Problems in Discrete and Continuous Geometry

Tony Zeng

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Stefan Steinerberger, Chair

Ricky Liu

Rekha Thomas

Program Authorized to Offer Degree:

Mathematics

University of Washington

Abstract

Problems in Discrete and Continuous Geometry

Tony Zeng

Chair of the Supervisory Committee:

Stefan Steinerberger

Department of Mathematics

This dissertation summarizes our work in various problems that are inherently geometric or geometrically motivated. The unifying theme of these problems is that they are easily stated but rife with sizable mathematical complications tied to various different areas. In Chapter 1, we consider the decomposition of graphs into a union of regular graphs. Chapter 2 explores variants of a dynamical system that stabilizes into a set of concentric circles. In Chapter 3, we study the geometry of certain matrices and how it acts on the $\{\pm 1\}^n$ cube. Chapter 4 serves as a collection of smaller scale projects that consist of investigating the Hausdorff dimension of a one-parameter family of iterated function systems and the similarity of curves to their evolutes and evolutoids. In Chapter 5, we explore the combinatorial game Juniper Green and its connection to the Gallai-Edmonds Decomposition. We conclude in Chapter 6 with a numerical evidence for a conjecture that a stellated tetrahedron is not Rupert.

TABLE OF CONTENTS

INTRODUCTION

Acknowledgments

I am of course grateful to my advisor, Professor Stefan Steinerberger for his unwavering support for the many problems I have studied and his invaluable insight and guidance in areas that do not necessarily fall under his primary research interests. He encouraged me to pursue a graduate education in mathematics during my undergraduate studies and provided much advice even before we worked together. Over the years, he has told many stories about other mathematicians, providing me a view of the history of mathematics and reminding me that our field, which has a reputation of being cold and austere, is, nevertheless, full of life.

A number of thanks are owed to those who guided and helped me learn about and appreciate mathematics beyond the grade school classroom: Guanghua Chen who strengthened my problem solving abilities through competition training, Chris Michael who supported me in all my mathematical endeavors throughout high school, and Vinton Geistfeld who mentored me during my first foray into math research. That project in fact eventually resulted in my first publication with my undergraduate advisor Pat Devlin, who made the decision to seek a PhD an easy one to make.

My fellow graduate students form a community that has made the past five years a wonderful time that I will forever cherish. There are the Plushie Pirates, who stood victorious over our opponents. We will remember our fallen computers, whose battery lives collapsed when running code for too many hours. There are the members of the gaming community who are inspiring in both skill and creativity. And there are all those who commune in the department lounge discussing any myriad of topics be they mathematical or not, bringing sorely needed renewed life to a department in a post-pandemic world. The friends I have made here have pushed me to be a better mathematician and person.

Over the past two and half decades, my parents have shown unfaltering patience and wisdom as they have helped shape me into the person I am today. I would not be here today if not for them. Finally, I give my most heartfelt thanks to Cherlin Zhu for being a part of my life.

Summary

A fox knows many things, but a hedgehog knows one big thing. (Archilochus)

It is common for those pursuing academia in mathematics to focus their attentions in a specific narrow direction. The breadth and depth of our subject makes it difficult, though not impossible, to become an expert in multiple subfields. I have found that I am motivated not by any particular area of mathematics, but rather by specific problems independent of the subject area. Consequently, my experience has been a broad rather than deep foray into research. This thesis, in a reflection of my nature and interests, includes a number of distinct problems, and they cannot be unified under the umbrella of any one single field. They do, however, share the fundamental property of being geometrically inspired and motivated.

Chapter 1 discusses a notion of graph boundary that is, to the best of our knowledge, novel. Note that when we refer to the boundary of a graph, we do not mean the boundary of a subset of the vertex set, which is well established. Rather, we are interested in almost viewing a graph as a topological space and considering whether there is an analogous definition of a boundary. Others have discussed this, and a number of other definitions do exist [80, 25]. These definitions, however, require global knowledge of the graph. We provide an entirely local definition of graph boundary which is fast to compute and has connections to the computation of graph edit distance and regularity.

The early half of **Chapter 2** explores the properties of a particular dynamical system whose trajectories seem to always converge to concentric circles. We show that if a trajectory stays in a set of concentric circles for sufficiently long, the orbit remains in that set of concentric

circles. We also explore a notion called ambiguity, which measures how close the system is to being undefined, in a sense. Regions of positive ambiguity in the parameter space of this dynamical system correspond to certain equivalence classes of trajectories. These regions we show are the finite union of finite intersections of circles and half spaces.

The later half of **Chapter 2** and **Chapter 3** cover work done in collaboration with undergraduate students through the math department's Washington Experimental Math Lab (WXML) program. Variants of the dynamical system in **Chapter 2** are explored with connections to the Farey sequence and we show a certain kind of Gaussian convergence in a randomized setting. Motivated by fair statistical tests producing false positive results, **Chapter 3** studies extreme behaviors of a certain class of matrices on all $\{\pm 1\}^n$ vectors.

In **Chapter 4**, we discuss several projects that are either smaller in scale or did not result in papers for various reasons. They were, however, significant in our studies and worthy of inclusion. We explore the evolute and its seemingly forgotten cousin, the evolutoid. Studied since antiquity by Apollonius, there a number of curves that are similar to their evolutes. These curves are the members of the cycloid family and the logarithmic spiral. One can observe that the evolutoids of these curves are also similar. We provide a sufficient condition for which one of these similarities implies the other and find other curves with these properties. This work was independent, but we later discovered that our results are not novel[55]. Then, we move on to a particular one-parameter family of iterated function systems inspired by a particular variant of the Chaos Game. In a subset of the parameter space, we are able to apply the Open Set Condition to show that the Hausdorff dimensions of the attractors are constant over this subset. Outside this subset, however, we suspect that the Open Set Condition does not hold and in fact the family of iterated function systems may suffer from overlap.

Chapter 5 describes our study of a particular combinatorial game popularized by Ian Stewart known as Juniper Green [85]. It is known that for all large n, the first player has a forced win. We considered extending the game to the Gaussian Integers and concluded that a similar statement should still hold. We then discovered that the original game has some interesting properties, which are, as far as we are aware, unexplored. While most work in combinatorial games simply concerns which player has a forced win, because the move set of Juniper Green is $[n]$, we are interested in which opening moves allow the first player to win and how many such moves there are. This set and quantity evolve in interesting ways with respect to n, but a concise statement describing this behavior yet evades us. We do, however, show that the set of winning moves corresponds to part of the Gallai-Edmonds Decomposition.

Finally, **Chapter 6** explores the Rupert property of convex polyhedra, a topic which gained a significant amount of attention in late 2025 when a preprint [84] came out claiming to disprove the longstanding conjecture that all convex polyhedra are Rupert. The polyhedron constructed by Steininger and Yurkevich has nice properties but 90 vertices and thus rather complicated. We propose an 8 vertex polyhedron as a candidate and provide some numerical evidence that eliminates most pairs of orientations from yielding a valid Rupert passage.

Papers and Preprints

Some of the work presented in this thesis has been published or is in preparation for submission.

- Tony Zeng. Graph decomposition via edge edits into a union of regular graphs. *Discrete Mathematics*, 349(1), 2026.
- Stefan Steinerberger, Tony Zeng. A curious dynamical system in the plane. [https://arxiv.org/abs/2409.08961]
- Alex Albors, Hisham Bhatti, Lukshya Ganjoo, Raymond Guo, Dmitriy Kunisky, Rohan Mukherjee, Alicia Stepin, Tony Zeng. On the structure of bad science matrices. *Involve*, To appear.
- Alex Albors, François Clément, Shosuke Kiami, Braeden Sodt, Ding Yifan, Tony Zeng. Approximately Jumping Towards the Origin. [https://arxiv.org/abs/2412.04284]
- Tony Zeng. Winning openings of Juniper Green and the Gallai-Edmonds Decomposition [https://arxiv.org/abs/2604.19721]
- Tony Zeng. A stellated tetrahedron that is probably not Rupert [https://arxiv.org/abs/2604.26531]

Chapter 1

GRAPHS

1.1 Introduction and Results

One might ask the simple question of what the "boundary" of a particular graph is. At first glance, this question may seem a bit strange. Often, the notion of an object's boundary requires the object to live within some ambient space. A graph, however, has no such thing. Indeed, for a given graph G, $G \cap G^C$ is completely meaningless. That said, when we are in an enclosed room with no windows, we do not require an understanding of the outside universe to recognize the boundary of the room.

To do so, we will adopt the notation used by Steinerberger in [80]. One notion of boundary devised by Chartrand et al. in [25] is defined as

$$(\partial G)^* := \{u \in V \mid \exists v \in V \ \forall (u,w) \in E : d(u,w) \leq d(u,v)\}.$$

This definition has a couple of properties which we find undesirable. First, on some rather tame graphs, this boundary would conflict with human intuition. One can check that for G the grid graph, $(\partial G)^*$ consists of the four corner vertices, whereas we would prefer the boundary to also include the edge vertices. We remark that this is an entirely subjective and aesthetic judgment, though we think a reasonable one. Second, this definition is global. In order to check if any particular vertex u is in the boundary, one must a priori check the above condition on all $v \in V$. Being in the boundary of a space should be a local condition that can be verified simply by inspecting the neighborhood.

Another notion of boundary defined by Steinerberger in [80] is

$$\partial G = \left\{ u \in V \mid \exists v \in V : \frac{1}{\deg(u)} \sum_{(u,w)\in E} d(w,v) < d(u,v) \right\},$$

an adjustment of $(\partial G)^*$ via averaging distances to v over the neighborhood of u. This definition improves over the Chartrand-Erwin-Johns-Zhang boundary in that ∂G for the grid graph is indeed the set of all edge and corner vertices. Unfortunately, this definition is again global. We would like an alternative definition to address both of these issues.

Figure 1.1: The boundaries $(\partial G)^*$ (left) and ∂G (right) of a particular grid graph G.

To do so, we draw on some intuition for (nice) subsets of $\mathbb{R}^n$. For a point x in the interior of some set $S \subset \mathbb{R}^n$, ε-balls centered on x have "full volume" for sufficiently small ε. For y on the boundary of S, ε-balls centered on y will not have "full volume" regardless of how small ε is. In order to relate this idea to graphs, we consider the cardinalities of neighborhoods as analogous to the volumes of ε-balls in order to attempt to define a notion of boundary. In some sense, this was unsuccessful, as our proposed definition does not satisfy any kind of isoperimetric inequality. In other ways, it does provide a satisfying notion of boundary and has interesting properties relating to degree and regularity.

The above "drawbacks" of $(\partial G)^*$ and ∂G have been presented merely as motivations for the main definition of this chapter. There are other properties of those definitions that make them better candidates for a graph boundary than ours. For more details, we refer the reader to their original manuscripts.

1.1.1 Graph decompositions

If $G = (V, E)$ is a graph obtained by taking some disjoint regular graphs (possibly of different degrees) with many vertices, adding a few edges to connect these components, and then possibly removing a few edges, G is still 'approximately' regular in the sense that most vertices $v \in V$ have the same degree as all their neighbors $u \in N(v)$ (see Figure 1.2).

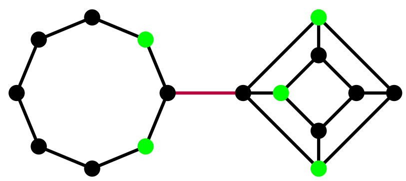


Figure 1.2: A 2-regular and 3-regular graph connected by an edge (in purple). Deleting this edge yields a union of regular graphs. The graph has 5 irregular vertices (shown in green). Later diagrams will color irregular vertices in green, deleted edges in red, and added edges in blue. We will study whether having a small number of irregular vertices is indicative of G being close to a union of regular graphs.

We define a vertex $v \in V$ to be an **irregular vertex** if it has a neighbor w with $\deg(w) > \deg(v)$. As an additional piece of terminology, we will speak of **modifying** an edge which will mean that we add or remove an edge. Throughout the paper, all our graphs are simple, i.e. we do not consider multiple edges between the same pair of vertices. If we start with a regular graph and modify some edges, that is add or remove some edges, the number of irregular vertices that arise in the modified graph can be bounded from above by the number of modifications.

Fact. *Let $G = (V, E)$ be a d-regular graph. Let G' be G after adding or removing k edges. G' has at most $2dk$ irregular vertices.*

We emphasize that the bound is completely independent of the actual size of the graph $|V|$. This is not too surprising: adding or removing edges has a local effect on the degrees of vertices and their neighbors. We are interested in understanding whether some kind of converse statement is true.

> **Question.** Suppose $G = (V, E)$ is a graph with a small number of irregular vertices. Can G, by adding or removing a few edges, be turned into a disjoint union of regular graphs?

We show that results of this type indeed exist: a small number of irregular vertices is indicative of G being close (up to adding or deleting a few edges) to a disjoint union of regular graphs (possibly of different degrees).

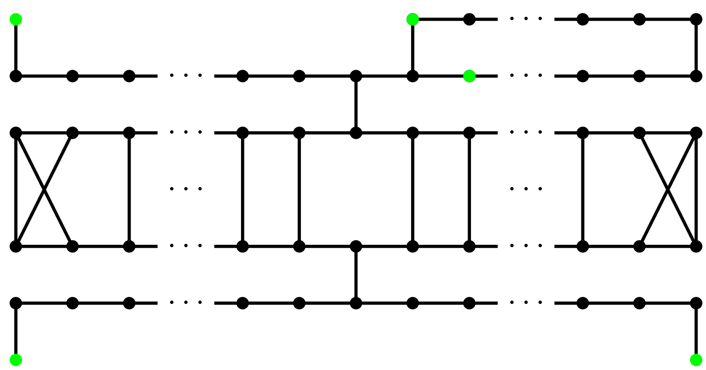


Figure 1.3: A family of graphs G with maximal degree 3 and 5 irregular vertices (green). Can it be decomposed into a union of regular graphs by adding or deleting few edges? Can this be done in a number of edits independent of $|V|$?

To make things precise, we will introduce a metric $d(G, H)$ on graphs. In our setting, each graph G is uniquely determined by its adjacency matrix $A \in \mathbb{R}^{n \times n}$, where $A_{ij} = 1$ if $(i, j) \in E$ and $A_{ij} = 0$ otherwise. We use adjacency matrices to induce a natural metric on graphs: if G and H are two graphs on n vertices, we could wonder in how many places their adjacency matrices A_1, A_2 differ. We never care about the actual labeling of the vertices, so the natural definition is thus up to a permutation π of the entries of the matrix, and therefore

$$d(G,H) = \frac{1}{2}\min_{\pi\in S_n}\sum_{i,j=1}^{n} |(A(G))_{i,j} - (A(H))_{\pi(i),\pi(j)}|$$
$$= \frac{1}{2}\min_{\pi\in S_n}\|A(G) - P_\pi A(H) P_\pi^T\|_F^2,$$

where P_π is the permutation matrix associated to π and $\|\cdot\|_F$ denotes the Frobenius norm, i.e.

$$\|A\|_F^2 = \sum_{i,j} A_{i,j}^2.$$

Notice that $\|A(G) - A(H)\|_F^2$ counts the number of "edge disagreements" between G and H. More precisely, assuming G, H share the same vertex set, $\|A(G) - A(H)\|_F^2$ counts the number of pairs of vertices v_1, v_2 such that (v_1, v_2) is an edge in G but not in H or vice versa. One can think about this expression as providing instructions for deleting or adding edges to get from G to H.

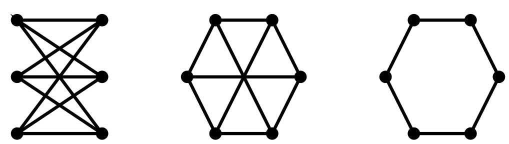


Figure 1.4: The standard embedding of $K_{3,3}$ in the plane (left) can be seen as equivalent to an unusual embedding (middle), which allows for a nice argument that $d(K_{3,3}, C_6) = 3$.

It is easy to see that this induces a metric on the graphs on n vertices. As an example, $d(P_n, C_n) = 1$, where P_n denotes the path graph on n vertices and C_n denote the cycle graph on n vertices. Notice that though P_n and C_n are close in this metric, they are very different graphs. As another example, consider $K_{3,3}$ and C_6. Since $K_{3,3}$ has 9 edges and C_6 has 6 edges, they must be at least 3 apart in this metric. It turns out that they are exactly a distance of 3 apart. The standard drawing of $K_{3,3}$ makes this not so clear, but $K_{3,3}$ can actually be visualized as C_6 with edges added between antipodal vertices, which we show in Figure 1.4. With this metric, we can now illustrate the spirit of our main result in a very simple special case.

Proposition 1. *Let G be a graph with maximal degree $\Delta = \max_{v \in V} \deg(v) = 3$. If G has k irregular vertices, then there exists a graph H that is the union of regular graphs which satisfies*

$$d(G, H) \leq 4k + 10.$$

It is immediately clear that this result is optimal up to constants, as there must be some modification affecting each irregular vertex. It is also clear that, in general, these constants are going to depend on the maximal degree, which is necessary since the maximal degree controls how many neighboring vertices can be affected by modifying an edge.

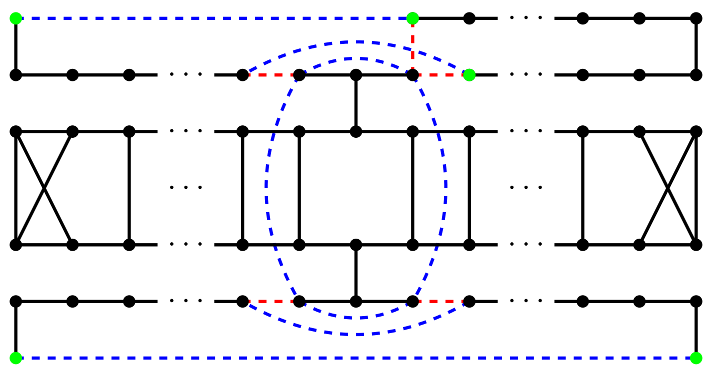


Figure 1.5: Proposition 1 applied to the example from Figure 1.3. Adding the blue edges and deleting the red edges shows that there exists a union of regular graphs H with $d(G, H) = 13$. We note that the distance is independent of the number of vertices $|V|$.

1.1.2 Main Result

Our main results states, informally, that if a graph G has few irregular vertices and a sufficient number of vertices of small degrees then there is a graph H that is the union of regular graphs that is close to G.

Theorem. *Suppose we have a graph $G = (V, E)$ whose vertices have maximal degree Δ. If G*

has k irregular vertices and at least $2d+d^2+d^3+k\cdot(\Delta!)$ vertices of degree d for $3<d\leq\Delta$, then there exists a union of regular graphs H satisfying

$$d(G,H)\leq 4(\Delta+1)!\cdot k+5\Delta^2.$$

Remarks.

1. The statement does not in any way depend on the total number of vertices $|V|$, as all conditions and conclusions only depend on Δ and k.

2. Simple examples show that the best one can hope for is a linear estimate in k and in this sense our result is optimal. However, how the constants depend on Δ is surely suboptimal. It is easy to see that in the statement $d(G,H)\leq Ak+B$, one needs the constant A to grow at least linearly in Δ. It is not clear to us what one might hope to expect.

3. It is an interesting question whether the assumption of having sufficiently many (depending on Δ and k) vertices of 'intermediate' degree, or some variant of it, is necessary or could be weakened.

4. A related question is whether there are any good heuristics to produce H. Following the proof strategy that will be used to establish Proposition 1 and applying it to the example in Figure 1.3 leads to the outcome illustrated in Figure 1.5. However, in that special case, a much closer graph (shown in Figure 1.6) exists.

1.1.3 Related results.

We are not aware of any directly related results. One can show that the metric d on graphs used above is equivalent to graph edit distance, which has been studied in applications such as pattern recognition, unsupervised neural networks, and similarity of seriated graphs. We refer to the survey [38]. Bunke describes a notion of graph edit distance whose computation is equivalent to the maximum common subgraph problem [23]. It has been shown that graph edit distance computation itself is NP-hard [96] though some work has been done toward approximations and exact computations [27, 68, 73, 75]. One way of describing our problem is to view graphs as almost component-wise regular but for a few misplaced edges. In this sense, a component-wise regular graph H close to G can be viewed as an edited decomposition of G into regular components. Graph decompositions typically refer to edge or vertex decompositions, possibly into regular or irregular structures [4, 5, 14, 21, 65, 71]. A key difference between these works and ours is that we are allowed to add or delete a small number of edges.

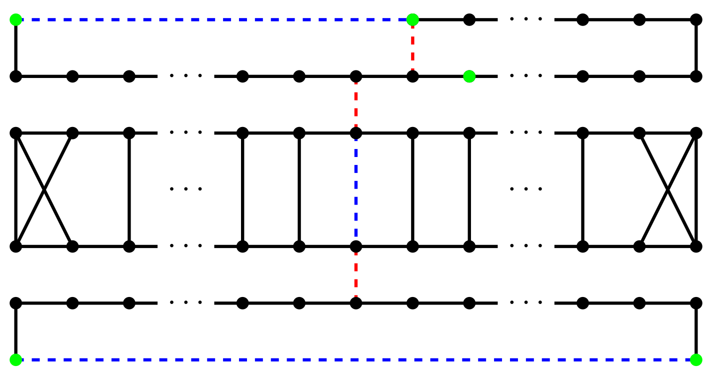


Figure 1.6: Another example of a graph H, a union of regular graphs, this time with $d(G, H) = 6$, where G is the example from Figure 1.3.

1.2 *Proof of Proposition 1*

1.2.1 *Preparatory Results.*

Our proofs use the following terminology and structure. Given a graph G, one can obtain a graph H by making a sequence of k edge addition or deletions on G. Then it must be that $d(G, H) \leq k$, as we can iteratively apply triangle inequality. When counting edge additions or deletions, we use the word 'edits'. We address Proposition 1 by considering maximal degree 3 graphs in 3 steps.

- Graphs with at most two degree 2 vertices and no degree 1 vertices.

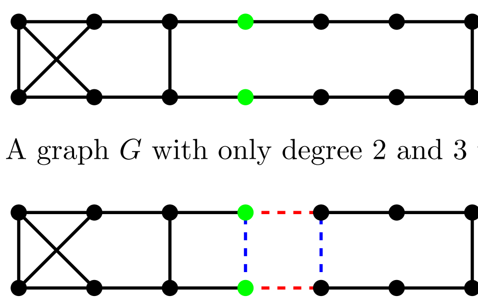
- Graphs with arbitrarily many degree 2 vertices and no degree 1 vertices.
- Graphs with arbitrarily many degree 2 and degree 1 vertices.

Lemma 1. *Suppose we have an almost regular graph G where at most 2 vertices have degree 2, and all other vertices have degree 3. Then there exists a union of regular graphs H such that $d(G, H) \leq 5$.*

Proof. If we have no degree 2 vertices, there is nothing to do. If we have a single degree 2 vertex, delete its 2 edges, thereby reducing to the case with two degree 2 vertices, as the disconnected vertex now forms a 0-regular component. This deletion costs 2 edits. Now suppose G has 2 vertices v_1, v_2 of degree 2. If they are not connected, add an edge between them, and we have arrived at a component-wise regular graph a distance of at most 3 (1 edge addition and possibly 2 edge deletions) away from G. Otherwise, the idea will be to delete a sufficiently far away edge (u_1, u_2) and add edges (u_1, v_1) and (u_2, v_2). Suppose we have $2m$ degree 3 vertices. Compute that G has $(2m \cdot 3 + 2 \cdot 2)/2 = 3m + 2$ edges. We may ignore $m = 1$ since the only such graph (sometimes called the diamond graph) falls under the previous case. Let u_1, u_2 be the respective higher degree neighbors of v_1, v_2. If we enumerate the edges incident to these labeled vertices, we have $(v_1, v_2), (u_1, v_1), (u_2, v_2)$, and at most four other edges incident to u_1 and u_2. Since we need only consider $m \geq 2$, G has at least 8 edges. In particular, there exists an edge in G not incident to any of the previously labeled vertices. Let this edge be (w_1, w_2). Delete it and add edges (v_1, w_1) and (v_2, w_2). Since v_1, v_2 were the only vertices of degree 2, this yields a 3-regular graph, costing 3 edits. In total, we have spent at most 5 edits. □

Lemma 2. *Suppose we have an almost regular graph G with no vertices of degree greater than 3, at least 2 vertices of degree 3, at least 1 vertex of degree 2, and k irregular vertices. Then there exists a union of regular graphs H such that $d(G, H) \leq 3k + 5$.*

Proof. Let U be the set of irregular vertices that have a degree 2 neighbor which is not an irregular vertex. We claim that members of U come in pairs in the following way. Take $u \in U$, with v its degree 2 neighbor. Consider the simple path starting from u, going to v, and terminating when we reach another irregular vertex u'. This termination condition must occur since G is finite, we have no degree 1 vertices, and we must encounter an irregular vertex before a degree 3 vertex.



(a) A graph G with only degree 2 and 3 vertices.

(b) Peeling away the purely degree 2 pieces of G.

Figure 1.7: Addressing degree 2 irregular vertices with a non-irregular degree 2 neighbor in the manner described in the proof of Lemma 2.

