Journal of Statistical Physics, Vol. 38, Nos. 5/6, 1985

The Study of Periodic Orbits of Dynamical Systems. The Use of a Computer

S. De Gregorio[1]

Received June 19, 1984

We examine the role of a computer to prove the existence of periodic orbits of nonquadratic dynamical systems on the lines of the work of Vul and Sinai [1] for quadratic systems. We show that, in principle, the work can be successful. Results, simpler in the case of quadratic systems, are applied to the well studied Lorenz model.

KEY WORDS: Periodic orbits; Poincaré map; Newton method; integration errors; round-off errors; fundamental matrix solution.

1. INTRODUCTION

A previous paper[2] considered the problem of the role of a computer to prove the existence of periodic orbits in the systems of quadratic differential equations, a problem already studied by Vul and Sinai.[1]

The relevance of the study of periodic orbits of dynamical systems can be deduced from the works of many authors, for example Ref. 3. In this work we generalize the study to nonquadratic systems and, perhaps even more interesting, we obtain bounds that are more effective also in the case of quadratic systems. We recall that the problem we face is how to use "*a priori*" bounds and evaluation by means of a computer to prove the existence of a "true" periodic orbit in some neighborhood of a numerical periodic orbit. We show that, with some hypotheses on the system of differential equations and with a convenient choice of a numerical procedure, the use of a computer is successful: in principle, periodic orbits can be "exactly" detected by a computer if the computer is sufficiently precise.

[1] Dipartimento di Matematica, Università di Roma "La Sapienza," 00185 Roma, Italy, and Dipartimento di Matematica, Università de L'Aquila, 67100 l'Aquila, Italy.

There are many numerical studies of quadratic dynamical systems.[4] Their relevance derives, for example, from the fact that truncated Euler or Navier–Stokes equations are systems of quadratic ordinary differential equations. We note that these systems have some very interesting behavior. Consider the system

$$\dot{x}_i = \sum_{k,l} a_{ikl} x_k x_l - \sum_k b_{ik} x_k + c_i, \qquad i = 1,\ldots, n$$

where

$$\sum_{i,k} b_{ik} x_i x_k > 0 \tag{1}$$

for $x = (x_1, x_2,\ldots, x_n) \neq 0$.

If we denote by α the maximun of $\sum_{i,k,l} a_{ikl} x_i x_k x_l$ on the unit sphere, by β the minimun of (2) and by γ the maximun of $\sum_i c_i x_i$, we have $\alpha,\ \gamma \geqslant 0$ because they refer to odd functions, and $\beta > 0$ by definition. It then follows that

$$\frac{d}{dt}\frac{1}{2}\sum_i x_i^2 = \sum_{i,k,l} a_{ikl} x_i x_k x_l - \sum_{i,k} b_{ik} x_i x_k + \sum_i c_i x_i$$
$$\leqslant \alpha R^3 - \beta R^2 + \gamma R, \qquad R = \left[\sum_i x_i^2\right]^{1/2}$$

so

$$\frac{dR}{dt} \leqslant \alpha R^2 - \beta R + \gamma \tag{2}$$

From (2) we have that

$$\frac{dR}{dt} < 0 \qquad \text{if} \quad R_1 < R < R_2$$

where

$$R_{1,2} = \frac{\beta \pm \sqrt{(\beta^2 - 4\alpha\gamma)^{1/2}}}{2\alpha} \tag{3}$$

and R_1, R_2 are real if $\beta^2 > 4\alpha\gamma$.

This means that if the flow $x(t)$ at some time gets into the sphere of radius $R_2 - \varepsilon$, $\varepsilon > 0$, it cannot escape from it any more and gets into the sphere of radius $R_1 + \varepsilon$. If $\alpha = 0$, then dR/dt is less than zero if $R > \gamma/\beta$: the flow always comes into the sphere of radius $R = \gamma/\beta + \varepsilon$.

When the system is of the kind we are discussing, this *a priori* knowledge of the region of the flow (in the first case only for the flows starting inside the outer sphere) can be used, if useful, in what follows.

We also note, incidently, that if

$$a_{iki} = a_{iil} = 0 \qquad \forall i, k, l$$

then

$$\operatorname{div} v = \sum_i \frac{\partial \dot{x}_i}{\partial x_i} = -\sum_i b_{ii} < 0$$

and the volume in the phase space decreases at an exponential constant rate

$$v(t) = v(0) \exp(-\operatorname{div} vt)$$

and the bounded component of the attractor can evidently have only Lebesgue measure zero. But nevertheless it can be "strange," that is, very complicated.

The work is obviously a balanced connection of analytical and numerical results.

In the first part we pose the notations, recall the result of Vul and Sinai, and prove the main properties of the flow and of the Poincaré map we need later. The rest of the paper is devoted to the analysis of errors of integration (of the computer and of the numerical procedure), and to the way of controlling them. Finally we give two applications to the Lorenz model.

2. NOTATIONS AND MAIN RESULT (THEOREM 1)

Let

$$\dot{x} = F(x) \tag{4}$$

be an autonomous system ordinary differential equations in R^n, with

$$x = (x_1, x_2, \ldots, x_n), \qquad F = (f_1, f_2, \ldots, f_n)$$

We suppose that $F(x)$ is regular, at least C^3, in the region we are interested in, and consider the following two cases:

(a) $F(x)$ is quadratic.

(b) The third derivatives of $F(x)$ are uniformly bounded in the region containing the flow.

Remark. The condition (b) is not so bad for the definition of boundedness we will give: for example a cubic term is uniformly bounded in the whole space (see next definition of $\|F^3\|$).

We will see that in some cases it will be sufficient also for a rough bound of the third derivatives to have useful conclusions.

Other simplifying conditions could also be considered, such as the Lipschitz condition on the first derivatives.

The solutions of (4) determine a flow in R^n.

We denote by S_t the corresponding one parameter semigroup. So, as usual, the solution that for $t=0$ is in x is denoted by $S_t x$, i.e.,

$$S_t x = x(t) \qquad \text{if} \quad x(0) = x$$

Given the trajectory

$$\gamma = \{S_t x_0, 9 \leqslant t \leqslant T\}$$

for an arbitrary fixed $T > 0$ we consider also the linear system associated to (4) along γ, that is, the system

$$\dot{z} = F'(S_t x_0) z \tag{5}$$

where F' is the matrix of derivatives $\partial f_i / \partial x_j$ evaluated along γ.

Denote by $\mathscr{L}(s, t)$ the fundamental matrix solution of (5) such that $\mathscr{L}(s, s) = E$, with E the identity matrix. Given the hypersurface Γ, $x_d = \text{const} = c$, we consider a trajectory that starts in a point x_0 of the plane and, after a time T, crosses again the plane in the same direction in a point x' near the initial point. x' is by definition the Poincaré map of x_0 and we denote it by $x' = P(x_0)$. Obviously we suppose that the vector field x is not tangential to the plane in a neighborhood of x_0, that is, $f_d(x) \neq 0$, so the Poincaré map is defined for every point of this neighborhood.

Denote by $\mathscr{U}_\rho(x_0)$ the cylindrical neighborhood of x_0

$$\mathscr{U}_\rho(x_0) = \left\{ x \colon \sum_{i \neq d}^{n} |x_i - x_{0i}|^2 < \rho^2;\ |x_d - c| < \rho \right\}$$

by $\mathscr{W}_\rho(\gamma)$ the ρ neighborhood of γ and by $\{x_k, k = 0, 1, ..., N\}$ the pseudotrajectory obtained by numerical integration of (4) starting in x_0.

Our problem is to find fixed points of the Poincaré map [and so a periodic trajectory for the flow defined by (4)]. We will see that the most important procedure to be successful is to separate the linear part of the Poincaré map from the nonlinear part.

Introduced the Euclidean norm $|x|_A$ for vectors x, we note that the norm $\|A\|$ of a matrix is then defined by the square root of the maximum eigenvalue of the matrix $A \cdot A^*$.

