Rigorous Shadowing of Numerical Solutions of Ordinary Differential Equations by Containment

by

Wayne Brian Hayes

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Rigorous Shadowing of Numerical Solutions
of Ordinary Differential Equations by Containment

Wayne Brian Hayes
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2001

An exact trajectory of a dynamical system lying close to a numerical trajectory is called a *shadow*. We present a general-purpose method for proving the existence of finite-time shadows of numerical ODE integrations of arbitrary dimension in which some measure of hyperbolicity is present and there is either 0 or 1 expanding modes, or 0 or 1 contracting modes. Much of the rigor is provided automatically by interval arithmetic and validated ODE integration software that is freely available. The method is a generalization of a previously published *containment* process that was applicable only to two-dimensional maps. We extend it to handle maps of arbitrary dimension with the above restrictions, and finally to ODEs. The method involves building n-cubes around each point of the discrete numerical trajectory through which the shadow is guaranteed to pass at appropriate times. The proof consists of two steps: first, the rigorous computational verification of an *inductive containment property*; and second, a simple geometric argument showing that this property implies the existence of a shadow. The computational step is almost entirely automated and easily adaptable to any ODE problem. The method allows for the rescaling of time, which is a necessary ingredient for successfully shadowing ODEs. Finally, the method is local, in the sense that it builds the shadow inductively, requiring information only from the most recent integration step, rather than more global information typical of several other methods. The method produces shadows of comparable length and distance to all currently published results.

Acknowledgements

Our strength as a species comes from our ability to communicate with each other. Very few feats, scholarly or otherwise, can be accomplished in a vacuum. Without the ideas, help, and challenges from professional collegues, and without the warmth, care, and compassion of friends and family, our work would be non-existent or meaningless.

On the professional side, I thank my committee. Their doors were always open, both for professional consultation, and for occasional personal discussion. The guidance of my supervisor, Ken Jackson, both around obstacles and out of dead ends, is much appreciated. On many occasions Wayne Enright provided much-needed guidance when my understanding of accuracy and stability issues went awry, and his keen eye for precise use of terms tightened my presentation in several places in the thesis. Tom Fairgrieve's experience with nonlinear chaotic systems helped me to view the larger picture, and his probing questions always made me think carefully about what was important. Ted Shepherd's practical experience in numerical techniques for physics problems and deep understanding of classical mechanics provided me with an even wider view than I would have ever thought possible. I cannot thank them enough.

This thesis was inspired by a desire to question, and ultimately to validate, the reliability of graviational n-body integrations. Of course, that turns out to be far too large and idealistic a goal, but along the way I had the aid of several astronomers and physicists who helped keep me honest and informed about how my work related to that goal. Gerald Quinlan and Scott Tremaine, in particular, wrote the paper that ultimately led to this work, and they paid me the highest compliment that could be paid to a graduate student at the beginning of his research career: initially enthusiastic and continuing interest in my work and how it related to their own.

I thank Paul Selick and Robert Corless for agreeing to be external examiners of my thesis, and for helpful comments on content and presentation of the ideas.

Other professional colleagues who helped along the way include: Ned Nedialkov, for guidance into the use of his validated ODE integrator; Jeff Tupper, for insightful late-night discussions about interval arithmetic and philosophy; James Stewart, for convincing me that inequalities were necessary in the definition of the inductive containment property; John Pryce, for the chance to give an impromptu talk on shadowing at a SIAM conference when a slot suddenly opened up; John Pryce and Jens von Bergmann for several days of discussions in trying to extend the proofs to arbitrary dimension; Danny House, for realizing that my problems were related to homotopy theory, and for pointing me towards Paul Selick; Jim Clarke, for putting up with an annoying young Lecturer and teaching assistant for several years; and Martha Hendriks, for being a den mother to all of us Comp. Sci. students.

My office-mate, Luis Dissett, was a great companion these past years, and his incredible ability to immediately provide small proofs and counter-examples guided me through several minor emergencies. His deep yet joyfully held views on life, philosophy, mathematics, and religion provided many needed hours of distraction.

I had the pleasure of lecturing several courses during my graduate career. I found the experience extremely rewarding, for I believe that there is no greater gift to a teacher than to be given the privilege to teach students who are willing to learn. Watching your faces as confusion turned to insight will be a sight I'll not soon forget.

Friends and family are indirectly part of any endeavor. I would like to thank my mother, E. Merle Hayes, for single-handedly raising a fine boy (or so I'm often told), and for supporting me and encouraging me to explore the world, and for sharing that exploration, throughout my childhood, and extending into my adult life. Without her there's no telling where I'd be or what (different kinds of) trouble I'd be in.

Each of us has a certain group of friends that are likely to last a lifetime. For me, it is the group I met as an undergraduate in the Department of Computer Science at the University of Toronto in the late 1980's and early 1990's. You are affectionately known as the "Cabal", and you know who you are.

The number of friends I've made in graduate school boggles the mind. They are too numerous to mention here. For many, the road has been long and hard, and graduation is close-at-hand, or a goal already reached. For the rest of you, the adventure has just begun!

Life outside of CS must exist as well, for a balanced life. I made many friends at the University of Toronto Outing Club. Some of them even aided my research. I thank Martin and Marlene for a front seat with a map light during a long, dark drive home, so that I could work on several inspirations that occurred after a long weekend of canoeing in Killarney.

Finally, I would like to thank the people of Canada, and of Ontario, for the public funding that supported my research and its publication. In the end, research is about improving people's lives and the environment in which we live. This must never be forgotten, and public funding of research must never die, for without it we will be left only with the views of corporations and governors.

Of course, any remaining errors or omissions are solely my responsibility.

Contents

Chapter 1

Introduction, motivation and background

1.1 Ordinary differential equations

The subject of *ordinary differential equations* (ODEs) concerns the study of solutions to equations of the form

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \tag{1.1}$$

where $\mathbf{y}(t) = (y_1(t), \ldots, y_n(t))^T$ is an n-dimensional vector, $\mathbf{y}'(t) = \frac{d\mathbf{y}(t)}{dt}$, and $\mathbf{f} : \mathbf{R} \times \mathbf{R}^n \to \mathbf{R}^n = (f_1(t, \mathbf{y}(t)), \ldots, f_n(t, \mathbf{y}(t)))^T$ is a vector-valued function. Good introductory texts on the subject abound; see for example Braun (1983, 1993) for a pleasant and readable undergraduate-level introduction. If $\mathbf{f}$ depends on t as above, the ODE is called *nonautonomous*; otherwise it is called *autonomous*. The nonautonomous ODE (1.1) can be converted into an autonomous one by adding one more variable, say $y_{n+1}(t)$, and letting $y'_{n+1}(t) = 1$, $y_{n+1}(t_0) = t_0$, then substituting y_{n+1} wherever t appears on the right hand side. We will concern ourselves in this thesis mostly with autonomous ODEs, keeping in mind that we can solve the nonautonomous case either by using the above substitution, or by straightforward extensions to our algorithms.

The subject of *initial value problems* (IVPs) for autonomous ODEs concerns the solution of

$$\begin{aligned} \mathbf{y}'(t) &= \mathbf{f}(\mathbf{y}(t)), && (1.2) \\ \mathbf{y}(t_0) &= \mathbf{y}_0, && (1.3) \end{aligned}$$

where (1.3) is called the *initial condition.* If $\mathbf{f}$ is bounded and Lipschitz continuous in a domain D[1], then the solution to (1.2,1.3) exists and is unique while it remains in D (Ascher, Mattheij, and Russell 1988, §3.1). Let $\mathbf{y}(t; t_0, \mathbf{y}_0)$ be the solution of (1.2,1.3). Define the time-h solution operator φ_h to be

$$\varphi_h(\mathbf{x}) \equiv \mathbf{y}(h; 0, \mathbf{x}), \tag{1.4}$$

keeping in mind that $\mathbf{y}(t; t_0, \mathbf{x}) \equiv \mathbf{y}(t - t_0; 0, \mathbf{x})$ for autonomous systems.

[1]A function $\mathbf{f}(\mathbf{y})$ is Lipschitz continuous in a domian D if $\forall \mathbf{x}, \mathbf{y} \in D$, $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| < C\|\mathbf{x} - \mathbf{y}\|$ for some constant C.

1.1.1 Error analysis of numerical solutions to ODEs

It is not possible, in general, to solve (1.2,1.3) in closed form (Braun 1983, §1.9). In fact, most initial value problems cannot be solved in closed form. Thus, approximate methods for the solution of (1.2,1.3) must be used. We restrict our discussion to one-step methods. A one-step method consists of building an approximation $\tilde{\varphi}_h$ to φ_h for small h, and then computing a sequence of discrete points $\mathbf{y}_{i+1} = \tilde{\varphi}_{h_i}(\mathbf{y}_i)$ representing approximations to $\mathbf{y}(t_{i+1}; t_0, \mathbf{y}_0)$ where $t_{i+1} = t_i + h_i$. See, for example, Dahlquist and Björck (1974) or Kahaner, Moler, and Nash (1989) for an undergraduate-level introduction, or Hairer, Nørsett, and Wanner (1993) for a more advanced exposition. We will term such a discrete sequence of points a *pseudo-trajectory.* If the pseudo-trajectory satisfies a *local error tolerance* of δ such that $\|\mathbf{y}_{i+1} - \varphi_{h_i}(\mathbf{y}_i)\| \leq \delta$, then we will call it a *δ-pseudo-trajectory.*

The natural first question to ask about pseudo-trajectories is how accurately they approximate the exact solution. Several approaches have been developed to aid in answering this question. *Forward error analysis* is the most straightfoward, and refers to the evolution of $\|\mathbf{y}_i - \mathbf{y}(t_i; t_0, \mathbf{y}_0)\|$. In general the best bound one can put on this *forward error* is an exponential one,

$$\|\mathbf{y}_i - \mathbf{y}(t_i; t_0, \mathbf{y}_0)\| \leq \frac{\delta}{hL}(e^{L|t_i - t_0|} - 1), \tag{1.5}$$

where δ is the local error and L is a bound on the *logarithmic norm* of the Jacobian of $\mathbf{f}$ (Dahlquist and Björck 1974, §§8.1.2, 8.3.6). If L is negative, then the error is uniformly bounded; otherwise, the error may be unbounded, and more sophisticated methods of error analysis must be used to gain insight into the value of the numerical solution. If a numerical method of order p has a stepsize bounded by h, then δ is $O(h^{p+1})$, and the right side of equation (1.5) becomes $O(e^{Lt}h^p/L)$ for time t.

Backward error analysis is a general term applied to methods of error analysis that relate the pseudo-trajectory to the exact solution of a nearby problem (Corless 1994a). *Defect based* backward error analysis requires a piecewise differentiable interpolant $\mathbf{x}(t)$ of the pseudo-trajectory, and then defines the *defect* as

$$\delta(t) = \mathbf{x}'(t) - \mathbf{f}(\mathbf{x}(t)). \tag{1.6}$$

If for some input tolerance ε we can show that $\|\delta(t)\| \leq \varepsilon$ wherever $\mathbf{x}'(t)$ is defined over the whole of the interpolated solution, then the interpolated solution is the exact solution to an ε-close problem (Corless and Corliss 1992; Corless 1994a). The method of *modified equations* is a special case of defect analysis in which, given a particular numerical method and a particular problem, we can write down an algorithm to *compute* $\delta(t)$ using a series expansion, although it is extremely tedious. Ahmed and Corless (1997) have implemented a prototype with the aid of the *Maple* symbolic manipulation package (Char 1993).

Defect-based and other backward error analysis methods modify (1.2) but leave (1.3) untouched. In contrast, shadowing is a method of backward error analysis in which (1.2) remains fixed while (1.3) is allowed to change. In other words, a *shadow* is an exact solution to (1.2) that remains close to the pseudo-trajectory, but that has slightly different initial conditions than the pseudo-trajectory. Shadowing is thus best applied to systems in which the governing equations are extremely well-known, and virtually all error is introduced by imprecise knowledge of initial conditions and/or by numerical error in the computation of the solution. It is less applicable to systems in which the mathematics only approximately model the truth.

1.2 Motivation

This thesis was inspired by a study of the reliability of the numerical simulation of physical systems, particularly simulations of the gravitational n-body problem. Many physical systems under active study today can be modelled using ODEs; however, many of them display *sensitive dependence on initial conditions*, which means that two solutions that are initially close to each other tend to diverge exponentially with time. Since numerical methods introduce small errors that produce a pseudo-trajectory rather than an exact solution, it is virtually guaranteed that a pseudo-trajectory of such an ODE will diverge exponentially away from the exact solution with the same initial conditions. Although this is widely recognized, its impact on the qualitative properties of a pseudo-trajectory of an ODE is not well understood.

Corless (1994a) argues that backward error analyses that modify (1.2) are often adequate because mathematical modelling *always* requires approximation and neglect of small effects:

> One neglects, for example, the effect of the gravitational attraction of Jupiter on one's earthbound experiment .... Similarly, one ignores 'small' stochastic terms in ordinary differential equation models of many phenomena, or 'small' nonautonomous perturbations of physics experiments (such as the effect of passing trucks). So a numerical analysis of methods of solving ODEs which puts [numerical] errors on the same basis as modelling, measurement, and data errors would be a completely successful analysis .... We [have to] study the effects of perturbations, of course, but we have to do this even if we know the exact solution of the specified problem. [Corless (1994a)]

Although these are good points, this author is not convinced for the following reasons. Numerical errors may be biased in qualitatively different ways than natural perturbations, and may introduce biases into the numerical solution that cause it to behave in a nonphysical manner. Corless (1992b) has himself noted this. For example, the perturbations mentioned above would not appreciably change the energy of the system under study, whereas spurious energy dissipation can be a major problem in long numerical integrations of systems which should be conservative (Channell and Scovel 1990; Sanz-Serna 1992). Although symplectic integrators (Channell and Scovel 1990; Sanz-Serna 1992) and other types of conservative integrators (Shadwick, Bowman, and Morrison 1999) may conserve certain quantities, it is not clear that they do not introduce new biases, such as nonphysical energy transport. Physical systems often satisfy properties such as symplecticness, conservation of energy, conservation of phase-space volume and conservation of various types of momentum. Many of these are well-conserved in real systems that experience perturbations, but are not well-conserved by many otherwise well-behaved numerical methods.[2] This has been confirmed with several symplectic maps using a fixed-timestep 4th-order Runge-Kutta integrator (Channell and Scovel 1990), and by this author using the n-body problem and comparing a 7/8 order Runge-Kutta pair (Enright 1993), two Adams's methods (Hindmarsh 1980; Kahaner, Moler, and Nash 1989), and a Bulirsch-Stoer method (Press, Teukolsky, Vetterling, and Flannery 1992) to the leapfrog method, which is a 2nd-order symplectic method. In general, we want to ensure that changing the field represented by (1.2) does not affect any quantities of interest (Skeel 1996; Skeel 1999).

[2] The effect on simulations of numerical error can be much greater than actual perturbations, even if those perturbations are larger. For example, nearby stars and the Galaxy at large exert forces on the Solar System that are at least 10^{-12} as large as the forces from our Sun. It is not difficult to create integrations with numerical errors several orders of magnitude smaller than this, and yet unless these integrations somehow account for symplectic structure or energy conservation, they produce an integration of the Solar System which quickly and clearly diverges from the behaviour of the real Solar System.

As a subtly different example, a close encounter between particles in a gravitational n-body integration involves forces between the participating particles which are so great that physical perturbations from other parts of the system are negligible. The numerical errors introduced during the encounter can have a far greater effect than any physical perturbations. Close encounters are very hard to integrate numerically with precision, and are well-known to be the bane of gravitational n-body integration (Aarseth 1999).

Despite all of this, numerical solutions often appear to mimic with astounding accuracy the phenomena they purport to simulate. Simulations of galaxies often closely resemble real galaxies (see almost any paper on galaxy simulation in Clarke and West (1997) or Merritt, Sellwood, and Valluri (1999)). Even galaxy collisions can be modelled in a convincing manner (Struck 1997). More generally, exponential divergence of nearby trajectories implies that an initially dense ensemble of points will disperse into a uniform distribution in a relatively short time (Skeel 1996; Skeel 1999). This effect is also seen in numerical simulations of chaotic systems (Merritt and Valluri 1996; Merritt 1999). The natural question to ask is whether these simulations are behaving in a fashion similar to real systems, or if they only superficially mimic real systems but are in fact behaving incorrectly at a more fundamental level. If this were the case, then we could be lulled into a false sense of security whilst our understanding of these systems becomes compromised.

Since shadowing disallows changes in the model (1.2), some of these kinds of insidious errors can be ruled out. Furthermore, if the problem (1.2) arises from a purely mathematical context and not a physical one, we may be earnestly interested in the properties of exact solutions of (1.2), in which case a recourse to shadowing may be the only option. The only remaining question would then be whether shadows are typical of exact solutions chosen at random.

On the other hand, rigorous shadowing as presented in this thesis and elsewhere is extremely expensive. Whereas defect controlled methods are of roughly equal expense compared to more traditional integration methods, rigorous shadowing requires validated ODE integration, which at present tends to be several orders of magnitude more expensive in both time and memory than non-validated methods, even for low-dimensional problems.

1.3 Background

This section covers material that is required to understand later chapters. It may be omitted by those already familiar with the concepts of interval arithmetic and validated ODE integration.

1.3.1 Interval arithmetic

The proofs that will be elucidated in Chapter 3 are computer aided proofs. In particular, they rely on the rigorous computational verification of some properties that are computed using floating-point arithmetic. Floating-point arithmetic is inexact; it is impractical in general to store the exact result of the addition of even two machine-representable numbers, much less compute complicated functions of them. *Error analysis* is the study of how these floating-point errors, and other numerical errors, affect the results of numerical computations (Dahlquist and Björck 1974; Hager 1989; Kahaner, Moler, and Nash 1989). *Computational interval arithmetic* is the study of how to automatically maintain rigorous yet tight bounds on the errors incurred in computations involving floating-point numbers, and software packages exist to implement computational interval arithmetic algorithms. In this thesis, we use the packages described in Nedialkov (1999). More background on interval arithmetic may be found in Moore (1966) and Alefeld and Herzberger (1983). An up-to-date, practical implementation is discussed at a very abstract level in Tupper (1996).

Interval arithmetic packages maintain upper and lower bounds on floating-point computations that are guaranteed to enclose the exact value of a computation. Let the symbol $+^{\uparrow}$ represent floating-point addition in which the machine *rounds up* the answer (towards $+\infty$) to the closest machine-representable number. Similarly, let $+_{\downarrow}$ represent floating-point addition in which the machine *rounds down* (towards $-\infty$). To compute an interval that is guaranteed to enclose the exact sum of two floating-point numbers a and b, we compute

$$\underline{c} = a +_{\downarrow} b,$$

$$\overline{c} = a +^{\uparrow} b.$$

Then the interval $[\underline{c}, \overline{c}]$ is guaranteed to enclose the exact sum, $a + b$. Similarly, we can add two intervals $c = [\underline{c}, \overline{c}]$ and $d = [\underline{d}, \overline{d}]$ by performing

$$[\underline{e}, \overline{e}] := [\underline{c}, \overline{c}] \oplus [\underline{d}, \overline{d}] \equiv [\underline{c} \ +_{\downarrow} \ \underline{d}, \overline{c} \ +^{\uparrow} \ \overline{d}].$$

Similar operations can be provided for subtraction, multiplication, and division, although the latter two are slightly more complicated. Interval arithmetic subroutines can also be provided for computing sin, cos, tan, log, exp, and all the standard elementary functions. These can be implemented, for example, by summing a Taylor series expansion in interval arithmetic and then bounding the remainder term in interval arithmetic as well, giving a rigorous bound on the total error. To obtain a *tight* bound, however, often requires more sophistication.

An interval whose upper and lower bounds are equal is called a *point interval*, sometimes refered to as a *thin interval* by other authors. Interval vectors and matrices are represented by vectors and matrices of intervals. An interval vector may be thought of as an axis-aligned "box" that represents all the real numbers inside the box.

The *width* of an interval is

$$w([a]) = \overline{a} - \underline{a}.$$

An important goal of interval arithmetic packages is to keep the width of the intervals as small as possible while still enclosing the exact solution. If the width of an interval becomes too large, virtually no information remains about the value of the exact solution, even though the interval encloses the exact solution. The *midpoint* of an interval $[a]$ is

$$m([a]) = (\overline{a} + \underline{a})/2.$$

The width and midpoint of interval vectors and matrices are defined component-wise.

In general, let $\mathbf{g} : \mathbf{R}^n \to \mathbf{R}^m$ and let $\bar{\mathbf{g}}$ be an interval arithmetic algorithm that attempts to compute $\mathbf{g}$. Then given any n-dimensional input interval vector $[\mathbf{x}]$, $\bar{\mathbf{g}}$ must either fail explicitly or produce as output an m-dimensional interval vector $[\mathbf{y}]$ satisfying

$$\forall \mathbf{x} \in [\mathbf{x}],\ \mathbf{g}(\mathbf{x}) \in [\mathbf{y}].$$

Although interval techniques are very powerful, they cannot be used in a blind or naïve manner, and some caveats must be noted. For example, although interval addition and multiplication are associative, the distributive law does not hold in general. That is, we can easily find three intervals $[a]$, $[b]$, and $[c]$ for which

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

Some special cases of the distributive law hold, for example if $[b][c] \geq 0$, if $[a]$ is a point interval, or if both $[b]$ and $[c]$ are symmetric about 0 (Nedialkov 1999, §2.1). Moreover, the *subdistributive law*

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

always holds. In general, interval methods can be tricky, and consequently their use requires some degree of sophistication from the user.

1.3.2 Validated ODE integration

A validated ODE integrator is an algorithm that uses, among other things, interval arithmetic to produce an interval vector that, in the forward error sense, is guaranteed to enclose the exact solution of the initial value problem (1.2,1.3). There are exactly two sources of error in numerical integrations: roundoff error, which can be accounted for by interval arithmetic, and truncation error, for which we need some theoretical bound on the error in the method. Most validated ODE integrators use a Taylor series to approximate φ_h, in which the remainder term computed in interval arithmetic bounds the truncation error (Nedialkov, Jackson, and Corliss 1999). An impressive accomplishment in this direction has been the development of software that can automatically differentiate code lists and compute Taylor series at run time (Bendtsen and Stauling 1996; Bendtsen and Stauling 1997), making it almost trivial to automatically generate Taylor series.

If the interval vector representing an enclosure of a validated integration is an axis-aligned box at time t_0, its image under the evolution of the ODE is unlikely to remain axis-aligned for time $t \neq t_0$. Many methods have been devised to account for this so-called *wrapping effect*. We use the software developed by Nedialkov (1999). The following description derives from Nedialkov (1999) and Nedialkov, Jackson, and Corliss (1999), which may be consulted for further details. The enclosure of a solution is represented by

$$\hat{\mathbf{y}}_i + A_i[\mathbf{r}_i] \equiv \{\mathbf{y}_i\} \supseteq \varphi_h(\{\mathbf{y}_{i-1}\}), \tag{1.7}$$

where $\hat{\mathbf{y}}_i$ is a point vector representing the approximate solution, $[\mathbf{r}_i]$ is an interval vector enclosing the zero vector, A_i is a point matrix providing a linear transformation (rotation, scaling, skewing, *etc.*) to $[\mathbf{r}_i]$, $\{\mathbf{y}_i\}$ is a shorthand for $\hat{\mathbf{y}}_i + A_i[\mathbf{r}_i]$, φ_h is the solution operator (1.4), and $\varphi_h(\{\mathbf{y}_{i-1}\})$ is the pointwise application of φ_h to $\{\mathbf{y}_{i-1}\}$. Thus, $\{\mathbf{y}_i\}$ can be thought of as a point vector numerical solution $\hat{\mathbf{y}}_i$ with a linearly transformed bounded error box $A_i[\mathbf{r}_i]$ around it. In our implementation, $[\mathbf{r}_i]$ is evolved using

$$[\mathbf{r}_i] = A_i^{-1}([S_{i-1}]A_{i-1})[\mathbf{r}_{i-1}] \quad + \quad A_i^{-1}([\mathbf{z}_i] - \mathbf{z}_i), \tag{1.8}$$

where $[S_{i-1}]$ is an interval approximation to the solution of a variational equation from t_{i-1} to t_i. Note that although $[S_{i-1}]$ as used in (1.8) rigorously evolves $[\mathbf{r}_i]$ to maintain enclosure of the solution, in general $[S_i]$ does not provide a rigorous bound on the solution of the variational equation. The first term of (1.8) is responsible for propagating the error from the previous step, while the second term is the new error introduced at the current step, with $[\mathbf{z}_i]$ being derived from the remainder term of the Taylor expansion for step i and $\mathbf{z}_i$ being the midpoint of $[\mathbf{z}_i]$. The matrix A_i is meant to provide a linear transformation to $[\mathbf{r}_i]$ that reduces wrapping, and is currently computed by performing a QR factorization

$$Q_i R_i = m([S_{i-1}])A_{i-1},$$

(where Q_i is an orthogonal matrix and R_i is an upper triangular matrix) and setting $A_i \equiv Q_i$. As a side remark, note that if we were to take $A_i = m(S_{i-1})A_{i-1}$, then $A_i^{-1}([S_{i-1}]A_{i-1})$ would approximate the identity, to within the width of $[S_{i-1}]$ times a constant depending on A_i^{-1} and A_{i-1}. In fact, A_i can be *any* nonsingular matrix, although some choices are better than others for providing tight enclosures. That is, any nonsingular choice of A_i will produce an $[\mathbf{r}_i]$ which is a proper enclosure of the solution as long as $[\mathbf{r}_i]$ is computed with equation (1.8). The orthogonalization is performed only because empirical evidence has demonstrated that the condition number of A_i has a large effect on the evolution of the width of $[\mathbf{r}_i]$, and orthogonalizing A_i appears to reduce this width. Future analyses may provide better choices for A_i or even entirely different ways of computing $[\mathbf{r}_i]$ (Nedialkov 1999; Nedialkov, Jackson, and Corliss 1999; Nedialkov and Jackson 2000).

The enclosures constructed by any interval arithmetic package are unlikely to be optimal. The amount by which they over-estimate the error is called the *excess*. It is worth noting that the excess for the methods used in this thesis is probably quite large. For example, this author has computed pseudo-trajectories for the initial value problems discussed in Chapter 4 using many diverse numerical methods including the three Runge-Kutta methods (i) the classic 4th order one (Press, Teukolsky, Vetterling, and Flannery 1992), (ii) one order 5/6 pair (Hull, Enright, and Jackson 1976), and (iii) one order 7/8 pair (Enright 1993); two Adams methods (Hindmarsh 1980; Kahaner, Moler, and Nash 1989); and a Bulirsch-Stoer method (Press, Teukolsky, Vetterling, and Flannery 1992). In most cases, these integrators all agreed with each other, and with the approximate solution $\hat{\mathbf{y}}$ provided by Nedialkov (1999), to a precision several orders of magnitude higher than the width of $[\mathbf{r}_i]$. Although all of these methods are ultimately based on local Taylor series approximations, they are algorithmically very diverse and we believe they are unlikely all to be biased in a similar fashion. Thus, we consider this to be strong evidence that the width of $[\mathbf{r}_i]$ is a gross upper bound on the error, and that much further work is needed in the area of providing tight enclosures of solutions to IVPs.

Although packages exist that can handle the wrapping effect more effectively, they are considerably more expensive. For example, if n is the number of equations in the system and k is the order of the Taylor series used to approximate one integration step, then the computational complexity of COSY INFINITY (Berz 1997; Berz and Makino 1998) is $\binom{n+k}{n}$ per step. This is a high-degree polynomial if either n or k is fixed and either is of nontrivial size, and is exponential if both n and k are allowed to grow, whereas the compuational complexity of our code is approximately $O(Nnk^2)$ (Nedialkov 1999), where N is the number of operations needed to compute the Jacobian of (1.2).

Finally, it is interesting to note that shadowing can be thought of as a generalization to validated ODE integration. Traditional validated ODE integration involves finding a bound enclosing *the* exact solution that starts at $t = 0$ at the exact point initial condition given. Since numerical solutions are not exact any time after time zero, it seems rather artificial to insist that they be exact *at* time zero. Considering that shadows can last many orders of magnitude longer than validated ODE integrations[3], perhaps it is better to say that numerical trajectories are valid approximations of exact trajectories as long as we do not require that they exactly satisfy the initial conditions, and instead treat time zero on an equal footing with all other times (Murdock 1995).

[3]For example, Nedialkov (1999) typically finds a validated solution of the Lorenz equations lasting about 25 time units, whereas shadows of the Lorenz system can last for anywhere from hundreds to *millions* of time units.

1.4 Thesis outline

In Chapter 2, we present a tutorial introduction, a survey of previous work, and some discussion of shadowing. Chapter 3 contains the bulk of the original work of the thesis, detailing our algorithms and theorems for proving the existence of shadows. Chapter 4 contains results of our shadowing experiments and comparison of our results to previously published work. Chapter 5 contains a brief discussion of future directions for this research. Finally, the Glossary contains definitions of terms with which the reader may not be familiar.

Chapter 2

A brief survey of shadowing results

Dynamical systems often display sensitive dependence on initial conditions: a small change at any point in an orbit produces a new orbit that tends to exponentially diverge from the original one, leading to a vastly different solution a short time later. Since a numerical method introduces small perturbations arising from roundoff and truncation error, we must naturally ask what effect these errors have on the validity of numerical solutions.

2.1 Introduction

2.1.1 Definitions

In this thesis, an *orbit* is a discrete sequence of points, a *solution* is a continuous curve, and a trajectory more generally refers to either an orbit or a solution depending upon context. The prefix *pseudo-* will be used to denote an approximate orbit, solution, or trajectory, although sometimes it will be omitted if the meaning is clear from the context. We assume a well-scaled problem where all macroscopic quantities of interest are of order unity; $|\cdot|$ denotes the magnitude of a scalar, while $\|\cdot\|$ denotes a norm of a vector or matrix. We use the max norm unless otherwise noted.

Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a function.

In this thesis, $\varphi(\mathbf{x})$ will usually be a diffeomorphism representing the one-timestep flow through $\mathbf{x}$ of the solution to an ODE. If the timestep is fixed, then φ is the same function on each step, but if the timestep is allowed to vary, φ may change from step to step and we will introduce a subscript, using φ_i for step i. For now, we leave φ unsubscripted. For a discrete map, φ may be a simple equation, such as the logistic equation $\varphi(x) = 1 - 2x^2$, which maps the interval $[-1, 1]$ onto itself.

Definition. The *iterated map* $\varphi^i(\mathbf{x})$ is the result of repeatedly composing φ with itself i times, *i.e.*, $\varphi^i(\mathbf{x}) = \underbrace{\varphi(\varphi(\ldots\varphi}_{i \text{ times}}(\mathbf{x})\ldots))$.

Definition. An *exact orbit* $\{\mathbf{x}_i\}_{i=j}^{k}$ of φ satisfies $\mathbf{x}_{i+1} = \varphi(\mathbf{x}_i)$, *i.e.*, $\mathbf{x}_i = \varphi^{i-j}(\mathbf{x}_j)$, for $j \le i < k$. We allow $j = -\infty$ and $k = \infty$.

Definition. $\{\mathbf{y}_i\}_{i=j}^{k}$ is a *δ-pseudo-orbit* or *noisy orbit* for φ if $\|\mathbf{y}_{i+1} - \varphi(\mathbf{y}_i)\| \le \delta$ for $j \le i < k$, where δ is called the *noise amplitude*.

For a discrete map, δ can be as small as the machine epsilon; for both discrete maps and ODE systems, it is a bound on the one-step error.

