Journal of Statistical Physics, Vol. 41, Nos. 5/6, 1985

Narrow Bounds for Numerical Integration of Differential Equations[1]

Salvatore De Gregorio[2,3]

Received January 11, 1985

Through a detailed analysis of the properties of a system of differential equations, bounds are given for the error affecting the final result of a numerical integration. These bounds appear to be narrower than those obtained with other methods. The key procedure is to consider carefully the linear part of the system and to bound it taking account of all possible errors. No very significant restriction is made on the system.

KEY WORDS: Propagation of errors; interval analysis; fundamental matrix solution; pseudotrajectory; exact solution.

1. INTRODUCTION

One of the biggest problems of numerical analysis is to find narrow bounds to the results of the numerical integration of systems of differential equations. For a single step of integration, using a step of integration sufficiently small, the different methods at hand give the possibility of having an error as small as we like, until the round-off error of the machine. But when we go ahead in the integration, the error due to the not exact knowledge of the variables propagates and grows up tendentially in an exponential way. So care has to be taken to evaluate the final error of integration. One of the most popular methods in recent time is to use interval analysis to obtain the bounds for these errors. Essentially this method substitutes numbers with intervals, defines the arithmetic operations with such intervals, and at the end gives the interval containing the exact solution. But such method works well only in very special cases (short time of integration, motion in a small region around the origin) because of the

[1] This work was partially supported by the Ministero della Pubblica Istruzione.

[2] Dipartimento di Matematica, Università di Roma "La Sapienza," 00185 Roma, Italy.
[3] Istituto di Matematica, Università de L'Aquila, 67100 L'Aquila, Italy.

0022-4715/85/1200-0865$04.50/0 © 1985 Plenum Publishing Corporation

exponential growth of the amplitude of the intervals. The fact is that the method does not take sufficient consideration of the properties of the system. And so, for example, even if the system has a strong atracting stationary solution, the interval continues to grow exponentially.

Here we present a method that considers very carefully the properties of the system and give bounds which formally are still exponentially divergent in time, but in fact, in many cases, nearly linearly divergent.

We apply the method to a very simple dynamical system; comparison with the results of interval analysis shows that the present method is much better.

The system we studied has the phase space volume contraction at constant rate typical of the Fourier truncations of Navier–Stokes equations. At the end we add some further observations about the consequence of this property for the quantities we have to control.

2. NOTATIONS AND RESULT

We refer heavily to the previous paper.[1][4] The result contained there that we exploit in the present context is Proposition I of Section 5.

We now give some notations and definitions in such a way that the meaning of the proposition can be clear and easily understood. The reader is referred to Ref. 1 for all proofs and a detailed discussion.

Let

$$\dot{x} = F(x) \tag{1}$$

be an autonomous system of ordinary differential equations in R^n:

$$x = (x_1, \ldots, x_n), \qquad F(x) = (f_1(x), \ldots, f_n(x))$$

Let $x_0(t)$ be the solution of (1) such that $x_0(t_0) = x_0$ and consider the linear differential equation

$$\dot{z} = F'(x_0(t))z \tag{2}$$

where $F'(x_0(t))$ is the matrix

$$\left(\frac{\partial f_i(x_0(t))}{\partial x_j}\right)$$

Denote by $\mathscr{L}(s, t)$ the fundamental matrix solution of (2) such that $\mathscr{L}(s, s) = E$, where E is the identity matrix; it means that the solution of (2), which at time t_0 is in Z_0, is given by

$$z(t) = \mathscr{L}(t_0, t)\, z_0, \qquad t \geqslant t_0$$

[4] Corrected proofs did not arrive in time and so in Ref. 1 x_i denotes both the i-component of x and a point of the pseudotrajectory. Here we prefer not to change notations knowing that either meaning will be clear from the context.

