A Rigorous ODE Solver and Smale's 14th Problem

Warwick Tucker*
IMPA, Est. D. Castorina 110
Jardim Botânico, 22460-320
Rio de Janeiro, RJ, Brazil
warwick@impa.br

July 26, 2000

2000 Mathematics Subject Classification: Primary 37C10; Secondary 37D45, 65G30

Abstract

We present an algorithm for computing rigorous solutions to a large class of ordinary differential equations. The main algorithm is based on a partitioning process and the use of interval arithmetic with directed rounding. As an application, we prove that the Lorenz equations support a strange attractor, as conjectured by Edward Lorenz in 1963. This conjecture was recently listed by Steven Smale as one of several challenging problems for the 21st century. We also prove that the attractor is robust, i.e., it persists under small perturbations of the coefficients in the underlying differential equations. Furthermore, the flow of the equations admits a unique SRB measure, whose support coincides with the attractor. The proof is based on a combination of normal form theory and rigorous computations.

*Current address: Department of Mathematics, Malott Hall, Cornell University, Ithaca, NY, 14853-4201, USA; email: warwick@math.cornell.edu

Contents

1 Introduction

Here, we give a brief description of and the background to the problem concerning the existence of the Lorenz attractor. For precise definitions, we refer the reader to the Appendix. A rather comprehensive overview of this problem can be found in Collin Sparrow's book, [23].

1.1 Background to the problem

The following non-linear system of differential equations,

$$\begin{aligned} \dot{x}_1 &= -\sigma x_1 + \sigma x_2 \\ \dot{x}_2 &= \varrho x_1 - x_2 - x_1 x_3 \\ \dot{x}_3 &= -\beta x_3 + x_1 x_2, \end{aligned} \tag{1}$$

was introduced in 1963 by Edward Lorenz, see [9]. As a crude model of atmospheric dynamics, these equations led Lorenz to the discovery of sensitive dependence of initial conditions - an essential factor of unpredictability in many systems. Numerical simulations for an open neighbourhood of the classical parameter values $\sigma = 10, \beta = 8/3$ and $\varrho = 28$ suggest that almost all points in phase space tend to a strange attractor - *the Lorenz attractor.*

We first note that the system (1) (and thus its solution) is invariant under the transformation $S(x_1, x_2, x_3) = (-x_1, -x_2, x_3)$. This means that any trajectory that is not itself invariant under S must have a "twin trajectory".

For $\varrho > 1$, there are three fixed points: the origin and the two "twin points"

$$C^{\pm} = (\pm\sqrt{\beta(\varrho - 1)}, \pm\sqrt{\beta(\varrho - 1)}, \varrho - 1).$$

For the parameter values we are considering, $C^{\pm}$ have a pair of complex eigenvalues with positive real part, and one real, negative eigenvalue. The origin is a saddle point with two negative and one positive eigenvalue satisfying

$$0 < -\lambda_3 < \lambda_1 < -\lambda_2.$$

Thus, the stable manifold of the origin $W^s(0)$ is two-dimensional, and the unstable manifold of the origin $W^u(0)$ is one-dimensional.

It is also worth mentioning that the flow contracts volumes at a significant rate. As the divergence of the vector field is given by

$$\frac{\partial \dot{x}_1}{\partial x_1} + \frac{\partial \dot{x}_2}{\partial x_2} + \frac{\partial \dot{x}_3}{\partial x_3} = -(\sigma + \beta + 1),$$

we see that the volume of a solid at time t can be expressed as

$$V(t) = V(0)e^{-(\sigma+\beta+1)t} \approx V(0)e^{-13.7t},$$

for the classical parameter values. This means that the flow contracts volumes almost by a factor *one million* per time unit, which is quite extreme.

There appears to exist a forward invariant open set U containing the origin but bounded away from $C^{\pm}$. The set U is a torus of genus two, with its holes centered around the two excluded fixed points. If we let φ denote the flow of (1), we can form the maximal invariant set

$$\mathcal{A} = \bigcap_{t \geq 0} \varphi(U, t).$$

Due to the flow being dissipative, the attracting set $\mathcal{A}$ must have zero volume. It must also contain the unstable manifold of the origin $W^u(0)$, which seems to spiral around $C^{\pm}$ in a very complicated, non-periodic fashion, see Figure 1. In particular, $\mathcal{A}$ contains the origin itself, and therefore the

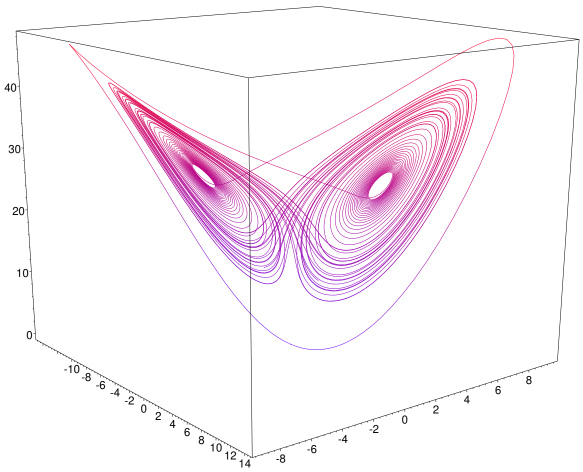

Figure 1: A part of the unstable manifold of the origin.

flow on $\mathcal{A}$ can not have a hyperbolic structure. The reason is that fixed points of the vector field generate discontinuities for the return maps, and as a consequence, the hyperbolic splitting is not continuous. Apart from this, the attracting set appears to have a strong hyperbolic structure as described below.

As it was very difficult to extract rigorous information about the attracting set $\mathcal{A}$ from the differential equations themselves, a *geometric model* of the Lorenz flow was introduced by John Guckenheimer in the late sixties, see [4]. This model has been extensively studied, and it is well understood today, see e.g. [5], [25], [23], [17], [19], [20]. Oddly enough, the original equations introduced by Lorenz have remained a puzzle. A few computer-assisted proofs, however, have quite recently been announced, see [3], [6], and [12]. These articles deal with subsets of $\mathcal{A}$ which are not attracting, and therefore only concern a set of trajectories having measure zero. Despite this, it has always been widely believed that the flow of the Lorenz equations has the same qualitative behaviour as its geometric model.

The geometric model is made up of two pieces: one piece dealing with all trajectories passing near the origin, and one piece taking care of the global aspects of the flow. We consider a flow with a fixed point at the origin with eigenvalues just as the Lorenz flow. We also assume that there exists a unit rectangle $\Sigma \subset \{x_3 = 1\}$ which is transversal to the flow, such that the induced return map R acts on Σ as illustrated in Figure 2.

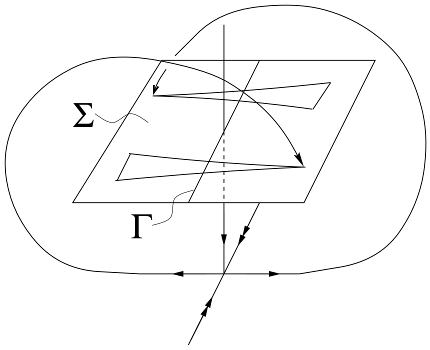

Figure 2: The return map acting on Σ.

Note that R is not defined on the line $\Gamma = \Sigma \cap W^s(0)$: these points tend to the origin, and never return to Σ. We will assume that $R(\Sigma \setminus \Gamma) \subset \Sigma$, to ensure that the flow has an attracting set with a large basin of attraction. We can now decompose the return map: $R = D \circ P$, where D is a diffeomorphism corresponding to the flow outside a unit cube centered at the origin, and P describes the flow inside the cube. By assuming that the flow is linear in the cube, we can explicitly find P:

$$P(x_1, x_2, 1) = \big(\mathrm{sgn}(x_1), x_2|x_1|^{-\lambda_2/\lambda_1}, |x_1|^{-\lambda_3/\lambda_1}\big).$$

Seeing that $-\lambda_3/\lambda_1 < 1 < -\lambda_2/\lambda_1$, we have very strong expansion in the x_1-direction, and an even stronger contraction in the x_2-direction:

$$\lim_{|x_1|\to 0} \frac{\partial P_3}{\partial x_1} = \mathcal{O}(|x_1|^{|\lambda_3|/\lambda_1 - 1}) \qquad \text{and} \qquad \lim_{|x_1|\to 0} \frac{\partial P_2}{\partial x_2} = \mathcal{O}(|x_1|^{|\lambda_2|/\lambda_1}).$$

The model assumes that the flow outside the cube preserves the x_2-direction, i.e., that D takes the horizontal lines $\ell(t) = (\pm 1, t, c)$ into lines $\tilde{\ell}(t) = (\tilde{c}, t, 1)$, $t \in [-1, 1]$. This ensures that the contracting direction is preserved, and it also implies that the first component of the return map is independent of x_2. Therefore, we can write $R = \big(R_1(x_1), R_2(x_1, x_2)\big)$. Further assumptions are that $\frac{\partial R_2}{\partial x_2} \leq \mu < 1$ and $R_1'(x_1) > \sqrt{2}$ for all $x_1, x_2 \in \Sigma$. The return map now has a hyperbolic splitting $\mathbb{E}^s_x \oplus \mathbb{E}^u_x$, with $\mathbb{E}^s_0 = \Gamma$, and the *stable leaves* $\tilde{\ell}(t)$ foliate Σ. Since all points on a stable leaf share a common future, we may form an equivalence class of such points. By taking the quotient, we get an interval map f (note that $f = R_1$), which is assumed to satisfy the following conditions:

1. f has a unique singularity at 0 with $f(0^-) = 1$, and $f(0^+) = -1$,
2. $f\colon [-1, 1] \setminus \{0\} \to [-1, 1]$,
3. f is C^1 on $[-1, 1] \setminus \{0\}$, and $f'(x) > \sqrt{2}$ for $x \neq 0$,

This suffices to prove that almost all points in [-1,1] have dense orbits under f. It is also clear that f exhibits exponential sensitivity. By pulling the information back to the original return map, it is possible to prove that the attracting set of the model flow is a generalized non-trivial hyperbolic attractor (also known as a *singular* hyperbolic attractor).

Before we close this section, let us make some simplifying remarks. By a linear change of variables, the Lorenz equations can be put in their Jordan normal form $\dot{x} = Ax + F(x)$:

$$\begin{aligned} \dot{x}_1 &= \lambda_1 x_1 - k_1(x_1 + x_2)x_3 \\ \dot{x}_2 &= \lambda_2 x_2 + k_1(x_1 + x_2)x_3 \\ \dot{x}_3 &= \lambda_3 x_3 + (x_1 + x_2)(k_2 x_1 + k_3 x_2). \end{aligned} \tag{2}$$

When we want to be very brief, we use the notation $\dot{x} = f(x)$, where $f(x)$ is the right-hand side of (2). Note that the parameters k_1, k_2 and k_3, and the eigenvalues λ_1, λ_2 and λ_3 only depend on the parameters σ, β and ϱ appearing in (1). By inserting the classical parameter values, we get the following approximate system:

$$\begin{aligned} \dot{x}_1 &= 11.8x_1 - 0.29(x_1 + x_2)x_3 \\ \dot{x}_2 &= -22.8x_2 + 0.29(x_1 + x_2)x_3 \\ \dot{x}_3 &= -2.67x_3 + (x_1 + x_2)(2.2x_1 - 1.3x_2). \end{aligned}$$

From now on, we will always refer to (2) as the Lorenz equations.

1.2 The main result

In a recent issue of *the Mathematical Intelligencer* the Fields medalist Steven Smale presented a list of challenging problems for the 21th century, see [22]. Problem number 14 reads as follows:

> *Is the dynamics of the ordinary differential equations of Lorenz that of the geometric Lorenz attractor of Williams, Guckenheimer, and Yorke?*

A historical remark is perhaps in order here. James Yorke was *not* involved in the actual work on the geometric attractors. He should, however, be credited for introducing Lorenz' original article to the mathematical community. Apparently, Yorke had written his name on his copy of the article, and when he faxed it to colleagues, his name became associated with the Lorenz attractor. Yorke also published several articles on the matter, see e.g. [26].

As an affirmative answer to Smale's question, we are now ready to state the sole theorem of this paper:

Main Theorem *For the classical parameter values, the Lorenz equations support a robust strange attractor $\mathcal{A}$. Furthermore, the flow admits a unique SRB measure μ_φ with* $\mathrm{supp}(\mu_\varphi) = \mathcal{A}$.

In fact, we prove that the attracting set is a singular hyperbolic attractor. Almost all nearby points separate exponentially fast until they end up on opposite sides of the attractor. This means that a tiny blob of initial values rapidly smears out over the entire attractor, as observed in numerical experiments.

The existence of the SRB measure is equivalent to saying that, for Lebesgue almost all points in the basin of attraction $B(\mathcal{A})$, and for all $h \in C^0(B(\mathcal{A}), \mathbb{R})$, the time and space averages coincide:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T h(\varphi(x,t))dt = \int h(x)d\mu_\varphi,$$

where μ_φ is an φ-invariant probability measure.

It is perhaps worth pointing out that the Lorenz attractor does not act quite as the geometric model predicts. The latter can be reduced to an interval map which is everywhere expanding. This is not the case for the Lorenz attractor: there are large regions in Λ that are contracted in all directions under the return map. Such regions, however, are pre-compensated by iterates having a large associated expansion. This corresponds to the interval map being eventually expanding, and does not lead to any different qualitative long-time behaviour.

Apart from this, the Lorenz attractor is just as the geometric model predicts: it contains the origin, and thus has a very complicated Cantor book structure as described in [25].

Acknowledgements

Most of this work is based on my Ph.D. thesis, which was carried out at Uppsala University, Uppsala, Sweden, under the supervision of Professor Lennart Carleson. The RODES program was developed during my stay at IMPA, Rio de Janeiro, Brazil, which was financed by a STINT fellowship through KTH, Stockholm, Sweden.

2 Outline of the approach

In this section, we will briefly describe the main phases of our approach. Basically, it can be broken down into two main sections: one global part, which involves rigorous computations, and one local part, which is based on normal form theory. A more detailed description of all necessary steps will be given in subsequent sections.

The novelty of the method of proof lies in that, rather than producing a traditional mathematical proof, we construct an algorithm which, if successfully executed, proves the existence of the strange attractor. This algorithm is put into effect via several C++ programs, all of which use interval arithmetic with directed rounding, provided by the PROFIL/BIAS package, [8]. The source codes and initial data used in the proof are available from the author's home page: `http://www.math.uu.se/~warwick/papers.html`.

2.1 Goals

In all attempts to prove Lorenz's conjecture, one main obstruction has been to gain useful global information about the flow far away from the origin. Locally, the evolution of a trajectory near a fixed point can, in principle, be examined in detail using standard linearization techniques. Outside a small neighbourhood of the fixed point, however, we are usually completely at a loss. This is exactly the situation that led Guckenheimer and Williams to construct the geometric Lorenz flows, which were *defined* to have exactly those global properties that were not easily attained from the original equations.

The global properties we will prove are the following:

- The return map R exists, and it is well defined in the sense of the geometric model.
- There exists a compact subset of the return plane, $N \subset \Sigma$, such that $N \setminus \Gamma$ is *forward invariant* under R, i.e., $R(N \setminus \Gamma) \subset N$. This ensures that the flow has an attracting set $\mathcal{A}$ with a large basin of attraction. We can then form a cross-section of the attracting set: $\mathcal{A} \cap \Sigma = \cap_{n=0}^{\infty} R^n(N) = \Lambda$. In particular, Λ is an attracting set for R.
- On N, there exists a cone field $\mathfrak{C}$ which is mapped strictly into itself by DR, i.e., for all $x \in N$, $DR(x) \cdot \mathfrak{C}(x) \subset \mathfrak{C}(R(x))$. The cones of $\mathfrak{C}$ are centered along an approximation of Λ, and each cone has an opening of at least $5°$.
- The tangent vectors in $\mathfrak{C}$ are eventually expanded under the action of DR: there exists $C > 0$ and $\lambda > 1$ such that for all $v \in \mathfrak{C}(x)$, $x \in N$, we have $|DR^n(x)v| \geq C\lambda^n|v|$, $n \geq 0$. In fact, the expansion is strong enough to ensure that R is topologically transitive on Λ. This is equivalent to having a dense orbit, and therefore proves that Λ is an attractor.

We will use these properties to prove that Λ (and thus $\mathcal{A}$) carries a singular hyperbolic structure.

2.2 Normal form theory

As mentioned earlier, some regions of the return plane will flow close to the origin. These regions consist of all points in a neighbourhood of the intersection between the stable manifold of the origin and the forward invariant set, see Figure 2.

Once a rectangle has flowed close to the origin, we interrupt the computations, and introduce a close to identity change of coordinates $\Phi(\zeta) = \zeta + \phi(\zeta)$. This change of coordinates will deform the rectangle and its cone field slightly, but in a controllable way. In the new coordinates, the vector field assumes a carefully designed normal form, which is virtually linear (although it is crucial that it need not be completely linear). This permits us to estimate the evolution of the rectangle and its cone field analytically, and thereby avoid the problem of having to use computers. When changing back to the original coordinates, we once again deform the out-going rectangle and its cone field, but still in a controllable fashion.

The change of coordinates is attained by a method developed by H. Poincaré which, at first, seems fairly straight-forward. Basically, we construct ϕ formally by choosing the desired normal form mentioned above. The question of convergence, however, involves a small divisor problem, and is somewhat non-trivial. Furthermore - in order to be able to interrupt the computations before the vector field has become so small that the numerical process breaks down - we need convergence on a relatively large neighbourhood of the origin. This requires explicit knowledge of a large number of coefficients of ϕ, and thus forces us to perform some rather involved induction based proofs.

2.3 Rigorous numerics

The trapping region N consists of two disjoint components, N^- and N^+, each made up of several adjacent rectangles belonging to the return plane $x_3 = 27$ $(= \varrho - 1)$. We will call these small rectangles $N_i^{\pm}$, and write

$$N = N^- \cup N^+ = \Big(\bigcup_{i=1}^{n_0} N_i^-\Big) \cup \Big(\bigcup_{i=1}^{n_0} N_i^+\Big).$$

The two components of N have the same symmetry as the Lorenz equations, i.e., $N_i^+ = S(N_i^-)$, where $S(x_1, x_2, x_3) = (-x_1, -x_2, x_3)$. Thanks to this symmetry, we only have to perform the computations on one component of N. When it is not relevant which component we are considering, we omit the $\pm$ labeling of the small rectangles.

Dealing with one N_i at a time, we compute a pseudo-path that strictly contains the flow of N_i. The pseudo-paths are obtained by introducing several intermediate return planes $\Sigma^{(k)}$, which are either x_1x_2-, x_1x_3-, or x_2x_3-planes. At each step, the plane is chosen so that its normal direction e_i has the same direction as the strongest component of the flow: $|f_i(x)| \geq |f_j(x)|$, $j = 1, 2, 3$. The initial rectangle N_i is flowed to the first plane $\Sigma^{(1)}$ by using an Euler method with rigorous error estimates. In the plane $\Sigma^{(1)}$, we take the rectangular hull of the largest image of N_i, giving us a new starting rectangle $\mathcal{R}^{(1)}$. This rectangle is then flowed to $\Sigma^{(2)}$ and so on. If a rectangle $\mathcal{R}^{(k)}$ has grown too large it is partitioned into smaller rectangles, which are then treated separately. This whole procedure is repeated until we return to Σ from above, as illustrated in Figure 2. Due to the contracting forces, the pseudo-return of N_i will consist of many overlapping rectangles $Q_{i,j}$ $j = 1, \ldots, k(i)$, whose union strictly contains $R(N_i)$.

The use of rectangles significantly simplifies the computations: when flowing between two intermediate planes $\Sigma^{(k)}$, $\Sigma^{(k+1)}$, it is generically the corners of $\mathcal{R}^{(k)}$ that yield the largest rectangular hull $\mathcal{R}^{(k+1)} \subset \Sigma^{(k+1)}$. This fact allows us to reduce the error analysis to small pieces of $\mathcal{R}^{(k)}$, which greatly reduces the local errors. With only finite precision, however, this property becomes "pseudo-generic", and has to be confirmed at every stage. The exceptional cases are treated slightly differently.

Turning to the question concerning the cone field, we define the field by equipping each N_i with an initial cone. Each cone is represented by the two angles α_i^-, α_i^+ its boundary vectors $u^{(0)}, v^{(0)}$ make with the positive x_1-axis. We then use similar techniques as just described: when a rectangle has been flowed from $\Sigma^{(k)}$ to $\Sigma^{(k+1)}$, we are provided with a box containing the path of the rectangle. The algorithm also gives us upper and lower bounds on the flow time involved. By solving the nine equations governing the partial derivatives of the flow, we obtain rigorous bounds on the evolution of the tangent vectors flowing through the box. By translating the flowed vectors onto the intermediate plane $\Sigma^{(k+1)}$, and by selecting (incorporating the errors) the pair of vectors $u^{(k+1)}, v^{(k+1)}$ making the largest angle $\theta^{(k+1)}$, we ensure that the resulting cone contains all images of tangent vectors from the initial cone. At the return, each rectangle $Q_{i,j}$ is thus equipped with a cone represented, as above, by two angles $\beta_{i,j}^-, \beta_{i,j}^+$, $j = 1, \ldots, k(i)$.

When computing the minimal expansion in each cone, we start with the widest pair of vectors, $u^{(k)}, v^{(k)}$ at each intermediate plane $\Sigma^{(k)}$, as described above. If $\theta^{(k+1)} \leq \theta^{(k)}$, the minimal expansion $\varepsilon^{(k)}$ is attained on the boundary of the cone, i.e., $\varepsilon^{(k)}$ is the smallest growth factor of the images of $u^{(k)}, v^{(k)}$. If $\theta^{(k+1)} > \theta^{(k)}$, however, we must adjust this estimate by a factor which is quadratically close to unity in $\theta^{(k+1)}$. At the return, each rectangle $Q_{i,j}$ is thus equipped with an expansion estimate $\mathcal{E}_{i,j} = \Pi_{k=0}^{n(i,j)} \varepsilon_{i,j}^{(k)}$, and $\mathcal{E}_i = \min_j \mathcal{E}_{i,j}$ gives an estimate for all vectors of the cone associated with N_i.

One major advantage of our numerical method is that we totally eliminate the problem of having to control the global effects of rounding errors due to the computer's internal floating point representation. This is achieved by using interval arithmetic with directed rounding. Each object Ξ subjected to computation is equipped with a maximal absolute error Δ_Ξ, and can thus be represented as a product of intervals $\Xi \pm \Delta_\Xi = [\Xi_1 - \Delta_{\Xi_1}, \Xi_1 + \Delta_{\Xi_1}] \times \cdots \times [\Xi_n - \Delta_{\Xi_n}, \Xi_n + \Delta_{\Xi_n}]$. When performing any operation with such objects, we compute upper bounds on the images of $\Xi_i + \Delta_{\Xi_i}$, and lower bounds on the images of $\Xi_i - \Delta_{\Xi_i}$, $i = 1, \ldots, n$. This results in a new box $\tilde{\Xi} \pm \Delta_{\tilde{\Xi}}$, which strictly contains the exact image of $\Xi \pm \Delta_\Xi$. To ensure that we have strict inclusion, we use directed rounding on the upper and lower bounds.

As long as we do not flow close to a fixed point, the local return maps are well defined diffeomorphisms, and the computer can handle all calculations. Some rectangles, however, will approach the origin (which is a fixed point), and then the computations must be interrupted as discussed in the previous section.