Definition. For $j \le i < k$, the *one-step error* made between step i and step $i+1$ of the pseudo-orbit $\{\mathbf{y}_i\}_{i=j}^k$ is $\mathbf{e}_{i+1} = \mathbf{y}_{i+1} - \varphi(\mathbf{y}_i)$.

Thus, an exact trajectory is one whose one-step errors are identically zero, and a δ-pseudo-orbit is one whose one-step errors satisfy $\|\mathbf{e}_i\| \leq \delta$ for $j < i \leq k$.

Definition of shadowing. An exact trajectory $\{\mathbf{x}_i\}_{i=j}^k$ ε-*shadows* a pseudo-trajectory $\{\mathbf{y}_i\}_{i=j}^k$ if $\|\mathbf{y}_i - \mathbf{x}_i\| \leq \varepsilon$ for $j \leq i \leq k$.

Definition. The δ'-pseudo-trajectory $\mathbf{Z} = \{\mathbf{z}_i\}_{i=j}^k$ is a *numerical shadow* of the δ-pseudo-trajectory $\mathbf{Y} = \{\mathbf{y}_i\}_{i=j}^k$ if their one-step errors are tightly bounded by δ' and δ, respectively, and $\delta' < \delta$.

In practice, a numerical shadow usually only has smaller error *bounds* than the original noisy orbit, because in most cases neither orbit has rigorously computed error bounds. To have confidence in the value of a numerical shadow, we like its noise to be as small as possible. For a map, the noise should ideally be the machine precision. For an ODE solution, the noise is "as small as possible" using some accurate integrator with its error tolerance set very stringently.

A pleasant introduction to shadowing is provided by Sanz-Serna and Larsson (1993).

Definition. The pseudo-orbit $\{\mathbf{y}_i\}_{i=j}^k$ has a *glitch* at point $i = G_0 < k$ if for some relevant ε there exists an exact trajectory that ε-shadows $\{\mathbf{y}_i\}_{i=j}^{G_0}$, but no exact trajectory exists that ε-shadows $\{\mathbf{y}_i\}_{i=j}^{G}$ for $G > G_0$ (Grebogi, Hammel, Yorke, and Sauer 1990).

Although rigorously disproving the existence of shadows of particular numerical trajectories is a virtually untouched area of research, the failure of a particular method to find a shadow is often cited as evidence that an actual glitch occurs somewhere in the vicinity of the computed end-of-shadow (Grebogi, Hammel, Yorke, and Sauer 1990; Sauer and Yorke 1991; Dawson, Grebogi, Sauer, and Yorke 1994; Sauer, Grebogi, and Yorke 1997; Quinlan and Tremaine 1992; Hayes 1995). This conclusion is not always valid, however (see the discussion following Theorem 2.4, p. 17, in this thesis), and so this author proposes two different terms. The term glitch, or *hard glitch*, should be reserved for the case in which the above definition can be verified, *i.e.*, non-existence of shadows can be *proved*. For example, a function $\varphi : X \to X$ which maps an interval onto itself may produce a numerically generated orbit of the iterated map which lies outside this interval. If a numerically generated point, say x_i, moves more than ε away from the interval X then a glitch is guaranteed. However, the failure of a particular method to find a shadow is a different matter, and for this case the author proposes the term *soft glitch*. For systems such as the n-body problem, the notion of a hard glitch cannot be used without proof because there is *no* point in phase space that is unphysical; that is, in a Newtonian, Euclidian space, particles can have any position and any velocity. Furthermore, small numerical errors are constantly occuring, and if the system is integrated carefully and local errors remain small, there is no obvious point at which one can say, "this behaviour is nonphysical". One can arbitrarily decide, for example, that when the total computed energy of the numerical solution has diverged from the known energy of the system by some chosen amount, the solution is no longer valid. But this is not the spirit of the term "glitch". The spirit of the term seems to be "a point at which all exact trajectories diverge from a numerical one", and currently this can only be proved for simple systems such as the system discussed above.

2.1.2 Tutorial

A simple example of a shadow is provided by Quinlan and Tremaine (1992), hereafter referred to as QT. Let $y'' = y$, which can be re-written as a pair of first order equations as $y' = v,\ \ v' = y$, where v is velocity. If $y(t_0) = v(t_0) = 0$ for any t_0, then the exact solution is $y = v = 0\ \forall t$. Now, assume that $y = v = 0$ for $t < 0$, and assume that the system is solved exactly for all

$t \neq 0$. Introducing a perturbation of size $\Delta v = \varepsilon$ at $t = 0$ gives the following "noisy" solution:

$$y(t) = \begin{cases} 0, & t < 0, \\ \varepsilon \frac{e^t - e^{-t}}{2}, & t > 0. \end{cases}$$

A shadow of this noisy solution is

$$x(t) = \varepsilon e^t/2,$$

which remains within $\varepsilon/\sqrt{2}$ (in phase space) from $y(t)$ for all t.

Next we offer a proof of an almost "trivial" theorem: if a map is contracting, then noisy orbits are shadowed.

Theorem 2.1 (Contracting map shadowing theorem). *Let X be a metric space and let $\varphi : X \to X$ be a continuous, uniformly contracting map, i.e., $\exists \rho < 1$ s.t. $\forall \mathbf{x}, \mathbf{y} \in X$, $\|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\| \leq \rho \|\mathbf{x} - \mathbf{y}\|$. Then for every $\varepsilon > 0$ there exists $\delta > 0$ such that every δ-pseudo orbit remaining in X is ε-shadowed.*

Proof. Assume we are given $\varepsilon > 0$. Let $\delta = \varepsilon(1 - \rho)$. Suppose $\{\mathbf{y}_i\}_{i=j}^{\infty}$ is a δ-pseudo-orbit that remains in X. Let $\mathbf{x}_j = \mathbf{y}_j$ and let $\mathbf{x}_{i+1} = \varphi(\mathbf{x}_i)$ for $i \geq j$, *i.e.,* $\{\mathbf{x}_i\}_{i=j}^{\infty}$ is an exact orbit. We will show by induction on i that $\|\mathbf{x}_i - \mathbf{y}_i\| \leq \varepsilon$ for $i \geq j$.

Base case: $\|\mathbf{x}_j - \mathbf{y}_j\| = 0 < \varepsilon$, by our choice of $\mathbf{x}_j$.

Induction step: Assume $\|\mathbf{x}_i - \mathbf{y}_i\| \leq \varepsilon$ for $i \geq j$. Then

$$\begin{aligned} \|\mathbf{x}_{i+1} - \mathbf{y}_{i+1}\| &= \|\varphi(\mathbf{x}_i) - \varphi(\mathbf{y}_i) + \delta_1\|, \qquad \|\delta_1\| \leq \delta \\ &\leq \|\varphi(\mathbf{x}_i) - \varphi(\mathbf{y}_i)\| + \delta \\ &\leq \rho\varepsilon + \delta \\ &= \varepsilon \end{aligned}$$

□

Remark: Notice that the closer ρ is to zero, the more contractive φ is, so that it can accomodate a larger noise amplitude δ.

Remark: If φ were uniformly expanding, then we would expect pseudo-orbits to exponentially diverge from each other, and from the exact solution. In this case, it is φ^{-1} that is contracting, and we can apply the above theorem in reverse time, as long as $\mathbf{y}_i \subset X \; \forall i \geq a$.

Another instructive way to look at shadowing is in terms of its relation to finding the zero of a function. To wit, let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^{N}$ be a δ-pseudo-trajectory in $\mathbf{R}^n$, and let $\mathbf{E} = \{\mathbf{e}_i\}_{i=1}^{N}$ be the set of one-step errors $\mathbf{e}_{i+1} = \mathbf{y}_{i+1} - \varphi(\mathbf{y}_i)$. Let $\mathbf{g} : \mathbf{R}^{(N+1)n} \to \mathbf{R}^{Nn}$ be a function that takes as input the entire orbit $\mathbf{Y}$ and produces an output which is the set of one-step errors $\mathbf{E}$, *i.e.,* $\mathbf{g}(\mathbf{Y}) = \mathbf{E}$. Since the one-step errors are assumed to be small, $\|\mathbf{E}\|$ is small. That is, $\mathbf{Y}$ may be close to a zero of $\mathbf{g}$, if one exists. A zero of $\mathbf{g}$ would represent an orbit with zero one-step error, *i.e.,* an exact orbit. This is an ideal situation in which to apply a zero-finding method such as Newton's method. If the method converges to an orbit $\mathbf{X}$ which is ε-close to $\mathbf{Y}$, then $\mathbf{X}$ ε-shadows $\mathbf{Y}$. This is the idea behind *refinement* (Grebogi, Hammel, Yorke, and Sauer 1990; Quinlan and Tremaine 1992) which will be discussed in more detail below.

A simple example of a system which is *not* shadowable (by the definitions seen thus far—cf. §2.2.5) is $y'' = 0$, the solution of which is straight-line motion at constant velocity v_0. Assume $v_0 = y'(-\infty) \neq 0$. If noise of size $\delta > 0$ in y' is added at $t = 0$, then the noisy solution has velocity $y' = v_0$ for $t < 0$, and a *different* velocity $y' = v_0 + \delta$ for $t \geq 0$. It is easy to see that any exact solution $y(t)$ with $y'(t) = \hat{v}_0$ for all t will diverge linearly away from the noisy solution inside at least one of the intervals $(-\infty, 0)$ or $(0, \infty)$. Thus, no exact solution exists that remains close to the noisy solution for both $t < 0$ and $t \geq 0$.

2.1.3 Hyperbolicity

One of the most important concepts in shadowing is that of *hyperbolicity*, which is related to *exponential dichotomy.* The following definitions are commonly used in the shadowing literature. See for example Palmer (1988), on which the following description is based. In this section, we will concentrate on *maps*, keeping in mind that we can translate between maps and solutions of ODEs by looking at the time-h solution operator $\varphi_h(x)$ defined in equation (1.4).

Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a diffeomorphism. Let $D\varphi(\mathbf{x})$ be the Jacobian of $\varphi(\mathbf{x})$, which exists, is unique, and is invertible since φ is a diffeomorphism. Every orbit of φ has associated with it a linear difference equation called the *linear variational equation*,

$$\mathbf{z}_{i+1} = D\varphi(\mathbf{x}_i)\mathbf{z}_i. \tag{2.1}$$

A sequence of Jacobians along an orbit can be multiplied together to produce a Jacobian of the corresponding sequence of applications of the map,

$$\Phi(i,j) = \begin{cases} D\varphi(\mathbf{x}_{i-1})\cdots D\varphi(\mathbf{x}_j), & \text{if } i > j. \\ I, & \text{if } i = j. \\ D\varphi(\mathbf{x}_i)^{-1}\cdots D\varphi(\mathbf{x}_{j-1})^{-1}, & \text{if } i < j. \end{cases} \tag{2.2}$$

The linear variational equation (2.1) is said to have an *exponential dichotomy* if there are positive constants K, α and a family of projections P_i such that

$$P_{i+1}D\varphi(\mathbf{x}_i) = D\varphi(\mathbf{x}_i)P_i \qquad \text{for all } i, \tag{2.3}$$

$$\|\Phi(i,j)P_j\| \le Ke^{-\alpha(i-j)} \qquad \text{for } i \ge j, \tag{2.4}$$

$$\|\Phi(i,j)(I-P_j)\| \le Ke^{-\alpha(j-i)} \qquad \text{for } j \ge i. \tag{2.5}$$

By repeated application of (2.3) we obtain the identity

$$P_i\Phi(i,j) = \Phi(i,j)P_j.$$

This means that the projections P_i are invariant with respect to equation (2.1). That is, if $\{\mathbf{z}_i\}_{i=j'}^{k}$ is a solution to (2.1) such that $\mathbf{z}_j$ is in the range (resp. nullspace) of P_j for some j then $\mathbf{z}_i$ is in the range (resp. nullspace) of P_i for all i. Inequalities (2.4–2.5) say firstly, that the P_i are bounded (proof: set $i = j$ in (2.4)) and secondly, that the solutions $\mathbf{z}_i$ of equation (2.1) which lie in the range of P_i decay exponentially in forward time, while those in the nullspace of P_i decay exponentially in backward time (Palmer 1988).

Definition. A trajectory $\mathbf{X} = \{\mathbf{x}_i = \varphi^i(\mathbf{x})\}_{i=j'}^{k}$ for some $\mathbf{x}$ is said to be *hyperbolic under* φ if the linear variational equation

$$\mathbf{z}_{i+1} = D\varphi(\mathbf{x}_i)\mathbf{z}_i \tag{2.6}$$

along $\mathbf{X}$ has an exponential dichotomy. Equivalently, we say that φ is *hyperbolic along* $\mathbf{X}$.

Definition. A set $S \subset \mathbf{R}^n$ is said to be *invariant* under φ if $\varphi(S) = S$.

Definition. A compact invariant set S is said to be *hyperbolic under* φ if every trajectory $\mathbf{X}$ in S is hyperbolic with the same constants K, α, and the projection matrices P_i have a rank which is independent of $\mathbf{X}$. Equivalently, we say that φ *is hyperbolic on* S, or that S and φ form a *hyperbolic system.*

If a system is hyperbolic, then the angle between the stable and unstable subspaces is always bounded away from 0 (Grebogi, Hammel, Yorke, and Sauer 1990).

2.1.4 Pseudo-hyperbolicity

This thesis deals not with hyperbolic systems, but with systems whose pseudo-trajectories are shadowable for finite but nontrivial lengths of time even though they are not hyperbolic. For this to occur, a system must display pseudo-hyperbolicity. We say that a system is *pseudo-hyperbolic* if trajectories of the system tend to have solutions to the variational equation which can be split into two classes, one of which tends to expand exponentially, while the other tends to contract exponentially, both simultaneously and for nontrivial lengths of time. This notion could be made more formal by, for example, attempting to find the two classes of solutions using the common methods described in the next section, and then performing least-squares fits of these solutions to exponential curves.

2.2 Survey

2.2.1 Hyperbolic systems

Shadowing was first discussed by Anosov (1967) and Bowen (1975), in relation to hyperbolic systems. Let S and φ be the invariant set and the map of a hyperbolic system, respectively. In such systems, Anosov (1967) proved that $\forall \varepsilon > 0 \; \exists \; \delta > 0$ such that every infinite-length δ-pseudo orbit remaining in S is ε-shadowed by a true trajectory in S. Bowen (1975) proved that the same result holds if the map is required to be hyperbolic only along trajectories in the vicinity of the pseudo-orbit. Palmer (1988) proved a similar theorem along the way towards using the theory of exponential dichotomies to prove Smale's Theorem (Smale 1965, 1967).

Theorem 2.2 (Hyperbolic set shadowing theorem). *Let S be a compact hyperbolic set for the C^1 diffeomorphism $\varphi : \mathbf{R}^n \to \mathbf{R}^n$. Then given any $\varepsilon > 0$ sufficiently small there exists $\delta > 0$ such that every doubly-infinite δ-pseudo-orbit in S has a unique ε-shadowing orbit.*

Proof. See Palmer (1988), Theorem 3.5. □

Chow and Van Vleck (1992) proved a similar theorem in the case that the function φ is allowed to change at each step. We omit the (rather long and involved) specifications of the hyperbolicity conditions of the following theorem, except to note that when the conditions hold, the difference equation

$$\mathbf{z}_{i+1} = D\varphi_i(\mathbf{x}_i)\mathbf{z}_i$$

has an exponential dichotomy for *all* sequences of functions $\{\varphi_i\}_{i=0}^{\infty}$ if $\mathbf{x}_{i+1} = \varphi_i(\mathbf{x}_i)$. These conditions, of course, tightly restrict the classes of sequences of functions whose orbits can be shadowed; otherwise, shadowing of numerical solutions of ODEs would be trivial!

Theorem 2.3 (Random Diffeomorphism Shadowing Lemma). *Let M be a smooth compact k-dimensional Riemannian manifold and let Diff(M) represent the set of all diffeomorphisms from M to M. Assume further that the [omitted] hyperbolicity conditions are satisfied. Let $\{\mathbf{y}_i\}_{i=0}^{\infty}$ be a sequence of points in M. Then for all $\varepsilon > 0$ sufficiently small $\exists \delta > 0$ such that if there exists a sequence of functions $\{\varphi_i \in \mathit{Diff}(M)\}_{i=0}^{\infty}$ satisfying $\|\mathbf{y}_{i+1} - \varphi_i(\mathbf{y}_i)\| \le \delta$ then there exists a unique sequence $\{\mathbf{x}_i\}_{i=0}^{\infty}$ such that $\mathbf{x}_{i+1} = \varphi_i(\mathbf{x}_i)$ and $\|\mathbf{x}_i - \mathbf{y}_i\| \le \varepsilon$ for all i.*

Proof. See Chow and Van Vleck (1992). □

2.2.2 Containment

For systems that are not hyperbolic, but whose trajectories display pseudo-hyperbolicity for a finite number of iterations of φ, we must be satisfied with proving the existence of finite-length shadows. The first studies of shadows for non-hyperbolic systems appear to be Beyn (1987) and Hammel, Yorke, and Grebogi (1987). Hammel, Yorke, and Grebogi (1988) and Grebogi, Hammel, Yorke, and Sauer (1990) (hereafter GHYS) provide the first proof of the existence of a shadow for a non-hyperbolic system over a non-trivial length of time. Their method consists of two parts. First, they *refine* a noisy trajectory using an iterative method that produces a nearby trajectory with less noise. This procedure will be discussed in more detail below. When refinement converges to the point that the noise is of order the machine precision, they invoke *containment*, which can prove the existence of a nearby exact trajectory. Their method, which we now describe, can be applied only to two-dimensional maps.

Let $\{\mathbf{y}_i\}_{i=a}^{b} \subset \mathbf{R}^2$ be a two-dimensional δ-pseudo-orbit of φ for integers a and b. As i increases, orbits separated from each other by a small distance along the expanding direction diverge on average away from each other, while orbits separated by a small distance along the contracting direction approach each other, on average. The containment process consists of building a parallelogram M_i around each point $\mathbf{y}_i$ of the pseudo-orbit such that two sides $C_i^{\pm 1}$ are separated from each other along the contracting direction, while the other two sides $E_i^{\pm 1}$ are separated along the expanding direction.[1] In order to prove the existence of a shadow, the image of M_i under φ must intersect M_{i+1} such that $\varphi(M_i)$ makes a "plus sign" with M_{i+1} (Figure 2.1).

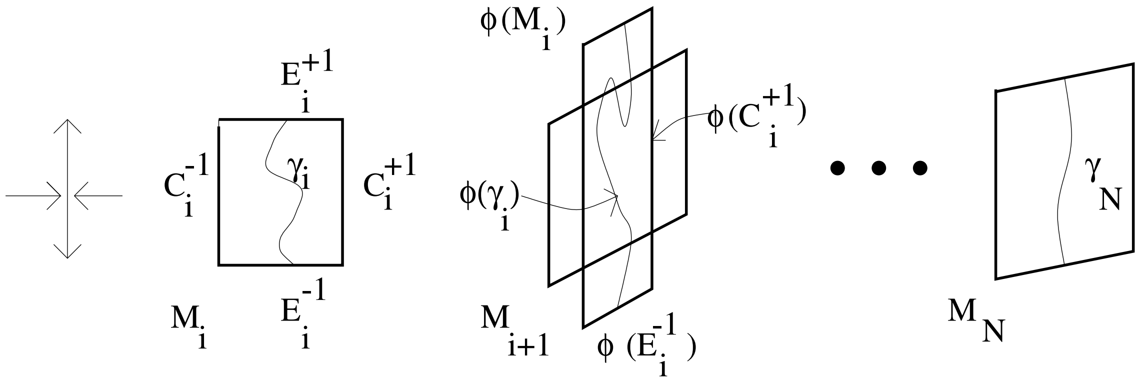

Figure 2.1: Containment in two dimensions, reproduced from GHYS. The horizontal direction is contracting, and the vertical direction is expanding.

The property that GHYS define as a "plus sign" is

$$\varphi(E_i^j) \cap M_{i+1} = \emptyset, \qquad \varphi(M_i) \cap C_{i+1}^j = \emptyset, \;\; j = \pm 1. \tag{2.7}$$

To ensure this occurs, GHYS require a bound on the second derivative of φ, and the expansion and contraction amounts need to be resolvable by the machine precision. The proof of the existence of an exact orbit then relies on the following argument. Let γ_0 be a continuous curve in M_0 connecting the expanding sides E_0^{-1} and E_0^{+1}. Its image $\varphi(\gamma_0)$ is then stretched such that

[1] Note that this naming convention is exactly opposite to that of GHYS, because in two dimensions they emphasized the direction to which the sides of M_i were *parallel*. In higher dimensions, the faces of an n-cube are not parallel to a unique direction, and it is the direction along which a face is separated from the centre of the n-cube that matters. We change the naming convention now to avoid confusion later.

there is a section of $\varphi(\gamma_0)$ lying wholly within M_1, and in particular $\varphi(\gamma_0)$ leaves M_1 through the expanding sides $E_1^{\pm 1}$ at both ends. Let γ_1 be a section of $\varphi(\gamma_0)$ lying wholly within M_1. Now look at $\varphi(\gamma_1)$ in M_2. Repeat this process along the orbit, producing γ_N lying wholly within the final parallelogram M_N. Then any point lying along γ_N, traced backwards, represents an exact orbit that stays within M_i, $i = N, N-1, \ldots, 1, 0$, and we are done (Grebogi, Hammel, Yorke, and Sauer 1990).

With this picture, there is a nice geometric interpretation of the requirement that the angle between the stable and unstable directions be bounded away from 0: if the angle gets too small, then the parallelogram essentially loses a dimension, and $\varphi(M_i)$ can not make a "plus sign" with M_{i+1}. Practically speaking, this occurs when the angle becomes comparable with the noise amplitude of the refined orbit. Hence, the more accurate the orbit, the longer it can be shadowed (Grebogi, Hammel, Yorke, and Sauer 1990; Quinlan and Tremaine 1992).

2.2.3 Refinement

Definition. Refinement (Hammel, Yorke, and Grebogi 1987, 1988; Grebogi, Hammel, Yorke, and Sauer 1990; Quinlan and Tremaine 1991, 1992; Hayes 1995) is a numerical procedure similar to Newton's Method (and also analogous to iterative improvement methods for solving linear systems (Golub and Van Loan 1991)) that takes a noisy orbit as input and attempts to produce a nearby orbit with less noise, *i.e.,* one with smaller one-step errors. A refinement iteration is *successful* if before the iteration the trajectory has noise tightly bounded by δ^0, after the iteration it has noise tightly bounded by δ^1, and

$$\delta^1 < \mu\delta^0 \text{ for some practical } \mu \in [0, 1). \tag{2.8}$$

Otherwise the refinement iteration is *unsuccessful.* Here, a "practical" μ is one that will allow a noisy trajectory to be refined to noise levels near the machine precision in a small number of refinement iterations.

The refinement procedure of GHYS is analagous to Newton's method for finding a zero of a function. GHYS presented their method for the two-dimensional case. (The basic idea was described on page 11 of this thesis immediately following Theorem 2.1.) Assume we have a noisy n-dimensional orbit $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^N$, $\mathbf{y}_i \in \mathbf{R}^n$, and it has a shadow $\{\mathbf{x}_i\}_{i=0}^N$, $\mathbf{x}_i \in \mathbf{R}^n$. Then $\mathbf{x}_{i+1} = \varphi(\mathbf{x}_i)$ and $\mathbf{y}_{i+1} = \tilde{\varphi}(\mathbf{y}_i) = \varphi(\mathbf{y}_i) + \mathbf{e}_{i+1}$, where $\tilde{\varphi}$ is an approximation to φ with noise bounded by δ. Now suppose we approximate the one-step errors $\mathbf{e}_{i+1} = \mathbf{y}_{i+1} - \varphi(\mathbf{y}_i)$ using a method with noise significantly less than δ. Let $\hat{\mathbf{c}}_i \equiv \mathbf{x}_i - \mathbf{y}_i$ represent a correction term that perturbs $\mathbf{y}_i$ towards $\mathbf{x}_i$. Then

$$\hat{\mathbf{c}}_{i+1} = \mathbf{x}_{i+1} - \mathbf{y}_{i+1} = \varphi(\mathbf{x}_i) - \varphi(\mathbf{y}_i) - \mathbf{e}_{i+1} = D\varphi(\mathbf{y}_i)\hat{\mathbf{c}}_i - \mathbf{e}_{i+1} + O(\|\hat{\mathbf{c}}_i\|^2). \tag{2.9}$$

In the spirit of Newton's method, we ignore the $O(\|\hat{\mathbf{c}}_i\|^2)$ term, and so one refinement iteration defines the corrections along the entire orbit:

$$\mathbf{c}_{i+1} := D\varphi(\mathbf{y}_i)\mathbf{c}_i - \mathbf{e}_{i+1}. \tag{2.10}$$

For a discrete map, $D\varphi(\mathbf{y}_i)$ is just the Jacobian of the map at step i. For a system of ODEs,

$D\varphi(\mathbf{y}_i)$ is the Jacobian of the solution of the ODE from step i to step $i+1$.[2] For simplicity of explanation, we assume an $n = 2$ dimensional problem for the remainder of this subsection. For a generalization to arbitrary n, see Quinlan and Tremaine (1992) or Hayes (1995).

If the problem did not display pseudo-hyperbolicity, then the correction terms $\mathbf{c}_i$ could be computed directly from (2.10). But since $D\varphi$ displays an approximate exponential dichotomy, it tends to amplify any numerical errors in $\mathbf{c}_i$ not lying in the stable direction. Thus computing the $\mathbf{c}_i$'s by iterating (2.10) forward will amplify errors and typically produce nothing but noise; iterating backwards suffers the same problem. Therefore, GHYS split the error and correction terms into components in the stable ($\mathbf{s}_i$) and unstable ($\mathbf{u}_i$) directions at each timestep:

$$\mathbf{e}_i = e_{u_i}\mathbf{u}_i + e_{s_i}\mathbf{s}_i, \qquad \mathbf{c}_i = c_{u_i}\mathbf{u}_i + c_{s_i}\mathbf{s}_i. \tag{2.12}$$

Since it is not known *a priori* which direction is unstable at each timestep, the unstable vector $\mathbf{u}_0$ at time t_0 is initialized to an arbitrary unit vector. The linearized map is then iterated forward with

$$\bar{\mathbf{u}}_{i+1} = D\varphi(\mathbf{y}_i)\mathbf{u}_i, \qquad \mathbf{u}_{i+1} = \bar{\mathbf{u}}_{i+1}/\|\bar{\mathbf{u}}_{i+1}\|. \tag{2.13}$$

Since $D\varphi(\mathbf{y}_i)$ magnifies any component that lies in the unstable direction, and assuming we are not so unlucky to choose a $\mathbf{u}_0$ that lies too close to the stable direction, then after a few iterations $\mathbf{u}_i$ will point roughly in the unstable direction at t_i. Similarly, the stable unit direction vectors $\mathbf{s}_i$ are computed by initializing $\mathbf{s}_N$ to an arbitrary unit vector and iterating backward,

$$\bar{\mathbf{s}}_i = D\varphi(\mathbf{y}_i)^{-1}\mathbf{s}_{i+1}, \qquad \mathbf{s}_i = \bar{\mathbf{s}}_i/\|\bar{\mathbf{s}}_i\|. \tag{2.14}$$

Substituting (2.12) into (2.10) yields

$$c_{u_{i+1}}\mathbf{u}_{i+1} + c_{s_{i+1}}\mathbf{s}_{i+1} = D\varphi(\mathbf{y}_i)(c_{u_i}\mathbf{u}_i + c_{s_i}\mathbf{s}_i) - (e_{u_{i+1}}\mathbf{u}_{i+1} + e_{s_{i+1}}\mathbf{s}_{i+1}). \tag{2.15}$$

While $D\varphi(\mathbf{y}_i)$ magnifies errors in the unstable direction, it damps them in the stable direction. Likewise, $D\varphi(\mathbf{y}_i)^{-1}$ damps errors in the unstable direction and magnifies errors in the stable direction. Thus the c_u terms should be computed backward, and the c_s terms forward. Taking components of (2.15) in the unstable direction at step $i+1$, we iterate backward on

$$c_{u_i} = (c_{u_{i+1}} + e_{u_{i+1}})/\|\bar{\mathbf{u}}_{i+1}\|, \tag{2.16}$$

and taking components in the stable direction, we iterate forward on

$$c_{s_{i+1}} = \|D\varphi(\mathbf{y}_i)\mathbf{s}_i\| c_{s_i} - e_{s_{i+1}}. \tag{2.17}$$

[2] In other words, let

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}(t)) \tag{2.11}$$

be the first-order ODE. Note that $\mathbf{y}_{i+1} = \varphi(\mathbf{y}_i)$ is the solution of (2.11) using $\mathbf{y}_i$ as the initial condition and integrating $\mathbf{f}$ to time t_{i+1}. The Jacobian $D\mathbf{f}(\mathbf{y}_i)$ measures how $\mathbf{y}'$ changes if $\mathbf{y}$ is changed by a small amount. The *resolvent* $R(t_{i+1}, t_i)$ is the integral of $D\mathbf{f}(\mathbf{y})$ along the path $\mathbf{y}(t)$, and describes how a small perturbation $\delta\mathbf{y}$ of $\mathbf{y}_i$ at time t_i gets mapped to a perturbation of $\mathbf{y}_{i+1}$ at time t_{i+1}. $R(t_{i+1}, t_i)$ is the solution of the *variational equation*

$$\frac{\partial R}{\partial t} = D\mathbf{f}(\mathbf{y}(t))R(t, t_i), \quad R(t_i, t_i) = I,$$

where I is the identity matrix. The reason the arguments to R seem to be reversed is for notational convenience: they satisfy the identity $R(t_2, t_0) = R(t_2, t_1)R(t_1, t_0)$, and so a perturbation $\delta\mathbf{y}$ at time t_0 gets mapped to a perturbation at time t_2 by the matrix-matrix and matrix-vector multiplication $R_2\delta\mathbf{y} = R_1R_0\delta\mathbf{y}$ (Hairer, Nørsett, and Wanner 1993). Finally, the linear map in the GHYS refinement procedure, if φ is the time-h solution operator for (2.11), is $D\varphi(\mathbf{y}_i) = R(t_{i+1}, t_i)$.

The initial choices for c_{s_0} and c_{u_N} are arbitrary as long as they are small — smaller than the maximum shadowing distance — because (2.17) damps initial conditions and (2.16) damps final conditions. GHYS and QT choose them both as 0. This choice is probably as good as any, but it can be seen here that, if one shadow exists, there are infinitely many of them.[3] Another way of looking at these initial choices for c_{s_0} and c_{u_N} is that they "pinch" the growing components at the end point, and the backward-growing components at the initial point, to be small. That is, *boundary conditions* are being forced on the problem so that the exponential divergence is forcibly masked, if possible, making the solution of (2.10) numerically stable.

The refinement algorithm of GHYS as originally presented (Hammel, Yorke, and Grebogi 1987; Hammel, Yorke, and Grebogi 1988; Grebogi, Hammel, Yorke, and Sauer 1990) was not rigorous; if it worked at all, it only produced a new pseudo-trajectory with less noise than the original. Refinement was made rigorous by Sauer and Yorke (1991) with the following theorem:

Theorem 2.4 (Sauer and Yorke 1991). *Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^{N}$ be an $n \geq 2$ dimensional δ-pseudo-orbit of the map φ. Assume further that the local stable and unstable subspaces, $\mathbf{S}_i$ and $\mathbf{U}_i$, respectively, at each step are known to a tolerance of δ. Let θ_i be the angle between the stable and unstable subspaces at step i.[4] Let $\|D\varphi(\mathbf{z})\| \leq r_i\|\mathbf{z}\|$ for $\mathbf{z} \in \mathbf{S}_i$, and let $\|D\varphi(\mathbf{z})^{-1}\| \leq t_i\|\mathbf{z}\|$ for $\mathbf{z} \in \mathbf{U}_{i+1}$. Let $C_0 = D_N = 0$, and recursively define $C_{i+1} = \csc\theta_{i+1} + r_iC_i$ for $i = 0, \ldots, N-1$ and $D_{i-1} = \csc\theta_{i-1} + t_{i-1}D_i$ for $i = 1, \ldots, N$. Let B be a bound on $D\varphi, D\varphi^{-1}, D^2\varphi$, and $D^2\varphi^{-1}$. If $\delta < \frac{1}{20n^2}$ and*

$$\max\{C_i, D_i\} \leq \left(n^{5/2}B^2\sqrt{\delta}\right)^{-1}$$

for all $i = 0, \ldots, N$, then $\mathbf{Y}$ has an ε-shadow of φ such that $\varepsilon = \sqrt{\delta}$.