For any fixed T define

$$\mathscr{C}_1(T) = \sup_{0 \leqslant s \leqslant t \leqslant T} \|\mathscr{L}(s, t)\|$$

Denote then by $\{x_K, k = 0, 1, ..., N\}$ the pseudotrajectory obtained integrating (1) numerically (N is given by T/Δ where Δ is the integration step). Finally define

$$\|F\| = \sup_K |F(x_K)|$$

$$\|F^{(1)}\| = \sup_K \|F'(x_K)\|$$

$$\|F^{(2)}\| = \sup_K \|F^{(2)}(x_K)\|$$

$$\|F^{(3)}\| = \sup_x \|F^{(3)}(x)\|$$

The first two definitions are clear while the norm $\|F^{(2)}(x_K)\|$ is defined by

$$|F''(x_K)\, XY| \leqslant \|F^{(2)}(x_K)\|\, |X|\, |Y|$$

valid for every choice of vectors X and Y; we used the shortened notation

$$F''(x_K)\, XY = \left(\sum_{i,j} \frac{\partial^2 f_1(x_K)}{\partial x_i\, \partial x_j} X_i Y_j, ..., \sum_{i,j} \frac{\partial^2 f_n(x_K)}{\partial x_i\, \partial x_j} X_i Y_j \right)$$

The norm $\|F^{(3)}(x)\|$ is defined similarly by

$$|F'''(x)\, XYZ| \leqslant \|F^{(3)}(x)\|\, |X|\, |Y|\, |Z|$$

with an obvious meaning for the vector $F'''(x)\, XYZ$.

Remark. Note that the norms are evaluated along the pseudotrajectory, except $\|F^{(3)}(x)\|$ that has to be considered on the whole space or, better, in a region that certainly contains the motion. Observe also that if $F(x)$ is a polynomial of the third degree, $\|F^{(3)}(x)\|$ is constant, while, obviously, if $F(x)$ is a quadratic polynomial , $\|F^{(3)}(x)\|$ is identically zero.

In any case we are not restricted to polynomial expressions.

We can now state the following fundamental propostion.

Proposition. Denoting by α the global error in one step of integration, i.e., the sum of the round-off error and the error due to the numerical procedure of integration (we always use Taylor expansion for the sake of simplicity), if

$$N e^{\rho\gamma(T)T} \mathscr{C}_1(T) \alpha \leqslant \rho \tag{3}$$

where ρ is such that

$$2\mathscr{C}_1^2(T)\,\rho[\|F^{(2)}\| + \tfrac{2}{3}\|F^{(3)}\|\,\mathscr{C}_1(T)\rho]T = 2\beta T \leqslant 1 \tag{4}$$

and

$$\gamma(T) = 2\mathscr{C}_1^2(T)\,e^{\beta T}[\|F^{(2)}\| + \tfrac{3}{4}e^{\beta T}\,\|F^{(3)}\|\,\mathscr{C}_1(T)\rho] \tag{5}$$

then

$$|x_K - x(k\Delta)| \leqslant e^{\rho\gamma(K\Delta)K\Delta}\left[\sum_0^{K-1}{}_j \|\mathscr{L}(j\Delta, k\Delta)\|\right]\delta(k\Delta)\alpha \tag{6}$$

and

$$|x_N - x(T)| \leqslant e^{\rho\gamma(T)T}\left[\sum_0^{N-1}{}_j \|\mathscr{L}(j\Delta, N\Delta)\|\right]\delta(T)\alpha \tag{7}$$

where

$$\delta(t) = 1 + \rho\gamma(t)\,te^{\|F^{(1)}\|\Delta}$$

Remark 1. The bound (7) for the distance of the pseudotrajectory from the exact solution is formally exponential in time, as expected, but a ρ satisfying (3) can be much smaller than the value giving equality in (4), so the bound (7) in fact depends linearly on T for a large interval of values of T, if $\mathscr{C}_1(T)$ does not grow exponentially, which is often the case.