Let v' be the degree 2 neighbor of u' (noting $v = v'$ is possible). Delete edges (u, v) and (u', v'), and add edge (u, u'). This converts u, u' into (neighboring) irregular vertices, and leaves u', v' as either an isolated vertex or a path component. Connect v', v with an edge if the path component contains more than 2 vertices (see Figure 1.7). This entire procedure costs at most $4k/2 = 2k$ edits. Notice that the degree 3 neighbors of u, u' are unaffected; in particular, we have introduced no new irregular vertices. Now, the only degree 2 vertices remaining are irregular vertices. Observe that we cannot have 3 irregular vertices all adjacent to each other. This is because such a structure would need to be an isolated K_3 which violates the fact that they need to be irregular vertices. Therefore, so long as we have at least 3 irregular vertices remaining, there always exists a pair of vertices between which we can add an edge. This allows us to use at most $k/2$ edits to leave us with either 1 or 2 irregular vertices remaining. We may then apply Lemma 1 on the component containing the remaining irregular vertices (this is important so as not to introduce new irregular vertices), leaving us with a component-wise regular graph. In total, we use no more than $2k + k/2 + 5 \leq 3k + 5$ edits. □

We note that the ideas here make use of a concept sometimes called an edge swap [74]. Formally, suppose $G = (V, E)$ has vertices u_1, u_2, v_1, v_2 such that $(u_1, u_2), (v_1, v_2) \in E$ and G does not have any of the 4 edges (u_i, v_j). Then an edge swap at u_1, u_2, v_1, v_2 is the deletion of the edges $(u_1, u_2), (v_1, v_2)$ and the addition of the edges $(u_1, v_1), (u_2, v_2)$ (see Figure 1.8). Notice that in a regular graph, an edge swap does not change the degree of any vertex.

(a) A 2-regular graph.

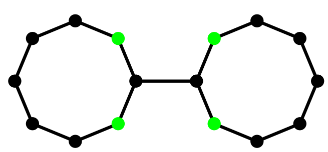
(b) The result of an edge swap.

Figure 1.8: An edge swap in a 2-regular graph.

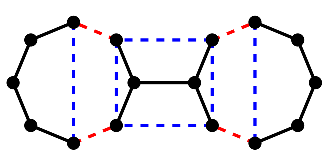
1.2.2 Proof of Proposition 1

Proof. G has at most $k_1 \leq k$ irregular vertices of degree 1. Arbitrarily pair up $2\lfloor k_1/2 \rfloor$ of them and add an edge within each pair. This is possible since degree 1 irregular vertices cannot be neighbors. If k_1 is even, we are done. Otherwise, we have one more degree 1 irregular vertex u, and there exists another vertex v of odd degree (strictly greater than 1). Delete an edge (v, w) (where $w \neq u$ if (v, u) exists), and add the edge (u, w). This procedure of addressing degree 1 irregular vertices requires at most $k/2 + 2$ edits. The neighbors of the first $2\lfloor k/2 \rfloor$ vertices we considered all had degree at least 2, so if any were not irregular vertices, they certainly do not become irregular vertices. Only vertex v might become a new irregular vertex since its degree decrements by 1. Consequently, our new graph has at most $k + 1$ irregular vertices. Consider the union of components that do not contain any degree 1 vertices. By Lemma 2, we need $3(k+1)+5$ edits to reach a component-wise regular graph. In total, we have used no more than $(k/2+2)+3(k+1)+5 \leq 4k+10$ edits. □

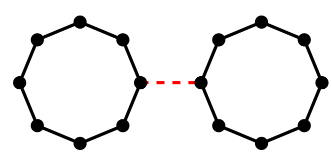
In some cases, while Proposition 1 gives something not far from optimal, it seems to miss the truth. Consider (see Fig. 1.9) the graph on $2n$ vertices consisting of two n-cycles connected by a single edge, for large n. Our procedure will give 10 edits, resulting in two $(n-3)$-cycles, and a 6-vertex 3-regular component. This completely misses the obvious (and much more elegant) choice of simply deleting the single edge connecting the 2 cycles to yield two disjoint n-cycles. This makes some sense, as this graph is almost regular in the sense that it is almost 2-regular, while Proposition 1 is designed to handle almost 3-regular graphs.



(a) Two cycles connected by an edge.



(b) Edits according to the proof of Proposition 1.



(c) Delete the single red dashed edge.

Figure 1.9: Two component-wise regular graphs close to a given graph.

One cannot easily amend the algorithm to this example, as the edge we would like to delete is not incident to an irregular vertex. What this seems to suggest is that our notion of irregular is incomplete. It does not seem to be able to nicely detect whether certain parts of a graph are 'essentially' degree 2 or 'essentially' degree 3. It might be necessary to consider vertices with neighbors of lower degree as well.

One interesting feature of this proof is that it very nearly does not depend on the non-irregular vertices being degree 3. In fact, the only place we use that is in the application of Lemma 1, when we show that there exists a sufficiently distant edge we can delete. Morally speaking, having vertices of arbitrary degree should not affect the existence of a sufficiently distant edge for us to delete. We use this idea to generalize to graphs of arbitrary maximal degree.

1.3 Proof of Main Result

The proof of Proposition 1 uses the idea of cutting away all portions of the graph that are component-wise 2-regular to split the graph into a component-wise 3-regular and component-wise 2-regular portion. This makes analysis more straightforward, and we generalize this approach for our main result.

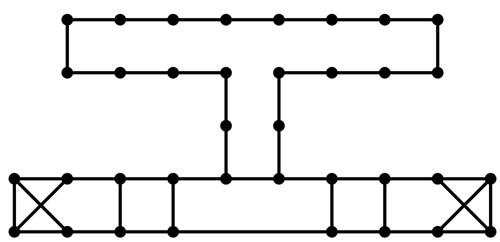
1.3.1 Preparatory Statements.

First, we take G and "cut away" the regions of the graph that have maximal degree strictly less than Δ (see Figure 1.10), which we formalize in the following lemma.

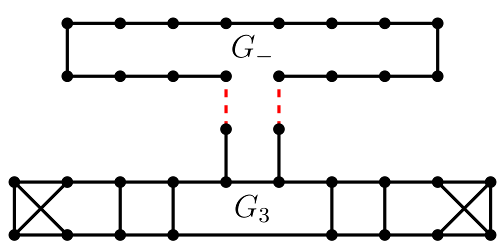
Lemma 3. *Suppose we have a graph $G = (V, E)$ whose vertices have maximal degree Δ. If G has k irregular vertices, there exists a union of two disjoint graphs $H = G_- \cup G_\Delta$ such that the following are true.*

1. *All vertices of G_Δ with degree less than Δ are irregular vertices.*
2. *G_Δ has at most k irregular vertices.*
3. *All vertices of G_- have degree strictly less than Δ.*
4. *G_- has at most $k(\Delta - 1)$ irregular vertices.*
5. *$d(G, H) \leq k(\Delta - 2)$.*

Proof. Let $V_0 \subset V$ denote the set of all irregular vertices in G adjacent to a vertex with degree Δ. Delete all edges of the form (u, v) where $u \notin V_0$, $\deg u < \Delta$, and $v \in V_0$. Consider the subgraph G_Δ of the resulting graph containing all vertices of maximum degree and their neighbors. Any vertex of degree less than Δ in G_Δ must be an irregular vertex adjacent to a vertex of maximum degree. Let V_1 denote the set of such vertices in G_Δ. For any edge (v_1, v) with $v_1 \in V_1$, we then must have that $\deg v = \Delta$ or $v \in V_1$. All irregular vertices in G_Δ were irregular vertices in our original graph G, so G_Δ has at most k irregular vertices. Let subgraph G_- contain all other vertices. By construction, there are no edges between G_Δ and G_-. Our edge deletions introduce at most $k(\Delta - 2)$ new irregular vertices, so G_- has no more than $k(\Delta - 1)$ irregular vertices. By construction, G_- has no vertex with degree Δ. All deletions occur only on edges incident to an original irregular vertex (though not to a vertex of maximum degree), so we have used at most $k(\Delta - 2)$ edits. $\square$



(a) A graph G with $\Delta = 3$.



(b) A graph G' obtained from G via Lemma 3.

Figure 1.10: An illustration of Lemma 3 in the $\Delta = 3$ case.

We remark that G_Δ satisfies the assumptions of the following proposition.

Proposition 2. *Suppose we have a graph $G = (V, E)$ whose vertices have maximal degree*

Δ. Further suppose that any vertex of degree strictly less than Δ is an irregular vertex and that G has k irregular vertices. Then if G has at least $2\Delta + \Delta^2 + \Delta^3$ vertices, there exists an almost regular graph H whose vertices all have degree Δ or $\Delta - 1$ such that $d(G, H) \leq 3k\Delta$.

Proof. Observe that we cannot have Δ irregular vertices all mutually adjacent to each other. This means we can always add edges between irregular vertices until we have at most $\Delta - 1$ irregular vertices in the entire graph, all of which are mutually adjacent. There are two things to note about this statement.

- If we start with $k > \Delta - 1$ irregular vertices, much like in the proof of Lemma 2, this procedure of adding edges between irregular vertices necessarily turns some irregular vertices into degree Δ vertices.
- The remaining irregular vertices may have any degree $d < \Delta$.

The added edges cost at most $k\Delta$ edits.

We proceed as follows. Define

$$V_1 = \left(\bigcup_{v_0 \in V_0} N(v_0) \right) \setminus V_0, \quad \text{and} \quad V_2 = \left(\bigcup_{v_1 \in V_1} N(v_1) \right) \setminus (V_0 \cup V_1).$$

Notice that

$$|V_0| = k_0 \leq \Delta - 1, \quad |V_1| \leq k_0(\Delta - 1) \leq (\Delta - 1)^2, \quad |V_2| \leq k_0(\Delta - 1)^2 \leq (\Delta - 1)^3.$$

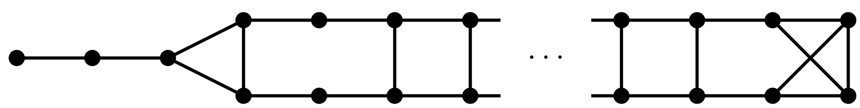
Edges with vertex in these sets could be too close to the irregular vertices. Any edge of G with at least 1 vertex not in V_0, V_1, or V_2, however, is guaranteed to have both its vertices not adjacent to any irregular vertex. To count such edges, we first enumerate the vertices of G not in V_0, V_1, or V_2. By assumption,

$$|V| \geq 2\Delta + \Delta^2 + \Delta^3,$$

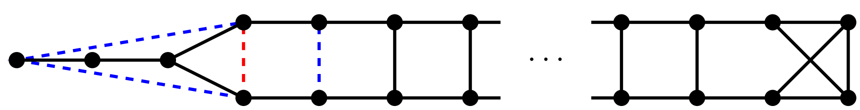
so

$$|V| - (|V_0| + |V_1| + |V_2|) \geq 2\Delta + \Delta^2 + \Delta^3 - k_0 - (\Delta - 1)^2 - (\Delta - 1)^3$$
$$\geq 2\Delta - k_0 \geq k_0.$$

Therefore, there must exist at least $k_0\Delta/2$ edges with at least 1 vertex not in V_0, V_1, V_2. Let E_0 denote this set of edges. For each irregular vertex v with degree d strictly less than $\Delta - 1$, let $l = \lfloor(\Delta - 1 - d)/2\rfloor$. Delete l sufficiently distant edges $u_i = (u_i^{(1)}, u_i^{(2)}) \in E_0$ and add edges $(v, u_i^{(1)})$ and $(v, u_i^{(2)})$, which costs $3l$ edits (see Figure 1.11). In total, this costs no more than $k_0 \cdot 3\Delta/2$ edits. Notice that this is possible since $l \leq \Delta/2$, and we ensured that $|E_0| \geq k_0\Delta/2$. Following this, every irregular vertex will have become a degree Δ vertex or a degree $\Delta - 1$ irregular vertex. None of the originally degree Δ vertices have had their degrees modified. Enumerating all our edits yields $k\Delta + 3k\Delta/2 \leq 3k\Delta$. $\square$



(a) A graph G satisfying the assumptions of Proposition 2.



(b) Almost regular graph H obtained from G as described by the proof of Proposition 2.

.

Figure 1.11: An illustration of Proposition 2 for a graph with maximal degree $\Delta = 3$.

Proposition 3. *Suppose we have an almost regular graph $G = (V, E)$ whose vertices have degree Δ or $\Delta - 1$ and that G has k vertices of degree $\Delta - 1$, all of which are irregular vertices.Then if G has at least $\Delta + \Delta^2 + \Delta^3 + 2$ vertices, there exists a component-wise regular graph H such that $d(G, H) \leq k + 4\Delta$.*

Proof. The opening of this proof will follow similarly to the beginning of the proof of Proposition 2. Observe that we cannot have Δ vertices of degree $\Delta - 1$ all adjacent to each other. This is because we would have an isolated K_Δ in G, which would violate these degree $\Delta - 1$ vertices being irregular vertices. Therefore, if we have at least Δ irregular vertices, there always exists a pair of irregular vertices between which we can add an edge. This allows us to use at most $k/2$ edits to connect pairs of irregular vertices such that we are left with at most $\Delta - 1$ mutually adjacent irregular vertices in the entire graph.

Let $k_0 \leq \Delta - 1$ denote the number of remaining irregular vertices. Once again, the idea is to delete sufficiently far away edges and connect the irregular vertices to the vertices of the just deleted edges. We need to check that G has sufficiently far away edges. Let V_0 denote all irregular vertices. Let V_1 denote the neighbors of V_0, excluding elements of V_0 itself. Let V_2 denote the neighbors of V_1 (excluding elements of both V_0 and V_1). Notice that

$$|V_0| = k_0 \leq \Delta - 1, \quad |V_1| \leq k_0(\Delta - 1) \leq (\Delta - 1)^2, \quad |V_2| \leq k_0(\Delta - 1)^2 \leq (\Delta - 1)^3.$$

Edges incident to these sets could be too close to the irregular vertices. Any edge of G with at least 1 vertex not in V_0, V_1, V_2, however, is guaranteed to have both its vertices not adjacent to any irregular vertex. We would like G to have at least $k_0/2$ sufficiently far edges. If G has x additional vertices, there are at least $\Delta x/2$ additional edges. Setting $x = 2$ is sufficient.

Therefore, if G has at least $\Delta + \Delta^2 + \Delta^3 + 2 \geq k_0 + k_0(\Delta - 1) + k_0(\Delta - 1)^2 + 2$ vertices, we may proceed with the following procedure. Arbitrarily pair up irregular vertices. For each pair u_1, u_2, delete a sufficiently far edge (v_1, v_2). Add edges (u_1, v_1) and (u_2, v_2). This costs at most $3(\Delta - 1)/2 \geq 3k_0/2$ edits. If k_0 is even, we are done. Otherwise, we have one remaining irregular vertex (which also implies that Δ is odd). Delete all edges incident to this remaining irregular vertex (costing Δ edits), leaving it as a 0-regular component. Then G is left with $\Delta - 1$ irregular vertices of degree $\Delta - 1$. This allows us to go back, and repeat this previously described procedure once, costing at most $3(\Delta - 1)/2$ edits.

In total, we have used no more than $k + 4\Delta$ edits. $\square$

1.3.2 Proof of the Main Result

Proof. We proceed by induction. As a base case, recall from Proposition 1 that any graph with maximal degree 3 is at most $4k + 10$ edits from a component-wise regular graph.

Apply Lemma 3, costing $k(\Delta - 2)$ edits. This yields G_Δ, a graph with maximal degree Δ such that all vertices with degree strictly less than Δ are irregular vertices, and G_-, a graph with maximal degree strictly less than Δ. Since G_Δ has at most k irregular vertices and has at least $2\Delta + \Delta^2 + \Delta^3$ vertices (by assumption), by Proposition 2, it costs at most $3k\Delta$ edits to obtain an almost regular graph with all vertices of degree $\Delta, \Delta - 1$. Note that the only degree $\Delta - 1$ vertices that arise in this manner must come from one of the k irregular vertices of G_Δ. Then by Proposition 3, we need at most $k + 4\Delta$ additional edits to obtain a Δ-regular graph.

It remains to address G_-. Lemma 3 causes G_- to have at most $k(\Delta - 1)$ irregular vertices. By assumption, G_- has at least $2d + d^2 + d^3 + k\Delta! - k(\Delta - 2)$ vertices of each degree $d < \Delta$, where the subtraction term arises from Lemma 3 reducing the degrees of at most $k(\Delta - 2)$ vertices. Let $K = k(\Delta - 1)$, $d_- = d - 1$, and $\Delta_- = \Delta - 1$. G_- has maximal degree Δ_- and at most K irregular vertices. Then notice that for all $d < \Delta$,

$$\begin{aligned} 2d + d^2 + d^3 + k\Delta! - k(\Delta - 2) &> 2d_- + d_-^2 + d_-^3 + k\left(\Delta! - (\Delta - 2)\right) \\ &> 2d_- + d_-^2 + d_-^3 + k\left(\Delta! - (\Delta - 1)!\right) \\ &= 2d_- + d_-^2 + d_-^3 + k(\Delta - 1)! \cdot (\Delta - 1) \\ &= 2d_- + d_-^2 + d_-^3 + K(\Delta_-)! \end{aligned}$$

which shows that G_- itself satisfies our hypothesis. We have so far used $k(\Delta - 1)$ edits to form G_Δ, G_- and $3k\Delta + k + 4\Delta$ edits to address G_Δ. Summing these yields

$$k(\Delta - 1) + 3k\Delta + k + 4\Delta = 4k\Delta + 4\Delta.$$

By induction, we apply the same argument to G_-. Since we have an explicit relation between the maximal degrees of G and G_- and between the number of irregular vertices of G and

G_-, we may compute that this entire procedure requires

$$
\begin{aligned}
&4k + 10 + \sum_{i=0}^{\Delta-4} \left((\Delta - i) \cdot 4k \prod_{j=1}^{i} (\Delta - j) + 4(\Delta - i) \right) \\
&\leq 4k + 10 + \sum_{i=0}^{\Delta-4} 4k\Delta! + 4\Delta \leq 4k + 10 + 4k\Delta \cdot \Delta! + 4\Delta^2 \\
&= (4 + 4\Delta \cdot \Delta!)\, k + (10 + 4\Delta^2)
\end{aligned}
$$

edits. Since $\Delta \geq 4$, we may bound this by $4(\Delta + 1)! \cdot k + 5\Delta^2$. □

The assumption of Theorem 1.1.2 requires that our input graphs have enough vertices of all degrees less than or equal to Δ. This condition arises from guaranteeing that there exist sufficiently many distant edges to delete for Propositions 2 and 3, but it is difficult to gauge how necessary this condition is.

Chapter 2

DYNAMICAL SYSTEMS

2.1 Introduction and Results

Consider approximating $x \in \mathbb{R}$ via the following method. Let $x_1 = 1$, and define

$$x_n = x_{n-1} + \frac{c_n}{n},$$

where

$$c_n = \begin{cases} 1 & x_n \leq x \\ -1 & x_n > x \end{cases}.$$

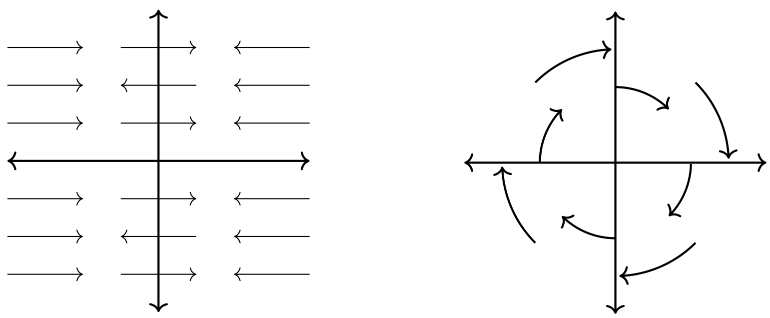
This is simply greedily approximating x by greedily assigning signs to the terms of the harmonic series. Some work has been done regarding convergence rates of this sequence. In fact, our original interest was sparked by a beautiful result of Bettin-Molteni-Sanna [17] who prove, among many other things, the following: pick some $x \in \mathbb{R}$, let $x_1 = 1$ and

$$x_n = \begin{cases} x_{n-1} + \frac{1}{n} & \text{if } x_n \leq x \\ x_{n-1} - \frac{1}{n} & \text{if } a_n > x. \end{cases}$$

This corresponds to taking the standard harmonic series $\sum 1/n$ and inserting signs in a greedy way to get as close as possible to x. It is easy to see that $|x_n - x| \leq 2/n$ for n sufficiently large. Bettin-Molteni-Sanna [17] show the existence of subsequence that converge *much* faster, even faster than a polynomial rate in n, for generic x. We discovered our system by looking at hypothetical two-dimensional analogues of this result.

A number of (vaguely) related results come from the abstract theory of rearrangements of series in Banach spaces (see Kadets-Kadets [54]), for example the results of Calabi-Dvoretzky

[24] or Dvoretzky-Hanani [31]. After completion of the project we learned that our results can be thought of as results in the area concerned with the *dynamics of piecewise isometries* (see Figure 2.1). Its setting is as follows: given a subset $X \subset \mathbb{R}^d$ and a finite partition $X = P_1 \cup P_2 \cup \ldots P_r$, a map $T : X \to X$ is a piecewise isometry if is restriction to P_i is an isometry for all $1 \leq i \leq r$.



$f(x, y) = (x - \text{sgn}(x), y)$ $\quad$ g = rotation by α around origin

Figure 2.1: $g \circ f$ is a piecewise isometric map whose orbits, suitably rotated, correspond to orbits of our procedure.

Many of the fundamental results in the area have been pioneered by Arek Goetz, starting with his PhD thesis [45] and the subsequent work [40, 41, 42, 43] as well as Ashwin-Goetz [9, 10], Ashwin-Goetz-Peres-Rodrigues [11], Boshernitzan-Goetz [19], Cheung-Goetz-Qyas [28] and Goetz-Quas [44]. There have been many subsequent developments and applications, we refer to Bruin-Deane [20], Deane [29], Park et al [70], Smith et al [78] and Sturman [90].

A natural question is whether one can extend this to a two-dimensional analog. We attempt this generalization by fixing some $z_{-1} \in \mathbb{C}$ and defining

$$z_n^+ = z_{n-1} + e^{2\pi i a_n}, \quad z_n^- = z_{n-1} - e^{2\pi i a_n},$$

where a_n is some sequence. Then our true sequence is given by

$$z_n = \begin{cases} z_n^+ & |z_n^+| < |z_n^-| \\ z_n^- & |z_n^-| < |z_n^+| \end{cases}.$$

Intuitively, we can say that a_n is the driving sequence for z_n. Indeed, the choice of a_n very much determines how z_n behaves.

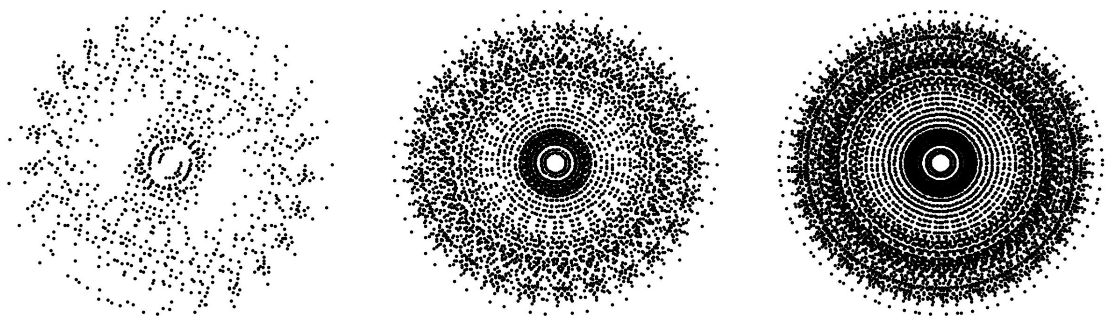


Figure 2.2: The first 1000 (left), 5000 (middle) and 10000 (right) terms for $z_{-1} = 0.0001 + 5i$ and $\alpha = 1.0415/\sqrt{2\pi^2}$. The sign pattern is periodic with period 222.

At first glance, our definition can be thought of as a piecewise isometric map with respect to half-spaces, defined by $e^{2\pi i n\alpha}$, that rotate. By introducing an additional rotation, we obtain a piecewise isometric map (see Fig. 2.1) that is given as the composition of a translation on two half-spaces and a global rotation. Its iterates y_n, when rotated by $e^{2\pi i n\alpha}$, correspond to iterates x_n of our procedure. This type of piecewise isometric map seems to not have been studied before and it stands to reason that our greedy sign choice perspective may arguable be easier to work with. Moreover, it appears that our main results (Theorem 2 and Theorem 3) appear to be new within the context of the dynamics of piecewise isometric maps. We explore z_n when $a_n = \alpha n$, for some irrational α. We make this choice for two reasons. First, if α were rational, $e^{2\pi i a_n}$ would be periodic and empirically lead to uninteresting sequences z_n. Second, this sequence is the natural choice for a sequence whose fractional part is equidistributed on $(0, 1)$. Later, we will consider other sequences.

2.1.1 Introduction

Let $z_{-1} \in \mathbb{C}$ and let $\alpha > 0$ be an irrational number. We define a sequence of complex numbers $(z_n)_{n=0}^{\infty}$ by setting, for $n \geq 0$,

$$z_n = z_{n-1} \pm e^{2\pi i \alpha n} \tag{2.1}$$

where the sign is chosen so as to minimize the absolute value. If both choices of sign lead to a complex number with the same absolute value, i.e. $|z_{n-1} + e^{2\pi i \alpha n}| = |z_{n-1} - e^{2\pi i \alpha n}|$, then we say the sequence is only defined up to z_{n-1} (this is of minor importance as it happens very rarely). We were motivated by the remarkably intricate complicated behavior that this dynamical system is able to exhibit. Examples are shown throughout starting with Fig. 2.2 and Fig. 2.3.

