CONTROLLING CHAOS, TARGETING, AND TRANSPORT

by

ERIK M. BOLLT

B.S., University of California at Berkeley, 1990

M.S., University of Colorado at Boulder, 1992

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Program in Applied Mathematics

1995

This thesis for the Doctor of Philosophy degree by

Erik M. Bollt

has been approved for the

Program in

Applied Mathematics

by

James D. Meiss

Robert W. Easton

James H. Curry

Elizabeth Bradley

Steve McCormick

Date _____

Bollt, Erik M. (Ph. D., Program in Applied Mathematics)

Controlling Chaos, Targeting, and Transport

Thesis directed by Professor James D. Meiss

The sensitivity that defines chaotic dynamics makes accessible a wide range of behaviors using arbitrarily small control signals. "Controlling chaos" attempts to cause large changes in the dynamics using only small perturbations. In *targeting*, one attempts to find a fast path from an initial condition $\boldsymbol{a}$ to a target point $\boldsymbol{b}$ by exploiting the fact that transport times for a chaotic system are highly sensitive to initial conditions and parameter values. The main difficulty is finding the *switching points*, the times and places to apply judiciously chosen perturbations.

I present a new technique to find rough orbits (epsilon chains) that rapidly achieve a desired transport. The strategy is to build the epsilon chain from segments of a long orbit. In two-dimensional maps, long orbits have recurrences in neighborhoods where faster orbits must also pass. Long orbits of higher dimensional maps are likely to have recurrences, albeit less frequently. The recurrences are used as switching points between segments. If a local hyperbolicity condition is satisfied, then a nearby shadow orbit might be constructed.

In one example, I show that transport times for the standard map can typically be reduced by a factor of 10^4. In another example, I apply the technique to the restricted three-body problem from which I find a low energy Earth-Moon transfer orbit which requires 38% less characteristic velocity than a comparable Hohmann transfer orbit.

In yet another example, a symbol dynamics model has a closed-form expression for the optimal transporting orbit from near $\boldsymbol{a}$ to near $\boldsymbol{b}$. I compare

the optimal orbit to the targeted orbit resulting from removing recurrences, which also takes a particularly simple form in symbol dynamics.

The techniques developed here do not require a closed-form representation of the map. Using the standard map as an example, I demonstrate that predictions from a time series may be sufficient for targeting.

Finally, as a contribution to the understanding of barriers in high-dimensional Hamiltonian maps, I present a technique to investigate the breakup of invariant tori with fixed frequency of a four-dimensional generalization of the complex, semi-standard map.

I dedicate this dissertation to my wife Elizabeth who has supported me, helped me, and endured me through this work.

ACKNOWLEDGMENTS

I am deeply indebted to my advisor James D. Meiss for his time and patience in helping me to achieve this important goal in my life. Early in our working together, there were many times when it probably would have been easier and faster, for him, to cut me out of the loop, but he never did; he always patiently waited and helped me to come to speed. Later, he always helped my self confidence by expecting my best from me, but never in demanding fashion; in his quiet reassuring way, he exudes confidence. When I recently thanked him, he simply said, "I'm just doing my job." Well then, I would like to say that he is very good at his job.

I would like to thank the Mathematics Research Centre, and particularly Robert Mackay and Mark Muldoon, at the University of Warwick for their hospitality during the fall of 1993. I am indebted to the several people who have helped me edit this dissertation: Carol Werckman, my wife Elizabeth Bollt, my father Dr. Steven Bollt, David Sholl, and David Sterling. I have enjoyed friendships and working relationships with fellow graduate students Cesar Ocampo, David Sholl, David Sterling, and Peter Staab. I would also like to thank professors Elizabeth Bradley, S. Aubry, and Mark Muldoon for their interest and valuable discussions on my work. I especially would like to thank my parents for their support and continued love. This work has been partly funded by NSF grants DMS-9305847, an NSF Graduate Research Traineeship, DMS-9208685, and a NATO collaborative research grant CRG 921181.

CONTENTS

TABLES

TABLE

FIGURES

FIGURE

0.1 Forward

In this thesis I present techniques to exploit the sensitivity that defines chaotic systems. Small, well-placed controlling perturbations in chaos can often suffice for a wide range of system states. Locally calculating control perturbations requires the application of standard linear control theory. Since control signals are typically small, the methodology of "controlling chaos" is the collection of techniques that selectively chooses amongst nearby behaviors which are concatenated to best exploit the available dynamics.

"Targeting" is the task of finding small control perturbations to rapidly direct an orbit to a specified system state using the nonlinear nature of the dynamics. Targeting may be described as having three aspects:

1. Local calculations to choose the correct perturbations to alter the dynamics.
2. An information data set formed from a long pre-study or observation period in which observed orbit segments are recorded and organized for quick access in the form of a library of known behaviors.
3. A general understanding of the transport mechanisms typical of chaotic dynamical systems.

Hamiltonian dynamics are particularly difficult to target because partial barriers layer the phase space into resonance regions with slow average transport between resonances. In this sense, there is an "order" in the "chaos". Previous techniques have been unable to find paths through such layered phase spaces due to difficulty in finding orbit segments between distinct layers. The targeting technique of this thesis excels in identifying the switching points at the

recurrences in turnstiles between layers; such identification proves particularly effective in improving slow transport.

The organization of this thesis is outlined below. Chapter 1.1 reviews previous work in chaos control. In Chapter 1.2, I discuss known chaotic transport mechanisms, a definition of the term "barrier", a description of the homoclinic tangle in two dimensions. Chapter 1 closes with a numerical demonstration of the localization implied by transport across a barrier in the standard map.

In Chapter 2, a numerical experiment demonstrates that transport across a barrier in the Hénon map can be performed solely by monitoring activity in the barrier's small turnstile. This demonstration is followed by the technique which constitutes a major part of this thesis: "Controlling Chaotic Transport through Recurrence." I argue that localization of a transporting orbit across a barrier can distinguish slow orbits from fast orbits and identifies the control switching points. An inefficient orbit typically crosses a barrier several times, while an efficient orbit crosses only once. Thus a recurrence in the turnstile during successive crossings identifies an inefficient orbit. By removing (frequently long) recurrent loops a locally efficient orbit may be created. Loops may be removed by shooting at the orbit's stable manifold after the recurrence from the unstable manifold of the orbit before the recurrence, hence constructing a shadow orbit. This targeting technique is demonstrated for the standard map, for which we show at least a factor of four improvement over the numerically calculated average transport rate. The reliance on hyperbolicity of constructing a shadow orbit is discussed We numerically investigate this reliance in the "chaotic" region accessible to our starting point $\boldsymbol{a}$ and target point $\boldsymbol{b}$ by testing if the stable and unstable directions are bounded away from zero. Chapter 2 closes with a modification of the control algorithm can be readily modified for

targeting *on the fly* using small parameter perturbations, which we demonstrate by numerical experiment.

In Chapter 3, we apply the control algorithm of the previous chapter to the planar restricted three-body problem. This has been shown to be a good first-order model of a spacecraft in the Earth-Moon system. After deriving the time-independent two degrees-of-freedom Hamiltonian and equations of motion we investigate dynamical properties of the flow which is restricted to a three-dimensional submanifold. Numerical integration of these equations is nontrivial because they are stiff near the singularities. Performing a Poincaré section yields an area preserving map in two dimensions. The resonance layered phase space structure of the restricted three body problem is qualitatively similar to that of the standard map. Hence, the targeting problem is like that of the standard map. We also discuss the special issue of finding the most stable and unstable directions in the full four-space, correctly restricted to the surface of section. The chapter is concluded by numerically demonstrating a new chaotic orbit from a parking orbit around the Earth to a Moon-orbiting KAM circle which requires only small transfer manoeuvres. The chaotic orbit requires 38% less characteristic velocity when compared to a classical Hohmann transfer orbit.

Chapter 4 is devoted to finding a closed form solution to the optimal control problem of transport in a horseshoe map model dynamics. We write the closed form optimal transfer orbit between a given starting point $\boldsymbol{a}$ and target point $\boldsymbol{b}$ to specified precision which we then compare to a closed-form path calculated by the technique in Chapter 2. We find that our targeted path is nonoptimal, but nonetheless good. Finally, we discuss the applicability of the model.

In Chapter 5 we discuss control where a global model is not available,

but a large data set of observed iterates can be substituted to make nearby predictions. The goal is to control chaotic dynamics using only a map reconstructed by a delay embedding of an experimental time series. First we review the technique of time series embedding. Then we formulate the equations required to least-squares fit a local affine model for nearest-neighbor prediction. Numerical results are presented for the standard map targeting problem where moderate success is demonstrated. We close the chapter with a discussion of the serious limitations of conceivable computer resources presented by increasing dimension.

Chapter 6 represents a contribution to the understanding of barriers and their breakup in higher-dimensional Hamiltonian dynamics. We compute the domain of existence of two-dimensional invariant tori with fixed frequency vectors for a four-dimensional, complex, symplectic map. The map is a generalization of the semi-standard map studied by Greene and Percival. It has three parameters: a_1 and a_2 represent the strength of the kicks in each degree of freedom, and ϵ represents the coupling. The domain of existence of a torus in (a_1, a_2) is shown to be complete and log-convex for fixed $k = \frac{\epsilon}{a_1 a_2}$. Explicit bounds on the domain for fixed k are obtained. Numerical results show that quadratic irrationals can be more robust than the cubic irrational, "the spiral mean".

Finally, there are several further explanations and directions for future research presented in the appendices. In particular, the orbit-restriction method is presented as an alternative technique to shooting from the unstable manifold to the stable manifold. It uses the action of a pseudo-orbit to relax the entire orbit patch simultaneously. Shooting between higher-dimensional manifolds is also investigated along with a discussion of Bennitin's algorithm for constructing higher-dimensional stable and unstable manifolds. Lastly there is a discussion

of the targeting technique which I presented as my thesis proposal: the decomposition of the large targeting problem in the standard map into smaller subproblems by targeting between switching points which can be found in the turnstiles corresponding to "noble" cantori. After several months of difficulties, I achieved a working solution to this approach just several hours before realizing the more elegant and widely applicable solution of removing recurrences. The last appendix contains this solution of my original approach, for completeness.

CHAPTER 1

INTRODUCTION

1.1 Using Chaos

Chaos in a nonlinear dynamical system is roughly defined as the presence of extreme sensitivity to perturbations. Specifically, this means that two nearby initial conditions, can quickly part company to evolve along completely separate and seemingly unassociated trajectories. This property seems to make chaotic systems difficult to forecast, and hopeless to control. Historically, a typical approach for predicting any dynamical system was to attempt to fit global linear models, even if the basic dynamics were nonlinear (e.g., regression analysis is still used widely today in many sciences). A control theorist, when presented with a possibly chaotic system, would try to avoid the chaos either by re-engineering all together or by making large and often costly changes in the system using large perturbations. Generally, chaos was a pitfall.

This attitude has been changing. The very sensitivity of chaos presents a unique opportunity not available to non-chaotic dynamics: a deliberate small perturbation can result in drastically different system performance, often with rapid response. Even if a budget restricts one to small perturbations, a slightly perturbed initial condition can access virtually all phase space behaviors. Characteristically, every small neighborhood of a given point in a chaotic region of phase space contains points on orbits that will visit arbitrarily closely to all other regions of chaotic and accessible phase space. Near any given point, orbits exist

with drastically different behaviors and rates to destinations.

Controlling chaos is the exercise of encouraging the dynamics to follow one particular behavior of its many regular behaviors. This is typically done by first learning (i.e. recording by observation) the behaviors. Standard linear control theory may then be applied locally near observed dynamical behavior. In this way we can concatenate a pseudo-orbit from the library of observed data by only small alterations to the dynamics. A pseudo-orbit[1] which makes use of the nonlinear aspects of the dynamics is achieved by keeping the control orbit within the limit of validly applying linear control to the pre-recorded orbits.

Many practical uses of chaos control have already been achieved. These include:

1. The synchronization of the output of chaotic electronic circuits [19].
2. The increase of the power output of a broad banded laser by selecting the unstable frequency output peak [88].
3. A pacemaker which successfully stabilizes a periodic heart beat of a rabbit with an arrhythmia (aperiodic heart beat) [36].
4. A technique for secure communications by encoded electronic messages [43].
5. Maneuvering the spacecraft ISEE-3/ICE on an extremely limited fuel budget more than fifty million miles across space to an otherwise impossible comet rendezvous [32, 25].

1.1.1 Stabilization One of the major tools borrowed from traditional control theory is the linear feedback control loop. In this section I will

[1] A pseudo-orbit $\{\boldsymbol{z}_i\}$ of a map T such that $\|T(\boldsymbol{z}_i) - \boldsymbol{z}_{i+1}\| < \epsilon$, $\forall i$ is defined an *ϵ-chain*.

present the well-known pole placement solution, and then the special solution which is most useful to chaotic control.

We begin by constructing a Poincaré map from a flow

$$\dot{\boldsymbol{z}} = F_k(\boldsymbol{z}),\ \boldsymbol{z} \in \mathbb{R}^{d+1}, \tag{1.1}$$

yielding a map

$$\boldsymbol{z}_{n+1} = T_k(\boldsymbol{z}_n),\ \boldsymbol{z} \in \mathbb{R}^{d}. \tag{1.2}$$

Note that I have explicitly written the dynamics in a form that expresses dependence on an adjustable parameter $k \in \mathbb{R}$. We allow only small variations δk_{max} around the nominal value k_0.

I will describe stabilizing an unstable fixed point for ease of exposition. The straightforward generalization to periodic points will be discussed at the end of this section. Suppose the dynamics (1.2) has an unstable fixed point $\boldsymbol{z}^*_{k_0}$ at the nominal parameter value. We will stabilize $\boldsymbol{z}^*_{k_0}$ by adding a small forcing term, known as a feedback gain, to (1.2). We linearize (1.2) near $\boldsymbol{z}^*_{k_0}$,

$$(\boldsymbol{z}_{n+1} - \boldsymbol{z}^*) = A \cdot (\boldsymbol{z}_n - \boldsymbol{z}^*) + \delta k B, \tag{1.3}$$

where $A = \frac{\partial T_i}{\partial \boldsymbol{z}_j}|_{\boldsymbol{z}=\boldsymbol{z}^*}$ is the $d \times d$-dimensional Jacobian matrix and $B = \frac{\partial T_i}{\partial k}|_{k=k_0}$ is the $d \times 1$-dimensional matrix of parameter variations of the map. If a linear time-dependence of the parameter k of the form

$$\delta k = -K^t \cdot (\boldsymbol{z}_n - \boldsymbol{z}^*), \tag{1.4}$$

is assumed, in terms of a $1 \times d$ "feedback gain matrix" K^t, on substitution into Eq. (1.3) we obtain

$$(\boldsymbol{z}_{n+1} - \boldsymbol{z}^*) = (A - B \cdot K^t) \cdot (\boldsymbol{z}_n - \boldsymbol{z}^*). \tag{1.5}$$

This is a linear transformation from a state $\boldsymbol{z}_n$ near $\boldsymbol{z}^*$ to the new state $\boldsymbol{z}_{n+1}$.

To stabilize $\boldsymbol{z}^*$, we must choose K such that $A - B \cdot K^t$ has eigenvalues of modulus less than one. This is sufficient for Eq. (1.5) to be a contraction mapping with stabilized fixed point $\boldsymbol{z}^*$. The correct choice of the matrix K, and the requirements on the original dynamics Eq. (1.2) such that the matrix exists, is the issue of "controllability" resolved by the well-known "pole placement technique" from control theory [78].

We can uniquely choose K such that the eigenvalues ("poles") of $A - B \cdot K^t$ have any specified values, with corresponding consequences to the stability of the dynamics in Eq. (1.5), if and only if A and B satisfy the controllability condition that the $n \times n$ matrix

$$C = [B|AB|...|A^{n-1}B], \tag{1.6}$$

has full rank. If C is of sufficient rank we may determine K with Ackerman's formula [78] in terms of C, A, the original unaltered poles of A, and the desired "regulator" poles of $A - B \cdot K^t$. Note that the "gain" matrix is fixed, independently of $\boldsymbol{z}_n$, once $\boldsymbol{z}^*$ and the regulator poles have been chosen. See Appendix A.1 for further discussion of the pole placement technique and derivation of Ackerman's formula.

An immediate result of Eq. (1.4) is a description of the controllable region around $\boldsymbol{z}^*$ in terms of a maximum perturbation limit $|\delta k_{max}|$.

$$|\delta k| = |K^t \cdot (z_n - z^*)| < |\delta k_{max}|. \tag{1.7}$$

This defines a region in which the limited perturbation still allows the desired stabilizing poles, but only in terms of a linear theory. For large δk_{max}, a linearization of Eq. (1.2) can break down. However, this type of control has been successfully used in conjunction with the technique of E. Ott, C. Grebogi, and J. Yorke

[79], commonly known as "OGY" control. Their technique has been demonstrated in many examples, including stabilization of points in the Hénon map and in the double rotor (see [86]).

A special case of pole placement applies when the stability matrix at $\boldsymbol{z}^*$ has a hyperbolic structure, as is common for chaotic systems. A periodic point $\boldsymbol{z}^*$ is defined as *hyperbolic* if the tanget manifold of the map T at $\boldsymbol{z}^*$ (given by the Jacobian matrix A) has a decomposition into stable and unstable subspaces, identified E^s and E^u. That is, there is no center space, E^c. These subspaces are defined by the eigenvalues and eigenvectors of A as follows:

$$\begin{aligned}
\boldsymbol{E^s} &\equiv \{\boldsymbol{z} : \boldsymbol{z} \in \textstyle\bigoplus_i \boldsymbol{v_{s,i}} \text{ where } \boldsymbol{v_{s,i}} \text{ correspond to } \boldsymbol{|\lambda_{s,i}| < 1}\}, \\
\boldsymbol{E^u} &\equiv \{\boldsymbol{z} : \boldsymbol{z} \in \textstyle\bigoplus_i \boldsymbol{v_{u,i}} \text{ where } \boldsymbol{v_{u,i}} \text{ correspond to } \boldsymbol{|\lambda_{u,i}| > 1}\}, \\
\boldsymbol{E^c} &\equiv \{\boldsymbol{z} : \boldsymbol{z} \in \textstyle\bigoplus_i \boldsymbol{v_{c,i}} \text{ where } \boldsymbol{v_{c,i}} \text{ correspond to } \boldsymbol{|\lambda_{c,i}| = 1}\}.
\end{aligned}$$

For a hyperbolic point, a natural choice for the regulator poles is to cause the λ_s to remain unaltered, and the λ_u to become 0. This was the choice made in the original OGY paper [79]. For this special case, it is possible to derive the necessary perturbations to the original dynamics by geometric arguments. I will describe this derivation below for a simple model in which E^s and E^u are both one-dimensional. The generalization to higher dimensions will be described briefly at the end of this section.

For this simplified model, we assume a change of coordinates so as to cause $\boldsymbol{z}^*_{k_0} = 0$ and $k_0 = 0$. Again, we linearize Eq. (1.2) near the unstable fixed point $\boldsymbol{z}^*_{k_0=0}$, but this time we do not explicitly include the variations in the parameter:

$$(\boldsymbol{z}_{n+1} - \boldsymbol{z}^*_{\delta k}) \approx A \cdot (\boldsymbol{z}_n - \boldsymbol{z}^*_{\delta k}). \tag{1.8}$$

The matrix A may be written in the form

$$\boldsymbol{A = \lambda_u \hat{u} f_u + \lambda_s \hat{s} f_s,} \tag{1.9}$$

where $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{s}}$ are the stable and unstable unit eigenvectors corresponding to the eigenvalues $\lambda_u > 1$ and $\lambda_s < 1$. f_u and f_s are the contravariant basis vectors defined by $\boldsymbol{f_u \cdot \hat{u} = f_s \cdot \hat{s} = 1}$ and $\boldsymbol{f_u \cdot \hat{s} = f_s \cdot \hat{u} = 0}$.

Since there is only one unstable direction, we generically need only one control parameter δk. Hence, B is an $n \times 1$ matrix, which may be approximated by

$$B \equiv \frac{\partial T}{\partial k}|_{k=0} \approx \frac{\boldsymbol{z}^*_{\delta k}}{\delta k}, \tag{1.10}$$

for small δk.

We need to choose the adjustable parameter's perturbation δk that causes $\boldsymbol{z}_n$ to iterate onto the stable direction of $\boldsymbol{z}^*_0$. The vector $\hat{\boldsymbol{u}}(\boldsymbol{z}^*_0)$ lies in the stable subspace $E^s(\boldsymbol{z}^*_0)$, which, in turn, approximates the full stable manifold $W^s(\boldsymbol{z}^*_0)$ close to $\boldsymbol{z}^*_0$. Once $\boldsymbol{z}_{n+1} = T_{\delta k}(\boldsymbol{z}_n)$ lies on the stable manifold of $\boldsymbol{z}^*_0$ at $k = 0$, applying T_0 causes future iterates to march exponentially towards $\boldsymbol{z}^*_0$.[2] We may write the geometric statement that $\boldsymbol{z}_{n+1}$ has zero component in the unstable direction:

$$\boldsymbol{f}_u \cdot \boldsymbol{z}_{n+1} = 0. \tag{1.11}$$

figure=pic1.eps,height=3.5in

Figure 1.1. A geometric interpretation of local stabilizing control of a hyperbolic fixed point. Variations δk of the control parameter cause movement of the fixed point, which can be placed to cause the point $\boldsymbol{z}_n$ to iterate onto the stable manifold of the fixed point at the original parameter value.

[2]Actually, small errors, included in the linearization, require that the stabilization control be frequently reapplied.

Fig. (1.1) displays parameter variations which effectively cause the fixed point under the new parameter value to move to where $\boldsymbol{z}_n$ will iterate onto $E^s(\boldsymbol{z}_0^*)$.

Substituting Eqs. (1.10) and (1.9) into (1.8) and dotting with $\boldsymbol{z}_{n+1}$ yields the expression for the parameter perturbation:

$$\delta k = \frac{\lambda_u}{(\lambda_u - 1)} \frac{\boldsymbol{z}_n \cdot \boldsymbol{f}_u}{B \cdot \boldsymbol{f}_u}. \tag{1.12}$$

From this equation, quickly follows an expression for the size of the "capture window" parallelogram, based on the maximum allowed perturbation δk_{max}.

$$\|\boldsymbol{z}_n\| \leq \delta k_{max} |(1 - \frac{1}{\lambda_u}) B \cdot \boldsymbol{f}|. \tag{1.13}$$

Geometrically, this describes a situation similar to "jiggling" a horse saddle to balance a ball-bearing near the moving unstable fixed point.

To derive a higher-dimensional version of Eq. (1.12) we require neutralization of all unstable vectors spanning $E^u(\boldsymbol{z}^*_{\boldsymbol{k}_0})$. This can be done in one step if m variations of the available control parameters span the m-dimensional unstable subspace $E^u(\boldsymbol{z}^*_{\boldsymbol{k}_0})$. Alternately, if the variation of the single parameter has a component in each of the unstable directions, we can perform the control in m iterations by neutralizing one unstable direction at a time. This is discussed more fully in Appendix A.4.

Both of these methods of stabilizing control are also readily applicable to unstable orbits which are not necessarily fixed points. A period p point can be controlled by applying the above techniques to the fixed point of the composed map T^p. However, this is not a stable solution for long periods p, since control is only possible every p^{th} step [86]. A better alternative is to successively cause iterates of $\boldsymbol{z}_n$ to land on the stable manifold of the corresponding iterates of

$\boldsymbol{z}_{k_0}$. The appropriate equations of the parameter perturbations are found in an analogous fashion to the above. In fact, such a derivation also works when the goal is to stabilize an open ended orbit which is not necessarily periodic. We will see in Chapter 2 that for an open-ended path, it is easier to stabilize by the equivalent technique of shooting at the (future) stable manifold of the path.

We do not need to have an analytic model of a physical system to apply these techniques. All the necessary quantities $\boldsymbol{z}, \boldsymbol{f}_u,$ and, B are accessible from a time series through time-delay embedding [1, 80]. Delay embedding offers the laboratory experimenter, who can measure only one of the system variables, the full dynamics, but in a transformed coordinate system. Further discussion on delay coordinates can be found in Sec. 5.2. In Sec. 5.3, I will explain the technique of Eckmann and Ruelle [26] to predict the Jacobian matrix allowing only observed dynamics.

It may seem surprising that the stability of the d-dimensional dynamics can be so easily altered. We resolve this issue by considering the dimension of the phase space in which the dynamics takes place. The dynamics of $\boldsymbol{z}$ are defined on a d-dimensional hyperplane with fixed k. This is a codimension one slice of the augmented $d+1$-dimensional space $(\boldsymbol{z}, k)$. The controlled dynamics is defined on the $d+1$-dimensional parameter extended space, whose dynamics effectively includes shifting up or down the stack of constant k sheets, as governed by Eq. (1.12).

1.1.2 The OGY Technique The most famous application of chaos control is the OGY technique [79]. An earlier paper due to A.M. Bloch and J.E. Marsden [10] outlines a similar strategy. The goal is to stabilize an unstable periodic orbit of the map (1.2) using only small parameter perturbations

$\delta k \leq \delta k_{max}$. The first step is to find the periodic orbit (perhaps numerically) experimentally by an averaging process near recurrences, thus closing the recurrent loop. Alternately, the constructive closing lemma, referred to again in Sec. 2.2.1, is applicable.

In addition to finding the periodic points, we need to collect data corresponding to unstable directions and parameter derivatives as called for by Eq. (1.12). By definition, a chaotic attractor has embedded within it an infinite number of unstable points [86]. Unstable periodic orbits are dense in the attractor and have important implications for the attractor's ergodic properties.

Most chaotic orbits (accessible to $\boldsymbol{z}^*$) eventually enter the capture window. The idea is to choose any initial condition, on the chaotic attractor, in the subset of phase space accessible to the periodic point. There is a typically long, chaotic transient for the orbit to enter the control window described by Eq. (1.13). Once within reach, the orbit can be controlled by applying formula (1.12). The global control program δk_n has the form

$$\delta k_n = \begin{cases} k_0 & \text{if } \|\boldsymbol{z}_n - \boldsymbol{z}^*_{k_0}\| > \delta k_{max} |(1 - \frac{1}{\lambda_u}) B \cdot \hat{u}| \\ \frac{\lambda_u}{\lambda_u - 1} \frac{\boldsymbol{z}_n \cdot \boldsymbol{f}_u}{B \cdot \boldsymbol{f}_u} & \text{otherwise} \end{cases} . \tag{1.14}$$

Hence, an important point is that the global control strategy relies only on the inherent, chaotic properties of the dynamical system that cause the orbit to wander into the control window. The only active control is local. Thus the only information needed, besides the knowledge that the dynamics is chaotic, is local (derivative) information at $\boldsymbol{z}^*$. This technique is particularly accessible to laboratory experiments, since only limited data must be collected.

The technique is robust to system noise. The complication is that unaccounted for perturbations can cause the captured orbit to be kicked outside

the control window. The probability of being kicked out of the control window is a function of the ratio of noise amplitude to control window size. An orbit lost by a noise perturbation can be reacquired, after a possibly lengthy wait, just as any other initial condition is acquired.

An important feature of the control scheme is the acquisition time of an initial condition to wander into the neighborhood of $\boldsymbol{z}^*$, which is unbounded as $\delta k_{max} \to 0$. At any given iteration, the probability of falling within the control window is approximately the natural measure of the window on the uncontrolled chaotic attractor [28].

Transport in an area preserving map is particularly slow; it proceeds only algebraically, due roughly to the "stickiness" effect of resonant island chains. A point initially near a KAM surface has a survival probability, $F(t)$, asymptotic to t^{-z}, that the point will still be near the surface at large time t [42, 71]. We therefore say that, in the presence of KAM surfaces, we find long correlations and, hence, roughly a power law decay. This contrasts to the exponential spreading expected on a typical chaotic attractor. See Sec. 2.2.6 for further discussion.

Hence, we have the potential for unacceptably long transients, and so we are lead to search for more active target acquisition techniques. These techniques will use the full nonlinear aspects of the dynamics. The cost of using the full nonlinear structure of the dynamics is a much higher overhead on learning time and information storage.

1.1.3 Targeting Stabilization, in its various forms, is always a local issue. If, however, a system starts unacceptably far from the "to be stabilized" state, global issues become important. In the previous section, we saw

that the OGY technique avoids the full brunt of this global problem by using the fact that a chaotic orbit will densely fill a chaotic phase space. This solution is costly in that the initial transient may be extremely long, (see [11, 52]), and in particular, Hamiltonian transport is extremely slow. Nonetheless, an advantage of the OGY technique is that only knowledge about the stabilization neighborhoods is needed.

In a chaotic system, small perturbations can steer an orbit to a desired state where it can be stabilized. The time optimal control problem for a chaotic dynamical system is known as "targeting". A new targeting solution is one of the main objectives of this thesis.[3] The goal of targeting is to steer a dynamical system from near an initial condition $\boldsymbol{a}$ to near the target $\boldsymbol{b}$ in the shortest time possible. Since the objective is to use only small perturbations, the target is basically achieved within variations on the dynamic's original behaviors. This is an important contrast to other techniques of control which may resort to large alterations in the dynamics to achieve that goal.

There are several different approaches to targeting, but they all have certain common characteristics. A typically long, global pre-study of the chaotic dynamics is required to learn "what goes where." For this reason, we can only hope to apply these methods to a compact attractor. The cost of targeting is the increased knowledge needed, and pre-processing time. The payoff is that transients are drastically reduced by choosing the best orbit segments amongst those stored in the "library" of already observed dynamics.

Targeting starts with a map which may be in the form of Eq. (1.2), possibly having been derived from a flow such as in Eq. (1.1). The problem is

[3] In the literature, simply improving transport time is commonly called "targeting" [94]; no optimal solution is discussed.

to find an initial condition $\boldsymbol{z}_a$ near a starting point $\boldsymbol{a}$ and a control strategy $\{k_i\}$ that will cause the point to iterate near a target point $\boldsymbol{b}$ as quickly as possible, using only small perturbations. Thus, with $\epsilon > 0$ and $\Delta > 0$ given, we wish to minimize the cost function

$$I(\{k_i\}, \boldsymbol{z}_a) \equiv n, \tag{1.15}$$

where n is the first time that

$$\boldsymbol{z}_b = [\Pi_{i=0}^{n-1} \circ T_{k_i}](\boldsymbol{z}_a) \in B_\epsilon(\boldsymbol{b}), \tag{1.16}$$

subject to the constraints

$$\|k_i - k_0\| < \Delta,\ \boldsymbol{z}_a \in B_\epsilon(\boldsymbol{a}). \tag{1.17}$$

The epsilon ball around $\boldsymbol{a}$ is defined by $B_\epsilon(\boldsymbol{a}) \equiv \{\boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{a}\| < \epsilon\}$. Minimization of I with respect to the constraints is the minimum-time control problem [46, 102]. In general, the minimum of I occurs not at a fixed value of k, but for a sequence of parameter values $\{k_i\}_{i=1}^n$.

I will now briefly review several techniques that have been applied to this problem. The first technique of targeting is by T. Shinbrot *et. al* [92], which can be illustrated with the following simple example from their paper. Consider the logistic map

$$z_{n+1} = T_k(z_n) = k z_n (1 - z_n), \tag{1.18}$$

with a nominal parameter value $k_0 = 3.9$, and starting point $a = z_1 = 0.4$. The goal is to reach a neighborhood of the target $z_n = b = 0.8$ as quickly as possible by varying k in the range 3.8 to 4.0. After one iteration of the entire range $k \in [3.8, 4.0]$, we can have the second iterate in the interval $z_2 \in [0.91, 0.96]$. Only a single perturbation is needed for this example, so all subsequent iterations are made at the parameter k_i set to the nominal value 3.9. Now we iterate the entire

interval of possible z_2's, which grows to $z_3 \in [0.15, 0.31]$. On the third iterate, we find $z_4 \in [0.5, 0.84]$. Since this range of third iterates of z_1 brackets the target $b = 0.8$, we are done. The value of $k \in [3.8, 4.0]$, which iterates directly to b, can be found by Newton's method.[4]

The lesson learned here is that the small perturbation quickly grows to fill the compact phase. As we will see, the fact that the phase space was only one-dimensional plays a large part in making targeting so simple.

This technique can be made to allow for noise, or small modeling errors, by retargeting at each step along the way. Shinbrot *et al.* [91] successfully applied their technique, in a laboratory experiment, to a time series derived model of the dynamics of a magnetoelastic ribbon with a varying Young's modulus. Their time series allowed for a one-dimensional model, and so targeting was similar to targeting the logistic map. In their experiments, small noise and modeling errors were present and unavoidable. They improved typical orbits, which required 500 iterations, to an average of 20 iterations.

Now we will see some of the complications that arise from generalizing the technique to two dimensions [93]. First note that it is no longer typical for $\boldsymbol{z}_a = \boldsymbol{a}$ and $\boldsymbol{z}_b = \boldsymbol{b}$. For the moment, we will continue to restrict ourselves to a single initial-parameter perturbation. Even though one perturbation does not typically yield an optimal I, there may nonetheless exist a $\delta k \in [-\delta k_{max}, \delta k_{max}]$ which iterates to within ϵ of $\boldsymbol{b}$ faster than the original map. As before, the answer can be found by testing the entire possible range of δk, on the computer.

The first iteration of $\boldsymbol{a}$ for the range of possible one-dimensional parameter perturbations results in the line segment[5] $\delta \boldsymbol{z}_{max} = T_{[k_0 - k_{max}, k_0 + k_{max}]}(\boldsymbol{a})$,

[4] Zeroing $T_{k_0}(T_{k_0}(T_k(a))) = b$ as a function of k is straightforward once the number of iterates needed is known. For the above example, $k_0 + \delta k = 3.83189...$

[5] only when $\delta k_{max} \ll 1$

represented by a grid on k. The length of the vector $\delta \boldsymbol{z}_{max}$, the phase space variation due to the small parameter perturbation range $2\delta k_{max}$, can be approximated

$$\|\delta \boldsymbol{z}_{max}\| \approx 2\frac{\partial T}{\partial k}|_{(\boldsymbol{z}^*,k_0)}\delta k. \tag{1.19}$$

The parameter is now returned to k_0, and the entire interval is iterated. The interval quickly grows in length[6] and becomes curved.[7] When, on the n^{th} iterate, the interval intersects the ϵ ball around $\boldsymbol{b}$, we may find the δk in the allowed range, since n is now known.

One of the problems with this technique comes from the fact that we are attempting to fill two-dimensional space with a line. We will see this problem again. Another problem comes from representing the perturbation with a grid, which makes the exponential growth, necessary to quickly fill the space, a problem.

Consider that if m points representing the grid have $\|\delta \boldsymbol{z}\| = \frac{\|\delta \boldsymbol{z}_{max}\|}{m}$ spacing between each point after the first iteration, then after n iterations, the space between grid points grows roughly according to the Lyapunov number

$$\lambda^n \|\delta \boldsymbol{z}\|. \tag{1.20}$$

If the phase space is of order one, we see that

$$n \approx \frac{-\ln \delta z}{\ln \lambda} \tag{1.21}$$

is the number of iterations required for the typical distance between grid points to grow to the size of the entire phase space. Hence, exponentially more grid points are required, as the number of iterations increases, to maintain a minimum

[6]The growth rate is approximately exponential according to the largest Lyapunov exponent.

[7]Typically, we will see folds which are one of the ingredients of chaos.

useful distance between the grid points. This distance should be at least 2ϵ to resolve the intersection with $B_\epsilon(b)$. Thus we can solve for the number of grid points, m, required at the first iteration:

$$m \approx \frac{\|\delta \boldsymbol{z}_{max}\|}{2\epsilon}\lambda^n. \tag{1.22}$$

By simultaneously back-iterating the target ball $B_\epsilon(\boldsymbol{b})$ using the inverse map,[8] the situation can be improved by a factor of two. If the resulting improved time from $\boldsymbol{a}$ to $\boldsymbol{b}$ is still large, the need to store the entire grid becomes restrictive.[9] The whole grid must be stored and evolved because, all grid points are equally likely candidates to intersect the target neighborhood. We can see that the technique, though plausible, would be quite difficult in higher dimensions.

The above technique attempts to find a faster path from start to finish by storing all possible alternatives along the way. This is brute-force and requires a lot of computer memory. The following alternative technique attempts to find a balance between knowing all orbits and using the chaos to "forget" the lack of information.

For higher-dimensional dynamics, filling the phase space with all possible alternatives, to bracket the best one, is impractical. The statement follows by considering that a unit hyper box requires a cover of typically $N = \epsilon^{-d}$ balls of radius ϵ. E. Kostelich *et al.* [51] realized that for a higher-dimensional system, they could not hope to find enough orbits to choose the best. Instead, they set

[8]The ball quickly stretches along the direction of the maximum Lyapunov exponent of T^{-1}, which is in fact the minimum Lyapunov exponent along the forward orbit T.

[9]In Sec. 2.2.7, we will see an example for the standard map where, for k=1.25, we find an orbit which is n=131 steps long. For this parameter value, $\lambda \approx 1.2$ and so $\lambda^{131} \sim 10^{10}$. For an initial segment $\|\boldsymbol{z}_{max}\| = 0.1$ (this is large in an order one phase space), according to Eq. (1.22), we require $m \approx 10^{10}$ grid points for a final average intergrid spacing $\|\delta \boldsymbol{z}\| \sim 0.05$.

out a net of known paths leading to the target $\boldsymbol{b}$. The chaotic nature of the dynamics guarantees that the orbit of $\boldsymbol{a}$ will eventually wander into a neighborhood of the known paths. In this sense, the technique is similar to OGY, except that control is activated if the test orbit falls within the control window of *any one* of the points in the network's data set, compared to monitoring a single OGY window.

Kostelich's net is organized in a tree hierarchy defined so that there is a set of paths that lead to a set of paths, etc., that lead to the predetermined target point $\boldsymbol{b}$. This tree is stored as a library of known paths. The tree branches provide a set of epsilon chains leading to $\boldsymbol{b}$ (pseudo-orbits which miss being exact orbits by a small phase space error at the junctions).

The first step in building the tree, in their example, requires (arbitrarily) a path of 20 iterates, from anywhere, to near $\boldsymbol{b}$. Searching for a point from anywhere to near $\boldsymbol{b}$, rather than fixing the starting point near $\boldsymbol{a}$, tremendously reduces the difficulty of the problem. Such paths could be found by using the inverse map, when available. Alternately, one could wait for a random initial condition to eventually iterate close enough to $\boldsymbol{b}$, and then store only the last 20 iterates. This transient is a necessary part of the long pre-study typical of targeting. This 20 step path represents the first level (or trunk) of the tree.

The second level of the tree consists of 20 step paths which lead to near any one of the 20 points on the first level, and can be found as described above. Once an n^{th} level has been found, the $(n+1)^{th}$ level can be found by keeping the last 20 iterates of a randomly chosen orbit leading to any one of the points on an n^{th} branch. Of course, if an orbit leads first to an earlier branch, it should be stored in the appropriate level of the tree.

The hope is that once the tree has been built, a significant portion of

the phase space will be within only a few iterates from one of the branches. Then, just as in the OGY technique, we again rely on chaos when patiently waiting for a given real time initial condition to wander into the realm of known behaviors. In the OGY technique, the known behaviors only consist of a small box around the target point $\boldsymbol{b}$. The idea here is that if there are enough branches, then any arbitrary initial condition $\boldsymbol{a}$ will rapidly iterate close to the tree. Once there, it could be stabilized to reach $\boldsymbol{b}$.

The "thickness" of the branches (i.e. the size of the control windows around each point of the tree) depends on the size of the maximum allowed parameter perturbation δk_{max}. Once a real time orbit (a "test orbit") wanders close to any point on the tree, it can be stabilized *on the fly* onto the known path by the techniques described in Sec. 1.1.1. Specifically, the authors use an open-ended path technique Eq. (A.40) similar to the *on the fly* control described in Sec. 2.2.2.

Notice that, unlike previous techniques, Kostelich allows for a program of parameter perturbations, $\{k_i\}$, starting at the switching points between branches and continuing beyond for stabilization. The switching points and branches however are independent of the starting point $\boldsymbol{a}$. It is possible for $\boldsymbol{a}$ and $\boldsymbol{b}$ to be in regions of phase space separated by (perhaps many) "partial barriers"[10]. In such a situation, the branches of the tree may be confined to $\boldsymbol{b}$'s section of phase space, hence offer little improvement of transport from $\boldsymbol{a}$'s section of phase space. This sort of complication will be addressed in later sections of this thesis, specifically with regard to Hamiltonian maps.

In the example presented by Kostelich, the authors successfully targeted the kicked double rotor map, which is a four-dimensional version of the

[10]These structures will be described in Sec. 1.2.2.

coupled dissipative standard map. They built a three level tree consisting of 10^4 points on 500 paths leading to $\boldsymbol{b}$. Once within stabilization range of any of the 10^4 points, $\boldsymbol{b}$ is guaranteed to be within 60 iterates. Stabilization was achieved by parameter perturbations whenever the test point wandered within 0.05 of the tree.[11] They found that they needed an average of only 35 iterates to target their example $\boldsymbol{b}$, in contrast to the 10^{11} iterates they argued would be required without control.

Dimension is an important aspect in targeting. It dictates the amount of information, represented by the already observed orbits in the pre-study, required to cover a compact attractor. The dimension of the chaotic attractor is germane. For the kicked double rotor, the attractor dimension is approximately 2.8. An ergodicity assumption implies that a large number of iterates of the orbit can be described approximately as being uniformly distributed with respect to the ergodic measure.[12] Therefore, the average distance between nearest neighbors on a subset of N points on the attractor scales as $N^{-1/2.8}$. This indicates the number of points in the library required to adequately fill the space. It also describes the number of iterations required for a test orbit to wander near an N point library by considering the probability that the test orbit does not land near any one point of the N points on a given iteration.

The final technique of targeting I will describe involves dividing the global dynamics into cell maps. The objective is to learn the dynamics by putting it on a grid, and refining where detail is needed. A complete description of how to learn a dynamical system using a grid analysis can be found in the book by C. Hsu [45].

[11] The phase space of the double rotor is a $(2\pi \times \mathbb{R}) \times (2\pi \times \mathbb{R})$ 2-cylinder.

[12] Specifically, a dynamics is defined as ergodic if there exists a measure such that the phase space averages of orbits are equal to the time averages of the orbit.