Remark 2. In a sense (7) is the best possible bound for $|x_N - x(T)|$ and in some cases can be a good improvement of simply taking $N\mathscr{C}_1(T)$ instead $\sum \|\mathscr{L}(j\Delta, N\Delta)\|\,\delta(T)$, especially if $\|\mathscr{L}(s, T)\|$ is much smaller than its maximum value in a large part of its trajectory. We did not try to optimize the expression $\exp[\rho\gamma(T)T]\,\delta(T)$ taking the right dependence on time, because we have in mind small values of ρ, in such a way that this term is in any case nearly 1. We also observe that we can in fact substitute (3) with a recursive application of (6): the Proposition is valid if

$$e^{\rho\gamma(T)T}\sup_{K\leqslant N}\left[\sum_0^{K-1}{}_j \|\mathscr{L}(j\Delta, k\Delta)\|\right]\delta(T)\alpha \leqslant \rho$$

The present formulation of the Proposition comes from the proof of Proposition 1 in Ref. 1 once we note the following:

(a) In Lemma 1 we can substitute $\sup_{0\leqslant r\leqslant t}\|\mathscr{L}(0, r)\|\,e^{\beta t}$ with

$$\|\mathscr{L}(0, t)\| + \beta\int_0^t \|\mathscr{L}(0, s)\|\,ds\cdot e^{\beta t} \leqslant \left[\|\mathscr{L}(0, t)\| + \beta\int_0^t \|\mathscr{L}(0, s)\|\,ds\right]e^{\beta t}$$

operating directly in the integral inequality (12), using $|x(t)-x_0(t)| \leqslant 2\mathscr{C}_1(T)\,|x-x_0|$ to linearize it.

(b) In Lemma 2, instead of $\sup_{0\leqslant r\leqslant t} \|\mathscr{L}(0, r)\|$, we now have a factor

$$\|\mathscr{L}(0, t)\| + \rho\gamma(t) \int_0^t \|\mathscr{L}(0, s)\|\, ds$$

and an analogous factor in its corollary.

We observe that if we take the supremum in t of the expressions we obtained, the results are worse than those in Ref. 1, but as we observed, with βT very small (in our case βT is firstly 10^{-5} and then 10^{-2}, and $\rho\gamma(T)\, Te^{\|F^{(1)}\|\Delta} \sim 2\beta T$), the present bounds are in fact better.

In Ref. 1 is also contained a discussion of the biggest numerical problem connected with the present context, the problem of the determination of a rigorous, not large bound, for $\mathscr{C}_1(T)$.

3. APPLICATIONS

We first consider a very simple nonlinear differential system just to control if the results are as we expect them. Let it be

$$\begin{aligned} \dot{x} &= xy - x + R \\ \dot{y} &= -x^2 - y \end{aligned} \tag{8}$$

For such a system the solution is always bounded: if (x_0, y_0) is the initial point, the solution is contained in the sphere of radius r,

$$r = \max[R, (x_0^2 + y_0^2)^{1/2}] \qquad \text{for} \quad R > 0$$

It can be shown that the system (8) for $R>0$ has always only one stationary solution which is stable.

We apply interval analysis for different values of the forcing R to have different sizes of the integral curves.

While it is obvious that the final interval containing the exact solution of the equation $\dot{x} = x$ grows exponentially, it is very inconvenient that the same happens applying interval analysis to differential systems with negative velocity terms. It is easy to see that, apart from less important factors, the amplitude of the interval containing the exact solution for a system like (8) grows with the number N of steps of integration as

$$\alpha_0(1+\alpha)^N(1+2\Delta l)^N \sim \alpha_0(1+2\Delta l)^N \tag{9}$$

where Δ is the step of integration, α_0 is the initial size of the interval, α is the total error (round-off and Taylor truncation) of one step of integration and l is a geometrical parameter that gives the size of the orbit. This is the contribution of the quadratic terms which are predominant if l is bigger than 1. If the coefficients were different from ± 1, they would appear as a factor in the term $2\Delta l$. We expect for example that nothing essentially changes if we stay near the stationary solution (x_0, y_0) and substitute (8) with

$$\begin{aligned} \dot{x} &= x_0 y - x + R \\ \dot{y} &= -x_0 x - y \end{aligned} \tag{10}$$

apart from the suppression of the factor 2 due to the linearization of the system and a consequent doubling of the time needed for the "explosion" of the intervals. It is to be stressed that, according to (9), if for some orbit there is the explosion of intervals at time T, this time increases very little improving α, for example from 10^{-15} to 10^{-18}. We note furthermore that also doing the integration operations in the best order to decrease the error, or applying different (reasonable) definitions of multiplication of intervals, there is no essential change.