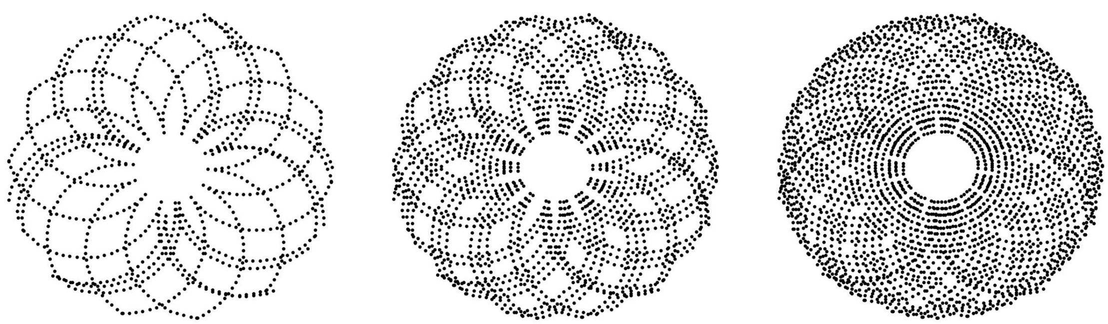


Figure 2.3: The system with initial value $z_{-1} = 2.0176 + 4.8585i$ and $\alpha = 0.00702367$. The pictures show the terms z_n for $2000 \leq n \leq 3000$ (left), $2000 \leq n \leq 4000$ (middle) and $2000 \leq n \leq 5000$ (right). The sequence of signs is periodic with period 51.

Numerical examples suggest that 'most' initial values z_{-1} and 'most' values of α lead to a sequence $(z_n)_{n=0}^{\infty}$ all of whose elements are eventually trapped in a single circle (by which we mean that all elements of the sequence z_n for $n \geq N$ are contained in a circle), see Fig. 2.5 or Fig. 2.6. However, exceptions are not infrequent (especially when α is close to $0, 1/2$ or 1): most of the examples shown in the various Figures were discovered by looking at many random choices for α and z_{-1}.

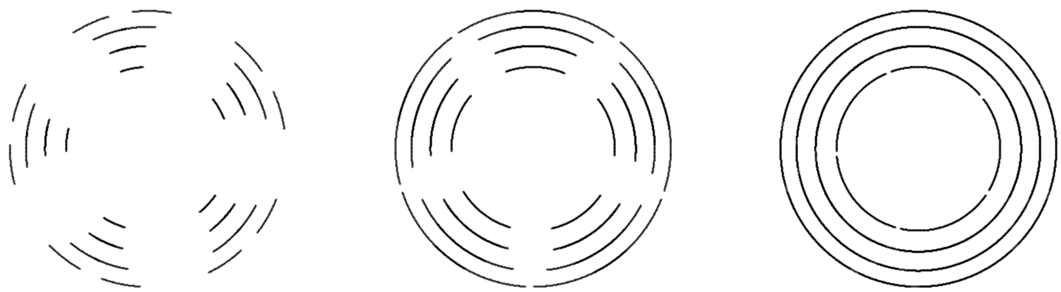


Figure 2.4: The system with initial value $z_{-1} = 1 - i$ and $\alpha = \sqrt{2}/3$. The pictures show the terms z_n for $100 \leq n \leq 2000$ (left), $100 \leq n \leq 5000$ (middle), and $100 \leq n \leq 8000$ (right). The sequence of $\pm$ signs is periodic with period 14.

2.1.2 Periodic signs implies concentric circles.

We start with a basic observation: when looking at the dynamical system for 'generic' initial values, the sequence of signs ± 1 that are chosen appears to become periodic. A periodic choice of signs corresponds to a simple geometric pattern: the sequence is ultimately contained in a union of finitely many concentric circles (see, for example, Fig. 2.4).

Theorem 1. *If the sequence of signs is eventually periodic with period p, eventually all elements of the sequence are contained in a union of at most p concentric circles.*

The inequality is necessary: it is fairly common to see signs with period 2 ($+1$ and -1 alternating) where the associated sequence is contained in a single circle. It is not clear to us how the period of the signs is connected to the number of circles beyond the one-sided inequality: Fig. 2.4 shows an example where signs are periodic with period 14 but all elements are ultimately contained in only 4 concentric circles.

2.1.3 Long-time behavior: a single circle.

One natural question is whether the figures shown throughout the paper are (a) numerically accurate and (b) reflect the overall long-time behavior. Two instructive examples are given in Fig. 2.5: one observes rather intricate behavior for the first few elements of the sequence after which the sequence suddenly becomes very simple and all subsequent elements are contained in a single circle.

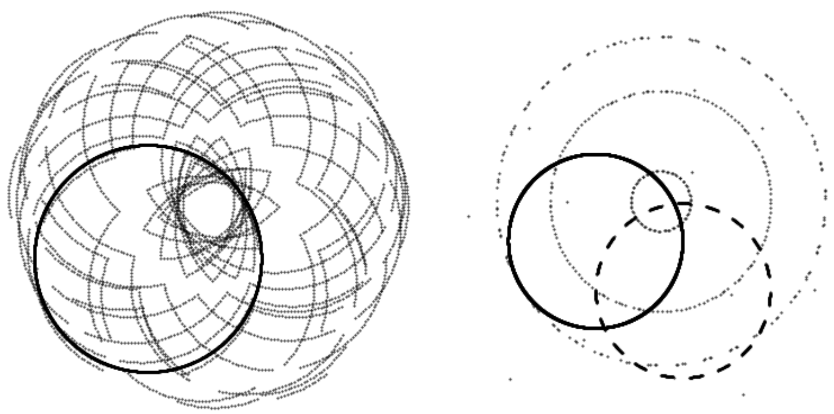


Figure 2.5: Left: system with $z_{-1} = 2 + 5i$ and $\alpha = 1.1054\sqrt{2}/(2\pi)$. Right: system with $z_{-1} = 5e^{2\pi i \cdot 0.00883}$ and $\alpha = \sqrt{7}/(2\pi)$. Both sequences are quite intricate before settling down in a single circle.

This raises a natural question: when observing more complicated behavior as in Fig. 2.2 and Fig. 2.3, are we guaranteed that they remain at that level of complexity? Maybe everything always ends up in a single circle (after possibly millions and millions of terms)? Before explaining our main results, we introduce a useful concept: the ambiguity sequence. Given such a sequence $(z_n)_{n=0}^{\infty}$, the associated ambiguity sequence $(a_n)_{n=0}^{\infty}$ is the sequence of real numbers

$$a_n = \left| |z_{n-1} + e^{2\pi i n\alpha}| - |z_{n-1} - e^{2\pi i n\alpha}| \right|.$$

One can think of $(a_n)_{n=0}^{\infty}$ as a measure of ambiguity that quantifies the difference between the two choices of sign: if a_n is large, then one sign very clearly leads to a complex number with smaller norm. If a_n is small, the difference is less pronounced. One could also think of very small values of a_n as a possible sign of danger: higher numerical accuracy may be required. By assumption, we always have $a_n > 0$ since z_n is not defined if both choices of sign lead to a complex number with the same absolute value. We can now state the first half of our main result. Informally, it will end up saying

> if one observes a sequence always choosing the same sign for a while and if the levels of ambiguity a_n never get too small, then the sign sequence will be periodic for all time.

This informal statement is in need of some additional precision. We first state the result for a special case where the signs end up being constant: in that case, all the contributing components can be made explicit. The general periodic case (Theorem 3) will be based on Theorem 2 and uses very similar ideas.

Theorem 2 (Single circle)**.** *Let $\alpha > 0$ be irrational. Then, using $(p_\ell/q_\ell)_{\ell=1}^{\infty}$ to denote the convergents, the following holds: if, for some $k, \ell \in \mathbb{N}$ all the signs for $k \le n \le k + q_\ell$ are the same and if*

$$\min_{k \le n \le k+q_\ell} a_n \ge \left(4\pi + \frac{4\pi}{|e^{2\pi i\alpha} - 1|}\right) \frac{20}{q_\ell},$$

then the sign is constant for all $n \ge k$.

Remarks.

1. Theorem 1 then implies that, in that case, all sufficient large elements of the sequence are contained in a single circle.

2. k and ℓ are arbitrary parameters. k allows for different starting indices, ℓ can be seen as a measure of the scale. A larger scale requires a weaker inequality to be true for a longer time.

3. Generically, if the sequence ends up choosing the same sign, the ambiguity sequence is bounded away from 0. This means that, generically, Theorem 2 will eventually be able to certify periodicity (see the proof for details).

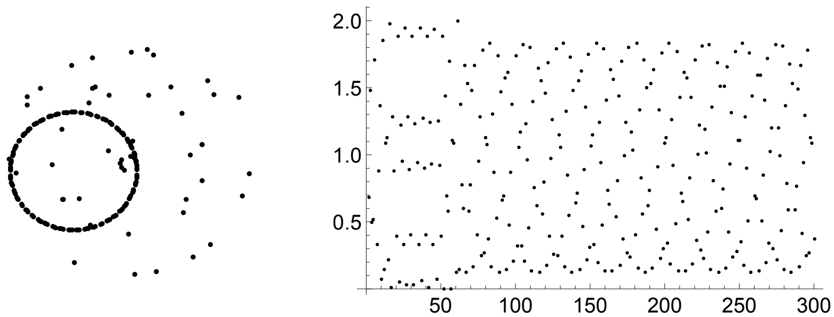


Figure 2.6: Left: system with $z_{-1} = -1/2 - i$ and $\alpha = 1/\sqrt{6}$. Right: the value of a_n along the evolution of the system

We quickly illustrate Theorem 2 for a concrete example (see Fig. 2.6): we start with $z_{-1} = -1/2 - i$ and $\alpha = 1/\sqrt{6}$ and observe that the sequence initially jumps around before, around the 60th term, starting to always choose the sign $+1$ and ending up in a circle. Computing the convergents p_ℓ/q_ℓ for $\alpha = 1/\sqrt{6}$, the sequence of denominators q_ℓ is given by

$$1, 2, 5, 22, 49, 218, 485, 2158, 4801, 21362, \ldots$$

Picking $q_9 = 4801$, an explicit computation shows that

$$\left(4\pi + \frac{4\pi}{|e^{2\pi i\alpha} - 1|}\right)\frac{20}{4801} = 0.0796\ldots$$

Simultaneously, we observe, numerically, that $a_n \geq 0.12$ for all $100 \leq n \leq 100 + 4801$. Theorem 2 applies and guarantees that the sequence will always keep choosing the sign $+1$ and remain in a single cycle. There is a little bit of flexibility in how to choose k and q_ℓ: the sequence of ambiguities satisfies $a_n \geq 0.1$ for all $n \geq 100$. The appearance of convergent fractions $(p_\ell/q_\ell)_{\ell=1}^{\infty}$ is related to the assumption of α being irrational: if α is irrational but extremely close to a rational number p/q, then $e^{2\pi i n\alpha}$ will be extremely close to a $q-$periodic sequence over extremely large periods of time. This is captured by the growth of q_ℓ. For example, if $\alpha = 10/17 + 10^{-7}\sqrt{2}$ then the sequence of q_ℓ is $1, 2, 5, 17, 415944, \dots$ which shows that the ambiguity may need to be checked for a large number of consecutive terms.

2.1.4 Long-time behavior: general periodicity.

Arguably the most interesting case is not covered by Theorem 2: examples where the sequence exhibits an interesting periodic pattern as seen in Figure 2.2 and Figure 2.3. We can now state our main result: if one observes periodic signs for a sufficiently long period of time, then the signs remain periodic for all time.

Theorem 3. *Let $\alpha > 0$ be irrational. There exists an (explicit) function*

$$g : (0, 1) \times \mathbb{N} \to \mathbb{N}$$

depending on α and z_{-1} such that for all $\eta \in (0, 1)$ the following holds: if

1. *the sequence of signs is $p-$periodic for $k \leq n \leq k + g(\eta, p)$ and*
2. *the ambiguity sequence is not too small in that range*

$$\min_{k \leq n \leq k+g(\eta,p)} \left| |z_{n-1} + e^{2\pi i n\alpha}| - |z_{n-1} - e^{2\pi i n\alpha}| \right| \geq \eta,$$

then the sequence of signs is $p-$periodic for all $n \geq k$.

Remarks.

1. Theorem 3 can be summarized as: if one observes periodic signs for a while and does not run into issues of numerical accuracy, the sequence of signs is going to be periodic for all time.
2. As in Theorem 2, the result is nearly sharp in the sense that for 'generic' examples with periodic signs the ambiguity sequence remains bounded away from 0 and Theorem 3 applies for a suitable choice of k and η.
3. The function $g(\eta, p)$ could be made explicit in terms of p and α: as in Theorem 2, the

convergents coming from the continued fraction expansion play a role in how quickly $g(\eta, p)$ tends to infinity as $\eta \to 0^+$ (see proof).

2.1.5 Stability and Symmetries

We start by describing an important notion of stability: if we have a sequence starting z_{-1} and the minimal ambiguity is uniformly bounded away from 0, then changing z_{-1} a little does not substantially change the behavior of the sequence.

Proposition 4. *Let $\alpha > 0$ and $z_{-1} \in \mathbb{C}$. Replacing z_{-1} by $z_{-1} + w$ with*

$$|w| < \frac{1}{2} \inf_{n \geq 0} a_n = \frac{1}{2} \inf_{n \geq 0} \left| |z_n + e^{2\pi i n \alpha}| - |z_n - e^{2\pi i n \alpha}| \right|$$

shifts the entire sequence by w, i.e. $z_n' = z_n + w$ for all $n \geq 0$.

Proposition 4 naturally suggests that the minimal ambiguity along a sequence should be a natural way to classify the behavior of sequences according to their starting value z_{-1}: if the minimal ambiguity is large, then nearby starting values are going to behave very much like the sequence arising from z_{-1} does. Small values of the minimal ambiguity suggest that a small change in the initial value z_{-1} can lead to very different dynamical behavior. Plotting the ambiguity function

$$z_{-1} \to \inf_{n \geq 0} \left| |z_n + e^{2\pi i n \alpha}| - |z_n - e^{2\pi i n \alpha}| \right|$$

shows the most surprising pictures (see Figure 2.7 and Figure 2.8). The dependency of $\inf_{n \geq 0} a_n$ on z_{-1} seems to be highly nontrivial.

Fig. 2.7 suggests the presence of symmetries: we note three such symmetries.

Proposition 5. *Let z_n be given by $0 < \alpha < 1$ and z_{-1}, with sign sequence ε_n.*

1. *The sequence arising from $1 - \alpha$ is $(\overline{z_n})$, with signs ε_n.*
2. *The sequence arising from $\alpha \pm 0.5 \pmod 1$ is (z_n) with signs $\varepsilon_n \cdot (-1)^n$.*
3. *The sequence arising from $-z_{-1}$ is $-z_n$, with signs $-\varepsilon_n$.*

While not appearing to be self-similar in the way classical fractals are, Fig. 2.8 shows that several layers of self-similarity are possible. Fig. 2.7 appears to feature balls (as well as balls cut by lines). Balls show up naturally as part of the dynamical system: there are two disks with radius 1/2 whose interior, when used as initial values, lead to a sequence with constant signs.

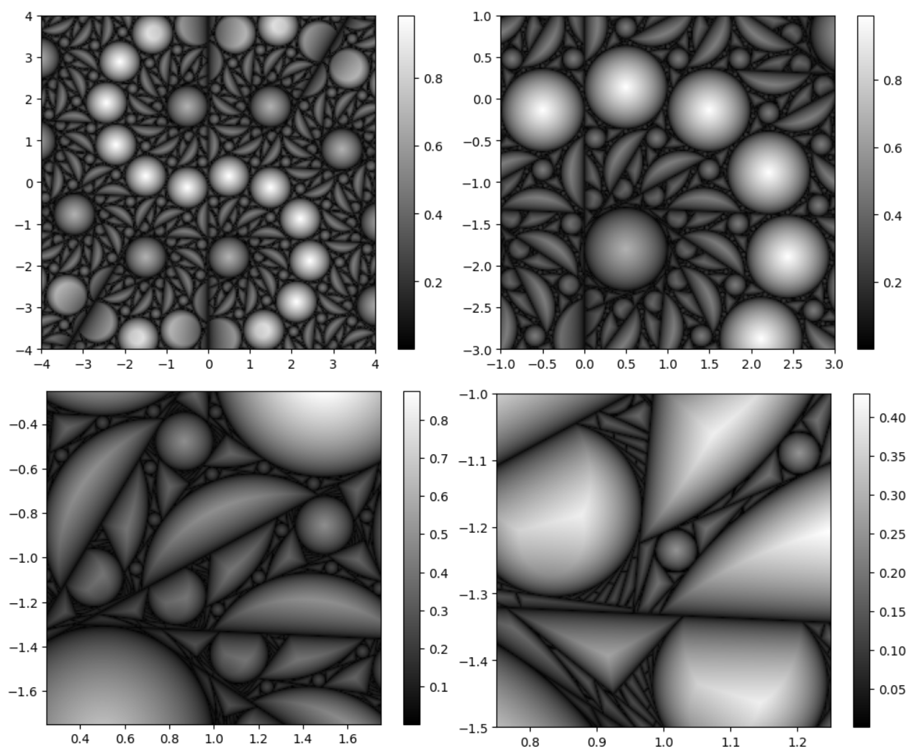


Figure 2.7: With $\alpha = \sqrt{2}$, (square roots of) minimum ambiguities of the first one thousand terms of sequences plotted against z_{-1}.

Proposition 6. *Let $\alpha \in (0, 1)$. If*

$$\left| z_{-1} - \frac{1}{1 - e^{2\pi i \alpha}} \right| < \frac{1}{2},$$

then the arising sequence always picks sign -1 (and ends up in a circle). Likewise, if $|z_{-1} - 1/(e^{2\pi i\alpha} - 1)| < 1/2$, the sequence always pick sign $+1$.

These two disks (which are clearly visible in Fig. 2.7 and Fig. 2.8) naturally have to be disjoint; this is reflected in the identity

$$\forall\, \alpha \notin \mathbb{N} \qquad \operatorname{Re} \frac{1}{e^{2\pi i \alpha} - 1} = \frac{1}{2}.$$

Other disks in the figures seem to appear for similar reasons: characterizing them all seems to be non-trivial and potentially an interesting area for further research.

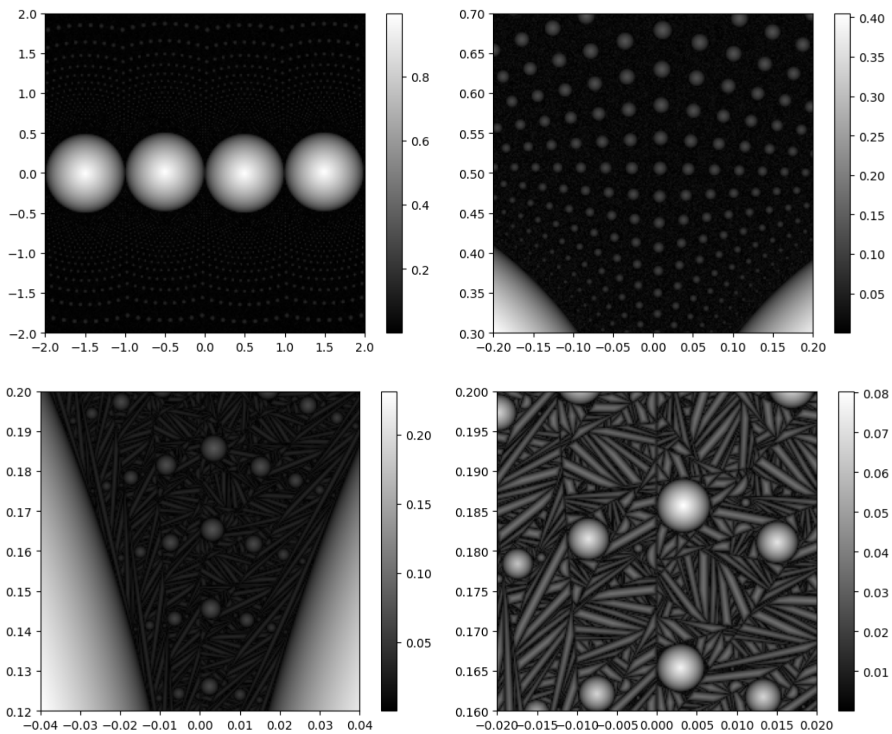


Figure 2.8: With $\alpha = 0.5 + \sqrt{3}/300$, (square roots of) minimum ambiguities of the first ten thousand terms plotted against z_{-1}.

The next natural question is whether the non-circular mysterious shapes in Fig. 2.7 and Fig. 2.8 can be explained and this will be covered by the next result.

Theorem 4. *Let α be irrational and let $z_{-1} \in \mathbb{C}$ lead to a sequence whose sign pattern is eventually periodic. The set of initial values that eventually exhibit the same (periodic) sign* *pattern is contained in a finite union of a finite intersection of disks and half spaces.*

The proof shows slightly more: the interior of such regions is given by an intersection of (open) disks and half spaces. If we use $z_{-1} \to \inf_{n\geq 0} \left||z_n + e^{2\pi i n\alpha}| - |z_n - e^{2\pi i n\alpha}|\right|$ to partition $\mathbb{C}$ into regions with similar dynamical behavior (separated by the zeros of that function) and if every sequence is eventually periodic, then Theorem 4 would have some more implications: the plane would be tiled by intersections of disks and half spaces (up to a set of measure 0).

2.1.6 Open problems

While we have a relatively good understanding of the case where one observes periodic sign patterns, the inverse problem is completely open.

> **Problem.** Is the sign pattern always eventually periodic?

We do not know how difficult that problem is. For fixed α, Proposition 4 shows that the set of initial values z_{-1} that lead to periodic orbits tend to have a certain 'openness' condition in terms of the ambiguity sequence: this might be an indication that hypothetical non-periodic orbits, if they exist at all, must be rare. A second natural question concerns the length of periods.

> **Long periods.** If a sequence eventually stabilizes into a periodic choice of signs, how long can this period be depending on α?

It seems that sequences can have arbitrarily long periods. A nice example is a sequence with period 874 given by

$$\alpha = 0.5010866 \quad \text{and} \qquad z_{-1} = 0.747467 + 0.445271i$$

whose associated ambiguity sequence is bounded below by $6 \cdot 10^{-6}$. The example is shown in Fig. 2.9. Moreover, it appears as if having a large period requires $\alpha \mod 1$ to be close to $0, 1/2, 1$ and one might be tempted to ask whether, for some universal $c > 0$, an inequality along the lines of

$$\text{length of period} \lesssim \frac{c}{\|2\alpha\|}$$

might be true (where $\|x\| = \inf_{n\in\mathbb{Z}} |x-n|$ is the distance to the nearest integer). It is not clear to us how difficult to this question is. It would be interesting to see whether there is a way to construct sequences with arbitrarily large periods: we observe, as a consequence of the symmetries, the following period-doubling trick.

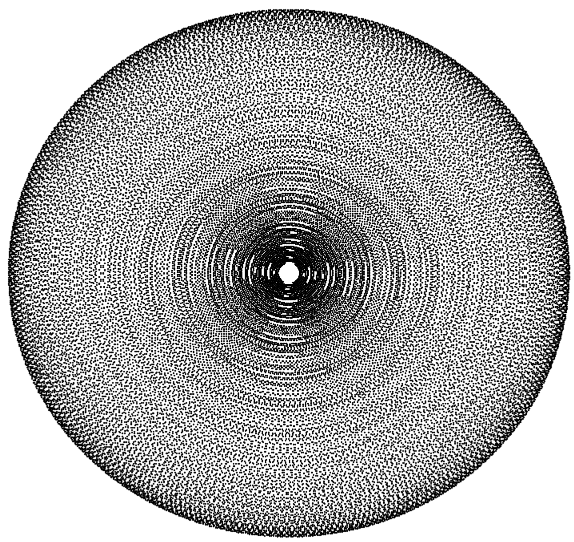


Figure 2.9: A sequence with period 874 (first million terms, only every 20th term displayed).

Proposition 7. *Let (z_n) given by α and z_{-1}. If the sign sequence ε_n is p-periodic with p odd, the sequence with initial point z_{-1} and $\alpha + 0.5 \pmod 1$ is $2p$-periodic.*

The largest period we are currently aware of is a sequence with period 2258 given by $\alpha = 0.50015827$ and initial value

$$z_{-1} = 0.5761982862055985 + 0.9356408428886818i.$$

However, we have no reason to believe that this is in any way extremal; it appears that, by choosing α close to $1/2$, arbitrarily large periods are possible. Another innocent question is as follows.

> **Which numbers are periods?** Which numbers can arise as periods of the dynamical system for some α and z_{-1}?

It appears, judging purely from empirical observations, as if not all integers can arise as a period. For example, we have observed empirically that periods are not divisible by 4. Is this always the case?