Transitions rules between the cells are similar to those of a Markov tree model [71] where, instead of probabilities of transitions, only 0's or 1's, representing "yes" and "no", occupy the matrix of allowed transitions between states. Once we categorize the parameter extended dynamics,

$$\boldsymbol{z}_{n+1} = T(\boldsymbol{z}_n, \boldsymbol{k}), \tag{1.23}$$

into corresponding cell maps, we can choose a final path by using segments of the possible trajectories, for the recorded parameter values, represented by cell sequences of the cell map. We can represent an entire epsilon-chain orbit by a sequence of cells (at the switching points) where each cell represents a true trajectory.[13] Note that learning the dynamics of Eq. (1.23) on a grid does not necessarily require availability of an analytic form of the map. Again, time series embedding of observed data can replace Eq. (1.23), making the technique applicable to real systems.

In this framework, there are typically a huge number of permutations of cell orbits[14] between a given $\boldsymbol{a}$ cell and $\boldsymbol{b}$ cell. Hence, finding the optimum cell path involves more than simply choosing the fastest from a finite list. A technique of finding a "highway system" of orbits is described by E. Bradley [15], where the search is broken into a hierarchy of paths, first from the vicinity of $\boldsymbol{a}$ to the vicinity of $\boldsymbol{b}$ (the major highway), and then from a vicinity of $\boldsymbol{a}$ (and also $\boldsymbol{b}$) to the vicinity of the major highway (the secondary roads), etc., until sufficient refinement has been achieved.

This technique also requires a high overhead pre-study to completely classify the dynamics into discrete cells. Nonetheless, it has proven successful

[13] The trajectory represented by a cell is that of the point contained in its center.
[14] Allowing cycles generates an infinite number of orbits.

for targeting a variety of systems [15], including experimental chaotic circuits, an experimental pendulum, and the Lorenz system.

1.2 Transport Through Chaos

Control of chaos attempts to make use of only the available dynamics, since only small controls are allowed. This means that targeting is a transport issue when we look for pseudo orbits extending from near $\boldsymbol{a}$ to near $\boldsymbol{b}$. In this section, after a few preliminaries, I will discuss chaotic transport, to better address the targeting problem. Transport is best understood for chaotic maps of the plane, which I will explain in some detail. Further detail can be found in the book by S. Wiggins devoted to the topic [104] or in the review article by J.D. Meiss [70]. Some of the ideas are applicable to higher-dimensional maps as well. We will see that a crucial part of understanding transport across a barrier is defining the term "barrier."

1.2.1 Preliminaries and the Basic Transport Mechanism

First recall the map from Eq. (1.2), which is an example of a dynamical system. Define a map T as a function of a differential manifold back into itself. To understand transport, we will restrict the discussion to continuous dynamical systems of orientable manifolds.[15] The reason for this will become clear shortly. In addition, we will discuss orientation preserving maps, which are so defined, in terms of the tangent map $DT|_{\boldsymbol{z}}$, if $\det(DT|_{\boldsymbol{z}}) > 0, \forall \boldsymbol{z}$. We can also develop transport mechanisms for other types of maps, such as orientation reversing maps ($\det(DT|_{\boldsymbol{z}}) < 0, \forall \boldsymbol{z}$), as long as the property is consistent.

Two-dimensional transport is particularly well understood and will be

[15] Two-dimensional, orientable manifolds include the plane, the sphere, and the torus, but not the Klein bottle or the Mobius strip. (Q: Why did the chicken cross the Mobius strip? A: To get to the same side.)

the subject of most of this discussion. A two-dimensional map can, as usual, result from a flow on a three-dimensional manifold. I will start by presenting the following example. Given a Jordan curve C enclosing a region A, we want to investigate the relative orientation of forward and backward iterations of these sets (see Fig. (1.2a)). There are four basic types of iterations:

1. $T(A) \cap A$ is empty.
2. $T(A)$ is completely contained in A.
3. A is completely contained in $T(A)$.
4. $T(A) \cap A$ is nonempty and neither set is completely contained in the other.

I choose to illustrate the final variety because it is typical of the "nice" barriers we will define in the next section. For example, if there exists a fixed point $\boldsymbol{z}^*$ on C, then $T(C) \cap C \neq \emptyset$. Similarly, a fixed point $\boldsymbol{z}^* \in A$ is sufficient, but not necessary.

figure=figabcd.eps,width=6.0in

Figure 1.2. a: Jordan curve C enclosing a region A. b: The first iterates of C and A, intersect C and A respectively. c: The region $B = T(A) - T(A) \cap A$ contains all points which will enter A on one application of the inverse map. d: The region $E_x = T^{-1}(B)$ contains all the points in A which will leave A upon one application of the map, and hence will be called the "exit region." e: The "entrance region", $E_n = T^{-1}(A) - T^{-1}(A) \cap A$, contains all points which will enter A upon one iteration of the map.

To define the subset of A that leaves A on one iteration of the map, consider the first iterate of the curve, $T(C)$, enclosing $T(A)$. This is drawn in Fig. (1.2b), illustrating the fourth of the four possible orientations above. Notice in particular the region $B = T(A) - T(A) \cap A$, shaded in Fig. (1.2c). The set

B contains all those points that left A after one iteration. Alternately, it is the set that will enter A in one iteration of the inverse map T^{-1}. Thus, B defines the entrance set of T^{-1}. In this sense, we can say that the points in B cross the barrier C.

The inverse iterate of B is shown in Fig. (1.2d), and is labeled $E_x = T^{-1}(B)$. E_x is the subset of A that will leave A on one iteration of T and hence is called the exit set. The only way for an orbit initially contained in A to leave A is for an iterate of the orbit to land in E_x. The map moves all the contents of E_x outside the closed curve on each iteration.

We may similarly construct the entrance set E_n outside of C, which is defined as $T^{-1}(A) - T^{-1}(A) \cap A$. It is the set outside of C which is moved inside of C on each iteration. It is the only way in, across C.

In summary, the entrance and exit sets are defined

$$\begin{aligned} E_n &= T^{-1}(A) - T^{-1}(A) \cap A, \\ E_x &= T^{-1}[T(A) - T(A) \cap A]. \end{aligned} \tag{1.24}$$

These definitions apply to all four of the intersection types listed. The fourth is shown in Fig. (1.2d), but the other three are just as valid. In the first case, for example, if $T(A)$ is disjoint from A, then $E_x = A$. In the second case where $T(A)$ is completely contained in A, we see that $T(A) - T(A) \cap A = T(A) - T(A) = \emptyset$, and therefore $E_x = \emptyset$.

There are certainly more complicated configurations possible for $T(A)$, relative to a general set A, than are implied by the previous figures. Some of these are indicated by Fig. (1.3a). Nonetheless, we can uniquely define E_n as the set that enters A in one iteration, and E_x as the set that leaves A in one iteration. Eq. (1.24) defines the entrance and exit lobes with no limitations on

figure=fig3.eps,height=2.5in

Figure 1.3. a: A possible, more complicated iterate (and back-iterate) of the region A. The implied entrance and exit regions in this example can intersect, which simply means that some subset of points entering A will immediately exit A on the next iterate. b: This configuration of $T(C)$ is not possible because it violates continuity. c: This configuration of $T(C)$ overlaps itself, and so violates single valuedness of $T^{-1}(T(C))$.

the amount of folding possible. A configuration such as Fig. (1.3a) presents no contradictions; it simply implies that once leaving A, the subset $T(E_x) \cap E_n \cap \bar{A}$ will immediately re-enter A on the next iteration. Eq. (1.24) makes no statement regarding two iterations. Configurations such as in Figs. (1.3b) and (1.3c), which may present problems, are not possible due to violations of continuity and single-valuedness.

On certain manifolds, it is possible to describe transport across a barrier C which is not a Jordan curve. The role which the closed curve serves in the above discussion is that it divides the space in two - an inside and an outside. Hence, it makes an appropriate barrier. If a curve does not completely divide the space, transport can occur "across" the barrier by going around it (or by going the "other way" around the cylinder $S^1 \times \mathbb{R}$ to avoid an infinite line "barrier" in the case of a cylinder). The surface of a cylinder can be divided in two (a top part and bottom part) by a closed curve (a "belt" wrapped around the "waist").

The surface of any two-dimensional manifold can be divided in two by any closed curve that is homotopic[16] to the point. These are the Jordan curves. On a sphere, all closed curves are Jordan curves because the sphere is of a trivial homotopy type; it has no holes. A torus, however, has a hole. Some closed curves are not Jordan curves since they contract to the circle through the

[16]A curve $f(\boldsymbol{z})$ is homotopic to $g(\boldsymbol{z})$ iff there exist a continuous deformation of f transforming it to g.

hole. These curves do not serve as barriers since we can go the other way around the torus. Higher genus two-dimensional manifolds have even more holes and so admit more non-Jordan closed curves with similar complications. However, we saw that the cylinder admits a barrier which is not homotopic to the point. To summarize, we can state that on all two-dimensional manifolds, a closed curve homotopic to the point is Jordan, and hence is a good barrier. On a compact two-manifold, a closed curve not homotopic to the point is not a barrier, and finally, on an unbounded two-manifold, a noncontractable curve may or may not form a barrier depending on whether it goes through a hole.

The description of transport across any barrier is made by forward and backward iterating the barrier, then finding the regions bounded by C and $T^{-1}(C)$ (or C and $T(C)$) and asking "Which region crosses the barrier on the next iterate (back-iterate)?" We found "lobe-like" structures in Figs. (1.2)-(1.3) because we illustrated the situation where $T(A) \cap A \neq \emptyset$ and neither set is completely contained in the other. We will see this situation in the next section where there will typically be a fixed point $\boldsymbol{z}^*$ on C.

1.2.2 Chaotic Transport Mechanism In the previous section, we saw that transport across any appropriately chosen barrier can be characterized. In this section, we ask the question, "What are the most natural barriers in chaotic transport?" The arbitrarily chosen barriers in the previous section move upon iteration. The entire barrier typically deforms upon iterations. The situation is even worse with continued iterations. A natural choice of barriers is one that has "minimal alterations" upon iteration.

We will see below that a natural barrier can be constructed of segments of stable and unstable manifolds on a homoclinic orbit. Given a period-n point

$\boldsymbol{z}$ of Eq. (1.2), define $W^s(\boldsymbol{z})$, the stable subspace, and $W^u(\boldsymbol{z})$, the unstable subspace, as follows:

$$\begin{aligned} W^s(\boldsymbol{z}) &\equiv \{\boldsymbol{x} : \ T^{jn}(\boldsymbol{x}) \to \boldsymbol{z} \text{ as } j \to \infty\}, \\ W^u(\boldsymbol{z}) &\equiv \{\boldsymbol{x} : \ T^{jn}(\boldsymbol{x}) \to \boldsymbol{z} \text{ as } j \to -\infty\}. \end{aligned} \tag{1.25}$$

In Sec. 1.1.1, a point was defined to be hyperbolic when the tangent space at that point is decomposable as the direct sum

$$M = E^s(\boldsymbol{z}) \oplus E^u(\boldsymbol{z}), \tag{1.26}$$

where $E^s(\boldsymbol{z})$ (or $E^u(\boldsymbol{z})$) is the linear subspace of the tangent space at $\boldsymbol{z}$, spanned by the eigenvectors corresponding to eigenvalues with modulus strictly less (or greater) than one. The stable manifold theorem [85] implies that these eigenvectors can be continued to the global stable (unstable) manifolds. The Hartman-Grobman theorem [3] states that, for a diffeomorphism T^n and a small enough neighborhood U of $\boldsymbol{z}$, there is a homeomorphism between the dynamics of DT^n on $E^s(\boldsymbol{z}) \oplus E^u(\boldsymbol{z})$ and $T^n|_U$. A hyperbolic saddle point is categorized by having all of the eigenvalues λ_i of the tangent map at $\boldsymbol{z}$ such that $\lambda_i \in \mathbb{R}$, $|\lambda_i| \neq 1$, $\forall i$. Hence, points not on an eigendirection move along a hyperbola upon application of DT^n, and so do orbits close enough to $\boldsymbol{z}$, upon application of T^n. In Sec. 2.2.3, we will discuss further the stable and unstable manifolds and the extension of the concept to nonperiodic points.

A hyperbolic saddle fixed point of a two-dimensional map is shown in Fig. (1.1). It should be stressed that the smooth curves shown are not flows of a single point. Each point "jumps" upon application of the map to another location on the curve. Continuity implies that a nearby point jumps nearby.[17]

[17]It may seem paradoxical that a chaotic dynamical system can nonetheless be continuous. Sensitivity to initial conditions and exponential spreading of nearby points

Certain rules must be obeyed by such manifolds. By definition, a point on the stable (unstable) manifold remains on the manifold. Single valuedness forbids that a stable (unstable) manifold intersects itself or the stable (unstable) manifold of another point. It is allowed, however, for the stable manifold to intersect the unstable manifold. A point $\boldsymbol{p}$ on the intersection of $W^s(\boldsymbol{z}_i)$ and $W^u(\boldsymbol{z}_j)$ is called a homoclinic point if $i = j$ or a heteroclinic point if $i \neq j$. By definition, the orbit of $\boldsymbol{p}$ accumulates on $\boldsymbol{z}_i$ in forward time, and on $\boldsymbol{z}_j$ in backward time. Thus, iterates of homoclinic (heteroclinic) points are homoclinic (heteroclinic) points. The existence of one intersection implies infinitely many intersections.

As one varies the parameters, the manifolds $W^s(\boldsymbol{z}_i)$ and $W^u(\boldsymbol{z}_j)$ may intersect either transversally or tangentially. The tangent-type intersections are not generic because they are not structurally stable, but the transverse-type are, and so will be the subject of the discussion to follow.

figure=fig4.eps,width=5.5in

Figure 1.4. a: A transverse homoclinic connection at point $\boldsymbol{p}$, and a few of its iterates and pre-iterates. b: A single "lobe" between $\boldsymbol{p}$ and $T(\boldsymbol{p})$ causes an illegal orientation change from $\boldsymbol{p}, \boldsymbol{x}, \boldsymbol{z}$ to $T(\boldsymbol{p}),\ T(\boldsymbol{y}),\ T(\boldsymbol{x})$. c: The "orientation of surface" (or "signed area") of the parallelogram described by the vectors $\boldsymbol{p} - \boldsymbol{y}$ and $\boldsymbol{p} - \boldsymbol{x}$ has opposite sign to that of parallelogram $T(\boldsymbol{p}) - T(\boldsymbol{y}),\ T(\boldsymbol{p}) - T(\boldsymbol{x})$. d: Inserting one more transverse homoclinic point $\boldsymbol{q}$ yields a legally oriented image of $\boldsymbol{p}, \boldsymbol{x}, \boldsymbol{y}$. e: Here we can see that the sign of the area of the nearby parallelogram is preserved by T.

Fig. (1.4a) shows a homoclinic orbit with transverse intersection. Also shown is part of the family of points corresponding to the orbit of $\boldsymbol{p}$. The stable and unstable manifolds must intersect at each point in this family, but

seems to exclude continuity. However, continuity is a property of *single* applications of the map, and sensitivity to initial conditions describes the evolution of nearby points under *many* applications of the map.

Poincaré showed that there exists another homoclinic point $\boldsymbol{q}$ between $\boldsymbol{p}$ and $T(\boldsymbol{p})$, due to orientation preservation. Consider two arbitrary nearby points, $\boldsymbol{x}$ near $\boldsymbol{p}$ where $\boldsymbol{x}$ is on $W^u(\boldsymbol{z})$ "farther" along $W^u(\boldsymbol{z})$ before $T(\boldsymbol{p})$,[18] and $\boldsymbol{y}$, also near $\boldsymbol{p}$, but on $W^s(\boldsymbol{z})$ closer to $\boldsymbol{z}$, but again before $T(\boldsymbol{p})$. The relative configurations of $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{p}$ are drawn in Fig. (1.4b). Reading clockwise around, the order is $\boldsymbol{p}$, $\boldsymbol{x}$, and $\boldsymbol{y}$. $T(\boldsymbol{x})$ must still be farther along than $T(\boldsymbol{p})$, and likewise so must $T(\boldsymbol{y})$ occur after $T(\boldsymbol{p})$. Again, reading clockwise around, we get $T(\boldsymbol{p})$, $T(\boldsymbol{y})$, and $T(\boldsymbol{x})$, which is in violation of orientation preservation. We can see this in Fig. (1.4c), where the area of the parallelogram, described by the vectors $\boldsymbol{p}-\boldsymbol{y}$ and $\boldsymbol{p}-\boldsymbol{x}$, has opposite to the parallelogram $T(\boldsymbol{p})-T(\boldsymbol{y})$, $T(\boldsymbol{p})-T(\boldsymbol{x})$. However, we can see in Fig. (1.4d) that inserting an additional transverse homoclinic intersection at $\boldsymbol{q}$ preserves the orientation, shown in Fig. (1.4e). Hence, there must be at least one more homoclinic point $\boldsymbol{q}$.

It is convenient to choose $\boldsymbol{p}$ to be what Wiggins defines as a principle intersection point (or p.i.p.). Any point on $W^s(\boldsymbol{z}_i)\cap W^u(\boldsymbol{z}_j)$ is a heteroclinic (or homoclinic) point. Using the ordering implicit along these invariant manifolds, we can define a p.i.p. as a heteroclinic (homoclinic) point for which the stable manifold segment between $\boldsymbol{z}_i$ and $\boldsymbol{p}$ has no previous intersections with the unstable manifold segment between $\boldsymbol{z}_j$ and $\boldsymbol{p}$. These segments of the stable and unstable manifolds are called "initial segments" [27]. Iterates of p.i.p.'s are also p.i.p.'s. Both families of points, shown in Fig. (1.4d), are examples of p.i.p.'s. Starting with p.i.p.'s, non-principle intersection points arise from the stretching and folding typical with transverse heteroclinic (homoclinic) intersections. The

[18] An ordering on $W^u(\boldsymbol{z})$ is possible since the invariant manifold is one-dimensional. A point is defined as farther away from $\boldsymbol{z}$ than another in the sense of the arc length along the unstable manifold. An ordering on $W^s(\boldsymbol{z})$ can be similarly defined in terms of the arc length closeness to $\boldsymbol{z}$.

resulting "tangle" quickly generates infinitely many other families of heteroclinic (homoclinic) points which are not p.i.p.'s. More will be said about the tangling process in the next section.

In Fig. (1.5c), the shaded regions are labeled "E_x" and "E_n", describing their transport roles. These "lobes" have infinitely many (pre)images, whose end points are the (pre)images of $\boldsymbol{p}$ and $\boldsymbol{q}$.

We may now define a Jordan curve C using the unstable manifold initial segment between $\boldsymbol{z}$ and $\boldsymbol{p}$, and the stable manifold initial segment between $\boldsymbol{z}$ and $\boldsymbol{p}$, for any p.i.p. $\boldsymbol{p}$[19]. See Fig. (1.5a). There is a well defined inside and outside, for this barrier C. Eq. (1.24), defining transport across an arbitrary barrier, applies to this special choice of C.

figure=fig5.eps,width=5.5in

Figure 1.5. a: Defining the barrier by initial segments of the stable and unstable manifold between the fixed point $\boldsymbol{z}$ and p.i.p. $\boldsymbol{p}$. b: The iterate C lies largely on top of C, as much of the curve stretches over itself. c: The exit and entrance lobes E_x and E_n, which together are called the "turnstile".

The claim is that we have chosen the most natural barrier C because orbits on the manifolds stay on the manifolds. Following the discussion in the previous section, in Fig. (1.5b) we draw $T(C)$, and in Fig. (1.5c), we draw $T^{-1}(C)$. In terms of the original barrier C, we see that the shaded region E_n in Fig. (1.5c) iterates to the region $T(E_n)$ inside C (which we easily see by following the iterates if $\boldsymbol{p}$, $\boldsymbol{q}$, and the manifold segments in between).

The only alteration in the overall form of C is the "growth" of the lobes E_n and E_x upon application of T^{-1}. In this sense the choice of the barrier C is "minimal." Mackay, Meiss, and Percival coined the term "turnstile" to describe

[19]In fact, as long as $\boldsymbol{p}$ is a p.i.p., any of its iterates are just as legitimate, and the resulting entrance and exit lobes can be used to define transport.

the two lobes E_n and E_x in that they act like rotating doors, transporting area across C.

An important point of the previous section, summarized by Eq. (1.24), is that *all orbits which transport across C must be localized in $E_n \cup E_x$*. Specifically, if we choose a fixed iteration of the turnstile, then we will find in it all orbits which transport. In the next section, I will describe how the lobes stretch in forward and backward time. The choice of the p.i.p. which is most appropriate to this thesis will be the one in which localization in E_n or E_x also implies the most localization in terms of the Euclidean norm.[20] This is the iteration where the diameters of E_x and E_n are minimal. (The diameter of a set is defined by the supremum of the Euclidean norm between any two points in the set.) This localization statement will be key in finding transporting orbits (i.e., targeting), which will be discussed in the next chapter.

Studying Fig. (1.5), there is another perspective on "transport" to be made. Forgetting our barrier C for a moment, let us focus on a point in the entrance set E_n "outside"[21] of the manifold segment of $W^s(\boldsymbol{z}_i)$ between $\boldsymbol{p}$ and $\boldsymbol{q}$. The role of iterating the map is to cause that segment of $W^s(\boldsymbol{z})$ to push in (relative to C). Points outside that segment may be viewed as still outside. In this perspective, there is no transport at all; it is just an illusion of the outside punching in further and further. This description only makes use of the stable manifold. Of course, only in terms of the full barrier C can we truly describe transport across the barrier.

I conclude this section by stressing that, while the topology of a single

[20] If E_n is a "long and narrow" shaped set, nonetheless, with perhaps small measure, two points could conceivably be localized (in terms of measure) in E_n, and yet far apart in terms of norm.

[21] Of course, a curve segment is not enough to define a barrier; a fully closed curve is required.

entrance and single exit lobe which do not intersect, is implied by the figures of this section, the situation is by no means exclusive[22]; defining the exit and entrance lobes across a barrier according to Eq. (1.24) avoids any such complications.

1.2.3 The Homoclinic Tangle What is the long-term fate of C in forward and backward time, and what is the fate of the points in E_n and E_x? These are the questions we address in this section. Their answers will lead us to horseshoes, the prototypical example of chaos.

In terms of the simple two p.i.p. family generated by $\boldsymbol{q}$ and $\boldsymbol{p}$ above, we see that the arc length between $\boldsymbol{q}$ and $\boldsymbol{p}$ along $W^s(\boldsymbol{z})$, labeled $\ell(W^s_{[\boldsymbol{q},\boldsymbol{p}]})$, must eventually (though not immediately) shrink upon repeated applications of the map, as the two points eventually accumulate at the fixed point. The arc length at time n is the line integral of the n^{th} iterate of $W^s_{[\boldsymbol{q},\boldsymbol{p}]}$. Likewise, the arc $W^u_{[\boldsymbol{q},\boldsymbol{p}]}$ iterates with $\boldsymbol{q}$ and $\boldsymbol{p}$. Hence, the curve $W^u_{[\boldsymbol{q},\boldsymbol{p}]} \cup W^s_{[\boldsymbol{q},\boldsymbol{p}]}$ is a dynamically varying boundary of $T(E_n)$.

In the case of area preserving maps, so defined when $\lambda_u = \lambda_s^{-1}$, the area of a region is constant upon iteration. So while $T^n(E_n)$ may stretch exponentially, the area is preserved for all time, resulting in a long and narrow lobe for large n. Stretching is one of the main components that can result in chaos.

In the area preserving case, it is easy to see that the finite area of region A, bounded by C, cannot completely contain all future iterates of E_n. There is

[22]More intermediate p.i.p.'s other than just the required transversal type point $\boldsymbol{q}$ are possible. They may be of either transversal or tangential type. We do not define the "lobes" so formed as *the* exit or entrance lobes unless they contain the entire entering or exiting set.

a time m when

$$\sum_{i=1}^{m} \mu(T^i(E_n)) \geq \mu(A), \tag{1.27}$$

i.e., the first time

$$m \geq \frac{\mu(A)}{\mu(E_n)}. \tag{1.28}$$

In terms of transport, some of the points in E_n which enter A must leave A by the m^{th} iterate, implying that there exists an $r \leq m$ such that $T^r(E_n) \cup E_x \neq \emptyset$. Almost all of the points must eventually leave. Once this intersection occurs, a new family of homoclinic points is implied. Considering the history of the lobe E_x, which also becomes long and narrow (as $n \to \infty$), we see that a homoclinic point is implied each of the times m and $-n$ that $T^m(E_n) \cap T^{-n}(E_x) \neq \emptyset$. The segment $W^u_{[\boldsymbol{q},\boldsymbol{p}]}$ accumulates at $\boldsymbol{z}$ as $n \to -\infty$, and $W^s_{[\boldsymbol{q},\boldsymbol{p}]}$ accumulates at $\boldsymbol{z}$ as $n \to \infty$. Of course, a "new" family of intersections implies infinitely more intersections as the homoclinic point iterates in forward and backward time. This is the "homoclinic tangle."

The horseshoe construction implies a set which remains trapped inside the region A for all time. S. Smale [95, 104] showed that for a diffeomorphism T with a transverse homoclinic point $\boldsymbol{p}$, there exists $m > 0$ such that the composition map T^m has an invariant Cantor set $\Lambda \in A$. He also showed that there exists a conjugacy

$$h : \Lambda \to \Sigma \tag{1.29}$$

such that

$$h \circ T^m|_\Lambda = \alpha \circ h. \tag{1.30}$$

The conjugacy is with the dynamics of the Bernoulli shift map α on the space of bi-infinite sequences of countably many symbols σ_i. In the simple horseshoe,

we let

$$\sigma_i = 0 \text{ or } 1. \tag{1.31}$$

By now, the horseshoe construction for the homoclinic tangle and corresponding symbol dynamics is quite standard, which I will only briefly review. A bi-infinite sequence of the two symbols consists of two infinite sequences $\{\sigma_i\}_{i=0}^{\infty}$ and $\{\sigma_j\}_{j=-1}^{-\infty}$ appended at their starting members, where the second sequence is written right to left as follows:

$$\sigma = ...\sigma_{-2}\sigma_{-1}.\sigma_0\sigma_1\sigma_2... \tag{1.32}$$

The "decimal point" can be thought of as marking "now" from which the bi-infinite sequence is to be read. The action of α is to shift the period to the right,[23] bringing new symbols into attention. The norm on this sequence space $\|\cdot\| = \sum_{i=-\infty}^{\infty} \frac{\sigma_i}{2^{|i|}}$ serves to weight the symbols near the period (middle).[24] Two symbol sequences are close if they agree over their middle portions. The norm implies a topology on Σ, where a neighborhood in Σ is defined by $2k+1$ symbols centered around the decimal. The map α serves to shift these known symbols out of focus and to bring "beyond accuracy", or "unspecified" symbols, into focus. These unspecified symbols are as random as a coin toss.[25] Hence, α serves as a "forgetting" process.

It is not difficult to show that two properties which define "chaos", sensitivity to initial conditions[26] and transitivity,[27] both hold for the Bernoulli

[23] Alternately, we can imagine holding the period fixed and shifting the symbols to the left, bringing new symbols into focus from the right.

[24] In the case of an area preserving map, if we define the norm using λ instead of 2, the norm can serve as the conjugacy h between Λ and Σ.

[25] "Heads" and "tails" also define a legitimate symbol space.

[26] A map T on a metric space is said to have *sensitive dependence on initial conditions* if there is an $r > 0$ such that, given a point $\boldsymbol{x}$ and arbitrary $\epsilon > 0$, there is a point $\boldsymbol{y}$ such that $d(\boldsymbol{x}, \boldsymbol{y}) < \epsilon$ and a time k when $d(T^k(\boldsymbol{x}), T^k(\boldsymbol{y})) \geq r$ [85].

[27] A map T is *transitive* if there exists a dense orbit.

shift on bi-infinite sequences of two symbols. Any dynamics conjugate to $\alpha|_\Lambda$ can be defined as "Bernoulli," which is one of the stronger notions of chaos, though some only define a chaotic system as Bernoulli if there exists a conjugacy to the shift map valid for a set of nonzero measure (such as the Arnold cat map).

The horseshoe may be constructed for Fig. (1.5) by drawing a thin curved strip S over $W^s_{[\boldsymbol{z},\boldsymbol{p}]}$ as shown in Fig. (1.6). As $\boldsymbol{p}$ iterates closer to $\boldsymbol{z}$, it drags the strip with it. Meanwhile, the point $\boldsymbol{s}$, defined as the intersection $W^u_{[\boldsymbol{z},\boldsymbol{p}]} \cap S$, marches away from $\boldsymbol{z}$. Define m as the first time that $T^m(\boldsymbol{s})$ is *after* $\boldsymbol{p}$. By time m, the short side of the strip has stretched and folded over to the strip $T^m(S)$ along $W^u_{[\boldsymbol{z},\boldsymbol{p}]}$ which intersects S by construction. Here we see the stretch and fold, which can be thought of as the ingredients for chaos.

figure=fig6.eps,width=3.5in

Figure 1.6. Constructing a horseshoe on a homoclinic orbit. The strip S contracts along the stable manifold, and expands along the unstable manifold to the shorter, wider strip $T(S)$. By the m^{th} iterate, the point $T(s)$ has passed p; the long and short sides of the strip $T^m(S)$ are reversed from the long and short sides of the original strip S. The invariant sets H_0 and H_1 are the first steps in generating the invariant Cantor set Λ.

Define

$$H_0 = T^m(S) \cap S \tag{1.33}$$

at $\boldsymbol{p}$ and

$$H_1 = T^m(S) \cap S \tag{1.34}$$

at $\boldsymbol{z}$, and define

$$H = H_0 \cup H_1. \tag{1.35}$$

By inspection, we see that the invariant set of T^m is contained in $T^m(H) \cap H$, which defines two vertical strips in H_0, and the two vertical strips in H_1.

Similarly, the invariant set of T^{-m} is $T^{-m}(H) \cap H$, which forms two horizontal strips in H_0 and H_1. Define

$$\Sigma = \bigcap_{i=-\infty}^{\infty} T^{im}(H). \tag{1.36}$$

Σ is the invariant set of the horseshoe,[28] which we see is the product of two Cantor sets, one in forward and one in backward time. For a thin enough strip S, the invariant set is hyperbolic [85, 57].

The "address" of a point in "H" can be labeled ".0" if it is in H_0 or ".1" if it is in H_1. On iteration, the point (say it is .0) lands in either H_0 or H_1, and hence is labeled ".00" or ".01", that defines which vertical strip in H_0 contains the point. Similarly, the address to the left of the decimal determines in which square the point lands, H_0 or H_1.

The construction directly implies that h is onto, and thus a semiconjugacy $h : \Lambda \to \Sigma$. By showing that $T^m|_\Sigma$ has a hyperbolic structure (see [85, 95]), it can be seen that h is also one-to-one. $T^{-m}|_\Sigma$ causes two horizontal strips to expand to either fill H_0 or H_1. The $2n + 1$ symbols near the decimal in Λ have a one-to-one correspondence with approximately rectangular neighborhoods in Σ. The dynamics $T^m|_\Sigma$ expands the strip until it fills H which is similar to the forgetting process in $\alpha|_\lambda$.

figure=homoc5.ps,height=5.5in

Figure 1.7: A "typical" homoclinic tangle.

It is possible to construct a horseshoe explicitly for the Hénon map, or for the standard map when $k > 2\pi$ (see [57]). Horseshoes can also be constructed

[28] The Smale horseshoe is so named because *the horseshoe* is constructed by stretching and folding a square into horseshoe shapes (again and again and...). The process is perhaps more akin to building a Japanese Samurai sword whose building process includes thousands of stretches and folds.

for heteroclinic cycles. Other "grammars", besides the simple left shift[29], on many symbols can also be useful in certain situations.

In terms of transport, the horseshoe makes a disappointing model for two reasons. Typically, one may be interested in the transport of more than a measure zero set of points. More seriously, the horseshoe models those points invariant to the horseshoe, i.e., those points which never transport out of the horseshoe set. Transport within Σ is completely described within the horseshoe, but no more. Given a complicated transport problem from $\boldsymbol{a}$ to $\boldsymbol{b}$, where only a long, convoluted, heteroclinic connection may exist, one may be successful in finding a complex grammar rule on a long list of symbols, if $\boldsymbol{a}$ and $\boldsymbol{b}$ happen to be in some invariant set of the dynamics. But, in general, only heteroclinic cycles are homeomorphic to the horseshoe, and hence have a reasonably easy-to-find symbol dynamics.

1.2.4 Transport Localization and Control The lesson learned is that transport across a heteroclinic or homoclinic barrier between $\boldsymbol{a}$ and $\boldsymbol{b}$ is localized to E_n and E_x. For a concrete example, I choose the much studied standard map

$$\boldsymbol{z}' = \begin{pmatrix} y' \\ x' \end{pmatrix} = T(\boldsymbol{z}) \equiv \begin{pmatrix} y - \frac{k}{2\pi}\sin(2\pi k x) \\ y - \frac{k}{2\pi}\sin(2\pi k x) + x \end{pmatrix}, \tag{1.37}$$

which has many complex and intriguing properties, some to be discussed later in Sec. 2.2.4. For now we need only observe that there is a periodicity in x, with period 1. The phase space is the cylinder.

We will be studying transport from a region near $\boldsymbol{a} = (x_a, y_a) =$

[29]The Bernoulli left shift grammar on two symbols can be described by the directed graph $0 \leftrightarrow 1$ which is equivalent to the 2×2 identity transition matrix. Other grammars on n symbols have directed graphs describable by more general $n \times n$ identity matrices.

$(0.5, 0.0)$ to near $\boldsymbol{b} = (x_b, y_b) = (0.5, 1.0)$, for $k = 1.25$. The prescription for this algorithm is as follows. First, we randomly choose an initial condition $\boldsymbol{z}_0^1$ from a small box around $\boldsymbol{a}$. (Define the "a box" as $\{\boldsymbol{z} : \|\boldsymbol{z} - \boldsymbol{a}\|_1 \leq l\}$ where here we choose $l = 0.01$.) Then iterate until $\boldsymbol{z}_n^1 \in$ "b" box, recursively defining $\boldsymbol{z}_i^1 = T(\boldsymbol{z}_{i-1}^1)$. We expect there to be such an n for at least some of the initial conditions in the a box, because numerical evidence indicates that both the a box and b box intersect a connected chaotic region.[30] If the randomly chosen initial condition never, in fact, reaches the b box, or simply takes too long (defined by $n \geq 10^7$ in this study), then we simply choose another initial condition, until a satisfactory $\boldsymbol{z}_0^1$ is found.

This orbit will be used as the primary list, and each of its entries are center points. We then choose a "common radius" $\delta > 0$. The trick is to find the points on the primary list which are in regions important to the transport. We define importance by the observation that other transporting orbits also pass within δ of the primary passing $\boldsymbol{z}_i^0$.

Now we find a second transporting orbit, again by choosing $\boldsymbol{z}_0^2 \in a$ box and iterating until $\boldsymbol{z}_m^2 \in b$ box. The points $\boldsymbol{z}_i^1$ of the primary orbit are arbitrarily treated as centers for this localization test. In turn, we compare each point $\boldsymbol{z}_i^1$ to every point on the second list $\boldsymbol{z}_j^2$. If

$$\|\boldsymbol{z}_i^1 - \boldsymbol{z}_j^2\| > \delta \quad \forall j, \tag{1.38}$$

then $\boldsymbol{z}_i^1$ is removed from the primary list. This is repeated for all i, which leaves a somewhat restricted primary list.

The process is performed again for 3^{rd} and 4^{th},...,etc., randomly chosen transporting initial conditions. Any primary point $\boldsymbol{z}_i^1$ is removed if it is not

[30] There must be a heteroclinic connection.

within δ for some $\boldsymbol{z}_j^k$ on every list for $k = 2, 3, 4, ...$

We continue to restrict the primary list until comparing it to new transporting orbits no longer removes any $\boldsymbol{z}_i^1$. For a pre-chosen δ, this shortens the original transporting orbit until we are left with just a handful of points in regions, which, by construction, all the tested orbits must have occupied while making the a box to b box trip.

The number of "common regions" left on the primary list depends on the value of δ chosen. If δ is too small, then the entire primary list will evaporate. There is a threshold δ_1, the minimum value of δ, which will cause the list not to evaporate, leaving at least one important transport region. Thus, δ_1 represents a "radius" of these important regions. Likewise, there exist other threshold values δ_2, δ_3,..., etc. which cause more and more centers on the primary list to survive comparison with other transporting orbits. Given n, we are able to find δ_n by a bisection scheme.

Fig. (1.8) shows just such a region. We can observe what appear to be bands of various lobes which intersect through this main lobe in what we know to be the classic rainbow-like cantor structure.

figure=clust2.eps,height=2.0in

Figure 1.8. A "common lobe" for the standard map. A region which 2000 different orbits had in common while transporting from near $\boldsymbol{a} = (0.5, 0.0)$ to near $\boldsymbol{b} = (0.5, 1.0)$. Displayed are the 2000 points in the common lobe, each from a different orbit.

Other studies of the standard map seem to indicate that our common region must correspond to the turnstile through a cantorus [70]. The cantori with the smallest turnstiles (which can be calculated by an action principle [61]) are those with golden mean frequencies.[31] We observe empirically this exact

[31] See Sec. 2.2.4 for more discussion of the phase space structures of the standard map.

phenomenon. The one lobe shown corresponds to the $\frac{1}{\gamma}$ golden mean frequency. With $n = 2$, the second lobe we found corresponds to the $\frac{1}{\gamma^2}$ frequency. These are the most important barriers to transport. Note that we find the lobes in a "minimal diameter" state. Iterates of the lobes are just as valid for transport considerations, yet even though all iterates of the lobe have the same area, they become long and narrow under forward and backward iterations. Hence, they become increasingly difficult to resolve on a computer. The minimal diameter state is most natural in this sense.

Other important barriers (low area turnstiles) correspond to other "noble frequencies."[32] We managed to verify the hypothesis that our common lobes appear to correspond to the noble lobes up to $n = 6$.

An important point of this experiment is that it can be used as a tool to learn about transport in the absence of a good analytic model. Not much was assumed. It is valid to reconstruct the important barriers in a time series embedding representation of a dynamics where no model is available. It also may be useful to test for localization of transport for higher-dimensional transport as well.

[32]The golden mean, or $\gamma = [1, 1, 1, ...]$ in continued fraction notation, is the "most" noble number and has the the most persistent KAM curves (the last to become cantori). The 1's in the tail of the continued fraction make γ hard to approximate by a rational number, which is responsible for the persistence of the frequency. Any number with all 1's in the tail of its continued fraction expansion is defined as noble and satisfies a Diophantine condition. The nobles tend to have particularly persistent KAM curves. See Chapter 6 for more details on the relationship between the breakup of barriers and Diophantine frequencies.

CHAPTER 2

CONTROLLING CHAOS THROUGH RECURRENCE

2.1 Monitoring the Turnstiles

A major theme of the previous chapter was that transport across a barrier is localized to the barrier's turnstiles. Before presenting my technique to make use of this fact, I will discuss the following example to demonstrate the power in the statement of localization and its pertinence to control.

As an example, let us take the area-preserving Hénon map

$$\boldsymbol{z}_{i+1} = T(\boldsymbol{z}_i) = \begin{pmatrix} x_i \cos\alpha - y_i \sin\alpha + x_i^2 \sin\alpha \\ \\ x_i \sin\alpha + y_i \cos\alpha - x_i^2 \cos\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} x_i \\ \\ y_i - x_i^2 \end{pmatrix}, \tag{2.1}$$

where we write $\boldsymbol{z}_i = (x_i, y_i)$. We may interpret this mapping according to Moser [73] as the composition of a rotation and a shear. This map is perhaps the prototypical example for its conjugacy to the horseshoe, whose stretch and fold dynamics are visibly apparent and proven in [57].

Let us define our barrier by the stable and unstable manifold segments from the hyperbolic fixed point $\boldsymbol{z}$ to their intersections at the p.i.p. $\boldsymbol{p}$, displayed in Fig. (2.1a). This is similar to the barrier shown in Fig. (1.5a). We have seen that it is possible to nonlinearly continue the unstable and stable directions to the full manifolds $W^u(\boldsymbol{z})$ and $W^s(\boldsymbol{z})$ by iterating a fine grid of points started near $\boldsymbol{z}$ on the eigenvectors of the stability matrix of Eq. (2.1). By doing this, we can find the two p.i.p.'s $\boldsymbol{p}$ and $\boldsymbol{q}$. Actually, there are two potentially serious complications to accurately finding $\boldsymbol{p}$ and $\boldsymbol{q}$ as I have drawn them. First, in

order to find the "symmetric" intersection at $\boldsymbol{p}$, I shoot each manifold at the symmetry line $S : (x, y) \rightarrow (y, x)$ through which the dynamics are identical. This avoids the need to address the issue of balancing time to guarantee a p.i.p. Second, the shooting algorithm, which finds the parameterization t on $\boldsymbol{z} + tf_s$ (or $\boldsymbol{z} + tf_u$) which iterates to $y = x$, has infinitely many zeroes. At least one zero exists for each decade on the stable (unstable) direction since if $\boldsymbol{z} + t_0 f_s$ (or $\boldsymbol{z} + t_0 f_s$) is a zero, then so are $\boldsymbol{z} + \lambda_s^n t_0 f_s$ (and $\boldsymbol{z} + \lambda_u^{-n} t_0 f_u$) $\forall n$. This is only one family of zeros. Infinitely many zeros exist in a given λ_s (λ_u) decade as well, corresponding to nonprincipal intersection points. I mention this complication because we will see it again in subsequent sections.

Between the two p.i.p.'s, I represent the exit lobe E_x by a grid of points on $W^s(\boldsymbol{z})$ and $W^u(\boldsymbol{z})$ ordered sequentially between $\boldsymbol{p}$ and $\boldsymbol{q}$. This grid representation of E_x is drawn in Fig. (2.1b).

figure=snap.eps,height=3.5in

Figure 2.1. a: The homoclinic tangle of the Hénon map. Shown are the fixed point $\boldsymbol{z}$, the two p.i.p.'s $\boldsymbol{p}$ and $\boldsymbol{q}$, and the exit lobe E_x. b: The exit lobe E_x of the Hénon map. c: Controlled dynamics. We reflect a point $\boldsymbol{z}_i$ through the symmetry $S : y = x$ into the entrance lobe whenever it enters the exit lobe. Hence, no escape is allowed.