2.4 Topological transitivity and SRB measures

Since the flow of (1) is uniformly volume-contracting and transversal to N, a finite iterate of the return map R is area-contracting on N. This property together with the existence of a forward invariant unstable cone field implies that R admits an invariant stable foliation with $C^{1+\alpha}$ leaves, see [18] or [7, §3]. The singular map f induced by taking quotients along the stable leaves acts on an interval $I = [-a, a]$, and satisfies the following properties:

- The restriction of f to $[-a, 0)$ and $(0, a]$ is of class $C^{1+\alpha}$ with $f'(x) \geq K > 0$ for all $x \neq 0$.
- There exists $C > 0$, $\lambda > 1$ such that $(f^n)'(x) \geq C\lambda^n$, for all $n \geq 0$.
- For any interval $J \subset I$ there exists $n \geq 0$ such that $f^n(J) = I$, i.e., f is *locally eventually onto*.

The last property is not immediate, but can be proved by using Proposition 5.1, see below.

It follows that f admits a unique finite SRB measure μ_f with $\mathrm{supp}(\mu_f) = I$. From this measure it is possible to construct an SRB measure μ_R with $\mathrm{supp}(\mu_R) = \Lambda$ for the return map R, and also an SRB measure μ_φ for the flow, see [1] or [24].

We will now prove that f is locally eventually onto: the map is singular at the origin since Γ itself is a stable leaf, and projects to $\{0\}$. The discontinuity acts as a razor blade, and can cut a passing line segment in half. If neither of the two halves have doubled their lengths before returning to $\{0\}$, they could be cut in half again, and thus we could end up with loads of tiny shreds of the initial segment, all of which continue to hit $\{0\}$ until we are left with nothing but a fine dust.

By Proposition 5.1, however, this can not happen. Any small segment that is cut over Γ will have expanded by more than a factor 2 before returning to Γ. By always selecting the larger half, we get a sequence of longer and longer segments. This will continue until one of them is mapped totally across a *fundamental domain*. In our particular situation, this means a union of adjacent rectangles $F_D = \cup_{i=n_1}^{n_2} N_i$ such that no orbit can cross F_D without having at least one iterate in F_D. Of course, any set containing a fundamental domain will also do. One such example is the set F used in Proposition 5.1.

When a segment stretches entirely across a fundamental domain, so will its projection along the stable leaves. This means that the projection covers an interval on the form $[x, f(x)]$. This set is mapped onto the entire interval I within a finite number of iterates. It follows that f is locally eventually onto.

3 Local theory

In this section we will construct the local change of variables which straightens out the stable and unstable manifolds, and linearizes the flow on these. We will also obtain estimates on the change of variables. For convenience, we will often use the vector notation $\xi = (\xi_1, \xi_2, \xi_3)$ combined with the multi-index notation $\xi^n = \xi_1^{n_1}\xi_2^{n_2}\xi_3^{n_3}$.

3.1 Flattening out the invariant manifolds

In order that the stable and unstable manifolds should coincide with the coordinate planes, it is necessary that these are invariant under the flow. To ensure this, we need a change of variables which, in a small neighbourhood of the origin, transforms the Lorenz equations $\dot{\xi} = A\xi + F(\xi)$ into $\dot{\zeta} = A\zeta + G(\zeta)$, where $G = (G_1, G_2, G_3)$, satisfies the following conditions:

$$G_1(0, \zeta_2, \zeta_3) = 0, \quad \text{and} \quad G_2(\zeta_1, 0, 0) = G_3(\zeta_1, 0, 0) = 0.$$

In these new coordinates, the unstable manifold coincides with the ζ_1-axis, and the stable manifold coincides with the $\zeta_2\zeta_3$-plane. However, this will not linearize the flow on the invariant manifolds. For this, we need to impose the condition

$$G_i \in \mathcal{O}(\zeta_1) \cap \mathcal{O}(\zeta_2, \zeta_3) \qquad (i = 1, 2, 3).$$

This simply means that if a point ζ is close to the ζ_1-axis (the unstable manifold) or the $\zeta_2\zeta_3$-plane (the stable manifold), then the perturbation $G(\zeta)$ is linearly small, i.e.,

$$\min\left\{|\zeta_1|, \max\{|\zeta_2|, |\zeta_3|\}\right\} = \mathcal{O}(\varepsilon) \Rightarrow |G_i(\zeta)| = \mathcal{O}(\varepsilon) \qquad (i = 1, 2, 3).$$

However, we will need to do better than this: we actually want to flatten out the invariant manifolds even more. Flatness of order p is given by requiring that $G \in \mathcal{O}^p(\zeta_1) \cap \mathcal{O}^p(\zeta_2, \zeta_3)$, i.e.,

$$\min\left\{|\zeta_1|, \max\{|\zeta_2|, |\zeta_3|\}\right\} = \mathcal{O}(\varepsilon) \Rightarrow |G_i(\zeta)| = \mathcal{O}(\varepsilon^p) \qquad (i = 1, 2, 3).$$

To be able to talk about smallness, we need a norm to work with. We will work in a complex neighbourhood of the origin, and use the following notations:

$$|\zeta| = \max\{|\zeta_i| \colon i = 1, 2, 3\} \qquad \|f\|_r = \sup\{|f(\zeta)| \colon |\zeta| \le r\}.$$

Our Ansatz is to do the calculations with formal vector-valued polynomials. The following proposition states that we obtain not only a formal change of variables, but that the formal power series actually converges in a fixed neighbourhood of the origin.

Proposition 3.1 *There exists a close to identity change of variables $\xi = \zeta + \phi(\zeta)$ with*

$$\|\phi\|_r \le \frac{r^2}{2} \qquad r \le 1$$

such that the Lorenz equations, $\dot{\xi} = A\xi + F(\xi)$, are transformed into the normal form $\dot{\zeta} = A\zeta + G(\zeta)$, where $G(\zeta) \in \mathcal{O}^{10}(\zeta_1) \cap \mathcal{O}^{10}(\zeta_2, \zeta_3)$, and satisfies

$$\|G\|_r \le 7 \cdot 10^{-9} \frac{r^{20}}{1 - 3r} \qquad r < \frac{1}{3}.$$

Before getting involved in the proof of the proposition, we highlight some important consequences of the statement.

Lemma 3.2 *For any ρ satisfying $0 \le \rho \le \frac{1}{2}$, we have*

$$\|D\phi\|_\rho \le 2\rho.$$

Proof: We use a classical argument in function theory: If $h(z)$ is a regular function of a complex variable z in the disk $|z| \leq r$ where it satisfies $|h(z)| \leq M$, then for $|z| \leq \rho < r$ Cauchy's integral formula gives

$$h'(z) = \frac{1}{2\pi i} \int_\gamma \frac{h(w)}{(w-z)^2} dw,$$

where the path of integration γ can be taken as the circle $|w - z| = r - \rho$. Since the circle lies within $|w| \leq r$, this leads to the estimate

$$|h'(z)| \leq M/(r-\rho) \qquad |z| \leq \rho.$$

Now, given w and ζ such that $|w| = 1$ and $|\zeta| = \rho$, we define

$$h(z) = \phi_i(\zeta_1 + w_1 z, \zeta_2 + w_2 z, \zeta_3 + w_3 z).$$

For $|z| \leq r - \rho$, we clearly have $|h(z)| \leq \|\phi\|_r$, and so $|h'(z)| \leq \|\phi\|_r/(r - \rho - |z|)$. In particular, for $z = 0$, we have $|h'(0)| \leq \|\phi\|_r/(r-\rho)$. On the other hand, we have

$$|h'(0)| = \Big|\sum_{j=1}^{3} \frac{\partial \phi_i}{\partial \zeta_j}(\zeta) w_j\Big| = \sum_{j=1}^{3} \Big|\frac{\partial \phi_i}{\partial \zeta_j}(\zeta)\Big|,$$

by taking $w_j = \mathrm{sgn}(\frac{\partial \phi_i}{\partial \zeta_j}(\zeta))$. Since $\|D\phi\|_\rho = \max\{|D\phi(\zeta)w| \colon |\zeta| \leq \rho, |w| = 1\}$, we immediately have the following estimates of $\|D\phi\|_\rho$ for $\rho < r \leq 1$:

$$\|D\phi\|_\rho \leq \frac{\|\phi\|_r}{r-\rho} \leq \frac{r^2}{2(r-\rho)}. \tag{3}$$

It is easy to see that the optimal bound is given by substituting r for 2ρ. Inserting this value directly gives the result. □

We also get estimates for the norm of the inverse change of variables:

Lemma 3.3 *For $|\zeta| \leq r < \frac{1}{2}$, the change of variables $\xi = \zeta + \phi(\zeta)$ has a well defined inverse $\zeta = \xi + \psi(\xi)$ in the ball $|\xi| \leq \tilde{r} = r - \|\phi\|_r$ satisfying*

$$\|\psi\|_{\tilde{r}} \leq \|\phi\|_r \qquad \|D\psi\|_{\tilde{r}} \leq \frac{\|D\phi\|_r}{1 - \|D\phi\|_r}.$$

Proof: Set $\Phi(\zeta) = \zeta + \phi(\zeta)$, $\Psi(\xi) = \xi + \psi(\xi)$ and let D denote the ball $|\zeta| \leq r$. By Proposition 3.1, it is clear that $\Phi(D)$ must contain a ball $\tilde{D}$, centered at the origin, and of radius $\tilde{r} = r - \|\phi\|_r$, see Figure 3.

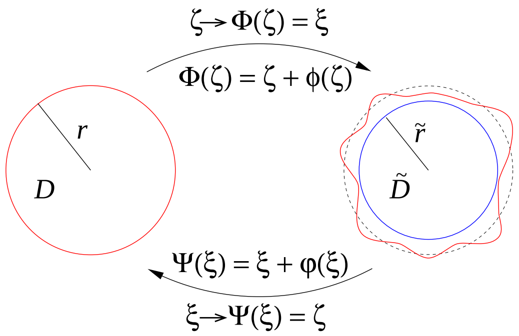

Figure 3: The change of variables and its inverse.

Let $\zeta_1, \zeta_2 \in D$. Then we have

$$|\Phi(\zeta_1) - \Phi(\zeta_2)| \geq |\zeta_1 - \zeta_2| - |\phi(\zeta_1) - \phi(\zeta_2)| \geq (1 - \|D\phi\|_r)|\zeta_1 - \zeta_2|.$$

Hence, if $\|D\phi\|_r < 1$, then Φ is injective on the whole of $\Phi(D)$, and in particular on $\tilde{D}$. By Lemma 3.2, we know that $\|D\phi\|_r \leq 2r$ for all $r \leq \frac{1}{2}$. This proves that the inverse Φ is well defined in the ball $|\xi| \leq \tilde{r} = r - \|\phi\|_r$ for $r < \frac{1}{2}$.

Combining the two coordinate changes, we immediately arrive at

$$\psi(\xi) = -\phi(\xi + \psi(\xi)),$$

which gives the first estimate of the lemma. Differentiating gives

$$D\psi(\xi) = -D\phi(\xi + \psi(\xi))(I + D\psi(\xi)),$$

which, after solving for $D\psi(\xi)$, becomes

$$D\psi(\xi) = -[I + D\phi(\xi + \psi(\xi))]^{-1} D\phi(\xi + \psi(\xi)).$$

The second estimate now follows by using the well known estimate

$$\left\|[I + D\phi]^{-1}\right\|_r \leq \frac{1}{1 - \|D\phi\|_r}.$$

□

Let us conclude this section by explicitly computing the maximal numerical values of the coordinate changes that may appear in the program. We will change coordinates in the cube centered at the origin, and having radius $\frac{1}{10}$.

Lemma 3.4

$$\|\phi\|_{\frac{1}{10}} \leq \frac{1}{200} = 0.005 \qquad \|D\phi\|_{\frac{1}{10}} \leq \frac{2}{10} = 0.2,$$
$$\|\psi\|_{\frac{1}{10}} < 0.00557281 \qquad \|D\psi\|_{\frac{1}{10}} < 0.26766109.$$

Proof: The first line of the lemma follows immediately from Proposition 3.1 and Lemma 3.2, respectively. The second line requires a little more work. Suppose that we want to apply the inverse change of variables at the distance r from the origin. By Lemma 3.3, we must find r^* such that $r = r^* - \|\phi\|_{r^*}$. Once we know r^*, we can use the formulæ

$$\|\psi\|_r \leq \|\phi\|_{r^*} \qquad \|D\psi\|_r \leq \frac{\|D\phi\|_{r^*}}{1 - \|D\phi\|_{r^*}}$$

to obtain the desired estimates. Solving for r^* gives $r^* = 1 - \sqrt{1-2r}$, and plugging this into the estimates above gives

$$\|\psi\|_r \leq \tfrac{(r^*)^2}{2} = 1 - r - \sqrt{1-2r} \qquad \|D\psi\|_r \leq \frac{2r^*}{1-2r^*} = \frac{2(1-\sqrt{1-2r})}{1-2(1-\sqrt{1-2r})}.$$

The numerical values (rounded up) of these estimates for $r = \frac{1}{10}$ appear in the statement. □

This lemma tells us several things: when we enter the small cube, we must increase the radii of the rectangle by the amount 0.005. We also lose 20 per-cent of the cone field information. When we leave the small cube, we must increase the radii of the outgoing rectangle by the amount 0.00557281, and now we lose roughly 26.77 per-cent of the cone field information. This may seem like a large loss of information, but thanks to the strong hyperbolicity near the origin, the cones are very narrow when entering the cube, and therefore the widening due to the change of coordinates is affordable.

Lemma 3.5 *When the almost horizontal vector $(1, \varepsilon)$ is distorted by a change of variables $\xi = \zeta + h(\zeta)$, its slope will increase by at most a factor*

$$\frac{1 + \|Dh\|}{1 - \|Dh\|}.$$

Proof: The worst case is attained when the vector $(1, \varepsilon)$ is mapped to $(1 - \|Dh\|, \varepsilon(1 + \|Dh\|))$. This vector has exactly the slope stated in the lemma. □

3.2 Proof of Proposition 3.1

First, we need to know how the vector field (2) is affected by the close to identity change of variables $\xi = \zeta + \phi(\zeta)$. We have the following:

$$\dot{\xi} = A(\zeta + \phi(\zeta)) + F(\zeta + \phi(\zeta)) = A\zeta + A\phi(\zeta) + F(\zeta + \phi(\zeta)). \tag{4}$$

On the other hand, we also have

$$\begin{aligned}\dot{\xi} &= \frac{d}{dt}(\zeta + \phi(\zeta)) = (I + D\phi(\zeta))\dot{\zeta} = (I + D\phi(\zeta))(A\zeta + G(\zeta)) \\ &= A\zeta + D\phi(\zeta)A\zeta + G(\zeta) + D\phi(\zeta)G(\zeta).\end{aligned} \tag{5}$$

Comparing the two right-hand sides of (4) and (5) gives

$$D\phi(\zeta)A\zeta - A\phi(\zeta) = F(\zeta + \phi(\zeta)) - D\phi(\zeta)G(\zeta) - G(\zeta). \tag{6}$$

For shorthand, we will use the following notation

$$L_A\phi(\zeta) = D\phi(\zeta)A\zeta - A\phi(\zeta).$$

The operator L_A is linear, and it acts on the space of formal vector fields. It leaves the spaces of homogeneous vector-valued polynomials of any degree invariant. Looking at (6) on component level, we have

$$L_{A,i}\phi_i(\zeta) = F_i(\zeta + \phi(\zeta)) - \sum_{j=1}^{3} \frac{\partial \phi_i}{\partial \zeta_j}(\zeta)G_j(\zeta) - G_i(\zeta) \qquad (i = 1, 2, 3), \tag{7}$$

where

$$L_{A,i}\phi_i(\zeta) = \sum_{j=1}^{3} \frac{\partial \phi_i}{\partial \zeta_j}(\zeta)\lambda_j\zeta_j - \lambda_i\phi_i(\zeta) \qquad (i = 1, 2, 3).$$

Note that

$$L_{A,i}(a_n\zeta^n) = (n_1\lambda_1 + n_2\lambda_2 + n_3\lambda_3 - \lambda_i)a_{n_1,n_2,n_3}\zeta_1^{n_1}\zeta_2^{n_2}\zeta_3^{n_3} = (n\lambda - \lambda_i)a_n\zeta^n.$$

The crux is now to choose ϕ so that we have $G_i \in \mathcal{O}^p(\zeta_1) \cap \mathcal{O}^p(\zeta_2, \zeta_3)$. This means that $G_i(\zeta)$, must not contain elements on the form $\zeta^n = \zeta_1^{n_1}\zeta_2^{n_2}\zeta_3^{n_3}$, where the natural numbers n_i satisfy $n_1 < p$ or $n_2 + n_3 < p$. By (7), elements on this form can only come from $F_i(\zeta + \phi(\zeta))$, and any such term can be canceled by an appropriate choice of ϕ_i if the corresponding *divisor* $(n\lambda - \lambda_i)$ does not vanish. Thus the component functions ϕ_i $(i = 1, 2, 3)$, need only consist of the undesired elements just described.

We start by splitting the 3-space of natural numbers into two disjoint sets: $\mathbb{N}^3 = \mathbb{U}_p \cup \mathbb{V}_p$, where

$$\mathbb{V}_p = \{(n_1, n_2, n_3) \in \mathbb{N}^3 : n_1 < p \text{ or } n_2 + n_3 < p\}.$$

Next, we define the following filters, which act on formal vector-valued polynomials: Consider $H = (H_1, H_2, H_3)$, where

$$H_i(\zeta) = \sum_n a_{i,n}\zeta^n = \sum_n a_{i,n_1,n_2,n_3}\zeta_1^{n_1}\zeta_2^{n_2}\zeta_3^{n_3} \qquad (i = 1, 2, 3).$$

Then we define

$$\{H_i(\zeta)\}_{\mathbb{U}_p} = \sum_{n \in \mathbb{U}_p} a_{i,n}\zeta^n, \quad \text{and} \quad \{H_i(\zeta)\}_{\mathbb{V}_p} = \sum_{n \in \mathbb{V}_p} a_{i,n}\zeta^n.$$

We extend the definition of these filters so that they act not only on components, but also on the whole formal vector-valued polynomial: $H = \{H\}_{\mathbb{U}_p} \oplus \{H\}_{\mathbb{V}_p}$. Note that $\{G\}_{\mathbb{U}_p} = G$, $\{G\}_{\mathbb{V}_p} = 0$, $\{\phi\}_{\mathbb{U}_p} = 0$, $\{\phi\}_{\mathbb{V}_p} = \phi$, and $\{MG\}_{\mathbb{V}_p} = 0$ for any 3×3-matrix M with formal polynomial entries. By filtering (7), we get

$$L_{A,i}\phi_i(\zeta) = \{F_i(\zeta + \phi(\zeta))\}_{\mathbb{V}_p} \qquad (i = 1, 2, 3), \tag{8}$$

and

$$G_i(\zeta) = \{F_i(\zeta + \phi(\zeta))\}_{\mathbb{U}_p} - \sum_{j=1}^{3} \frac{\partial \phi_i}{\partial \zeta_j}(\zeta) G_j(\zeta) \qquad (i = 1, 2, 3). \tag{9}$$

3.2.1 The change of variables

The recursive scheme (8) can be formally solved by a power series

$$\phi_i(\zeta) = \sum_{|n|=2}^{\infty} a_{i,n}\zeta^n \qquad (i = 1, 2, 3),$$

where the coefficients are determined by inserting this expression into (8). The existence of a solution ϕ is given by comparing both sides of (8) : If $a_{i,n}\zeta^n$ is a term of $\phi_i(\zeta)$ with $|n| = n_1 + n_2 + n_3$, the comparison gives

$$(n\lambda - \lambda_i)a_{i,n} = \gamma,$$

where γ is a polynomial in the coefficients of the terms in ϕ_1, ϕ_2, ϕ_3 of degree less than $|n|$. Thus the existence is proved if we show that the divisors $(n\lambda - \lambda_i)$ do not vanish. As ϕ does not contain constant or linear terms, and $\{\phi\}_{\mathbb{V}_p} = \phi$, the only divisors we need to consider are on the form $n\lambda - \lambda_i$ $(i = 1, 2, 3)$, where $n \in \mathbb{V}_p$ and $|n| \geq 2$. The following computer-aided lemma proves the existence of a formal series for ϕ.

Lemma 3.6 *For any $p \in [2, 10]$, $n \in \mathbb{V}_p$ and $|n| \geq 2$, the divisors*

$$n\lambda - \lambda_i \qquad (i = 1, 2, 3)$$

are bounded away from zero. Furthermore, after a finite time of fluctuations, there exists a sharp lower bound on the modulus of these divisors:

$$|n\lambda - \lambda_i| \geq \big|(p-1)\lambda_1 + \big(|n| - (p-1)\big)\lambda_3 - \lambda_i\big| \qquad (i = 1, 2, 3).$$

This bound is valid when $|n| \geq \mathcal{N}(p)$, where $\mathcal{N}(p) = \frac{11p+6}{2}$ if p is even, and $\mathcal{N}(p) = \frac{11p+7}{2}$ if p is odd.

Proof: Take $|n|$ large. Since $n \in \mathbb{V}_p$, there are two cases to consider:
(1) $n_1 < p$: This means that $n_2 + n_3$ is large, i.e, the divisor $n\lambda - \lambda_i$ is large and negative. Recalling that the eigenvalues satisfy $0 < -\lambda_3 < \lambda_1 < -\lambda_2$, we clearly minimize the divisor's modulus by choosing $n_1 = p - 1$, $n_2 = 0$ and $n_3 = |n| - (p - 1)$, which gives

$$|n\lambda - \lambda_i| \geq \big|(p-1)\lambda_1 + \big(|n| - (p-1)\big)\lambda_3 - \lambda_i\big| \qquad (i = 1, 2, 3).$$

(2) $n_2 + n_3 < p$: This means that n_1 is large, i.e, the divisor $n\lambda - \lambda_i$ is large and positive. Its modulus is minimized by choosing $n_1 = |n| - (p - 1)$, $n_2 = p - 1$ and $n_3 = 0$, which gives

$$|n\lambda - \lambda_i| \geq \big|\big(|n| - (p-1)\big)\lambda_1 + (p-1)\lambda_2 - \lambda_i\big| \qquad (i = 1, 2, 3). \tag{10}$$

Comparing these two candidates for the smallest divisor, we find that case (1) yields the optimal lower bound, as stated in the lemma. This completes the proof of the asymptotic lower bound.

To see when the lower bound becomes valid, we note that the choice $i = 2$ in the right-hand side of (10) gives the smallest divisor for $|n|$ large, i.e.,

$$|n\lambda - \lambda_i| \geq \big|(p-1)\lambda_1 + \big(|n| - (p-1)\big)\lambda_3 - \lambda_2\big| \qquad (i = 1, 2, 3).$$

The expression on the right-hand side has a minimum when its positive part has roughly the same modulus as its negative part. This occurs when

$$(p-1)\lambda_1 + |\lambda_2| \approx \big(|n| - (p-1)\big)|\lambda_3|,$$

or, equivalently, when

$$|n| \approx \frac{1}{|\lambda_3|}\big((p-1)(\lambda_1 + |\lambda_3|) + |\lambda_2|\big).$$

By defining

$$\mathcal{N}(p) = \left\lceil \frac{1}{|\lambda_3|}\big((p-1)(\lambda_1 + |\lambda_3|) + |\lambda_2|\big) \right\rceil,$$

where $\lceil x \rceil = \min\{k \colon k \in \mathbb{N}, k \geq x\}$, we have $\mathcal{N}(p) = \frac{11p+6}{2}$ if p is even, and $\mathcal{N}(p) = \frac{11p+7}{2}$ if p is odd for $p \in [2, 10]$, as stated in the lemma. That this is the appropriate choice for $\mathcal{N}(p)$ when $p \in [2, 10]$ (for $p \geq 11$ it is not) can be checked by explicit calculations.