The proof of the theorem (see Sauer and Yorke (1991), Theorem 3.3) is constructive, in the sense that it uses the procedure for refining noisy orbits originally given in Hammel, Yorke, and Grebogi (1988). The essential point of the proof is to show that under the conditions of the theorem, the iterated application of the refinement procedure beginning with the pseudo-orbit results in a sequence of refined pseudo-orbits with decreasing noise level whose limit is an exact orbit. Furthermore, the exact orbit is not too far from the original pseudo-orbit.

Sauer and Yorke (1991) considered this theorem as a justification for the non-rigorous refinement procedure. Conversely, QT argued that if the refinement algorithm fails then there is good reason to believe that no shadow exists, for two reasons. First, from the more rigorous study of simpler systems, glitches are known to exist and are not just a failure of any particular refinement algorithm. Second, QT's results are consistent with a conjecture by GHYS on the frequency of glitches. However, there is no guarantee that refinement converges towards an exact orbit. In fact, even if some refinements are successful, numerical refinement alone does not prove that an *exact* shadow exists; it only proves the existence of a numerical shadow, *i.e.,* a trajectory with less noise than the original. Hayes (1995) frequently saw cases in which the refinement algorithm failed to find a numerical shadow for noisy orbits of length N, but

[3] For any system, *even a chaotic one*, given any exact orbit of fixed length, a small enough perturbation in the initial condition in any direction produces a small change in the final condition, although for chaotic systems this perturbation must be exponentially small in the length of the orbit. (If the perturbation is restricted to the stable subspace, then obviously a similar solution will be obtained.) Thus given any exact orbit that δ-shadows a noisy orbit, there exist infinitely many exact orbits nearby that also shadow it. However, it may be that all the exact orbits are packed into a space unresolved by the machine precision.

[4] Sauer and Yorke (1991) do not specify if θ_i should be an upper or lower bound; presumably it is a lower bound, since we want the system to be as pseudo-hyperbolic as possible.

succeeded in finding a numerical shadow for the superset of length $2N$. Hence, the algorithm failed to find a numerical shadow of length N, even though one clearly exists. On the other hand, this author often observed the code from his Master's thesis "converge" to an arbitrary precision several orders of magnitude less precise than the machine precision but then fail to converge any further, even though the algorithm is usually capable of converging very close to the machine precision. This would seem to imply that in these cases, refinement can not reduce the errors any further, implying that no shadow exists. However, the algorithm fails to "blow up". This leads us to ask the question of whether convergence to machine precision is enough: is it possible that refinement, if continued in higher precision, would stop before converging to an exact orbit (Hayes 1995)? Despite these objections, this author believes that refinement to machine precision implies with reasonable probability that a shadow exists whose length is comparable to that of the numerical shadow, although this evidence should not be taken as conclusive.

This author's Master's Thesis (Hayes 1995) provided empirical evidence that supports a conjecture that shadow lengths in "unsoftened" n-body systems scale as $O(1/n)$. However, more careful analysis (Hayes and Jackson 1997) has revealed that the scaling is much better described by a $O(1/n^2)$ law. No simple explanation is available for this scaling, although one could conjecture that there is some relation to the fact that there are $O(n^2)$ possible interactions between n particles in an n-body system, and that these interactions somehow conspire to cause glitches. One could argue that the algorithm itself is at fault: however, Hayes and Jackson (1996) showed that an artificially created nonlinear pseudo-hyperbolic system with 180 dimensions was easily shadowed by the refinement algorithm. Furthermore, Hayes and Jackson (1997) demonstrated that shadow lengths of "softened" n-body systems, in which the interaction between particles is decreased, scale more optimistically even than $O(1/n)$, in fact almost $O(1)$. On the other hand, the theorem of Sauer and Yorke (1991) also contains a factor of $O(1/n^2)$ in the length of shadows, even though their theorem deals with general n-dimesional systems in which there is no clear association between dimensions and physical objects like particles. Apparently, all one can conclude from this discussion is that high-dimensional shadowing is an area ripe for further study.

In terms of computational cost, note that a resolvent has $O(n^2)$ elements in it, and is generally expensive to compute. Hayes (1995) and Hayes and Jackson (1996) list several optimizations to the procedure that increase the speed of GHYS/QT refinement by about two orders of magnitude. If one is interested in studying high-dimensional systems, a chaotic map would be a better test problem than an ODE system, because no variational equation integration is needed. We note that the GHYS/QT refinement algorithm is trivially parallelizable, since the computation of each $D\varphi(\mathbf{y}_i)$ is completely independent of all the others. For the same reason, it also has excellent locality of reference in a serial implementation, so virtual memory paging is minimized. Once the $D\varphi(\mathbf{y}_i)$'s are computed, it may also be worth parallelizing the recurrence (2.10) (Jackson and Pancer 1992). Finally, we note that $D^2\varphi$ has $O(n^3)$ elements so, unless significant sparsity is present, actually applying Theorem 2.4 is impractical for any but small n.

Refinement is closely related to the problem of *noise reduction.* Farmer and Sidorowich (1991) make the distinction between *observational noise* and *dynamical noise.* The former occurs when one is observing a physical process, which inherently involves observational noise. One can attempt to dampen the noise by applying a technique similar to refinement in which one searches the nearby phase space for an exact solution. The basic idea behind their noise-reduction scheme is illustrated in figure 2.2. One can further attempt to find the *closest* exact solution to the observations using a least-squares constraint, presumably giving a good approx-

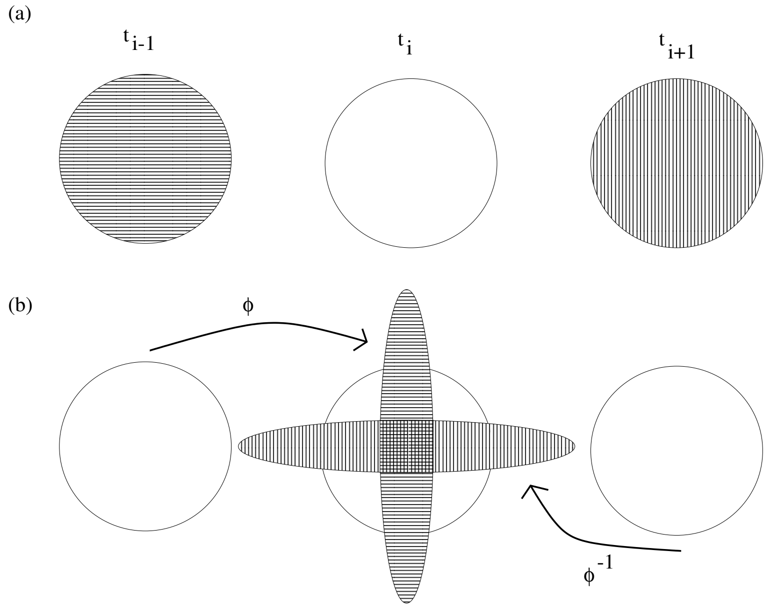

Figure 2.2: Schematic representation of the noise reduction technique of Farmer and Sidorowich (1991). (a) The circles represent noisy measurements of a deterministic trajectory at three different times. (b) As successive measurements are transported to the same point in time (the middle circle at the bottom), the associated noise probability distributions distort according to the local derivatives of the dynamical system. The true state should lie somewhere in the intersection of the three regions (the square region lying in the intersection of the two ellipses). Averaging the transported measurements at time t_i makes it possible to produce a better estimate of the true state at time t_i.

imation to the actual trajectory followed by the process.[5] This is in contrast to a numerically generated pseudo-trajectory, in which the noise is injected into the dynamics and affects the future evolution of the system. Although the problems are clearly similar, and refinement can be used as a noise reduction technique, Quinlan and Tremaine (1992) found that some "tricks" often used when applying noise reduction failed to work when adapted to the refinement algorithm and applied to the shadowing problem.

A tangentially related work (Fryska and Zohdy 1992) proved that numerical solutions of piecewise linear ODEs can sometimes introduce statistical biases, causing numerical solutions to have substantially different statistical properties than the closed-form solution. In an ironic twist to the whole concept of shadowing, they found that the correct statistical properties could be recovered by *injecting* uniformly distributed random noise into the numerical solution. The apparent explanation is that the injected random noise somehow masks the statistical bias introduced by the numerical method.

2.2.4 Results by bounding non-hyperbolicity

We make the distinction between *rigorous results* and *nonrigorous results*. We call a result rigorous if the method used to produce it is entirely rigorous from start to finish: for example, if a computational component rigorously bounds numerical errors, and then a theorem is used to show that the computed properties imply the existence of a shadow. Some results are

[5]Note that there is no need to make this method rigorous, because it is *known* that an exact trajectory exists near the observed one, namely the exact trajectory that is being obscured by noise.

partially rigorous, in that floating-point computations without rigorous error bounds are used in combination with a theorem; such results could easily be made rigorous with the application of interval arithmetic. Finally, non-rigorous results use convincing numerical experiments to infer properties of noisy trajectories and their purported shadows.

The original results presented in this thesis are rigorous.

Rigorous results

The procedures of containment and refinement do not make explicit use of the hyperbolicity of the system, although they work only if some measure of hyperbolicity is present (Chow and Palmer 1991). In contrast, Chow and Palmer (1991, 1992) make explicit use of the hyperbolicity of the system, and use the ideas of the traditional Shadowing Lemma (Anosov 1967; Bowen 1975; Palmer 1988) to estimate how far a shadow is from a pseudo-orbit. Chow and Palmer (1991) discussed the one-dimensional case, and Hadeler (1996) made explicit the relationship between the one-dimensional case and Kantorovich's Theorem, which lays out conditions under which Newton's method will converge. We omit detailed discussion of the one-dimensional case because later work by the same authors (Chow and Palmer 1992) subsumes it, except to note one very interesting fact: Chow and Palmer (1991) proved that in the one-dimensional case, the shadowing distance not only has an upper bound, but a *lower* bound as well. That is, they proved that the shadow must maintain a minimum distance from the noisy orbit; it cannot approach the noisy orbit arbitrarily closely. It is not clear if this result is extendible to higher dimensions, nor is it clear exactly what the significance of this result is; however, it is certainly interesting. The high-dimensional theorem and its proof by Chow and Palmer (1992) is so concise and elegant that we now include it in its entirety.

Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a C^2 function and let $\{\mathbf{y}_i\}_{i=0}^N$ be a δ-pseudo-orbit of φ. Given any sequence $(\mathbf{h}_i)_{i=0}^{N-1}$ in $\mathbf{R}^{nN}$, the difference equation

$$\mathbf{z}_{i+1} = D\varphi(\mathbf{y}_i)\mathbf{z}_i + \mathbf{h}_i$$

has many solutions. So the linear operator $L : \mathbf{R}^{n(N+1)} \to \mathbf{R}^{nN}$ defined for $\mathbf{Z} = \{\mathbf{z}_i\}_{i=0}^N$ by

$$(L\mathbf{Z})_i = \mathbf{z}_{i+1} - D\varphi(\mathbf{y}_i)\mathbf{z}_i$$

is onto and so has right inverses. For the following theorem, we choose any such right inverse.

Theorem 2.5 (Chow and Palmer 1992). *Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a C^2 function and let $M = \sup\{\|D^2\varphi(\mathbf{x})\| : \mathbf{x} \in \mathbf{R}^n\}$. Let $\{\mathbf{y}_i\}_{i=0}^N$ be a δ-pseudo-orbit of φ with $2M\|L^{-1}\|^2\delta \le 1$, where L^{-1} is a right inverse of L. Then there is an exact orbit $\{\mathbf{x}_i\}_{i=0}^N$ of φ such that*

$$\|\mathbf{x}_i - \mathbf{y}_i\| \le \frac{2\|L^{-1}\|\delta}{1 + \sqrt{1 - 2M\|L^{-1}\|^2\delta}}, \qquad i = 0, \ldots, N.$$

Proof. **(Chow and Palmer 1992)** The sequence $\{\mathbf{x}_i\}_{i=0}^N$ satisfies $\mathbf{x}_{i+1} = \varphi(\mathbf{x}_i),\ i = 0, \ldots, N-1$. If we set $\mathbf{x}_i = \mathbf{y}_i + \mathbf{z}_i$, we find that $\mathbf{z}_i$ satisfies

$$\mathbf{z}_{i+1} = D\varphi(\mathbf{y}_i)\mathbf{z}_i + \mathbf{g}_i(\mathbf{z}_i), \tag{2.18}$$

where

$$\mathbf{g}_i(\mathbf{z}) = \varphi(\mathbf{y}_i) - \mathbf{y}_{i+1} + \varphi(\mathbf{y}_i + \mathbf{z}) - \varphi(\mathbf{y}_i) - D\varphi(\mathbf{y}_i)\mathbf{z}. \tag{2.19}$$

Remark: Equation (2.18) is the analog of the correction term in the refinement algorithm, equation (2.10) on page 15. The first two terms in equation (2.19) represent the one-step error at step i, while the last three terms describe the amount of nonlinearity in φ, i.e., the $O(\|\mathbf{c}\|^2)$ terms that were ignored in equation (2.9) of the refinement algorithm, which are bounded by $\frac{1}{2}M\|\mathbf{z}\|^2$.

So our task is to solve equation (2.18) for a sequence $\mathbf{z}_i$ such that for $i = 0, \ldots, N$,

$$\|\mathbf{z}_i\| \le \varepsilon := \frac{2\|L^{-1}\|\delta}{1 + \sqrt{1 - 2M\|L^{-1}\|^2\delta}}.$$

Let $\mathbf{Z} = (\mathbf{z}_i)_{i=0}^N \in \mathbf{R}^{n(N+1)}$ and let $\|\mathbf{Z}\| \equiv \max_{i=0}^N \|\mathbf{z}_i\|$. Denote by Υ the set of sequences with max norm ε, $\Upsilon = \{\mathbf{Z} \mid \|\mathbf{Z}\| \le \varepsilon\}$. Υ is a compact convex subset of $\mathbf{R}^{n(N+1)}$. We define a mapping T on Υ. Note that we can write equation (2.18) as $L\mathbf{Z} = \mathbf{g}(\mathbf{Z})$ where

$$(L\mathbf{Z})_i = \mathbf{z}_{i+1} - D\varphi(\mathbf{y}_i)\mathbf{z}_i, \qquad (\mathbf{g}(\mathbf{Z}))_i = \mathbf{g}_i(\mathbf{z}_i).$$

Then we define $T\mathbf{Z} = L^{-1}\mathbf{g}(\mathbf{Z})$, where L^{-1} is the given right inverse of L.

Since the $\mathbf{g}_i$'s are continuous, T is a continuous mapping of Υ into $\mathbf{R}^{n(N+1)}$. We show that T maps Υ into itself. First observe that

$$\begin{aligned}\|\mathbf{g}_i(\mathbf{z})\| &\le \|\mathbf{g}_i(0)\| + \|\varphi(\mathbf{y}_i + \mathbf{z}) - \varphi(\mathbf{y}_i) - D\varphi(\mathbf{y}_i)\mathbf{z}\| \\ &\le \delta + \frac{1}{2}M\|\mathbf{z}\|^2.\end{aligned}$$

Then, if $\mathbf{Z} \in \Upsilon$,

$$\|(T\mathbf{Z})_i\| \le \|L^{-1}\|(\delta + \frac{1}{2}M\varepsilon^2) = \varepsilon,$$

where the middle term is shown equal to ε by moving ε into the middle term and solving the resulting quadratic equation in ε. So T maps Υ into itself. By Brouwer's fixed point theorem, it has a fixed point $\mathbf{Z} = \{\mathbf{z}_i\}_{i=0}^N$. Then $T\mathbf{Z} = \mathbf{Z}$ and so, since LL^{-1} is the identity, $L\mathbf{Z} = \mathbf{g}(\mathbf{Z})$. That is, $\mathbf{Z}$ is a solution of equation (2.18) satisfying $\|\mathbf{z}_i\| \le \varepsilon$ for $i = 0, \ldots, N$. Then $\mathbf{x}_i = \mathbf{y}_i + \mathbf{z}_i$ is the shadow. □

Remark 1: It is not actually necessary to assume that $D^2\varphi(\mathbf{x})$ is bounded over $\mathbf{R}^n$ because usually $\mathbf{y}_i$ would be restricted to a bounded set and M could be replaced by a bound for $\|D^2\varphi(\mathbf{x})\|$ over that set (Chow and Palmer 1992).

Remark 2: $\|L^{-1}\|$ is the "magnification factor". If δ is the local error made in computing the orbit, then $\|L^{-1}\|\delta$ is approximately the distance to the shadow.

The next step is to choose L^{-1} in such a way that $\|L^{-1}\|$ is minimized. Not surprisingly, the best L^{-1} to choose is one whose components are as aligned as possible with the stable and unstable subspaces at each step, computed in a fashion similar to the refinement algorithm. Finally, computing an upper bound for $\|L^{-1}\|$ involves noting that even though the orbit $\{\mathbf{x}_i\}$ is not hyperbolic under φ, it may be hyperbolic under φ^p for some integer $p > 1$. If such a p is found, it allows explicit bounds to be computed on the hyperbolicity constants for the orbit $\{\mathbf{x}_i\}$ under φ^p using the ideas of the traditional Shadowing Lemma (Anosov 1967; Bowen 1975; Palmer 1988), leading to an upper bound on $\|L^{-1}\|$. Chow and Palmer demonstrate their method on a δ-pseudo-orbit of the Hénon map with $\delta = 2^{-54} \approx 10^{-15.5}$. For a particular orbit of $N = 333,000$ iterates of the map, they find that $p = 40$ guarantees hyperbolicity of the orbit under φ^p and that $\|L^{-1}\| \le 113277 \approx 10^5$. This means that the shadowing distance is about 10^5 times the size of the one-step errors, giving a shadow distance of about $10^{-10.5}$.

Non-rigorous results

This "magnification factor", the ratio between the shadow distance and the local error, is termed the "brittleness" of an orbit by Dawson, Grebogi, Sauer, and Yorke (1994).[6] If the brittleness is of order the inverse of the machine epsilon or larger, then all accuracy is lost as the shadowing error is comparable to the size of the variables themselves. They show that if the number of positive and negative Lyapunov exponents changes, or if a Lyapunov exponent fluctuates about zero, then the brittleness can blow up. The effect of a fluctuating exponent is depicted in Figure 2.3. However, Dawson, Grebogi, Sauer, and Yorke (1994) make the strong claim that they

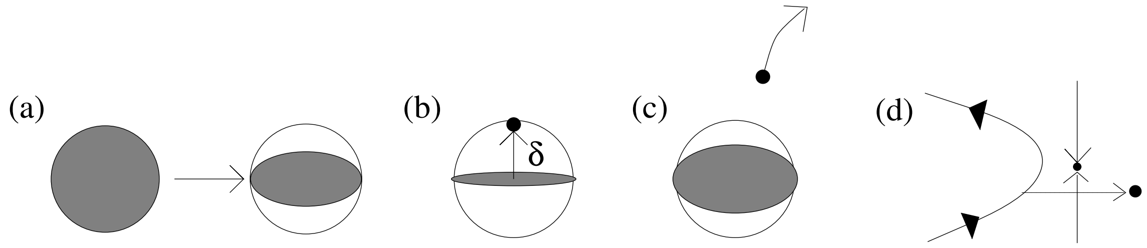

Figure 2.3: Fluctuating Lyapunov exponent in the "vertical" direction, reproduced from Dawson *et al.* (1994). δ is the local error, and the vertical direction is initially contracting, but then becomes expanding. (a) an ensemble of trajectories that starts off in an ε-ball is first compressed into a sheet. (b) If the local error steps outside this sheet, and then the direction becomes an expanding direction, then (c) the numerical trajectory diverges away from all exact trajectories that started in the original ε-ball, (d) possibly entering regions of phase space with qualitatively different behaviour than the exact trajectories.

believe this kind of fluctuating Lyapunov exponent is "common" in high dimensional systems, with the only justification apparently being that there are so many dimensions that there must be a fluctuating exponent *somewhere*. Although this argument is not formally compelling, it may have some merit. On the other hand, Hayes and Jackson (1996) demonstrated numerical shadowing of a 180-dimensional non-hyperbolic system, although that system was artificially constructed to have pseudo-hyperbolicity.

Systems which possess such fluctuating Lyapunov exponents are termed *hyperchaotic* by Sauer, Grebogi, and Yorke (1997). Let $\mathbf{z}_i$ be the displacement from the pseudo-orbit to the shadow at step i. Sauer, Grebogi, and Yorke (1997) observe that the evolution of $\mathbf{z}_i$ with i is similar to a biased random walk. A glitch occurs when the random walk pushes the numerical orbit further away from the shadow than the hyperbolicity can correct for. They model the random walk formally as a *Kolmogorov diffusion process* and demonstrate that the distribution of shadowing distances using this model closely resembles actual shadowing distance distributions. Furthermore, they compute how often glitches occur, based on the behaviour of the fluctuating Lyapunov exponent which is closest to zero. They show that the expected time $\langle\tau\rangle$ for the shadowing distance to become the same size as the variables is proportional to

$$\langle\tau\rangle \sim \delta^{-2\lambda_0/\sigma_0^2},$$

where λ_0 and σ_0 are the mean and standard deviation, respectively, of the fluctuating Lyapunov exponent closest to zero. Finally, they demonstrate that when the fluctuations are sufficiently badly behaved, the length of the shadow is virtually independent of the local error — in other

[6]The terms "modulus of continuity" and "condition number" are commonly used in the literature for this ratio.

words, in a sufficiently badly behaved system, the shadow length will never get very long for any practical local error.

Methods have also been developed to shadow one-dimensional lattice maps, typically discretizations of partial differential equations (Chow and Van Vleck 1993, 1994b), and for problems that are piecewise hyperbolic in which the number of stable dimensions is monotonically increasing with time (Chow and Van Vleck 1994a).

2.2.5 Shadowing lemmas designed explicitly for ODE systems

This thesis concerns the problem of shadowing of ODE systems, including the rescaling of time (defined below), which is the topic of this subsection.

Introduction

There is a fundamental difference between a discrete map and a discrete solution to an ODE. Local errors of the former are restricted to being "space like" — there is no notion of the passage of time between iterations of the map. The latter, however, can have errors in space *as well as time*. The numerical error in the length of each timestep can accumulate, leading the numerical solution to have a slightly different time scale than the real system. In the integration of periodic or almost periodic systems like the solar system, this is also known as phase error, because the numerical solution may have a slightly different period than the exact solution. Thus, although the orbit of a planet may be reproduced correctly by the numerical trajectory, the time at which a real and simulated planet pass through a fixed plane perpendicular to the orbit may differ. This is the case even if the integrator is symplectic (Stuart and Gonzalez 1996; Gonzalez and Stuart 1996). Thus, when attempting to shadow a numerical solution of an ODE, it may be necessary to "rescale" time (Coomes, Koçak, and Palmer 1994b, 1995a, 1995b; Van Vleck 1995). To take this into account, we redefine a shadow of an ODE system as follows:

Definition of ODE shadowing: A pseudo-trajectory $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^{N}$ with timesteps $\{h_i\}_{i=0}^{N-1}$ is ε-shadowed by an exact trajectory $\mathbf{X} = \{\mathbf{x}_i\}_{i=0}^{N}$ with timesteps $\{\tau_i\}_{i=0}^{N-1}$ if $\mathbf{x}_{i+1} = \phi_{\tau_i}(\mathbf{x}_i)$, where $\|\mathbf{y}_i - \mathbf{x}_i\| \le \varepsilon$, and $|h_i - \tau_i| \le \varepsilon$.

Remark: In the above definition, we assume that $\varepsilon \ll h_i$, that is, the shadowing distance is significantly smaller than the timesteps. In practice, this appears sufficient for the systems we have studied. If this were not the case, the above definition could be modified to include some notion of global time error per-unit-step.

In other words, the numerical trajectory is shadowed if it closely follows the *path* of an exact solution, but at time t it is allowed to be a little ahead of or behind the exact solution. This linear growth of time errors is due to a lack of hyperbolicity in the direction of the flow in phase space (Van Vleck 1995). For large $|t - t_0|$ this can be a significant difference, so a shadowing method which does not take the rescaling of time into account is likely to grossly underestimate the length of the shadow. Coomes, Koçak, and Palmer (1994b, 1995a, 1995b) dramatically demonstrate this when they show that a rescaling of time allows the Lorenz equations to be shadowed for almost 10^5 time units, while the *map method*, which does not rescale time, finds shadows lasting only 10 time units—an astounding increase in shadow length of a factor of 10^4!

Finally, note that the non-shadowable example given in the tutorial ($y'' = 0$, page 11) *is* shadowable if time is rescaled. This matches what our intuition would say: as long as we care only about qualitative properties of the solution, it should not matter if the numerical trajectory traverses the path at a slightly different velocity than the exact solution, as long as the trajectories, taken as a whole, remain near to each other.

Explicitly rescaling time in Newton's method

Errors in time manifest themselves as errors directed along the direction of $\mathbf{y}'$, and so one way to account for these errors is to explicitly perturb the noisy solution along the $\mathbf{y}'$ direction. These perturbations translate back into a rescaling of time. To this end, Van Vleck (1995) proves a theorem similar to that of Chow and Van Vleck (1993, 1994b) in which time is explicitly added to the variational equation of the one-step error function. To wit, if $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^{N}$ is a δ-pseudo trajectory with associated timesteps $\{h_i\}_{i=0}^{N-1}$, then let $\mathbf{z}_i = (\mathbf{y}_i, h_i)$ and $\mathbf{Z} = \{\mathbf{z}_i\}_{i=0}^{N}$ and compute the one-step error by $\mathbf{g}(\mathbf{Z})_i = \mathbf{y}_{i+1} - \varphi_{h_i}(\mathbf{y}_i)$. Then the first variational equation $D\mathbf{g}(Z) : \mathbf{R}^{n(N+1)} \times \mathbf{R}^N \to \mathbf{R}^{nN}$ including the effects of time is

$$\begin{aligned}(D\mathbf{g}(\mathbf{Z})\Delta\mathbf{Z})_i &= \Delta\mathbf{y}_{i+1} - \frac{\partial\varphi_{h_i}(\mathbf{y}_i)}{\partial\mathbf{y}_i}\Delta\mathbf{y}_i - \theta\frac{\partial\varphi_{h_i}(\mathbf{y}_i)}{\partial h_i}\Delta h_i \\ &\equiv \Delta\mathbf{y}_{i+1} - \frac{\partial\varphi_{h_i}(\mathbf{y}_i)}{\partial\mathbf{y}_i}\Delta\mathbf{y}_i - \theta f(\varphi_{h_i}(\mathbf{y}_i))\Delta h_i\end{aligned}$$

where θ is a user-input parameter controlling the amount of time rescaling which is allowed.[7] More formally, we are changing the norm with respect to which the variation is performed: $\theta = 0$ corresponds to the norm in which variations with respect to time are not considered at all, whereas $\theta = 1$ corresponds to the norm in which variations with respect to time are fully considered. Choosing $\theta \in [0, 1]$ allows the scale of variations in time to be *different* from the scale of variations in space, which is precisely what we need in order to perform a rescaling of time. Then we have the following theorem.

Theorem 2.6 (Van Vleck 1995). *Given constants $\delta, c > 0$ and $\eta \geq 0$ suppose L is an approximation to $D\mathbf{g}(\mathbf{Z})$ such that*

(i) a right inverse L^{-1} of L satisfies $\|L^{-1}\| \leq c$.

(ii) $\|L^{-1} - D\mathbf{g}(\mathbf{Z})^{-1}\| \leq \eta$ for some right inverse $D\mathbf{g}(\mathbf{Z})^{-1}$ of $D\mathbf{g}(\mathbf{Z})$.

Assume that $\|\mathbf{g}(\mathbf{Z})\| \leq \delta$ and let $\varepsilon := 2\delta(\eta + c)$. If $\|D\mathbf{g}(\mathbf{Z}) - D\mathbf{g}(\mathbf{W})\| \leq 1/(2(\eta + c))$ for $\|W - Z\| \leq \varepsilon$, then $\mathbf{g}$ has a solution $\mathbf{W}$ of $\mathbf{g}(\mathbf{W}) = 0$ such that $\|\mathbf{W} - \mathbf{Z}\| \leq \varepsilon$.

Proof. See Van Vleck (1995), Theorem 2.2, which quotes a theorem from Chow, Lin, and Palmer (1989). □

For problems that lack hyperbolicity in the direction of motion, Van Vleck (1995) demonstrates that non-zero values of θ are capable of finding shadows between 10 and 100 times longer than if $\theta = 0$, with shadow lengths for the Lorenz system lasting up to about 10^4 time units. However, good values for θ must be found by trial and error.

Implicitly rescaling time

Coomes, Koçak, and Palmer (1994b, 1995a) provide the most impressive results to date on shadowing numerical solutions to ODEs. They detail a rigorous method allowing for the rescaling of time that finds shadows for the Lorenz system longer and with a smaller global error than any other published work (except this thesis, which matches their results). Their method relies upon building a hyperplane $\mathcal{H}_i$ perpendicular to $\mathbf{f}(\mathbf{y}_i)$ and containing $\mathbf{y}_i$, and then finding a sequence of points $\mathbf{x}_i \in \mathcal{H}_i$ such that $\mathbf{x}_{i+1} = \varphi_{\tau_i}(\mathbf{x}_i)$ and $|\tau_i - h_i| < \varepsilon$. (See Figure 2.4.) In

[7] This is the only place in this thesis where $D\varphi$ includes a differentiation with respect to h_i. It is this term which allows a rescaling of time by allowing an adjustment along $\mathbf{y}'$.

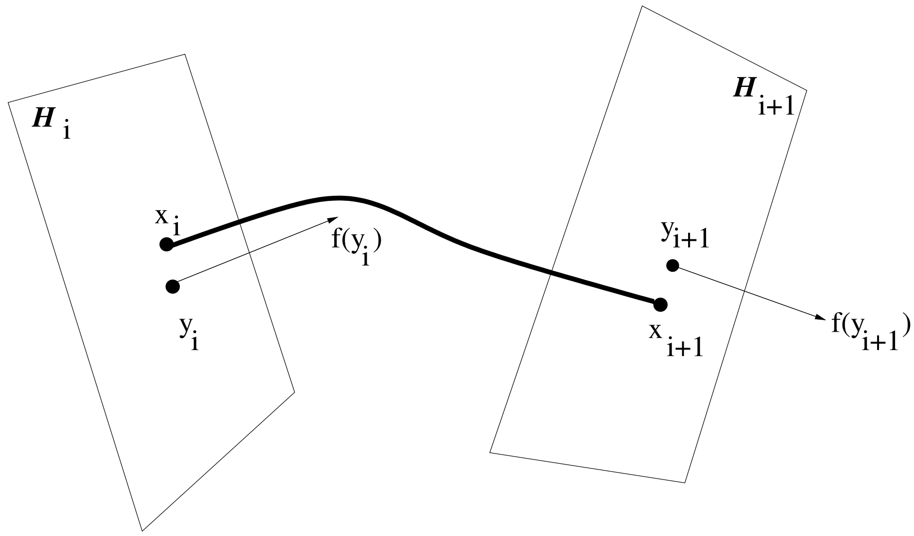

Figure 2.4: Pseudo-orbit $\mathbf{y}_i$ and the shadowing orbit $\mathbf{x}_i$ in hyperplane $\mathcal{H}_i$ (Coomes, Koçak and Palmer 1994).

this way, they avoid having to find τ_i explicitly, as opposed to Van Vleck (1995) who computed the τ_i explicitly as part of a Newton's method. The statement of their theorem requires some introductory notation.

Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^{N}$ be a δ-pseudo orbit with associated stepsizes $\{h_i\}_{i=0}^{N-1}$. Also suppose that we have a sequence $\{Y_i\}_{i=0}^{N-1}$ of $n \times n$ matrices such that

$$\|Y_i - D\varphi_{h_i}(\mathbf{y}_i)\| \leq \delta, \qquad i = 0, \ldots, N-1.$$

Now, let S_i be an $n \times (n-1)$ matrix chosen so that its columns form an almost-orthonormal basis for the subspace orthogonal to $\mathbf{f}(\mathbf{y}_i)$,

$$\|S_i^T \mathbf{f}(\mathbf{y}_i)\| \leq \delta_1, \qquad \|S_i^T S_i - I\| \leq \delta_1,$$

for some positive number δ_1. Now, we compute $(n-1) \times (n-1)$ matrices A_i satisfying

$$\|A_i - S_{i+1}^T Y_i S_i\| \leq \delta_1, \qquad i = 0, \ldots, N-1.$$

Geometrically, A_i is Y_i restricted to the subspace orthogonal to $\mathbf{f}(\mathbf{y}_i)$ and then projected to the subspace orthogonal to $\mathbf{f}(\mathbf{y}_{i+1})$. Next, define a linear operator $L : (\mathbf{R}^{(n-1)})^{(N+1)} \to (\mathbf{R}^{(n-1)})^N$ in the following way: If $\Xi = \{\xi_i\}_{i=0}^{N}$ is in $(\mathbf{R}^{(n-1)})^{(N+1)}$, then we take $L\Xi = \{(L\Xi)_i\}_{i=0}^{N-1}$ where

$$(L\Xi)_i = \xi_{i+1} - A_i \xi_i, \qquad i = 0, \ldots, N-1.$$

The operator L has right inverses and we choose one such right inverse L^{-1}. We now define several constants. Let U be a convex subset of $\mathbf{R}^n$ containing $\{\mathbf{y}_i\}_{i=0}^{N}$ in its interior. For such U, we define

$$M_0 = \sup_{\mathbf{x} \in U} \|\mathbf{f}(\mathbf{x})\|, \qquad M_1 = \sup_{\mathbf{x} \in U} \|D\mathbf{f}(\mathbf{x})\|, \qquad M_2 = \sup_{\mathbf{x} \in U} \|D^2\mathbf{f}(\mathbf{x})\|.$$

Then we define

$$\overline{h} = \sup_{0 \leq i \leq N-1} h_i, \qquad \underline{h} = \inf_{0 \leq i \leq N-1} h_i.$$

Next, we choose a positive number $\varepsilon_0 \leq \underline{h}$ such that for $i = 0, \ldots, N-1$ and all $\mathbf{x}$ satisfying $\|\mathbf{x} - \mathbf{y}_i\| \leq \varepsilon_0$, the solution $\varphi_t(\mathbf{x})$ is defined and remains in U for $0 \leq t \leq h_i + \varepsilon_0$. Finally, we define

$$\underline{M}_0 = \inf_{0 \leq i \leq N} \|\mathbf{f}(\mathbf{y}_i)\|, \quad \overline{M}_0 = \sup_{0 \leq i \leq N} \|\mathbf{f}(\mathbf{y}_i)\|, \quad \overline{M}_1 = \sup_{0 \leq i \leq N} \|D\mathbf{f}(\mathbf{y}_i)\|, \quad \Theta = \sup_{0 \leq i \leq N-1} \|Y_i\|.$$

Then, we have the following theorem.

Theorem 2.7 (Coomes, Koçak, and Palmer 1994b). *Let*

$$C = \max\left\{\underline{M}_0^{-1}(\Theta\, \|L^{-1}\|\, (1+\delta_1)+1),\ \ \|L^{-1}\|\sqrt{1+\delta_1}\right\},$$

$$\delta_{\mathcal{H}} = C((M_1 + \sqrt{1+\delta_1})\delta + (3\delta_1(\sqrt{1+\delta_1} + \underline{M}_0^{-1}))/(1 - \delta_1(1+\underline{M}_0^{-2})),$$

$$\overline{M} = (\overline{M}_1 + M_2\nu\delta)\left(\overline{M}_0 + M_1\nu\delta + 2e^{M_1(\overline{h}+\varepsilon_0)}\sqrt{1+\delta_1}\right) + M_2(\overline{h}+\varepsilon_0)(1+\delta_1)e^{2M_1(\overline{h}+\varepsilon_0)},$$

where

$$\nu = 2C(e^{M_1(\overline{h}+\varepsilon_0)}\sqrt{1+\delta_1} + M_0)(1-\delta_{\mathcal{H}})^{-1} + 1.$$

If these quantities together with δ, δ_1 and ε_0 satisfy the inequalities

(i) $\delta_1(1+\underline{M}_0^{-2}) < 1$

(ii) $\delta_{\mathcal{H}} < 1$

(iii) $2C(1-\delta_{\mathcal{H}})^{-1}\delta\sqrt{1+\delta_1} < \varepsilon_0$

(iv) $2\overline{M}C^2(1-\delta_{\mathcal{H}})^{-2}\delta \leq 1$,

then $\mathbf{Y}$ *is ε-shadowed with shadowing distance*

$$\varepsilon \leq 2C(1-\delta_{\mathcal{H}})^{-1}\delta\sqrt{1+\delta_1}.$$

Proof. See Coomes, Koçak, and Palmer (1994b, 1995a). □

Coomes, Koçak, and Palmer use a Taylor series integration method with interval arithmetic (see, for example, Nedialkov 1999) to produce a rigorously bounded local error of their numerical trajectory, and also require the computation of an integer p identical to the p in Chow and Palmer (1992) (cf. p. 21).

As Coomes, Koçak, and Palmer state, "Admittedly, the statement of the theorem seems rather imposing." The proof, which spans some 9 pages, quotes several other nontrivial theorems and lemmas from other papers, and omits many details, also appears imposing to this author. It is also of practical importance to note that M_0, M_1, and M_2 are bounds on $\mathbf{f}$ and its derivatives over the *entire* convex set U containing the pseudo-trajectory. This makes the theorem inapplicable to problems which may contain poles in U, such as the unsoftened gravitational n-body problem. By contrast, containment only requires bounds over a much smaller volume, essentially a (possibly self-intersecting) "tube" containing the pseudo-trajectory and its shadow. If the requirement that U be convex were withdrawn, perhaps this restriction could be lifted. Furthermore, the bound on the second derivative of $\mathbf{f}$ over U could be very expensive to compute if a closed form bound is not available. However, requiring bounds on the first and second derivatives of $\mathbf{f}$ is a significant improvement over requiring bounds on the first and second derivatives of φ, as required by Theorem 2.4.

For a local error of about 10^{-13}, Coomes, Koçak, and Palmer were able to find shadows for the Lorenz system lasting 10^5 time units, with a shadowing distance of about 10^{-9}. Adding to this the fact that their results are entirely rigorous, it is this author's opinion that Coomes, Koçak, and Palmer have the best results in the field thus far. As we will see later, the results of this thesis are comparable.

Periodic shadowing

The problem of errors in time is exacerbated when attempting to shadow periodic solutions of ODEs, because any non-zero error in time is repeated *ad infinitum*. Thus, a rescaling of time is absolutely necessary to shadow periodic solutions of ODEs.

The idea for shadowing periodic solutions is simple. Given a pseudo-trajectory $\{\mathbf{y}_i\}_{i=0}^N$ with timesteps $\{h_i\}_{i=0}^{N-1}$, we require not only that the local error $\|\mathbf{y}_{i+1} - \varphi_{h_i}(\mathbf{y}_i)\|$ is small, but also that $\|\mathbf{y}_0 - \varphi_{h_N}(\mathbf{y}_N)\|$ is small. This gives a periodic pseudo-orbit. Then, only minor modifications are required to non-periodic shadowing theorems to produce a periodic shadowing theorem (Van Vleck 1995; Coomes, Koçak, and Palmer 1994a). It is also possible to use refinement-like algorithms to produce accurate pseudo-trajectories from remarkably *in*accurate ones, allowing one to prove the existence of very long periodic trajectories (Coomes, Koçak, and Palmer 1997).

2.2.6 Shadowing conservative integrations

As described in Chapter 1, much attention has recently been devoted to integrators that preserve various quantities such as symplectic structure (Channell and Scovel 1990; Sanz-Serna 1992) and energy (Stuart and Gonzalez 1996; Gonzalez and Stuart 1996; Shadwick, Bowman, and Morrison 1999). Coomes (1997) demonstrates that such integrations are often shadowable. In particular, if $\mathcal{M}$ is the submanifold of interest (*eg.,* symplectic manifold or energy surface) on which the initial condition $\mathbf{y}_0$ lies, then a shadow of the pseudo-orbit $\mathbf{Y} = \{\mathbf{y}_i\}_{i=0}^N$ exists in $\mathcal{M}$ if $\mathbf{Y}$ has sufficiently small local error, remains close to $\mathcal{M}$, avoids the neighborhood of fixed points of $\mathbf{f}$, and the variational equation along $\mathbf{Y}$ exhibits sufficient hyperbolicity. This is a very significant result for problems in which such submanifolds occur, most notably Hamiltonian systems.

2.2.7 Are shadows typical of true orbits chosen at random?

The presence of a shadowing orbit does not imply that the statistical properties of the numerical orbit are typical of those of true orbits chosen at random; the shadowing orbit might be atypical (Quinlan and Tremaine 1992). This observation is perhaps the most fundamental open question remaining for shadowing research. Although it is not directly related to the work in this thesis, it is important enough to discuss briefly.

For example, consider the binary shift map $x_{i+1} = 2x_i \bmod 1$. Iteration on a computer that uses binary floating point arithmetic always results in $x_i = 0$ after a finite (and relatively small) number of iterations. Although $\{x_i = 0\}_{i=m}^{\infty}$ for some m is a valid exact orbit, it is highly atypical, with misleading statistical properties (Farmer and Sidorowich 1991). Fryska and Zohdy (1992) proved that numerical simulation of a simple piecewise linear ODE sometimes produces solutions with substantially different statistical properties than the closed-form solution. This idea is taken further by Corless (1994b) (see also Corless 1992a), who studies the Gauss map,

$$G(x) = \begin{cases} 0, & \text{if } x = 0, \\ x^{-1} \bmod 1, & \text{otherwise.} \end{cases} \tag{2.20}$$

This well-known map has several properties which make it very interesting, especially from the shadowing viewpoint (Corless 1994b):

1. The orbit $\{x_i\}$ (where $x_{i+1} = G(x_i),\ i = 0 \ldots \infty$) of every rational initial point x_0 goes to zero in a finite number of iterations. The rationals are dense in [0,1].
2. An orbit is ultimately periodic if and only if it starts from a *quadratic irrational* or, trivially, a rational initial point. Quadratic irrationals are roots of quadratics with integer coefficients, and are dense in [0,1]. Like the rationals, they are countable, and hence of measure zero. There are an infinite number of orbits with each period.
3. The map is ergodic, meaning almost all initial points have orbits that are dense in [0,1].
4. The Lyapunov exponent of this map is, for almost all initial points, $\pi^2/(6\log 2) \approx 2.3731$, but is *undefined* for rational initial points and is *different* for each quadratic irrational initial point.

As Corless (1994b) states, we see that there are "formidable numerical difficulties in simulating this map." From point #1, we see that unless our numerical orbit converges to zero in a finite number of iterations, it is not representing the properties of the exact orbit starting at our (numerically represented) initial point. Since any numerical orbit must ultimately be periodic, and if our numerical orbit does *not* converge to zero, we see from point #2 that we can only shadow periodic solutions whose initial points are unrepresentable. From point #4 we see that a numerically computed Lyapunov exponent may be completely unrepresentative of almost all orbits. Paradoxically, the numerically computed Lyapunov exponent *does* give a good approximation to the almost-sure value. In fact, a very strong shadowing result can be proved (Corless 1992a; Corless 1997). However, from point #2, we see that, ultimately, we can shadow only periodic orbits, and thus the shadow that follows our numerical solution has a quadratic irrational initial point, and thus does not have a dense orbit (point #3) or the "correct" Lyapunov exponent. The final resolution of this paradox must account for the fact that the true shadowing orbit behaves like a typical orbit, even though it is not. An analysis of this behaviour is provided by Corless (1994b), based upon Góra and Boyarsky (1988).

On the other hand, Góra and Boyarsky (1988) showed that long pseudo-trajectories of a one-dimensional map τ satisfying some special properties have densities which approach that of τ itself. This is an exciting result, and if it can be generalized to continuous systems of arbitrary dimension, it may go a long way towards answering this question.

A weak result concerning this question can be abstracted from Coomes, Koçak, and Palmer (1997). The paper is chiefly concerned with shadowing long periodic orbits, and they use the Lorenz equations as their example. Long-term solutions to the Lorenz system are confined approximately to two disks in three-space (cf. Figure 4.1, p. 56 and §4.1.1, p. 55), and solutions generally jump between the two disks chaotically. If a revolution around one disk is labelled '0' and a revolution around the other is labelled '1', Coomes, Koçak, and Palmer (1997) demonstrated that they were able to build pseudo-trajectories with an arbitrary sequence of '0's and '1's, and then prove the existence of periodic shadows for these pseudo-trajectories. This eliminates at least one simple kind of bias: if we assume that true periodic orbits of the

Lorenz system chosen at random can produce arbitrary sequences of 0s and 1s, it appears that we can build pseudo-trajectories that possess each sequence, and so shadows of the Lorenz system are not biased in such a way as to disallow certain sequences. Palmer and Stoffer (1995) demonstrate a similar result for the Hénon map.

Note that if shadows are generally atypical of true orbits chosen at random, then the properties of the original pseudo-trajectories that produce the shadows are also atypical. This conclusion would have grave implications for the vast quantities of literature over the past several decades that have studied problems numerically. If otherwise reliable-looking pseudo-trajectories *are* atypical, they must be atypical in an extremely subtle way, because researchers have been making apparently reliable, self-consistent, peer-reviewed conclusions based on numerical simulations for decades. Considering that shadowing is only one of many available methods of error analysis, it would be very surprising (to say the least!) if shadows and their otherwise reliable-looking parent pseudo-trajectories were atypical in a substantial way. This does not mean that the problem should not be studied, of course; the apparently small chance that pseudo-trajectories are substantially atypical is balanced by the importance of proving that they are not.

Finally, we would like to point out that similar criticisms can be levelled against *all* forms of backward error analysis. For example, defect analysis says that the solution obtained by a defect-controlled method is the exact solution to a nearby problem in which the right-hand-side of the ODE suffers a small time-varying perturbation. We can then ask, "Is this slightly perturbed problem typical of nearby problems chosen at random?" Or even more pointedly, we can ask if the perturbations are typical of perturbations suffered by a real-life system? We argue in section 1.2 that the answer is sometimes "no". This criticism can also be levelled at the method of modified equations. Even symplectic integrations, which have received much attention recently, suffer the same problem: a solution to a Hamiltonian problem integrated with a symplectic integrator is guaranteed to be exponentially close to the exact solution of a nearby Hamiltonian problem; but is that nearby Hamiltonian problem typical of (pertinent) nearby Hamiltonian problems chosen at random?

This discussion illustrates that answering the question, "Are shadows typical of exact solutions chosen at random?" may be a very difficult one to answer, and that to be fair, we must ask similar questions of other forms of backward error analysis.

Chapter 3

Containment

3.1 Introduction

The method in this thesis relies on a very simple geometrical argument, which is a generalization of the *containment* process first introduced by Grebogi, Hammel, Yorke, and Sauer (1990), hereafter refered to as GHYS.

Although containment was the first method introduced for proving the existence of finite-time shadows of numerical orbits, and even though it is in this author's opinion the most intuitive and easily understood method for proving the existence of shadows, it has not, to this author's knowledge, been pursued beyond its initial conception. This thesis fills that gap, and demonstrates that at least in the cases of no more than one contracting or expanding dimension, containment is about as strong a method as any currently in the literature.

3.1.1 Chapter outline

We first present the proofs that are central to the thesis in section 3.2. Formally, these proofs break into two steps. First, we must prove that $\varphi(M_i)$ and M_{i+1} satisfy the property analogous to the "plus sign" of GHYS (cf. Figure 2.1 on page 14). We call this property the (n,k)-*Inductive Containment Property* (ICP for short), and it is formalized in n dimensions for k expanding directions and $n-k$ contracting directions in section 3.3. The Inductive Containment Property can be proven computationally using a validated ODE integrator; we defer discussion of how to prove ICP until section 3.5. Second, we must show that the Inductive Containment Property implies the existence of a shadow. We prove this in n dimensions for the cases that there are either no more than one expanding direction (§3.2.3), or no more than one contracting direction (§3.2.4). We present our ideas for extending these proofs to the general case in n dimensions in section 3.4.2. For now, we assume that φ is simply a map; extending it to apply effectively to ODE integrations requires a modification dealing with the rescaling of time, which is presented in section 3.6.

3.2 Containment theorems and proofs

3.2.1 Containment in two dimensions

For an introduction to containment applied to two-dimensional maps, the reader is referred to section 2.2.2, starting on page 14. Here we provide a proof of what we call the $(2,1)$-Inductive Containment Theorem, *viz.* the two-dimensional case in which one direction is expanding and the other is contracting. The proof is more rigorous and formal than previous containment proofs that have appeared in the literature, and demonstrates some of the ideas used in the higher-dimensional proofs that appear in the following sections. Furthermore, previous proofs of containment required explicit *a priori* bounds on spatial derivatives, whereas our proof requires no such bounds.[1]

Let M_i be a parallelogram in $\mathbf{R}^2$ with sides oriented in the order $E_i^{-1}, C_i^{-1}, E_i^{+1}, C_i^{+1}$, for $i = 0, \ldots, N$. We denote the union of a set of faces by listing multiple integers in the superscript. Let $\partial_X M_i \equiv E_i^{-1} \cup E_i^{+1} \equiv E_i^{\pm 1}$, and $\partial_C M_i \equiv C_i^{-1} \cup C_i^{+1} \equiv C_i^{\pm 1}$. Let $\varphi : \mathbf{R}^2 \to \mathbf{R}^2$ be a homeomorphism. Let *int* X represent the interior of X. Then M_i and M_{i+1} satisfy the $(2,1)$-*Inductive Containment Property* if

(1) $\varphi(E_i^{\pm 1}) \cap M_{i+1} = \emptyset$, and $\varphi(E_i^{-1})$ and $\varphi(E_i^{+1})$ are on opposite sides of M_{i+1}, *i.e.,* on opposite sides of the infinite slab between the lines containing E_{i+1}^{-1} and E_{i+1}^{+1}.

(2) $\exists Q_{i+1}$, a compact convex set s.t. $\varphi(M_i) \subset \mathit{int}\ Q_{i+1}$, $Q_{i+1} \cap E_{i+1}^{j} \neq \emptyset$ for $j = \pm 1$, and $Q_{i+1} \cap C_{i+1}^{\pm 1} = \emptyset$.

Let $\gamma_0 \subset M_0$ be a simple curve joining E_0^{-1} to E_0^{+1} and remaining in the interior of M_0, *i.e.,*

$$\mathit{int}\ \gamma_0 \subset \mathit{int}\ M_0 \ \wedge\ \gamma_0 \cap E_0^{-1} \neq \emptyset \ \wedge\ \gamma_0 \cap E_0^{+1} \neq \emptyset,$$

where $\wedge$ means "and".

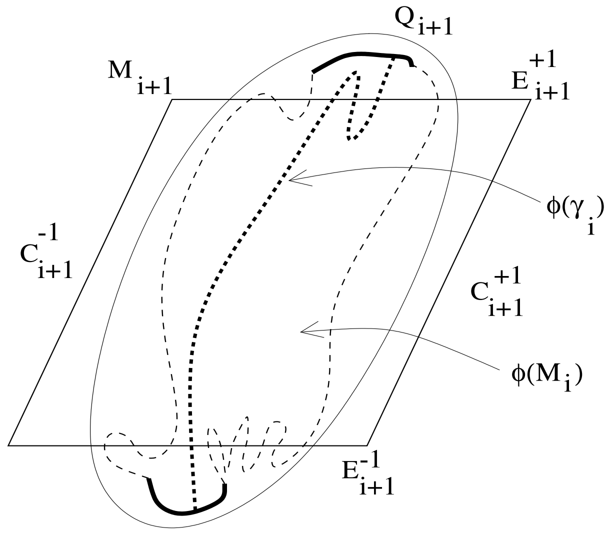

Figure 3.1: The image $\varphi(M_i)$ and M_{i+1}. The solid dark curves at the bottom and top are $\varphi(E_i^{-1})$ and $\varphi(E_i^{+1})$, respectively. The dashed curves at the left and right are $\varphi(C_i^{-1})$ and $\varphi(C_i^{+1})$, respectively.

Theorem 3.1 ((2,1)-Inductive Containment Theorem). *If M_i and M_{i+1} satisfy (2,1)-ICP $\forall i = 0, \ldots, N-1$, then*

$$\forall i = 0, \ldots, N\ \exists \textit{ simple curve } \gamma_i \subseteq \varphi^i(\gamma_0) \textit{ s.t. } \mathit{int}\ \gamma_i \subset \mathit{int}\ M_i \ \wedge\ \gamma_i \cap E_i^{-1} \neq \emptyset \ \wedge\ \gamma_i \cap E_i^{+1} \neq \emptyset,$$

i.e., γ_i touches the boundary of M_i in precisely two places, connecting E_i^{-1} to E_i^{+1}, and otherwise remains entirely inside M_i.

[1] Of course, our algorithm (Nedialkov 1999) must compute bounds on derivatives in order to compute enclosures, but these bounds are not *a priori*; they are computed on-the-fly, and if a bounds check fails, we can always try a smaller timestep to compensate.

Proof. By induction on i. The proof of the base case $i = 0$ is immediate, by the definition of γ_0. For the inductive case, assume $\exists$ simple curve $\gamma_i \subseteq \varphi^i(\gamma_0)$ s.t. $int\ \gamma_i \subset int\ M_i \ \wedge \ \gamma_i \cap E_i^{-1} \neq \emptyset \ \wedge \ \gamma_i \cap E_i^{+1} \neq \emptyset$. First, since Q_{i+1} is convex and intersects E_{i+1}^{-1} and E_{i+1}^{+1} but not $C_{i+1}^{\pm 1}$, $Q_{i+1} - M_{i+1}$ is disconnected into two components, say Q_{i+1}^{-1} and Q_{i+1}^{+1}, and Q_{i+1} encloses $\varphi(E_i^{-1})$ and $\varphi(E_i^{+1})$, which are on opposite sides of M_{i+1} with $\varphi(E_i^{\pm 1}) \cap M_{i+1} = \emptyset$, by ICP(1). Without loss of generality, assume $\varphi(E_i^j) \subset Q_{i+1}^j$, $j = \pm 1$. Now, consider one of the components, say Q_{i+1}^{-1}. It contains one of the two endpoints of $\varphi(\gamma_i)$ since that endpoint is a point of $\varphi(E_i^{-1}) \subset Q_{i+1}^{-1}$, while the other endpoint of $\varphi(\gamma_i)$ is in $\varphi(E_i^{+1}) \subset Q_{i+1}^{+1}$. Since γ_i is a simple curve and φ is a homeomorphism, $\varphi(\gamma_i)$ is a simple curve. Now, $Q_{i+1}^{-1} \cap Q_{i+1}^{+1} = \emptyset$, and $\varphi(\gamma_i)$ connects the two. Thus, $\varphi(\gamma_i)$ must cross the boundary of Q_{i+1}^{-1}. This boundary consists of exactly two contiguous segments, one of which is a segment of ∂Q_{i+1}, while the other is a segment of E_{i+1}^{-1}. Since $\varphi(\gamma_i) \subset \varphi(M_i) \subset int\ Q_{i+1}$, $\varphi(\gamma_i) \cap \partial Q_{i+1} = \emptyset$, and so $\varphi(\gamma_i)$ leaves Q_{i+1}^{-1} through E_{i+1}^{-1}. A similar argument shows that $\varphi(\gamma_i)$ leaves Q_{i+1}^{+1} through E_{i+1}^{+1}. Thus, $\varphi(\gamma_i) \cap E_{i+1}^k \neq \emptyset$ for $k = \pm 1$.

Since $\varphi(\gamma_i)$ is a simple curve, by definition there exists a parameterization $\gamma(t)$ for $t \in [0, 1]$ s.t. $\gamma([0, 1]) = \varphi(\gamma_i)$ and $\gamma(t)$ is a homeomorphism(Munkres 1975). Let $s^{-1} = \varphi(\gamma_i) \cap E_{i+1}^{-1}$ and $s^{+1} = \varphi(\gamma_i) \cap E_{i+1}^{+1}$. Now, s^{-1} and s^{+1} are disjoint since $E_{i+1}^{-1} \cap E_{i+1}^{+1} = \emptyset$, they are compact because E_{i+1}^k and γ_i are compact and φ is a homeomorphism and the intersection of two compact sets in $\mathbf{R}^n$ is compact. Finally, $\gamma^{-1}(s^{\pm 1})$ is compact because γ is a homeomorphism. To prove that there exists a simple curve $\gamma_{i+1} \subset \varphi(\gamma_i)$ s.t. $int\ \gamma_{i+1} \subset int\ M_{i+1}$, we need to show that there exist two points in $[0, 1]$, one each from $\gamma^{-1}(s^{-1})$ and $\gamma^{-1}(s^{+1})$, such that no points from either set are between them. This will prove that there exists a simple curve, which is a section of $\varphi(\gamma_i)$, that connects E_{i+1}^{-1} to E_{i+1}^{+1} without otherwise intersecting ∂M_{i+1}. The following lemma completes the proof. $\square$

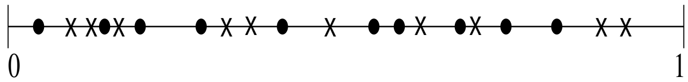

Figure 3.2: Schematic representation of the sets $\gamma^{-1}(s^{-1})$ (dots) and $\gamma^{-1}(s^{+1})$ ($\times$'s).

Lemma 3.2. *Let G and R be (possibly infinite) disjoint compact nonempty subsets of* $[0, 1]$. *Then $\exists g \in G$, $r \in R$ s.t. the open interval $(g, r) \cap (G \cup R) = \emptyset$.*

Proof. Consider the function $f(x, y) = |x - y|$ over the subset $G \times R$ of the plane. f is continuous and $G \times R$ is compact. Thus, f attains its minimum at some point $(g, r) \in G \times R$, *i.e.*, $|g - r| \leq |g' - r'|$ for any other $g' \in G$, $r' \in R$. Thus, $\nexists$ an element of either set between g and r, so the open interval (g, r) is disjoint from $G \cup R$. $\square$

Theorem 3.3 (Shadowing Containment Theorem). *Let $\{M_i\}_{i=0}^N$ be a sequence of parallelepipeds enclosing a pseudo-trajectory $\{\mathbf{y}_i\}_{i=0}^N$. Let ε be the maximum diameter of M_i over i. Let $\gamma_i \subset M_i, \gamma_i \neq \emptyset$, $i = 0, \ldots, N$ and let $\gamma_{i+1} \subseteq \varphi(\gamma_i)$, $i = 0, \ldots, N - 1$. Then $\exists$ an ε-shadow $\{\mathbf{x}_i\}_{i=0}^N$ of $\{\mathbf{y}_i\}_{i=0}^N$, i.e., $|\mathbf{x}_i - \mathbf{y}_i| \leq \varepsilon$, $i = 0, \ldots, N$.*

Proof. Pick any point $\mathbf{x}_N \in \gamma_N$, and recursively define $\mathbf{x}_i = \varphi^{-1}(\mathbf{x}_{i+1})$, $i = N-1, N-2, \ldots, 0$. Since φ is a homeomorphism, it is uniquely invertible, and so $\mathbf{x}_i \in \gamma_i$, $i = 0, \ldots, N$ since

$$\gamma_{i+1} \subseteq \varphi(\gamma_i) \Longrightarrow \varphi^{-1}(\gamma_{i+1}) \subseteq \gamma_i \text{ and } \mathbf{x}_{i+1} \in \gamma_{i+1} \Longrightarrow \mathbf{x}_i = \varphi^{-1}(\mathbf{x}_{i+1}) \in \gamma_i.$$

Since $\mathbf{y}_i \in M_i$ and $\mathbf{x}_i \in \gamma_i \subset M_i$, $|\mathbf{y}_i - \mathbf{x}_i| < \text{diam}(M_i) \leq \varepsilon$. $\square$

Thus, applying Theorem 3.3 to an orbit satisfying the (2,1)-Inductive Containment Property implies the existence of a shadow.

Remark: Note that Theorem 3.3 is independent of the number of dimensions n, and independent of the number of expanding and contracting directions, because the only parts of the Inductive Containment Theorem that are used are the conclusions that $\gamma_{i+1} \subseteq \varphi(\gamma_i)$ for $i = 0, 1, \ldots, N-1$ and $\gamma_i \subset M_i$ for $i = 0, \ldots, N$. As will be seen, the $(n,1)$ and $(n, n-1)$ Inductive Containment Theorems also assert this property. The 0-expanding and 0-contracting directions are handled separately. We conjecture that the general (n,k)-Inductive Containment Theorem will also assert this property, so that the above Shadowing Containment Theorem is applicable to the general (n,k) case.

3.2.2 Informal description of containment in 3 dimensions

The process described by GHYS and rigorously proved above is not directly applicable to systems with more than 2 dimensions, and GHYS provided no argument for how it could be extended beyond 2 dimensions. We describe the method in 3 dimensions, in which there are precisely two interesting cases:

(i) 1 expanding direction, and 2 contracting (Figure 3.3). Assume that the z direction is expanding, while the x and y directions are contracting. (We assume, for simplicity of exposition and for ease of drawing, that these three directions are roughly orthogonal, although in practice they need only be resolvable from each other.) Then, analogous to the 2 dimensional argument, we draw *cubes* M_i around the noisy points $\mathbf{y}_i$, and require that $\varphi(M_i)$ maps over M_{i+1} so that φ stretches M_i into a long, thin tube, a segment of which lies wholly in M_{i+1}. Then, precisely as in the 2-dimensional case, we introduce

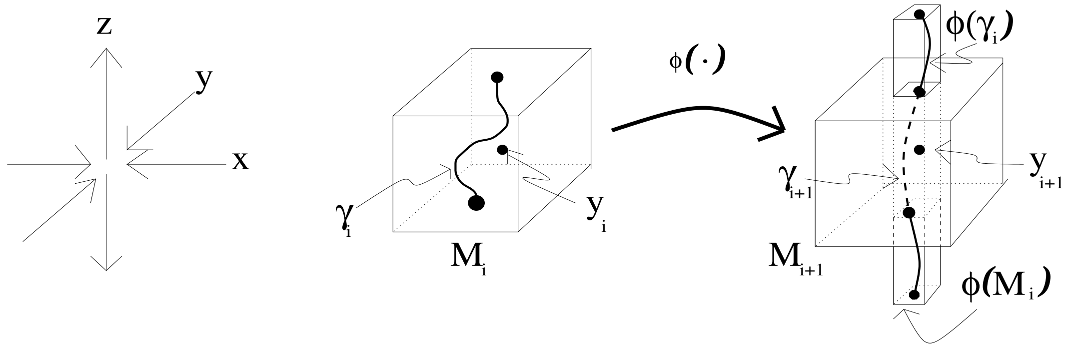

Figure 3.3: Containment in 3D, case (i): 1 expanding direction and 2 contracting.

a curve γ_i that runs approximately along the expanding (vertical) direction from any point on the top of M_i to its bottom. If $\varphi(M_i)$ maps over M_{i+1} as in Figure 3.3, then we are guaranteed that a contiguous section of $\varphi(\gamma_i)$ lies inside M_{i+1}, connecting its top and bottom along the expanding direction. This becomes γ_{i+1}, and by induction γ_N lies inside M_N, and any point $\mathbf{x}_N$ on it can be traced backwards to a point $\mathbf{x}_i \in M_i$ for $i = 0, 1, \ldots, N-1$.