The control objective is to prevent transport across the barrier by monitoring *only* the small region E_x. The "control" action[1] will be to kick the point $\boldsymbol{z}_i$ back inside the barrier before it escapes, whenever $\boldsymbol{z}_i \in E_x$. In this case I choose to do this operation by reflecting $\boldsymbol{z}_i$ through the symmetry line S, which lands it in the entrance lobe. Therefore, $T(S(\boldsymbol{z}_i))$ is back inside the barrier.

In Fig. (2.1c), we see the orbits of several initial conditions under the

[1]I put quotes around the word "control" because this control typically has a large perturbation magnitude. It therefore is no surprise that we can drastically alter the dynamics. I will address smaller controls in the subsequent sections.

influence of the altered Hénon dynamics

$$\boldsymbol{z}_{i+1} = \begin{pmatrix} T(S(\boldsymbol{z}_i)) & \text{if } \boldsymbol{z}_i \in E_x \\ T(\boldsymbol{z}_i) & \text{otherwise} \end{pmatrix}. \tag{2.2}$$

An interesting consequence of the symmetric nature of both the dynamics and the control is that new periodic orbits appear.[2] As an aside, let us recall the definition (see [57]) that a map T has a symmetry S iff S is an orientation reversing involution such that

$$S^2 = (TS)^2 = I. \tag{2.3}$$

It follows that

$$T^{-1} = STS^{-1}. \tag{2.4}$$

So we see that in fact our particular choice of control is equivalent to inverting the map for the image flipped through the symmetry line. All those points which eventually escape the barrier under the Hénon map, many of which are shown in Fig. (2.1c), are now periodic under Eq. (2.2). All those points which are forever bounded inside the barrier, including the quasi-periodic invariant circles, the regions they bound, and other chaotic orbits which may wander indefinitely without escaping, remain aperiodic because Eq. (2.2) allows them to follow their original Hénon dynamics.

Of course, it may not always be desirable to apply such a large control. Fortunately, such large controls are not necessary. The size ϵ of the perturbation to $\boldsymbol{z}_i$ required to kick $\boldsymbol{z}_i$ out of E_x depends on the shape of the set E_x. The particular iteration of $\boldsymbol{p}$ and $\boldsymbol{q}$ chosen makes E_x as "evenly proportioned" as possible. The lobes $T^n(E_x), n > 0$ have the same area, but become long and

[2] If we fold the paper through the line $y = x$, we have still captured the whole orbit.

narrow under the influence of the unstable and stable manifolds respectively. Likewise, the lobes in backward time $T^{-n}(E_x), n > 1$ are narrow and long. *To stop transport* across the barrier, it is just as valid to *monitor any of the lobes* $T^{-n}(E_x)$. The key point is that the original required perturbation, ϵ, diminishes approximately proportionally as

$$\lambda_s^n \epsilon, \tag{2.5}$$

providing that we choose the direction carefully.[3] Hence we see that it is possible to prevent transport across the barrier with an arbitrarily small control as long as we are willing to monitor a sufficiently early pre-iterate of E_x.

The control I have described so far was to prevent transport, and so I call it a "negative" control. It is also possible to encourage transport across a barrier, which is the goal of targeting. In this case, we wish to perturb points into E_x. If we wish to do so with a small control ϵ, then we can imagine the controllable set as an ϵ-thickened boundary region around E_x (and inside the transport region). An improvement to the scheme would be to keep several ϵ-thickened pre-iterates $\{T^{-j}(E_x)\}_{j=0}^n$. This description of targeting is similar in character to the web built by Kostelich *et al.* [51].

A complete charting of transport by keeping iterates of the lobes (on a grid) of the barriers in phase space seems promising. It was my original approach to the targeting problem in the standard map. I had planned to put all the lobes of all the important barriers on a grid. This would put targeting squarely in the intersection with pattern matching and artificial intelligence. I targeted the standard map following this type of approach, the discussion of which can be

[3]The "round" lobe becomes long and narrow under the influence of the strong growth and contraction directions. The small control Eq. (2.5) is sufficient only if it is applied in the stable direction (the unstable direction of the inverse map) in which the lobe becomes thinnest.

found in Appendix B.1.

There is a fatal flaw with this scheme. The same Lyapunov contraction which allowed us to use arbitrarily small controls in Eq. (2.5) makes it impossible to store the lobe on an adequately fine grid on a computer. The number of grid points N required to preserve a *maximum* intergrid distance of l (on the unstable sides of the lobe) scales approximately proportionally as

$$N \sim \frac{\lambda_u^n}{l}. \tag{2.6}$$

Meanwhile, two opposite ends of the lobe rapidly get so close, the grid seems to lie on a single line. It becomes almost impossible to distinguish inside the lobe from outside the lobe.[4] Another problem is that actually finding the turnstiles requires special knowledge about the dynamics, either from an enormous pre-study by the technique of common lobes, described in Sec. 1.2.4, or by having the map in analytic form (which nonetheless also requires a large computational pre-study).

In the rest of this chapter, I present an alternative approach which automatically uses the switching points (the turnstiles) without explicitly finding them, and uses the Lyapunov error expansion only to our advantage.

2.2 Controlling Chaotic Transport Through Recurrence

We have seen that transport times for a chaotic system are highly sensitive to initial conditions and parameter values. In the subsequent sections of this chapter, I present a new technique to find rough orbits (epsilon chains) that achieve a desired transport rapidly and which can be stabilized with small parameter perturbations [12]. The strategy is to build the epsilon chain from

[4]No matter what the grid density is in fact, there is always the possibility of inside-outside classification errors for a point near the boundary due to the representation of a curved lobe boundary by straight line segments.

segments of a long orbit; the point is that long orbits have recurrences in neighborhoods where faster orbits must also pass. The recurrences are used as the switching points between segments. The resulting epsilon chain can be refined by gluing orbit segments over the switching points, provided that a local hyperbolicity condition is satisfied. As an example, we show that transport times for the standard map can be reduced by factors of 10^4. The techniques presented here can be easily generalized to higher dimensions and to systems where the dynamics is known only as a time-series.

2.2.1 Introduction This chapter addresses the problem of time-optimal control, or targeting as stated by (1.15)-(1.17). This, I remind the reader in brief, is the technique to steer a dynamical system from near an initial condition $\boldsymbol{a}$ to near a target point $\boldsymbol{b}$ in the shortest possible time. The major difficulty in targeting is to find a scheme to decide when and where judiciously chosen perturbations should be applied. All of the techniques discussed in Chapter 1.1 involve directly looking for these paths and suitable switching points. This can be impractical, we saw in the previous section, because the switching points can become effectively invisible in a computer calculation due to the stretching typical in chaotic dynamics. This problem becomes increasingly important when even the fast orbit is not particularly short and incorporates many switching points. The standard map Eq. (1.37) is just such a case. The technique I present avoids this issue with the alternative approach of letting the short path reveal itself as the shadow of a longer orbit.

In order to best choose the orbit segments of the epsilon chain, we would like to know where to most efficiently switch between the segments. I present here an alternative approach for building and managing a library of numerically

known orbit behaviors so that this information can be quickly accessed to build fast-transporting epsilon chains. Our main result is that by studying the path of a non-optimal orbit that, nonetheless, eventually achieves the desired target objective, the switching points reveal themselves.

In Sec. 2.2.2, I argue that recurrences are common in slow orbits, and that these should be tested as switching point candidates. The resulting epsilon chain can by refined, as we show in Sec. 2.2.3, by gluing patches across the switching points (points where orbit segments are joined). This requires finding the stable and unstable directions along the original orbit, and provides a patch orbit segment that:

1. Skips the recurrent loop, often representing of the bulk of the orbit's length.
2. Converges to the original orbit backward in time.
3. Converges to the original orbit forward in time from the point of recurrence.

The patch size can be chosen to meet the control saturation bound. Hence, we can effectively pick and choose desired segments of a slow orbit using hyperbolicity to our advantage to leverage away the error upon gluing in an orbit patch. Gluing has been used, for example, in proving the shadowing theorem for Axiom A systems [14] as well as in other contexts [98].

The obvious advantage here is the possibility of constructing fast orbits between any two points in accessible phase space. By following an arbitrary orbit for a long time and recording its local stability properties, any two points near the observed dynamics can now be reached by an epsilon chain constructed

from segments of the observed dynamics. A chaotic orbit will cover all of its accessible phase space and so will have most starting and target points close to the accessible set somewhere within its length. We propose that this technique is also applicable in the case where we only have an approximate model of the dynamics formed by a time series of data from a real world system, and from which we can make local predictions according to the work of Farmer and Sidorowich [29]. This is possible since no inverse image of the map is necessary for our method.

We use a local linear controller at each step of the predicted orbit to diminish the effects of modeling error and system noise. Local linear controllers have been demonstrated using accessible dynamic parameters for a number of chaotic systems, and have also been shown to be effective even for dynamics specified only by time series. An effective method is to use accessible parameters to cause the image of an initial condition to have no component on the unstable manifold of the target point [52, 79]; hence, knowledge pertaining to the map's parameter derivatives and the unstable directions is required. More traditional "pole placement" techniques yield much the same result [86].

In Sec. 2.2.4, there is a demonstration of the method for the standard map, which has notoriously slow transport. We investigate transport distributions before and after control. We also investigate the hyperbolicity of our trajectories before and after control by computing the distribution of angles between the stable and unstable manifolds [53].

2.2.2 Chaos and Recurrence In this section we will discuss the difference between an optimal trajectory and a non-optimal trajectory of dynamics arising from a given map. We write a map in the form of Eq. (1.2).

which could have been derived from the continuous time flow of a differential equation by Poincaré section. We demonstrate rather general conditions under which a non-optimal path has a nearby path that reduces the time-optimal cost function.

Our problem is to find an initial condition $\boldsymbol{z}_a$ near the starting point $\boldsymbol{a}$, and a control strategy consisting of a set of parameter values $\{k_i\}$, that will cause $\boldsymbol{z}_a$ to iterate near the target point $\boldsymbol{b}$ as quickly as possible.

figure=sm.eps,height=4.0in

Figure 2.2. A phase space portrait of the Standard map for the range $0 \leq x < 1$ and $0 \leq y < 1$. The o's are centered on the 131 step path between $\boldsymbol{a}$ and $\boldsymbol{b}$ constructed by cutting the recurrences from a 80307 step orbit. The $\frac{1}{\gamma}$ and $\frac{1}{\gamma^2}$ golden mean cantori are represented by the gray squares in the middle region of the picture. The point $\boldsymbol{a}$ is located at $(0.5, 0.0)$ on the $(0, 1)$ hyperbolic point, and $\boldsymbol{b}$ is located at $(0.5, 1.0)$ on the $(1, 1)$ hyperbolic point.

Fig. (2.2) displays the orbit of such a control solution. Thus, we wish to minimize a cost function I in Eq. (1.15) subject to Eqs. (1.16) and (1.17), known as the minimum time control problem.

In general, the minimum occurs not at a fixed value of k, but for a program of parameter values. Knowing when and where to vary k leads us to consider whether there might be regions of phase space through which transport must occur. These regions are analogous to hub airports used in deciding which trajectory an airplane takes between L.A. and Boston. Finding the actual route seems hopeless at first, until we realize that there are just a few possible switching points: e.g., Denver, Chicago, and Houston. With this sort of model, we reduce the infinite-dimensional search in all of "phase space" for places to program k to a few small switching regions. Targeting can then successfully be performed between a finite, hopefully complete, set of such regions. The problem then

becomes just one of permutations among the fastest orbits between switching regions; but it still may be intractable if the number of switching regions is high and there is no obvious order of their importance. See Appendix B.1 for further discussion about this sort of approach.

To find these switching regions seems at first to require detailed knowledge of the transport properties of the chaotic system; yet these are only understood for the case of two-dimensional mappings (see Chapter 1.2). The solution is that transport between two regions separated by a homoclinic orbit occurs by landing in the exit lobe delineated by intersections of stable and unstable manifold segments of the homoclinic orbit. Thus, a transiting orbit must always have a point in the lobe between the two regions, regardless of the orbit's complexity. The lobes are examples of switching regions. The most efficient transporting orbit lands on a subset of the exit lobe that does not intersect the lobe again before intersecting the target $\boldsymbol{b}$. The inefficient orbit will in fact recur in the exit lobe. A complete description of the transport may even be reducible to a shift on a set of symbols such as the Smale horseshoe example [95, 104] which we will discuss as a model for our targeting algorithm. Unfortunately, using such a description to quantitatively define transport between $\boldsymbol{a}$ and $\boldsymbol{b}$ requires knowing the stable and unstable manifolds that delineate the important switching lobes. In addition, the geometry for higher-dimensional phase space is not yet well understood.

Even though the lobe structure is special for the case of two dimensions, there is an important aspect of it that has much wider application: recurrence.

Lemma 2.2.1 (Poincaré) *Let Eq. (1.2) represent a continuous map on a compact, finite-dimensional phase space* $T : \Omega \to \Omega$*, with metric* $\rho(\cdot)$*. Given an*

initial condition $\boldsymbol{z}_0 \in \Omega$ *and a* $\delta > 0$*, then there exists an* $m \geq 0$ *such that there are infinitely many times* q, p *where* $q > p$*,* $q \geq m$ *and such that the orbit of* $\boldsymbol{z}_0$ *will recur with itself at these times to within* δ*. Hence,* $\rho(T^q(\boldsymbol{z}_0) - T^p(\boldsymbol{z}_0)) < \delta$*.*

The reader will note that this lemma is similiar to the Poincaré recurrence theorem [83]. The proof of the lemma is quite simple, using the pigeonhole principle. Assume that the lemma is false. Recall that if we cover a compact set with δ balls $\{B_\delta(\boldsymbol{z})\}$, then we may take a finite subcover $\bigcup_{j=1}^{m} B_{\delta,j} \supseteq \Omega$. By assumption, $T^i(\boldsymbol{z}_0)$ must each lie in a distinct ball of the subcover. However, if $i = q - 1$ and $q \geq m$, then there are no balls left to accommodate another iterate; all the pigeonholes are filled. Hence the q^{th} iterate must fall in an already occupied p^{th} ball, indicating a recurrence. We see that there must be infinitely many such instances if we shift $t = 0$ to $t = q$. $\square$

Now we consider what this implies in terms of minimum-time control. If $\boldsymbol{z}_a$ is in a hyperbolic set and if the recurrence distance δ between $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i+s}$ is small enough, hyperbolicity implies that there exists a real orbit that converges to that of $\boldsymbol{z}_i$ backward in time and converges to that of $\boldsymbol{z}_{i+s}$ forward in time. Thus our original orbit could not have been time-optimal since the loop $\{\boldsymbol{z}_{i+1}, ..., \boldsymbol{z}_{i+s-1}, \boldsymbol{z}_{i+s}\}$ only serves to increase $I(k)$. We can only determine if a given recurrence is in fact close enough on a case-by-case basis by successfully constructing a patch of the shadow orbit. A technique to cut the loop and re-glue using an orbit segment is described in Sec. 2.2.3 below. Such a construction, when successful, monotonically decreases the value of I. Since the patched segment asymptotically converges to the original orbit, we can also satisfy the constraints Eqs. (1.16) and (1.17) as well.

The important point here is that it is very difficult to find fast orbits

or to know *apriori* when and how to apply a control sequence $\{k_i\}_{i=1}^{n}$. On the other hand, it is relatively easy to find slow orbits. However, slow orbits tend to waste time on long, sometimes extremely long, recurrent loops. These loops serve little more than to bide time until an appropriately aligned pass through the ball of recurrence has been achieved.

Slow orbits, while not useful in themselves, tell the story of how to find the switching points. Furthermore, the orbit segments between recurrences which do not themselves recur are assumed to be locally optimal. This assumption relies on having correctly chosen the preassigned recurrence threshold δ so that all possible patches are glued. Choosing the threshold too large, however, wastes time checking "recurrences" that have no chance of being patched.

There are similar ideas to this in classical control literature. Dynamic programming, based on Bellman's principle of optimality, asserts that a globally optimal orbit (and its associated control function) must also necessarily be locally optimal for segments of the orbit [46, 102]. While local optimization does not in general imply global optimization, a huge improvement may nonetheless be achieved with a small amount of computer work, in the process of satisfying Bellman's necessary condition. For the two-dimensional case, if we correctly choose δ as the radius of a lobe, and if each region only has one family of lobes, I believe that the restricted orbit found must be close to optimal.

In order to efficiently find the recurrences, we use the following algorithm. Given an orbit $\{\boldsymbol{z}_0, ..., \boldsymbol{z}_j, ..., \boldsymbol{z}_N\}$, where $\boldsymbol{z}_0 = \boldsymbol{z}_a$ is close to $\boldsymbol{a}$ and $\boldsymbol{z}_N = \boldsymbol{z}_b$ is close to $\boldsymbol{b}$, we perform the following:

```
for i=0 to N
    for j=N to i+1, step -1
        if ρ(z_i − z_j) ≤ δ  /*Is there a recurrent loop?*/
```

```
            then
            attempt to remove the loop /*Can a patch be found?*/
            if recurrent loop can be removed,
                 then
                 cut it and glue in patch
                 let i=j
            end if
         end if
     loop
loop
```

This algorithm automatically considers the largest recurrences for removal first by working forward from $\boldsymbol{z}_0$ and backward from $\boldsymbol{z}_N$. Shorter, intermediate recurrences that occur inside of a bigger recurrence are automatically removed without ever being considered. This represents an improvement over a purely forward search which might require an ordering according to lengths of loops.

It may not always be necessary to have a single orbit from $\boldsymbol{a}$ to $\boldsymbol{b}$ to use this algorithm. Two regions of phase space can be explored separately by starting separate initial conditions and concatenating their resulting orbits. This can be a useful way to explore the phase space near $\boldsymbol{a}$ and $\boldsymbol{b}$ separately when a single orbit between them is particularly difficult to find. If the two orbits closely approach each other, then it might be possible to patch from one to the other. The above algorithm, applied to the concatenated orbit, will automatically test all such possibilities. If, however, such a patch is not possible, then the end of the first orbit will be reached with no connection to the second orbit. In contrast, when a single orbit between the two regions can be found, the algorithm is

robust, because the original, albeit slow orbit is always available as the path.

The prerecorded orbit represents known information about transport in the visited phase space. How we manage this information depends on our assigned task. If we are likely to be presented with a variety of initial conditions $\boldsymbol{a}$ and targets $\boldsymbol{b}$, then the following model may be used. Ergodicity causes a long orbit to cover accessible phase space. The longer the orbit, the better the cover. Any target point close to the known orbit is feasible, and initial conditions close to the orbit allow for immediately starting stabilization. Alternatively, if no points of the known orbit are close enough to the initial condition, then uncontrolled iteration will quickly cause it to come close. This model requires directly stabilizing the initial condition to the known orbit and restricting its length at points of recurrence, *on the fly*.

Stabilization can be performed by shooting the initial condition at the stable manifold of the known path using Newton's method to find the correct perturbations to the internal parameters. Details are discussed in [51] and represent only a slight modification to the gluing algorithm presented in the next section. Recurrences are detected and cut according to the above algorithm, where stabilization can immediately be used to skip a loop by shooting at the path after the loop. A successful cut is one in which the recurrence is close enough so that stabilization works with a sufficiently small parameter perturbation.

Another possible control task is one where $\boldsymbol{a}$ and $\boldsymbol{b}$ are fixed in advance. We may be presented with such a model either when just a few objectives are likely, or perhaps a decision tree is to be built and even the segments of the tree are difficult to find. In this case, time can be spent to find a more optimal solution achieving the transport. The resulting epsilon chain can be stabilized

later, in real-time as above. This model tends to build faster orbits since the patches are built forward and backward from the switching points, as compared to the *on the fly* model described above where only the future points can be modified.

2.2.3 Cutting Recurrent Loops Suppose that we find a recurrence between $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i+s}$, s steps later, i.e. $\|\boldsymbol{z}_{i+s} - \boldsymbol{z}_i\| < \delta$. Already, it is possible to skip $(s-1)$ iterates of the orbit by making the appropriate δ perturbation from $\boldsymbol{z}_i$ to $\boldsymbol{z}_{i+s}$. Even better, it may be possible to find a patch consisting of a nearby orbit $\boldsymbol{z}'_i$ with the property that $\|\boldsymbol{z}_{i-m} - \boldsymbol{z}'_{i-m}\|$ and $\|\boldsymbol{z}_{i+s+m} - \boldsymbol{z}'_{i+m}\|$ are both small enough to satisfy the control constraint. If the orbit is hyperbolic, then we can remove the $(s-1)$ step loop using an exponentially smaller total perturbation. We find a patch $\{\boldsymbol{z}'_{i-m}, ..., \boldsymbol{z}'_{i+m}\}$ consisting of a nearby orbit segment which is close to the pre-orbit of $\boldsymbol{z}_i$ before the recurrence, and close to the orbit of $\boldsymbol{z}_{i+s}$ after the recurrence, and which completely avoids the unwanted loop $\{\boldsymbol{z}_{i+1}, ..., \boldsymbol{z}_{i+s-1}, \boldsymbol{z}_{i+s}\}$. We choose m so that the perturbation onto the patch from the natural orbit is as small as we require. The existence of such an orbit patch is guaranteed if the recurrent points are hyperbolic and the recurrence distance δ is small enough, but the condition is not necessary. For the hyperbolic case, the size of δ depends on the geometry and angle of intersection between $W^s(\boldsymbol{z}_i)$ and $W^u(\boldsymbol{z}_{i+s})$. We will start by describing a point $\boldsymbol{p}$ on the patch which is between $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i+s}$; then the rest of the patch is formed by forward and backward iteration. The point $\boldsymbol{p}$ lies on the intersection of the stable manifold of $\boldsymbol{z}_{i+s}$, denoted $W^s(\boldsymbol{z}_{i+s})$, and the unstable manifold of $\boldsymbol{z}_i$, similarly denoted as $W^u(\boldsymbol{z}_i)$ and therefore has the property that

$$\|T^n(p) - T^n(\boldsymbol{z}_{i+s})\| \to 0 \text{ and } \|T^{-n}(p) - T^{-n}(\boldsymbol{z}_i)\| \to 0 \text{ as } n \to \infty. \qquad (2.7)$$

figure=ums.eps,height=2.5in

Figure 2.3. Construction of a patch. When the point $\boldsymbol{z}_i$ recurs with $\boldsymbol{z}_{i+s}$, the point of principal intersection $\boldsymbol{p}$ between $W^u(\boldsymbol{z}_i)$ the unstable manifold of $\boldsymbol{z}_i$ and $W^s(\boldsymbol{z}_{i+s})$ the stable manifold of $\boldsymbol{z}_{i+s}$ converges to the orbit of $\boldsymbol{z}_{i+s}$ under applications of the map T, and converges to the pre-orbit of $\boldsymbol{z}_i$ under applications of the inverse map T^{-1}.

By the proximity of $\boldsymbol{z}_i$ to $\boldsymbol{z}_{i+s}$, $\boldsymbol{p}$ is within $h\delta$ of both $\boldsymbol{z}_i$ and $\boldsymbol{z}_{i+s}$, where h is a constant that depends on the geometry of the intersection between $W^s(\boldsymbol{z}_{i+s})$ and $W^u(\boldsymbol{z}_i)$. To lowest order, h depends on θ, the angle of intersection between the local linear approximations to the manifolds. We expect that our technique will be less effective when θ is small; the resulting triangle implies that $\boldsymbol{p}$ will be far from the point of recurrence which invalidates the locality assumptions. By construction, we expect that

$$\begin{aligned} \|T^m(\boldsymbol{z}_{i+s}) - T^m(\boldsymbol{p})\| &< h\lambda_s^m\delta \\ \|T^{-m}(\boldsymbol{z}_i) - T^{-m}(\boldsymbol{p})\| &< h\lambda_u^{-m}\delta \end{aligned} \tag{2.8}$$

where $\lambda_s < 1$ is the local stable Lyapunov number at $\boldsymbol{z}_{i+s}$ and $\lambda_u > 1$ is the local unstable Lyapunov number at $\boldsymbol{z}_i$ (see Fig. (2.3)).

In principle, it should be possible to choose m so that the perturbations from the original orbit onto the patch, and then back onto the original orbit are as small as we like. However, numerically finding points on the stable (unstable) directions becomes increasingly ill-conditioned if m is too large. In practice finding the complete manifolds $W^s(\boldsymbol{z}_{i+s})$ and $W^u(\boldsymbol{z}_i)$ in order to find $\boldsymbol{p}$ is not practical, efficient, or even important. Instead we find $\boldsymbol{p}$ indirectly by making the approximation that $T^{-m}(\boldsymbol{p})$ lies in the tangent space of $W^u(\boldsymbol{z}_{i-m})$, and likewise that $T^m(\boldsymbol{p})$ lies in the tangent space of $W^s(\boldsymbol{z}_{i+s+m})$. Hence finding $T^{-m}(\boldsymbol{p})$ can

be reduced to a problem of shooting.

In two dimensions we can parameterize the unstable direction with the vector $f_u(\boldsymbol{z}_{i-m})$, a unit vector in the tangent space of $W^u(\boldsymbol{z}_{i-m})$, with the variable t. An initial condition is chosen using

$$\boldsymbol{z}_0(t) = \boldsymbol{z}_{i-m} + t f_u(\boldsymbol{z}_{i-m}). \tag{2.9}$$

The success of an initial condition can be measured by how closely $T^{2m}(\boldsymbol{z}_0(t))$ lands on the line $\boldsymbol{z}_{i+s+m} + \tau f_s(\boldsymbol{z}_{i+s+m})$. We write components of the vectors $\boldsymbol{z}(t) = (x(t), y(t))$ and $f_s = (f_{s,x}, f_{s,y})$. The roots of the expression,

$$F(t) = f_{s,y}(x(t) - x_{i+s+m}) - f_{s,x}(y(t) - y_{i+s+m}) = 0, \tag{2.10}$$

can be found quickly using a Newton-secant method. We need only make an appropriate initial guess so that the point we find will in fact be a principal intersection point.[5] As a rough guess, we can use Eq. (2.8) to write

$$t_0 = \lambda_u^{-m} \delta \tag{2.11}$$

where we have assumed that $h \approx 1$. Extension to higher dimensions is straightforward. The number of variables needed to parameterize the initial condition must equal the dimension spanned by the unstable subspace of the tangent space at $\boldsymbol{z}_{i-m}$. Likewise, examining the related problem of shooting at the stable manifold using the parameters k provides the controllability condition that perturbations to k must span the unstable subspace of the tangent space at $\boldsymbol{z}_{i+s+m}$.

[5] There are in fact an infinite number of zeroes to (2.10) due to the stretching and folding of the line (2.9) by the chaotic map. There is however a first zero t corresponding to the p.i.p. $\boldsymbol{p}$. If Newton's method bounces chaotically between basins of attraction, even when given a "good" guess (2.11) it may be necessary to use a sure bracketing algorithm such as the bisection method until the parameter is sufficiently close to use the faster Newton's method.

Finding the stable direction f_s and the unstable direction f_u at a point $\boldsymbol{z}$, which is not necessarily periodic, from a chaotic set first requires the complete orbit $\{..., \boldsymbol{z}_{-n}, ..., \boldsymbol{z}_{-1}, \boldsymbol{z}_0, \boldsymbol{z}_1, ..., \boldsymbol{z}_n, ...\}$. Recall that the Jacobian matrix rotates a vector in the tangent space towards the unstable direction, and the Jacobian matrix of the inverse map T^{-1} rotates a vector towards the stable direction. Therefore, in practice, we choose an arbitrary unit vector $\boldsymbol{u}$ and forward multiply, starting at $\boldsymbol{z}_{-n}$, the Jacobian matrices along the orbit to $\boldsymbol{z}$, normalizing the vector at each step:

$$DT^n|_{z_{-n}} \cdot \boldsymbol{u} \equiv DT|_{z_{-1}} \cdot DT|_{z_{-2}} \cdot ... \cdot DT|_{z_{-n}} \cdot \boldsymbol{u} \rightarrow f_u(\boldsymbol{z}) \text{ as } n \rightarrow \infty. \quad (2.12)$$

Likewise, the stable direction is formed from the inverse Jacobian starting at $T^n(\boldsymbol{z})$.

$$DT^{-n}|_{z_n} \cdot \boldsymbol{u} \equiv DT^{-1}|_{z_1} \cdot DT^{-1}|_{z_2} \cdot ... \cdot DT^{-1}|_{z_n} \cdot \boldsymbol{u} \rightarrow f_s(\boldsymbol{z}) \text{ as } n \rightarrow \infty. \quad (2.13)$$

Convergence is exponential; in practice we find that $n = 20$ gives an error of 10^{-5}. We use $n = 40$ which we expect is more than adequate considering the scale of other errors, (see [53]). At the same time, as in the above calculation, we calculate the corresponding Lyapunov multipliers, also by the power method.

All of the required quantities for cutting and gluing are in fact accessible to a model of a dynamics formed by time-series embedding. The primary piece of information, a recurrence, requires no modeling to identify. To form the patch, however, we need to fit a piecewise model of the data in order that predictions may be formulated between known data points [29]. In such a case, a more accessible numerical technique to form the unstable and stable directions at a point $\boldsymbol{z}$ is to consider the recorded histories of nearby clusters of points in forward and backward time. Likewise, from $T^{-n}(\boldsymbol{z})$, nearby points orient along

the unstable axis in forward time at $\boldsymbol{z}$. In addition, partial derivatives of the map T in each of the parameter directions may be approximated by interpolating between three models of separate data sets which bracket the range of each parameter. More details on modeling can be found in Chapter 5.

Note that further refinement to the epsilon chain can be achieved by running a second pass of the gluing algorithm, by treating chain errors as the δ and finding a patch over it to further reduce the error by a factor of λ^m. Hence a smaller error epsilon chain can be achieved with a modest m by redistributing the points of the error to the ends of the new patch.

2.2.4 The Standard Map, an Example:

2.2.5 Area Preserving Transport We now demonstrate our method for the case of the much studied standard map. The standard map, also known as the kicked rotor, is an area preserving twist map of the plane:

$$\boldsymbol{z}' = \begin{pmatrix} y' \\ x' \end{pmatrix} = \begin{pmatrix} y - \frac{k}{2\pi}\sin(2\pi x) \\ y - \frac{k}{2\pi}\sin(2\pi x) + x \end{pmatrix}. \tag{2.14}$$

The phase space structures and transport characteristics are typical of Hamiltonian systems with two degrees of freedom. There is periodicity in both x and y with period 1, so the phase space is the torus. Hence according to the Poincaré recurrence lemma 2.2.1 every orbit must eventually recur.

As a concrete example, we investigate transport from a neighborhood of the $(0, 1)$ hyperbolic point of (2.14), to a neighborhood of the $(1, 1)$ resonance. The notation (p, q) denotes the frequency of an orbit, i.e. q iterations of the map results in exactly p wraps around the cylinder: $T^q(\boldsymbol{z}) = \boldsymbol{z} + p$. The starting point $(0, 1)$ $\boldsymbol{a}$ is located at $(x_a, y_a) = (0.5, 0.0)$, and (1,1), our target point $\boldsymbol{b}$, at $(x_b, y_b) = (0.5, 1.0)$.

It is not possible to find such an orbit if $k < k_c \approx 0.97163540631...$ [70]; k_c is the parameter value at which the last invariant curve dividing phase space between $(0,1)$ and $(1,1)$ becomes a cantorus. The most robust curves between $(0,1)$ and $(1,1)$ are the circles with rotation frequencies $\frac{1}{\gamma}$ and $\frac{1}{\gamma^2}$, where $\gamma = \frac{1+\sqrt{5}}{2}$ is the golden mean.[6]

The standard map is an example of an area preserving twist diffeomorphism. Therefore, a theorem of Mather [69] allows us to conclude that if there are no separating invariant curves between any two unstable periodic orbits, then there exists a heteroclinic connection between these points. Specifically, if $k > k_c$, there are no invariant curves separating vertical transport of the cylinder, and therefore according to Mather's theorem, there exists a heteroclinic connection between the $(0,1)$ and $(1,1)$ orbits for which we will search.

For $k > k_c$, the golden mean invariant curves become cantori. They have the smallest lobe areas, and hence, represent the most difficult barriers to transport. An arbitrary orbit will typically pass through these lobes many times before finally reaching the target point $\boldsymbol{b}$. This effect also occurs when an orbit is trapped near an island, and near islands around islands, and so on. This phenomenon has been successfully modeled using Markov trees [42, 71]. It was found that a point initially "near" a KAM surface has a survival probability $F(t)$ is asymptotic to $t^{-\alpha}$ and that the orbit will still be near the surface at large time t with small constant α [71]. Therefore in the presence of KAM surfaces, we find long correlations and, hence, roughly power law decay. However the important point is that, without knowing where the lobes are located, recurrence is a way of locally detecting globally inefficient orbits.

[6] See Sec. 6.3 for further discussion of noble numbers such as the golden mean and their role as robust frequencies in the breakup of invariant KAM tori.

2.2.6 Transport Time Distributions Before we demonstrate the improvements made by cutting recurrences, we first investigate the natural transport time distributions for a range of parameter values. We performed a Monte Carlo study within a box of size 0.1 centered around our starting point $\boldsymbol{a}$ at $(-0.5, 0.0)$, from which we randomly choose an ensemble of 10^4 initial conditions, and bin them according to how long they took to arrive in a similar box around the target $\boldsymbol{b}$ at $(-0.5, 1.0)$. Two of the resulting histograms are displayed together in Fig. (2.4) for the parameter values $k = 1.01$, just above k_c, and the moderately high value $k = 1.25$.

figure=Fig3.eps,height=3.5in

Figure 2.4. Histograms of transport time for the standard map between the box of side 0.1 centered on the starting point $\boldsymbol{a}$ at $(0.5, 0.0)$ to a similar box centered on the target point $\boldsymbol{b}$ at $(0.5, 1.0)$. 10^4 initial conditions are randomly chosen from the $\boldsymbol{a}$ box and binned according to time to transport into the $\boldsymbol{b}$ box. The number of bins allotted is 100 chosen as the square root of the population. The solid curve is for $k = 1.01$, and the dashed curve is for $k = 1.25$. The maximum iterations performed on an initial condition is $5 \cdot 10^6$ before cut-off. Points requiring more iterations are found in the last bin.

We find results similar to the experiment performed by Chirikov [21] who observed that transport time from $y \approx 0$ to $y \approx 1$ obeyed a power law. Fig. (2.5) displays average transport time on a log log plot versus $(k - k_c)$, demonstrating that the average crossing time is indeed well approximated by the singular power law $(k - k_c)^\eta$ [21, 23].

figure=Fig4.eps,height=3.5in

Figure 2.5. Transport times from the $\boldsymbol{a}$ box to the $\boldsymbol{b}$ box as a function of $(k-k_c)$. The top curve shows the average time, calculated amongst 10^4 randomly chosen initial conditions from the $\boldsymbol{a}$ box iterated until first intersection with the $\boldsymbol{b}$ box. The bottom curve shows optimized transport times from $\boldsymbol{a}$ to $\boldsymbol{b}$ resulting from cutting recurrent loops from slower orbits.

It is possible that many of the initial conditions are chosen in the subset of the box that is inaccessible[7] to $\boldsymbol{z}_b$. These, and other initial conditions have such long transport times, that we choose a cut-off time of $5 \cdot 10^6$ iterates, which is increasingly a problem for k approaching k_c. This causes the value of η that we calculate to be somewhat less than the value 3.012 predicted [70]. The variance of the distribution is large, as seen in Fig. (2.4), but impossible to measure due to the large fraction of the box with transport times in the last bin, for any reasonable finite cut-off. Indeed, we observed that the transport rate is extremely sensitive to the initial condition, so that the average transport rate is not an indicative measure of a "typical" rate. Fig. (2.6) displays the percentage of initial conditions from the $\boldsymbol{a}$ box that never reach $\boldsymbol{b}$.

figure=Fig5.eps,height=3.5in

Figure 2.6. The percentage of the initial conditions from the $\boldsymbol{a}$ box which take longer than $5 \cdot 10^6$ iterations to reach the $\boldsymbol{b}$ box.

Here we find that as k approaches k_c, most of the transport times are actually larger than our cut-off. Thus, our computed statistics are only lower bounds on the actual transport statistics. Nonetheless, the point is that transport is slow as $k \to k_c$, indicating that our efforts to find faster orbits are worthwhile.

2.2.7 Cutting and Gluing Slow Orbits From the same box around the initial point $\boldsymbol{a}$ described above, we choose an initial condition which iterates eventually to near the target $\boldsymbol{b}$. We restrict ourselves to the orbits of randomly chosen initial conditions that perform the transport required in

[7] In order for an initial condition to be accessible to $\boldsymbol{z}_b$ it must lie in a pre-iterate of the box around $\boldsymbol{b}$ which is contained in a pre-iterate of each of the lobes of each of the barriers in between. Another way of saying this is that there must be a heteroclinic connection.

less than 10^6 iterations so that we do not strain the memory capacity of our computing resources by storing uninteresting information.

As a concrete example, consider a 80307 step orbit for $k = 1.25$. Recurrences are sought following Sec. 2.2.2 searching for $\boldsymbol{z}_i$ from the start of the orbit, and the last recurrence $\boldsymbol{z}_{i+s}$ (largest value of s) from the end of the orbit. A certain amount of space must be reserved in order to fit the patches.[8] We used patches consisting of $(2m+1) = 31$ steps so $m = 15$ steps must be allowed for on either side of the recurrence in order that the error may have time to contract sufficiently that the constraint (1.17) is satisfied. The rate of contraction is determined by the Lyapunov exponents, according to Eq. (2.8). Therefore a strict lower bound on the cost function I for our technique is $(2m+1)+1$, the space required for one patch. If patches are forced to not overlap, then q recurrences imply that I is bounded by $q(2m+1)+1$. In principle, arbitrarily small constraints (1.17) can be met, but in practice, solutions become numerically ill-conditioned as m gets large. We chose a modest value $m = 15$ for this example, although $m = 25$ and $m = 30$ were successfully tested.

The first recurrence that we can successfully remove from our 80307 step orbit is between $\boldsymbol{z}_{16}$ and $\boldsymbol{z}_{78704}$ which recurs to a distance $\delta = 0.08$. The cut and glue algorithm allows us to construct an orbit patch $\{\boldsymbol{z}'_1, ..., \boldsymbol{z}'_{31}\}$ such that the error to perturb on to the orbit patch is only $\|\boldsymbol{z}'_1 - \boldsymbol{z}_1\| = 0.002$, and the error to perturb back off of the orbit patch is $\|\boldsymbol{z}'_{31} - \boldsymbol{z}_{78720}\| = 0.002$. Fig. (2.7), showing the error between the patch orbit segment and the two ends of the recurrence on the orbit, displays how hyperbolicity is used to diminish the recurrence error. The slopes of the decaying error on either side of the recurrence

[8] We will see in Chapter 4 by examining a symbol dynamics model how this required space implies that we do not actually get an optimal orbit, but we are nonetheless close.

figure=Fig6.eps,height=3.5in

Figure 2.7. The error between a patch orbit $\{\boldsymbol{z}'_n\}_{n=-15}^{15}$ and an original orbit at each of the points $\boldsymbol{z}_{i+s+n}$ for $n \geq 0$, and $\boldsymbol{z}_{i+n}$ for $n \leq 0$.

represent the corresponding stable and unstable Lyapunov exponents.

With this single patch, we have already demonstrated a 1619 step epsilon chain orbit near our original orbit. By finding every recurrence within a threshold $\delta = 0.1$ and cutting those that can be patched within the error $\epsilon = 0.005$ we eventually construct a 131 step epsilon chain orbit including 13 overlapping patches. The largest error found in this example was $\|T(\boldsymbol{z}_{92}) - \boldsymbol{z}_{93}\| = 0.003$, but there were several others of the same order. For this case, 13 important switching points in 13 important lobes are inferred, and orbit segments between the switching points are automatically found by keeping those segments that do not recur close enough to be further cut.

To further demonstrate manageability of the errors, we ran a second pass of the patching algorithm over the error points of the newly formed epsilon chain. Pushing the errors back along the unstable manifold, and forward along the stable manifold, applies just as readily to new ϵ-chain as it did to the δ-chain. Redistribution of the errors yields a further reduction by a factors of 50 to 1000.

A phase space portrait of the final path is displayed in Fig. (2.2). The black regions in the plot represent inaccessible regions of phase space, corresponding to elliptic islands with rational frequencies between 0 and 1. The 131 step epsilon chain orbit, marked by black circles, manages to transport through all the resonances without getting caught in their periodicity, even though orbits trapped in a given resonance layer must move at that frequency. The efficiency

of transport can be observed by the lack of getting bogged down in almost periodicities as revealed by the lack of corresponding recurrences.[9]

Several different runs with $k = 1.25$ for various initial conditions ranging up to a length of 10^6, and various recurrence thresholds from $\delta = 0.02$ to 0.07 yield epsilon chain orbits of lengths $n = 131$ to 251. There is a trade off when choosing the recurrence threshold value δ. Recall that the rate at which a recurrence error δ can be reduced to the tolerance ϵ is governed by inequalities (2.8) which we adapt here to require that

$$\epsilon > h\lambda_{s,u}^{\pm m}\delta, \tag{2.15}$$

where we choose $m = 15$. Close recurrences are more likely to be cut. The resulting t and τ, the perturbations along the vectors $f_u(\boldsymbol{z}_{i-m})$ and $f_s(\boldsymbol{z}_{i+s+m})$ both denoted ϵ, are small according to (2.15) for fixed m. In addition, it is often possible for (2.10) to be solved even if t and τ are large, but the linearization of the stable (unstable) manifolds is not valid, and so contraction according to (2.15) is not expected. A successful cut is one where a given tolerance ϵ can be satisfied. Since zeroing (2.10) is relatively cheap[10], it is practical to set δ, the recurrence testing threshold, to a relatively high value where most recurrences are not successfully cut, but no opportunities are missed.

Table (2.1) shows that higher δ can yield faster orbits, but at a cost of many more trials. We observe in fact that $\delta \leq 0.1$ and $\delta \leq 0.2$ yield the same paths; considering $\delta \leq 0.2$ recurrences yields no extra successful patches. This is reflected in the success rate column of Table (2.1). All increases in δ

[9]This observation can perhaps be used to loosely distinguish slow transporting orbits from faster orbits by the degree of their quasi-periodicity.

[10]It is cheap when we have a closed form representation of the map. It may not be as cheap if we must calculate the map from a flow by Poincaré section. Hence we will need to take more care in the next chapter.