We need to estimate the following parameters:

$$c_1(t) = \sup_{0 \leqslant \tau \leqslant t} \|\mathscr{L}(0, \tau)\|, \qquad c_1 = \sup_{0 \leqslant s \leqslant t \leqslant T} \|\mathscr{L}(s, t)\|$$

$$\|F^i\| = \sup_k \|F^i(x_k)\|, \qquad i = 0, 1, 2 \tag{6}$$

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

where $\{x_k\}$ is the pseudotrajectory and $\|F^i(x_k)\|$ is the norm of tensors obtained deriving i times the vector F, the case $i = 0$ corresponding to the vector F itself. More precisely,

$$\|F\| = \|F^0\| = \sup_k |F(x_k)|, \qquad \|F^1\| = \sup_k \|F'(x_k)\|$$

and $\|F^i(x)\|$, $i = 2, 3$, are defined by

$$\left| \sum_{i,j} F_{ij}(x_k)\, X_i Y_j \right| \leqslant \|F^2(x_k)\|\ |X|\ |Y|$$

where

$$F_{ij}(x_k) = \left[\frac{\partial^2 f_1(x_k)}{\partial x_i\, \partial x_j}, \ldots, \frac{\partial^2 f_n(x_k)}{\partial x_i\, \partial x_j} \right], \qquad X = (X_1, \ldots, X_n), \ldots,$$

and

$$\left| \sum_{i,j,r} F_{ijr}(x)\, X_i Y_j Z_r \right| \leqslant \|F^3(x)\|\ |X|\ |Y|\ |Z|$$

where

$$F_{ijr}(x) = \left[\frac{\partial^3 f_1(x)}{\partial x_i\, \partial x_j\, \partial x_r}, \ldots, \frac{\partial^3 f_n(x)}{\partial x_i\, \partial x_j\, \partial x_r} \right]$$

We need also

$$c_3 = \inf_{x \in \mathscr{U}_{\rho_1}(x_0)} |f_d(x)|$$

$$c_4 = \sup_{x \in \mathscr{U}_{\rho_1}(x_0)} \left[\sum_{i \neq d} f_i^2(x) \right]^{1/2} = \sup_x |F_{\hat{d}}(x)|$$

and

$$c_5 = \sup_{x \in \mathscr{U}_{\rho_1}(x_0)} \|F'(x)\|$$

where F_{∂} denotes the vector $F_{\partial} = (f_1, ..., f_{d-1}, f_{d+1}, ..., f_n)$. It is convenient to use $|F_{\partial}|$ instead of $|F|$ only if $|f_d| \gg |f_i|$, $i \neq d$. ρ_1 has to satisfy the condition: $2c_1\rho_0(1 + c_4/c_3) + \varepsilon \leqslant \rho_1$ with the value of ρ_0 given in Theorem 2 and ε defined in Theorem 1 (see the end of the proof of Lemma 3).

The previous hypothesis (b) means that $\|F^3\|$ is bounded. We now recall, for completeness, the fundamental theorem given in Ref. 1. Let the Poincaré map P be defined in the neighborhood $\mathscr{U}_\rho(x_0) \cap \Gamma$ and suppose that $x' = P(x_0)$ is still in $\mathscr{U}_\rho(x_0) \cap \Gamma$. Develop the Poincaré map around x_0 for $x \in \mathscr{U}_\rho(x_0) \cap \Gamma$ in the following way:

$$P(x) = x' + L(x - x_0) + Q(x, x_0)$$

so L is a linear matrix and $Q(x, x_0)$ contains the nonlinear terms of the map. Then we have the following:

Theorem 1 (Ref. 1). Suppose (i) there exist constants ρ_0 and K_0 such that

$$|Q(x, x_0) - Q(y, x_0)| \leqslant K_0 \rho \, |x - y| \tag{7}$$

for every $\rho \leqslant \rho_0$ and $|x - x_0| \leqslant \rho$, $|y - x^0| \leqslant \rho$.

(ii) For $\varepsilon = |x' - x_0|$ and for some $\bar{\rho} \leqslant \rho_0$ let the following inequality be satisfied:

$$\|(L - E)^{-1}\| \left(\frac{\varepsilon}{\bar{\rho}} + K_0 \bar{\rho}\right) \leqslant 1 \tag{8}$$

then in a $\bar{\rho}$ neighborhood of x_0 there exists a unique fixed point of the Poincaré map.

We give here, for completeness, a short proof. If x^* is a fixed point of P, it is

$$x^* = x' + L(x^* - x_0) + Q(x^*, x_0)$$

from which

$$x^* - x_0 = -(L - E)^{-1}[x' - x_0 + Q(x^*, x_0)]$$

So it is natural to check if the inductive sequence

$$x^n - x_0 = -(L - E)^{-1}[x' - x_0 + Q(x^{n-1}, x_0)]$$

will converge to x^*. First of all, if $x^{n-1} \in \mathscr{U}_{\bar{\rho}}(x_0) \cap \Gamma_1$ also x_n stays in the same neighborhood

$$|x^n - x_0| \leqslant \|(L-E)^{-1}\| \; (\varepsilon + K_0 \bar{\rho}^2) \leqslant \bar{\rho}$$

Furthermore the sequence is contractive

$$\begin{aligned} |x^{n+1} - x^n| &\leqslant \|(L-E)^{-1}\| \; |(Q(x^n, x_0) - Q(x^{n-1}, x_0))| \\ &\leqslant \|(L-E)^{-1}\| \; K_0 \bar{\rho} \; |x^n - x^{n-1}| < |x^n - x^{n-1}| \end{aligned}$$

and so there exists a unique fixed point.

Remark. Even if $\|(L-E)^{-1}\|$ and K_0 are relatively large, the condition (8) of the Theorem can be satisfied if ε is small enough. Denoting by $\bar{x}$ the numerical image of x_0 on Γ, in the relation

$$\varepsilon = |x' - x_0| \leqslant |x' - \bar{x}| + |\bar{x} - x_0|$$

the first term on the right contains only errors of the integration procedure and round off errors, and can be reduced conveniently using a higher order of integration and a more precise computer, or using multiple or infinite precision programs. The second term on the contrary tells us how near we are numerically to a fixed point of the Poincaré map: if x_0 is a true fixed point also this term can be made conveniently small. So, if the system has periodic orbits, we can find them by means of the Newton method, applicable even if the orbit is not attractive, and can, in principle at least, if $\|(L-E)^{-1}\|$ is not very big or infinite, satisfy the condition (8) of the theorem. What remains is the condition (7), but we are able to prove the following theorem.

Theorem 2. Condition (7) of Theorem 1 is satisfied with

$$K_0 \simeq \left(1 + \frac{c_4}{c_3}\right) T\gamma(T) \, c_1(T) \, e^{\rho_0 \gamma(T) T}$$

where ρ_0 is such that

$$\beta = c_1^2 \rho_0 (\|F^2\| + \tfrac{2}{3} \|F^3\| \; c_1 \rho_0) \leqslant 1/2T$$

and

$$\gamma(t) = 2c_1^2 e^{\beta t} (\|F^2\| + \tfrac{3}{4} e^{\beta t} \; \|F^3\| \; c_1 \rho_0)$$

The exact value of K_0 is at the end of Remark (3) of Lemma 2.

Remark. We are not saying that we can certainly find all periodic orbits by computer: $\|(L-E)^{-1}\|$ and K_0 can be very big and, despite of the fact that the Newton method is applicable also to nonattractive orbits, we have to start near them to have convergence. In any case, a lot of work is necessary.

3. PROPERTIES OF THE FLOW

We start with some useful lemmas. We suppose that $\|F^3\|$ is bounded (if instead the system is quadratic, in all what follows $\|F^3\|$ has to be considered zero).

