DISSERTATION

DISCRETE-TIME TOPOLOGICAL DYNAMICS, COMPLEX HADAMARD MATRICES, AND OBLIQUE-INCIDENCE ION BOMBARDMENT

Submitted by

Francis Charles Motta

Department of Mathematics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2014

Doctoral Committee:

Advisor: Patrick D. Shipman

Gerhard Dangelmayr
Chris Peterson
R. Mark Bradley

ABSTRACT

DISCRETE-TIME TOPOLOGICAL DYNAMICS, COMPLEX HADAMARD MATRICES, AND OBLIQUE-INCIDENCE ION BOMBARDMENT

The topics covered in this dissertation are not unified under a single mathematical discipline. However, the questions posed and the partial solutions to problems of interest were heavily influenced by ideas from dynamical systems, mathematical experimentation, and simulation. Thus, the chapters in this document are unified by a common flavor which bridges several mathematical and scientific disciplines.

The first chapter introduces a new notion of orbit density applicable to discrete-time dynamical systems on a topological phase space, called the linear limit density of an orbit. For a fixed discrete-time dynamical system, $\Phi(x) : M \to M$ defined on a bounded metric space, we introduce a function $E : \{\gamma_x : x \in M\} \to \mathbb{R} \cup \{\infty\}$ on the orbits of Φ, $\gamma_x \doteq \{\Phi^t(x) : t \in \mathbb{N}\}$, and interpret $E(\gamma_x)$ as a measure of the orbit's approach to density; the so-called **linear limit density** (LLD) of an orbit. We first study the family of dynamical systems $R_\theta : [0,1) \to [0,1)$ ($\theta \in (0,1)$) defined by $R_\theta(x) = (x + \theta) \mod 1$. Utilizing a formula derived from the Three-Distance theorem, we compute the exact value of $E(\{R_\phi^t(x) : t \in \mathbb{N}\}, x \in [0,1))$, where $\phi = \left(\sqrt{5} - 1\right)/2$. We further compute $E(\{R_\theta^t(x) : t \in \mathbb{N}\}, x \in [0,1))$ for a class of irrational rotation angles $\theta = [j, j, \ldots]$ with period-1 continued fraction expansions and discuss how this measure distinguishes the topologically transitive behavior of different choices of θ. We then expand our focus to a much broader class of orientation-preserving homeomorphisms of the circle and extend a result of R. Graham and J.H. van Lint about optimal irrational rotations. Finally, we consider the LLD of orbits of the Bernoulli shift map acting on sequences defined over a finite alphabet and prove bounds for a class of sequences built by recursive extension

of de Bruijn sequences. To compute approximations of $E(\gamma_x)$ for orbits of the Bernoulli shift map, we develop an efficient algorithm which determines a point in the set of all words of a fixed length over a finite alphabet whose distance to a distinguished subset is maximal.

Chapter two represents a departure from a dynamical systems problem by instead exploring the structure of the space of complex Hadamard matrices and mutually unbiased bases (MUBs) of complex Hilbert space. Although the problem is not intrinsically dynamical, our mechanisms for experimentation and exploration include an algorithm which can be viewed as a discrete-time dynamical system as well as a gradient system of ordinary differential equations (ODEs) whose fixed points are dephased complex Hadamards. We use our discrete system to produce numerical evidence which supports existing conjectures regarding complex Hadamards and mutually unbiased bases, including that the maximal size of a set of 6×6 MUBs is four. By applying center-manifold theory to our gradient system, we introduce a novel method to analyze the structure of Hadamards near a fixed matrix. In addition to formalizing this technique, we apply it to prove that a particular 9×9 Hadamard does not belong to a continuous family of inequivalent matrices, despite having a positive defect. This is the first known example of this type.

The third chapter explores the phenomenon of pattern formation in dynamical systems by considering a model of off-normal incidence ion bombardment (OIIB) of a binary material. We extend the Bradley-Shipman theory of normal-incidence ion bombardment of a binary material by analyzing a system of partial differential equations that models the off-normal incidence ion bombardment of a binary material by coupling surface topography and composition. In this chapter we perform linear and non-linear analysis of the equations modeling the interaction between surface height and composition and derive a system of ODEs which govern the time-evolution of the unstable modes, allowing us to identify parameter

ranges which lead to patterns of interest. In particular, we demonstrate that an unusual "dots-on-ripples" topography can emerge for nonzero angles of ion incidence θ. In such a pattern, nanodots arranged in a hexagonal array sit atop a ripple topography. We find that if dots-on-ripples are supplanted by surface ripples as θ or the ion energy are varied, the transition is continuous.

ACKNOWLEDGMENTS

Had you asked the author at the beginning of his PhD program how many hands would eventually make significant contributions to this document, his estimation would undoubtedly have fallen far short of the actual number. What follows is the culmination of a great effort on the part of the author and a multitude of collaborators, buttressed by the the ever-present support of friends and family. First I would like to thank my advisor, Patrick Shipman. I dare not try and list all of the adjectives which comprise his efficacy as an advisor. Instead I will simply say that I know how lucky I am to have Patrick as my advisor. He is tremendous. I am also deeply grateful to Mark Bradley who has not only been a constant mentor and collaborator, but has also been a source of tremendous insight into the academic world. Much thanks goes to Chris Peterson for being unfailingly generous with his ideas and his time. Chris is a fountainhead of mathematical intimations and I owe him for more than one research topic. Thank you to the other members of my committee, Gerhard Dangelmayr and Renzo Cavalieri, for useful suggestions during my preliminary examination and the latter for being an exceptional instructor and a kindred spirit on the issues of mathematical instruction and the power of visualization as a path to understanding.

I would like to acknowledge my peers, Bethany Springer, Eric Hanson, and Lori Ziegelmeier. Each of you is a collaborator and a friend. I have been privileged to share with you this experience.

In addition to those people who directly contributed to this dissertation, I would like to thank the slew of teachers that crafted me into the academic I am today. There are too many, and too many I will no doubt forget, to list here. To each of you who influenced my trajectory I am indebted in a profound way, such that reaching this point serves only as a partial repayment. I expect to work off the rest with a lifetime of likewise mentorship.

Thank you to Colorado State University: Within your walls I grew up and was changed for the better by the pressure of this endeavor.

Finally, let me express my love and gratitude to my family. My mom and dad, Karen Ciesielski-Motta and Francis Motta, for instilling in me a deep curiosity and a love of knowledge. Thanks to my mother-in-law, Terri Leboeuf, for always supporting me like I was her own son and my father-in-law, Michael Leboeuf, who called me "Doctor Francis" even ten years ago when I was just a dreadlocked undergraduate, falling for his daughter. Which brings me to my wife Jordan; thank you, thank you, thank you for everything there is to thank you for.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Optimally Topologically Transitive Orbits in Discrete-Time Dynamical Systems

1.1. Introduction

A discrete-time dynamical system $f : M \to M$ defined on a topological phase space M is said to be **topologically transitive** on M if for all open $U, V \subseteq M$

$$f^m(U) \cap V \neq \emptyset$$

for some $m \in \mathbb{N}$. Many discrete-time dynamical systems possess this transitivity property, however, we will devote much of our attention to the classical and well-studied family of rigid rotations of the circle by an irrational angle.

For a fixed $\theta \in (0, 1)$, and for $S^1 \subset \mathbb{C}$, the complex unit circle, define $R_\theta : S^1 \to S^1$ as

$$R_\theta(z) = e^{i\theta} z.$$

If we think of the unit interval $[0, 1) \cong \mathbb{R}/\mathbb{Z}$, as being a circle with circumference 1 (having identified the endpoints) we can interpret R_θ as the translation of the interval by θ, reducing mod 1 to ensure the image after translation remains [0,1). Explicitly $R_\theta(x) : [0, 1) \to [0, 1)$ is defined to be

$$R_\theta(x) = (x + \theta) \mod 1.$$

We will refer to both maps as R_θ and call them rotations (either of the circle or unit interval), often without explicit mention of which domain is under consideration, unless the identification improves clarity or precision.

For any irrational θ, not only is R_θ topologically transitive, but the forward orbit of any point $x \in [0,1)$,

$$\gamma_x \doteq \{R_\theta^m(x) : m \in \mathbb{N} \cup \{0\}\} = \{\{x + m\theta\} : m \in \mathbb{N} \cup \{0\}\}$$

(where $\{z\} \doteq z - \lfloor z \rfloor$, the fractional part of z), is a dense subset of $[0,1)$. That is, for any initial $x \in [0,1)$ and any $\epsilon > 0$, given $y \in [0,1)$ there exists an $n \in \mathbb{N}$ such that $|R_\theta^n(x) - y| < \epsilon$. However, consider the first 45 points in the forward orbit of $x = 0$ under rotation by the golden number $\phi \doteq \left(\sqrt{5} - 1\right)/2$ compared to rotation by $\theta = 4 - \pi$, as shown in Figure 1.1. While it is true that the full orbits of 0 are dense for both maps, we notice a striking difference in the distribution of the points in these truncated orbits. In both cases, there exist (many) points y in [0,1) and $\epsilon(y) > 0$ such that $B_{\epsilon(y)}(y) \cap \{0, R_\theta(0), \ldots, R_\theta^{44}(0)\} = \emptyset$. However, for $\theta = 4 - \pi$, the maximal ϵ for which this intersection is trivial for some point $y \in [0,1)$ is clearly larger than the maximal ϵ for which there exists a point in [0,1) whose ϵ-neighborhood is disjoint from $\{R_\phi^i(0) | i = 0, 1, \ldots, 44\}$.

It is this observation which motivates us to extend the notion of the transitivity of a map to the (topological) transitivity of an orbit, and to quantify the efficiency of an orbit's transitivity.

Definition 1.1.1. An orbit $\gamma_x = \{f^m(x) | m \geq 0\}$ of a discrete-time dynamical system $f : M \to M$ will be called a **topologically transitive orbit** if for every open subset $U \subseteq M$, there exists $m \in \mathbb{N}$ such that $f^m(x) \in U$.

If M is a metric space, with a topology inherited from the metric, then the orbit γ_x is topologically transitive if for every $y \in M$ and every $\epsilon > 0$, there exists $m \in \mathbb{N}$ such that $f^m(x) \in B_\epsilon(y)$. Clearly, an orbit is topologically transitive if and only if it is dense. It is,

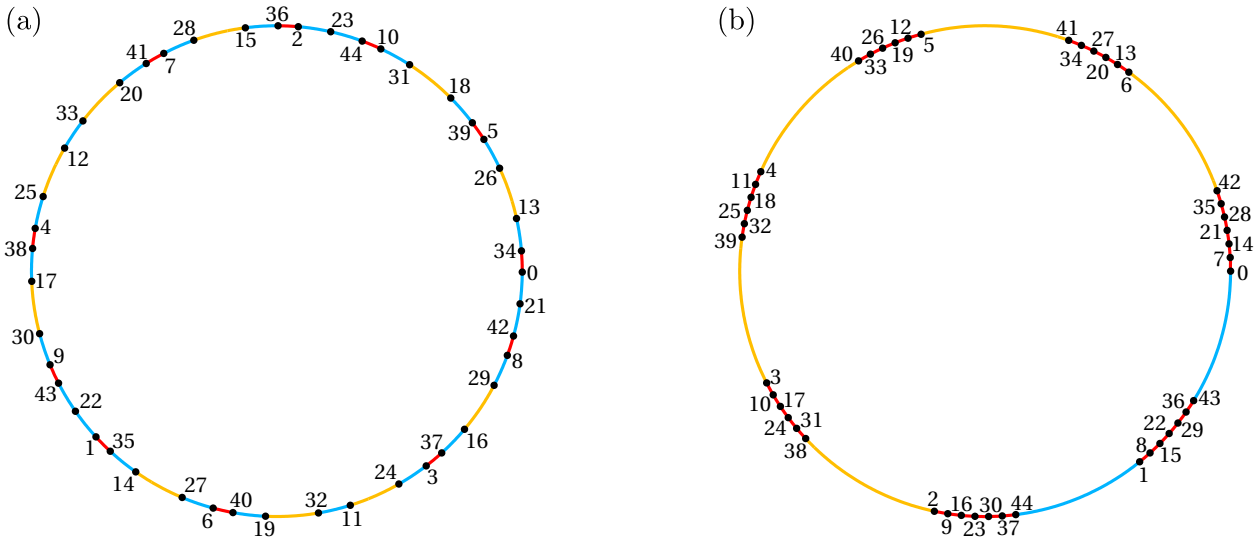

FIGURE 1.1. (a) The first 45 iterates of $x = 0$ under R_ϕ for $\phi = \left(\sqrt{5} - 1\right)/2$. (b) The first 45 iterates of $x = 0$ under R_θ for $\theta = 4 - \pi$. Iterates are labelled and arcs between consecutive points in each orbit are colored according to their relative length.

however, the behavior of the orbit's approach to density (considering truncated orbits at finite times) which is of interest in this paper. Using this language, the distinction between R_ϕ and $R_{4-\pi}$ can be made more precise. Fix $\epsilon > 0$ and initial condition $x \in S^1$. If $m_\phi \in \mathbb{N}$ and $m_{4-\pi} \in \mathbb{N}$ are the smallest integers satisfying the definition of orbit transitivity for R_ϕ and $R_{4-\pi}$ respectively, then we expect that $m_{4-\pi} \gg m_\phi$. That is, we expect that the smallest m for which $B_\epsilon(y) \cap \{R^i_{4-\pi}(x) | 0 \leq i \leq m\} \neq \emptyset$ for every $y \in S^1$ is larger than the smallest m for which $B_\epsilon(y) \cap \{R^i_\phi(x) | 0 \leq i \leq m\} \neq \emptyset$ for every $y \in S^1$. This distinction is how we will ultimately define optimality of the transitivity of an orbit, refining the notion of an orbit being dense by assigning to it a numerical value derived from its asymptotic approach to density.

In Section 1.2 of this paper, we make precise a measure of the optimality of an orbit's transitivity. We then apply this general definition to the particular family of systems introduced above: rigid rotations of the circle by an irrational angle. Using some elementary tools, we explicitly compute this measure for a class of irrational numbers. We then recall a result

of R.L. Graham and J.H. van Lint [1] which essentially answers the question of which irrational rotation has optimally topologically transitive orbits. Finally, we provide an extension of their result to a much broader class of dynamical systems defined by homeomorphisms of the circle, namely those that are diffeomorphically conjugate to a rotation.

In Section 1.3, we turn our attention to another classical family of dynamical systems with transitive orbits, namely Bernoulli shift maps of sequence spaces over finite alphabets. The question of which initial conditions have optimally transitive orbits leads us to an exploration of de Bruijn sequences and their orbits under the shift map.

1.2. Maps of the Circle or Unit Interval

Throughout this section, we focus our attention on maps of the circle. However, we begin in Section 1.2.1 by making a general definition which provides a quantitative measure of the optimality of an orbit's topological transitivity. In Sections 1.2.2 and 1.2.3, we explicitly compute this measure for several members of a family of irrational rotations and relate it to a result of Graham and van Lint which shows that amongst all rotations, the rotation by the golden number ϕ has optimal orbit transitivity. In Section 1.2.4, we show (Theorem 1.2.10) that nearly all diffeomorphisms of the circle define dynamical systems that have orbits with suboptimal transitivity (that is, less optimal than any orbit of rotation by ϕ).

1.2.1. A Measure of the Optimality of Orbit Transitivity.

In the preceding section, we observed that an orbit under the rotation R_ϕ approached density in a more uniform manner than R_θ for $\theta = 4-\pi$. More precisely, for $\theta = 4-\pi$ the largest positive number ϵ for which $B_\epsilon(y) \cap \{R_\theta^i(0) | i = 0, 1, \ldots, 44\}\}$ is trivial for some point $y \in [0, 1)$ is clearly larger than the maximal ϵ for which there exists a point in [0,1) whose ϵ-neighborhood is disjoint

from $\{R^i_\phi(0)|i = 0, 1, \ldots, 44\}$. Motivated by this observation, we make the following (very general) definition of a measure of a subset's failure to be dense:

DEFINITION 1.2.1. Let M be a bounded metric space equipped with the metric $d : M \times M \to \mathbb{R}$. Fix a subset $D \subseteq M$, and for each $y \in M\backslash D$ define the **density failure epsilon** of D at y to be

$$\epsilon_D(y) \doteq \sup\{\epsilon \geq 0 | B_\epsilon(y) \cap D = \emptyset\}.$$

Further define the **density failure** of D in M to be

$$\mathcal{E}(D) \doteq \sup_{y \in M\backslash D} \{\epsilon_D(y)\}.$$

Although we are interested in countable dense subsets of M, the existence of which does not require boundedness (e.g. $\mathbb{Q} \subset \mathbb{R}$), assuming M is bounded ensures it has a finite diameter with respect to d. If the diameter of M is $m < \infty$, then m is an upper bound for the density failure epsilon of any subset of M at any point. The axiom of completeness of $\mathbb{R}$ thus guarantees that the density failure is defined for every subset of a bounded metric space M.

PROPOSITION 1. *D is dense in M iff $\mathcal{E}(D) = 0$.*

PROOF. First observe that D is dense in M iff $\forall x \in M\backslash D$, $\epsilon_D(x) = 0$. To see this, assume that D is dense in M. Then for any $x \in M\backslash D$ and any $\epsilon > 0$ there exists $y \in D$ such that $d(x, y) < \epsilon$. $B_0(x) \cap D = x \cap D = \emptyset$ since we assumed $x \notin D$. Thus, $\epsilon_D(x) = 0$ for every $x \in M\backslash D$.

Now assume $\epsilon_D(x) = 0$ for each $x \in M \backslash D$. Let $\epsilon > 0$ be given and $x \in M$ fixed. If $x \in D$, then $x \in B_\epsilon(x) \cap D$, and so the intersection is nonempty. If $x \in M \backslash D$, then since $\epsilon > 0 = \epsilon_D(x) = \sup\{\epsilon \geq 0 | B_\epsilon(x) \cap D = \emptyset\}$ it follows that $B_\epsilon(x) \cap D \neq \emptyset$, i.e. there exists $y \in D$ such that $d(x, y) < \epsilon$. Therefore, D is dense in M.

Finally, observe that $\mathcal{E}(D) = 0$ iff $\forall x \in M \backslash D$, $\epsilon_D(x) = 0$. □

We will therefore interpret a small density failure to mean that a subset is close to achieving density. We do not define what we mean by "small," or "close" in this context, as it will depend on the metric space under consideration. One could modify this definition to normalize by the diameter of M to help dissipate this ambiguity.

Although we have defined $\mathcal{E} : \mathcal{P}(M) \to \mathbb{R}$ on the entire power set of M, we are particularly interested in those subsets which are the orbits of discrete time dynamical systems acting on the space. For example, if we choose $M = [0, 1)$ (the circle of radius $1/2\pi$) the observation that $\gamma_x \doteq \{R_\theta^t(x) | t \in \mathbb{N}\}$ is dense in M (for irrational choices of θ) can be stated in terms of our density failure.

PROPOSITION 2. *Define* $\gamma_x(m) \doteq \{R_\theta^t(x) | 0 \leq t \leq m\}$ *to be the truncated orbit of length* $m + 1$ *for fixed* θ *and* x. *Then,* $\lim_{m \to \infty} \mathcal{E}(\gamma_x(m)) = 0$.

PROOF. Let $\epsilon > 0$ be given. Since γ_x is dense in M, there exists $m \in \mathbb{N}$ such that for any $y \in M, B_{\epsilon/2}(y) \cap \gamma_x(m) \neq \emptyset$. Thus $\mathcal{E}(\gamma_x(m)) \leq \epsilon/2 < \epsilon$. □

As an exercise, let us compute the density failure, $\mathcal{E}(\gamma_0(49))$, of the truncated orbits of length 50 for each of the irrational rotations $\phi = (1 - \sqrt{5})/2$ and $\theta = 4 - \pi$. Since $M = [0, 1)$ is 1-dimensional, and $\gamma_0(49)$ is a finite list of points, the midpoint of the largest gap between neighboring points in $\gamma_0(49)$ is the point $y \in M \backslash \gamma_0(49)$ for which the density failure epsilon is maximized. The density failure is therefore half the maximum distance between any two

neighboring points. Thus, for R_ϕ, the density failure of the truncated orbit is approximately 0.01722, while for R_θ the density failure of the truncated orbit is approximately 0.04424. Here we are measuring distance along geodesics (as opposed to the Euclidean distance between points embedded in the plane), a point worth mentioning since the value of the density failure of a subset is not independent of the choice of metric.

We now define a measurement of the optimality of an orbit's transitivity:

DEFINITION 1.2.2. Let (M, d) be a bounded metric space and let $\Phi : M \to M$ be a function. Define the **linear limit density** (LLD) function $E : \{\gamma_x | x \in M\} \to \mathbb{R} \cup \{\infty\}$ on the orbits, γ_x, of Φ (the dynamical system defined by $x_{n+1} = \Phi(x_n)$) to be

$$E(\gamma_x) \doteq \limsup_{m\to\infty} \left[(m+1)\mathcal{E}(\gamma_x(m))\right].$$

Put more precisely, the LLD is the limit superior of the linear-time rescaled density failure of a truncated orbit. Intuitively, a large LLD corresponds to an orbit which (infinitely often) takes a "long" time to achieve a specified level of density failure. In this sense, the LLD is a measure of an orbit's topological transitivity: the smaller the LLD the more optimal the orbit's transitivity. If an orbit is not dense, then certainly its LLD will be equal to infinity. We will see that the converse is not in general true.

The asymptotic behaviors of several related quantities were explored in a paper by R. L. Graham and J.H. van Lint in 1966 [1]. Motivated by the question of the existence of an ergodic stationary stochastic process with zero entropy [2], Graham and van Lint discussed the function

$$d_\theta(m) = \max_{1\le i\le m+1} a_i - a_{i-1},$$

where $0 = a_0 < a_1 < a_2 < \ldots a_n < a_{m+1} = 1$ is a relabeling of the $m+1$ points in $\gamma_0(m)$ associated to the irrational rotation by an angle θ (plus the point $a_{m+1} \doteq 1$). In other words, $d_\theta(m)$ is the largest gap between neighboring points in the finite forward orbit (of 0) consisting of $m+1$ points. In the language introduced here,

$$d_\theta(m) = 2\,\mathcal{E}(\gamma_x(m)).$$

In [1], the following results are proved:

THEOREM 1.2.3 (**Graham and van Lint**).

$$\sup_{\theta\in(0,1)} \liminf_{m\to\infty} [m d_\theta(m)] = \frac{1+\sqrt{2}}{2},$$

$$\inf_{\theta\in(0,1)} \limsup_{m\to\infty} [m d_\theta(m)] = 1 + \frac{2\sqrt{5}}{5},$$

and

$$\limsup_{m\to\infty} [m d_\theta(m)] = \infty \iff \limsup_{m\to\infty} b_m = \infty,$$

where $\theta = [b_1, b_2, b_3, \ldots]$ *is the simple continued fraction expansion of* θ.

Results of this kind are of interest, in part, because very little is known about the nature of the integer sequences given by the continued fraction expansion of irrational numbers other than quadratic surds, which are exactly those numbers with (infinite) periodic continued fractions [3]. It is this fact which will motivate our approach to understanding the asymptotic behavior of these ergodic systems and ultimately will lead us to consider certain subsequences of the sequence $((m+1)\mathcal{E}(\gamma_x(m)))_{m\in\mathbb{N}}$.

It is not surprising that the proof of these results relies on a closed-form expression for $d_\theta(m)$. For that, we must introduce the Steinhaus Conjecture, turned theorem (see [4], [5],

[6], [7]), which gives a beautiful description of the way in which points are distributed on the circle via rotation by an irrational number. First, we formally fix our notation.

Let

$$[b_1, b_2, b_3, \ldots] = \cfrac{1}{b_1 + \cfrac{1}{b_2 + \cfrac{1}{b_3 + \ddots}}}$$

be the simple continued fraction expansion (CFE) of a number $\theta \in (0, 1)$.

DEFINITION 1.2.4. We refer to b_k as the kth **partial quotient** of θ. The **partial convergents** of θ are the ratios $p_k/q_k \doteq [b_1, b_2, ..., b_k]$ obtained by computing (in lowest terms) the rational number defined by the first k terms in the truncated CFE of θ.

It is well known (see [3]) that the partial convergents of θ are the **best rational approximates** of θ, in the sense that $|\theta - p_n/q_n| < |\theta - p/q|$ for any rational p/q with $q < q_n$. Of course, not every integer q can be the denominator of a best rational approximate. As it happens, the partial convergents of θ are all of the best rational approximates of θ. Additionally, the partial convergents satisfy the following recurrence relations:

$$\begin{aligned} &p_{-1} = 1, \quad p_0 = b_0, \quad p_k = b_k\, p_{k-1} + p_{k-2}, \quad k \geq 1, \\ &q_{-1} = 0, \quad q_0 = 1, \quad q_k = b_k\, q_{k-1} + q_{k-2}, \quad k \geq 1. \end{aligned} \tag{1}$$

We are now prepared to state the theorem which will empower us to compute the linear limit densities of irrational rotation maps.

THEOREM 1.2.5 (**The Three-Gap/Distance Theorem**). *Fix* $\theta = [b_1, b_2, ...]$. *For* $m+1$ *points distributed by rotations of the unit interval by* θ, *(i.e.* $\gamma_x(m), \forall x \in (0,1)$*), there are at most three distinct distances between adjacent points in this truncated orbit. In particular, one can uniquely write* $m \geq 1$ *as* $m = sq_k + q_{k-1} + r$, *where* $1 \leq s \leq q_{k+1}$ *and* $0 \leq r \leq q_k$. *Then there are exactly*

$$s + 1 - q_k \text{ gaps between successive points of length } d1_k,$$
$$r + 1 \text{ gaps between successive points of length } d2_k, \text{ and}$$
$$q_k - r - 1 \text{ gaps between successive points of length } d3_k,$$

where $d1_k < d2_k < d3_k$ *and* $d3_k = d2_k + d1_k$*: the largest gap is the sum of the smaller two gaps.*

While we will not present a proof of the Three-Distance Theorem here, the truthfulness of the theorem is made quite apparent by employing a powerful tool from ergodic theory, referred to as interval "cutting and stacking," which we develop in the next section. Cutting and stacking will also reveal an observation on the position of gaps of maximal length which we will utilize in generalizing the result of Graham and van Lint to other discrete-time dynamical systems of the circle.

1.2.2. CUTTING AND STACKING. The method of cutting and stacking the unit interval is well-known in ergodic theory as a way to construct and visualize transformations [8]. We first illustrate cutting and stacking with a simple example: the binary odometer. The standard description for the binary odometer is symbolic – the phase space is $\Sigma_2 \doteq \{0,1\}^{\mathbb{N}} = \{(s_n)_{n\in\mathbb{N}} = (s_1, s_2, \ldots) | s_n \in \{0,1\}\}$, and the transformation is addition by the sequence $(1,0,0,0,\ldots)$ with roll-over. The cutting and stacking description, which is not unique, merely acts as a tool for visualization. Here we follow the method presented by Rudolph

and Dykstra [9]. Start with the unit interval $[0,1]$. Now, cut the interval in half to form $I_0 = \left[0, \frac{1}{2}\right]$ and $I_1 = \left[\frac{1}{2}, 1\right]$. Stack the right-hand interval on top of the left, and define the transformation at this stage to be the upward translation from $\left[0, \frac{1}{2}\right]$ to $\left[\frac{1}{2}, 1\right]$. This collection of intervals stacked one atop the other is called a **stack** or a **tower**. For the second stage of construction, cut the stage-one tower (both I_0 and I_1) in half, forming two stacks, and then stack the right-hand side upon the left. The dynamics at this stage is again upward translation, where

$$\left[0, \frac{1}{4}\right] \to \left[\frac{1}{2}, \frac{3}{4}\right] \to \left[\frac{1}{4}, \frac{1}{2}\right] \to \left[\frac{3}{4}, 1\right],$$

as illustrated in Figure 1.2 (a). The process of cutting in half and stacking the right-hand side upon the left, and defining the transformation to be upward translation on all but the top-most interval continues indefinitely. The limit of this process defines the binary odometer.

For irrational rotations the tower consists of two stacks at each stage. Begin with the unit interval $[0,1]$ with the usual identification, $0 \equiv 1$. Rotating by irrational $\theta \in (0,1)$, there is a maximal number of times that length-θ intervals can fit into $[0,1)$, namely $\lfloor 1/\theta \rfloor = b_1$, the first term in the continued fraction expansion of θ. Cut the unit interval into b_1 intervals of length $\theta_0 \doteq \theta$,

$$[0, \theta_0), [\theta_0, 2\theta_0), \ldots, [(b_1 - 1)\theta_0, b_1\theta_0),$$

and one remaining interval of length $\theta_1 \doteq 1 - b_1\theta_0$, namely $[1 - \theta_1, 1)$. We then stack so that upward translation (as described in the case of the binary odometer) respects the dynamics

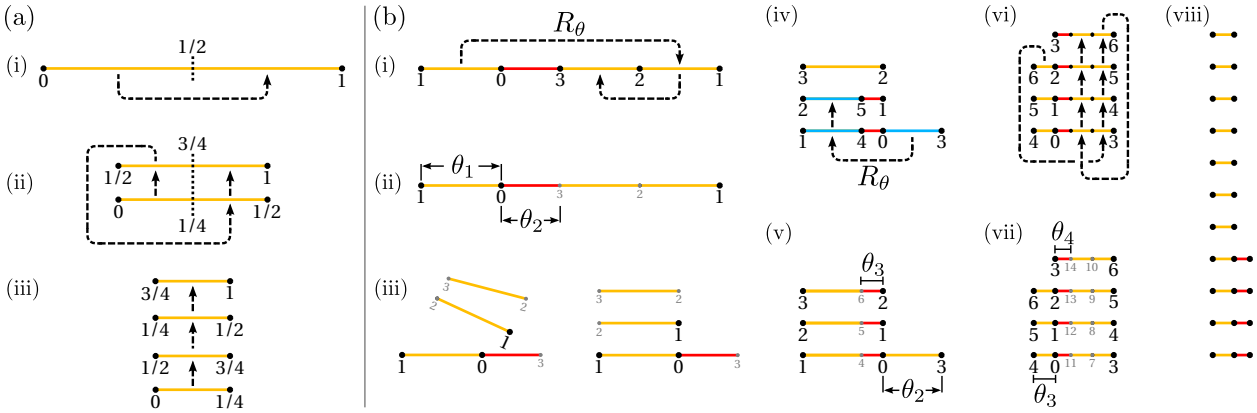

FIGURE 1.2. (a, i-iii) Stages of the tower construction for the binary odometer. Dotted lines indicate the action of the transformation on intervals. (b, i) Stage-one tower for an irrational rotation of $x = 0$ by $\theta = \sqrt{3} \mod 1$ with iterates 1-3 labelled. (ii) Stage-one tower with θ_1 and θ_2 identified and color coded. (iii) Visualization of the process of cutting and stacking the interval to form the stage-two tower. (iv) Stage-two tower with iterates 1-5 labelled. The maximum length interval is colored yellow, the medium length interval is blue, and the smallest gap between iterates is colored red. (v) Stage-two tower with θ_2 and θ_3 identified and color coded. (vi) Stage-three tower with iterates 1-14 identified. (vii) Stage-three tower with θ_3 and θ_4 identified. (viii) Stage-four tower with short/narrow stack and tall/wide stack distinguished by color. Dotted lines indicate the action of R_θ on an interval.

of the rotation, forming two stacks.

$$[0, \theta_0) \to [\theta_0, 2\theta_0) \to \ldots \to [(b_1 - 1)\theta_0, b_1\theta_0)$$

forms the taller, wider stack of height b_1 and width θ_0, while

$$[1 - \theta_1, 1)$$

forms the shorter, narrower stack of height 1 and width θ_1. If we define $q_{-1} = 0$ and $q_0 = 1$ then the height of the short stack is q_0 and the height of the tall stack is $b_1 = b_1 q_0 + q_{-1} = q_1$ according to the recurrence relation given in Equation 1. As part of the visualization for this first stage, we place the shorter, narrower stack to the left of the taller, wider stack so that 0 and 1 are essentially joined, as seen in Figure 1.2 (b, i-ii).

To move to the second-stage tower, we first imagine following the path of the interval $[1-\theta_1, 1)$ through the stage-one tower as we rotate by θ. Under rotation, $[1-\theta_1, 1) \to [\theta_0 - \theta_1, \theta_0)$, which is the right-most side of the wide/tall tower. The images of $[1-\theta_1, 1)$ under rotation continue up the rightmost side of the tall stack, then appear again in the base at $[\theta_0 - 2\theta_1, \theta_0 - \theta_1)$, which continues up the stack and appears again in the base at $[\theta_0 - 3\theta_1, \theta_0 - 2\theta_1)$, etc, until the remaining gap length in the base of the current tall/wide stack is smaller than θ_1. We define θ_2 to be this remainder gap length, $\theta_2 \doteq \theta_0 - b_1\theta_1$. The second term in the continued fraction expansion of θ, which we denoted b_2, gives the number of times that θ_1 divides θ_0, so in all we shall cut the tall/wide tower into b_2 columns of width θ_1, starting from the right and moving towards the left (see Figure 1.2 (b,i)). We then stack these columns, starting from the right-most and working to the left, on top of the old short/narrow stack from the previous stage in compliance with the dynamics of the system (see Figure 1.2 (b,iii)). The result is a new tall/wide stack of width θ_1 and height $q_2 = b_2q_1 + q_0$, and a new short/narrow stack of width $\theta_2 = \theta_0 - b_2\theta_1$ and height q_1. Together these stacks form the stage-two tower. Note that now, the tall/wide stack is on the left, the short/narrow stack is on the right, and that the point $0 \equiv 1$ appears in the base where the towers join. The transformation is again upward translation (which corresponds to rotation by θ).

For the third stage, we first imagine following the path of the interval $[0, \theta_2)$ throughout the tower as we rotate by θ. The interval first moves up through the short/narrow stack and then appears at the left side of the base of the tall/wide stack (Figure 1.2 (b,iv)). At this stage, we cut each of the intervals of length θ_1 of the tall/wide stack from left to right into columns of width θ_2. We then stack these columns, starting from the left and moving to the right, on top of the short/narrow stack. The wide/tall stack from the third stage has height

$q_3 = b_3q_2 + q_1$ and width θ_2, while the short/narrow stack from the third stage has the height of the stage-two tall/wide tower q_2 and width $\theta_3 \doteq \theta_1 - b_3\theta_2$, as seen in Figure 1.2 (b,vii).

We continue this process indefinitely. For odd-stage towers, the tall/wide stack appears on the right and the short/narrow stack on the left. For even-stage towers, the short/narrow stack appears on the right and the tall/wide stack on the left. To move from an odd stage to an even stage, one cuts the tall/wide stack from right to left into columns the width of the narrow stack. To move from an even stage to an odd stage, one cuts the tall/wide stack from left to right into columns with width equal to that of the narrow stack. Comparing the above description of the cutting and stacking to the definition of q_k in the partial convergents given in Equation 1 it becomes clear that, for the stage-k tower, $q_k = b_kq_{k-1} + q_{k-2}$ is the height of the tall/wide stack and q_{k-1} is the height of the short/narrow stack.

The process of cutting and stacking also allows us to observe the genesis of the Three-Distance Theorem. As we cut the tall/wide stack into columns, each time we make a cut in the base, we introduce an interval of a new size. There are exactly two different gap sizes for each of the first b_1 iterates, θ_0 and $1 - j\theta_0$ $(j = 1, 2, \ldots, b_1)$. Looking at moving from stage one to stage two, first we only have two interval lengths: θ_0 and θ_1. After we make the first cut in the base interval $[0, \theta_0)$, we now have three interval lengths: θ_0 (for the intervals above which have not been cut yet), θ_1, and $\theta_0 - \theta_1$. The top of the tower contains the longest interval, and we finish the first cut after we move all the way up through the tower and cut the top-most level, and then, again, have only two gap lengths between consecutive iterates: θ_1 and $\theta_0 - \theta_1$. However, the next iterate (which again cuts the base) creates a gap between consecutive points of a third length: $\theta_0 - 2\theta_1$ to go along with θ_1 and $\theta_0 - \theta_1$ (for the intervals above the base which have not yet been cut). The process continues, and so we have at least two and at most three interval lengths as we shift 0 by θ and reduce mod 1,

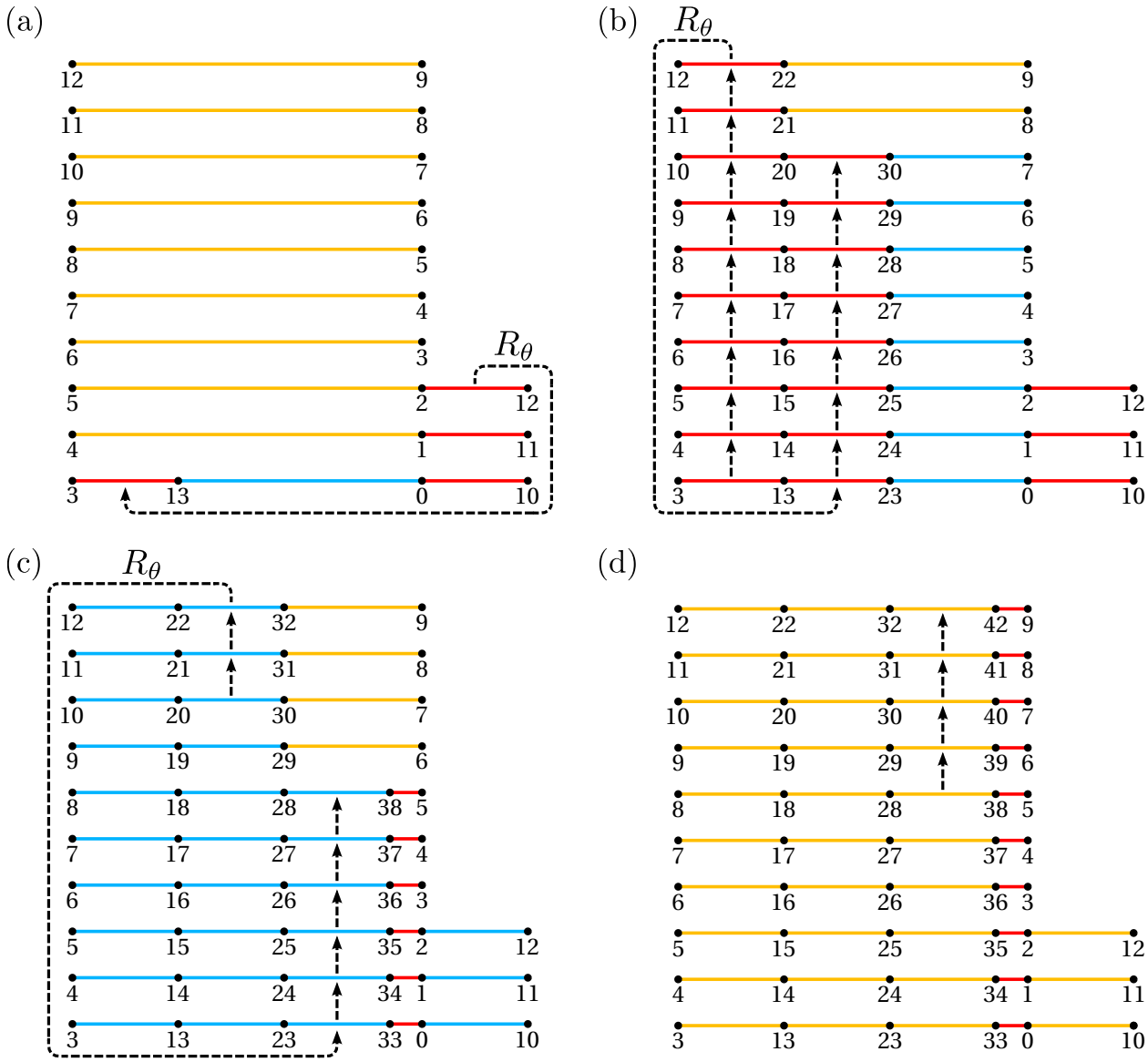

FIGURE 1.3. (a) Stage-two tower of height $q_k = 10$ with the first 13 iterates of the point 0 under R_θ with $\theta = \left(\sqrt{13} - 3\right)/2$. $R_\theta^{13}(0)$ divides one of the 10 longest gaps of length θ_1 into intervals of length θ_2 and $\theta_1 - \theta_2$. (b) Stage-two tower along with iterate numbers 14 through 30. The new, now smaller, longest gap between points in the orbit is $\theta_1 - \theta_2$. (c) Stage-two tower along with iterate numbers 31 through 38. The new longest gap is $\theta_1 - 2\theta_2$. (d) Stage-two tower along with iterate numbers 39 through 42. There are exactly 2 different gap sizes, the longest of which is now θ_2 since $\theta_1 - 3\theta_2 < \theta_2$. Arrows indicate the action of R_θ on intervals defined by consecutive iterates, as ordered by their position in the unit interval.

or equivalently rotate the circle by θ. Figure 1.3 demonstrates this observation for rotation by the angle $\left(\sqrt{13}-3\right)/2 = [3,3,\ldots]$, whose continued fraction expansion is period-one, consisting entirely of threes. Here we illustrate the $k=2$ stage tower of height $q_k = 10$, and each of the $b_k q_k = 30$ iterates between $q_k + q_{k-1} = 13$ and 42, at which point a new tower of height $b_k q_k + q_{k-1} = 33$ would be created. Within each tower, the smallest gap (of length $d1_k$) is colored red, the medium gap (of length $d2_k$) is colored blue, and the longest gap (of length $d3_k$) is colored yellow. We can see that the narrower intervals of the short tower will cut the longer intervals of the tall tower exactly $b_k = 3$ times before the introduction of a new smallest gap. Thus one can infer those numbers of iterates, not explicitly shown in this figure (namely 22 and 32), which will result in precisely two different gap lengths.

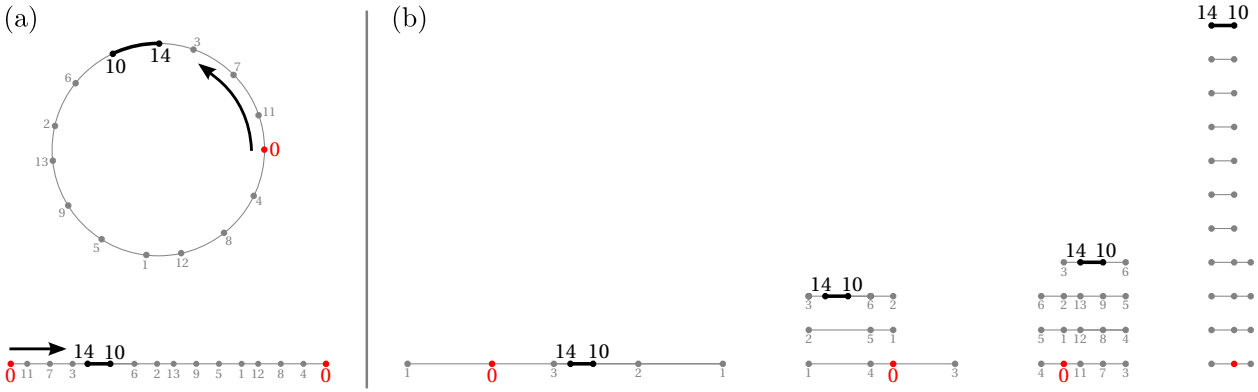

FIGURE 1.4. (a) S^1 and the unit interval with the first 14 iterates of $R_\theta(0)$, where $\theta = \sqrt{3} \mod 1 = \left[0,\overline{1,2}\right]$. Arrows indicate orientation. (b) The first four towers of $R_\theta(0)$. The interval $I \doteq [R^{14}(\theta), R^{10}(\theta)]$, always appears at the top of the towers. A subinterval of I will be the top of the next tower.

Another observation we can make from this construction, and which will prove important in generalizing the result of Graham and van Lint in Section 1.2.4, is that the top level of the tall tower always contains the widest interval because it is the last interval to be divided into narrower intervals as R_θ cuts the columns. Also, the topmost levels of both the tall/wide and short/narrow stage-$(k+1)$ towers are always subintervals of the top-most level of the

stage-k tall/wide tower. We illustrate this fact in Figure 1.4. The subinterval which becomes the top of the stage-$(k+1)$ tall/wide tower is the interval at the top of the stage-k tall tower whose endpoints are iterate numbers $(b_k+1)q_k+q_{k-1}-1$ and $b_kq_k+q_{k-1}-1$, formed by the cut and one endpoint of the top of the tall tower if $b_k=1$, or by the second-to-last and last cuts of the tall tower if $b_k>1$. The tops of the tall towers form a nested sequence of intervals which always contain the largest gap between points in the truncated orbit of 0 as we make the cuts during the transitions from the kth tower to the $(k+1)$th tower.