Table 2.1. Various recurrence thresholds. Values of δ are tested on a single 109594 step orbit between the $\boldsymbol{a}$ box and $\boldsymbol{b}$ box for $k = 1.25$ and patch size $m = 15$ to achieve an $\epsilon = 0.005$ tolerance. This shows how increasing computer work, to a point, yields faster paths by considering unlikely patchable recurrences. Tabulated quantities are: the threshold tested δ, the resulting epsilon chain length n, the number of loops successfully cut, and percentage ratio of successfully cut loops to those attempted when a δ recurrence was detected.

δ	n	*loops cut*	*% success rate*
0.005	748	15	100.0
0.01	597	13	3.96
0.025	236	8	2.020
0.05	177	14	1.552
0.08	173	14	1.14
0.09	156	16	1.16
0.1	156	16	0.920
0.2	156	16	0.096
1.0	156	16	0.035

up to 0.09 did in fact yield faster paths. By contrast, if we allow ourselves to use longer patches by increasing m to allow more space to contract, in principle, we expect that higher values of δ are likely to be successful, but at the cost of ill-conditioning the solution of (2.10). We find a good balance at $m = 15$, but this choice of m is arbitrary. We note that the solution of the long time targeting problem would be trivial if there were no ill-conditioning problem with growing m, since we could simply choose $\delta = 1$. There would be no need to consider intermediate switching points, the location of which is the source of the major difficulty to slow transport problems. In that case $\boldsymbol{a}$ would be recurrent with $\boldsymbol{b}$ and by choosing a very large value of m (such as $m = 65$ as indicated by our 131 step orbit above) we could shoot directly from beginning to end.

To find the optimum value it is best to choose δ equal to the "diameter" of the important lobes between $\boldsymbol{a}$ and $\boldsymbol{b}$, in their most round iteration as they become long and thin in both forward and backward time. Define the most round iteration of a lobe as the iteration with minimum diameter, where diameter is defined by the supremum of the distance between any two points in the set. Unfortunately, in general we do not know this value *apriori*. An efficient technique would involve several passes for increasing values of δ, first forming an epsilon chain orbit that always stays below control saturation (1.17), and then removing any possible recurrences for the next value of δ. This would represent a multi-pass algorithm since the length of the epsilon chain orbit decreases monotonically.

The lower curve of Fig. (2.5) displays "optimized" transport times as a function of $(k - k_c)$, for a one pass optimization. These values can be compared to the uncontrolled transport time averages also displayed in the same figure. We see an improvement by a factor of almost 10^4 on average for the lower values

figure=Graph.eps,height=4.5in

Figure 2.8. Histograms of angles between stable and unstable manifolds. In (a) the distribution shows angles for each of the 975760 points along an uncontrolled orbit between the $\boldsymbol{a}$ box and $\boldsymbol{b}$ box for $k = 1.01$. In (c) the angles along an uncontrolled orbit for $k = 1.25$ are shown. In (b) the 975760 step orbit for $k = 1.01$ has been restricted to a 778 step epsilon chain. Similarly in (d) for $k = 1.25$, a 80307 steps orbit has been restricted to a 131 step epsilon chain.

of k.

2.2.8 Hyperbolicity in the Standard Map The construction of an orbit patch as explained in Sec. 2.2.3 is guaranteed to work when the orbit is a hyperbolic saddle. A hyperbolic invariant set is defined as a compact, nonattracting, invariant set with a dense orbit such that each point of the set has a stable direction and unstable direction. A hyperbolic saddle has all the angles between stable and unstable manifolds bounded away from zero, and a nonhyperbolic saddle may have angles that approach zero. In order that the intersection point $\boldsymbol{p}$ is close to the δ ball containing the recurrent points, the constant h, depending on the angle between stable manifold at $\boldsymbol{z}_{i+s}$ and unstable manifold at $\boldsymbol{z}_i$, must be large. In this section, we discuss the validity of this assumption by numerically investigating the angle distribution between stable and unstable directions, calculated according to Eqs. (2.12)-(2.13), at each point along a slow orbit between the objectives $\boldsymbol{a}$ and $\boldsymbol{b}$. This is in the spirit of the recent paper by Lai *et al.* [53] who perform similar calculations for the Hénon map.

Fig. (2.8) show angle distributions for $k = 1.01$ and $k = 1.25$. Distributions found were similar in shape, average, and peak for each value k, independent of initial condition. In Figs. (2.8a) and (2.8c) we can see a definite spike at the peak in the angle distributions. From there the probability falls off

smoothly with increasing angle. It is not clear whether the angles are bounded away from zero or not, since the curve falls off smoothly, seemingly toward zero. This situation for the standard map is qualitatively quite different from that of the Hénon map, where the distributions show "complicated structure without much regularity near $\theta = 0$." The lowest angle for the figures displayed is $\theta = 0.0009$ for Fig. (2.8a) and $\theta = 0.12$ for Fig. (2.8c). Results are similar when other transporting initial conditions from the $\boldsymbol{a}$ box are chosen.

The point that concerns us here is that the probability of finding angles below any reasonable value is quite low. According to our algorithm, a recurrence is only cut when an orbit patch within the control tolerance can be found on a trial and error basis. So, we are only concerned here with the probability of small angles: for example in Fig. (2.8c) $P(\theta < 8^o) \cong x = 0.7\%$.

It is interesting to compare the angles between $f_u(\boldsymbol{z}_i)$ and $f_s(\boldsymbol{z}_{i+s})$ of recurrences that cannot be successfully mended to those that can be mended. For $k = 1.25$ and $\delta = 0.01$ we recorded separately the angles of successes and failures. The failure category includes the entire range of angles. We expect a problem with small angles, but even large angles can be a problem for recurrences that are not close enough or when the manifolds curve sharply away from the linear approximations. In contrast, the smallest successfully mended angle found was 11^o, and the angles tend to be much higher than that in general.

Figures (2.8b) and (2.8d) show the angle distributions for the mended epsilon chain orbits. Angles are now calculated using the epsilon chain rather than following the natural orbit of the point $\boldsymbol{z}$ which may quickly diverge from the predicted pseudo-orbit. The main feature we observe in the restricted orbits is that the average angle invariably increases. For example, the average for Fig. (2.8a), $< \theta >= 29.9^o$ was increased to $< \theta >= 50.9^o$ along the 778 step

epsilon chain. Similarly, Fig. (2.8d) reveals a change from $< \theta >= 40.5^o$ to $< \theta >= 46.9^o$ along the 131 step path.

2.2.9 Stabilization We next demonstrate that an initial condition can indeed be stabilized onto the epsilon chain with small parameter perturbations as described in Sec. 2.2.2. As an illustration consider the same orbit as in Sec. 2.2.7 for which we found a nearby 131 step pseudo-orbit $\{\boldsymbol{z}_1, ..., \boldsymbol{z}_{131}\}$ with 13 overlapping patches. The parameter perturbation size required for stabilization depends directly on the phase space error found.

We demonstrated that initial conditions close to $\boldsymbol{z}_1$ can be stabilized along the epsilon chain to $\boldsymbol{z}_{131}$. Parameter perturbations were used to shoot at the stable manifold further down the path, and were calculated whenever a point of error was predicted on the epsilon chain, or the test point drifted outside a predetermined tolerance of the known path. For this particular example, a maximum value $\Delta k = 0.016$ was required, but most perturbations were several orders of magnitude smaller. Stabilization was successful for all the epsilon chains tested, for various values of k_0.

For comparison, we also tested *on the fly* stabilization directly to the long 80307 step orbit. Again, we used a randomly chosen initial condition near the known orbit and stabilization was turned on whenever the test orbit drifted outside a set tolerance. In addition, whenever a recurrence in the pre-recorded orbit was detected along the way, stabilization was attempted by shooting at the end of the loop. We expect that the length of the test orbit will be longer, because with this method, only future events can be modified. In this example, the test orbit achieves the final destination in 464 iterations. In spite of being typically slower, *on the fly* control can be more flexible since one can rapidly

retarget[11] as needed, as the whole pre-recorded orbit is available.

2.2.10 Conclusions We have discussed the time-optimal control problem for chaotic regions, presenting a method to find paths that quickly achieve transport goals and which can be stabilized with small parameter perturbations. A recurrent orbit necessarily violates Bellman's condition for optimality; we eliminate recurrences by using them as switching points between orbit segments. The resulting epsilon chain is refined by smoothing with a patch that shoots from the unstable manifold of the orbit before the recurrence onto the stable manifold of the orbit after the recurrence. The method was demonstrated on the standard map for which uncontrolled transport is extremely slow and the statistics of the transit times are anomalous. The technique, when applied to orbits of up to 10^6 iterates, typically reduces transport times by factors of up to 10^4, even for k close to k_c where previous targeting schemes are unsuccessful [52]. Finally, we demonstrated that small parameter perturbations can be used to stabilize a test orbit onto either a pre-calculated fast epsilon chain, or alternatively onto a long known orbit, eliminating recurrences *on the fly*.

In straight forward manner, the techniques in this chapter can be extended in a number of directions. Though we used a Hamiltonian mapping as an example, the method makes no assumptions as to the nature of the dynamics; indeed since the inverse map is not needed, the dynamics can even be noninvertible. The use of recurrences as switching points requires no assumptions as to the dimensionality of the phase space, though recurrences will be less common in higher dimensions. As well as optimizing transport between two separated points, the method could also be used to eliminate escape from a region, using

[11] We can choose a new $\boldsymbol{b}$ near our long orbit library at any time.

the recurrences to construct a rough periodic orbit. This might require smaller and less frequent control than stabilizing a fixed point in a given region. Finally, the techniques presented in this chapter are also applicable to a piecewise local model of a time series built through embedding, which we will explore in more detail in Chapter 5.

CHAPTER 3

TARGETING CHAOTIC TRANSFER ORBITS TO THE MOON THROUGH RECURRENCE

3.1 Introduction

In this chapter, we will apply the targeting technique of the previous section to another well known area preserving map: the circular[1] restricted three-body problem. As with the standard map, we will see the resonance layered phase space portrait typical of Hamiltonian maps, which makes targeting particularly difficult.

After a brief overview of the restricted three-body problem and derivation of its equations of motion, I will discuss some problems special to integrating this flow. Then I will discuss the process of building a Poincaré map of the flow, and some special issues involved with finding the stable and unstable manifolds restricted to the surface of section. Finally, I will perform "targeting through recurrence." Fixing the the mass ratio to that appropriate for the Earth - Moon system, orbits found have real life application of sending a rocket to the Moon with small perturbations (rocket burns) in minimal time. In the final discussion of the chapter, we will compare these low energy orbits to the classic Hohmann transfer orbits.

[1]In the literature we find the description "elliptic restricted three-body problem." The convention is that if the word "elliptic" is omitted, then it is understood that we are discussing the circular problem.

3.2 The Restricted Three-Body Problem

The full three body problem Hamiltonian may be written [72]

$$H(\boldsymbol{P},\boldsymbol{Q}) = T(\boldsymbol{P}) + V(\boldsymbol{Q}) = \sum_{i=1}^{3} \frac{\|\boldsymbol{P}_i\|^2}{2m_i} - \sum_{1\le i<j\le 3} \frac{Gm_i m_j}{r_{ij}}, \tag{3.1}$$

where

$$r_{ij} = \|\boldsymbol{Q}_i - \boldsymbol{Q}_j\|. \tag{3.2}$$

We see that this Hamiltonian is in the form of kinetic energy plus a potential. Each $\boldsymbol{Q}_i \in \mathbb{R}^3$ and has a conjugate momentum $\boldsymbol{P}_i \in \mathbb{R}^3$. Thus the general problem requires an 18-dimensional phase space. Poincaré ended a long standing search for a closed form solution to the general n-body problem by proving that standard perturbation theory does not converge [83, 56]. In fact, the only general solutions to this problem are the Kepler solutions applicable to the two body case [99].

The "restricted" problem is a special case of (3.1) first formulated by the American astronomer G.W. Hill[2] in 1878 [100]. Here one of the n masses is assumed to be vanishingly small, and so has no influence on the remaining bodies, the "primaries", which are free to follow Keplerian motion. For the three body case, the Hamiltonian for the small mass becomes

$$H(\boldsymbol{P},\boldsymbol{Q}) = \frac{\|\boldsymbol{P}\|^2}{2} - \Phi(\boldsymbol{Q}), \tag{3.3}$$

where

$$\Phi(\boldsymbol{Q}) = (\frac{m_1}{r_1} + \frac{m_2}{r_2}), \tag{3.4}$$

and

$$\begin{aligned} r_1^2 &= (Q_{1,x} - Q_x)^2 + (Q_{1,y} - Q_y)^2, \\ r_2^2 &= (Q_{2,x} - Q_x)^2 + (Q_{2,y} - Q_y)^2. \end{aligned} \tag{3.5}$$

[2] In this paper, Hill argued that the Moon is permanently bounded near the Earth by using his method of "curves of zero velocity."

Here $(Q_{1,x}, Q_{1,y})$ and $(Q_{2,x}, Q_{2,y})$ are the time dependent positions of the two primary masses m_1 and m_2 which may be solved for exactly, given the primaries' orbital elements.

We choose to normalize units such that the sum of the two masses is one,

$$\begin{aligned} m_1 &= 1-\mu, \\ m_2 &= \mu, \end{aligned} \tag{3.6}$$

the distance between the primaries is one and the gravity constant is one. We will derive the conversion between these units and the physical units at the end of the section. We may also restrict the small mass to the plane of the primaries' orbits by assuming that $Q_z = 0$ and P_z=0. For simplification, it is assumed that the two primaries orbit each other with zero eccentricity. This is typically a well founded approximation for many of the orbits in our solar system. The eccentricity of the lunar orbit is $e = 0.0549$, and that of the Earth's solar orbit is $e = 0.0167$, to cite two examples. The exact Hamiltonian of the elliptic restricted problem may be expressed perturbatively in terms of the circular restricted Hamiltonian with small parameter e (see [72]).

We may change coordinates to a rotating frame with the generating function

$$S_2(\boldsymbol{p}, \boldsymbol{Q}) = \boldsymbol{p}^t \cdot R \cdot \boldsymbol{Q}, \tag{3.7}$$

expressed in terms of the old position and the new momentum, where

$$R = \begin{pmatrix} \cos t & -\sin t \\ \sin t & \cos t \end{pmatrix}, \tag{3.8}$$

is a rotation matrix. In Goldstein [37] we learn that choosing the function

$$S = S_2(\boldsymbol{p}, \boldsymbol{Q}) + \boldsymbol{p} \cdot \boldsymbol{q}, \tag{3.9}$$

is appropriate for the second type of function S_2, so that we may match the correct terms. The canonical change of coordinates follows from Hamilton's least action principle which states that the action is stationary,

$$\delta \int_{t_0}^{t_f} (\boldsymbol{P} \cdot \dot{\boldsymbol{Q}} - H) dt = 0. \tag{3.10}$$

The statement is true after the canonical transformation,

$$\delta \int_{t_0}^{t_f} (\boldsymbol{p} \cdot \dot{\boldsymbol{q}} - K) dt = 0, \tag{3.11}$$

which implies that the actions in both coordinate frames agree up to the derivative of a perfect derivative because

$$\delta \int_{t_0}^{t_f} \frac{dS}{dt} dt = \delta(S(t_f) - S(t_0)) = 0. \tag{3.12}$$

Here S is the generating function bridge between the two coordinates and is itself a Hamiltonian. Therefore we may compare the two actions using the generating function (3.9),

$$\boldsymbol{P} \cdot \dot{\boldsymbol{Q}} - H(\boldsymbol{P}, \boldsymbol{Q}, t) = \boldsymbol{p} \cdot \dot{\boldsymbol{q}} - K(\boldsymbol{p}, \boldsymbol{q}, t) + \frac{dS}{dt}, \tag{3.13}$$

written in differential form. Substituting the generating function (3.9) into (3.13) and expanding, we find

$$\boldsymbol{P} \cdot \dot{\boldsymbol{Q}} - H(\boldsymbol{P}, \boldsymbol{Q}, t) = -\dot{\boldsymbol{p}} \cdot \boldsymbol{q} - K(\boldsymbol{p}, \boldsymbol{q}, t) + \frac{\partial S_2}{\partial \boldsymbol{p}} \cdot \dot{\boldsymbol{p}} + \frac{\partial S_2}{\partial \boldsymbol{Q}} \cdot \dot{\boldsymbol{Q}} + \frac{\partial S_2}{\partial t}. \tag{3.14}$$

Linear independence of the $\dot{\boldsymbol{Q}}$, $\dot{\boldsymbol{p}}$, and 1 terms allows us to compare their coefficients separately yielding the transformation equations

$$\boldsymbol{P} = \frac{\partial S_2}{\partial \boldsymbol{Q}} = \frac{\partial}{\partial \boldsymbol{Q}} (\boldsymbol{p}^t \cdot R \cdot \boldsymbol{Q}) = R^{-1} \cdot \boldsymbol{p}, \tag{3.15}$$

$$\boldsymbol{q} = \frac{\partial S_2}{\partial \boldsymbol{p}} = \frac{\partial}{\partial \boldsymbol{p}} (\boldsymbol{p}^t \cdot R \cdot \boldsymbol{Q}) = (R \cdot \boldsymbol{Q}), \tag{3.16}$$

and

$$K(\boldsymbol{q},\boldsymbol{p},t) = H(\boldsymbol{Q},\boldsymbol{P},t) + \frac{\partial S_2}{\partial t}. \tag{3.17}$$

First inverting equation (3.16) to eliminate the old coordinate

$$\boldsymbol{Q} = R^{-1}\boldsymbol{q}, \tag{3.18}$$

we may then write

$$\frac{\partial S_2}{\partial t} = \frac{\partial}{\partial t}(\boldsymbol{p}^t \cdot R(t) \cdot \boldsymbol{Q}) = \boldsymbol{p}^t \cdot \dot{R}(t) \cdot R^{-1}(t) \cdot \boldsymbol{q}. \tag{3.19}$$

The reader may check the identity

$$\dot{R} \cdot R^{-1} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = -J. \tag{3.20}$$

Substituting (3.18) and (3.20) into (3.19) yields the term

$$\frac{\partial S_2}{\partial t} = -\boldsymbol{p}^t \cdot J \cdot \boldsymbol{q}, \tag{3.21}$$

corresponding to coriolis of the rotating coordinates.

R is a unitary matrix and therefore it follows from Eqs. (3.15) and (3.18) that

$$\|\boldsymbol{p}\| = \|\boldsymbol{P}\|, \tag{3.22}$$

$$\|\boldsymbol{q}\| = \|\boldsymbol{Q}\|. \tag{3.23}$$

Assuming that we start the two primaries m_1, and m_2 at $(Q_{1,x}, Q_{1,y}) = (-m_2, 0)$ and $(m_1, 0)$ respectively, and that they rotate around their center of mass at $(0,0)$, we find

$$r_1^2 = (Q_x - m_2\cos t)^2 + (Q_y - m_2 \sin t)^2 = (q_x - m_2)^2 + q_y \tag{3.24}$$

$$r_2^2 = (Q_x + m_1\cos t)^2 + (Q_y + m_1 \sin t)^2 = (q_x + m_1)^2 + q_y. \tag{3.25}$$

Defining

$$\phi(\boldsymbol{q}) \equiv \frac{m_1}{r_1(\boldsymbol{q})} + \frac{m_2}{r_2(\boldsymbol{q})}, \tag{3.26}$$

we may now write the new Hamiltonian in its canonical coordinates

$$K(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{2}(p_x^2 + p_y^2) - \boldsymbol{p}^t \cdot J \cdot \boldsymbol{q} - \phi(\boldsymbol{q}). \tag{3.27}$$

The Hamiltonian equations of motion are

$$\begin{aligned} \dot{q_x} &= \tfrac{\partial K}{\partial p_x} = p_x + q_y, \qquad & \dot{p_x} &= -\tfrac{\partial K}{\partial q_x} = p_y + \tfrac{\partial \phi}{\partial q_x}, \\ \dot{q_y} &= \tfrac{\partial K}{\partial p_y} = p_y - q_x, \qquad & \dot{p_y} &= -\tfrac{\partial K}{\partial q_y} = -p_x + \tfrac{\partial \phi}{\partial q_y}. \end{aligned} \tag{3.28}$$

The time independent form of the Hamiltonian (3.27) implies that the flow in this four-dimensional phase space is restricted to a three-dimensional submanifold of constant energy. Choosing the surface q_y=0 yields a two-dimensional Poincaré map.

The standard practice in the astronomy literature is to work with a Lagrangian form of the equations of motion, in which the coordinates $(\boldsymbol{q}, \dot{\boldsymbol{q}})$ are used in favor of the canonical variables $(\boldsymbol{q}, \boldsymbol{p})$. Eliminating $\boldsymbol{p}$ from (3.28) we find

$$\begin{aligned} \dot{q_x} &= u, \\ \dot{q_y} &= v, \\ \dot{u} &= q_x + 2v + \frac{\partial \phi}{\partial q_x}, \\ \dot{v} &= q_y - 2u + \frac{\partial \phi}{\partial q_y}. \end{aligned} \tag{3.29}$$

On the Poincaré section $q_y = 0$ we see in (3.28) that the velocity corresponds to the momentum because $\dot{q_x} = p_x + 0$. There are two area preserving maps corresponding to $q_y = 0$, one for each branch of $\pm\dot{q_y}$. We choose the branch $q_y > 0$.

Finally we may rewrite the Hamiltonian (3.27), replacing $(\boldsymbol{q}, \boldsymbol{p})$ in favor of $(\boldsymbol{q}, \dot{\boldsymbol{q}})$ which yields Jacobi's constant[3]

$$J(q_x, q_y, \dot{q_x}, \dot{q_y}) = \|\dot{\boldsymbol{q}}\|^2 - \|\boldsymbol{q}\|^2 - 2\phi(\boldsymbol{q}). \tag{3.30}$$

figure=otmp.eps,height=3.5in

Figure 3.1. A phase space portrait of the Poincaré mapping of a 10^5 iterate test orbit for the restricted three-body problem with $J = -3.177$. The point $(q_x, \dot{q_x})$ is plotted every time the flow pierces the surface $q_y = 0$ with positive $\dot{q_y}$. The Earth and Moon are clearly labeled at their fixed locations in the rotating frame. The phase space locations of the starting point $\boldsymbol{a}$ near the Earth, and the target point $\boldsymbol{b}$, near a Moon orbiting invariant torus, are also labeled. Several invariant tori are also shown to improve features.

An application of the model (3.29) is to look for a low energy transfer orbit to the Moon. To this end, we set $\frac{m_1}{m_2} = 0.0123$. Fig. (3.1) shows a phase space portrait of a long chaotic orbit of a small mass in the Earth - Moon system. In deriving (3.29) we normalized our units by setting $G = 1$, $m_1 + m_2$=1, and $|q_{1,x} - q_{2,x}| = 1$ which can be used to find the normalized time and distance. The normalized q_x unit is $L = 3.844 \cdot 10^5 km$, the time unit[4] is $T = 104h$ and therefore the velocity is $V = 1024 m/sec$. For comparison we calculate that the Earth - Sun restricted three-body problem has normalized units $L = 1.496 \cdot 10^8 km$, the time unit[5] is $T = 1395h$, and therefore $V = 29.8 km/sec$.

The Earth - Moon system has a very low eccentricity, and so is well approximated by the circular problem. An orbit which becomes a real mission is typically obtained first by such an approximate system, and then later refined

[3]Jacobi's constant is often defined as the negative of two times the energy. I have chosen to define it here $J = 2H$ to preserve its relation to energy.

[4]To find the time unit, we make use of the fact that $m_1 + m_2 = 1$ corresponds to the actual value of $5.9742 \cdot 10^{24} Kg + 7.3483 \cdot 10^{22} Kg$. We may then eliminate the mass and distance units from the gravity constant $G = 6.672 \cdot 10^{-11} m^3 Kg^{-1} sec^{-2}$ which also corresponds to one.

[5]Now we have $m_1 + m_2 \approx m_1 = 1.9891 \cdot 10^{30} Kg$.

through more and more precise models which eventually include small effects such as the influence of other planets, solar wind, high altitude atmospheric drag, etc. In addition, there is a limited precision to which a rocket can be positioned and thrusted thus requiring occasional corrective manoeuvres. With this in mind, (3.29) is considered a good starting model [22, 87].

3.3 Hill's Curves of Zero Velocity - Accessible Regions

Although long time solutions of (3.29) are difficult to obtain except for certain special cases (i.e. the Lagrange points). Hill first realized that the form of the equation for J implies a restriction on the configuration space in which a particle with fixed energy may visit [44]. Within these bounds, the curves of zero velocity, Hill's technique has little to say about exactly where the particle may be found. Nonetheless, this represents an important achievement as he was able to show, for example, that the Moon (the small mass!) is permanently bounded to the Earth in the Earth - Sun - Moon three body system as approximated by (3.29). We will also see in this section that the idea may be modified to find an upper bound on accessible phase space.

A requirement on the velocity (u, v) is that it remain real valued. Therefore $u^2 \geq 0$ and $v^2 \geq 0$. At the boundary of the closed set of allowed states in configuration space, $u = v = 0$ and so following Eq. (3.30) Hill's curve of zero velocity, for fixed J, bounds Hill's region, given by the relation

$$C + (x^2 + y^2) + 2(\frac{m_1}{r_1} + \frac{m_2}{r_2}) \geq 0, \tag{3.31}$$

where we take C as a fixed value of J (a negative quantity).

In Fig. (**??**) we see the surface over (q_x, q_y) given by $J(q_x, q_y, 0, 0)$. Each of the five Lagrange points are critical points on this surface. L_1, L_2, and L_3 can be found at the saddle points along the line through the two masses on

the $q_y = 0$ axis, and L_4 and L_5 are located at the maximum points located symmetrically off the $q_y = 0$ axis.

A slice at constant C yields the curve bounding the accessible region for a given orbit. We see that a low value of C causes the slice to cut through the wells of the two primaries. A point starting near a primary stays within a "circle" around the primary forever,[6] which we see portrayed in Fig. (3.3a).

Increasing C to $J_2 = -3.1883...$ causes the circles to enlarge (the wells widen) until they touch at L_2.[7] For a slightly higher value of J, a particle may pass between the two primaries through the neck, and is bounded within a "dumbbell" shaped region (see Fig. (3.3b)). We also see in a configuration space plot Fig. (3.6) showing that an orbit with such a value of Jacobi's constant stays within the dumbbell shaped region.

figure=Hill-0.eps,height=1.5in

Figure 3.3. a: A sketch of a slice of the zero velocity surface for a value of J below the level at which the two primaries are accessible. A particle starting near one of the primaries at this energy level is deep down the primary's well and inaccessible to the other primary. b: The critical value J_2 at which the two primaries become accessible. The two regions connect through a "neck" at the critical point of the zero velocity surface at L_2.

For a still higher value of $C = J_1 = -3.174...$, we reach the "top of the hill" at the second saddle point L_1 on the outside of the smaller primary. Now the dumbbell opens, allowing access to the region away from the primaries. A particle with this higher level of energy is no longer bounded to the primary system for long times.

[6]The surface also intersects the plane of constant C well away from both primaries. Thus a particle starting outside that "circle" stays outside. The region between is inaccessible for that value of C as that would require imaginary velocity.

[7]At the libration point L_2, we see a saddle on the $C(x, y)$ surface where a fixed value of C allows the "neck" to open.

The targeting technique of the previous chapter calls for a long orbit which explores accessible phase space. We choose J larger than the critical value J_2 at which the neck opens [22], so that the spacecraft may drift between the primaries. We must also choose J less than the critical value J_1 where the spacecraft may escape as we require a long bounded orbit for the library.[8] This long orbit may explore an accessible phase space which includes proximity to both primaries. The range of $J_2 \leq J \leq J_1$ within which we will work turns out to be quite small because the ratio $\frac{m_1}{m_2}$ is small.[9]

A curve of zero velocity may also be found on the section $q_y = 0$, $v > 0$. We may modify the expression for J to yield u as a graph over q_x. Again assuming that $v = 0$ at the boundary of real valued states, solving (3.30) for u yields

$$|u| \leq \sqrt{|J + x^2 + 2(\frac{m_1}{|x + m_2|} + \frac{m_2}{|x - m_1|})|}. \tag{3.32}$$

Fig. (**??**) is the plot of this graph for $J = -3.177$ in which we see that the neck is open, but the orbit is bounded.[10] This region we see closely approximates the upper bound on the region of accessible phase space revealed by Fig. (3.1).

3.4 From Flow to Map, Special Issues

In this section, we will discuss how to accurately integrate (3.29) between sections and through the singularity at close approaches to m_1 or m_2. Then we will find the stable and unstable directions of the full four-dimensional flow, restricted to the two-dimensional surface of section.

[8] We are also motivated to choose a small value of energy J since we wish to find a low energy orbit.

[9] However, $\frac{m_1}{m_2}$ for the Earth Moon system is in fact the largest such ratio in our solar system.

[10] The orbit on the section is unbounded in velocity u, as arbitrarily close approaches are possible which cause correspondingly large increases in velocity. Nonetheless, the particle must remain bounded in q_x within the curves of zero velocity.

To find the map, we integrate the flow until we cross the section in a given step. Then by "backing-up" we shoot at the surface with smaller and smaller values of h until $|q_y| < 10^{-16}$. A cleaner alternative would be to change the independent variable in (3.29) from t to q_y, so that each integration step automatically increments q_y, and we may find $q_y = 0$ exactly. For (3.29) however the scheme is not so pleasing and the shooting method works quite well.

We see that if $r_1 \ll 1$ (or $r_2 \ll 1$), then (3.29) becomes singular. Conservation of energy implies that the velocities become correspondingly large. Hence, special precautions must be taken.

If possible, it would seem to be favorable to make use of the symplectic structure definitive in Hamiltonian systems such as (3.27). For example, the "leap-frog" integrator calls for adding the potential change and then the momentum change in successive half steps to exactly conserve a nearby Hamiltonian of a discretized separable Hamiltonian (see [107]).

Unfortunately, symplectic integrators are not well suited to stiff problems such as those derived from $\frac{1}{r}$ potentials. The dynamics of the Hamiltonian that the discrete scheme conserves in fact may not be very close to the original dynamics.[11] The singularities make the use of a fixed step size of even a high degree scheme eventually inaccurate for a close enough approach (which typically occurs, eventually, along a general orbit). Meanwhile, a small step size may not

[11]The discretization may in fact "add new dynamics" if too large a step size is chosen. A nice example [105] is given by the standard map (1.37) which is a first order leap-frog discretization of a simple pendulum flow, $H = \frac{1}{2}p^2 + \epsilon\cos\theta$. Alternatively one can use the discretized Hamiltonian $H = \frac{1}{2}p^2 + 2\pi\delta_{2\pi}(\Omega t)\epsilon\cos\theta$. It is straight forward to exactly integrate the momentum change through the delta functions, which yields the standard map where the parameter is $k = (\frac{2\pi}{\Omega})^2\epsilon$. If the time between the delta function potential bursts is very small, then k is small, and so we know that the standard map phase space is filled with KAM circles which bound a given numerical solution and keep it from from wandering far from the corresponding continuous solution. However, for larger values of k, in particular for $k > k_c \approx 0.971$, solutions may wander tremendously.

be so well suited for a "flat" potential found away from the primaries. Hence, an adaptive scheme is called for. There is, however, no adaptive symplectic integrator [89]. A variable step size integrator, which we would require here, would specify the step size as a function of the position, $dt(\boldsymbol{z})$. This we see alters the Jacobian along an integrated orbit perhaps resulting in non-conservation of the symplectic forms.

Instead, an adaptive RK4 [31] scheme can be used to conserve J within a preset tolerance. We wish to find J to within 10^{-12} for each iteration flowing between section piercings, keeping in mind that an error in J relates to an error in u on the section according to

$$|\delta u| \leq \frac{2\delta J}{J}, \tag{3.33}$$

found by implicit differentiation of (3.30) and realizing that the maximum variation in u occurs when $\delta q_x = \delta q_y = 0$ and $\delta v = 0$. We analogously find a similar expression for δq_x. If a given step size h is insufficient to conserve J, then we halve h until the tolerance is met. Because RK4 has an $O(h^5)$ error term, each halving of h results in decreasing the error term by at least a factor of 32.

We still need a good first guess at the step size to minimize time spent testing overly large stepping attempts. To this end we follow an idea in [47] to choose each new integration step size as

$$h = c \min_{i=1,2} \frac{|r_i|}{\sqrt{u^2 + v^2}}, \tag{3.34}$$

where the authors advise that one arbitrarily choose $c = 0.1$.

Typically, these procedures might be enough for many tasks, but we need to integrate (3.29) for 10^5+ iterates of the map, where each iterate represents many integration steps, while conserving J to a high degree of accuracy over the entire orbit.

To further improve accuracy we make use of a manifold correction technique [76] to take a "Newton" step back onto the manifold of constant J after a (sequence of) integration step(s) moves us off. Referring to Eq. (3.30), we may define a function f as follows,

$$f(\boldsymbol{w}) = u^2 + v^2 - (q_x^2 + q_y^2) - 2(\frac{m_1}{r_1} + \frac{m_2}{r_2}) - J = 0, \tag{3.35}$$

where $\boldsymbol{w} = (q_x, q_y, u, v)$ is a point exactly on the manifold of fixed J. Rather than finding $\boldsymbol{w}$ on the manifold, a numerical error yields $\boldsymbol{w}'$ which is on another J' manifold. We may write an error vector $\boldsymbol{e}$, which brings us from $\boldsymbol{w}'$ back to $\boldsymbol{w}$.

$$\boldsymbol{w} = \boldsymbol{w}' + \boldsymbol{e}. \tag{3.36}$$

Using Eq. (3.35), we may expand a Taylor's series of f to first order around $\boldsymbol{w}'$ as follows,

$$f(\boldsymbol{w}) = f(\boldsymbol{w}') + \boldsymbol{e} \cdot \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{w}') + O(|\boldsymbol{e}|^2), \tag{3.37}$$

from which we can solve for $\boldsymbol{e}$, the correction back to $\boldsymbol{w}$.

Nonetheless, it is difficult to integrate through the singularities for too close an approach. There exists an exact coordinate transformation to a regularized[12] coordinate system, due to Birkhoff, which eliminates the singularities at the primaries. However it can be extremely cumbersome due to the

[12]Regularization is a class of exact coordinate transformations which eliminate a removable singularity of a point mass, sometimes at the cost of creating new singularities, such as at infinity in the case of the Birkhoff transformation. The change of coordinates can be cumbersome to work with. We can get the flavor of regularization by considering the case of a straight line two body problem $\ddot{x} = \mp 1/x^2$. A change of the independent variable t to a "slow time" variable τ can be made where the old and new times are related differentially by $dt = d\tau/x$. Performing the change of coordinates requires liberal use of the chain rule. The Birkhoff transformation removes simultaneously both singularities of the restricted three-body problem. Details of this and other regularizations can found in Szebehely.

appearance of fractions with high degree rational terms in the vector field (see Szebehely [100]).

Close approaches the the center of masses are not physical because the surface of the Earth is at a radius of 0.0156 normalized units above its center and that of the Moon is at 0.0035 above its center. Integrating long orbits of (3.29) is mathematically interesting regardless of the approach, however not important when we consider our final goal here. I find that my integrator with the above special improvements allows me to integrate quickly and to within the required accuracy for an approach of $\sim 10^{-6}$ to a center. When we convert this value to real units, we find that it implies less than a one hundred meter approach to the center, which is not practical.

In generating the long test orbit, we avoid the issue of close approaches with the following scheme. Whenever an approach closer than $10^{-4}L$ is detected, we "backup" 25 iterates, and make a 10^{-8} perturbation in a random direction on the surface. The procedure is used to build a 10^5 iterate pseudo-orbit, with several small errors, which conserves the integral J to within 10^{-12}. Note that the test orbit needs to be generated only once and then stored as a "library" of known behaviors. It may be used for generating paths for any mission in the accessible phase space.

Our targeting algorithm requires us to find the stable and unstable directions at a point. This requires that we integrate the variational equations of (3.29), written in the form $\dot{\boldsymbol{w}} = F(\boldsymbol{w})$, which are

$$\dot{\delta\boldsymbol{w}} = DF(\boldsymbol{w}_b) \cdot \delta\boldsymbol{w}. \tag{3.38}$$

Evaluating the derivative $DF(\boldsymbol{w}_b)$ along the base flow $\boldsymbol{w}_b$, (3.38) yields the total change of a variation $\delta\boldsymbol{w}(t_0)$ at $\boldsymbol{w}_b(t_0)$ until the finish of integration $\delta\boldsymbol{w}(t_f)$ at

$\boldsymbol{w}_b(t_f)$. The variation for an iterate of the Poincaré map is found simply by integrating (3.38) from an initial $\boldsymbol{w}_b(t_1)$ on the section, and a variation $\delta\boldsymbol{w}(t_1)$ on the section to $\boldsymbol{w}_b(t_2)$, the next iterate on the section, yielding $\delta\boldsymbol{w}(t_2)$. A basis of variations at t_1 may be evolved to t_2 to generate the matrix $M(t_1, t_2)$. This fundamental solution of (3.38) which is often called the "monodromy" matrix serves the role to evolve an arbitrary variation at $\boldsymbol{w}_b(t_1)$ to its image at $\boldsymbol{w}_b(t_2)$. We find the stable and unstable manifolds at a nonperiodic point analogously to the power method based technique defined by Eqs. (2.12)-(2.13).

Since the stable and unstable directions found, by integrating along the base flow $\boldsymbol{w}(t)$ from section to section, have components not on the surface, we must project the flow from a perturbation of the base curve back onto the surface. Projecting the (un)stable direction of the full 4-dimensional dynamics back along the flow onto the surface of section recovers the (un)stable direction of the 2-dimensional Poincaré map. The linear approximation of this statement is portrayed in Fig. (3.5), and expressed by the first order equations

$$\begin{aligned}
\delta q_x &= s_{q_x} - (\frac{\dot{q_x}}{\dot{q_y}})|_w s_{q_y}, \\
\delta \dot{q_x} &= s_{\dot{q_x}} - (\frac{\dot{u}}{\dot{q_y}})|_w s_{q_y}, \\
\delta \dot{q_y} &= s_{\dot{q_y}} - (\frac{\dot{v}}{\dot{q_y}})|_w s_{q_y}, \\
\delta q_y &= 0,
\end{aligned} \tag{3.39}$$

where the stable vector is written in the form $\boldsymbol{s} = (s_{q_x}, s_{q_y}, s_{\dot{q_x}}, s_{\dot{q_y}})$, and $\dot{\boldsymbol{w}} = (\dot{q_x}, \dot{q_y}, u, v)$ denotes the vector field of the base point $\boldsymbol{w}_b$ at time t. These formulas follow immediately from the approximation that the perturbation $\epsilon\boldsymbol{u}$ flows back along the vector field locally along straight lines for small enough ϵ. Eqs. (3.39) are written in slope intercept form on each of the three cross

sections, and may be represented succinctly by the vector equation

$$\boldsymbol{S} = \boldsymbol{s} - (\frac{s_{q_y}}{\dot{q_y}})\dot{\boldsymbol{w}}. \tag{3.40}$$

Eq. (3.40) has the alternative geometric interpretation that on each of the four projections, the length l along the flowback vector $\dot{\boldsymbol{w}}$ such that the vector sum $\boldsymbol{s} - l\dot{\boldsymbol{w}}$ is in the q_y section may be found by similar triangles. The right triangle labeled $< a, b, c >$ where $a = \boldsymbol{w}_b + \boldsymbol{S}$, $b =$ the projection of $\boldsymbol{w}_b + \boldsymbol{s}$ onto $q_y = 0$ (the height), and $c = \boldsymbol{w}_b + \boldsymbol{s}$, is similar to the triangle outlined by the incidence of the flow labeled $< d, e, f >$ where $d = \boldsymbol{w}_b$, $e =$the projection of $\boldsymbol{w}_b + \dot{\boldsymbol{w}}$ onto $q_y = 0$ (again the height), and $f = \boldsymbol{w}_b + \dot{\boldsymbol{w}}$. Hence by elementary Euclidean geometry the hypotenuses of the two triangles are proportional to their heights.

$$l = \frac{s_{q_y}}{\dot{q_y}}. \tag{3.41}$$

figure=flowback.eps,height=1.5in

Figure 3.5. A projection in the q_x, q_y plane of the flowback of a perturbation along the stable manifold back along the vector field $\dot{\boldsymbol{w}}$ onto the section $q_y = 0$.

In Fig. (3.5), we see that on the (q_y, q_x) cross section, the point (s_{q_x}, s_{q_y}) away from the base point $\boldsymbol{w}$ flows back to the section $q_y = 0$ at $(\delta q_x, 0)$ with the vector field slope $\frac{1}{m} = (\frac{\dot{q_x}}{\dot{q_y}})$ according to the slope intersect formula $s_{q_x} = (\frac{\dot{q_x}}{\dot{q_y}})|_w u_y + \delta x$ from which the formulas (3.39) follow immediately. Hence, we find that the stable direction on the section is $(\delta q_x, \delta \dot{q_x})$. Invisible to the map, but necessary to integrate the flow, we set $\delta q_y = 0$ on the section and $\delta \dot{q_y}$ is fixed by (3.30) for constant J. The unstable direction of the map can be found analogously.

3.5 A New Orbit to The Moon

In this section we will now target a chaotic orbit to the Moon [13]. In evaluating this new orbit, the goal is to beat the energy requirements of the standard Hohmann transfer from a parking orbit around the Earth to a parking orbit around the Moon. This transfer typically takes only a few days, depending on the altitude of the initial parking orbit. It requires two large rocket thrusts (perturbations), one parallel to the motion to leave the Earth, and one anti-parallel to the motion to capture the rocket around the Moon. The size of these perturbations, measured by the velocity boost ΔV, depends again on the altitudes of the Earth and Moon orbits. Fig. (3.6) shows our reference Hohmann

figure=Hoh.ps,height=2.5in

Figure 3.6. A Hohmann transfer from the parking orbit at the location of $\boldsymbol{a}$ near the Earth to the location of $\boldsymbol{b}$ near the Moon. Flight time between the two orbits is approximately 6.67 days and requires $\Delta V = 1250$ m/sec specific impulse. Note that the first manoeuvre is parallel the Earth parking orbit, the second is anti-parallel to the Moon parking orbit, and the spacecraft is allowed to drift between.

transfer orbit. We will see that the chaotic orbit will eliminate the need for the large deceleration at the Moon and reduce required initial boost.