Lemma 1. If $|x - x_0| \leqslant \rho$ where ρ is such that

$$\beta = c_1^2 \rho(\|F^2\| + \tfrac{2}{3}\|F^3\| \, c_1 \rho) \leqslant 1/2T \tag{9}$$

then

(i) $$|x(t) - x_0(t)| < e^{\beta t} c_1(t) \, |x - x_0| \qquad \text{for} \quad t \leqslant T$$

and

(ii) $$\begin{aligned} &|x(t) - x_0(t) - \mathscr{L}(0, t)(x - x_0)| \\ &\quad \leqslant \tfrac{1}{2} c_1^3 e^{2\beta t} t(\|F^2\| + \tfrac{1}{3} e^{\beta t} \, \|F^3\| \, c_1 \rho) \, |x - x_0|^2 \end{aligned}$$

Remark. We see that, if ρ is small, it is not necessary to have a narrow bound for $\|F^3\|$ to satisfy (9). In our applications we find for ρ values of the order of 10^{-6}, 10^{-7}.

Proof. (i) Expanding $F(x(t))$ we have

$$\begin{aligned} \dot{x}(t) - \dot{x}_0(t) = F'(x_0(t))[x(t) - x_0(t)] + \tfrac{1}{2} F''(x_0(t))[x(t) - x_0(t)]^2 \\ + \tfrac{1}{6} F'''(\xi(t))[x(t) - x_0(t)]^3 \end{aligned} \tag{10}$$

where the points $\xi(t)$ depend on the components. The notation has been simplified; for example $F''x^2$ stays for

$$F''x^2 = \left(\sum_{j,l} \frac{\partial f_1}{\partial x_j \, \partial x_l} x_j x_l, \sum_{j,l} \frac{\partial f_2}{\partial x_j \, \partial x_l} x_j x_l, \ldots, \sum_{j,l} \frac{\partial f_n}{\partial x_j \, \partial x_l} x_j x_l \right)$$

The solution of (10) satisfies the integral equation

$$\begin{aligned} x(t) - x_0(t) = \mathscr{L}(0, t)(x - x_0) + \tfrac{1}{2} \int_0^t \mathscr{L}(s, t) \, F''(x(s))[x(s) - x_0(s)]^2 \, ds \\ + \tfrac{1}{6} \int_0^t \mathscr{L}(s, t) \, F'''(\xi(s))[x(s) - x_0(s)]^3 \, ds \end{aligned} \tag{11}$$

from which have

$$|x(t)-x_0(t)| \leqslant c_1(t)\,|x-x_0| + \tfrac{1}{2}c_1\,\|F^2\| \int_0^t |x(s)-x_0(s)|^2\,ds$$
$$+\tfrac{1}{6}c_1\,\|F^3\| \int_0^t |x(s)-x_0(s)|^3\,ds \tag{12}$$

Consider the differential equation

$$\dot{z}(t) = \dot{c}_1(t)\,z(0) + \tfrac{1}{2}c_1\,\|F^2\|\,z^2 + \tfrac{1}{6}c_1\,\|F^3\|\,z^3, \qquad z(t) = |x(t)-x_0(t)| \tag{13}$$

If $z(t) < \alpha z(0)$ for $t \leqslant T$, where $\alpha > 1$ is to be specified later, it is

$$\dot{z}(t) < \dot{c}_1(t)\,z(0) + [\tfrac{1}{2}c_1\,\|F^2\|\,\alpha z(0) + \tfrac{1}{6}\,\|F^3\|\,\alpha^2 z^2(0)]z$$

so

$$z(t) \leqslant c_1(t)\,z(0)\,e^{bt}$$

where

$$b = \tfrac{1}{2}c_1 \alpha z(0)[\|F^2\| + \tfrac{1}{3}\|F^3\|\,\alpha z(0)]$$

If

$$bT \leqslant \tfrac{1}{2}$$

or

$$c_1 \alpha z(0)[\|F^2\| + \tfrac{1}{3}\|F^3\|\,\alpha z(0)]T \leqslant 1 \tag{14}$$

it follows that

$$z(t) \leqslant e^{1/2} c_1(T)\,z(0) \qquad \text{for} \quad t \leqslant T$$

so the procedure is consistent if $\alpha \geqslant e^{1/2}c_1(T)$, for example, $\alpha = 2c_1$. The condition (14) is always satisfied for $z(0) = |x-x_0| \leqslant \rho$ if

$$2c_1^2\rho(\|F^2\| + \tfrac{2}{3}\|F^3\|\,c_1\rho)T \leqslant 1 \tag{15}$$

so (i) follows with the given value of β.

(ii) From (11) it comes

$$
\begin{aligned}
&|x(t)-x_0(t)-\mathscr{L}(0,t)(x-x_0)| \\
&\quad\leqslant \tfrac{1}{2}c_1\,\|F^2\|\int_0^t |x(s)-x_0(s)|^2\,ds+\tfrac{1}{6}c_1\,\|F^3\|\int_0^t |x(s)-x_0(s)|^3\,ds \\
&\quad\leqslant \tfrac{1}{2}c_1^3\,\|F^2\|\,e^{2\beta t}t\,|x-x_0|^2+\tfrac{1}{6}c_1^4\,\|F^3\|\,e^{3\beta t}t\,|x-x_0|^3 \\
&\quad\leqslant \tfrac{1}{2}c_1^3e^{2\beta t}t(\|F^2\|+\tfrac{1}{3}e^{\beta t}\,\|F^3\|\,c_1\rho)\,|x-x_0|^2
\end{aligned}
$$

The interesting fact is that relations like those of Lemma 1 are valid also for any two points in a ρ neighborhood of x_0. ▮

Lemma 2. If $|x_1-x_0|\leqslant\rho$ and $|x_2-x_0|\leqslant\rho$ with ρ satisfying (15), then, for $t\leqslant T$

$$\text{(i)}\quad |x_2(t)-x_1(t)|\leqslant e^{\rho\gamma(t)t}c_1(t)\,|x_2-x_1| \tag{16}$$

and

$$\text{(ii)}\quad |x_2(t)-x_1(t)-\mathscr{L}(0,t)(x_2-x_1)|\leqslant k\rho\,|x_2-x_1| \tag{17}$$

where $\gamma(t)$ is the monotone increasing function

$$\gamma(t)=2c_1^2e^{\beta t}(\|F^2\|+\tfrac{3}{4}e^{\beta t}\,\|F^3\|\,c_1\rho)$$

and

$$k=T\gamma(T)\,e^{\rho T\gamma(T)}c_1(T)$$

Proof. (i) We have

$$
\begin{aligned}
&\dot{x}_2(t)-\dot{x}_1(t) \\
&\quad=F'(x_1(t))[x_2(t)-x_1(t)]+\tfrac{1}{2}F''(\xi(t))[x_2(t)-x_1(t)]^2 \\
&\quad=F'(x_0(t))[x_2(t)-x_1(t)]+F''(x_0(t))[x_1(t)-x_0(t)][x_2(t)-x_1(t)] \\
&\qquad+\tfrac{1}{2}F'''(\xi_1(t))[x_1(t)-x_0(t)]^2[x_2(t)-x_1(t)] \\
&\qquad+\tfrac{1}{2}F''(x_0(t))[x_2(t)-x_1(t)]^2+\tfrac{1}{2}F'''(\xi_2(t))[\xi(t)-x_0(t)](x_2-x_1)^2
\end{aligned} \tag{18}
$$

with the same observation about the points $\xi(t)$ as before. For later use we simply denote $H(x_1(t),x_2(t))$ the nonlinear terms on the right side of (18), so

$$x_2(t)-x_1(t)=\mathscr{L}(0,t)(x_2-x_1)+\int_0^t \mathscr{L}(s,t)\,H(x_1(s),x_2(s))\,ds \tag{19}$$