2.2 Proofs

2.2.1 Proof of Theorem 1

Proof. Introducing the sign sequence $(\varepsilon_n)_{n=0}^{\infty}$, we have for all $n > N$ that

$$z_n = z_N + \sum_{j=N+1}^{n} \varepsilon_j e^{2\pi i \alpha j}.$$

Assume the sign sequence is periodic with period p, then, for all $\ell \in \mathbb{N}$,

$$\begin{aligned} \sum_{j=k+(\ell-1)p+1}^{k+\ell p} \varepsilon_j e^{2\pi i j\alpha} &= \sum_{j=k+1}^{k+p} \varepsilon_{j+\ell p} e^{2\pi i (j+\ell p)\alpha} \\ &= \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha} e^{2\pi i \ell p\alpha} = e^{2\pi i \ell \alpha p} \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha}. \end{aligned}$$

This last sum is independent of ℓ: it only depends on the period p and the starting point k (this observation will occur repeatedly in subsequent proofs). Abbreviating

$$q_k = \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha}, \tag{2.2}$$

we deduce that for $k \geq N$ and $\ell \geq 0$,

$$\begin{aligned} z_{k+\ell p} &= z_k + \sum_{j=k+1}^{k+\ell p} \varepsilon_j e^{2\pi i j\alpha} = z_k + \sum_{s=0}^{\ell-1} \sum_{j=k+sp+1}^{k+(s+1)p} \varepsilon_j e^{2\pi i j\alpha} \\ &= z_k + q_k \sum_{s=0}^{\ell-1} e^{2\pi i\alpha ps} = z_k + q_k \frac{e^{2\pi i\alpha p\ell}-1}{e^{2\pi i\alpha p}-1} \\ &= \left[z_k - \frac{q_k}{e^{2\pi i\alpha p}-1} \right] + q_k \frac{e^{2\pi i\alpha p\ell}}{e^{2\pi i\alpha p}-1} \end{aligned}$$

For $\ell \in \mathbb{N}$, this subsequence of points lies on a circle with center

$$c_k = z_k - \frac{q_k}{e^{2\pi i\alpha p}-1}.$$

Repeating the above argument for $k \in \{N, \dots, N+p-1\}$ shows that all z_n for $n > N$ lie on one of p circles. It remains to prove that these circles have a common center. By using the definition of the sequence and iterating the recursion (2.2)

$$\begin{aligned} c_{k+1} &= z_{k+1} - \frac{q_{k+1}}{e^{2\pi i\alpha p}-1} \\ &= z_k + \varepsilon_{k+1} e^{2\pi i\alpha(k+1)} - \frac{q_k + \varepsilon_{k+p+1} e^{2\pi i\alpha(k+p+1)} - \varepsilon_{k+1} e^{2\pi i\alpha p(k+1)}}{e^{2\pi i\alpha p}-1} \\ &= z_k - \frac{q_k}{e^{2\pi i\alpha p}-1} = c_k. \end{aligned}$$

Therefore all the circles share the same center. □

2.2.2 A Sampling Lemma

The purpose of this section is to prove a Lemma that will be useful in the proof. We were unable to find the Lemma in the literature but all its ingredients are standard and it is presumably fair to say that the Lemma is well-known 'in spirit'. We recall that if α is an irrational number, then it has an infinite continued fraction expansion: truncating this expansion leads to a rational number p_ℓ/q_ℓ that is very close to α.

Lemma 4. *Suppose α is an irrational real with convergents $(p_\ell/q_\ell)_{\ell=1}^{\infty}$. If*

$$f : \{z \in \mathbb{C} : |z| = 1\} \to \mathbb{R}$$

is Lipschitz-continuous with Lipschitz constant $L > 0$ and $k, \ell \in \mathbb{N}$ is such that

$$\min_{k \le n \le k+q_\ell} f(e^{2\pi i n \alpha}) > \frac{20L}{q_\ell}, \tag{2.3}$$

then

$$\min_{0 \le t \le 1} f(e^{2\pi i t}) > 0.$$

Conversely, if $\min_{0 \le t \le 1} f(e^{2\pi i t}) > 0$, then there exists $\ell_0 \in \mathbb{N}$ such that (2.3) holds for all $k \ge 0$ and all $\ell \ge \ell_0$.

Proof. We first show that it suffices to prove the statement for $k = 0$. For arbitrary $k \ge 0$, we can define $g : \{z \in \mathbb{C} : |z| = 1\} \to \mathbb{R}$ via

$$g(z) = f(e^{2\pi i k \alpha} z)$$

which corresponds to a rotation and does not change the Lipschitz constant. Then

$$\min_{k \le n \le k+q_\ell} f(e^{2\pi i n \alpha}) = \min_{0 \le n \le q_\ell} g(e^{2\pi i n \alpha})$$

and we see that one implies the other. We will now argue that $n\alpha$, when interpreted modulo 1, is rather close to a set of equispaced points provided n is chosen at the correct scale (suggested by the denominators of the convergents). The crucial ingredient is (see [22]) the inequality

$$\left| \alpha - \frac{p_\ell}{q_\ell} \right| \le \frac{1}{q_\ell^2}.$$

Therefore, for all $1 \le n \le q_\ell$, we have

$$\left| n\alpha - \frac{p_\ell n}{q_\ell} \right| \le \frac{n}{q_\ell^2} \le \frac{1}{q_\ell}.$$

Since p_ℓ and q_ℓ are always coprime (see [22]), we have

$$\left\{ \frac{p_\ell n}{q_\ell} \bmod 1 : 1 \le n \le q_\ell \right\} = \left\{ \frac{n}{q_\ell} : 0 \le n \le q_\ell - 1 \right\}.$$

Let now $0 \le t \le 1$ be arbitrary. Then there exists $1 \le n \le q_k$ such that

$$\left| \left(\frac{p_\ell n}{q_\ell} \mod 1 \right) - t \right| \le \frac{1}{q_\ell}.$$

We deduce that $|(n\alpha \mod 1) - t| \leq 3/q_\ell$ and

$$|(2\pi n\alpha \mod 2\pi) - 2\pi t| \leq \frac{2\pi \cdot 3}{q_\ell} \leq \frac{19}{q_\ell}.$$

Using Lipschitz-continuity, we have

$$\begin{aligned} f(e^{2\pi i t}) &\geq f(e^{2\pi i n\alpha}) - L\,|2\pi n\alpha - 2\pi t| \\ &\geq \frac{20L}{q_k} - \frac{19}{q_k} \cdot L \geq \frac{L}{q_k} > 0. \end{aligned}$$

The second part of the statement is easy: f, if positive, is a continuous function on a compact interval and thus bounded from below by a positive constant: the sequence q_ℓ is unbounded and the conclusion follows. $\square$

2.2.3 Proof of Theorem 2

Proof. We assume α, z_{-1} are fixed. We also suppose that, starting at z_k, the chosen signs are constant for a large number of iterations (we assume at least until z_N where $N > k$) and that without loss of generality, the sign that ends up being chosen is always $+$ (the case where the sign is always $-$ is completely analogous). Then, for $k < n \leq N$

$$z_n = z_k + \sum_{j=k+1}^{n} e^{2\pi i j\alpha} = z_k + \frac{e^{2\pi i\alpha(n+1)} - e^{2\pi i(k+1)\alpha}}{e^{2\pi i\alpha} - 1}.$$

Introducing the map $h_k : \mathbb{S}^1 \to \mathbb{C}$

$$h_k(z) = z_k + \frac{z - e^{2\pi i(k+1)\alpha}}{e^{2\pi i\alpha} - 1}$$

we have, for $k < n \leq N$,

$$z_n = h_k(e^{2\pi i\alpha(n+1)}). \tag{2.4}$$

We would like to guarantee that, for all $n > k$,

$$\left|z_n + e^{2\pi i\alpha(n+1)}\right| < \left|z_n - e^{2\pi i\alpha(n+1)}\right|$$

since that would then enforce that the sign chosen at each point is always $+$. Using (2.4), this is equivalent to

$$\left|h_k(e^{2\pi i\alpha(n+1)}) + e^{2\pi i\alpha(n+1)}\right| < \left|h_k(e^{2\pi i\alpha(n+1)}) - e^{2\pi i\alpha(n+1)}\right|.$$

This inequality would be implied by the truth of the inequality

$$\forall\ |z| = 1 \qquad |h_k(z) + z| < |h_k(z) - z|\,.$$

Every time we pick another element of the sequence, we certify the truth of that inequality at another point on $\mathbb{S}^1$. Moreover, the sequence $e^{2\pi i\alpha(n+1)}$ is dense in $\mathbb{S}^1$. We apply Lemma 4 to to the function

$$f(e^{2\pi it}) = \left|h_k(e^{2\pi it}) - e^{2\pi it}\right| - \left|h_k(e^{2\pi it}) + e^{2\pi it}\right|.$$

Strict positivity of f would guarantee stability of the sign choice for all time. Observing that

$$\left|\frac{d}{dt}h_k(e^{2\pi it})\right| = \left|\frac{d}{dt}\frac{e^{2\pi it}}{e^{2\pi i\alpha} - 1}\right| = \frac{2\pi}{|e^{2\pi i\alpha} - 1|}$$

we note that f is Lipschitz continuous with Lipschitz constant satisfying

$$L \leq 4\pi + \frac{4\pi}{|e^{2\pi i\alpha} - 1|}.$$

At this point, we can invoke Lemma 4. The desired stability is implied by the existence of a convergent p_k/q_k such that

$$\min_{1 \leq n \leq q_k} f(e^{2\pi in\alpha}) > \frac{20}{q_k}\left(4\pi + \frac{4\pi}{|e^{2\pi i\alpha} - 1|}\right)$$

which is the desired statement. □

Remark. The truth of the inequality

$$\forall\ |z| = 1 \qquad |h_k(z) + z| < |h_k(z) - z|$$

is necessary but not sufficient for the sign choice to be stable. However, it is *almost* necessary in a way that renders Theorem 1 applicable for virtually all choices of initial data. We will now explain this in greater detail. First, suppose that there exists $|z_*| = 1$ such that the inequality *strictly fails*

$$|h_k(z_*) + z_*| > |h_k(z_*) - z_*|\,,$$

then there exists an entire open interval $J \subset \mathbb{S}^1 \subset \mathbb{C}$ containing z_* for which the inequality fails. Since $e^{2\pi i \alpha n}$ is dense on $\mathbb{S}^1$, at some point in the future we are bound to encounter an index n where one would end up choosing the $-$ sign. In that case, the statement we are trying to prove is actually false. This means that the only remaining case is

$$\forall\ |z| = 1 \qquad |h_k(z) + z| \leq |h_k(z) - z|$$

with equality for some z_*. This case is subtle because the relevant question is then whether an equation of the type $z_* = e^{2\pi i \alpha n}$ has a solution in $n \in \mathbb{N}$: if it does, then there exists a point at which the sequence becomes undefined. If not, then the $+$ sign is stable. However, distinguishing between the case of strict inequality and strict failure from a finite number of samples is impossible. In summary, we expect Theorem 1 to be applicable in all practically arising cases.

Remark. We also note that the inequality

$$\forall\ |z| = 1 \qquad |h_k(z) + z| < |h_k(z) - z|$$

is the type of statement that is frequently used in complex analysis to deduce that two functions have the same number of roots inside the unit disk (via Rouché's theorem). It is not clear to us whether this is a coincidence.

2.2.4 Proof of Theorem 3

Proof. The main idea behind the proof is to show that the periodic setting can be reduced to the setting of Theorem 2 by arguing along a subsequence. Let us again assume that $\alpha > 0$ is a given irrational number. We assume that, starting at z_k, the arising sequence of signs is $p-$periodic where $p \in \mathbb{N}$ at least in the range $k \leq n \leq N$ where N is large. Then, for

$k \leq n \leq N$ we have

$$z_n = z_k + \sum_{j=k+1}^{n} \varepsilon_j e^{2\pi i j\alpha},$$

where $\varepsilon_j \in \{-1, 1\}$ and the sequence of signs $(\varepsilon_j)_{j\in\mathbb{N}}$ is $p-$periodic for $k \leq j \leq N$. This implies, just as in the proof of Theorem 1, that

$$\begin{aligned}\sum_{j=k+(\ell-1)p+1}^{k+\ell p} \varepsilon_j e^{2\pi i j\alpha} &= \sum_{j=k+1}^{k+p} \varepsilon_{j+\ell p} e^{2\pi i (j+\ell p)\alpha} \\ &= \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha} e^{2\pi i \ell p\alpha} = e^{2\pi i \ell\alpha p} \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha}.\end{aligned}$$

Observe that this last sum is independent of ℓ: it only depends on the period p and the starting point k. Abbreviating

$$q = \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha},$$

we deduce

$$\begin{aligned}z_{k+\ell p} &= z_k + \sum_{j=k+1}^{k+\ell p} \varepsilon_j e^{2\pi i j\alpha} = z_k + \sum_{s=0}^{\ell-1} \sum_{j=k+sp+1}^{k+(s+1)p} \varepsilon_j e^{2\pi i j\alpha} \\ &= z_k + q \sum_{s=0}^{\ell-1} e^{2\pi i\alpha ps} = z_k + q \frac{e^{2\pi i\alpha p\ell} - 1}{e^{2\pi i\alpha p} - 1}.\end{aligned} \tag{2.5}$$

By definition of the sign sequence $(\varepsilon_n)_{n=0}^{\infty}$,

$$z_{k+\ell p+1} = z_{k+\ell p} + \varepsilon_{k+\ell p+1} e^{2\pi i (k+\ell p+1)\alpha}$$

which implies (since that sign has been chosen so as to minimize the absolute value)

$$\left| z_{k+\ell p} + \varepsilon_{k+\ell p+1} e^{2\pi i (k+\ell p+1)\alpha} \right| < \left| z_{k+\ell p} - \varepsilon_{k+\ell p+1} e^{2\pi i (k+\ell p+1)\alpha} \right|. \tag{2.6}$$

The next step consists of observing that $\varepsilon_{k+\ell p+1}$ is independent of p (by p-periodicity for small indices). Thus

$$\varepsilon = \varepsilon_{k+\ell p+1} e^{2\pi i (k+1)\alpha}$$

is independent of ℓ as long as $k \leq k + \ell p + 1 \leq N$. Plugging (2.5) into (2.6), we observe that, as long as $k \leq k + \ell p + 1 \leq N$,

$$\left| z_k + q\frac{e^{2\pi i\alpha p\ell} - 1}{e^{2\pi i\alpha p} - 1} + \varepsilon e^{2\pi i\ell p\alpha} \right| < \left| z_k + q\frac{e^{2\pi i\alpha p\ell} - 1}{e^{2\pi i\alpha p} - 1} - \varepsilon e^{2\pi i\ell p\alpha} \right|.$$

It would be nice if that inequality were true for all ℓ sufficiently large and, much like in the proof of Theorem 2, we see that this would be implied by the truth of the inequality

$$\forall\ t \in \mathbb{R} \qquad \left| z_k + q\frac{e^{it} - 1}{e^{2\pi i\alpha p} - 1} + \varepsilon e^{it} \right| < \left| z_k + q\frac{e^{it} - 1}{e^{2\pi i\alpha p} - 1} - \varepsilon e^{it} \right|.$$

Applying the Sampling Lemma now, just as in the proof of Theorem 2, shows that if one observes strict (quantitative) inequality for a sufficiently long time, we can deduce the truth of the inequality. The choice of the subsequence was completely arbitrary and thus, by repeating the argument $p-1$ more times, we get the desired description. □

2.2.5 Proof of Proposition 4

Proof. Note that if $z_n' = z_n + w$ with $|w| < m/2$, then, by the triangle inequality,

$$|z_n + a_n| - |w| < |z_n + w + a_n| < |z_n + a_n| + |w|,$$

as well as

$$|z_n - a_n| - |w| < |z_n + w - a_n| < |z_n - a_n| + |w|.$$

Therefore,

$$\begin{aligned} |z_n + a_n| - |z_n - a_n| - m &< |z_n + w + a_n| - |z_n + w - a_n| \\ &< |z_n + a_n| - |z_n - a_n| + m. \end{aligned}$$

Put another way, $|z_n + w + a_n| - |z_n + w - a_n|$ is not more than m away from $|z_n + a_n| - |z_n - a_n|$. Since $|z_n + a_n| - |z_n - a_n|$ is at least m in magnitude, it must be the same sign as $|z_n + w + a_n| - |z_n + w - a_n|$, which means that $z_{n+1}' = z_{n+1} + w$. We assume that $z_{-1}' = z_{-1} + w$, so by induction, we have the desired result. □

2.2.6 Proof of Proposition 5

Proof. Note that

$$e^{2\pi i(1-\alpha_0)n} = e^{2\pi i n} \cdot e^{-2\pi i\alpha_0 n} = \overline{e^{2\pi i\alpha_0 n}}.$$

Since both the initial point and the additive sequence $e^{2\pi i(1-\alpha_0)n}$ are entirely conjugated, and the Euclidean norm is unaffected by conjugation, the new sign sequence must be identical to the original. Therefore, the resulting sequence must also be entirely conjugated. Note that

$$e^{2\pi i(\alpha_0 \pm 0.5)n} = e^{2\pi i\alpha_0 n} \cdot e^{\pm\pi i n} = e^{2\pi i\alpha_0 n} \cdot (-1)^n.$$

Since the initial point is unchanged and the additive sequence $e^{2\pi i(\alpha_0 \pm 0.5)n}$ is only different by a sign, at each step of the sequence, the same two choices are presented, giving the same values of z_n. In particular, because the additive sequence alternates in sign from the original, the new sign sequence must also alternate in sign relative to the original. If $|z+w| < |z-w|$, then of course $|-z-w| < |-z+w|$. Then by induction we obtain the third statement immediately. □

2.2.7 Proof of Proposition 6

The proof of Proposition 6 will make use of a simple Lemma that will also be used in the proof of Theorem 4. We present it separately.

Lemma 5. *Let $A, B, C \in \mathbb{C}$ with $C \neq 0$. The set of complex numbers z_{-1} with the property that for all $|z| = 1$*

$$|z_{-1} + A - zB - Cz| < |z_{-1} + A - zB|$$

is an open ball of radius $-|C|/2 - \mathrm{Re}\, B\overline{C}/|C|$ centered at $-A$.

Proof. For $x, y \in \mathbb{C}$, the inequality $|x + y| < |x|$ implies, after squaring and using that $|z|^2 = z\overline{z}$, that

$$\mathrm{Re}\; x\overline{y} < -\frac{|y|^2}{2}.$$

We apply this inequality with $x = z_{-1} + A - zB$ and $y = -Cz$ Since $y = -Cz$ and $|z| = 1$, we have $-|y|^2/2 = -|C|^2/2$. This inequality can be written as

$$\text{Re }\left[(z_{-1} + A - zB)(-\overline{Cz})\right] < -\frac{|C|^2}{2}.$$

Using $(-z)(-\overline{z}) = 1$, this is

$$\text{Re }\left[(z_{-1} + A)(-\overline{Cz})\right] < -\frac{|C|^2}{2} - \text{Re } B\overline{C}$$

This inequality needs to hold for all $|z| = 1$, thus this is equivalent to

$$\max_{|z|=1} \quad \text{Re }\left[(z_{-1} + A)(-\overline{Cz})\right] < -\frac{|C|^2}{2} - \text{Re } B\overline{C}$$

We note that since z ranges over all complex numbers on the unit circle, so does $-\overline{z}$. We now argue that for any complex $w = |w|e^{i\phi}$, one has

$$\max_{0\le t\le 2\pi} \text{Re}\left[we^{it}\right] = \max_{0\le t\le 2\pi} \text{Re}\left[|w|e^{i\phi}e^{it}\right] = |w|.$$

Thus the inequality turns into

$$|z_{-1} + A| \cdot |C| < -\frac{|C|^2}{2} - \text{Re } B\overline{C}$$

or, equivalently,

$$|z_{-1} + A| < -\frac{|C|}{2} - \text{Re } B\frac{\overline{C}}{|C|}.$$

□

Proof. We start with the case where all the signs are -1, the other case will be very similar. Suppose $z_{-1} \in \mathbb{C}$ and the sequence always chooses the sign -1. Then

$$z_n = z_{-1} - \sum_{k=0}^{n} e^{2\pi i\alpha k} = z_{-1} - \frac{e^{2\pi i\alpha(n+1)} - 1}{e^{2\pi i\alpha} - 1}.$$

Since the sign is always -1, for all $n \ge -1$

$$|z_n - e^{2\pi i\alpha(n+1)}| < |z_n + e^{2\pi i\alpha(n+1)}|.$$

Plugging in the formula for z_n, we deduce $A_n < B_n$ where

$$A_n = \left| z_{-1} + \frac{1}{e^{2\pi i\alpha} - 1} - e^{2\pi i\alpha(n+1)}\left(\frac{1}{e^{2\pi i\alpha} - 1} + 1\right)\right|$$

and

$$B_n = \left| z_{-1} + \frac{1}{e^{2\pi i\alpha} - 1} - e^{2\pi i\alpha(n+1)} \left(\frac{1}{e^{2\pi i\alpha} - 1} - 1 \right) \right|.$$

α is irrational and $e^{2\pi i\alpha n}$ is dense. By density, the truth of $A_n < B_n$ for all $n \geq N$ requires, for all $|z| = 1$, the inequality $C \leq D$ where

$$\begin{aligned} C &= \left| z_{-1} + \frac{1}{e^{2\pi i\alpha} - 1} - z \left(\frac{1}{e^{2\pi i\alpha} - 1} - 1 \right) - 2z \right| \\ D &= \left| z_{-1} + \frac{1}{e^{2\pi i\alpha} - 1} - z \left(\frac{1}{e^{2\pi i\alpha} - 1} - 1 \right) \right| \end{aligned}$$

Appealing to the Lemma, the set of z_{-1} for which this holds is a disk of radius $-\mathrm{Re}(1/(e^{2\pi i\alpha} - 1)$ with center $1/(1 - e^{2\pi i\alpha})$. This real part can be computed via

$$\begin{aligned} \mathrm{Re}\, \frac{1}{1 - e^{2\pi i\alpha}} &= \mathrm{Re}\, \frac{1}{1 - \cos(2\pi\alpha) - i\sin(2\pi\alpha)} \\ &= \mathrm{Re}\, \frac{1 - \cos(2\pi\alpha) + i\sin(2\pi\alpha)}{2 - 2\cos(2\pi\alpha))} = \frac{1}{2} \end{aligned}$$

This is the desired claim. Let us now suppose the sign is always $+1$. The argument is virtually identical and after repeating the same computation with flipped signs, we arrive at

$$\left| z_{-1} - \frac{1}{e^{2\pi i\alpha} - 1} \right| \leq \frac{1}{2}.$$

$\square$

2.2.8 Proof of Proposition 7

Proof. By Proposition 5 (b), we know that our sign sequence is given by $\varepsilon_n \cdot (-1)^n$. Since ε_n is p-periodic and $(-1)^n$ is 2-periodic, we certainly have that the period of $\varepsilon_n \cdot (-1)^n$ is a factor of $2p$. In particular,

$$\varepsilon_n \cdot (-1)^n = \varepsilon_{n+2p} \cdot (-1)^{n+2p}$$

for all n. Notice that period is not a factor of 2 or p, as

$$\varepsilon_n \cdot (-1)^n = \varepsilon_{n+2} \cdot (-1)^{n+2} \Rightarrow \varepsilon_n = \varepsilon_{n+2}$$

cannot be true for all n, and

$$\varepsilon_n \cdot (-1)^n = \varepsilon_{n+p} \cdot (-1)^{n+p} \Rightarrow (-1)^n = (-1)^{n+1}$$

is a contradiction. Thus, we see that our period is indeed $2p$. $\square$

2.2.9 Proof of Theorem 4

Proof. Suppose the sequence (z_n) has signs that are $p-$periodic starting at index z_k. Then, recalling the proof of Theorem 3 and with the choice of

$$q = \sum_{j=k+1}^{k+p} \varepsilon_j e^{2\pi i j\alpha},$$

we have, for all $\ell \in \mathbb{N}$, that the subsequence $z_{k+\ell p}$ is given by

$$z_{k+\ell p} = z_k + q \sum_{s=0}^{\ell-1} e^{2\pi i \alpha p s} = z_k + q \frac{e^{2\pi i \alpha p \ell} - 1}{e^{2\pi i \alpha p} - 1}.$$

By definition of the signs $\varepsilon_n \in \{-1, 1\}$

$$z_{n+1} = z_n + \varepsilon_{n+1} e^{2\pi i (n+1)\alpha}$$

we have

$$z_{k+\ell p+1} = z_{k+\ell p} + \varepsilon_{k+\ell p+1} e^{2\pi i (k+\ell p+1)\alpha}.$$

Since the signs are chosen so as to minimize the absolute value

$$\left| z_{k+\ell p} + \varepsilon_{k+\ell p+1} e^{2\pi i (k+\ell p+1)\alpha} \right| < \left| z_{k+\ell p} - \varepsilon_{k+\ell p+1} e^{2\pi i (k+\ell p+1)\alpha} \right|.$$

Again, as in the proof of Theorem 3, we recall that $\varepsilon_{k+\ell p+1}$ is independent of ℓ because the signs are, by assumption $p-$periodic and that thus

$$\varepsilon = \varepsilon_{k+\ell p+1} e^{2\pi i (k+1)\alpha} \qquad \text{is independent of } \ell.$$

Hence, for all ℓ and with that choice of ε,

$$\left| z_k + q \frac{e^{2\pi i \alpha p \ell} - 1}{e^{2\pi i \alpha p} - 1} + \varepsilon e^{2\pi i \ell p \alpha} \right| < \left| z_k + q \frac{e^{2\pi i \alpha p \ell} - 1}{e^{2\pi i \alpha p} - 1} - \varepsilon e^{2\pi i \ell p \alpha} \right|. \tag{2.7}$$

Since α is irrational, so is αp which forces $e^{2\pi i \alpha p \ell}$ to be dense as a sequence of ℓ and we can thus deduce that

$$\forall\, t \in \mathbb{R} \qquad \left| z_k + q\frac{e^{it}-1}{e^{2\pi i \alpha p}-1} + \varepsilon e^{it} \right| \leq \left| z_k + q\frac{e^{it}-1}{e^{2\pi i \alpha p}-1} - \varepsilon e^{it} \right|. \tag{2.8}$$

This is exactly the type of inequality covered by Lemma 2 and we deduce that z_k is contained in a ball centered around some point y_k and with a certain radius r_k

$$z_k \in B(y_k, r_k).$$

The very same argument can now be carried out for z_{k+1} for which we deduce that $z_{k+1} \in B(y_{k+1}, r_{k+1})$ and so on. After we have completed one period and arrive at z_{k+p}, we stop getting any new information: one could run the argument for z_{k+p} and one would arrive at $z_{k+p} \in B(y_{k+p}, r_{k+p})$ but one could alternatively just rerun the argument for z_k and force $\ell \geq 2$. By periodicity, one would arrive at Equation (2.7) with an additional restriction on the choice of ℓ. However, since Equation (2.7) is implied by Equation (2.8), the restriction does not play any further role. We are therefore interested in the set

$$\left\{ z_{-1} \in \mathbb{C} : \forall\, k \leq n \leq k+p-1 : z_n \in B(y_n, r_n) \right\}.$$

At this point, we argue backwards: suppose $X \subset \mathbb{C}$ and $z_n \in X$. What does this tell us about z_{n-1}? We always have $z_n = z_{n-1} \pm e^{2\pi i \alpha n}$. For a given point $w \in \mathbb{C}$, the regions of z for which $|z+w| < |z-w|$ and $|z+w| > |z-w|$ are two half-spaces H_1 and H_2 and the region where $|z+w| = |z-w|$ is a line going through the origin and separating the two half spaces. Thus, the relevant 'pre-image', meaning the set of z_{n-1} which, under the procedure, would be mapped to z_n is given by

$$\left\{ z \in H_1 : z + w \in X \right\} \cup \left\{ z \in H_2 : z - w \in X \right\},$$

where $w = e^{2\pi i \alpha n}$, and H_1, H_2 are defined with respect to this w. This can also be written as $(H_1 \cap (X-w)) \cup (H_2 \cap (X+w))$. If X can be written as a finite union of finite intersections of disks and half-spaces, then the same is true for $X - w$ (by translation invariance) and the same is true for $H_1 \cap (X-w)$ which is just one more intersection with a half space.