(ii) 2 expanding and 1 contracting direction (Figure 3.4). Assume now that the z (vertical) direction is contracting, while the x and y directions are expanding. We again draw a cube M_i around each noisy point $\mathbf{y}_i$, except now $\varphi(M_i)$ maps over M_{i+1} so that φ flattens M_i into a thin slice, cutting M_{i+1} into 3 pieces, the middle piece of which contains a contiguous section of $\varphi(M_i)$. Now, γ_i must be a *surface*, whose boundary connects all of

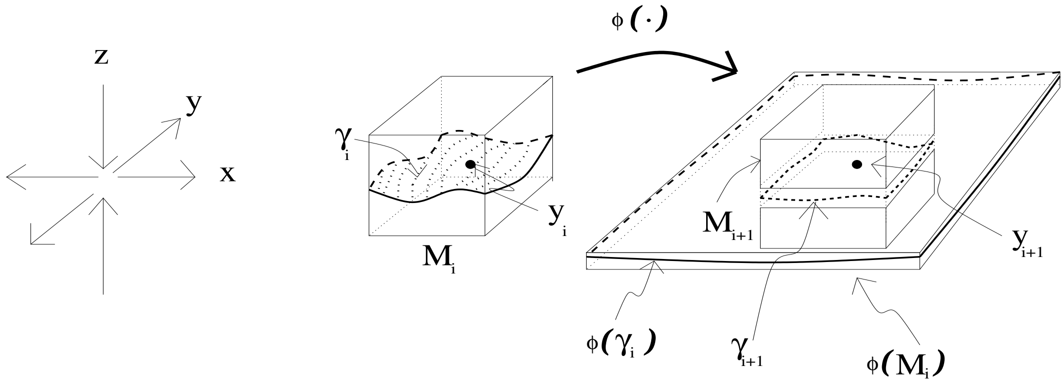

Figure 3.4: Containment in 3D, case (ii): 2 expanding directions and 1 contracting.

the expanding sides, so that under the mapping, $\varphi(\gamma_i)$ is stretched in all the directions it has extent (both horizontal directions), and is "compressed" along the direction it has measure zero (vertical). Then, we are guaranteed that there is a contiguous segment of $\varphi(\gamma_i)$ lying wholly in M_{i+1} and connecting all of its expanding sides. We call this surface γ_{i+1}, and by induction γ_N lies wholly within M_N, and any point on γ_N, traced backwards to a point $\mathbf{x}_i \in M_i$ for $i = 0, 1, \ldots, N-1$.

It seems intuitively clear that we can replace "cube" with "n-cube", "surface" with "manifold", and the above argument still applies in arbitrarily high dimension. The crucial points appear to be that γ has dimension equal to the number of expanding directions, and that its border must "wrap around" all the expanding sides of M_i.[2]

3.2.3 Containment in n dimensions with one expanding direction

Let M_i be a parallelepiped in $\mathbf{R}^n$ with faces F_i^j, for $i = 0, \ldots, N$ and $j = \pm 1, \ldots, \pm n$, with opposite signs in the superscript representing opposite faces of a parallelepiped. Without loss of generality, we assume that the first direction is the "expanding" one. We will denote the union of a set of faces by listing all of them in the superscript; for example, $F_i^{\pm 1, \ldots, \pm(n-1)}$ represents the set of all the faces of M_i except F_i^{-n} and F_i^{+n}. Let $\partial_X M_i \equiv F_i^{-1} \cup F_i^{+1} \equiv F_i^{\pm 1}$, and $\partial_C M_i \equiv \bigcup_{j=2}^n F_i^{-j} \cup F_i^{+j} \equiv F_i^{\pm 2, \ldots, \pm n}$. Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a homeomorphism. Let *int* X represent the interior of X. Then M_i and M_{i+1} satisfy the $(n,1)$-*Inductive Containment Property* (called ICP for short throughout this section) if

(1) $\varphi(F_i^{\pm 1}) \cap M_{i+1} = \emptyset$, and $\varphi(F_i^{-1})$ and $\varphi(F_i^{+1})$ are on opposite sides of the infinite slab between the two hyperplanes containing F_{i+1}^{-1} and F_{i+1}^{+1}, respectively.

(2) $\exists Q_{i+1}$, a parallelepiped in $\mathbf{R}^n$ with faces G_{i+1}^j parallel to the faces F_{i+1}^j of M_{i+1} for $j = \pm 1, \ldots, \pm n$ s.t.

   (i) $\varphi(M_i) \subset int\ Q_{i+1}$,

   (ii) $F_{i+1}^{\pm 2, \ldots, \pm n} \cap Q_{i+1} = \emptyset$, and $\forall j \in \{2, \ldots, n\}$, F_{i+1}^{-j} and F_{i+1}^{+j} are on opposite sides of the infinite slab between the two hyperplanes containing G_{i+1}^{-j} and G_{i+1}^{+j}, respectively.

[2]Discussion of how the the phrase "wrap around" generalizes to higher dimensions is beyond the scope of this thesis, although it can be defined precisely by means of homotopy theory (see for example Munkres 1975).

Let $\gamma_0 \subset M_0$ be a simple curve joining F_0^{-1} to F_0^{+1} and remaining in the interior of M_0, *i.e.*,

$$int\ \gamma_0 \subset int\ M_0 \ \wedge\ \gamma_0 \cap F_0^{-1} \neq \emptyset \ \wedge\ \gamma_0 \cap F_0^{+1} \neq \emptyset.$$

Theorem 3.4 ($(n,1)$-Inductive Containment Theorem). *If M_i, M_{i+1} satisfy ICP $\forall i = 0,\ldots,N-1$, then $\forall i = 0,\ldots,N$*

$$\exists \text{ simple curve } \gamma_i \subseteq \varphi^i(\gamma_0) \text{ s.t. } int\ \gamma_i \subset int\ M_i \ \wedge\ \gamma_i \cap F_i^{-1} \neq \emptyset \ \wedge\ \gamma_i \cap F_i^{+1} \neq \emptyset, \quad (3.1)$$

i.e., γ_i touches the boundary of M_i in precisely two places, connecting F_i^{-1} to F_i^{+1}, and otherwise remains entirely inside M_i.

Proof. By induction on i. The proof of the base case $i = 0$ is immediate, by the definition of γ_0. For the inductive case, assume $\exists$ a simple curve $\gamma_i \subseteq \varphi^i(\gamma_0)$ s.t. $int\ \gamma_i \subset int\ M_i \ \wedge\ \gamma_i \cap F_i^{-1} \neq \emptyset \ \wedge\ \gamma_i \cap F_i^{+1} \neq \emptyset$. From ICP(1) and $\varphi(F_i^{\pm 1}) \subset Q_{i+1}$ and the fact that Q_{i+1} is convex,

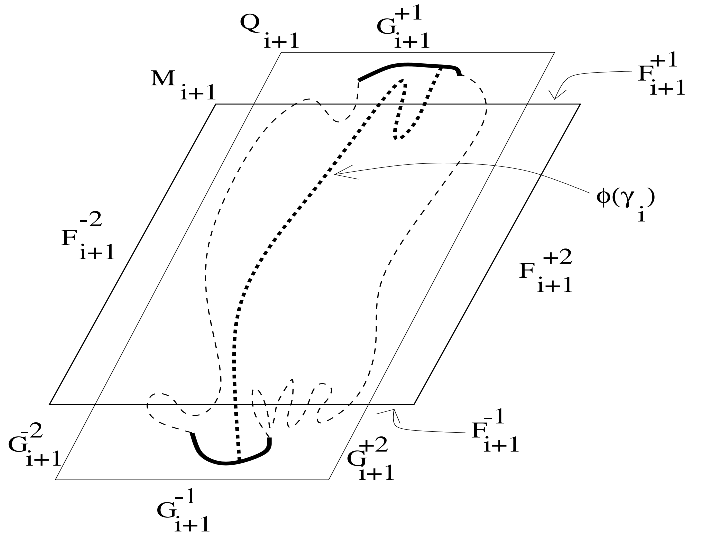

Figure 3.5: The image $\varphi(M_i)$ and M_{i+1} for 2 dimensions. The dark curves at the bottom and top are $\varphi(F_i^{\pm 1})$. The dashed curves at the left and right are $\varphi(F_i^{\pm 2})$.

we know that Q_{i+1} intersects both F_{i+1}^{-1} and F_{i+1}^{+1}; and from ICP(2ii), Q_{i+1} does not intersect $F_{i+1}^{\pm 2,\ldots,\pm n}$. Thus, since Q_{i+1} is convex, $Q_{i+1} - M_{i+1}$ is disconnected by the slab defined in ICP(1) into two disjoint components[3] say Q_{i+1}^{-1} and Q_{i+1}^{+1}, each containing one of $\varphi(F_i^{\pm 1})$, by ICP(1). Without loss of generality, assume $\varphi(F_i^j) \subset Q_{i+1}^j$, $j = \pm 1$. Now, consider one of the components, say Q_{i+1}^{-1}. It contains one of the two endpoints of $\varphi(\gamma_i)$ since that endpoint is a point of $\varphi(F_i^{-1}) \subset Q_{i+1}^{-1}$, while the other endpoint of $\varphi(\gamma_i)$ is in $\varphi(F_i^{+1}) \subset Q_{i+1}^{+1}$. Since γ_i is a simple curve and φ is a homeomorphism, $\varphi(\gamma_i)$ is a simple curve. Now, $Q_{i+1}^{-1} \cap Q_{i+1}^{+1} = \emptyset$, and $\varphi(\gamma_i)$ connects the two. Thus, $\varphi(\gamma_i)$ must cross the boundary of Q_{i+1}^{-1}. This boundary consists of exactly two mutually exclusive patches, one of which is a subset of ∂Q_{i+1}, the other a subset of F_{i+1}^{-1}. Since $\varphi(\gamma_i) \subset \varphi(M_i) \subset int\ Q_{i+1}$, this implies $\varphi(\gamma_i) \cap \partial Q_{i+1} = \emptyset$, and so $\varphi(\gamma_i)$ leaves Q_{i+1}^{-1} through F_{i+1}^{-1}. A similar argument shows that $\varphi(\gamma_i)$ leaves Q_{i+1}^{+1} through F_{i+1}^{+1}. Thus,

[3] This is because F_{i+1}^{-1} and F_{i+1}^{+1} are each patches of an $n-1$ dimensional hyperplane residing in n dimensions, and so they each disconnect any convex set they intersect, as long as that convex set does not intersect their boundaries ∂F_{i+1}^{-1} and ∂F_{i+1}^{+1}, respectively.

$\varphi(\gamma_i) \cap F_{i+1}^j \neq \emptyset,\ j = \pm 1$. It remains to show that there exists a segment γ_{i+1} of $\varphi(\gamma_i)$ which is a simple curve and maintains the property defined in (3.1).

Since $\varphi(\gamma_i)$ is a simple curve, there exists a parameterization $\gamma(t)$ for $t \in [0,1]$ s.t. $\gamma([0,1]) = \varphi(\gamma_i)$ and $\gamma(t)$ is a homeomorphism (Munkres 1975). Let $s^j = \varphi(\gamma_i) \cap F_{i+1}^j$, $j = \pm 1$. Now, s^{-1} and s^{+1} are disjoint since $F_{i+1}^{-1} \cap F_{i+1}^{+1} = \emptyset$, they are compact because F_{i+1}^j for $j = \pm 1$ and γ_i are compact and φ is a homeomorphism and the intersection of two compact sets in $\mathbf{R}^n$ is compact. Finally, $\gamma^{-1}(s^{\pm 1})$ are compact because γ is a homeomorphism. To prove that there exists a simple curve $\gamma_{i+1} \subset \varphi(\gamma_i)$ s.t. $int\ \gamma_{i+1} \subset int\ M_{i+1}$, we need to show that there exist two points in $[0,1]$, one each from $\gamma^{-1}(s^{-1})$ and $\gamma^{-1}(s^{+1})$, such that no points from either set are between them. This will prove that there exists a simple curve, which is a section of $\varphi(\gamma_i)$, that connects F_{i+1}^{-1} to F_{i+1}^{+1} without otherwise intersecting ∂M_{i+1}. Let $G = \gamma^{-1}(s^{-1})$ and $R = \gamma^{-1}(s^{+1})$ and note that G and R are compact, disjoint, non-empty subsets of $[0,1]$. Applying Lemma 3.2 completes the proof. $\square$

Applying Theorem 3.3 proves that a shadow exists for any noisy trajectory $\{\mathbf{y}_i\}_{i=0}^N$ s.t. $\mathbf{y}_i \in M_i,\ \ i = 0, \ldots, N$.

3.2.4 Containment in n dimensions with one contracting direction

Remark: This case could immediately be proved by applying the one-expanding-direction theorem in backward time, since a system with one contracting direction in forward time is equivalent to a system with one expanding direction in backward time. However, we provide a different proof because although it is not applicable to the general case, we believe a proof of the following form is more likely to be generalizable (cf. section 3.4.2).

Remark: In the one expanding direction case, γ is one dimensional; in the one contracting direction case, γ is $(n-1)$-dimensional.

Let M_i be a parallelepiped in $\mathbf{R}^n$ with faces F_i^j, for $i = 0, \ldots, N$ and $j = \pm 1, \ldots, \pm n$, with opposite signs of j representing opposite faces of a parallelepiped. Without loss of generality, let the nth direction be the "contracting" one. We will denote the union of a set of faces by listing all of them in the superscript; for example, $F_i^{\pm 1, \ldots, \pm(n-1)}$ represents the set of all faces except F_i^{-n} and F_i^{+n}.

Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a homeomorphism. Let $int\ X$ represent the interior of X. Then M_i and M_{i+1} satisfy the $(n, n-1)$-*Inductive Containment Property* (called ICP for short throughout this section) if

(1) $\varphi(F_i^{\pm 1, \ldots, \pm(n-1)}) \cap M_{i+1} = \emptyset$ and $\forall j \in \{1, \ldots, n-1\}$ $\varphi(F_i^{-j})$ and $\varphi(F_i^{+j})$ are on opposite sides of the infinite slab between the two hyperplanes containing F_{i+1}^{-j} and F_{i+1}^{+j}, respectively.

(2) $\exists Q_{i+1}$, a parallelepiped in $\mathbf{R}^n$ with faces G_{i+1}^j parallel to the faces F_{i+1}^j of M_{i+1} for $j = \pm 1, \ldots, \pm n$ s.t.

    (i) $\varphi(M_i) \subset int\ Q_{i+1}$,

    (ii) $F_{i+1}^{\pm n} \cap Q_{i+1} = \emptyset$ and F_{i+1}^{-n} and F_{i+1}^{+n} are on opposite sides of the infinite slab between the two hyperplanes containing G_{i+1}^{-n} and G_{i+1}^{+n}, respectively.

Lemma 3.5. *In an n-dimensional cube, the border of each face is contained in the union of all the other faces except the one opposite itself, i.e.,* $\partial F_i^j \subset \bigcup_{\substack{k \in \{\pm 1, \ldots, \pm n\} \\ k \neq \pm j}} F_i^k$.

Proof. Without loss of generality, we will look at the n-dimensional unit cube $C = [0,1]^n$. The border ∂C of C consists of all points $\mathbf{p} = (p_1, \dots, p_n)^T$ that have $p_k = 0$ or 1 for some $k \in \{1, \dots, n\}$ and $p_j \in [0,1]$ for all $j \in \{1, \dots, n\}$. A face F of C is defined by assigning 0 or 1 to precisely one of the co-ordinates p_k of $\mathbf{p}$ (with the choice between 0 and 1 being the choice between opposite faces), and freeing all the other co-ordinates $p_j, j \neq k$ to roam in $[0,1]$. Without loss of generality, let $p_k = 0$. Note that the associated face F is a hypercube of dimension $n-1$. Thus, ∂F consists of all points $\mathbf{q} = (q_1, \dots, q_n)^T$ with $q_k = 0$ and $q_j = 0$ or 1 for $j \neq k$. However, if a co-ordinate q_j for some $j \neq k$ is 0 or 1, then in addition to being in F, $\mathbf{q}$ is in some other face F' of C. Furthermore, F' can be any face of C except F or the one opposite F by choosing j and $q_j \in \{0,1\}$ appropriately. $\square$

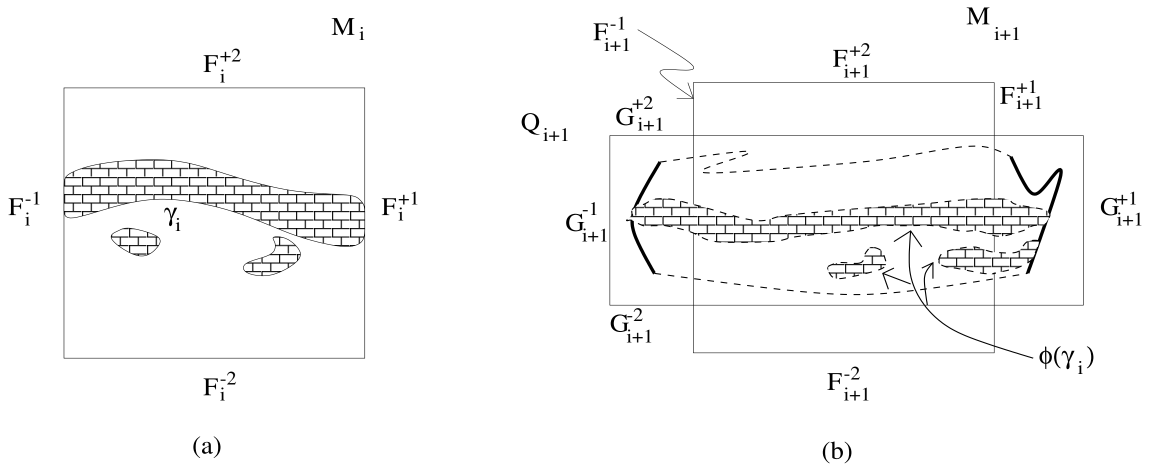

Figure 3.6: Schematic representation of the one-contracting-dimension proof in two dimensions. (a) γ_i, M_i and the faces of M_i. (b) The dark curves at the left and right represent $\varphi(F_i^{\pm 1})$. The dashed curves at the top and bottom are $\varphi(F_i^{\pm 2})$. The proof of Theorem shows that no path in M_{i+1} connects F_{i+1}^{-2} to F_{i+1}^{+2} without intersecting $\gamma_{i+1} = \varphi(\gamma_i) \cap M_{i+1}$.

Lemma 3.6. *Both $\varphi(F_i^{-n})$ and $\varphi(F_i^{+n})$ path disconnect F_{i+1}^{-n} from F_{i+1}^{+n} in M_{i+1}.*

Proof. By Lemma 3.5, the border of F_i^{-n} is contained in $F_i^{\pm 1, \dots, \pm(n-1)}$. Since φ is a homeomorphism, the border of $\varphi(F_i^{-n})$ is contained in $\varphi(F_i^{\pm 1, \dots, \pm(n-1)})$, which by ICP(1) is disjoint from M_{i+1}. However, by ICP(2i), $\varphi(F_i^{-n})$ is contained in the slab between G_{i+1}^{-n} and G_{i+1}^{+n}. Since $\varphi(F_i^{-n})$ is homeomorphic to a face of M_{i+1} and it lies in the slab between G_{i+1}^{-n} and G_{i+1}^{+n} of M_{i+1} and its border lies outside M_{i+1}, it path disconnects M_{i+1} somewhere inside the slab between G_{i+1}^{-n} and G_{i+1}^{+n}. Furthermore, since by ICP(2ii) the slab between G_{i+1}^{-n} and G_{i+1}^{+n} is contained in the slab between F_{i+1}^{-n} and F_{i+1}^{+n}, then F_{i+1}^{-n} and F_{i+1}^{+n} must be path disconnected in M_{i+1} by $\varphi(F_i^{-n})$. The same argument applies to $\varphi(F_i^{+n})$. $\square$

Let $\gamma_0 \subset M_0$ be a set constructed such that γ_0 path-disconnects M_0 in such a way that no path exists in M_0 that connects F_0^{-n} to F_0^{+n} without intersecting γ_0.

Theorem 3.7 (**($n, n-1$) Inductive Containment Theorem**). *If M_i, M_{i+1} satisfy ICP for all $i = 0, \dots, N-1$, then*

$\forall i = 0, \dots, N$ $\exists$ *a set* $\gamma_i \subseteq \varphi^i(\gamma_0)$ *s.t.* $\gamma_i \subset M_i$ *and* γ_i *path disconnects* F_i^{-n} *from* F_i^{+n} *in* M_i.

Proof. By induction on i. The proof of the base case $i = 0$ is immediate, by the definition of γ_0. For the inductive case, assume there exists a set $\gamma_i \subseteq \varphi^i(\gamma_0)$ s.t. $\gamma_i \subset M_i$ and γ_i path disconnects F_i^{-n} from F_i^{+n} in M_i. Assume to the contrary that there exists a path β in M_{i+1} from F_{i+1}^{-n} to F_{i+1}^{+n} that does not intersect $\varphi(\gamma_i)$. By Lemma 3.6, $\varphi(F_i^{-n})$ and $\varphi(F_i^{+n})$ each path disconnect F_{i+1}^{-n} from F_{i+1}^{+n} in M_{i+1}. Thus, β intersects both $\varphi(F_i^{-n})$ and $\varphi(F_i^{+n})$, and connects the two, but by assumption does not intersect $\varphi(\gamma_i)$. Since $\beta \subset M_{i+1}$, ICP(1) implies $\beta \cap \varphi(F_i^{\pm 1,\dots,\pm(n-1)}) = \emptyset$. Thus, β intersects $\varphi(\partial M_i)$ only in $\varphi(F_i^{\pm n})$. Without loss of generality, assume that β intersects each only once, *i.e.,* it enters $\varphi(M_i)$ through $\varphi(F_i^{-n})$ and leaves through $\varphi(F_i^{+n})$. Thus, β connects $\varphi(F_i^{-n})$ to $\varphi(F_i^{+n})$ without leaving $\varphi(M_i)$, and, by assumption, without intersecting $\varphi(\gamma_i)$. However, φ is a homeomorphism, and applying φ^{-1} to $\varphi(F_i^{\pm n}), \varphi(M_i)$, and β, we see that $\varphi^{-1}(\beta)$ is a path from F_i^{-n} to F_i^{+n} that remains in M_i and does not intersect γ_i, contradicting our inductive hypothesis. Thus, any path in M_{i+1} that connects F_{i+1}^{-n} to F_{i+1}^{+n} must intersect $\varphi(\gamma_i)$. Let $\gamma_{i+1} = \varphi(\gamma_i) \cap M_{i+1}$. Then no path exists from F_{i+1}^{-n} to F_{i+1}^{+n} in M_{i+1} that does not intersect γ_{i+1}. $\square$

Thus, $\gamma_{i+1} \subseteq \varphi(\gamma_i)$ and $\gamma_{i+1} \cap M_{i+1} \neq \emptyset$ for all i. Applying Theorem 3.3 proves that a shadow exists for any noisy trajectory $\{\mathbf{y}_i\}_{i=0}^N$ s.t. $\mathbf{y}_i \in M_i, \;\; i = 0, \dots, N$.

3.2.5 Containment with zero contracting or expanding directions

For completeness, we mention the trivial cases in which all directions are contracting, or all directions are expanding. We call these the $(n, 0)$ and (n, n) cases, respectively. The former case is entirely trivial, because the problem is stable: if $\varphi(M_i) \subset \bar{\varphi}(M_i) \subset M_{i+1}$ for all i, then clearly any exact solution starting in M_0 will be in M_i for all $i > 0$. Similarly, if all directions are expanding, then we apply the same argument in the reverse direction: if $\varphi^{-1}(M_{i+1}) \subset \bar{\varphi}^{-1}(M_{i+1}) \subset M_i$ for all i, then any exact solution *finishing* in M_N, traced backwards, lies in M_i for $i = N-1, N-2, \dots, 0$.

3.2.6 Discussion

The four cases $(n, 0), (n, 1), (n, n-1)$, and (n, n) cover *all* cases when $n = 3$. That is, the theorems in this thesis can prove the existence of shadows for any n-dimensional system, $n \leq 3$, in which some measure of pseudo-hyperbolicity is present. Furthermore, although the proofs, for simplicity, only deal with a single function φ, the induction argument could just as easily use a *different* φ at each step, so the proofs work just as well if each step uses a different function φ_i. In particular, φ_i could be the ODE time-h_i solution operator φ_{h_i} from equation 1.4. Thus, modulo a rescaling of time (which we discuss later), the above proofs are valid for use in finding shadows of noisy trajectories of ODE systems, as well as maps, with up to three dependent variables. They can also be used in the case of n dependent variables, with the restriction that solutions have no more than one expanding direction, or no more than one contracting direction.

Ideally, of course, we would like to be able to use containment to prove the existence of shadows for *any* system which displays pseudo-hyperbolicity, regardless of the number of expanding and contracting directions; certainly no other rigorous method currently in the literature has such restrictions on the hyperbolicity. Unfortunately, after months of diligent searching, this author has been unable to prove the general case, even after consulting many other people. We tried about a dozen distinct methods of proof of the general case, without success. Ideas that appear promising at first always evaporate under closer scrutiny. Despite these failures, this author remains optimistic that a proof of the general case exists. At the very least, we know that other, completely different proofs of high-dimensional shadows exist, namely those of Coomes, Koçak, and Palmer. The theorems and proofs presented in this thesis seem so simple and elegant that we are compelled to believe that a similar proof must exist for the general case.

3.3 The general Inductive Containment Property

The essence of the Inductive Containment Property can be explained by looking at a simplified homeomorphism $\psi : \mathbf{R}^n \to \mathbf{R}^n$ with the following properties. Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ and let $\mathbf{x}|_{x_j=a}$ be $\mathbf{x}$ with its jth component replaced with the value $a \in \mathbf{R}$. Let $\psi = (\psi_1(\mathbf{x}), \ldots, \psi_n(\mathbf{x}))^T$. Let $\{1, \ldots, k\}$, $k \in \{1, \ldots, n\}$ be the *expanding directions*. If for all $\mathbf{x}$ in the unit cube $[0,1]^n$, ψ satisfies

$$\forall j \in \{1, \ldots, k\} \quad \psi_j(\mathbf{x}|_{x_j=0}) < 0 \text{ and } \psi_j(\mathbf{x}|_{x_j=1}) > 1,$$

$$\forall j \in \{k+1, \ldots, n\} \quad \psi_j(\mathbf{x}|_{x_j=0}) > 0 \text{ and } \psi_j(\mathbf{x}|_{x_j=1}) < 1,$$

then ψ maps the unit cube $I^n = [0,1]^n$ over itself in such a way that I^n satisfies the Inductive Containment Property with itself. The rightmost diagrams in Figures 3.3 and 3.4, respectively, are schematic representations of the Inductive Containment Property if $n = 3$ and $M_i = M_{i+1} = [0,1]^3$ in each Figure.

It is not hard to see how the Inductive Containment Properties defined in Theorems 3.4 and 3.7 can be generalized. The Inductive Containment Property is topologically unchanged from the above if we compose an arbitrary homeomorphism with ψ. In particular, if $\varphi = L \circ \psi$ where L is an arbitrary linear transformation, then the Inductive Containment Property can be more generally stated as follows.

The (n,k)-Inductive Containment Property Let M_i be a parallelepiped in $\mathbf{R}^n$ with faces F_i^j, for $i = 0, \ldots, N$ and $j = \pm 1, \ldots, \pm n$, with opposite signs in the superscript representing opposite faces of a parallelepiped. Let the first k directions be the *nominally expanding directions*,[4] while the remainder are called *nominally contracting directions*. We will denote the union of a set of faces by listing all of them in the superscript; for example, $F_i^{\pm 1, \ldots, \pm(n-1)}$ represents the set of all faces except F_i^{-n} and F_i^{+n}. Let

$$\partial_X M_i \equiv \bigcup_{j=1}^{k} \left(F_i^{-j} \cup F_i^{+j} \right) \equiv F_i^{\pm 1, \ldots, \pm k}$$

be the set of expanding faces, and

$$\partial_C M_i \equiv \bigcup_{j=k+1}^{n} \left(F_i^{-j} \cup F_i^{+j} \right) \equiv F_i^{\pm(k+1), \ldots, \pm n}$$

be the set of contracting faces. Let $\varphi : \mathbf{R}^n \to \mathbf{R}^n$ be a homeomorphism. Let *int* X represent the interior of X. Then M_i and M_{i+1} satisfy the (n,k)-*Inductive Containment Property* if

(1) $\varphi(\partial_X M_i) \cap M_{i+1} = \emptyset$ and $\forall j \in \{1, \ldots, k\}$, $\varphi(F_i^{-j})$ and $\varphi(F_i^{+j})$ are on opposite sides of the infinite slab between the two hyperplanes containing F_{i+1}^{-j} and F_{i+1}^{+j}, respectively.

[4] *Nominally* because they do not expand uniformly for all time; otherwise the system would be hyperbolic.

(2) $\exists Q_{i+1}$, a parallelepiped in $\mathbf{R}^n$ with faces G^j_{i+1} parallel to the faces F^j_{i+1} of M_{i+1} for $j = \pm 1, \dots, \pm n$ s.t.

 (i) $\varphi(M_i) \subset int\ Q_{i+1}$,

 (ii) $Q_{i+1} \cap \partial_C M_{i+1} = \emptyset$ and $\forall j \in \{k+1, \dots, n\}$, F^{-j}_{i+1} and F^{+j}_{i+1} are on opposite sides of the infinite slab between the two hyperplanes containing G^{-j}_{i+1} and G^{+j}_{i+1}, respectively.

Remark: ICP(1) is probably stronger that we need, because we do not generally care where each of the expanding faces maps to, provided that they "pull" the border of γ_i outside of M_{i+1}. A more topologically sophisticated proof might be constructed assuming only that $\varphi(\partial_X M_i) \cap M_{i+1} = \emptyset$ along with some statement about $\varphi(\partial_X M_i)$ "wrapping around" M_{i+1} in some topological sense. A similar remark may hold for the contracting faces described by ICP(2ii).

3.4 Discussion of containment in the general case

3.4.1 A simplistic linear example

In this subsection, we present an n-dimensional proof in the case that φ is a simple linear function. The intent is to show that there exists an n-dimensional system with an arbitrary number k of expanding directions that we can shadow using containment.

Let $\rho > 1$, let $\mathbf{x} = (x_1, \dots, x_n)^T$, and let $\varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \dots, \varphi_n(\mathbf{x}))^T$, where

$$\varphi_j(\mathbf{x}) = \begin{cases} \rho x_j, & j = 1, \dots, k. \\ \frac{1}{\rho} x_j, & j = k+1, \dots, n. \end{cases}$$

That is, $\varphi(\mathbf{x})$ is linearly expanding about the origin in the first k directions, linearly contracting about the origin in the remaining directions, and each direction is orthogonal to all the others. Note also that φ is clearly hyperbolic along any exact orbit, and so shadows of pseudo-orbits certainly exist for sufficiently small local error. Let M be the n-dimensional cube centred at the origin with maximum diameter $\varepsilon > 0$, *i.e.*, $M = [-\mu, \mu]^n$ where $\mu = \varepsilon/(2\sqrt{n})$, and let $M_i = M$ for all i. We want to demonstrate how (n, k)-containment is applied to this system.