Of course, the required energy for the transfer is $J_2 = -3.1883$, which is that of the Lagrange point L_2. This is the minimum energy for which an orbit could possibly move between the primaries. For our mission we set $J = J_0 = -3.177$ slightly above J_2, but below the critical value J_1 at which orbits may escape, so that we may have a long bounded test orbit. This energy we imagine is attained by an impulsive boost, ΔV, of a spacecraft in a parking orbit around the Earth to the energy J_0. Fig. (3.1) shows a phase space plot of a single "chaotic" test orbit with 10^5 iterates. This test orbit may be stored as

a "library" of known behaviors, and used for generating many missions in the accessible portion of phase space. Certain islands in phase space are inaccessible; these are bounded by invariant tori, some of which are shown. At the center of each island is a periodic orbit.

We choose the point $\boldsymbol{a} = (x_0, u_0)$ to achieve a fast chaotic orbit. A trial and error search for various x_0 along the line segment $u_0 = 0$ gave the best results for an orbit at an altitude of 57310 km above the Earth's center. As our target, we choose the outermost invariant torus, marked "b" in Fig (3.1), corresponding to a quasi-periodically precessing "ellipse" around the Moon. For the actual target point $\boldsymbol{b}$, we use the point of closest approach of our test orbit to $\boldsymbol{b}$, at an altitude of 13970 km above the Moon's center. From $\boldsymbol{b}$ a tiny perturbation will move the orbit onto the torus, thus achieving a state bound to the Moon without the large deceleration required by a Hohmann transfer.

Our targeting algorithm calls for finding the stable and unstable directions of a point. This requires that we integrate the variational equations, $\dot{\delta \boldsymbol{w}} = DF(\boldsymbol{w}) \cdot \delta \boldsymbol{w}$. Numerical errors in both the calculation of the stable (unstable) directions and the integration of the flow, both of which require many integration steps between surface intersections, limit the accuracy of the calculation to about 25 iterates, and the total compression to factors ranging from 10^2 to 10^3. This therefore limits the extent to which a recurrence δ can be compressed to meet a tolerance ϵ, and the length of the patch, $2m + 1$. We choose to make the patch length 25 steps by setting $m = 12$. A better integrator would presumably allow us to choose a larger patch, thus allowing smaller ϵ.

The 10^5 iterate test orbit has a 15037 iterate segment which goes from $\boldsymbol{a}$ to $\boldsymbol{b}$. By fixing the recurrence distance at $\delta = 0.01$, which considering the size of (X, L) is quite large, we achieved a 117 iterate pseudo-orbit by cutting out

4 recurrence loops and requiring a maximum perturbation of $\epsilon = 2.14 \times 10^{-4}$. Note that this implies perturbations to the real coordinates of $\delta x \leq 82.3$km and $\delta u \leq 0.219$m/sec. The actual time along this orbit is $T = 325.6 = 3.87$ years.

Standard optimal control theory can be applied to our pseudo-orbit to find the orbital maneuvers to transfer between each segment. Finite parameter optimization theory techniques can be used to determine the optimal solution (in a local sense) for the locations and times of the deterministic ΔV manoeuvres between nearby orbit segments. The *strength* of our algorithm is to find a globally improved path. Finding these "drift" orbit segments solves what is known as Lambert's problem. Once found, we have a first guess to prime an optimization technique to find the local optimum.

figure=orbit.to.moon.ps,height=4.0in

Figure 3.7. A configuration space plot (q_x, q_y) of the first ten iterates (dashed) around the Earth and the last twenty iterates (solid) of the 117 iterate transfer to the Moon. The final state at $\boldsymbol{b}$ is a precessing ellipse around the Moon corresponding to the targeted invariant torus.

The first several (dashed) and last several (solid) rotations of our orbit are shown in configuration space in Fig. (3.7). On the solid segment, we can see the accelerating boosts of the Moon's gravitational pull as the earth orbiting spacecraft swings by (in the inertial frame). These boosts perturb the spacecraft into just the right orientation to pass through the neck around L_2 exactly once with the correct speed and position so that it is captured by the Moon near the chosen invariant torus. In Fig. (3.8) we see the phase space plot of the 117 iterates between $\boldsymbol{a}$ and $\boldsymbol{b}$. Of interest to note is the lack of recurrences of this fast orbit as it avoids getting trapped in any given resonance layer. We saw exactly the same property in the fast transporting standard map orbit displayed

in Fig. (2.2). In a sense these fast transporting orbits are the opposite of an orbit which displays any given frequency.

figure=otm.eps,height=3.5in

Figure 3.8. A phase space plot $(q_x, \dot{q_x})$ of the 117 iterate orbit to the Moon. This is the restricted three-body analogue of Fig. (2.2). Note that the fast orbit spends a minimum of time in resonance layers, which would be characterized by recurrences, which we have removed.

The boosts required for our chaotic trajectory can be compared to those of a corresponding Hohmann like, two impulse transfer (the classic mission). Both orbits start at the (almost circular) parking orbit around the Earth at the starting altitude 57310 km with Jacobi's constant $J = -7.431084$. An initial impulsive thrust is required for both transfers to increase the energy such that the zero velocity curves permit the transfer, $J > J_2$. The chaotic transfer requires an initial boost of $\Delta V = 771.8$ m/sec to attain $J_0 = -3.177$. Additionally, it requires 4 patches with $\epsilon \leq 2.14 \cdot 10^{-4}$, and therefore the total change in velocity is bounded by $\Delta V \leq 4 \cdot 0.219$m/sec $= 0.876$m/sec. Finally, to jump from $\boldsymbol{b}$ to the targeted invariant torus requires $\epsilon = 9.96 \cdot 10^{-4}$ and therefore $\Delta V \leq 1.02$m/sec. Thus the total perturbation required by the chaotic transfer is 773.7m/sec. By contrast, the Hohmann transfer requires an initial parallel burn of $\Delta V = 846.6$m/sec boosting the energy to attain $J = -2.747$. At this energy level, there are no longer any zero velocity curves restricting the motion. The spacecraft is then allowed to coast until it arrives near the Moon parking orbit at the same altitude as $\boldsymbol{b}$, where a deceleration of $\Delta V = 403.5$ m/sec is applied. Therefore the total boost required for this Hohmann transfer is 1250.1 m/sec, but the transfer requires only 6.67 days.

Therefore we find that the ratio between the impulses is 1.616, or a

38% advantage over the Hohmann orbit. This is a significant improvement, but at the cost of a much longer (and circuitous) transfer. In terms of transferring passengers, the extra time is probably not worth the savings. However, for transferring freight, the ΔV savings of our orbit translates directly to a considerably smaller fuel requirement and therefore allows the transfer of a larger payload.

Recently, another approach due to Belbruno was used to find chaotic transfer orbits to the moon utilizing the so called "fuzzy boundary" [5, 6]. This method was successfully applied to send the spacecraft Hiten to the Moon, thus saving an otherwise failed mission when the original Moon probe was lost. The Hiten orbit requires a restricted four-body model, including the Sun. The technique is to send the spacecraft to the fuzzy boundary between the Earth and Sun, where their gravitational effects balance, so that only a small perturbation is necessary to reach the Moon in a "ballistic capture orbit". This orbit is much less circuitous than our transfer Fig. (3.7) and requires approximately 4.6 months. However, a larger rocket burn ΔV is required to escape the Earth in order to reach the fuzzy boundary, well away from the Earth - Moon zero velocity curve at J_c.

The technique developed here could also be applied to the restricted four-body problem (with the added complication that the dimension of the phase space is increased since time cannot be eliminated by going to a rotating frame), to provide a systematic method for finding fast orbits in this case as well.

CHAPTER 4

A CLOSED FORM OPTIMAL ORBIT

4.1 A Symbol Dynamics Description of Optimal Chaotic Transport

In Sec. 1.2.3 we saw that there is a conjugacy between $T|_\Lambda$, the invariant set of a horseshoe arising from a transverse homoclinic intersection, and $\alpha|_\Sigma$ the left shift map on the space of all bi-infinite sequences of 0's and 1's. In this section we will discuss optimal paths of chaotic targeting for the model dynamics $\alpha : \Sigma \to \Sigma$. In the next section we will then examine the cut and glue optimizing technique specifically applied to this model.

Our problem is as always to find a fast epsilon chain path from a to b with small epsilon. Given a and b with specific symbol representations:

$$a = ...\sigma^a_{-2}\sigma^a_{-1}.\sigma^a_0\sigma^a_1\sigma^a_2... \tag{4.1}$$

and

$$b = ...\sigma^b_{-2}\sigma^b_{-1}.\sigma^b_0\sigma^b_1\sigma^b_2..., \tag{4.2}$$

we wish to find a point $a_\epsilon \in B_\epsilon(a)$, defined in terms of the Σ topology. Here, the topology is induced by the norm as usual: $B_\epsilon(a) \equiv \{z : \|a - z\|_\Sigma < \epsilon\}$. The definition of

$$\|a - b\|_\Sigma = \sum_{i=-\infty}^{\infty} \frac{\delta_i(a,b)}{\lambda^{|i|}}, \tag{4.3}$$

where

$$\delta_i(a,b) = \left\{ \begin{array}{l} 0 \text{ if } \sigma^a_i = \sigma^b_i \\ 1 \text{ if } \sigma^a_i \neq \sigma^b_i \end{array} \right\}, \tag{4.4}$$

implies that two symbol sequences are close if they agree on their "middle sections" near the decimal point. It follows that for any particular choice of ϵ, there exists the corresponding k, where $k \geq -\frac{\ln \epsilon}{\ln \lambda}$ such that the symbols of $z \in B_\epsilon(a)$ must agree with at least the $2k+1$ symbols of a before and after the decimal point. This correspondence lets us write "$B_k(a)$."

As is well known, there is a point d dense in Σ and nonperiodic. We use this as the test point required by the recurrence cutting algorithm. Here, d is the library of known behaviors of orbits encoded in 0's and 1's. We choose to write d in the form

$$d = ...0.1011011000100010001... \tag{4.5}$$

which consists of all permutations of n symbols, followed by all permutations of $n+1$ symbols, etc., for all n. This particular choice of d is obviously nonperiodic. We see that d is dense in Σ since for any ϵ (with corresponding k) and z, $\exists$ an m such that $\alpha^m(d) \in B_k(z)$. The $2k+1$ "centered" symbols of z must be amongst the $2k+1$ permutations of d, and will be shifted to the decimal point ("now") under enough applications of the map α.

It is convenient to define a truncation operator

$$\hat{z}^k = \sigma^z_{-k}\sigma^z_{-(k-1)}..\sigma^z_{-1}.\sigma^z_0\sigma^z_1..\sigma^z_k, \tag{4.6}$$

which can be used to define an equivalence class amongst bi-infinite sequences as those points with the same "k hat" truncation. Such points are in the same "k ball." In terms of targeting, and epsilon chains, it is the equivalence class in which we are most interested. After all, a path is said to reach b if it gets close enough, and the path consists of a pseudo-orbit which is exact in terms of the equivalence class.

We now have the tools to address the issue of the optimal path. Given

ϵ, the control tolerance threshold, we have the corresponding k, in targeting bi-infinite sequences. Then by definition of the point d, there exists a time $m_a \geq 0$ such that m_a applications of the shift operator causes the point d to iterate to a in terms of "k hat" equivalence classes. Hence

$$\widehat{a}^k = \widehat{\alpha^{m_a}(d)}^k. \tag{4.7}$$

Likewise, there exists a time $m_b \geq 0$ when

$$\widehat{b}^k = \widehat{\alpha^{m_b}(d)}^k. \tag{4.8}$$

Note that d recurs with a and b infinitely many times since all sequences are found as subsequences amongst all permutations of longer sequences. Nonetheless m_a and m_b do have finite values since d contains all subsequences in increasing order of complexity by construction, including the $2k+1$ symbol length sequences $\widehat{a}^k$ and $\widehat{b}^k$. We wish to find an orbit from a and then to b; hence we consider only $m_b > m_a$.

The optimal time of transport for the orbit of d, in terms of k hat equivalence classes, is

$$n = \min |m_b - m_a| \tag{4.9}$$

which is simply the smallest integer amongst the finite values of m_b and m_a.

Theorem 4.1.1 *Given a control tolerance $\epsilon > 0$, and therefore the corresponding $k > 0$ symbol space tolerance, with implied $\widehat{\cdot}^k$ operator, the point $a_\epsilon \in B_\epsilon(a)$ with the fastest orbit to near b has the form*

$$a_\epsilon = \widehat{a}^k\widehat{b}^k = ...\sigma^a_{-k}..\sigma^a_{-1}.\sigma^a_0\sigma^a_1..\sigma^a_k\sigma^b_{-k}..\sigma^b_{-1}\sigma^b_0\sigma^b_1..\sigma^b_k... \tag{4.10}$$

and requires $2k+1-o$ applications of the shift map, where o is the number of symbols which agree at adjoining ends of $\widehat{a}^k$ and $\widehat{b}^k$.

Proof: All points in $B_\epsilon(a)$, have the same $\widehat{a}^k$, and so agree in the $2k+1$ symbols around the decimal. Likewise, all $b_\epsilon \in B_\epsilon(b)$ have the same $\widehat{b}^k$. Shifting the $\widehat{b}^k$ in $\widehat{a}^k\widehat{b}^k$ to the present by applying α^{2k+1} causes the $\widehat{a}^k$ part to be "forgotten" in the k equivalency. This represents a path between $\widehat{a}^k$ and $\widehat{b}^k$ which is the fastest such path possible in general since $\widehat{a}^k\widehat{b}^k$ contains the minimum information necessary to represent the two neighborhoods. There can however be the special case of a faster path when an 0 symbol overlap between the two ends of the form $\{\sigma_i^a\}_{i=k-o+1}^k = \{\sigma_{-i}^b\}_{i=k-o+1}^k$ exists. Hence, $4k+2-o$ symbols are sufficient to represent a_ϵ which iterates to b_ϵ in $2k+1-o$ shifts, where o is the number of overlap symbols. □

Note that neighborhoods of the horseshoe are squares, which can be seen by examining what the conjugacy $h : \Sigma \to \Lambda$ does to neighborhoods. The k symbols to the right of σ_0^a restrict the location of a point to lie in a λ^{-k} wide vertical strip. The k symbols to the left of the decimal further restrict the point to also lie in a λ^{-k} wide horizontal strip. So the $2k+1$ symbols in $\widehat{a}^k$ tell us in which $\lambda^{-k} \times \lambda^{-k}$ box, amongst the 2^{2k+1} choices our point lies. The $2k+1$ symbols to the right of σ_k^a specify the added information as to which $\lambda^{-[k+(2k+1)]}$ wide vertical strip inside the $\widehat{a}^k$ box our fast transporting orbit lies.

In the next section we will examine the cut and glue strategy at recurrences as formulated for the symbol dynamics. We will then compare results to the optimal orbit which we now know for this simple dynamics.

4.2 Cutting Recurrences in Σ

We now reprise the cut and glue process to targeting in Σ. Just as in Sec. 2.2.3, we will start with a long chaotic test orbit. Given an arbitrary $k \geq 0$, the orbit d, defined in Eq. (4.5) will eventually visit $B_k(a)$ at time m_a and then

$B_k(b)$ at time m_b according to Eqs. (4.7) and (4.8). We rewrite d so as identify the inclusion of the symbols of $\widehat{a}^k$ and $\widehat{b}^k$ somewhere in its length.

$$d = ... \overbrace{\sigma_{-k}^{a}..\sigma_{k}^{a}}^{\widehat{a}^k} .. \overbrace{\underbrace{\sigma_{-m}..\sigma_{m}}_{\text{s iterate loop}} ... \sigma_{-m}..\sigma_{m}}^{\text{2m+1 symbol recurrence}} .. \overbrace{\sigma_{-k}^{b}..\sigma_{k}^{b}}^{\widehat{b}^k} ... \tag{4.11}$$

Let N be the (first) number of iterates along d to shift from $\widehat{a}^k$ to $\widehat{b}^k$. An s iterate recurrent loop is also explicitly shown, where s left shifts brings the same $2m+1$ symbol subsequence to the center. This recurrence is $\delta = 2m+1$ symbols close. Hence there exist a point $\tilde{d}$ whose orbit agrees with that of d between $\widehat{a}^k$ and $\widehat{b}^k$, but skips the s step recurrent loop. Only a $\delta = 2m+1$ perturbation is required at the first time the recurrence is encountered to perturb from d to $\tilde{d}$ to save s iterates. In terms of the horseshoe, this corresponds to a perturbation within the $\lambda^{-m} \times \lambda^{-m}$ box implied by the $2m+1$ centered symbols.

In the notation of Chapter 2, to cut and glue a delta recurrence requires us to shoot from $f^u(z_{i-n})$ to $f^s(z_{i+s+n})$. We need to exercise care to correctly interpret these expressions in symbol dynamics since d represents the point and its orbit at the same time. Only the location of the decimal distinguishes z_i from z_{i+s}. Let us define z_0 to be the m_a^{th} iterate of d when $\widehat{a}^k$ is prominent, z_N the m_b^{th} iterate when $\widehat{b}^k$ is prominent, z_i when the $2m+1$ symbol recurrence first occurs, and z_{i+s} when the $2m+1$ symbols last occur.[1] Define i as the time (left shifts) to cause $\widehat{z}_i{}^k$ to have the decimal point centered, starting from $\widehat{z}_0{}^k = \widehat{a}^k$. Let us choose n, the patch length, so as to not alter the endpoints $\widehat{a}^k$ and $\widehat{b}^k$. Therefore n may be chosen $n + m = \min\{i, N - (i+s)\}$.

We may define a patch orbit d' such that $\widehat{d'}^{n+m}$ agrees with the $n+m$

[1] The phrases "now", "centered", and "prominent" tell us that given a point, (i.e. d) the decimal point is centered so as to make a given iteration of the point prominent. Similarly, the symbols of a given point along an orbit "occur" at the appropriate iteration when those symbols are centered.

symbols before the recurrence and the $n+m$ symbols after the recurrence. We see that shooting is quite simple in symbol space. By altering n symbols of d before $\widehat{z}_i^m$ and n symbols after $\widehat{z_{i+s}}^m$, we have a patch d' such that only an $\epsilon = \delta\lambda^{-n}$ perturbation is required at z_{i-n} to jump onto the patch d' and then another ϵ perturbation $2n$ iterates later to jump back onto d after the recurrence (near z_{i+s+n}). As before, we demand

$$\epsilon > \delta\lambda^{-n}, \tag{4.12}$$

which we now rewrite in symbol space, making use of the norm based conjugacy,

$$\lambda^{-k} \geq \lambda^{-m}\lambda^{-n}. \tag{4.13}$$

This translates to the following form, relating recurrence size and the required patch length to meet a desired tolerance, all in terms of the symbol count,

$$k \leq m + n. \tag{4.14}$$

This allows us to investigate the best path achieved in general by cutting and gluing all recurrences possible to the optimal transport orbit written in Eq. (4.10). To cut all recurrences "possible" we need to set the recurrence δ as large as possible so that no possibilities are missed, an issue we have discussed previously (see Table (6.6)). We consider a trivial recurrence of only one symbol. Say we have a 0 symbol at time i, and then another 0 symbol at some time $i+s$ later. Such a one symbol recurrence is guaranteed, allowing for recurrences of 1's as well. Hence $m = 1$, and so according to Eq. (4.14), the patch length n must be at least $k-1$. We see that removing the trivial recurrence leads to at best a $2k$ step orbit segment (patch) between $\widehat{a}^k$ and $\widehat{b}^k$. The conclusion then is that recurrence removal, using cut and glue to satisfy the tolerance, results in at

best an orbit which is $2k$ iterates longer than the optimal achieved by the point $\widehat{a}^k\widehat{b}^b$ as in theorem 4.1.1.

4.3 Validity of the Model

Unfortunately, the simple horseshoe dynamics is not such a good model for general transport even in two dimensions, except for special examples. We know that heteroclinic cycles have a horseshoe structure [104]. A heteroclinic cycle is defined by the heteroclinic orbit between the points of a periodic hyperbolic orbit. In another example, Mackay [57] explicitly writes down a form of the standard map and restricts it to a deformed square to demonstrate the horseshoe structure when $k \geq 2\pi$. It may even be possible to demonstrate other restrictions of the standard map to subsets with the horseshoe structure for compositions of the map for $k > k_c$. The horseshoe is the prototypical example of chaotic dynamics. The shift is a forgetting process.

The horseshoe, however, is only valid for the map T restricted to a subset of the phase space Λ with zero measure. The Arnold cat map is an example where the full phase space is hyperbolic, and there exists a conjugacy to a symbol dynamics valid for the full phase space. (Hence the cat map is defined as Bernoulli.) We certainly want a more general description of transport, at least on a subset of full measure. Demonstrating a conjugacy to the horseshoe implies hyperbolicity, but Λ is not necessarily the largest hyperbolic subset of the phase space; nor is hyperbolicity a requirement to discuss transport.

It is often possible to describe the dynamics of a homoclinic tangle not representable by a shift on 0's and 1's by a more general grammar of a finite set of symbols. Likewise, higher-dimensional dynamics can be conjugate to a symbol dynamics of several symbols. It may be possible to find a conjugacy of

the map with a symbol dynamics for a subset of chaotic phase space containing many a and b general targeting goals.

Suppose a and b are contained in a subset of T's phase space which is conjugate to a more general symbol dynamics with a finite grammar. It is likely that a version of the optimal symbol path theorem 4.1.1 is still valid with trivial modification. The main conclusion here is that the optimal symbol orbit from a to b with precision k is the minimum number of symbols required to represent a and then b. The recurrence skipping algorithm does not yield this optimal orbit in general, but always yields a very fast path between these two states.

CHAPTER 5

CONTROL WITHOUT GLOBAL MODELING

5.1 A New Twist on the Old Problem

In previous chapters we targeted the standard map, the restricted three-body problem, and dynamics conjugate to the horseshoe. We used the method of controlling through recurrences, which has two key elements. First, there is the identification of switching points with recurrences. Second, there is either a patching or a stabilization step, both of which we perform by shooting at the stable manifold. The first step makes no use of the analytical representation of the map. Although our technique for the first step requires observation of a long test orbit so that we may infer the switching points, once we have collected the test orbit data, the map is not needed. However, the second step, shooting at the stable manifold, utilizes the map to advance successive trials and the derivative of the map to find the linear approximations of the (un)stable manifold.

In this chapter we will play a new game. We will reprise the problem of targeting the standard map. This time however, we assume that we do not know the map in the form Eq. (1.37), and therefore neither do we know its derivative. Instead, we will put the test orbit data to a second use. We will reconstruct the map by nearest neighbor piecewise linear models. This relies on the availability of enough data to insure that any predicted orbit will have enough close neighbors so that the linear approximation is valid. The chaotic test orbit will hopefully be long enough and visit enough of its accessible phase space so

that good predictions can be made. It must be cautioned that the conclusion of this chapter is somewhat negative because the techniques developed here require too much resources in general as I will discuss further in the final section.

5.2 Time Series Delay Embedding

In real world applications, one does not always have the dynamics available in the form of a flow (1.1) or a map (1.2). Without knowledge of the correct dimension of the vector $\boldsymbol{z}$ and which quantities to measure, it is still possible to infer topological information pertaining to the ω-limit[1] set of the flow from the measurement of a single smooth scalar function of the state vector,

$$x(t) = f(\boldsymbol{z}(t)). \tag{5.1}$$

In 1980, Packard *et al.,* [81] proposed that the "time series embedding" vector,

$$\boldsymbol{X}(t) = (x(t), x(t-\tau), x(t-2\tau), ..., x(t-(D-1)\tau)), \tag{5.2}$$

for a delay time τ, yields a state vector which (as Takens [101] proved) generically provides an orbit homeomorphic to the orbit of $\boldsymbol{z}(t)$. Takens' argument suggests that the choice of time lag τ is not important for almost all τ (τ not rationally related to periods of orbits on the ω-limit set). The choice of delay however is quite important in practice.

Almost all choices yield a topologically valid coordinate system. Yet we see that if τ is quite small, then the dynamics have little time to evolve and so coordinates are nearly singular:

$$x(t) \approx x(t-\tau) \approx ... \approx x(t-(D-1)\tau). \tag{5.3}$$

[1] A sequence $\{\boldsymbol{z}_n\}_{n=0}^{\infty}$ has a limit point $\boldsymbol{z}$ iff there exists a subsequence $\{\boldsymbol{z}_{n_k}\}$ such that $\boldsymbol{z}_{n_k} \to \boldsymbol{z}$ as $n_k \to \infty$. The ω-limit set is defined as the set of all limit points of $\{\boldsymbol{z}_n\}$.

Similarly if τ is chosen too large, each of the coordinates becomes completely uncorrelated. An embedding dimension D, larger than $2r+1$, is sufficient where r is the box counting dimension of the attractor (see [101, 90]).

An interesting problem is the following: given a time series (from a black box) deduce the minimal embedding dimension D and the "best" delay time τ. This turns out to be a rather difficult problem to define and solve. An entire "embedding community" has arisen with literature devoted to this and the related problem of noise reduction and filtering. A comprehensive overview of some of the many approaches can be found in the review article by Abarbanel *et al.* [1]. In particular, I would like to mention the work of Broomhead and King [17] who by singular value decomposition compute the spectrum of the covariance matrix $(x_i(t)x_j(t))$. The k^{th} eigenvalue is the average root mean square projection of the D-dimensional time series onto the k^{th} eigenvector. Their numerical calculations demonstrate that by ordering the eigenvalues according to size, one finds that after an exponential fall off, a floor value is reached due to the projection of the dynamics onto the higher-dimensional noise. Their method is to discard all the projections below the noise floor by a Gram-Schmidt process, thus automatically keeping the cleaned data projected onto the D axes. Another well known competing approach due to Fraser and Swinney [33] is to maximize the mutual information inherent in the data. They argue that their information theoretic approach is more valid than diagonalizing the covariance matrix.

In this chapter I will not worry about the many issues involved with finding a good embedding of a time series and filtering noise. Instead I will imagine that we already have available a sequence of vectors $\boldsymbol{X}_i$ which may have been derived from a delay embedding where $\boldsymbol{X}_i = \boldsymbol{X}(t+i\tau)$, a Poincaré section,

or directly from a map of a chaotic dynamics[2]. The rules of this game dictate that we do not know the underlying dynamics generating the time series, but that we may nonetheless assume that there is a deterministic dynamical system which generated the data. The goal will be to predict the dynamics in order to infer short orbits. To simulate this game, I will generate a 10^6 iterate orbit of the standard map. After storing just the x values, the game is to repeat Sec. 2.2 without referring to the analytic form of the map Eq. (1.37) or its derivative. Instead, we can predict these quantities.

5.3 Prediction

In Weigend and Gershenfeld [103] we find a distinction between weak modeling (data-rich and theory-poor) and strong modeling (data-poor and theory-rich). This is related to "...the distinction between memorization and generalization..." It is always nice to have a general theory from which we may write down a global set of equations of motion. However, this is not always necessary.

Predicting the future evolution of dynamical systems has been a main goal of scientific modeling for centuries. The classic approach has been to build a global model based on fundamental laws yielding a differential equation or otherwise describing the motion of the states. "This requires strong assumptions. A good fit of the data to the model validates the assumptions." [103].

In the absence of a good model, accurate predictions may be made using a large collection of data. Given the sequence of state vectors $\{\boldsymbol{X}_i\}_{i=1}^N$, which evolves according to some (unknown) dynamics

$$\boldsymbol{X}_{i+1} = T(\boldsymbol{X}_i), \tag{5.4}$$

[2]Actually, for our limited application $\boldsymbol{X}$ will have be the x components of an orbit of the standard map. It is generally accepted [33] that the correct delay for a map is one, which we will adopt for our "model" problem.

we may approximate the function T at $\boldsymbol{X}_i$ *locally.* The map T is replaced by a patchwork of "charts." Local modeling depends on finding k near neighbors to $\boldsymbol{X}$. The "first order approximation," according to the classification of Farmer and Sidorowich [30] the simplest local modeling technique called the "method of analogs", approximates that given the nearest neighbor $\boldsymbol{X}_j$ to $\boldsymbol{X}_i$, that $\boldsymbol{X}_{j+1}$ is the prediction for $\boldsymbol{X}_{i+1}$. This is how weather prediction according to tables was performed for years. A higher level of accuracy and sophistication is gained by the "second order approximation" to T locally by the affine model

$$\boldsymbol{X}' = T(\boldsymbol{X}) \approx \boldsymbol{A} \cdot \boldsymbol{X} + \boldsymbol{b}. \tag{5.5}$$

$\boldsymbol{X}$ evolves according to an interpolation of the evolutions of its nearest neighbors. Given $k \geq D+1$ nearest neighbors, we may least squares fit the Jacobian matrix $\boldsymbol{A}$ and the shift vector $\boldsymbol{b}$ by minimizing the chi-square function

$$\chi^2(A, \boldsymbol{b}) = \sum_{j=1}^{k} \|\boldsymbol{A} \cdot \boldsymbol{X}_j + \boldsymbol{b} - \boldsymbol{X}_{j+1}\|^2, \tag{5.6}$$

for each of the $j \leq k$ near neighbors.

First let us investigate how to least squares fit a linear model of the form

$$\boldsymbol{u}' = \boldsymbol{u} \cdot \boldsymbol{M}. \tag{5.7}$$

Given $\boldsymbol{X}$ we wish to predict its local linear behavior, i.e. the Jacobian of T near $\boldsymbol{X}$. In this form, $\boldsymbol{M}$ is the transpose of the Jacobian matrix. We use this transposed form of the usual equations to emphasize that $\boldsymbol{u}$ and $\boldsymbol{u}'$ are the known data and we wish to solve for the unknown $\boldsymbol{M}$. We require $\{\boldsymbol{X}_i, \boldsymbol{X}'_i\}_{i=1}^{k+1}$, the $k \geq D$ near neighbors and their iterates from which we form the variation vectors $\{\boldsymbol{u}_i, \boldsymbol{u}'_i\}_{i=1}^{k}$ that we define as the row vectors

$$\boldsymbol{u}_{i-1} \equiv (\boldsymbol{X}_i - \boldsymbol{X}_1)^T,$$

$$\boldsymbol{u}'_{i-1} \equiv (\boldsymbol{X}'_i - \boldsymbol{X}'_1)^T. \tag{5.8}$$

The matrix $\boldsymbol{M}$ must simultaneously fit each of these k transformations from $\boldsymbol{u}_i$ to $\boldsymbol{u}'_i$ as equations of the form (5.7) which may be "stacked" as k row vector equations by writing the matrix equation

$$\overset{k}{\downarrow}\left[\begin{array}{c} \overset{D}{\rightarrow} \\ \boldsymbol{u}'_i \end{array}\right] = \overset{k}{\downarrow}\left[\begin{array}{c} \overset{D}{\rightarrow} \\ \boldsymbol{u}_i \end{array}\right] \cdot \left[\begin{array}{c} \overset{D}{\rightarrow} \\ M_{i,j} \end{array}\right] \overset{D}{\downarrow}. \tag{5.9}$$

If $k = D$, then the least square problem reduces to an exact change of basis whose solution is $\boldsymbol{M} = [\boldsymbol{u}]^{-1} \cdot [\boldsymbol{u}']$, providing the $\boldsymbol{u}_k$ are linearly independent.

figure=trans.eps,height=2.0in

Figure 5.1. Nearby $\boldsymbol{X}_1$, DT rotates the variation vectors $\boldsymbol{u}_i$ to $\boldsymbol{u}'_i$. Given at least D transformations $\{\boldsymbol{u}_i, \boldsymbol{u}'_i\}$, the Jacobian $DT \in \mathbb{R}^{D\times D}$ may be least square fitted.

Usually, least squares linear modeling in the literature (e.g., see Stoer-Bulirsch [97]) is presented for a vector equation of the form

$$\left[\boldsymbol{u}'\right]_{\boldsymbol{j}} = [\boldsymbol{u}] \cdot [M_{i,j}]_{\boldsymbol{j}}\,, \tag{5.10}$$

where the symbol "$[\ ]_{\boldsymbol{j}}$" denotes the j^{th} column vector of the bracketed matrix. The solution of Eq. (5.10) in the least square sense

$$\min \sum_{i=1}^{k} \| \left[\boldsymbol{u}'\right]_{\boldsymbol{j}} - [\boldsymbol{u}] \cdot [M_{i,j}]_{\boldsymbol{j}} \|^2, \tag{5.11}$$

strives to find the minimizing column vector of parameters $[M_{i,j}]_{\boldsymbol{j}}$. The minimum is stationary, thus

$$\nabla_M((\left[\boldsymbol{u}'\right]_{\boldsymbol{j}} - [\boldsymbol{u}]\cdot[M_{i,j}]_{\boldsymbol{j}})^T \cdot (\left[\boldsymbol{u}'\right]_{\boldsymbol{j}} - [\boldsymbol{u}]\cdot[M_{i,j}]_{\boldsymbol{j}})) = 2\,[\boldsymbol{u}]^T\cdot[\boldsymbol{u}]\cdot[M_{i,j}]_{\boldsymbol{j}} - 2\,[\boldsymbol{u}]\cdot\left[\boldsymbol{u}'\right]_{\boldsymbol{j}} = 0, \tag{5.12}$$

yielding the "normal equations,"

$$([\boldsymbol{u}]^T \cdot [\boldsymbol{u}]) \cdot [M_{i,j}]_{\boldsymbol{j}} = [\boldsymbol{u}]^T \cdot \left[\boldsymbol{u}'\right]_{\boldsymbol{j}}\,, \tag{5.13}$$

with the solution

$$[M_{i,j}]_{\boldsymbol{j}} = ([\boldsymbol{u}]^T \cdot [\boldsymbol{u}])^{-1} \cdot [\boldsymbol{u}]^T \cdot [\boldsymbol{u}']_{\boldsymbol{j}} \,. \tag{5.14}$$

Note we can see that the D least squares fitting problems for each of the $[M_{i,j}]_{\boldsymbol{j}}$, $j = 1, 2, .., D$ all rely on the same $[\boldsymbol{u}]$ data (called the "design matrix" in *Numerical Recipes* [31]) and independent $[\boldsymbol{u}']_{\boldsymbol{j}}$ data (called the "measured values of the dependent variable"). Each of the D solutions in the form Eq. (5.14) may be stacked side by side to form the full solution to Eq. (5.11) written compactly as a single matrix equation,

$$\boldsymbol{M} = ([\boldsymbol{u}]^T \cdot [\boldsymbol{u}])^{-1} \cdot [\boldsymbol{u}]^T \cdot [\boldsymbol{u}'] \,. \tag{5.15}$$

We may similarly solve the affine least squares problem Eq. (5.6) which we see requires at least one more data vector to fit the extra parameter vector corresponding to the shift $\boldsymbol{b}$. The following augmentation by a column of "1's" allows for the shift $\boldsymbol{b}$ to be incorporated into a linear model yielding an equation analogous to Eq. (5.9),

$$\overset{k+1}{\downarrow}\left[\begin{array}{c} \overset{D\ dim}{\rightarrow} \\ \boldsymbol{X}_i'^T \end{array}\right] = \overset{k+1}{\downarrow}\left[\begin{array}{cc} & \overset{D+1}{\rightarrow} \\ 1 & \boldsymbol{X}_i^T \\ \vdots & \vdots \end{array}\right] \cdot \left[\begin{array}{c} \overset{D}{\rightarrow} \\ N_{i,j} \end{array}\right] \overset{D+1}{\downarrow} , \tag{5.16}$$

where $\left[1\ \boldsymbol{X}^T\right]$ is the design matrix, $\left[\boldsymbol{X}'^T\right]_{\boldsymbol{j}}$ are the measured values of the dependent vectors, and $[N]_{\boldsymbol{j}}$ are the vectors of parameters to be fitted for $1 \leq j \leq D+1$. The least squares solution to this equation follows similarly to that for (5.9) and has normal equations analogous to (5.15).

In theory, we may solve the normal equations in straight forward manner by LU decomposition. However, according to the authors of *Numerical Recipes* [31], "..the solution of a least squares problem directly from the normal

equations is rather susceptible to roundoff error." There may exist several different fits of the model to the data, all of which are close to optimal; this causes singular normal equations. Singular value decomposition ("SVD") is well suited for dealing with sets of equations that are numerically almost singular and over determined. Thus SVD is the preferred solution to our problem; an appropriate routine SVDFIT is supplied in *Numerical Recipes*, which we need to apply for each j.

Higher order models other than those mentioned above may be fitted by least squares regression as long as the fitting parameters are the quantities of a linear combination of basis functions. A trade off exists (see [29]) in that high order polynomials give better predictions when fitted over data in a given neighborhood size. However, since more near neighbors are required to determine the increased number of parameters of a high order polynomial, the neighborhood size tends to increase.

An advantage of polynomial models is that one may perform an error analysis of their likely accuracy based on a Taylor series of the map T around the point of prediction (see [1, 30]). An important issue for us to decide is how to iterate a prediction which we require to shoot from the unstable manifold before a recurrence to the stable manifold after the recurrence. Skipping the lengthy details which may be found in the above citations, I would nonetheless like to mention the results comparing the worthiness of "direct forecasting," (a single model is made directly to the k near neighbors to predict the evolution under T^n in one step) in contrast to "iterative forecasting," (n models of T are composed to evolve a prediction for $\boldsymbol{X}$ under T^n). The leading order error is the $(q+1)^{th}$ term of the Taylor series for a q^{th} degree polynomial interpolation. Several assumptions are used including that the difference between the forecast orbit and

the true orbit separate approximately as the most unstable Lyupanov exponent; thus $< DT^n > \sim \exp(\lambda n)$ where n is the expansion time (or prediction interval). Farmer and Sidorowich argue that higher order derivatives scale similarly, $< D^q T^n > \sim \exp(q\lambda n)$. Making further assumptions on the average data density for an ergodic map based on a data set size N and attractor dimension d_A yields an RMS error estimate for a direct forecast,

$$E \sim N^{\frac{-q}{d_A}} \exp(q\lambda n). \tag{5.17}$$

Contrast this to the iterative forecast error

$$E \sim N^{\frac{-q}{d_A}} \exp(\lambda n), \tag{5.18}$$

in which there is absent the q term in the exponent. Therefore, iterative forecasts can be expected to give better predictions. This is sensible in that iterating n one step models allows for the information from up to nk different near neighbors (k new near neighbors for each of n iterates), whereas a direct forecast allows for only k near neighbors.

We also predict the (un)stable directions with the iterative process of composing the tangent maps attached along an orbit by replacing predictions of the Jacobian matrices for the exact matrices specified in the defining equations for the (un)stable directions along a nonperiodic orbit, Eqs. (2.12)-(2.13), following a suggestion of Eckmann and Ruelle [26].

Since we wish to apply prediction near recurrent points, we are in the fortunate situation that there will be near neighbor data where we need it (assuming a long enough test orbit). Another issue of concern is that there will only be near neighbors in regions of phase space accessible to the test orbit. A prediction in some other region of phase space is impossible, no matter how

long the test orbit. This potential problem does not arise because fast orbits optimizing the slow test orbit are in the same accessible phase space, and this is where shooting by prediction will be required.

For the *on the fly* control performed in Sec. 2.2.9, we additionally require information pertaining to the parameter derivative $\frac{\partial T}{\partial k}$. A minimum model of this requires a difference quotient between predictions of models of two time series data sets, one for each parameter value k_l, k_h where $\Delta k = |k_h - k_l|$ is small. Call $PT_k(\boldsymbol{X})$ the predicted value of $T_k(\boldsymbol{X})$. In [91] the authors modeled the parameter derivatives of a one-dimensional map by difference quotients between independent predictions formed from different data sets measured at several nearby parameter values, $k_l < ... < k_i < k_{i+1} < ... < k_h$. The idea extends directly to higher-dimensional predictions where we may approximate

$$\frac{\partial T}{\partial k}(\boldsymbol{X})|_{k_i} \approx \frac{PT_{k_{i+1}}(\boldsymbol{X}) - PT_{k_i}(\boldsymbol{X})}{k_{i+1} - k_i}. \tag{5.19}$$

For $\frac{\partial T}{\partial k}(\boldsymbol{X})|_k$ where $k_i \leq k \leq k_{i+1}$, a linear interpolation should yield a reasonable answer for Δk small. Of course, more sophisticated models with several different parameter valued data set predictions using a higher order polynomial interpolation should yield improved accuracy.

A philosophical note is in order on the validity of using predictions to find paths which may be longer than our ability to make a prediction with confidence. Recall that for $k = 1.25$ the fastest orbit we found was $n = 131$. Fig. (5.3) (below) implies that about twenty iterates is the longest time we can expect accurate predictions (for a million iterate data set). Extrapolating the error growth rate, we expect errors of order one (the size of the phase space) by about fifty iterates. Nonetheless, the idea is valid since we do not predict an entire path. What we do is skip long recurrent loops of a known path. The

predictions are only between sections of known orbit. Between patches, the original exact path is used. For the nonoverlapping patch technique described in Sec. 2.2.7, the known test orbit is always available as the reference from which to make relatively short predictions for the patches.

5.4 Targeting Revisited - Numerical Results

We have seen above how we may replace the map and derivative with predictions formed from observations of the test orbit which we already have available. We may now repeat controlling chaos through recurrence.

First we must address a slightly contrived problem to which I propose an equally slightly contrived solution. Since we know that the phase space topology of the standard map is the torus[3] which lives in $\mathbb{R}^3$, we know that we generally require a sufficient embedding dimension $D = 2 \cdot 2 + 1$. However, as we will discuss in more detail in the next section, computer memory restrictions make it difficult to achieve the data density required for good predictions, as outlined by Eq. (5.18). Lower-dimensional embeddings alleviate this problem which motivates us to make use of our special knowledge about the form of the standard map phase space which we know to be the torus. The range of the map (1.37) is a unit square $(x, y) \in [0, 1) \times [0, 1)$. Therefore the range of the time series "embedding" of the x variable is $(x_i, x_{i-1}) \in [0, 1) \times [0, 1)$.

Local linear model predictions as described in the previous section can cause the iterate of a predicted point to leave the unit square under the mapping twist. On the next iterated prediction, we find an unfortunate situation where the closest k "neighbors" to the point $\boldsymbol{X}$ are on the nearest edge of the square.

[3]or the cylinder depending on whether we modulus the momentum. There is an aliasing effect in the momentum direction which allows us to identify 2π segments of momentum.

Without some insight into the submanifold's topology the computer program may (and did for a certain "buggy" version I made) attempt to fit a linear model to these "neighbors" and then proceed to make an absolutely meaningless prediction.