The same argument applies to $H_2 \cap (X + w)$. Taking then an additional intersection with $B(y_{n-1}, r_{n-1})$ does not change the desired property either and we can induct backwards to z_{-1}. $\square$

In joint work with François Clément and undergraduate students through WXML, we study z_n as defined in 2.1 when a_n takes on other sequences. The key feature of a_n we would like to preserve is that its fractional part is equidistributed on $(0, 1)$. An obvious, if uninspired, choice is of course to choose a_n to be sampled from $(0, 1)$ uniformly at random. In fact, choosing a direction at random in this manner yields a generalization to $\mathbb{R}^n$ by sampling uniformly at random from $\mathbb{S}^{n-1}$. More interesting choices for a_n (still in the 2-dimensional case) are the Farey and van der Corput sequence, both of which are equidistributed on $(0, 1)$. These have rich and interesting properties that yield interesting sequences z_n in our setting.

2.3 Introduction and Results

2.3.1 Introduction

We now considering the natural and interesting generalization. Given $x \in \mathbb{R}^d$ and a sequence of vectors $v_1, v_2, \dots \in \mathbb{R}^d$, suppose we consider

$$x_n = \begin{cases} x_{n-1} + v_n & \text{if} \quad \|x_{n-1} + v_n\| < \|x_{n-1} - v_n\| \\ x_{n-1} - v_n & \text{if} \quad \|x_{n-1} - v_n\| < \|x_{n-1} + v_n\|. \end{cases}$$

What can we expect? It is clear, a priori, that one might expect a general trend towards the origin. It is not clear whether one would expect much more than that.

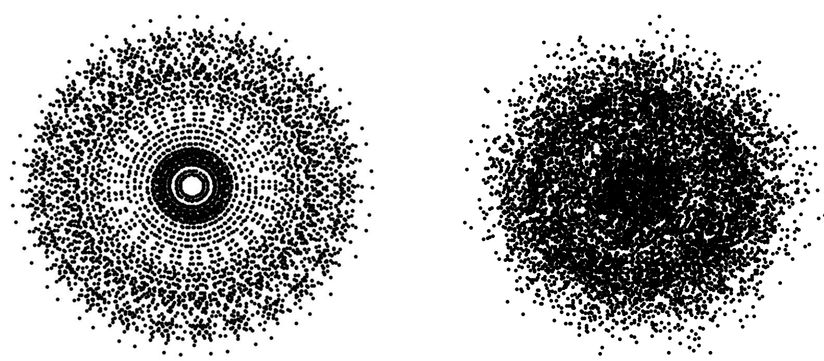


Figure 2.10: Left: starting in $0.0001 + 5i$ with $\alpha = 1.0415\sqrt{2}$ and $v_n = e^{i\alpha n}$. Right: using the vector $v_n = e^{in^2}$.

This is already highly nontrivial in the two-dimensional case. Restricting ourselves to vectors (complex numbers) of unit length one could consider sequences $(a_n)_{n=1}^{\infty}$ and then set $v_n = e^{2\pi i a_n}$. Fig. 2.11 shows three such examples. The first one comes from concatenating the Farey fractions which results in

$$0, 1, 0, \frac{1}{2}, 1, 0, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, 1, 0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, 1, 0, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \dots$$

The other two examples are derived from the van der Corput sequence which is defined as follows: to compute $x_n \in [0,1]$, write n in base b, reflect its digits around the comma and turn it back into a real number. The van der Corput sequence in base 2 begins

$$\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \dots$$

Three arising dynamical systems are shown in Figure 2.11. We will be able to say some things about the van der Corput system, see §2.3.3, but we do not have any rigorous results of any type for the Farey system.

Figure 2.11: Using the Farey sequences (left), the van der Corput sequence in base 3 (middle) and in base 8 (right).

2.3.2 The random case

It is a natural assumption that for 'many' choices of vectors, the arising dynamical system will not have any type of structure: it will be 'random'. This leads to a natural question: what do we observe when the sequence is truly random? To motivate the question, we consider the two sequences

$$z_n = z_{n-1} \pm e^{2\pi i \sqrt{2} n^2} \qquad \text{and} \qquad z_n = z_{n-1} \pm e^{2\pi i \sqrt{2} n^3}.$$

When drawing a great many points, it is difficult to see a difference, they each approximately fill out a disk. When keeping track of the distance to the origin, however, a difference seems to reveal itself (see Fig. 2.12).

As it turns out, the distribution for $\pm e^{2\pi i \sqrt{2} n^3}$ coincides with what one gets if one replaces

$\pm e^{2\pi i \sqrt{2} n^3}$ by $\pm e^{iX_n}$ where X_n is chosen uniformly at random with respect to $[0, 2\pi]$. The distribution of $\sqrt{2}n^3 \mod 2\pi$ is sufficiently 'random' to emulate the random case while, in contrast, the sequence $\sqrt{2}n^2 \mod 2\pi$ is not. Formally, we define the random sequence by $x_0 \in \mathbb{R}^d$ and

$$x_n = \begin{cases} x_{n-1} + v_n & \text{if} \quad \|x_{n-1} + v_n\| < \|x_{n-1} - v_n\| \\ x_{n-1} - v_n & \text{if} \quad \|x_{n-1} - v_n\| < \|x_{n-1} + v_n\|, \end{cases}$$

where v_n is chosen uniformly at random from $\mathbb{S}^{d-1}$. We are interested in the asymptotic density of x_n. By decoupling into radial and angular part, it is clear that if we prove sufficiently strong upper bounds on $\|x_n\|$, then the distribution is going to be radial (the angular part can be considered to be a random walk on $\mathbb{S}^{d-1}$), it thus suffices to understand $\|x_n\|$.

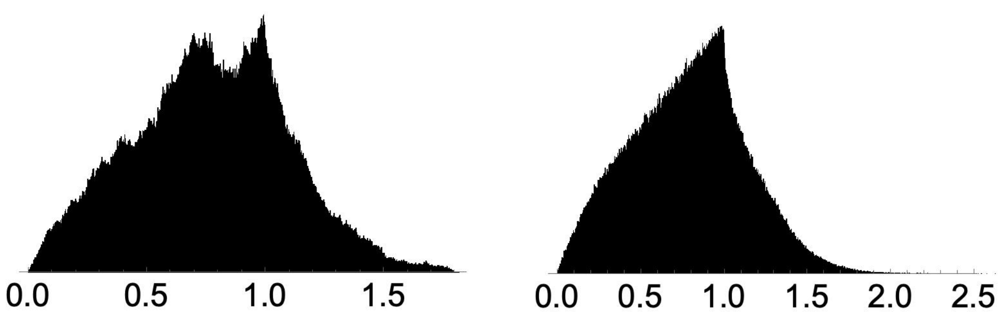


Figure 2.12: Distribution of distances from the origins for the first million terms of $\pm e^{2\pi i\sqrt{2}n^2}$ (left) and $\pm e^{2\pi i\sqrt{2}n^3}$ (right).

Theorem 5. *Suppose v_n is chosen uniformly at random from $\mathbb{S}^{d-1}$. For each dimension $d \geq 2$, for any starting point x_1, the random variable $\|x_n\|$ has a unique, invariant measure π_d on $[0, \infty]$. This invariant measure π_d has expectation*

$$\int_0^\infty x \, d\pi_d(x) = \frac{\sqrt{\pi}}{2} \frac{\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d}{2})} = \sqrt{\frac{\pi}{8}}\sqrt{d} + o(\sqrt{d})$$

which is asymptotically $(\sqrt{\pi/8} + o(1)) \cdot \sqrt{d}$. The invariant measure π_d has faster than

Gaussian decay: for every $\alpha > 0$

$$\int_0^\infty e^{\alpha x^2} d\pi_d(x) < \infty.$$

Remarks. Several remarks are in order.

1. This result can be seen as yet another illustration of the concentration of measure phenomenon [58]. When we are in $x_n \neq 0$, then most random vectors v are actually going to be 'nearly' orthogonal to $-x_n$ and thus $x_n \pm v$ is going to be typically further away from the origin than x_n as long as $\|x_n\|$ is small. Once $\|x_n\|$ is large, even a small inner product can usually be used to decrease the norm. This process self-stabilizes around $\|x_n\| \sim c_d\sqrt{d}$ and this is where $\|x_n\|$ ends up spending most of its time.

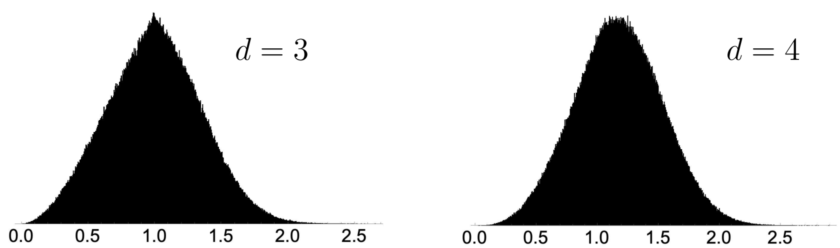


Figure 2.13: Distribution of distances to the origin for $d = 3, 4$.

2. We may think of the process as a random walk on $[0, \infty]$. The main difficulties are that it's not infinitesimally generated, it jumps, and that the jumps are of bounded size. We will prove that it is an aperiodic, recurrent Harris chain; the decay estimate is achieved by the construction of a suitable Lyapunov function.

3. The Lyapunov function implies that π_d decays quite quickly once one exceeds the expectation; one could build a similar Lyapunov function to show that it also decays quite quickly for values smaller than the expectation. We have not pursued this line of research because something stronger is true: none of the distributions π_d are Gaussian,

however, it appears that, even in relatively low dimensions, one has

$$\pi_d \quad \sim \quad \frac{e^{-(x-\mathbb{E}\pi_d)^2}}{\sqrt{\pi}}.$$

In particular, it appears as if the limiting distribution has a variance that is (asymptotically) constant. It would be interesting to make this precise.

2.3.3 *Deterministic Constructions.*

We will now move towards the deterministic setting. Steinerberger-Zeng [82] showed a remarkably degree of regularity arising from the choice of vectors $v_n = \exp(2\pi i\alpha n)$ with α irrational. The sequence $n\alpha$ mod 1 is also known as the Kronecker sequence and is one of the best known examples of a sequence with a very regular distribution [69]. This motivated us to consider the canonical example of such a sequence $(a_n)_{n\in\mathbb{N}}$: the van der Corput sequence. This classical sequence has been extensively studied since its introduction by Johannes van der Corput [92] in the 1930s. It is traditionally defined in base 2, but can be defined in any base b in the following way. Given an integer n, write

$$n = \sum_{k=0}^{\ell} a_k b^k \qquad \text{where the 'digits' satisy} \quad a_k \in \{0, \dots, b-1\}.$$

The n-th point $\mathrm{V}_b(n)$ of the base b van der Corput sequence is defined as

$$\mathrm{V}_b(n) := \sum_{k=0}^{\ell} a_k b^{-k-1}.$$

This is easiest to understand with an example. We already saw the first few terms of van der Corput sequence in base 2 above

$$\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \dots$$

More generally, we see that $a_0 = 0$, $a_1 = 1/b$ and $a_b = 1/b^2$. As above, we now consider the sequence

$$z_n = \begin{cases} z_{n-1} + e^{2\pi i\,\mathrm{V}_b(n)} & \text{if } |z_{n-1} + e^{2\pi i\,\mathrm{V}_b(n)}| < |z_{n-1} - e^{2\pi i\,\mathrm{V}_b(n)}| \\ z_{n-1} - e^{2\pi i\,\mathrm{V}_b(n)} & \text{if } |z_{n-1} + e^{2\pi i\,\mathrm{V}_b(n)}| > |z_{n-1} - e^{2\pi i\,\mathrm{V}_b(n)}| \end{cases}$$

We start by showing that if we start anywhere in the plane and use the van der Corput sequence in base 2, then one reaches a fixed neighborhood of the origin in a number of steps proportional to the original distance from the origin.

Theorem 6. *Let $b = 2$ and $z_{-1} \in \mathbb{R}$. If the sequence is not indeterminate by then, it takes at most $O(|z|)$ steps of the sequence to reach $B(0, \sqrt{2})$.*

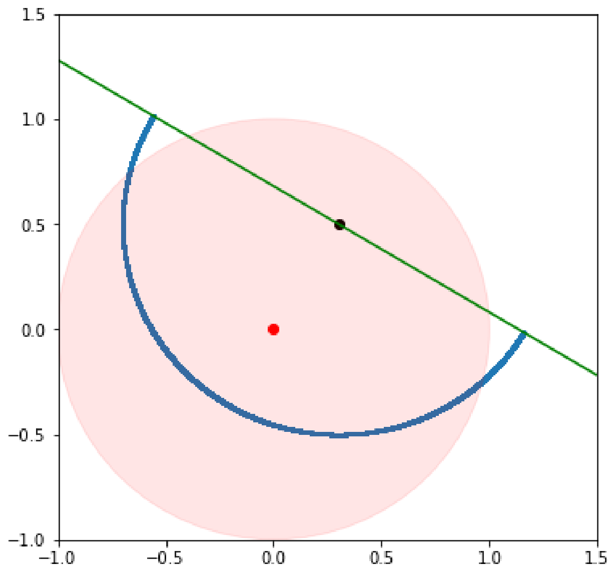


Figure 2.14: The sequence associated to base 2 for $z_{-1} \in B(0,1)\backslash\{(0,0)\}$. In red, the origin and $B(0,1)$. z_{-1} is in black, and all other distinct points in the sequence are in blue. All points are in the semi-circle defined by the green line.

While similar results with varying ball sizes could be obtained for different b, we limit our proofs to base 2 and some insights into the base 3 case. We call *periodic starts* initial positions $z_{-1} \in \mathbb{R}^2$ such that $z_{kb-1} = z_{-1}$ for any $k \in \mathbb{N}$ and the sign choice at all steps is always the same. We also show that no even b other than $b = 2$ has periodic starts, while all odd b and $b = 2$ have such periodic starts.

Theorem 7. *Let $b = 2$. Any initial value in $B(0,1)\backslash\{(0,0)\}$ is a periodic start. When z_n enters the unit ball and n is odd, then subsequent elements of the sequence are contained in* *a point and a semicircle centered on that point of radius 1.*

As the base b increases, things become more complicated. When b is even, we can prove that no periodic starts exist. In that sense, the case $b = 2$ is special. If $b \geq 5$ is odd, then periodic starts always exist and we can identify two regions that yield periodic starts.

Theorem 8. *Let $b \geq 4$. Then, considering $(a_n)_{n\in\mathbb{N}} = \mathrm{V}_b(n)$,*

1. *if b is even, no periodic starts exist.*

2. *if b is odd, the periodic starts form two open triangles.*

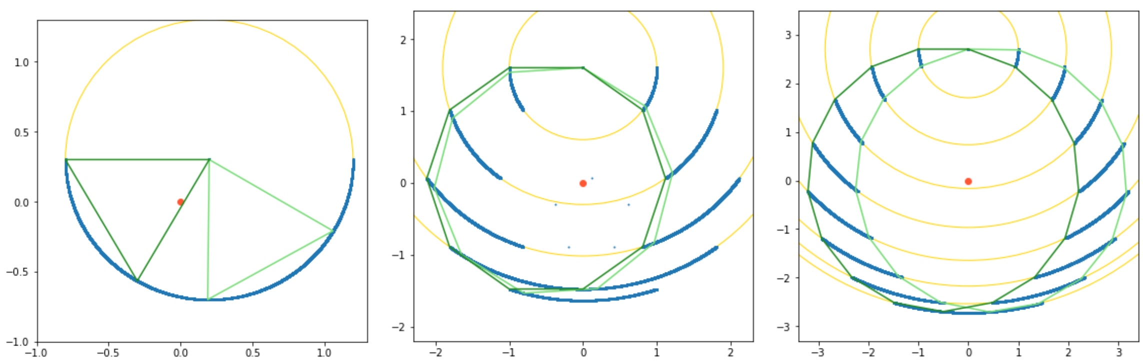


Figure 2.15: The polygons corresponding to the first (dark green) and 100^{th} (light green) set of b points of the sequence associated with base b van der Corput, for $b = 3$, 10 and 17 (left to right), with a fixed $z{-}1$ at the center of the orange circles. 200 000 points are computed. The blue arcs correspond to all the points of the sequence, while the orange circles correspond to the different distances from z_{-1} for these points, as described in the proof of Theorem 9.

Empirically, periodic starts also exist when $b = 3$ but that argument would require additional work. We conclude by showing that periodic starts lead to relatively uninteresting sequences: they trace out a regular polygon in a periodic manner, partially rotated around z_{-1}.

Theorem 9. *Let $b \geq 2$. For any periodic start $z_{-1} \in \mathbb{R}^2$, the arising sequence forms rotated copies of a regular $b-$sided polygon.*

In the even b case, this characterization holds for a certain number of steps in the sequence that depends on z_{-1}, given in Proposition 9.

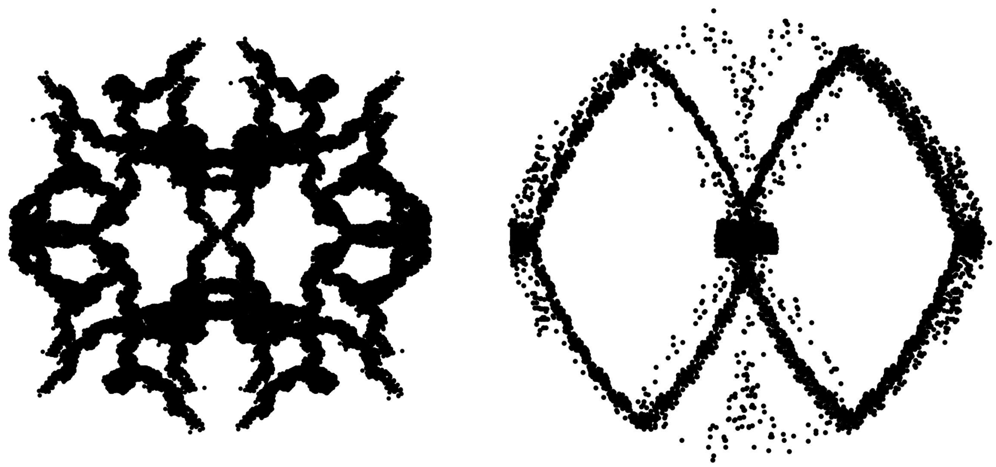


Figure 2.16: Left: $v_n = \exp(2\pi i \|\sqrt{3}n\|)$. Right: $v_n = \exp(2\pi i \|\sqrt{10}n\|)$.

2.3.4 More open problems.

Needless to say, the results presented up to now contain their fair number of associated problems. However, there are more. Indeed, it is worth noting that these types of greedy sequences

$$x_n = \begin{cases} x_{n-1} + v_n & \text{if} \quad \|x_{n-1} + v_n\| < \|x_{n-1} - v_n\| \\ x_{n-1} - v_n & \text{if} \quad \|x_{n-1} - v_n\| < \|x_{n-1} + v_n\| \end{cases}$$

appear to give rise to the most peculiar patterns. An example, starting in $x_0 = (0.01, 0.01)$ and using the sequence of vectors $v_n = \exp(2\pi i \|\sqrt{3}n\|)$ with $\|\cdot\|$ denoting the distance to the nearest integer gives rise to points in a strange shape shown in Figure 2.16 (left). Taking $v_n = \exp(2\pi i \|\sqrt{10}n\|)$ gives rise to the shape shown in Figure 2.16 (right), both remain unexplained.

All of these examples are two-dimensional with a sequence of vectors $(v_n)_{n=1}^{\infty}$ that has unit norm. We conclude by giving two examples that do not follow these constraints. First, we consider the sequence of vectors $v_n = \sqrt{n}\exp(2\pi i\sqrt{2}n)$ when starting in $v_0 = (0, 0.1)$. The second example takes place in three dimensions with $v_n = (\cos n, \sin n, \cos\sqrt{n})$ and $v_0 = (0, 0, 0.1)$. We believe that these examples are indicative of a surprising richness and versatility of dynamical systems of this type, many open problems remain.

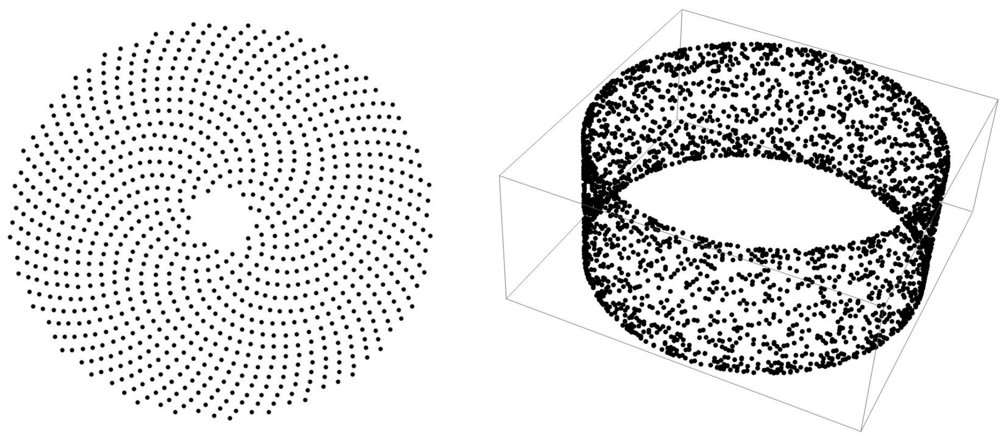


Figure 2.17: Left: $v_n = \sqrt{n}\exp(2\pi i\sqrt{2}n)$. Right: $v_n = (\cos n, \sin n, \cos\sqrt{n})$.

2.4 *Proof of Theorem 5*

2.4.1 A stochastic representation

The purpose of this section is to give a slightly different way of phrasing the process. We can assume, using invariance under rotations, that

$$x_n = (\|x_n\|, 0, 0, 0, \dots)$$

and that $x_{n+1} = x_n \pm v$ where $v = (v_1, \dots, v_d)$ is chosen uniformly at random from $\mathbb{S}^{d-1}$. Then

$$x_n \pm v = (\|x_n\|, 0, 0, \dots, 0) \pm (v_1, v_2, v_3, \dots, v_d)$$

and

$$\|x_n \pm v\|^2 = (\|x_n\| \pm v_1)^2 + v_2^2 + \cdots + v_d^2.$$

In this representation, it is clear that we will choose the sign so that

$$\begin{aligned}\|x_{n+1}\|^2 &= (\|x_n\| - |v_1|)^2 + v_2^2 + \cdots + v_d^2 \\ &= \|x_n\|^2 - 2 \cdot |v_1| \cdot \|x_n\| + 1.\end{aligned}$$

A classic way to sample uniformly at random from $\mathbb{S}^{d-1}$ is to pick d independent $\mathcal{N}(0,1)$ Gaussians $(\gamma_1, \ldots, \gamma_d)$ and then to normalize that vector. The first coordinate is distributed like

$$\frac{\gamma_1}{(\gamma_1^2 + \gamma_2^2 + \cdots + \gamma_d^2)^{1/2}} = \frac{\gamma_1}{(\gamma_1^2 + \chi^2(d-1))^{1/2}},$$

where $\chi^2(d-1)$ is a chi-squared random variable of degree $d-1$. We may thus equivalently rewrite the process as

$$\|x_{n+1}\| = \left(\|x_n\|^2 - 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\| + 1\right)^{1/2}$$

which is a useful way of rewriting things.