From (18) it follows that

$$
\begin{aligned}
&|x_2(t)-x_1(t)| \\
&\quad \leqslant c_1(t)\,|x_2-x_1| + c_1^2\,\|F^2\|\,e^{\beta t}\rho\int_0^t |x_2(s)-x_1(s)|\,ds \\
&\qquad + \tfrac{1}{2}c_1^3\,\|F^3\|\,e^{2\beta t}\rho^2\int_0^t |x_2(s)-x_1(s)|\,ds \\
&\qquad + \tfrac{1}{2}c_1\,\|F^2\|\int_0^t |x_2(s)-x_1(s)|^2\,ds + \tfrac{1}{2}c_1^2\,\|F^3\|\,e^{\beta t}\rho\int_0^t |x_2(s)-x_1(s)|^2\,ds \\
&\quad \leqslant c_1(t)\,|x_2-x_1| + 2e^{\beta t}c_1^2\rho(\|F^2\| + \tfrac{3}{4}e^{\beta t}\,\|F^3\|\,c_1\rho)\int_0^t |x_2(s)-x_1(s)|\,ds \\
&\quad = c_1(t)\,|x_2-x_1| + \gamma(t)\rho\int_0^t |x_2(s)-x_1(s)|\,ds
\end{aligned}
$$

and, proceeding as usual, we have

$$|x_2(t)-x_1(t)| \leqslant c_1(t)\,e^{\rho\gamma(t)t}\,|x_2-x_1| \tag{20}$$

If

$$\|F^3\|\,c_1\rho \ll \|F^2\| \tag{21}$$

for example, $\|F^3\|\,c_1\rho \leqslant \frac{1}{10}\|F^2\|$, (20) gives

$$|x_2(t)-x_1(t)| \leqslant c_1(T)\,e^2\,|x_2-x_1| \leqslant 8c_1(T)\,|x_2-x_1|$$

(ii) Subtracting the linear part and applying (20) it follows that

$$
\begin{aligned}
&|x_2(t)-x_1(t)-\mathscr{L}(0,t)(x_2-x_1)| \\
&\quad \leqslant 2e^{\beta t}c_1^2\rho(\|F^2\| + \tfrac{3}{4}e^{\beta t}\,\|F^3\|\,c_1\rho)\int_0^t |x_2(s)-x_1(s)|\,ds \\
&\quad \leqslant \gamma(t)\,\rho c_1(t)\,e^{\rho\gamma(t)t}t\,|x_2-x_1| \\
&\quad \leqslant k\rho\,|x_2-x_1|
\end{aligned} \tag{22}
$$

Remarks. (1) Proceeding more carefully the factor 2 in $\gamma(t)$ could be substituted by $\frac{3}{2}$ if ρ is small.

(2) With $\|F^3\|=0$ [or the condition (21)] it is

$$\rho\gamma(T)T \leqslant 2$$

and so

$$\rho\gamma(t)\, te^{\rho\gamma(t)t} \leqslant 2 \cdot e^2 < 2 \cdot 8$$

We did not use these fixed bounds because ρ can be much less than the value given by equality in (15), and so the bounds (16) and (17) are better. In our applications ρ is in fact small and we needed these finer results.

(3) The bound (22) controls the nonlinear terms of the flow. It is easy to understand that, apart from some corrections which we will consider in the next section, of which the principal one is a factor $(1 + c_4/c_3)$, (22) is also the principal part of the nonlinear terms of the Poincaré map: that is, K_0 of Theorem 2 is nearly equal to $K(1 + c_4/c_3)$ if T is the time of the Poincaré map.

(4) It is to be stressed that the inequality (22) is really quadratic for its dependence on ρ and $|x_2 - x_1|$.

As a consequence, while $|x_2(t) - x_1(t)|$ is in fact expanding (at least the function that gives the superior bound is expanding), on the contrary $|x_2(t) - x_1(t) - \mathscr{L}(0, t)(x_2 - x_1)|$, in virtue of its intrinsic quadratic character, taking ρ much smaller than the value given in (15) so that $K\rho$ is less than 1, can become a contraction with respect to $|x_2 - x_1|$: the distance of the points, subtracted the linear part, is less than its initial value. The main criterion says essencially that, strengthening enough this contraction character on the nonlinear part of the Poincaré map so that $k\bar{\rho}$ is not only less than 1, but satisfies the condition (8), then the Poincaré map has a unique fixed point.

Corollary. If x_1 and x_2 are in a ρ neighborhood of γ and $x_0(t^*)$ is one of the points of γ at a distance less than ρ from x_1 and x_2, then

$$|s_t x_2 - s_t x_1| \leqslant e^{\rho\gamma(T)T} c_1(t^*, t^* + t)\, |x_2 - x_1|$$

with

$$c_1(t^*, t^* + t) = \sup_{0 \leqslant \tau \leqslant t} \|\mathscr{L}(t^*, t^* + \tau)\|$$

Proof. It follows from the autonomous character of the system. ∎

4. POINCARÉ MAP

First of all we give two simple properties of the Poincaré map: the uniform continuities of the map.

Lemma 3. If x_1 and x_2 are two points in the ρ neighborhood of x_0 and $x_1' = P(x_1)$, $x_2' = P(x_2)$ are the corresponding Poincaré maps obtained from the flow after a time T_1 and T_2, respectively, then, supposing $T_2 > T_1$

$$\text{(i)}\quad \Delta T = T_2 - T_1 \leqslant \frac{e^{\rho\gamma(T_1)T_1}}{c_3} c_1(T_1)\,|x_2 - x_1| \tag{23}$$

and

$$\text{(ii)}\quad |x_2' - x_1'| = |P(x_2) - P(x_1)|$$

$$\leqslant e^{\rho\gamma(T_1)T_1} c_1(T_1)\left(1 + \frac{c_4}{c_3}\right)|x_2 - x_1| \tag{24}$$

Proof. From Lemma 1, applied to $|x_i(T) - x_0(T)|$ where T is the time of the Poincaré map of x_0, we have

$$|x_i(T) - x_0(T)| \leqslant e^{\beta T} c_1(T)\,|x_i - x_0| \tag{25}$$

and, while $x_0(T)$ is on the plane $x_d = x$ by definition, (25) gives a bound for the distance of $x_i(T)$ from that plane. In any case, the points $x_i(T)$ are in the $\mathcal{U}_{2\rho c_1(T)}$ neighborhood of $x_0(T)$. In this neighborhood the minimum value of the velocity orthogonal to the plane is c_3 and so

$$\Delta T_i = |T_i - T| \leqslant e^{\beta T}\frac{c_1(T)}{c_3}|x_i - x_0| \tag{26}$$

Obviously if ρ is small, also ΔT is small, for, example, of the order of numerical integration step. The maximum value of the distance of $x_i(T_i)$ from $x_i(T)$ projected on the plane $x_d = c$ is obviously

$$c_4\,\Delta T_i \leqslant e^{\beta T} c_1(T)\,|x_i - x_0|\,\frac{c_4}{c_3}$$

and so finally

$$|P(x_i) - P(x_0)| = |x(T_i) - x_0(T)| \leqslant e^{\beta T} c_1(T)\,|x_i - x_0|\left(1 + \frac{c_4}{c_3}\right)$$

$$\leqslant 2c_1\rho\left(1 + \frac{c_4}{c_3}\right) = \rho^* \tag{26'}$$

Now, applying Lemma 2 in the same way and noting that in any case the points $x_i(t)$ are in a $\mathcal{U}_{\rho_1}(x_0)$ neighborhood of x_0, $\rho_1 = \rho^* + \varepsilon$, for t from T to T_i, as a consequence of (16) we obtain the results.

Remark. Continuing Remark (3) of Lemma 2 and considering the result (24), we expect that, if we remove the linear part of the Poincaré map, the inequality we will obtain is

$$|P(x_2)-P(x_1)-L(x_2-x_1)| \leqslant k\rho\, |x_2-x_1| \left(1+\frac{c_4}{c_3}\right) \tag{27}$$

in analogy to (16), (17), apart from some minor corrections due to the fact that now the time of the flow of the Poincaré map depends on the point. We consider (27) well established by previous considerations and only for completeness do we add a more formal derivation of it; obviously, the reader who is not interested in the details can skip the rest of this Remark.

In fact, in what follows we also obtain an expression for the matrix L, the linear part of the Poincaré map in X_0.