Finally, to show that the low-order divisors are non-zero, we note that there are only a finite number of them that we have to check. This can be done by explicit calculations carried out on a computer, and gives the desired result. In Table 1, we list the values of $\Omega(k) = \min_{i=1,2,3} \min\{|\lambda n - \lambda_i| \colon |n| = k, n \in \mathbb{V}_{10}\}$. This required the computation of 19,386 divisors. The C++ program `smalldiv.cc` handles all necessary computations. All floating point operations are performed in interval arithmetic with directed rounding (see Section 4.2) which guarantees the correctness of the given lower bounds. □

$\Omega(2) > 2.6667$	$\Omega(3) > 3.4944$	$\Omega(4) > 0.8277$	$\Omega(5) > 1.8389$
$\Omega(6) > 1.1611$	$\Omega(7) > 1.5056$	$\Omega(8) > 1.0112$	$\Omega(9) > 1.1723$
$\Omega(10) > 0.6779$	$\Omega(11) > 0.1835$	$\Omega(12) > 0.3446$	$\Omega(13) > 0.1498$
$\Omega(14) > 0.0112$	$\Omega(15) > 0.4832$	$\Omega(16) > 0.9776$	$\Omega(17) > 0.8165$
$\Omega(18) > 1.1947$	$\Omega(19) > 1.1498$	$\Omega(20) > 0.8614$	$\Omega(21) > 0.3670$
$\Omega(22) > 0.5280$	$\Omega(23) > 0.6891$	$\Omega(24) > 0.1947$	$\Omega(25) > 0.3558$
$\Omega(26) > 0.1386$	$\Omega(27) > 0.6330$	$\Omega(28) > 0.4720$	$\Omega(29) > 0.9663$
$\Omega(30) > 0.8053$	$\Omega(31) > 1.2997$	$\Omega(32) > 1.3670$	$\Omega(33) > 1.6330$
$\Omega(34) > 1.0337$	$\Omega(35) > 0.5393$	$\Omega(36) > 0.7003$	$\Omega(37) > 0.2059$
$\Omega(38) > 2.7941$	$\Omega(39) > 0.1274$	$\Omega(40) > 2.5393$	$\Omega(41) > 0.4607$
$\Omega(42) > 0.9551$	$\Omega(43) > 1.7115$	$\Omega(44) > 1.2885$	$\Omega(45) > 1.3782$
$\Omega(46) > 1.6218$	$\Omega(47) > 1.0449$	$\Omega(48) > 3.7115$	$\Omega(49) > 2.4495$
$\Omega(50) > 0.2172$	$\Omega(51) > 2.7828$	$\Omega(52) > 0.1162$	$\Omega(53) > 2.5505$
$\Omega(54) > 5.2172$	$\Omega(55) > 6.6106$	$\Omega(56) > 3.9439$	$\Omega(57) > 1.2772$
$\Omega(58) > 1.3894$	$\Omega(59) > 4.0561$	$\Omega(60) > 6.7228$	$\Omega(61) > 9.3895$

Table 1: The smallest absolute values of low-order divisors for $p = 10$.

Remark: For a given p, the non-vanishing of the divisors is an open condition. Thus the lemma is valid for an open neighbourhood of the classical parameter values of the Lorenz equations.

Remark: The lemma gives an asymptotic estimate on the growth of the modulus of the smallest divisors. For large $|n|$, we have $|n\lambda - \lambda_i| \sim |n||\lambda_3| \approx \frac{8|n|}{3}$.

Now that we know that the formal power series for ϕ defined by (8) exists, we want to show that it also actually converges. We follow [21], and use the methods of majorants. If

$$f(\zeta) = \sum_n a_{n_1,n_2,n_3}\zeta_1^{n_1}\zeta_2^{n_2}\zeta_3^{n_3} \qquad g(\zeta) = \sum_n b_{n_1,n_2,n_3}\zeta_1^{n_1}\zeta_2^{n_2}\zeta_3^{n_3}$$

are two formal power series, g is said to be a majorant of f, which we denote $f \prec g$, if

$$|a_{n_1,n_2,n_3}| \leq b_{n_1,n_2,n_3}$$

for all the coefficients. Note that the coefficients of g must be real and non-negative, which implies that f must have at least as large radius of convergence as g. We will also use the notation

$$[f(\zeta)]_k = \sum_{|n|=k} a_{n_1,n_2,n_3}\zeta_1^{n_1}\zeta_2^{n_2}\zeta_3^{n_3}.$$

Now suppose that we find a function $\tilde{F}\colon \mathbb{C}^3 \to \mathbb{C}$ such that $F_i \prec \tilde{F}$ $(i = 1, 2, 3)$ and, together with (8), consider the majorant system

$$\tilde{L}_A\varphi_i(\zeta) = \left\{\tilde{F}(\zeta + \varphi(\zeta))\right\}_{\mathbb{V}_p} \qquad (i = 1, 2, 3), \tag{11}$$

where $\tilde{L}_A(\zeta^n) = \Theta(n)\zeta^n$ and $\Theta\colon \mathbb{N}^3 \to \mathbb{R}$ is defined by $\Theta(n) = \min_{i=1,2,3}|n\lambda - \lambda_i|$. This can again be solved formally by a power series

$$\varphi_i(\zeta) = \sum_{|n|=2}^{\infty} b_{i,n}\zeta^n, \tag{12}$$

and it is clear that φ_i is a majorant of ϕ_i. Furthermore, since the right-hand side of (11) is independent of i, we have $\varphi_1 = \varphi_2 = \varphi_3 = \tilde{\varphi}$. If we set $\zeta_1 = \zeta_2 = \zeta_3 = \zeta$, and find a function $\hat{F}\colon \mathbb{C} \to \mathbb{C}$ such that $\tilde{F}(\zeta, \zeta, \zeta) \prec \hat{F}(\zeta)$, we may, together with (11), consider the majorant system

$$\hat{L}_A\psi(\zeta) = \hat{F}(\zeta + \psi(\zeta)), \tag{13}$$

where $\hat{L}_A(\zeta^k) = \Omega(k)\zeta^k$ and $\Omega\colon \mathbb{N} \to \mathbb{R}$ is defined by $\Omega(k) = \min\left\{\Theta(n)\colon |n| = k, \quad n \in \mathbb{V}_p\right\}$. Again, this can be solved formally by a power series

$$\psi(\zeta) = \sum_{k=2}^{\infty} c_k\zeta^k, \tag{14}$$

and it is clear that $\tilde{\varphi}(\zeta, \zeta, \zeta) \prec \psi(\zeta)$. Note that this implies that $\|\phi\|_r \leq \psi(r)$ in the region of convergence. Thus it suffices to prove the convergence of ψ.

Recall that the non-linear part of the Lorenz equations is given by

$$F(\zeta) = \begin{pmatrix} -k_1\zeta_1\zeta_3 - k_1\zeta_2\zeta_3 \\ k_1\zeta_1\zeta_3 + k_1\zeta_2\zeta_3 \\ k_2\zeta_1^2 + (k_2 + k_3)\zeta_1\zeta_2 + k_3\zeta_2^2 \end{pmatrix} \approx \begin{pmatrix} -0.29\zeta_1\zeta_3 - 0.29\zeta_2\zeta_3 \\ 0.29\zeta_1\zeta_3 + 0.29\zeta_2\zeta_3 \\ 2.2\zeta_1^2 + 0.9\zeta_1\zeta_2 - 1.3\zeta_2^2 \end{pmatrix}.$$

Thus if we set

$$\tilde{F}(\zeta) = k_2\zeta_1^2 + (k_2 + k_3)\zeta_1\zeta_2 + k_1\zeta_1\zeta_3 - k_3\zeta_2^2 + k_1\zeta_2\zeta_3,$$

we clearly have $F_i \prec \tilde{F}$ $(i = 1, 2, 3)$. Using the exceptional fact that F only contains quadratic terms, we can find a particularly simple majorant of $\tilde{F}(\zeta, \zeta, \zeta)$:

$$\tilde{F}(\zeta, \zeta, \zeta) = k_2\zeta^2 + (k_2 + k_3)\zeta^2 + k_1\zeta^2 - k_3\zeta^2 + k_2\zeta^2 \prec 5\zeta^2 = \hat{F}(\zeta).$$

Combining this with (13) and (14), we get the following recursive scheme for the coefficients of $\psi(r)$:

$$c_m r^m = \frac{5}{\Omega(m)}\left[\left(r + \sum_{k=2}^{m-1} c_k r^k\right)^2\right]_m \qquad (m = 2, 3, \ldots). \tag{15}$$

Unfortunately, by Lemma 3.6, we have no uniform estimates on $\Omega(m)$ for $m < \mathcal{N}(p)$. This is due to low-order resonances between the eigenvalues, and therefore we must calculate these divisors explicitly. We did this already in the proof of Lemma 3.6, and found that the smallest modulus of such a divisor was $|\lambda_1 + 13\lambda_3 - \lambda_2| \approx 0.0112$. This divisor appears already for $p = 2$, and it remains minimizing for all $p \in [2, 10]$. Using this a an crude estimate on $\Omega(m)$ for $m < \mathcal{N}(p)$ results in very poor bounds on the coefficients of ψ, and thus gives a very small radius of convergence. This problem can be avoided by postponing the use of the recursive scheme (15) until $m \geq \mathcal{N}(p)$. We can achieve this if we can estimate the leading coefficients of ψ by other means. This will produce a smaller function (which we still call ψ) than had we used only (15), but the important fact is that it will still majorize the ϕ_i's.

The way we will proceed is to explicitly compute $a_{i,n}$ $(i = 1,2,3)$ for $|n| = 2, ..., n_1$ by using (8). This is possible by Lemma 3.6. Then, if we define the leading coefficients of φ by $b_n = \max_{i=1,2,3} |a_{i,n}|$ for $|n| = 2, ..., n_1$, we clearly have $[\phi_i(\zeta)]_j \prec [\varphi(\zeta)]_j$ $(i = 1,2,3)$ for $j = 2, ..., n_1$. Continuing, we may define the leading coefficients of ψ by

$$c_j = \sum_{|n|=j} b_n = \sum_{|n|=j} \max_{i=1,2,3} |a_{i,n}| \qquad (j = 2, ..., n_1).$$

Naturally, we then have $[\varphi(r,r,r)]_j = [\psi(r)]_j$ $(= c_j)$ for $j = 2, ..., n_1$. Now assume that we fix a positive integer n_0 and find two positive constants C and M such that the following induction assumption holds: $c_j \leq CM^j$ for $j = n_0 + 1, ..., 2n_0, ..., n_1$, where $n_1 \geq \mathcal{N}(p)$. We then want to prove that $c_j \leq CM^j$ for all $j > n_1$. This we do by induction: by defining $c_1 = 1$, the recursive scheme (15), gives

$$c_{n_1+1} r^{n_1+1} = \frac{5}{\Omega(n_1+1)} \left[\left(\sum_{k=1}^{n_1} c_k r^k \right)^2 \right]_{n_1+1} = \frac{5}{\Omega(n_1+1)} \left(\sum_{k=1}^{n_1} c_k c_{n_1+1-k} \right) r^{n_1+1},$$

so the coefficients of ψ satisfy

$$\begin{aligned} c_{n_1+1} &= \frac{5}{\Omega(n_1+1)} \left(\sum_{k=1}^{n_0} c_k c_{n_1+1-k} + \sum_{k=n_0+1}^{n_1-n_0} c_k c_{n_1+1-k} + \sum_{k=n_1+1-n_0}^{n_1} c_k c_{n_1+1-k} \right) \\ &= \frac{5}{\Omega(n_1+1)} \left(2\sum_{k=1}^{n_0} c_k c_{n_1+1-k} + \sum_{k=n_0+1}^{n_1-n_0} c_k c_{n_1+1-k} \right). \end{aligned}$$

Although we know nothing about the n_0 first coefficients, we know that $c_j \leq CM^j$ for $j = n_0 + 1, \ldots, n_1$ by the induction hypothesis. Using this gives

$$\begin{aligned} c_{n_1+1} &\leq \frac{5}{\Omega(n_1+1)} \left(2\sum_{k=1}^{n_0} c_k CM^{n_1+1-k} + \sum_{k=n_0+1}^{n_1-n_0} C^2 M^{n_1+1} \right) \\ &= \frac{5}{\Omega(n_1+1)} \left(2\sum_{k=1}^{n_0} c_k M^{-k} + (n_1 - 2n_0)C \right) CM^{n_1+1}. \end{aligned}$$

Since we chose n_1 large enough for the lower bound of $\Omega(n_1 + 1)$ (see Lemma 3.6) to be valid, we have

$$c_{n_1+1} \leq \frac{5\left(2\sum_{k=1}^{n_0} c_k M^{-k} + (n_1 - 2n_0)C\right)}{|(p-1)\lambda_1 + (n_1 + 1 - (p-1))\lambda_3 - \lambda_2|} CM^{n_1+1}, \tag{16}$$

which, for n_1 large, gives the asymptotic estimate

$$c_{n_1+1} \sim \frac{5C}{|\lambda_3|} CM^{n_1+1}.$$

From this it follows that, in order to prove the induction step, we must choose $C \leq \frac{|\lambda_3|}{5} \approx \frac{8}{15}$ ($C \leq \frac{1}{2}$ works nicely for an open neighbourhood of the classical parameter values). Then, if we take n_1 sufficiently large, we will have completed the induction step.

The following computer-aided lemma[1] proves the induction hypothesis needed in the above argument, and gives an estimate on the sum appearing in (16).

Lemma 3.7 *For $p \in [2, 10]$, we have the estimates*

$$c_j < 5 \cdot 10^{-6} \left(\frac{5}{9}\right)^j \qquad (j = 11, \ldots, 70) \quad \text{and} \quad \sum_{j=1}^{10} c_j \left(\frac{5}{9}\right)^{-j} < 3.54.$$

Furthermore, $\mathcal{N}(p) < 70$ for $p \in [2, 10]$.

Proof: A computer program that computes the coefficients $a_{i,n}$, and then calculates the sums $c_j = \sum_{|n|=j} \max_{i=1,2,3} |a_{i,n}|$ was constructed, see `coeff.cc`. Again, all floating point operations are performed in interval arithmetic with directed rounding (see Section 4.2) which guarantees the correctness of the given upper bounds. Thanks to the simple form of the non-linear terms appearing in the Lorenz equations, the program only needs to handle two sums and two products. If we define $\Phi(\zeta) = \zeta + \phi(\zeta)$, the two sums we need to compute are of the type:

$$S_1 = [\Phi_1(\zeta) + \Phi_2(\zeta)]_n \quad \text{and} \quad S_2 = [k_2\Phi_1(\zeta) + k_3\Phi_2(\zeta)]_n,$$

and the two products are of the type:

$$P_1 = [(\Phi_1(\zeta) + \Phi_2(\zeta))\Phi_3(\zeta)]_n \quad \text{and} \quad P_2 = [(\Phi_1(\zeta) + \Phi_2(\zeta))(k_2\Phi_1(\zeta) + k_3\Phi_2(\zeta))]_n.$$

The results are presented in Table 2, and it is simple to check that the coefficients c_j satisfy the conditions in the lemma. Since, for $p \in [2, 10]$, $\mathcal{N}(p) \leq 58$, the final statement in the lemma is verified. □

If we assign the constants n_0, C and M the same values as in this lemma, it is plain to see that the expression appearing in (16) is decreasing in n_1, for $n_1 \geq 70$. Thus, if c_{71} satisfies the desired estimate, all coefficients of higher order will too. Considering the worst case, $p = 10$, we have

$$\begin{aligned} c_{71} &\leq \frac{5\left(2 \cdot 3.54 + 5(70-20) \cdot 10^{-6}\right)}{|(10-1)\lambda_1 + (71-(10-1))\lambda_3 - \lambda_2|} CM^{71} \\ &\leq \frac{5\left(7.08 + 2.5 \cdot 10^{-4}\right)}{|9\lambda_1 + 62\lambda_3 - \lambda_2|} CM^{71} \leq 0.982 CM^{71}, \end{aligned}$$

which completes the induction. Hence, for $r < 9/5$, we arrive at

$$\begin{aligned} \psi(r) &\leq \sum_{j=2}^{10} c_j r^j + \sum_{j=11}^{\infty} CM^j r^j \leq r^2 \sum_{j=2}^{10} c_j r^{j-2} + 5 \cdot 10^{-6} \sum_{j=11}^{\infty} \left(\frac{5r}{9}\right)^j \\ &\leq r^2 \sum_{j=2}^{10} c_j r^{j-2} + 5 \cdot 10^{-6} \left(\frac{5r}{9}\right)^2 \sum_{j=9}^{\infty} \left(\frac{5r}{9}\right)^j \\ &\leq r^2 \sum_{j=2}^{10} c_j r^{j-2} + 5 \cdot 10^{-6} \left(\frac{5}{9}\right)^2 r^2 \frac{(5r/9)^9}{1 - 5r/9} \\ &\leq r^2 \sum_{j=2}^{10} c_j r^{j-2} + 1.4 \cdot 10^{-5} \frac{r^2}{9 - 5r}. \end{aligned}$$

As we will restrict our attention to the case $r < 1$, we can replace the estimate on $\psi(r)$ by

$$\psi(r) \leq \Big(\sum_{j=2}^{10} c_j + \frac{1.4 \cdot 10^{-5}}{9-5}\Big) r^2 \leq \frac{r^2}{2}.$$

This completes the proof of the first part of Proposition 3.1.

[1] The reader may be interested in knowing that verifying this induction hypothesis requires knowledge of the first 186,576 coefficients of ϕ.

$c_2 <$ 4.372e-01	$c_3 <$ 4.320e-02	$c_4 <$ 4.928e-03
$c_5 <$ 5.702e-04	$c_6 <$ 7.196e-05	$c_7 <$ 1.095e-05
$c_8 <$ 2.249e-06	$c_9 <$ 3.761e-07	$c_{10} <$ 7.073e-08
$c_{11} <$ 7.458e-09	$c_{12} <$ 1.091e-09	$c_{13} <$ 1.221e-10
$c_{14} <$ 1.549e-11	$c_{15} <$ 1.950e-12	$c_{16} <$ 2.849e-13
$c_{17} <$ 5.197e-14	$c_{18} <$ 8.166e-15	$c_{19} <$ 1.342e-15
$c_{20} <$ 1.520e-16	$c_{21} <$ 1.894e-17	$c_{22} <$ 2.016e-18
$c_{23} <$ 1.484e-19	$c_{24} <$ 1.626e-20	$c_{25} <$ 1.533e-21
$c_{26} <$ 1.383e-22	$c_{27} <$ 1.632e-23	$c_{28} <$ 1.522e-24
$c_{29} <$ 2.457e-25	$c_{30} <$ 2.118e-26	$c_{31} <$ 2.952e-27
$c_{32} <$ 3.418e-28	$c_{33} <$ 3.513e-29	$c_{34} <$ 3.177e-30
$c_{35} <$ 3.107e-31	$c_{36} <$ 2.775e-32	$c_{37} <$ 2.241e-33
$c_{38} <$ 1.701e-34	$c_{39} <$ 1.227e-35	$c_{40} <$ 9.372e-37
$c_{41} <$ 6.903e-38	$c_{42} <$ 5.016e-39	$c_{43} <$ 3.378e-40
$c_{44} <$ 2.713e-41	$c_{45} <$ 2.265e-42	$c_{46} <$ 2.109e-43
$c_{47} <$ 1.936e-44	$c_{48} <$ 1.868e-45	$c_{49} <$ 1.734e-46
$c_{50} <$ 1.511e-47	$c_{51} <$ 1.217e-48	$c_{52} <$ 9.286e-50
$c_{53} <$ 6.715e-51	$c_{54} <$ 4.711e-52	$c_{55} <$ 3.300e-53
$c_{56} <$ 2.311e-54	$c_{57} <$ 1.542e-55	$c_{58} <$ 1.052e-56
$c_{59} <$ 7.838e-58	$c_{60} <$ 5.701e-59	$c_{61} <$ 4.452e-60
$c_{62} <$ 3.834e-61	$c_{63} <$ 3.422e-62	$c_{64} <$ 3.180e-63
$c_{65} <$ 2.873e-64	$c_{66} <$ 2.498e-65	$c_{67} <$ 2.121e-66
$c_{68} <$ 1.759e-67	$c_{69} <$ 1.419e-68	$c_{70} <$ 1.109e-69

Table 2: The leading coefficients of ψ

3.2.2 The normal form

Now, we turn our attention to the existence and size of the normal form $G(\zeta)$. Recall that G is defined recursively by

$$G_i(\zeta) = \{F_i(\zeta + \phi(\zeta))\}_{\mathbb{U}_p} - \sum_{j=1}^{3} \frac{\partial \phi_i}{\partial \zeta_j}(\zeta) G_j(\zeta) \qquad (i = 1, 2, 3). \tag{17}$$

The existence of a formal solution is immediate as we have no divisors whatsoever. Turning to the question of convergence, we use similar majorization techniques as in the previous section. We arrive at the estimate $\|G\|_r \leq \hat{G}(r)$, by considering the following majorant systems together with (17):

$$\tilde{G}(\zeta) = \Big\{\tilde{F}(\zeta + \varphi(\zeta))\Big\}_{\mathbb{U}_p} + \sum_{j=1}^{3} \frac{\partial \tilde{\varphi}}{\partial \zeta_j}(\zeta) \tilde{G}_j(\zeta)$$

(recall that $\varphi = (\tilde{\varphi}, \tilde{\varphi}, \tilde{\varphi})$ and compare with (11)), and

$$\begin{aligned} \Big[\hat{G}(r)\Big]_n &= \Big[\hat{F}(r + \psi(r)) + 3\psi'(r)\hat{G}(r)\Big]_n \\ &= \Big[5\big(r + \psi(r)\big)^2 + 3\psi'(r)\hat{G}(r)\Big]_n \qquad (n \geq 2p), \end{aligned} \tag{18}$$

compare with (13). By the last equation, it is clear that $\hat{G}$ does not contain terms of degree less than $2p$, and that its lowest order term is given by

$$\Big[\hat{G}(r)\Big]_{2p} = \Big[5\big(r + \psi(r)\big)^2\Big]_{2p}. \tag{19}$$

As in the previous section, the recursive equation for $\hat{G}$ can be solved formally by a power series

$$\hat{G}(r) = \sum_{k=20}^{\infty} \hat{g}_k r^k,$$

and we will prove, by induction, that the coefficients of $\hat{G}$ satisfy $\hat{g}_n \leq DK^n$, where we can take the constants $D = 2 \cdot 10^{-18}$ and $K = 3$. This will then immediately give

$$\|G\|_r \leq \hat{G}(r) \leq 2 \cdot 10^{-18} \sum_{k=20}^{\infty} (3r)^k \leq 2 \cdot 10^{-18} \frac{(3r)^{20}}{1-3r} \leq 7 \cdot 10^{-9} \frac{r^{20}}{1-3r},$$

which will complete the proof of Proposition 3.1.