2.4.2 Existence of a limiting measure.

Proof. Here we make use of the Harris chain framework [48]. We fix the dimension d. A Markov chain on $\Omega = [0, \infty)$ with a stochastic transition kernel is called a Harris chain if there exists a set $A \subset \Omega$, a constant $\varepsilon > 0$ and a probability measure $\rho(\Omega) = 1$ such that

1. the hitting time $\tau_A = \inf\{n \geq 0 : X_n \in A\}$ is finite almost surely

$$\mathbb{P}(\tau_A < \infty | x_0 = x) = 1 \qquad \text{for all } x \in \Omega.$$

2. For any $x \in A$ and any $C \subset \Omega$, we have

$$\mathbb{P}\left(x_{n+1} \in C \middle| x_n \in A\right) \geq \varepsilon \rho(C).$$

We choose

$$A = \left[\sqrt{d}, \sqrt{d} + \frac{1}{2}\right].$$

The first step is to prove that when starting in x, the expected hitting time τ_A is finite. We first argue that if $0 \leq \|x_n\| \leq \sqrt{d}$, then there is a positive probability to hit A in the next $2d$ steps. If we do not hit A, then either $0 \leq \|x_{n+2d}\| \leq \sqrt{d}$ in which case we can run the argument again or $\|x_{n+2d}\| \geq \sqrt{d} + 1/2$. We will then show, by means of a separate argument, that if $\|x_n\| \geq \sqrt{d} + 1/2$, then there is drift back to the origin that leads us back to the regime $\|x_{n+k}\| \leq \sqrt{d}$. Let us first assume that $0 \leq \|x_n\| \leq \sqrt{d}$. We make use of the explicit representation

$$\|x_{n+1}\| = \left(\|x_n\|^2 - 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\| + 1\right)^{1/2}.$$

For each point $0 \leq \|x_n\| \leq \sqrt{d}$ there is a nonzero probability that the next $\sim 2d$ steps have $|\gamma_1| \leq 1/100$ and $|\chi^2(d-1) - d| \leq \sqrt{d}$ which leads to growth of the sequence. Note that

$$\|x_{n+1}\| \leq \left(\|x_n\|^2 + 1\right)^{1/2} \leq \|x_n\| + \frac{1}{2\|x_n\|}.$$

This means that if we take $\sim 2d$ steps, we will surpass $\sqrt{d}$. Moreover, when $\|x_n\| \sim \sqrt{d}$, the jumps are of size $\lesssim d^{-1/2} < 1/2$ and we are sure to hit A. Let us now assume that $\|x_n\| > \sqrt{d} + 1/2$. It is clear from the representation and the argument above that then

$$\|x_n\| - 1 \leq \|x_{n+1}\| \leq \|x_n\| + \frac{1}{2\|x_n\|}.$$

We argue that the process is much more likely to decrease. One has

$$\begin{aligned}\|x_{n+1}\|^2 - \|x_n\|^2 &= -2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\| + 1 \\ &\leq -2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \sqrt{d} + 1.\end{aligned}$$

Thus, conditioning on $\|x_n\| \geq \sqrt{d}$, we have

$$\mathbb{E}\left[\|x_{n+1}\|^2 - \|x_n\|^2\right] \leq 1 - 2 \cdot \mathbb{E}\, \frac{|\gamma_1|\sqrt{d}}{(\gamma_1^2 + \chi^2(d-1))^{1/2}}.$$

In large dimensions, we can use that $\chi^2(d-1) = d \pm \mathcal{O}(\sqrt{d})$ to deduce that

$$\mathbb{E}\left[\frac{|\gamma_1|\sqrt{d}}{(\gamma_1^2 + \chi^2(d-1))^{1/2}}\right] = \sqrt{\frac{2}{\pi}} + o(1) > \frac{1}{2}$$

which implies negative drift back to the origin. In small dimensions, one can replace $||x_n|| > \sqrt{d}$ by $||x_n|| > 10000\sqrt{d}$ and obtain the same estimate with fairly crude bounds on the $\chi^2(d-1)$ distribution. (Here, 10000 is a concrete number that will work, in practice one could also choose this number depending on the dimension).

As for the second part of the construction, we use a standard approach in the area and let ρ be a multiple of the Lebesgue measure localized to an interval

$$\rho = 1_{\sqrt{d}-1/4<x<\sqrt{d}} \cdot 4dx.$$

The representation

$$\|x_{n+1}\| = \left(\|x_n\|^2 - 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\| + 1\right)^{1/2}$$

shows that $\|x_{n+1}\|$ can be anywhere in $[\|x_n\| - 1, \sqrt{\|x_n\|^2 + 1}]$ and it does so with a density that is uniformly bounded away from 0 in any strict subinterval.

This shows that if $\sqrt{d} < \|x_n\| < \sqrt{d} + 1/2$, then the likelihood of $\|x_{n+1}\|$ being somewhere in $[\sqrt{d} - 1/4, \sqrt{d}]$ is bounded away from 0 with positive probability. This shows that the process is a recurrent Harris chain. It is easy to see from the definition that it is aperiodic. The Harris chain is recurrent and thus, see [66, Theorem 10.0.1], it has a unique (up to multiples) invariant measure. Moreover, the argument also shows that it is positively recurrent. $\square$

2.4.3 A Lyapunov Lemma.

The purpose of this section is to establish a simple inequality for a Lyapunov function. We will then use it in the subsequent section to establish existence of all moments and fast decay of π_d.

Lemma 6. *Let $\alpha > 0$. Then, for some $c'_d > 0$ depending only on the dimension*

$$\mathbb{E}\left[e^{\alpha\|x_{n+1}\|^2}\right] \le \frac{e^{\alpha}c'_d}{2\alpha\|x_n\|}\cdot e^{\alpha\|x_n\|^2}.$$

Proof. We use

$$\|x_{n+1}\|^2 = \|x_n\|^2 - 2\cdot\frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}\cdot\|x_n\| + 1.$$

Rewriting this equality, we get

$$\begin{aligned}\mathbb{E}\left[e^{\alpha\|x_{n+1}\|^2}\right] &= \mathbb{E}\left[\exp\left(\alpha\|x_n\|^2 - 2\alpha\cdot\frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}\cdot\|x_n\| + \alpha\right)\right]\\ &= e^{\alpha\|x_n\|^2}\ e^{\alpha}\cdot\mathbb{E}\left[\exp\left(-2\alpha\|x_n\|\cdot\frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}\right)\right].\end{aligned}$$

We will try to bound this rather precisely by thinking of

$$X = \frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}$$

as the absolute value of a coordinate of a randomly chosen point on the sphere. To find the probability density of X, we let $A_d(r) = a_d r^d$ denote the surface area of a d-sphere with radius r, where $a_d \in \mathbb{R}$ is a dimension-dependent constant. Firstly,

$$A_d(1) = \int_0^{\pi} A_{d-1}(\sin\theta)d\theta = \int_0^{\pi} a_{d-1}\sin^{d-1}\theta d\theta$$

Due to the fact there's an absolute value, we integrate solely on the top cap of the sphere, hence from $\arccos x$ to $\pi/2$.

$$\begin{aligned}\mathbb{P}_d(X\le x) &= \frac{\int_{\arccos x}^{\frac{\pi}{2}} A_{d-2}(\sin\theta)d\theta}{\int_0^{\frac{\pi}{2}} A_{d-2}(\sin\theta)d\theta} = \frac{\int_{\arccos x}^{\frac{\pi}{2}}\sin^{d-2}\theta d\theta}{\int_0^{\frac{\pi}{2}}\sin^{d-2}\theta d\theta}\\ &= \frac{2\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}\int_{\arccos x}^{\frac{\pi}{2}}\sin^{d-2}\theta d\theta = c_d\int_{\arccos x}^{\frac{\pi}{2}}\sin^{d-2}\theta d\theta.\end{aligned}$$

Thus,

$$p_d(x) = \frac{d}{dx}\mathbb{P}_d(X\le x) = c_d(1-x^2)^{\frac{d-3}{2}}.$$

Therefore, for any parameter $t > 0$ and $d \geq 3$, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(-t\frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}\right)\right] &= \int_0^1 e^{-tx}p_d(x)dx \\ &\leq c_d\int_0^1 e^{-tx}dx \\ &\leq c_d\int_0^\infty e^{-tx}dx = \frac{c_d}{t}.\end{aligned}$$

In the $d = 2$ case, for any parameter t, we have that

$$\begin{aligned}\mathbb{E}\left[\exp\left(-t\frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}\right)\right] &= \int_0^1 e^{-tx}p_d(x)dx \\ &= c_d\int_0^1 \frac{e^{-tx}}{\sqrt{1-x^2}}dx.\end{aligned}$$

If $t \geq 1$, this integral is less than $2/t$ and we get the desired result up to the replacement of c_d by some other dimension-dependent constant c'_d. For $t < 1$, the crude bound $e^{-tx} \leq 1$ gives that the integral is less than $2/t$. Thus

$$\mathbb{E}\left[\exp\left(-2\alpha\|x_n\|\cdot\frac{|\gamma_1|}{(\gamma_1^2+\chi^2(d-1))^{1/2}}\right)\right] \leq \frac{c'_d}{2\alpha\|x_n\|}.$$

□

2.4.4 *Boundedness of Exponential Moments*

We make use of the following lemma from Meyn-Tweedie [66]. We state it as it appears in [66], they use the abbreviations

$$PV(x) = \int_\Omega V(y)P(x,dy) \qquad \text{and} \qquad \pi(f) = \int_0^\infty f(x)\pi(dx) = \mathbb{E}_\pi[f(X)].$$

With this notation, the statement reads as follows.

Lemma 7 (Theorem 14.3.7 [66])**.** *Suppose that a Markov chain Φ is positive recurrent with invariant probability π, and suppose that V, f and s are non-negative, finite-valued functions on X such that*

$$PV(x) \leq V(x) - f(x) + s(x)$$

for every $x \in X$. Then $\pi(f) \leq \pi(s)$.

In our case, as in the previous proof, Φ is the Markov chain given by the $||x_n||$, with x_n defined just before Theorem 5. We will now use this to prove that for every $\alpha > 0$

$$\int_0^\infty e^{\alpha x^2} d\pi_d(x) < \infty.$$

Proof. We choose $V(x) = e^{\alpha x^2}$. The lemma in the previous section shows that for all x sufficiently large there exists a c'_d (depending only on the dimension) such that

$$PV(x) \leq \frac{e^\alpha c'_d}{2\alpha x} \cdot V(x).$$

This estimate is not particularly good when $0 \leq x \leq 1$ in which case, we can always use that $x_n \leq 2$ implies that $x_{n+1} \leq 3$ and thus $x \leq 2 \implies PV(x) \leq C$. Let $z = e^\alpha c'_d/\alpha$. Then, for $x \geq z$, we have $PV(x) \leq V(x)/2$. Thus, after possibly increasing the constant c'_d, there exists a constant c''_d depending only on the dimension such that

$$PV(x) \leq \frac{e^\alpha c''_d}{2\alpha \max\{x,1\}} \cdot V(x).$$

This means that

$$f(x) = \frac{V(x)}{2} 1_{x>z} \qquad \text{and} \qquad s(x) = \frac{e^\alpha c''_d}{2\alpha \max\{x,1\}} \cdot V(x) 1_{x<z}$$

are valid choices to ensure

$$PV(x) \leq V(x) - f(x) + s(x)$$

for all $x \geq 0$. The lemma implies that

$$\pi_d(f) = \frac{1}{2}\int_z^\infty V(x) d\pi_d(x) \leq \int_0^z \frac{e^\alpha c''_d}{2\alpha \max\{x,1\}} V(x) d\pi_d(x) < \infty.$$

However,

$$\int_0^\infty e^{\alpha x^2} d\pi_d(x) = \int_0^z e^{\alpha x^2} d\pi_d(x) + \int_z^\infty e^{\alpha x^2} d\pi_d(x)$$

and the first integral is $\leq e^{\alpha z^2} < \infty$ and the second integral is finite and we have the desired result. $\square$

2.4.5 *Computing the mean (informally).*

The decay estimate from the preceding section implies that the second moment exists. We can use this to determine the mean exactly. We first give an informal estimate. Using

$$\|x_{n+1}\|^2 = \|x_n\|^2 - 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\| + 1$$

and iterating this recursion, we deduce

$$\|x_{n+k}\|^2 - \|x_n\|^2 = k - \sum_{\ell=0}^{k-1} 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\|.$$

Taking an average, we have

$$\frac{\|x_{n+k}\|^2 - \|x_n\|^2}{k} = 1 - \frac{1}{k}\sum_{\ell=0}^{k-1} 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\|.$$

Since the second moment exists, we can take an expectation on both sides and let $k \to \infty$ to deduce that

$$0 = 1 - \mathbb{E}\left[2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}}\right] \mathbb{E}\left[\|x_n\|\right]$$

which suggests that the first moment is

$$\int_0^\infty x \; d\pi_d = \left(\mathbb{E}\left[\frac{2|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}}\right]\right)^{-1}.$$

This formula can also be used to deduce the asymptotic behavior: as d increases, we get tight concentration of the χ^2-random variable $\chi^2(d-1) \sim d \pm \sqrt{d}$ suggesting

$$\mathbb{E}\left[\frac{2|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}}\right] \sim \mathbb{E}\left[\frac{2|\gamma_1|}{\sqrt{d}}\right] = \frac{2}{\sqrt{d}}\sqrt{\frac{2}{\pi}}$$

which suggests that, as $d \to \infty$,

$$\mathbb{E}\left[\|x_n\|\right] = \frac{\sqrt{\pi}}{\sqrt{8}}\sqrt{d} + o(\sqrt{d}).$$

This is indeed the case, and we will later determine the mean exactly.

2.4.6 *Another stochastic representation*

As is perhaps unsurprising, we could also try to formulate many of these arguments using an explicit representation of the transition kernel.

Lemma 8. *The stationary distribution* π *satisfies*

$$\pi(y) = \begin{cases} \int_{1-y}^{y+1} P_d(x,y)\pi(x)dx & \text{if } 0 \leq y < 1 \\ \int_{\sqrt{y^2-1}}^{y+1} P_d(x,y)\pi(x)dx & \text{if } y \geq 1 \end{cases}$$

where P_d *is the transition kernel in* d *dimensions:*

$$P_d(x,y) = \frac{2\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \frac{y}{x} \left(1 - \left(\frac{x^2 - y^2 + 1}{2x}\right)^2\right)^{\frac{d-3}{2}}.$$

Proof. By definition of stationarity, π must satisfy

$$\pi(y) = \int_{-\infty}^{\infty} P_d(x,y)\pi(x)dx.$$

As a brief note on bounds of integration, if $0 \leq y < 1$ then we can either step down to it from $y+1$ or step through the origin from $1-y$, and if $y \geq 1$ then we can again either step down to it from $y+1$ or step up to it from $\sqrt{y^2-1}$. Recall that we can interpret $P_d(Y \leq y | X = x)$ geometrically as the surface area ratio of a spherical cap to a hemisphere (in $d-1$ dimensions). The angle defining this cap is given by the law of cosines to be

$$\theta(x,y) = \arccos\left(\frac{x^2 - y^2 + 1}{2x}\right).$$

Using the same construction as we did for Lemma 1, we have

$$\mathbb{P}_d(Y \leq y | X = x) = c_d \int_0^{\theta(x,y)} \sin^{d-2}\theta d\theta.$$

This gives

$$P_d(x,y) = \frac{d}{dy} P_d(Y \leq y | X = x) = c_d \sin^{d-2}\theta(x,y) \frac{\partial}{\partial y}\theta(x,y)$$

where

$$\sin\theta(x,y) = \left(1 - \left(\frac{x^2 - y^2 + 1}{2x}\right)^2\right)^{\frac{1}{2}},$$

$$\frac{\partial}{\partial y}\theta(x,y) = \frac{y}{x}\left(1 - \left(\frac{x^2 - y^2 + 1}{2x}\right)^2\right)^{-\frac{1}{2}}.$$

Combining this all together, we obtain

$$P_d(x,y) = c_d \frac{y}{x}\left(1 - \left(\frac{x^2 - y^2 + 1}{2x}\right)^2\right)^{\frac{d-3}{2}}.$$

□

Remark. There is one particularly nice special case: when $d = 3$, the transition kernel is

$$P_3(x,y) = \frac{y}{x}$$

and the invariant measure satisfies, for $y \geq 1$,

$$\pi_3(y) = \int_{\sqrt{y^2-1}}^{y+1} \frac{y}{x}\pi_3(x)dx.$$

This can be rewritten as the curious identity

$$\forall\ y \geq 1 \qquad \frac{\pi_3(y)}{y} = \int_{\sqrt{y^2-1}}^{y+1} \frac{\pi_3(x)}{x}dx.$$

It is a natural question whether this identity could be used to derive more precise results when $d = 3$.

2.4.7 Computing the mean (formally).

As we did with our previous mean calculation, we begin by setting up

$$\|x_{n+1}\|^2 = \|x_n\|^2 - 2 \cdot \frac{|\gamma_1|}{(\gamma_1^2 + \chi^2(d-1))^{1/2}} \cdot \|x_n\| + 1.$$

Taking the expectation we get

$$\mathbb{E}[\|x_{n+1}\|^2] = \mathbb{E}[\|x_n\|^2] - 2\mathbb{E}[X]\mathbb{E}[\|x_n\|] + 1$$

where X is shown in the proof for Lemma 1 to satisfy

$$p_d(x) = c_d(1-x^2)^{\frac{d-3}{2}}$$

Letting $n \to \infty$, since we have stationarity and finite second moments, we can write

$$\mathbb{E}[\|x_n\|] = \frac{1}{2\mathbb{E}[X]}.$$

We have

$$\mathbb{E}[X] = c_d \int_0^1 x(1-x^2)^{\frac{d-3}{2}} dx = \frac{c_d}{2}\int_0^1 (1-u)^{\frac{d-3}{2}} du = \frac{c_d}{2}\frac{2}{d-1}$$

Thus, reinserting the previous expression for c_d, we obtain

$$\mathbb{E}[X] = \frac{2\Gamma(\frac{d}{2})}{\sqrt{\pi}(d-1)\Gamma(\frac{d-1}{2})}$$

Putting this all together and using $\Gamma(x+1) = x\Gamma(x)$,

$$\mathbb{E}[\|x_n\|] = \frac{\sqrt{\pi}\Gamma(\frac{d+1}{2})}{2\Gamma(\frac{d}{2})}.$$

2.5 Van der Corput: base 2

For this section, $(a_n)_{n\in\mathbb{N}}$ corresponds to the base 2 van der Corput sequence.

2.5.1 Understanding the different regimes

This subsection focuses on proving a set of results that we require for the proof of Theorem 6. These results will be slightly more general than necessary to prove Theorem 6, in particular Lemma 9, but provide a complete picture of the behavior of the sequence. We point out to the reader that while the van der Corput sequence is defined from 0 onwards, our process begins in z_{-1}. As such, there is an offset of 1 in all the modulo conditions: the first step from z_{-1} to z_0 uses $\mathrm{V}_2(0)$. We begin by showing that every 2 points, the norm cannot have increased.

Proposition 8. *For all $z_{-1} \in \mathbb{R}^2$, for all $n \in \mathbb{N}$, either the sequence is indeterminate or $|z_{2n-1}| \geq |z_{2n+1}|$, with equality if and only if $z_{2n-1} = z_{2n+1}$.*

Proof. Consider any starting position $z_{-1} \in \mathbb{R}$ and $n \in \mathbb{N}$. By definition of the van der Corput sequence, $\mathrm{V}_2(2n) = \mathrm{V}_2(2n+1) - 1/2$. z_{2n-1}, z_{2n} and z_{2n+1} are aligned, and the two possible choices for $+$ and $-$ in z_{2n} are exactly z_{2n-1} and its symmetric relative to z_{2n}. If the symmetric has strictly smaller norm, it is chosen as z_{2n+1}. If it has equal norm, the sequence is no longer determined. If it has strictly higher norm, then $z_{2n+1} = z_{2n-1}$. These three cases prove the desired result. □

To obtain a guarantee on the decrease of the norm, we need to study more than 2 consecutive steps. The following lemmas study blocks of 4 or 8 steps to cover all possible cases. This relies on the following observation: for $k = 0$ modulo b^j, $\mathrm{V}_b(k + b^{j-1}) - \mathrm{V}_b(k) = 1/b^j$.

Lemma 9. *Let $k \equiv -1$ modulo 4. Then if z_k exists and $|z_k| < \sqrt{2}$, we have at least one of the following:*

1. $z_k = z_{k+2}$
2. $z_{k+2} = z_{k+4}$.

Proof. Let $k \equiv -1$ modulo 4. By the definition of the van der Corput sequence, we know $\mathrm{V}_b(k+1) = \mathrm{V}_b(k) + 1/2$, $\mathrm{V}_b(k+3) = \mathrm{V}_b(k+2) + 1/2$ and $\mathrm{V}_b(k+2) = \mathrm{V}_b(k) + 1/4$. In other words, the first two steps will be along the same line D_1, and the next two steps along the line D_2 orthogonal to D_1 and intersecting it in z_{k+2}. We will prove the result by contradiction. If $z_k \neq z_{k+2}$ and $z_{k+2} \neq z_{k+4}$, then $|z_{k+4}| < |z_{k+2}| < |z_k|$ by Proposition 8. Since none of z_k, z_{k+2} and z_{k+4} are equal to each other, the steps in each pair of steps were done in the same direction. We have $|z_{k+2} - z_k| = 2$ and $|z_{k+4} - z_{k+2}| = 2$ and the three points z_k, z_{k+2} and z_{k+4} form a triangle with $\widehat{z_{k+4} z_{k+2}} z_k = \pi/2$ modulo π. We therefore have $|z_{k+4} - z_k| = 2\sqrt{2}$ and, by hypothesis, $|z_k| < \sqrt{2}$. This gives $|z_{k+4}| > \sqrt{2} > |z_k|$, a contradiction. □

Lemma 10. *Let $k \equiv -1$ modulo 4. Suppose that z_k exists and that $|z_k| = r > 1$. If $z_k = z_{k+2}$, then*

$$|z_k| - |z_{k+4}| > r - \sqrt{r^2 + 4 - 4\sqrt{r^2 - 1}}.$$

Furthermore, if $\sqrt{2} < |z_k| < 2$, then $z_k \neq z_{k+4}$ and $|z_{k+4}| < \sqrt{2}$.

Proof. Let $k \equiv -1$ modulo 4 such that $|z_k| = r > 1$ and $z_k = z_{k+2}$. Firstly, we suppose without loss of generality that z_k is given by $(r, 0)$. We are only trying to determine distances, and these are invariant by rotation. As illustrated in Figure 2.18, if $z_{k+2} = z_k$, this means that the intersections between the circle C_z of radius 2 centered in z_k and the line D_1 defined by the step with $\mathrm{V}_2(k+1)$ are outside the circle C_o of radius r centered in the origin. We also know that the direction D_2 of the next two steps is orthogonal to D_1. We will now determine the worst norm possible for z_{k+4}.

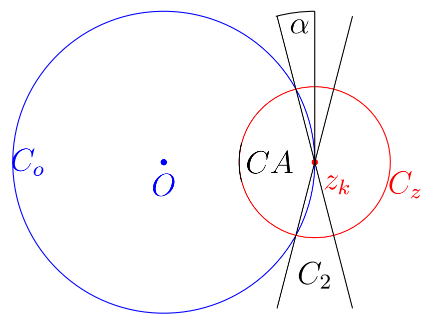


Figure 2.18: Given a failed step in the double cone C_2, the next step has to be in the arc CA.

Let α be the angle between the vertical line $y = r$ and the line passing through z_k and the upper intersection of C_o and C_z. The directions that would lead to $z_k = z_{k+2}$ are in the "double" cone C_2 of angle 2α with summit z_k. The possible next steps are formed by the circle arc CA defined by the intersection between C_z and the lines orthogonal to those contained in C_2. We first determine α. Let (x, y) be the upper intersection of C_o and C_z. We have $x^2+y^2 = r^2$ and $(x-r)^2+y^2 = 4$. Solving this gives us $x = r-2/r$ and $y = 2\sqrt{1-1/r^2}$. We then obtain $\sin(\alpha) = 1/r$. The point p on CA furthest from the origin O is such that $\widehat{Oz_kp} = \alpha$. Its y-coordinate p_y is such that $\sin(\alpha) = p_y/2 = 1/r$. As such, $p_y = 2/r$. Its x-coordinate is $r - 2\sqrt{1-1/r^2}$. We can now determine the distance d to the origin,

$$d = \sqrt{r^2 + 4 - 4\sqrt{r^2-1}}.$$

Among all the points on the arc CA, this is the furthest away from the origin, as CA is symmetric relative to the x-axis. Finally, we obtain that $|z_k| - |z_{k+4}| > r - \sqrt{r^2+4-4\sqrt{r^2-1}}$, which is the first part of the desired result. We now suppose that $\sqrt{2} < |z_k| < 2$. Since

$z_k = z_{k+2}$, the origin has to be in a band of width 1 orthogonal to the direction D_1 imposed by $\mathrm{V}_2(k+1)$. Let D_2 be the direction of this band. Since $k \equiv -1$ modulo 4, the next two steps of the process are along this direction D_2. If $z_{k+4} = z_k$, then the origin must be strictly inside a square of side 1 with one corner equal to z_k. This would mean $|z_k| < \sqrt{2}$, a contradiction. As such, $z_{k+4} \neq z_k$. In this case, since $|z_k| < 2$, the origin has to be in a rectangle R of side $[z_k, z_{k+4}]$ of length 2 and whose adjacent side is of length 1. Since $|z_k| > \sqrt{2}$, we also know that it will be strictly inside the square S_2 of size 1 that forms half of R and has z_{k+4} as a summit. We conclude $|z_{k+4}| < \sqrt{2}$. □

Lemma 11. *Let $k \equiv -1$ modulo 8. If z_k exists and z_k, z_{k+2} and z_{k+4} are all distinct, then $|z_k| - |z_{k+8}| \geq 2(\sqrt{2}-1)$ if $|z_k| > 2$, and $|z_{k+8}| < 1 + (\sqrt{2}-1)^2$ otherwise.*

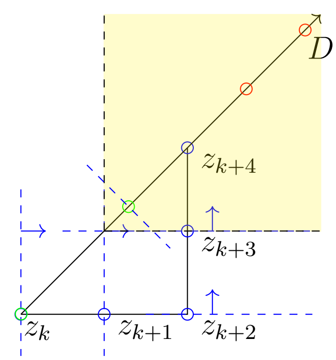


Figure 2.19: The possible positions of z_{k+5} and z_{k+6} are in red and green, and we could have $z_{k+6} = z_{k+4}$. In yellow, the cone C where the origin could be based on the first 4 steps. The non-dashed edges indicate the sides where the region is infinite.