A cursory glance at φ shows that M_i, M_{i+1} satisfy the (n, k)-Inductive Containment Property under φ for all i. Let

$$\gamma_0 = \{\mathbf{z} \in \mathbf{R}^n \mid z_j \in [-\mu, \mu] \text{ for } j \in \{1, \dots, k\}, \text{ and } z_j = 0 \text{ for } j \in \{k+1, \dots, n\}\},$$

and $\gamma_{i+1} = \varphi(\gamma_i) \cap M$. Now pick $\mathbf{x} \in \gamma_0$. Then clearly,

$$\varphi^m(\mathbf{x}) = (\rho^m x_1, \dots, \rho^m x_k, 0, \dots, 0)^T.$$

Theorem 3.8 (Containment in n dimensions for a linear φ). *Let φ, $\{M_i\}_{i=0}^{\infty}$ and γ_0 be defined as above. Then*

$$\forall m \geq 0\ \exists \mathbf{x} \in \gamma_0 \ s.t.\ \varphi^i(\mathbf{x}) \in M \ for\ i = 0, \dots, m.$$

Thus $\{\varphi^i(\mathbf{x})\}_{i=0}^m$ is an ε-shadow of any pseudo-orbit $\{\mathbf{y}_i\}_{i=0}^m$ of φ that remains in M.

Proof. Pick a $\mathbf{z} = (z_1, \dots, z_n)^T \in M_m = M$ such that $z_j \in [-\mu, \mu], j = 1, \dots, k$ and $z_j = 0, j = k+1, \dots, n$. Clearly $z \in M$ and $\left(\varphi^i(\mathbf{z})\right)_j = 0$, for all i and for $j = k+1, \dots, n$, where $(\mathbf{w})_j$

extracts the jth component of the vector $\mathbf{w}$. Looking at the first k (expanding) directions, we want first to show that $\mathbf{z} = \varphi^m(\mathbf{x})$ for some $\mathbf{x} \in \gamma_0$. Pick $x_j = \frac{z_j}{\rho^m}, j = 1, \ldots, k,\ x_j = 0, j = k+1, \ldots, n$. Clearly $\mathbf{x} \in \gamma_0 \subset M$, since $z_j \in [-\mu, \mu]$ and $\rho^m \geq 1$. Then,

$$\begin{aligned} \varphi^m(\mathbf{x}) &= (\rho^m \frac{z_1}{\rho^m}, \ldots, \rho^m \frac{z_k}{\rho^m}, 0, \ldots, 0)^T \\ &= \mathbf{z}. \end{aligned}$$

Finally,

$$\varphi^i(\mathbf{x}) = (\rho^{i-m} z_1, \ldots, \rho^{i-m} z_k, 0, \ldots, 0)^T \in M,$$

since $z_j \in [-\mu, \mu]$ and $\rho^{i-m} \leq 1$ for $i \leq m$. Since the maximum diameter of M_i is ε, and $\mathbf{y}_i \in M_i$ for $i = 0, \ldots, m$, $\{\varphi^i(\mathbf{x})\}_{i=0}^m$ ε-shadows $\{\mathbf{y}_i\}_{i=0}^m$. □

Thus, this system satisfies the (n,k)-Inductive Containment Property, and is shadowable. What remains to prove is that the former always implies the latter.

Although this argument is not very useful in itself, it is still interesting. Note that we can apply an arbitrary homeomorphism to the objects in the theorem, producing a theorem that is applicable to the general n-dimensional case in which $\varphi(\gamma_i) \cap M_{i+1}$ is topologically identical to γ_i. In particular, the topological aspects of the argument are unchanged in the case that we can prove that $\varphi(\gamma_i)$ never causes $\delta\gamma_i$ to "loop back" on itself and intersect M_{i+1} in a manner that produces topological entities often described as "handles", "ears", or "fingers". In fact, the original proof of two-dimensional containment (Grebogi, Hammel, Yorke, and Sauer 1990) contained the restriction that the first and second spatial derivatives of φ needed to be bounded. We showed in section 3.2.1 that the two-dimensional theorem still holds when such restrictions are lifted. It seems plausible to conjecture that the same could be true in the general n-dimensional case.

3.4.2 Ideas for proving the general case

The fundamental reason we believe the general (n,k) case to be provable is because φ is a homeomorphism, and thus introduces no holes into $\varphi(\gamma_i)$ that did not exist in γ_i. So if a k-dimensional γ_i "covers" M_i in the k expanding directions, and $\varphi(M_i)$ stretches M_i over M_{i+1} in those same directions, then $\varphi(\gamma_i)$ will "cover" M_{i+1} in the k expanding directions. The only place holes are introduced is where $\varphi(\gamma_i)$ intersects $\delta_X M_{i+1}$, at which point fingers, ears, and handles may be cut off, introducing holes of unknown topology where $\varphi(\gamma_i)$ meets $\partial_X M_{i+1}$. Now, recall that $\partial_X M_i$ is the set of expanding faces of M_i, so that $\varphi(\partial_X M_i) \cap M_{i+1} = \emptyset$. However, these new holes in $\gamma_{i+1} = \varphi(\gamma_i) \cap M_{i+1}$ are of no consequence because $\varphi(\partial_X M_{i+1})$ in turn lies outside M_{i+2}; the only part of $\varphi(\gamma_{i+1})$ that is inside M_{i+2} was also inside M_{i+1}, inside of which there were no holes in γ_{i+1}. Thus, there are no new holes inside $\varphi(\gamma_{i+1}) \cap M_{i+2}$, and by induction, no holes inside any $\gamma_i \cap M_i$, $\quad i = 0, \ldots, N$. Even more succinctly, the only holes in γ_{i+1} are in its border $\partial\gamma_{i+1} \equiv \varphi(\gamma_i) \cap \partial_X M_{i+1}$, and the Inductive Containment Property ensures that $\varphi(\partial\gamma_{i+1}) \subset \varphi(\partial_X M_{i+1})$ is outside M_{i+2}. Here, a "hole" may be defined as something through which an $(n-k)$-dimensional manifold β can pass, analogous to the simple curve β created to induce a contradiction in the proof of Theorem 3.7. Note that this argument does not claim that no holes (fingers, handles, or ears) exist in γ_i; it merely says that they are of no consequence, because if a hole exists in γ_{i+1} through which a β can pass, then a similar hole must have existed in γ_i. If we start with a γ_0 with no such holes, then a contradiction results, and a proof analogous to that of Theorem 3.7 would hold.

Unfortunately, formalizing this argument has proved surprisingly difficult. A start may be to generalize Lemmas 3.5 and 3.6.

3.5 Four ways to verify the Inductive Containment Property

We have devised four different methods of verifying that the general Inductive Containment Property holds for a given pseudo-trajectory derived from the numerical solution of an ODE. We note in passing that all of these schemes could easily be adapted to the simpler problem of maps. Each has strengths and weaknesses, which we will discuss. Each one requires the use of interval arithmetic, or a validated ODE integrator (cf. §1.3.2) if φ derives from an ODE. The validated ODE integrator that we use is called VNODE (Nedialkov 1999). VNODE works with n-dimensional parallelepipeds, and satisfies the following property: given an n-dimensional parallelepiped A and a timestep h, VNODE will return an n-dimensional parallelepiped B such that $\varphi_h(A) \subset B$, where φ_h is the solution operator for the ODE defined in equation (1.4). For the purposes of this description, we will denote the output B as $\bar{\varphi}_h(A)$,

$$\varphi_h(A) \subset \bar{\varphi}_h(A).$$

We will usually omit the timestep parameter h; we will talk only of φ, keeping in mind that in the induction, φ can be different for each step.

3.5.1 Direct integration of all $2n$ faces

The most direct method of building a pair of parallelepipeds M_i, M_{i+1} satisfying the Inductive Containment Property is to integrate each face of M_i individually, building M_{i+1} explicitly to satisfy the Inductive Containment Property. See Figure 3.7. A separate validated ODE

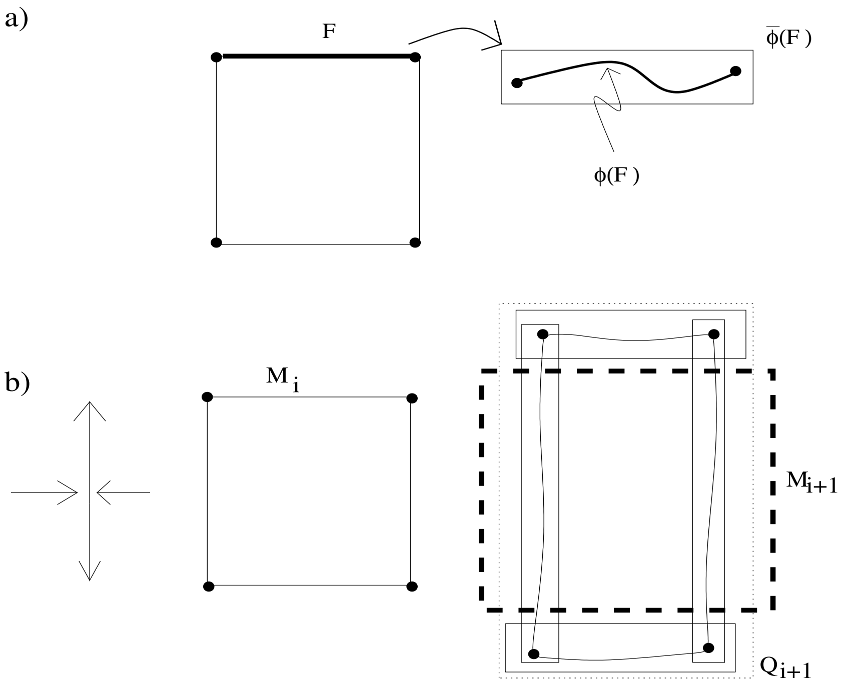

Figure 3.7: Schematic diagram of the direct $2n$ face-integration approach to proving the Inductive Containment Property. (a) An enclosure of the image of each individual face is built. (b) From these enclosures, M_{i+1} is built. Q_{i+1} is a parallelepiped enclosing $\varphi(M_i)$ by enclosing the enclosures of all its faces. For simplicity, M_i, Q_{i+1} and M_{i+1} are all drawn axis-aligned, although in reality they may be arbitrarily oriented (as long as they are identically oriented).

integration is performed on each of the $2n$ faces of M_i, as shown in Figure 3.7.a. A face F is simply represented by an n-dimensional parallelepiped which has zero width in one particular dimension. The image of F under φ lies inside the box $\bar{\varphi}(F)$. In Figure 3.7.b, the vertical direction is depicted as expanding, while the horizontal direction is contracting. In the rightmost section of the Figure, the rectangles with thin solid boundaries depict $\bar{\varphi}(F)$ for each face F of M_i. The thin dotted box is the boundary of Q_{i+1}, which is a parallelepiped convex hull enclosing the images of all the faces, thus ensuring ICP(2i). M_{i+1} (heavy dashed line) is built as follows: for the expanding directions $j = \pm 1, \ldots, \pm k$, F^j_{i+1} is aligned one machine number inside the inner boundary of $\bar{\varphi}(F^j_i)$, thus enforcing ICP(1). For the contracting directions $j = \pm(k+1), \ldots, \pm n$, F^j_{i+1} is aligned strictly outside the boundary of Q_{i+1}, thus enforcing ICP(2ii). Note that it is probably sufficient if the contracting faces of M_{i+1} are aligned strictly outside the outer boundary of $\bar{\varphi}(F^j_i), j = \pm(k+1), \ldots, \pm n$. This choice would give a slightly smaller M_{i+1}, and thus tighter containment, although it would complicate the proof of the Inductive Containment Theorem slightly.

As with any shadowing method, the system becomes hard to shadow when it lacks pseudo-hyperbolicity. This means either that the nominally contracting directions fail to contract enough for us to detect contraction, or the nominally expanding directions fail to expand enough for us to resolve the two faces opposite each other in that direction. In the former case, the width of M_i grows in the contracting directions as i increases, eventually resulting in a shadowing distance which grows without bound. In the latter case, we cannot resolve the two opposite faces because the bounding boxes for their images overlap; see Figure 3.8.

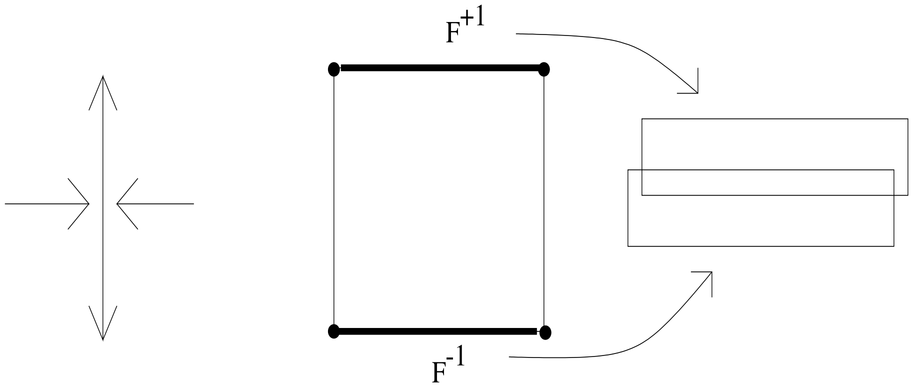

Figure 3.8: How the $2n$ direct face integration method can fail when not enough expansion occurs. Here, the vertical direction is the nominally expanding direction.

The direct $2n$ face integration method was the first method of proving the Inductive Containment Property that we implemented in our code. However, we found that this method had several drawbacks. The most catastrophic problem is that the current implementation of VNODE is not designed to handle parallelepipeds that have some dimensions initially of zero thickness, and thus it is not always capable of providing tight enclosures of the images of faces. This problem is particularly bad if the initially zero width dimension lies along an expanding direction. Then, numerical errors very quickly compound to give an enclosure which is useless for containment purposes (see Figure 3.9). The amount by which a validated ODE enclosure over-estimates the error is called the "excess". It is well-known that most current implementations of validated ODE integration have a large excess. It may be possible to decrease the excess for this particular application (Nedialkov, personal communication), although it may not be worth the effort given that we introduce below more efficient and accurate methods for

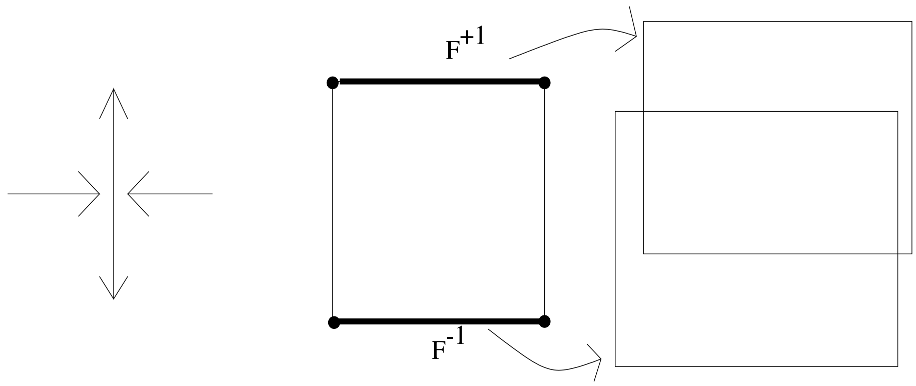

Figure 3.9: How the $2n$ direct face integration method *actually* fails, because the current implementation of VNODE cannot produce tight enclosures of the images of faces. Again, the vertical direction is the nominally expanding direction.

proving ICP.

Finally, we note that the $2n$ direct face integration method is expensive: it requires $2n$ validated ODE integrations per shadow step; we will show below that there exist ways of verifying ICP with fewer validated integrations per shadow step. On the other hand, each of the $2n$ validated integrations is independent of all the others, so they could be performed in parallel.

3.5.2 Direct integration of $n+1$ corners of M_i

If a bound similar to those in Theorem 2.4 (page 17) on the first and second spatial derivatives of φ can be found either analytically or computationally, then we can bound the curvature of $\varphi(F)$ for a face F of M_i. This allows us to compute an enclosure of $\varphi(F)$ by integrating only the corners of F. Furthermore, we can implicitly get the positions of all 2^n corners of M_i using the positions of only $n+1$ corners by choosing one corner c as an origin and finding the lengths and directions of the n distinct edges emanating from c. Given the bound on the first and second spatial derivatives of φ, this allows us to compute enclosures on all $2n$ faces of $\varphi(M_i)$ using only $n+1$ validated integrations of the appropriate corners of M_i (Nedialkov, Jackson, and Corliss 1999). From this, it is easy to build an M_{i+1} that satisfies ICP, as shown in Figure 3.10. In particular, from the enclosures of the corners and a bound on the derivatives of φ, we can build parallelepipeds Q_{i+1} and R_{i+1} such that $R_{i+1} \subset \varphi(M_i) \subset Q_{i+1}$. Then, the faces of M_{i+1} can be chosen to be strictly outside those of Q_{i+1} in the contracting directions, and strictly inside those of R_{i+1} in the expanding directions.

This method has the disadvantage that the second spatial derivatives need to be computed, which can be expensive if done computationally, and tedious if done analytically. On the other hand, if the analytical bounds can be computed *a priori*, they may be computationally cheap. The $n+1$ validated integrations can also be performed cheaply because they are *point* integrations. That is, $[\mathbf{r}_{i-1}]$ in equation (1.8) is zero and so, if the validated integrator can integrate from t_i to t_{i+1} in one step, $[S_{i-1}]$, which is the expensive part of a validated integration, need not be computed. The $n+1$ point integrations are also independent, so they can be done in parallel. Since point integrations are cheap and produce very tight enclosures, if one can also produce *a priori* tight bounds on the first and second spatial derivatives of φ, and these bounds are small, then one can efficiently produce very tight enclosures of the faces of $\varphi(M_i)$.

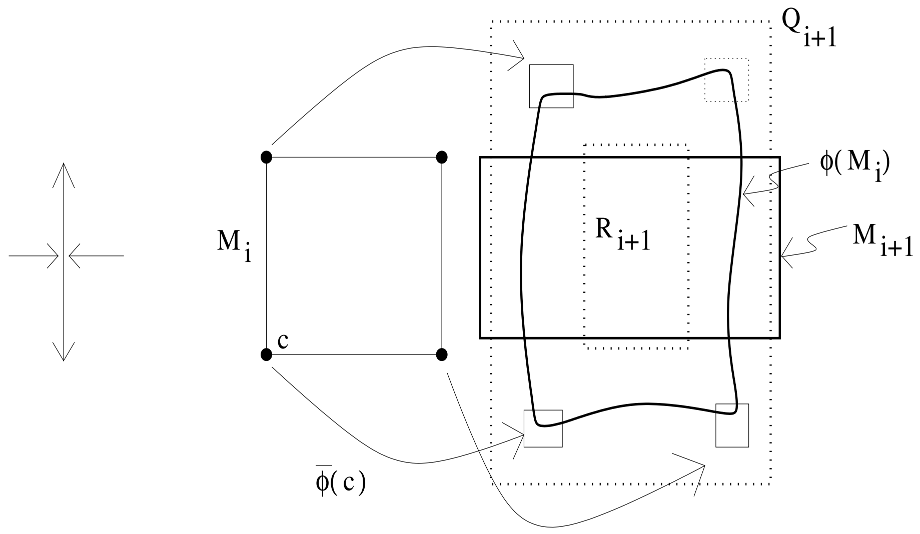

Figure 3.10: The direct $n+1$ corner integration method of proving ICP. In the depicted 2-dimensional case, 3 corners (c and the two adjacent to it) are integrated using VNODE; the top right corner is not integrated explicitly. Q_{i+1} (outer dotted box) is the outer bounding box of $\varphi(M_i)$, and R_{i+1} (inner dotted box) is the inner bounding box. As before, M_{i+1} is aligned so that its expanding faces are strictly inside R_{i+1}, while its contracting sides are strictly outside Q_{i+1}. The sizes of the enclosures of the corners are highly exaggerated; in practice, they can have a diameter approaching the machine precision.

This implies that this method is both efficient for small n and likely to produce the tightest containment boxes of all the methods discussed in this thesis, as long as φ is not too far from being linear, and we can compute *a priori* bounds on the derivatives. This can usually be arranged by choosing a sufficiently small timestep, although too small a timestep can lead to other complications. We have not pursued this idea beyond this discussion.

Finally, this method also has the advantage that, if one does not require rigor, then the $n+1$ integrations can be done with a non-validated integrator, and a local error estimate can be used to build Q_{i+1} and R_{i+1}. This method was actually employed by the author during a very early prototyping phase of code development, using LSODE (Hindmarsh 1980) as the non-rigorous integrator. The results are beyond the scope of this thesis, but were sufficiently encouraging to demonstrate that containment could work as a method of finding high-dimensional shadows.

3.5.3 Forward-backward iterative method

The above two methods can be used on arbitrary n-dimensional systems, but require order n validated ODE integrations for each containment step. Since validated ODE integrations are very expensive, we would like to find a way to ensure that the Inductive Containment Property holds using fewer integrations, especially for large n. The following method demonstrates that it is possible to verify ICP using an iterative method that we have found empirically to require about 3–4 validated integrations per step on average, independent of n. This method rigorously verifies ICP in the cases for which we have proven the Inductive Containment Theorem. We are not sure if it verifies ICP in the general case, and would need to perform further work to ensure that it does before using it in the general problem. However, considering that this method of verifying ICP holds is cheaper and easier to work with than the previous two methods and that it verifies ICP holds in exactly the cases for which we can prove the Inductive Containment Theorem, we are content to leave further exploration of this method until later.

In this paragraph, we look at the simple two-dimensional case in which one of the directions is expanding, while the other is contracting. First, assume that the only information provided by our validated ODE integration is an outer bound $\bar{\varphi}(M_i)$ on $\varphi(M_i)$. Then, it is *not* possible to verify $(2,1)$-ICP with only one validated integration, because this information can only prove contraction, not expansion. Refer to Figure 3.11. In both Figures 3.11.a and 3.11.b,

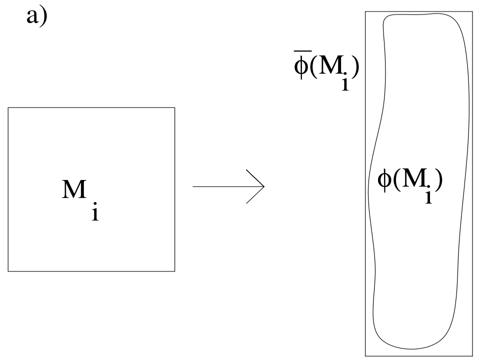

Figure 3.11: Enclosure methods prove contraction, but not expansion.

$\bar{\varphi}(M_i)$ is a valid enclosure of $\varphi(M_i)$. In both Figures, $\bar{\varphi}(M_i)$ can be used to prove that $\varphi(M_i)$ has contracted in the horizontal direction. However, enclosure methods cannot directly prove expansion, as Figure 3.11.b demonstrates: although $\bar{\varphi}(M_i)$ is a valid enclosure of $\varphi(M_i)$, it is not a very good one, because the actual image $\varphi(M_i)$ of M_i has not expanded in any direction. To solve this problem, we perform two validated integrations. Refer to Figure 3.12.a. The

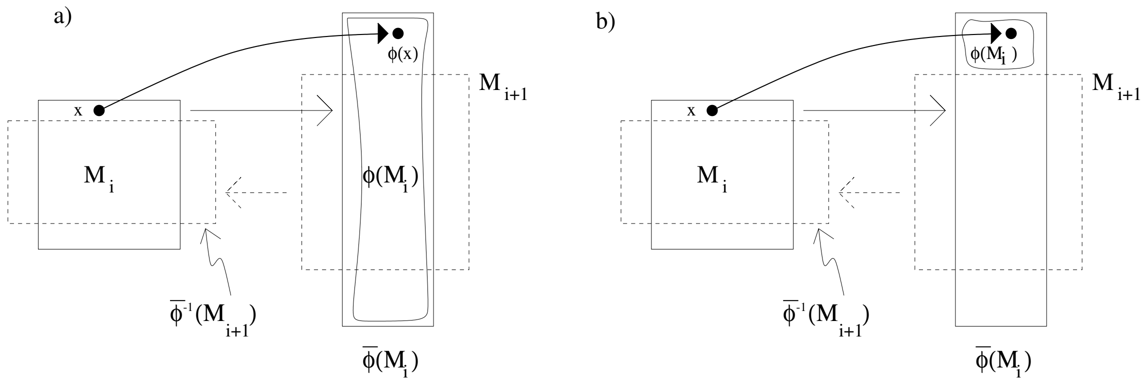

Figure 3.12: (a) The two validated integrations required to prove $(2,1)$-ICP. (b) A potential problem, which is solved by doing a (cheap) point integration of one point on each expanding face, to verify there are points of $\varphi(M_i)$ on both side of M_{i+1}.

first integration (solid rectangles) is a forward integration that provides $\bar{\varphi}(M_i)$, which in turn gives us a bound on the size of $\varphi(M_i)$ in the contracting directions (depicted as the horizontal direction in the Figure). Now, assume we can find an M_{i+1} which satisfies ICP not with $\varphi(M_i)$, but with $\bar{\varphi}(M_i)$. (If we cannot find such an M_{i+1}, then our method fails and we cannot prove the existence of a shadow beyond step i.) A validated integration *backwards* (dashed rectangles) is then performed on M_{i+1}, giving $\bar{\varphi}^{-1}(M_{i+1})$. If $\bar{\varphi}^{-1}(M_{i+1})$ proves that *contraction* has occured in the nominally expanding directions when moving back from M_{i+1} to M_i, then we argue that expansion in forward time has occurred, as follows. Choose any $\mathbf{x} \in M_i - \bar{\varphi}^{-1}(M_{i+1})$. Since

$\mathbf{x} \notin \bar{\varphi}^{-1}(M_{i+1}) \supset \varphi^{-1}(M_{i+1})$, this implies $\varphi(\mathbf{x}) \in \varphi(M_i) - M_{i+1}$. Since $F_i^{\pm 1} \subset M_i - \bar{\varphi}^{-1}(M_{i+1})$, this tells us that $\varphi(F_i^{\pm 1}) \cap M_{i+1} = \emptyset$. This is insufficient to prove ICP(1), as illustrated in Figure 3.12.b: perhaps $\bar{\varphi}(M_i)$ is a loose enclosure of $\varphi(M_i)$, and all of $\varphi(M_i)$ is actually on one side of M_{i+1}. To verify that this is not the case, we pick one point on each of F_i^{+1} and F_i^{-1} and perform a validated point integration (which can be done cheaply, as described above) of each to verify that they land on opposite sides of M_{i+1}.[5] Since there is exactly one expanding direction, M_{i+1} cuts $\bar{\varphi}(M_i)$ into two disjoint sets, and a simple continuity argument shows that the two faces in their entirety land on opposite sides of M_{i+1}, thus verifying ICP(1). A similar argument in reverse time shows that the chosen M_{i+1} also verifies ICP(2ii).

The argument of the previous paragraph clearly applies just as well in n dimensions when there is exactly one expanding direction, for the same reasons that Theorem 3.1 is easily transformed into Theorem 3.4. To prove that it also works when there is exactly one contracting direction, note that there is a precise symmetry between the two cases (one expanding *vs.* one contracting): if we simultaneously reverse the order of $\{M_i\}_{i=0}^N$ giving $L_i = M_{N-i}$ and let $\psi = \varphi^{-1}$, then the above argument applies to the sequence $\{L_i\}_{i=0}^N$ using ψ as the homeomorphism. Thus, by symmetry, this method is also rigorous in the case that there is exactly one contracting direction.

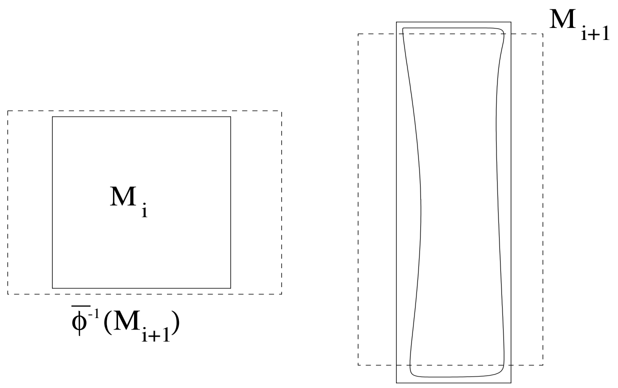

Figure 3.13: Shortcomings of the two-integration method: sometimes it can not prove expansion even if the M_{i+1} is valid.

Figure 3.13 demonstrates that it is possible to choose an M_{i+1} that satisfies ICP, but for which it we cannot *verify* ICP holds. This occurs when M_{i+1} is chosen to be "almost as large" as $\bar{\varphi}(M_i)$ in the expanding directions; then, the excess when computing $\bar{\varphi}^{-1}(M_{i+1})$ swamps the contraction that occurs when integrating the expanding direction backwards in time. We solve this problem by iteratively shrinking M_{i+1} in the nominally expanding directions until $\bar{\varphi}^{-1}(M_{i+1})$ fits inside M_i in those directions. If we shrink M_{i+1} to size zero in the expanding direction without being able to integrate it backwards to fit inside M_i, then the method fails, and we cannot prove the existence of a shadow past step i. We have found empirically that, when the algorithm is succeeding, no more than 2 to 3 backward integrations are usually required, independent of n. The number of backwards integrations is occasionally significantly larger, when the system encounters areas of non-hyperbolicity.

If the system were hyperbolic, then the nominally expanding directions would always expand, and the nominally contracting directions would always contract, on average. However, in

[5]We have found empirically that this problem must be very rare, because it has not happened even once during our experiments. We suspect that it may be possible to prove ICP without this extra point integration, but we have not devoted much thought towards how to avoid it.

systems that are only pseudo-hyperbolic, the nominally expanding directions may expand most of the time, but not always; and similarly for the contracting directions. One of the reasons our shadowing method can fail is if a nominally expanding direction contracts too much or for too long a time (Figure 3.14). Then, the expanding dimensions of M_i can become so small that no

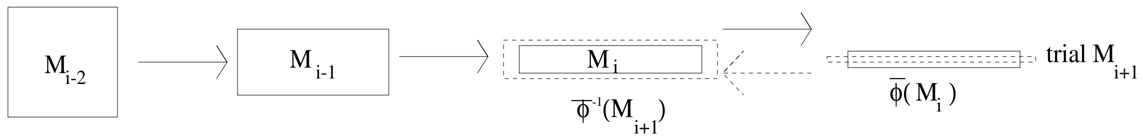

Figure 3.14: Example of the nominally expanding direction contracting too much for our integrator to prove contraction in the backwards direction.

backwards integration from M_{i+1} can fit inside M_i in the nominally expanding directions.

We note that this method is *not* parallelizable across dimensions in the fashion that the previous two methods are, so that a parallel implementation may be faster using one of the previous methods, if it can overcome the other shortcomings mentioned for those methods.

3.5.4 Single integration method

There may be more efficient ways to implement the verification of ICP. For example, it may be possible to prove both expansion and contraction using a single forward integration if we take advantage of knowledge of the second term in equation (1.8), which essentially tells us how much uncertainty is introduced to the boundary of the image of $[\mathbf{r}_i]$ as a result of new error introduced on this step, both inwards and outwards. This would allow us to build both the outer bound Q_{i+1} on $\varphi(M_i)$ and the inner bound R_{i+1} as depicted in Figure 3.10 using only one validated integration. This would be a tremendous improvement over the current methods which all require several validated integrations per shadow step. Alternatively, there are other implementations of validated ODE integration that we could use that are more expensive, but provide tighter bounds on the solution. We have not yet explored any of these options.