Starting with (19), and using the computed numerical values of $\{c_j\}_1^{19}$, we have (recall that we defined $c_1 = 1$)

$$\hat{g}_{20} r^{20} = \left[\hat{G}(r)\right]_{20} \leq \left[5(r + \psi(r))^2\right]_{20} = 5 \sum_{j=1}^{19} c_j c_{20-j} r^{20} \leq 3 \cdot 10^{-13} r^{20}$$

which clearly satisfies our induction hypothesis, as $3 \cdot 10^{-13} < 2 \cdot 10^{-18} \cdot 3^{20} = DK^{20}$. We now proceed with the induction step: assume that we have proved that $\hat{g}_j \leq DK^k$ for $j = 20, ..., n$. Then, by (18), we have

$$\hat{g}_{n+1} = 5 \sum_{j=1}^{n} c_j c_{n+1-j} + 3 \sum_{j=2}^{n-18} j c_j \hat{g}_{n+2-j} = \Sigma_1 + \Sigma_2,$$

where

$$\begin{aligned}
\Sigma_1 &\leq 5 \left(2 \sum_{j=1}^{10} c_j M^{-j} + (n-20)C \right) CM^{n+1} \\
&\leq \left[5 \left(2 \sum_{j=1}^{10} c_j M^{-j} + (n-20)C \right) \frac{C}{D} \left(\frac{M}{K} \right)^{n+1} \right] DK^{n+1} \\
&\leq \left[5 \left(7.08 + (n-20) \cdot 5 \cdot 10^{-6} \right) \frac{5 \cdot 10^{-6}}{2 \cdot 10^{-18}} \left(\frac{5}{27} \right)^{n+1} \right] DK^{n+1} \\
&\leq 0.0369 DK^{n+1},
\end{aligned}$$

and

$$\begin{aligned}
\Sigma_2 &\leq 3D \left(\sum_{j=2}^{10} j c_j K^{n+2-j} + C \sum_{j=11}^{n-18} j M^j K^{n+2-j} \right) \\
&\leq \left[3 \left(\sum_{j=2}^{10} j c_j K^{1-j} + CM \sum_{j=11}^{\infty} j \left(\frac{M}{K} \right)^{j-1} \right) \right] DK^{n+1} \\
&\leq \left[3 \left(0.31 + \frac{25}{9} \cdot 10^{-6} \cdot \left(\frac{5}{27} \right)^{10} \frac{11 - 50/27}{(1-5/27)^2} \right) \right] DK^{n+1} \\
&\leq 3 \left(0.31 + 2 \cdot 10^{-12} \right) DK^{n+1}.
\end{aligned}$$

Summing up, gives the following estimate

$$\hat{g}_{n+1} \leq 0.97 DK^{n+1},$$

which proves the induction step. This completes the final part of Proposition 3.1.

3.3 The dynamics inside the cube

In this section, we will show that the normal form flow, $\psi(\zeta,t)$, i.e., the solution to the equations $\dot{\zeta} = A\zeta + G(\zeta)$, where G is as in Proposition 3.1, acts much like its linear counterpart used in the geometric model of the Lorenz flow. The geometric model uses the linearity near the origin to obtain estimates on trajectories passing near the origin. These estimates, however, are not valid for the original Lorenz flow without an analog of the change of variables described in Proposition 3.1. Finding an analytic change of variables which completely linearizes the Lorenz equations in a neighbourhood of the origin poses two major difficulties:

- As we must remove all non-linear terms of the vector field, we will encounter all possible divisors, having modulus $|n\lambda - \lambda_i|$, where $n \in \mathbb{N}^3$ and $|n| \geq 2$. These are of course not bounded away from zero, so unless we impose a Diophantine condition[2] on the eigenvalues, the linearizing change of variables will not converge. Even if we manage to get a positive radius of convergence, it is likely that the radius is too small to be of any practical use.
- Although the set of all eigenvalues satisfying any Diophantine condition has full measure, the set of resonant eigenvalues[3] is everywhere dense. As resonant eigenvalues produce vanishing divisors, we must exclude a dense set of parameters to avoid this situation. In doing so, we lose the robustness of our statements.

This is why we bring the Lorenz equations into a carefully selected normal form instead of the linear one. The price we have to pay is that it is a little more difficult to gain estimates on the normal form flow. On the other hand, once we have these estimates, we may effectively link them to the original Lorenz flow by using Proposition 3.1. As mentioned in the outline, we will make the change of coordinates when a trajectory hits the lid of a small cube centered around the origin. Inside the cube, we use our estimates on the normal form flow to find the exit point of the trajectory. We then change back to the original coordinates and carry on with the numerical computations.

3.4 The linear flow

To begin with, we will extract the properties of the linear flow that really are used in the geometric model of the Lorenz flow. Let $\phi(\zeta,t)$ denote the flow of the linearized Lorenz equations $\dot{\zeta} = A\zeta$. These can be solved explicitly: $\phi(\zeta,t) = e^{At}\zeta$, i.e., $\phi_i(\zeta,t) = e^{\lambda_i t}\zeta_i$ $(i = 1,2,3)$. Consider a small cube centered at the origin, with radius r, and take a trajectory starting from the interior of the cube, say at the point ζ, with $|\zeta_1| \neq 0$. Since the eigenvalues satisfy $0 < -\lambda_3 < \lambda_1 < -\lambda_2$, it is plain to see that the trajectory exits the cube when $\phi_1(\zeta,t) = \mathrm{sgn}(\zeta_1)r$. Solving for the exit-time $\tau_e(\zeta)$ gives $\tau_e(\zeta) = \frac{1}{\lambda_1}\log\frac{r}{|\zeta_1|}$. Note that $\lim_{\zeta_1\to 0}\tau_e(\zeta) = \infty$. This is one of the reasons why numeric calculations break down near the origin. Inserting τ_e in the other coordinate functions gives the location of the exit

$$\phi(\zeta,\tau_e(\zeta)) = \left(\mathrm{sgn}(\zeta_1)r, \zeta_2\left(|\zeta_1|/r\right)^{-\frac{\lambda_2}{\lambda_1}}, r\left(|\zeta_1|/r\right)^{-\frac{\lambda_3}{\lambda_1}}\right). \tag{20}$$

Since $-\lambda_2/\lambda_1 \approx 1.93$ and $-\lambda_3/\lambda_1 \approx 0.225$, a line segment lying in the ζ_2 direction will be strongly contracted, whereas a line segment lying in the ζ_1 direction will have expanded on its exit. An explicit calculation gives

$$\frac{\phi_2(\zeta,t)}{\phi_3(\zeta,t)} = \frac{\zeta_2}{r}e^{(\lambda_2-\lambda_3)t}.$$

Therefore, since $\lambda_2 - \lambda_3 < 0$, the lid of the cube will exit as two cusp shaped regions (in Figure 4 one of the cusps is illustrated). In the cube, the $\zeta_1\zeta_3$-plane acts as a separatrix, and all trajectories approach this exponentially fast.

[2] We say that the eigenvalues $\lambda = (\lambda_1,\lambda_2,\lambda_3)$ satisfy a Diophantine condition of type (κ,τ) if there exists positive κ,τ such that for $i = 1,2,3$, we have $|n\lambda - \lambda_i| \geq \kappa|n|^{-\tau}$ for all $n \in \mathbb{N}^3$ with $|n| \geq 2$.

[3] We say that the eigenvalues $\lambda = (\lambda_1,\lambda_2,\lambda_3)$ are resonant if there exists an $i \in \{1,2,3\}$ such that $|n\lambda - \lambda_i| = 0$ for some $n \in \mathbb{N}^3$ with $|n| \geq 2$.

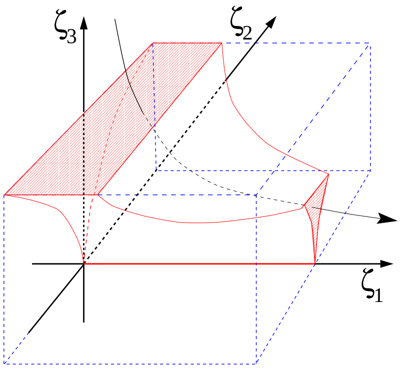

Figure 4: The linear flow inside the cube for $\zeta_1 > 0$.

We are also interested in the evolution of tangent vectors following a trajectory inside the cube. Since $\phi(\zeta,t) = e^{At}\zeta$, it is clear that $D\phi(\zeta,t) = e^{At}$. Hence we have

$$D\phi(\zeta,\tau_e(\zeta)) = \begin{pmatrix} (|\zeta_1|/r)^{-1} & 0 & 0 \\ 0 & (|\zeta_1|/r)^{-\frac{\lambda_2}{\lambda_1}} & 0 \\ 0 & 0 & (|\zeta_1|/r)^{-\frac{\lambda_3}{\lambda_1}} \end{pmatrix}.$$

Any three-dimensional cone field centered around $e_1 = (1,0,0)$ is taken into itself under $D\phi$. This is due to the fearsome hyperbolicity (expansion and contraction) $D\phi$ exhibits near the origin. This property and (20), which gives rise to the cusps, are the two features used in the geometric models.

3.5 The normal form flow

Our main objective now is to show that the normal form flow and the linear flow have very similar behaviour. Although we use a cube having radius $\frac{1}{10}$ in the computer program, we do all estimates for a radius of $\frac{1}{4}$. This allows for a slight deformation of the cube as we change coordinates, see Lemma 3.4. We will need the following estimate on G:

Lemma 3.8 *In the cube $\{\zeta\colon |\zeta| \le \frac{1}{4}\}$, we have*

$$|G(\zeta)| \le 3\cdot 10^{-8}|\zeta_1|^{10}\max\{|\zeta_2|^{10},|\zeta_3|^{10}\}.$$

Proof: Recall that we arranged for $G(\zeta) \in \mathcal{O}^{10}(\zeta_1)\cap\mathcal{O}^{10}(\zeta_2,\zeta_3)$. This means that for any term $g_{i,n}\zeta^n$ of G_i, there exist $\tilde{n} \in \mathbb{N}^3$ and $k \in [0,10]$ such that we can factor the term as

$$g_{i,n}\zeta^n = g_{i,n}\zeta_1^{10}\zeta_2^{10-k}\zeta_3^k\zeta^{\tilde{n}}. \tag{21}$$

The estimate on G in Proposition 3.1 implies that $|G(\zeta)| \le 3\cdot 10^{-8}|\zeta|^{20}$ for $|\zeta| \le \frac{1}{4}$, but we also have some additional information: since the estimate was obtained by using majorants, it is also valid for the majorants themselves. The estimate is naturally valid for all smaller majorants of G than the ones we used and, in particular, for the smallest majorants of G. Let

$$G_i(\zeta) = \sum_{\substack{n_1\ge 10\\ n_2+n_3\ge 10}} g_{i,n}\zeta^n, \quad m(\zeta) = \sum_{\substack{n_1\ge 10\\ n_2+n_3\ge 10}} m_n\zeta^n, \quad \text{and} \quad M(r) = \sum_{k\ge 20} M_k r^k,$$

where $m_n = \max_{i=1,2,3}|g_{i,n}|$, and $M_k = \sum_{|n|=k} m_n$. Then clearly $G_i \prec m \prec M$, and the functions m and M are the smallest majorants of G on their corresponding levels. By levels we mean the following: we may view G as a function $G\colon \mathbb{C}^3 \to \mathbb{C}^3$. The first level majorant m can be viewed as a function $m\colon \mathbb{C}^3 \to \mathbb{C}$, and the second level majorant M can be viewed as a function $M\colon \mathbb{C} \to \mathbb{C}$.

Thus the majorants not only provide bounds, they also successively reduce the dimension of the domain and range of the original function G. Both m and M are the smallest functions, on their respective levels, that majorize G.

Now, set $\omega = |\zeta_1| \max\{|\zeta_2|, |\zeta_3|\}$, and suppose that $|\zeta| \leq \frac{1}{4} = r$. We have

$$\begin{aligned}
|G_i(\zeta)| &= \Big| \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} g_{i,n}\zeta^n \Big| \leq \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} |g_{i,n}||\zeta_1|^{n_1}|\zeta_2|^{n_2}|\zeta_3|^{n_3} \\
&\leq \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} m_n|\zeta_1|^{n_1}|\zeta_2|^{n_2}|\zeta_3|^{n_3} \leq \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} m_n|\zeta_1|^{n_1}\big(\max\{|\zeta_2|, |\zeta_3|\}\big)^{n_2+n_3} \\
&\leq |\zeta_1|^{10}\big(\max\{|\zeta_2|, |\zeta_3|\}\big)^{10} \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} m_n|\zeta_1|^{n_1-10}\big(\max\{|\zeta_2|, |\zeta_3|\}\big)^{n_2+n_3-10} \\
&\leq \omega^{10} \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} m_n\big(\max\{|\zeta_1|, |\zeta_2|, |\zeta_3|\}\big)^{n_1+n_2+n_3-20} \\
&= \omega^{10} \sum_{\substack{n_1 \geq 10 \\ n_2+n_3 \geq 10}} m_n|\zeta|^{|n|-20} \leq \omega^{10} \sum_{k \geq 20} \Big(\sum_{|n|=k} m_n \Big) r^{k-20} = \omega^{10} \sum_{k \geq 20} M_k r^{k-20} \\
&= \omega^{10} r^{-20} \sum_{k \geq 20} M_k r^k \leq \omega^{10} r^{-20} \cdot 3 \cdot 10^{-8} r^{20} = 3 \cdot 10^{-8} \omega^{10},
\end{aligned}$$

which completes the proof. □

3.5.1 C^0-properties

Starting with the topological properties, we will examine how the orbits of the linear and normal form flows differ. Using Lemma 3.8, we shall first prove that the modulus of the ζ_1-component of the normal form flow is monotonically increasing.

Lemma 3.9 *In the cube $\{\zeta \colon |\zeta| \leq r\}$, where $r \leq \frac{1}{4}$, we have*

$$(\lambda_1 - \kappa)\psi_1(\zeta, t) \leq \dot{\psi}_1(\zeta, t) \leq (\lambda_1 + \kappa)\psi_1(\zeta, t)$$

and

$$|\zeta_1|e^{(\lambda_1 - \kappa)t} \leq |\psi_1(\zeta, t)| \leq |\zeta_1|e^{(\lambda_1 + \kappa)t},$$

where $\kappa = 2 \cdot 10^{-19}$.

Proof: We just have to check the differential equation for the ζ_1-component of the normal form flow: $\dot{\psi}_1(\zeta, t) = \lambda_1\psi_1(\zeta, t) + G_1(\psi(\zeta, t))$. By Lemma 3.8,

$$\begin{aligned}
|\dot{\psi}_1(\zeta, t) - \lambda_1\psi_1(\zeta, t)| &= |G_1(\psi(\zeta, t))| \\
&\leq 3 \cdot 10^{-8}|\psi_1(\zeta, t)|^{10} \max\{|\psi_2(\zeta, t)|^{10}, |\psi_3(\zeta, t)|^{10}\} \\
&\leq 3 \cdot 10^{-8} \cdot 4^{-10}|\psi_1(\zeta, t)|^{10} \leq 3 \cdot 10^{-8} \cdot 4^{-19}|\psi_1(\zeta, t)|.
\end{aligned}$$

Setting $\kappa = 2 \cdot 10^{-19} > 3 \cdot 10^{-8} \cdot 4^{-19}$ completes the proof. □

Next, we prove that the ζ_3-component of the normal form flow dominates the ζ_2-component.

Lemma 3.10 *For all trajectories starting from the lid of the cube $\{\zeta \colon |\zeta| \leq r\}$, where $r \leq \frac{1}{4}$, we have $\psi_3(\zeta, t) \geq |\psi_2(\zeta, t)|$. These estimates hold throughout the cube.*

Proof: The differential equations for ψ_2 and ψ_3 can be expressed as differential inequalities:

$$\begin{aligned}
|\dot{\psi}_2(\zeta, t) - \lambda_2\psi_2(\zeta, t)| &= |G_2(\psi(\zeta, t))| \leq 3 \cdot 10^{-8} \cdot 4^{-10} \max\{|\psi_2(\zeta, t)|^{10}, |\psi_3(\zeta, t)|^{10}\} \\
|\dot{\psi}_3(\zeta, t) - \lambda_3\psi_3(\zeta, t)| &= |G_3(\psi(\zeta, t))| \leq 3 \cdot 10^{-8} \cdot 4^{-10} \max\{|\psi_2(\zeta, t)|^{10}, |\psi_3(\zeta, t)|^{10}\}.
\end{aligned}$$

Initially, we have $\frac{1}{4} = \psi_3(\zeta, 0) \geq |\psi_2(\zeta, 0)|$, and by the differential inequalities it is clear that, if $\psi_3(\zeta, 0) = |\psi_2(\zeta, 0)|$, then $|\psi_2(\zeta, t)|$ decreases faster than $\psi_3(\zeta, t)$, since $\lambda_2 < \lambda_3 < 0$. So suppose that after some time t^* we have $\psi_3(\zeta, t^*) = |\psi_2(\zeta, t^*)|$. Then we can rewrite the differential inequalities as

$$|\dot{\psi}_2(t^*, \zeta) - \lambda_2\psi_2(t^*, \zeta)| \leq 3 \cdot 10^{-8} \cdot 4^{-10}|\psi_2(t^*, \zeta)|^{10} \leq \kappa|\psi_2(t^*, \zeta)|$$
$$|\dot{\psi}_3(t^*, \zeta) - \lambda_3\psi_3(t^*, \zeta)| \leq 3 \cdot 10^{-8} \cdot 4^{-10}|\psi_3(t^*, \zeta)|^{10} \leq \kappa|\psi_3(t^*, \zeta)|,$$

and, again, $|\psi_2(t^*, \zeta)|$ decreases faster than $\psi_3(t^*, \zeta)$. Hence $\psi_3(\zeta, t) \geq |\psi_2(\zeta, t)|$ throughout the whole cube. □

It is now easy to show that the ζ_3-component of the flow is monotonically decreasing:

Lemma 3.11 *For all trajectories starting from the lid of the cube $\{\zeta\colon |\zeta| \leq r\}$, where $r \leq \frac{1}{4}$, we have*

$$(\lambda_3 - \kappa)\psi_3(\zeta, t) \leq \dot{\psi}_3(\zeta, t) \leq (\lambda_3 + \kappa)\psi_3(\zeta, t)$$

and

$$re^{(\lambda_3 - \kappa)t} \leq \psi_3(\zeta, t) \leq re^{(\lambda_3 + \kappa)t},$$

where $\kappa = 2 \cdot 10^{-19}$.

Proof: Using Lemma 3.10, we just copy the proof of Lemma 3.9 by changing the roles of $\psi_1(\zeta, t)$ and $\psi_3(\zeta, t)$. □

Now, given an initial point in the lid of the cube $\{\zeta\colon |\zeta| \leq r\}$ we know that $|\psi_1(\zeta, t)|$ is increasing, whereas $\psi_3(\zeta, t)$ is decreasing and majorizing $|\psi_2(\zeta, t)|$. Since $\lambda_1 > |\lambda_3|$, we also know that $|\psi_1(\zeta, t)|$ increases faster than $\psi_3(\zeta, t)$ decreases. Therefore, it is clear that the orbit exits the cube through one of the sides $\{\zeta\colon |\zeta_1| = r, |\zeta_2|, |\zeta_3| \leq r\}$. A trivial calculation enables us to control the exit-time of the normal form flow:

Lemma 3.12 *For all trajectories starting from the lid of the cube $\{\zeta\colon |\zeta| \leq r\}$, where $r \leq \frac{1}{4}$, the normal form flow exits the cube at time $\tau_e(\zeta)$, where*

$$\frac{1}{\lambda_1 + \kappa} \log \frac{r}{|\zeta_1|} \leq \tau_e(\zeta) \leq \frac{1}{\lambda_1 - \kappa} \log \frac{r}{|\zeta_1|},$$

and $\kappa = 2 \cdot 10^{-19}$.

Proof: We just have to solve $|\psi_1(\zeta, \tau_e)| = r$ for τ_e. Using Lemma 3.9, we get

$$|\zeta_1|e^{(\lambda_1 - \kappa)\tau_e} \leq r \leq |\zeta_1|e^{(\lambda_1 + \kappa)\tau_e},$$

which immediately gives the desired result. □

Note that we have $\lim_{\zeta_1 \to 0} \tau_e(\zeta) = \infty$, just as in the linear case.

Turning to the ζ_2-component of the normal form flow, we have the following differential inequality

$$|\dot{\psi}_2(\zeta, t) - \lambda_2\psi_2(\zeta, t)| \leq 3 \cdot 10^{-8} \cdot r^{10}|\psi_3(\zeta, t)|^{10}$$

for all trajectories starting from the lid of the cube $\{\zeta\colon |\zeta| \leq r\}$. This does not imply that $|\psi_2(\zeta, t)|$ decreases exponentially, so there is a slight discrepancy between the normal form and linear flows in this sense. However, just as in the linear case, there exists a surface acting as a separatrix. It is a slight deformation of the $\zeta_1\zeta_3$-plane, and all orbits tend to this separatrix exponentially fast. The important property of both the normal form and linear flows is that the quotient of the ζ_2- and ζ_3-components tends to zero exponentially fast with respect to the exit-time. This gives rise to the nice cusp shaped image of the cube's lid, as illustrated in Figure 4.

Lemma 3.13 *For all trajectories starting from the lid of the cube $\{\zeta\colon |\zeta| \le r\}$, where $r \le \frac{1}{4}$, we have*

$$\big(\zeta_2 - \kappa r(1-e^{-3t})\big)e^{\lambda_2 t} \le \psi_2(\zeta,t) \le \big(\zeta_2 + \kappa r(1-e^{-3t})\big)e^{\lambda_2 t},$$

where $\kappa = 2\cdot 10^{-19}$. These estimates hold throughout the cube.

Proof: Using the second part of Lemma 3.11, we can rewrite the differential inequality for $\dot{\psi}_2(\zeta,t)$ as

$$|\dot{\psi}_2(\zeta,t) - \lambda_2\psi_2(\zeta,t)| \le 3\cdot 10^{-8}\cdot 4^{-19} r e^{10(\lambda_3+\kappa)t} \le \kappa r e^{10(\lambda_3+\kappa)t}.$$

A straight-forward calculation yields the following bounds on $\psi_2(\zeta,t)$:

$$|\psi_2(\zeta,t) - \zeta_2 e^{\lambda_2 t}| \le \frac{\kappa r\big(e^{\lambda_2 t} - e^{10(\lambda_3+\kappa)t}\big)}{\big||\lambda_2| + 10(\lambda_3+\kappa)\big|} \le \kappa r(1-e^{-3t})e^{\lambda_2 t},$$

using the fact that $10(\lambda_3+\kappa) - \lambda_2 < -3$. □

Combining this lemma with the second part of Lemma 3.11 gives

$$\left|\frac{\psi_2(\zeta,t)}{\psi_3(\zeta,t)}\right| \le \frac{|\zeta_2| + \kappa r(1-e^{-3t})}{r e^{(\lambda_3-\kappa)t}} e^{\lambda_2 t} \le (1+\kappa)e^{(\lambda_2-\lambda_3+\kappa)t},$$

which proves that the lid of the cube will exit as two cusp shaped region.

Remark: Had we flattened out the invariant manifolds to an order $p < 9$, the situation would be slightly different: We would then have $p(\lambda_3+\kappa) - \lambda_2 > 0$, which would result in a blunter cusp.

Combining and summarizing the results of this section, we achieve very tight bounds on the trajectories leaving the cube.

Lemma 3.14 *For all trajectories starting from the lid of the cube $\{\zeta\colon |\zeta| \le r\}$, where $r \le \frac{1}{4}$, we have the following interval-valued enclosures:*

$$\psi_2(\zeta,\tau_e(\zeta)) \in [\zeta_2]\Big(\tfrac{|\zeta_1|}{r}\Big)^{\frac{|[\lambda_2]|}{[\lambda_1]}}; \qquad \psi_3(\zeta,\tau_e(\zeta)) \in r\Big(\tfrac{|\zeta_1|}{r}\Big)^{\frac{|[\lambda_3]|}{[\lambda_1]}},$$

where $[\lambda_i] = [\lambda_i - \kappa, \lambda_i + \kappa]$, $[\zeta_2] = [\zeta_2 - \kappa, \zeta_2 + \kappa]$, and $\kappa = 2\cdot 10^{-19}$.

For a precise definition of interval-valued enclosures, see Sections 4.2 and 4.3.