Recall that the modulus function was used in the standard map to "roll" the unit square into a torus, so we never see the third dimension of this two-dimensional submanifold. Another choice of measurement function such as $y = f(x) = cos(2\pi x)$ would preserve the natural periodicity of the map. A physical example of this would be the measurement of the angle θ, the rotation of a pendulum arm. It has values $\theta \in \mathbb{R}$, if we keep track of winding numbers, but there is a natural 2π periodicity. If instead we measure a value y as the projection of the tip of the pendulum arm (of length l), i.e., $y = lcos(2\pi\theta)$, then y never experiences the discontinuity at the boundaries $\theta = 2k\pi$. Typically, we would never encounter this problem for an attractor derived from a real world time series for a generic measurement function. With this in mind, the predictions which follow were made as described above, followed by an application of the modulus function. If we nonetheless want to know the topology of the attractor using the time series embedding, there is available a technique of Muldoon *et al* [75] which totals the Euler index using a complex of symplices.

figure=embedd.eps,height=3.0in

Figure 5.2. Left: A phase space plot of the canonical variables (x_i, y_i) of the standard map at $k = 1.25$ for a 10^6 iterate test orbit. The range is $(x, y) \in [0, 1) \times [0, 1)$. Right: A delay plot of the same orbit using the coordinates (x_{i-1}, x_i) in place of the canonical variable.

In Fig. (5.2) we see a delay plot (x_i, x_{i-1}) of the standard map data side by side with the usual phase space data (x_i, y_i) for a 10^6 iterate test orbit

with parameter value $k = 1.25$. We can see clearly that, albeit distorted, both plots display the same structures or holes which correspond to elliptic islands inaccessible to our long chaotic test orbit. Besides being promised by embedding theorems, it should come as no surprise that x data alone is enough exhibit the topological structures of the standard map when we recall that orbits are described by stationary action $\delta W(x_0, x_1, ..., x_n) = 0$ which depends only on x data, as described in Appendix A.2.

Fig. (5.3) shows a graph of the error of a predicted orbit of (x_{i-1}, x_i) against the prediction interval n for several data set sizes including $N = 10^3, 10^4, 10^6$. For short time predictions, smaller data set sizes give good results. However, we can see that for a reasonable prediction error of $\sim 10^{-5}$ we do require the full one million data points. This algorithm is indeed data hungry!

In Chapter 2, we set the recurrence threshold to a relatively high value $\delta = 0.1$ in order to guarantee that no opportunity to remove a patchable recurrence was missed, at the cost of making more unsuccessful trials. Table (2.1) displayed the trade off between δ, patching success, and the resulting orbit length. Now the cost of unsuccessful trials is much higher; predicted shooting and predicted calculation of local linear manifolds f_u and f_s is computationally intensive. In addition, numerical inaccuracies become more of a problem; patching may now fail due to the predicted orbit failing to land on the predicted stable manifold. We choose here to set $\delta = 0.02$ to reduce work at the cost of a possibly longer path. The goal is to achieve an $\epsilon = 0.005$ tolerance for $k = 1.25$. Following Fig. (5.3), we choose the patch size $2m+1 = 21$, a length for which we expect reasonable predictions. This reduced space to contract to ϵ was another motivating factor in choosing the smaller δ. Examining Fig. (2.7), revealing a rough estimate for the Lyupanov contraction rate λ, we see that this is almost

the best contraction for which we can hope.

figure=predict.ps,height=3.0in

Figure 5.3. The error of a predicted orbit as a function of time for the standard map, $k = 1.25$. The top two curves are for $N = 10^3$ and 10^4 respectively, while the bottom solid is for $N = 10^6$.

The best path found takes 618 steps requiring 5 patches with up to 0.003 perturbations from patch to orbit. A second path of 760, and a third of 1024 were found. These computations require several computer days. Compare this to version in Chapter 2, using the analytic map, which took tens of seconds. Besides the successful paths, there were many trials where I was forced to kill the job after several days under the persuasions of other users of the computer lab who were unhappy of my use of too much CPU and all the memory. In the next section, I will discuss reasons for the difficulty of prediction control, especially in higher dimensions, and the lessons learned.

5.5 Conclusions - Memory Limitations

In the end, time series prediction as described above is more akin to memorization than modeling. My original intention on undertaking the work of this chapter was to build a generic black box predictor controller. To this end, I programmed all of the prediction models above in general-dimensional form and built correspondingly general code. The hope was that once the code was written, it could just churn out controlled paths. This thesis was to read as a list of problems to which it applied, but there was to be no such luck.

Let us first consider what can be done. In [91], Shinbrot *et al.* targeted a magnetoelastic ribbon which they found could be represented by a one-dimensional map. Here $\boldsymbol{X}_{n+1}$ is a graph over $\boldsymbol{X}_n$. The authors made good predictions by spline fitting 500 observed data points. Parameter variations

were predicted as described above. Recall for a moment the logistic map example in Sec. 1.1.3 where we saw that a small parameter perturbation quickly fills the entire accessible one-dimensional range. Such is typical of one-dimensional maps. The magnetoelastic ribbon was successfully controlled in real time.

The two-dimensional standard map above requires a two-dimensional model. We used 10^6 points to guarantee sufficient data density, a little more than the square of the number of points used for sufficient one-dimensional density. This is exactly as we should expect.

A basic definition relating dimension and volume is

$$V = L^d. \tag{5.20}$$

This tells us how many parts on a side L fill a d-dimensional volume. In particular an average distance ϵ between data points, assuming uniform density, translates to

$$N = \epsilon^{-d}, \tag{5.21}$$

points are required to fill a unit hypercube. So in terms of a given data accuracy ϵ for good predictions, we expect the number of points required for enough two-dimensional coverage to be roughly the square of the one-dimensional value. This trend is bad news for higher-dimensional prediction control.

Consider that on the workstation that I use, 8 bytes are required to represent one double precision entry $\boldsymbol{X}_n$ in a time series. The five hundred points required for a 1-D model fit in requires 40Kbytes; no problem for an average PC. The one million points for a 2-D model fit in 8 Mbytes, which fits nicely on a work station. A corresponding 3-D model would require about $(10^6)^{\frac{3}{2}} = 10^9$bytes= 1Gbytes. Computers this big exist now, but are very expensive and hard to find. Continuing, we see that a four-dimensional model

requires 1000 Gbytes, bigger than anything around now, but still conceivable. A ten-dimensional model requiring 10^{30}bytes= 10^{21} Gbytes is out of the question. I was forced to stop with two-dimensional prediction control because that is all that fits on the computers to which I have access. Clearly the brute-force memorization approach will almost surely never be workable for anything over five (or certainly six?!) dimensions.

This story is not all bleak. There may be special cases for which higher-dimensional prediction is possible. A typical special case occurs when the ω-limit set lives on a lower-dimensional attractor of dimension d_A. In that case the data requirement Eq. (5.21) depends on d_A. To give an example, I was able to make pretty good predictions for the Lorenz attractor which lives in $\mathbb{R}^3$, but has a box counting dimension only slightly bigger than 2. Memory requirements are similar to those for the standard map. Another dynamical system I worked with is the kicked double rotor map,[4] [86]. The phase space is four-dimensional, but its ω-limit set lives on a 2.8-dimensional attractor. I was able to make reasonable predictions (but not accurate enough for control) for this map as well.

It is fairly common for a high-dimensional dynamics to contract to an attractor with a much smaller box counting dimension, often small enough that we can expect to make good predictions on available computers. The box counting dimension d_A speaks to us of the set covered and therefore predictability, given N. However, it does not tell us about the dynamics on the d-dimensional manifold. For example, it turns out that the kicked double rotor has a two-dimensional unstable manifold and a two-dimensional stable manifold at a given point [51]. In this case, shooting from the unstable manifold before a recurrence

[4]The double rotor models a two armed pendulum under a kick potential with dissipation. It is a four-dimensional generalization of the dissipative standard map.

to the stable manifold after the recurrence requires a more general form than Eq. (2.10) to include the extra parameters. A discussion of the issues involved in shooting between higher-dimensional saddles can be found in Appendix A.4.

The transport mechanism in higher dimensions is not clear. We cannot be as sure that if an orbit goes from point $\boldsymbol{a}$ to point $\boldsymbol{b}$ that it was necessarily localized in an intermediate turnstile. We saw in Sec. A.5 that there can be higher-dimensional analogues of turnstiles, but the general case is not as clear. Nonetheless, we have the recurrence lemma 2.2.1 for dynamics on a compact manifold of arbitrary dimension. Thus we expect that long orbits will recur with themselves, and so if they are part of a hyperbolic set, they are likely shadowed by much faster orbits that skip the recurrent loops. However, we expect that recurrences are less frequent along such orbits since exponentially more pigeonholes are needed in higher dimensions.[5] Experimenting with the double rotor map does seem to reveal recurrences, albeit less frequently. Let us also realize that in higher dimensions an orbit having no recurrences is not necessarily optimal. In the end, there may be substantial savings as a result of removing all unnecessary recurrences even if this does not always yield the optimal orbit.

We can see that there are severe limitations to the amount of memorization which can be done for higher-dimensional dynamics. For the special case that there is a sufficiently low-dimensional attractor, prediction control may still be workable. Higher-dimensional targeting still requires some straightforward extensions of already developed algorithms to account for generalized saddles. Putting this all together can be the subject of some interesting future

[5] In terms of probabilities of a recurrence, it is the box counting dimension d_A, not the manifold dimension d which is important, good news for targeting a dynamics which collapses significantly.

research.

CHAPTER 6

BREAKUP OF INVARIANT TORI FOR THE FOUR-DIMENSIONAL SEMI-STANDARD MAP

6.1 Introduction

Stability of motion in Hamiltonian systems and symplectic mappings is of great interest in many physical situations such as plasma and accelerator confinement and stellar and planetary dynamics; an understanding of stability is also of intrinsic theoretical interest. The primary stability result is the KAM theorem which asserts that most of the invariant tori of a nonlinear integrable Hamiltonian survive upon a *small*, smooth perturbation [2]. The robust tori, according to the theorem, are those that have sufficiently incommensurate frequency vectors (they satisfy a Diophantine condition, see Sec. 6.3). As a practical result, however, the KAM theorem has several drawbacks. The first is that estimates of the perturbation size for the destruction of tori are typically extremely small: much smaller than the size indicated by numerical computations for specific perturbations on specific tori (of course the theorem guarantees the survival of any Diophantine torus for any small enough perturbation). The second is that the theorem guarantees stability only for systems of two degrees of freedom since the invariant tori have half the dimension of phase space. Nonetheless, computations indicate that while a system of three degrees of freedom may not be rigorously stable, it exhibits a "practical stability" since orbits appear to remain trapped near invariant tori for extremely long periods. To some extent

this is addressed by the Nekhoroshev theorem [77], though this theorem requires extremely small perturbation sizes as well.

For the case of two degrees of freedom, or equivalently area preserving mappings, much progress has been made in determining the existence of invariant tori. Three basic techniques have been used. The first is to examine the stability of a sequence of periodic orbits whose frequencies limit on the irrational frequency of interest—this gives rise to the *residue criterion* [38, 58]. This technique yields extremely accurate values for the parameters at which invariant circles are destroyed and can be made rigorous [59]. The second method is a nonexistence criterion for twist maps, called converse KAM theory [68, 63]. The final technique is numerical computation (in some cases using interval arithmetic) of the conjugacy to pure rotation [82, 24, 9, 67]. These methods can give accurate nonrigorous values for the critical parameter for essentially arbitrary Diophantine frequencies, and can also give reasonable rigorous values.

Though many have attempted to generalize these techniques to Hamiltonian systems with more than two degrees of freedom, or equivalently, symplectic maps of four or more dimensions, there has been limited success in determining the existence of invariant tori. Periodic orbit approximations to invariant tori have been obtained [50, 74], and computations reveal that the stability domains of periodic orbits limiting on an incommensurate frequency vector may be converging for a volume preserving example [65]. However the existence of the limit is difficult to prove [8], primarily because the ordering property of periodic orbits on the circle no longer applies on the torus. Converse KAM theory can be generalized to higher dimensions [62], though in this case one must assume that the tori are Lagrangian graphs.

One of the fundamental problems in these studies is number theoretic:

there is no satisfactory generalization of the continued fraction theory to simultaneous approximation of several irrationals (perhaps the most promising is that of Brentjes [16]). In the case of the residue criterion, it is the best approximants (convergents of a continued fraction) whose properties converge to those of the invariant circle. Furthermore, quadratic irrationals play a large role in these studies because their continued fraction expansions are eventually periodic (these give rise to self-similar structures). Finally, the most robust tori appear to correspond to the class of quadratic irrationals known as the *noble* numbers; these have a continued fraction expansion with a tail of all one's. Roughly speaking, the explanation for this is that the noble numbers are the most difficult to approximate in the sense of Diophantine. The generalization of this class to higher dimensions is unknown.

There has been some speculation that for four-dimensional mappings, cubic irrationals will replace the quadratics. One reason for this is that a periodic approximation scheme based on a Farey tree construction necessarily leads to a frequency which is the eigenvalue of a 3×3 matrix, and is therefore cubic [41, 49]. However, even in this case it has been difficult to determine if there is self-similar behavior near breakup [65], and there is no evidence that cubic irrationals are more robust than others.

In this chapter we study the four-dimensional, complex, symplectic map corresponding to the coupling of two semi-standard maps, as introduced in [40]. This map is the complex version of a mapping introduced by Froeshlé [34, 50]—we call it the semi-Froeshlé map. We generalize the method of Percival and Greene [40] to this case and find recursion formulae for the Fourier coefficients of an invariant two torus with a fixed frequency vector in Sec. 6.4. Existence of such a torus for small enough parameter values is guaranteed providing the

frequency vector satisfies a Diophantine condition; we discuss this in Sec. 6.3. Because of the simple structure of the Fourier series for the semi-Froeshlé map, we are able to apply some results from the theory of holomorphic functions of several complex variables in Sec. 6.5, and show that the domain of convergence of the Fourier series has a particular form; it is complete and log-convex. Finally in Sec. 6.6 we compute these convergence domains for several example frequency vectors, including quadratic and cubic irrationals.

6.2 Coupling of Two Semi-standard Maps

The *semi-standard*, area preserving map was introduced by Greene and Percival [40] as a numerically simpler model than the standard map for the investigation of the analytic properties of invariant circles. In Lagrangian form, the semi-standard map takes $\{x_{t-1}, x_t\} \mapsto \{x_t, x_{t+1}\}$ and is defined by

$$\delta^2 x_t \equiv x_{t+1} - 2x_t + x_{t-1} = iae^{ix_t} \; ; \tag{6.1}$$

this is a map on $\mathbb{C}^2$. The notation δ^2 is reminiscent of the second derivative operator.

In this paper we study a four-dimensional generalization, analogous to the map introduced by Froeshlé [34, 50]. Letting $\boldsymbol{x}_t \in \mathbb{C}^2$, the semi-Froeshlé map is

$$\delta^2 \boldsymbol{x}_t \equiv \boldsymbol{x}_{t+1} - 2\boldsymbol{x}_t + \boldsymbol{x}_{t-1} = \mathbf{F}(\boldsymbol{x}_t) \; , \tag{6.2}$$

where

$$\mathbf{F}(\boldsymbol{x}) \equiv i \begin{pmatrix} a_1 e^{ix^{(1)}} & + & \epsilon e^{ix^{(1)}+ix^{(2)}} \\ a_2 e^{ix^{(2)}} & + & \epsilon e^{ix^{(1)}+ix^{(2)}} \end{pmatrix} \; . \tag{6.3}$$

There are three parameters, the strength of the kicks for each component semi-standard map (a_1, a_2) and ϵ, the strength of the coupling of the two maps.

Equation (6.2) is symplectic since $\mathbf{F}$ is the gradient of a scalar potential (see for example [50]).

We are looking for solutions $\boldsymbol{x}_t$ of Eq. (6.2) which lie on an invariant two-torus homotopic to the trivial torus defined by the momentum $\mathbf{y}_t \equiv \boldsymbol{x}_t - \boldsymbol{x}_{t-1}$ being constant. In fact, we demand that this torus be analytically conjugate to a uniform rotation on the angle variable $\boldsymbol{\theta}$ with a given frequency vector $\boldsymbol{w}$. These tori include those found by KAM theory. The conjugacy is represented by the following commuting diagram

$$\begin{array}{ccc} \boldsymbol{x}_t & \longrightarrow & \boldsymbol{x}_{t+1} \\ \downarrow & & \downarrow \\ \boldsymbol{x}(\boldsymbol{\theta}) & \longrightarrow & \boldsymbol{x}(\boldsymbol{\theta} + 2\pi\boldsymbol{w}) \; . \end{array} \tag{6.4}$$

Thus, for a given $\boldsymbol{w}$, an invariant torus for Eq. (6.2) is given by

$$\boldsymbol{x}_t = \boldsymbol{x}(\boldsymbol{\theta} + 2\pi\boldsymbol{w}t) \; , \tag{6.5}$$

for $\boldsymbol{\theta} \in \mathrm{T}^2$. The homotopy condition implies that

$$\boldsymbol{x}(\boldsymbol{\theta} + 2\pi\mathbf{m}) = \boldsymbol{x}(\boldsymbol{\theta}) + 2\pi\mathbf{m} \quad \forall \mathbf{m} \in \mathbb{Z}^2 \; , \tag{6.6}$$

thus $\boldsymbol{x}(\boldsymbol{\theta})$ is coperiodic with $\boldsymbol{\theta}$;

$$\boldsymbol{x}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{\chi}(\boldsymbol{\theta}) \tag{6.7}$$

where $\boldsymbol{\chi}(\boldsymbol{\theta})$ is doubly 2π periodic. If we suppose that $\boldsymbol{x}$ is analytic, it can be expanded in a Fourier series

$$\boldsymbol{x}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \sum_{\mathbf{n} \in \mathbb{Z}^2} \boldsymbol{\chi}_{\mathbf{n}} e^{i\mathbf{n}\cdot\boldsymbol{\theta}} \tag{6.8}$$

Inserting Eq. (6.5) into Eq. (6.2) yields the Percival form of the mapping

$$\delta^2 \boldsymbol{x}(\boldsymbol{\theta}) \equiv \boldsymbol{x}(\boldsymbol{\theta} + 2\pi \boldsymbol{w}) - 2\boldsymbol{x}(\boldsymbol{\theta}) + \boldsymbol{x}(\boldsymbol{\theta} - 2\pi \boldsymbol{w}) = \mathbf{F}(\boldsymbol{x}(\boldsymbol{\theta})) \; . \tag{6.9}$$

Inserting the series (6.8) into Eq. (6.9) will yield equations determining the Fourier coefficients $\boldsymbol{\chi}_{\mathbf{n}}$; these will be obtained in Sec. 6.4.

6.3 Incommensurate Frequencies

The convergence of the Fourier series for the semi-standard map has been studied extensively in [40, 82, 67]. In particular, rather sophisticated techniques for computing convergence of this series were developed in [82]; these give accurate results for quite general frequencies. In general one determines a parameter interval $|a| < a^{ss}(\omega)$ for which there is an analytic invariant circle with frequency ω. Here a^{ss}, the critical function, is zero for every rational value and exhibits a maximum for

$$\omega = \gamma \equiv \frac{1 + \sqrt{5}}{2} \; . \tag{6.10}$$

The critical function appears to have a local maximum at each of the *noble frequencies*: those equivalent to γ under a modular transformation, or equivalently that have a continued fraction expansion whose elements are all 1 beyond some level.

These results also apply to the semi-Froeshlé map when $\epsilon = 0$. Thus an invariant torus of frequency $\boldsymbol{w} = (\omega_1, \omega_2)$ exists within the rectangle $\{(a_1, a_2, \epsilon) : |a_1| < a^{ss}(\omega_1), |a_2| < a^{ss}(\omega_2), \epsilon = 0\}$. Furthermore, since the semi-Froeshlé map is an analytic perturbation of a twist map, KAM theory implies that for sufficiently small values of the three parameters (a_1, a_2, ϵ) there exists an invariant torus analytically conjugate to the rotation $\boldsymbol{\theta} \mapsto \boldsymbol{\theta} + 2\pi \boldsymbol{w}$ providing the frequency vector satisfies a Diophantine condition.

For d-dimensions, the set of Diophantine vectors $\mathcal{D}_\mu$ consists of those $\boldsymbol{w} \in \mathbb{R}^d$ for which there exists a $C > 0$ such that for all $(\mathbf{p}, q) \in \mathbb{Z}^{d+1}$

$$|\mathbf{p} \cdot \boldsymbol{w} - q| \geq \frac{C}{||\mathbf{p}||^\mu} \ , \tag{6.11}$$

where $||\mathbf{p}|| = \max(|p_1|, ..., |p_d|)$. It is easy to see that if $\mu > d$, the measure of $\mathcal{D}_\mu$ approaches one as $C \to 0$; however, the measure of $\mathcal{D}_d$ is zero.

Certainly if $\boldsymbol{w} \in \mathcal{D}_\mu$, then it is *incommensurate*, that is 1 and $\omega_1, ..., \omega_d$ are linearly independent over the rationals. For $d > 1$ one must distinguish between commensurate vectors and *resonant* vectors. While the former satisfy *some* rational relation $\mathbf{p} \cdot \boldsymbol{w} = q$, the latter have *all* components rational and correspond to periodic orbits. A straightforward generalization of Greene's method [38] to higher dimensions would use resonant vectors, e.g. [41], instead of commensurate vectors. However, in KAM theory it is commensurabilities which cause the problems, not just resonances.

Though there exist many Diophantine vectors, a result of Minkowski implies that every $\boldsymbol{w}$ can be closely approximated in a certain sense [18]:

Theorem 6.3.1 *For any $\boldsymbol{w} \in \mathbb{R}^d$ there are infinitely many integer vectors $(\mathbf{p}, q)$ such that when $K = 1$*

$$|\mathbf{p} \cdot \boldsymbol{w} - \mathrm{q}| < \frac{\mathrm{K}}{||\mathbf{p}||^{\mathrm{d}}} \ . \tag{6.12}$$

If $d = 1$ then K can be replaced by $1/\sqrt{5}$ but nothing smaller.

To our knowledge, the minimal value of K for $d > 1$ is not known.

One class of frequency vectors which are Diophantine are those constructed from algebraic irrationals [18]:

Theorem 6.3.2 *If the components of $\boldsymbol{w}$ are incommensurate and elements of a real algebraic field of degree $d + 1$, then $\boldsymbol{w} \in \mathcal{D}_d$.*

Recall that an algebraic field generated by $\xi \in \mathbb{R}$ of degree n is defined as the set of numbers of the form

$$R(\xi) = \frac{P(\xi)}{Q(\xi)}$$

where P and Q are polynomials of degree n with integer coefficients.

One would expect that a frequency vector $\boldsymbol{w}$ which is more incommensurate, in the sense of having a larger Diophantine constant C and smaller exponent μ would tend to persist for higher perturbations. This is numerically verified for the standard and semi-standard maps where the noble numbers give local maxima of a^{ss}, and are also the "most" irrational in the sense of Diophantine. Unfortunately, to our knowledge, there are no results in the theory of simultaneous approximations which determine a class of frequency vectors analogous to the noble numbers. Indeed one of the main reasons for our numerical investigation is to attempt to develop a technique for determining this class.

We will choose several simple frequency vectors as examples for our study. In addition to the golden mean, we will use the quadratic irrationals

$$\sigma \equiv \sqrt{2} = [2, 2, 2, 2....] \equiv [2^{\infty}] \; ; \quad \zeta \equiv \frac{1+\sqrt{2}}{5+4\sqrt{2}} = [0, 4, 2^{\infty}] \; . \tag{6.13}$$

The expressions on the right above give the continued fraction expansions. Setting $\boldsymbol{w} = (\gamma, \sigma)$ or (γ, ζ) yields two incommensurate frequency vectors since $\sqrt{\frac{5}{2}}$ is irrational. Furthermore by Theorem 6.3.2, both of these vectors are in $\mathcal{D}_2$, since they are elements of the algebraic field of degree three generated by $\xi = \sqrt{2} + \sqrt{5}$. This is easy to see, since any cubic polynomial in ξ has the form $P(\xi) = a + b\sqrt{2} + c\sqrt{5} + d\sqrt{10}$ for $a, b, c, d \in \mathbb{Z}$. Thus γ, σ, and ζ are all in $R(\xi)$.

Finally we consider a cubic irrational, the real solution of

$$\begin{aligned} \tau^3 &= \tau + 1 \\ \tau &\simeq 1.324717957244746025960908854478097 34 \\ &\simeq [1, 3, 12, 1, 1, 3, 2, 3, 2, 4, 2, 141, 80, 2, 5, 1, 2, 8, 2, 1, 1, 3, 1, ...] \ . \end{aligned} \tag{6.14}$$

This so called "spiral mean" frequency, was introduced in [49] as a possible analogue of the golden mean since in the Ostlund-Kim version of the Farey tree, τ has a simple periodic construction. The number τ is Diophantine since according to a theorem of Roth, every algebraic irrational is in $\mathcal{D}_{1+\delta}$ $\forall \delta > 0$. Thus the critical function $a^{ss}(\tau) \neq 0$; however, determining its value is difficult because the continued fraction elements appear unbounded [66]. Nonetheless, for the four-dimensional case we will study the vector (τ, τ^2) which is in the cubic field generated by τ, and so an element of $\mathcal{D}_2$. Furthermore, τ is the smallest of the "PV numbers," which implies that the rational vectors on the Farey sequence approaching (τ, τ^2) converge more slowly than any other algebraic pair [49].

As we will see in Sec. 6.6, the frequencies enter the Fourier expansion for $\boldsymbol{x}(\boldsymbol{\theta})$ solely in terms of the small denominators

$$D_{\mathbf{n}} = 4 \sin^2(\pi \mathbf{n} \cdot \boldsymbol{w}) \ ; \tag{6.15}$$

that is, the $\mathbf{n}^{th}$ Fourier coefficient of $\boldsymbol{x}(\boldsymbol{\theta})$ is divided by $D_{\mathbf{n}}$. For Diophantine frequency vectors, $D_{\mathbf{n}}$ is bounded from below; in fact Eq. (6.11) implies that if $\boldsymbol{w} \in \mathcal{D}_2$, then $1/D_{\mathbf{n}} < \mathcal{O}(||\mathbf{n}||^4)$. Unfortunately, there is no theory analogous to the continued fraction theory which provides the values of $\mathbf{n}$ for which there are large peaks in $1/D_{\mathbf{n}}$. In Fig. (6.1) we show a plot of the values of $\mathbf{n}$ for which $(D_{\mathbf{n}}(\gamma, \sigma)||\mathbf{n}||^4)^{-1} > 1.0 \times 10^{-2}$ and 5.0×10^{-5}. As this figure shows, these peaks are quite isolated and rare. Thus, following the results for the semi-standard map, one would expect the Fourier coefficients to have similar isolated

figure=fig1k2.eps,height=3.5in

Figure 6.1. Peaks in the inverse of the small denominator where $\boldsymbol{w} = (\gamma, \sigma)$. Values of $(D_{(m.n)}(\gamma, \sigma)||(m, n)||^4) > 1.0 \times 10^{-2}$ are displayed as bold dots and those $> 5.0 \times 10^{-5}$ as light dots.

peaks, and for the convergence determination of the Fourier series to be quite delicate. However, as we will see in Sec. 6.6, this is fortuitously not the case.

6.4 Recursion Relation

In this section we will derive recursion relations for the Fourier coefficients of $\mathbf{x}(\boldsymbol{\theta})$, the solution to Eq. (6.9). For the semi-standard map it was possible to find a solution analytic in the upper half θ plane. In this case only the positive Fourier coefficients are nonzero. This is one advantage over the series for real mappings where all the Fourier coefficients must be considered. In the case at hand, since the force, Eq. (6.3), has only positive imaginary exponentials, we can also find solutions analytic in the domain $\{(\theta_1, \theta_2) : \mathrm{Im}(\theta_1) \geq 0, \mathrm{Im}(\theta_2) \geq 0\}$, so that only positive Fourier coefficients are needed

It is convenient to define $\boldsymbol{u} \in \mathbb{C}^2$ as

$$\boldsymbol{u} = \begin{pmatrix} a_1 e^{i\theta_1} \\ a_2 e^{i\theta_2} \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} . \tag{6.16}$$

The advantage of this definition, is that the parameters a_1 and a_2 will not appear in any of our expansions. Further, using Eq. (6.7) we define

$$\mathbf{g}(\boldsymbol{u}) = i(\mathbf{x}(\boldsymbol{\theta}) - \boldsymbol{\theta}) = i\boldsymbol{\chi}(\boldsymbol{\theta}) . \tag{6.17}$$

Since by ansatz, only the positive coefficients will be needed in the Fourier expansion of $\boldsymbol{\chi}(\boldsymbol{\theta})$, $\mathbf{g}(\boldsymbol{u})$ has a Taylor expansion

$$\mathbf{g}(\boldsymbol{u}) = \sum_{\mathbf{n} \in \mathbb{N}^2} \boldsymbol{b}_{\mathbf{n}} \boldsymbol{u}^{\mathbf{n}} \equiv \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \begin{pmatrix} b^{(1)}_{(n_1,n_2)} \\ b^{(2)}_{(n_1,n_2)} \end{pmatrix} u_1^{n_1} u_2^{n_2} . \tag{6.18}$$

Here we use standard multi-index notation for the vector exponentiation: while $\boldsymbol{u} \in \mathbb{C}^2$ and $\mathbf{n} \in \mathbb{N}^2$, $\boldsymbol{u}^{\mathbf{n}} \equiv u_1^{n_1} u_2^{n_2} \in \mathbb{C}$. In addition to the expansion of $\mathbf{g}$, we will need the expansions of its exponential as well:

$$e^{g_i(\boldsymbol{u})} = \sum_{\mathbf{n} \in \mathbb{N}^2} c_{\mathbf{n}}^{(i)} \boldsymbol{u}^{\mathbf{n}} \tag{6.19}$$

where $i = 1, 2$. In terms of the new variables, the map Eq. (6.9) takes the form

$$\delta^2 \mathbf{g}(\boldsymbol{u}) = -\begin{pmatrix} u_1 e^{g_1(\boldsymbol{u})} \\ u_2 e^{g_2(\boldsymbol{u})} \end{pmatrix} - k \begin{pmatrix} u_1 u_2 e^{g_1(\boldsymbol{u})+g_2(\boldsymbol{u})} \\ u_1 u_2 e^{g_1(\boldsymbol{u})+g_2(\boldsymbol{u})} \end{pmatrix} \tag{6.20}$$

where

$$k = \frac{\epsilon}{a_1 a_2} \tag{6.21}$$

is the coupling parameter. Note that these equations depend upon the three parameters a_1, a_2 and ϵ solely through k.

Substituting Eqs. (6.18)-(6.19) into Eq. (6.20) and noting that for a term in the Fourier series the operator δ^2 becomes $-D_{\mathbf{n}}$, as defined by Eq. (6.15), yields the recursion relation for $\boldsymbol{b}_{\mathbf{n}}$:

$$D_{\mathbf{n}} \boldsymbol{b}_{\mathbf{n}} = \begin{pmatrix} c_{\mathbf{n}-(1,0)}^{(1)} \\ c_{\mathbf{n}-(0,1)}^{(2)} \end{pmatrix} + k \sum_{\mathbf{m}=(0,0)}^{\mathbf{n}-(1,1)} \begin{pmatrix} c_{\mathbf{m}}^{(1)} c_{\mathbf{n}-\mathbf{m}-(1,1)}^{(2)} \\ c_{\mathbf{m}}^{(1)} c_{\mathbf{n}-\mathbf{m}-(1,1)}^{(2)} \end{pmatrix} . \tag{6.22}$$

If $\boldsymbol{w}$ is incommensurate, then $D_{\mathbf{n}}$ is nonzero, so that Eq. (6.22) defines $\boldsymbol{b}_{\mathbf{n}}$. In fact $\boldsymbol{b}_{\mathbf{n}}$ is a convolution sum of $\{\mathbf{c}_{\mathbf{m}}\}$ for those $\mathbf{m} \prec \mathbf{n}$ where we define the partial order $\prec$ on integer vectors by $\mathbf{m} \prec \mathbf{n}$ if $m_i \leq n_i$, and $\mathbf{m} \neq \mathbf{n}$.

A simple derivative identity allows us to find the $\mathbf{c}_{\mathbf{m}}$ coefficients.

$$\frac{d}{du_j} e^{g_i(\boldsymbol{u})} = [\frac{d}{du_j} g_i(\boldsymbol{u})] e^{g_i(\boldsymbol{u})}, \tag{6.23}$$

which upon substitution of Eqs. (6.18)-(6.19) yields

$$n_j c_{\mathbf{n}}^{(i)} = \sum_{\mathbf{m} \neq (0,0)}^{\mathbf{n}} m_j b_{\mathbf{m}}^{(i)} c_{\mathbf{n}-\mathbf{m}}^{(i)} . \tag{6.24}$$

Note that Eq. (6.24) allows the two forms, j=1 or 2, for $\mathbf{n}$ off the axis (these are equivalent), but for $\mathbf{n}$ on the axis, only one is valid because of a required division by a zero value of n_j.

Examining Eq. (6.24) reveals that $\mathbf{c_n}$ is a function of strictly previous $\mathbf{c_m}$, but up to current $\boldsymbol{b}_\mathbf{n}$, therefore the process must be started by generating $\boldsymbol{b}_\mathbf{n}$. Since the choice of initial phase $\boldsymbol{\theta}$ is arbitrary, we can set

$$\boldsymbol{b}_\mathbf{0} = \mathbf{0} \; . \tag{6.25}$$

Examination of the mapping Eq. (6.20) yields in addition

$$b^{(2)}_{(n_1,0)} = 0 \; ; \quad b^{(1)}_{(0,n_2)} = 0 \; . \tag{6.26}$$

Similarly, Eq. (6.25) and Eq. (6.19) imply that $\mathbf{c_0} = \mathbf{1}$, and Eq. (6.24) yields

$$c^{(2)}_{(n_1,0)} = 0 \; ; \quad c^{(1)}_{(0,n_2)} = 0 \; . \tag{6.27}$$

Finally, the recursion (6.22) implies that the values $b^{(1)}_{(n_1,0)}$ and $b^{(2)}_{(0,n_2)}$ are identical to those for the semi-standard map with frequencies ω_1 and ω_2, respectively.

This completes the recursion algorithm which allows $\boldsymbol{b}_\mathbf{n}$ to be built as an explicit function of previous $\boldsymbol{b}_\mathbf{n}$ and $\mathbf{c_n}$ coefficients. Note that if $k > 0$ then $\boldsymbol{b}_\mathbf{n}$ is positive and real, a big advantage in their computation. Since equation Eq. (6.18) actually represents two series, one in each component of the vector $\mathbf{g}$, the domain of convergence of $\mathbf{g}(\boldsymbol{u})$ is the intersection of the domains of convergence of each component's series.

6.5 The Domain of Convergence

In this section, we review some relevant results on the domain of convergence for power series in several complex variables. Let $\boldsymbol{z} = (z_1, ...z_d) \in \mathbb{C}^d$,

and for $\mathbf{m} \in \mathbb{N}^d$, define $\mathbf{z}^m = z_1^{m_1} z_2^{m_2} .... z_d^{m_d} \in \mathbb{C}$. We consider a power series,

$$S = \sum_{\mathbf{m} \in \mathbb{N}^d} b_{\mathbf{m}} \boldsymbol{z}^{\mathbf{m}} \,, \tag{6.28}$$

similar to the series obtained in the previous section. We denote the radii by $r_j = |z_j|$. The projection onto the radius space is denoted $\Pi : \Pi(\boldsymbol{z}) = (r_1, r_2, ...r_d)$. The subset $\mathbb{C}^{d*} = \{\boldsymbol{z} : z_j \neq 0\}$ excludes points for which any component of $\boldsymbol{z}$ is zero.

Several types of subsets of $\mathbb{C}^d$ are of interest. The *domain of convergence* of a series is the interior of the set of points for which it converges absolutely. A *polydisk* is the appropriate generalization of a disk: $P(a) = \{\boldsymbol{z} : |z_j| < |a_j|,\ j = 1, ..., d\}$. A *Reinhardt domain* is a domain R such that $R = \Pi^{-1}(\Pi(R))$; that is, if it contains a point with radii r_j, then it must contain every point with those same radii, regardless of phases. Reinhardt domains are conveniently pictured in the radius space $\Pi(\mathbb{C}^d) = \mathbb{R}^d$. A Reinhardt domain is *complete* if for every $\boldsymbol{z} \in R$, the polydisk $P(\boldsymbol{z}) \subset R$; thus a complete domain contains all points with smaller radii. Finally a domain D is *log-convex* if the set

$$\log(\Pi(D)) \equiv \{(\log(r_1), \log(r_2), ... \log(r_d)) : z \in \mathbb{C}^{d*} \cap D\} \tag{6.29}$$

is a convex subset of $\mathbb{R}^d$.

We will use the following theorem [84, 48]:

Theorem 6.5.1 *If S converges for all orderings of its terms at a point $\boldsymbol{z}$ then it converges absolutely to a holomorphic function. The domain of convergence, D, of S is the interior of the set for which $|b_{\mathbf{m}} \boldsymbol{z}^{\mathbf{m}}|$ is bounded. Furthermore D is a log-convex, complete Reinhardt domain. Conversely, if $|b_{\mathbf{m}} \boldsymbol{z}^{\mathbf{m}}|$ is unbounded then there is an ordering of the terms in S for which it diverges.*

The proof of this theorem is straightforward. Its most unusual aspect is that the domain of convergence is log-convex, which we will discuss in more detail. Suppose $\boldsymbol{z}, \boldsymbol{x} \in \mathbb{C}^{d*} \cap D$. Then for $\alpha + \beta = 1$ let $\boldsymbol{u}$ be any point in $\mathbb{C}^{d*}$ such that

$$\Pi(\boldsymbol{u}) = (r_1^\alpha s_1^\beta, ..., r_d^\alpha s_d^\beta) \,, \tag{6.30}$$

where r_j and s_j are the radii of $\boldsymbol{z}$ and $\boldsymbol{x}$, respectively. Then, since S converges at both $\boldsymbol{z}$ and $\boldsymbol{x}$, $B = \sup(|b_{\mathbf{m}} \boldsymbol{z}^{\mathbf{m}}|, |b_{\mathbf{m}} \boldsymbol{x}^{\mathbf{m}}|)$ exists, and

$$|b_{\mathbf{m}} \boldsymbol{u}^{\mathbf{m}}| = |b_{\mathbf{m}}| \prod_{i=1}^{d} r_i^{m_i \alpha} s_i^{m_i \beta} \leq B^{\alpha+\beta} = B \tag{6.31}$$

is bounded as well. Thus S converges at $\boldsymbol{u}$. Now since

$$\log(\Pi(\boldsymbol{u})) = \alpha \log(\Pi(\boldsymbol{z})) + \beta \log(\Pi(\boldsymbol{x})) \,, \tag{6.32}$$

we have shown that D is log-convex. □

The application of Theorem 6.5.1 to our system is straightforward since the series Eq. (6.18) has the desired form; it yields the interesting result

Corollary 6.5.2 *For fixed k defined by Eq. (6.21), an analytic invariant torus with Diophantine frequency $\boldsymbol{w}$ of the semi-Froeshlé map exists in a parameter domain in (a_1, a_2) which is complete and log-convex.*

In particular, for fixed k, completeness implies that the domain of convergence is simply connected, and its boundary projected onto the radius space can be expressed as a graph of a function $r_1(r_2)$ or $r_2(r_1)$.

As we will see in the next section, the calculation of these domains is possible with reasonable accuracy using the requirement that the terms in the series must be bounded.

6.6 Numerical Results

Determination of the sequence of $\{\boldsymbol{b}_{\mathbf{m}}\}$ of Fourier coefficients of $\boldsymbol{\chi}(\boldsymbol{\theta})$ using the recursion algorithm of Sec. 6.4 is straightforward, since they are real and positive for $k \geq 0$. The next issue is to numerically find the domain of absolute convergence, which from Sec. 6.5, is the set of $\boldsymbol{u} \in \mathbb{C}^2$ for which $|\boldsymbol{b}_{\mathbf{m}}\boldsymbol{u}^{\mathbf{m}}|$ is bounded. We begin by noting that the series (6.18)

$$\mathbf{g}(\boldsymbol{u}) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \boldsymbol{b}_{m,n} u_1^m u_2^n \tag{6.33}$$

converges absolutely in the polydisk $P(\boldsymbol{u})$ if the reordered series

$$\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \boldsymbol{b}_{m,n} r_1^m r_2^n = \sum_{n=0}^{\infty} r_1^n \mathbf{B}_n(s)\ . \tag{6.34}$$

converges. Here we define the slope $s \equiv r_2/r_1$, and the diagonal coefficient,

$$\mathbf{B}_n(s) \equiv \sum_{m=0}^{n} \boldsymbol{b}_{n-m,m} s^m \tag{6.35}$$

which, most importantly, is expressed as a finite sum. It follows that the radius of convergence for the i^{th} component is

$$\log(r_1^{(i)}(s)) = -\lim_{n\to\infty} \frac{\log B_n^{(i)}(s)}{n} \tag{6.36}$$

for each fixed s. Since the domain of convergence is complete according to Theorem 6.5.1, $r_1(s)$ is a single valued function.

To avoid numerical overflow for large s^m, we use Eq. (6.35) when $s \leq 1$, and use a corresponding formula with the slope defined as r_1/r_2 otherwise. Furthermore, we can take advantage of definition (6.35) by computing the coefficients $\boldsymbol{b}_{\mathbf{m}}$ in a triangular domain $m+n \leq N$. This saves approximately a factor of ten in computing time over using the square domain. Computer memory constraints lead us to chose $N = 255$ as our matrix dimension.