If T_i is the time of the Poincaré map of X_i (suppose $T_2 > T_1$), we have

$$\begin{aligned} P(x_2)-P(x_1) &= x_2(T_2)-x_1(T_1) \\ &= x_2(T_2)-x_2(T_1)+x_2(T_1)-x_1(T_1) \\ &= F(x_2(\bar{t}))(T_2-T_1)+\mathscr{L}(0, T_1)(x_2-x_1) \\ &\quad + \int_0^{T_1} \mathscr{L}(s, T_1)\, H(x_1(s), x_2(s))\, ds \end{aligned} \tag{28}$$

By definition $[x_2(T_2)-x_1(T_1)]_d=0$, so

$$\begin{aligned} T_2-T_1 = -\frac{1}{f_d(x_2(\bar{t}_d))} \bigg[& \mathscr{L}(0, T_1)(x_2-x_1) \\ & + \int_0^{T_1} \mathscr{L}(s, T_1)\, H(x_1(s), x_2(s))\, ds \bigg]_d \end{aligned}$$

and, substituting in (28) we have

$$\begin{aligned} &[P(x_2)-P(x_1)]_i \\ &\quad = \sum_k \left[\mathscr{L}_{ik}(0, T_1) - \frac{f_i(x_2(\bar{t}_i))}{f_d(x_2(\bar{t}_d))} \mathscr{L}_{dk}(0, T_1) \right] (x_1-x_2)_k \\ &\quad\quad + \int_0^{T_1} \sum_k \left[\mathscr{L}_{ik}(s, T_1) - \frac{f_i(x_2(\bar{t}_i))}{f_d(x_2(\bar{t}_d))} \mathscr{L}_{dk}(s, T_1) \right] H(x_1(s), x_2(s))_k\, ds \end{aligned}$$

From this expression it follows easily (taking $X_1 = X_0$ and $X_2 \Rightarrow X_0$) that

$$L_{ik} = \mathscr{L}_{ik}(0, T) - \frac{f_i(x_0(T))}{f_d(x_0(T))} \mathscr{L}_{dk}(0, T), \qquad i \neq d \tag{28'}$$

We then have

$$
\begin{aligned}
&[P(x_2)-P(x_1)-L(x_2-x_1)]_i \\
&\quad =\sum_k [\mathscr{L}_{ik}(0, T_1)-\mathscr{L}_{ik}(0, T)](x_2-x_1)_k \\
&\quad +\sum_k \left[\frac{f_i(x_0(T))}{f_d(x_0(T))}-\frac{f_i(x_2(\bar t_i))}{f_d(x_2(\bar t_d))}\right] \mathscr{L}_{dk}(0, T)(x_2-x_1)_k \\
&\quad +\sum_k \frac{f_i(x_2(\bar t_i))}{f_d(x_2(\bar t_d))} [\mathscr{L}_{dk}(0, T)-\mathscr{L}_{dk}(0, T_1)](x_2-x_1)_k \\
&\quad +\int_0^{T_1} \sum_k \left[\mathscr{L}_{ik}(s, T_1)-\frac{f_i(x_2(\bar t_i))}{f_d(x_2(\bar t_d))} \mathscr{L}_{dk}(s, T_1)\right] H(x_1(s), x_2(s))_k \, ds
\end{aligned}
$$

From this, recalling Lemma 2, (ii), and the previous bounds (26) and (26′), we obtain

$$
\begin{aligned}
&|P(x_2)-P(x_1)-L(x_2-x_1)| \\
&\quad \leqslant c_1 c_5 \, \Delta T_1 \, e^{c_5 \Delta T_1} \, |x_2-x_1| \\
&\quad\quad + c_1^2 e^{\beta T} \frac{c_5}{c_3} \rho \left(1+\frac{c_4}{c_3}\right)^2 |x_2-x_1| \\
&\quad\quad + c_1 \frac{c_4}{c_3} c_5 \, \Delta T_1 \, e^{c_5 \Delta T_1} \, |x_2-x_1| \\
&\quad\quad + \left(1+\frac{c_4}{c_3}\right) \frac{K(T)}{T} (T+\Delta T_1) \, e^{c_5 \Delta T_1} \rho \, |x_2-x_1| \\
&\quad \leqslant \left(1+\frac{c_4}{c_3}\right)\left[K(T)\left(1+\frac{\Delta T_1}{T}\right) e^{c_5 \Delta T_1} + e^{\beta T} c_1^2 \frac{c_5}{c_3} e^{c_5 \Delta T_1} \right. \\
&\quad\quad \left. + e^{\beta T} c_1^2 \frac{c_5}{c_3}\left(1+\frac{c_4}{c_3}\right)\right] \rho \, |x_2-x_1|
\end{aligned}
$$

so

$$
\begin{aligned}
K_0 = &\left(1+\frac{c_4}{c_3}\right)\left\{\left[K(T)\left(1+\frac{e^{\beta T}}{T}\frac{c_1}{c_3}\rho\right)+e^{\beta T} c_1^2 \frac{c_5}{c_3}\right] \exp\left(c_5 e^{\beta T} \frac{c_1}{c_3}\rho\right) \right. \\
&\left. + \, e^{\beta T} c_1^2 \frac{c_5}{c_3}\left(1+\frac{c_4}{c_3}\right)\right\}
\end{aligned}
$$

and Theorem 2 of the Introduction is proved, because, as we already noted, the principal part in K_0 is

$$K_0 \simeq \left(1+\frac{c_4}{c_3}\right) K(T) = \left(1+\frac{c_4}{c_3}\right) c_1(T)\, T\gamma(T)\, e^{\rho\gamma(T)T}$$

5. ERRORS OF INTEGRATION. THE PSEUDOFLOW

In each step of integration we introduce an error that is due to two different causes.

First of all there is the error due to the numerical method of integration, and then the round-off error of the computer. While the first can be reduced with a procedure of integration more precise, the second is an intrinsic limit of the machine (at least if one does not use multiple or infinite precision programs).

We assume knowledge of the region in which the solution is contained either by virtue of considerations like those of the Introduction or by some numerical method. This is just to suppose that the round-off error can be uniformly bounded. Call α_c the error in one step of integration. It can be evaluated considering the number of elementary operations needed for one step of integration. α_c can also be determined as a result of interval analysis applied to a single step of integration.

With regard to the numerical integration error α_I, we note that, using for example Taylor expansion of order n, the error is bounded by

$$\alpha_I \leqslant \frac{1}{n+1} \left(\sup_{i \leqslant n+1} \|F^i\| \right)^{n+1} \cdot \Delta^{n+1} \tag{29}$$

with $\|F^i\|$ having the meaning of paragraph 1.

It is understood that, using again the knowledge of the region of the motion, we can reduce this error as we like choosing small Δ and high n, for example until it is smaller than the computer error. We observe that also now we do not need a careful knowledge of the region of the motion. We denote with α the sum of the two errors, $\alpha = \alpha_I + \alpha_c$.

Remark. We are supposing that the derivatives of F greater than 3 have bounds not greater than $\|F^3\|$. The result (29) can be improved, if some of the $\|F^i\|$ is much bigger than the others, taking the right power of it in the rest of Taylor.

Let now $\{x_k\}$ be the pseudotrajectory obtained by numerical integration. If R_Δ is the operator of numerical integration, that, is the operator that had to give x_{k+1} in terms of x_k, we really have

$$|x_{k+1} - R_\Delta x_k| \leqslant \alpha_c$$

for the round-off error of the computer. So

$$x_{k+1} = R_\Delta x_k + \xi_k$$

where $|\xi_k| \leqslant \alpha_c$.

Consider also the operator R_t with $0 \leqslant t \leqslant \Delta$, $R_0 = E$, obtained by substituting in the numerical integration operator the step of integration Δ with the continuous parameter t, and define

$$\bar{x}_k(t) = R_t x_k + \frac{t}{\Delta}\,\xi_k, \qquad 0 \leqslant t \leqslant \Delta,\ k = 0, 1,..., N-1$$

so that $\bar{x}_k(\Delta) = R_\Delta x_k + \xi_k = x_{k+1}$. That means that we are considering the pseudoflow, that is, the continuous curve connecting x_k with x_{k+1} which, in our case, is the analog of connecting x_k and x_{k+1} by means of segments in the case of linear approximation.