Figure 1.5 refers back to Figure 1.1 by comparing towers for the rigid rotation of the circle by $\phi=\left(\sqrt{5}-1\right)/2$ and $\theta=4-\pi$. Using this construction, it becomes strikingly clear why ϕ is more consistently uniformly distributing points than θ; after only 7 iterations θ had very nearly uniformly distributed points in the circle, leaving a gap between its 7th iterate and its initial condition of only $\theta_1=7\theta-6\approx 0.008851$. As a result, each of the longest gaps (indicated in yellow) require 15 iterations before the introduction of a new smallest gap size, corresponding to $b_3=15$ in the continued fraction expansion of $4-\pi$. If one were to build the next tower for θ they would observe that the 106 largest gaps of length $\theta_1=7\theta-6\approx 0.008851$ are nearly the exact length of the medium gaps $\theta_2=91-106\theta\approx 0.008821$, which sets up the successive tower to require 292 cuts of each of the 113 intervals of length θ_2.

1.2.3. Computation of Linear Limit Density for Rigid Rotations. Corollary 1 follows from the Three-Distance Theorem [4] and provides a means to compute the maximum gap size present in $\gamma_x(m)$. This allows us to compute the density failures along subsequences of truncated orbit lengths where $m\mathcal{E}(\gamma_x(m))$ is maximized or minimized. This known result allows us to better understand how $(m+1)\mathcal{E}(\gamma_x(m))$ behaves in finite time and, ultimately, to compute the linear limit densities of rotations by quadratic surds with period-1 continued fraction expansions.

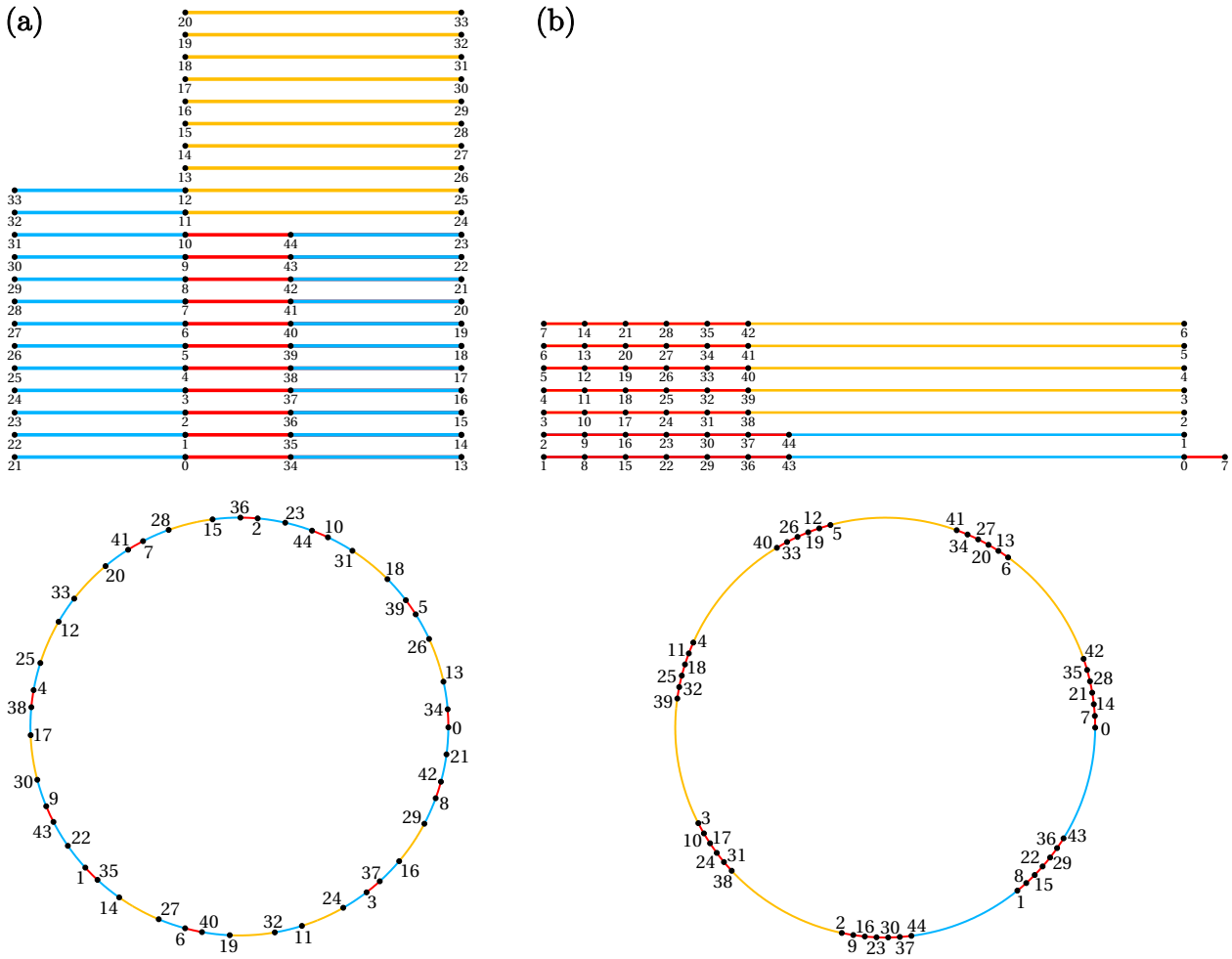

FIGURE 1.5. Comparison of the tower construction to the distribution of points on the circle via a rigid rotation. (a) The 8th tower for the map $R_\phi(x)$ of the unit interval starting at $x = 0$ along with the additional iterates numbers 34 through 44. (b) The 3rd tower for the map $R_\theta(x)$ of the unit interval starting at $x = 0$ along with the additional iterates 8 through 44.

COROLLARY 1. *$\mathcal{E}(\gamma_x(m)) = d_\theta(m)/2 = |q_k\theta - p_k + \alpha(q_{k+1}\theta - p_{k+1})|/2$, where $0 \le \alpha \le b_{k+2} - 1$ and q_n are chosen so that*

$$q_k + (\alpha+1)q_{k+1} \le m+1 \le q_k + (\alpha+2)q_{k+1} - 1.$$

Table 1.1: The columns labeled '$m+1$ range' specifies those number of points (counting the initial point 0) distributed on the circle for which the associated $d_\theta(m)$ is the largest gap size. The smallest number of points for which $d_\theta(m)$ is the largest gap size is given by $q_n + (\alpha+1)q_{n+1}$, while the largest number of points for which $d_\theta(m)$ is the largest gap size is $q_n + (\alpha+2)q_{n+1} - 1$, as stated in Corollary 1.

			$\alpha = 0$		$\alpha = 1$		$\alpha = 2$	
n	p_n	q_n	$d_\theta(m)$	$m+1$ range	$d_\theta(m)$	$m+1$ range	$d_\theta(m)$	$m+1$ range
-1	1	0	1	$1-1$	0.6972	$2-2$	0.3944	$3-3$
0	0	1	0.3028	$4-6$	0.2111	$7-9$	0.1194	$10-12$
1	1	3	0.0917	$13-22$	0.0639	$23-32$	0.0362	$33-42$
2	3	10	0.0278	$43-75$	0.0194	$76-108$	0.0109	$109-141$

Corollary 1 not only provides us with an explicit formula for the size of the largest gaps, but also specifies the range of iterates over which the largest gaps persist. In effect, both the range and size of a largest gap are computable from the denominators of the partial convergents to θ. We call the sequence of truncated orbit lengths when new largest gaps are introduced the **full sequence**.

Let us consider a particular example to emphasize the significance of the integer ranges over which a particular largest gap size persists. Let $\theta = \left(\sqrt{13}-3\right)/2 = [3,3,3,\ldots] \approx .3028$ (the irrational number with a simple continued fraction expansion consisting of all threes). For this choice of rotation angle, we construct Table 1.1 and a plot of $m\mathcal{E}(\gamma_x(m-1))$ for truncated orbits of size $m = 1,\ldots,1189$, shown in Figure 1.6. Since a largest gap size remains constant over a range of iterations of R_θ, the quantity

$$(m+1)\mathcal{E}(\gamma_x(m))$$

grows linearly with m between terms of the full sequence. Along the full sequence one observes sudden drops in the maximum gap size, and thus also in the linear-time rescaled density failure.

Although the range and size of a largest gap is computable, it remains recursive since it depends on the denominators of the partial convergents. By recasting the defining relations of p_k and q_k (the numerator and denominator of the kth partial convergent to θ) into matrix equations, we can determine closed-form expressions for q_k and p_k.

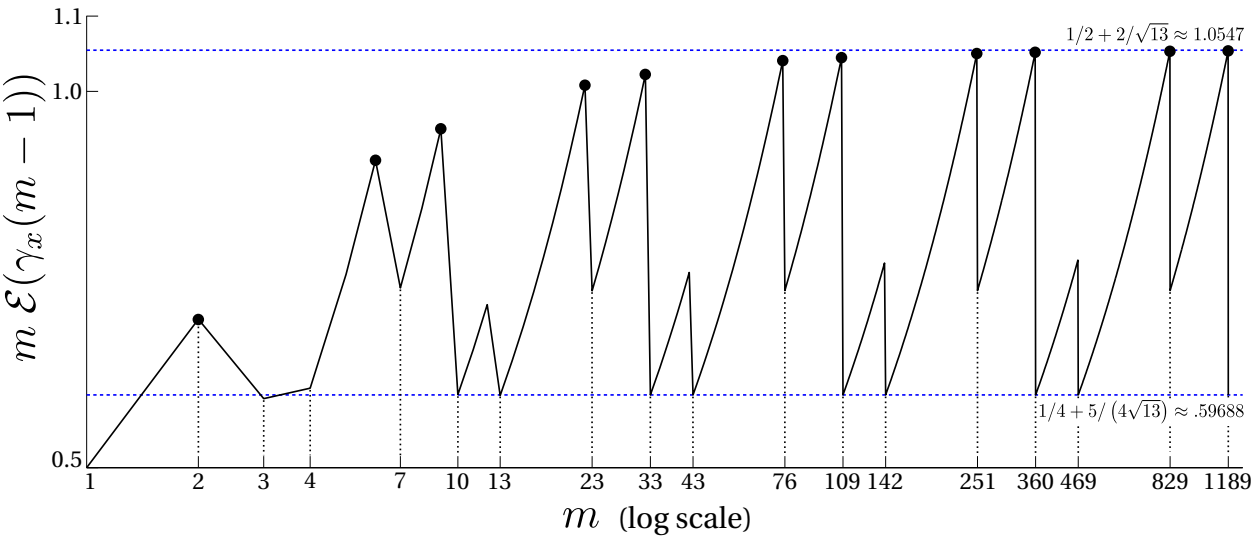

FIGURE 1.6. A plot of $m\mathcal{E}(\gamma_x(m-1))$ versus m, where $\gamma_x(m-1) = \{0, R_\theta(0), \ldots, R_\theta^{m-1}(0)\}$ for $\theta = \left(\sqrt{13}-3\right)/2$. Largest gaps are introduced along the full sequence, indicated by vertical dotted lines. This figure also demonstrates the linear growth of $m\mathcal{E}(\gamma_x(m-1))$ between terms of the full sequence. Horizontal dashed lines indicate $\liminf_{m\to\infty} m\mathcal{E}(\gamma_x(m-1))$ and $\limsup_{m\to\infty} m\mathcal{E}(\gamma_x(m-1))$, as given in Table 1.2.

If $\theta = [b_1, b_2, \ldots]$, then p_k and q_k satisfy the following matrix equations:

$$\begin{pmatrix} p_k \\ p_{k-1} \end{pmatrix} = \begin{pmatrix} b_k & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_{k-1} \\ p_{k-2} \end{pmatrix}, \quad \begin{pmatrix} p_0 \\ p_{-1} \end{pmatrix} = \begin{pmatrix} b_0 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} q_k \\ q_{k-1} \end{pmatrix} = \begin{pmatrix} b_k & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} q_{k-1} \\ q_{k-2} \end{pmatrix}, \quad \begin{pmatrix} q_0 \\ q_{-1} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The sequence of integer values at which new maximum gap sizes are first introduced by iterating R_θ is, in fact, a superset of the sequence of denominators $(q_k)_{k\in\mathbb{N}}$ (which we refer to as the **outer sequence**) derived from the CFE of θ. To understand the full sequence, we first decompose the matrices of the form

$$\begin{pmatrix} b_k & 1 \\ 1 & 0 \end{pmatrix}$$

into a product of simpler matrices.

DEFINITION 1.2.6. Define the matrix $\mathcal{F} \doteq \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ and the matrix $\mathcal{G} \doteq \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$.

PROPOSITION 3. *For any $j \in \mathbb{N}$, $\begin{pmatrix} j & 1 \\ 1 & 0 \end{pmatrix} = \mathcal{G}^{j-1}\mathcal{F}$, where $\mathcal{G}^0 = \mathcal{I}$, the 2×2 identity matrix.*

PROOF. The base case, $j = 1$, is clearly true. Assume the proposition holds for some $j \in \mathbb{N}$. Then,

$$\begin{pmatrix} j+1 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} j & 1 \\ 1 & 0 \end{pmatrix}$$

$$= \mathcal{G}\mathcal{G}^{j-1}\mathcal{F}$$

$$= \mathcal{G}^{(j+1)-1}\mathcal{F}$$

□

Returning our attention to the example $\theta = \left(\sqrt{13} - 3\right)/2$, recall that the matrix relation which defines q_k is

$$\begin{pmatrix} q_k \\ q_{k-1} \end{pmatrix} = (\mathcal{G}^2\mathcal{F})^k \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$= \mathcal{G}^2\mathcal{F} \ldots \mathcal{G}^2\mathcal{F} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

$$= \mathcal{G}\mathcal{G}\mathcal{F} \ldots \mathcal{G}\mathcal{G}\mathcal{F} \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

If we let s_i be the first component of the vector which results from multiplication (on the left) of the column vector $(1, 0)^\top$ by the first i matrices given in the above decomposition of $(\mathcal{G}^2\mathcal{F})^k$, we observe that the first few terms of s_i agree with the full sequence. Consider again the truncated orbit, $\gamma_x(m)$ consisting of $m+1$ points in the forward orbit of $x \in [0, 1)$ under the map R_θ. By the Three-Distance Theorem, a new largest gap size will be introduced precisely when $m + 1 = q_k + (\alpha + 1)q_{k+1}$. Our observation is that each $s_i = q_k + (\alpha + 1)q_{k+1}$ for some q_k and some α. We formalize this observation in the following corollary to the Three-Distance Theorem.

Corollary 2. *Let $\theta = [b_1, b_2, \ldots]$. Consider the truncated orbit, $\gamma_x(m)$, consisting of $m + 1$ points in the forward orbit of $x \in [0, 1)$ under the map R_θ. Then a new largest gap size will be introduced at each s_i, where*

$$\begin{pmatrix} s_i \\ t \end{pmatrix} = \overbrace{\underbrace{\mathcal{G} \ldots \mathcal{G}}_{\alpha \text{ terms}} \mathcal{F}\mathcal{G}^{b_j - 1}\mathcal{F} \ldots \mathcal{G}^{b_2 - 1}\mathcal{F}\mathcal{G}^{b_1 - 1}\mathcal{F}}^{i \text{ terms}} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

PROOF. Fix an integer $i \geq 1$ and let $k+2$ be the number of times the matrix $\mathcal{F}$ appears in the defining matrix equation for s_i. Furthermore, let α be the number of copies of $\mathcal{G}$ which appear after all copies of $\mathcal{F}$ (reading right to left). Then we can express $i = (k+2) + (b_1 - 1) + \ldots + (b_j - 1) + \alpha$ and $s_i = q_k + (\alpha + 1)q_{k+1}$. □

Using this result one can compute the LLD of any quadratic surd by deriving closed-form formulas for each s_i. We proceed with this derivation for the family of irrational numbers with period-1 continued fraction expansions.

Let $\theta_j = [j, j, j, \ldots]$ be the irrational number whose partial quotients are all equal to j. First observe that the denominators of the partial convergents, q_k, satisfy

$$\begin{pmatrix} q_k \\ q_{k-1} \end{pmatrix} = \begin{pmatrix} j & 1 \\ 1 & 0 \end{pmatrix}^k \begin{pmatrix} q_0 \\ q_{-1} \end{pmatrix}$$

$$= (\mathcal{G}^{j-1}\mathcal{F})^k \begin{pmatrix} q_0 \\ q_{-1} \end{pmatrix}$$

for each $k \in \mathbb{N}$.

Diagonalization of the matrices $\mathcal{G}^{j-1}\mathcal{F}$ yields:

PROPOSITION 4. $\begin{pmatrix} j & 1 \\ 1 & 0 \end{pmatrix} = \Sigma\Lambda\Sigma^{-1}$, *where*

$$\Sigma = \begin{pmatrix} \frac{j-\sqrt{j^2+4}}{2} & \frac{j+\sqrt{j^2+4}}{2} \\ 1 & 1 \end{pmatrix} \text{ and } \Lambda = \begin{pmatrix} \frac{j-\sqrt{j^2+4}}{2} & 0 \\ 0 & \frac{j+\sqrt{j^2+4}}{2} \end{pmatrix}.$$

Thus $(\mathcal{G}^{j-1}\mathcal{F})^k = \Sigma\Lambda^k\Sigma^{-1}$. Using this expression, we can derive a closed-form formula for q_k. However, since we would like to compute the density failure along the full sequence it would be useful to derive a formula for s_i. Notice that in this case $s_{tj} = q_t$ for each $t \geq 0$. Given the periodic structure of the sequence of matrices ($\mathcal{F}$ and $\mathcal{G}$) in the decomposition of the defining equation for q_k, it is clear that if we can derive a closed-form formula for $\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1}$ for $p = 1, 2, \ldots j$, we will have an equation for each s_i.

Notice that $\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1} = \mathcal{G}^{-1}(\mathcal{G}^{j-p+1}\mathcal{F}\mathcal{G}^{p-2})\mathcal{G}$. Therefore $\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1}$ is $\mathcal{G}^{j-(p-1)}\mathcal{F}\mathcal{G}^{(p-1)-1}$ conjugated by $\mathcal{G}^{-1}$.

PROPOSITION 5. *If $\vec{v}$ is an eigenvector of $\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1}$, then $\mathcal{G}^{-1}\vec{v}$ is an eigenvector of $\mathcal{G}^{j-(p+1)}\mathcal{F}\mathcal{G}^{(p+1)-1}$.*

PROOF.

$$\begin{aligned}
\mathcal{G}^{j-(p+1)}\mathcal{F}\mathcal{G}^{(p+1)-1}\left(\mathcal{G}^{-1}\vec{v}\right) &= \mathcal{G}^{j-(p+1)}\mathcal{F}\mathcal{G}^{(p+1)-2}\vec{v} \\
&= \mathcal{G}^{-1}\left(\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1}\vec{v}\right) \\
&= \mathcal{G}^{-1}\left(\lambda\vec{v}\right) \\
&= \lambda\left(\mathcal{G}^{-1}\vec{v}\right).
\end{aligned}$$

□

If we let $\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1} = \Sigma_p\Lambda_p\Sigma_p^{-1}$ define Λ_p and Σ_p for each $p = 1, 2, \ldots, j$, then Proposition 5 shows that $\Sigma_{p+1} = \mathcal{G}^{-1}\Sigma_p$ for $p = 1, 2, \ldots, j-1$. Since conjugation preserves eigenvalues, $\Lambda_{p+1} = \Lambda$ for $p = 1, 2, \ldots, j-1$. With this in hand, we can derive closed-form expressions for $\Sigma_p(\Lambda_p)^n\Sigma_p^{-1}, p = 1, 2, \ldots, j$ which will give j Binet's formulas - that is, closed expressions for each s_i along the full sequence.

Corollary 3.

$$\Sigma_p = \begin{pmatrix} \frac{j-\sqrt{j^2+4}-2(p-1)}{2} & \frac{j+\sqrt{j^2+4}-2(p-1)}{2} \\ 1 & 1 \end{pmatrix} \text{ and } \Sigma_p^{-1} = \begin{pmatrix} \frac{-1}{\sqrt{j^2+4}} & \frac{j+\sqrt{j^2+4}-2(p-1)}{2\sqrt{k^2+4}} \\ \frac{1}{\sqrt{j^2+4}} & \frac{j-\sqrt{j^2+4}+2(p-1)}{2\sqrt{k^2+4}} \end{pmatrix}$$

The proof follows from induction on p and Proposition 5.

We now compute

$$\left(\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1}\right)^n = \Sigma_p \Lambda^n \Sigma_p^{-1} = \begin{pmatrix} \sigma_{p-} & \sigma_{p+} \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_-^n & 0 \\ 0 & \lambda_+^n \end{pmatrix} \begin{pmatrix} \sigma_{p-} & \sigma_{p+} \\ 1 & 1 \end{pmatrix}^{-1}$$

for each $n \in \mathbb{N}$ and each $p = 1, 2, \ldots j$.

By reindexing to keep track of p (which subsequence of the outer sequence we are computing) we define

$$\begin{pmatrix} s_{p,n+1} \\ s_{p,n} \end{pmatrix} \doteq \left(\mathcal{G}^{j-p}\mathcal{F}\mathcal{G}^{p-1}\right)^n \begin{pmatrix} s_{p,1} \\ s_{p,0} \end{pmatrix}$$

and finally compute

$$s_{p,n+1} = \frac{s_{p,1}}{\sqrt{j^2+4}}\left(\sigma_{p+}\lambda_+^n - \sigma_{p-}\lambda_-^n\right) + \frac{s_{p,0}}{\sqrt{j^2+4}}\left(\sigma_{p+}\sigma_{p-}\lambda_-^n - \sigma_{p+}\sigma_{p-}\lambda_+^n\right) \tag{2}$$

for $p = 1, 2, \ldots j$ and $n \in \mathbb{N}$.

Using Equation 2 along with the observations that the maximum values of $m\mathcal{E}(\gamma_x(m-1))$ occur at those numbers of iterations equal to $q_k + (\alpha+2)q_{k+1} - 1$, for $\alpha = \lfloor j/2 \rfloor - 1$, and the minimum values $m\mathcal{E}(\gamma_x(m-1))$ occur along $q_k + (\alpha+1)q_{k+1}$, for $\alpha = 0$, we construct Table 1.2. This table recalls the results of Graham and van Lint and goes further by explicitly

Table 1.2: The linear limit densities for a few choices of irrational θ with period-1 continued fraction expansions. The quantity $\liminf_{m\to\infty} m\mathcal{E}(\gamma_x(m-1))$ is computed by finding the limit along the subsequence of the full sequence given by $q_k + (\alpha+1)q_{k+1}$, for $\alpha = 0$. The quantity $\limsup_{m\to\infty} m\mathcal{E}(\gamma_x(m-1))$ is computed by finding the limit along the subsequence of the full sequence given by $q_k + (\alpha+2)q_{k+1} - 1$, for $\alpha = \lfloor j/2 \rfloor - 1$.

j	$\theta = \frac{\sqrt{j^2+4}-j}{2}$	$\liminf_{m\to\infty} [m\mathcal{E}(\gamma_x(m-1))]$	$\limsup_{m\to\infty} [m\mathcal{E}(\gamma_x(m-1))]$
1	$\frac{\sqrt{5}-1}{2}$	$\frac{1}{4} + \frac{3\sqrt{5}}{20} \approx 0.585410$	$\frac{1}{2} + \frac{1}{\sqrt{5}} \approx 0.947214$
2	$\sqrt{2}-1$	$\frac{1}{4} + \frac{\sqrt{2}}{4} \approx 0.603553$	$\frac{1}{2} + \frac{3\sqrt{2}}{8} \approx 1.030330$
3	$\frac{\sqrt{13}-3}{2}$	$\frac{1}{4} + \frac{5}{4\sqrt{13}} \approx 0.596688$	$\frac{1}{2} + \frac{2}{\sqrt{13}} \approx 1.054700$
4	$\sqrt{5}-2$	$\frac{1}{4} + \frac{3\sqrt{5}}{20} \approx 0.585410$	$\frac{1}{2} + \frac{3\sqrt{5}}{10} \approx 1.170820$
5	$\frac{\sqrt{29}-5}{2}$	$\frac{1}{4} + \frac{7}{4\sqrt{29}} \approx 0.574967$	$\frac{1}{2} + \frac{4}{\sqrt{29}} \approx 1.242781$
100	$\sqrt{2501}-50$	$\frac{1}{4} + \frac{51}{4\sqrt{2501}} \approx 0.504949$	$\frac{1}{2} + \frac{\sqrt{2501}}{4} \approx 13.002499$

computing the linear limit densities for several choices of rotation angle. For example, if we consider the subsequence of orbit lengths equal to $q_k + q_{k-1} - 1$ (for all k) under rotation by the golden number, then the associated density failure is $(q_{k-1}\phi - p_{k-1})/2$, and thus

$$\begin{aligned}
&\limsup_{m\to\infty} [(m+1)\mathcal{E}(\gamma_x(m+1))] \\
&= \lim_{k\to\infty} (q_k + q_{k-1} - 1)\frac{q_{k-1}\phi - p_{k-1}}{2} \\
&= \lim_{k\to\infty} \frac{(\sqrt{5}-1)^k}{5\cdot 2^{2k+1}} \left(\left(5+2\sqrt{5}\right)\left(1+\sqrt{5}\right)^k + \left(5-2\sqrt{5}\right)\left(1-\sqrt{5}\right)^k - 5\cdot 2^k \right) \\
&= \frac{1}{2} + \frac{1}{\sqrt{5}} \approx 0.947214.
\end{aligned}$$

We note that the size of $\liminf_{m\to\infty}(m+1)\mathcal{E}(\gamma_x(m))$ approaches $1/2$ as j increases. To understand why, recall that the larger the partial quotients in the CFE of an irrational number θ, the more rapid the convergence of the rational convergents (those truncated CFEs) to θ. For example, the error $|1/j-\theta|$, for $\theta=[j,j,\ldots]$, is strictly decreasing in j. Thus the distribution of $j+1$ points under rotation by R_θ will approach uniformity as j increases. The cost such a rotation pays is that the number of points which are needed to achieve this first nearly uniform distribution is $j+1$. Similarly, the sequence along which these most uniform distributions occur grows quickly for large values of j. It is this trade-off which is, in the sense of the two quantities computed here, best balanced by the golden ratio.

A natural question, then, is: How small can $\liminf_{m\to\infty}\left[(m+1)\mathcal{E}(\gamma_x(m))\right]$ be? Moreover, concerning our quantity of interest – What is the smallest LLD achievable by some homeomorphism of the circle? The answer to the first question has a straightforward solution, while the latter question will take significantly more effort to resolve.

PROPOSITION 6. *Let $T:[0,1)\to[0,1)$ be a map such that γ_0 is dense in $[0,1)$. For all $t\in\mathbb{N}$*

$$t\mathcal{E}(\gamma_0(t-1))\geq 1/2.$$

PROOF. Label the t points in $\gamma_0(t-1)$,

$$0=a_0<a_1<\ldots<a_{t-1}<a_t=1\equiv 0.$$

The density failure of this set is

$$\mathcal{E}(\gamma_0(t-1))=\frac{1}{2}\max_{1\leq i\leq t}\{a_i-a_{i-1}\}.$$

If

$$t\mathcal{E}(\gamma_0(t-1)) = t\frac{1}{2}\max_{1\le i\le t}\{a_i - a_{i-1}\} < \frac{1}{2},$$

then

$$\max_{1\le i\le t}\{a_i - a_{i-1}\} < \frac{1}{t}$$

which implies that

$$\sum_{i=1}^{t} a_i - a_{i-1} < t\,\frac{1}{t} < 1,$$

a contradiction. □

Proposition 6 simply points out that the minimum density failure for a finite subset of the circle (or interval) of fixed size occurs when the points are uniformly distributed. As a consequence of this fact, it is clear that

$$\liminf_{m\to\infty}\left[(m+1)\mathcal{E}(\gamma_x(m))\right] \ge 1/2,$$

not only for irrational rotations, but for any topologically transitive orbit in the circle. So it would appear that for an irrational rotation the answer to our minimization problem is: arbitrarily close to 1/2. Can 1/2 be achieved by some dense orbit on the circle if we do not restrict ourselves to rigid rotations or even continuous maps? The answer is yes.

Let $x_0 = 0$ and $x_1 = 1/2$ in $[0, 1)$ (having identified the points 0 and 1). Now recursively define the rest of the orbit of 0 by letting x_{n+1} be the midpoint of some largest gap between adjacent points in $\gamma_0(n-1)$, the orbit with n points. By always choosing the midpoints we are guaranteed to have uniform distributions for each 2^n points, for all $n \in \mathbb{N}$. The density failure of such a distribution is then $\frac{1}{2^{n+1}}$ and the resulting product is exactly $\frac{1}{2}$. Since this happens infinitely often $\liminf_{m\to\infty}\left[(m+1)\mathcal{E}(\gamma_x(m))\right] = 1/2$ for this dense orbit. However,

the density failure just prior to a new largest gap size being introduced (i.e. $2^n - 1$ points) is 2^n and thus $\limsup_{m\to\infty}[(m+1)\mathcal{E}(\gamma_x(m))] = 1$, which is certainly not minimal. We have already seen a map on the circle whose LLD is less than 1, namely the irrational rotation by the golden ratio, $\phi = \left(\sqrt{5}-1\right)/2$. The question of whether the maximal LLD of an orbit under R_ϕ is minimal among all homeomorphisms of the circle will take more sophisticated tools to answer, and is the focus of the next section.

Before proceeding to consider general homeomorphisms of the circle, we mention a best-possible result concerning any sequence of points in S^1. Recall that the orbit built be recursively choosing the nth iterate maximally far from the previous $n-1$ iterates is not the most dense orbit – as measured by the LLD – one can construct in the circle. So, among all possible sequences of points in the circle, what is the smallest achievable LLD? This question was answered by M. Boshernitzan and J. Chaika in [10].

THEOREM 1.2.7 (**Boshernitzan and Chaika**). *For any sequence of points in the circle,* $\boldsymbol{x} = \{x_k\}_{k\in\mathbb{N}}$,

$$\limsup_{n\to\infty} n\mathcal{E}(\boldsymbol{x}[n]) \geq \frac{1}{2ln(2)} \approx 0.72134,$$

where $\boldsymbol{x}[n] = \{x_1, \ldots, x_n\}$. *This bound is achieved by the sequence* $x_k = log_2(2k-1)$ *mod* 1.

1.2.4. AN EXTENSION OF LLD CALCULATIONS TO CIRCLE DIFFEOMORPHISMS. In the preceding section, we restricted our attention to rigid rotations of the circle. These are, in a sense, the simplest members of a much larger family of orientation-preserving homeomorphisms of the circle. Since one of our goals is to understand how small the LLD of an orbit in the circle can be, we now widen our search to include maps which are not simply rotations. In this section we provide an extension of the results of Graham and van

Lint by showing that the largest LLD of any orbit under the rigid rotation of the circle by the golden number is smaller than the largest LLD for some orbit of any member in a broad family of circle maps.

Efforts to completely characterize the dynamics of orientation-preserving circle homeomorphisms date back to Poincaré who, in the 1880's [11], introduced a topological invariant called the rotation number of a homeomorphism.

DEFINITION 1.2.8. Let $f : S^1 \to S^1$ be an orientation-preserving homeomorphism of the circle and let $F : \mathbb{R} \to \mathbb{R}$ be a lift of f. Define

$$\rho(F, x) = \lim_{n\to\infty} \frac{F^n(x) - x}{n}.$$

Then the **rotation number** of f is

$$\rho(f) = \rho(F, x) \mod 1,$$

for some lift F and point $x \in \mathbb{R}$.

It can be shown that $\rho(F, x)$ is independent of the choice of initial point x, and that $\rho(F_1, x) - \rho(F_2, x) \in \mathbb{Z}$ for any lifts F_1 and F_2 of f and so $\rho(f)$ is a well defined. Furthermore, Poincaré was able to show that any orientation-preserving homeomorphism of the circle is at least semi-conjugate to rotation of the circle by its rotation number. Since the main result of this section will rely on sufficient smoothness of a conjugacy between a circle homeomorphism and a rigid rotation, we will require more than Poincaré combinatorial equivalence. We first recall a result which essentially states that the rationality of a map's rotation number governs many of its important dynamical properties.

PROPOSITION 7. *$\rho(f) \in \mathbb{Q}$ if and only if f has periodic points, in which case all orbits are asymptotic to a periodic orbit. If $\rho(f) \notin \mathbb{Q}$, then there exists a unique minimal set, K, that is either all of S^1 or a Cantor set to which all points asymptotically approach. If $K = S^1$ then f is conjugate to the irrational rotation $R_{\rho(f)}$, otherwise f is semi-conjugate to $R_{\rho(f)}$.*

Efforts were made to characterize the smoothness of conjugating maps over the next 100 years ([12], [13]), culminating with the work of Michael Herman [14] and Jean-Christophe Yoccoz [15]) who provided an an answer to the question: Under what conditions is a diffeomorphism of the circle conjugated by a diffeomorphism to a rigid rotation? For a more complete treatment of the historical development of this topic see [11].

THEOREM 1.2.9 (**Herman and Yoccoz**). *Let $f : S^1 \to S^1$ be a C^r diffeomorphism, with $r \geq 3$ and assume $\rho(f)$ is an irrational number satisfying*

$$\left|\rho(f) - \frac{p}{q}\right| > \frac{K}{q^{2+\beta}}$$

for all rational numbers $\frac{p}{q}$ and some positive constants K and β. If $r > 2\beta + 1$, then there exists a C^1 conjugacy, $h \circ f = R_{\rho(f)} \circ h$, with $h \in C^{r-1-\beta-\epsilon}$ for every $\epsilon > 0$.

This result was strengthened Katznelson and Ornstein [16] who relaxed the condition on f being at least three times continuously differentiable to being in the differentiability class C^r, for any $r > 0$ and relaxed the hypothesis that $r > 2\beta + 1$ to only requiring that $r > \beta + 2$. In either case, the differentiability of the conjugation map is guaranteed for a class of sufficiently differentiable diffeomorphisms with badly approximable rotation numbers (bounded partial quotients). As it happens, only a Lebesgue measure 0 set of irrational numbers do not satisfy this diophantine condition.

We are now ready to present an extension of Graham and van Lint's result to circle diffeomorphisms that are conjugate to rotation maps, which Theorem 1.2.9 has established exist in abundancy.

THEOREM 1.2.10. *Let $f : S^1 \to S^1$ be a C^r-diffeomorphism satisfying the hypotheses of Theorem 1.2.9, such that $f \circ h = h \circ R_{\rho(f)}$ for a C^α-diffeomorphism h, with $\alpha \geq 1$. Then, for $\gamma_x(m) = \{f^n(x) | 0 \leq n \leq m\}$,*

$$\sup_{x \in S^1} \limsup_{m \to \infty} [m\mathcal{E}(\gamma_x(m-1))] \geq \frac{1}{2} + \frac{1}{\sqrt{5}}.$$

The significance of Theorem 1.2.10 is that it says that any circle diffeomorphism which is diffeomorphically conjugate to an irrational rotation will have an orbit that has a larger linear limit density than any orbit of R_ϕ, the rigid rotation by the golden ratio. The reason is that conjugation of a rotation by a non-identity continuously differentiable function must stretch some interval of the circle. By choosing an appropriate initial condition, we can guarantee that the last remaining maximal gap is (eventually) always contained in that interval. Since these gaps are necessarily larger than they would be for the rigid rotation, the density failure epsilon for a truncated orbit of length m under f is always larger than for the truncated orbit of length m under $R_{\rho(f)}$, for sufficiently large m. This, together with Graham and van Lint's result (which essentially says that the minimum LLD for any rigid rotation of the circle is achieved by the golden ratio), will prove the claim.

PROOF. If f has periodic points then, according to Proposition 7, all orbits are asymptotic to a periodic orbit and so $\limsup_{m \to \infty} [m\mathcal{E}(\gamma_x(m-1))]$ is necessarily equal to infinity. If f does not have periodic points, then h is a conjugacy between f and an irrational rotation.

We first note that any C^1-diffeomorphism $h : S^1 \to S^1$ of the circle may be identified with a C^1-diffeomorphism, $\hat{h} : [0,1] \to [0,1]$ of the unit interval by an appropriate composition with an initial rotation. In other words, we choose our 0 point in S^1 so that $\hat{h}(0) = 0$, $\hat{h}(1) = 1$ and $\hat{h}$ is monotonic on $[0,1]$. By continuity of the derivative, we further know $\hat{h}'(x)$ exists for all $x \in [0,1]$ and $\hat{h}'(0) = \hat{h}'(1)$. Therefore we will recast h as a diffeomorphism of the closed unit interval and prove the theorem by constructing an initial condition $x^* \in [0,1]$ such that the longest-surviving maximum gaps in the truncated orbits of x^* (under rotation by $R_{\rho(f)}$) can eventually be stretched infinitely often by h.

In the case that h is the identity function, f is a rigid rotation and the claim reduces to the case proven by Graham and van Lint. So assume that $h(x) \neq x$ for some $x \in (0,1)$. Then, $(h(1)-h(x))/(1-x) > 1$ or $(h(x)-h(0))/(x) > 1$, if $h(x) < x$ or $h(x) > x$ respectively. In either case, the Mean Value Theorem asserts the existence of a point $a \in (0,1)$ such that $h'(a) > 1$. Since h' is continuous, there exists an open interval $I = (u,v) \subset (0,1)$ containing a such that $h'(x) > 1$ for all $x \in I$. We call I an interval of stretching since for any $x, y \in I$, $|h(x) - h(y)| > |x - y|$.

For $x_0 \in [0,1)$, define $x_i \doteq R^i_{\rho(f)}(x_0)$ for $i \geq 1$ and define $y_0 \doteq h(x_0)$. Then $y_i \doteq f^i(y_0) = h \circ R^i_{\rho(f)} \circ h^{-1}(y_0) = h \circ R^i_{\rho(f)}(x_0) = h(x_i)$. Thus, if we can show that there exists an initial condition with a forward orbit under $R_{\rho(f)}$ having the property that, eventually, all of the largest gaps which survive the longest lie within I, then, since I is an interval of stretching by h, the gaps between iterates under f in I will be strictly larger than the gaps between iterates under $R_{\rho(f)}$.

Towards that goal, we return to the description of the rotation map via cutting and stacking, as described in Section 2.2. Let the stage-k tower be the first even tower such that the taller/wider stack (on the left) has width $\theta_{k-1} < (v-u)/2$. We know such a tower

exists since $\rho(f)$ is necessarily irrational. The top-most level of this stack is the interval $[R^{r_k}(0), R^{s_k}(0)]$, where $r_k = q_{k-1} + q_k - 1$ and $s_k = q_k - 1$. Define $\delta \doteq (v-u)/2 - \theta_{k-1} > 0$ and define $x^* = R^{-s_k}(v-\delta)$, so that $R^{s_k}(x^*) = v - \delta$. Since $v - \delta < v$ and $v - \delta = v - ((v-u)/2 - \theta_{k-1}) = (u+v)/2 + \theta_{k-1} > u$, we have $R^{s_k}(x^*) \in I$. Furthermore, $R^{r_k}(x^*) = R^{s_k}(x^*) - \theta_{k-1} = v - \delta - \theta_{k-1} = (u+v)/2 \in I$. Therefore, $[R^{r_k}(x^*), R^{s_k}(x^*)] \subset I$.

We observed in Section 1.2.2 that the interval $[R^{r_k}(x^*), R^{s_k}(x^*)]$ contains every maximal gap size less than or equal to θ_{k-1}. Furthermore, if $d3_n$ is a maximal gap that exists for some truncated orbit, and $d3_n \leq \theta_{k-1}$, then a gap of length $d3_n$ contained in $[R^{r_k}(x^*), R^{s_k}(x^*)]$ will survive longer than any other gap of length $d3_n$, as more iterates of x^* are computed under $R_{\rho(f)}$.

Now define $y^* \doteq h(x^*)$. Letting $x_0 = x^*$ and $y_0 = y^*$, every truncated orbit under $R_{\rho(f)}$, $\gamma_{x^*}(m) = \{x_i | 0 \leq i \leq m\}$, has the property that $h(\gamma_{x^*}(m)) = \{y_i | 0 \leq i \leq m\} = \{f^i(y^*) | 0 \leq i \leq m\}$. By the above discussion, if $m \geq r_k = q_{k-1} + q_k - 1$, we know $\gamma_{x^*}(m)$ contains a maximal gap that is a subset of I, the interval of stretching under h. Therefore, $\mathcal{E}(h(\gamma_{x^*}(m))) > \mathcal{E}(\gamma_{x^*}(m))$. Since this holds for all $m \geq q_{k-1} + q_k - 1$, we have

$$\limsup_{m\to\infty} [m\mathcal{E}(h(\gamma_{x^*}(m-1))] > \limsup_{m\to\infty} [m\mathcal{E}(\gamma_{x^*}(m-1))] \geq \frac{1}{2} + \frac{1}{\sqrt{5}}.$$

□

We note that we did not require the interval of stretching to contain a point where h' is maximal, only that h' be greater than 1 on an interval.

The reader may recognize that a similar argument shows there must exist an interval of shrinking, where $h'(x) < 1$. We can force infinitely many maximal gaps (which persist the longest) into this interval by an appropriate choice of initial condition, so it may appear

that the measure of density failure of all truncated orbits for this initial condition after some number of iterations could be made smaller than for the rigid rotation by the golden number. However, imagine a map with rotation number $\rho(f)$ whose interval of compression, J, is approximately an angle $\rho(f)$ from an interval of stretching, I, such that eventually the maximal gap (in the orbit of x) in the interval of stretching is always the gap which survives nearly as long as the maximal gap in the interval of compression (i.e. if $R^m_{\rho(f)}(x) \in J$ then $R^{m-1}_{\rho(f)}(x) \in I$).

If at $m-2$ iterations of $R_{\rho(f)}$ the maximal gap in the interval of stretching is length ξ and the maximal gap in the interval of compression is length η, then the quantities

$$(m-1)\mathcal{E}(\gamma_x(m-2)) = (m-1)\,\xi$$

and

$$m\mathcal{E}(\gamma_x(m-1)) = m\,\eta$$

(provided that the derivative of h in I is not so large that the maximal gap remains in I even after the $(m-1)$th iterate divides the maximal gap there). Therefore, the amount that this quantity decreases from the $(m-1)$th iterate to the mth iterate is $m\,(\xi-\eta)-\xi$ which could, in principle, be quite small relative to the size of $(m-1)\,\xi$. Computing $\frac{m\,(\xi-\eta)-\xi}{(m-1)\,\xi} = 1-\frac{m\eta}{(m-1)\xi}$ we see this ratio is small when $\eta \approx \xi$ for large values of m. Thus we cannot make ξ large relative to η, which we would like to do to easily show that $(m-1)\,\xi$ can be large enough that $m\,\eta$ stays large. A more complicated analysis of the interactions of the intervals of stretching and compressing appear to be required to understand if *all* orbits of a circle diffeomorphism have LLD larger than the rotation by the golden number.

1.3. Symbolic Dynamics

We now shift our focus from maps of the circle to a class of discrete-time dynamical systems with a very different metrizable phase space: the interesting and theoretically useful Bernoulli shift maps. In Section 1.3.1 we outline a few well-known preliminaries to motivate our consideration of maps on symbol spaces. A first approach to effective computation of the LLD in this context will demand the development of an algorithm for determining the epsilon failure density of a distinguished subset in the set of all words of fixed length over a fixed alphabet, and is the content of Section 1.3.2. Finally, in our search for optimally topologically transitive orbits, we make a connection to de Bruijn sequences in Section 1.3.3.