3.6 Rescaling time

3.6.1 Informal description

Containment as presented thus far has put no restrictions on φ other than that it is a homeomorphism. As has also been mentioned, all of our theorems and proofs have been based on a single application of φ, and there is no explicit connection between the φ used at one step, and the one used on the next. Thus, everything said thus far is also applicable if we allow φ to change between steps. In particular, at each step we could use φ_{h_i} as defined in equation (1.4) with $h = h_i$ being the length of the ODE integration timestep taken at step i. The resulting method for shadowing numerical ODE integrations has been dubbed the *Map Method* by Coomes, Koçak, and Palmer (1994b, 1995a, 1995b). As described in section 2.2.5, however, ODE integrations suffer from errors in time. For systems in which the $\mathbf{y}'$ direction lacks even pseudo-hyperbolicity, errors in time (which manifest themselves in phase space as errors in the $\mathbf{y}'$ direction) can lead to short shadowing times that can be dramatically increased if time is *rescaled*. In this section, we describe how the rescaling of time can be applied to containment.

Our idea for rescaling time in containment was inspired in part by the rescaling of time of Coomes, Koçak, and Palmer (1994b, 1995a) as depicted in Figure 2.4 (although our proofs are profoundly different), and partly by the idea of the *Poincaré section*, also known as a *Poincaré map* or *return map*. There are several variations on this idea, but the one that concerns us is the following. Assume that the solution to an ODE is "almost periodic", in the sense that the solution passes through some given plane $\mathcal{H}$ approximately every T time units, where $\mathcal{H}$ is approximately perpendicular to the orbit at the point the orbit crosses the plane. The Poincaré map generates the sequence of points where the orbit intersects $\mathcal{H}$. To accomplish the general rescaling of time, we modify this idea to remove the almost-periodic requirement of the orbit, and simply place a plane $\mathcal{H}_i$ in the vicinity of the solution at time t_i, placed so that $\mathcal{H}_i$ is approximately perpendicular to $\mathbf{y}'(t_i)$.

To facilitate containment, we must extend the idea of the Poincaré section to encompass a small ensemble of solutions. To that effect, we wish to take a set $M_{i-1} \subset \mathcal{H}_{i-1}$, where the diameter of M_{i-1} is small, and place a plane $\mathcal{H}_i$ in the vicinity of $\varphi_{h_{i-1}}(M_{i-1})$. Then we define the Poincaré section of the set $\varphi_{h_{i-1}}(M_{i-1})$ pointwise as follows. Let Δh_{i-1} bound the time interval over which the ensemble $\varphi_{h_{i-1}}(M_{i-1})$ crosses $\mathcal{H}_i$; *i.e.*,

$$\forall \mathbf{x} \in M_{i-1} \exists h \in [h_{i-1} - \Delta h_{i-1}, h_{i-1} + \Delta h_{i-1}] \text{ s.t. } \varphi_h(\mathbf{x}) \in \mathcal{H}_i,$$

where we assume that for each $\mathbf{x}$, the h chosen is unique. That is, we take the point-by-point Poincaré section of the points in M_{i-1} with respect to the plane $\mathcal{H}_i$. We call this a *splash* operation, because we imagine that the points in M_{i-1}, evolving via φ_h for $h \in [h_{i-1} - \Delta h_{i-1}, h_{i-1} + \Delta h_{i-1}]$, "splash" through $\mathcal{H}_i$ approximately simultaneously, and we assume that each trajectory intersects $\mathcal{H}_i$ precisely once during that interval. See Figure 3.15.

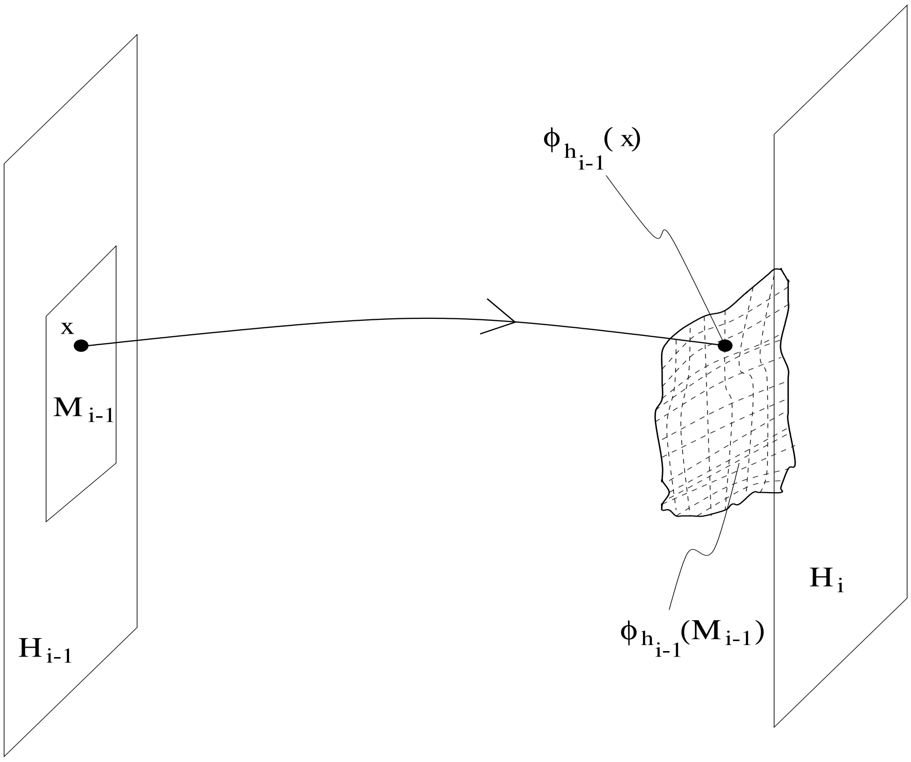

Figure 3.15: The "splash" operation depicted for a two-dimensional ensemble evolving in a three-dimensional configuration space. M_{i-1} is embedded in the plane $\mathcal{H}_{i-1}$, and evolves through one timestep to $\varphi_{h_{i-1}}(M_{i-1})$. As depicted, the ensemble is about to splash through $\mathcal{H}_i$.

Our intent in this endeavor is to build our parallelepipeds M_i inside $\mathcal{H}_i$, and then show that the point-by-point Poincaré section at $\mathcal{H}_i$, *i.e.*, the splash operation, is a homeomorphism. We can then directly apply the previously proved containment theorems to the $n-1$-dimensional M_i's which are each contained in the $n-1$-dimensional hyperplane $\mathcal{H}_i$, for an ODE system of n equations.

We note that since rescaling time via the splash operation effectively deletes one dimension from the problem, and our map containment theorems are rigorous in three dimensions, this means that the methods presented in this thesis are capable of rigorously shadowing ODE solutions of up to four dimensions, as long as a rescaling of time is applied.

3.6.2 Theorem: splash is a homeomorphism

Refer to Figure 3.16. Let Q_i be a parallelepiped. Let $F_i^{\pm 1}$ be two opposing faces of Q_i that are approximately normal to $\mathbf{y}'$ inside Q_{i+1}, and let $\mathbf{v}_i$ be the normal vector to these two faces, with $\mathbf{v}_i$ pointing from F_i^{-1} to F_i^{+1}. That is, $\mathbf{v}_i$ is approximately parallel to $\mathbf{y}'$ inside Q_{i+1}. Let D be the distance between F_i^{-1} and F_i^{+1} along $\mathbf{v}_i$. Let Z_i be the closed infinite slab between the two hyperplanes containing F_i^{-1} and F_i^{+1}, and let the infinite planes be H_i^{-1} and H_i^{+1}. Let B_i be a parallelepiped with faces parallel to Q_i satisfying $Q_i \subset B_i \subset Z_i$, with two of the faces of B_i contained in $H_i^{\pm 1}$. Let $\{\mathbf{f}(\mathbf{x}) \cdot \mathbf{v}_i \mid \mathbf{x} \in B_i\} \subset [v_0, v_1]$, and assume $0 < v_0 \leq v_1$.

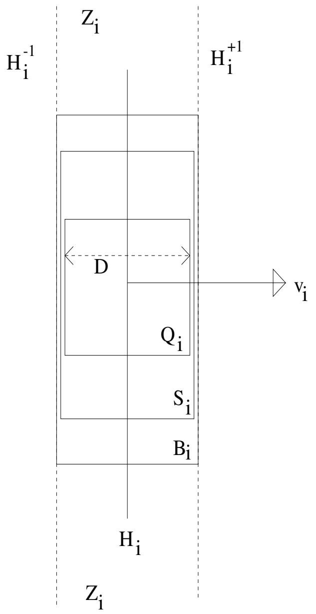

Figure 3.16: The objects used in Lemmas 3.9–3.12. Note that the left and right sides of Q_i, S_i, B_i, and Z_i are all in the planes H_i^{-1}, H_i^{+1}, respectively; they have been drawn distinct for illustrative purposes only.

Lemma 3.9. *If a trajectory remains in B_i while it is in Z_i, then it remains in Z_i for at least time $\underline{\varepsilon}_i^t \equiv D/v_1$ and at most $\bar{\varepsilon}_i^t \equiv D/v_0$.*

Proof. Let $\mathbf{y}(t)$ be a trajectory that remains in B_i while it is in Z_i. Let $z(t) = \mathbf{y}(t) \cdot \mathbf{v}_i$. Since $0 < v_0 \leq z'(t) \leq v_1$, and the width of B_i in the $\mathbf{v}_i$ direction is D, the maximum time to cross B_i is D/v_0, while the minimum time to cross is D/v_1. □

Let $\bar{\mathbf{f}}(B_i)$ be an enclosure of $\{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in B_i\}$. Let S_i be a parallelepiped enclosure of $\left\{Z_i \cap (Q_i + h\bar{\mathbf{f}}(B_i)) \mid h \in [-\bar{\varepsilon}_i^t, \bar{\varepsilon}_i^t]\right\}$ and assume $S_i \subseteq B_i$.

Remark: S_i is intended to enclose how far a trajectory can drift from Q_i along the direction approximately perpendicular to $\mathbf{y}'$ as it travels across Z_i. This is required because a point in Q_i may not remain in Q_i when it is "splashed" onto $\mathcal{H}_i$. The following lemma formalizes this statement.

Lemma 3.10. *Any trajectory intersecting Q_i remains in S_i while in Z_i, and thus remains in B_i as well.*

Proof. Since $S_i \subseteq B_i$, $\bar{\mathbf{f}}(B_i)$ bounds $\mathbf{y}' \equiv \mathbf{f}$ inside S_i. Since $\bar{\varepsilon}_i^t$ is the maximum time a trajectory remains in Z_i, and since $S_i \subset Z_i$, $\{h\bar{\mathbf{f}}(B_i) \mid h \in [-\bar{\varepsilon}_i^t, \bar{\varepsilon}_i^t]\}$ encloses the maximum possible distance from Q_i that a trajectory can travel in time $|\bar{\varepsilon}_i^t|$ while it remains in B_i. Thus, since $Q_i \subset S_i \subseteq B_i$, $\{Q_i + h\bar{\mathbf{f}}(B_i) \mid h \in [-\bar{\varepsilon}_i^t, \bar{\varepsilon}_i^t]\}$ encloses the position of any trajectory $\mathbf{y}(t)$ that is within time $\bar{\varepsilon}_i^t$ of intersecting Q_i, unless $\mathbf{y}(t)$ leaves Z_i during that time. Intersecting with Z_i completes the proof. $\square$

Let $\mathcal{H}_i$ be any plane perpendicular to $\mathbf{v}_i$ which intersects Q_i.

Lemma 3.11. *Every trajectory intersecting Q_i intersects $\mathcal{H}_i$ at precisely one point while it crosses Z_i.*

Proof. Let $\mathbf{y}(t)$ be a trajectory that intersects Q_i. By Lemma 3.10, $\mathbf{y}(t)$ remains in $S_i \subseteq B_i$ while it crosses Z_i. Let $z(t) = \mathbf{y}(t) \cdot \mathbf{v}_i$. Let the z co-ordinates of $H_i^{-1}, \mathcal{H}_i, H_i^{+1}$ be z_{-1}, z_0, z_{+1}, respectively. While the trajectory remains in $S_i \subseteq B_i$, $z'(t) \geq v_0 > 0$, and, since $z(t)$ is continuous, it increases monotonically while $\mathbf{y}(t)$ remains in S_i, taking on every value between z_{-1} and z_{+1} precisely once, by the Intermediate Value Theorem. In particular, it takes on the value z_0 precisely once, and thus crosses $\mathcal{H}_i$ precisely once. $\square$

Assume Q_i is an enclosure of $\varphi_{h_{i-1}}(M_{i-1})$. Lemma 3.11 implies that every trajectory through Q_i crosses $\mathcal{H}_i$ precisely once while in S_i. For a point $\mathbf{x} \in M_{i-1}$, let $\varphi_{i-1}(\mathbf{x})$ be this unique point in $\mathcal{H}_i$. Let $\bar{M}_i = S_i \cap \mathcal{H}_i$. Clearly, $\bar{M}_i$ is an enclosure of $\varphi_{i-1}(M_{i-1})$.

To show that φ_{i-1} applied to M_{i-1} is a homeomorphism, we need to show it is continuous and one-to-one. We will prove it is continuous below, and by Lemma 3.11, it is at worst many-to-one.

Let $\varepsilon^t > 0$ be given.

Assumption 1: Assume $\bar{\varepsilon}_i^t < \varepsilon^t$ and $\nexists$ distinct $\mathbf{x}, \mathbf{y} \in M_{i-1}$ s.t. $\mathbf{y} = \varphi_t(\mathbf{x})$ for $|t| < \varepsilon^t$.

Each of the *Assumptions* introduced in this section are assumed to hold throughout the remainder of section, once they are introduced.

Lemma 3.12. *φ_{i-1} applied to M_{i-1} is one-to-many.*

Proof. Assume to the contrary that there exist distinct $\mathbf{x}, \mathbf{y} \in M_{i-1}$ s.t. $\varphi_{i-1}(\mathbf{x}) = \varphi_{i-1}(\mathbf{y}) = \mathbf{z} \in \bar{M}_i$. Since $\varphi_{h_{i-1}}(\mathbf{x}), \varphi_{h_{i-1}}(\mathbf{y})$ both splash to $\mathbf{z}$, they are on the same trajectory, and since they are both in Q_i, the time-shift between them is $\leq \bar{\varepsilon}_i^t$. Thus, $\exists t_1, t_2$ s.t. $\varphi_{t_1}(\mathbf{x}) = \mathbf{z} = \varphi_{t_2}(\mathbf{y})$ with $|t_1 - t_2| \leq \bar{\varepsilon}_i^t$. Then $\mathbf{x} = \varphi_{t_2 - t_1}(\mathbf{y})$, contradicting Assumption 1. $\square$

Theorem 3.13. *φ_{i-1} applied to M_{i-1} is one-to-one.*

Proof. Lemma 3.11 proves that $\varphi_{i-1}(M_{i-1})$ is many-to-one, and Lemma 3.12 proves it is one-to-many. Thus, it is actually one-to-one. $\square$

Assumption 2: $\varphi_t(\mathbf{x})$ exists and is continuous in both t and $\mathbf{x}$ for all $\mathbf{x} \in M_{i-1}$ and t s.t. $\varphi_t(\mathbf{x}) \in B_i$. Note that this is true as long as $\mathbf{f}$ is Lipschitz continuous (Stuart and Humphries 1996, Theorem 2.1.12).

Recall that a fundamental tenet of the definition of continuity is that a function $\mathbf{f}$ is continuous at a point $\mathbf{x}$ only if $\mathbf{f}(\mathbf{y})$ exists in an open neighborhood $\mathcal{N}(\mathbf{x})$ around $\mathbf{x}$, $\lim_{\mathbf{y}\to\mathbf{x}} \mathbf{f}(\mathbf{y}) = \mathbf{f}(\mathbf{x})$, and this limit is the same regardless of the path taken by $\mathbf{y}$ as it approaches $\mathbf{x}$.

In the following, we assume that, in order for two things to be equal, they must both exist.

Lemma 3.14. *$\varphi_{i-1}(\mathbf{x})$ is continuous for all $\mathbf{x} \in M_{i-1}$.*

Proof. Assume to the contrary that φ_{i-1} is not continuous at $\mathbf{x}_0 \in M_{i-1}$. That is,

$$\begin{aligned} &\lim_{\mathbf{y}\to\mathbf{x}_0} \varphi_{i-1}(\mathbf{y}) \neq \varphi_{i-1}(\mathbf{x}_0) \\ \Longrightarrow \quad &\lim_{(h,\mathbf{y})\to(t_0,\mathbf{x}_0)} \varphi_h(\mathbf{y}) \neq \varphi_{t_0}(\mathbf{x}_0), \end{aligned}$$

where t_0 is chosen to put $\varphi_{t_0}(\mathbf{x}_0)$ in $\mathcal{H}_i$, and for each $\mathbf{y}$ the h is chosen to put $\varphi_h(\mathbf{y})$ in $\mathcal{H}_i$. (h is unique for each $\mathbf{y}$, by Theorem 3.13.) This means that the limit as $\mathbf{y}$ approaches $\mathbf{x}_0$ along a path remaining in $\mathcal{H}_i$ is not equal to $\varphi_{t_0}(\mathbf{x}_0)$, *i.e.*, either $\varphi_h(\mathbf{y})$ is discontinuous at $(t_0, \mathbf{x}_0)$, or either $\lim_{(h,\mathbf{y})\to(t_0,\mathbf{x}_0)} \varphi_h(\mathbf{y})$ or $\varphi_{t_0}(\mathbf{x}_0)$ does not exist. This contradicts Assumption 2, and so $\varphi_{i-1}(\mathbf{x})$ is continuous at $\mathbf{x}_0$, and is thus continuous for all $\mathbf{x} \in M_{i-1}$. □

Let W_i be an infinite slab with width $E > D$ in the $\mathbf{v}_i$ direction, parallel to Z_i such that $Z_i \subset W_i$. Let C_i be a parallelepiped with sides parallel to Q_i, also with a width of E in the $\mathbf{v}_i$ direction, satisfying $M_i \subset C_i \subset W_i$, where M_i is built inside $\mathcal{H}_i$ to satisfy ICP with M_{i-1} under φ_{i-1}. Let $\{\mathbf{f}(\mathbf{x}) \cdot \mathbf{v}_i \mid \mathbf{x} \in C_i\} \subset [u_0, u_1]$, and assume $0 < u_0 \leq u_1$. Let $\bar{\mathbf{f}}(C_i)$ be an enclosure of $\{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in C_i\}$. Let T_i be a parallelepiped enclosure of $\{W_i \cap (M_i + h\bar{\mathbf{f}}(C_i)) \mid h \in [-\varepsilon^t, \varepsilon^t]\}$, and assume $T_i \subseteq C_i$.

Assumption 3: Assume $E/u_1 > \varepsilon^t$, *i.e.*, the minimum crossing time of C_i is greater than ε^t.

Lemma 3.15. *$\nexists$ distinct $\mathbf{x}, \mathbf{y} \in M_i$ s.t. $\mathbf{y} = \varphi_t(\mathbf{x})$ for $|t| < \varepsilon^t$.*

Proof. Substituting M_i for Q_i, W_i for Z_i, T_i for S_i, and C_i for B_i in Lemmas 3.9–3.11, we see that

1) If a trajectory remains in C_i while it is in W_i, then it remains in W_i for at least time E/u_1 and at most E/u_0.

2) Any trajectory intersecting M_i remains in T_i while it is in W_i, and thus remains in C_i.

3) Every trajectory intersecting M_i intersects $\mathcal{H}_i$ at precisely one point while it remains in W_i, where $\mathcal{H}_i \subset W_i$ and $\mathcal{H}_i$ is parallel to the planes enclosing W_i.

Thus, by point (3), to intersect $\mathcal{H}_i$ more than once inside M_i, a trajectory must, at least, first traverse the distance from $\mathcal{H}_i$ to ∂C_i, exit and then re-enter C_i, and traverse the distance from ∂C_i back to the same point on $\mathcal{H}_i$. By point (1), it takes time at least E/u_1 to do so. By Assumption 3, $E/u_1 > \varepsilon^t$. Thus, no trajectory can intersect M_i, exit T_i, and then re-enter T_i to again intersect the same point of M_i in time less than ε^t. □

Remark: It is Lemma 3.15 at step $i-1$ that gives us the second part of Assumption 1 at step i.

Remark: The base case of the induction is produced by substituting M_0 for M_i in Lemma 3.15, after building suitable W_0, C_0, and T_0.

3.6.3 Algorithmic details

Algorithmic verification of the requirements for the above theorems and lemmas are fairly straightforward: Q_i is simply the enclosure of $\varphi_{h_{i-1}}(M_{i-1})$ given to us by VNODE; the size of B_i is computed heuristically in an effort to ensure that $S_i \subseteq B_i$, and if our first guess is incorrect we simply increase its size until $S_i \subseteq B_i$, or fail if increasing the size of B_i results in $0 \in \{\mathbf{f}(\mathbf{x}) \cdot \mathbf{v}_i \mid \mathbf{x} \in B_i\}$; ε^t, which is an upper bound on the time error introduced at each step by the rescaling of time, must currently be pre-chosen by trial and error, although this author believes that good, simple heuristics for choosing it probably exist. The sole complication is to maintain the property that Q_i has a pair of faces approximately normal to $\mathbf{y}'$ inside Q_i. Recall from section 1.3.2 that VNODE maintains a rotation matrix A_i which represents the orientation of the parallelepiped Q_i. Let the columns of A_i be $\mathbf{a}_i^j, j = 1, \ldots, n$. We simply assign $\mathbf{a}_i^1$ to be parallel to our best estimate of $\mathbf{y}'(t_i)$. VNODE then ensures that $\mathbf{a}_{i+1}^1$ evolves via the variational equation to be approximately parallel to $\mathbf{y}'(t_{i+1})$. To account for the slow buildup of error that would allow $\mathbf{a}_i^1$ to drift away from $\mathbf{y}'(t_i)$, we reset $\mathbf{a}_i^1$ to be parallel to the computed $\mathbf{y}'(t_i)$ at each timestep. This corresponds to rotating Q_i about its centre by a small angle θ, computed by solving

$$\cos(\theta) = \frac{\mathbf{a}_i^1 \cdot \mathbf{y}'(t_i)}{\|\mathbf{a}_i^1\| \, \|\mathbf{y}'(t_i)\|},$$

where $\mathbf{a}_i^1$ is the vector computed via evolution of the ODE from the previous timestep, and $\mathbf{y}'(t_i)$ is the value of $\mathbf{y}'$ computed directly from the right hand side of the ODE at the current timestep. The largest distance a point in Q_i will move as a result of this rotation is $r\theta$, where r is the distance of the furthest corner in Q_i from its centre. Thus, after rotating Q_i by θ, we increase its size by $r\theta$ in all directions, thus ensuring that it still encloses $\varphi_{h_{i-1}}(M_{i-1})$.

A simple variable stepsize algorithm was used: whenever containment of a particular step succeeds, we increase the stepsize by a small factor; whenever it fails, we decrease the stepsize by a factor of 2. We do not explicitly fail due to small stepsize, because too small a stepsize results in failures in other parts of the method, for example as depicted in Figure 3.14.

Finally, we note that the rescaling of time theorems presented in this section are independent of the containment results of previous sections, and thus do not need to be modified if and when our proofs are extended to cover the general case (c.f. §§3.3, 3.4.2).

Chapter 4

Results and discussion

In this Chapter, we will present results of our containment method for ODEs, compare our results to those of others, discuss some of the interesting implementation details of our method, and comment on observations of the behaviour of our method including how it fails and some improvements that were discovered by accident.

4.1 Quantitative comparisons with other methods

4.1.1 The Lorenz system of equations

The Lorenz equations (Lorenz 1963),

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \sigma(y-x) \\ \rho x - y - xz \\ xy - \beta z \end{pmatrix}, \tag{4.1}$$

define a dissipative dynamical system (*i.e.,* energy is not conserved) which was originally constructed to be a very simplified weather model. It can be shown (Coomes, Koçak, and Palmer 1995a) that under the Lorenz equations, the set

$$U = \{(x,y,z) : \rho x^2 + \sigma y^2 + \sigma(z-2\rho)^2 \leq \sigma\rho^2\beta^2/(\beta-1)\}$$

is *forward invariant*, *i.e.,* any solution that is in U at time t_0 remains in U for all time $t \geq t_0$. We, and the authors we compare against in this thesis, solve the Lorenz equations using the classical parameter values $\sigma = 10, \rho = 28, \beta = 8/3$ (Lorenz 1963). It is easy to show that for these parameter values, the cube $[0,15]^3$ lies in U, and so for our experiments we chose initial conditions randomly inside this cube. A set of initial conditions in this cube will invariably produce a solution whose three-dimensional shape has been dubbed the "Lorenz butterfly" (Figure 4.1). Schematically, the Lorenz butterfly consists of two two-dimensional disks in three-space with a "bridge" between them. The two disks together are termed a "chaotic attractor", because solutions tend to remain in the disks, but jump chaotically from one to the other and back again. Solutions lack pseudo-hyperbolicity in the direction of the flow (Van Vleck 1995; Coomes, Koçak, and Palmer 1994b, 1995a), and so a rescaling of time is required to shadow them effectively. As should be clear from Figure 4.1 and the above description, in addition to the $\mathbf{y}'$ direction, at any given point a solution has one contracting direction, which is perpendicular to the disk currently housing the solution, and one expanding direction, directed radially from the centre of the disk. Provided a rescaling of time is employed, solutions to the

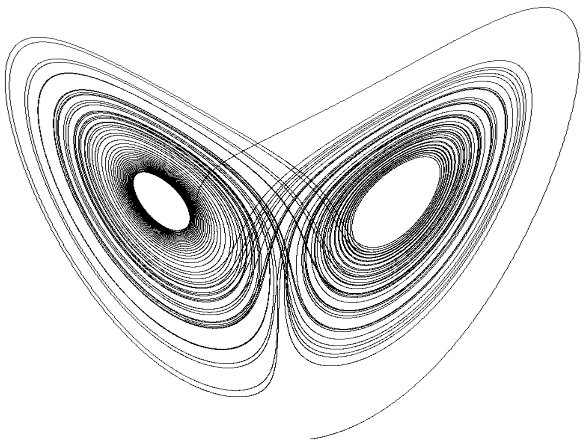

Figure 4.1: THE "LORENZ BUTTERFLY".

Lorenz equations display remarkable pseudo-hyperbolicity for extremely long periods of time. Thus, this system is a prime first candidate for testing shadowing methods.

We will compare our results to the only other published results on shadowing the Lorenz equations using a rescaling of time: Van Vleck (1995), whose results could be made rigorous but currently are not; and Coomes, Koçak, and Palmer, (1994b, 1995a), whose results are completely rigorous.

First, with *no* rescaling of time (the "map method"), Van Vleck gives two examples of shadows with a local error[1] of about 10^{-5} lasting 1.04 and 1.38 time units; Coomes *et al.* have six examples with local error of about 10^{-13} lasting 9.7, 9.8, 9.9, 9.9, 86, and 126 time units. For this thesis, we have simulated hundreds of shadows with various local errors. We have found that with local errors of about 10^{-5}, containment finds shadows that last between 1 and 30 time units, with a median and mean of about 20. With local errors of 10^{-13}, we find shadows lasting between 10 and 1000 time units, again with a mean and median about halfway through that range. Thus, it appears that, without a rescaling of time, the containment method is capable of finding shadows that are about an order of magnitude longer than other methods.

With a rescaling of time, Van Vleck gives many examples of shadows (with a local error of about 10^{-6}) ranging from 10^2 to 10^4 time units. Coomes *et al.* (with a local error of 10^{-13}) give six examples of shadows lasting at least 10^5 time units; they do not attempt finding longer shadows, so in fact their method may be capable of finding shadows longer than 10^5. The corresponding numbers for containment are 10^2 to 10^5 for local errors of 10^{-6}, and 10^2 to almost 10^6 for local errors of 10^{-13}. The results are summarized in Table 4.1.[2] It is clear that containment is at least as powerful as the other methods. It is worth noting that our results for local errors of 10^{-13} were produced using only a 17th order Taylor series, whereas Coomes *et al.* used a Taylor series of 31st order.

Figure 4.2 shows two sets of results of shadow lengths, including the rescaling of time. The

[1] All authors other than that of this thesis used constant timesteps, and so the local errors are implicitly *per-unit-step*. The local errors used in the current thesis were normalized to have comparable size per-unit-step, even though variable stepsize methods were used both for the validated ODE integration (Nedialkov 1999), and for choosing the size of shadow steps.

[2] Our attempts to find the longest possible shadows for the latter case have been repeatedly confounded by having either workstation or disk crashes while our simulations were running. The longest shadow we've observed is thus 7.7×10^5, even though, had our machines not crashed, the shadows may have been longer.

Author	local error	global error	Map Method	Rescaling Time
VV	10^{-6}	10^{-5}	1–2	$10^2 \sim 10^4$
Hayes	10^{-6}	10^{-5}	$10 \sim 50$	$10^3 \sim 10^5$
CKP	10^{-13}	10^{-9}	$10 \sim 100$	$\geq 10^5$
Hayes	10^{-13}	10^{-9}	$10 \sim 1000$	$\geq 7.7{\times}10^5$

Table 4.1: COMPARISON OF SHADOW LENGTHS FOR THE LORENZ SYSTEM.

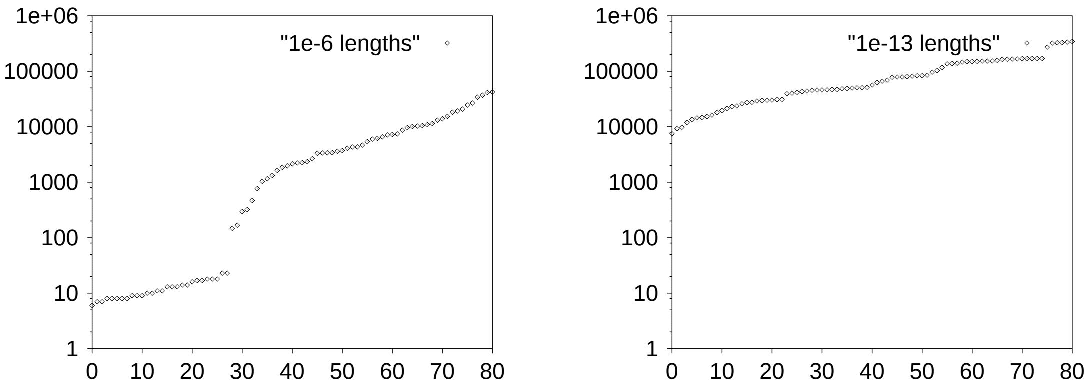

Figure 4.2: Distribution of shadow lengths computed by containment with a rescaling of time. Each Figure shows a sorted list of shadow lengths for 80 simulations of the Lorenz equations. The horizontal axis is simply a label for each shadow; the vertical axis is its length. The magnitude of the noise (*i.e.,* the local error) in the noisy orbits is about 10^{-6} in the left graph, and 10^{-13} in the right.

first is for eighty solutions with local error of approximately 10^{-6}, and the second for eighty solutions with local error of approximately 10^{-13}. The sharp increase in shadow lengths occuring just left of centre in the first Figure is probably due to the fact that, other than choosing $\mathbf{v}_0$ (cf. Figure 3.16 on page 51) to be parallel to $\mathbf{y}'(t_0)$, the directions of the faces of M_0 are currently chosen at random. This means that we sometimes choose nominally expanding and contracting directions that are not sufficiently close to the actual expanding and contracting directions. Thus, many shadows fail early on due to this problem. However, if our nominally chosen directions are (by luck) close enough to the actual ones, then we get over this hump to find much longer shadows. There is probably a more clever way to choose the initial M_0, but we have not yet studied this problem closely. This problem becomes less pronounced as the local error decreases, and is virtually absent in the right figure, which has local error $\delta = 10^{-13}$.