Let $\bar{x}(t)$ be this pseudoflow, for which then

$$\bar{x}(t) = \bar{x}_k(t') \quad \text{with } k = \left[\frac{t}{\Delta}\right] \text{ and } t' = t - k\Delta$$

[ ] denoting the integer part.

We observe first that we can estimate

$$\|\bar{F}'\| = \sup_{0 \leqslant t \leqslant T} |F'(\bar{x}(t))| = \sup_k \sup_{0 \leqslant t \leqslant \Delta} |F'(\bar{x}_k(t))|$$

along the pseudoflow by

$$F'(\bar{x}_k(t)) = F'(k_k) + F''(x_k)[\bar{x}_k(t) - x_k] + \tfrac{1}{2}F'''(\xi_k)[\bar{x}_k(t) - x_k]^2$$

[recall that we need $F'''(\xi_k)$ only if $F''(x)$ is not constant or well bounded] and then

$$\|\bar{F}'\| \leqslant \|F'\| + \|F^2\| \cdot \sup_{k, 0 \leqslant t \leqslant \Delta} |\bar{x}_k(t) - x_k| + \tfrac{1}{2}\|F^3\| \sup_{k, 0 \leqslant t \leqslant \Delta} |\bar{x}_k(t) - x_k|^2$$

Bounds for $|\bar{x}_k(t) - x_k|$ are easily obtained. For Δ conveniently small we can have for example

$$\sup_k \sup_{0 \leqslant t \leqslant \Delta} |\bar{x}_k(t) - x_k| \leqslant 2\,\|F\|\,\Delta$$

We are now ready to give a bound to the distance of the pseudoflow from the flow.

Proposition 1. If $(N+1)\, e^{\rho\gamma(T)T} c_1\, \alpha \leqslant \rho$, where ρ is such that

$$2c_1^2 \rho(\|F^2\| + \tfrac{2}{3}\|F^3\|\, c_1 \rho) T \leqslant 1 \tag{15}$$

and

$$N = \left[\frac{T}{\Delta}\right]$$

defining

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

then

$$|\bar{x}(t) - x(t)| \leqslant e^{\rho\gamma(T)T}(k+1)\, \mathscr{C}_1(t)\alpha$$

with

$$k = \left[\frac{t}{\Delta}\right] \leqslant N$$

and

$$\bar{x}(t) = \bar{x}_k(t), \qquad 0 \leqslant \tau \leqslant \Delta$$

Proof. By definition of α we have, for every k,

$$|\bar{x}_k(\tau) - s_\tau x_k| \leqslant \alpha \tag{30}$$

For $\tau = \Delta$ (30) means

$$|x_{k+1} - s_\Delta x_k| \leqslant \alpha \tag{31}$$

The corollary of Lemma 2 is now applicable:

$$|s_{j\Delta+\tau} x_{k-j} - s_{j\Delta+\tau}(s_\Delta x_{k-j-1})| \leqslant e^{\rho\gamma(T)T} \mathscr{C}_1(k\Delta + \tau)\alpha$$

Then by

$$\bar{x}_k(\tau) - s_{k\Delta+\tau} x_0 = \sum_{0j}^{k-1} [s_{j\Delta+\tau} x_{k-j} - s_{j\Delta+\tau}(s_\Delta x_{k-j-1})] + (\bar{x}_k(\tau) - s_\tau x_k)$$

it follows

$$\begin{aligned} |\bar{x}(t) - x(t)| &= |\bar{x}_k(\tau) - s_{k\Delta+\tau} x_0| \\ &\leqslant |\bar{x}_k(\tau) - s_\tau x_k| + \sum_{0j}^{k-1} |s_{j\Delta+\tau} x_{k-j} - s_{j\Delta+\tau}(s_\Delta x_{k-j-1})| \\ &\leqslant (k+1)\, \mathscr{C}_1(t)\alpha \times e^{\rho\gamma(T)T} \end{aligned}$$

A control of the procedure can be obtained with the interval analysis applied to the whole orbit, but the final error of the interval analysis can be bigger of this bound for the propagated error.

Remark. Knowing that the flow is contained in some bounded region, we can estimate uniformly the error due to truncation of Taylor expansion and to the computer. Then, from Proposition 1 it follows that the flow $x(t)$ is contained in the $\mathscr{W}_\rho$ neighborhood of the pseudoflow $\bar{x}(t)$ with

$$\bar{\rho} = (N+1)\, e^{\rho\gamma(T)T} c_1 \alpha$$

We can then evaluate again $\|F^3\|$, the unique norm depending not only on the pseudotrajectory, in this smaller region, obtaining hopefully a narrower bound for it to use in the previous expressions.

6. A CRITICAL CONSTANT

We come now to the evaluation of the most important constant of the criterion, the constant C_1 (Note that K_0 depends on C_1^3 and C_1 appears in all bounds of the Lemmas).

Introduce the matrices:

$$\tilde{\mathscr{L}}(i, i) = E, \qquad i = 0, 1, ..., N$$

$$\tilde{\mathscr{L}}(0, l) = \prod_{0^k}^{l-1} [E + \Delta F'(x_k)], \qquad i = 1, ..., N$$

$$\tilde{\mathscr{L}}(i, l) = \prod_{i^k}^{l-1} [E + \Delta F'(x_k)], \qquad i < l,\ l = 1, ..., N$$

and, for P a fixed integer,

$$\tilde{\mathscr{L}}(I, L) = \prod_{IP^k}^{LP-1} [E + \Delta F'(x_k)]$$

Define

$$\tilde{C}_1(T) = \sup_l \|\tilde{\mathscr{L}}(0, l)\|, \qquad \tilde{C}_1 = \sup_{I,L} \|\tilde{\mathscr{L}}(I, L)\| \tag{32}$$

Our aim is to use the matrices $\tilde{\mathscr{L}}(i, l)$ and the value $\tilde{C}_1$ to bound the norms of the fundamental matrix solution $\mathscr{L}(s, t)$ of the linear equation

$$\dot{A} = F'(x_0(t))A$$

along the flow $x_0(t)$ satisfying the condition

$$\mathscr{L}(s, s) = E$$

We have to stress that there are many causes of errors and precisely:

(i) $\tilde{\mathscr{L}}(i, k+1)$ is not exactly given in terms of $\tilde{\mathscr{L}}(i, k)$, but the round-off errors of the computer are introduced.

(ii) $\tilde{\mathscr{L}}(i, l)$ is not evaluated for every i and l when $i \neq 0$ [see (32)].

(iii) $\tilde{\mathscr{L}}(i, l)$ is not the fundamental matrix solution along the pseudotrajectory.

(iv) $\tilde{\mathscr{L}}(i, l)$ is evaluated in fact along the pseudotrajectory and not along the trajectory.

Nevertheless, we try to bound the norms of $\mathscr{L}(s, t)$ in terms of the norms of $\tilde{\mathscr{L}}(I, L)$. First of all consider the computer's error.

(i) The matrices we found by means of the computer are not $\tilde{\mathscr{L}}(i, l) = \prod_{ik}^{i-1} [E + \Delta F'(x_k)]$ but the matrices $\tilde{\tilde{\mathscr{L}}}(i, l)$ such that

$$\tilde{\tilde{\mathscr{L}}}(i, l) = [E + \Delta F'(x_l)]\, \tilde{\tilde{\mathscr{L}}}(i, l-1) + \delta_{i,l}$$

The errors considered here are those the computer adds once the pseudotrajectory $\{x_k\}$ is given. Later we will consider the problem of the pseudotrajectory as a cource of error.