1.3.1. Preliminaries.

For each $k = 2, 3, \ldots$ define

$$\Sigma_k \doteq \{0, 1, \ldots, k-1\}^{\mathbb{N}} = \{(s_n)_{n\in\mathbb{N}} = (s_1, s_2, \ldots) | s_n \in \{0, 1, \ldots, k-1\}\}.$$

Then a Bernoulli (left) shift map acting on $(s_n)_{n\in\mathbb{N}} = (s_1, s_2, \ldots)$ is

$$\sigma((s_n)_{n\in\mathbb{N}}) = (s_{n+1})_{n\in\mathbb{N}} = (s_2, s_3, \ldots).$$

The sets Σ_k can be endowed with a natural measure, the product measure derived by assigning to each letter in the alphabet a measure p_i such that $\sum_i p_i = 1$. Furthermore, by defining $d : \Sigma_k \times \Sigma_k \to \mathbb{R}$ by

$$d(s, t) = \sum_{i\in\mathbb{N}} \frac{|s_i - t_i|}{k^i},$$

Σ_k is endowed with a metric structure such that the maximum distance between any two points in Σ_k is 1. We will also extend this metric function to allow distances from points $y \in \Sigma_k$ to subsets $D \subset \Sigma_k$ by defining $d(y, D) = \inf_{x\in D} d(y, x)$, which is well defined since

the distance function is bounded below by 0. To avoid notational clutter, we refer to all such maps as d, as the domain will be clear from context. Note, one could also define a consistent metric across all of these sequence spaces by taking

$$d(s,t) = \sum_{i \in \mathbb{N}} \frac{\rho(s_i, t_i)}{2^i},$$

where $\rho(\cdot, \cdot)$ is the discrete metric. This metric de-emphasizes the relative size of the letters, which we need not think of as integers carrying with them their usual ring properties. If one wishes to additionally place an ordering on the sequences, an ordering on the letters is then required.

The simplicity of the Bernoulli shift map masks its rich dynamics, which include sensitivity to initial conditions and the existence of dense orbits. It is this final property which we address here for the set of all binary sequences, Σ_2.

PROPOSITION 8. *$\sigma : \Sigma_k \to \Sigma_k$ is topologically transitive.*

The proof of Proposition 8 follows from constructing a sequence with a topologically transitive orbit by concatenating each of the k^M words of length M, for all $M \in \mathbb{N}$. The resulting element of Σ_k gets arbitrarily close to every k-ary sequence under iterations of the shift map. In contrast to our investigation of the family of irrational rotations of the circle, not every k-ary sequence has a dense orbit, and here we are considering only one map instead of a family of maps. Nonetheless, we are driven to ask similar questions: Among the initial k-ary sequences with dense orbits, which ones fill out Σ_k most efficiently as measured by the LLD? In other words, what is the minimum LLD attainable by a k-ary sequence with a dense orbit under the Bernoulli shift map?

At this point, it is not clear that the density failure for truncated orbits will shrink to 0 at a rate proportional to the growth of the number of iterations required to achieve that density failure. The existence of dense orbits may not guarantee a finite LLD; as was mentioned, the orbit of a point under the rotation by an irrational number with unbounded partial quotients in its continued fraction expansion has an unbounded LLD. Towards understanding and computing the limit density of a binary sequence with dense orbits, we develop an efficient algorithm for numerically approximating the LLD of a specified initial sequence. We then use combinatorial results concerning so-called de Bruijn sequences to give bounds on the LLD of binary sequences under the shift map.

1.3.2. Numerical Considerations and an Algorithm for Computing Epsilon Failure Density. Let $x = (x_1, x_2, x_3 \ldots) \in \Sigma_2$ be an initial binary sequence whose orbit is generated by Bernoulli shifts. Even though the points in this orbit are infinite binary sequences, and thus perfect measures of distance will in general be impossible, we can achieve approximations with arbitrary accuracy by considering sufficiently many terms of our sequence, since the metric more heavily weights differences in earlier terms. Explicitly, if we know the first n terms of each element of the $m+1$ points in $\gamma_x(m)$, then we can approximate the density failure of $\gamma_x(m)$ to within $1/2^n$, which can be made arbitrarily small by taking n sufficiently large.

For example, let $n = 4$ and $x = 00110001 \ldots \in \Sigma_2$, such that we know the first 4 digits of, say, the first 5 points in the orbit of x, namely

$$D \doteq \{0011, 0110, 1100, 1000, 0001\}.$$

Then we can search among all $2^4 = 16$ sequences of length four to find the one which is farthest from the set D. In this case one such sequence is 0100, a distance 5/16 from the distinguished set D. The tail of a sequence which begins 0100 can contribute at most 1/16 to the density failure of D, and thus 5/16 is within 1/16 of the actual density failure. However, if we wish to approximate the density failure to within $1/2^n$ by naively searching through the space of all sequences of length n for the one whose minimum distance to a specified distinguished set is maximal, then for large values of n, the search space is prohibitively large. In particular, $n \geq 53$ is required to guarantee double-precision accuracy in our approximation of the density failure. Thus our first goal is to establish an algorithm which can find, among the set of all binary sequences of length n, a sequence which maximizes the minimum distance from a distinguished subset of sequences of length n in better than exponential time.

For $m + 1 = 1$ the density failure of the set consisting of our initial point is clearly 1 since $y = x + 1 \mod (2)$ (the binary sequence farthest from x) is a distance 1 from x. The problem of determining the density failure of $\gamma_x(1)$ is slightly more subtle, so we proceed with an example.

Let $x = 011010001\ldots$, and $\sigma(x) = 11010001\ldots$. Our goal is to find the binary sequence $y \in \Sigma_2$ with maximal distance from the distinguished set $D \doteq \gamma_x(1) = \{x, \sigma(x)\}$. Certainly y begins either with a 0 or a 1 and thus its distance to D cannot be larger than $\sum_{i=2}^{\infty} 1/2^i = 1/2$; it necessarily shares its first digit in common with either x or $\sigma(x)$. Notice that the first two digits of x and $\sigma(x)$ are 01 and 11 respectively. If our goal is to maximize the distance of y from D then obviously we should not choose y to begin with 01 or 11 (if we can avoid it) since then the maximum distance y could be from D would be $\sum_{i=3}^{\infty} 1/2^i = 1/4$. This leaves two other choices for the first two digits of y, either 00 or 10. If we choose y to begin 00,

then it agrees with x in the first digit and is within a distance $1/2$ of x, while it disagrees with $\sigma(x)$ in the first two digits and is thus at least a distance $3/4$ from $\sigma(x)$. Therefore, we need only to ensure that we maximize the distance from y and x, which is achieved if we choose y to be the unique sequence defined by

$$y_i = \begin{cases} 0 & \text{for } i = 1 \\ x_i + 1 \bmod 2 & \text{for } i > 1. \end{cases}$$

Similarly, if we choose y to begin 10, we are within a distance $1/2$ of $\sigma(x)$ but are at least a distance $3/4$ from x and so in this case the sequence whose distance from D is the density failure of the set D is defined by

$$y_i = \begin{cases} 1 & \text{for } i = 1 \\ \sigma(x)_i + 1 \bmod 2 & \text{for } i > 1. \end{cases}$$

This example suggests an algorithm for finding a point whose distance to a distinguished set is maximal, without searching through the entire space of words. What follows are two propositions which, together, formalize this algorithm. Before presenting these results, let us fix some notation. We have previously defined $\Sigma_k \doteq \{0, 1, \ldots k-1\}^{\mathbb{N}} = \{(s_n)_{n\in\mathbb{N}} | s_n \in \{0, 1, \ldots k-1\}\}$. Define $\Sigma_k[M] = \{0, 1, \ldots, k-1\}^M$, the collection of all k^M words of length M formed from an alphabet of k letters. For any word $s = (s_1, s_2, \ldots, s_M) \in \Sigma_k[M]$ and for any integers n and m with $1 \leq n < m \leq M$, define $s[n, m] \doteq (s_n, s_{n+1}, \ldots, s_m)$ to be the subword consisting of all letters between and including the nth and the mth letters of s. Notice that $\Sigma_k[M]$ also can be given a metric structure for each k, M, by defining a distance

function $d : \Sigma_k[M] \times \Sigma_k[M] \to [0, \infty)$ by

$$d(s,t) = \sum_{i=1}^{M} \frac{|s_i - t_i|}{k^i}.$$

PROPOSITION 9. *Let $D \subseteq \Sigma_k[M]$ (called the distinguished set) and let $1 \leq n < M$ be an integer such that for some $w = (w_1, w_2, \ldots, w_n) \in \Sigma_k[n]$ (called a missing leading word) and for every $x = (x_1, x_2, \ldots, x_M) \in D$ we have $(x_1, \ldots, x_n) \neq (w_1, \ldots, w_n)$. That is, $x[1,n] \neq w$ for any $x \in D$. Let $y, y' \in \Sigma_k[M]$ be any words of the form $y = (x_1, \ldots, x_n, y_{n+1}, \ldots, y_M)$ and $y' = (w_1, \ldots, w_n, y'_{n+1}, \ldots y'_M)$. Then $d(y, D) \leq d(y', D)$.*

PROOF. Let $x \in D$ be arbitrarily chosen and $y \in \Sigma_k[M]$ be any element of the form $y = (x_1, \ldots, x_n, y_{n+1}, \ldots, y_M)$. We will say that the n-leading terms of y are equal to the n-leading terms of x, or, in our notation, $y[1,n] = x[1,n]$. Then $d(y, D) \leq d(y, x) = \sum_{i=n+1}^{M} |y_i - x_i|/k^i$. But $|y_i - x_i| \leq k - 1$ for each i since $x_i, y_i \in \{0, 1, \ldots, k-1\}$. Therefore,

$$d(y, D) \leq d(y, x) = \sum_{i=n+1}^{M} \frac{|y_i - x_i|}{k^i} \leq \sum_{i=n+1}^{M} \frac{k-1}{k^i} = \frac{1}{k^n} - \frac{1}{k^M}.$$

Let $y' = (w_1, \ldots w_n, y'_{n+1}, \ldots, y'_M)$ (where $y'_{n+1}, \ldots, y'_M \in \{0, 1, \ldots, k-1\}$ are arbitrary). Since for each $x \in D$ there exists $i, 1 \leq i \leq n$ such that $w_i \neq x_i$ (by assumption that $x[1,n] \neq w[1,n]$) it follows that $d(y', x) \geq d(y'[1,n], x[1,n]) \geq 1/k^n$ for each $x \in D$. Therefore,

$$d(y', D) \geq \frac{1}{k^n} > \frac{1}{k^n} - \frac{1}{k^M} \geq d(y, D). \tag{3}$$

□

Said another way, every y' which starts with some missing leading word, $w = (w_1, \ldots, w_n)$ is farther from D than any word y whose n-leading terms are equal to the n-leading terms

of some distinguished word, $x \in D$. It is worth noting here that the strict inequality in Equation 3 becomes greater-than-or-equal-to in the limit as M is taken to infinity, and the above proposition is still true.

To explain the significance of Proposition 9, we present Figure 1.7, which provides an illustration of a k-ary tree ($k = 2$) whose leaves are the k^M elements of $\Sigma_k[M]$. The circled leaves indicate elements of some distinguished set D. $w = (1, 1)$, shown here as a depth $n = 2$ node, is a missing leading word. That w is a missing leading word can be seen from the tree since no distinguished element is rooted in word $(1, 1)$. Proposition 9 simply says that every length-M word which is rooted in w is farther from D than any length-M word which is rooted in a depth-n node that has a distinguished element as a leaf. In other words, an element of $\Sigma_2[5]$ whose distance to the distinguished set is maximal, among all elements of $\Sigma_2[5]$, will be a word rooted in (that is, beginning with) $w = (1, 1)$.

This missing leading word w is a depth-n node (it is a length-n word). It follows from this proposition that if there were a length-$(n-1)$ word which was also a missing leading word we should look in the leaves rooted in this length-$(n-1)$ word for a length-M word which is at least as far from the distinguished set as any other. Simply put, we can start by searching the distinguished set for missing leading words of length one and if none are found, we can search for missing leading words of length two, and so on until at least one missing word (of minimal length) is found. Proposition 9 guarantees that a word whose distance from D is maximal will begin with one of the missing leading words of minimal length. This is, of course, an improvement over searching the entire space of k^M words. If P minimal length missing words appear at depth n, we have (at this point) reduced our search space to the space of all Pk^{M-n} tails. The following proposition shows we can, in most cases, do much better.

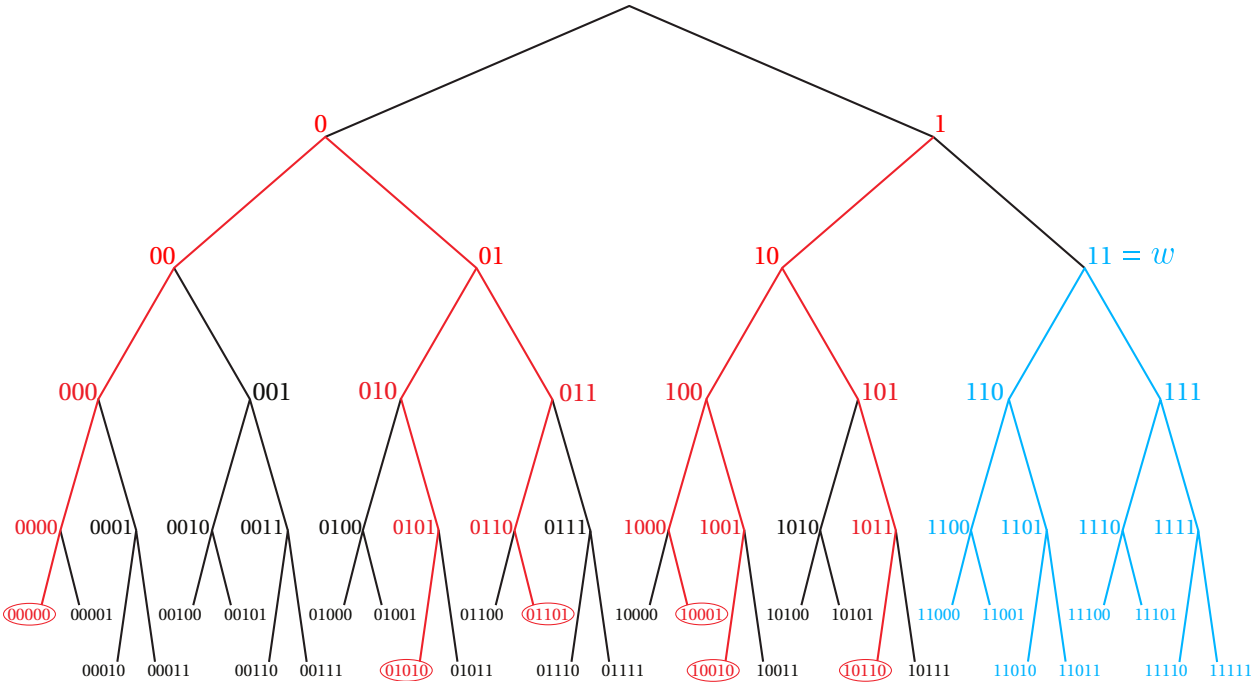

FIGURE 1.7. Binary tree representing leading words of all lengths in $\Sigma_2[5]$. Circled words represent elements of an example distinguished set $D \subset \Sigma_2[5]$, while those leading words (nodes) which are highlighted with red indicate leading words which are present in the distinguished set. The word $w = (1,1) = 11$ (highlighted in blue along with all of its branches), is a missing leading word of minimal length.

PROPOSITION 10. *Let $x, z \in D$ be such that $d(x[1,n], w) < d(z[1,n], w)$ where w is a length-n missing leading word. Then $d(x,y) < d(z,y)$ for all*

$y = (w_1, \ldots, w_n, y_{n+1}, \ldots, y_M)$.

PROOF. Define $\Delta_{\text{lead}} \doteq d(z[1,n], w) - d(x[1,n], w) > 0$. Assume $\Delta_{\text{lead}} < 1/k^n$. This says that

$$0 < \left(\frac{|z_1 - w_1|}{k} + \ldots + \frac{|z_n - w_n|}{k^n} \right) - \left(\frac{|x_1 - w_1|}{k} + \ldots + \frac{|x_n - w_n|}{k^n} \right) < \frac{1}{k^n},$$

which implies that

$$0 < \left(k^{n-1}(|z_1 - w_1|) + \ldots + |z_n - w_n| \right) - \left(k^{n-1}(|x_1 - w_1|) + \ldots + |x_n - w_n| \right) < 1,$$

a contradiction since $\sum_{i=1}^{n} k^{n-i}(|z_i - w_i| - |x_i - w_i|)$ is an integer. Therefore $\Delta_{\text{lead}} \geq 1/k^n$.

Further define

$$\begin{aligned}
\Delta_{\text{tail}} &= |d(z[n+1,M], y[n+1,M]) - d(x[n+1,M], y[n+1,M])| \\
&= \left| \sum_{i=n+1}^{M} \frac{|z_i - y_i| - |x_i - y_i|}{k^i} \right| \\
&\leq \sum_{i=n+1}^{M} \frac{\big||z_i - y_i| - |x_i - y_i|\big|}{k^i} \\
&\leq \sum_{i=n+1}^{M} \frac{\big|(z_i - y_i) - (x_i - y_i)\big|}{k^i} \\
&= \sum_{i=n+1}^{M} \frac{\big|z_i - x_i\big|}{k^i} \\
&\leq \sum_{i=n+1}^{M} \frac{k-1}{k^i} \\
&= \frac{1}{k^n} - \frac{1}{k^M}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
d(x,y) &= d(x[1,n], w) + d(x[n+1,M], y[n+1,M]) \\
&= d(z[1,n], w) - \Delta_{\text{lead}} + d(x[n+1,M], y[n+1,M]) \\
&\quad + d(z[n+1,M], y[n+1,M]) - d(z[n+1,M], y[n+1,M]) \\
&= d(z[1,n], w) + d(z[n+1,M], y[n+1,M]) \\
&\quad + d(x[n+1,M], y[n+1,M]) - d(z[n+1,M], y[n+1,M]) - \Delta_{\text{lead}}
\end{aligned}$$

$$\leq \quad d(z,y) + (\Delta_{\text{tail}} - \Delta_{\text{lead}})$$

$$\leq \quad d(z,y),$$

since $\Delta_{\text{tail}} - \Delta_{\text{lead}} < 0$ because $\Delta_{\text{tail}} \leq 1/k^n - 1/k^M < 1/k^n \leq \Delta_{\text{lead}}$. □

Proposition 10 says that if the first n terms of a word, x, are closer to some length-n missing leading word, w, than the first n terms of another word, z, then all words which begin with w will be closer to x than they are to z.

To review, we now know that a word whose distance to a distinguished set is maximal will have leading terms equal to some missing leading word, w, of minimal length. Furthermore, the only distinguished words which may be closest to a word starting with w are those whose first n terms are closest to w. Together these results allow us to associate to each missing word of minimal length n a new (hopefully much smaller) distinguished set consisting of the tails (of length $M - n$) of those distinguished words whose n leading terms are closest (among all distinguished words) to w. We can then apply these results again to each of these new distinguished sets, searching for missing leading terms to tack onto the current missing leading terms until we have constructed a collection of candidate words whose distance to the original distinguished set can be computed. The above propositions guarantee that among these candidates will be a length-M word whose distance to the distinguished set D is maximal among all length-M words. For example, say y (length-j) is the unique, shortest missing leading word from a distinguished set D, and associated to it is a refined distinguished set D_y consisting of words of length $M - j$. Assume further that D_y has three missing leading words, y'_1, y'_2, and y'_3 (each with their own associated distinguished sets), of minimal length, j'. Then we know that a word which maximizes its minimum distance to D will have the word (y, y'_1), (y, y'_2), or (y, y'_3) as its first $j + j'$ letters.

In Algorithm 1, we present pseudo-code for the algorithm described by Propositions 9 and 10, which determines the density failure epsilon of a distinguished set $D \subset \{0, \ldots, k-1\}^M$. The function $W = \mathit{missingLeads}(D)$ takes in a list of k-ary words of length M and determines the smallest integer j for which there exists a length-j word which does not appear as the first j letters of any element $x \in D$. For each such missing lead, $w = (w_1, \ldots, w_j)$, the distance to the leading j letters of each element of the distinguished set is computed $(\{d(w, x[1,j]) | x \in D\})$. We associate to each missing lead a refined distinguished set consisting of the length-$(M-j)$ tails of those elements $x \in D$ with $d(w, x[1,j]))$ minimal among all $x \in D$. All missing leads with minimal distance to D are retained in the output structure W, along with their associated refined distinguished sets. The function *missingLeads* is then recursively applied to each refined distinguished set. This process naturally terminates when no refined distinguished set has any missing leading words, at which point we will have constructed a list of candidate words whose distance to the original distinguished set D can be computed. The candidate with maximal distance to D achieves the density failure epsilon of D in $\Sigma_k[M]$.

The key benefit of this algorithm is that run-time effectively does not grow with word length. Instead, the run-time should grow exponentially with the lengths of the missing leading words. This is because a brute-force search of $\Sigma_k[m]$ is performed for each $m \leq j$, where j is the shortest missing leading word. Naturally, as the distinguished set grows, the length of the shortest missing lead tends to grow as well. This shifts the difficulty of solving this problem in a space of long words to solving the problem for a subset consisting of many words. Since our application demands we consider large word lengths, and thus very large search spaces relative to the size of the distinguished sets, the payoff of this shift in complexity is enormous.

Algorithm 1: $W = missingLeads(D)$

input : D, a distinguished set of n words in $\Sigma_k[M]$.
output: W, a structured list consisting of missing leads, $W[\,].missLead$, and their associated refined distinguished sets, $W[\,].refinedD[\,]$

```
j = 1
leads[ ] = an empty array
W[ ].* = an empty structure with components missLead and refinedD[ ]
while leads[ ] is empty do
    // Checks if D consists of all k-ary words of length j.
    // If so, returns a random element of D as a missing lead.
    if j < M then
        DLeads[ ] = the list of the j leading terms in D
        missLeads[ ] = the list of possible j-length words not in DLeads[ ]
        j = j + 1
    else
        W[ ].missLead = a random element of D
        W[ ].refinedD[ ] = an empty array
        return

// Computes the distance from each missing lead to each lead in D.
for i = 1 to size of missLeads[ ] do
    for k = 1 to size of DLeads[ ] do
        distances[i, k] = distance from ith element of missLeads[ ] to kth element of
        DLeads[ ]

// Associates to each missing lead the tails of those elements of D
// whose leading terms are closest to that missing lead.
for i = 1 to missLeads[ ] do
    currentMissLead = ith element of missLeads[ ]
    W[i].missLead = currentMissLead
    minDist = the minimum distance from currentMissLead to DLeads[ ]
    allMinDist[ ] = the list of all elements of DLeads[ ] a distance minDist to
    currentMissLead
    for r = 1 to size of allMinDist[ ] do
        for k = 1 to size of D do
            if the leading j terms of the k-th element of D match the r-th element of
            allMinDist[ ] then
                append to W[i].refinedD, the M − j trailing terms of the k-th element
                of D
remove duplication from W[i].refinedD
```

In Figure 1.8, we present a run-time analysis of Algorithm 1. For each distinguished set size ($500 \leq n \leq 3000$) we generated, at random, 5000 distinguished subsets of $\Sigma_2[50]$. The epsilon failure density of each distinguished set was determined using a Matlab implementation of our algorithm. The average, minimum, and maximum run-time over the 5000 randomly sampled distinguished sets is given in Figure 1.8 (a), (b) and (c) respectively. It is somewhat surprising that the average run-time appears to be non-monotonic in the size of the distinguished set.

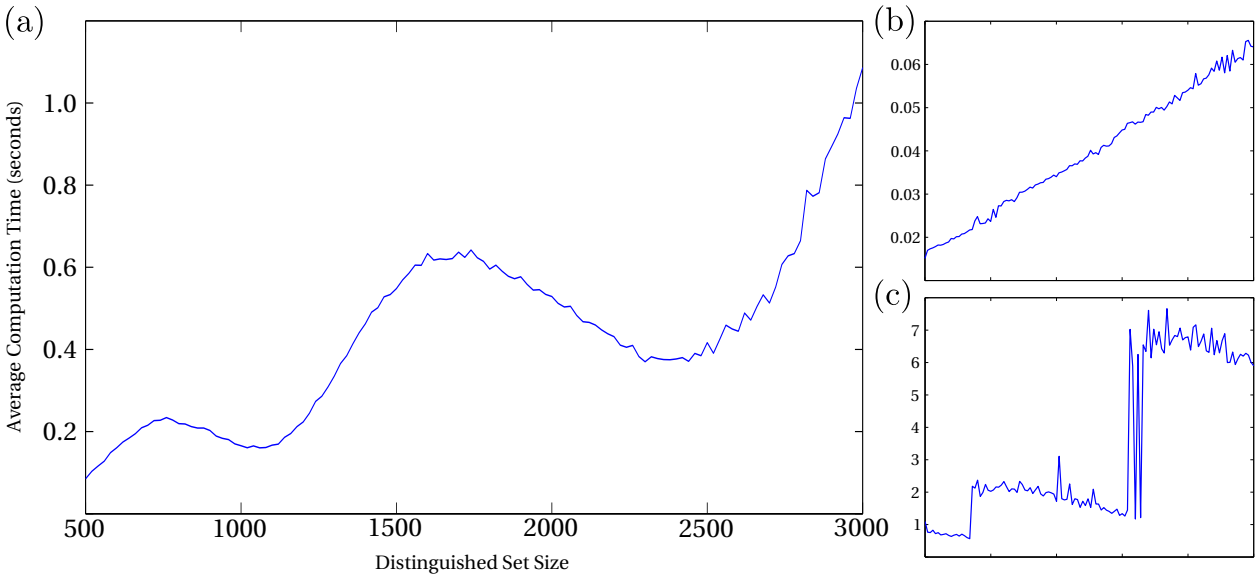

FIGURE 1.8. Timings of the Matlab implementation of Algorithm 1 were performed using a single-thread instance of Matlab running on an AMD Opteron 6276 processor with a 2300.110 MHz clock speed. The operating system used was CentOS Linux version mockbuild@c6b8.bsys.dev.centos.org. (a) The average time (over 5000 distinguished sets, sampled at random for each distinguished set size in the range 500 to 3000) required to determine a point in the space $\Sigma_2[50]$ whose distance to a distinguished set is maximal. (b) The minimum run-time over all 5000 distinguished sets for each distinguished set size in the range 500 to 3000. (b) The maximum run-time over all 5000 distinguished sets for each distinguished set size in the range 500 to 3000.

1.3.3. DE BRUIJN SEQUENCES. In the proof of Proposition 8, we noted that one can prove that the Bernoulli shift map is topologically transitive by constructing a sequence whose orbit is topologically transitive in Σ_k. One such infinite word, or sequence, can be

built by recursively concatenating all k-ary words of length one, followed by all k-ary words of length two, and so on.

For example, let $\mathcal{S} \doteq \bigcup_{i=1}^{\infty} \Sigma_s[i]$ have the ordering, for all $s, t \in \mathcal{S}$, $s < t$ if s is shorter than t or if s has the same length as t and $n(s) < n(t)$, where $n : \Sigma_k \to \mathbb{N}$ is the function taking a binary word to the integer value it represents. Then the sequence built by concatenating all elements of $\mathcal{S}$ in increasing order begins

$$x = (\underbrace{0,1,}_{\text{length-one}} \underbrace{0,0,0,1,1,0,1,1,}_{\text{length-2}} \underbrace{0,0,0,0,0,1,0,1,0,0,1,1,1,0,0,1,0,1,1,1,0,}_{\text{length-3}} \ldots)$$

We know that this element has a dense orbit. However, as we begin shifting this sequence it becomes clear that the way this sequence fills the space is suboptimal. For example, it takes 6 shifts before every binary word of length two appears as the leading term of an element in the orbit of x ((1,0) first appears as a leading word in $\sigma^6(x)$). This means that the density failure of the first seven truncated orbits is at least $1/4$. Had we instead chosen x to begin

$$x = (0, 1, 1, 0, 0, \ldots),$$

all words of length two (and thus also of length one) appear within the first 3 shifts.

In the example above, we showed that it is possible to construct a binary word of length five which contains all subwords of length two. Five is, of course, the shortest that such a word could be since, in general, if a k-ary word is to contain all k-ary subwords of length M, it must have length at least $k^M + M - 1$.

DEFINITION 1.3.1. A **k-ary de Bruijn sequence** $B(k, M)$ of order M is a word over the alphabet with k letters, $A = \{0, 1, \ldots, k-1\}$, for which every possible subsequence of

length M in A appears as a sequence of consecutive letters exactly once. Each $B(k, M)$ will necessarily have length $k^M + M - 1$.[1]

Although these objects carry the name de Bruijn, after Dutch mathematician Nicolaas Govert de Bruijn (see [17]), he acknowledges in [18] that binary de Bruijn sequences were first studied by Camille Flye Sainte-Marie in 1894 [19]. M. H. Martin [20], I.J. Good [21], and D. Rees [22] generalized the binary case to larger (finite) alphabets and proved the existence of k-ary de Bruijn sequences of all orders.

We now present a proof of this result, which relies on a well known theorem regarding the existence in directed graphs of **Euler cycles**: a tour which starts and ends at the same vertex and visits each edge in the graph exactly once. The theorem states that a directed graph has an Euler cycle if and only if it is connected and the indegree of each vertex equals its outdegree. This classical result provides the backbone of the proof of the existence of k-ary de Bruijn sequences.

PROPOSITION 11. *For each pair of positive integers k, M there exists a k-ary de Bruijn sequence of order M.*

PROOF. We will prove the claim by constructing a connected directed graph, called a de Bruijn graph, that has the property that the indegree of each vertex equals its outdegree. The existence of an Euler cycle in this graph will then be used to prove the existence of a k-ary de Bruijn sequence of order $M + 1$.

Let M and k be positive integers and let the vertices of a graph be all k-ary words of length M, $V = \Sigma_k[M]$. We will draw a directed edge from vertex x to vertex y if $(x[2, M], r) = y$ for some letter $r \in \{0, 1, \ldots, k-1\}$. Since there are k choices for r, each vertex x will have

[1]Traditionally a de Bruijn sequence $B(k, M)$ is taken to be a cyclic word of length k^M.

an outdegree equal to k. Similarly there are k vertices which could have been shifted by one letter into x, namely the vertices $y = (r, x[1, M-1])$ for some letter r, of which there are k. Thus each vertex has an indegree equal to k as well. Note that there is a natural bijection between the edges of this graph and the words of length $M+1$ built by labeling the edge connecting x to y by the word $(x[1, M], y_M)$.

To see that this defines a connected graph let $x, y \in V$ and construct the path

$$x \to (x[2, M], y_1) \to (x[3, M], y[1, 2]) \to \ldots \to (x_M, y[1, M-1]]) \to y.$$

Because the graph is connected and each vertex has an indegree equal to its outdegree, we know there exists an Euler cycle. Say this tour starts at vertex u and first crosses the edge connecting to v.

By following this tour, we can construct a word z of length $k^M + M - 1$ by starting with the word $(u[1, M], v_1)$ (i.e. the edge connected u to v) and appending to it the last letter of each edge in the order they are traversed. By the observation that the edges of this graph are in bijection with $\Sigma_k[M+1]$, visiting each edge in the graph exactly once amounts to seeing each word in $\Sigma_k[M+1]$ exactly once as a subword of z.

We have shown that de Bruijn sequences of all orders $M \geq 2$ exist. The sequence $(0, 1, \ldots, k-1)$ is de Bruijn of order 1. $\square$

From our perspective, the existence of de Bruijn sequences of all orders is significant since for each integer $M \geq 1$ there exists a sequence $x \in \Sigma_k$ whose truncated orbit after $k^M - 1$ shifts has a density failure no larger than $\sum_{i=M+1}^{\infty}(k-1)/k^i = 1/k^M$. This is because the distinguished set $\gamma_x(k^M - 1)$ of a point x whose leading $k^M + M - 1$ terms is a k-ary de Bruijn sequence of order M, has the property that every k-ary word of length M appears as

the leading M terms of one of its element. In fact, each k-ary word of length M appears as the leading M terms of *exactly* one of its elements, since $\gamma_x(k^M - 1)$ has size k^M. Thus, for such a sequence, we know that

$$k^M \, \mathcal{E}(\gamma_x(k^M - 1)) \leq 1. \tag{4}$$

This discussion bears no mention of the tail of such an x and so gives us no deeper insight into the possible limiting behaviors of $(m+1)\mathcal{E}(\gamma_x(m))$. However, if we could show the existence of a sequence whose first $k^{M_i} + M_i - 1$ terms is a k-ary de Bruijn sequence of order M_i, for infinitely many M_i, then the bound given in (4) would apply to every tail of such a sequence, and would ultimately provide an upper bound on the smallest possible $\liminf_{m\to\infty}[(m+1)\mathcal{E}(\gamma_x(m))]$ over Σ_k.

In 2011, V. Becher and P. A. Heiber [23], provided complete answers to this line of questioning by proving the following theorem:

THEOREM 1.3.2 (**Becher and Heiber**). *Every de Bruijn sequence of order M in at least three symbols can be extended to a de Bruijn sequence of order $M+1$. Every de Bruijn sequence of order $M \geq 2$ in two symbols cannot be extended to order $M+1$, but can be extended to order $M+2$.*

The heart of the proof is the observation that a de Bruijn graph on 2^{M+1} vertices is necessarily disconnected if one removes the edges visited by a path derived from any de Bruijn word of order M. In particular, the vertex consisting of all ones or the vertex consisting of all zeros (which both have an edge pointing from themselves to themselves) must be disconnected from the rest of the graph. Therefore no such path can be extended to on

Euler cycle on the original graph. This does not occur for graphs built over larger alphabets, nor does a path derived from a binary de Bruijn word of order M disconnect the de Bruijn graph on 2^{M+2} vertices.

This theorem allows us to define another class of elements of Σ_k with dense orbits.

Definition 1.3.3. An **infinite k-ary de Bruijn sequence** is an element of Σ_k built by the recursive extension of finite de Bruijn sequences.

The existence of infinite de Bruijn sequences immediately provides the following upper bound:

Proposition 12.

$$\inf_{x\in\Sigma_k} \liminf_{m\to\infty} [(m+1)\mathcal{E}(\gamma_x(m))] \leq 1.$$

Proof. Let $k \geq 2$ and $x \in \Sigma_k$ be an infinite de Bruijn sequence built by extension starting with a de Bruijn sequence of order n. For every $M = n + 2r$ ($r \in \mathbb{N}$), after $k^M - 1$ shifts of x, each element of $\Sigma_k[M]$ appears as the leading term of exactly one element in $\gamma_x(k^M - 1)$ (for $k > 2$ this holds for all positive integers M, not only every other integer, as in the binary case $k = 2$). Then for every point $y \in \Sigma_k$,

$$\begin{aligned} d(y, \gamma_x(k^M - 1)) &= \min_{0\leq j\leq k^M-1} \left\{ \sum_{i=1}^{\infty} \frac{|y_i - x_i^j|}{k^i} \right\} \\ &= \min_{0\leq j\leq k^M-1} \left\{ \sum_{i=1}^{M} \frac{|y_i - x_i^j|}{k^i} + \sum_{M+1}^{\infty} \frac{|y_i - x_i^j|}{k^i} \right\} \\ &= \min_{0\leq j\leq k^M-1} \left\{ \sum_{M+1}^{\infty} \frac{|y_i - x_i^j|}{k^i} \right\} \end{aligned}$$

$$\leq \sum_{M+1}^{\infty} \frac{k-1}{k^i} = \frac{1}{k^M},$$

which implies

$$k^M \, \mathcal{E}(\gamma_x(k^M - 1)) \leq k^M \, \frac{1}{k^M} = 1,$$

for all $M = n + 2r$. Therefore $\liminf_{m\to\infty} [(m+1)\mathcal{E}(\gamma_x(m))] \leq 1$. □

Since the quantity $(m+1)\mathcal{E}(\gamma_x(m))$ grows linearly with m until the introduction of a new density failure epsilon, it is easy to see that for $k > 2$, the largest that this quantity can get is certainly bounded. In fact,

PROPOSITION 13. *For any infinite de Bruijn sequence over an alphabet of size $k > 2$,*

$$\limsup_{m\to\infty} (m+1)\mathcal{E}(\gamma_x(m)) \leq k.$$

For any infinite binary de Bruijn sequence,

$$\limsup_{m\to\infty} (m+1)\mathcal{E}(\gamma_x(m)) \leq 4.$$

PROOF. According to Proposition 12, for each $m = k^M$, $(m+1)\mathcal{E}(\gamma_x(m)) \leq 1$, and $\mathcal{E}(\gamma_x(m)) \leq 1/k^M$. The quantity $(m+1)\mathcal{E}(\gamma_x(m))$ could grow at most linearly at a rate of $1/k^M$ until $m = k^{M+1}$, at which point a new smaller density failure epsilon must be introduced, since the sequence is an infinite de Bruijn sequence built from extensions. Therefore, over the $k^M(k-1) - 1$ integers in the range $\left[k^M, k^{M+1} - 1\right]$, the quantity $(m+1)\mathcal{E}(\gamma_x(m))$

could increase by at most

$$\frac{1}{k^M}\left(k^M(k-1)-1\right) = (k-1) - \frac{1}{k^M}.$$

This implies that the largest $(m+1)\mathcal{E}(\gamma_x(m))$ could be on the interval $\left[k^M, k^{M+1}-1\right]$ is $(k-1) - 1/k^M + 1 = k - 1/k^M$. Since this is true for each $M \in \mathbb{Z}$, the first part of the claim is proven.

An infinite binary de Bruijn sequence could maintain a linear growth rate of at most $1/2^M$ over the iterate range $m \in \left[2^M, 2^{M+2}-1\right]$, and so $(m+1)\mathcal{E}(\gamma_x(m))$ is bounded above by $4 - 1/2^M$ on this interval. □

In the proof of Proposition 13 we claim that a de Bruijn sequence could at most grow linearly over the entire interval $\left[k^M, k^{M+1}-1\right]$. This is, a priori, true. We hypothesize, however, that no de Bruijn extension actually grows linearly over the entire range between extensions. In other words, new density failure epsilons must be introduced earlier (after fewer shifts) and more often than along powers of k. How early and how often will govern how large $m\mathcal{E}(\gamma_x(m-1))$ actually becomes and thus how small the LLD of a particular infinite de Bruijn sequence actually is.

We return now to our endeavor to compute numerical approximations of the quantity $m\mathcal{E}(\gamma_x(m-1))$. By utilizing the algorithm developed in Section 1.3.3, we are able to compute $\mathcal{E}(\gamma_x(m))$ for very long de Bruijn sequences built by extension of x, to arbitrarily high precision, in reasonable time. To accomplish this, we first must generate an infinite de Bruijn extension, or at least a reasonably long finite approximation of one. To this end, we begin with a de Bruijn sequence of order 2 and think of it as the beginning of an Eulerian cycle in a de Bruijn graph of order 3 (or order 4 if $k = 2$). We then extend this to a

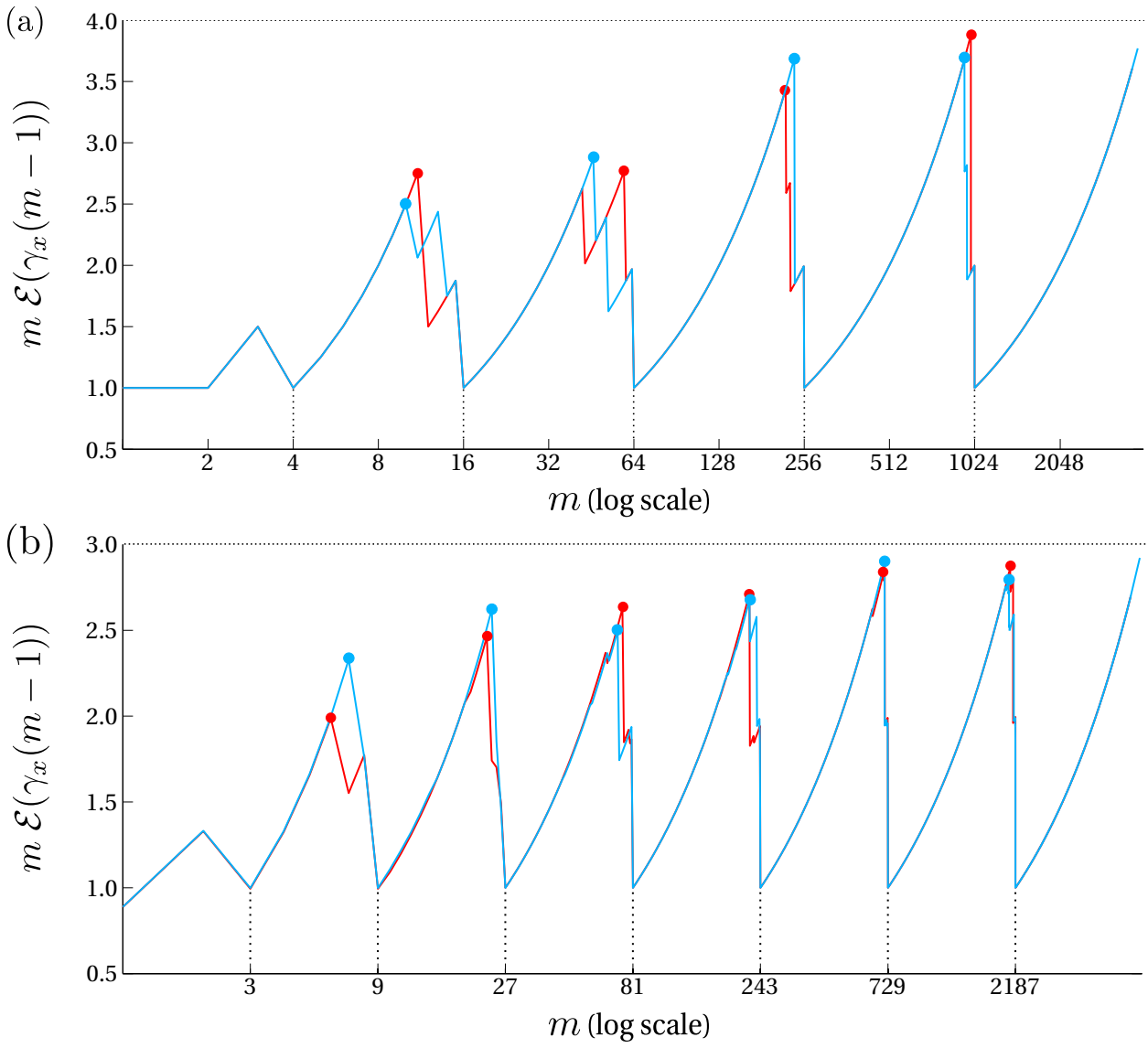

FIGURE 1.9. (a) Plots of linear-time-rescaled density failure $m\mathcal{E}(\gamma_x(m-1))$ versus the number of iterations (m) of the shift map, applied to two choices of binary de Bruijn sequences (x_1 and x_2) built by extension using the order 2 de Bruijn sequence 1 0 0 1 1 as a seed. The leading words $x_i[1, 2^M + M - 1]$ are de Bruijn sequences of order M, for even integers M. (b) Plots of the linear-time rescaled density failure for two ternary de Bruijn sequences.

complete Eulerian cycle using a modified Hierholzer's algorithm, being careful to maintain the beginning of the cycle, which represents the order-2 de Bruijn sequence. The process of embedding an Eulerian cycle of order M into a path in the order $M + 1$ graph and then

extending it to a full Eulerian cycle in the order $M+1$ graph is repeated as many times as desired, creating a de Bruijn extension, x.

Since the distance between two infinite sequences is approximated to within $1/k^M$ by the distance between the first M leading terms, we generate truncated orbits consisting of m points in $\Sigma_k[60]$, where m ranges from 1 to $k^n + n - 1 - 59$ by sliding a window of length 60 along the de Bruijn extension x.

Figure 1.9 shows plots of a numerical approximation of the quantity $m\mathcal{E}(\gamma_x(m-1))$ versus the number of iterations (m) of the shift map for several choices of initial (a) binary and (b) ternary de Bruijn extensions. The largest values within each of the intervals $\left[k^M, k^{M+1}\right]$ are highlighted. Notice these peaks need not occur at the iteration immediately preceding the first drop in density failure epsilon. These numerical approximations support the belief that k (or 4, in the binary case) need not be a strict upper bound. These computations also suggest that, not surprisingly, the LLD of the orbits of two different infinite de Bruijn sequences over the same alphabet are likely to be different.