The efficacy of this method depends upon estimating Eq. (6.36), the asymptotic growth rate of the Fourier coefficients. We first consider $\boldsymbol{w} = (\gamma, \sigma)$ where the components were defined by Eqs. (6.10) and (6.13). Fig. (6.2) is a

figure=fig2k2.eps,height=3.5in

Figure 6.2. Logarithmic contour plot of the Fourier coefficients $b^{(1)}_{\mathbf{n}}$ where $\boldsymbol{w} = (\gamma, \sigma)$ and $k = 0.2$. Here $0 \leq b^{(1)}_{\mathbf{n}} \leq 10^{150}$.

logarithmic contour plot of $b^{(1)}_{\mathbf{m}}$. It can be seen that $\boldsymbol{b}_{\mathbf{m}}$ grows rapidly as $\mathbf{m}$ grows, indeed the maximal values of $b_{\mathbf{m}}$ in the figure are $\mathcal{O}(10^{150})$. This is a result of the recursion algorithm, which shows that $\boldsymbol{b}_{\mathbf{m}}$ is a combination of all the previous $\mathbf{c}_{\mathbf{m}}$. When $k \geq 0$, the $\mathbf{c}_{\mathbf{m}}$ coefficients are positive and $\mathbf{c}_{\mathbf{m}} > \mathbf{c}_{\mathbf{n}}$ for $\mathbf{m} \succ \mathbf{n}$. Whenever there is a near commensurability, $D_{\mathbf{m}}$ is small, and $\boldsymbol{b}_{\mathbf{m}}$ takes a sudden jump. This can be seen in the contour plot as a serrating of the contour lines. The neighboring coefficients for greater $\mathbf{m}$ are influenced by this jump, but the recursion algorithm serves to spread and dissipate the extra height. In other words, the coupling serves to dampen the commensurabilities. This partly accounts for the stepping up nature of the contour plot.

Fortunately, the limit (6.36) is not as difficult to evaluate for this problem as it could be in general. It turns out that $\log B_n$ behaves quite linearly as a function of n; this can be seen clearly in Fig. (6.3). The small spikes visible

figure=fig3k2.eps,height=3.5in

Figure 6.3. Log $B^{(1)}_n(s)$ vs. n for various values of the slope s, where s has the range $1 \leq \tan^{-1} s \leq 83$ and every odd degree angle is displayed.

on a given "line" of $(\log[B_n(s)], n)$ are due to near commensurabilities.

By contrast, the Fourier coefficients for the semi-standard map, b_n, depend only on the small denominator and the single previous coefficient c_{n-1},

(which is in turn implicitly a function of the coefficients $(b_1, ..., b_{n-1})$. Resonances are extremely important, and primary, secondary, and even tertiary prominences can be observed, so that the Fourier coefficients have an extremely spiked profile. In the four-dimensional case the coupling between the frequencies $\boldsymbol{w}$ appears to play the dominant role. Resonances gain and lose prominence in a delicate balancing of the coupling between frequencies, which can be seen as shadows of vertical lines in Fig. (6.3).

To determine the radius in Eq. (6.36) we performed a least squares fit of variable data sets. The top and bottom ends of the fit were allowed to float by $n = 10$ points each, and the fit with the lowest residual was automatically chosen. This eliminates the problem of a given fit falling just above or below a resonance spike. RMS errors in the slope fit are typically $\sigma_c = 0.003 \pm 0.001$, which leads us to expect at least 2 decimal accuracy in the r_i values.

figure=fig4k2.eps,height=3.5in

Figure 6.4. Boundary of the domain of convergence for $\boldsymbol{w} = (\gamma, \sigma)$.and various fixed values of k. Curves for $B_n^{(1)}$ are solid, and are dashed for $B_n^{(2)}$. Values of k are $10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 1.0, 10.0$. The vertical axis corresponds to the frequency $\omega_1 = \gamma$, and the horizontal to $\omega_2 = \sigma$.

figure=fig5k2.eps,height=3.5in

Figure 6.5: Same as Fig. (6.4) with $\boldsymbol{w} = (\gamma, \zeta)$.

Using the three frequency pairs, $\boldsymbol{w} = (\gamma, \sigma)$, (γ, ζ), and (τ, τ^2), we generate the respective $(r_1, r_2) = (a_1, a_2)$ curves for various coupling constants k, and for each of the $B_n^{(i)}$ components. These domains of convergence are displayed in Figs. (6.4)-(6.6). $B_n^{(1)}$ is represented as solid curves, and $B_n^{(2)}$ as dashed curves.

Fig. (6.7) displays the (a_1, a_2) curves for $\boldsymbol{w} = (\gamma, \sigma)$ on a log- log scale.

figure=fig6k2.eps,height=3.5in

Figure 6.6: Same as Fig. (6.4) with $\boldsymbol{w} = (\tau, \tau^2)$.

This example shows that D is log-convex in accord with Theorem 6.5.1. The sharp bends seen in some of the curves are due to the regular spacing of angles on a grid, which the log scale makes especially prevalent near the axes.

Here we will discuss the behavior of the curves for the first component, $B_n^{(1)}$ (solid curves). When $a_2 \to 0$, r_1 must approach $a^{ss}(\omega_1)$ since the map Eq. (6.20) becomes uncoupled in this limit, and $B_n^{(1)}$ becomes the coefficients of the semi-standard map with frequency ω_1. We call the r_1 axis the "dominant axis" for $B_n^{(1)}$; similarly, the r_2 axis will be the dominant axis for $B_n^{(2)}$. This behavior can be seen in Figs. (6.4)-(6.6) as all the various curves intersect the dominant axis at $a^{ss}(\omega_1)$. For reference, Table 1 gives the critical values of the semi-standard map for the various frequencies. Note that the curves in Fig. (6.4) and (6.6) actually overestimate the correct values on the axis; for example in Fig. (6.4), the intersection with the r_2 axis occurs near 0.985, while Table 1 implies that the correct value is 0.966. This overestimate is due to the fact that we compute the coefficients only out to the 255^{th} Fourier coefficient, and that near the axes the spikes in the B_n curves become more prominent (see further discussion of this below). For the semi-standard mapping, more sophisticated fitting techniques (e.g. [82]) are required for an accurate evaluation of the critical function. For our mapping we believe that, away from the axes, the radius curves are actually more accurate than this indicates. In Fig. (6.5), the intersection with the r_1 axis appears to be much lower than the value $r_1 = 0.979$ given

figure=fig7k2.eps,height=3.5in

Figure 6.7: Same as Fig. (6.4) except on a log-log scale.

Table 6.1: Critical values for the semi-standard map.

ω	a^{ss}
γ	0.979661
σ	0.966165
ζ	0.833726
τ	0.657
τ^2	0.660
$\gamma+\sigma$	0.09
$\gamma+\zeta$	0.66
$\tau+\tau^2$	0.33

in Table 1; however these curves actually rise rapidly to the correct (actually overestimated) value as $r_2 \to 0$. It is interesting that in this case, even though the values on axis are quite different, the convergence boundary has adjusted itself to be nearly square for small k. Finally, this rapid rise—approaching $a^{ss}(\omega_1)$ at a sharp angle, does not violate log-convexity, as required by Theorem 6.5.1.

The figures also show that the solid curves limit to $a^{ss}(\omega_2)$ on the r_2 axis, which we call the "subdominant" axis for $B_n^{(1)}$. This phenomena requires some explanation. When $\epsilon \equiv 0$, the boundary of domain of convergence for $B_n^{(1)}$ is $r_1^{(1)} = a^{ss}(\omega_1)$, independent of r_2; the numerical results for nonzero ϵ, however, imply that r_2 limits to $a^{ss}(\omega_2)$ on the r_2 axis. This also occurs for the domain of convergence of the second component $B_n^{(2)}$: $r_1 \to a^{ss}(\omega_1)$ as $r_2 \to 0$. To explain this phenomena, consider for example the small slope limit of $B_n^{(2)}(s)$. Equation (6.35) implies that

$$B_n^{(2)}(s) = b_{n,1}^{(2)}\ s + \ \mathcal{O}(s^2) \tag{6.37}$$

where the s^0 term vanishes according to Eq. (6.26). Using the recursion relation (6.22) implies

$$B_n^{(2)}(s) \simeq sk\frac{D_{(n,0)}}{D_{(n,1)}} b_{(n,0)}^{(1)} \tag{6.38}$$

Thus using Eq. (6.36) the radius of convergence is

$$\log(r_1^{(2)}) \simeq \log\left[a^{ss}(\omega_1)\right] \; - \; \lim_{n \to \infty} \frac{1}{n} \log\left(\frac{D_{(n,0)}}{D_{(n,1)}}\right) \tag{6.39}$$

The last limit in fact is zero, since by the Diophantine condition Eq. (6.11), the ratio of the denominators is bounded by $\mathcal{O}(n^4)$. A similar result holds for the first component of the mapping along the r_2 axis, so we have shown that

$$\lim_{s \to 0} \; r_1^{(2)}(s) \; = \; a^{ss}(\omega_1), \quad \lim_{s \to \infty} \; r_2^{(1)}(s) \; = \; a^{ss}(\omega_2) \tag{6.40}$$

Furthermore, Eq. (6.40), together with completeness, implies that the the domain of convergence is bounded by the rectangle

$$a_1 \leq a^{ss}(\omega_1), \quad a_2 \leq a^{ss}(\omega_2) \tag{6.41}$$

In fact the figures show that as $k \to 0$ the domain of convergence approaches this rectangle. Our interpretation of this is that for small but nonzero k, the singularity corresponding to $r_2 = a^{ss}(\omega_2)$ is still present, though weakened (the "residue" of this singularity, limits to zero as $k \to 0$, but it is still present for any nonzero k). This causes a difficulty with our numerical scheme for finding $r_1(r_2)$ when k is small; we discuss this further below.

figure=fig8k2.eps,height=3.5in

Figure 6.8. $\mathbf{B}_n(s)$ *vs.* n where $s = 10^{-25}$ and $\boldsymbol{w} = (\gamma, \sigma)$. The upper plot displays $B_n^{(1)}(s)$ vs. n; with $s \ll 1$, γ is the dominant frequency, and so $B_n^{(1)}$ approaches the γ semi-standard map coefficients. The lower plot displays $B_n^{(2)}$ *vs.* n; again with $s \ll 1$, σ is the subdominant frequency, and so $B_n^{(2)}$ decays.

Fig. (6.8) displays the coefficients $B_n^{(1)}$ and $B_n^{(2)}$ for $s = 10^{-25}$ and $\boldsymbol{w} = (\gamma, \sigma)$. In the limit of small slope $B_n^{(1)} \approx b_{n,0}^{(1)}$ which are the Fourier coefficients for the semi-standard map [40]. Thus the upper plot is indistinguishable from that for the semi-standard map. The lower half of Fig. (6.8) shows $B_n^{(2)}$ for small

slope. The profile exhibits spikes and valleys corresponding to a complicated coupling between σ resonances and the still important γ resonances, as shown in Eq. (6.38). Furthermore, Eq. (6.38) implies that the profile approaches a limiting form as $s \to 0$, even though the magnitude of $B_n^{(2)}$ approaches zero. Likewise, $B_n^{(2)}$ near the r_2, axis yields the semi-standard coefficients for $\omega_2 = \sigma$, while $B_n^{(1)}$ goes to zero, while similarly converging to a fixed profile.

As the domain of convergence plots show, the rectangular domain for small k contains the domain for any finite k. This follows from the completeness of the domain of convergence, and the fact that the curves limit to the semi-standard values on the axes. This fact can be used as an upper-bound when discussing the question of which torus is "last."

We also computed $r_1(r_2)$ curves for negative values of k. By the same argument as above, the negative k curves intersect the axis at $a^{ss}(\omega_1)$ and $a^{ss}(\omega_2)$. Otherwise the curves are qualitatively similar to those shown in Figs. (6.4)-(6.6), so we omit the plots. Since the domain of convergence depends only on k, these curves provide the boundary of existence in four of the octants in (a_1, a_2, ϵ) space, the other four being determined by the positive k results.

As k increases all of the boundaries in Figs. (6.4)-(6.6) become hyperbolic in shape. This can be seen most clearly in Fig. (6.5), for $\boldsymbol{w} = (\gamma, \zeta)$. The large k limit corresponds to

$$\epsilon \gg (a_1, a_2). \tag{6.42}$$

Taking this to the extreme, we set $a_1 = a_2 = 0$, then Eqs. (6.2)-(6.3) have the form

$$\delta^2 \boldsymbol{x} = i\epsilon \begin{pmatrix} e^{ix_1 + ix_2} \\ e^{ix_1 + ix_2} \end{pmatrix}. \tag{6.43}$$

Defining the new variables

$$\xi_1 = x_1 + x_2 \tag{6.44}$$

$$\xi_2 = x_1 - x_1, \tag{6.45}$$

and adding and subtracting the components of Eq. (6.43) yields a new map.

$$\begin{aligned} \delta^2\xi_1 &= 2i\epsilon e^{i\xi_1} \\ \delta^2\xi_2 &= 0. \end{aligned} \tag{6.46}$$

Thus, there exists an invariant torus for (ξ_1, ξ_2) up to some critical value

$$2\epsilon = a^{ss}(\omega_1 + \omega_2). \tag{6.47}$$

Now Eq. (6.43) is approximately valid for small a_1 and a_2, so we expect that as $k \to \infty$, using $\epsilon = ka_1a_2$, the fixed k boundary will limit to

$$r_1 r_2 = \frac{a^{ss}(\omega_1 + \omega_2)}{2k}, \tag{6.48}$$

which defines a hyperbola.

Thus we have three analytic bounds on the domain of existence of a torus:

$$a_1 < a^{ss}(\omega_1), \quad a_2 < a^{ss}(\omega_2), \quad \epsilon < 0.5a^{ss}(\omega_1 + \omega_2) \quad , \tag{6.49}$$

though the last equation is not rigorously derived. As a confirmation, Table 1 shows that $a^{ss}(\gamma + \sigma)$ is much smaller than $a^{ss}(\gamma + \zeta)$ and $a^{ss}(\tau + \tau^2)$. Thus, Eq. (6.48) predicts that the curves for $\boldsymbol{w} = (\gamma, \sigma)$ should become hyperbolae more quickly than for other $\boldsymbol{w}$ curves, as we do in fact observe.

As mentioned earlier, the scheme (6.34)-(6.36) for finding $r_1(s)$ has numerical problems when $k \ll 1$. For such small k, the singularity on one axis is dominant over the singularity on the other axis. To illustrate the problem,

consider a simple example which has a similar imbalance in the prominence of its singularities. Let

$$S(r_1, r_2) = \frac{\alpha}{\alpha - r_1} + \frac{\delta\beta}{\beta - r_2} = \sum_{m,n} b_{m,n} r_1^m r_2^n \quad . \tag{6.50}$$

Here small values of δ simulate small values of k; however, for any nonzero δ, the domain of convergence of this series is the rectangle $\{(r_1, r_2) : r_1 < \alpha, r_2 < \beta\}$.

We examine the behavior of equations (6.34)-(6.36) when applied to Eq. (6.50) by a perturbation analysis near $s = 0$. For a finite n, the algorithm gives an error in r_1 of

$$\Delta r_1 \ \sim \ - \frac{\delta\alpha}{\mathrm{n}} \left(\frac{\alpha s}{\beta}\right)^n \quad . \tag{6.51}$$

Thus the method works well provided $s < \beta/\alpha$, but fails drastically for larger s. In our computations, the slope is never larger than one; we switch to the inverse of the slope when $s = 1$. Thus, supposing $\beta < \alpha$ the method fails in a cone $\beta/\alpha < s < 1$. So for the Froeshlé mapping, we also expect that slopes within a similar cone will give bad results if k is too small. That this is true can be seen as a slight loss of convexity for the smallest values of k along the subdominant axis in Figs. (6.4)- (6.6). In practice we are unable to lower k below 10^{-5} in the computations.

figure=fig9k2.eps,height=3.5in

Figure 6.9. Radii of convergence curves in (a_1, a_2, ϵ) space for $B_n^{(1)}$ and $\boldsymbol{w} = (\tau, \tau^2)$ where the r_1 axis is represented as the τ axis, and r_2 as the τ^2 axis.

figure=fig10k2.eps,height=3.5in

Figure 6.10: Same as Fig. (6.9) with $\boldsymbol{w} = (\gamma, \zeta)$.

Finally, our $r_1(r_2)$ data can be displayed in terms of the coupling parameter ϵ, instead of k. Figs. (6.5) and (6.6) are converted via Eq. (6.21) to

the three-dimensional graphs seen in Figs. (6.9) and (6.10). Here we see in a new way the importance of the sum frequency $(\omega_1 + \omega_2)$ through Eq. (6.49). Numerical overflow for large k prevents us from calculating the curves for ϵ too close to its maximum value.

In many ways, it is these three-dimensional plots that are most useful when deciding a partial order to determine the "last invariant torus." One concept of ordering of the domains of convergence is to choose a directed curve in (a_1, a_2, ϵ) beginning at the origin. One could linearly order the domains of convergence in terms of the order of intersection of the domain boundaries with this curve. This motivates the following local definition of order.

Curve based Order: *An $\boldsymbol{w}$ torus persists longer than a $\boldsymbol{\mu}$ torus along a curve $\boldsymbol{\xi}(t)$ for which $\boldsymbol{\xi}(0) = \mathbf{0}$, if $\boldsymbol{\xi}(t)$ intersects the boundary of the domain of convergence of the $\boldsymbol{\mu}$ torus first.*

The simplest example of a parameterized family is a line emanating from the origin. Another example is a parabolic ray $a_1 = t, a_2 = st, \epsilon = kst^2$ for fixed s and k. Figs. (6.4)-(6.6) order domains in this sense.

In general, one wants to do more than compare two surfaces using a single point from each surface, which is all a curve based order allows. In some sense, one may want to incorporate the information of the entire surface in a comparison. This motivates the definitions of the following global comparisons.

Metric based Order: *For a given metric, an $\boldsymbol{w}$-torus persists longer than a $\boldsymbol{\mu}$-torus if the boundary of the domain for $\boldsymbol{w}$ has a point farther from the origin than that for $\boldsymbol{\mu}$.*

This definition for ordering is limited in that it requires the choice of a metric.

If one surface is completely contained inside another, then that torus

is more persistent than the other according to any definition, since containment is a topological notion. Thus we define the *partial ordering*

Topological Order: *An* $\boldsymbol{w}$ *torus persists longer than a* $\boldsymbol{\mu}$ *torus if the domain for* $\boldsymbol{w}$ *contains that of* $\boldsymbol{\mu}$*.*

Of course, the surfaces for two different frequencies will intersect in general, and then the topological ordering does not apply. In our examples, the surface for (τ, τ^2) is completely contained inside that of (γ, ζ), and therefore the (γ, ζ) torus is more persistent. The complete containment of the (γ, ζ) surface is partly due to the fact that each of $(a^{ss}(\gamma), a^{ss}(\zeta), a^{ss}(\gamma + \zeta))$ are greater than their counterparts $(a^{ss}(\tau), a^{ss}(\tau^2), a^{ss}(\tau + \tau^2))$. On the other hand, in order to compare the (γ, σ) and (γ, ζ) tori, note that though $a^{ss}(\sigma) > a^{ss}(\zeta)$, $a^{ss}(\gamma + \sigma) < a^{ss}(\gamma + \zeta)$. Thus the surfaces must intersect, and therefore there can only be parameterized comparisons.

6.7 Conclusions

We have determined the domain of existence of invariant two-tori analytically conjugate to a rotation for the semi-Froeshlé mapping by expanding the conjugacy function in a Fourier series in the angle variables. The semi-Froeshlé mapping has the advantage that two of the parameters can be eliminated in the Fourier series, so that the boundary of existence of the tori in all three parameters can be obtained with a single parameter sweep. We have studied the boundary of the domain for several frequency vectors, all of which are elements of a cubic algebraic field, and therefore satisfy Diophantine conditions. The boundary of these domains appears to be smooth; rather surprisingly, it appears smooth even when the parameters have opposite signs (i.e. negative values of k). We have shown that when projected on the parameters (a_1, a_2) for fixed

$k = \epsilon/a_1 a_2$, the boundary is log-convex and complete, and that as $k \to 0$ the domain limits to the rectangle corresponding to the domain for the uncoupled mappings. Furthermore, numerical results imply that the domain is bounded by the critical function for the sum frequency, as shown by Eq. (6.47).

The methods and theorems of this paper are not restricted to the four-dimensional version of Eq. (6.2). They also apply to the 2d-dimensional complex semi-Froeshlé map, providing only that each occurrence of $x^{(j)}$ in an exponential, $\exp(imx^{(j)})$, in the force, has the same sign. The main bottleneck is computing the Fourier coefficients recursively which involves a (d-1) degree iterated convolution sum, where d is the dimension. Computing the m^d coefficients would take $\mathcal{O}(m^{2d})$ steps, making computer time a major problem in practice. In the same vein, more complicated forcing terms in Eq. (6.3) could also be considered, but similar time constraints may be a problem.

There are a number of open questions left by our study.

1) When the Fourier series does not converge, does there exist an invariant Cantor set for the mapping (a cantorus)? Results for twist mappings near the anti-integrable limit show the existence of cantori for all frequencies [60]. What is the nature of the invariant set when the Fourier coefficients for $x^{(2)}$ converge, but those for $x^{(1)}$ do not, as seen especially in Fig. (6.6)? One is tempted to think it is a Cantor set of circles.

2) Are all invariant tori for the semi-Froeshlé mapping analytically conjugate to a rotation? Perhaps all tori with Diophantine frequency vectors?

3) Is there an extension of the converse KAM theory of [62] to complex mappings?

4) Is there an extension of some of the results of Theorem 6.5.1 to real valued four-dimensional mappings of some class? It is possible that such a map

may also have a log-convex domain in the proper coordinates.

5) Can one use similar techniques to study the existence of invariant circles for a four-dimensional mapping? In [62] it was suggested that circles may last longer than any tori.

6) Which class of frequency vectors correspond to the most persistent invariant tori? In this paper we compare several likely candidates, but do not present evidence that there are not more persistent tori. In searching for a particularly persistent torus, a first step might be to maximize the values of $a^{ss}(\omega_1), a^{ss}(\omega_2)$, and $a^{ss}(\omega_1 + \omega_2)$. Which class of frequency vectors does this? Of course, since denominators containing all $\mathbf{m}\cdot\boldsymbol{w}$ occur, the most persistent class of frequencies may be that with maximal Diophantine constant C in Eq. (6.11). Since incommensurate algebraic frequency vectors form a field, any elements of such a field will have the same C. Moreover, since a degree three algebraic field has the minimal exponent μ in Eq. (6.11), it seems reasonable that it is such a field which will be most persistent. Of course the definition of persistence will depend on the choice of a partial ordering, and even then it is not clear how dependent upon the specific model the results would be. An enlightening discussion of these issues is given by Lochak [55].

CHAPTER 7

SUMMARY

In this thesis I demonstrated that the sensitivity defining chaotic dynamics makes accessible a wide range of behaviors with an arbitrarily small control signal. In targeting, finding long paths may be broken into smaller problems of finding short paths between switching points. Finding the switching points between orbit segments is the key difficulty.

Given a test orbit which explores its accessible phase space, I presented an algorithm to remove the long recurrent loops resulting from several passes through a barrier's turnstile by inefficient multiple passes across the barrier. Multiple crossings of a barrier distinguish an inefficient orbit from the efficient orbit, which crosses a barrier exactly once. The point is that it is easy to find an an inefficient orbit which eventually achieves the desired transport from near $\boldsymbol{a}$ to near $\boldsymbol{b}$. Using recurrence to indicate multiple crossings of intermediate barriers, we can construct a pseudo-orbit consisting of only the segments of the test orbit required for the efficient single crossings of barriers. The required orbit segments of a pseudo-orbit are those in which we truncate (long) recurrent loops from the test orbit. Given a mild hyperbolic splitting hypothesis between the stable and unstable manifolds along the orbit segments of the test orbit we wish to keep, we present an algorithm to find a nearby shadow orbit to the pseudo-orbit. The algorithm reduces the error of truncating recurrent loops, by shooting from the unstable manifold of the orbit before the recurrence to the stable manifold of

the orbit after the recurrence. Thus the error of the pseudo-orbit can be pushed backwards in time along the unstable manifold, and forwards in time along the stable manifold to achieve arbitrary control constraints on the error's size.

In two dimensions, the switching point may be recognized automatically and model independently by recurrence in the barrier's turnstile. In higher dimensions, we have the recurrence lemma from which we conclude that long orbits must have recurrences and hence may be shortened by their removal. We showed that this algorithm allows us to manage efficiently *all* of the test orbit data simultaneously, as a library of known behaviors, to target between any $\boldsymbol{a}$ and $\boldsymbol{b}$ in the accessible phase space.

Targeting through recurrences requires a chaotic dynamical system. To shoot from the unstable manifold before a recurrence to the stable manifold after the recurrence seems to require that the test orbit belong to a uniformly hyperbolic subset of the accessible phase space. We constructively build an ϵ shadow to the δ shadow of the test orbit with the recurrences deleted. We saw, however that the standard map test orbit is not uniformly hyperbolic; angles between stable and unstable manifolds do not appear to be numerically bounded away from zero. Thus, a "glitch" may occur. Nonetheless, the algorithm seems to work on an "accept on success" basis. Small angles that occupy a small measure on the angle frequency distribution do not present a problem; larger angles prove to be frequent enough that successful patches are found.

We also require that there exist a test orbit which visits an accessible phase space with large measure. This seems to require a transitivity condition. An ergodic orbit is certainly sufficient. However, such a condition is not necessary, and unproven for the standard map. A mixing property is also sufficient but unnecessary and unproven.

The algorithm has allowed us to find fast paths in:

1. The standard map (an example of a discrete dynamical systems in which the typical "layered phase space" of area preserving maps due to KAM curves and residual cantori makes competing targeting techniques ineffective).
2. The restricted three body problem (a flow which yields and area preserving map by Poincaré section).
3. The Bernoulli shift map (this model of chaotic transport allowed us to find closed form paths which we compared to the closed form time-optimal orbit from $\boldsymbol{a}$ to $\boldsymbol{b}$).
4. A map which we recovered by prediction of phase space data from time-series delay embedding techniques.

The main goal of the targeting algorithm presented here has been to make targeting widely accessible, in a manner independent of special knowledge about the dynamics, and in some cases, independent of even an analytic representation of the dynamics (i.e. only using observed time-series data). Our technique is widely applicable, and has achieved a high degree of success towards these goals. It remains to continue exploring ever more (applied) examples to which the technique may prove useful.

Finally, on a separate note, we have presented techniques to find the breakup of invariant tori due to a Hamiltonian perturbations to the four-dimensional semi-standard map. This allows for discussion of finding frequency vectors $\boldsymbol{w}$ with the most robust invariant tori and comparison of more persistent frequencies by several definitions of the (partial) orders which we present.

7.1 Problems for the Future

There are a number of directions in which research continuing the work of this thesis may proceed. Some of these are enumerated as follows.

- Extend the model of the Earth - Moon spacecraft system. The elliptic restricted three-body problem is time dependent as is the restricted four-body problem including the Sun. A four-dimensional Poincaré map results. There are likely other interesting and efficient chaotic transfer orbits.
- Build a better RTBP integrator using Birkhoff's regularization transformation so that test orbits with close approaches can be generated. In principle, a starting point $\boldsymbol{a}$ located on the Earth's surface is possible, requiring an initial δV to jump onto the correct energy level J_2.
- Apply optimal control theory to connect the epsilon chain transfer orbit to the moon with small manoeuvres. Finite parameter optimization can be used to find the locations and times of ΔV manoeuvres between nearby segments.
- Investigate further the recurrence frequencies in higher dimensional maps and resulting path lengths.
- Shoot between higher dimensional saddles as described in Appendix A.4. To this end, find full stable and unstable manifolds, and generalize Eq. (2.11) for a good Newton-secant primer.
- Investigate further the possibilities of prediction control as formulated in Chapter 5 for a high dimensional dynamics that collapses sufficiently to a low dimensional attractor.

- Add noise. Kostelich has noted that his tree based targeting algorithm is quite sensitive to noise; a 1% noise signal is enough to lose a targeted path.
- Do the orbit restriction method described in Appendices A.2-A.3.
- Investigate further the drift through momentum "barriers" caused by coupling between two standard maps as modeled in Appendix A.5, by Eqs. (A.42) and (A.43), a Hénon map which drives a simple twist map (see Fig. (A.3)). The more complex phenomenon of "Arnold Diffusion" allows for back coupling, i.e. in the Froeshle' map, two standard maps simultaneously drive each other. There is the possibility of a picture analogous to Fig. (A.3) for each commensurate vector frequency $\boldsymbol{w}$. The important $\boldsymbol{w}$ have important resonant channels. This is the so called "Arnold web". An interesting question concerns what route an orbit travels through the web, and why it takes a given direction at a fixed junction $\boldsymbol{w}$. A first approach is to investigate the path of the unstable manifold at a fixed commensurance $\boldsymbol{w}$ both by writing down the unstable manifold of the linearized coupled map near the commensurance and by then following the unstable manifold by evolving a nearby cluster of points on the unstable eigenvector. This is analogous to seeing how the unstable manifold of the fixed point of Eq. (A.42) tips in the I direction in the coupled map. This should give us some idea of the vertical drift.

BIBLIOGRAPHY

[1] H. Abarbenel, R. Brown, J. Sidorowich, L. Tsimring, "The analysis of observed chaotic data in physical systems", *Rev. of Mod. Phys.*, Vol. 65, No. 4, October 1993.

[2] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, (Springer, New York), 462pp. (1978).

[3] D.K. Arrowsmith, and C.M. Place, *An Introduction to Dynamical Systems*, (Cambridge University Press, New York 1990).

[4] S. Aubry, private communications.

[5] E. Belbruno, "Ballistic Lunar Capture Transfers Using the Fuzzy Boundary and Solar Perturbations: A Survey", *J. of Interplanetary Soc.*, Vol. 47, 73-80, (1994).

[6] E. Belbruno, "The Dynamical Mechanism Of Ballistic Lunar Capture Transfers In The Four-Body Problem From The Perspective Of Invariant Manifolds And Hill's Regions", (preprint n.270, Centre de Recerca Matematica Institut D'Estudis Catalans Apartat 50, E-08193 Bellaterra, Barcelona, Spain, 1994).

[7] G. Benettin, L. Galgani, A. Giorgilli, J. Strelcyn, "Lyupanov Characteristics Exponents for Smooth Dynamical Systems and For Hamiltonian Systems; A Method for Computing All of Them", *Meccanica* V14-15, Mar. (1980).

[8] D. Bernstein and A. Katok, "Birkhoff Periodic Orbits for Small Perturbations of Completely Integrable Hamiltonian Systems with Convex Hamiltonians", *Invent. Math.* **88**, 225 (1987).

[9] A. Berretti and S. Marmi, "Standard Map at Complex Rotation Numbers: Creation of Natural Boundaries", *Phys. Rev. Let.* **68**, 1443-1446 (1992).

[10] A.M. Bloch and J.E. Marsden, "Controlling Homoclinic Orbits", *Theoretical and Computational Fluid Dynamics,* 1, 179-190 (1989).

[11] E. Bollt and J. Meiss, "Break-up of invariant tori for the four-dimensional semi-standard map", *Physica* **D** (1993) 282-297.

[12] E. Bollt and J. Meiss, "Controlling Chaos Through Recurrence", *Physica* **D** Vol. 81, No. 3, 280-294, (1995).

[13] E. Bollt and J. Meiss, "Targeting Chaotic Orbits to the Moon Through Recurrence", (in press).

[14] R. Bowen, "On axiom A diffeomorphisms", *J. Diff. Eq.* **18** 333-339 (1975).

[15] E. Bradley, "Taming Chaotic Circuits", M.I.T. Ph.D. thesis, (M.I.T. A–TR 1338 1992).

[16] A.J. Brentjes, "Multidimensional Continued Fraction Algorithms", Ph.D. thesis, (Mathematisch Centrum, Amsterdam, 1981).

[17] D.S. Broomhead and G.P. King, "Extracting Qualitative Dynamics from Experimental Data", *Physica* **D** (1986) 217-236.

[18] J.W.S. Cassels, *An Introduction to Diophantine Approximation*, (Cambridge University Press, Cambridge), 168 pp. (1965).

[19] G. Chen and X. Dong, "Controlling Chua's Circuit", *J. Circ. Syst. and Computers,* 3, (1993).

[20] Q. Chen, J.D. Meiss and I.C. Percival, "Orbit Extension Method for Finding Unstable Periodic Orbits", *Physica* 29**D** (1987) 143-154.

[21] B.V. Chirikov, "A universal instability of many dimensional ocs illator systems", *Physics Reports* 52 265 (1979).

[22] C.C. Conley, "Low Energy Transit Orbits in the Restricted Three-Body Problem," SIAM J. Appl. Math. Vol. 16, No. 4, (1968).

[23] I. Dana and S. Fishman, "Diffusion in the standard map", *Physica* 17**D** 63 (1985).

[24] R. de la Llave and D. Rana, "Accurate Strategies for Small Divisor Problems", *Bull.AMS* **22**, 85-90(1990).

[25] D.W. Dunham and S.A. Davis, "Optimization of a Multiple Lunar-Swingby Trajectory Sequence", *J. Astronautical Sci.* **33**, 275-278 (1985).

[26] J. Eckmann and D. Ruelle, "Ergodic theory of chaos and strange attractors", *Rev. of Mod. Phys.* **57,** 617 (1985).

[27] R. Easton, "Trellises formed by stable and unstable manifolds in the plane," Trans. Am. Math. Soc. **294** 719-31.

[28] J. Farmer, E. Ott, and J. Yorke, *Physica* 7**D**, 153 (1983).

[29] J. Farmer and J. Sidorowich, "Predicting chaotic time series", *Phys. Rev. Lett.* 59 845 (1987).

[30] J. Farmer and J. Sidorowich, "Exploiting Chaos to Predict the Future and Reduce Noise", appears in *Evolution, Learning, and Cognition,* ed. Y.C. Lee, World Scientific Press, pp. 277, (1988).

[31] W. Press, B. Flannery, S. Teukolsky and W. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing*, Cambridge University Press, New York, (1988).

[32] R. Farquhar, D. Muhonen and L.C.Church, "Trajectories and Orbital Maneuvers for the ISEE-3/ICE Comet Mission", *J. Astronautical Sci.* **33**, 235-54 (1985).

[33] A.M. Fraser and H.L. Swinney, "Independent coordinates for strange attractors from mutual information", *Phys. Rev., 33A*, 1134-1140, (1986).

[34] C. Froeschle, "Numerical Study of a Four-Dimensional Mapping", *Astron. and Astrophys* **16**, 172-189(1972).

[35] C. Froeschle and J.-P. Scheidecker, "Numerical Study of a Four Dimensional Mapping II", *Astron. and Astrophys.* **22**, 431-436 (1973).

[36] A. Garfinkel, M.L. Spano, W.L. Ditto, and J.N. Weiss, "Controlling Cardiac Chaos", *Science* 257, 1230-1235 (1992).

[37] H. Goldtein, *Classical Mechanics, Second Edition,* Addison-Wesley Publishing Company, Reading, Mass. 1980.

[38] J.M. Greene, "A Method for Computing the Stochastic Transition", *J. Math. Phys.* **20**, 1183-1201(1979).

[39] J.M. Greene and J.S. Kim, "The Calculation of Lyupanov Spectra", *Physica* **D** (1986).

[40] J.M. Greene and I. C. Percival, "Hamiltonian Maps in the Complex Plane", *Physica* 3**D**, 530-548(1981).

[41] J. Guckenheimer, B. Hu and J. Rudnick, "Quasiperiodic Transition to Chaos with Three Incommensurate Frequencies", unpublished, (1983).

[42] J. Hanson, J. Cary and J. Meiss, "Algebraic Decay in Self-Similiar Markov Chains", *J. Stat. Phys.*, Vol. 39, Nos. 3/4 (1985) 327.

[43] S. Hayes, C. Grebogi and E. Ott, "Communicating with Chaos", (preprint 1992).

[44] G.W. Hill, "Researches in the Lunar Theory", *Amer. J. Math.* **1** (1878)

[45] C. Hsu, *Cell-to-Cell Mapping*, (Springer-Verlag, New York, 1987).

[46] J. Hsu, and A.J. Meyer, *Modern Control Principles and Applications*, (McGraw-Hill, New York, 1968).

[47] P. Hutt and J. Bahcall, " Binary-Single Star Scattering. I. Numerical Experiments for Equal Masses", *Astrophys. J.* **268**: 319-341, (1983).

[48] L. Kaup and B. Kaup, *Holomorphic Functions of Several Variables*, (W. de Gruyter,Berlin), 349 pp. (1983).

[49] S. Kim and S. Ostlund, "Simultaneous Rational Approximations in the Study of Dynamical Systems", *Phys. Rev. A* **34**, 3426-3434(1986).

[50] H.T. Kook and J.D. Meiss, "Periodic Orbits for Reversible, Symplectic Mappings", *Physica* 35**D**, 65-86(1989).

[51] Kostelich, E., Grebogi, C., Ott, E., and Yorke, J.A., "Higher-dimensional targeting", Phys. Rev. E, Vol. 47, No. 1, 305-10 (1993).

[52] Y. Lai, M. Dong, C. Grebogi, "Controlling Hamiltonian Chaos", *Phys. Rev.* **E** 47 1 86-92 (1992).

[53] Y.C. Lai, C. Grebogi, J.A. Yorke, and I. Kan, "How often are chaotic saddles nonhyperbolic", *Nonlinearity* **6** 779 (1993).

[54] A.J. Lichtenberg and M.A. Lieberman, *Regular and Stochastic Motion,* (Springer-Verlag, New York, 1983).

[55] P. Lochack, "Canonical perturbation theory via simultaneous approximation", *Russ. Math. Surv.*, (in press).

[56] E.N. Lorenz, *The Essence of Chaos,* (University of Washington Press, Seattle, 1993).

[57] R.S. Mackay, *Renormalization in Area Preserving Maps,* PhD. thesis, (Princeton University, 1981).

[58] R.S. MacKay, "A Renormalisation Approach to Invariant Circles in Area-Preserving Maps", *Physica* 7**D**, 283-300(1983).

[59] R.S. MacKay, "On Greene's Residue Criterion", University of Warwick, (1991).

[60] R.S. MacKay and J.D. Meiss, "Cantori for Symplectic Maps near the Anti-integrable Limit", *Nonlinearity* **5**, 149-160(1992).

[61] R.S. MacKay, J.D. Meiss and I.C. Percival, "Transport in Hamilton ian Systems", *Physica* 13**D**, 55-81 (1984).

[62] R. S. MacKay, J. D. Meiss and J. Stark, "Converse KAM Theory for Symplectic Twist Maps", *Nonlinearity* **2**, 555-570(1989).

[63] R. S. MacKay and I. C. Percival, "Converse KAM: Theory and Practice", *Comm. Math. Phys.* **98**, 469-512(1985).

[64] M. Malavasi and S. Marmi, "Critical KAM Circles and the Brjuno Function", *J. Phys. A* **22**, L563-L569(1989).

[65] J. M. Mao and R. H. G. Helleman, "Existence of KAM Tori for Four-Dimensional Volume Preserving Maps", *N. Cimento* **104B**, 177-183(1989).

[66] J. M. Mao and R. H. G. Helleman, "Break-up of KAM Tori of Cubic Irrational Winding Number", Univ. of Houston, 1989).

[67] S. Marmi and J. Stark, "On the Standard Map Critical Function", *Nonlinearity*, (in press).

[68] J.N. Mather, "Non-Existence of Invariant Circles", *Erg. Th. Dyn. Sys.* **2**, 301-309 (1984).

[69] J.N. Mather, "Variational Construction of Orbits of Twist Diffeomorphisms", J. Am. Math. Soc., 4 (1991), 207-263.

[70] J.D. Meiss, "Symplectic maps, variational principles, and transport", *Rev. Mod. Phys.* **64** 795 (1992).

[71] J.D. Meiss and E. Ott, " Markov Tree Model of Transport in Area-Preserving Maps", *Physica* **D** (1986) 387-402.

[72] K.R. Meyer and G.R. Hall, *Intoduction to Hamiltonian Dynamical Systems and the N-Body Problem, Appl. Math. Sci.*, Vol. 90, Springer-Verlag, New York, (1992).

[73] J. Moser, "On Quadratic Symplectic Mappings", Eidgenössische Tecnische Hochschule, ETH Zürich, (1991).

[74] M. Muldoon, "Ghosts of Order on the Frontier of Chaos" Ph.D. thesis, (California Institute of Technology, 1989).

[75] M. Muldoon, , "Topology from a Timeseries", Physica **D** 65 1-16 (1993).

[76] M.A. Murison, "On An Efficient and Accurate Method to Integrate Restricted Three-Body Orbits", *Astron. J. Vol.* **97**, No. 5 (1989).

[77] N.N. Nekhoroshev, "An Exponential Estimate for the Stability Time of Hamiltonian Systems Close to Integrable Ones", *Russ. Math. Surveys* **32:6**, 1-65(1977).
G. Benettin, L. Galgani and A. Giorgilli, "A Proof of Nekhoroshev's Theorem for the Stability Times in Nearly Integrable Hamiltonian Systems", *Celest. Mech.* **37**, 1-25(1985).

[78] K. Ogata, *Control Engineering*, (second edition Prentice-Hall, Englewood, NJ, 1990), 782-784.

[79] E. Ott, C. Grebogi and J. Yorke,"Contolling chaotic dynamical systems", Chaos: Soviet-American Perspective on Nonlinear Science(American Institute of Physics, N.Y., 1990) 153-172.

[80] E. Ott, C. Grebogi and J. Yorke, "Controlling Chaos", *Phys. Rev. Lett.*, Vol. **63**, No. 11, 1196-1199, (1990).

[81] N.H. Packard, J.P. Crutchfield, J.D. Farmer and R.S. Shaw, "Geometry from a time series, *Phys. Rev. Lett.* **45**, 712, (1980).

[82] I.C. Percival, "Chaotic Boundary of a Hamiltonian Map", *Physica* 6**D**, 67-77(1982).

[83] H. Poincaré , *Les methodes nouvelles de la mechanique celeste,* (Gauthier-Villar, Paris, 1892).

[84] R.M. Range, *Holomorphic Functions and Integral Representations in Several Complex Variables*, (Springer-Verlag, New York), pp. 386 (1986).

[85] C. Robinson, *Dynamical Systems: Stability, Symbol Dynamics, and Chaos,* (CRC Press, Ann Arbor 1995).

[86] F. Romerias, C. Grebogi, E. Ott and W. Dayawansa, "Contolling chaotic dynamical systems", *Physica* **D** 58 (1002) 165-192 (1992).

[87] A.E. Roy, *Orbital Motion, third edition*, (Adam Hilge, Philadelphia 1988)

[88] R. Roy, T.W. Murphy, T.D. Maier, A. Gills and E.R.Hunt, "Dynamical Control of a Chaotic Laser: Experimental Stabalization of a Globally Coupled System", *Phys. Rev. Lett.*, 68, 1259-62 (1992).