If we assume that β is a uniform bound for $\delta_{i,l}$ in the case under consideration, $\|\delta_{i,l}\| \leqslant \beta$, we can easily compute how the error propagates. We have

$$\tilde{\tilde{\mathscr{L}}}(i, l) = \prod_{ik}^{l-1} [E + \Delta F'(x_k)] + \sum_{i+1j}^{l-1} \prod_{j+1k}^{l-1} [E + \Delta F'(x_k)]\, \delta_{i,j}$$

and then, denoted by $\tilde{\tilde{C}}_1$ the sup of the norm of $\tilde{\tilde{\mathscr{L}}}(i, l)$,

$$\begin{aligned} \|\tilde{\mathscr{L}}(i, l)\| &\leqslant \|\tilde{\tilde{\mathscr{L}}}(i, l)\| + \beta \sum_{i+1j}^{l-1} \|\tilde{\mathscr{L}}(j+1, l)\| \\ &\leqslant \|\tilde{\tilde{\mathscr{L}}}(i, l)\| + (l-i)\beta \sup_{i \leqslant j \leqslant l} \|\tilde{\mathscr{L}}(i, l)\| \end{aligned}$$

and

$$\begin{aligned} \tilde{C}_1 = \sup_{i,l} \|\tilde{\mathscr{L}}(i, l)\| &\leqslant \sup_{i,l} \|\tilde{\tilde{\mathscr{L}}}(i, l)\| + N\beta \sup_{j,l} \|\tilde{\mathscr{L}}(i, l)\| \\ &\leqslant \tilde{\tilde{C}}_1 + N\beta \tilde{C}_1 \end{aligned}$$

If $N\beta < 1$ it follows

$$\tilde{C}_1 \leqslant \frac{\tilde{\tilde{C}}_1}{1 - N\beta} \tag{33}$$

In our case, for the elementary error given in paragraph 6, and for the orbits considered, β is certainly smaller than 10^{-8}, so this correction is not very significant.

We now go on to the second point.

(ii) It is easily understood that, knowing

$$\tilde{C}_1 = \sup_{I,L} \|\tilde{\mathscr{L}}(I, L)\|$$

the corresponding value with the sup done over every i and l is given by

$$\begin{aligned} \tilde{C}_1^* &= \sup_{i,l} \|\tilde{\mathscr{L}}(i, l)\| \leqslant \tilde{C}_1 \cdot e^{\|F'\| \Delta P} \\ &= \tilde{C}_1 \cdot e^{\|F'\| \Delta_1} \qquad \text{with} \quad \Delta_1 = \Delta P \end{aligned}$$

Remark. This is, in a sense, the most critical correction, meaning that knowing a first check of $\|F'\|$, we have to choose Δ; so that $\|F'\|\ \Delta_1$ is, for example, not greater than $\frac{1}{2}$, this value already giving a correction to the point (ii) of more than the 50%. But the number of operations needed to evaluate $\tilde{C}_1$ grows quadratically when Δ_1 is reduced. All other corrections depend linearly on Δ.

We come then to consider the error due to the approximation we used of the fundamental matrix solution along the pseudotrajectory. Furthermore we pass from discrete to continuous values of time.

(iii) Introduce the matrices $\bar{\mathscr{L}}(s, t)$, fundamental matrix solution of the linear equation

$$\dot{B} = F'(\bar{x}(t))B$$

with the condition $\bar{\mathscr{L}}(s, s) = E$ and define

$$\bar{C}_1 = \sup_{i,l} \|\bar{\mathscr{L}}(i\Delta, l\Delta)\|$$

We have

$$\bar{C}_1 \leqslant \sup_{i,l} \|\tilde{\mathscr{L}}(i, l)\| + \sup_{i,l} \|\tilde{\mathscr{L}}(i, l) - \bar{\mathscr{L}}(i\Delta, l\Delta)\|$$

$$\leqslant \tilde{C}_1^* + \sup_{i,l} \left\| \sum_{i^r}^{l-1} \tilde{\mathscr{L}}(i, r) \right.$$

$$\left. \times [(E + \Delta F'(x_r)) - \bar{\mathscr{L}}(r\Delta, (r+1)\Delta)] \times \bar{\mathscr{L}}((r+1)\Delta, l\Delta) \right\|$$

$$\leqslant \tilde{C}_1^* + N\tilde{C}_1^* \Delta^2 \|F'\|^2 \bar{C}_1 = \tilde{C}_1^* + \tilde{C}_1^* T\Delta \|F'\|^2 \bar{C}_1 \tag{34}$$

Finally, if

$$\tilde{C}_1^* T\Delta \|F'\|^2 < 1 \tag{35}$$

(34) gives a useful inequality

$$\bar{C}_1 \leqslant \frac{\tilde{C}_1^*}{1 - \Delta \tilde{C}_1^* T \|F'\|^2}$$

In (34) we used the condition $\Delta \|F'\| \leqslant \frac{1}{2}$ contained in (35). When the condition (35) is not satisfied, or the value is for example greater than $\frac{1}{2}$, we can improve the procedure without changing Δ, defining the matrix $\tilde{\mathscr{L}}(i, l)$ in the following way:

$$\tilde{\mathscr{L}}(i, l) = \prod_{i^k}^{l-1} \{E + \Delta F'(x_k) + \tfrac{1}{2}\Delta^2 [F'^2(x_k) + F''(x_k) \cdot F(x_k)]\}$$

This gives another factor Δ in the left member of (35) which now becomes $\tilde{C}_1^* T\Delta^2 (\|F'\|^3 + \|F\|\ \|F'\|\ \|F^2\| + \|F^2\| \cdot \|F^3\|)$ To go from discrete to continuous values of time it is sufficient to multiply for the factor $e^{\|\bar{F}'\|\Delta}$ so that

$$\sup_{0 \leqslant s \leqslant t \leqslant T} \|\bar{\mathscr{L}}(s, t)\| = \bar{C}_1^* \leqslant \bar{C}_1 e^{\|\bar{F}'\|\Delta} \qquad (\|\bar{F}'\| \text{ on the pseudoflow})$$

We come now to the final correction. We have to bound $\|\mathscr{L}(s, t)\|$, where $\mathscr{L}(s, t)$ is the fundamental matrix solution of the linear variational equation along the flow $x_0(t)$, in terms of $\|\bar{\mathscr{L}}(s, t)\|$ with $\bar{\mathscr{L}}(s, t)$ the same matrix along the pseudoflow $\bar{x}(t)$.

(iv) From the definition of $\mathscr{L}$ and $\bar{\mathscr{L}}$, and their properties, the following integral relation can be deduced:

$$\mathscr{L}(s, t) = \bar{\mathscr{L}}(s, t) + \int_s^t \bar{\mathscr{L}}(\tau, t)\{F''(\bar{x}(\tau))[\bar{x}(\tau) - x_0(\tau)]$$

$$+ \tfrac{1}{2}F'''(\xi(\tau))[\bar{x}(\tau) - x_0(\tau)]^2\}\ \mathscr{L}(s, \tau)\, d\tau$$

which gives

$$\|\mathscr{L}(s,t)\| \leqslant \|\bar{\mathscr{L}}(s,t)\| + \bar{C}_1^* \, \|\bar{F}^2\| \int_s^t |\bar{x}(\tau) - x_0(\tau)| \cdot \|\mathscr{L}(s,\tau)\| \, d\tau$$

$$+ \tfrac{1}{2}\bar{C}_1^* \, \|F^3\| \int_s^t |\bar{x}(\tau) - x_0(\tau)|^2 \, \|\mathscr{L}(s,\tau)\| \, d\tau$$

Using now Proposition 1, with ω denoting the coefficient of $\mathscr{C}_1(t)$,

$$\omega = (N+1)\, e^{\rho\gamma(T)T}\alpha$$

we have

$$\mathscr{C}_1(t) \leqslant \bar{C}_1^* + \bar{C}_1^* \, \|\bar{F}^2\| \, \omega \int_0^t \mathscr{C}_1^2(\tau)\, d\tau + \tfrac{1}{2}\bar{C}_1^* \, \|F^3\| \, \omega^2 \int_0^t \mathscr{C}_1^3(\tau)\, d\tau$$

where $\|\bar{F}^2\|$ refers to the pseudoflow and is evaluated like $\|\bar{F}'\|$.