3.5.2 C^1-properties

We will now prove that, in our small cube, the normal form flow expands and contracts tangent vectors at almost the same rate as the linear flow does. The first variational equations for the normal form flow are

$$\frac{d}{dt}D\psi(\zeta,t) = AD\psi(\zeta,t) + DG(\psi(\zeta,t))D\psi(\zeta,t), \qquad D\psi(\zeta,0) = I,$$

and the solution is formally given by

$$D\psi(\zeta,t) = e^{At}D\psi(\zeta,0) + \int_0^t e^{A(t-s)}DG(\psi(\zeta,s))D\psi(\zeta,s)ds.$$

We will prove the following lemma, which states that $D\psi$ is virtually linear:

Lemma 3.15 *For all trajectories starting from the lid of the cube $\{\zeta\colon |\zeta| \le r\}$, where $r \le \frac{1}{4}$, we have*

$$\left|\frac{\partial\psi_i}{\partial\zeta_j}(\zeta,t) - \frac{\partial\phi_i}{\partial\zeta_j}(\zeta,t)\right| \le \kappa e^{[9(\lambda_3+\kappa)+\lambda_j]t} \qquad (i,j = 1,2,3),$$

where $\kappa = 2\cdot 10^{-19}$. These estimates hold throughout the cube.

Remark: The key point is that $9(\lambda_3+\kappa)+\lambda_j$ is negative. This means that the error decreases as the exit-time increases, i.e., as we take $|\zeta_1|$ small. This is expected seeing how we constructed G: the perturbation of the linearized equations is small enough to completely dampen out the bad effects due to having to spend a long time near the origin.

Proof: Since $D\psi(\zeta,0)=D\phi(\zeta,0)=I$, the inequality is trivially true for small t, say for $t\in[0,t^*]$. We will prove that we can take $t^*=\tau_e(\zeta)$, which will prove the lemma. For $t\in[0,t^*]$, we have

$$\begin{aligned}|D\psi(\zeta,t)-D\phi(\zeta,t)| &= \left|\int_0^t e^{A(t-s)}DG(\psi(\zeta,s))D\psi(\zeta,s)ds\right| \\ &\leq \int_0^t \left|e^{A(t-s)}DG(\psi(\zeta,s))D\phi(\zeta,s)\right| ds \\ &+ \int_0^t \left|e^{A(t-s)}DG(\psi(\zeta,s))\big(D\psi(\zeta,s)-D\phi(\zeta,s)\big)\right| ds.\end{aligned} \tag{22}$$

By using the facts that A is diagonal, and $D\phi(\zeta,s)=e^{As}$, a simple calculation gives that

$$\left|\left\{e^{A(t-s)}DG(\zeta)e^{As}\right\}_{ij}\right| = e^{\lambda_i(t-s)}e^{\lambda_j s}\left|\frac{\partial G_i}{\partial\zeta_j}(\zeta)\right| \qquad (i,j=1,2,3).$$

If we set $E(\zeta,t)=D\psi(\zeta,t)-D\phi(\zeta,t)$, we can treat each matrix entry separately: For $i,j=1,2,3$, we have

$$\begin{aligned}|E_{ij}(\zeta,t)| \leq \int_0^t e^{\lambda_i(t-s)}e^{\lambda_j s}\left|\frac{\partial G_i}{\partial\zeta_j}(\psi(\zeta,s))\right| ds \\ + \int_0^t e^{\lambda_i(t-s)}\sum_{k=1}^{3}\left|\frac{\partial G_i}{\partial\zeta_k}(\psi(\zeta,s))E_{kj}(\zeta,s)\right| ds.\end{aligned}$$

To carry on further, it is clear that we need some estimates on the partial derivatives of G.

Lemma 3.16 *In the cube $\{\zeta\colon |\zeta|\leq\frac{1}{4}\}$, we have*

$$\left|\frac{\partial G_i}{\partial\zeta_j}(\zeta)\right| \leq 9\cdot10^{-8}|\zeta_1|^9\max\{|\zeta_2|^9,|\zeta_3|^9\}.$$

Proof: By Proposition 3.1, the functions G_i $(i=1,2,3)$ can be extended to a ball in $\mathbb{C}^3$ such that they are analytic in each of the three variables ζ_i $(i=1,2,3)$. Furthermore, for $|\zeta|\leq\frac{3}{10}$, the functions satisfy

$$|G_i(\zeta)| \leq \frac{7\cdot10^{-9}|\zeta|^{20}}{1-\frac{9}{10}} = 7\cdot10^{-8}|\zeta|^{20},$$

and the argument used in the proof of Lemma 3.2 shows that in the complex ball $|\zeta|\leq\frac{1}{4}$, we have the estimate

$$\left|\frac{\partial G_i}{\partial\zeta_j}(\zeta)\right| \leq \frac{7\cdot10^{-8}|\zeta|^{20}}{\frac{3}{10}-\frac{1}{4}} = 1.4\cdot10^{-6}|\zeta|^{20} < 9\cdot10^{-8}|\zeta|^{18} \qquad (i,j=1,2,3). \tag{23}$$

Since we arranged for $G(\zeta)\in\mathcal{O}^{10}(\zeta_1)\cap\mathcal{O}^{10}(\zeta_2,\zeta_3)$, we know that terms of the partial derivatives of G_i belong to $\mathcal{O}^9(\zeta_1)\cap\mathcal{O}^9(\zeta_2,\zeta_3)$. Thus, for any term $g_{i,n}\zeta^n$ of a partial derivative of G_i, there exists $\tilde{n}\in\mathbb{N}^3$ and $k\in[0,9]$ such that we can factor the term as

$$g_{i,n}\zeta^n = g_{i,n}\zeta_1^9\zeta_2^{9-k}\zeta_3^k\zeta^{\tilde{n}}. \tag{24}$$

Applying the methods used in the proof of Lemma 3.8 combined with the estimate (23) gives the desired result. □

From the previous section, we know that $|\psi_2(\zeta,t)| \le \psi_3(\zeta,t)$ for all trajectories starting from the lid of the cube. Using this with Lemmas 3.9 and 3.11, gives

$$\left|\frac{\partial G_i}{\partial \zeta_j}(\psi(\zeta,t))\right| \le 9\cdot 10^{-8}|\psi_1(\zeta,t)\psi_3(\zeta,t)|^9 \le 4\cdot 10^{-13}|\zeta_1|^9 e^{9(\lambda_1+\lambda_3+2\kappa)t}.$$

We are now prepared to continue our estimates. We will use the facts that $|\zeta_1 e^{(\lambda_1+\kappa)t}| \le r \le \frac{1}{4}$ and $|E_{kj}(\zeta,s)| \le \kappa e^{[9(\lambda_3+\kappa)+\lambda_j]s} \le \kappa e^{\lambda_j s}$. For $i,j = 1,2,3$, we have

$$\begin{aligned}
\big|E_{ij}(\zeta,t)\big| \le & \int_0^t e^{\lambda_i(t-s)}e^{\lambda_j s}\left|\frac{\partial G_i}{\partial \zeta_j}(\psi(\zeta,s))\right| ds \\
& \qquad + \int_0^t e^{\lambda_i(t-s)}\sum_{k=1}^{3}\left|\frac{\partial G_i}{\partial \zeta_k}(\psi(\zeta,s))\right| \kappa e^{\lambda_j s} ds \\
& \le 4\cdot 10^{-13}|\zeta_1|^9(1+3\kappa)e^{\lambda_i t}\int_0^t e^{[9(\lambda_1+\lambda_3+2\kappa)+\lambda_j-\lambda_i]s} ds \\
& = 4\cdot 10^{-13}|\zeta_1|^9(1+3\kappa)e^{\lambda_i t}\left[\frac{e^{[9(\lambda_1+\lambda_3+2\kappa)+\lambda_j-\lambda_i]s}}{9(\lambda_1+\lambda_3+2\kappa)+\lambda_j-\lambda_i}\right]_0^t \\
& \le \frac{4\cdot 10^{-13}|\zeta_1|^9(1+3\kappa)}{9(\lambda_1+\lambda_3)+\lambda_2-\lambda_1} e^{[9(\lambda_1+\lambda_3+2\kappa)+\lambda_j]t} \\
& \le \frac{4\cdot 10^{-13}(1+3\kappa)}{47}|\zeta_1 e^{(\lambda_1+\kappa)t}|^9 e^{[9(\lambda_3+\kappa)+\lambda_j]t} \\
& \le \frac{4\cdot 10^{-13}(1+3\kappa)}{47\cdot 4^9} e^{[9(\lambda_3+\kappa)+\lambda_j]t} \le \kappa e^{[9(\lambda_3+\kappa)+\lambda_j]t},
\end{aligned}$$

for $\kappa = 2\cdot 10^{-19}$. This completes the proof, since all estimates are valid in the cube. $\square$

4 Rigorous numerics - details

We will now outline the main underlying algorithms used to compute the return map of the Lorenz flow. These algorithms are extremely general, and apply to virtually any vector field in any dimension. We should, however, point out that the algorithms used for following the cone fields have only been developed for two-dimensional cones.

The real implementations of the algorithms differ slightly from what will be presented below. This is because the program code has been optimized to minimize the number of floating point operations, and an exact listing would be both tedious and confusing. Nonetheless, the underlying mathematical reasoning is the same, and any interested reader can study the actual source code which is available at `http://www.math.uu.se/~warwick/papers.html`.

Recall that the algorithms described below are used *outside* the small cube centered at the origin. Whenever a trajectory hits the lid of the cube, the program computes the image of the trajectory leaving the cube (see Section 4.9) before resuming the main algorithm. This "cube-part" of the program is strictly three-dimensional, and relies heavily on the computations performed in Chapter 3. A higher dimensional saddle fixed point would certainly increase the complexity of the computations performed in Chapter 3, but should in principle not introduce any additional difficulties. Regardless of the dimension, these computations are performed a priori, and the necessary constants are hand coded into the program. Thus this part of the code depends on the particular vector field at hand, as opposed to the part that computes the enclosures of the actual trajectories and the partial derivatives along them.

4.1 Ordinary differential equations

As a model problem, we will consider the general initial value problem:

$$\dot{x} = f(x); \qquad x(0) = x^{(0)}, \tag{25}$$

where $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$. We will denote the solution of (25) by $\varphi(x,t)$, with $\varphi(x,0) = x(0)$. For real-valued functions, this setting is classical and much studied in standard text books on ordinary differential equations. It is, however, not difficult to find situations where having a whole set of initial values is natural. Indeed, any model of a physical system always has some uncertainty concerning the measured initial values. Furthermore, we are seldom sure of the exact appearance of the vector field modeling our system. The natural thing do is to enclose the initial value x_0 in a box $[x_0]$ whose side lengths reflect the maximal error made in the measurements of the initial data, and to replace f in (25) by a function F, whose components are interval valued and contain the values of f. The problem we then face is to find the solution of the following system:

$$\dot{x} \in F([x]); \qquad x(0) \in [x^{(0)}], \tag{26}$$

Our objective is to compute a set that is guaranteed to contain all the solutions of (26) with respect to some given stopping condition. Before presenting such algorithms solving (26), we will outline the basics of interval arithmetic.

4.2 Interval arithmetic

In this section, we will briefly describe the fundamentals of interval arithmetic. For a concise reference on this topic, see [14], [15].

Let $\mathbb{IR}$ denote the set of all closed intervals of the real line. For any element $[a] \in \mathbb{IR}$, we adapt the notation $[a] = [\underline{a}, \overline{a}]$. If $\odot$ is one of the operators $+, -, \times, \div$, we define arithmetic operations on elements of $\mathbb{IR}$ by

$$[a] \odot [b] = \{a \odot b \colon a \in [a], b \in [b]\},$$

except that $[a] \div [b]$ is undefined if $0 \in [b]$. Working exclusively with closed intervals, we can describe the resulting interval in terms of the endpoints of the operands:

$$\begin{aligned}
[a] + [b] &= [\underline{a} + \underline{b}, \overline{a} + \overline{b}] \\
[a] - [b] &= [\underline{a} - \overline{b}, \overline{a} - \underline{b}] \\
[a] \times [b] &= [\min\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}, \max\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}] \\
[a] \div [b] &= [a] \times [1/\overline{b}, 1/\underline{b}], \quad \text{if } 0 \notin [b].
\end{aligned}$$

For practical implementations, speed is important. Therefore it is customary to break the formula for multiplication into nine cases (depending of the signs of the endpoints), where only one case involves more than two multiplications. Moreover, when computing with finite precision, the formula for division can be modified for improved accuracy, and directed rounding must be taken into account, see e.g. [2], [14], [15].

It follows immediately from the definitions that addition and multiplication are both associative and commutative. The distributive law, however, does *not* always hold. As an example, we have

$$[-1, 1]([-1, 0] + [3, 4]) = [-1, 1][2, 4] = [-4, 4]$$

whereas

$$[-1, 1][-1, 0] + [-1, 1][3, 4] = [-1, 1] + [-4, 4] = [-5, 5].$$

This unusual property is important to keep in mind when representing functions as part of a program. Interval arithmetic satisfies a weaker rule than the distributive law, which we shall refer to as *sub-distributivity*:

$$[a]([b] + [c]) \subseteq [a][b] + [a][c].$$

Another key feature of interval arithmetic is that it is *inclusion monotonic*, i.e., if $[a] \subseteq [a']$, and $[b] \subseteq [b']$, then

$$[a] \odot [b] \subseteq [a'] \odot [b'],$$

where we demand that $0 \notin [b']$ for division. This is the single most important property of interval arithmetic, it allows us to prove open conditions in a robust way.

We can turn $\mathbb{IR}$ into a metric space by equipping it with the Hausdorff distance:

$$d([a], [b]) = \max\{|\underline{a} - \underline{b}|, |\overline{a} - \overline{b}|\}.$$

For dealing with higher dimensional problems, we define the arithmetic operations to be carried out component-wise. We then talk about an *interval vector* or, more simply, a *box*. The metric on the space $\mathbb{IR}^n$ is defined by

$$d([a], [b]) = \max_{1 \le i \le n} \{d([a_i], [b_i])\}.$$

Matrix operations are defined analogously to the real case.

When implementing interval arithmetic on a computer, we no longer work over the space $\mathbb{R}$, but rather $\mathbb{F}$ - the floating points of the computer. This is a finite set, and thus so is the set of all intervals with floating point endpoints $\mathbb{IF}$. When performing arithmetic on intervals in $\mathbb{F}$ we must round the resulting interval outwards to guarantee inclusion of the true result. This is because, fixing the set $\mathbb{F}$, the sum of two floating points may not be a floating point. The same holds for the other arithmetic operations. As an example, adding two intervals $[a], [b] \in \mathbb{IF}$, becomes

$$[a] + [b] = [\downarrow \underline{a} + \underline{b} \downarrow, \uparrow \overline{a} + \overline{b} \uparrow],$$

where $\downarrow x \downarrow$ is the largest floating point in $\mathbb{F}$ that is strictly less than x (called x rounded down), and $\uparrow x \uparrow$ is the smallest floating point in $\mathbb{F}$ that is strictly greater than x (called x rounded up). This type of arithmetic is called interval arithmetic with *directed rounding*.

4.3 Interval-valued functions

Consider a function $f\colon \mathbb{R}^n \to \mathbb{R}^n$. Given a box $[a]$ we define the *range* of f over $[a]$ by

$$R(f;[a]) = \{f(x)\colon x \in [a]\}.$$

As mentioned earlier, it is often desirable in applications to exchange the function f for an *interval extension* F.

Definition 4.1 *A function $F\colon \mathbb{IR}^n \to \mathbb{IR}^n$ is an interval extension of $f\colon \mathbb{R}^n \to \mathbb{R}^n$ if for all boxes $[x] \in \mathbb{IR}^n$, we have $R(f;[x]) \subseteq F([x])$.*

It should be noted that many interval extensions are possible for a real-valued function f. As an example, given $f(x,y) = 2x - \pi y$, both $F_1(x,y) = [1.99, 2.01]x - [3.14, 3.15]y$ and $F_2(x,y) = [1.9, 2.1]x - [3.1, 3.2]y$ are interval extension of f. From now on, any expression on the form $f([x])$ should always be interpreted as to be an interval extension of f evaluated in interval arithmetic. When we are interested in the range of (the $\mathbb{R}^n$-valued) f over $[x]$, we explicitly state so.

If we fix a representation of f (which we also denote f), and evaluate it in interval arithmetic, we always have

$$R(f;[a]) \subseteq f([a]),$$

due to the inclusion monotonic property. From this property, it also follows that by splitting the box $[a]$ into smaller pieces $[a_0], \ldots, [a_n]$, we have

$$R(f;[a]) \subseteq \bigcup_{i=0}^{n} f([a_i]) \subseteq f([a]).$$

It is clear that, by splitting $[a]$ into many small pieces, we can approximate the true range of f over $[a]$ with any desired accuracy. If, however, f is differentiable, then there are better ways to approximate the range of f: let $m([a])$ denote the midpoint of $[a]$. By the Mean Value Theorem, we have the following relation:

$$R(f;[a]) \subseteq f_{MV}([a]) := f(m([a])) + Df([a])([a] - m([a])).$$

Let $\|[a]\|$ denote the maximal diameter of $[a]$. It is easy to show that

$$d(R(f;[a]), f([a])) = \mathcal{O}(\|[a]\|),$$

whereas

$$d(R(f;[a]), f_{MV}([a])) = \mathcal{O}(\|[a]\|^2).$$

It is obvious that the latter version is preferred, seeing that we have a quadratically small error. This assumes, however, that we only deal with intervals of small widths. The most fundamental part of our algorithm – the partitioning process – guarantees that this indeed will be the case, and thus allows us to attain a quadratic approximation of the vector field range $R(f;[a])$.

For computer applications, we consider interval extensions $F\colon \mathbb{IF}^n \to \mathbb{IF}^n$, and perform all operations with directed rounding.

4.4 The Euler method for a time-t map

The solution of (25) is formally given by

$$\varphi(x, t^{(k+1)}) = \varphi(x, t^{(k)}) + \int_{t^{(k)}}^{t^{(k+1)}} f(\varphi(x,s))ds, \tag{27}$$

where $\varphi(x, t^{(0)}) = x^{(0)}$. Approximating the integrand in (27) by $f(\varphi(x, t^{(k)}))$, we arrive at the classical Euler method, which gives the iterative scheme

$$x^{(k+1)} = x^{(k)} + \Delta t^{(k)} f(x^{(k)}) \qquad k \geq 0$$

for an approximate solution to (25), i.e., $x^{(k)} \approx \varphi(x, t^{(k)})$. Here we have used the notation $\Delta t^{(k)} = t^{(k+1)} - t^{(k)}$. The error we are making is in assuming that the vector field f is constant over each time step. With interval arithmetic this can be overcome by using the following algorithm:

Algorithm 1. *For $k \geq 0$ do the following:*

1. *Enclose the computed solution at step k in a box: $[x^{(k)}] \subset [\tilde{x}^{(k)}]$;*
2. *Compute a time step $\Delta t^{(k)}$ such that $[x^{(k)}] + \Delta t^{(k)} f([\tilde{x}^{(k)}]) \subseteq [\tilde{x}^{(k)}]$;*
3. *If $t^{(k)} + \Delta t^{(k)} > T$, set $\Delta t^{(k)} = T - t^{(k)}$;*
4. *Set $[x^{(k+1)}] = [x^{(k)}] + \Delta t^{(k)} f([\tilde{x}^{(k)}])$, and $t^{(k+1)} = t^{(k)} + \Delta t^{(k)}$;*
5. *If $t^{(k+1)} = T$, break.*

This algorithm produces a box-valued solution that is guaranteed to contain the true solution of the time-t map, i.e., $\varphi([x^{(0)}], t^{(k)}) \subseteq [x^{(k)}]$. As promised above, it also covers the case when the initial value is a solid box, rather than simply being a point. The reason why this method works is that, instead of evaluating the vector field at a single point $x^{(k)}$, we evaluate over a whole box which is constructed to contain all trajectories between $[x^{(k)}]$ and $[x^{(k+1)}]$.

Of course, this algorithm is too simple to produce useful results in most cases. Indeed, note that the successive enclosures $[x^{(k)}]$ are increasing in width, even if the actual images $\varphi([x^{(0)}], t^{(k)})$ of the initial box are uniformly shrinking in size. Also, we have not indicated how to construct the widened box $[\tilde{x}^{(k)}]$. This can be quite delicate, and if not dealt with carefully, may produce gross over-estimates of the true images $\varphi([x^{(0)}], t^{(k)})$. Finally, it is not obvious that time-steps $\Delta t^{(k)}$, uniformly bounded away from zero, can be attained at every step. This requires a few additional conditions on the vector field. All of these issues are properly dealt with in the following sections.

4.5 The Euler method for a distance-d map

There are, however, many situations when we are more interested in how far we have flowed the initial values rather than how long time has passed. In dynamical systems, the main example is the study of Poincaré maps, which play a crucial role in determining the dynamics of many systems. In a local setting, a Poincaré map P can be thought of as a map between two very close codimension-one surfaces, $P\colon \Sigma^{(k)} \to \Sigma^{(k+1)}$. For simplicity, we will always demand that the sides of the surfaces be parallel to the coordinate axes. Then, if the two surfaces are at a distance d from each other, we call the (local) Poincaré map a distance-d map.

In what follows, we will restrict out attention to case $n = 3$, as is the case for the Lorenz equations. Note, however, that the methods described can be extended to any finite dimension. With our restriction, we will now study maps between rectangles and/or planes. Since we only consider planes that are cross products of the coordinate axes, there are exactly three planes through each point (the x_1x_2-, x_1x_3-, and the x_2x_3-planes).

Given any point $x \in \mathbb{R}^3$ that is not a fixed point of the vector field, there is always at least one plane that is transversal to the flow at the point x. We will always choose the plane whose normal vector corresponds to the component of the vector field having the largest modulus. In other words, if $|f_j(x)| = \max\{|f_i(x)|\colon i = 1, \dots, 3\}$, then we select the plane whose unit normal is e_j. By abuse of notation, this direction will often be referred to as the *transversal direction*. The remaining directions will be called the *non-transversal directions*.

For the sake of concreteness, let us now assume that $f_3(x)$ is negative and has the largest modulus. Then there exists a rectangle $\mathcal{R}^{(k)} = [x_1] \times [x_2] \times \{x_3^{(k)}\}$ containing x so that $f_3(x)$ is negative on $\mathcal{R}^{(k)}$. Geometrically, this means that the flow is passing through the rectangle from above. By continuity, the flow will pass through a plane $\Sigma^{(k+1)} = \{x\colon x_3 = x_3^{(k+1)}\}$ situated slightly beneath $\mathcal{R}^{(k)}$. The two questions we now pose are the following: (1) how do we determine the distance d to be travelled? (2) how do we estimate the rectangular hull $\mathcal{R}^{(k+1)}$ of the flow of $\mathcal{R}^{(k)}$ passing through $\Sigma^{(k+1)}$? The problems are illustrated in Figure 5.

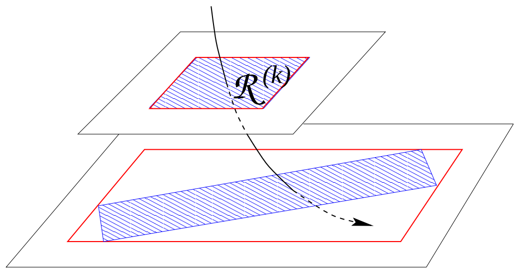

Figure 5: Finding the rectangular hull of a propagated surface.

If we denote the flow by $\varphi(x,t)$, we may define a local Poincaré map $\Pi\colon \mathcal{R}^{(k)} \to \Sigma^{(k+1)}$ by

$$\Pi(x) = \big(\Pi_1(x), \Pi_2(x)\big) = \big(\varphi_1(x,\tau(x)), \varphi_2(x,\tau(x))\big),$$

where $\tau(x)$ is the solution to $\varphi_3(x,\tau(x)) = x_3^{(k+1)}$. We will sometimes view $x_3^{(k+1)}$ and $x_3^{(k)}$ as being fixed, and consider the Poincaré map to be a function of two variables.