[89] J.M. Sanz-Serna, "Symplectic integrators for Hamiltonian problems: an overview", *Acta Numerica* (1991) 243-286.

[90] T. Sauer, J. Yorke and M. Casdagli, "Embedology", *J. Stat. Phys.*, Vol. **65** Nos. 3/4. (1991).

[91] T. Shinbrot, W. Ditto, C. Grebogi, E. Ott, M. Spano and J. Yorke, "Using the Sensitive Dependence of Chaos(the "Butterfly Effect") to Direct Trajectories in an Experimental Chaotic System", *Phys. Rev. Lett.*, Vol. **68**, No. 19, 2863-2866, (1992).

[92] Shinbrot, T., Grebogi, C., Ott, E. and Yorke, J.A., "Using Chaos to Direct Trajectories to Targets in Systems Describable by a One-Dimensional Map", *Phys. Rev.* **A** 45 4165-8 (1992).

[93] Shinbrot, T., Grebogi, C., Ott, E. and Yorke, J.A., "Using Chaos to Direct Trajectories to Targets", *Phys. Rev. Lett.* Vol. **65**, No. 26 3215-18 (1990).

[94] Shinbrot, T., Grebogi, C., Ott, E. and Yorke, J.A., "Using Small Perturbations to Control Chaos", (preprint).

[95] S. Smale, "Differentiable dynamical systems", *Bull. Amer. Math. Soc.*, **73**, 747-817, (1967).

[96] I. Starobinets and A. Pikovsky, "Multistep method for controlling chaos", *Phys. Lett.* **181A** 149 (1993)

[97] J. Stoer and R. Bulirsh, *Intoduction to Numerical Analysis,* (Springer-Verlag, New York, 1980).

[98] H. Swinnerton-Dyer and C. Sparrow, "The Falkner-Skan equation I: The creation of Strange invariant Sets," DAMTP Cambridge report (1993).

[99] V. Szebehely, *Adventures in Celestial Mechanics, A First Course in the Theory of Orbits*, (the University of Texas Press, Austin TX., 1989).

[100] V. Szebehely, *Theory of Orbits The Restricted Problem of Three Bodies*, (Academic Press, Inc. N.Y., 1967).

[101] F. Takens, in *Dynamical Systems and Turbulence,* edited by D. Rand and L.-S. Young, Lecture Notes in Mathematics No. 898 (Springer, Berlin), p.366. (1980).

[102] J. Tou, *Modern Control Theory*, (McGraw Hill, New York 1964).

[103] A.S. Weigenbend and N.A. Gershenfeld, *Times Series Prediction: Forcasting the Future and Understanding the Past,* eds. Reading, MA: Addison-Wesley, (1993).

[104] S. Wiggins, *Chaotic Transport in Dynamical Systems*, (Springer-Verlag, New York, 1992).

[105] J. Wisdom and M. Holmanm "Symplectic Maps for the n-Body Problem: Stability Analysis", *Astron. J.* Vol **104,** No. 5, (1992).

[106] A. Wolf, J. Swift, H. Swinney and J. Vastano, " Determining Lyupanov Exponants from a Time Series", *Physica* 16**D** 285-317 (1985).

[107] H. Yoshida, "Recent Progress in the Theory and Application of Symplectic Integrators", *Cel. Mech. D. Aston.* **56**:27-43, (1993).

APPENDIX A

A.1 Pole Placement

In this section we give a brief example of how the poles (eigenvalues) of Eq. (1.3) might be arbitrarily chosen with a well defined gain matrix K. For more details of this derivation (see Douglas Miron). We rewrite Eq. (1.3) in the form

$$\boldsymbol{z}' = A \cdot \boldsymbol{z} + Bu, \ \boldsymbol{z} \in \mathbb{R}^d, \tag{A.1}$$

with one signal ($u \in \mathbb{R}$). B distributes the signal to each of the d states. The stability of $\boldsymbol{z}^* = 0$ is governed by the eigenvalues of the matrix A. The goal is to effectively alter the dynamics (A.1) to a contraction mapping by altering the Bu term. We will see that if we choose a gain matrix K and feedback rule of the form

$$u = -K^t \cdot \boldsymbol{z}. \tag{A.2}$$

then we may be able to alter the eigenvalues of the resulting dynamics

$$\boldsymbol{z}' = (A - B \cdot K^t)\boldsymbol{z}, \tag{A.3}$$

arbitrarily.

For ease of exposition, we restrict the discussion to $d = 3$. Assume that we have performed a linear change of coordinates of (A.1) to "control canonical

form."

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a_3 & -a_2 & -a_1 \end{bmatrix}, \; B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \text{ and } K = \begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix}. \tag{A.4}$$

The gain matrix k_i holds the free parameters to alter the dynamics by the rank one modification $-B \cdot K^t$. We obtain the characteristic equation

$$\begin{aligned} |(A - B \cdot K^t) - sI| &= \begin{vmatrix} -s & 1 & 0 \\ 0 & -s & 1 \\ -a_3 - k_1 & -a_2 - k_2 & -a_1 - k_3 - s \end{vmatrix} \\ &= s^3 + (a_1 + k_3)s^2 + (a_2 + k_2)s + (a_3 + k_1) = 0. \end{aligned} \tag{A.5}$$

If we wish to obtain the poles $(\lambda_1, \lambda_2, \lambda_3)$, then the characteristic equation must be of the form

$$(s-\lambda_1)(s-\lambda_2)(s-\lambda_3) = s^3 - \underbrace{(\lambda_1 + \lambda_2 + \lambda_3)}_{\alpha_1} s^2 + \underbrace{(\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3)}_{\alpha_2} s + \underbrace{\lambda_1\lambda_2\lambda_3}_{\alpha_3}. \tag{A.6}$$

Matching with the controlled characteristic equation (A.5) we obtain

$$\begin{aligned} \alpha_1 &= a_1 + k_3 = -(\lambda_1 + \lambda_2 + \lambda_3) \\ \alpha_2 &= a_2 + k_2 = (\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3) \\ \alpha_3 &= a_3 + k_1 = -\lambda_1\lambda_2\lambda_3, \end{aligned} \tag{A.7}$$

from which we can solve for the gains

$$\begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix} = \begin{bmatrix} -a_1 - (\lambda_1 + \lambda_2 + \lambda_3) \\ -a_2 - (\lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_2\lambda_3) \\ -a_3 - \lambda_1\lambda_2\lambda_3 \end{bmatrix}. \tag{A.8}$$

Hence we see that we may arbitrarily choose the poles $\{\lambda_i\}$ less than 1 to force a contraction if we can find a change of coordinates to the form (A.4).

The more general result, of course, is Ackermann's formula for the necessary and sufficient condition which are already stated (see Eq. (1.6)). The derivation of the Ackermann formula bears a similar flavor to the above calculation which is made much cleaner for arbitrary dimension d by making use of the Caley-Hamilton theorem which states that a matrix C satisfies its own characteristic equation.

$$C^d + \alpha_1 C^{d-1} + ... + \alpha_{d-1} C + \alpha_d I = 0. \tag{A.9}$$

(For details see [78]).

A.2 Action of a "Pseudo-orbit" the Orbit Restriction Method

In this section, we present an alternative method to that presented in Sec. 2.2.3 to find a real orbit patch through a recurrence. The technique was developed in Chen *et al.* [20] to find and track periodic orbits for the special case of area preserving maps, such as Eq. (1.37). In this paper they developed what they call an "orbit extension method" to find a high order periodic orbit nearby a lower order periodic orbit. Using a numeric rule, they were able to identify for "ordered orbits" the points of the low order orbit which are "parents" to several iterates of the higher order orbit. These points they identified as the iterates of a pseudo-orbit near the real high order orbit. Under the assumption that they were close enough to have a convergent Newton's method, they then used the Lagrangian rule that a real orbit has zero variation of its action.

Our goal is quite similar in targeting. We wish to build a real orbit segment (patch) given a nearby δ pseudo-orbit. The method of Sec. 2.2.3 was similar to the closing lemma [14] which uses hyperbolicity to constructively prove

the existence of a periodic orbit, only we constructed an orbit which avoided the periodic part. Similarly, we modify the orbit extension method to again build a non-periodic path. By analogy, we can call this targeting technique the "orbit restriction method" as it throws away the periodic part.

Given a pseudo-orbit $\{\boldsymbol{z}_i\}_{i=0}^{n+1}$ where $\boldsymbol{z}_i = (x_i, y_i)$, with epsilon errors

$$\|T^i(\boldsymbol{z}_{i-1}) - \boldsymbol{z}_i\| < \epsilon,\ i = 1, 2, ..., n+1, \tag{A.10}$$

one approach to find a true nearby orbit (which we expect to exist in a hyperbolic set due to the shadowing theorem [14]) is to simultaneously zero all $2n$ equations implied by (A.10). The existence of a Lagrangian for the dynamics T allows to alternatively require that the real orbit have stationary action. This will yield a n^{th} order tri-diagonal system.

The action of the pseudo-orbit is simply

$$W(x_0, x_1, ..., x_n) = \sum_{i=1}^{n} F(x_{i-1}, x_i), \tag{A.11}$$

where F is the Lagrangian generating function. The basis of our relaxation scheme will be to require that the "test" orbit $\{\boldsymbol{z}_i\}_{i=0}^{n+1}$ have stationary action

$$\delta W = 0. \tag{A.12}$$

Therefore, we obtain

$$\frac{\partial W}{\partial x_j} = \frac{\partial}{\partial x_j}(\sum F(x_{i-1}, x_i)) = F_2(x_{j-1}, x_j) + F_1(x_j, x_{j+1}) = 0, \tag{A.13}$$

where

$$F_1(a, b) = \frac{\partial F(a, b)}{\partial a} \text{ and } F_2(a, b) = \frac{\partial F(a, b)}{\partial b}. \tag{A.14}$$

We see that (A.13) defines a tri-diagonal system.

Once we have found the n x coordinates of the real orbit, we may recover the other n y coordinates using the action:

$$\begin{aligned} y &= -F_1(x, x') \\ y' &= F_2(x, x'). \end{aligned} \tag{A.15}$$

We will solve (A.13) by a Newton scheme when the ϵ-chain errors are small, thus giving a good initial guess. We rewrite each of the n equations of (A.13) in the form

$$f_j(x_{j-1}, x_j, x_{j+1}) = 0, \ j = 1, ..., n. \tag{A.16}$$

Note that we expect for a hyperbolic orbit small errors at the ends between the test orbit $\tilde{\boldsymbol{z}_j}$ and the real orbit $\boldsymbol{z}_j$. The errors we expect are proportional to $\|\boldsymbol{z}_0 - \tilde{\boldsymbol{z}_0}\| \approx k\epsilon\lambda_u^{-\frac{n}{2}}$ and $\|\boldsymbol{z}_n - \tilde{\boldsymbol{z}_n}\| \approx k\epsilon\lambda_s^{\frac{n}{2}}$.

Linearizing (A.16), we obtain

$$f_j(\boldsymbol{x} + \delta\boldsymbol{x}) = f_j(\boldsymbol{x}) + \sum_{i=1}^{n-1} \frac{\partial f_j}{\partial x_i}\delta x_i, \tag{A.17}$$

where we have written $\boldsymbol{x}^t = (x_0, ..., x_{n+1})$.

One Newton iteration is to solve the above linear system for the corrections $\delta\boldsymbol{x}$ which cause $f(\boldsymbol{x} + \delta\boldsymbol{x}) = 0$.

$$\sum_{i=0}^{n} (\frac{\partial f_j}{\partial x_i})\delta x_i = -f_i(x_{i-1}, x_i, x_{i+1}), \tag{A.18}$$

may be inverted with an efficient tri-diagonal solver.

For the standard map (1.37), the action is

$$F(x, x') = \frac{1}{2}(x - x')^2 + \frac{k}{4\pi^2}\cos(2\pi x_i), \tag{A.19}$$

and therefore we find

$$F_1(x, x') = (x' - x) - \frac{k}{2\pi}\sin(2\pi x), \tag{A.20}$$

and

$$F_2(x, x') = (x' - x). \tag{A.21}$$

Hence, we wish to zero the functions

$$f_j(x_{j-1}, x_j, x_{j+1}) = x_{j+1} - 2x_j + x_{j-1} - \frac{k}{2\pi}\sin(2\pi x_j) = 0. \tag{A.22}$$

Equation (A.18) calls for finding the derivatives $\frac{\partial f_j}{\partial x_i}$ of (A.22). We may now write (A.18) in its matrix form

$$A \cdot \delta \boldsymbol{x} = \boldsymbol{b}, \tag{A.23}$$

where

$$A_{m,n} = \frac{\partial f_m}{\partial x_n} = \left\{ \begin{array}{lll} 1 & \text{if} & n = m+1, \\ -\frac{k}{2\pi}\cos(2\pi x_m) & \text{if} & n = m, \\ -1 & \text{if} & n = m-1, \\ 0 & \text{otherwise}, & \end{array} \right\} \tag{A.24}$$

and,

$$b_n = f_n(x_{n-1}, x_n, x_{n+1}). \tag{A.25}$$

Some remarks are in order comparing this technique to that of Sec. 2.2.3. By shooting from $W^u(\boldsymbol{z}_0)$ to $W^s(\boldsymbol{z}_{n+1})$ we had only a first order system to zero by Newton's method. However, we only had errors at the end points $\boldsymbol{z}_0$ and $\boldsymbol{z}_{n+1}$. The orbit between has zero chain error by construction. Alternatively, minimizing (A.10) may result in an epsilon chain with the cumulative error distributed evenly through the patch and hence perhaps with much smaller maximum error as compared to the technique with error only at the ends. Ultimately, I have not pursued the variational formulation of patching at this time due to both its limited applicability and increased complexity. Shooting seems more natural and the model seems realizable.

A.3 An Initial Condition for Orbit Restriction

The standard map in its Hamiltonian form (1.37) may be rewritten in Lagrangian form

$$(x_{i+1} - x_i) - (x_i - x_{i-1}) + \frac{k}{2\pi}\sin(2\pi x_i) = 0. \tag{A.26}$$

This version of the standard map has an interpretation in solid state physics called the Frenkel-Kontorova model in which a chain of molecules is deposited on a one-dimensional periodic crystal lattice and experience nearest neighbor simple spring potentials of the form $V(x) = \frac{k}{4\pi^2}\cos(2\pi x)$ [70].

In this section we present an initial condition (a "pseudo-orbit") between the $\boldsymbol{a}$ box at $(0,1)$ to the $\boldsymbol{b}$ box at $(1,1)$, which using the orbit restriction method of the previous section, may be convergent to a single real orbit segment. A chain of molecules in the $(0,1)$ state has the orbit

$$x_j = 0,\ \forall j, \tag{A.27}$$

meaning they are all to be found in the same lattice site. The chain of molecules at the $(1,1)$ state has the orbit

$$x_j = j,\ \forall j, \tag{A.28}$$

which means that each consecutive molecule is to be found at the bottom of each consecutive potential well.

The idea is to write the initial guess for the pseudo-orbit as a concatenation of a long time of state (A.27) followed immediately by state (A.28).

$$\{x_j\}_{j=-n}^{n} = \left\{ \begin{array}{ll} 0 & j < 0 \\ j & j \geq 0 \end{array} \right\} \text{ where } n \to \infty. \tag{A.29}$$

This initial state is shown in Fig. (A.1). We know such an orbit exists near the anti-integrable limit (i.e. $k \to \infty$), and so we continue the orbit downward (lowering k) using the orbit at the previous value of k to prime the convergence towards an orbit at the next lower value of k.

Once given the initial guess, it may be possible to use (A.18) to relax the state to a real orbit between $\boldsymbol{a}$ and $\boldsymbol{b}$. A reasonable hypothesis [4] for the relaxed state is shown by the dashed hyperbolic curve also in Fig. (A.1), which represents a maximum smoothing of the discontinuous derivative change between states (A.27) and (A.28).

It is uncertain whether (A.29) is in the basin of attraction of a real orbit. It might be analogous to shooting from the unstable manifold of (near) $\boldsymbol{a}$ to (near) $\boldsymbol{b}$. In that case, we found that there are too many iterates in between, so zeroing equation (2.10) was impractical. The orbit restriction method does however have the advantage over the shooting method in that it simultaneously relaxes the entire middle portion of the pseudo-orbit, whereas the shooting method "flies blind" in the middle portion (and may numerically drift off the calculated manifold segments).

A.4 Shooting Between Higher-Dimensional Saddles

Here we will discuss the higher-dimensional generalization of Eq. (2.10), the shooting equation between the unstable manifold along the test orbit before

figure=F-K.eps,height=3.5in

Figure A.1. An initial configuration of the Frenkel-Kantorova model for the orbit restriction method. This represents a real orbit at the anti-integrable limit. The dashed hyperbolic curve is a reasonable guess at the form of the relaxed state. Left: molecules number i in their well sites x plotted next to a potential diagram $V(x)$. Right: Well site x_i as a function of molecule number i.

the recurrence to the stable manifold of the test orbit after the recurrence. This now requires the full Lyapunov spectrum.

First let us note that Eq. (2.10) is no longer sufficient on a $d > 2$ dimensional manifold. For $d > 2$, Eqs. (2.12)-(2.13) define the most (un)stable direction, not necessarily the complete (un)stable manifold. Eq. (2.10) attempts to shoot from the most unstable direction to the most stable direction. In general, such an intersection does not exist, therefore two curves in $d \geq 3$ generically do not intersect.

A codimension q hyperplane generically intersects a codimension p hyperplane if $q + p = d$. Recall the general definition of the stable and unstable manifold Eq. (1.25) and that the orbit through a point $\boldsymbol{z}$ is called hyperbolic if its tangent space is decomposable into the stable and unstable tangent subspaces, $E^s(\boldsymbol{z})$ and $E^u(\boldsymbol{z})$ as in Eq. (1.26). A hyperbolic saddle may be identified by a Lyapunov spectrum of q positive and p negative values on a d-dimensional manifold. Let us assume that the tangent subspace $E^s(\boldsymbol{z})$ is spanned by the vectors $\{\boldsymbol{v}_{s,l}\}_{l=1}^{p}$ and that $E^u(\boldsymbol{z})$ is spanned by the vectors $\{\boldsymbol{w}_{u,l}\}_{l=1}^{q}$.

In this setting we may generalize shooting a $2m+1$ step patch from the unstable manifold of $\boldsymbol{z}_{i-m}$ before a recurrence to the stable manifold of $\boldsymbol{z}_{i+s+m}$ after the s step recurrence. Eq. (2.10) generalizes by requiring that a point with initial condition

$$\boldsymbol{z}_0(\rho_1, \rho_2, .., \rho_q) = \boldsymbol{z}_{i-m} + \sum_{l=1}^{q} \rho_l \boldsymbol{w}_{u,l}, \tag{A.30}$$

has a $2m + 1^{th}$ iterate, $\boldsymbol{z}_{2m+1}(\rho_1, \rho_2, .., \rho_q)$ on the hyperplane

$$\boldsymbol{z}_{i+m+s} + \sum_{l=1}^{p} \varrho_l \boldsymbol{v}_{s,l}, \tag{A.31}$$

We may solve the d equation

$$\boldsymbol{z}_{2m+1}(\rho_1, \rho_2, .., \rho_q) - \boldsymbol{z}_{i+s+m} - \sum_{l=1}^{p} \varrho_l \boldsymbol{v}_{s,l} = 0, \tag{A.32}$$

for the d values $\{\rho_l\}_{l=1}^{q}$ and $\{\varrho_l\}_{l=1}^{p}$ by the Newton-secant method. This, however, is an overdetermined solution.[1] It should be possible to solve for only the independent variables $\{\rho_l\}_{l=1}^{q}$. For example, if $E^s(\boldsymbol{z}_{i+s+m})$ is two-dimensional and $d = 3$, we may write the iterate $\boldsymbol{z}_{2m+1}(\rho_1)$ on this tangent plane in the form

$$\boldsymbol{n} \cdot (\boldsymbol{z}_{2m+1}(\rho_1) - \boldsymbol{z}_{i+s+m}) = 0, \tag{A.33}$$

where

$$\boldsymbol{n} = \boldsymbol{v}_{s,1} \times \boldsymbol{v}_{s,2}, \tag{A.34}$$

is the normal to the plane. This eliminates the dependent variables ϱ_1 and ϱ_2.

More generally, we see that a p-dimensional hyperplane in d dimensions is of codimension q meaning that q equations are needed to specify the object. In other words, the d Eqs. (A.32) are over specified. We may solve this by considering q vectors which span $E^u(\boldsymbol{z}_{i+s+m})$:

$$span(\boldsymbol{w}_{u,1}, .., \boldsymbol{w}_{u,q}) = E^u(\boldsymbol{z}_{i+s+m}). \tag{A.35}$$

Since the tangent space is a direct product of stable and unstable linear subspaces, we may specify the stable subspace by stating that a vector

$$(\boldsymbol{z}_{2m+1}(\rho_1, \rho_2, .., \rho_q) - \boldsymbol{z}_{i+s+m}) \in E^s(\boldsymbol{z}_{i+s+m}) \tag{A.36}$$

[1] Recall that in Eq. (2.10), there was the t parameterization along the unstable direction which resulted in a τ parameterization along the stable direction. We could have solved for t and τ simultaneously by a two-dimensional Newton's method. Instead we were able to eliminate the dependent variable τ, leaving the solution of the independent variable t to a one-dimensional Newton's method.

has no component in any of q directions spanning $E^u(\boldsymbol{z}_{i+s+m})$. Hence we may replace the d equations (A.32) with the q equations,

$$(\boldsymbol{z}_{2m+1}(\rho_1, \rho_2, .., \rho_q) - \boldsymbol{z}_{i+s+m}) \cdot \boldsymbol{w}_{u,l} = 0,\ l = 1, .., q. \tag{A.37}$$

Let us consider the example of shooting from a one-dimensional unstable manifold ($span(\boldsymbol{v}_s) = E^s(\boldsymbol{z}_{i-m})$) to a one-dimensional stable manifold ($span(\boldsymbol{w}_u) = E^u(\boldsymbol{z}_{i+s+m})$) as in Eq. (2.10). Here Eq. (A.37) becomes the single equation

$$(\boldsymbol{z}_{2m+1}(\rho) - \boldsymbol{z}_{i+s+m}) \cdot \boldsymbol{w}_u = 0. \tag{A.38}$$

But any vector in $span(\boldsymbol{w}_u)$ is sufficient, including $(v_{s,y}, -v_{s,x})$ which is *orthogonal* to $\boldsymbol{v}_s = (v_{s,x}, v_{s,y})$. Thus we may write

$$(\boldsymbol{z}_{2m+1}(\rho) - \boldsymbol{z}_{i+s+m}) \cdot (v_{s,y}, -v_{s,x}) = (\boldsymbol{z}_{2m+1}(\rho) - \boldsymbol{z}_{i+s+m}) \times \boldsymbol{v}_s = 0, \tag{A.39}$$

which is an equivalent form to Eq. (2.10) which we included in [13].

Kostelich *et al.* [51] performed a similar shooting algorithm for *on the fly* control of the kicked double rotor map in four dimensions. After resolving hyperbolic saddles into two stable and two unstable directions, it was determined that two parameter perturbations generically specify a point intersection with the two-dimensional tangent stable manifold. The authors of [51] shoot from $\boldsymbol{z}_0$ to $E^s(\boldsymbol{z}_n)$ by making two successive perturbations of a single accessible system parameter k. The equation

$$T^{-(n-2)}(\boldsymbol{z}_n + \varrho_1 v_{s,1} + \varrho_2 v_{s,2}) = T_{k_0}(T_{k_1}(\boldsymbol{z}_0)), \tag{A.40}$$

was solved simultaneously for the four[2] variables $\varrho_1, \varrho_2, k_0$, and k_1.

[2]According to the above argument, four variables over-specifies the solution. Only two variables are required since $E^u(\boldsymbol{z}_n)$ is 2 dimensional.

A numerical estimation of the full Lyapunov spectrum is a delicate problem. In higher dimensions, Eqs. (2.12) and (2.13) define the most stable and the most unstable directions. These definition of f_u and f_s work analogously to the power method which rotates almost all arbitrary vectors to the dominant eigendirection.

Following the algorithm of Benettin *et al,* [7], higher order Lyapunov exponents and directions may be found in a stable fashion by monitoring the long term growth of volume elements under application of the tangent maps along the orbit. A spheroid evolves to an ellipsoid. A given volume element outlined by the two vectors v_1, v_2 has area $\mathrm{Det}([v_1, v_2])$ which evolves to $\alpha_1 = \mathrm{Det}([DT \cdot v_1, DT \cdot v_2])$. Analogously for (2.12)-(2.13) we can evolve these two vectors along a long orbit, rescaling by the area α_i at each step. The limit

$$\lambda_1 + \lambda_2 = \lim_{i \to \infty} \frac{\ln \alpha_i}{i}, \tag{A.41}$$

yields the second Lyapunov exponents, having previously calculated the first exponent λ_1. This technique may be performed recursively to find all the positive exponents in turn. Wolf *et al.* [106] successfully applied such an algorithm to extract exponents from experimental time-series data. The negative exponents may be extracted similarly to Eq. (2.13) by performing the process to the inverse transformations.

For shooting we require a set of vectors spanning the (un)stable manifold. These may be found recursively by a Gram-Schmidt process of following the evolution of n vectors simultaneously and then subtracting off components in the already found more dominant directions. Thus by evolving n vectors and subtracting off $n-1$ vectors spanning the $n-1$ more dominant (larger positive

λ_i) directions, we may find the n^{th} orthonormal vector to span[3] an n-dimensional unstable submanifold. Again, stable directions may be found by applying the process in reverse to the inverse transformation. There exists another competing theory of Greene and Kim [39] which extracts similar information pertaining to "magnitudes and directions of stretching and shrinking of an initial small sphere of realizations", based on the singular values of the linearization along the orbit.

Higher-dimensional targeting is a yet unfinished work. The discussion of this section furnishes some of the tools required to connect path segments once they have been identified.

A.5 An Example of Higher Dimensional "Transport Mechanism"

A useful technique, but of only limited success, is to understand complicated, higher-dimensional dynamics as the cross-product of low-dimensional, well understood dynamics. Consider, for example, the cross-product of the area-preserving Hénon map with rotation of constant twist:

$$\begin{aligned} x' &= y, \\ y' &= y(k-y) - x, \qquad \text{(A.42)} \\ I' &= I + \epsilon F(y), \\ \theta' &= [\theta + \omega(I)] mod 1, \qquad \text{(A.43)} \end{aligned}$$

where $F(\alpha) = F(\alpha + 1)$. For the uncoupled case ($\epsilon = 0$), each of these two-dimensional dynamical systems is well understood. The Hénon map, Eq. (A.42), makes a nice homoclinic tangle in $\mathbb{R}^2$, while the trivial twist map Eq. (A.43), on $S^1 \times \mathbb{R}$, in action angle form, is effectively represented by a map of the line alone. The full cross-product map can be viewed in $\mathbb{R}^3$ where we see stacked leaves in I

[3]These n directions do not generally point along the n unstable directions, which in fact are not well defined. However, a spanning set is sufficient for our purpose.

of homoclinic tangles. Each point in the (I, x, y) volume has a circular rotation in θ attached to it. Meanwhile, the Hénon dynamics is obeyed independently in the (x, y) plane (see Fig. (A.2)).

A barrier, formed of stable and unstable manifold segments between a p.i.p. and the fixed point as in Sec. 1.2.2, and then crossed with the action, forms a normally hyperbolic cylinder.[4] The transport dynamics are exactly as before, with the twist playing no influence.

figure=fig7.eps,width=5.0in

Figure A.2. The cross-product of a homoclinic tangle with a uniform twist map, in the presence of a small coupling ϵ. The transverse intersections of the homoclinic tangle persist for small enough ϵ, but the stable and unstable manifolds are free to wander between the various leaves, in the I dimension.

For small, nonzero coupling $0 < \epsilon \ll 1$, the transverse intersections of the Hénon map persist due to a local structural stability property of hyperbolic points under sufficiently small C^1 perturbations (see [3]). However, the (un)stable manifold can possibly wander between leaves, as depicted in Fig. (A.2). This in turn implies that there is now transport between the leaves. Hence, we see a complication in higher dimensions.[5] Properly defining a barrier may allow us to find the transport according to Eq. (1.24), but we see that a "nice" barrier, as in Sec. 1.2.2, may be more difficult to define.

[4]The unstable manifold segment, in $\mathbb{R}^2$ cross the twist direction, forms an unstable manifold "sheet" in $\mathbb{R}^3$, with points moving with component normal to its surface due to the influence of the transverse stable manifold intersections. Similarly, the stable manifold forms a sheet which, together they form a cylinder.

[5]This makes a nice model for the mechanism of Arnold diffusion through a given (vector) resonance channel. In the full Arnold diffusion, we see this picture for every resonance corresponding to the dynamics of rational relations between two coupled Hamiltonian maps, forming the interconnected "Arnold web." See Lichtenberg and Lieberman [54] for a description of transport rates through the various channels.

figure=diff.eps,height=3.5in

Figure A.3. Driving diffusion of a coupled map. The outlines of entrance lobes of the Hénon map Eq. (A.42) are shown in the (x, y) plane. Below, the projection of an above iterate of the entrance lobe in the (x, I) plane. Note that the coupling drive of the Hénon map causes the "constant" twist map Eq. (A.43) to have drift in I. Without the coupling, we would expect that the simple twist map has constant value I. This partly models the mechanism of "Arnold Diffusion".

APPENDIX B

B.1 Knowing the Switching Points Apriori

The following represents some early work toward a more straightforward search for paths through an area preserving map such as the standard map. For the discussion of this section, we will reprise the problem of Chapter 2 to find a path between the $\boldsymbol{a}$ box near the $(0,1)$ resonance to the $\boldsymbol{b}$ box near the $(1,1)$ resonance.

The basic targeting mechanism is an extension of the Shinbrot technique discussed in Sec. 1.1.3. Recall briefly that the idea is to iterate the initial point $\boldsymbol{a}$ by the map T with a small range of variations such as $\delta k \in [-\delta k_{max}, \delta k_{max}]$ which are represented by a grid of m points. Iterating these m grid points forward by T_{k_0} causes them to quickly align themselves along $W^u(\boldsymbol{a})$. Back-iterating a grid of points on a small variation (this time in phase space) around the target point $\boldsymbol{b}$ causes them to quickly align themselves along $W^s(\boldsymbol{b})$. Generically, there is a transverse intersection between the two intervals, and so by bisection, we would hope to be able to bracket the correct δk which iterates to the neighborhood of $\boldsymbol{b}$. However, we also saw that the intergrid spacing δz between grid points grows roughly as the Lyapunov number according to $\lambda^n \|\delta z\|$. The inverse of this quantity tells us the scaling of the number of grid points required to maintain a given spacing despite the unstable manifold's stretching and folding. Targeting a long orbit directly by this technique (such as for an area preserving map were transport is typically slow) is impossible as the number of

grid points required grows exponentially with the length of the path according to Eq. (1.22).

Breaking up the targeting problem into smaller pieces keeps n and therefore m to reasonable proportions. We achieve this by targeting between successive switching regions that lay between $\boldsymbol{a}$ and $\boldsymbol{b}$. Regions between $\boldsymbol{a}$ and $\boldsymbol{b}$ consist of resonance layers corresponding to all the frequencies ω where $\frac{0}{1} \leq \omega \leq \frac{1}{1}$. Inefficient orbits tend to spend a long time trapped in a given layer, mostly rotating with a period approximately that of the major periodic elliptic island chain. Since the most irrational tori (in a Diophantine sense) break-up last, their residual cantori tend to have smaller turnstiles and thus offer the most difficult barriers to transport. We have seen that the way between the resonance layers is through the turnstile of one of the bounding cantori. Between the $\frac{0}{1}$ layer and the $\frac{1}{1}$ layer one finds that the most difficult barriers correspond to

$$\frac{0}{1} < \frac{1}{\gamma^2} < \frac{1}{\gamma} < \frac{1}{1}, \tag{B.1}$$

where γ is the golden mean and $<$ describes the order in which one finds the various resonances in phase space.

The main difficulty here is to find an appropriate ordering on the importance of the possible switching points amongst the infinite number of irrational frequency cantori. The nobles[1] are the last Diophantine frequencies to break up and may be ordered accordingly. The critical function $a^{ss}(\omega)$ provides a partial order on the persistence between the noble frequencies in the standard map and may be used to rank switching points. The frequencies $\boldsymbol{w}$ may then be found by

[1]Noble numbers ω have continued fraction expansions ending in all ones after an arbitrary finite beginning. This makes these numbers particularly difficult to find a rational approximation, and so they have correspondingly large Diophantine constants. The size of the Diophantine constants govern the persistence of the tori, and so indicate the importance of the remaining partial barriers.

one of several specialized algorithms (such as orbit extension method [20]).

In an attempt to avoid the specialization to the standard map inherent in the above switching point identification and location scheme, we may turn to the method of common lobes introduced in Sec. 1.2.4. Recall that the technique allows us to identify the most important switching points by collecting a large number of orbits (2000 in this case) which achieve the studied transport and asking the question, "What regions do all the orbits have in common?" We have verified up to the 6 most important regions, and these do in fact correspond to prominent noble frequencies. The importance of this technique is that it is model free, requiring only a large data set of sample iterations which may be obtained even by time series embedding.

The costly step of targeting between switching points is testing the intersection of the forward iterating grid of variations against the backward iterating grid of variations. An $O(m^2)$ operation is required to test whether the line segment $\boldsymbol{z}_i, \boldsymbol{z}_{i+1}$ from the first grid intersects $\boldsymbol{z}'_j, \boldsymbol{z}'_{j+1}$ from the second grid at a point *inside* the trapezoid formed by these four points. Additionally, there is a clipping problem, as the phase space of the standard map is a torus, which on the computer is represented by unit square with opposite ends identified using a modulus function. Considering that this test must be performed for $i = 1, .., m-1$ against $j = 1, .., m-1$, we see the cost of the operation.

figure=C.F.Intersect.eps,height=2.0in

Figure B.1. a: For the line segment $l_1 = (\boldsymbol{z}_{i+1} - \boldsymbol{z}_i)$ to intersect $l_2 = (\boldsymbol{z}'_{j+1} - \boldsymbol{z}'_j)$ it is necessary that the line segment l_2 "straddles" the line continued from l_1. The condition may be checked by the sign of cross-products test. b: The condition however is not sufficient as is portrayed by this image. Reapplying the test after reversing the roles of primed and unprimed variables completes the sufficiency test.

It is worthwhile to write down a clean test for the intersection between the two grids which avoids too many "if then" statements. The following is just such a test. In Fig. (B.1) we have drawn a configuration of four grid points $\boldsymbol{z}_i, \boldsymbol{z}_{i+1}, \boldsymbol{z}'_j, \boldsymbol{z}'_{j+1}$ which intersect inside their trapezoid. A requirement for their intersection is that the line segment $\boldsymbol{z}'_{j+1} - \boldsymbol{z}'_j$ straddles the line through the line segment $\boldsymbol{z}_{i+1} - \boldsymbol{z}_i$ between $\boldsymbol{z}_{i+1}$ and $\boldsymbol{z}_i$. The reader may check that this is equivalent to the statement that

$$v_1 = |\boldsymbol{z}'_j - \boldsymbol{z}_i||\boldsymbol{z}_{i+1} - \boldsymbol{z}_i| \sin\alpha, \tag{B.2}$$

has a sign opposite to the sign of

$$v_2 = |\boldsymbol{z}'_{j+1} - \boldsymbol{z}_i||\boldsymbol{z}_{i+1} - \boldsymbol{z}_i| \sin\beta. \tag{B.3}$$

We see that (B.2) and (B.3) may be evaluated as the cross products so that we equivalently require that

$$\begin{aligned} v_1 &= (\boldsymbol{z}'_j - \boldsymbol{z}_i) \times (\boldsymbol{z}_{i+1} - \boldsymbol{z}_i), \\ v_2 &= (\boldsymbol{z}_{j+1} - \boldsymbol{z}_i) \times (\boldsymbol{z}_{i+1} - \boldsymbol{z}_i), \end{aligned}$$

and

$$sgn(v_1)sgn(v_2) < 0. \tag{B.4}$$

Since this still allows the situation portrayed in Fig. (B.1), we need to reverse (rotate) the roles of the primed and the unprimed variables to further require that

$$\begin{aligned} v_3 &= (\boldsymbol{z}_i - \boldsymbol{z}'_j) \times (\boldsymbol{z}'_{j+1} - \boldsymbol{z}'_j), \\ v_4 &= (\boldsymbol{z}_{i+1} - \boldsymbol{z}'_j) \times (\boldsymbol{z}'_{j+1} - \boldsymbol{z}'_j) \end{aligned}$$

and

$$sgn(v_3)sgn(v_4) < 0. \tag{B.5}$$

Together, (B.4) and (B.5) represent a cheap and easy test for intersection which may be easily performed along the entire grids.

Now putting this all together, we will target between the first switching point, which we label "$\boldsymbol{a}$" and the second switching point, which we label "$\boldsymbol{b}$" in Fig. (B.2a), both found by common lists. In the Fig. (B.2a), we have also shown the golden mean cantori. We can see that the common region $\boldsymbol{b}$ is located in a gap in the $\frac{1}{\gamma^2}$ cantorus, indicating that the technique indeed has found what we know to be an important turnstile. Starting a parameter perturbation of $\delta k = 0.02$ around the nominal value $k = 1.25$ at $\boldsymbol{a}$ we iterate a grid of 500 points forward until we detect an intersection with an inverse image of the 500 point grid on the $\delta z = 0.02$ variation around $\boldsymbol{b}$. We can see in the figure how the two variations spend time twisting around the torus, as they stretch and fold, until they are both large enough and in the same vicinity in which an intersection occurs. The detected intersection is denoted by a solid box. Since 10 forward iterations and 8 backward iterations were required, we have shown a 18 step path between $\boldsymbol{a}$ and $\boldsymbol{b}$ requiring the above perturbations.

I must mention that the situation is not always as nice as I have indicated in the Fig. (B.2a). If the perturbation ranges started at $\boldsymbol{a}$ and $\boldsymbol{b}$ are too small, then by the time the two grids have reached the same resonance layer (in approximately 10 and 8 iterations respectively) they may not have stretched sufficiently so that an intersection between them is likely. Typically, a small grid of points around a center point can wander chaotically between layers with the center point. Only when the grid has grown enough so that different grid points can incorporate different behaviors can a new path be found. The trick is to hope that the two grids grow to a "reasonable size" in the same resonance layer. This need is contrary to the competing need that there are enough grid points

figure=2in1k2.eps,height=3.0in

Figure B.2. Building a path for the standard map between known switching points by following variations on a grid. Left: A small variation in a cantorus at $\boldsymbol{a}$ is iterated forward, while a small variation at $\boldsymbol{b}$ in a nearby cantorus is iterated backward until an intersection between the two curves of perturbations is detected, shown as a box between. Right: This process is repeated for successive nearby "important" resonances until a path all the way from the $\boldsymbol{a}$ box to the $\boldsymbol{b}$ box is found.

on a given range so that the intergrid point distance remains small. Otherwise, the grid grows incoherent, and the grid ordering is lost. Having enough switching points so that the big targeting problem is sufficiently broken into smaller targeting problems plays an important role. Some trial and error is required to choose the grid sizes, perturbation ranges, and number of switching points so that the "intermediate" range may be achieved.

By trial and error, we can find workable grid sizes and perturbations to target all the switching points between the $\frac{0}{1}$ box and the $\frac{1}{1}$ box. For $k = 1.25$, 9 switching regions were required. Thus 9 $\delta k < 0.02$ perturbations were required to target a 250 step path. All the variations tested are depicted by the "scribbles" strewn across the standard map portrait in Fig. (B.2b).

This work demonstrates a completely different approach to the same targeting problem for which we used cut and glue at recurrences in previous chapters. Comparing the two, we see that the forward iteration of a variation around $\boldsymbol{a}$ aligns itself along $W^u(\boldsymbol{a})$ and inverse iteration of a variation around $\boldsymbol{b}$ aligns itself along $W^s(\boldsymbol{b})$. Hence the point we find and refine in the intersection of the two grids lies on (nearly) the intersection of $W^u(\boldsymbol{a})$ and $W^s(\boldsymbol{b})$. So in this sense, the Shinbrot technique is similar in its end result to the shooting technique developed here.

In terms of computational efficiency, using a grid is inferior. Typically

mn map applications are required to advance the two m point grids in an n step path, and approximately $\frac{n}{2}m^2$ checks for grid intersections are required along the way. In contrast, the shooting technique requires just a handful of map iterations for an intelligent initial guess (according to Eq. (2.11)). When a good guess is chosen, convergence is extremely fast. Table (2.1) address the issue of success ratios.

The grid based targeting method is very similar in computational requirements and in methodology to manifold continuation which is more easily compared to shooting. Recall that the technique in Chapter 2 strives to find $\boldsymbol{p}$ on the intersection of $W^u(T^{\frac{n}{2}}(\boldsymbol{a}))$ and $W^s(T^{-\frac{n}{2}}(\boldsymbol{b}))$, by shooting from $W^u(\boldsymbol{a})$ to $W^s(\boldsymbol{b})$. The point $\boldsymbol{p}$ may be found alternatively by continuing the manifolds $W^u(\boldsymbol{a})$ and $W^s(\boldsymbol{b})$ as defined by Eqs. (1.25)-(1.26) using the stable manifold theorem. Continuing the (un)stable manifolds requires starting a grid of points along the (un)stable direction, and then iterating. This process is continued until we find a $\boldsymbol{p}$. The steps to perform this task are analogous to Shinbrot's technique and even many of the grid intersection subroutines can be recycled. Since we still want a p.i.p., the special guess, Eq. (2.11), helps us identify where we should look for $\boldsymbol{p}$.

Note that techniques presented in this section are not as good (in terms of generality and computer speed) as cutting loops at recurrences. This approach however, was in my thesis proposal and so I feel obliged to include it here. I opted in this thesis for the Newton based targeting method for reason of computational efficiency of shooting at a single point instead of brute force tracking a whole grid. I also chose the more natural approach of allowing the dynamics to show us the switching points through recurrence instead of looking for them at resonances or by using common lobes. Nonetheless, I hope that the techniques in this section

have provided an interesting alternative to the targeting problem. In addition, this work is the solution of the problem and approach that I presented in my thesis proposal, before I realized the more efficient method of constructing a shadow orbit by removing recurrences.