Another obvious consequence is that if we make the transformation $x \to x/C$, $y \to y/C$, $R \to R/C$, $C > 1$, in the new variables the orbit is a factor C smaller, but a factor C appears in the equations and so nothing changes.

The results of interval analysis that we obtained with different values of R, and consequently different size of the orbits, are in good agreement qualitatively and quantitatively with these predictions. For example for $R = 10$, the orbit tends to the stationary solution $(2, -4)$; with $T = 1$ and $\Delta = 1/N$ we have

$$\left(1 + \frac{2}{N} l\right)^N \sim e^9 \sim 10^3$$

so, if $\alpha \sim 10^{-15}$, we can expect that for $T = 5$ the interval is large. We found that, starting from (5, 8), for $T = 4$ the interval amplitude is of the order of the unity. For $R = 130$ the stationary solution is $(5, -25)$, and starting still from (5, 8) the system explodes at $T \sim 1.3$–1.4. Obviously, in all these cases, the usual integration gives the stationary solution for any time, as long as we like, because the attractivity compensates for the errors of integration. Observe also that $\mathscr{C}_1(T)$ for the system (10) goes to a constant value.

Another interesting fact is that the result does not improve decreasing the step of integration and considering the same final time of integration,

but, on the contrary, there is some light worsening; this happens to be connected to the fact that

$$\left(1+\frac{2}{N}l\right)^N$$

is increasing in N.

For the sake of simplicity we are here reporting what happens in the case of a stationary solution of a simple system. In fact all these behaviors were found applying interval analysis to the periodic orbits of the Lorenz system studied in Ref. 1.

We go now to the method proposed. We apply it to the following system of differential equations:

$$\begin{aligned}\dot{y}_1 &= Sy_2\, y_3 - By_1 + R\\ \dot{y}_2 &= -Sy_1\, y_3 - By_2 + Cy_3\\ \dot{y}_3 &= -3By_3 - \tfrac{9}{25}Cy_2\end{aligned} \tag{11}$$

It comes out, for some particular choice of orography and forcing, from Fourier truncation of a partial differential equation, the quasigeostrophic equation, which describes the long-time and large-scale atmosphere dynamics.

The system is quadratic so $\|F^{(3)}\| = 0$, $\|F^{(2)}\| = \text{constant} = s$ and the conditions of the proposition simplify to

$$2\mathscr{C}_1^2(T)\rho\, \|F^{(2)}\|\, T = 2\beta T \leqslant 1, \qquad \gamma(T) = 2\mathscr{C}_1^2(T)\, e^{\beta T}\, \|F^{(2)}\|$$

The system has still only bounded solutions. The boundness of the orbit is necessary to have uniform bounds for the rest of Taylor and for the round off error.

The value we take for the parameters, suggested by their physical meaning, are

$$S = \frac{8}{9}\sqrt{3}, \qquad B = \frac{1}{48}, \qquad C = \frac{1}{16}S, \qquad R = \frac{0.1}{9}$$

We always take the initial condition: $y_{10} = y_{20} = y_{30} = 0.1$. For these values of the parameters the motion is chaotic. We take different values for Δ and in correspondence different orders of Taylor truncation.

Applying the method of interval analysis for the integration of the system (11), the interval grows rapidly and, for a value of T less than 200 it explodes.

On the other side, for $T=200$, $\mathscr{C}_1(200)$ is less than 12, after having applied the procedure described in Ref. 1 to take account of all possible errors affecting it. A value needed for these corrections is $\|F^{(1)}\|$. We have $\|F^{(1)}\| \leqslant 0.46$. For the global error of one step of integration we take $\alpha = 10^{-15}$ (we used the Univac 1100 of the University of Rome scientific computer center: with double precision its round-off error is of the order of 0.2×10^{-17}). The value ρ of the proposition is then

$$\rho \leqslant \rho_1 = \frac{1}{2T\mathscr{C}_1^2(T)\,\|F^{(2)}\|} \sim 10^{-5}$$