Proof. Let $k \equiv -1$ modulo 8, such that z_k, z_{k+2} and z_{k+4} are all distinct. By contrapositive of Lemma 9, we have $|z_k| > \sqrt{2}$. The reasoning used in the proof of Lemma 9 relative to the positioning of z_k, z_{k+2} and z_{k+4} still holds: they form a triangle with right-angle for $z_{k+4}\widehat{z_{k+2}}z_k$. Furthermore, the region where the origin can be is a cone C with summit in the middle of $[z_k, z_{k+4}]$ and with sides parallel to, and with same directions as, (z_k, z_{k+2})

and (z_{k+2}, z_{k+4}). An illustration is provided in Figure 2.19. We now use the van der Corput sequence's definition: if $k \equiv -1$ modulo 8, $\mathrm{V}_2(k+5) = \mathrm{V}_2(k) + 1/8$. There are now two possibilities, depending on whether or not z_k, z_{k+4}, z_{k+5} and z_{k+6} are aligned. We will treat the aligned case in detail, the other works in a very similar way. Let D be that line. By Proposition 8, $|z_{k+6}| \leq |z_{k+4}|$, with equality iff the two points are equal.

We distinguish three cases. If $z_{k+6} = z_{k+4}$, then the origin is in an open band of width 1 orthogonal to (z_{k+4}, z_k) where one of the two side-edges passes through z_{k+4}. We also had previously that the origin is in the cone C. The point in this convex set that minimizes $|z_k| - |z_{k+6}|$ is on $[z_k, z_{k+6}]$, and by construction is at distance 1 of z_{k+6}. We also know that $|z_{k+6} - z_k| = \sqrt{8}$. For this point we have $|z_k| - |z_{k+6}| = 2(\sqrt{2} - 1)$.

In the second case, z_{k+6} is on $(z_{k+4}, z_k]$ but not between z_k and z_{k+4}. The acceptable region for the origin is the intersection of C and the halfplane defined by the bisector of $[z_{k+6}, z_{k+4}]$. Since $|z_{k+4} - z_k| = \sqrt{8}$ and z_k is to the right of z_{k+4} on D, we have that $|z_k| - |z_{k+6}| \geq 1$. In the final case, z_{k+6} is on D and between z_k and z_{k+4}. The possible positions for the origin are all in the triangle formed by the top of the cone C, cut off at the halfplane orthogonal to D and at distance 1 to z_{k+4}. We necessarily have $|z_k| < 2$. We now need to find the smallest $|z_k| - |z_{k+6}|$ possible. If the origin is on one of the outer edges of the cone,

$$|z_k| = \sqrt{2 + 2\sqrt{2}a + 2a^2}$$

and

$$|z_{k+6}| = \sqrt{6 - 4\sqrt{2} + 2(2 - \sqrt{2})a + 2a^2},$$

for $a \in [0, \sqrt{2} - 1)$. The minimal $|z_k| - |z_{k+6}|$ is obtained for $a = \sqrt{2} - 1$ as the derivative is always negative on $[0, 2 - \sqrt{2}]$. This corresponds to $|z_{k+6}| < 1 + (\sqrt{2} - 1)^2 < \sqrt{2}$. Since this third case can only happen if $|z_k| < 2$, we get the desired result in this case too. Grouping all three cases together, we obtain that $|z_k| - |z_{k+6}| > 2(\sqrt{2} - 1)$ if $|z_k| \geq 2$ when $k \equiv -1$ modulo 8, and $|z_{k+6}| < 1 + (\sqrt{2} - 1)^2$ if $|z_k| < 2$. Proposition 8 gives the desired result in the aligned case. When they are not aligned, if $z_{k+4} = z_{k+6}$, then z_k, z_{k+4}, z_{k+7} and z_{k+8} are aligned and we use the reasoning above. If $z_{k+4} \neq z_{k+6}$, the origin has to be in cones pointing away from the origin that are much closer to the line containing z_{k+4} and z_{k+6}. We omit the calculation details here, they represent a better situation than that proved in the next lemma. $\square$

Lemma 12. *Let $k \equiv -1$ modulo 8. If z_k exists, $z_{k+2} \neq z_k$, and $z_{k+4} = z_{k+2}$, then:*

1. *If $|z_k| < 2$ then $|z_{k+4}| < \sqrt{2}$.*
2. *If $2 \leq |z_k| \leq \sqrt{5}$ then $|z_{k+4}| < \sqrt{8(1-\sqrt{3}/2)}$.*
3. *If $|z_k| > \sqrt{5}$ then $|z_k| > |z_{k+4}| + \sqrt{5} - 1$.*

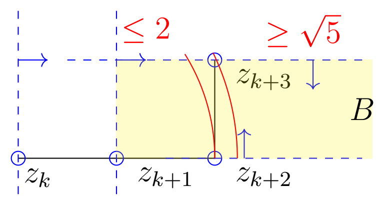


Figure 2.20: Without loss of generality, one of two acceptable regions for the origin after 4 steps. The other is its symmetric relative to (z_k, z_{k+2}). The origin has to be in B, and depending on $|z_k|$, it is in one of the three regions defined by the red arcs.

Proof. Let $k \equiv -1$ modulo 8. We suppose that we have $z_{k+2} \neq z_k$ and $z_{k+2} = z_{k+4}$. Since $k \equiv 1$ modulo 8, the third and fourth steps are on a line orthogonal to the one the first two steps are on, and they intersect in z_{k+2}. Since we are back in z_{k+2} after the fourth step, this implies that the origin is in a band B of width 1, and one of its sides is defined by (z_k, z_{k+2}). Firstly, if $|z_k| < 2$, and $z_{k+2} \neq z_k$, then $|z_k| > 1$. We also have that the origin is in a square of side 1, S_1, and one of the corners of the square is z_{k+2}, see Figure 2.20 (whether we are to the left or right of (z_k, z_{k+2}) corresponds exactly to the side z_{k+3} is on). As such $|z_{k+4}| < \sqrt{2}$. If $2 < |z_k| < \sqrt{5}$, then the origin has to be within $(B(z_k, \sqrt{5}) \backslash B(z_k, 2)) \cap B$. The point furthest away from z_{k+2} in that case is on the outer edge of S_1. We clearly have $|z_{k+4}| < \sqrt{2}$, but this can be further improved by considering this point when $|z_k| = 2$. In this case, we have

$$|z_{k+4}| = \sqrt{8(1 - \sqrt{3}/2)}.$$

The intersection point between the circle $C(z_k, |z_k|)$ and B will only be closer to z_{k+4} for higher norms of z_k.

Finally, if $|z_k| > \sqrt{5}$, the origin will still be in B but all we can guarantee is that, when splitting the band in 2 with the orthogonal line to B passing through z_{k+4}, it is not in the same part as z_k. For a given distance $|z_k|$, the furthest point from z_{k+4} is once again the intersection of the circle $C(z_k, |z_k|)$ and one of the outer boundary of B. In this case, we obtain

$$|z_{k+4}| = \sqrt{4 + |z|^2 - 4\sqrt{1 + |z|^2}}.$$

The value $|z_k| - |z_{k+4}|$ is strictly increasing in $|z_k|$, and is minimized for $|z_k| = \sqrt{5}$. This corresponds exactly to $|z_{k+4}| = 1$ (and z_{k+3} is equal to the origin, as such we require $|z_k| > \sqrt{5}$). This gives us the desired result, and concludes the proof. $\square$

2.5.2 Proof of Theorem 6

Let $z_{-1} \in \mathbb{R}$. Firstly, we set aside the indeterminate case over which we have no control. Suppose that (z_n) is defined at least for $8|z_{-1}|/(2(\sqrt{2} - 1))$ steps. We will show that for any $k \equiv -1$ modulo 8, $|z_k| > |z_{k+8}| + 2(\sqrt{2} - 1)$. For such a k, we first consider the case $|z_k| > 2$. There are three possibilities:

- If $z_k = z_{k+2}$, then by Lemma 10, $|z_k| - |z_{k+4}| > r - \sqrt{r^2 + 4 - 4\sqrt{r^2 - 1}} > 2(\sqrt{2} - 1)$ since $r > 2$.
- If $z_k \neq z_{k+2}$ and $z_{k+2} = z_{k+4}$ then, by Lemma 12, if $|z_k| < \sqrt{5}$ then $|z_{k+6}| < \sqrt{2}$ and $|z_{k+6}| < |z_k| - (\sqrt{5} - 1)$ otherwise.
- If z_k, z_{k+2} and z_{k+4} are all distinct then Lemma 9 directly states $|z_k| > |z_{k+8}| + 2(\sqrt{2} - 1)$.

Using Proposition 8, we have that $|z_{k+4}|$ and $|z_{k+6}|$ are greater than or equal to $|z_{k+8}|$. As such, while $|z_k| > 2$, we have the guarantee that $|z_k| > |z_{k+8}| + 2(\sqrt{2}-1)$. It therefore takes at most $8(|z_{-1}| - 2)$ steps to reach the ball of radius 2. With the special cases for $|z_k| < 2$ from Lemmas 9, 10 and 12, it takes exactly one more block of 8 steps to reach $B(0, \sqrt{2})$, proving the desired result.

2.5.3 Periodicity

Proof of Theorem 7. We want to show that, for n odd and $z_n \in B(0,1)$, $z_{n+2j} = z_n$ and z_{n+2j+1} is on a half-circle centered in z_n of radius 1. This half circle is exactly that contained in the halfplane defined by the line orthogonal to $((0,0), z_n)$ and that contains the origin. We will show this result holds once z_n for n odd is in $B(0,1)$ and this will directly prove the first statement on $B(0,1)\backslash\{(0,0)\}$. We first prove that the result is true for the first two steps, i.e. $j = 0$ for the first even and odd point after z_n.

Even case. We have that $z_{n+1} = z_n \pm e^{2\pi i \,\mathrm{V}_2(n+1)}$. By definition, z_{n+1} is on the circle C_n of radius 1 centered in z_n, and is on the line D_n defined by $z_n + xe^{2\pi i \,\mathrm{V}_2(n+1)}$. There are exactly two such points since D_n contains the center of C_n. In the first case, if D_n is orthogonal to $((0,0), z_n)$, both points have same norm. The sequence is indeterminate and stops. If not, one of the two elements has norm strictly smaller than the other and is selected. Since any D_n such that the sequence is properly defined cannot be equal to $((0,0), z_n)^{\perp}$, the element with smaller norm is in the desired halfplane.

Odd case. By definition, $z_{n+2} = z_{n+1} \pm e^{2\pi i \,\mathrm{V}_2(n+2)}$. Since n is even, $\mathrm{V}_2(n+2) = 1/2 + \mathrm{V}_2(n+1)$. This gives us $z_{n+2} = z_{n+1} \pm e^{2\pi i(\mathrm{V}_2(n+1)+1/2)} = z_{n+1} \pm e^{2\pi i \,\mathrm{V}_2(n+1)}$. z_{n+2} is therefore on a circle of radius 1 centered on z_{n+1} and on D_n since z_{n+1} is on it. There are two such points separated by a distance of 2, one of which is z_n. Since only one of the two can have norm smaller than 1, it will necessarily be selected as z_{n+2}. By hypothesis, $|z_n| < 1$, therefore $z_{n+2} = z_n$.

By induction, this can be generalized to any $j \in \mathbb{N}$. $\square$

This theorem proves two different elements: firstly any $z_{-1} \in B(0,1)$ is a periodic start position in base 2. Secondly, once we have an odd-numbered $z_k \in B(0,1)$ then it will act as a periodic start position for the rest of the sequence. Gaps in the semi-circle then correspond exactly to previously taken directions since no two points of the van der Corput sequence are equal.

2.5.4 The uncertain regime: 1 *to* $\sqrt{2}$

Theorem 7 shows that once an odd point is in $B(0,1)$, we have a perfect understanding of the future points. Theorem 6 also allows us to understand the behavior of the sequence as long as $|z_n| > \sqrt{2}$. However, apart from the limited information given by Lemma 9, none of the results presented so far give information on what can happen between 1 and $\sqrt{2}$. Studying bigger blocks of consecutive steps could improve the different lemmas shown previously. However, the following proposition shows that it is impossible to guarantee that we will reach $B(0,1)$ for any finite number of steps n.

Proposition 9. *For any* $n \in \mathbb{N}$*, there exist convex sets of starting positions in the set* $B(0,\alpha_n)\backslash B(0,1)$ *such that* $z_{-1} = z_{2n-1}$*, where* $\alpha_n > 1$ *is some constant depending on* n *such that* $\lim_{n\to\infty} \alpha_n = 1$*.*

Proof. Let $n \in \mathbb{N}$ and let $k = \lfloor \log_2(n) \rfloor$. When considering the first 2^{k+1} elements of the sequence, the smallest difference between any two elements is 2^{-k-1}. Given a point $z_k \notin B(0,\sqrt{2})$, we have $z_k \neq z_{k+2}$ if and only if $C(z_k,2)$, the circle centered in z_k of radius 2, has a non-empty intersection with $B(0,|z_k|)$. When $|z_k| = 1$, this intersection is a single point. As $|z_k|$ increases, the circle arc forming $C(z_k,2) \cap B(0,|z_k|)$ grows. We now need to find positions for which this circle arc corresponds to an angle of less than $2\pi 2^{-k-1}$ in z_k. Without loss of generality, we suppose that z_k is on the x-axis (starting there is obviously incorrect as it would improve in the first two steps). The edge of the circle arc, z_k and the origin form a triangle with known distances. Using the law of cosines, we have that

$\alpha/2 = \arccos((3+r^2)/(4r))$ and we want $\alpha \le 2\pi 2^{-k-1}$. In the case of equality, this gives us

$$3 + r^2 - 4\cos(\pi 2^{-k-1})r = 0.$$

This has two real roots for $k > 1$, both of which are above 1. The smallest of the two is the desired solution. We have

$$\alpha_n = \frac{4\cos(\pi 2^{-k-1}) - \sqrt{16\cos^2(\pi 2^{-k-1}) - 12}}{2}.$$

Any point on $[O, z_k]$ within that distance of the origin will be such that $z_{-1} = z_{2i-1}$ for $i \le n$. With a small rotation around the origin to guarantee that none of the first n pairs of steps will improve on $|z_{-1}|$, this proves the desired result. □

The precise regions are larger than simply a line, and their number depends on $\log(n)$. We leave their computation to an interested reader. This result suffices to show that there cannot be a guaranteed improvement for the norm of $|z_k|$ within a fixed number of steps, regardless of that number. This concludes our analysis of the base 2 case.

2.6 Higher bases

2.6.1 General behavior: a discussion

Before proving Theorem 8, we give some quick elements to explain the behavior of the sequence in higher dimensions. There are two main differences with base 2. The first is that we no longer have an equivalent to Proposition 8: we cannot guarantee that every j steps we have a strictly decreasing norm, for some well-chosen value of j. With a similar case decomposition as in Lemmas 9, 10 and 12, one can show that in base 3 we can lower-bound $|z_k| - |z_{k+3}|$ by a constant for $k \equiv -1$ modulo 3 as long as $|z_k| > \sqrt{1.5} + c$ where c is a small constant. The $\sqrt{1.5}$, improving on the $\sqrt{2}$ in base 2, corresponds to the maximum distance between the origin and z_k, while being equidistant to z_{k+3}. This improvement is a side effect of the second difference: we have access to more information on consecutive steps, and we do not have consecutive colinear steps. To obtain $\sqrt{1.5}$ as the threshold where we can no longer guarantee improvement, it suffices to consider all the possible choices in 3 consecutive steps, whereas 2^2 were necessary in the base 2 case to get $\sqrt{2}$ (and even 2^3 to obtain precise values). As such, z_k and z_{k+3} cannot be too far away from each other, and in particular the origin has to be closer to z_{k+3} if it is outside of a small ball centered on z_k. For the four different cases to consider here, the maximum distance $|z_k|$ such that z_k and z_{k+3} are equidistant to the origin is respectively 1, $2/\sqrt{3}$ and $\sqrt{1.5}$. The size of the ball should be seen as a balance between the angle of consecutive steps and the information their greater number provides. As b increases, Theorem 8 shows that the minimal size of the ball we are guaranteed to reach also increases. We do not present calculations for higher bases here, as they provide relatively little information while requiring a case distinction of 2^{b-1} cases.

2.6.2 The polygonal structure

Base 2 presents a specific behavior as the regular polygon defined by 2 points, if one accepts it as such, is simply a line. For higher bases, we will show that the repeating pattern formed by b consecutive steps forms a regular b-sided polygon, which we will abbreviate to b-gon. We begin by giving a characterization of all periodic starts for $b > 2$. Our results will be shown with the hypothesis that the steps are associated with "-" signs. By central symmetry of the sequential construction, any valid z_{-1} obtained for a periodic start with "-" will be associated to $-z_{-1}$, a valid periodic start for "+".

Proof of Theorem 9. Without loss of generality, suppose that the periodic sequence z_n is associated with $-$ signs. Let $n \equiv j$ modulo b. Firstly, if $j = -1$, then $z_n = z_{-1}$ by definition of a periodic start and they are indeed on the circle of radius 0. We now suppose $j \neq -1$, and let b_0 be the closest multiple of b smaller than n. Using the periodicity of the sequence, we have

$$
\begin{aligned}
z_{n,x} &= z_{n-1,x} - \cos(2\pi \underset{b}{\mathrm{V}}(n)) \\
&= z_{b_0-1,x} - \sum_{i=0}^{j} \cos(2\pi \underset{b}{\mathrm{V}}(b_0+i)) \\
&= z_{-1,x} - \sum_{i=0}^{j} \cos(2\pi \underset{b}{\mathrm{V}}(b_0) + 2i\pi/b) \\
&= z_{-1,x} - \frac{\sin(j\pi/b)}{\sin(\pi/b)} \cos(2\pi \underset{b}{\mathrm{V}}(b_0) + j\pi/b).
\end{aligned}
$$

Similarly, we can obtain for the y-coordinate

$$
z_{n,y} = z_{-1,y} - \frac{\sin(j\pi/b)}{\sin(\pi/b)} \sin(2\pi \underset{b}{\mathrm{V}}(b_0) + j\pi/b).
$$

This gives us

$$
\begin{aligned}
|z_n - z_{-1}| &= \sqrt{(z_{n,x} - z_{-1,x})^2 + (z_{n,y} - z_{-1,y})^2} \\
&= \frac{\sin(j\pi/b)}{\sin(\pi/b)} \sqrt{\cos(2\pi V_b(b_0) + j\pi/b)^2 + \sin(2\pi V_b(b_0) + j\pi/b)^2} \\
&= \frac{\sin(j\pi/b)}{\sin(\pi/b)}.
\end{aligned}
$$

We obtain the first part of the desired result. Furthermore, for $k \equiv -1$ modulo b and $0 \leq j \leq b-1$, $\mathrm{V}_b(k+j) = \mathrm{V}_b(j) + \mathrm{V}_b(k)$ and, for $j > 0$, $\mathrm{V}_b(k+j+1) - \mathrm{V}_b(k+j) = 1/b$. Since we are supposing that all signs are picked to be "-", the b steps associated to $\mathrm{V}_b(k), \ldots, \mathrm{V}_b(k+b)$ are exactly the same as $\mathrm{V}_b(0), \ldots, \mathrm{V}_b(b)$ with an extra $\mathrm{V}_b(k)$ rotation. This rotation is the same for all of the steps, therefore both sequences of steps describe the same polygonal shape. The second relation imposes that the angle between two consecutive steps is exactly $2\pi/b$ for each of the $b-1$ steps after the first one. We therefore have that $z_k = z_{k+b}$, that each side length is 1, and that each other angle of the b-sided polygon is $2\pi/b$. Each sequence of b steps starting in z_k for $k \equiv -1 \bmod b$ must be a b-gon. Finally, the angle between (z_{-1}, z_0) and (z_k, z_{k+1}) is $\mathrm{V}_b(k)$, this concludes the proof. $\square$

Each b-gon corresponding to b consecutive steps will be a rotated copy of the b-gon formed by the points $(z_{-1}, \ldots, z_{b-1})$, with a fixed point in z_{-1}. For each copy, the angle of that rotation is given by the first element of the van der Corput sequence in that block of b elements: the $\mathrm{V}_b(bj-1)$ for all j determine the rotations. Since these values are all in $[0, 1/b)$, the maximum rotation possible is of angle $2\pi/b$. Figure 2.21 shows these arcs and the circles they lie on. For a rotation of angle α, we will write P_α for the associated polygon. For such an infinite sequence to be valid with our construction methods, it requires that the norm associated with $z_n - e^{2\pi i V_b(n+1)}$ is smaller than that of $z_n + e^{2\pi i V_b(n+1)}$ for all n (the role of the signs could be reversed, we study this case without loss of generality). This means that the scalar product of $(z_n, (0,0))$ and (z_n, z_{n+1}) is strictly positive. At each point in the sequence,

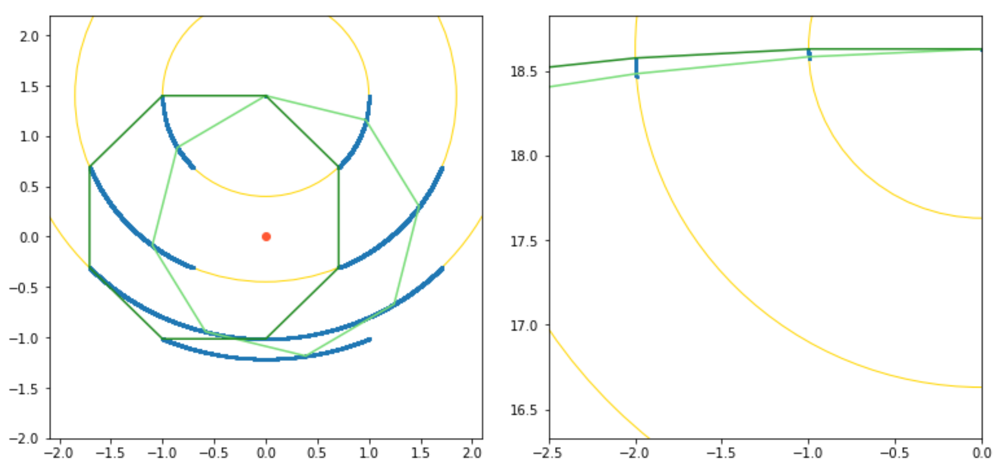


Figure 2.21: On the left, the first 200 000 points of the base 8 Van der Corput sequence with periodic start, with supporting circles. The first and 102^{nd} polygons are drawn. The arcs correspond exactly to those of Figure 2.11. On the right, the zoomed in points in first, second and third position of the 117-gon for the base 117 van der Corput sequence, again for 200 000 points.

$(0,0)$ must be in the open halfplane containing all the x such that the scalar product of (z_n, x) and (z_n, z_{n+1}) is strictly positive to keep the polygonal structure. In other words, we are considering the halfplane defined by the orthogonal line to the next side of the b-gon that passes through z_n. For each b-gon, we are interested in the intersection of these halfplanes, that forms another b-gon for any regular b-gon (smaller for $b \geq 5$, equal for $b = 4$ and greater for $b = 3$). The origin must be inside this smaller b-gon, for all possible rotations. For a polygon P, we will denote P_i its inner polygon obtained this way.

Lemma 13. *The intersection of $P_{0,i}$ and $P_{2\pi/b,i}$ is a triangle.*

Proof. We will show that the intersection of $P_{0,i}$ and $P_{2\pi/b,i}$ is exactly a triangle formed by the leftmost submit of $P_{2\pi/b,i}$ cutoff by a vertical edge. We will call p_{left} this summit. Since this structure is invariant by translation, we will consider that $z_{-1,x} = 0$. By construction, we know that the rightmost side of $P_{0,i}$ lies on $x = 0$. We also have that $P_{2\pi/b,i}$ is a copy of $P_{0,i}$ translated by 1 to the right. Finally, we also know that the inner polygon's axis of symmetry is horizontal as the rightmost side is vertical. As such, it suffices to show that exactly one summit of $P_{2\pi/b,i}$ has its x-coordinate strictly negative and that its adjacent edges both intersect the vertical edge of $P_{0,i}$. The center c_i of $P_{0,i}$ is on the $y = 0$ line by symmetry. We can compute its distance from the right vertical edge, it is given by

$$-c_{i,x} = \frac{\sec(\pi/2 - 2\pi/b) - \tan(\pi/2 - 2\pi/b)}{2\tan(\pi/b)}.$$

This is equal to

$$-c_{i,x} = \frac{(1 - \cos(2\pi/b))\cos(\pi/b)}{2\sin(\pi/b)\sin(2\pi/b)}.$$

Expressing all the terms as functions of $\sin(\pi/b)$ and $\cos(\pi/b)$ gives us $c_{i,x} = -1/2$. Translated by 1 to the right, the center of the circumscribed circle of $P_{2\pi/b,i}$ is in $1/2$, as such the x-coordinate of p_{left} is

$$\frac{1}{2} - \frac{\sec(\pi/2 - 2\pi/b) - \tan(\pi/2 - 2\pi/b)}{2\sin(\pi/b)}.$$

This can be reformulated to

$$\frac{1}{2} - \frac{\sin(\pi/b)}{\sin(2\pi/b)} = \frac{1}{2} - \frac{1}{2\cos(\pi/b)}.$$

This is strictly negative for all integers $b \geq 5$. The x-coordinate of the adjacent summits is given by

$$\frac{1}{2} - \frac{\cos(2\pi/b)}{2\cos(\pi/b)} = \frac{1}{2} - \frac{\cos(\pi/b)}{2} + \frac{\sin^2(\pi/b)}{2\cos(\pi/b)},$$

which is positive for $b \geq 5$.