The first problem now is to find bounds on the various flow times. In order to this, we must restrict the flow to a compact set connecting $\mathcal{R}^{(k)}$ and $\Sigma^{(k+1)}$. The simplest way to do this is to define a box B as follows:

(1) Stretch the sides of $\mathcal{R}^{(k)}$ by a factor $\gamma > 1$, and call the new rectangle $\tilde{\mathcal{R}}^{(k)} = [\tilde{x}_1] \times [\tilde{x}_2] \times \{x_3^{(k)}\}$;

(2) Set $B = \tilde{\mathcal{R}}^{(k)} \times [x_3^{(k+1)}, x_3^{(k)}]$, where $|x_3^{(k+1)} - x_3^{(k)}| = d$.

Our hope is to be able to flow the rectangle $\mathcal{R}^{(k)}$ all the way down to the bottom of the box B whilst staying completely inside of B. For an arbitrary distance d, this is unfortunately not always possible. It is, however, clear that given $\mathcal{R}^{(k)}$ and B, there exists a positive distance d' for which the distance-d' map is well defined. Before we demonstrate this, we will introduce some auxiliary interval functions:

$$\begin{aligned}
\mathrm{Abs}([a]) &= \{|x|\colon x \in [a]\},\\
\mathrm{Mag}([a]) &= \max\{x\colon x \in \mathrm{Abs}([a])\},\\
\mathrm{Mig}([a]) &= \min\{x\colon x \in \mathrm{Abs}([a])\}.
\end{aligned}$$

Note that the function Abs is interval-valued, whereas Mag and Mig are real-valued.

Returning to the distance-d map, we first compute the minimal flow time for the non-transversal coordinates:

$$\Delta t = \min_{i=1,2}\{t > 0\colon [\mathcal{R}_i^{(k)}] + tf_i(B) \not\subset [B_i]\}. \tag{28}$$

This is the first time the image of $\mathcal{R}^{(k)}$ can possibly intersect a non-transversal side of the cube B. Next, we see how far the rectangle flows in the transversal direction during this time, with the restriction that we may not flow further than distance d:

$$d' = \min\{d, \Delta t \cdot \mathrm{Mig}(f_3(B))\}. \tag{29}$$

If $d' = d$, the rectangle flows all the way down to the bottom of the box B without touching the sides of B. If this is not the case, we define a smaller cube B' by trimming B in the transversal direction:

$$B' = [B_1] \times [B_2] \times [x_3^{(k)} - d', x_3^{(k)}].$$

This ensures that, when we repeat the procedure above using the trimmed box B', we will flow all the way down to its bottom. Note that this re-defines $\Sigma^{(k+1)}$.

Recall that $\tilde{\mathcal{R}}^{(k)}$ was constructed by stretching the sides of $\mathcal{R}^{(k)}$ by a constant factor γ. From a computational point of view, this is not optimal. Instead, we may now trim $\tilde{\mathcal{R}}^{(k)}$ (and thus B') by defining

$$\begin{aligned}
[B_i''] &= [\mathcal{R}_i^{(k)}] + [0, \Delta t] \cdot f_i(B') \qquad (i = 1, 2),\\
[B_3''] &= [B_3'].
\end{aligned}$$

If we now evaluate over B'', we get a tight bound on the flow times for the distance-d' map from $\mathcal{R}^{(k)}$:

$$[t] = \frac{d'}{\mathrm{Abs}(f_3(B''))}. \tag{30}$$

Finally, plugging the flow times into the Euler step gives a rigorous bound on the image of $\mathcal{R}^{(k)}$ under the distance-d' map:

$$\Pi_i(\mathcal{R}^{(k)}) \subseteq [\mathcal{R}_i^{(k+1)}] := [\mathcal{R}_i^{(k)}] + [t] \cdot f_i(B'') \qquad (i = 1, 2). \tag{31}$$

By composing several distance-d maps, we may flow an initial rectangle until the vector field shifts its dominating direction.

We can guarantee that the successive step sizes d' do not approach zero under very mild conditions: (1) the sizes of the flow-boxes B must be uniformly bounded from above; (2) the flow-boxes B must be uniformly bounded away from fixed points of the vector field: (3) the vector field evaluations must be finite. These conditions ensure that the minimal flow time given by (28) is uniformly "large", and that the transversal component of the vector field enclosure does not contain a zero, which would give $d' = 0$ in (29). We ensure condition (1) holds by imposing a *fixed scale*, which is described in Section 4.6. Condition (2) is taken care of by interrupting the computations whenever we come too close to the origin. The other two fixed points of the system are never approached, which the program checks for before computing a d'-step. Finally, condition (3) is valid since the vector field is finite in any compact region of phase space.

4.5.1 Fine-tuning the Euler method

There is, however, one major flaw in the Euler method: even if the true solution set is shrinking, the computed boxes $\mathcal{R}^{(k)}$ are always non-decreasing in k. This is because we always have the equality $\|[a] + [b]\| = \|[a]\| + \|[b]\|$ for any two intervals $[a]$ and $[b]$. The problem is that we are using interval arithmetic on a far too large scale in (31). We will now show how to eliminate this problem.

Generically, the trajectories maximizing the rectangular hull of $\mathcal{R}^{(k+1)}$ originate from the corners $\{p^{(i)}\}_{i=1}^4$ of $\mathcal{R}^{(k)}$ as illustrated in Figure 5. This happens when the partial derivatives of Π are non-zero in $\mathcal{R}^{(k)}$. There are, however, rare occasions when this is not the case: sometimes a mixed derivative $(\Pi_i)'_{x_j}$, $(i \neq j)$ may vanish (the fact that the non-mixed derivatives are always positive is confirmed by the computer program). If, say, $0 \in (\Pi_1)'_{x_2}(\mathcal{R}^{(k)})$, we must consider the whole line segment connecting[4] $p^{(4)}$ and $p^{(1)}$ in order to estimate the upper bound of $[\mathcal{R}_1^{(k+1)}]$, see Figure 6(b).

Assuming the generic case for now, the maximizing trajectories are confined to small boxes contained in sub-boxes with lids centered on the corners of $\mathcal{R}^{(k)}$, and extending down to $\Sigma^{(k+1)}$, see Figure 6(a).

Given a corner p of $\mathcal{R}^{(k)}$, we define its associated flow box B^p by

$$\begin{aligned}
[B_i^c] &= p_i + [0, \overline{t}] f_i(B'') \qquad (i = 1, 2),\\
[B_3^c] &= [B_3''].
\end{aligned} \tag{32}$$

Recall that $[t] = [\underline{t}, \overline{t}]$ are the flow times derived in the previous section. By definition, the trajectory connecting p and $\Pi(p)$ is contained in B^p. Using this flow box, we can compute tighter

[4]We label the corners as quadrants.

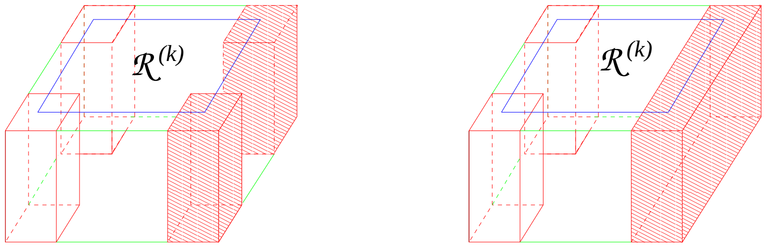

Figure 6: Restricting the computations: (a) the generic case (b) the non-generic case.

bounds on the flow times associated to the initial point p:

$$[t^p] = \frac{d'}{\mathrm{Abs}(f_3(B^p))}. \tag{33}$$

As before, this gives an enclosure of the distance-d' map of p:

$$\Pi_i(p) \subseteq [S_i^p] := p_i + [t^p] \cdot f_i(B^p) \qquad (i = 1, 2), \tag{34}$$

and by taking the convex hull (denoted by $\sqcup$) of the components of S^p over all corners, we get an enclosure of the image of $\mathcal{R}^{(k)}$:

$$\Pi_i(\mathcal{R}^{(k)}) \subseteq [\mathcal{R}_i^{(k+1)}] := \bigsqcup_p [S_i^p] \qquad (i = 1, 2). \tag{35}$$

The main reward is that now the image of $\mathcal{R}^{(k)}$ is free to contract in any direction, see Figure 7.

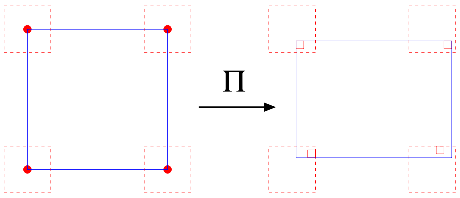

Figure 7: The image may now contract in some directions.

Now considering the case when one or several mixed derivatives vanish, we may repeat the procedure outlined above by substituting some of the corners with appropriate line segments. This results in larger flow boxes, and therefore gives a somewhat less tight bound on the image of $\mathcal{R}^{(k)}$.

4.5.2 Computing the partial derivatives

We will now provide an algorithm for computing enclosures of the partial derivatives of the local Poincaré map $\Pi\colon \mathcal{R}^{(k)} \to \Sigma^{(k+1)}$ as defined in the previous sections.

Consider the partial derivatives of Π:

$$\begin{aligned}
\frac{\partial \Pi_i}{\partial x_j}(x) &= \frac{\partial}{\partial x_j}\big[\varphi_i(x, \tau(x))\big] = \frac{\partial \varphi_i}{\partial x_j}\big(x, \tau(x)\big) + \frac{\partial \tau}{\partial x_j}(x)\frac{d\varphi_i}{dt}\big(x, \tau(x)\big) \\
&= \frac{\partial \varphi_i}{\partial x_j}\big(x, \tau(x)\big) + \frac{\partial \tau}{\partial x_j}(x) f_i\big(\varphi(x, \tau(x))\big) \\
&= \frac{\partial \varphi_i}{\partial x_j}\big(x, \tau(x)\big) + \frac{\partial \tau}{\partial x_j}(x) f_i\big(\Pi(x)\big) \qquad (i, j = 1, 2, 3).
\end{aligned} \tag{36}$$

The partial derivatives of $\tau(x)$ are obtained by noting that one $\Pi_i(x)$ is constant. Continuing our example (i.e. assuming that we still are flowing between x_1x_2-planes), we have that $\Pi_3(x)$ is constant, i.e,

$$0 = \frac{\partial \Pi_3}{\partial x_j}(x) = \frac{\partial \varphi_3}{\partial x_j}\big(x, \tau(x)\big) + \frac{\partial \tau}{\partial x_j}(x) f_3\big(\Pi(x)\big) \qquad (j = 1,2,3),$$

and solving for $\frac{\partial \tau}{\partial x_j}$ yields

$$\frac{\partial \tau}{\partial x_j}(x) = -\big[f_3\big(\Pi(x)\big)\big]^{-1}\frac{\partial \varphi_3}{\partial x_j}\big(x, \tau(x)\big) \qquad (j = 1,2,3).$$

Inserting this expression into (36) gives

$$\frac{\partial \Pi_i}{\partial x_j}(x) = \frac{\partial \varphi_i}{\partial x_j}\big(x, \tau(x)\big) - \frac{\partial \varphi_3}{\partial x_j}\big(x, \tau(x)\big)\frac{f_i\big(\Pi(x)\big)}{f_3\big(\Pi(x)\big)} \qquad (i, j = 1,2,3). \tag{37}$$

Note that the components with $i = 3$ vanish, just as we desired.

Since we already have an estimate on $\Pi(x)$, we can easily estimate the right-most factor in (37). The partial derivatives of the flow require some work, though. First, we need the differential equations for the partial derivatives. These are attained simply by differentiating the equations for the flow, $\frac{d}{dt}\varphi_i(x,t) = f_i(\varphi(x,t))$ $(i = 1,2,3)$, and changing the order of differentiation. On component level, this gives

$$\frac{d}{dt}\frac{\partial \varphi_i}{\partial x_j}(x,t) = \sum_{k=1}^{3}\frac{\partial f_i}{\partial x_k}(\varphi(x,t))\frac{\partial \varphi_k}{\partial x_j}(x,t) \qquad (i, j = 1,2,3),$$

or, in matrix form, $\frac{d}{dt}D\varphi(x,t) = Df(\varphi(x,t))D\varphi(x,t)$, with the initial condition $D\varphi(x,0) = I$, where I is the identity matrix. This translates into the following integral formula:

$$D\varphi(x,t) = I + \int_0^t Df(\varphi(x,s))D\varphi(x,s)ds. \tag{38}$$

We will now state a simple lemma used to compute the enclosure of $D\varphi(x,t)$.

Lemma 4.2 *Let A be an $n \times n$ interval matrix containing zero, i.e., $0 \in [A_{i,j}]$ for $i, j = 1, \dots, n$. If $I - \frac{1}{2}A$ is invertible, then the exponential of A satisfies*

$$e^A \subseteq I + [I - \tfrac{1}{2}A]^{-1}A.$$

Proof: By Taylor's formula, we know that

$$e^A = \sum_{k=0}^{\infty}\frac{A^k}{k!} = I + A + \frac{A^2}{2!} + \frac{A^3}{3!} + \dots$$

Now, for any interval $[a]$ with $0 \in [a]$, it is clear that $[a]/k! \subseteq [a]/2^{k-1}$ for all $k \geq 1$, seeing that $k! \geq 2^{k-1}$. Note that it is crucial that the interval $[a]$ contains zero for this condition to hold. Another fact we will use is that if B is an interval matrix containing zero, then $0 \in B^k$ for all $k \geq 1$. Therefore it follows that $A^k/k! \subseteq A^k/2^{k-1}$ for all $k \geq 1$, so we have

$$e^A \subseteq I + A + \frac{A^2}{2^1} + \frac{A^3}{2^2} + \dots = I + \sum_{k=1}^{\infty}\frac{A^k}{2^{k-1}} = I + 2\sum_{k=1}^{\infty}\left(\frac{A}{2}\right)^k = I + [I - \tfrac{1}{2}A]^{-1}A,$$

which concludes the proof. □

We are now ready to compute the enclosure of $D\varphi(x,t)$.

Lemma 4.3 *Define $A(t) = [0,t] \cdot Df(B'')$, and let Δ be the interval matrix defined by*

$$\Delta = Df(B'') \cdot \left(I + [I - \tfrac{1}{2}A(t)]^{-1} \cdot A(t)\right).$$

Then the solution to (38) satisfies $D\varphi(x,s) \subseteq I + s\Delta$ for all $(x,s) \in \mathcal{R}^{(k)} \times [0,t]$.

Proof: By the construction in the previous section, we know that $\varphi(x,s) \subseteq B''$ for all $(x,s) \in \mathcal{R}^{(k)} \times [0,t]$. Therefore, we have the following differential inclusion:

$$\tfrac{d}{dt} D\varphi(x,t) \in Df(B'')D\varphi(x,t). \tag{39}$$

Since $Df(B'')$ is a constant interval matrix, we can enclose the solution of (39) by taking the exponential:

$$D\varphi(x,t) \in e^{tDf(B'')} \subseteq e^{[0,t]\cdot Df(B'')}.$$

Since the interval matrix $A(t) = [0,t] \cdot Df(B'')$ obviously contains zero, Lemma 4.2 applies, and so

$$D\varphi(x,s) \in e^{A(t)} \subseteq I + [I - \tfrac{1}{2}A(t)]^{-1} \cdot A(t), \tag{40}$$

for all $(x,s) \in \mathcal{R}^{(k)} \times [0,t]$. We will now use this enclosure of $D\varphi(x,t)$, together with the enclosure of $Df(\varphi(x,s))$, in the right-hand side of (38). This gives

$$\begin{aligned} D\varphi(x,t) &\in I + \int_0^t Df(B'')\left(I + [I - \tfrac{1}{2}A(t)]^{-1} \cdot A(t)\right) ds \\ &\subseteq I + t \cdot Df(B'')\left(I + [I - \tfrac{1}{2}A(t)]^{-1} \cdot A(t)\right), \end{aligned} \tag{41}$$

which completes the proof. □

Note that we may intersect the right-hand sides of (40) and (41) to get a (possibly) tighter enclosure since they are both valid enclosures of $D\varphi(x,t)$.

Returning to the partial derivatives of Π given by (37), we now have enclosures of all appearing elements, which gives the following estimates:

$$\frac{\partial \Pi_i}{\partial x_j}(\mathcal{R}^{(k)}) \subseteq D\varphi_{i,j}\left(\mathcal{R}^{(k)}, [t]\right) - D\varphi_{3,j}\left(\mathcal{R}^{(k)}, [t]\right)\frac{f_i\left(\mathcal{R}^{(k+1)}\right)}{f_3\left(\mathcal{R}^{(k+1)}\right)} \qquad (i,j = 1,2,3).$$

Here, the flow times $[t]$ are given by (30), and all appearing elements are interval-valued. Note that since (in our example) we are flowing between two x_1x_2-planes, we are only interested in the four partial derivatives indexed by $i,j = 1,2$.

4.6 Partitioning

Although the method outlined in Section 4.5.1 may increase the accuracy on a local level, we are still left with a global problem: if the flow of the system under consideration rotates boxes, the strongest expanding (or least contracting) direction will *contaminate* all other directions. By this, we mean that the computed enclosures $\mathcal{R}^{(k)}$ will expand in all directions, although the true solution may contract in several directions. This phenomena is often referred to as the *wrapping effect*, see Figure 8(a).

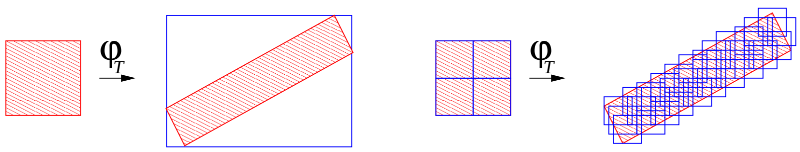

Figure 8: (a) The wrapping effect, and (b) how to overcome it.

Fortunately, we can reduce the wrapping effect by enforcing a *fixed scale*: if an element of any intermediate solution set (including the initial set) attains a width larger than a predetermined constant `max_size`, it is bisected along the directions that are too wide. Thus, the computed solution set will be made up of several small boxes, all having widths less than `max_size`. If the system has contracting directions, these will now show up in the solution set. This is due to the fact that elements squeeze together in the contracting directions, which results in an overlapping effect as illustrated in Figure 8(b). The global error is now comparable to `max_size`, and the contamination is greatly reduced. The following pseudo-code outlines an implementation of the algorithm just described:

Algorithm 2.

```
Initialize Stack with a box [x]
while Stack is not empty {
    Get a box [x] from Stack
    if [x] is too large {
        Bisect [x] in all directions that are wider than max_size
        Put the partitioned boxes in Stack
    }
    else {
        if [x] satisfies the stopping condition
            Put [x] in OutStack
        else
            Compute [x'], the propagation of [x], using your favorite algorithm
        Put [x'] in Stack
    }
}
Output OutStack
```

The concept of partitioning along the flow is probably the most fundamental idea in the entire program. Although the rectangles produced by our previous algorithms may expand in both directions, we can force their returns to be confined to a thin strip by partitioning often. The partitioning process just described is self-adaptive: there is no need to know in advance where the expansion is strong, or in what directions it may act. Each partitioned rectangle feels the pull in the attractor's normal direction (that is why we call it an attractor), and is therefore forced to center itself along the attractor. This results in a considerable overlapping, and is precisely why we can see contraction in one direction at the return, see Figure 8(b). Also, as mentioned earlier, we can attain quadratically close approximations of the interval-valued vector field f by choosing `max_size` small.

Again, using the conditions mentioned at the end of Section 4.4, we ensure that the number of partitioned elements remains finite. Indeed, since the vector field is finite and bounded away from zero, both the return time and the expansion along the flow are finite. Thus an initial rectangle can only expand by so much during its return, and therefore only a finite number of partitionings are required. Of course, if the vector field was extremely small (or our floating point system extremely coarse), the effects of the directed rounding might dominate over the actual expansion due to the vector field. In this case only the successful termination of the program would verify that a finite number of partitionings was required.

4.7 Switching the transversal direction

Once we have found rigorous bounds on the image of $\mathcal{R}^{(k)}$, we can restart the whole procedure with $\mathcal{R}^{(k+1)}$ as the initial rectangle. This can be repeated as long as the vector field does not vanish in the direction that we are flowing (the transversal direction). If we stay away from fixed points, there will always be at least one component of the vector field that is non-zero. Therefore, if the transversal component of the vector field becomes small, we switch to planes whose normal vector corresponds to the strongest component of the vector field. We do this by a transition procedure described below. Having switched the transversal direction, we can continue to flow the surface using the methods described in the previous sections. By switching between various planes when appropriate, we can follow the initial surface a whole lap up to its complete return to the global Poincaré section Σ. As mentioned above, the only exception to this rule is when a trajectory comes close to a fixed point. As we will not be performing any numerical computations near a fixed point, we may disregard this particular situation.

Let us continue our example from the previous sections. Suppose that we have followed the initial rectangle $\mathcal{R}^{(0)} \subset \Sigma$ by composing several distance-d maps, and suppose that we, at stage k, have

$$\mathrm{Mig}(f_1(\mathcal{R}^{(k)})) \geq C \cdot \mathrm{Mag}(f_3(\mathcal{R}^{(k)})), \tag{42}$$

for some $C > 1$. Without loss of generality, we may assume that f_1 is positive on $\mathcal{R}^{(k)}$. This means that the flow is turning to the right. Suppose $\mathcal{R}^{(k)} = [x_1] \times [x_2] \times \{x_3\}$. Instead of flowing to a plane $\Sigma^{(k+1)}$ situated slightly beneath $\mathcal{R}^{(k)}$, we flow to $\Sigma^{(k+1)} = \{x \colon x_1 = \overline{x}_1\}$. Just as before, we first must construct a flow box B which contains all trajectories from $\mathcal{R}^{(k)}$ to $\Sigma^{(k+1)}$. Let $\Delta x_1 = \overline{x}_1 - \underline{x}_1$, and set

$$B = [x_1] \times [\underline{x}_2 - \Delta x_1, \overline{x}_2 + \Delta x_1] \times [x_3 - \Delta x_1, x_3].$$

This is a candidate for the flow box, but we must confirm that no trajectory leaks out of B except through $\Sigma^{(k+1)}$. An easy way of checking this is simply to compute the range of the vector field over B, and then compare the components. If

$$\mathrm{Mig}(f_1(B)) \geq \mathrm{Mag}(f_i(\mathcal{R}^{(k)})) \qquad (i = 2, 3), \tag{43}$$

then no initial point in $\mathcal{R}^{(k)}$ is displaced by more than Δx_1 in any direction, and thus B qualifies as a flow box. If, however, (43) does not hold, we partition $\mathcal{R}^{(k)}$ into smaller pieces and start the switching procedure all over with each piece individually. Since the constant appearing in (42) is strictly larger than one, this procedure will succeed sooner or later.

Given a flow box B, we start by computing its associated flow times $[t]$: the largest flow time is given by $\overline{t} = \Delta x_1 / \mathrm{Mig}(f_1(B))$, and since some points of $\mathcal{R}^{(k)}$ are also points of $\Sigma^{(k+1)}$, the smallest flow time $\underline{t}$ is zero. We can now construct a (possibly) tighter flow box B':

$$\begin{aligned} [B'_1] &= [B_1] \\ [B'_i] &= [\mathcal{R}_i^{(k)}] + [t] \cdot f_i(B) \qquad (i = 2, 3), \end{aligned}$$

Computing the range of f over B' gives a tighter enclosure of the flow times $[t]$, and finally, we get bounds on the image of $\mathcal{R}^{(k)}$:

$$\begin{aligned} [\mathcal{R}_1^{(k+1)}] &= \{\overline{x}_1\}, \\ [\mathcal{R}_i^{(k+1)}] &= [\mathcal{R}_i^{(k)}] + [t] \cdot f_i(B') \qquad (i = 2, 3). \end{aligned} \tag{44}$$

4.8 Cone field propagation and expansion estimates

Now that we can rigorously compute the propagation of a rectangle travelling along the flow, we want to compute the evolution of tangent vectors associated to the rectangle[5]. To be more specific, each rectangle $\mathcal{R}^{(k)}$ is equipped with a cone $\mathfrak{C}^{(k)}$ spanned between two unit vectors $u^{(k)}$ and $v^{(k)}$:

$$\mathfrak{C}^{(k)} = \{w^{(k)}(t) \colon w^{(k)}(t) = u^{(k)} \cos t + v^{(k)} \sin t, \quad t \in [0, \tfrac{\pi}{2}]\}.$$

Our aim is two-fold:

(1) we want to compute an enclosure $\mathfrak{C}^{(k+1)}$ of the image of $\mathfrak{C}^{(k)}$ under $D\Pi$;

[5] The algorithms that will be outlined in this section have only been implemented in three dimensions. Thus we will restrict our discussion to cones that are two-dimensional, which makes life somewhat easier.