1.4. Conclusions and Future Work

We were first motivated to consider optimal transitivity in rotation maps by studies of phyllotaxis, the arrangement of elements such as leaves, seeds, or florets at plant meristems. Observations of plants as well as analysis of PDE models of plant growth [24] lead to a description of the positioning of phyllotactic elements around a plant meristem via a set $S(\theta, N) \subset \mathbb{R}^2$ given by

$$S(\theta, N) = \left\{ \left(\sqrt{n}\cos(2\pi n\theta), \sqrt{n}\sin(2\pi n\theta) \right) : n = 1 \dots N \right\},$$

where, for most plants, θ is the golden number $\phi = \left(\sqrt{5}-1\right)/2$. Plots of the set $S(\theta, N)$ for $N = 400$ and (a) $\theta = \phi = \left(\sqrt{5}-1\right)/2$ and (b) $\theta = 4 - \pi$ are shown in Figure 1.10 and have distinctly different characters: for $\theta = \phi = \left(\sqrt{5}-1\right)/2$ the pattern is that of a sunflower head with its evenly spaced seeds, whereas for $\theta = 4 - \pi$, there are large gaps between the spirals that dominate the pattern. Note that the angular positions of the points in $S(\theta, N)$ are the orbit of the rotation map with rotation angle θ, so that Figures 1.10 (a) and (b) are respectively analogous to Figures 1.1 (a) and (b), but with the addition of a radial coordinate that is a monotonically increasing function of iteration time n. One visualizes, even in a plot of $S(\theta, N = \infty)$, a dependence on θ of the optimality of packing. A similar visualization may also prove to be interesting for other dynamical systems defined by circle diffeomorphisms, as discussed in Section 1.2.4, or other discrete-time dynamical systems on the circle.

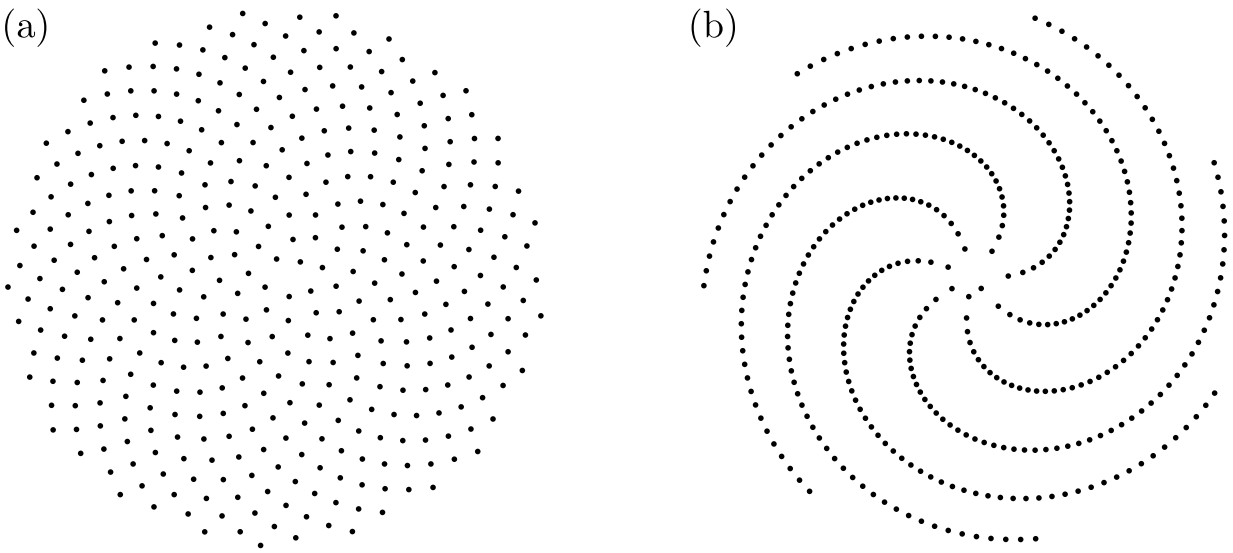

FIGURE 1.10. The set $S(\theta, N)$ for $N = 400$ and (a) $\theta = \phi = \left(\sqrt{5}-1\right)/2$ and (b) $\theta = 4 - \pi$.

Among linear circle maps the sequence of points $x_k = k\phi \bmod 1$ $\left(\phi = \left(\sqrt{5}-1\right)/2\right)$ achieves the minimal possible LLD. Theorem 1.2.7 states that the LLD of any sequence of points in the circle is bounded below by $\frac{1}{2\ln(2)}$ and that this is achieved by the sequence

$x_k = \log_2(2k-1) \bmod 1$. Note that the former is a linear map, while the latter is non-polynomial. One might search between these extremes for the optimal LLD among sequences given by nonlinear polynomials of fixed degree. As the degree increases, does the smallest possible LLD decrease towards the bound achieved by $x_k = \log_2(2k-1) \bmod 1$? Does there exist a sequence of polynomials which converges to this bound? For the time being, these questions will have to remain enticing but unexplored.

Concerning optimally topologically transitive orbits for Bernoulli shift maps, many questions remain open. Of combinatorial interest: how many de Bruijn extensions are there at each order? Is this number constant or does it depend on which finite de Bruijn sequence is being extended? If one minimally extends a binary de Bruijn word of order M so that it contains all subwords of length $M+1$ (necessarily some more than once), can this word still be extended to a de Bruijn sequence of order $M+2$? And of dynamical interest: on a fixed interval $\left[k^M, k^{M+1}-1\right]$, how soon, how often, and how large can the drops in the size of the density failure epsilon be? Among all infinite de Bruijn sequences which has minimal LLD? And finally, since the LLD depends on the tails of the sequence $(m+1)\mathcal{E}(\gamma_x(m))$, does the optimal orbit need to be de Bruijn at earlier stages, or might it be that a non-de Bruijn leading term allows for earlier and more frequent drops in the density failure epsilon at larger finite times?

The notion of optimal transitivity put forward in this paper may be extended to consider discrete-time dynamical systems with higher dimensional phase spaces, such as linked twist maps on tori, which are employed to model fluid mixing in DNA microarrays [25]. Separately, de Bruijn sequence are of interest in the design of DNA microarrays that are used to determine the binding preferences of transcription factors [26]. In this context, it may be appropriate to to consider a new type of mathematical object: minimal length words which

contain all words of some length as subwords written either forward or backward. As far as the authors are aware these so-called **unoriented de Bruijn sequences** have not been studied in the literature. Thus, many of the questions which have been resolved for de Bruijn sequences – such as existence, abundance and extendibility – remain relevant and open for unoriented de Bruijn sequences. We have begun searching for answers to these questions and will present our findings in future work.

CHAPTER 2

Complex Hadamard Matrices and Mutually Unbiased Bases

2.1. Introduction

In this chapter the principal objects of interest are **complex Hadamard matrices**; matrices $H \in M^{d\times d}(S^1)$ having the property that

$$HH^* = dI_d,$$

where I_d is the $d \times d$ identity matrix, * denotes conjugate transpose, and $S^1 \subset \mathbb{C}$ is the complex unit circle. These are natural generalizations of their real analogue: square matrices with entries in $\{-1, 1\}$ with mutually orthogonal rows and columns, discovered – by Hadamard in 1893 [27] – to have the largest determinant among all real matrices with entries whose absolute values are bounded by one.

While both real and complex Hadamards are, in their own right, mathematically interesting objects, they also have a wide range of applications: Real Hadamards have been used in coding theory [28] and the design of statistical experiments [29], while complex Hadamards are found in numerous constructions in theoretical physics ([30], [31], [32]) and quantum information theory [33].

It is an open conjecture of Jacques Hadamard that real Hadamards of size $d \times d$ exist for all dimensions $d > 2$ which are multiples of 4. However, the existence of complex Hadamards of all orders is settled by explicit construction of the so-called Fourier matrices:

$$[F_d]_{j,k} = e^{\mathrm{i}(j-1)(k-1)2\pi/d} \text{ with } j, k \in \{1, 2, \ldots, d\}.$$

Numerous, often nontrivial constructions have led to the discovery of many other complex Hadamards, including continuous, parametrized families. Thus the focus of much of the current mathematical research ([34], [35], [36], [37]) is aimed at complete classification of complex Hadamards, at least for small dimensions.

In dimensions $d \leq 5$ complete classification has been established [38]. However, for 6×6 matrices, the question remains open if not on the precipice of resolution [39]. For small dimensions ($d \geq 6$), although many methods of construction have been exploited to give explicit families of matrices ([40],[41],[42]), complete identification and classification appears difficult to obtain.

It is our intent to explore novel ways of thinking about complex Hadamard matrices by bringing in techniques from dynamical systems to provide both numerical and analytical evidence of existing conjectures. Our hope is that, ultimately, these approaches will develop into fruitful methods of attack which will resolve some open problems. In Section 2.2 we establish the necessary preliminaries and introduce basic definitions. We then proceed in Section 2.3 to construct a novel discrete-time dynamical system which elicit complex Hadamards, and closely-related objects known as mutually unbiased bases. In Section 2.3.3, we utilize this algorithm as a tool to generate collections of numerical matrices and, by adopting techniques from geometric data analysis, we buttress several existing conjectures regarding dimension-6 Hadamard matrices.

The main contribution of this chapter rests in Section 2.4, which is dedicated to crafting a novel approach for determining the local dimension of the space of inequivalent Hadamards using center-manifold theory. We first present evidence of the applicability and scope of our technique by working with an example for which some of the properties our method elucidates have already been established by other means. We conclude this chapter by applying our

technique to answer an open problem regarding the local structure of the space of Hadamards near a known 9×9 Hadamard.

2.2. Preliminaries

In this section we fix our notation, establish key definitions, and focus our discussion to particular problems of interest. Without exception we will differentiate vectors from scalars with the use of bold-face font: $\mathbf{v} = [v_1, \ldots, v_n]$. We will denote the row-i, column-j entry of an $m \times n$ complex matrix $X \in M^{m\times n}(\mathbb{C})$ by $[X]_{i,j}$, unless the entries have been specified otherwise. X^* will be used to denote the conjugate-transpose of the matrix X, while $\overline{w}$ will denote the complex conjugate of a the scalar $w \in \mathbb{C}$. The complex dot product between two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{C}^n$ is defined to be

$$\mathbf{v} \cdot \mathbf{w} \doteq v_1\overline{w_1} + v_1\overline{w_1} + \ldots + v_1\overline{w_1},$$

and we will use the notation $X \circ Y$ and $\mathbf{v} \circ \mathbf{w}$ to denote componentwise multiplication of matrices and vectors.

It will often be necessary for us to compare the distance between complex matrices. For this purpose we will employ several commonly-used matrix norms:

Definition 2.2.1. Define the **Frobenius norm** $\|\cdot\|_F : M^{m\times n}(\mathbb{C}) \to [0, \infty)$ by

$$\|X\|_F \doteq \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|[X]_{i,j}\right|^2} = \sqrt{\mathrm{tr}(X^*X)}.$$

Note that the Frobenius norm is merely the Euclidean norm on $\mathbb{C}^{mn}$, thinking of an $m \times n$ matrix X as a complex vector of length mn. Another natural norm is derived by considering

a matrix X as a linear operator on $\mathbb{C}^n$. The resulting operator norm induced by the standard L^2-norm on a vector space is called the spectral norm.

DEFINITION 2.2.2. Define the **spectral norm** $\|\cdot\|_2 : M^{m\times n}(\mathbb{C}) \to [0,\infty)$ by

$$\|X\|_2 \doteq \sup_{\mathbf{v}\neq 0}\left\{\frac{\|X\mathbf{v}\|_2}{\|\mathbf{v}\|_2}\right\} = \sqrt{\lambda_1(X^*X)} = \sigma_1(X),$$

where $\lambda_1(M)$ and $\sigma_1(M)$ are, respectively, the largest eigenvalue and singular value of a matrix M.

2.2.1. COMPLEX HADAMARD MATRICES. Let $\mathcal{H}_d$ denote the subset of complex Hadamards in $M^{d\times d}(\mathbb{C})$: that is the matrices H which satisfy

$$\left|[H]_{i,j}\right| = 1, \text{ for } 1 \leq i,j \leq d \text{ and}$$

$$HH^* = dI_d. \tag{5}$$

Obviously each entry of an element of $\mathcal{H}_d$ is a complex number of the form $e^{\mathrm{i}\theta}$, for some phase $\theta \in [0, 2\pi)$. Note that $\mathcal{H}_d$ is closed under permutations of matrix rows and columns. In other words P_1HP_2 is Hadamard for any $H \in \mathcal{H}_d$ and permutation matrices P_1 and P_2. We will say P_1HP_2 and H are **permutation equivalent**. Likewise, shifting the phase of a row or column of a matrix $H \in \mathcal{H}_d$ does not destroy the defining conditions 5. Thus, a natural equivalence relation is placed on $\mathcal{H}_d$: For $H, K \in \mathcal{H}_d$, we say $H \approx K$ if there exists unitary, diagonal matrices D_1 and D_2 and permutation matrices P_1 and P_2 such that

$$H = D_1P_1KP_2D_2.$$

It can be shown that for any $H \in \mathcal{H}_d$ there are unique unitary, diagonal matrices D_r and D_c which "rotate" H so that the entries in the first column and first row are all equal to 1. In particular, let $D_r = \mathbf{diag}\left(\overline{[H]_{1,1}}, \overline{[H]_{2,1}}, \ldots, \overline{[H]_{d,1}}\right)$ and $D_c = \mathrm{diag}\left(1, [H]_{1,1}\overline{[H]_{1,2}}, \ldots, [H]_{1,1}\overline{[H]_{1,d}}\right)$. Then $H' \doteq D_r H D_c \in \mathcal{H}_d$ and $H'_{1,i} = H'_{i,1} = 1$ for $i = 1, 2, \ldots, d$.

Definition 2.2.3. A complex Hadamard matrix, $H \in \mathcal{H}_d$, is said to be **dephased** if

$$H_{1,i} = H_{i,1} = 1 \text{ for } i = 1, 2, \ldots, d,$$

otherwise it is said to be **enphased**. After dephasing, the lower $(d-1) \times (d-1)$ submatrix is called the **core** of the matrix.

Thus we reduce the classification problem to the space of equivalence classes of complex Hadamards. In dimensions 2 and 3 it is straightforward to verify that the sets of equivalence classes are finite. In dimension 5 it is known that there also is only one equivalence class [38]. This observation leads to the following notion:

Definition 2.2.4. If there exists a neighborhood of a dephased Hadamard matrix H which does not contain any other dephased Hadamard matrix, H is said to be **isolated**.

If a dephased Hadamard matrix H is not isolated, then every neighborhood of H contains other dephased Hadamards. The unimodularity condition requires these Hadamards (and in fact all Hadamards) to be of the form

$$H \circ \exp(\mathrm{i}R) \tag{6}$$

where $R \in M^{d\times d}(\mathbb{R})$ is a matrix encoding the phases of the entries of the complex matrix $\exp(\mathrm{i}R)$: $[\exp(\mathrm{i}R)]_{i,j} = e^{\mathrm{i}[R]_{i,j}}$. If $H \circ \exp(\mathrm{i}R)$ is a dephased Hadamard, then the unitary

condition, together with the dephased property require the $d^2 + d - 1$ equations

$$
\begin{aligned}
R_{i,1} &= 0, \text{ for } 1 \leq i \leq d \\
R_{1,j} &= 0, \text{ for } 2 \leq j \leq d \\
\sum_{k=1}^{d} [H]_{i,k}[H^*]_{j,k} e^{\mathrm{i}\left([R]_{i,k} - [R]_{j,k}\right)} &= 0, \text{ for } 1 \leq i < j \leq d
\end{aligned} \tag{7}
$$

to be satisfied.[1] By computing the Jacobian of the above non-linear system at a particular Hadamard H, one derives a system of linear equations which has a solution space whose dimension is an upper bound for the dimension of the manifold of dephased Hadamards in a neighborhood of H.

DEFINITION 2.2.5. The **defect** $d(H)$ of a $d \times d$ Hadamard matrix $H \in \mathcal{H}_d$ is the dimension of the solution space of the real linear system

$$
\begin{aligned}
R_{i,1} &= 0, \text{ for } 1 \leq i \leq d \\
R_{1,j} &= 0, \text{ for } 2 \leq j \leq d \\
\sum_{k=1}^{d} [H]_{i,k}[H^*]_{j,k} \left([R]_{i,k} - [R]_{j,k}\right) &= 0, \text{ for } 1 \leq i < j \leq d
\end{aligned} \tag{8}
$$

where $R \in M^{d \times d}(\mathbb{R})$ is a matrix of variables.

Further define $H(\mathcal{R}) = \{H \circ \exp(\mathrm{i}R) \in \mathcal{H}_d | R \in \mathcal{R}\}$, where $\mathcal{R}$ is the subspace of $d \times d$ matrices satisfying 8 and 5 . Then $H(R)$ is a submanifold of complex Hadamards referred to as an **affine Hadamard family** (or orbit) stemming from H. The notion of the defect,

[1] For any matrix $X \in M^{d \times d}(\mathbb{C})$ with unit entries, the diagonal entries of XX^* are automatically equal to d.

introduced in [43], has been used extensively to compute bounds on the dimension of affine orbits stemming from dephased Hadamards [44]. In particular the following lemma is used:

LEMMA 2.2.6 (Tadej [43]). *The dimension of a continuous Hadamard orbit stemming from a dephased Hadamard matrix H is not greater than the defect, $d(H)$.*

As a consequence, a matrix must be isolated if its defect is 0.

In dimension 4 it has been shown that every complex Hadamard is equivalent to a member of a continuous 1-parameter family [45]. This orbit is

$$F_4^{(1)}(a) = F_4 \circ \exp\left(\mathrm{i}R_{F_4^{(1)}}(a)\right),$$

stemming from the Fourier matrix F_4, where $R_{F_4^{(1)}}(a), a \in [0, \pi]$ parametrizes the subspace of real matrices

$$R_{F_4^{(1)}}(a) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & a & 0 & a \\ 0 & 0 & 0 & 0 \\ 0 & a & 0 & a \end{bmatrix}.$$

$F_4^{(1)}(a)$ provides us with a first example of a 1-dimensional space of inequivalent complex Hadamards. In higher dimensions (in particular in dimension 6) efforts to classify Hadamards have lead to the construction of numerous 1, 2, and 3 dimensional inequivalent families ([35], [36], [37], [46], [47], [48], [49], [50]), as well as an isolated matrix called $S_6^{(0)}$, attributed independently to both T. Tao [51] and E. Moorhouse [52].

Strong numerical evidence [53] supports the conjecture [54] that there exists a 4-parameter family of inequivalent 6×6 Hadamards which contains the other known families. A very

recent paper gives a method of constructing general inequivalent Hadamards with four degrees of freedom, denoted $G_6^{(4)}$ [39]. Thus, the current conjecture is that all 6 by 6 complex Hadamards belong either to this 4-parameter family or are equivalent to $S_6^{(0)}$. This is the strongest conjecture of this type in dimension 6 and all numerical evidence suggests we are close to a complete classification of 6×6 complex Hadamards.

2.2.2. Mutually Unbiased Bases. A question related to the classification of complex Hadamards comes from quantum tomography and asks: in what dimensions can the precision of a measurement scheme used to determine all $d^2 - 1$ parameters that characterize a $d \times d$ density matrix be optimal? The answer to this question lies in determining for which dimensions there exists a set of $d + 1$ mutually unbiased bases (MUBs) for $\mathbb{C}^d$.

Definition 2.2.7. A collection of $d \times d$ unitary matrices (orthogonal bases) $\{A_1, ..., A_n\}$ is **mutually unbiased** if for all $1 \leq i \neq j \leq n$, and for all $1 \leq p, q \leq d$, we have

$$\left| \left[A_i A_j^* \right]_{p,q} \right| = \frac{1}{\sqrt{d}}.$$

Notice that multiplication of a set of dimension-d MUBs $\{A_1, ..., A_n\}$ by A_1^* gives rise to another set of MUBs: $\{I_d, A_2 A_1^*, \ldots, A_n A_1^*\}$. By definition, a matrix which is unbiased with respect to the identity is a unitary matrix whose entries are all of magnitude $1/\sqrt{d}$. Therefore, a complex Hadamard matrix may be regarded as $\sqrt{d}$ times a matrix that belongs to a collection of MUBs which includes the identity. In this way, searching for a collection of $d+1$ MUBs amounts to finding a collection of d **mutually unbiased Hadamards** (MUHs).

If we denote the maximum size of a set of MUBs in dimension $d = p_1^{e_1} \cdot ... \cdot p_r^{e_r}$ by $\mathcal{M}(d)$, (where $p_1^{e_1} \cdot ... \cdot p_r^{e_r}$ is the prime factorization of d, written so that $p_1 < p_2 < ... < p_r$), then it is known that $p_1^{e_1} + 1 \leq \mathcal{M}(d) \leq d + 1$ [55], [56]. Thus the quantum tomography

problem is solved for prime-power dimensions. However, for non-prime-power dimensions little progress has been made. In particular, for the smallest non-prime-power example (dimension 6) no improvement to the lower bound has been proposed and so it is conjectured that $\mathcal{M}(6) = 3$ [46]. While numerical evidence suggests that this conjecture is correct (see [57]), a proof has never been given. In the following sections we further support this conjecture with new numerical evidence produced by a discrete-time dynamical system that can converge to sets of MUBs.

2.3. Hadamard Fixed Points of a Discrete-Time Dynamical System

2.3.1. Generating Complex Hadamards. In this section we define a simple discrete-time dynamical system on $M^{d\times d}(C^*)$[2], whose unitary fixed points demonstrate remarkable stability. In effect, we introduce an algorithm for generating (numerical) $d \times d$ complex Hadamard matrices since a Hadamard matrix is just a scalar multiple away from being unitary.

For brevity we will refer to a matrix $X \in M^{d\times d}(C^*)$ s.t. $\left|[X]_{i,j}\right| = 1/\sqrt{d}$ for $i, j = 1, 2, \ldots, d$ as a **flat** matrix and denote the subset of all such matrices $\mathfrak{F}(d) \subset M^{d\times d}(C^*)$. Define the **flattening function** $\mathcal{F}_d : M^{d\times d}(\mathbb{C}^*) \to \mathfrak{F}(d)$ entry by entry by

$$[\mathcal{F}_d(X)]_{i,j} = \frac{[X]_{i,j}}{\left|[X]_{i,j}\right|\sqrt{d}}.$$

Also, we define the **uniterizing function**, $\mathcal{U}_d : M^{d\times d}(\mathbb{C}^*) \to \mathfrak{U}(d)$, as

$$\mathcal{U}_d(X) = UV^*,$$

[2] $\mathbb{C}^*$ is the punctured complex plane: $\mathbb{C} - \{0\}$. Since we will be choosing initial matrices at random, entries will all be non-zero with probability one.

where $X = U\Sigma V^*$ is the singular value decomposition of X[3] and $\mathfrak{U}(d)$ is the degree-d unitary group. By composing these two functions we define our dynamical system on the space of square complex matrices.

DEFINITION 2.3.1. Define $\Phi_d : M^{d\times d}(\mathbb{C}^*) \to M^{d\times d}(\mathbb{C}^*)$ by $\Phi_d(X) = \mathcal{U}_d(\mathcal{F}_d(X))$

Of course $\mathcal{U}_d$ likely destroys flatness and $\mathcal{F}_d(X)$ is unlikely to remain unitary for a general unitary X. However, enormous numerical evidence supports the following conjecture: For any $X \in M^{d\times d}(\mathbb{C}^*)$, the sequence ${\Phi_d}^n(X)$ converges (in any unitarily-invariant norm) to a flat and unitary matrix. As a first step towards proving this conjecture, let as show that flat, unitary matrices are fixed points of Φ_d.

LEMMA 2.3.2. *If $X \in M^{d\times d}(\mathbb{C}^*)$ is flat and unitary then $\Phi_d(X) = X$.*

PROOF. If $X \in M^{d\times d}(\mathbb{C}^*)$ is flat and unitary then

$$[\mathcal{F}_d(X)]_{i,j} = \frac{[X]_{i,j}}{\left|[X]_{i,j}\right|\sqrt{d}} = \frac{[X]_{i,j}}{\frac{1}{\sqrt{d}}\sqrt{d}} = [X]_{i,j} \implies \mathcal{F}_d(X) = X.$$

Also, X being unitary implies the singular value decomposition of is $X = UI_dV^*$ for some unitary matrices U and V. Therefore $\mathcal{U}_d(X) = UV^* = X$ and so X is fixed by Φ_d. □

The next step in proving the stability properties of Φ_d would be to demonstrate that its *only* fixed points are, in fact, matrices which are simultaneously flat and unitary. A priori it is conceivable that, in some dimension, there exists a flat matrix F and a unitary matrix $U \neq F$ which are related in just the right way so that $F = \mathcal{F}_d(U)$ and $U = \mathcal{U}_d(F)$, as illustrated in Figure 2.1.

[3] $\mathcal{U}(X)$ is the nearest unitary matrix to X as measured by any unitarily-invariant norm [58].

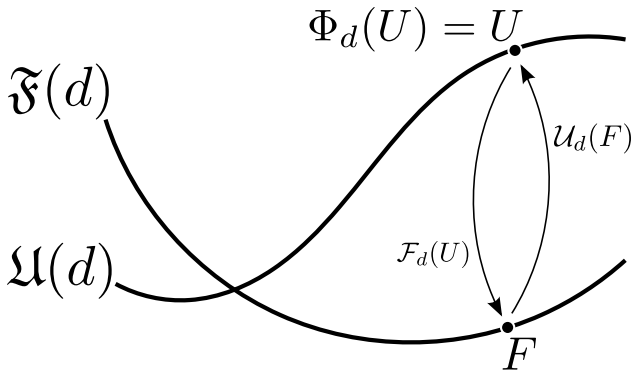

FIGURE 2.1. Illustration of a fixed point of Φ_d, U, which is unitary but not flat – a potentiality for $d > 2$.

The proof of the nonexistence of such strangely-related $d \times d$ matrices has eluded us for $d > 2$. We have, however, shown that the only fixed points of Φ_2 are those matrices which are are both flat and unitary. The reason for this is as surprising as it is unilluminating to the general case. The proof hinges on the remarkable observation that the the nearest 2×2 unitary matrix, $U \doteq \mathcal{U}_2(F)$, to any flat matrix F is itself flat. Therefore, if we assume that $\mathcal{F}_2(U) = F$ and $\mathcal{U}_2(F) = U$, so that U is a fixed point of Φ_2, it must have been that the flat matrix $U = \mathcal{F}_2(U) = F$ – where the first equality comes from the fact that a flat matrix is fixed by $\mathcal{F}_d$ and the second comes from our assumption of the relationship between U and F.

PROPOSITION 14. *Let $X \in M^{2\times 2}(\mathbb{C}^*)$. If $\Phi_2(X) = X$, then X is both flat and unitary.*

PROOF. We proceed by direct verification that the nearest unitary matrix to any flat matrix $F \in M^{2\times 2}(\mathbb{C}^*)$ is flat as well. Let

$$F \doteq \begin{bmatrix} w & x \\ y & z \end{bmatrix} \in M^{2\times 2}(\mathbb{C}^*)$$

be a flat matrix. Then the covariance matrix

$$FF^* = \begin{bmatrix} 1 & \alpha \\ \overline{\alpha} & 1 \end{bmatrix},$$

where $\alpha = w\overline{y} + x\overline{z}$. If $\alpha = 0$ then F is unitary and the claim is vacuous. So we assume $\alpha \neq 0$. The eigenvalues of the covariance matrix are $\lambda_\pm \doteq 1 \pm |\alpha|$ which implies the singular values of F are $\sigma_\pm \doteq \sqrt{1 \pm |\alpha|}$. Solving for the left-eigenvalues of FF^* yields the orthonormal left-singular-vectors of F

$$\mathbf{v}_+ \doteq \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{\alpha}{|\alpha|} \\ 1 \end{bmatrix}$$

$$\mathbf{v}_- \doteq \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{-\alpha}{|\alpha|} \\ 1 \end{bmatrix}.$$

Let

$$L \doteq \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{\alpha}{|\alpha|} & \frac{-\alpha}{|\alpha|} \\ 1 & 1 \end{bmatrix}$$

and

$$\Sigma \doteq \begin{bmatrix} \sigma_+ & 0 \\ 0 & \sigma_- \end{bmatrix}.$$

By the singular value decomposition ($F = L\Sigma R^*$), we have $R^* = \Sigma^{-1}L^{-1}F$, and so

$$U = \mathcal{U}(F) = LR^* = L\Sigma^{-1}L^{-1}F = \begin{bmatrix} \frac{\alpha}{|\alpha|} & \frac{-\alpha}{|\alpha|} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2\sigma_+} & 0 \\ 0 & \frac{1}{2\sigma_-} \end{bmatrix} \begin{bmatrix} \overline{\alpha}/|\alpha| & 1 \\ -\overline{\alpha}/|\alpha| & 1 \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix}$$

$$= \begin{bmatrix} wk_+ + y\frac{\alpha}{|\alpha|}k_- & xk_+ + z\frac{\alpha}{|\alpha|}k_- \\ yk_+ + w\frac{\overline{\alpha}}{|\alpha|}k_- & zk_+ + x\frac{\overline{\alpha}}{|\alpha|}k_- \end{bmatrix}, \tag{9}$$

where $k_\pm \doteq 1/2\sigma_+ \pm 1/2\sigma_-$.

With a little effort one can show that the square of the magnitude of each entry of matrix 9 is $1/2$, which means U is flat. For example,

$$\begin{aligned}
\left|[U]_{1,1}\right|^2 &= \left(wk_+ + y\frac{\alpha}{|\alpha|}k_-\right)\left(\overline{w}k_+ + \overline{y}\frac{\overline{\alpha}}{|\alpha|}k_-\right) \\
&= w\overline{w}{k_+}^2 + y\overline{y}{k_-}^2 + \frac{k_+k_-}{|\alpha|}w\overline{y\alpha} + \frac{k_+k_-}{|\alpha|}\overline{w}y\alpha + \\
&= \frac{1}{2}\left({k_+}^2 + {k_-}^2\right) + 2\mathrm{Re}\left(\overline{w}y(w\overline{y} + x\overline{z})\right)\frac{k_+k_-}{|\alpha|} \\
&= \frac{1}{2}\left({k_+}^2 + {k_-}^2\right) + 2\mathrm{Re}(1/4 + xy\overline{wz})\frac{k_+k_-}{|\alpha|} \\
&= \frac{1}{2}\left({k_+}^2 + {k_-}^2\right) + \frac{k_+k_-}{|\alpha|}\left(\frac{1}{2} + 2\mathrm{Re}(xy\overline{wz})\right) \\
&= \frac{1}{2}\left({k_+}^2 + {k_-}^2\right) - \frac{|\alpha|^2}{2-2|\alpha|^2} \quad \left(\text{since } k_+k_- = \frac{-|\alpha|}{2-2|\alpha|^2} \text{ and } 2\mathrm{Re}(xy\overline{wz}) = |\alpha|^2 - \frac{1}{2}\right) \\
&= \frac{1}{4{\sigma_+}^2} + \frac{1}{4{\sigma_-}^2} - \frac{-|\alpha|}{2-2|\alpha|^2} \\
&= \frac{2(1-|\alpha|^2)}{4(1-|\alpha|^2)} \\
&= \frac{1}{2}.
\end{aligned}$$

Therefore, the nearest unitary 2×2 matrix to a flat 2×2 matrix is itself not only unitary but also flat, which implies the claim. □

We see that in $M^{2\times 2}(\mathbb{C}^*)$, the collection of flat matrices and the space of unitary matrices are intertwined in a most delicate fashion: such that projection from $\mathfrak{F}(2)$ into $\mathfrak{U}(2)$ remains in $\mathfrak{F}(2)$. As a consequence, Φ_2 converges to a fixed point in a single iteration starting at any matrix in $M^{d\times d}(\mathbb{C}^*)$. Unfortunately, this enviable quality appears to belong uniquely to dimension-2.[4]

Thus far we have been unsuccessful in proving that Proposition 14 holds in any dimension greater than two. Until we verify this conjecture we are at a loss as to how to prove that flat and unitary matrices are global attractors of Φ_d. Nonetheless, we believe this statement to be true because of overwhelming numerical evidence.

Let us define a function that serves as a measure of how close a matrix is to being flat:

Definition 2.3.3. Define $\mathcal{N}_d : M^{d\times d}(\mathbb{C}^*) \to [0,\infty)$ by

$$\mathcal{N}_d(X) = \sum_{r,s=1}^{d} \left(\left| [X]_{r,s} \right|^2 - \frac{1}{d} \right)^2$$

Clearly, $\mathcal{N}_d(X) = 0$ if and only if X is flat. If X is assumed to be unitary (which it will be if $X = \Phi_d(Y)$ for some matrix Y), then $\mathcal{N}_d(X)$ can be thought of as a deviation of $\sqrt{d}X$ from being Hadamard.

Figure 2.2 shows the function $\mathcal{N}_d({\Phi_d}^n(X))$ as we iterate Φ_d starting with matrices in dimensions 6, 7, 8, and 9 whose entries were drawn uniformly at random from the complex unit circle. Note the strictly decreasing sequences of values which appear to be exponentially decaying in a neighborhood of 0. One way of interpreting this plot is to suggest that the sequence of unitary matrices $(\Phi_d(X), {\Phi_d}^2(X), \ldots, {\Phi_d}^n(X), \ldots)$ always approaches $\mathfrak{F}(d)$, and does so exponentially-fast within a neighborhood of $\mathfrak{F}(d)$. This consistent, qualitative

[4] Experiments in dimensions $d > 2$ suggest that Φ_d always converges to a matrix which is both flat and unitary, but generally not in a single iteration.

behavior begs for rigorous explanation. One might first try to show that $\mathcal{N}_d(\Phi_d{}^n(X))$ is strictly decreasing in n. If one can bound the size of $\mathcal{N}_d(\Phi_d{}^n(X))$ by an appropriately chosen function of n, it may follow immediately that Φ_d converges to $\mathfrak{F}(d) \cap \mathfrak{U}(d)$.

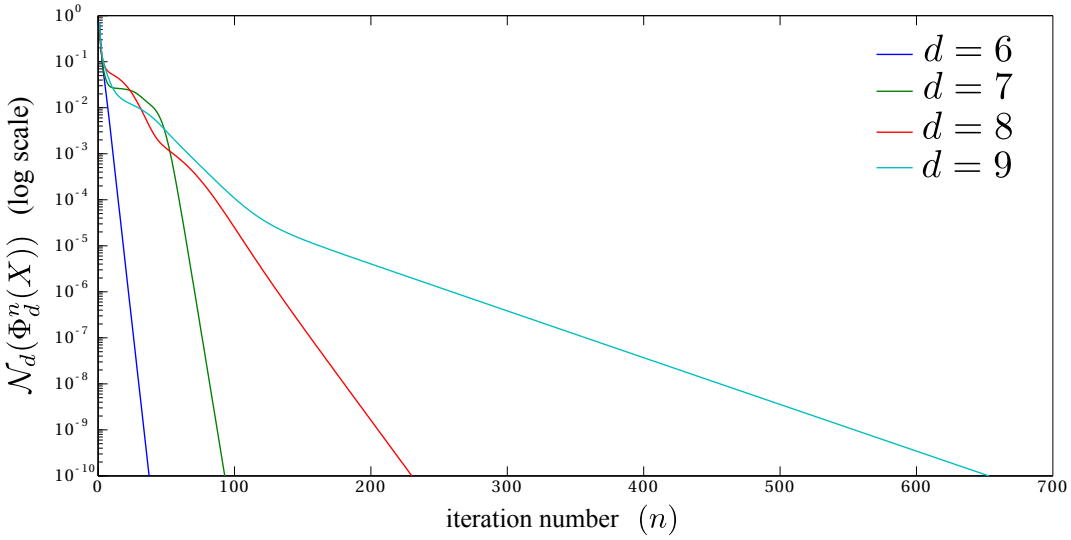

FIGURE 2.2. The function $\mathcal{N}_d(\Phi_d{}^n(X))$ for $n = 0, 1, \ldots, n_d$ for random starting matrices in dimensions $d = 6, 7, 8$ and 9. Here n_d is the smallest iteration, n, where $\mathcal{N}_d(\Phi_d{}^n(X)) < 10^{-10}$.

Towards proving the conjecture that $\mathcal{N}_d(\Phi_d{}^n(X)) \to 0$ as $n \to \infty$, we have derived several elementary propositions which may (or may not) help elicit a solution. The first is a simple, yet surprising consequence of viewing the flattening of a unitary matrix by the function $\mathcal{F}$ as a perturbation of its entries.

PROPOSITION 15. *Let*

$$U = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_d \end{bmatrix} \in \mathfrak{U}(d),$$

be any unitary matrix, where $\boldsymbol{u}_i = [u_{i,1}, \ldots, u_{i,d}]$ *is the* i*th row of* U*. Then*

$$\sum_{i,j=1}^{d} |u_{i,j}| \leq d\sqrt{d}.$$

PROOF. Express $\mathcal{F}(U) = U + E$ as the sum of U plus some perturbation matrix E. Solving for E we find

$$[E]_{i,j} = \begin{cases} u_{i,j}\left(\frac{1}{\sqrt{d}|u_{i,j}|} - 1\right), & \text{if } u_{i,j} \neq 0 \\ 0, & \text{if } u_{i,j} = 0. \end{cases}$$

Note that here we have extended the definition of the flattening function$\mathcal{F}$ to include the possibility of U having entries equal to 0. Let $\mathcal{D}$ be the linear indices $\{1, 2, \ldots, d^2\}$ of the elements of a $d \times d$ matrix produced by the indexing function $\sigma_d : \{1, \ldots, d\}^2 \to \{1, \ldots, d^2\}$ defined as

$$\sigma_d(i, j) = d(i - 1) + j.$$

Let $\mathcal{P} \doteq \{\sigma_d(i, j) | [E]_{i,j} \neq 0\}$ and $\mathcal{Z} \doteq \{\sigma_d(i, j) | [E]_{i,j} = 0\}$. Clearly $\mathcal{D} = \mathcal{P} \cup \mathcal{Z}$. We denote $[E]_{i,j}$ by $[E]_s$ if $s = \sigma_d(i, j)$. A straightforward calculation shows

$$\begin{aligned} \|E\|_F^2 &= \sum_{i,j=1}^{d} \left|[E]_{i,j}\right|^2 \\ &= \sum_{p \in \mathcal{P}} \left|[E]_p\right|^2 + \sum_{z \in \mathcal{Z}} \left|[E]_z\right|^2 \\ &= \sum_{p \in \mathcal{P}} \left|[E]_p\right|^2 \\ &= \sum_{p \in \mathcal{P}} \left| [U]_p \left(\frac{1}{\sqrt{d}\left|[U]_p\right|} - 1 \right) \right|^2 \end{aligned}$$

$$
\begin{aligned}
&= \sum_{p\in\mathcal{P}} \Big|[U]_p\Big|^2 \left(\frac{1}{d\Big|[U]_p\Big|^2} - \frac{2}{\sqrt{d}\Big|[U]_p\Big|} + 1 \right) \\
&= \sum_{i,j=1}^{d} \left(\frac{1}{d} - \frac{2|u_{i,j}|}{\sqrt{d}} + |u_{i,j}|^2 \right) \\
&= d^2\frac{1}{d} + \sum_{i,j=1}^{d} |u_{i,j}|^2 - \frac{2}{\sqrt{d}} \sum_{i,j=1}^{d} |u_{i,j}| \\
&= d + \sum_{i,j=1}^{d} \mathbf{u}_i \cdot \mathbf{u}_j - \frac{2}{\sqrt{d}} \sum_{i,j=1}^{d} |u_{i,j}|.
\end{aligned}
$$

Note that $\sum_{i,j=1}^{d} \mathbf{u}_i \cdot \mathbf{u}_j = d$ since $UU^* = I_d$ and thus

$$
\mathbf{u}_i \cdot \mathbf{u}_j = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}
$$

Since $0 \leq \|E\|_F^2$ it follows that

$$
\begin{aligned}
0 \leq \|E\|_F^2 &= 2d - \frac{2}{\sqrt{d}} \sum_{i,j=1}^{d} |u_{i,j}| \\
&\implies \frac{2}{\sqrt{d}} \sum_{i,j=1}^{d} |u_{i,j}| \leq 2d \\
&\implies \sum_{i,j=1}^{d} |u_{i,j}| \leq d\sqrt{d}.
\end{aligned}
$$

□

Note that the bound given in Proposition 15 is strict since flat, unitary matrices achieve it.

Because it appears that iterations of Φ_d monotonically approach the intersection of $\mathfrak{F}(d)$ and $\mathfrak{U}(d)$, we imagine it may be fruitful to consider the action of $\mathcal{U}_d$ on a flat matrix as a matrix perturbation, and likewise the action of $\mathcal{F}_d$ as a perturbation of a unitary matrix. At

least then we may be able to identify a neighborhood of $\mathfrak{F}(d) \cap \mathfrak{U}(d)$ where convergence is guaranteed. For now we simply state two theorems which bound the change in the singular values when one perturbs a matrix.

THEOREM 2.3.4 (Weyl [59], Mirsky [60]). *Let $X = U\Sigma V^*$ and $X + E = \tilde{U}\tilde{\Sigma}\tilde{V}^*$ be the singular value decompositions of $X, X + E \in M^{d\times d}(\mathbb{C})$, where $\Sigma = diag(\sigma_1, \ldots, \sigma_d)$ and $\tilde{\Sigma} = diag(\tilde{\sigma}_1, \ldots, \tilde{\sigma}_d)$. Then*

$$|\tilde{\sigma}_i - \sigma_i| \leq \|E\|_2$$

for each $i = 1, 2, \ldots, d$. Also,

$$\sqrt{\sum_{i=1}^{d} (\tilde{\sigma}_i - \sigma_i)^2} \leq \|E\|_F.$$

By considering $E_i \doteq \mathcal{F}({\Phi_d}^i(X)) - {\Phi_d}^i(X)$, we could prove convergence of Φ_d by showing that the perturbation matrices which correspond to flattening (E_i) shrink to the zero matrix. It is our hope that we can use the inequalities in Theorem 2.3.4 to prove this. It remains to be seen if the required effort is justified by the payoff.

2.3.2. GENERATING MUTUALLY UNBIASED BASES. By design, Φ_d fixes flat, unitary matrices. By applying this methodology to pairs of flat and unitary matrices, we utilize and extend Φ_d to an algorithm which generates a pair of mutually unbiased bases. Think of two matrices $A, B \in \mathfrak{F}(d) \cap \mathfrak{U}(d) \subset d \times d(\mathbb{C}^*)$ as flat, orthogonal bases of $\mathbb{C}^d$. It need not be the case that A and B are mutually unbiased. In other words, BA^* need not be flat and unitary.

However, if $\{A, B\}$ happens to be a MUB pair then

$$\Phi_d(BA^*) = BA^*,$$

by Lemma 2.3.2. More to the point, if A and B are mutually unbiased then

$$\Phi_d(\Phi_d(BA^*)A) = \Phi_d(BA^*A) = \Phi_d(B) = B.$$

Thus, the hope is that by allowing $B \in \mathfrak{F}(d) \cap \mathfrak{U}(d)$ to evolve under the influence of another flat, unitary matrix A (and A^*), it will converge to a matrix which is flat, unitary, and mutually unbiased with respect to A. Remarkably, this process has been very effective at generating pairs of mutually unbiased bases. In theory, this serialized procedure can be extended to any number of matrices by fixing $k-1$ MUBs and allowing a kth flat, unitary matrix to evolve under the influence of the first $k-1$ matrices (in some fixed order). Algorithm 2 provides pseudo-code of this serialized procedure. A related, "parallelized" process is given in Algorithm 3, in which all matrices in a set of fixed size are allowed to evolve simultaneously towards being mutually unbiased.

These algorithms are numerical in practice and so we are compelled to define the notion of matrix being numerically-Hadamard and a collection of matrices being numerically-MUB. So, following the example of [57], we will utilize a function whose value is 0 precisely when the input is a set of MUBs and positive otherwise: Let $k \geq 2$ and define $\mathfrak{N}_{d,k} : M^{d\times d}(\mathbb{C}^*)^k \to \mathbb{R}$ by

$$\mathfrak{N}_{d,k}(X_1, ..., X_k) = \sum_{1\leq n<m\leq k} \sum_{r,s=1}^{d} \left(\left| [X_n X_m^*]_{r,s} \right|^2 - \frac{1}{d} \right)^2$$

Note that $\mathfrak{N}_{d,2}(I_d, X) = \mathcal{N}_d(X)$.

