ by a piecewise rational function,

$$x(t_i + h) \approx p(h) \quad \text{for } h \le \overline{h}.$$

(To get higher accuracy, the step size of the approximation may well be smaller than $\overline{h}$.)

(iii) enclose $\{p(h) \mid 0 \le h \le \overline{h}\}$ by an interval vector $[p]$ using a piecewise boundary value form (NEUMAIER [12], if necessary with extra subdivisions, and calculate

$$[x_j] = [p_j] + [-\omega, \omega]\nu_j,$$

where ν_j is the 2-norm of the jth row of S. This implements (18).

(iv) Calculate the remaining quantities of Proposition 4.3, and find the smallest positive solution $\overline{h}$ of the equation

$$\varphi(\overline{h}) = \omega.$$

This verifies existence of a solution with (3) and t_i in place of t_0.

(v) Set $t_{i+1} = t_i + \overline{h}$, $\xi_{i+1} = p(\overline{h})$. Replace i by $i + 1$, find a suitable value for the new $\overline{h}$ and continue with step (i).

Of course, this still leaves many details open, which will be discussed in a subsequent paper.

References

1. W.A. Coppel, Stability and Asymptotic Behaviour of Differential Equations, Boston 1965.

2. G. Dahlquist, Stability and error bounds in the numerical integration of ordinary differential equations, Trans. Royal Inst. Technology, No. 130, Stockholm 1959.

3. P. Eijgenraam, The solution of initial value problems using interval arithmetic, Math. Centre Tracts 144, Amsterdam 1981.

4. P. Hartman, Ordinary Differential Equations. Wiley, New York, 1964.

5. K.G. Guderley and C.L. Keller, A basic theorem in the computation of ellipsoidal error bounds, Numer. Math. 19 (1972), 218-229.

6. E. Hairer, G. Bader and C. Lubich, On the stability of semi-implicit methods for ordinary differential equations, BIT 22 (1982), 211-232.

7. P. Hartman, Ordinary Differential Equations. Wiley, New York, 1964.

8. R. Krawczyk and A. Neumaier, Interval slopes for rational functions and associated centered forms, SIAM J. Numer. Anal. 22 (1985), 604-616.

9. R. Lohner, Enclosing the solution of ordinary initial- and boundary-value problems, pp. 255-286 in: Computer Arithmetic (E. Kaucher et al., eds.), Teubner, Stuttgart 1987.

10. S.M. Lozinskij, Error estimate for numerical integration of ordinary differential equations, Part I, Izv. Vysš. Učebn. Zaved. Matematica 6 (1958), 52-90.

11. R. E. Moore, Interval arithmetic and automatic error analysis in digital computing. Ph. D. thesis, Appl. Math. Statist. Lab. Rep. 25, Stanford Univ., 1962.

12. A. Neumaier, Interval Methods for Systems of Equations, Cambridge Univ. Press, Cambridge 1990.

13. A. Neumaier, The wrapping effect, ellipsoid arithmetic, stability and confidence regions, pp. 175-190 in: Validation Numerics (R. Albrecht et al., eds.), Computing Suppl. 9, Springer, Wien 1993.

14. S.M. Rump, Validated solution of large linear systems, Computing Suppl. 9 (1993), in press.

15. B. Schmitt, Norm bounds for rational matrix functions, Numer. Math. 42 (1983), 379-389.

16. R.B. Schnabel and E. Eskow, A new modified Cholesky factorization, SIAM J. Sci. Stat. Comput. 11 (1990), 1136-1158.

17. T. Ström, On logarithmic norms, SIAM J. Numer. Anal. 12 (1975), 741-753.