(2) we want to estimate the amount elements of $\mathfrak{C}^{(k)}$ are expanded/contracted under $D\Pi$.
Starting with (1), we use the enclosures derived in Section 4.5.2 and compute

$$\tilde{u}^{(k+1)} = D\Pi(\mathcal{R}^{(k)})u^{(k)} \quad \text{and} \quad \tilde{v}^{(k+1)} = D\Pi(\mathcal{R}^{(k)})v^{(k)}.$$

Note that the components of $\tilde{u}^{(k+1)}$ and $\tilde{v}^{(k+1)}$ are interval-valued. We simply define $\mathfrak{C}^{(k+1)}$ to be the hull of $\tilde{u}^{(k+1)}\cos t + \tilde{v}^{(k+1)}\sin t$ where $t \in [0, \frac{\pi}{2}]$.
Turning to the second point, let

$$[\varepsilon_{prel}^{(k+1)}] = |\tilde{u}^{(k+1)}| \bigsqcup |\tilde{v}^{(k+1)}|$$

be a preliminary estimate on the expansion of vectors $w^{(k)}(t)$ flowing from $\mathcal{R}^{(k)}$ to $\mathcal{R}^{(k+1)}$. Here $|\cdot|$ denotes the Euclidean norm on $\mathbb{R}^2$. Also, let $\theta^{(k)}$ denote the angle between the boundary vectors $u^{(k)}$ and $v^{(k)}$, and let $[\tilde{\theta}^{(k+1)}]$ denote the set of angles between the interval vectors $\tilde{u}^{(k+1)}$ and $\tilde{v}^{(k+1)}$. Since we have not considered the interior vectors, we must correct the preliminary expansion estimate by the factor:

$$[f_c] = \sqrt{\left(\{1\} \bigsqcup \frac{1+\cos[\tilde{\theta}^{(k+1)}]}{1+\cos\theta^{(k)}}\right)}.$$

To see where this correcting factor comes from, we first note that, for $t \in [0, \frac{\pi}{2}]$, we have

$$|w^{(k)}(t)|^2 = |u^{(k)}|^2\cos^2 t + |v^{(k)}|^2\sin^2 t + 2\langle u^{(k)}, v^{(k)}\rangle \sin t\cos t = 1 + 2\sin t\cos t\cos\theta^{(k)},$$

and

$$\begin{aligned} |\tilde{w}^{(k+1)}(t)|^2 &= |\tilde{u}^{(k+1)}|^2\cos^2 t + |\tilde{v}^{(k+1)}|^2\sin^2 t + 2\langle \tilde{u}^{(k+1)}, \tilde{v}^{(k+1)}\rangle \sin t\cos t \\ &\subseteq |\tilde{u}^{(k+1)}|^2\cos^2 t + |\tilde{v}^{(k+1)}|^2\sin^2 t + 2|\tilde{u}^{(k+1)}||\tilde{v}^{(k+1)}|\sin t\cos t\cos[\tilde{\theta}^{(k+1)}] \\ &\subseteq [\varepsilon_{prel}^{(k+1)}]^2\big(1 + 2\sin t\cos t\cos[\tilde{\theta}^{(k+1)}]\big). \end{aligned}$$

Therefore, we clearly have

$$\left(\frac{|\tilde{w}^{(k+1)}(t)|}{|w^{(k)}(t)|}\right)^2 \subseteq [\varepsilon_{prel}^{(k+1)}]^2 \frac{1+2\sin t\cos t\cos[\tilde{\theta}^{(k+1)}]}{1+2\sin t\cos t\cos\theta^{(k)}} \subseteq \left(\varepsilon_{prel}^{(k+1)}\right)^2 \left(\{1\} \bigsqcup \frac{1+\cos[\tilde{\theta}^{(k+1)}]}{1+\cos\theta^{(k)}}\right).$$

Before continuing to the next plane, we multiply the expansion estimate with the previous one, taking the correcting factor into account:

$$[\varepsilon^{(k+1)}] = [\varepsilon_{prel}^{(k+1)}] \cdot [f_c] \cdot [\varepsilon^{(k)}].$$

Initially, we take $[\varepsilon^{(0)}] = \{1\}$.
Due to the partitioning process, an initial rectangle $\mathcal{R}^{(0)}$ returns to Σ as many overlapping rectangles $\{\mathcal{R}_i\}_{i=1}^n$ whose union strictly contains the exact return of $\mathcal{R}^{(0)}$. By the procedure just described, each one of these rectangles is equipped with an expansion estimate $[\varepsilon_i] = \Pi_{k=0}^{n(i)}[\varepsilon_i^{(k)}]$. By taking the hull $[\mathcal{E}] = \sqcup_i[\varepsilon_i]$, it is clear that any vector of the initial cone associated with $\mathcal{R}^{(0)}$ is expanded by some factor $e \in [\mathcal{E}]$.

4.9 Handling cube entries

Let us finally deal with the trajectories passing close to the fixed point at the origin. The sole purpose of Chapter 3 was to allow us to interrupt the numeric computations whenever a rectangle $\mathcal{R}^{(k)}$ comes close to the origin. To be more precise, we center a cube of radius $\frac{1}{10}$ at the origin, and we interrupt the computations if a rectangle hits the lid of the cube. At this stage, the rectangles are small compared to the lid of the cube, and therefore we need not interrupt the computations for partial hits. If, however, the rectangle is completely within the boundary of the lid, we change to the normal form coordinates. This will distort the rectangle and its tangent vectors as described in Lemma 3.4 and the discussion following it. Once we have taken the distortion into account, we may assume that the flow is totally linear in the cube. Indeed, the maximal error this assumption yields is of the same size as κ, and can thus be considered as taken into account via the distortion[6]. Thus we can explicitly compute where the rectangle exits the cube, and also how the tangent vectors are affected.

There are two different ways that a rectangle can pass through the cube. If the distorted rectangle intersects the stable manifold of the origin (which is the x_2x_3-plane), it is split along the line of intersection, and exits the cube in two directions. Otherwise the rectangle flows out in one piece, see Figure 9.

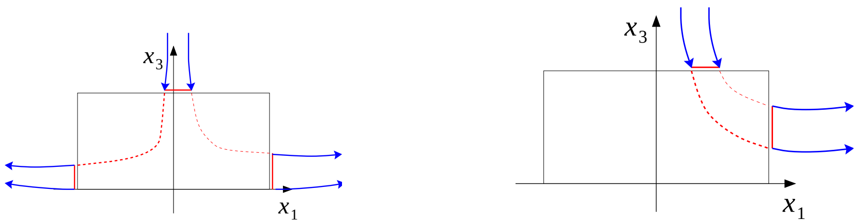

Figure 9: (a) Splitting along the stable manifold (b) No splitting.

Now we are ready to switch back to the original coordinates. Once again, the computed exit rectangle(s) and the associated tangent vectors will be distorted as described in Lemma 3.4. The distortion widens the rectangle in all directions, making it a solid box. Since our algorithms are tuned for flowing codimension one surfaces (e.g, rectangles in $\mathbb{R}^3$) we therefore flatten out the box by flowing it to its out-most slide, see Figure 10. After having completed these outlined steps, we can resume the numeric computations as described in the previous sections.

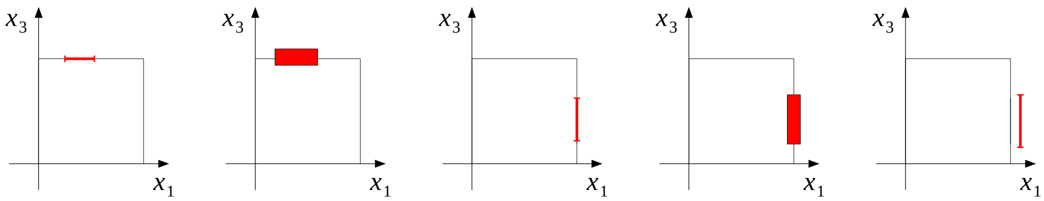

Figure 10: The five stages of passing through the cube: (1) hit the cube completely, (2) distort via change of coordinates, (3) compute the exit, (4) distort via inverse change of coordinates, and (5) flow the box into a codimension one surface.

[6] Recall that κ was used in the estimates on the normal flow. Its value is $2 \cdot 10^{-19}$.

5 The RODES program

The RODES (Rigorous ODE Solver) program is a highly adaptive, multi-processor program. As we pointed out earlier, the computations are performed in interval arithmetic with directed rounding when necessary. This functionality is provided by the PROFIL/BIAS package (see [8]) which is supported on all architectures utilized in the proof. The program was executed on 20 machines working in parallel. Data was passed between the processes via a common text file. All floating point numbers were passed with 17 digits of precision, which converts exactly according to the IEEE standard. The computers employed for the task were a variety of SUN Sparc stations, with models ranging from LX, Sparc 4 to Ultra 1. The total computational time in this setting was about 100 hours. Other setups with fewer computers equipped with stronger processors have been performed with similar results. In the sections to come, we will give an overview of the program's global structure and the computations carried out.

5.1 Fundamental classes and computational structure

Let us begin by explaining how we represent data in the program. First, we select our return plane Σ. In our case, it was chosen to be $\Sigma = \{x \in \mathbb{R}^3 : x_3 = 27\}$. This is the usual choice for the classical parameter values $(\sigma, \beta, \varrho) = (10, 8/3, 28)$. Note that the plane $\{x \in \mathbb{R}^3 : x_3 = \varrho - 1\}$ contains the two symmetric fixed points $C^{\pm}$ mentioned in Section 1.1. Next, we restrict the class of *admissible* initial rectangles to the ones that are representable in the following form:

$$\mathcal{R}^{(0)} = 2^{-P} * [u-1, u+1] \times [v-1, v+1] \qquad (u, v, P \in \mathbb{Z}).$$

We are now assuming that all rectangles lie in the plane Σ. Not only does this allow us to represent initial rectangles in a very compact way: $\mathcal{R}^{(0)} \sim \langle u, v, P \rangle$, we also have a very restricted number of rectangles (grids) to deal with later, when we will be following orbits under the return map.

The value $P = 8$ is used for the computations at hand. Also, we exclusively use odd integers for the grid coordinates u and v. This makes the admissible rectangles non-overlapping.

Once an initial rectangle has entered the program it is converted to a degenerate box which, in turn, is represented as the cross product of three intervals, one of them having zero diameter. Every initial box comes equipped with a cone, which is represented as the two angles its boundary vectors make with the positive x_1-axis. The initial box and cone are parts of a larger structure called a *parcel*. A parcel contains all information we need to perform the flowing procedures described in Chapter 4. The parcel structure (which really is a C++ class) contains the following elements:

```
class parcel
{
  BOX       box;              // The coordinates of all the variables.
  INTERVAL  angles;           // The boundary angles of the cone.
  INTERVAL  expansion;        // The enclosure of the expansion.
  short     trvl;             // The transversal coordinate: 1,...,DIM.
  short     sign;             // The sign of the flow: - 1 or + 1.
  INTERVAL  time;             // The "flow time" variable.
  short     message;          // Any message that needs to be passed on.
};
```

Once all the elements of the parcel are defined, the variable `max_size` is given an appropriate value, depending on the location of the initial rectangle. Recall that this variable determines the maximal diameter a box may attain before being partitioned into small pieces.

Next, we set the global stopping conditions. These make sure that we flow the initial parcel all the way back to Σ as illustrated in Figure 2. This information is stored in the variable `glob_stop_param`. We are now all set to call the core function described in Chapter 4 that computes the return of the initial parcel:

```
Flow_The_Parcel(current_pcl, Return_List, glob_stop_param, max_size);
```

Here, `Return_List` is a whole collection of parcels representing the return of the initial parcel.

5.2 The initial data

Let us recall our candidate for the trapping region N. This set consists of two disjoint branches, N^- and N^+, each made up of admissible rectangles belonging to the return plane Σ. We will call these small rectangles $N_i^\pm$, and write

$$N = N^- \cup N^+ = \Big(\bigcup_{i=1}^{n_0} N_i^-\Big) \cup \Big(\bigcup_{i=1}^{n_0} N_i^+\Big).$$

The two branches of N have the same symmetry as the Lorenz equations, i.e., $N_i^+ = S(N_i^-)$, where $S(x_1, x_2, x_3) = (-x_1, -x_2, x_3)$. Thanks to this symmetry, we only have to perform the computations on one branch of N. When it is not relevant which branch we are considering, we sometimes omit the $\pm$ labeling of the small rectangles. For quantifying the hyperbolic properties of the return map, each initial rectangle N_i comes with a cone $\mathfrak{C}([\alpha_i])$, where we use the notation

$$\mathfrak{C}([\alpha]) = \{v \in \mathbb{R}^2 \colon v\angle(0,1) \bmod 180 \in [\alpha]\}.$$

Initially, the candidate for the trapping region N consists of just one seed element, situated in the upper branch. It is represented as $\langle u, v, P\rangle = \langle 1255, 727, 8\rangle$, and its associated cone is spanned between the angles 0 and 10 (degrees). In the next section, we shall see how the program modifies N by gradually adding elements to it. The cones are also subject to modification as the program adds more elements to the the trapping region.

5.3 Forward invariance

As described in Chapter 4, the program computes C^0 and C^1 information about the return of the rectangles. The C^0 information gives us rigorous bounds on the entire orbit of an initial rectangle. In particular, given any N_i, the program produces a set of overlapping rectangles $\{Q_{i,j}\}_{j=1}^{k(i)}$ whose union strictly contains the return of N_i:

$$R(N_i) \subset \bigcup_{j=1}^{k(i)} Q_{i,j}.$$

Note that, although we demand that the initial rectangles are admissible, the rectangles $Q_{i,j}$ are generally not. By simply adding any admissible rectangle that has non-zero intersection with one of the $Q_{i,j}$'s, and which is already not a member of N, we ensure that $R(N_i \setminus \Gamma) \subset N$. When we have gone through the whole list of elements of N, and no more elements need to be added, we clearly have proved that $R(N \setminus \Gamma) \subset N$, and thus that the return map is well defined (in the sense of the geometric model) on the whole trapping region N.

Turning to the cone field, recall that each initial rectangle N_i comes with a cone $\mathfrak{C}([\alpha_i])$. The C^1 information provides us with $k(i)$ new cones $\mathfrak{C}([\beta_{i,j}])$ associated with the returns $Q_{i,j}$, $j = 1, \dots, k(i)$. If a $Q_{i,j}$ intersects an element N_k not already belonging to N, we simply add N_k to N and equip it with the cone of $Q_{i,j}$, i.e., $\mathfrak{C}([\beta_{i,j}])$. If, on the other hand, $Q_{i,j}$ intersects an element N_k already belonging to N, then we check if $\mathfrak{C}([\beta_{i,j}]) \subseteq \mathfrak{C}([\alpha_k])$. If this is the case, we do not need to take any action. If, however, $\mathfrak{C}([\beta_{i,j}]) \nsubseteq \mathfrak{C}([\alpha_k])$, we must widen the cone associated with N_k so that is contains the hull of both cones. If, furthermore, the C^0 and C^1 information for N_k has already been computed, we must recompute this information for N_k with its wider cone.

Again, when we have gone through the whole list of elements of N, and no more elements need to be added or recomputed, we clearly have proved that

$$R(N_i) \cap N_k \neq \emptyset \implies DR(N_i) \cdot \mathfrak{C}([\alpha_i]) \subset \mathfrak{C}([\alpha_k]).$$

Since this inclusion holds for all elements of N, we have proved the existence of a forward invariant cone field. This condition was satisfied with $n_0 = 7260$ for our initial seed.

A short remark is in order here: since most elements of N have several pre-images, it is highly unlikely that we would *not* have to recompute due to the cone being widened. Therefore, when a new element is added to N, we modify its associated cone and make it wider than strictly necessary. To be more precise, we first widen the cone by a factor of 1.5. If the cone opening is still less than 5 degrees, we widen it to an opening of 5 degrees. If an already existing element needs to be recomputed due to cone overflow, we also take the new cone to be wider than strictly necessary.

5.4 Expansion estimates

We now turn to the question of expansion. As described in Section 4.8, the computations carried out to prove forward invariance also provide us with an enclosure of the expansion of tangent vectors belonging to the initial cone: each $Q_{i,j}$ is associated with an interval $[\varepsilon_{i,j}]$ containing the range of the expansion a vector starting within $\mathfrak{C}([\alpha_i])$ can be subjected to. As we are primarily concerned with the *minimal* expansion a vector may have, we simple take

$$\mathcal{E}_i = \min\{\underline{\varepsilon}_{i,j} : j = 1, \ldots, k(i)\}$$

to be the minimal expansion estimate associated with all tangent vectors in $\mathfrak{C}([\alpha_i])$.

Each element of the trapping region will also have a corresponding *pre-expansion* estimate $\mathcal{E}_i^{(-1)}$ (which is *not* the reciprocal of $\mathcal{E}_i$). This is defined as follows:

$$\mathcal{E}_k^{(-1)} = \min\{\underline{\varepsilon}_{i,j} : Q_{i,j} \cap N_k \neq \emptyset\}.$$

Note that the returning elements $Q_{i,j}$ are to be taken over all intersecting images $R(N_i) \cap N_k \neq \emptyset$ Just as in the section above, this estimate will thus be modified during the computations of the trapping region.

Much to our surprise, we found regions in N which were contracted in all directions under R. In view of the geometric model, which is everywhere expanding in the unstable cones, this was not anticipated. We prove, however, that all tangent vectors within the cone field are *eventually* expanded under DR. More precisely, given any orbit $x_0, x_1, \ldots,$ where $x_j = R^j(x_0)$, we can divide it into non-overlapping pieces $[x_0, \ldots, x_{k_0}], [x_{k_0+1}, \ldots, x_{k_1}], \ldots$ where all but the first piece accumulate an expansion factor greater than 2.79. The fact that this number is greater than 2 is relevant when proving transitivity of the attracting set, see Section 2.4. We also show that $k_{i+1} - k_i \leq 31$, which gives a very crude lower estimate of the positive Lyapunov exponent of R:

$$\liminf_{n\to\infty} \frac{1}{n} \log(\min\{|DR^n(x)v| : v \in \mathfrak{C}, |v| = 1\}) \geq \sqrt[31]{2.79} > 1.033.$$

5.4.1 Getting oriented

Before we prove these facts, let us pause for a moment, and consider Figure 12. We will use this illustration as a dynamic road map. First, note that Λ consists of two main branches (upper and lower). Next, note that each branch is split in two by the line $\Gamma = W^s(0) \cap \Sigma$. This gives us four naturally defined regions, and we will refer to these as $A_1, \ldots, A_4$, where the labeling is analogous to that of quadrants. The dynamics of R is now easy to describe: any trajectory starting in A_1 tends to the left, and enters A_2 after a finite number of iterates. A trajectory starting in A_2, however, must immediately go to either A_3 or A_4. Using the symmetry of the Lorenz equations, we end up with the diagram illustrated in Figure 11.

Of course, we are deliberately ignoring the pre-images of Γ. These points have a finite number of iterates before hitting Γ and vanishing.

Let us now resume our discussion about the contracting regions. There is one such region in each A_i, and they are situated as follows: in A_1, the right-most part ($x_1 > 2.820$) is contracting. The

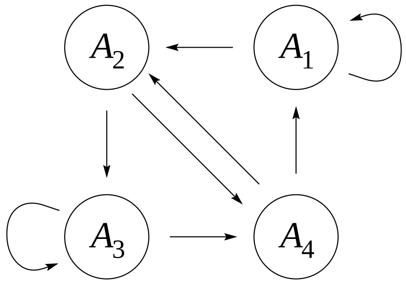

Figure 11: The simplified dynamics of the return map R.

strongest contraction[7] occurs to the far right, and is roughly by a factor 0.6804. All trajectories travel to the left in A_1, and thereby tend to the expanding region of A_1. The contracting region in A_2 is situated to the left ($x_1 < -0.5937$), and the strongest contraction is no less than by a factor 0.1186. This is quite a strong contraction rate, mostly due to the small angle between the attractor and the stable direction, see Figure 12. As always, these numbers should not be taken as absolute facts: they only reflect the outcome of our computations. Indeed, with higher accuracy, it is possible to get the expansion estimates for the region mentioned in A_1 to be almost one. The left-most region mentioned in A_2, however, does not seem to rise much when the accuracy is increased. This is expected, as there is nothing we can do about the close tangency present.

As the geometric model assumes that the attractor is virtually perpendicular to the stable directions, the problem of close tangencies is never discussed in the literature. It is, however, a serious obstruction that must be overcome to prove that there exists an expanding direction. Indeed, by increasing the parameter r slightly, the attractor appears to become tangent to the stable directions, and then there is no hope to save the robust persistence of the attractor, see [11] and [10].

We are saved by the fact that orbits entering the out-most parts of A_2 or A_4 have just been very close to the fixed point. Recall that the expansion is very large (unbounded, in fact) in a neighbourhood of the origin. Thus, before we enter a strongly contracting region, we have hopefully already pre-compensated for the coming contraction. We prove that this indeed is the case by confirming that the product $\mathcal{E}_i^{(-1)}\mathcal{E}_i$ always is large.

5.4.2 Forward iterates and accumulated expansion

Let us begin this section by introducing some notation: let $[R](N_i)$ denote the computed return of N_i, and let $\langle R\rangle(N_i)$ denote the set of admissible rectangles of N that intersect $[R](N_i)$, i.e.,

$$[R](N_i) = \bigcup_{j=1}^{k(i)} Q_{i,j} \quad \text{and} \quad \langle R\rangle(N_i) = \{N_k \colon N_k \cap [R](N_i) \neq \emptyset\}.$$

Note that $\langle R\rangle$ takes the set of admissible rectangles into itself. Therefore, we may consider higher iterates, $\langle R^2\rangle, \langle R^3\rangle, \dots$, by defining

$$\langle R^n\rangle(N_i) = \langle R\rangle(\langle R^{n-1}\rangle(N_i)),$$

where $\langle R^0\rangle(N_i) = N_i$. Thus, for any initial point $x_0 \in N_i$, we have $R^k(x_0) \in \langle R^k\rangle(N_i)$ for $k = 0, 1 \dots$. With each iterate, we can now also associate a lower expansion estimate:

$$\mathcal{E}_i^{(k)} = \min\{\mathcal{E}_j \colon N_j \in \langle R^k\rangle(N_i)\}.$$

[7]When we talk about contraction and expansion rates, we always mean the rates restricted to the unstable cone field.