DEFINITION 2.3.5. Fix $\epsilon > 0$. We will say that $X \in M^{d\times d}(\mathbb{C}^*)$ is **ϵ-Hadamard** if $\mathfrak{N}_{d,2}(I_d, X) \leq \epsilon$. Similarly, we will say that a collection of $d \times d$ unitary matrices $\{A_1, A_2, \ldots, A_k\}$ is **ϵ-mutually-unbiased** (ϵMUB) if $\mathfrak{N}_{d,k}(A_1, A_2, ..., A_k) \leq \epsilon$.[5]

With these definitions established let us formally introduce two algorithms for producing ϵMUBs:

Algorithm 2: Serialized algorithm for producing MUBs

Data: **1.** A collection of ϵ-Hadamard matrices: $A_1, \ldots, A_{k-2}, X$ such that $\{I_d, A_1, \ldots, A_{k-2}\}$ is a set of ϵ-MUBs.
2. Stopping criteria: δ_1, and δ_2.

Result: A δ_2-Hadamard matrix X such that $\{I_d, A_1, \ldots, A_{k-2}, X\}$ is a set of δ_1-MUBs (if possible).

```
while 𝔑_{d,k}(I_d, A_1, ..., A_{k-2}, X) > δ_1 do
    for i = 1 to k − 2 do
        X = X A_i^*;
        /* Iterate Φ_d on X until convergence to a δ_2-Hadamard.   */
        while 𝔑_{d,2}(I_d, X) > δ_2 do
            X = Φ_d(X);
        X = X A_i;
        while 𝔑_{d,2}(I_d, X) > δ_2 do
            X = Φ_d(X);
```

In the following section we develop several experiments which utilize these algorithms to generate and analyze numerical Hadamards and numerical MUBs.

2.3.3. NUMERICAL EVIDENCE OF A 4-DIMENSIONAL MANIFOLD IN DIMENSION 6. Numerical convergence of Φ_d depends on many factors, not the least of which is that it is defined in terms of the singular value decomposition. Theoretically we can think of $\mathcal{U}$ as a sort of projection onto the space of unitary matrices since it maps a matrix to its nearest unitary matrix. However, in practice, the precision and stability of one's chosen singular value decomposition algorithm cannot be ignored and will greatly influence the rate of convergence

[5] In experiments, one uses ϵ as a stopping criterion. The choice will depend on the dimension d, the numerical precision of one's computations, and of course the desired accuracy of the numerical matrices.

Algorithm 3: Parallelized algorithm for producing MUBs

Data: **1.** A collection of ϵ-Hadamard matrices: $A_1, \ldots, A_{k-1}$.
2. Stopping criteria: δ_1 and δ_2

Result: A collection of $k-1$, δ_2-Hadamard matrices $(A_1, \ldots, A_{k-1})$ such that $\{I_d, A_1, \ldots, A_{k-1}\}$ is a set of δ_1-MUBs (if possible).

```
while 𝔑_{d,k}(I_d, A_1, ..., A_{k-1}) > δ_1 do
    for i = 1 to k − 1 do
        for j = 1 to k − 1, j ≠ i do
            A_i = A_i A_j^*;
            /* Iterate Φ_d on A_i until convergence to a δ_2-Hadamard.   */
            while 𝔑_{d,2}(I_d, A_i) > δ_2 do
                A_i = Φ_d(A_i);
            A_i = A_i A_j;
            while 𝔑_{d,2}(I_d, A_i) > δ_2 do
                A_i = Φ_d(A_i);
```

to (and final accuracy of) one's ϵ-Hadamards. There are many tantalizing questions about the numerical convergence of Φ_d which will, undoubtedly, require an investigation into numerical SVD methods. Also, it is clear that if $k \in \mathbb{N}$ exceeds $\mathcal{M}(d)$, then neither Algorithm 2 nor Algorithm 3 has any hope of converging to a set of MUBs of size k. It is our hope that understanding the dynamics of these systems, when used to try and generate a set of MUBs of size k (where k exceeds the theoretical maximum) will help reveal why a full set of MUBs cannot exist in non-prime-power dimensions, if this is indeed the case.

We do not concern ourselves with these questions here. Instead, we simply use our discrete-system to generate dephased Hadamards and MUBs in small dimensions where we are comfortable with the accuracy of our numerical matrices. In particular, we focus our efforts on dimension-6 – the smallest composite dimension where classification of complex Hadamards remains incomplete and a full set of MUBs is not known to exist. We utilize a standard statistical procedure, principal component analysis (PCA) using the singular-value

decomposition (SVD), to study samples of dephased Hadamards in neighborhoods of fixed matrices. For a detailed description of PCA using the SVD see [61].

As previously stated, there exist several papers in the literature which provide numerical evidence of a 4-parameter family of 6×6 Hadamards [53], [57]. Ultimately, this section contributes new numerical evidence in support of existing conjectures by applying a novel method for generating Hadamards and statistical techniques not previously seen in the literature.

Many of the construction methods which have produced, 1, 2 or 3 dimensional families in $\mathcal{H}_6$ (usually parametrized by closed-form equations involving well-known trigonometric functions) impose strong assumptions on the structure of the matrix such as symmetry, being self-adjoint, a circulant block structure or H_2-reducibility [39]. These simplify the problem, restricting to subvarieties, allowing algebraic conditions to be solved explicitly. One such family, which we refer to as Karlsson's family and denote $K_6^{(3)}$ (see [49]), will be of particular interest to us.

$$K_6^{(3)}(\theta,\phi,\psi) \doteq \begin{bmatrix} F_2 & Z_1 & Z_2 \\ Z_3 & \frac{1}{2}Z_3AZ_1 & \frac{1}{2}Z_3BZ_2 \\ Z_4 & \frac{1}{2}Z_4BZ_1 & \frac{1}{2}Z_4AZ_2 \end{bmatrix},$$

where $A = \begin{bmatrix} A_{11} & A_{12} \\ \overline{A_{12}} & -\overline{A_{11}} \end{bmatrix}$, defining

$$A_{11} \doteq -\frac{1}{2} + \mathrm{i}\frac{\sqrt{3}}{2}[\cos(\theta) + e^{-\mathrm{i}\phi}\sin(\theta)] \text{ and}$$
$$A_{12} \doteq -\frac{1}{2} + \mathrm{i}\frac{\sqrt{3}}{2}[-\cos(\theta) + e^{\mathrm{i}\phi}\sin(\theta)],$$

for $\theta, \phi \in [0, \pi)$ and $B \doteq -F_2 - A$. For some $\psi \in [0, \pi)$ further define $z_1 \doteq e^{i\psi}$ and build

$$Z_1 \doteq \begin{bmatrix} 1 & 1 \\ z_1 & -z_1 \end{bmatrix}, \quad Z_2 \doteq \begin{bmatrix} 1 & 1 \\ z_2 & -z_2 \end{bmatrix}, \quad Z_3 \doteq \begin{bmatrix} 1 & z_3 \\ 1 & -z_3 \end{bmatrix}, \quad Z_4 \doteq \begin{bmatrix} 1 & z_4 \\ 1 & -z_4 \end{bmatrix},$$

where $z_3^2 = \dfrac{\alpha_A z_1^2 - \beta_A}{\overline{\beta_A} z_1^2 - \overline{\alpha_A}}$, $z_2^2 = \dfrac{\overline{\alpha_B} z_3^2 - \beta_B}{\overline{\beta_B} z_3^2 - \alpha_B}$ and $z_4^2 = \dfrac{\alpha_B z_1^2 - \beta_B}{\overline{\beta_B} z_1^2 - \overline{\alpha_B}}$ for $\alpha_A = A_{12}^2$, $\beta_A = A_{11}^2$ and $\alpha_B = B_{12}^2$, $\beta_B = B_{11}^2$.

There are several reasons for our interest in $K_6^{(3)}$: it contains all other known 1 and 2 dimensional families and members of this family have the strikingly beautiful structure that all 2×2 submatrices are themselves Hadamard (H_2-reducibility), which may be relevant to understanding MUBs in dimension six since all known MUB triplets are H_2-reducible. Lastly it is not known if $K_6^{(3)}$ is a subfamily of a 4-parameter family.

In a very recent paper, Szöllősi's presents a method of constructing inequivalent Hadamards with four degrees of freedom resulting in a set of inequivalent Hadamards called $G_6^{(4)}$ [39]. His approach is general in that it does not impose any restrictions on the form of the dephased matrix. Instead he starts with a 3×3 submatrix,

$$E \doteq \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & b \\ 1 & c & d \end{bmatrix}, \tag{10}$$

and asks when E can be extended to (embedded in) a dimension-6 dephased Hadamard matrix,

$$H = \begin{bmatrix} E & B \\ C & D \end{bmatrix},$$

for 3×3 matrices

$$B = \begin{bmatrix} 1 & 1 & 1 \\ e & s_1 & s_2 \\ f & s_3 & s_4 \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & g & h \\ 1 & t_1 & t_3 \\ 1 & t_2 & t_4 \end{bmatrix},$$

and

$$D \doteq -CE^* \left(B^{-1}\right)^*.$$

Provided certain conditions on the matrix E are satisfied, the variables e, f, g, h and $s_k, t_k (k = 1, 2, 3, 4)$ are shown to depend on the parameters a, b, c, and d through carefully chosen solutions to several sextic polynomials. In general B may not be invertible and a characterization of Hadamards with vanishing 3×3 minors excludes $K_6^{(3)}$ and $S_6^{(0)}$ from consideration.

Although the appearance of degree 6 polynomials suggests a parametrization in terms of radicals is not possible, Szöllősi's conjectures that he suspects that any complex Hadamard matrix in dimension 6 (except perhaps $S_6^{(0)}$ and $K_6^{(3)}$) can be recovered from his method of construction. The following steps outline our numerical investigation of this proposal.

Step 1: Starting with a 6×6 matrix whose entries are uniformly taken from the complex unit circle using Matlab's built-in, pseudo-random number generator **rand()**, we apply our DTDS Φ_6 until convergence to a numerical Hadamard matrix is achieved. Call the resulting matrix M.

Step 2: We extract from M the entries $a \doteq [M]_{2,2}$, $b \doteq [M]_{2,3}$, $c \doteq [M]_{3,2}$, and $d \doteq [M]_{3,3}$ to define the submatrix E, Equation 10.

Step 3: We attempt to embed the matrix E into a complex Hadamard matrix and check if the resulting matrix is equivalent to M. If so, we classify M as of the type able to be generated by the method of Szöllősi's and then repeat, starting with **Step 1**. If not we move on to **Step 4**.

Step 4: We check if M is equivalent to $S_6^{(0)}$ by comparing $P_i M P_j$ to $S_6^{(0)}$, for all 720 different 6×6 permutation matrices, P_i and P_j. If M is permutation equivalent to $S_6^{(0)}$ we classify it and begin again with **Step 1**. If not we move on to **Step 5**.

Step 5: We finally check if M is equivalent to an element of $K_6^{(3)}$. If so, we classify M and repeat, starting with **Step 1**.

The above procedure was repeated for 5741 "randomly" generated starting matrices. Of these, 1652 (nearly 29%), were found to be permutation equivalent to $S_6^{(0)}$. The remaining 4089 were found to be reproducible by Szöllősi's method of construction.

This experiment suggests, among other things, that there exist large basins of attraction for the permutations of $S_6^{(0)}$. Also, any other isolated matrices (if they exist) must have relatively small basins of attraction or may be unstable fixed points of our system. Naturally this supports Szöllősi's conjecture as no example was found, other than $S_6^{(0)}$, which was not a member of $G_6^{(4)}$.

It is curious that no member of $K_6^{(3)}$ was found. Towards understanding why we consider a perturbation K_ϵ of a member $K \in K_6^{(3)}$, where the entries of K_ϵ do not deviate from K by more than .01. Iterations of Φ applied to $1/\sqrt{6}K_\epsilon$ converge to a nearby, numerically-Hadamard matrix which is not found to not be a member of $K_6^{(3)}$. This suggests $K_6^{(3)}$ is a subfamily of a larger one. This observation, and this technique of perturbing and re-Hadamarizing a matrix leads us to our next experiment.

We have implemented in Matlab the following steps to produce a sample of the space dephased Hadamards near a fixed matrix.

Step 1: Start with a fixed 6×6 dephased matrix M either by choosing an element of a known Hadamard family or by producing one at random.

Step 2: Perturb the core of M, entry-by-entry, by choosing a 5×5 matrix of complex entries taken uniformly from the complex disk of some small radius, ϵ.

Step 3: Since this perturbed matrix is most likely no-longer Hadamard, apply Φ_6 until convergence to a numerical Hadamard occurs.

Step 4: Repeat steps 1 through 3 many times to obtain a sample of the space of complex Hadamards.

Step 5: Treat each member of the sample as an element of $\mathbb{C}^{25}$ by treating the core as a column vector.

Step 6: Mean-center the sample.

Step 7: Apply the singular value decomposition to the mean-centered sample and count the number of non-zero singular values.

In this experiment we are hopeful that the sample will closely resemble the tangent space at the matrix M of the space of *inequivalent*, dephased Hadamards, provided that we choose our maximum perturbation modulus small enough, since the non-zero singular values correspond to the left-singular vectors which span the range of the decomposed matrix (in our case, the sample of Hadamards). Thus the left-singular vectors corresponding to the non-zero singular values can be taken to be numerical approximations of tangent vectors to the space at M, provided that the tangent space is well defined at M. We must be careful to observe that we are not accounting for permutation equivalence with this sampling procedure. Thus we cannot immediately rule-out the possibility that the space near M is not a manifold at all,

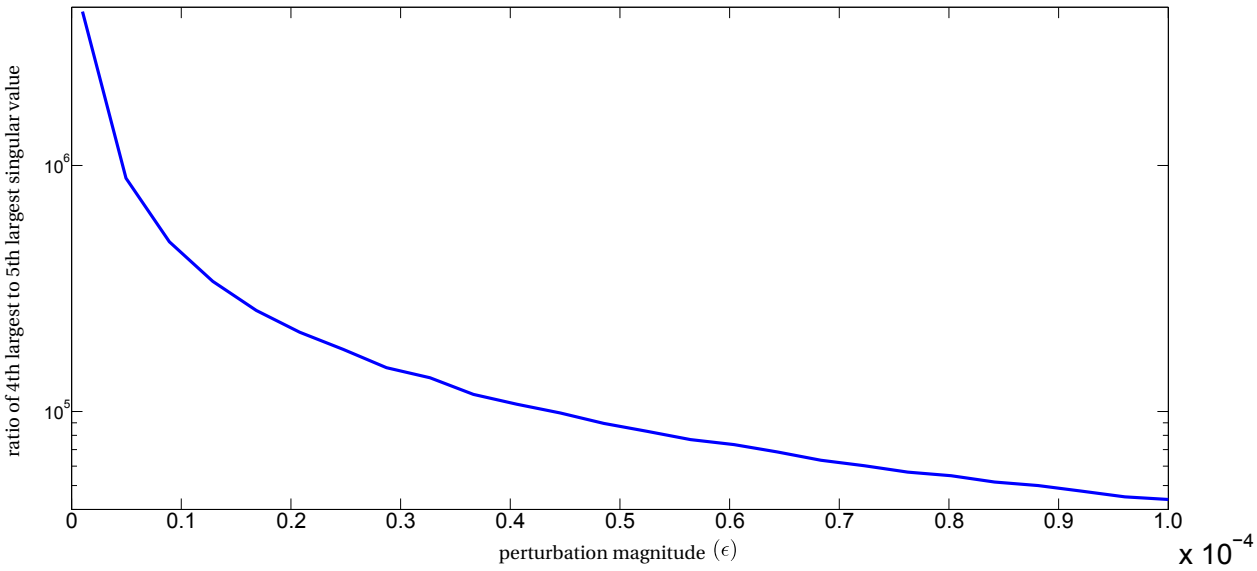

FIGURE 2.3. For each of 25 choices of ϵ (equally spaced between 10^{-6} and 10^{-4}), 10,000 numerical Hadamards were produced by perturbing the entries of a randomly chosen member of $F_6^{(2)}$ by a complex number of modulus no more than ϵ. The SVD was applied to each sample and the ratio of the 4th to the 5th largest singular value is shown.

but an intersection of 'permutation-equivalent' manifolds. In short, the number of nonzero singular values derived from PCA may be in excess of the local dimension of the space of inequivalent, dephased Hadamards.

For each of several randomly chosen members of known Hadamard families in dimension-6 (excluding $S_6^{(0)}$) the result of PCA is always the same: 4 non-zero singular values. Because our samples contain noise and live on a non-linear manifold, none of the singular values $(\sigma_i, i = 1, \ldots, 25)$ derived by this procedure are precisely zero. Thus we consider a sudden jump in the ratio of the nth singular value to the $(n+1)$th (ordered by decreasing magnitude) to indicate that $\sigma_k > 0$ for $k = 1, \ldots, n$, while $\sigma_j = 0$ for $j = n+1, \ldots, 25$. Figure 2.3 illustrates how the ratio of the 4th singular value to the 5th singular value grows rapidly as the maximum perturbation magnitude is reduced towards 0. This is the expected behavior as we anticipate the sample becoming a better approximation of the tangent space as ϵ is reduced.

The same qualitative picture emerges for members of each of the known families, including $K_6^{(3)}$, as well as all "randomly" generated Hadamards not equivalent to $S_6^{(0)}$. In fact, the only Hadamard we have found which does not give rise to four non-zero singular values is $S_6^{(0)}$, which is known to be isolated. As expected, applying the above procedure to $S_6^{(0)}$ results in 25 vanishing singular values. In other words, and more to the point, small perturbations of $S_6^{(0)}$ converge under Φ_6 to $S_6^{(0)}$.

To summarize, we sampled the space of inequivalent complex Hadamard matrices in a small region about a fixed matrix. We then applied PCA to this sample to determine the local dimension of the space of dephased Hadamards. For all randomly chosen complex Hadamards, and all randomly chosen members of known families, the number of nonzero singular values is always found to be 4 or 0. In the latter case the Hadamard is always found to be equivalent to Tao's isolated matrix. We go further by randomly sampling the space of complex Hadamards in dimension 6 and test their membership in known families. Up to equivalence, we only ever find Tao's matrix and matrices which can be constructed according to the method of Ferenc Szöllősi. This provides strong numerical evidence of the conjecture that these are all the complex Hadamards in dimension 6.

We now briefly mention our observations regarding MUBs in dimensions 2 through 10. As outlined in Section 2.3, the function Φ_d can be extended to produce "random" MUBs of a specified size. These algorithms have been very successful in creating MUBs of various sizes in small dimensions; however, the success of the algorithm is not absolute. For small prime-power dimensions ($d < 10$) the algorithm often converges to MUBs of maximum size $(d+1)$, while for non-prime power dimensions, only MUBs whose size is equal to the known lower bound ($d = p_1^{e_1}$) have been produced. In other words, up to this date, our algorithm merely supports existing conjectures that the maximal size of a set of MUBs in dimension

$d = p_1^{e_1} \dots p_k^{e_k}$ is this lower bound. In particular, we have not found a numerical set of MUBs of size 4 in dimension 6.

Many interesting (though not at all well-understood) behaviors of the discrete system constructed in Section 2.3 have been observed and demand further exploration. When one attempts to produce a set of MUBs of size 4 in dimension 6, the algorithm behaves in a manner consistent with the outcome of attempting to produce a set of 3×3 MUBs of size 5, or a set of size 6 in dimension 4, which are both known to not exist. We believe an explanation of these behaviors can be made rigorous and may elucidate, among other things, why some dimensions support full sets of MUBs while others (apparently) do not.

Another compelling observation is that the serialized algorithm for producing MUBs in dimension 6 appears to be unable to generate MUB-triplets, in contrast to the parallelized approach, which has proven quite capable of doing so. Recall the evidence that $K_6^{(3)}$ is contained in a larger 4-parameter family. If this is true, then we do not expect general numerical methods to converge to members of $K_6^{(3)}$, as it would then be a measure-0 submanifold of a more general family. This would explain the failing of the serialized algorithm to extend any set $\{I_6, H\}$ $(H \in \mathcal{H}_6)$ to a MUB-triple since it is believed that 6×6 MUH pairs are, in fact, comprised entirely of H_2-reducible Hadamards.

Ultimately there remain many unanswered questions and avenues of research, both experimental and theoretical, concerning the discrete system Φ_d. We hope we have laid the groundwork here for its continued use in the pursuit of understanding complex Hadamards and mutually unbiased bases.

2.4. Hadamard Fixed Points of a Continuous Dynamical System

In the preceding section we introduced discrete-time dynamical systems which converge to Hadamard matrices and mutually unbiased bases. It is Φ_d's discrete nature which endows it, paradoxically, with simplicity of form yet makes it difficult to analyze. We transition in this section to the use of dynamics that struggle less with these attributes. By the end of this chapter, we will have established new formalism which provides, among other things, the opportunity to improve upon the defect's estimate of the local dimension of dephased Hadamards. As a specific consequence, we construct a novel method of proving that a Hadamard is isolated and use it to establish the first-known example of an isolated Hadamard with a positive defect.

Fix $d \geq 2$ and consider a dephased matrix of the form

$$H_d(\boldsymbol{\theta}) \doteq \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & e^{\mathrm{i}\theta_1} & e^{\mathrm{i}\theta_2} & \cdots & e^{\mathrm{i}\theta_{d-1}} \\ 1 & e^{\mathrm{i}\theta_d} & e^{\mathrm{i}\theta_{d+1}} & \cdots & e^{\mathrm{i}\theta_{2(d-1)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e^{\mathrm{i}\theta_{(d-2)(d-1)+1}} & e^{\mathrm{i}\theta_{(d-2)(d-1)+2}} & \cdots & e^{\mathrm{i}\theta_{(d-1)^2}} \end{bmatrix}.$$

with $\boldsymbol{\theta} \doteq \left[\theta_1, \theta_2, \ldots, \theta_{(d-1)^2}\right]$ and $\theta_i \in [0, 2\pi)$. If $H_d(\boldsymbol{\theta})$ is a dephased Hadamard, then by definition $H_d(\boldsymbol{\theta})H_d(\boldsymbol{\theta})^* = dI_d$. Naturally this imposes conditions on the allowed phases in the entries of $H_d(\boldsymbol{\theta})$. In particular, $\boldsymbol{\theta}$ must be chosen to satisfy $d(d-1)$ equations stemming from the requirement that the off-diagonal entries of $H_d(\boldsymbol{\theta})H_d(\boldsymbol{\theta})^*$ must be identically 0:

$$[H_d(\boldsymbol{\theta})H_d(\boldsymbol{\theta})^*]_{i,j} = \begin{cases} 0, & \text{if } i \neq j \\ d, & \text{if } i = j \end{cases}. \tag{11}$$

Note that the conditions given in Equation 11, along with the choice that $H_d(\boldsymbol{\theta})$ is dephased, are the same as those expressed in the system 7, described in Section 2.2:

$$\begin{aligned} R_{i,1} &= 0, \text{ for } 1 \leq i \leq d \\ R_{1,j} &= 0, \text{ for } 2 \leq j \leq d \\ \sum_{k=1}^{d} [H]_{i,k}[H^*]_{j,k} e^{\mathrm{i}\left([R]_{i,k} - [R]_{j,k}\right)} &= 0, \text{ for } 1 \leq i < j \leq d, \end{aligned}$$

since the first $2d-1$ equations of 7 force $H \circ \exp(\mathrm{i}R)$ to be dephased, while the last $d(d-1)$ equations enforce the unitary condition.

Recall that the dimension of the linearization of equations 7 at a Hadamard, H, was called the defect of H. Our aim is to use dynamics to improve upon the estimation of the local dimension of dephased Hadamards made by the defect. Toward that end, we use the equations in 11 to define a scalar potential $\mathcal{V}_d : \mathbb{R}^{(d-1)^2} \to \mathbb{R}$, which can be thought of as measuring the extent of the failure of a matrix to be Hadamard,

$$\mathcal{V}_d(\boldsymbol{\theta}) \doteq \sum_{i \neq j}^{d} \left| [H_d(\boldsymbol{\theta}) H_d(\boldsymbol{\theta})^*]_{i,j} \right|^2, \tag{12}$$

and which vanish exactly when $H_d(\boldsymbol{\theta})$ is Hadamard. By computing the negative gradient of $\mathcal{V}_d$, we define a gradient system of ordinary differential equations,

$$\Phi_d(\boldsymbol{\theta}) \doteq -\nabla \mathcal{V}_d, \tag{13}$$

whose stationary points are the dephased complex Hadamard matrices.

Equation 13 can be thought of as defining a flow on the $(d-1)^2$-torus, $\mathbf{T}^{(d-1)^2}$, since we insist the matrices $H_d(\boldsymbol{\theta})$ be dephased. This restriction ensures that any set of non-isolated

fixed points sufficiently close to a Hadamard, H (to which H belongs) will be inequivalent to H, since any permutation of the rows and columns of the core of H must yield an equivalent Hadamard some non-zero distance from H, and because the dephasing matrices D_c and D_r are unique. That being said, this system does not take into account permutation equivalences, and so the set to which $\Phi_d(\boldsymbol{\theta})$ converges is not the space of inequivalent Hadamards, but rather the space of inequivalent Hadamards together with all copies of this space derived from core permutations of its members. This is a subtle and important distinction which we will address in Section 2.4.2. For now, we simply observe that at a dephased Hadamard the local structure of the space to which Φ_d converges is the local structure of the space of inequivalent Hadamards up to permutation equivalence.

As we have seen, there exists a significant collection of known manifolds of inequivalent complex Hadamard matrices. To rephrase this observation in the language presented here: we know that for certain integers d, $\Phi_d(\boldsymbol{\theta})$ will necessarily have non-isolated (degenerate) fixed points. Our goal is to exploit well-established theories of continuous dynamical systems to derive estimates of the dimensions of these spaces of fixed points.

Imagine a smooth vector field on $\mathbb{R}^N$ given by

$$\dot{\mathbf{x}} = \frac{d\mathbf{x}}{dt} = f(\mathbf{x})$$

with a k-dimensional ($k < N$) manifold of degenerate fixed points, $\mathcal{M}$. A small perturbation of a fixed point, $\mathbf{p} \in \mathcal{M}$ may land on a nearby fixed point, $\mathbf{x}(0) \in \mathcal{M}$, in which case the value $\mathbf{x}(t)$ will remain fixed for all time t. More likely, a random perturbation will not remain in $\mathcal{M}$, in which case $\mathbf{x}(t)$ will change in a manner determined by the flow defined by f. In

the case of f being gradient, any perturbation off a fixed point will result in negative flow towards a fixed point, perhaps different from $\mathbf{p}$.

In the same manner as for an isolated fixed point, one can linearize at the degenerate fixed point $\mathbf{p}$ and study the linear behavior of the nearby flow. Naturally there is no linear contribution to the flow on $\mathcal{M}$ since there is no flow on $\mathcal{M}$ at all! Thus, we expect no linear flow in the tangent space to $\mathcal{M}$ at $\mathbf{p}$. Said another way, if there *is* linear flow in some direction, $\mathbf{v} \in \mathbb{R}^N$, then $\mathbf{v} \notin T_{\mathbf{p}}(\mathcal{M})$. If we could detect those directions along which there is no linear flow, we would have an immediate upper bound for the dimension of $\mathcal{M}$. Thankfully there is a well-established theorem which will allow us to do just this and much more.

Let $\mathbf{0}$ be a fixed point of the nonlinear system

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x} + F(\mathbf{x}), \mathbf{x} \in \mathbb{R}^N \tag{14}$$

where A is an $N \times N$ matrix and F is C^r in a neighborhood of $\mathbf{0}$. Then it is known that the eigenvalues of A can be used to determine the local dynamics of the system near the origin. In particular, there exists stable (E^s), unstable (E^u), and center (E^c) subspaces spanned by the eigenvectors corresponding to the eigenvalues of A with negative, positive, and 0 real parts respectively. The generalized eigenvectors spanning the stable and unstable subspaces are tangent at the origin to the stable and unstable manifolds (W^s and W^u): invariant sets with consistent asymptotic behavior with respect to the origin. Similarly the center subspace is tangent to every center manifold, W^c: invariant sets on which $F(\mathbf{x})$ (the non-linear part of the vector field) governs the dynamics.

THEOREM 2.4.1 (Local Center Manifold Theorem). *Assume A has c eigenvalues with real part equal to zero and $N-c$ eigenvalues with negative real part. Then the system defined by Equation 14 can be written in diagonal form*

$$\frac{d\boldsymbol{x}}{dt} = C\boldsymbol{x} + G(\boldsymbol{x}, \boldsymbol{y})$$
$$\frac{d\boldsymbol{y}}{dt} = P\boldsymbol{y} + H(\boldsymbol{x}, \boldsymbol{y})$$

where $\boldsymbol{x} \in \mathbb{R}^c, \boldsymbol{y} \in \mathbb{R}^{N-c}$, C is a square matrix whose eigenvalues all have 0 real part, P is a square matrix whose eigenvalues have negative real part, and $G(\boldsymbol{0}) = H(\boldsymbol{0}) = DG(\boldsymbol{0}) = DH(\boldsymbol{0}) = \boldsymbol{0}$. Furthermore there exists $\delta > 0$ and $h \in C^r(B_\delta(\boldsymbol{0}))$ that defines the local center manifold $W^c(\boldsymbol{0}) = \{[\boldsymbol{x}, \boldsymbol{y}] \in \mathbb{R}^N | \boldsymbol{y} = h(\boldsymbol{x})$ for $|\boldsymbol{x}| < \delta\}$ and satisfies

$$Dh(\boldsymbol{x})[C\boldsymbol{x} + F(\boldsymbol{x}, h(\boldsymbol{x}))] = Ph(\boldsymbol{x}) + G(\boldsymbol{x}, h(\boldsymbol{x}))]$$

for $|\boldsymbol{x}| < \delta$. The flow on the center manifold is governed by the system

$$\frac{d\boldsymbol{x}}{dt} = C\boldsymbol{x} + F(\boldsymbol{x}, h(\boldsymbol{x}))$$

for all $\boldsymbol{x} \in \mathbb{R}^c$ with $|\boldsymbol{x}| < \delta$.

For a more complete treatment see [62] or another text on non-linear dynamics. Note that the linearized system given in Theorem 2.4.1 has no eigenvalues with positive real part, and thus no unstable subspace or manifold. This assumption is not required but is appropriate for us since $\Phi_d(\boldsymbol{\theta})$ is a gradient flow.

Theorem 2.4.1 provides a means to bound the local dimension of a manifold of fixed points, $\mathcal{M}$, containing $\mathbf{p}$ since the center manifold at $\mathbf{p}$ must contain $\mathcal{M}$. Thus, by applying 2.4.1 to the gradient system 12 we can estimate the local dimension of the space of dephased (possibly permutation-equivalent) complex Hadamards at $H \in \mathcal{H}_d$ by computing the dimension of the center manifold at H. It is not surprising that the dimension of the center manifold at a matrix $H \in \mathcal{H}(d)$ is actually equal to the defect $d(H)$ since we are essentially linearizing the Hadamard conditions in both cases. For example, recall that if the defect of a Hadamard matrix is 0, then it must be isolated. It is also plainly seen that if there does not exist a center subspace of Φ_d at $H_d(\boldsymbol{\theta})$, then the stable subspace is $(d-1)^2$-dimensional and all points sufficiently close to $H_d(\boldsymbol{\theta})$ must flow to it.

Since a center manifold is not necessarily (and not usually) comprised entirely of fixed points, Theorem 2.4.1 can be used to improve-on the initial overestimate of the local dimension of the space of complex Hadamards predicted by the the dimension of the center subspace and the defect. If $\Phi_d(\boldsymbol{\theta}) \neq 0$, then $H_d(\boldsymbol{\theta})$ is not a complex Hadamard matrix and therefore any flow on the center manifold, as slow as it may be, will shrink the bound given by the defect.

A center-manifold reduction is usually done by performing a change of coordinates on the system into eigen-coordinates so that the center manifold can be written as a graph over the center subspace. This becomes impractical for a high-dimensional systems, even with the aid of a computer algebra system, since the process first requires diagonalization of (then large) symbolic matrices. We will instead follow an alternative method in which the center subspace is written as an embedding over the center subspace [63]. Because this process is essentially a Taylor expansion of the center manifold by repeated differentiation of the vector field, the primary computational limitation is the memory requirements of storing high-order tensors,

which are required if one wishes to expand the manifold to high order. Since one of our present purposes in performing a center-manifold reduction is to identify flow on the center manifold (rather than well-approximate the manifold itself) we do not anticipate the need for high-order approximations. Thus, this coordinate-independent reduction is advantageous.

In [63], the coordinate-independent reduction is derived for the special case when the center subspace is one-dimensional. As noted in the original paper, the reduction process generalizes to a center subspace of any dimension. That being said, there are a few subtleties which separate the cases where the dimension is greater than one from the case where the center subspace is a line. In the following section we closely follow [63] and derive the coordinate-independent reduction for a two-dimensional center manifold.

2.4.1. Coordinate-Independent Center Manifold Reduction. We begin by fixing our notation. Throughout this section we will interchangeably denote the partial derivative of a scalar function F with respect to the variable x by $\frac{\partial F}{\partial x}$ or F_x, choosing one over the other as it assists clarity of the expression. Borrowing the notation in [63] we will denote the directional derivative of a vector field $f : \mathbb{R}^N \to \mathbb{R}^N$ ($f = [f_1, f_2, \ldots, f_N]$, where each f_i is a smooth, scalar-valued map on $\mathbb{R}^N$) in the direction of a vector $\mathbf{v}$ by $Df[\mathbf{v}]$. This is nothing more than standard multiplication of the $N \times N$ Jacobian matrix Df with a length-N column vector $\mathbf{v}$. Higher order derivatives of f will be expressed similarly: $D^2 f[\mathbf{v}_1, \mathbf{v}_2]$ is the action of the second derivative tensor $D^2 f$ on the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. For any positive integer k, $D^k f[\mathbf{v}_1, \ldots, \mathbf{v}_k]$ is a length N column vector whose ith component is

$$\left[D^k f[\mathbf{v}_1, \ldots, \mathbf{v}_k]\right]_i \doteq \sum_{j_1,\ldots,j_k} \frac{\partial^k f_i}{\partial x_{j_1} \ldots \partial x_{j_k}} [v_1]_{j_1} [v_2]_{j_2} \cdots [v_k]_{j_k}$$

In [63], it is assumed that the center subspace at the origin is the kernel of the Jacobian there: The eigenvalues with zero real-part are, in fact, identically zero. In this case the center manifold is sometimes referred to as the slow manifold. Furthermore, it is assumed that there is no unstable subspace, meaning all eigenvalues of $Df_{\mathbf{0}}$ are nonpositive. The latter condition is not essential to the derivation but, because our system is gradient, all eigenvalues will be nonpositive and real. The latter is a consequence of the fact that the Jacobian is necessarily symmetric for a gradient system.

Let $f : \mathbb{R}^N \to \mathbb{R}^N$ define a smooth flow $\dot{\mathbf{x}} = \dfrac{d\mathbf{x}}{dt} = f(\mathbf{x})$. We assume the origin is fixed, $(f(\mathbf{0}) = \mathbf{0})$ and the center subspace at $\mathbf{0}$ is

$$E^c \doteq \{a\mathbf{v}_1 + b\mathbf{v}_2 | a, b \in \mathbb{R}\},$$

the subspace spanned by $\mathbf{v}_1$ and $\mathbf{v}_2$. We parametrize the center manifold, W^c, as follows: Define $\mathbf{X} : \mathbb{R}^2 \to \mathbb{R}^N$ by

$$\mathbf{X}(t_1, t_2) = t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \mathbf{w}(t_1, t_2),$$

where $\mathbf{w} : \mathbb{R}^2 \to \mathbb{R}^N$ is the embedding function chosen in the orthogonal complement of E^c. This choice will lead to a particular parametrization of the center manifold and allow us to expand $\mathbf{w}$ in a Taylor series about the origin. In particular, we aim to compute the power series

$$\mathbf{w}(t_1, t_2) = \frac{1}{2}\frac{\partial^2 \mathbf{w}}{\partial t_1^2}\bigg|_{(0,0)} t_1^2 + \frac{1}{2}\frac{\partial^2 \mathbf{w}}{\partial t_2^2}\bigg|_{(0,0)} t_2^2 + \frac{\partial^2 \mathbf{w}}{\partial t_1 \partial t_2}\bigg|_{(0,0)} t_1 t_2 + \text{h.o.t.},$$

subject to the conditions $\mathbf{w}(t_1, t_2) \perp \mathbf{v}_1$ and $\mathbf{w}(t_1, t_2) \perp \mathbf{v}_2$. Note that the embedding function must satisfy the boundary conditions $\mathbf{w}(0, 0) = \mathbf{w}_{t_1}(0, 0) = \mathbf{w}_{t_2}(0, 0) = \mathbf{0}$. This embedding is illustrated in Figure 2.4.

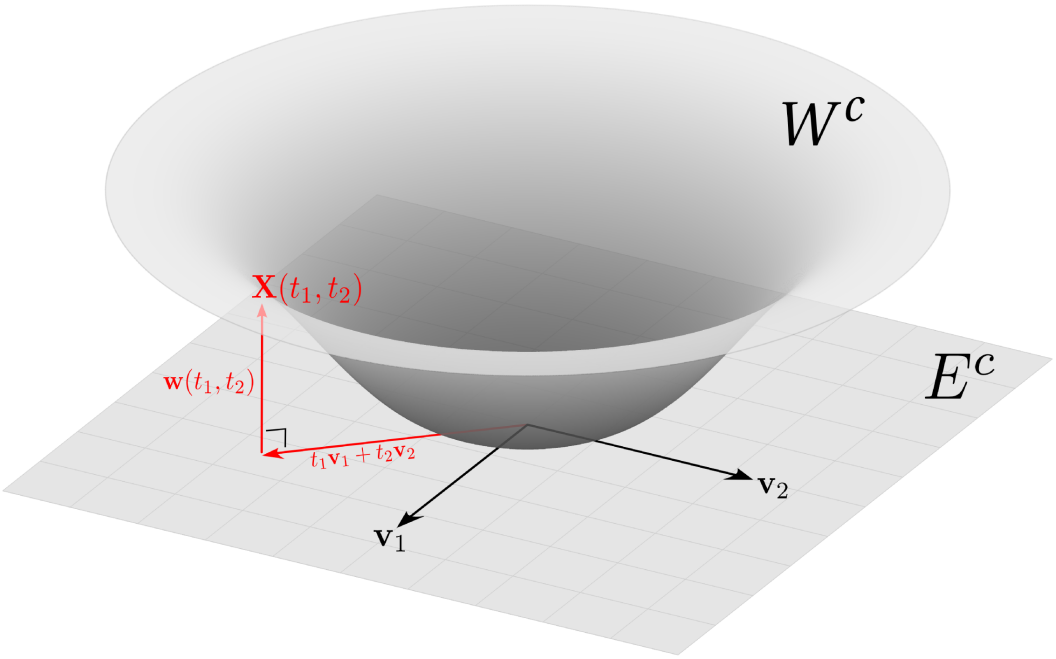

FIGURE 2.4. 2-Dimensional center manifold, W^c, written as an embedding, $\mathbf{X}(t_1, t_2)$, over the center subspace E^c.

Since W^c is an invariant manifold, the flow at a point on the center manifold is tangent to W^c there, and so

$$f(\mathbf{X}(t_1, t_2)) = \alpha_1(t_1, t_2)\frac{\partial \mathbf{X}}{\partial t_1} + \alpha_2(t_1, t_2)\frac{\partial \mathbf{X}}{\partial t_2}, \tag{15}$$

for all t_1 and t_2. The real-valued functions, $\alpha_1 : \mathbb{R}^2 \to \mathbb{R}$ and $\alpha_2 : \mathbb{R}^2 \to \mathbb{R}$, are implicitly defined by the observation that the partial derivatives, $\left.\frac{\partial \mathbf{X}}{\partial t_1}\right|_{(t_1,t_2)}$ and $\left.\frac{\partial \mathbf{X}}{\partial t_2}\right|_{(t_1,t_2)}$, form a basis for the tangent space $T_{\mathbf{X}(t_1,t_2)}(W^c)$. We note that $\alpha_1(0,0) = \alpha_2(0,0) = 0$, because $f(\mathbf{0}) = \mathbf{0}$.

Since any point, $\mathbf{X} \in W^c$ will evolve under the flow defined by f we have

$$\dot{\mathbf{X}} = \frac{d\mathbf{X}}{dt}$$

$$= \frac{dt_1}{dt}\frac{\partial \mathbf{X}}{\partial t_1} + \frac{dt_2}{dt}\frac{\partial \mathbf{X}}{\partial t_2}$$

$$= f(\mathbf{X})$$

$$= \alpha_1 \frac{\partial \mathbf{X}}{\partial t_1} + \alpha_2 \frac{\partial \mathbf{X}}{\partial t_2}$$

by equation 15. This implies

$$\left(\alpha_1 - \frac{dt_1}{dt}\right)\frac{\partial \mathbf{X}}{\partial t_1} + \left(\alpha_2 - \frac{dt_2}{dt}\right)\frac{\partial \mathbf{X}}{\partial t_2} = \mathbf{0}$$

which implies $\alpha_1 = \dot{t_1}$ and $\alpha_2 = \dot{t_2}$, since $\dfrac{\partial \mathbf{X}}{\partial t_1}$ and $\dfrac{\partial \mathbf{X}}{\partial t_2}$ are linearly independent. The observation is that the functions α_1 and α_2 describe the time rate-of-change of t_1 and t_2 respectively, as a point on the center manifold evolves under the flow. Thus, for example, if the center manifold consists entirely of fixed-points, we expect the functions α_1 and α_2 to be identically 0. However, if these functions are nonzero at some point on the center manifold, we must conclude that this point is not fixed. Note that if the center subspace were k-dimensional, we would find k functions $\alpha_i : \mathbb{R}^k \to \mathbb{R}$, each describing the time rate of change of one of the k parameters of the embedding.

By differentiating equation 15 with respect to t_1 and t_2 we find

$$\begin{aligned} Df|_{\mathbf{X}(t_1,t_2)}\left[\mathbf{X}_{t_1}(t_1,t_2)\right] &= \alpha_1(t_1,t_2)\mathbf{w}_{t_1t_1}(t_1,t_2) + \left(\mathbf{v}_1 + \mathbf{w}_{t_1}(t_1,t_2)\right)\frac{\partial \alpha_1}{\partial t_1}(t_1,t_2) \\ &\quad + \alpha_2(t_1,t_2)\mathbf{w}_{t_2t_1}(t_1,t_2) + \left(\mathbf{v}_2 + \mathbf{w}_{t_2}(t_1,t_2)\right)\frac{\partial \alpha_2}{\partial t_1}(t_1,t_2), \end{aligned} \tag{16}$$

and

$$Df|_{\mathbf{X}(t_1,t_2)}\left[\mathbf{X}_{t_2}(t_1,t_2)\right] = \alpha_1(t_1,t_2)\mathbf{w}_{t_1t_2}(t_1,t_2) + \left(\mathbf{v}_1 + \mathbf{w}_{t_1}(t_1,t_2)\right)\frac{\partial \alpha_1}{\partial t_2}(t_1,t_2)$$

$$+ \alpha_2(t_1,t_2)\mathbf{w}_{t_2t_2}(t_1,t_2) + (\mathbf{v}_2 + \mathbf{w}_{t_2}(t_1,t_2))\,\frac{\partial \alpha_2}{\partial t_2}(t_1,t_2). \tag{17}$$

Evaluating 16 and 17 at $(t_1,t_2) = (0,0)$ gives

$$\mathbf{0} = Df|_{\mathbf{0}}[\mathbf{v}_1] = \mathbf{v}_1\frac{\partial \alpha_1}{\partial t_1}(0,0) + \mathbf{v}_2\frac{\partial \alpha_2}{\partial t_1}(0,0) \text{ and}$$
$$\mathbf{0} = Df|_{\mathbf{0}}[\mathbf{v}_2] = \mathbf{v}_1\frac{\partial \alpha_1}{\partial t_2}(0,0) + \mathbf{v}_2\frac{\partial \alpha_2}{\partial t_2}(0,0),$$

which together imply that $\frac{\partial \alpha_1}{\partial t_1} = \frac{\partial \alpha_1}{\partial t_2} = \frac{\partial \alpha_2}{\partial t_1} = \frac{\partial \alpha_2}{\partial t_2} = 0$ at the origin. We expected this conclusion since there is no linear contribution from the flow along $\mathbf{v}_1$ nor $\mathbf{v}_2$.