With $\Delta = 10^{-2}$ we have

$$N\mathscr{C}_1(200)\alpha = 2.4 \times 10^{-10} = \rho$$

so the condition $\rho \ll \rho_1$ is satisfied and $\exp[\rho\gamma(T)T]$ is nearly 1. Evaluating the sum $\sum_{0\,K}^{N-1} \|\mathscr{L}(k\Delta, N\Delta)\|$ every ten steps in k [but the matrix $\mathscr{L}(k\Delta, l\Delta)$ is evaluated step by step], we obtain

$$\sum_{0\,K}^{N/P} \|\mathscr{L}(kP\Delta, N\Delta)\| \leqslant 10^4, \qquad P = 10$$

and the final error is

$$e^{\rho\gamma(T)T}\left(\sum \|\mathscr{L}(k\Delta, N\Delta)\|\right) e^{\|F^{(1)}\|\Delta_1}\,\delta(T)\alpha \sim 10 \times 10^4 e^{0.46 \times 0.1} 10^{-15}$$

$$\leqslant \sim 10^{-10}$$

The values found for the orbit parameters doing the integration with different steps Δ (and consequently different orders of Taylor truncation in order to have the error less than the computer round-off error) are in good agreement with this result. The difference between the different values is of the order of about 10^{-12}.

But we went ahead in the integration to $T=1000$. The value found for $\|F^{(1)}\|$ is now $\|F^{(1)}\| \leqslant 0.57$, while $\mathscr{C}_1(1000) \leqslant 10^2$ to which it corresponds

$$\rho_1 = 3 \times 10^{-8}$$

To have $\rho \ll \rho_1$ we need now to refer to the form of the proposition contained in Remark 2. It is

$$\sup_{K \leqslant N}\left(\sum_{0\,j}^{K-1} \|\mathscr{L}(j\Delta, k\Delta)\|\right) \leqslant 10N$$

and, from the value found for the sum

$$\sum_{0}^{N-1}{}_{j} \|\mathscr{L}(j\Delta, N\Delta)\| \leqslant 2.5N$$

it follows that the final error of integration is bounded by

$$|x_N - x(1000)| \leqslant 5 \times 10^{-10}$$

The values found for orbit are well within this bound, the difference being of the order of 10^{-11}. So while interval analysis cannot give any information for a time of integration longer than 200, with our method we can have sharp conclusions for a much longer time.

We observe that, starting from an error of 10^{-15} in a single step of integration, we can hardly hope to obtain a bound better than that reported, after 2×10^5 steps of integration, the final error being nearly twice the value we obtain simply with the linear propagation of the error.

We note that in principle, for a fixed time of integration T, better results can be obtained with a longer step of integration and higher order of Taylor, but obviously there is a limit in this procedure, because, with a reasonable number of terms in each step of integration, we have to reach the computer precision, and in any case the step of integration has to be such that the corrections to $\mathscr{C}_1(T)$ are not too large.

Remark. It is to be stressed that physical consideration suggested that small forcing which, by the way, produces a chaotic motion around the origin. The smallness of the orbit is the fundamental reason that allows to consider so long times of integration. We have to observe in any case that the limitations to the present method come from the condition

$$2\mathscr{C}_1^2(T)\rho \, \|F^{(2)}\| \, T \leqslant 1$$

and from the value of $\|F^{(1)}\|$, which controls the correction needed for $\mathscr{C}_1(T)$; smaller values of $\|F^{(1)}\|$ give better corrections and so allow a longer time of integration. But if we take a long T, not only can $\mathscr{C}_1(T)$ be bigger, but we have to take also ρ ($\rho \leqslant \rho_1$), bigger, and so at the end we cannot always have a result as good as that reported.