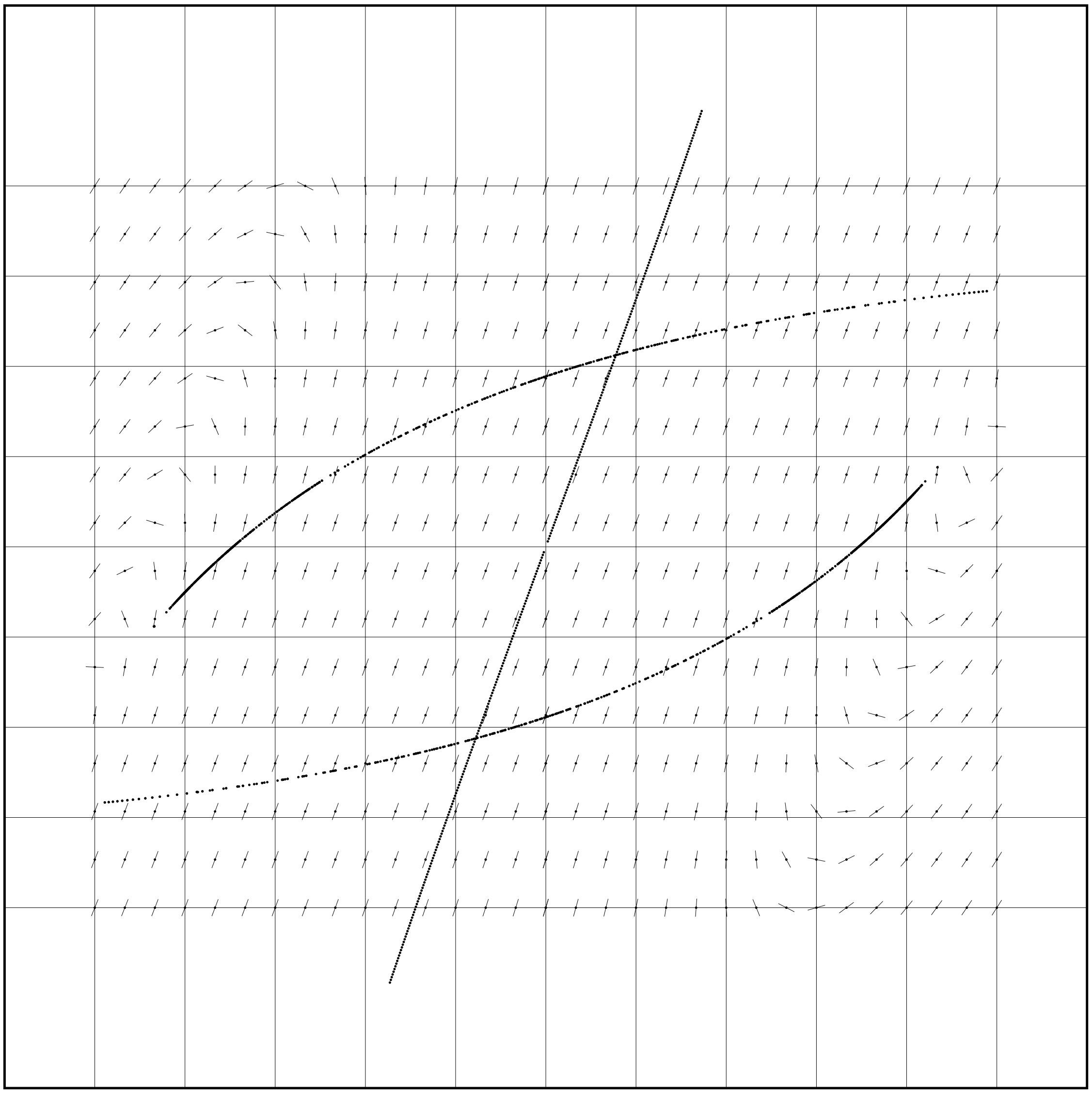

Figure 12: An approximation of Λ with the most contracting directions for one iterate of R. The (almost) straight line cutting across the two branches of Λ is the intersection between the stable manifold of the origin and the return plane. Note the close tangencies between the extreme tips of the attractor and the contracting directions. The bounding box is $[-6, 6] \times [-6, 6] \times \{27\}$.

Therefore any tangent vector $v \in \mathfrak{C}([\alpha_i])$ following the orbit of an initial point $x_0 \in N_i$ will satisfy

$$|DR^n(x_0)v| \geq |v| \prod_{k=0}^{n-1} \mathcal{E}_i^{(k)}.$$

We will now prove that the expansion along orbits in N grows exponentially with the number of iterates. First, we will coarsen the set of admissible rectangles making up the trapping region N. This is done by considering one branch of N at a time, and grouping all rectangles that have the same u-value into one larger rectangle. This makes the set N very similar to a one-dimensional set, which clarifies much of the arguments to come. We keep the same notation by also denoting these new (and larger) elements of N by $N_i^{\pm}$.

Since N is a trapping region with a forward invariant cone field, it is foliated by stable leaves. Let ℓ be any leaf of the foliation, and consider the set of points in N between (and including) ℓ and its image under the return map $R(\ell)$. We call such a set a *fundamental domain* for R. An important property of such a set is that an orbit can not cross a fundamental domain without having an iterate in it. Of course this property also holds for any set containing a fundamental domain.

We will produce a set F such that the following proposition holds:

Proposition 5.1 *There exists a set $F \subset N$ satisfying:*

(1) F contains a fundamental domain;

(2) F contains the leaves $\Gamma \cap N$;

(3) Any orbit with $x_0 \in F$, eventually leaving F, satisfies for every return $x_n \in F$

$$\min\{|DR^n(x_0)v| \colon v \in \mathfrak{C}\} > 2|v|;$$

(4) Any orbit completely contained in F satisfies

$$\min\{|DR^n(x_0)v| \colon v \in \mathfrak{C}\} > (\sqrt{2})^n|v|.$$

In particular, statement (3) is true for the *first* return to F. Therefore, a small line segment must more than double its length between two consecutive slicings over $\Gamma = \Sigma \cap W^s(0)$.

Once more, our claims will be proved by a computer program, `expansion.cc`. The underlying algorithm can be described as follows:

Algorithm 3.

Enter F
Set $\tilde{F} = \langle R\rangle(F) \cap (N \setminus F)$
For each $N_i \in \tilde{F}$ {
    Set `acc_exp` $= \mathcal{E}_i^{(-1)}$
    Insert N_i *in* `Stack`
    while `Stack` *is not empty* {
        For each N_k *in* `Stack` {
            if $N_k \in F$ {
                if `acc_exp` > 2
                    Remove N_k *from* `Stack`
                else
                    Signal an error
            }
        }

```
        Set local_exp = min{ℰ_j : N_j ∈ Stack}
        Set Stack = ⟨R⟩(Stack)
        Set acc_exp = local_exp * acc_exp
    }
}
```

Note that we start computing orbits from a set $\tilde{F}$ which is the image of F minus any overlaps with F itself. This is a technical trick motivated by the fact that the pre-expansion estimates in $\tilde{F}$ are much better than the corresponding forward estimates in F.

The reason we may "trim" away the overlapping elements $F \cap \langle R\rangle(F)$ is seen as follows: these points will either eventually leave the set F or remain in F forever. In the first case, we use the fact that the expansion in F is greater than one. This is verified by the program by checking that $\min\{\mathcal{E}_i \colon N_i \in F\} > 1$. Therefore we may consider these points to be taken care of by the main algorithm. In the second case, we simply verify that the points satisfy $\min\{|DR^2(x_0)v| \colon v \in \mathfrak{C}\} > 2|v|$. This is done by the program by checking that all $N_i \in F$ with $\langle R\rangle(N_i^{\pm}) \cap N_i^{\mp} \neq \emptyset$ satisfy $\mathcal{E}_i > \sqrt{2}$.

The program also verifies that F really contains a fundamental domain. This is easily done by checking that the the right-most boundary element N_i^+ of the upper branch part of F, called F^+, is mapped into F^+. Therefore, since F^+ is a union of adjacent rectangles, any element to the right of N_i^+ must also have an iterate in F^+.

Finally, the fact that the program does not signal an error proves that all points that leave F have accumulated an expansion factor of at least 2 on their first return to F. The fact that the program terminates proves that all points that leave F really do return to F.

The program was executed with F^+ selected to be the union of all rectangles with u-values in $[-128, 512]$. This corresponds to the set of rectangles in N^+ whose x_1-coordinates belong to $[-\frac{1}{2}, 2]$. The smallest accumulated expansion for orbits returning to F was found to be 2.7914; the smallest expansion factor for orbits confined to F was found to be 1.7526; the minimal expansion in F was found to be 1.06792. Furthermore, the longest number of iterates spent outside F was found to be 30. The time required for these computations was a couple of minutes on a *very* slow computer.

This completes the proof of Proposition 5.1, and also provides us with an algorithm for dividing an orbit into non-overlapping pieces $[x_0, \dots, x_{k_0}], [x_{k_0+1}, \dots, x_{k_1}], \dots$ where all but the first piece accumulate an expansion factor greater than 2.

A Chaos theory for pedestrians

In this section we will outline some basic definitions used in the theory of dynamical systems. The presentation is based mainly on [16], [18], and [24].

A.1 Hyperbolicity

Consider a C^k diffeomorphism ($k \geq 1$) of a compact manifold to itself, $f\colon M \to M$. The *forward orbit* of a point $p \in M$ under f is the set $\{f^i(p)\}_{i=0}^{\infty}$, where f^i is f composed with itself i times. A point p is a *periodic point of period* k provided $f^k(p) = p$ and $f^i(p) \neq p$ for $0 < i < k$. If p has period one, we call it a *fixed point* of f. We say that p is a *hyperbolic fixed point* of f if $f(p) = p$ and if Df_p has no eigenvalues on the unit circle. According to standard results in spectral theory, there then exists a splitting of the tangent space $T_pM = \mathbb{E}_p^s \oplus \mathbb{E}_p^u$, where the invariant subspaces $\mathbb{E}_p^s$ and $\mathbb{E}_p^u$ correspond to the spectrum inside and outside the unit circle, respectively. This means that we can find constants $\sigma \in (0,1)$ and $C > 0$ such that for all $n \in \mathbb{N}$,

$$\|Df_p^n|\mathbb{E}_p^s\| \leq C\sigma^n \quad \text{and} \quad \|Df_p^{-n}|\mathbb{E}_p^u\| \leq C\sigma^n$$

for some Riemannian norm $\|\cdot\|$ on T_pM. The subspaces $\mathbb{E}_p^s$ and $\mathbb{E}_p^u$ are called the *stable* and *unstable subspaces* for the fixed point p. Given a hyperbolic fixed point p, we define its stable and unstable manifolds:

$$W^s(p) = \{x \in M \colon \lim_{i\to\infty} f^i(x) = p\},$$
$$W^u(p) = \{x \in M \colon \lim_{i\to\infty} f^{-i}(x) = p\}.$$

These sets are injectively immersed C^k submanifolds of M, and have the same dimensions as their corresponding linear subspaces.

We can extend these definitions to the case when p is a periodic point of period k simply by replacing f by f^k. Also, we can extend the notion of a hyperbolic fixed point to a whole set.

Consider a compact set $\Lambda \subset M$ which is invariant under f, i.e., $f(\Lambda) = \Lambda$. We say that Λ is a *hyperbolic set* for f if there exists a splitting $T_xM = \mathbb{E}_x^s \oplus \mathbb{E}_x^u$ for each $x \in \Lambda$, such that

1. $\mathbb{E}_x^s$ and $\mathbb{E}_x^u$ vary continuously with x,
2. the splitting is invariant, i.e., $Df_x \cdot \mathbb{E}_x^s = \mathbb{E}_{f(x)}^s$ and $Df_x \cdot \mathbb{E}_x^u = \mathbb{E}_{f(x)}^u$,
3. there are constants $\sigma \in (0,1)$ and $C > 0$ such that for all $n \in \mathbb{N}$,

$$\|Df_x^n|\mathbb{E}_x^s\| \leq C\sigma^n \quad \text{and} \quad \|Df_x^{-n}|\mathbb{E}_x^u\| \leq C\sigma^n.$$

A.2 Robustness

In practice, it is impossible to explicitly find the invariant set Λ, not to mention the splitting $T_\Lambda M = \mathbb{E}^s \oplus \mathbb{E}^u$, except in the most trivial cases. Fortunately, we shall soon see that hyperbolicity is a robust property, and one can thus make do with pretty crude approximations of both Λ and the subbundles of the splitting. By *robust*, we mean that the defining hypotheses are open in the C^1-topology.

A compact region $N \subset M$ is called a *trapping region* for f provided $f(N) \subset N^\circ$, where N° denotes the interior of N. Given such a set, we can construct the *maximal invariant set* of N:

$$\Lambda = \bigcap_{i=0}^{\infty} f^i(N).$$

It is clear that any other invariant set in N must be a proper subset of Λ. Seeing that the sequence $\{f^i(N)\}_{i=0}^{\infty}$ is nested, we can approximate Λ by considering high iterates of N. Any property valid in an open neighbourhood of Λ will then also hold for $f^k(N)$ if we take k sufficiently large. Seeing that the image of a trapping region also is a trapping region, we may assume that $k = 1$ by taking N close to Λ.

Let $T_N M = \mathbb{F}^s \oplus \mathbb{F}^u$ be a continuous splitting approximating $\mathbb{E}^s \oplus \mathbb{E}^u$. Given $\alpha \geq 0$ we define the *stable* and *unstable cone fields*

$$C_x^s(\alpha) = \{v_1 + v_2 \in \mathbb{F}_x^s \oplus \mathbb{F}_x^u \colon |v_2| \leq \alpha |v_1|\},$$
$$C_x^u(\alpha) = \{v_1 + v_2 \in \mathbb{F}_x^s \oplus \mathbb{F}_x^u \colon |v_2| \geq \alpha |v_1|\}.$$

The following theorem provides a practical way of proving that a set is hyperbolic:

Theorem A.1 *Let N be a trapping region for a C^1 diffeomorphism f. Suppose that there exists a continuous splitting $T_N M = \mathbb{F}^s \oplus \mathbb{F}^u$, and that there are constants $\alpha \geq 0$, $C > 0$, and $\sigma > 1$ so that*

$$Df_x^{-1} \cdot C_x^s(\alpha) \subset C_{f^{-1}(x)}^s(\alpha) \quad \text{and} \quad Df_x \cdot C_x^u(\alpha) \subset C_{f(x)}^u(\alpha)$$

and

$$\|Df_x^{-n} | C_x^s(\alpha)\| \geq C\sigma^n \quad \text{and} \quad \|Df_x^n | C_x^u(\alpha)\| \geq C\sigma^n$$

for every $x \in N$. Then $\Lambda = \cap_{i=0}^{\infty} f^i(N)$ is hyperbolic for f.

It is clear that the hypotheses of this theorem are open in the C^1-topology, which proves that hyperbolicity is a robust property. In particular, if g is C^1 close to f, then $\Lambda_g = \cap_{i=0}^{\infty} g^i(N)$ is hyperbolic for g.

A.3 Strange attractors

We will fix the following notation[8]: a compact, invariant set Λ_f is called *attracting* if there exists an open neighbourhood U of Λ_f such that $\cap_{i=0}^{\infty} f^i(U) = \Lambda_f$. The largest such U is called the basin of attraction for Λ_f, and is denoted $B(\Lambda_f)$. In particular, the maximal invariant set of any trapping region is an attracting set. Even so, it may be the case that most points in $B(\Lambda_f)$ tend to a much smaller subset of Λ_f. As an example, consider the diffeomorphism with a phase portrait as illustrated in Figure 13.

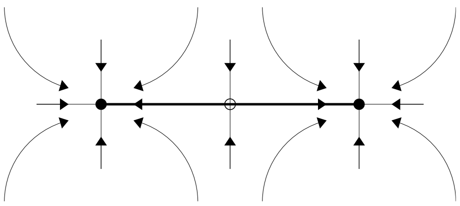

Figure 13: An attracting set which is not an attractor.

Although the whole interval I between the two filled fixed points is attracting with $B(I) = \mathbb{R}^2$, it is clear that most orbits tend to either one of the extreme points of I.

In order to avoid this kind of situation we restrict our attention to a subset of the attracting sets. An *attractor* is an attracting set which contains a dense orbit: $\Lambda_f = \overline{\cup_{i=0}^{\infty} f^i(x)}$ for some $x \in \Lambda_f$. This means that Λ_f is minimal in the sense that no proper subset of Λ_f is attracting. Clearly, the attracting set I in our example is not an attractor whereas the two extreme fixed points are. There is, however, nothing "chaotic" about the asymptotic behaviour of points tending to these attractors, and the situation is therefore dynamically uninteresting.

[8] The reader should be aware of that there are several different notions of a strange attractor, see [13]. We choose to use very strong (but natural) requirements seeing that the Lorenz attractor satisfies almost all existing definitions of a strange attractor.

From this point of view we would like to be able to distinguish attractors exhibiting interesting dynamical properties from those who do not. For this purpose, an attractor is called ***strange*** if for almost all pairs of different points in $B(\Lambda_f)$, their forward orbits eventually separate by at least a constant δ (only depending on Λ_f). Here, almost all pairs means with probability one in $B(\Lambda_f) \times B(\Lambda_f)$ with respect to Lebesgue measure. These attractors are sometimes called ***chaotic*** or ***sensitive*** seeing that, no matter how accurately we measure the initial conditions, we will eventually accumulate an error of size δ.

Sometimes, we can also say something about the speed at which nearby orbits separate. Indeed, if an attractor Λ_f is hyperbolic with a non-trivial unstable tangent bundle, we clearly have exponential divergence of almost all nearby orbits. Such an attractor is called ***non-trivial hyperbolic*** and, apart from being strange, it is also robust.

A.4 Flows and their return maps

We will now describe a useful relation between discrete-time (maps) and continuous-time (flows) dynamical systems. Consider the system of ordinary differential equations

$$\dot{x} = v(x), \tag{45}$$

where $x \in M$ and where the vector field v is a C^r function, $v\colon M \to TM$. Let $\varphi(x,t)$ denote the flow of (45), i.e.,

$$\tfrac{d}{dt}\varphi(x,t) = v(\varphi(x,t)),$$

and suppose that the system (45) has a periodic solution of period $T > 0$, containing the point x_0, i.e., $\varphi(x_0, T+t) = \varphi(x_0, t)$ for all $t \in \mathbb{R}$. Suppose that $\dim(M) = n$, and let Σ be an $(n-1)$-dimensional surface transverse to the vector field at x_0, see Figure 14. By this, we mean that $\langle v(x_0)\rangle \oplus T_{x_0}\Sigma = T_{x_0}M$. Then we can find an open set $U \subset \Sigma$ containing x_0 such that for all $x \in U$, there exists a $\tau(x)$ close to T such that $\varphi(x, \tau(x)) \in \Sigma$.

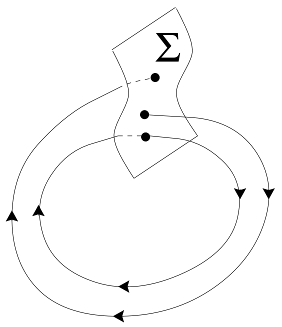

Figure 14: The surface Σ and two trajectories.

The point $\varphi(x,\tau(x))$ is called the *first return* of x, and the map R which associates a point with its first return is called the *return map*: $R(x) = \varphi(x, \tau(x))$. Note that, by construction, we have $\tau(x_0) = T$ and $R(x_0) = x_0$. Thus a fixed point of R corresponds to a periodic orbit of (45), and a periodic point of period k corresponds to a periodic orbit of (45) piercing Σ k times before closing. The following theorem states that the return map is as smooth as the vector field:

Theorem A.2 *Under these conditions, and for sufficiently small U, the return map is a C^r diffeomorphism of U onto a subset of Σ.*

This means that the partial derivatives of R are well defined. Once we have the return map and its derivative, we can employ the machinery described in the previous sections: we say that a periodic orbit of (45), γ, is hyperbolic if any member in $\Sigma \cap \gamma$ is a hyperbolic periodic point for return map R. Likewise, we say that a set $\mathcal{A}$, which is flow-invariant, is hyperbolic if its intersection with Σ is a hyperbolic set for the R. All definitions concerning attractors can be carried over to flows by substituting f^i, $i \in \mathbb{N}$ for $\varphi(\cdot, t)$, $t \geq 0$.

References

[1] C. Bonnati, A. Pumariño, and M. Viana, Lorenz attractors with arbitrary expanding directions, *C. R. Acad. Sci. Paris* **325**, Série I (1997), 883–888.

[2] L. H. de Figueiredo and J. Stolfi, *Métodos numéricos auto-validados e aplicações*, Braz. Math. Colloq. **21**, IMPA, Rio de Janeiro, 1997.

[3] Z. Galias and P. Zgliczyński, Computer assisted proof of chaos in the Lorenz equations, *Physica D* **115** (1998), 165–188.

[4] J. Guckenheimer, A strange, strange attractor. In *The Hopf Bifurcation and its Applications* (Marsden and McCracken, eds.), Springer-Verlag, New York, 1976.

[5] J. Guckenheimer and R. F. Williams, Structural Stability of Lorenz Attractors, *Publ. Math. IHES* **50** (1979), 307–320.

[6] S. P. Hasting and W. C. Troy, A Shooting Approach to the Lorenz Equations, *Bull. Amer. Math. Soc.* **27** (1992), 298–303.

[7] M. W. Hirsch, C. C. Pugh, and M. Shub, *Invariant Manifolds*, Lect. Notes in Math. **583**, Springer-Verlag, New York, 1977.

[8] O. Knüppel, PROFIL – Programmer's Runtime Optimized Fast Interval Library. Technical Report 93:4, Technical University Hamburg-Harburg, (1993). Available from `http://www.ti3.tu-harburg.de`

[9] E. N. Lorenz, Deterministic Non-periodic Flow, *J. Atmos. Sci.* **20** (1963), 130–141.

[10] S. Luzzatto and W. Tucker, Non-uniformly expanding dynamics in maps with singularities and criticalities, *Publ. Math. IHES* **89** (1999), 179-226.

[11] S. Luzzatto and M. Viana, Lorenz attractors without invariant foliations, In preparation.

[12] K. Mischaikow and M. Mrozek, Chaos in the Lorenz Equations: A Computer-Assisted Proof, *Bull. Amer. Math. Soc.* **32** (1995), 66–72.

[13] J. Milnor, On the Concept of Attractor, *Commun. Math. Phys.* **99** (1985), 177–195.

[14] R. E. Moore, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.

[15] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM Studies in Applied Mathematics, Philadelphia, 1979.

[16] J. Palis and F. Takens *Hyperbolicity & sensitive chaotic dynamics at homoclinic bifurcations*, Cambridge University Press, Cambridge, 1993.

[17] D. Rand, The topological classification of Lorenz attractors, *Math. Proc. Camb. Phil. Soc.* **83** (1978), 451–460.

[18] C. Robinson, *Dynamical Systems* [2nd edition], CRC Press, New York, 1995.

[19] C. Robinson, Homoclinic Bifurcation to a Transitive Attractor of Lorenz Type, *Nonlinearity* **2** (1989), 495–518.

[20] M. Rychlik, Lorenz attractors through a Sil'nikov-type bifurcation. Part 1, *Ergod. Th. & Dynam. Sys.* **10** (1989), 793–821.

[21] C. L. Siegel and J. K. Moser, *Lectures on Celestial Mechanics*, Springer-Verlag, New York, 1971.

[22] S. Smale, Mathematical problems for the next century, *Math. Intelligencer* **20**, No. 2 (1998), 7–15.

[23] C. Sparrow, *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*, Springer-Verlag, New York, 1982.

[24] M. Viana, *Stochastic Dynamics of Deterministic Systems*, Braz. Math. Colloq. **21**, IMPA, Rio de Janeiro, 1997.

[25] R. F. Williams, The Structure of Lorenz Attractors, *Publ. Math. IHES* **50** (1979), 321–347.

[26] J. A. Yorke and E. D.Yorke, Metastable chaos: The transition to sustained chaotic oscillations in a model of Lorenz, *J. Stat. Phys.* **21** *(1979), 263-277.*