The general process of further expanding $\mathbf{w}(t_1,t_2)$ and each $\alpha_i(t_1,t_2)$ now proceeds by taking partial derivatives of equations 16 and 17 with respect to t_1 and t_2. The result will be a set of four equations of the form

$$\begin{aligned} D^2f[\mathbf{X}_{t_i},\mathbf{X}_{t_j}] + Df[\mathbf{X}_{t_it_j}] &= \alpha_1\mathbf{X}_{t_1t_it_j} + \frac{\partial \alpha_1}{\partial t_j}\mathbf{X}_{t_1t_i} + \frac{\partial \alpha_1}{\partial t_i}\mathbf{X}_{t_1t_j} + \frac{\partial^2 \alpha_1}{\partial t_j \partial t_i}\mathbf{X}_{t_1} \\ &\quad + \alpha_2\mathbf{X}_{t_2t_it_j} + \frac{\partial \alpha_2}{\partial t_j}\mathbf{X}_{t_2t_i} + \frac{\partial \alpha_2}{\partial t_i}\mathbf{X}_{t_2t_j} + \frac{\partial^2 \alpha_2}{\partial t_j \partial t_i}\mathbf{X}_{t_2}, \end{aligned} \tag{18}$$

for $i,j \in \{1,2\}$. Upon substitution of $\mathbf{X}_{t_i} = \mathbf{v}_i + \mathbf{w}_{t_i}$ and $\mathbf{X}_{t_it_j} = \mathbf{w}_{t_it_j}$ into 18 and evaluation at the origin, we find

$$D^2f[\mathbf{v}_1,\mathbf{v}_1] + Df[\mathbf{w}_{t_1t_1}] = \frac{\partial^2 \alpha_1}{\partial {t_1}^2}\,\mathbf{v}_1 + \frac{\partial^2 \alpha_2}{\partial {t_1}^2}\,\mathbf{v}_2 \tag{19}$$

$$D^2f[\mathbf{v}_2,\mathbf{v}_1] + Df[\mathbf{w}_{t_1t_2}] = \frac{\partial^2 \alpha_1}{\partial t_2 \partial t_1}\,\mathbf{v}_1 + \frac{\partial^2 \alpha_2}{\partial t_2 \partial t_1}\,\mathbf{v}_2 \tag{20}$$

$$D^2f[\mathbf{v}_1,\mathbf{v}_2] + Df[\mathbf{w}_{t_2t_1}] = \frac{\partial^2 \alpha_1}{\partial t_1 \partial t_2}\,\mathbf{v}_1 + \frac{\partial^2 \alpha_2}{\partial t_1 \partial t_2}\,\mathbf{v}_2 \tag{21}$$

and

$$D^2 f[\mathbf{v}_2, \mathbf{v}_2] + Df[\mathbf{w}_{t_2 t_2}] = \frac{\partial^2 \alpha_1}{\partial {t_2}^2}\, \mathbf{v}_1 + \frac{\partial^2 \alpha_2}{\partial {t_2}^2}\, \mathbf{v}_2. \tag{22}$$

Note that all second order derivatives of each $\alpha_i(t_1, t_2)$ and $\mathbf{w}(t_1, t_2)$ are unknown in Equations 19-22. However, if $\mathbf{u}_1$ and $\mathbf{u}_2$ span the left kernel of $Df|_{\mathbf{0}}$, left-multiplication by each $\mathbf{u}_1$ and $\mathbf{u}_2$ eliminates $\mathbf{w}_{t_i t_j}$ and transforms each equation into a system of two linear-equations in only two unknowns:

$$\mathbf{u}_1 \cdot D^2 f[\mathbf{v}_i, \mathbf{v}_j] = \mathbf{u}_1 \cdot \mathbf{v}_1 \frac{\partial^2 \alpha_1}{\partial t_j \partial t_i} + \mathbf{u}_1 \cdot \mathbf{v}_2 \frac{\partial^2 \alpha_2}{\partial t_j \partial t_i}$$
$$\mathbf{u}_2 \cdot D^2 f[\mathbf{v}_i, \mathbf{v}_j] = \mathbf{u}_2 \cdot \mathbf{v}_1 \frac{\partial^2 \alpha_1}{\partial t_j \partial t_i} + \mathbf{u}_2 \cdot \mathbf{v}_2 \frac{\partial^2 \alpha_2}{\partial t_j \partial t_i},$$

for $i, j \in \{1, 2\}$. Solving these systems give unique solutions for $\dfrac{\partial^2 \alpha_1}{\partial t_j \partial t_i}$ and $\dfrac{\partial^2 \alpha_2}{\partial t_j \partial t_i}$:

$$\frac{\partial^2 \alpha_2}{\partial t_i \partial t_j}(0,0) = \frac{\left[\mathbf{u}_2 - \left(\dfrac{\mathbf{u}_2 \cdot \mathbf{v}_1}{\mathbf{u}_1 \cdot \mathbf{v}_1}\right) \mathbf{u}_1\right] \cdot D^2 f[\mathbf{v}_i, \mathbf{v}_j]}{\left[\mathbf{u}_2 - \left(\dfrac{\mathbf{u}_2 \cdot \mathbf{v}_1}{\mathbf{u}_1 \cdot \mathbf{v}_1}\right) \mathbf{u}_1\right] \cdot \mathbf{v}_2} \tag{23}$$

$$\frac{\partial^2 \alpha_1}{\partial t_i \partial t_j}(0,0) = \frac{\mathbf{u}_1 \cdot D^2 f[\mathbf{v}_i, \mathbf{v}_j] - \mathbf{u}_1 \cdot \mathbf{v}_2 \dfrac{\partial^2 \alpha_2}{\partial t_i \partial t_j}(0,0)}{\mathbf{u}_1 \cdot \mathbf{v}_1}. \tag{24}$$

If $Df|_{\mathbf{0}}$ is symmetric, $\mathbf{u}_i = \mathbf{v}_i$ for each i. In the case of a dimension-k center manifold, one would eliminate $\mathbf{w}_{t_i t_j}$ in the same manner as above, through left-multiplication by each of the k vectors in the span of the left kernel, leading to a system of k equations in (then) k unknowns: $(\alpha_1)_{t_i t_j}, \ldots, (\alpha_k)_{t_i t_j}$.

Finally, having found all second order partial derivatives of α_1 and α_2 at the origin, we can solve for each $\mathbf{w}_{t_i t_j}(0,0)$ by imposing the condition that $\mathbf{w}$ lie in the orthogonal complement

of E^c. Specifically,

$$\mathbf{w}_{t_i t_j}(0,0) = L^{-1}\left(\frac{\partial^2 \alpha_1}{\partial t_i \partial t_j}\mathbf{v}_1 + \frac{\partial^2 \alpha_2}{\partial t_i \partial t_j}\mathbf{v}_2 - D^2 f[\mathbf{v}_i, \mathbf{v}_j]\right) \tag{25}$$

for $i, j \in \{1, 2\}$, where L is the restriction of the Jacobian $Df|_{\mathbf{0}}$ to $(E^c)^{\perp}$. In practice, this amounts to solving (for each i, j) the system of linear equations:

$$Df[\mathbf{w}_{t_i t_j}] = \frac{\partial^2 \alpha_1}{\partial t_i \partial t_j}\mathbf{v}_1 + \frac{\partial^2 \alpha_2}{\partial t_i \partial t_j}\mathbf{v}_2 - D^2 f[\mathbf{v}_i, \mathbf{v}_j]$$

$$\mathbf{w}_{t_i t_j} \cdot \mathbf{v}_1 = 0$$

$$\mathbf{w}_{t_i t_j} \cdot \mathbf{v}_2 = 0,$$

for the unknown vector $\mathbf{w}_{t_i t_j} \in \mathbb{R}^N$.

The embedding process continues in this way. Having taken $m + 1$ derivatives of the tangency condition (Equation 15), with respect to each parameter in the embedding, and having recursively solved for all order-$(m-1)$ derivatives of $\mathbf{w}$ and each α_i at the origin, we are able to solve for the order-m derivatives in the Taylor expansion of the center manifold and the parametrized flow there.

We end this section with a simple yet illustrative example. Consider the scalar function

$$V(x, y) = x^2 y^2,$$

which is non-negative for all $[x, y] \in \mathbb{R}^2$ and 0 precisely on the axes $x = 0$ and $y = 0$. This potential defines the gradient system

$$f(x, y) = -\nabla V(x, y) = \begin{bmatrix} -2xy^2 \\ -2yx^2 \end{bmatrix},$$

whose degenerate fixed-points are comprised of the x- and y-axis. So, we proceed by performing coordinate-independent center-manifold reductions at the origin, where the axes intersect.

Since

$$Df = \begin{bmatrix} -2y^2 & -4xy \\ -4yx & -2x^2 \end{bmatrix},$$

$Df|_{\mathbf{0}}$ is the zero matrix, and the center subspace is all of $\mathbb{R}^2$. This was expected since the flow is entirely non-linear. Thus we expect the parametrized flow on the center manifold to be equivalent to the system itself. Since the kernel of $Df|_{\mathbf{0}}$ is the entire plane, we choose $\mathbf{v}_1 = [1, 0]$ and $\mathbf{v}_2 = [0, 1]$ as a basis for E^c. It is clear that if

$$\mathbf{X}(t_1, t_2) = t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + \mathbf{w}(t_1, t_2),$$

with $\mathbf{w}(t_1, t_2) \perp \mathbf{v}_1$ and $\mathbf{w}(t_1, t_2) \perp \mathbf{v}_2$, is the parametrization of the center manifold at the origin, then $\mathbf{w}(t_1, t_2) \equiv 0$. Still, we expand the functions $\alpha_1(t_1, t_2) = \dot{t_1}$ and $\alpha_2(t_1, t_2) = \dot{t_2}$ to determine the flow on the center manifold and illustrate the process outlined above. The

second derivative of f also vanishes at the origin, since

$$D^2 f = \begin{bmatrix} \begin{bmatrix} 0 \\ -4y \end{bmatrix} & \begin{bmatrix} -4y \\ -4x \end{bmatrix} \\ \begin{bmatrix} -4y \\ -4x \end{bmatrix} & \begin{bmatrix} -4x \\ 0 \end{bmatrix} \end{bmatrix},$$

which implies that the second order partials $(\alpha_1)_{t_i t_j} = (\alpha_2)_{t_i t_j} = 0$. The third-order derivative of f is the first which does not vanish at $[x, y] = [0, 0]$:

$$D^3 f|_{\mathbf{0}} = \begin{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -4 \end{bmatrix} & \begin{bmatrix} 0 & -4 \\ -4 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & -4 \\ -4 & 0 \end{bmatrix} & \begin{bmatrix} -4 & 0 \\ 0 & 0 \end{bmatrix} \end{bmatrix}.$$

All higher order derivatives vanish everywhere and so the only possible nonzero terms in the Taylor expansion of α_1 and α_2 are degree three monomials. The embedding process leads to the system of equations at the origin,

$$\mathbf{u}_1 \cdot D^3 f|_{\mathbf{0}}[\mathbf{v}_k, \mathbf{v}_j, \mathbf{v}_i] = \frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i} \mathbf{u}_1 \cdot \mathbf{v}_1 + \frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i} \mathbf{u}_1 \cdot \mathbf{v}_2$$
$$\mathbf{u}_2 \cdot D^3 f|_{\mathbf{0}}[\mathbf{v}_k, \mathbf{v}_j, \mathbf{v}_i] = \frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i} \mathbf{u}_2 \cdot \mathbf{v}_1 + \frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i} \mathbf{u}_2 \cdot \mathbf{v}_2$$

for $i, j, k \in \{1, 2\}$. Solving for the third order partials of α_1 and α_2, we find that only the mixed-partials

$$(\alpha_1)_{t_2 t_2 t_1} = (\alpha_1)_{t_2 t_1 t_2} = (\alpha_1)_{t_1 t_2 t_2} = -4$$

and

$$(\alpha_2)_{t_1t_1t_2} = (\alpha_2)_{t_1t_2t_1} = (\alpha_2)_{t_2t_1t_1} = -4$$

are nonzero. Therefore,

$$\frac{dt_1}{dt} = \alpha_1(t_1, t_2) = 3\frac{1}{3!}(-4t_1t_2^2) = -2t_1{t_2}^2$$
$$\frac{dt_2}{dt} = \alpha_2(t_1, t_2) = 3\frac{1}{3!}(-4t_2t_1^2) = -2t_2{t_1}^2.$$

By parametrizing the center manifold in this way, $\mathbf{X}(t_1, t_2) = t_1\mathbf{v}_1 + t_2\mathbf{v}_2 = [t_1, t_2]$, we have determined the flow on the center manifold to be $[\dot{t_1}, \dot{t_2}] = [-2t_1{t_2}^2, -2t_2{t_1}^2]$, as we knew all along.

The same calculations at a fixed point $[x, 0]$ away from the origin (similarly at any $[0, y], y \neq 0$) reveal a one dimensional center manifold spanned by the unit eigenvector $\mathbf{v}_1 = [1, 0]$ corresponding to the eigenvalue $\lambda_2 = 0$ of $Df|_{[x,0]}$. This center manifold consists entirely of the line of fixed points passing through $[x, 0]$. The other unit eigenvector, $[0, 1]$, corresponds to the eigenvalue $\lambda_1 = -2x^2$, which is nonzero for any choice of $x \neq 0$. As x vanishes, so to does λ_1. It is for this reason that the dimension of the center manifold jumps from one to two dimensions precisely at the origin; as the center manifold is forced to open-up to allow for the two independent directions of fixed-point degeneracy.

In addition to demonstrating that the center manifold at a non-isolated fixed point must contain those nearby fixed points, this example illustrates an obvious fact: the set of degenerate fixed points of a gradient system need not be a smooth manifold. In this example, the set of fixed points is an algebraic variety with a singular point at the origin where the two irreducible components $x = 0$ and $y = 0$ intersect. In the next section, we will discover this phenomenon again in the systems which flow towards complex Hadamards.

2.4.2. 4×4 Complex Hadamards. It is known that every 4×4 Hadamard matrix is equivalent to a member of the continuous 1-parameter family of inequivalent Hadamards,

$$F_4^{(1)}(a) \doteq \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \mathrm{i}e^{\mathrm{i}a} & -1 & -\mathrm{i}e^{\mathrm{i}a} \\ 1 & -1 & 1 & -1 \\ 1 & -\mathrm{i}e^{\mathrm{i}a} & -1 & \mathrm{i}e^{\mathrm{i}a} \end{bmatrix}.$$

Moreover, for all but one choice of parameter $a \in [0, \pi]$, the defect of $F_4^{(1)}$ coincides with the dimension of this topological-circle of inequivalent Hadamards. However, for $a = \pi/2$, when the Hadamard is real, the defect jumps from 1 to 3. Thus, this is an example where we know the defect overestimates the local freedom of inequivalent Hadamards. In this section we provide proof-of-concept of our center-manifold reduction by using it to show what is already known: The space of inequivalent Hadamards near $F_4^{(1)}(\pi/2)$ is one-dimensional, despite what the defect there might suggest.

Derivation of the gradient system (Equation 13) begins with the matrices

$$H_4(\boldsymbol{\theta}) \doteq \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{\mathrm{i}\theta_1} & e^{\mathrm{i}\theta_2} & e^{\mathrm{i}\theta_3} \\ 1 & e^{\mathrm{i}\theta_4} & e^{\mathrm{i}\theta_5} & e^{\mathrm{i}\theta_6} \\ 1 & e^{\mathrm{i}\theta_7} & e^{\mathrm{i}\theta_8} & e^{\mathrm{i}\theta_9} \end{bmatrix}.$$

As in Section 2.4, we define the potential function $\mathcal{V}_4(\boldsymbol{\theta})$ that vanishes when $H_4(\boldsymbol{\theta})$ is Hadamard and further define a 4×4 gradient system: $\Phi_4(\boldsymbol{\theta}) = -\nabla \mathcal{V}_4(\boldsymbol{\theta})$.

For the remainder of this section we will denote the one-parameter family of inequivalent 4×4 Hadamards as $F(a) \doteq F_4^{(1)}(a)$. We will interchangeably refer to elements of this space

as either the matrix $F(a)$, for a particular $a \in [0, \pi/2]$, or the vector

$$\boldsymbol{\theta}(a) \doteq \left[a - \frac{\pi}{2}, \pi, a + \frac{\pi}{2}, \pi, 0, \pi, a + \frac{\pi}{2}, \pi, a - \frac{\pi}{2}\right].$$

With the aid of a compute algebra system, we computed explicit formulae for the eigenvalues of the matrix $D\Phi_4|_{\boldsymbol{\theta}(a)}$ as functions of the parameter $a \in [0, \pi]$.

$$D\Phi_4|_{\boldsymbol{\theta}(a)} = \begin{bmatrix} -12 & 4 & 4 & 4 & -4\sin a & -4 & 4 & -4 & 4 \\ 4 & -12 & 4 & -4\sin a & 4 & 4\sin a & -4 & 4 & -4 \\ 4 & 4 & -12 & -4 & 4\sin a & 4 & 4 & -4 & 4 \\ 4 & -4\sin a & -4 & -12 & 4 & 4 & 4 & 4\sin a & -4 \\ -4\sin a & 4 & 4\sin a & 4 & -12 & 4 & 4\sin a & 4 & -4\sin a \\ -4 & 4\sin a & 4 & 4 & 4 & -12 & -4 & -4\sin a & 4 \\ 4 & -4 & 4 & 4 & 4\sin a & -4 & -12 & 4 & 4 \\ -4 & 4 & -4 & 4\sin a & 4 & -4\sin a & 4 & -12 & 4 \\ 4 & -4 & 4 & -4 & -4\sin a & 4 & 4 & 4 & -12 \end{bmatrix}.$$

The characteristic polynomial of $D\Phi_4|_{\boldsymbol{\theta}(a)}$ is

$$\begin{aligned} p(\lambda; a) = -\lambda[\lambda + 8][\lambda^2 + (32 - 8\sin a)\lambda + 128(1 - \sin a)] \\ [\lambda^2 + (32 + 8\sin a)\lambda + 128(1 + \sin a)] \\ [\lambda^3 + 36\lambda^2 + (352 - 64\sin^2 a)\lambda + 512(1 - \sin^2 a)]. \end{aligned}$$

Let $\lambda_1(a) = 0, \lambda_2(a) = -8, \lambda_3(a)$, $\lambda_4(a)$ and $\lambda_5(a)$ equal the three real roots of the cubic factor

$$c(\lambda; a) \doteq \lambda^3 + 36\lambda^2 + (352 - 64\sin^2 a)\lambda + 512(1 - \sin^2 a),$$

$\lambda_6(a)$ and $\lambda_7(a)$ the roots of the quadratic

$$q_1(\lambda; a) \doteq \lambda^2 + (32 - 8\sin a)\lambda + 128(1 - \sin a),$$

and $\lambda_8(a)$ and $\lambda_9(a)$ the roots of the quadratic

$$q_2(\lambda; a) \doteq \lambda^2 + (32 + 8\sin a)\lambda + 128(1 + \sin a).$$

The constant terms of $c(\lambda; a)$ and $q_1(\lambda; a)$ vanish precisely at $a = \pi/2$, increasing the multiplicity of the root $\lambda = 0$, of $p(\lambda; a)$, to three there. A plot of the eigenvalues of $D\Phi_4|_{\boldsymbol{\theta}(a)}$ is given in Figure 2.5.

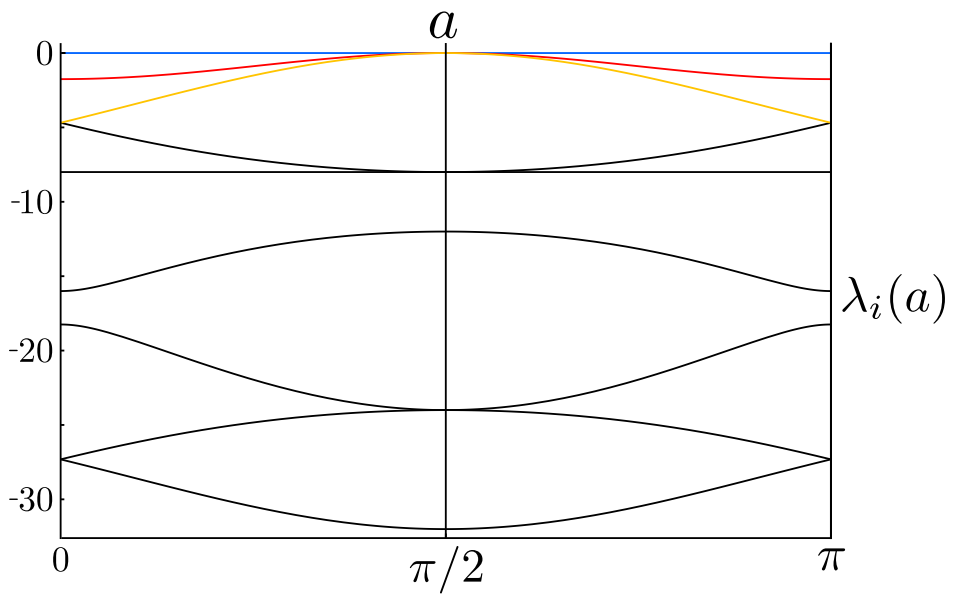

FIGURE 2.5. Plot of the eigenvalues of the linearization of Φ_4 at $\boldsymbol{\theta}(a)$ for $a \in [0, \pi]$. $\lambda_1(a)$ (blue), $\lambda_3(a)$ (red), and $\lambda_6(a)$ (yellow) simultaneously vanish at $a = \pi/2$, while all other eigenvalues (black) are strictly negative for all $a \in [0, \pi]$.

As stated, for every parameter value (other than $a = \pi/2$) there are exactly 8 negative eigenvalues and 1 eigenvalue equal to 0. The latter corresponds to the 1-dimensional manifold of fixed points parametrized by a, as its eigenvector is the tangent vector to the manifold

$F_4^{(1)}$ embedded in $\mathbf{T}^9$,

$$\mathbf{v}_1 = [1,0,1,0,0,0,1,0,1]\,.$$

At $a = \pi/2$, λ_3 and λ_6 also vanish, giving rise to a three dimensional center manifold spanned by $\mathbf{v}_1$ and the eigenvectors

$$\mathbf{v}_2 = [0,0,0,1,1,0,1,1,0]$$

and

$$\mathbf{v}_3 = [0,1,1,0,1,1,0,0,0]\,.$$

Note the similarity of this occurrence to the example we constructed at the end of Section 2.4.1. We anticipate nonlinear flow at all nearby points off of the lines spanned by each $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ since the space of inequivalent Hadamards is 1-dimensional at $F(\pi/2)$. Thus, we proceed by expanding the center manifold,

$$\mathbf{X}(t_1,t_2,t_3) \doteq t_1\mathbf{v}_1 + t_2\mathbf{v}_2 + t_3\mathbf{v}_3 + \mathbf{w}(t_1,t_2,t_3),$$

over $E^c = \text{span}(\mathbf{v}_1,\mathbf{v}_2,\mathbf{v}_3)$ in the hopes of finding nonlinear flow. Again, we choose $\mathbf{w}(t_1,t_2,t_3) \in (E^c)^\perp$.

Let the time-rates-of-change of the embedding parameters be $\alpha_1(t_1,t_2,t_3) = \dot{t_1}$, $\alpha_2(t_1,t_2,t_3) = \dot{t_2}$, and $\alpha_3(t_1,t_2,t_3) = \dot{t_3}$, which vanish (along with their first partials) at $[t_1,t_2,t_3] = [0,0,0]$. It happens that $D^2\Phi_4|_{\boldsymbol{\theta}(\frac{\pi}{2})}[\mathbf{v},\mathbf{w}] = \mathbf{0}$ for any $\mathbf{v},\mathbf{w} \in \mathbb{R}^9$. Thus, we find that

$$\begin{aligned}
0 = \mathbf{v}_1 \cdot D^2\Phi_4[\mathbf{v}_i,\mathbf{v}_j] &= \frac{\partial^2\alpha_1}{\partial t_j \partial t_i}\mathbf{v}_1\cdot\mathbf{v}_1 + \frac{\partial^2\alpha_2}{\partial t_j \partial t_i}\mathbf{v}_1\cdot\mathbf{v}_2 + \frac{\partial^2\alpha_3}{\partial t_j \partial t_i}\mathbf{v}_1\cdot\mathbf{v}_3 \\
0 = \mathbf{v}_2 \cdot D^2\Phi_4[\mathbf{v}_i,\mathbf{v}_j] &= \frac{\partial^2\alpha_1}{\partial t_j \partial t_i}\mathbf{v}_2\cdot\mathbf{v}_1 + \frac{\partial^2\alpha_2}{\partial t_j \partial t_i}\mathbf{v}_2\cdot\mathbf{v}_2 + \frac{\partial^2\alpha_3}{\partial t_j \partial t_i}\mathbf{v}_2\cdot\mathbf{v}_3 \\
0 = \mathbf{v}_3 \cdot D^2\Phi_4[\mathbf{v}_i,\mathbf{v}_j] &= \frac{\partial^2\alpha_1}{\partial t_j \partial t_i}\mathbf{v}_3\cdot\mathbf{v}_1 + \frac{\partial^2\alpha_2}{\partial t_j \partial t_i}\mathbf{v}_3\cdot\mathbf{v}_2 + \frac{\partial^2\alpha_3}{\partial t_j \partial t_i}\mathbf{v}_3\cdot\mathbf{v}_3,
\end{aligned}$$

for $i, j \in \{1, 2, 3\}$, which together imply that all second order partials of α_1, α_2 and α_3 vanish. Note that here we have used the fact that the left kernel of $D\Phi_4$ is spanned by $\mathbf{v}_1, \mathbf{v}_2$ and $\mathbf{v}_3$ since the Jacobian is symmetric.

By differentiating the tangency condition three times with respect to t_i, t_j and t_k and evaluating at $\boldsymbol{\theta}(\pi/2)$, we derive 27 equations of the form

$$D\Phi_4[\mathbf{w}_{t_i t_j t_k}] + D^3\Phi_4[\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k] = \frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i}\mathbf{v}_1 + \frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i}\mathbf{v}_2 + \frac{\partial^3 \alpha_3}{\partial t_k \partial t_j \partial t_i}\mathbf{v}_3, \tag{26}$$

for $i, j, k \in \{1, 2, 3\}$. Left multiplication by each $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$, expands each of the 27 equations (26) into a system of three equations with a unique solution for the third order partials of α_1, α_2 and α_3. These third-order Taylor coefficients are summarized in Table 2.1. Thus, up to third order, the motion of the point $\mathbf{X}(t_1, t_2, t_3)$ on W^c will be governed by the

Table 2.1: Third-order partial derivatives of α_1, α_2 and α_3 at $F(\pi/2)$, indicated by the indices of the parameters with which the partial derivatives are taken.

	$\{i,j,k\}$									
	$\{1,1,1\}$	$\{1,1,2\}$	$\{1,1,3\}$	$\{1,2,2\}$	$\{1,3,3\}$	$\{1,2,3\}$	$\{2,2,2\}$	$\{2,2,3\}$	$\{2,3,3\}$	$\{3,3,3\}$
$\frac{\partial^3\alpha_1}{\partial t_k \partial t_j \partial t_i}$	0	8/9	8/9	-40/9	-40/9	0	0	8/9	8/9	0
$\frac{\partial^3\alpha_2}{\partial t_k \partial t_j \partial t_i}$	0	-40/9	8/9	8/9	8/9	0	0	8/9	-40/9	0
$\frac{\partial^3\alpha_3}{\partial t_k \partial t_j \partial t_i}$	0	8/9	-40/9	8/9	8/9	0	0	-40/9	8/9	0

system of ODE's:

$$\begin{aligned}
\dot{t_1} &= -\frac{20}{9}t_1{t_2}^2 - \frac{20}{9}t_1{t_3}^2 + \frac{4}{9}t_2{t_1}^2 + \frac{4}{9}t_2{t_3}^2 + \frac{4}{9}t_3{t_1}^2 + \frac{4}{9}t_3{t_2}^2 \\
\dot{t_2} &= -\frac{20}{9}t_2{t_1}^2 - \frac{20}{9}t_2{t_3}^2 + \frac{4}{9}t_1{t_2}^2 + \frac{4}{9}t_1{t_3}^2 + \frac{4}{9}t_3{t_1}^2 + \frac{4}{9}t_3{t_2}^2
\end{aligned} \tag{27}$$

$$\dot{t_3} = -\frac{20}{9}t_3{t_1}^2 - \frac{20}{9}t_3{t_2}^2 + \frac{4}{9}t_1{t_2}^2 + \frac{4}{9}t_1{t_3}^2 + \frac{4}{9}t_2{t_1}^2 + \frac{4}{9}t_2{t_3}^2$$

Because every nonzero cubic-term in Equations 27 is a mixed monomial, setting any two embedding parameters to zero will result in no flow. For example, setting $t_2 = t_3 = 0$ amounts to choosing a point $\mathbf{X}(t_1, 0, 0) \in F_4^{(1)}$ on the manifold of fixed points (since moving in the direction of $\mathbf{v}_1$ amounts to varying the parameter a). Figure 2.6 depicts examples of flow lines of the system of equations governing the evolution of the embedding parameters, colored by the magnitude of the velocity field. We note the convergence to the axes, where two of the three embedding parameters vanish. The important point is that there is flow

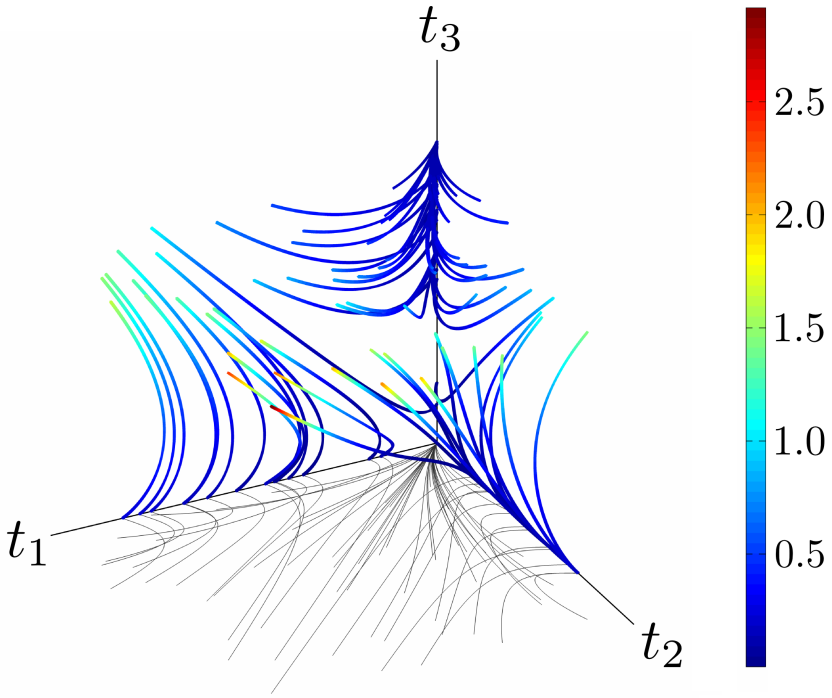

FIGURE 2.6. Flow lines of system of ODEs 27 colored by the magnitude of the velocity, along with the flow line projections into the $t_1 - t_2$ plane. Initial conditions chosen in the plane $t_3 = .75$ uniformly at random in the range $0 < t_1, t_2 < 1$.

for any choice of embedded point $\mathbf{X}(t_1, t_2, t_3)$ not directly over a lines spanned by $\mathbf{v}_1$, $\mathbf{v}_2$

or $\mathbf{v}_3$ (i.e. on an axis in $t_1 - t_2 - t_3$ space). Therefore, the local dimension of the space of inequivalent Hadamards at $F(\pi/2)$ cannot be greater than one.

Having found $(\alpha_m)_{t_i t_j t_k}$, one can solve the systems

$$D\Phi[\mathbf{w}_{t_i t_j t_k}] = \mathbf{v}_1(\alpha_1)_{t_i t_j t_k} + \mathbf{v}_2(\alpha_2)_{t_i t_j t_k} + \mathbf{v}_3(\alpha_3)_{t_i t_j t_k} - D^3\Phi[\mathbf{v}_k, \mathbf{v}_j, \mathbf{v}_i]$$

$$\mathbf{w}_{t_i t_j t_k} \cdot \mathbf{v}_1 = 0$$

$$\mathbf{w}_{t_i t_j t_k} \cdot \mathbf{v}_2 = 0$$

$$\mathbf{w}_{t_i t_j t_k} \cdot \mathbf{v}_3 = 0$$

for the unknown vectors $\mathbf{w}_{t_i t_j t_k}$. For reference, the values of the degree-three Taylor coefficients of $\mathbf{w}(t_1, t_2, t_3)$ at $F(\pi/2)$ are summarized in Table 2.2. Even to fifth order (the highest

Table 2.2: Third-order partial derivatives of $\mathbf{w}(t_1, t_2, t_3)$ at $F(\pi/2)$, indicated by the indices of the parameters with which the partial derivatives are taken.

	$\{i,j,k\}$									
	$\{1,1,1\}$	$\{1,1,2\}$	$\{1,1,3\}$	$\{1,2,2\}$	$\{1,3,3\}$	$\{1,2,3\}$	$\{2,2,2\}$	$\{2,2,3\}$	$\{2,3,3\}$	$\{3,3,3\}$
$\frac{\partial^3 \mathbf{w}}{\partial t_k \partial t_j \partial t_i}$	$\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}2/9\\2/9\\-1/3\\1/9\\-1/9\\2/9\\-1/9\\1/9\\2/9\end{bmatrix}$	$\begin{bmatrix}2/9\\1/9\\-1/9\\2/9\\-1/9\\1/9\\-1/3\\2/9\\2/9\end{bmatrix}$	$\begin{bmatrix}1/9\\2/9\\-1/9\\2/9\\-1/3\\2/9\\-1/9\\2/9\\1/9\end{bmatrix}$	$\begin{bmatrix}1/9\\2/9\\-1/9\\2/9\\-1/3\\2/9\\-1/9\\2/9\\1/9\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$	$\begin{bmatrix}2/9\\1/9\\-1/9\\2/9\\-1/9\\1/9\\-1/3\\2/9\\2/9\end{bmatrix}$	$\begin{bmatrix}2/9\\2/9\\-1/3\\1/9\\-1/9\\2/9\\-1/9\\1/9\\2/9\end{bmatrix}$	$\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\\0\\0\\0\end{bmatrix}$

order to which we performed the Taylor expansion of the flow on the center manifold), there remains no flow along the eigenvectors $\mathbf{v}_2$ and $\mathbf{v}_3$. There is, however, a simple explanation for this: Recall that Φ_4 does not converge to the space of inequivalent Hadamards, but rather

to the superset containing all row and column permutations of the core of dephased 4×4 Hadamards. Let $P_r(i,j)$ and $P_c(i,j)$ be the 4×4 permutation matrices which act to swap rows i and j and columns i and j respectively. There are exactly five unique row and column permutations of $F(a)$ which, for some choice of parameter a, are again equal to the matrix $F(\pi/2)$. In particular,

$$\begin{aligned} F(\pi/2) &= F(3\pi/2)P_c(2,4) \\ &= P_r(2,4)F(3\pi/2) \\ &= P_r(2,4)F(\pi/2)P_c(2,4) \\ &= P_r(2,3)F(\pi/2)P_c(3,4) \\ &= P_r(3,4)F(\pi/2)P_c(2,3) \end{aligned}$$

Any such core-permutation amounts to a permutation of the the coordinates $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_9]$ of $\mathbf{T}^9$. For example,

$$P_r(2,3)H_4(\boldsymbol{\theta})P_c(3,4) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{\mathrm{i}\theta_4} & e^{\mathrm{i}\theta_6} & e^{\mathrm{i}\theta_5} \\ 1 & e^{\mathrm{i}\theta_1} & e^{\mathrm{i}\theta_3} & e^{\mathrm{i}\theta_2} \\ 1 & e^{\mathrm{i}\theta_7} & e^{\mathrm{i}\theta_9} & e^{\mathrm{i}\theta_8} \end{bmatrix},$$

corresponds to the permutation

$$\sigma_2 \doteq (14)(26)(35)(7)(89) : [\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9] \mapsto [\theta_4, \theta_6, \theta_5, \theta_1, \theta_3, \theta_2, \theta_7, \theta_9, \theta_8].$$

Let $\sigma_3 \doteq (12)(3)(48)(57)(69)$; the permutation of the coordinates resulting from the action of $P_r(3,4)$ and $P_c(2,3)$. These permutations act on the vector $\mathbf{v}_1 = [1,0,1,0,0,0,1,0,1]$, which we recall is tangent to $F(a)$ at $a = \pi/2$, in the following ways:

$$[1,0,1,0,0,0,1,0,1] \overset{\sigma_2}{\longmapsto} [0,0,0,1,1,0,1,1,0] = \mathbf{v}_2, \text{ and}$$

$$[1,0,1,0,0,0,1,0,1] \overset{\sigma_3}{\longmapsto} [0,1,1,0,1,1,0,0,0] = \mathbf{v}_3.$$

Note that $[1,0,1,0,0,0,1,0,1]$ is fixed by both $P_r(2,4)$ and $P_c(2,4)$. Therefore, the three-dimensional center manifold emerges at the real Hadamard to account for three copies of the space of dephased, permutation-equivalent Hadamards intersecting here. We see that there are five total, and three distinct, directions of fixed-point degeneracy at $F(\pi/2)$ caused by the permutations of its core. One conclusion we can draw from this is that there will never be flow on the center manifold in the eigenvector directions (as was the case in the example in Section 2.4.1) because $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_3$ indicate three lines of (permutation-equivalent) complex Hadamards sharing a common point of intersection. A cartoon illustrating our conclusions about the local structure of dephased Hadamards at $F(\pi/2)$ is given in Figure 2.7.

2.4.3. An Isolated Matrix with Positive Defect. In the previous section we performed a coordinate-independent center manifold reduction on the gradient system $\Phi_4(\boldsymbol{\theta})$ at $\boldsymbol{\theta}(\pi/2)$. We were able to obtain explicit eigenvectors and eigenvalues for its linearization and explicit coefficients in the Taylor expansions of the center manifold and the time-rates-of-change of the embedding parameters, t_1, t_2, and t_3. However, it is worth noting that the success of this technique does not hinge on one's ability to perform symbolic computations which yield exact solutions. Had we performed our computations numerically, we would

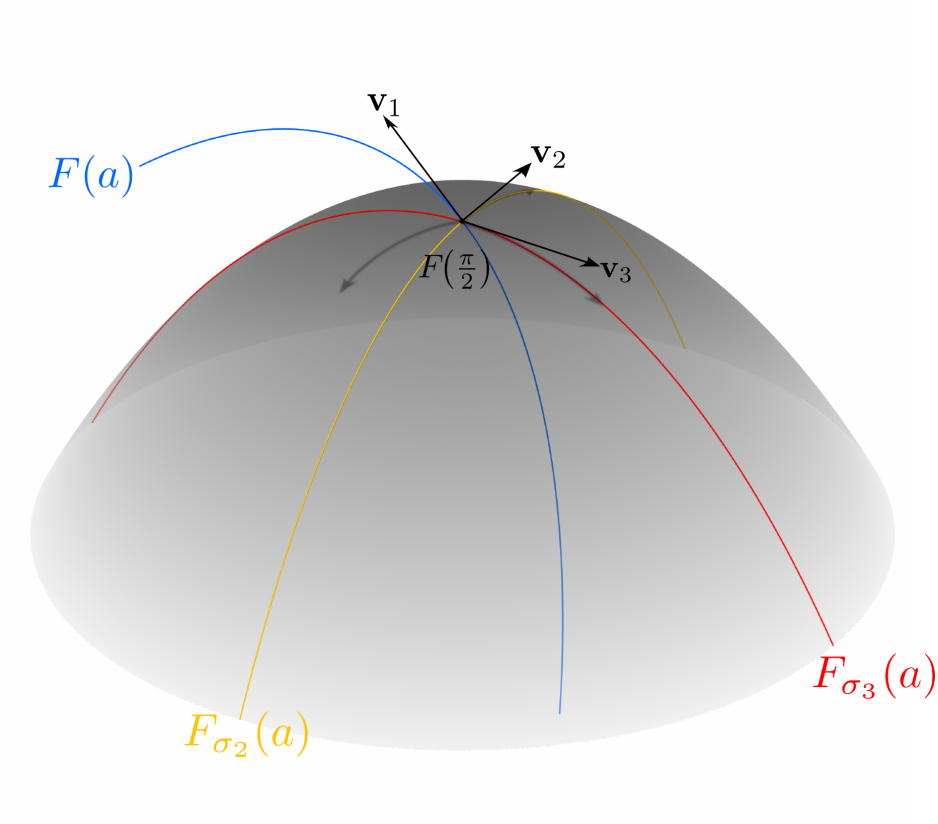

FIGURE 2.7. Cartoon of the local structure of the space of dephased, permutation-equivalent Hadamards in a neighborhood of the real matrix $F(\pi/2)$. $F_{\sigma_2}(a) \doteq P_r(2,3)F(a)P_c(3,4)$ (yellow curve), and $F_{\sigma_3}(a) \doteq P_r(3,4)F(a)P_c(2,3)$ (red curve) indicating copies of the space $F(a)$ (blue curve) of inequivalent Hadamards under the action of core permutations. A two-dimensional submanifold of W^c containing the lines of fixed points F_{σ_3} and F_{σ_2} is shaded grey.

have derived numerical approximations of the Taylor coefficients of each $\alpha_i(t_1, t_2, t_3)$, yielding approximations of the time-rates-of-change of the embedding parameters. Since these approximate coefficients would be bounded away from 0, we would deduce that there exists non-linear flow on parts of the center manifold, even without exact knowledge of that flow.

In this section we consider the 9×9 complex Hadamard matrix

$$B_9^{(0)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & \epsilon^3 & \epsilon^3 & -1 & \epsilon^9 & \epsilon^8 & \epsilon^7 & \epsilon \\ 1 & \epsilon^4 & -1 & \epsilon^7 & \epsilon & \epsilon^3 & -1 & \epsilon^9 & \epsilon^9 \\ 1 & \epsilon^3 & \epsilon^7 & -1 & \epsilon & \epsilon^8 & \epsilon^9 & \epsilon^3 & -1 \\ 1 & \epsilon^9 & \epsilon & -1 & -1 & \epsilon^3 & \epsilon^7 & \epsilon^2 & \epsilon^7 \\ 1 & \epsilon^9 & -1 & \epsilon & \epsilon^3 & -1 & \epsilon & \epsilon^7 & \epsilon^6 \\ 1 & \epsilon & \epsilon^7 & \epsilon^9 & \epsilon^6 & \epsilon & -1 & -1 & \epsilon^3 \\ 1 & \epsilon^7 & \epsilon^9 & \epsilon^4 & \epsilon^9 & -1 & \epsilon^3 & -1 & \epsilon \\ 1 & -1 & \epsilon^2 & \epsilon^9 & \epsilon^7 & \epsilon^7 & \epsilon^3 & \epsilon & -1 \end{bmatrix},$$

where $\epsilon = e^{2\pi \mathrm{i}/10}$, discovered by Beauchamp and Nicoara [36]. It is known that the defect $d\left(B_9^{(0)}\right) = 2$. Likewise, there is a 2-dimensional center subspace of the 64×64 gradient system, $\Phi_9(\boldsymbol{\theta})$, at the real vector of the phases in the core of $B_9^{(0)}$, $\boldsymbol{\theta}_B \in \mathbf{T}^{64}$. With the use of Mathematica's built-in function Nullspace[ ], we computed – to 200 decimal places of accuracy – numerical approximations of the basis vectors which span the center subspace. Again, we refer to these vectors as $\mathbf{v}_1$ and $\mathbf{v}_2$ and proceed to expand the center manifold

$$\mathbf{X}(t_1, t_2) \doteq t_1 \mathbf{v}_1 + t_2 \mathbf{v}_2 + \mathbf{w}(t_1, t_1)$$

as an embedding over the center subspace.

We find $D^2\Phi_9[\mathbf{v}_i, \mathbf{v}_j] \approx 0$[6], which we take to imply that the second-order partials of $\alpha_1(t_1, t_2)$ and $\alpha_2(t_1, t_2)$ vanish, along with second derivatives of the embedding function $\mathbf{w}(t_1, t_2)$. At third order – upon evaluation at $[t_1, t_2] = [0, 0]$ – we find 8 equations of the form

$$D\Phi_9[\mathbf{w}_{t_i t_j t_k}] + D^3\Phi_9[\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k] = \frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i} \mathbf{v}_1 + \frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i} \mathbf{v}_2, \tag{28}$$

[6] $D^2\Phi_9[\mathbf{v}_i, \mathbf{v}_j] = 0$ to the accuracy of our numerical approximation: 1e-200.

for $i, j, k \in \{1, 2\}$. Multiplication of Equation 28 on the left by each $\mathbf{v}_1$ and $\mathbf{v}_2$ yields eight linear systems:

$$
\begin{aligned}
\mathbf{v}_1 \cdot D^3\Phi_9[\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k] &= \mathbf{v}_1 \cdot \mathbf{v}_1 \frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i} + \mathbf{v}_1 \cdot \mathbf{v}_2 \frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i} \\
\mathbf{v}_2 \cdot D^3\Phi_9[\mathbf{v}_i, \mathbf{v}_j, \mathbf{v}_k] &= \mathbf{v}_2 \cdot \mathbf{v}_1 \frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i} + \mathbf{v}_2 \cdot \mathbf{v}_2 \frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i},
\end{aligned} \tag{29}
$$

which are solved for the third-order partials of the functions α_1 and α_2 governing the flow of the embedding-parameters. Table 2.3 gives approximate numerical values of the partials derivatives of α_1 and α_2 found by solving systems 29. For clarity we show only 6 decimal places of accuracy.