Having in mind a control of the way in which $\mathscr{C}_1(T)$ and $\sum \|\mathscr{L}(j\Delta, N\Delta)\|$ grow with T, we found, before applying the corrections,

$$\mathscr{C}_1(100) \leqslant 5.63, \qquad \mathscr{C}_1(200) \leqslant 10.3, \qquad \mathscr{C}_1(300) \leqslant 71.1$$

$$\mathscr{C}_1(400) \leqslant 71.1 \ldots \qquad \mathscr{C}_1(1000) \leqslant 71.1$$

$$\sum \|\mathscr{L}(j\Delta, 100)\| \leqslant 3.4\frac{100}{\Delta}, \qquad \sum \|\mathscr{L}(j\Delta, 200)\| \leqslant 4.3\frac{200}{\Delta}$$

$$\sum \|\mathscr{L}(j\Delta, 300)\| \leqslant 27\frac{300}{\Delta}, \qquad \sum \|\mathscr{L}(j\Delta, 1000)\| \leqslant 1.9\,N$$

so we see that $\mathscr{C}_1(T)$, after $T=300$, does not grow any more with T. Moreover the sum $\sum \|\mathscr{L}(j\Delta, N\Delta)\|$ is not necessarily increasing with time, even if we have to sum more terms: for $T=300$ it takes a value four times bigger than the corresponding value for $T=1000$.

4. SOME FURTHER CONSIDERATIONS

We discussed in Ref. 1 the errors affecting $\mathscr{C}_1(T)$; for example, the variation of time with step Δ_1 in making sup $\|\mathscr{L}(s, t)\|$ gives a factor $e^{\|F^{(1)}\|\Delta_1}$. But in systems like (11) we have a very important indirect control of the value found for $\mathscr{C}_1(T)$. Considering the quadratic system

$$\dot{x}_i = \sum_{jl} a_{ijl}x_j x_l - \sum_j b_{ij}x_j + c_i = f_i(x)$$

if $a_{iil} = a_{ili} = 0$ and $\operatorname{div} \dot{x} = -\sum_i b_{ii}$, we have not only the contraction of volumes of its phase space at constant rate, but the same identical contraction occurs in the linear system

$$\dot{z} = F'(x_0(t))z \tag{12}$$

This contraction for both systems is given by the Jacobian

$$J(t) = e^{\operatorname{div} \dot{x}t}$$

and the Jacobian of the system (12) is just the determinant of the matrix $\mathscr{L}(0, t)$.

So we know exactly which would be the value of the determinant of the matrix $\mathscr{L}(0, t)$ as a function of t, for any trajectory $x(t)$, and this gives an indirect exact control of the values we found for $\|\mathscr{L}(0, t)\|$. Obviously the norm of $\mathscr{L}(0, t)$, i.e., the square root of the maximum eigenvalue of $\mathscr{L}(0, t)\cdot\mathscr{L}(0, t)^*$, tells us nothing about volume contraction, but instead says how far are evolving points in a neighbourhood of the reference solution $x_0(t)$. In any case, if the determinant has exactly the value expected, we can have great confidence about the eigenvalues. In our case $\operatorname{div} \dot{x} = -5B = -5/48$. We compare $\exp(-5Bt)$ with the Jacobian J of the

flux defined by (11) and with the determinant of the matrix $\mathscr{L}(0, t)$. The values found for $T = 10$, 100, 200 are the following:

	$T = 10$	$T = 100$	$T = 200$
$J(t)$	0.124 514 471 447 0	$0.299\,294\,783\,321 \times 10^{-4}$	$0.895\,769\,1 \times 10^{-9}$
$\det \mathscr{L}(0, t)$	0.124 514 471 443 9	$0.299\,294\,783\,071 \times 10^{-4}$	$0.895\,773\,7 \times 10^{-9}$
$\exp[-(5/48)t]$	0.124 514 471 444 1	$0.299\,294\,783\,076 \times 10^{-4}$	$0.895\,773\,6 \times 10^{-9}$

For $\mathscr{L}(0, t)$ we used a third order expansion (see Ref. 1). The Jacobian J is evaluated instead with a completely different method, determining step by step the matrix of the transformation defined by the flux. In principle J had to be more precise than $\det \mathscr{L}(0, t)$ because the flux contains terms of higher order, but to evaluate J we have to make differences of numbers nearly equal, so we lose information. This is also the reason why the values of J after $T = 280$ become unstable and cease to be significant. Another reason that makes both J and $\det \mathscr{L}(0, t)$ unstable for long times is that the errors of the computer become now relatively large. The matrix of the flux and the matrix $\mathscr{L}(0, t)$ consist of elements of very different order of magnitude, some of the order of the unity and some of the order of the determinant of the matrix, so, as expected, when the value of the determinant has reached the precision of the computer, all further computations have no more meaning. For $T = 350$ we found $\det \mathscr{L}(0,350) = 0.12 \times 10^{-15}$, while $\exp[-(5/48)\,350] = 0.14 \times 10^{-15}$ and from then on the values found for $\det \mathscr{L}(0, t)$are random fluctuations around this value, a value that is in a good agreement with our statement about the computer error α.