In addition, our shadowing distances (*i.e.,* the maximum distance between the shadow and the numerical trajectory) are comparable to the above authors: for orbits with noise 10^{-6} and 10^{-13}, our method and those of Van Vleck and Coomes, Koçak and Palmer find shadowing distances of approximately 10^{-5} and 10^{-9}, respectively. For containment, these sizes are based on ε^t and the maximum size of M_i over all i, which are at least in part user-controlled. For Van Vleck and Coomes *et al.* , the shadowing distances are computed analytically based upon global bounds of various computed quantities.

4.1.2 Other systems of equations

We have reproduced the shadowing experiments of several other authors, usually getting comparable results, as illustrated in Table 4.2. We discussed results for the Lorenz system in the previous section. In this section, we provide results for three other problems.

System	Auth.	δ	ε	ε_t	L	Comment
Lorenz						
	VV	10^{-6}	10^{-5}		10^4	NR
	Hayes	10^{-6}	10^{-5}	2.5×10^{-5}	10^3–10^5	
	CKP	10^{-13}	10^{-9}		$\geq 10^5$	
	Hayes	10^{-13}	10^{-9}	2.5×10^{-9}	$\geq 7.7 \times 10^5$	
Forced Damped Pendulum						
	SY	10^{-18}	10^{-9}		3×10^4	high machine precision
	Hayes	10^{-15}	10^{-6}	10^{-3}	10^3–3×10^4	
	CVV	10^{-6}	10^{-3}		10^4	NR
	Hayes	10^{-6}	10^{-5}	10^{-3}	10^3	
	CVV	10^{-11}	10^{-8}		10^3	NR
	Hayes	10^{-11}	10^{-8}	10^{-3}	10^3	
Forced van der Pol						periodic attractor
	VV	10^{-5}	10^{-4}		10^4	NR
	Hayes	10^{-5}	10^{-6}	3×10^{-5}	$\geq 10^5$	
Logistic Equation						
	CVV	10^{-7}	5×10^{-6}		9.22	$y_0 = 0.01$, fixed L, NR
	Hayes	10^{-7}	10^{-6}		9.22	
	CVV	10^{-7}	5×10^{-6}		18.46	$y_0 = 10^{-4}$, fixed L, NR
	Hayes	10^{-7}	10^{-6}		18.46	

Table 4.2: Comparison of shadow lengths for four systems. For our results, the lengths shown are typical results after attempting many trials with the given local and global errors; the results of others are taken from their respective publications. Legend: δ = local error; ε = global space error; ε_t = global time error (if none is listed for this author, then we did not rescale time); L = shadow length; CKP = Coomes, Koçak, and Palmer (1994b, 1995a); SY = Sauer and Yorke (1991); CVV = Chow and Van Vleck (1994a); VV = Van Vleck (1995); NR=not rigorous.

Forced damped pendulum

We first compare our results for the forced damped pendulum problem,

$$y'' + ay' + \sin y = b \cos t,$$

to those of Grebogi, Hammel, Yorke, and Sauer (1990), Sauer and Yorke (1991), and Chow and Van Vleck (1994a). These authors use the values $a = 0.2, b = 2.4$ and $a = 1, b = 2.4$, with initial conditions $(y, y') = (0, 0)$, and mention that they get similar results with other values of a, b and initial conditions. We used the above two values of a, b and various random initial conditions in the unit square $[0, 1]^2$. We convert the second order equation to two first-order equations by assigning $y_1 = y, y_2 = y'$, giving

$$y_1' = y_2,$$

$$y_2' = b \cos t - \sin y_1 - ay_2.$$

Grebogi, Hammel, Yorke, and Sauer (1990) and Sauer and Yorke (1991) use an extended machine precision of 10^{-29} to generate a trajectory with local truncation error rigorously bounded by 10^{-18} per step, which allows them to find a shadow of length 3×10^4 and rigorous maximum distance 10^{-9} from their noisy trajectory. In comparison, we use standard IEEE754 floating-point numbers and arithmetic, and obtain a local truncation error of about 10^{-15} at best, so our shadow distances are significantly less stringent at 10^{-6}, and tend to be shorter, although in a few instances we successfully found shadows of length $\sim 3 \times 10^4$. Given that Sauer and Yorke used higher precision, we are not surprised that our shadows tend to be shorter and not as close as theirs. Comparing our results to Chow and Van Vleck (1994a), we see we are capable of rigorously proving the existence of a shadow which is closer, but lasts for a shorter time, than they do; on the other hand, our result is rigorous, whereas theirs is only partially rigorous, because they do not rigorously bound numerical errors before applying their theorem.

The primary problem with shadowing this system appears to be the fact that it is non-autonomous. We currently handle a non-autonomous system by converting it to an autonomous system with one component of our solution, y_1, representing time: $y_1(0) = t_0,\ \ y_1'(t) = 1$. This has several drawbacks: (1) assuming we can solve the linear system $y' = 1$ exactly, the interval representing y_1 then accumulates roundoff error and as time progresses, the error in y_1 grows; (2) this is exacerbated by the minimum absolute error in $y - 1$ increasing as $\varepsilon_{mach} t$, where ε_{mach} is the machine precision; (3) finally, the error in the computation of $\cos(y_1)$ adds to the error. These drawbacks, however, do not seem to adequately explain our poor shadowing results for this system. Perhaps the difficulties would vanish if a native procedure for validated integration of non-autonomous systems were used, or if we used higher precision, as did Sauer and Yorke (1991).

Forced van der Pol

The forced van der Pol equation,

$$x'' + \alpha(x^2 - 1)x' + x = \beta \cos(\omega t),$$

is studied by Van Vleck (1995). He defines the parameters implicitly with $\alpha = k = \sigma = 2/5$, where $k = \beta/(2\alpha)$ and $\sigma = (1 - \omega^2)/\alpha$, and uses the initial conditions $(x, x') = (0, 0)$. We try this initial condition, as well as others chosen randomly in the unit square $[0, 1]^2$, and we convert the second order equation to two first-order equations by assigning $y_1 = x, y_2 = x'$, giving

$$y_1' = y_2,$$
$$y_2' = \beta \cos(\omega t) - (y_1^2 - 1)\alpha y_2 - y_1.$$

This equation has a hyperbolic periodic attractor which all solutions approach asymptotically, and so this system is easy to shadow. With a local truncation error of 10^{-6}, Van Vleck found numerical shadows of length 10^4 and distance 10^{-4}, while we went significantly further, finding rigorous shadows lasting 10^5 and longer with a distance of 10^{-6}. Since solutions asymptotically approach a periodic solution that is hyperbolic, we conjecture that containment could be maintained indefinitely.

Logistic equation

Finally, the logistic equation,

$$y' = y(1-y), \quad y(0) = \zeta,\ 0 < \zeta \ll 1,$$

was studied by Chow and Van Vleck (1994a). In this problem, there is an unstable fixed point at $y = 0$ and a stable fixed point at $y = 1$. Chow and Van Vleck attempt shadowing two solutions, both starting at $y(0) = \zeta$ and integrating until $y(T) \approx 1 - \zeta$. If $\zeta = 10^{-2}$, then $T \approx 9.22$, and if $\zeta = 10^{-4}$, then $T \approx 18.46$. In both cases, we use a local truncation error of $\delta = 10^{-7}$. We find that we easily match their results, noting again that ours are rigorous, while theirs are not. In fact, we find that we can prove the existence of these shadows for $\varepsilon \approx 10\delta$ for δ down to about 10^{-14}.

4.2 Qualitative comparisons with other methods

Although containment is rigorous, it appears to be less robust than non-rigorous methods. For example, in two examples out of three, the non-rigorous results of Chow and Van Vleck (1994a) produced shadows that were about an order of magnitude longer than we could produce using containment. In addition, this author's Master's Thesis (Hayes 1995) demonstrated convincing evidence that the gravitational n-body problem is shadowable, and yet containment could prove the existence of shadows lasting only 1% as long as those (found nonrigorously) in Hayes (1995). Even worse, the VNODE package (Nedialkov 1999) is capable of providing a validated enclosure of an IVP for the n-body problem which is about ten times as long as the containment-produced shadow! Clearly, if an enclosure of an IVP exists, then a shadow exists for the associated point solution for at least as long. Thus, at least for some problems, this author's implementation of containment is incapable of finding shadows even if they exist. This does not necessarily imply that the theorems proved in Chapter 3 are deficient; it probably means that our implementation for verifying that the Inductive Containment Property holds can be improved, for example by reducing the excess of the validated numerical integrator.

Our method requires some *a priori* guesses; for example, the maximum and minimum sizes of the M_i, and the maximum time rescaling ε^t need to be chosen before the algorithm can run. We generally had to choose these numbers by trial and error for each problem; if a certain ε^t did not work, for example, we often found that increasing it or decreasing the maximum size of M_i would allow us to find longer or closer shadows, respectively. Van Vleck's (1995) method also requires some *a priori* guesswork to make a rescaling of time work. Although Coomes, Koçak and Palmer do not discuss their choice of parameters, it is likely that they require significant guesswork to find parameters that satisfy their theorems as well. Finally, *all* shadowing methods currently in the literature appear to require guesswork to discover the number of expanding and contracting dimensions, and to choose a local error δ which is stringent enough to satisfy their respective theorems.

It is also not trivial to see how containment could be parallelized, since each M_i depends on M_{i-1}. Possibly an iterative method that guesses all the $\{M_i\}_{i=0}^N$ and then iteratively refines them in parallel could be constructed; this may also be related to two-point boundary value problems (Ascher, Mattheij, and Russell 1988).

Finally, our method has only been proven to work in three dimensions and the other special cases noted in Chapter 3.

On the other hand, containment appears to have several advantages over other methods.

- First and foremost, the method of proof is simple and easy to understand. Improving our results reduces to the problem of producing the best possible Inductive Containment Property.

- We use an (almost) off-the-shelf validated integrator (Nedialkov 1999) to verify that ICP holds; this integrator is almost as easy-to-use as any standard integrator, and thus getting the code "up and running" on a new problem usually takes only a few minutes. Another advantage of this simplicity is that it requires the user to have no deeper understanding of the system than knowing the defining equations.[3]

- Although the success of containment may depend, of course, upon global properties of the system, the method itself is local. By that we mean that it requires information only from the previous step to extend the length of the shadow. Several other methods require computing, storing, and updating global information such as the extent of non-hyperbolicity (cf. Chow and Palmer's p parameter (1991, 1992), discussed on page 21).

4.3 Implementation issues

In the original paper that described containment, Grebogi, Hammel, Yorke, and Sauer (1990) appear to have used boxes M_i of fixed size, and found that smaller boxes seemed to work better. In contrast, our method dynamically grows and shrinks the M_i as i progresses, simply in an effort to maintain the Inductive Containment Property. In fact, we find it advantageous to allow the expanding dimension of M_i to be fairly large, to allow us to "absorb" possible future non-expansion, in an effort to avoid the situation depicted in Figure 3.14 (page 49). Simultaneously, the contracting dimension can be relatively small, in order to avoid the opposite effect (allowing us to "absorb" non-contraction without the nominal contracting dimensions becoming too large). Practically, we find that our "boxes" can be extremely long and thin: typically, they are of length 10^{-3}–10^{-6} in the expanding dimensions, and as small as 10^{-12}–10^{-14} in the contracting dimensions.

Referring once again to Figure 3.14 on page 49, we note that when containment fails, the "expanding" dimension of M_i has often shrunk to almost the same size as the contracting dimension, and both can be quite small (say, 10^{-12}), whereas when containment is "working", the expanding dimension of M_i can be several orders of magnitude larger. It is interesting to note that this implies that the hardest parts of an orbit to shadow are the places where our

[3] Some may consider this a disadvantage.

bounds on the distance between the noisy and shadow orbits are *smallest*, *i.e.,* where we can prove that they are unusually close together. This appears counter-intuitive, but may be related to the one-dimensional result of Chow and Palmer (1991), where they proved that shadows must maintain a minimum distance from the noisy orbit.

Chapter 5

Future work

There are several directions in which this research can be extended.

First and foremost, the author firmly believes that the general containment theorem (cf. §3.4.2) is true, and that containment can be extended to rigorously prove the existence of periodic shadows. Proving both of these results would add a measure of closure to the current work.

Second, our current implementation of ICP is tied intimately to the C++ implementation of VNODE (Nedialkov 1999). As such, the only pseudo-trajectories we can shadow are the ones produced by VNODE (cf. the $\hat{\mathbf{y}}_i$ in equation 1.7, page 6). In contrast, the refinement code of the author's Master's Thesis (Hayes 1995) could be given *any* noisy trajectory on which to perform refinement. Since there is no explicit dependence of our theorems on the algorithm that produces pseudo-trajectories, extending our code so it can be run on any given pseudo-trajectory would be a good practical improvement.

Software exists that produces so-called "continuous numerical solutions" to ODE problems (see for example Enright 1993). These methods use sophisticated interpolation techniques to allow the user to request the solution at any floating-point time t in the interval of integration. It should not be too difficult to extend our results to produce enclosures of these solutions, rather than the discrete sequence of points which we currently shadow.

The question of whether shadows are typical of true orbits chosen at random is a large open question, but we point out that the same question must be asked of other methods of backwards error analysis. A possible start would be to extend the work of Góra and Boyarsky (1988) to continuous systems in arbitrary dimension, as discussed in Chapter 2.

Currently, almost all shadowing work of which this author is aware consists of trying to prove that a shadow *exists*. However, failure to prove existence does not imply a shadow does not exist. Trying to prove that a shadow does not exist for a pseudo-trajectory is an interesting problem, because it would lead naturally to the question of, in what sense is a non-shadowable trajectory valid? Some convincing work has already been done in this direction (Dawson, Grebogi, Sauer, and Yorke 1994; Sauer, Grebogi, and Yorke 1997), although none of it is rigorous. Making the work rigorous could involve, for example, proving non-hyperbolicity via validated integration of the variational equation (which would be very expensive).

Note that it may not be possible to disprove the existence of shadows in general for any particular system of equations. For example, although we have found that the n-body problem is hard to shadow, and previous work (Quinlan and Tremaine 1992; Hayes 1995) suggests that n-body shadows do not last forever, there almost certainly exist pseudo-trajectories of the n-body problem which possess infinitely long shadows: *eg.,* any machine-representable periodic orbit with sufficiently small local noise, or even stable non-periodic solutions, are probably shadowable indefinitely.

For the most part, the current thesis expounds only a method of producing shadows. It would clearly be interesting to start *applying* this method to interesting problems. The author, for example, is very interested in determining whether long-term solar system integrations (Wisdom and Holman 1991; Wisdom 1992; Sussman and Wisdom 1992; Laskar 1994; Laskar 1997), or long-term three-body problem integrations which are known to be stable (Gladman 1993), are shadowable. More generally, the only work of which we are aware that deals with systems with more than a few dimensions is this author's Master's thesis; clearly, shadowing high-dimensional systems is an area ripe for further study.

Glossary

arc A topological term that is used to describe what is more commonly called a *simple curve*. Formally, an *arc*, or *simple, non-closed curve* is a one-dimensional space that is homeomorphic to the unit interval [0,1] (Munkres 1975). This implies that an arc can be parametrized by a variable $t \in [0, 1]$. (The "non-closed" adjective is to distinguish it from a *simple closed curve*, which is a curve that is homeomorpic to a circle.)

curve a line, either straight, or continuously bending; a path. Note that a curve, by definition, is continuous.

diffeomorphism a homeomorphism whose first derivative is also a homeomorphism.

ergodicity a map is ergodic if almost all initial conditions lead to solutions whose time distribution in the limit as $t \to \infty$ is independent of the initial state.

homeomorphism a map which is continuous, 1-to-1, and onto.

simple curve a non-self-intersecting curve. Also an *arc*.

Bibliography

Aarseth, S. (1999). From NBODY1 to NBODY6: The growth of an industry. In *AAS/Division of Dynamical Astronomy Meeting*, Volume 31.

Ahmed, M. O. and R. M. Corless (1997). The method of modified equations in maple. In M. Wester and S. Steiberg (Eds.), *Electronic Proceedings of the 3rd International IMACS conference on Applications of Computer Algebra, Maui, July 24–26*. Marcel Dekker, New York.

Alefeld, G. and J. Herzberger (1983). *Introduction to Interval Computations*. Academic Press, New York.

Anosov, D. V. (1967). Geodesic Flows and Closed Riemannian Manifolds with Negative Curvature. *Proc. Steklov Inst. Math 90*, 1.

Ascher, U. M., R. M. M. Mattheij, and R. D. Russell (1988). *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. Prentice-Hall Series in Computational Mathematics. Prentice-Hall.

Bendtsen, C. and O. Stauling (1996, August). FADBAD, a flexible C++ package for automatic differentiation, using the forward and backward methods. Technical Report IMM-REP-1996-17, Department of Mathematical Modelling, Technical University of Denmark, 2800 Lyngby, Denmark.

Bendtsen, C. and O. Stauling (1997, April). TADIFF, a flexible C++ package for automatic differentiation, using taylor series expansion. Technical Report IMM-REP-1997-07, Department of Mathematical Modelling, Technical University of Denmark, 2800 Lyngby, Denmark.

Berz, M. (1997). COSY INFINITY version 8 reference manual. Technical Report MSUCL-1088, National Superconducting Cyclotron Lab., Michigan State University, East Lansing, Mich. Available at `http://www.BeamTheory.nscl.msu.edu/cosy/`.

Berz, M. and K. Makino (1998). Verified integration of ODEs and flows using differential algebraic methods of high-order Taylor models. *Reliable Computing 4*, 361–369.

Beyn, W.-J. (1987). On Invariant Closed Curves for One-Step Methods. *Numer. Math. 51*, 103–122.

Bowen, R. (1975). ω-Limit Sets for Axiom A Diffeomorphisms. *Journal of Differential Equations 18*, 333.

Braun, M. (1983). *Differential Equations and Their Applications* (3rd ed.). Springer-Verlag.

Braun, M. (1993). *Differential Equations and Their Applications* (4th ed.). Springer-Verlag.

Channell, P. J. and C. Scovel (1990). Symplectic integration of Hamiltonian Systems. *Nonlinearity 3*, 231–259.

Char, B. W. (1993). *First leaves: a tutorial introduction to Maple V*. Springer-Verlag.

Chow, S. N., X. B. Lin, and K. J. Palmer (1989). A shadowing lemma for maps in infinite dimensions. In C. M. Daffermos, G. Ladas, and G. Papanicolaou (Eds.), *Differential Equations: Proceedings of the EQUADIFF Conference*, pp. 127–136. Marcel Dekker, New York.

Chow, S.-N. and K. J. Palmer (1991). On the numerical computation of orbits of dynamical systems: the one-dimensional case. *Dynamics and Differential Equations 3*, 361–380.

Chow, S.-N. and K. J. Palmer (1992). On the numerical computation of orbits of dynamical systems: the higher dimensional case. *Journal of Complexity 8*, 398–423.

Chow, S.-N. and E. S. Van Vleck (1992). A shadowing lemma for random diffeomorphisms. *Random & Computational Dynamics 1*(2), 197–218.

Chow, S. N. and E. S. Van Vleck (1993). Shadowing of Lattice Maps. In P. E. Kloeden and K. J. Palmer (Eds.), *Chaotic Numerics*, pp. 97–113. American Mathematical Society.

Chow, S.-N. and E. S. Van Vleck (1994a, July). A Shadowing Lemma Approach to Global Error Analysis for Initial Value ODEs. *SIAM Journal on Scientific Computing 15*(4), 959–976.

Chow, S. N. and E. S. Van Vleck (1994b). Shadowing of lattice maps. *Contemporary Mathematics 172*, 97–113.

Clarke, D. A. and M. J. West (Eds.) (1997). *The 12th "Kingston Meeting": Computational Astrophysics*, Volume 123 of *ASP Conference Series*. Astronomical Society of the Pacific.

Coomes, B. A. (1997, January). Shadowing orbits of ordinary differential equations on invariant submanifolds. *Transactions of the American Mathematical Society 349*(1), 203–216.

Coomes, B. A., H. Koçak, and K. J. Palmer (1994a). Periodic shadowing. In P. Kloeden and K. Palmer (Eds.), *Chaotic Numerics*, Volume 172 of *Contemporary Mathematics*, pp. 115–130. Amer. Math. Soc., Providence, RI.

Coomes, B. A., H. Koçak, and K. J. Palmer (1994b). Shadowing orbits of ordinary differential equations. *Journal of Computational and Applied Mathematics 52*, 35–43.

Coomes, B. A., H. Koçak, and K. J. Palmer (1995a). Rigorous computational shadowing of orbits of ordinary differential equations. *Numerische Mathematique 69*, 401–421.

Coomes, B. A., H. Koçak, and K. J. Palmer (1995b). A shadowing theorem for ordinary differential equations. *Z angew Math Phys (ZAMP) 46*, 85–106.

Coomes, B. A., H. Koçak, and K. J. Palmer (1997). Long periodic shadowing. *Numerical Algorithms 14*, 55–78.

Corless, R. M. (1992a). Continued fractions and chaos. *American Mathematical Monthly 99*(3), 203–215.

Corless, R. M. (1992b). Defect-controlled numerical methods and shadowing for chaotic differential equations. *Physica D 60*, 323–334.

Corless, R. M. (1994a). Error Backward. *Contemporary Mathematics 172*, 31–62.

Corless, R. M. (1994b). What good are numerical simulations of chaotic dynamical systems? *Computers Math. Applic. 28*(10–12), 107–121.

Corless, R. M. (1997). Continued fractions and chaos. In *Canadian Mathematical Society Conference Proceedings*, Volume 20. Reprinted from *American Mathematical Monthly* 99 (1992), no. 3, 203–215.

Corless, R. M. and G. F. Corliss (1992). Rationale for guaranteed ODE defect control. In L. Atanassova and J. Herzberger (Eds.), *Computer Arithmetic and Enclosure Methods*. Elsevier Science Publishers B. V. (North-Holland). 1992 IMACS.

Dahlquist, G. and Å. Björck (1974). *Numerical Methods*. Prentice-Hall series in Automatic Computation. Prentice-Hall.

Dawson, S., C. Grebogi, T. Sauer, and J. A. Yorke (3 Oct 1994). Obstructions to Shadowing When a Lyapunov Exponent Fluctuates about Zero. *Physical Review Letters 73*(14), 1927–1930.

Enright, W. H. (1993). The relative efficiency of alternative defect control schemes for high-order continuous Runge-Kutta formulas. *SIAM Journal on Numerical Analysis 30*(5), 1419–1445.

Farmer, J. D. and J. J. Sidorowich (1991). Optimal shadowing and noise reduction. *Physica D 47*, 373–392.

Fryska, S. T. and M. A. Zohdy (1992). Computer dynamics and the shadowing of chaotic orbits. *Physics Letters A 166*, 340–346.

Gladman, B. (1993). Dynamics of systems of two close planets. *Icarus 106*, 247–263.

Golub, G. H. and C. F. Van Loan (1991). *Matrix computations*. Johns Hopkins University Press.

Gonzalez, O. and A. M. Stuart (1996). Remarks on the Qualitative Properties of Modified Equations. Preprint.

Góra, P. and A. Boyarsky (1988). Why computers like Lebesgue measure. *Comput. Math. Applic. 16*(4), 321–329.

Grebogi, C., S. M. Hammel, J. A. Yorke, and T. Sauer (1990, 24 September). Shadowing of Physical Trajectories in Chaotic Dynamics: Containment and Refinement. *Physical Review Letters 65*(13), 1527–1530.

Hadeler, K. P. (1996). Shadowing orbits and Kantorovich's theorem. *Numerische Mathematique 73*, 65–73.

Hager, W. W. (1989). *Applied Numerical Linear Algebra*. Prentice-Hall.

Hairer, E., S. P. Nørsett, and G. Wanner (1993). *Solving Ordinary Differential Equations* (2nd ed.). Springer-Verlag. Two volumes.

Hammel, S. M., J. A. Yorke, and C. Grebogi (1987). Do Numerical Orbits of Chaotic Dynamical Processes Represent True Orbits? *Journal of Complexity 3*, 136–145.

Hammel, S. M., J. A. Yorke, and C. Grebogi (1988). Numerical Orbits of Chaotic Dynamical Processes Represent True Orbits. *Bull. Am. Math. Soc. 19*, 465–470.

Hayes, W. (1995, January). Efficient Shadowing of High Dimensional Chaotic Systems with the Large Astrophysical N-body Problem as an Example. Master's thesis, Dept. of Computer Science, University of Toronto. Available on the web at `http://www.cs.toronto.edu/NA/reports.html#hayes95` and via anonymous ftp from `ftp.cs.utoronto.ca/pub/reports/na/hayes-95-msc`.

Hayes, W. and K. R. Jackson (1996). A Fast Shadowing Algorithm for High Dimensional ODE Systems. Unpublished.

Hayes, W. and K. R. Jackson (1997). Global error measures for the large gravitational n-body problem. See Clarke and West (1997), pp. 237–239.

Hindmarsh, A. C. (1980). LSODE and LSODI, two new initial value ordinary differential equation solvers. *ACM-SIGNUM Newsletter 15*(4), 10–11.

Hull, T. E., W. H. Enright, and K. R. Jackson (1976, october). User's guide for DVERK — a subroutine for solving non-stiff ODEs. Technical Report UT-DCS-100, Dept. of Computer Science, University of Toronto. Source code available from the Authors, or from `http://www.netlib.org/` in the ODE directory.

Jackson, K. R. and R. N. Pancer (1992, May). The Parallel Soluton of ABD Systems Arising in Numerical Methods for BVPs for ODEs. Technical Report 255/91, Dept. of Computer Science, University of Toronto.

Kahaner, D., C. Moler, and S. Nash (1989). *Numerical Methods and Software*. Prentice-Hall series in Computational Mathematics. Prentice-Hall.

Laskar, J. (1994). Large-scale chaos in the solar system. *Astronomy and Astrophysics 287L*, 9–12.

Laskar, J. (1997). Large scale chaos and the spacing of the inner planets. *Astronomy and Astrophysics 317*, L75–L78.

Lorenz, E. N. (1963, March). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences 20*(2), 130–141. Also reprinted in *Chaos* (1984), by Hao Bai-Lin, World Scientific Publishing Co., Singapore.

Merritt, D. (1999, February). Elliptical galaxy dynamics. *Publications of the Astronomical Society of the Pacific 111*(756), 129–168.

Merritt, D., J. A. Sellwood, and M. Valluri (Eds.) (1999). *Galaxy Dynamics: a Rutgers Symposium*, Volume 182 of *ASP Conference Series*. Astronomical Society of the Pacific.

Merritt, D. and M. Valluri (1996). Chaos and mixing in triaxial stellar systems. *The Astrophysical Journal 471*, 82–105.

Moore, R. E. (1966). *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J.

Munkres, J. R. (1975). *Topology: a first course*. Prentice-Hall, Englewood Cliffs, New Jersey.

Murdock, J. (1995). Shadowing multiple elbow orbits: an application of dynamical systems to perturbation theory. *J. Diff. Eq. 119*(1), 224–247.

Nedialkov, N. S. (1999). *Computing rigorous bounds on the solution of an initial value problem for an ordinary differential equation*. Ph. D. thesis, Department of Computer Science, University of Toronto.

Nedialkov, N. S. and K. R. Jackson (2000). A new perspective on the wrapping effect in interval methods for initial value problems for ordinary differential equations. In preparation.

Nedialkov, N. S., K. R. Jackson, and G. F. Corliss (1999). Validated solutions of initial value problems for ordinary differential equations. *Applied Mathematics and Computation 105*(1), 21–68.

Palmer, K. J. (1988). Exponential Dichotomies, the Shadowing Lemma and Transversal Homoclinic Points. In U. Kirchgraber and H. O. Walther (Eds.), *Dynamics Reported*, Volume 1. Wiley and Teubner.

Palmer, K. J. and D. Stoffer (1995). Validated shadowing of numerically constructed pseudo-periodic orbits in chaotic systems. Talk given at Dynamical Numerical Analysis Conference, Georgia Tech, Atlanta, Georgia, December 1995.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C* (second ed.). Cambridge University Press.

Quinlan, G. D. and S. Tremaine (1991). Shadow orbits and the gravitational N-body problem. In B. Sundelius (Ed.), *Dynamics of Disc Galaxies*, pp. 143–148. Göteborg, Sweden.

Quinlan, G. D. and S. Tremaine (1992). On the reliability of gravitational N-body integrations. *Monthly Notices of the Royal Astronomical Society 259*, 505–518.

Sanz-Serna, J. M. (1992). Symplectic integrators for Hamiltonian problems: an overview. In A. Iserles (Ed.), *Acta Numerica 1992*, pp. 243–286. Cambridge University Press.

Sanz-Serna, J. M. and S. Larsson (1993). Shadows, Chaos, and Saddles. *Appld. Numer. Math. 13*, 181–190.

Sauer, T., C. Grebogi, and J. A. Yorke (7 July 1997). How long do numerical chaotic solutions remain valid? *Physical Review Letters 79*(1), 59–62.

Sauer, T. and J. A. Yorke (1991). Rigorous Verification of Trajectories for the Computer Simulation of Dynamical Systems. *Nonlinearity 4*, 961–979.

Shadwick, B. A., J. C. Bowman, and P. J. Morrison (1999). Exactly conservative integrators. *SIAM Journal on Applied Mathematics 59*(3), 1112–1133.

Skeel, R. (1996). The Meaning of Molecular dynamics. Unpublished, e-mail skeel@cs.uiuc.edu.

Skeel, R. D. (1999). Integration schemes for molecular dynamics and related applications. In M. Ainsworth, J. Levesley, and M. Marletta (Eds.), *The Graduate Student's Guide to Numerical Analysis*, SSCM, pp. 119–176. Springer-Verlag.

Smale, S. (1965). Diffeomorphisms with many periodic points. In S. Cairns (Ed.), *Differential and Combinatorial Topology*, pp. 63–80. Princeton Univ. Press.

Smale, S. (1967). Differentiable dynamical systems. *Bull. Amer. Math. Soc. 73*, 747–817.

Struck, C. (1997). Galaxy splashes: The effects of collisions between gas-rich galaxy disks. See Clarke and West (1997), pp. 225–230.

Stuart, A. M. and O. Gonzalez (1996). Comments on Qualitative Properties of Integrators. Talk given at SIAM Conference in Dynamical Systems, Snowbird, Utah, May 1995. See also Gonzalez and Stuart (1996).

Stuart, A. M. and A. R. Humphries (1996). *Dynamical Systems and Numerical Analysis.* Cambridge University Press.

Sussman, G. J. and J. Wisdom (1992). Chaotic Evolution of the Solar-System. *Science 257*, 56–62.

Tupper, J. A. (1996, January). Graphing Equations with Generalized Interval Arithmetic. Master's thesis, Dept. of Computer Science, University of Toronto. Available on the web at `http://www.dgp.toronto.edu/~mooncake/msc.html`.

Van Vleck, E. S. (1995). Numerical Shadowing Near Hyperbolic Trajectories. *SIAM Journal on Scientific Computing 16*(5), 1172–1189.

Wisdom, J. (1992). Long-Term Evolution of the Solar-System. *IAU Symposia*, 17–24.

Wisdom, J. and M. Holman (1991, October). Symplectic maps for the n-body problem. *The Astronomical Journal 102*, 1528–1538.