Table 2.3: Third-order partial derivatives of α_1 and α_2 at $B_9^{(0)}$, indicated by the indices of the parameters with which the partial derivatives are taken.

	$\{i, j, k\}$			
	$\{1, 1, 1\}$	$\{1, 1, 2\}$	$\{1, 2, 2\}$	$\{2, 2, 2\}$
$\frac{\partial^3 \alpha_1}{\partial t_k \partial t_j \partial t_i}$	-0.061053	0.015157	-0.040022	0.021325
$\frac{\partial^3 \alpha_2}{\partial t_k \partial t_j \partial t_i}$	0.015157	-0.040022	0.021325	-0.069138

Having found the third-order Taylor coefficients, we express the flow of the embedding parameters as a system of ODEs:

$$
\begin{aligned}
\frac{dt_1}{dt} = \alpha_1(t_1, t_2) &= \frac{1}{6}\frac{\partial^3 \alpha_1}{\partial t_1^3} t_1^3 + \frac{1}{6}\frac{\partial^3 \alpha_1}{\partial t_2^3} t_2^3 + \frac{1}{2}\frac{\partial^3 \alpha_1}{\partial t_1 \partial t_1 \partial t_2} t_1^2 t_2 + \frac{1}{2}\frac{\partial^3 \alpha_1}{\partial t_2 \partial t_2 \partial t_1} t_2^2 t_1 + \text{h.o.t.} \\
\frac{dt_2}{dt} = \alpha_2(t_1, t_2) &= \frac{1}{6}\frac{\partial^3 \alpha_2}{\partial t_1^3} t_1^3 + \frac{1}{6}\frac{\partial^3 \alpha_2}{\partial t_2^3} t_2^3 + \frac{1}{2}\frac{\partial^3 \alpha_2}{\partial t_1 \partial t_1 \partial t_2} t_1^2 t_2 + \frac{1}{2}\frac{\partial^3 \alpha_2}{\partial t_2 \partial t_2 \partial t_1} t_2^2 t_1 + \text{h.o.t.},
\end{aligned} \tag{30}
$$

which indicates nonlinear flow of the embedding parameters for all choices of $[t_1, t_2]$. In other words, the center manifold near $B_9^{(0)}$ contains no other complex Hadamards.

Observe that all sufficiently small choices of t_1 and t_2 must flow back to the origin of the center subspace since Φ_9 is a gradient system. One can verify the stability of $[t_1, t_2] = [0, 0]$ of 30 by computing

$$\begin{aligned}\frac{d\|[t_1, t_2]\|^2}{dt} &= 2t_1\frac{dt_1}{dt} + 2t_2\frac{dt_2}{dt} \\ &= 2t_1\alpha_1 + 2t_2\alpha_2.\end{aligned}$$

It is straightforward to show that the rate-of-change of the square of the magnitude of $[t_1, t_2]$ is nonpositive for all $[t_1, t_2] \in \mathbb{R}^2$ and vanishes only at $[0, 0]$. As a consequence of this computation we have proven Proposition 16.

PROPOSITION 16. *The matrix $B_9^{(0)} \in \mathcal{H}_9$ is isolated.*

It is a virtue of the formalism built in Section 2.4 that we need not have explicit values for the Taylor coefficients of the functions α_i to conclude that flow on the center manifold exists. Also, the computations are well-suited for implementation on a computer and can be readily performed on large, numerically-derived Hadamards as well as those given by explicit formulae. It is a shortcoming of this method that it cannot prove flow does not exist on some part of a center manifold, thereby proving the existence of a manifold of complex Hadamards. We are certain of this limitation in the construction presented here since, a priori, one has no knowledge of the smallest order in the Taylor expansion where one might first encounter nonzero contribution to the time-rates-of-change of the embedding parameters. We are not certain that the use of the gradient system 13 is bound with this deficiency and are hopeful that a deeper understanding of the derivatives of the vector field Φ_d may elicit answers to more difficult questions.

2.5. Conclusions and Future Work

Recently we linearized the 25×25 system $\Phi_6(\boldsymbol{\theta})$ about several explicitly defined matrices from known dimension-6 families and, using a computer algebra system we computed analytic eigenvalues of the linearized system. For all examples chosen, except for the isolated matrix $S_6^{(0)}$, the center and stable subspaces were found to be 4- and 21-dimensional respectively. In the case of $S_6^{(0)}$ the stable subspace spanned all of $\mathbb{R}^{25}$, which is expected since $d\left(S_6^{(0)}\right) = 0$.

A next step is apply Theorem 2.4.1 to perform a center-manifold reduction on members of $K_6^{(3)}$. Of course, we know that at least a three-dimensional submanifold of the four-dimensional center-manifold is comprised of fixed points, since $K_6^{(3)}$ is specified by three parameters. As we discussed at the end of Section 2.4.3, if $K_6^{(3)}$ belongs to a larger manifold of complex Hadamards, we will not be able to prove it using a center-manifold reduction unless we can somehow demonstrate that all terms in the Taylor expansions of each α_i vanish. At this time, we are not sure how to do this. Still, one can always expand the center manifold and the flow on it to any order one wishes to support (or refute, if flow is found) existing conjectures.

The impressive efficiency of existing algorithms for computing kernels of large symbolic and numerical matrices allows our process to be applied broadly to Hadamards of significant size. Additionally, our method not only presents a new interpretation of and approach to computing the defect of a Hadamard, but also provides a new avenue to improve the defect bound by providing deeper insight into the nature of matrices near Hadamards.

CHAPTER 3

Oblique-Incidence Ion Bombardment

3.1. Introduction

Experimental physicists have observed spontaneous pattern formation when a solid surface is bombarded with a broad ion beam([64], [65], [66], [67]). These patterns vary from parallel ripples to highly ordered hexagonal arrays of nanodots. While academic interest in pattern and defect formation has motivated research, much of the interest in these experimental observations is born from the potential to improve fabrication of nanostructures and provide effective alternatives to traditional fabrication methods. Ion bombardment has the potential to be a fast method of producing regular nanostructures which are on the order of tens of nanometers; a scale which matches that achieved by modern lithography techniques.

Several competing theories have emerged to model the hexagonal pattern formation observed in experiments; two of which attempt to model the scenario of bombardment of an elemental material ([68], [69]) and one (the Bradley/Shipman theory (BS) studied in detail in this paper) which emphasizes the necessity of a binary solid in the formation of hexagonal patterns [70]. The first two theories study the Kuramoto-Sivashinsky equation with the addition of either one or two nonlinear terms. The latter argues that generally within a binary material one material is preferentially sputtered[1] and that this results in a change in the relative concentrations of the two atomic species at the surface. The theory suggests that this change in surface stoichiometry, when coupled with the change in surface topography, is crucial in the formation of hexagonal arrays of nanodots [70].

[1]The effect of ejecting atoms from a solid by bombarding the surface with energetic particles.

In [70] coupled equations of motion describing the surface height and the relative concentrations of the components of the binary material under normal-incidence ion bombardment are derived and analyzed. Here we perform a similar treatment of the equations modeling off-normal incidence ion bombardment which introduces an anisotropy in the coupled system. In particular we consider the following system:

$$u_t = \phi - s(r_1 u_{xx} + u_{yy}) - \nabla^2\nabla^2 u + \lambda(r_3 u_x^2 + u_y^2) \tag{31}$$

$$\phi_t = -a\phi + b(r_2 u_{xx} + u_{yy}) + c\nabla^2\phi + \nu\phi^2 + \eta\phi^3, \tag{32}$$

where $u(x, y, t)$ is the surface height and $\phi(x, y, t)$ is the relative concentration of material species, and $s \in \{-1, 1\}$. Choosing $\phi = 0$ (and $b = 0$, $s = 1$) this system reduces to the anisotropic Kuramoto-Sivashinsky (AKS) equation [71],

$$u_t = -(r_1 u_{xx} + u_{yy}) - \nabla^2\nabla^2 u + \lambda(r_3 u_x^2 + u_y^2). \tag{33}$$

Choosing $r_1 = r_2 = r_3 = 1$ reduces our system to the normal incidence equations in [70].

In this chapter we study our generalization of the BS theory using both computer simulations and amplitude equations: ODEs governing the time-evolution of the unstable modes, which we find analytically in the weakly nonlinear regime. We observe that, for a given set of parameters, the surface may remain flat or develop ripples with their wavevector either parallel or perpendicular to the ion beam direction. Such ripple patterns are also observed

in oblique-incidence bombardment of elemental materials but, with our model of *binary*-material bombardment, we observe another possibility that does not arise for elemental materials: a "dots-on-ripples" topography. In such a pattern, dots that form a hexagonal array sit atop a ripple topography. We demonstrate that a transition between a ripple pattern and the dots-on-ripples morphology may occur as the angle of incidence or the ion energy is varied and we find that if such a transition does occur, it does so continuously.

The formation of defects in patterns is an obstacle to the use of ion bombardment as a tool for nanofabrication. In preliminary work on OIIB, [71] we numerically integrated the equations of motion for OIIB that we develop in detail in this paper and showed that a highly ordered ripple pattern may result. This motivates us to determine what parameter choices will lead to patterns with few defects. Numerical simulations in the present paper suggest that defects are more likely to persist for values of the parameters that result in a more quickly growing soft-mode amplitude.

In Section 3.2 we perform a linear stability analysis on Equations 31 and 32 to identify the types of bifurcations that can occur. In Section 3.3 we add nonlinear terms to these equations and the amplitude equations are constructed and analyzed. Finally, we perform numerical integrations of the nonlinear equations of motion and the amplitude equations in Section 3.4. We discuss our results and give our conclusions in Section 3.5.

3.2. Linear Stability Analysis

In this section we will perform linear stability analysis about the steady-state solution $u = \phi = 0$. First we fix the following notational conventions: We will denote wavevectors $\boldsymbol{k} = [k_x, k_y]$ with magnitude $k = \|\boldsymbol{k}\|$. Also, let $l^2 \doteq s(r_1 k_x^2 + k_y^2)$ and $m^2 \doteq r_2 k_x^2 + k_y^2$. We further define the anisotropic operators $\Delta_i \doteq r_i \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$ for $i = 1, 2, 3$.

The linear terms in Equations 31 and 32 can be expressed as a matrix equation:

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ \phi \end{pmatrix} = \begin{pmatrix} s\Delta_1 - \nabla^2\nabla^2 & 1 \\ b(\Delta_2) & c\nabla^2 - a \end{pmatrix}\begin{pmatrix} u \\ \phi \end{pmatrix}. \tag{34}$$

We assume Equation 34 has separable solutions of the form

$$\tilde{u} \doteq u_0 e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} e^{\sigma t}$$

$$\tilde{\phi} \doteq \phi_0 e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} e^{\sigma t},$$

where $\boldsymbol{x} = [x, y]$ and σ is a scalar-valued function of $\boldsymbol{k}$, and substitute these expressions into Equation 34. This ansatz yields

$$\sigma\tilde{u} = \tilde{u}l^2 - \tilde{u}k^4 + \tilde{\phi}, \tag{35}$$

and

$$\sigma\tilde{\phi} = -a\tilde{\phi} - b\tilde{u}m^2 - c\tilde{\phi}k^2. \tag{36}$$

Thus, the dispersion relation $\sigma(k_x, k_y)$ can be seen as a solution to the eigenvalue problem:

$$\begin{pmatrix} l^2 - k^4 & 1 \\ -bm^2 & -a - ck^2 \end{pmatrix}\begin{pmatrix} \tilde{u} \\ \tilde{\phi} \end{pmatrix} = \sigma\begin{pmatrix} \tilde{u} \\ \tilde{\phi} \end{pmatrix}. \tag{37}$$

Let $b_c(\boldsymbol{k})$ be the maximum value of the parameter b for which $\sigma(\boldsymbol{k}) = 0$. Then for $b = b_c(\boldsymbol{k})$ and $\boldsymbol{k} \neq \mathbf{0}$,

$$\begin{pmatrix} -k^4 - l^2 & 1 \\ -b_c(\boldsymbol{k})m^2 & -ck^2 - a \end{pmatrix} \begin{pmatrix} 1 \\ k^4 - l^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{38}$$

while for $\boldsymbol{k} = \mathbf{0}$ we find

$$\begin{pmatrix} 0 & 1 \\ 0 & -a \end{pmatrix} \begin{pmatrix} u_0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{39}$$

Equations 38 and 39 give rise to left eigenvectors $\boldsymbol{v}_1 \doteq [ck^2 + a, 1]$ and $\boldsymbol{v}_2 \doteq [a, 1]$ respectively which will be required in the nonlinear analysis in Section 3.3.

Equation 37 has solutions

$$\sigma_{\pm}(k_x, k_y) = \frac{T}{2} \pm \sqrt{\frac{T^2}{4} - D}, \tag{40}$$

where

$$\mathcal{T}(k_x, k_y) = r_1 k_x^2 + k_y^2 - k^4 - a - ck^2$$

and

$$\mathcal{D}(k_x, k_y) = -\left(r_1 k_x^2 + k_y^2\right)\left(a + ck^2\right) + b\left(r_2 k_x^2 + k_y^2\right) + k^4\left(a + ck^2\right).$$

Note that $\mathrm{Re}(\sigma)$ gives the rate with which the amplitude of the mode grows (for $\mathrm{Re}(\sigma) > 0$) or attenuates (for $\mathrm{Re}(\sigma) < 0$). We take the square root in Equation 40 to have a nonnegative real part, so that $\mathrm{Re}(\sigma_+) \geq \mathrm{Re}(\sigma_-)$.

The surface of the solid is unstable if and only if $\mathrm{Re}(\sigma_+)$ is positive for some $\boldsymbol{k}$. For $\boldsymbol{k} = \mathbf{0}$, $\sigma_+ = 0$ and $\sigma_- = -a < 0$, and so these modes are not unstable. We may therefore

limit the remainder of the linear stability analysis to nonzero wavevectors $\boldsymbol{k}$. The neutrally stable wavevector $\boldsymbol{k} = \mathbf{0}$ will, however, have an important effect coming from its nonlinear interactions with unstable modes, as we will discuss in Section 3.3.

If k is sufficiently small,

$$\sigma_+(k_x, k_y) = (1 - b/a)k_y^2 + (r_1 - br_2/a)k_x^2 + \mathcal{O}(k^4).$$

Recall that a is positive. Thus, if $b < a$, the steady state is unstable for wavevectors of small magnitude that are oriented along the k_y-axis. The solid surface is unstable for all $b \leq 0$ as a consequence, and so we will restrict our attention to $b > 0$. Similarly, if $r_2 \leq 0$ and $r_1 > 0$, then the steady state is unstable for all $b > 0$ and sufficiently small wavevectors oriented along the k_x-axis. The case in which r_1 and r_2 are both negative can be handled in exactly the same way as the case in which $r_2 > 0$, but the two cases must be dealt with separately. To avoid tedious repetition, we will only consider the case $r_2 > 0$. Thus, for the remainder of the paper, we will restrict our attention to $a > 0$, $b > 0$, $c > 0$ and $r_2 > 0$, with no restrictions on r_1.

The growth rate is positive for a given $\boldsymbol{k}$ if and only if $\mathcal{T}(\boldsymbol{k}) > 0$ or $\mathcal{T}(\boldsymbol{k}) \leq 0$ and $\mathcal{D}(\boldsymbol{k}) < 0$. In the regions of parameter space in which $\mathcal{T}$ is positive for some $\boldsymbol{k}$, the surface is unstable for all choices of b, since $\mathcal{T}$ does not depend on b. We refer to this region of parameter space as Region III.

Outside of Region III, for any choice of a, c, r_1 and r_2 and for sufficiently large b, the surface is stable (i.e., $\mathrm{Re}(\sigma_+) < 0$ for all nonzero $\boldsymbol{k}$). However, when we decrease b below a critical value which depends on a, c, r_1 and r_2, we find that $\mathrm{Re}(\sigma_+)$ becomes positive for some nonzero wavevectors $\boldsymbol{k}$. In particular, there exist unstable regions in a neighborhood

of the wavevectors with the largest linear growth rate

$$\boldsymbol{k} = \pm \left[(cr_1 - a)/2c\right]^{1/2} \hat{\boldsymbol{x}}$$

for choices of $b < b_{\mathrm{I}_x} \doteq (cr_1 + a)^2/4cr_2$, provided that $cr_1 > a$. Similarly, we find unstable regions in a neighborhood of the wavevector

$$\boldsymbol{k} = \pm \left[(c - a)/2c\right]^{1/2} \hat{\boldsymbol{y}}$$

for choices of $b < b_{\mathrm{I}_y} \doteq (c+a)^2/4c$, provided that $c > a$. In these cases, the system undergoes a Turing bifurcation at the critical value of b [72]. This leads to unstable regions in $\boldsymbol{k}$-space that are centered at the critical wavevectors and that are bounded away from the origin. We call these bifurcations I_x transitions if $b_{\mathrm{I}_x} > b_{\mathrm{I}_y}$ and I_y transitions if $b_{\mathrm{I}_y} > b_{\mathrm{I}_x}$. On the other hand, if $cr_1 < a$, the I_x instability does not occur. Instead, the necessary and sufficient condition for instability is $b < b_{\mathrm{II}_x} \doteq r_1 a/r_2$. Similarly, if $c < a$, then there exist unstable wavevectors for $b < b_{\mathrm{II}_y} \doteq a$. These bifurcations will be referred to as II_x transitions if $b_{\mathrm{II}_x} > b_{\mathrm{II}_y}$ and II_y transitions if $b_{\mathrm{II}_x} > b_{\mathrm{II}_y}$. For b just below b_{II_x} or b_{II_y}, there are unstable wavevectors of arbitrarily small magnitude. Figure 3.1 shows the nature of the unstable regions in $\boldsymbol{k}$-space for each of the four types of transitions associated with the four critical values of b.

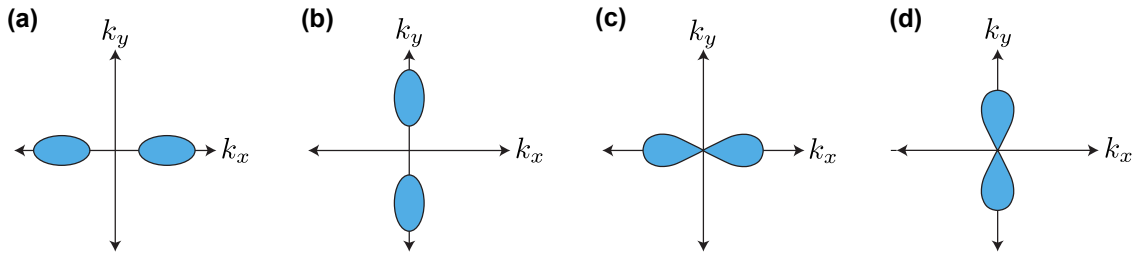

FIGURE 3.1. Typical shape and location of the unstable regions in $\boldsymbol{k}$-space associated with the four types of instabilities: (a) I_x, (b) I_y, (c) II_x, (d) II_y.

Consider a point in the parameter space with coordinates (a, c, r_1, r_2) that is outside Region III and suppose that b has a large enough value that the solid surface is stable. As b is reduced, at some point an instability occurs. If this instability is of Type X, we say that the point (a, c, r_1, r_2) is in Region X, where X can be I_x, I_y, II_x or II_y. Regions I_x and I_y will together form Region I. Similarly, the union of Regions II_x and II_y will be called Region II.

Partitioning the parameter space with coordinates (a, c, r_1, r_2) into regions of the various types is not as straightforward as in the isotropic case ($r_1 = r_2 = 1$). Rather than provide an exhaustive list of cases, we divide up the portion of the $r_1 - r_2$ plane with $r_2 < 1$ by way of example. For example, choosing $c > a$, Region III is the subset of the $r_1 - r_2$ plane $r_2 > 0$ and $r_1 > 2\sqrt{a} + c$, provided that $2\sqrt{a} + c > 1$; otherwise Region III is comprised of all r_2 with $r_1 > 1$. Taking $a/c \leq r_1 \leq 2\sqrt{a} + c$, the I_y transition will always exist (since $c > a$) as we decrease b below $b_{\mathrm{I}_y} = (c + a)^2/4c$. The I_x transition at $b_{\mathrm{I}_x} = (cr_1 + a)^2/4cr_2$ can also occur since $cr_1 > a$. In the region $a/c \leq r_1 \leq 2\sqrt{a} + c$, we observe that $b_{\mathrm{I}_y} > b_{\mathrm{I}_x}$ if and only if $r_2 > [(cr_1 + a)/(c + a)]^2$. If $r_1 < a/c$, then the I_x instability does not occur. We instead compare the relative size of b_{I_y} with b_{II_x} and find that $b_{\mathrm{I}_y} > b_{\mathrm{II}_x}$ if and only if $r_2 > r_1(4ac)/(a + c)^2$. It is easily verified that the line $r_2 = r_1(4ac)/(a + c)^2$ and the parabolic curve $r_2 = [(cr_1 + a)/(c + a)]^2$ are tangent at the value $r_1 = a/c$. These results are summarized in Figure 3.2.

This analysis allows us to identify points in parameter space where transitions between different types of instabilities occur. For example, we identify a switch between Type I_y and Type I_x instabilities along the curve in the $r_1 - r_2$ plane where $r_2 = [(cr_1 + a)/(c + a)]^2$ and $r_1 > a/c$, as shown in Figure 3.2.

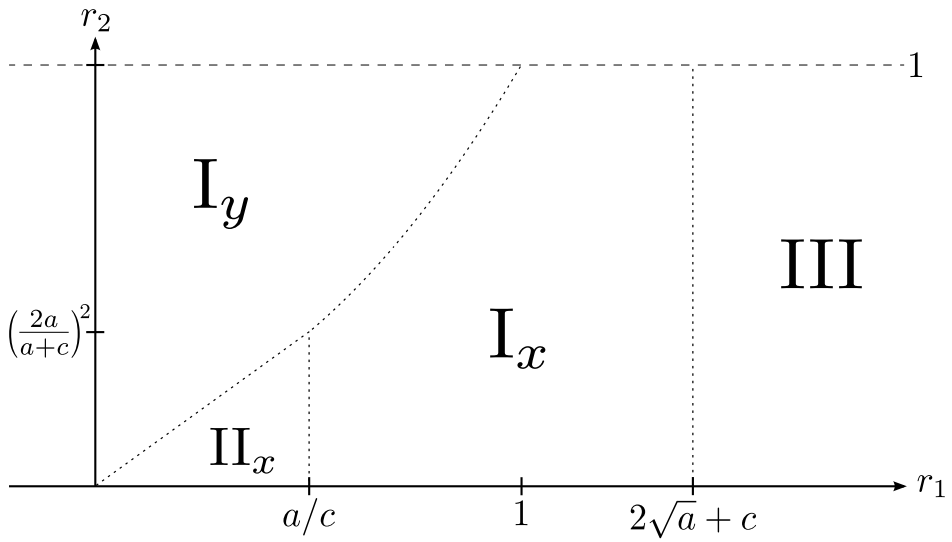

FIGURE 3.2. $r_1 - r_2$ plane partitioned into regions according to transition type assuming $c > a$. For the purposes of illustration, we have fixed c and a so that $2\sqrt{a} + c > 1$.

3.3. NON-LINEAR ANALYSIS

So far, we have developed the linearized equations of motion — we expanded about the steady state $u = \phi = 0$ and retained only terms of first order in u and ϕ. We found that for certain ranges of the parameters a, b, c, r_1, and r_2, the steady state is unstable against small perturbations. Once a linear instability has set in, the deviations from the steady state grow progressively larger and it becomes necessary to include nonlinear terms in the equations of motion. In this section we analyze the full nonlinear Equations 31 and 32 for parameter values a, c, r_1 and r_2 in Region I. The bifurcation parameter b is taken to be slightly below the critical value $b_T \doteq \max\left\{b_{\mathrm{I}_x}, b_{\mathrm{I}_y}\right\}$. We set $b = b_T - \epsilon b_1$, where $\epsilon > 0$ is small and b_1 is positive and of order 1. As discussed in Sec. 3.2, there are then small regions of wavevectors centered about the most unstable wavevectors (those with the largest linear growth rate), namely

$$\boldsymbol{k} = \pm[(cr_1 - a)/2c]^{1/2}\hat{\boldsymbol{x}},$$

if $b_{\mathrm{I}_x} > b_{\mathrm{I}_y}$ or

$$\boldsymbol{k} = \pm[(c-a)/2c]^{1/2}\hat{\boldsymbol{y}},$$

if $b_{\mathrm{I}_y} > b_{\mathrm{I}_x}$, for which the Fourier modes have positive linear growth rate $\mathrm{Re}(\sigma_+(\boldsymbol{k}))$. These modes interact with each other as well as with modes with a small (order ϵ) negative linear growth rate via the nonlinear terms in the equations. We call these interacting modes the *active modes* and denote the corresponding set of wavevectors by $\mathcal{A}$.

The analysis yields nonlinear ordinary differential equations for the time evolution of the amplitudes of the active modes and proceeds as follows: We expand u and ϕ in powers of ϵ and write

$$\boldsymbol{\rho} \doteq \begin{bmatrix} u \\ \phi \end{bmatrix} = \boldsymbol{\rho}_0 + \epsilon\boldsymbol{\rho}_1 + \epsilon^2\boldsymbol{\rho}_2 + \ldots, \tag{41}$$

where

$$\boldsymbol{\rho}_j = \begin{bmatrix} u_j \\ \phi_j \end{bmatrix}, \tag{42}$$

and $u_0 = \phi_0 = 0$ is the uniform steady-state solution. Since the linear growth rate $\mathrm{Re}(\sigma_+)$ is of order ϵ for the wavevectors with largest linear growth rate, we expect that the amplitudes of the unstable modes will evolve slowly in time. We therefore introduce multiple time scales $t_n \doteq \epsilon^n t$ with $n = 0, 1, 2, \ldots$ and treat these as independent variables, so that

$$\frac{\partial}{\partial t} = \frac{\partial}{\partial t_0} + \epsilon\frac{\partial}{\partial t_1} + \epsilon^2\frac{\partial}{\partial t_2} + \ldots. \tag{43}$$

After substituting Equations 41 and 43 into Equations 31 and 32, we collect like powers of ϵ and solve the systems of equations that arise at each order of ϵ.

For each $\boldsymbol{k} \in \mathcal{A}$, we write $b = b_T - \epsilon b_1 = b_c(\boldsymbol{k}) - \epsilon b_1(\boldsymbol{k})$, where $b_c(\boldsymbol{k})$ is the maximum value of b for which $\sigma(\boldsymbol{k}) = 0$ and $b_1(\boldsymbol{k})$ is of order 1. Note that, by definition, b_T is the maximum value of $b_c(\boldsymbol{k})$ for $\boldsymbol{k} \in \mathcal{A}$. Writing

$$L_{\boldsymbol{k}} \doteq \begin{bmatrix} \Delta^2 + \Delta_1 & -1 \\ -b_c(\boldsymbol{k})\Delta_2 & a - c\Delta \end{bmatrix}, \tag{44}$$

and expanding $\boldsymbol{\rho}_1$ as

$$\boldsymbol{\rho}_1 = \sum_{\boldsymbol{k}\in\mathcal{A}} \begin{bmatrix} u_{\boldsymbol{k}}^{(1)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \\ \phi_{\boldsymbol{k}}^{(1)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \end{bmatrix}, \tag{45}$$

at order ϵ the equations are

$$\sum_{\boldsymbol{k}\in\mathcal{A}} \left(\frac{\partial}{\partial t_0} + L_{\boldsymbol{k}} \right) \begin{bmatrix} u_{\boldsymbol{k}}^{(1)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \\ \phi_{\boldsymbol{k}}^{(1)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{46}$$

This system has solutions of the form

$$\boldsymbol{\rho}_1 = \sum_{\boldsymbol{k}\in\mathcal{A}} \begin{bmatrix} 1 \\ k^4 - l^2 \end{bmatrix} \left(A_{\boldsymbol{k}}(t_1) e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} + \text{c.c.} \right) + \begin{bmatrix} G(t_1) \\ 0 \end{bmatrix}. \tag{47}$$

The complex-valued amplitudes $A_{\boldsymbol{k}}(t_1)$ and real-valued amplitude $G(t_1)$ depend on the slow time t_1 but not on t_0. The soft mode with $u_1 = G$ and $\phi_1 = 0$ corresponds to the neutrally stable wavenumber $\boldsymbol{k} = \boldsymbol{0}$; physically, this mode is simply a vertical displacement of the surface. Although this mode is not linearly unstable, it may not be neglected, due to its interactions with the linearly unstable modes through the nonlinear terms in the equations of motion.

Equations for the time evolution of the amplitudes $A_{\boldsymbol{k}}$ and G come about as solvability conditions for the correction $\boldsymbol{\rho}_2$. Writing $\boldsymbol{\rho}_2$ as

$$\boldsymbol{\rho}_2 = \sum_{\boldsymbol{k}\in\mathcal{A}} \begin{bmatrix} u_{\boldsymbol{k}}^{(2)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \\ \phi_{\boldsymbol{k}}^{(2)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \end{bmatrix} \doteq \sum_{\boldsymbol{k}\in\mathcal{A}} \boldsymbol{\rho}_{\boldsymbol{k}}^{(2)} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}, \tag{48}$$

the expansion up to order ϵ^2 reads

$$\begin{aligned} \sum_{\boldsymbol{k}\in\mathcal{A}} L_{\boldsymbol{k}} \boldsymbol{\rho}_{\boldsymbol{k}}^{(2)} \mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} &= \sum_{\boldsymbol{k}\in\mathcal{A}} \begin{bmatrix} \frac{\partial}{\partial t_1} A_{\boldsymbol{k}}(t_1) e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \\ \frac{1}{\epsilon}[b - b_c(\boldsymbol{k})]\Delta_2 \left(A_{\boldsymbol{k}}(t_1) e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}\right) - \frac{\partial}{\partial t_1}(k^4 - l^2) A_{\boldsymbol{k}}(t_1) e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}} \end{bmatrix} \\ &+ \begin{bmatrix} \lambda \left[r_3 \left(\frac{\partial u_1}{\partial x}\right)^2 + \left(\frac{\partial u_1}{\partial y}\right)^2 \right] \\ \nu\phi_1^2 + \hat{\eta}\phi_1^3 \end{bmatrix} \\ &\doteq \sum_{\boldsymbol{k}\in\mathcal{A}} \boldsymbol{q}_{\boldsymbol{k}}^{(2)} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}, \end{aligned} \tag{49}$$

where $\hat{\eta} \doteq \epsilon\eta$ is assumed to be of order 1. This assumption is necessary in order for the cubic term to appear at order ϵ^2 in our analysis.

Collecting coefficients of $e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}$ in Equation 49 gives rise to equations $L_{\boldsymbol{k}}\boldsymbol{\rho}_{\boldsymbol{k}}^{(2)} = \boldsymbol{q}_{\boldsymbol{k}}^{(2)}$ for each $\boldsymbol{k} \in \mathcal{A}$. The operator $L_{\boldsymbol{k}}$ is not invertible, as it has the eigenvector $[1, k^4 - l^2]$ of eigenvalue 0 (see Equation 47). According to the Fredholm Alternative, the equation $L_{\boldsymbol{k}} = \boldsymbol{q}_{\boldsymbol{k}}^{(2)}$ has a solution if and only if $\boldsymbol{q}_{\boldsymbol{k}}^{(2)}$ is orthogonal to the kernel of the adjoint operator $L_{\boldsymbol{k}}^{\dagger}$ [73]. This means that $\left\langle \boldsymbol{v}_{\boldsymbol{k}} | \boldsymbol{q}_{\boldsymbol{k}}^{(2)} \right\rangle = 0$ for $\boldsymbol{k} \in \mathcal{A}$, where

$$\boldsymbol{v}_{\boldsymbol{k}} = \begin{bmatrix} ck^2 + a \\ 1 \end{bmatrix} e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}$$

belong to the kernel of $L_{\boldsymbol{k}}^{\dagger}$. The inner product $\langle \boldsymbol{v} | \boldsymbol{w} \rangle$ of $\boldsymbol{v} = [v_x, v_y] e^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}$ and $\boldsymbol{w} = [w_x, w_y] e^{\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{x}}$ is defined to be the average of $\boldsymbol{v}^* \cdot \boldsymbol{w}$ over all $\boldsymbol{x}$.

The nonlinear terms in Equation 49 contribute to $\boldsymbol{q}_{\boldsymbol{k}}^{(2)}$ in the following ways: For any given $\boldsymbol{k}_1 \in \mathcal{A}$, the quadratic terms proportional to λ and ν give rise to products of the form $e^{\mp \mathrm{i}\boldsymbol{k}_2\cdot\boldsymbol{x}} e^{\mp \mathrm{i}\boldsymbol{k}_3\cdot\boldsymbol{x}} = e^{\mp \mathrm{i}(\boldsymbol{k}_2+\boldsymbol{k}_3)\cdot\boldsymbol{x}}$, which equal $e^{\pm \mathrm{i}\boldsymbol{k}_1\cdot\boldsymbol{x}}$ for wavevectors $\boldsymbol{k}_j \in \mathcal{A}$ such that $\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3 = 0$. The cubic terms give rise to products $e^{\mathrm{i}(\boldsymbol{k}_1 - \boldsymbol{k}_1 + \boldsymbol{k}_1)\cdot\boldsymbol{x}} = e^{\mathrm{i}(\boldsymbol{k}_2 - \boldsymbol{k}_2 + \boldsymbol{k}_1)\cdot\boldsymbol{x}} = e^{\mathrm{i}\boldsymbol{k}_1\cdot\boldsymbol{x}}$. We therefore take $\boldsymbol{\rho}_1$ to be the sum over one triad of modes with wavevectors $\boldsymbol{k}_1$, $\boldsymbol{k}_2$, and $\boldsymbol{k}_3$ such that $\boldsymbol{k}_1 + \boldsymbol{k}_2 + \boldsymbol{k}_3 = 0$. We choose $\boldsymbol{k}_1$ to be the wavevector of the most linearly unstable mode–that is, the wavevector that maximizes $\mathrm{Re}(\sigma(\boldsymbol{k}))$.

The solvability conditions $\left\langle \boldsymbol{v}_{\boldsymbol{k}_j} | L_{\boldsymbol{k}_j} \boldsymbol{\rho}_2 \right\rangle = 0$ with $j = 1, 2, 3$ give three equations for the rescaled amplitudes $A_j \doteq A_{\boldsymbol{k}_j}/\epsilon$, namely

$$\frac{\partial A_j}{\partial t} = \sigma_j A_j + \tau_j A_p^* A_q^* + \gamma_j A_j \left[(k_j^4 - l_j^2)^2 |A_j|^2 + 2(k_p^4 - l_p^2)^2 |A_p|^2 + 2(k_q^4 - l_q^2)^2 |A_q|^2 \right], \tag{50}$$

where $j, p, q \in \{1, 2, 3\}$ are in cyclic order, $l_i^2 = r_1 k_{ix}^2 + k_{iy}^2$ for $i = 1, 2, 3$,

$$\sigma_j \doteq \sigma_+(\boldsymbol{k}_j), \tag{51}$$

$$\tau_j \doteq \frac{-2\lambda(ck_j^2 + a)(r_3 k_{px} k_{qx} + k_{py} k_{qy}) + 2\nu(k_p^4 - l_p^2)(k_q^4 - l_q^2)}{ck_j^2 + a + k_j^4 - l_j^2}, \tag{52}$$

and

$$\gamma_j \doteq \frac{3\eta(k_j^4 - l_j^2)}{ck_j^2 + a + k_j^4 - l_j^2}. \tag{53}$$

Applying the solvability condition $\langle \boldsymbol{v_0} | L_0 \boldsymbol{\rho}_2 \rangle = 0$, yields

$$\frac{\partial G}{\partial t} = \frac{2}{a} \sum_{j=1}^{3} \left[\lambda (r_3 k_{jx}^2 + k_{jy}^2) + \nu (k_j^4 - l_j^2)^2 \right] A_j A_j^* + \frac{6\eta}{a} (k_1^4 - l_1^2)(k_2^4 - l_2^2)(k_3^4 - l_3^2) \mathrm{Re}(A_1 A_2 A_3). \tag{54}$$

By setting $r_1 = r_2 = r_3 = 1$, we find that for $j = 1, 2$, and 3,

$$k_j^2 = \|\boldsymbol{k}_j\|^2 = (c - a)/2c$$

and so

$$(k_j^4 - l_j^2)^2 = \left[\frac{(a + c)(a - c)}{4c^2} \right]^2 .$$

Also, the expression $r_3 k_{px} k_{qx} + k_{py} k_{qy}$ appearing in the coefficient α_j becomes $\boldsymbol{k}_p \cdot \boldsymbol{k}_q = [(c - a)/2c] \cos(2\pi/3) = (a - c)/4c$ because $\boldsymbol{k}_1$, $\boldsymbol{k}_2$, and $\boldsymbol{k}_3$ are uniformly spaced on the circle of radius $[(c - a)/2c]^{1/2}$. Upon substitution of these expressions into Equations 50-53 the amplitude equations reduce to those for the isotropic case [74]:

$$\frac{\partial A_j}{\partial t} = \sigma A_j + \tau A_p^* A_q^* + \gamma A_j \left(|A_j|^2 + 2|A_p|^2 + 2|A_q|^2 \right),$$

where

$$\sigma \doteq \sigma_+(\boldsymbol{k}) = 2(b_T - b) \frac{c(c - a)}{(c + a)(2c^2 + a - c)}$$

$$\tau \doteq \tau_1 = \tau_2 = \tau_3 = \frac{2\lambda c^3 (c - a) + \nu (c + a)(c - a)^2}{4c^2 (2c^2 + a - c)}$$

and

$$\gamma \doteq \gamma_1 = \gamma_2 = \gamma_3 = \frac{3\eta(a-c)(a^2-c^2)^2}{16c^4(2c^2+a-c)}.$$

Without the anisotropy introduced by r_1, r_2 and r_3, there is a range of parameter values at which the solution $A_1 = A_2 = A_3$ is stable, which gives rise to hexagonal order. If we fix $r_2 = 1$, we expect a range of values of r_1 in a neighborhood of $r_1 = 1$ to give rise to patterns of nearly hexagonal order. However, for sufficiently large anisotropy, we expect the hexagonal order will be lost, and ripples oriented either parallel or perpendicular to the ion beam direction will develop. This is suggested by Figure 3.2, which illustrates that, for $r_2 > [2a/(a+c)]^2$, there is a value of r_1 separating the regions I_x and I_y. It is our goal to understand the stability of the solutions to the amplitude equations in the vicinity of this transition. Thus we will focus our analysis on the following steady-state solutions to the amplitude equations (50):

(1) *Homogeneous state*: $A_1 = A_2 = A_3 = 0$. This solution is the undisturbed steady state $u = \phi = 0$.

(2) *Ripple pattern*: $A_1 = \pm\sqrt{-\sigma_1/\gamma_1(k_1^4 - l_1^2)^2}$, $A_2 = A_3 = 0$. These solutions are surface ripples with wavelength $2\pi/k_1$.

(3) *Dots-on-ripples pattern*: $A_1 \geq A_2 = A_3 > 0$.

It is possible to analytically solve for the stationary solutions of the dots-on-ripples pattern. By requiring $A_2 = A_3 > 0$, we find that

$$A_2 = \sqrt{\frac{-\sigma_1 A_1 + \gamma_1 P_1 A_1^3}{\tau_1 + 4\gamma_1 P_2 A_1}} \tag{55}$$

where $P_1 \doteq (k_1^4 - l_1^2)^2$, $P_2 \doteq (k_2^4 - l_2^2)^2$, and A_1 is a solution to

(56) $(5\gamma_2\gamma_1 P_2 P_1)A_1^3 + (4\gamma_1 P_2\tau_2 + 2\gamma_2 P_1\tau_1)A_1^2 + (4\gamma_1 P_2\sigma_2 + \tau_2\tau_1 - 3\gamma_2 P_2\sigma_1)A_2 + \tau_1\sigma_2 = 0,$

noting that $\gamma_2 = \gamma_3$, $\tau_2 = \tau_3$, and $\sigma_2 = \sigma_3$. Thus, solving for the dots-on-ripples stationary solution involves find the roots of a cubic in A_1. Instead of doing this, we fix a set of parameters $r_2, r_3, a, c, \lambda, \nu, \eta$ and $b \doteq .96 b_c$, where $b_c = \max\left\{b_{\mathrm{I}_x}, b_{\mathrm{I}_y}\right\}$. We then integrate the amplitude equations numerically over a range of r_1 values using **ode45**, Matlab's built-in, variable-step Runge-Kutta Method. We also set

(57) $$\boldsymbol{k}_1 = \begin{cases} \sqrt{\dfrac{cr_1 - a}{2c}}\hat{\boldsymbol{x}}, & \text{if } b_c = b_{\mathrm{I}_x} \\ \sqrt{\dfrac{c - a}{2c}}\hat{\boldsymbol{y}}, & \text{if } b_c = b_{\mathrm{I}_y}, \end{cases}$$

(58) $$\boldsymbol{k}_2 = \begin{cases} -\dfrac{1}{2}k_{1x}\hat{\boldsymbol{x}} + \dfrac{\sqrt{3}}{2}k_{1x}\hat{\boldsymbol{y}}, & \text{if } b_c = b_{\mathrm{I}_x} \\ -\dfrac{\sqrt{3}}{2}k_{1y}\hat{\boldsymbol{x}} - \dfrac{1}{2}k_{1y}\hat{\boldsymbol{y}}, & \text{if } b_c = b_{\mathrm{I}_y}, \end{cases}$$

and $\boldsymbol{k}_3 = -\boldsymbol{k}_1 - \boldsymbol{k}_2$. This gives a triad of wavevectors that lie on vertices of a regular hexagon at 120° angles from each other. We will compare this choice with the results of our numerical integrations of Equations 31 and 32 in Section 3.4.

In Figure 3.3 we have fixed $r_2 = 1, r_3 = 1, a = 0.25, c = 1, b = 0.96 b_c, \nu = 1, \lambda = 0, \eta = 10$ and have plotted the stable solutions to the amplitude equations, found via integration with randomly chosen initial conditions $A_1, A_2, A_3 \in (0, 10^{-13})$, for a range of values of r_1 between 0.8 and 1.2. The solution $A_1 = A_2 = A_3$ at $r_1 = 1$ corresponds to hexagonal order in the

isotropic case. As r_1 moves away from unity, the amplitude corresponding to the most unstable wavevector (A_1) increases in magnitude, while the other two amplitudes ($A_2 = A_3$) decrease. We conclude that there is a region around the isotropic case where the dots-on-ripples solution is stable. At $|1 - r_1| \approx .045$ the pure ripple solution appears to become stable.

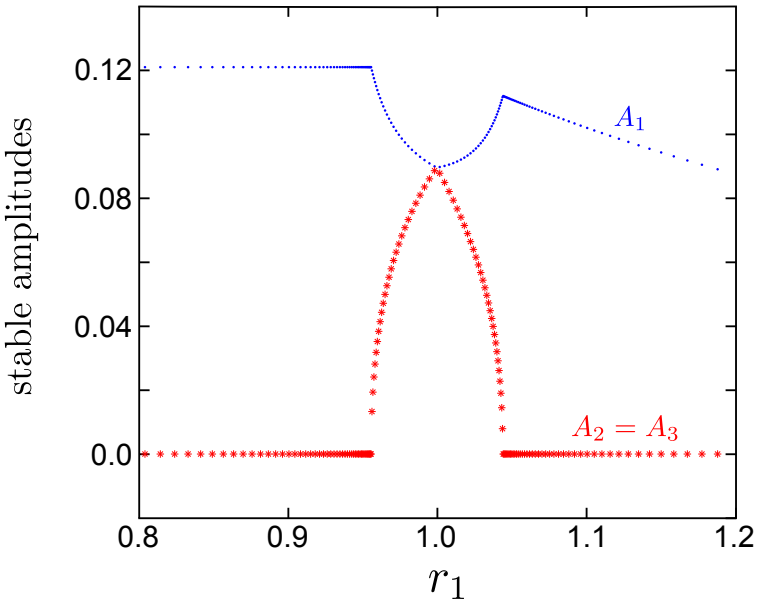

FIGURE 3.3. (color online) A plot of the stable amplitudes A_1 (blue dots) and $A_2 = A_3$ (red stars) for $r_2 = 1, r_3 = 1, a = .25, c = 1, b = .96b_c, \nu = 1, \lambda = 0$, and $\eta = 10$ and a sample of r_1 values between 0.8 and 1.2.