Remark. C_1 is contained also in ω, but for ρ small the result does not depend on that. In any case also this condition has to be checked at the end. Proceeding as in Paragraph 2 for the equation

$$\begin{aligned} \dot{z}(t) &= \alpha z^2 + \beta z^3 \\ &\leqslant [\alpha + \beta\gamma z(0)]\, z^2 \qquad \text{if} \quad z(t) \leqslant \gamma z(0) = \gamma \bar{C}_1^* \end{aligned}$$

we have

$$\mathscr{C}_1(t) \leqslant \frac{\bar{C}_1^*}{1 - [\alpha + \beta\gamma z(0)]\, \bar{C}_1^* t}$$

with the consistency condition

$$(\bar{C}_1^{*2} \, \|\bar{F}^2\| \, \omega + \tfrac{1}{2}\bar{C}_1^{*3} \, \|F^3\| \, \omega^2 \cdot 2)T = \delta \leqslant \tfrac{1}{2}$$

so that

$$C_1 \leqslant \frac{\bar{C}_1^*}{1-\delta} < 2\bar{C}_1^*$$

Come finally to the evaluation of $C_I(T)$. We need it to have a bound of L through the formula (28′). We see that we have to consider only the points (iii) and (iv) because (i) is negligible and (ii) does not occur because the supremum was done over all i

$$\tilde{C}_1(T) = \sup_i \|\tilde{\mathscr{L}}(0, i)\|$$

The correction (iii) is the same, but depends on the previous knowledge of $\bar{C}_1$. It is

$$\bar{C}_1(T) \leqslant \frac{\tilde{C}_1(T)}{1 - \Delta \tilde{C}_1 T \|F'\|^2}$$

Similarly, from the procedure in (iv), it follows that

$$C_1(T) \leqslant \frac{\bar{C}_1(T)}{1-\delta}$$

7. APPLICATIONS

We applied the previous results to two different versions of the Lorenz model. With the bounds in [2], especially with the corrections for C_1 used there, we could not conclude, as we do now, that the numerical closed orbits we studied satisfy the condition of the criterion.

First of all consider the system studied by Vul and Sinai[1]

$$\dot{x} = a_1 x + b_1 yz + b_1 xz$$

$$\dot{y} = a_2 y - b_1 yz - b_1 xz$$

$$\dot{z} = -a_3 z + (x+y)(b_2 x + b_3 y)$$

which by a linear transformation reduces to the Lorenz system (see below). With

$$a_1 = 9.700378782, \qquad b_1 = -0.227266206$$

$$a_2 = -16.700378782, \qquad b_2 = 2.616729797$$

$$a_3 = 2.666666667, \qquad b_3 = -1.783396463$$

the values of the parameters of the corresponding Lorenz equations are

$$r = 28, \qquad 6 = 6, \qquad b = 8/3$$

The fixed point found by Vul and Sinai

$$(3.5007872047249;\ 3.3303317970426;\ 27)$$

is not stable, but the Newton method converges the same rapidly. We obtain a value for $|\bar{x} - x_0|$ of the order of 10^{-14}. The intersection with the plane $z = 27$ was found reducing recursively by a factor 2 the step of integration until the desired precision was reached. The computer used is the Univac 1100 of the University of Rome. The elementary error of it, in double precision, is less than 3×10^{-18}. We can assume that the computer error in one step of integration is not greater than 10^{-14} (interval analysis gives a value less than 10^{-15}). We used the Taylor expansion scheme to integrate the system, and could change easily the order of the expansion with a recursive routine. The step of integration used was $\Delta = 10^{-5}$ and $P\Delta = \Delta_1 = 5 \times 10^{-3}$ so all previous critical conditions are satisfied, considering also that the value for ρ we obtain is $\rho = 0.7 \times 10^{-5}$. We observe that we do not need a careful knowledge of the constants: only ε and the corrections to c_1 deserve attention. The larger contribution to ε is due to $|x_N - x(N\Delta)|$, and it can be made smaller by increasing Δ and, correspondingly, the order of Taylor: for example, with $\Delta = 10^{-3}$, the number of steps of integration decreases by a factor 10^2 but the number of elementary operations in one step of integration does not grow by the same factor passing from 4 to 7 terms in the Taylor expansion. The corrections to $\tilde{C}$, on the contrary, depend critically on Δ and Δ_1, and we cannot easily satisfy the condition to have both small. We already noted that the evaluation of $\tilde{C}_1$ depends quadratically on T/Δ_1.

The parameters for the orbit of Vul and Sinai are

$$c_1 \leqslant 7, \quad \|F'\| \leqslant 45, \quad \|F^2\| \leqslant 6, \quad c_3 \geqslant 49, \quad c_4 \leqslant 16$$
$$c_5 \leqslant 27, \quad T = 0.68992, \quad \|(L-E)\|^{-1}\| \leqslant 20, \quad K_0 < 3 \times 10^3$$

so it is easily seen that the condition (8) of the criterion is satisfied with $\varepsilon = 1.5 \times 10^{-7}$, $\bar{\rho} = 0.7 \times 10^{-5}$. The bound of precision needed for the computer is

$$\alpha \leqslant \frac{\varepsilon}{N e^{\rho\gamma(T)T} c_1} \simeq 0.3 \times 10^{-12}$$

a value well within the limits of our computer. The other closed orbit studied is that found by Franceschini and Tebaldi in the Lorenz system[4]

$$\dot{x} = -sx + sy$$
$$\dot{y} = -y - (s+z)x$$
$$\dot{z} = -Bz + xy - R$$

for the following values of the parameters: $s = 10$, $B = 8/3$, $R = 294.13333$. The coordinates of the fixed point are

$$(-9.3249753511062; 7.70489278613672; -11)$$

In this case the values of $\tilde{C}_1$ and of $\|F'\|$ are large and we had to use, for example, the definition of $\tilde{\mathscr{L}}$ given in (iii) of Section 6 and $\Delta_1 = 2 \times 10^{-3}$.

The final value of the parameters, considering also Remark 1 of Lemma 2, is

$$c_1 \leqslant 89, \qquad \|F'\| \leqslant 64, \qquad \|F^2\| \leqslant \sqrt{2}, \qquad c_3 \geqslant 336, \qquad c_4 172$$
$$c_5 \leqslant 21, \qquad T = 1.09751, \qquad \|(L-E)^{-1}\| \leqslant 1, 2, \qquad K_0 < \tfrac{5}{2} \times 10^6$$

The condition of the criterion is satisfied with $\varepsilon = 0.9 \times 10^{-7}$, $\bar{\rho} = \frac{5}{2} \times 10^{-7}$. The precision needed for the computer is $\alpha \leqslant 0.9 \times 10^{-14}$ and we can say that, even if to the limit, the precision we have is the right one (recall anyway that also in this case the interval analysis gives an error of 10^{-15}).

The final conclusion is that in both systems there exists only one periodic orbit in the $\bar{\rho}$ neighborhood of the numerical periodic orbit.

REFERENCES

1. Ja. G. Sinai and E. E. Vul, *J. Stat. Phys.* **23**:27 (1980).
2. S. De Gregorio, E. Scoppola, and B. Tirozzi, *J. Stat. Phys.* **32**:25 (1983).
3. D. Ruelle, *Statistical Mechanics and Dynamical Systems* (Duke University Press, Durham, North Carolina 1976); J. E. Marsden and M. McCracken, *The Hopf Biforcation and Its Applications* (Springer, New York, 1976); M. J. Fiegenbaum, *J. Stat. Phys.* **19**:25 (1978).
4. C. Boldrighini and V. Franceschini, *Commun. Math. Phys.* **64**:159 (1979); V. Franceschini, *J. Stat. Phys.* **22**:397 (1980); J. Pedlosky and C. Frenzen, *J. Atmos. Sci.* **37**:1177 (1980).