We also note that to evaluate $\sum \|\mathscr{L}(j\Delta, N\Delta)\|$ we have to integrate until $t = T$, saving in the computer memory the values of the matrices at regular intervals of time. So it is true that the operation increases linearly with N, but we need the computer memory and cannot evaluate it for every j, if we have to do, for example, as we did, 10^5 steps of integration or more. So a good way of avoiding this difficulty would be to consider the relation

$$\mathscr{L}(t, T) = \mathscr{L}(0, T) \cdot \mathscr{L}(0, t)^{-1} \tag{13}$$

which would give

$$\sum \mathscr{L}(j\Delta, N\Delta) = \mathscr{L}(0, T) \sum \mathscr{L}(0, j\Delta)^{-1} \tag{14}$$

We could now evaluate the sum for every j without using the computer memory. Unfortunately in our case the relation (13) is not useful because the product $\|\mathscr{L}(0, T)\| \cdot \|\mathscr{L}(0, t)^{-1}\|$ can be much bigger, for example by a factor 10^3, than $\|\mathscr{L}(t, T)\|$. The explanation of this surprising behavior is very simple and depends on the fact, seen above, that $\det \mathscr{L}(0, t)$ goes

exponentially to zero in t. As a consequence, by increasing t, some eigenvalue of $\mathscr{L}(0, t) \cdot \mathscr{L}(0, t)^*$ becomes smaller and smaller, of the order of the value of $[\det \mathscr{L}(0, t)]^2$, and so the norm $\|\mathscr{L}(0, t)^{-1}\|$, given by the square root of the inverse of the smallest eigenvalue, becomes bigger and bigger.

5. CONCLUSIONS

We say that, in general, we cannot hope to have good results using general techniques that do not consider in some way the properties of the particular system at hand. The interval analysis gives nearly the best possible result for one step of integration, or for very short time of integration, but when the interval grows, it still considers the biggest possible error for the system in that interval, and so gives amplitudes growing rapidly. For example the attractivity of a stationary solution does not play any role. In this way only in very special cases we obtain good results.

The method proposed considers very carefully the linear part of the system, its fundamental matrix $\mathscr{L}(s, t)$ is evaluated with only the errors of the computer and of the substitution of the pseudotrajectory to the exact solution. Only after having considered the linear part, we grossly evaluate the higher terms, but now we already extracted the biggest part of information from the system and so the results at the end are still good.

It has to be stressed that all depends on the possibility, as shown in Ref. 1, of evaluating a true not exploding bound for the norms $\|\mathscr{L}(s, t)\|$. In any case the computation of $\mathscr{C}_1(T)$, for a fixed value of $\|F^{(1)}\|$, takes a computer time growing quadratically with T, as can be easily understood, once we have fixed the maximum value of the correction due to the factor $e^{\|F^{(1)}\|\Delta_1}$, Δ_1 indicating the time interval used for evaluate $\mathscr{C}_1(T)$. But if ρ can be taken much smaller than the value giving equality in (4), it is enough to have a rough evaluation of $\mathscr{C}_1(T)$; what is more important is the sum $\sum \|\mathscr{L}(j\Delta, N\Delta)\|$, and this is linear in T, even if, as we explained, in some cases we still cannot evaluate it for every j.

REFERENCES

1. S. De Gregorio, The study of periodic orbits of dynamical systems. The use of a computer, *J. Stat. Phys.* **38**:947 (1985).
2. E. N. Lorenz, Attractor sets and quasi-geostrophic equilibrium, *J. Atmos. Sci.* **37**:1685 (1980).
3. S. De Gregorio and G. A. Dalu, Barotropic (low-order) spectral model flow in β-plane, *Nuovo Cimento* **7C**:179 (1984).