Figure 3.3 suggests that the transition between the ripple pattern and the dots-on-ripples solution is continuous in r_1. In other words, it appears that the ripple pattern loses stability to the dots-on-ripples pattern at the same value of r_1 where the value of the solution with $A_2 = A_3 > 0$ (found by solving Equations 55 and 56) meets the ripple solution $A_2 = A_3 = 0$. A linear stability analysis of the pure ripple solution proves that this is indeed the case. Perturbing the ripple solution by setting $A_1 \doteq R(1 + \delta A_1)$, $A_2 \doteq \delta A_2$, $A_3 \doteq \delta A_3$, where

$R \doteq \sqrt{-\sigma_1/\gamma_1(k_1^4 - l_1^2)^2}$ and linearizing, we find that

$$
\begin{aligned}
&\tfrac{d}{dt}(\delta A_1 + \delta {A_1}^*) = -2\sigma_1(\delta A_1 + \delta {A_1}^*),\\
&\tfrac{d}{dt}(\delta A_1 - \delta {A_1}^*) = 0,\\
&\tfrac{d}{dt}\delta A_2 = \left(\sigma_2 - 2\sigma_1\tfrac{\gamma_2}{\gamma_1}\right)\delta A_2 + \tau_2 R\delta {A_3}^*,\\
&\tfrac{d}{dt}\delta A_3 = \left(\sigma_2 - 2\sigma_1\tfrac{\gamma_2}{\gamma_1}\right)\delta A_3 + \tau_2 R\delta {A_2}^*.
\end{aligned}
\tag{59}
$$

The growth rate eigenvalues for this linear system are $0, -2\sigma_1$, and twice each $\lambda_\pm \doteq \sigma_2 - 2\sigma_1\gamma_2/\gamma_1 \pm \tau_2 R$. Only λ_+ may be positive. In Figure 3.4, we plot λ_+ and the stable solutions to the amplitude equations over a range of values of r_1. Note that the ripple solution does, in fact, become unstable where the dots-on-ripples solution meets the ripple solution. Varying r_2, a, and c yields similar results.

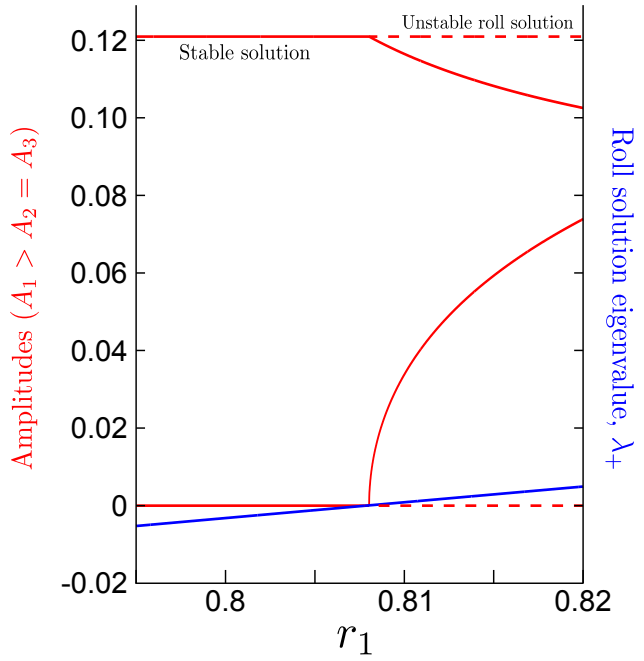

FIGURE 3.4. (color online) A plot of the stable amplitudes (solid red line) for $.795 \leq r_1 \leq .820$ and $r_2 = .75, r_3 = 1, a = .25, c = 1, b = .96b_c, \nu = 1, \lambda = 0$, and $\eta = 10$. Also shown are the unstable ripple solution (dashed red line) and the largest eigenvalue, λ_+, of the ripple solution (solid blue line).

Figure 3.5 shows the function $u_1(\boldsymbol{x}) = \sum_{j=1}^{3} A_j e^{\mathrm{i}\boldsymbol{k}_j \cdot \boldsymbol{x}}$ with amplitudes determined by finding the stable steady-state solutions of the amplitude equations with $r_2 = 0.75$ and for increasing values of r_1. This illustrates the stable patterns predicted by the amplitude equations as we vary r_1 from Region I_y ($r_1 < 0.835$) to Region I_x ($r_1 > 0.835$).

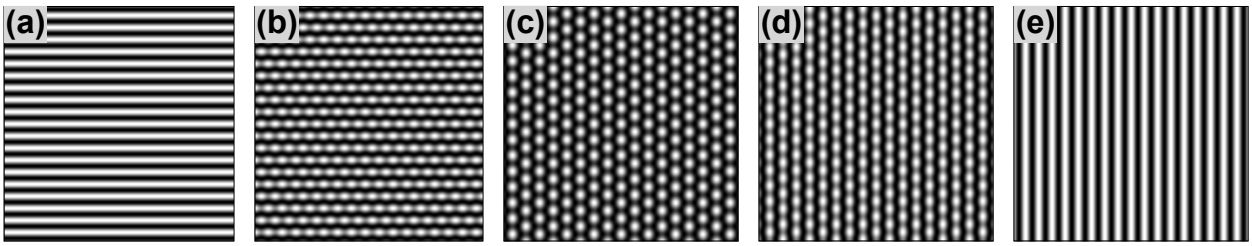

FIGURE 3.5. The function $u_1(\boldsymbol{x}) = \sum_{j=1}^{3} A_j e^{\mathrm{i}\boldsymbol{k}_j \cdot \boldsymbol{x}}$ for $r_2 = 0.75$, $a = 0.25, c = 1, \lambda = 0, \nu = 1, \gamma = 10$ and several values of r_1. For this choice of parameters, $b_{\mathrm{I}_x} = b_{\mathrm{I}_y}$ for $r_1 = 0.835$. **(a)** $r_1 = 0.78, A_1 \approx 0.1209, A_2 = A_3 = 0$, **(b)** $r_1 = 0.81, A_1 \approx 0.1161, A_2 = A_3 = 0.0271$, **(c)** $r_1 = 0.835, A_1 = A_2 = A_3 \approx 0.0895$, **(d)** $r_1 = 0.86, A_1 \approx 0.1123, A_2 = A_3 \approx 0.03706$, **(e)** $r_1 = 0.87, A_1 \approx 0.1209, A_2 = A_3 = 0$.

3.4. Numerical Simulations

We now compare the analytical results with numerical simulations of the full system of partial differential equations, Equations 31 and 32. The numerical technique for these simulations was a Fourier spectral method with periodic boundary conditions on a 256×256 spatial grid, with a fourth-order exponential time differencing Runge-Kutta method for the time stepping. [75, 76] The initial conditions for all simulations was low-amplitude white noise.

Numerical simulations using the same parameter values as for the analytical results of Figure 3.5 are shown in Figure 3.6. As predicted by the amplitude equation analysis, the simulations with $r_1 = 0.81$ (Figure 3.6 (b)) and $r_1 = 0.86$ (Figure 3.6 (d)) show a dots-on-ripples pattern; the ratio A_2/A_1 is 0.76 for the simulation of Figure 3.6 (b) and 0.46

for the simulation of Figure 3.6 (d). The orientation of the ripple patterns is vertical for $r_1 < 0.835$ (Region I_y) and horizontal for $r_1 > 0.835$ (Region I_x). Also consistent with the bifurcation analysis of the amplitude equations, the transition from ripples to a hexagonal pattern proceeds continuously through a dots-on-ripples pattern as r_1 increases or decreases to $r_1 = 0.835$.

Fourier transforms of $u(x, y, \cdot)$ are shown as insets in Figure 3.6. For the simulation of Figure 3.6 (d), the Fourier transform shows that the wavevectors do to a good approximation lie on the vertices of a regular hexagon, in accordance with the choices (57) and (58). However, for the simulation of Figure 3.6 (b), the angles between wavevectors are 57° and 63°; the wavevectors lie slightly off of a regular hexagon. For this simulation, the amplitude ratio $A_2/A_1 = A_3/A_1$ taken from the Fourier transform is 0.76. This is larger than the ratio 0.34 predicted by the amplitude equations with the wavevector choices (57) and (58). The solution to the amplitude equations with wavevectors taken from the Fourier transform of the simulation in Figure 3.6 (b) has a ratio $A_2/A_1 = A_3/A_1 = 0.83$ that is much closer to the value 0.76 obtained in our simulations.

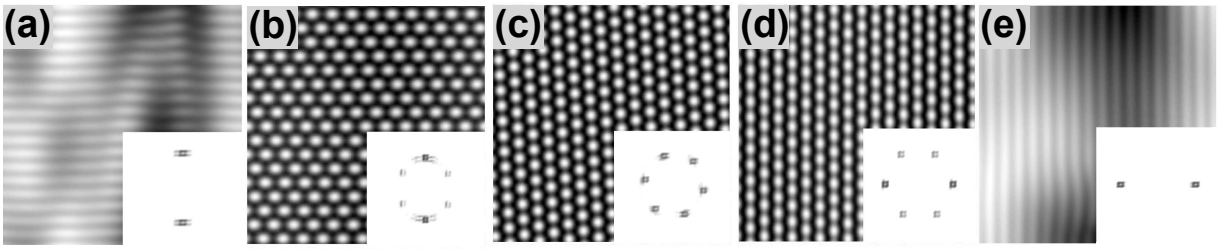

FIGURE 3.6. Numerical simulations of the system Equations 31 and 32 with parameter values $a = 0.25, c = 1, r_2 = 0.75, \lambda = 0, \nu = 1, \eta = 10$, and **(a)** $r_1 = 0.78$ **(b)** $r_1 = 0.81$, **(c)** $r_1 = 0.835$, **(d)** $r_1 = 0.86$, **(e)** $r_1 = 0.87$. For (a,b) $b = 0.96 b_{\mathrm{I}_y}$, for (c), $b = 0.96 b_{\mathrm{I}_x} = 0.96 b_{\mathrm{I}_y}$, and for (d,e), $b = 0.96 b_{\mathrm{I}_y}$. The times are (a) $t = 1500$, (b) $t = 30000$, (c) $t = 50000$, (d) $t = 20000$, (e) $t = 3500$. The magnitudes of the Fourier transforms of u are shown as insets, graphed on the domain $-20 \leq k_x, k_y \leq 20$ in reciprocal space.

In previous work, we showed that Equations 31 and 32 can yield nearly defect-free patterns [71]. In that paper, we chose a greater degree of anisotropy (namely $r_1 = 2$ and $r_2 = 0.2$) than was used in the simulations of Figure 3.6, which show significant defects. A time series of a simulation using the same parameters as in Figure 3.6 (a) is shown in Figure 3.7 (a-d). It reveals that these defects persist; a defect-free ripple pattern is not achieved over time. Note that as the pattern evolves in this simulation, the areas surrounding defects become increasingly deeper than the defect-free zones of the pattern. Equation 54, which governs the soft mode, suggests a reason that the average height of defect-free zones may evolve differently from the average height of defect zones: Near defects, the amplitudes A_1, A_2, and A_3 are close to zero, and the rate of change $\mathrm{d}G/\mathrm{d}t$ of the amplitude of the soft mode in these zones is approximately zero. In zones where a ripple pattern has developed, the amplitude A_1 is nonzero (whereas A_2 and A_3 are approximately zero), so the soft mode evolves according to

$$\frac{\mathrm{d}G}{\mathrm{d}t} \approx \frac{2}{a}\left[\lambda(r_3 k_{1x}^2 + k_{1y}^2) + \nu(k_1^4 - l_1^2)^2\right] A_1 A_1^*.$$

For the parameters chosen in our simulations, $\mathrm{d}G/\mathrm{d}t > 0$, so that the defect-free zones are expected to be higher than defect zones, as observed in the simulations. This difference in average height between zones of defects and zones that are defect-free evidently hinders the roll pattern from invading the defect zones. This suggests that choosing coefficients so as to decrease $\mathrm{d}G/\mathrm{d}t$ should facilitate the formation of defect-free patterns. Decreasing the parameter ν from $\nu = 1$ to $\nu = 0.5$ yields the time series shown in Figure 3.7 (e-h), which does indeed achieve a defect-free state. A similar scenario holds for hexagonal patterning: A defect-free hexagonal pattern is the end result of the simulation of Figure 3.8 (e-h) with

$\nu = 1$, but upon increasing ν to $\nu = 1.5$, defects persist, as shown in Figure 3.8 (a-d). Similar results hold for simulations in which λ is increased in magnitude instead of ν. These simulation results, together with Equation 54, suggest that the soft mode is a major player in the pattern formation, potentially producing a barrier to achieving a defect-free state.

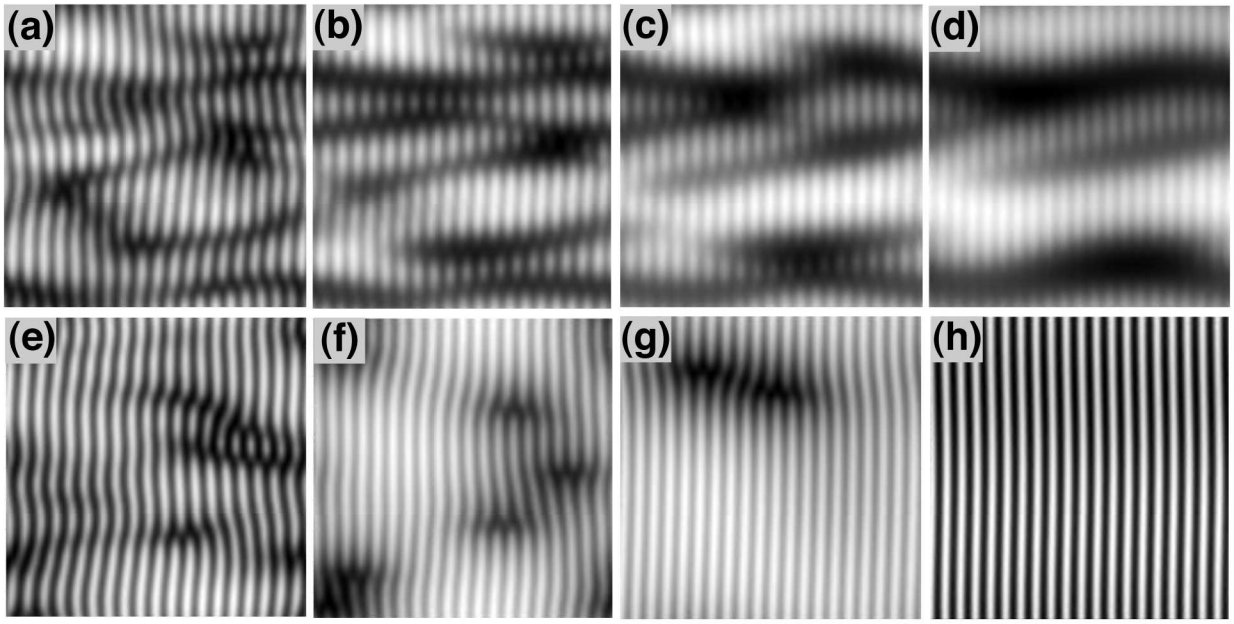

FIGURE 3.7. Numerical simulations of the system Equations 31 and 32 with parameter values $a = 0.25, c = 1, r_1 = 0.78, r_2 = 0.75, \lambda = 0, \eta = 10$, and **(a-d)** $\nu = 1$ **(e-h)** $\nu = 0.5$. The times are **(a,e)** $t = 4000$, **(b,f)** $t = 10000$, **(c,g)** $t = 24000$, and **(d,h)** $t = 50000$.

In the simulations shown in Figure 3.6, the parameter b was chosen to be slightly below threshold; $b = 0.96 b_{\mathrm{I}_x}$ or $b = 0.96 b_{\mathrm{I}_y}$. How does decreasing b further below threshold affect the pattern? Figure 3.9 (a,b,c) show the results of simulations with identical parameter values and times of integration, except that the parameter b decreases from $b = 0.99 b_{\mathrm{I}_x}$ in Figure 3.9 (a) to $b = 0.9 b_{\mathrm{I}_x}$ in Figure 3.9 (b) to $b = 0.8 b_{\mathrm{I}_x}$ in Figure 3.9 (c). More defects are apparent in the pattern for values of b further from threshold. This effect is more evident in simulations with larger coefficients of quadratic nonlinear terms: The simulations of Figure 3.9 (d,e,f) have the same parameter values as those of Figure 3.9 (a,b,c), except that

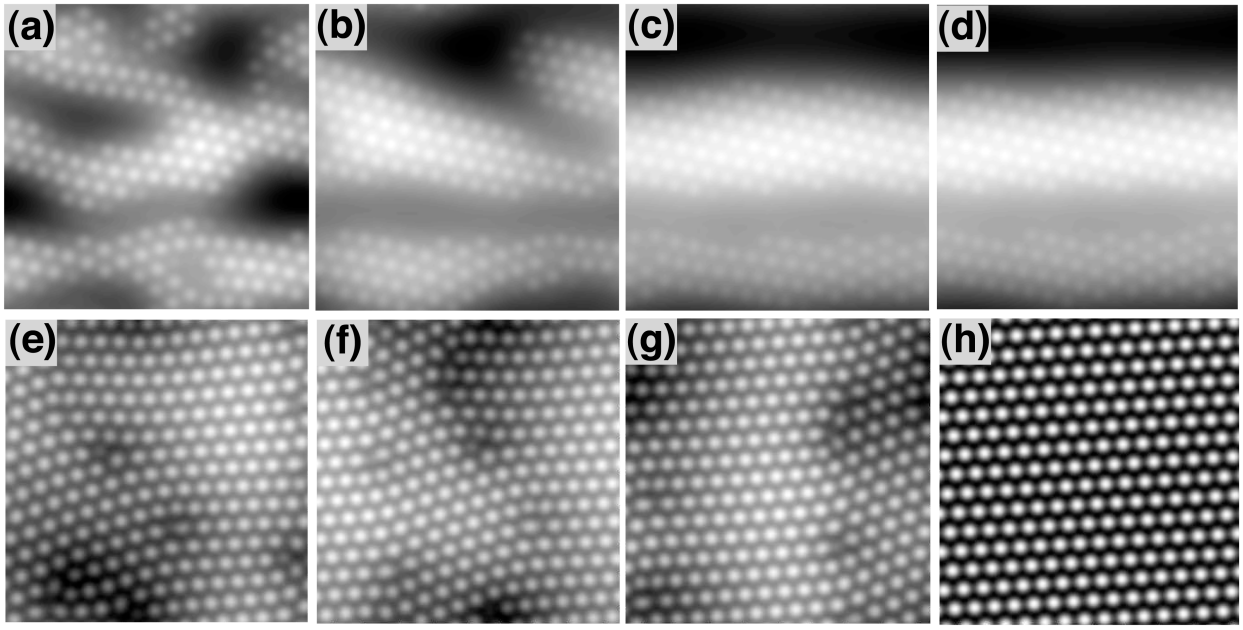

FIGURE 3.8. Numerical simulations of the system Equations 31 and 32 with parameter values $a = 0.25, c = 1, r_1 = 0.835, r_2 = 0.75, \lambda = 0, \eta = 10$, and **(a-d)** $\nu = 1.5$ **(e-h)** $\nu = 1$. The times are **(a,e)** $t = 5000$, **(b,f)** $t = 10000$, **(c,g)** $t = 20000$, and **(d,h)** $t = 50000$.

the parameter ν is increased from 0 to 1. Note that for the simulations of Figure 3.9 (c,f), $b_{\mathrm{I}_x} > b_{\mathrm{I}_y} > b$, so that the set $\mathcal{A}$ includes wavevectors on both the k_x and k_y axes.

Figure 3.10 gives a time-sequence of the surface $u(x, y, \cdot)$ with the same parameters as for the simulation of Figure 3.9 (a), except that $a = 1.5$ instead of $a = 0.25$. As predicted by our linear stability analysis, this gives a II_x instability, and the pattern coarsens with time.

3.5. Conclusions and Future Work

In this paper, we generalized the Bradley-Shipman theory [70, 74] for normal-incidence ion bombardment of a binary material to oblique incidence. We began by constructing the linearized equations of motion. These reduce to the equations of motion introduced by Shenoy, Chan and Chason [77] if the effect of momentum transfer from the incident ions to atoms near the solid surface is negligible.

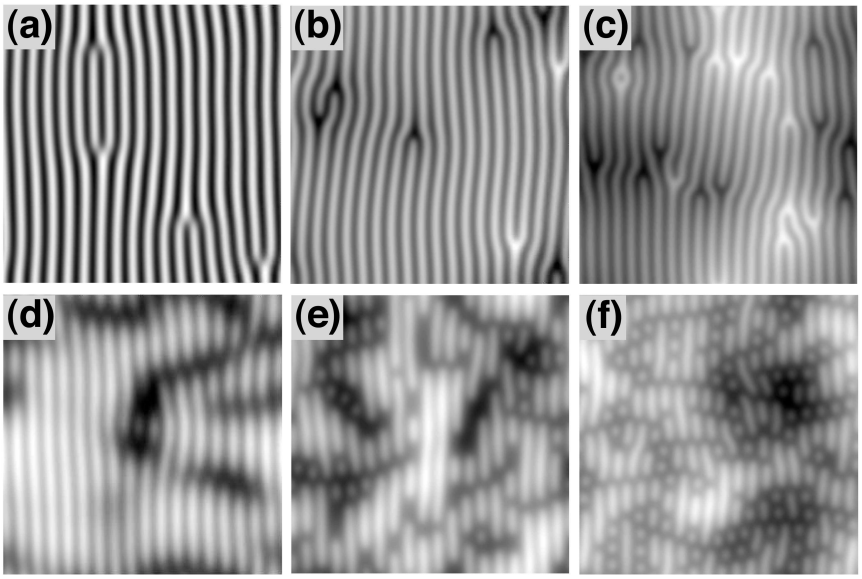

FIGURE 3.9. Numerical simulations of the system Equations 31 and 32 with parameter values $a = 1, c = 1.5, r_1 = 1, r_2 = 0.94, \eta = 10$, and **(a,d)** $b = 0.99 b_{\mathrm{I}_x}$ **(b,e)** $b = 0.9 b_{\mathrm{I}_x}$, **(c,f)** $b = 0.8 b_{\mathrm{I}_x}$, and **(a,b,c)** $\nu = 0$ or **(d,e,f)** $\nu = 1$. The time of integration is $t = 15000$ for all panels.

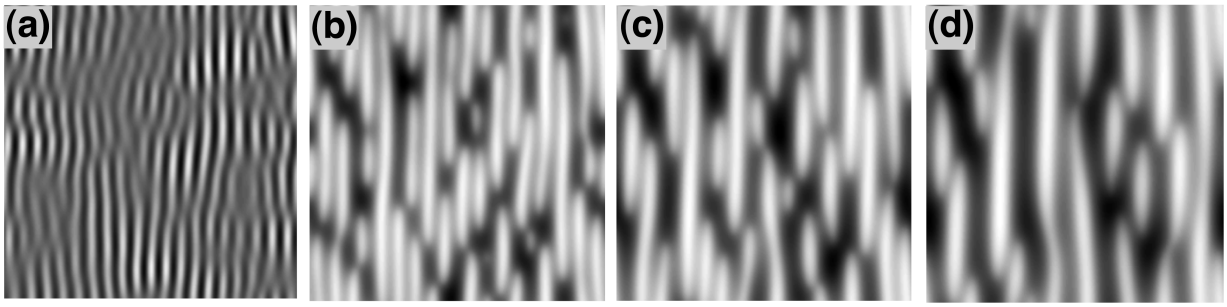

FIGURE 3.10. Numerical simulations of the system Equations 31 and 32 with parameter values $a = 1.5, c = 1, r_1 = 0.835, r_2 = 0.75, \eta = 10$, and $b = 0.96 b_{\mathrm{II}_x}$. The times are (a) $t = 3000$, (b) $t = 8000$, (c) $t = 13000$, and (d) $t = 21000$.

We adopted the same simple models for sputtering, mass redistribution and surface diffusion as BS. If more sophisticated descriptions of these phenomena become available for binary materials in the future, we expect that the form of the linearized equations of motion will remain unchanged, but the relations between the coefficients in these equations and the underlying physical parameters will be altered.

In our linear stability analysis, we showed that there are extended regions in parameter space in which the growth rate $\mathrm{Re}(\sigma(\boldsymbol{k}))$ is positive only in the immediate vicinity of two nonzero points in $\boldsymbol{k}$-space, $\pm\boldsymbol{k}_0$. These points lie on either the x- or the y-axis. In both cases, the theory of pattern formation tells us that patterns with a high degree of order will form.

To determine the kinds of order that emerge, nonlinear terms must be added to the equations of motion. These terms were chosen in precisely the same way as in the BS theory for normal-incidence bombardment of binary materials. Our analysis in the weakly nonlinear regime of Region I_x established that an unusual kind of order can develop for oblique-incidence bombardment – a "dots-on-ripples" pattern. In this type of pattern, ripples of three different orientations that are separated by 120° angles are superimposed. The ripple with its wavevector along the projected beam direction has an amplitude $|A_1|$ that is greater than the common amplitude $|A_2| = |A_3|$ of the other two ripples. Our numerical integration of the original equations of motion confirmed that dots-on-ripples patterns do indeed develop for certain choices of the model parameters. Besides this type of pattern, it is also possible for the surface to remain flat or for parallel-mode ripples to develop.

Our analysis of the amplitude equations we derived and our numerical simulations indicate that if a transition between a dots-on-ripples pattern and a ripple topography occurs as an arbitrary parameter p is varied, it occurs continuously. By this we mean that $|A_2| = |A_3|$ is a continuous function of p that tends to zero as a critical value of the parameter is approached; beyond this critical value $|A_2|$ and $|A_3|$ are identically equal to zero and parallel-mode ripples result.

In a real experiment, we expect that a transition between a dots-on-ripples pattern and parallel-mode ripples will occur as the angle of incidence θ is increased from a small initial

value. As θ is increased, our analysis indicates that the amplitude of the dots will decrease continuously until it drops to zero at a critical value of θ. Parallel-mode ripples will be found for θ exceeding this critical value. A second transition — this time to perpendicular-mode ripples — could be observed as the angle of incidence nears 90°.

Suppose that a highly ordered hexagonal array of nanodots is obtained for normal incidence and for a given ion energy. If the angle of incidence θ is increased, the degree of order may decline. This will occur if the bifurcation parameter b moves further away from its critical value. Therefore, it may be necessary to adjust the beam energy each time θ is increased if the same high degree of order is to be maintained.

Our simulations indicate that smaller values of the coefficients ν and λ of the quadratic nonlinearities favor patterns with fewer defects. For larger values of these coefficients, the portions of the surface that are free of defects are eroded at a very different overall rate than regions containing defects. The resulting wide disparity in surface heights seems to inhibit the spread of the defect-free pattern into defect-laden regions.

The experimental observation of a dots-on-ripples pattern would provide support for the BS theory. Additional support would be garnered if a continuous transition between a dots-on-ripples pattern and parallel-mode ripples was observed. These observations would not rule out the theory of Facsko *et al.*, [78] however, because simulations of the generalization of this theory to oblique-incidence bombardment have produced dots-on-ripples patterns [79]. Whether the transition between a dots-on-ripples pattern and parallel-mode ripples is continuous for that model is not currently known.

Recently, Kang *et al.* bombarded a InSb target with a rastered Ga^+ focused ion beam [80]. Although the beam was normally incident on the initially flat sample surface, Kang *et al.* argued that the rastering leads to an effective local angle of incidence that is nonzero.

Ripples formed at early times, but as the fluence increased, dots appeared on the ripples. The Fourier transform of the surface height at the highest fluence studied suggests that weak hexagonal order may have been present. However, further experimental evidence is needed before it can be concluded that a true dots-on-ripples pattern can be produced in this way.

In 1962, it was observed that nanoscale surface ripples can be produced on a solid surface by bombardment with a broad ion beam [81]. It was not until 1999 that it was discovered that another kind of pattern can develop: hexagonal arrays of nanodots [66]. Square arrays of nanodots were obtained shortly thereafter by bombarding binary materials with concurrent sample rotation [82]. Well-ordered hexagonal arrays of nanoholes were the next new type of pattern to be discovered [83]. Very recently, it was shown how non-overlapping patches of ripples at oblique angles to one another can be produced [84]. Clearly, the range of patterns that can be produced by bombardment with a broad ion beam is growing at an accelerating pace, further enhancing the prospects that ion sputtering will be widely adopted as a nanofabrication tool. The experimental observation of a dots-on-ripples morphology would add another pattern to that already impressive roster.Going forward we would like to identify a range of allowable parameter values which, adjusting the parameters continuously through this range, results in a transition between horizontal and vertical roll patterns. We would also like to explore in greater depth the effect the soft mode has on the severity and longevity of defects in observed pattern. To do this we will perform numerical simulations of a system whose parameter values force the soft mode to be near 0 and compare the evolution of defects to a systems with more active soft modes. We can also derive PDEs analogous to the ODE amplitude equations which allow for spatial variation in the amplitudes.

BIBLIOGRAPHY

[1] R.L. Graham and J.H. van Lint, *On the distribution of $n\theta$ modulo 1*, Canad. J. Math., **20** (1968), 1020–1024.

[2] H. Dym, *On the structure and the Hausdorff dimension of the support of a class of distribution functions induced by ergodic sequences*, Mitre Corp., T.M.-4155 (1965).

[3] G.H. Hardy and E.M. Wright, *"An Introduction to the Theory of Numbers"*, 6th edition, Oxford University Press, New York, 2008.

[4] V.T. Sós, *On the distribution mod 1 of the sequence $\eta\alpha$*, Ann. Univ. Sci. Budapest., Eötvös Sect. Math., 1 (1958), 127–134.

[5] N.B. Slater, *The distribution of the integer N for which $\theta N < \phi$*, Proc. Cambridge Philos. Soc., **46** (1950), 525–537.

[6] J.H. Halton, *The distribution of the sequence $\{n\xi\}(n = 0, 1, 2, \ldots)$*, Proc. Cambridge Philos. Soc., **61** (1965), 665–670.

[7] T. Van Ravenstein, *The three gap theorem (Steinhaus conjecture)*, J. Austral. Math. Soc. (Series A), **45** (1988), 360–370.

[8] C.E. Silva, *"Invitation to Ergodic Theory"*, Amer. Math. Soc., Student Mathematical Library, **42**, 2007.

[9] A. Dykstra and D. Rudolph, *Any two irrational rotations are nearly continuously Kakutani equivalent*, J. Anal. Math., **110** (2010), 339–384.

[10] M. Boshernitzan and J. Chaika, *The Densest Sequence in the Unit Circle*, unpublished.

[11] W. de Melo and S. van Strien, *"One dimensional dynamics"*, Springer Verlag, Ergebnisse (Series 25), 1993.

[12] A. Denjoy, *Sur les courbes définies par les équations différentielles à la surface du tore*, Journal de Mathématiques Pures et Appliquées, **11** (1932), 333–375.

[13] V.I. Arnol'd, *Small Denominators. I: Mapping the Circle onto Itself*, Am. Math. Soc. Transl. (Series 2), **46** (1965), 213–284. (translation from Izv. Akad. Nauk SSSR Ser. Mat., **25(1)** (1961), 21–86.)

[14] M.R. Herman, *Sur la conjugaison différentiable des difféomorphismes du cercle a des rotations*, Publ. Math. I.H.E.S., **49** (1979), 5–233.

[15] J.-C. Yoccoz, *Conjugaison différentiable des difféomorphismes du cercle dont le nombre de rotation vérfie une condition diophantienne*, Ann. Sci. École. Norm. Sup., **17** (1984), 333–359.

[16] Y. Katznelson and D. Ornstein, *The differentiability of conjugation of certain diffeomorphisms of the circle*, Ergod. Th. and Dynam. Sys., **9** (1989), 643–680.

[17] N.G. de Bruijn, *A combinatorial problem*, Nederl. Akad. Wetensch. Proc., **49** (1946), 758–764.

[18] N.G. de Bruijn, *Acknowledgement of priority to C. Fly Sainte-Marie on the counting of circular arrangements of 2^n zeros and ones that show each n-letter word exactly once*, Technical Report, Technische Hogeschool Eindhoven, (1975).

[19] C. Flye Sainte-Marie, *Question 48*, L'intermédiaire des mathématiciens, 1 (1894), 107–110.

[20] M.H. Martin, *A problem in arrangements*, Bull. Amer. Math. Soc., **40** (1934), 859–864.

[21] I.J. Good, *Normal recurring decimals*, J. London Math. Soc., **21** (1946), 167–169.

[22] D. Rees, *Notes on a paper by I.J. Good*, J. London Math. Soc., **21** (1946), 169–172.

[23] V. Becher and P. A. Heiber, *On extending de Bruijn sequences*, Information Processing Letters, **111** (2011), 930–932.

[24] P.D. Shipman, *Discrete and continuous invariance in phyllotactic tilings*, Phys. Rev. E, **81** (2010), .

[25] J.M. Hertzsch, R. Sturman and S. Wiggins, *DNA microarrays: design principles for maximizing ergodic, chaotic mixing*, Small, **3** (2007), 202–218.

[26] A. A. Phillippakis, A. M. Qureshi, M. F. Berger and M. L. Bulyk *Design of Compact, Universal DNA Microarrays for Protein Binding Microarray Experiments*, Journal of Computational Biology, **15**, no. 7 (2008), 655–665.

[27] J. Hadamard, *Resolution d?ne question relative aux determinants*, Bull. des Sci. Math., **17** (1893), 240–246.

[28] A.A. Agaian *Hadamard Matrices and their Applications*, Lecture Notes in Mathematics, **1168**, Berlin, Springer, 1985.

[29] A. S. Hedayat, N. J. A. Sloane and J. Stufken, *Orthogonal Arrays*, Springer Series in Statistics, New York, Springer, 1999.

[30] M. Reck M, A. Zeilinger, H.J. Bernstein and P. Bertani, *Experimental realization of any discrete unitary operator*, Phys. Rev. Lett., **73** (1994), 58–61.

[31] I. Jex, S. Stenholm and A. Zeilinger, *Hamiltonian theory of a symmetric multiport*, Opt. Commun., **117** (1995), 95-101.

[32] D. W. Leung, *Simulation and reversal of n-qubit Hamiltonians using Hadamard matrices*, J. Mod. Opt., **49** (2002), 1199–1217.

[33] R.F. Werner, *All teleportation and dense coding schemes*, J. Phys. A: Math. Gen., **34** (2001), 7081–7094 .

[34] G. Björck, Göran and B. Saffari, *New classes of finite unimodular sequences with unimodular Fourier transforms. Circulant Hadamard matrices with complex entries*, C. R. Acad. Sci. Paris Sér. I Math., **320** (1995), 319–324.

[35] P. Diţă, *Some results on the parametrization of complex Hadamard matrices*, J. Phys. A, **20** (2004), 5355-?374.

[36] K. Beauchamp and R. Nicoara, *Orthogonal maximal abelian $*$-subalgebras of the 6×6 matrices*, Linear Algebra Appl., **428** (2008), 1833–1853.

[37] F. Szöllősi and M. Matolcsi, *Towards a classification of 6×6 complex Hadamard matrices*, Open Syst. Inf. Dyn., **15** (2008), 93–108.

[38] U. Haagerup, *Orthogonal maximal abelian $*$-subalgebras of the $n \times n$ matrices and cyclic n-roots*, Operator Algebras and Quantum Field Theory (Rome), Cambridge, MA International Press, (1996), 296–322.

[39] F. Szöllősi, *Complex Hadamard matrices of order 6: a four-parameter family*, J. London Math. Soc., **85** (2012), 616?632.

[40] J. Williamson, *Hadamard? determinant theorem and the sum of four squares*, Duke Math. J. 11, (1944), 61–81.

[41] W. Tadej, *Permutation equivalence classes of Kronecker Products of unitary Fourier matrices*, Lin. Alg. Appl., **418** (2006), 719-?36.

[42] P. Diţă, *Some results on the parametrization of complex Hadamard matrices*, J. Phys. A: Math. Gen., **37** (2004), 5355–5374.

[43] W. Tadej and K. Życzkowski, *A concise guide to complex Hadamard matrices*, Open Syst. Inform. Dyn. **13** (2006) 13377.

[44] W. Tadej and K. Życzkowski, *Defect of a unitary matrix*, Linear Algebra and its Applications, **429** (2008), 447-81.

[45] R. Craigen, *Equivalence classes of inverse orthogonal and unit Hadamard matrices*, Bull. Austral. Math. Soc., **44** (1991), 109?115.

[46] G. Zauner, *Quantum designs: foundations of a noncommutative design theory*, Int. J. Quantum Inf., **9** (2011), 445–507.

[47] B. R. Karlsson, *Two-parameter complex Hadamard matrices for $N = 6$*, J. Math. Phys., **50** (2009), 082104.

[48] B. R. Karlsson, *H_2-reducible complex Hadamard matrices of order 6*, Linear Algebra Appl., **434** (2011) 239?246.

[49] B. R. Karlsson, *Three-parameter complex Hadamard matrices of order 6*, Linear Algebra Appl., **434** (2011), 247-?58.

[50] F. Szöllősi, *A two-parameter family of complex Hadamard matrices of order 6 induced by hypocycloids*, Proc. Amer. Math. Soc., **138** (2010), 921–928.

[51] T. Tao, *Fuglede's conjecture is false in 5 and higher dimensions*, Math Res. Lett., **11** (2004), 251?258.

[52] G. E. Moorhouse, The 2-Transitive Complex Hadamard Matrices, (preprint)

[53] A. J. Skinner, V. A. Newell and R. Sanchez *Unbiased bases (Hadamards) for six-level systems: four ways from Fourier*, J. Math. Phys., **50** (2009), 012107.

[54] I. Bengtsson, W. Bruzda, Åsa Ericsson, J. Larsson, W. Tadej and K. Życzkowski, *Mutually unbiased bases and Hadamard matrices of order six*, J. Math. Phys., **48** (2007), 052106.

[55] W.K. Wootters and B.D. Fields, *Optimal state-determination by mutually unbiased measurements*, Ann. Physics, **191** (1989), 363–381.

[56] A. Klappenecker and M. Rötteler, *Constructions of Mutually Unbiased Bases*, Lecture Notes in Comput. Sci., Berlin, Springer, 2004, 137–144.

[57] P. Butterley, W. Hall, *Numerical evidence for the maximum number of mutually unbiased bases in dimension six*, Physics Letters A, **369** (2007), 5–8.

[58] Ky Fan and A. J. Hoffman, *Some metric inequalities in the space of matrices*, Proceedings of the American Mathematical Society, **6** (1955), 111–116.

[59] H. Weyl, *Das asymptotische Verteilungsgestez der Eigenwert linearer partieller Differentialgleichungen (mit einer Anwendung auf der Theorie der Hohlraumstrahlung)*, Mathematische Annalen, **71** (1912), 441–479.

[60] L Mirsky, *Symmetric Gage Functions and Unitarily Invariant Norms*, Quarterly Journal of Mathematics, **11** (1960), 50–59.

[61] , Åke Björck and G.H. Golub, *Numerical Methods for Computing Angles Between Linear Subspaces*, Mathematics of Computation, **123** (1973), 579–594.

[62] F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer, 2, 2006.

[63] T. K. Leen, *A Coordinate-Independent Center Manifold Reduction*, Phys. Lett. A, **174** (1993), 89–93.

[64] E. Chason and W. L. Chan, *Spontaneous patterning of surfaces by low-energy ion beams*, Topics Appl. Physics, **116** (2010), 53–71.

[65] B. Ziberi, M. Cornejo, F. Frost and B. Rauschenbach, *Highly ordered nanopatterns on Ge and Si surfaces by ion beam sputtering*, J. Phys. Condens. Matter, **21** (2009), 224003.

[66] S. Facsko, T. Dekorsy, C. Koerdt, C. Trappe, H. Kurz, A. Vogt and H. L. Hartnagel, *Formation of ordered nanoscale semiconductor dots by ion sputtering*, Science, **285** (1999) 1551–1553.

[67] F. Frost, A. Schindler and F. Bigl, *Roughness evolution of ion sputtered rotating InP surfaces: Pattern formation and scaling laws*, Phys. Rev. Lett., **85** (2000), 4116–4119.

[68] M. Castro, R. Cuerno, L. Vázquez and R. Gago, *Self-Organized Ordering of Nanostructures Produced by Ion-Beam Sputtering*, Phys. Rev. Lett., **94**, 016102.

[69] S. Facsko, T. Bobek, A. Stahl and H. Kurz, *Dissipative continuum model for self-organized pattern formation during ion-beam erosion* Phys. Rev. B , **69** (2004).

[70] R. M. Bradley and P. D. Shipman, *Spontaneous Pattern Formation Induced by Ion Bombardment of Binary Compounds*, Phys. Rev. Lett., **105** (2010).

[71] F.C. Motta, P.D. Shipman and R.M. Bradley, *Highly Ordered Nanoscale Surface Ripples Produced by Ion Bombardment of Binary Compounds*, J. Phys. D: Appl. Phys., **45** (2012).

[72] M. Cross and H. Greenside, *Pattern Formation and Dynamics in Nonequilibrium Systems*, Cambridge University Press, Cambridge, England, 2009.

[73] R. Hoyle, *Pattern Formation: An Introduction to Methods*, Cambridge University Press, Cambridge, England, 2007.

[74] P. D. Shipman and R. M. Bradley, *Theory of Nanoscale Pattern Formation Induced by Normal-Incidence Ion Bombardment of Binary Compounds*, Phys. Rev. B, **84**, 085420 (2011).

[75] S. M. Cox and P. C. Matthews, *Exponential time differencing for stiff systems*, J. Comp. Phys. **176**, 430 (2002), 430–455.

[76] R. V. Craster and R. Sassi, *Spectral algorithms for reaction-diffusion equations*, Tech. Report No. 99, Università degli studi di Milano (2006).

[77] V. B. Shenoy, W. L. Chan and E. Chason, *Compositionally Modulated Ripples Induced by Sputtering of Alloy Surfaces*, Phys. Rev. Lett. **98**, 256101 (2007).

[78] S. Facsko, T. Bobek, A. Stahl, H. Kurz and T. Dekorsy, *Dissipative continuum model for self-organized pattern formation during ion-beam erosion*, Phys. Rev. B **69**, 153412 (2004).

[79] S. Vogel and S. J. Linz, *How ripples turn into dots: Modeling ion-beam erosion under oblique incidence*, Europhys. Lett. **76**, no. 5, 884 (2006).

[80] M. Kang, J. H. Wu, W. Ye, Y. Jiang, E. A. Robb, C. Chen and R. S. Goldman, *Formation and evolution of ripples on ion-irradiated semiconductor surfaces*, Appl. Phys. Lett. **104**, 052103 (2014).

[81] M. Navez, C. Sella and D. Chaperot, *Microscopie electronique-étude de lattaque du verrepar bombardement ionique*, C. R. Acad. Sci. **254**, 240 (1962).

[82] F. Frost, B. Ziberi, T. Höche and B. Rauschenbach, *The shape and ordering of self-organized nanostructures by ion sputtering*, Nucl. Inst. and Meth. Phys. Res. B, **216** (2004), 9–19.

[83] Q. Wei, X. Zhou, B. Joshi, Y. Chen, K.-D. Li, Q. Wei, K. Sun and L. Wang, *Self-Assembly of Ordered Semiconductor Nanoholes by Ion Beam Sputtering*, Adv. Mater., **21** (2009), 2865–2869.

[84] S. A. Mollick, D. Ghose, P. D. Shipman and R. M. Bradley, *Anomalous patterns and nearly defect-free ripples produced by bombarding silicon and germanium with a beam of gold ions*, Appl. Phys. Lett. **104**, 043103 (2014).