Since $P_{2\pi/b,i}$ is a convex b-gon, we have that $P_{2\pi/b,i}$ has a single summit with negative x-coordinate. Since this summit is aligned with the centers of the circumscribed circles of $P_{0,i}$ and $P_{2\pi/b,i}$ and is a translation of the leftmost summit of $P_{0,i}$, p_{left} is in $P_{0,i}$. It remains to show that its adjacent edges intersect the vertical edge of $P_{0,i}$. Without loss of generality, we suppose that $z_{-1,y}$ is such that the axis of symmetry of $P_{0,i}$ is on $y = 0$, i.e. $y = 0$ bisects the vertical edge of $P_{0,i}$. The upper summit u of this vertical edge has for y-coordinate $u_y = (\sec(\pi/2 - 2\pi/b) - \tan(\pi/2 - 2\pi/b))/2$, half the side length of $P_{0,i}$. Similarly, the lower vertex has for y-coordinate the opposite of that value. We know the angles of $P_{2\pi/b,i}$, as such the intersection of the top edge adjacent p_{left} has y-coordinate equal to $-p_{left,x}\tan(\pi/2-\pi/b)$. This is equal to

$$\left(\frac{1}{2} - \frac{1}{2\cos(\pi/b)}\right)\frac{\cos(\pi/b)}{\sin(\pi/b)}.$$

We want this to be less than u_y. After reformulating, we want

$$\frac{\cos(\pi/b) - 1}{2\sin(\pi/b)} + \frac{\cos(2\pi/b) - 1}{\sin(2\pi/b)} < 0.$$

This is clearly true for $b \geq 5$ since both cosines are strictly smaller than 1 and the sines are positive. The upper edge adjacent to p_{left} therefore intersects the vertical edge of $P_{0,i}$. By symmetry, we obtain that the lower edge from p_{left} also intersects this vertical edge. We therefore have that $P_{0,i} \cap P_{2\pi/b}$ is a triangular region with p_{left} as one of the summits, and the two other summits are determined above. $\square$

Proposition 10. *The intersection of all rotated inner b-gons $P_{\alpha,i}$ for $\alpha \in [0, 2\pi/b)$ is exactly the intersection of $P_{0,i}$ and $P_{2\pi/b,i}$.*

We note that while there is no polygon with exactly a rotation of angle $2\pi/b$, there is a sequence of polygons whose limit is this rotated polygon. Since we consider only the interior of each polygon as the acceptable region for the origin, this has no impact on our result.

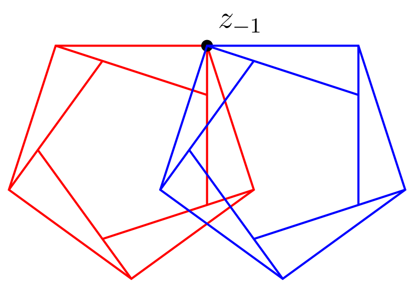


Figure 2.22: For $b = 5$, $P_{0,i}$ in red and $P_{2\pi/b}$ in blue, with their associated inner pentagons. The leftmost vertex of the inner blue pentagon is p_{left}.

Proof. We consider $P_{0,i}$ and $P_{2\pi/b,i}$ the inner polygons of P_0 and $P_{2\pi/b}$. We first note that $P_{2\pi/b}$ is exactly a copy of P_0, translated to the right by 1. We separate in two cases depending on the parity of b. Firstly, we consider an even b. In this case, there are exactly two horizontal sides of P_0: $[z_{-1}, z_0]$ and $[z_{b/2-1}, z_{b/2}]$. We also have $z_{-1,x} = z_{b/2,x}$ and $z_{0,x} = z_{b/2-1,x}$. The normals associated with z_{-1} and $z_{b/2-1}$ are vertical and the intersection of the desired halfplanes forms a vertical band $B_1 = \{(x, y) : z_{b/2-1,x} < x < z_{-1,x}\}$. $P_{2\pi/b}$ is the symmetric of P_0 with respect to the axis $x = z_{-1,x}$. It also has two horizontal segments, $[z'_0, z_{-1}]$ and $[z_{b/2}, z'_{b/2-1}]$, where z'_j is the symmetric of z_j with respect to the axis given above. For $P_{2\pi/b}$, its intersection polygon is contained in the band $B_2 = \{(x, y) : z_{-1,x} < x < z'_{0,x}\}$. We obtain $P_{0,i} \cap P_{2\pi/b,i} = \emptyset$. This implies that the intersection of all the rotated polygons is also empty.

We now turn to the odd b case, where the intersection is not empty. We need to show that for any $\alpha \in [0, 2\pi/b)$, $(P_{0,i} \cap P_{2\pi/b,i}) \subseteq P_{\alpha,i}$. For all of these polygons, the translation of the global polygon naturally leads to a translation of this center for $P_{2\pi/b}$. The length of each side of an inner polygon is given by $\sec(\frac{\pi}{2} - \frac{2\pi}{b}) - \tan(\frac{\pi}{2} - \frac{2\pi}{b})$. As illustrated in Figure 2.23 and shown in Lemma 13, the intersection between the two translated inner polygons $P_{0,i}$ and

$P_{2\pi/b,i}$ is non-empty.

Let c be a point in $P_{0,i} \cap P_{2\pi/b,i}$. Let A be the arc of angle $-2\pi/b$ centered in z_{-1} and with rightmost extremity in c. Since $c \in P_{2\pi/b,i}$, the leftmost extremity of that arc, d, is also in $P_{0,i}$. Observe that c and d can be written

$$c = \left(x_{-1} + r\cos(\theta + \frac{2\pi}{b}), y_{-1} + r\sin(\theta + \frac{2\pi}{b})\right)$$

and

$$d = (x_{-1} + r\cos(\theta), y_{-1} + r\sin(\theta)).$$

Where $r := |c - z_{-1}|$. Let $0 < \beta \leq \frac{2\pi}{b}$ be the angle of rotation of P_β with respect to z_{-1} (with the extremal case $\beta = \frac{2\pi}{b}$ corresponding to $P_\beta = P_{2\pi/b}$). Consider

$$c_\beta = \left(x_{-1} + r\cos(\theta + \frac{2\pi}{b} + \beta), y_{-1} + r\sin(\theta + \frac{2\pi}{b} + \beta)\right)$$

and

$$d_\beta = (x_{-1} + r\cos(\theta + \beta), y_{-1} + r\sin(\theta + \beta)).$$

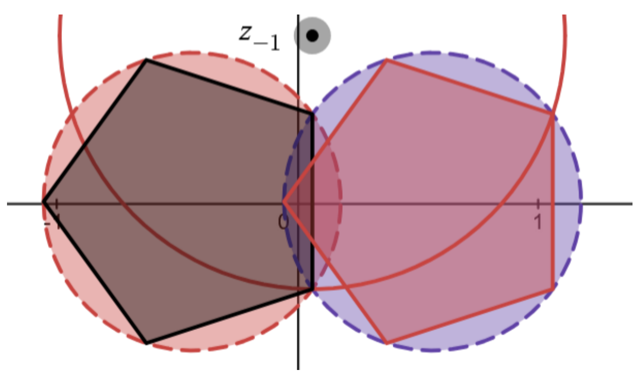


Figure 2.23: The intersection of the two inner polygons in the case of $b = 5$.

We have that c lies on the arc between d_β and c_β about z_{-1} (with $c = d_\beta$ for $\beta = \frac{2\pi}{b}$). If we suppose that the arc A lies within $P_{0,i}$, then the d_β to c_β arc lies within $P_{\beta,i}$. Since c lies within the second arc, we have that $c \in P_{\beta,i}$. It remains to show that the arc A is in

$P_{0,i}$. We will do this in two steps, first showing that the outermost arc does not intersect the outer edges apart from at its endpoints. This furthest arc $A_{\max}$ corresponds to the lowest intersection point between the circumscribed circle of $P_{0,i}$ and $P_{2\pi/b,i}$, and is at distance $K = \sec(\pi/2 - 2\pi/b)$ from z_{-1}. We now show that the intersection of $A_{\max}$ and $P_{0,i}$ forms a single arc.

As illustrated in Figure 2.24, let s_1 be the segment between the two intersection points of $A_{\max}$ and the circumscribed circle of $P_{0,i}$. Note that it is orthogonal to the left edge of the inner polygon it intersects. Let s_2 be the parallel segment passing through the lower summit of the side of $P_{0,i}$ that $A_{\max}$ intersects. The maximum distance between s_1 and $A_{\max}$ is reached in the middle of that segment, and is strictly smaller than

$$\frac{\sec(\pi/2 - 2\pi/b) - \tan(\pi/2 - 2\pi/b)}{2}.$$

The distance between s_1 and s_2 is equal to

$$\frac{\sec(\pi/2 - 2\pi/b) - \tan(\pi/2 - 2\pi/b)}{2}.$$

$A_{\max}$ is therefore always contained between s_1 and s_2. There are therefore only two sides of $P_{0,i}$ that can intersect $A_{\max}$, exactly those intersected by s_1 and s_2. We need to verify that each can be intersected only once by the arc. Both of these sides have one point inside the circle of radius K centered in z_{-1} and one outside. As such, there can be only one intersection point with each side. The intersection of $A_{\max}$ and $P_{0,i}$ therefore forms a single arc and since both extremal points of $A_{\max}$, d and c, are in $P_{0,i}$, we conclude that $A_{\max} \subseteq P_{0,i}$. For any other arc A_α, since c_α and d_α are in $P_{0,i}$ and s_1 is orthogonal to the left side of $P_{0,i}$, its orthogonal projection onto the line on which lies s_1 is in s_1. A_α is an arc between d_α and c_α for a circle centered in z_{-1}, as such it must lie on the same side of (c_α, d_α) as $A_{\max}$. This implies that A_α is in the convex region defined by the arc $A_{\max}$, c_α, d_α and the endpoints of s_1. All of these elements are in $P_{0,i}$, which is itself a convex region. We therefore have that $A_\alpha \subset P_{0,i}$. This concludes the proof. $\square$

2.6.3 *Proof of Theorem 8*

Proof. We first treat the much easier even b case. As Proposition 10 shows, the intersection of the acceptable bands is empty, $B_1 \cap B_2 = \emptyset$. However, we must have $(0,0) \in B_1$ and $(0,0) \in B_2$, we have a contradiction. It is therefore impossible for both these b-gons to appear. We do not have exactly $P_{2\pi/b}$ in our set of rotated polygons, but only a sequence of polygons converging to $P_{2\pi/b}$.

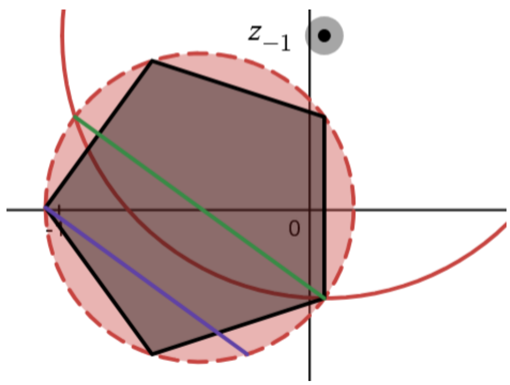


Figure 2.24: Base 5: in green s_1 and in blue s_2, the two segments that we use to bound the red arc in $P_{0,i}$.

For any p in $P_{0,i}$, there will exist $\delta > 0$, such that the polygon $P_{2\pi/b-\delta}$ does not contain p in its inner intersection polygon. As such there is no valid position for $(0,0)$ to obtain a periodic start. We now consider b odd. Using Lemma 13, we know that the intersection of $P_{0,i}$ and $P_{2\pi/b,i}$ is a triangle which we characterized in the proof of Lemma 13. By Proposition 10, we know that the intersection of all the rotated polygons is exactly $P_{0,i} \cap P_{2\pi/b,i}$. Given that the (open) inner polygons correspond to the acceptable regions for the origin, z_{-1} must be such that the origin is contained in the interior of $P_{0,i} \cap P_{2\pi/b,i}$. This corresponds to a first triangular region T_1 with $z_{-1} > 0$ coordinate-wise. Our construction corresponded to a first step to the left. In other words, all the sign choices during the construction of the sequence were "-". As mentioned previously, any sequence starting in z_{-1} with only "-" choices is the symmetric with respect to the origin of a sequence starting in $-z_{-1}$ with only "+" signs. A second triangular region $T_2 = \{(-x, -y) : (x, y) \in T_1\}$ is exactly all the valid periodic starting positions for "+" signs. $\square$

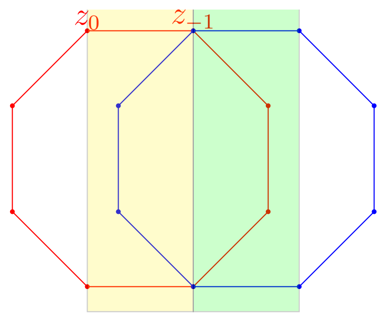


Figure 2.25: In red P_a and in blue P_b. In respectively yellow and green are B_1 and B_2, which do not intersect as they are open.

An example of these regions for $b = 5$ is provided in Figure 2.26. The only base not covered by Theorems 7 and 8 is base 3. The big difference with other higher bases is that the intersection of the halfplanes forms a triangle which contains the initial triangle, rather than contained in. A similar approach could be taken, but Proposition 10 is no longer verified. As such, the intersection of the larger valid triangles would need to be done over the rotation of all triangles. The edge of the valid region would then correspond to the union of three identical curves. One of these corresponds to the convex hull of the rotation of angle $2\pi/3$ around z_{-1} of the line passing through z_0 such that the angle between (z_0, z_{-1}) and this line (considered with positive slope) is $\pi/6$. A further case distinction would then need to be done depending on the sign of $z_{-1,x}$ and $z_{-1,y}$. We do not provide a formal proof of the region of periodic starts, only an illustration in Figure 2.26.

2.6.4 A limit of the characterization

Finally, even though there are no infinite periodic sequences for even b, it is possible to find periodic sequences of length greater than n, for any $n \in \mathbb{N}$. Proposition 11 characterizes the length of the sequence $(z_n)_{n\in\mathbb{N}}$ given a starting position close to the best possible choice.

Proposition 11. *Let b be a positive even integer. For $\varepsilon > 0$, if $z_{-1} = (\varepsilon, \frac{M}{2})$, where M is the height of a regular b-gon, then the b-gon rotation stops at the k-th cycle, where k is the first integer that satisfies $2\pi\,\mathrm{V}_b(kb) = \arctan(M/(2\varepsilon)) - \pi/2 + 2\pi/b$.*

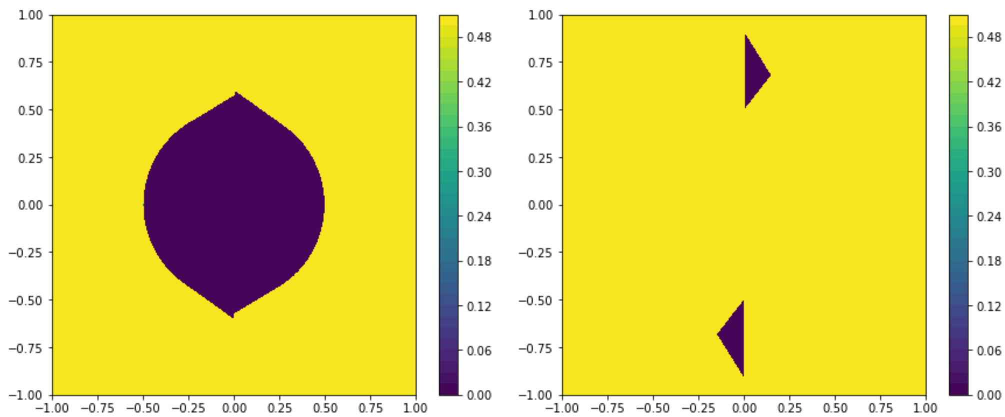


Figure 2.26: Left: base 3, right: base 5. In purple, regions of $[-1,1]^2$ where the starting position leads to a distance of less than 10^{-4} between z_{-1} and z_{149}. Other values set to 1/2 for simplicity.

Before we prove this proposition, we first verify that this inequality indeed has a solution. In our case, since M and ε are positive, we have $\arctan(\frac{M}{2\varepsilon}) \in (0, \frac{\pi}{2})$. This gives us

$$\arctan(M/(2\varepsilon)) - \pi/2 + 2\pi/b < 2\pi/b.$$

Since $\{\mathrm{V}_b(kb) : k \in \mathbb{N}\}$ is uniformly distributed in $[0, 1/b)$, there exists a k that satisfies the inequality. Figure 2.27 provides an illustration of this behavior. Now let us prove the proposition.

Proof of Proposition 11. In an even b-gon, the halfplane orthogonal to (z_i, z_{i+1}) passing through z_i also passes through $z_{i+b/2}$ for any i, where numbering is done modulo b. As such the halfplane associated with the $b/2+1$-th step passing through $z_{b/2-1}$ also passes through z_{-1}. When considering a maximal rotation $2\pi/b$, this halfplane is the one that will not contain $(0,0)$, as shown in the proof of Theorem 8. We therefore need to find the smallest angle defined by $\mathrm{V}_b(kb)$ for some $k \in \mathbb{N}$ such that the line $(z_{b/2-1}, z_{-1})$ crosses the $y = 0$ axis below 0. The intersection between this line and $y = 0$ is a decreasing function for $\theta \in [0, 2\pi/b)$. For the k-th polygon, the angle between $(z_{-1}, z_{b/2-1})$ and the horizontal line in z_{-1} is $\theta + \pi/2 - 2\pi/b$, where $\theta = 2\pi\,\mathrm{V}_b(kb)$. From this, $(z_{-1}, z_{b/2-1})$ intersects $x = 0$ in $(0,0)$ exactly when $\tan(\theta + \pi/2 - 2\pi/b) = M/(2\varepsilon)$. This corresponds to $2\pi\,\mathrm{V}_b(kb) = \arctan(M/(2\varepsilon)) - \pi/2 + 2\pi/b$. For any angle greater than this one, the origin will be not be in the acceptable halfplane obtained in the $(b/2-1)$-th summit of the rotated polygon, and the periodicity will end. $\square$

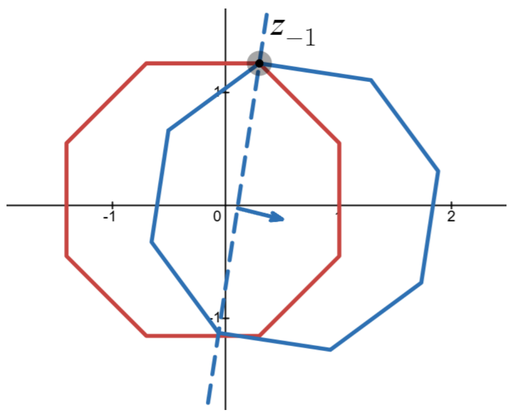


Figure 2.27: Example where the periodic rotations stop in base 8.

Once again, by central symmetry, an identical result can be obtained for $z_{-1} = (-\varepsilon, -M/2)$. There is an algorithmic way to find the smallest kb such that the inequality above holds. If we write kb in base b, then it is of the form $kb = d_L d_{L-1} \cdots d_1 0$, where the last digit is 0. Then $V_b(kb)$ is equal to

$$V_b(kb) = \frac{0}{b} + \frac{d_1}{b^2} + \cdots + \frac{d_{L-1}}{b^L} + \frac{d_L}{b^{L+1}}$$

Finding the first $d_i < b-1$ then gives us the first polygon which will not be completed. There will be at most, in b-ary notation, $(d_i + 1)d_{i-1} \ldots, d_1 0$ z_j following the polygonal structure.

Chapter 3

BAD SCIENCE MATRICES

3.1 Introduction

The bad science matrix problem [81] is concerned with understanding, for matrices $A \in \mathbb{R}^{n \times n}$ whose rows are normalized to have ℓ^2-length 1, the size of the quantity

$$\beta(A) = \frac{1}{2^n} \sum_{x \in \{1,1\}^n} \|Ax\|_\infty .$$

While originally only defined for square matrices in [81], its definition naturally makes sense for $A \in \mathbb{R}^{m \times n}$ where one measures the $\infty-$norm of the 2^n vertices of the cube which are mapped to 2^n vertices in $\mathbb{R}^m$. The problem is interesting for a variety of different reasons. We list four.

3.1.1 Bad Science.

The name is derived from the following thought experiment: we are given a sequence of coin tosses $x \in \{-1, 1\}^n$ and are trying to understand whether they come from a fair coin or not. One natural way of testing is to fix a vector $a_1 \in \mathbb{R}^n$, normalized to $\|a_1\|_2 = 1$, and consider the size of the inner product $\langle a_1, x \rangle$: Hoeffding's inequality guarantees that if, the coin is fair, then the inner product should be small. Similar reasoning suggests that we could take two vectors $\|a_1\|_2 = 1 = \|a_2\|_2$ and consider both inner products: they should both be small and if one of them is large, that is probably indicative of the coin not being fair. We may also think of this case as two statistical tests being run. Note that, since we do not make any assumptions about a_1 and a_2, these tests need not be independent. The question now is whether it is possible to construct n vectors $a_1, \ldots, a_n$ corresponding to n fair statistical tests such that, for a typical sequence of random coin tosses $x \in \{-1,1\}^n$, at least one of the statistical tests will typically end up with a large result even though the coin is fair. This motivates the name ('bad science matrix problem'): the bad scientist is running a battery of (fair) tests to find an abnormal result even though the coins are not abnormal.

3.1.2 Rotating the cube.

Since the rows are normalized to have length 1, we deduce that the Frobenius norm of the matrix is

$$\|A\|_F^2 = \sum_{i,j=1}^{n} A_{ij}^2 = n.$$

The Frobenius norm is simultaneously the sum of the squares of the singular values of the matrix and therefore

$$n = \sum_{i=1}^{n} \sigma_i(A)^2.$$

Geometrically, this means that the matrix cannot be too expansive. A very simple example of matrices with this property is the identity matrix $\mathrm{Id}_{n \times n}$ or, more generally, orthogonal matrices $Q \in \mathbb{R}^{n \times n}$. Given such a matrix A, we could now ask where it ends up sending the unit cube $\{-1,1\}^n$

$$A\{-1,1\}^n = \left\{ \sum_{i=1}^{n} \varepsilon_i c_i : \varepsilon_i \in \{-1,1\} \right\},$$

where $c_1, \dots, c_n \in \mathbb{R}^n$ are the columns of the matrix A. Using $a_1, \dots, a_n \in \mathbb{R}^n$ to denote the rows of the matrix A, we then have

$$\frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \|Ax\|_2^2 = \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \sum_{i=1}^{n} \langle a_i, x \rangle^2 = \sum_{i=1}^{n} \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \langle a_i, x \rangle^2.$$

For any arbitrary vector $y \in \mathbb{R}^n$, we have

$$\begin{aligned}\frac{1}{2^n}\sum_{x\in\{-1,1\}^n}\langle y,x\rangle^2 = \mathbb{E}[\langle y,x\rangle^2] = \mathbb{E}\left[\left(\sum_{i=1}^n y_i x_i\right)^2\right] = \mathbb{E}\left[\sum_{i,j=1}^n y_i y_j x_i x_j\right] \\ = \sum_{i,j=1}^n y_i y_j \mathbb{E}[x_i x_j] = \sum_{i=1}^n y_i^2 \underbrace{\mathbb{E}[x_i^2]}_{1} + \sum_{i\neq j} y_i y_j \underbrace{\mathbb{E}[x_i x_j]}_{0} \\ = \sum_{i=1}^n y_i^2 = \|y\|^2\end{aligned}$$

and therefore

$$\frac{1}{2^n}\sum_{x\in\{1,1\}^n}\|Ax\|_2^2 = \|A\|_F^2 = n.$$

One way of interpreting the results is that the normalization on the rows shows that A rotates the vertices of the cube to points whose average ℓ^2-norm remains fixed. The question is now: can these points have the property that the typical largest entry of such a vertex ends up being significantly larger than the average?

3.1.3 The Komlós conjecture.

Another motivation is given by the Komlós conjecture. It is known (see Banaszczyk [12]) that there exists a constant $C > 0$ such that for all matrices $A \in \mathbb{R}^{n\times n}$ whose *rows* are of size ≤ 1 in ℓ^2, there always exists a sign vector $\varepsilon \in \{-1,1\}^n$ such that

$$\|A\varepsilon\|_{\ell^\infty} \leq C\sqrt{\log n}.$$

The open question is whether the $\sqrt{\log n}$ term is necessary or whether the result remains true with a universal constant C. It is known that if this were true, the constant C cannot be too small: Kunisky [56] has proven that for each $\eta > 0$ there exists a matrix $A \in \mathbb{R}^{n\times n}$ so that

$$\forall\, \varepsilon \in \{-1,1\}^n \qquad \|A\varepsilon\|_{\ell^\infty} \geq 1+\sqrt{2}-\eta.$$

The bad science matrix problem is somewhat dual: one can think of the Komlós problem as the problems of balancing the columns of a matrix to ensure that the final sum is in a small $\ell^\infty-$box centered at the origin. Our problem is less a vector balancing problem and more of a 'functional-balancing' question: we have n functionals $\langle a_i, \cdot \rangle$ whose operator norm is 1 and ask whether they can be arranged so that a typical random ± 1 is outside a large $\ell^\infty-$box. To further underscore the similarity, an idea of Kunisky [56] shown to be useful in the setting of the Komlós problem, will also be useful in the setting of bad science matrices.

3.1.4 Extremizers.

One final possible motivation why this functional may be of intrinsic interest are the results: extremal matrices appear to have an interesting structure. We illustrate this with an example taken from [81]: the currently best known example for $n = 5$ is the matrix

$$A = \frac{1}{2\sqrt{3}} \begin{pmatrix} 2 & 2 & 0 & 0 & 2 \\ -2 & 2 & 0 & 2 & 0 \\ -2 & 0 & 0 & -2 & 2 \\ 0 & -\sqrt{3} & \sqrt{3} & \sqrt{3} & \sqrt{3} \\ 0 & \sqrt{3} & \sqrt{3} & -\sqrt{3} & -\sqrt{3} \end{pmatrix}.$$

This shows that

$$\max_{A \in \mathbb{R}^{5\times 5}} \beta(A) \geq \frac{2+3\sqrt{3}}{4} \approx 1.799\dots.$$

Moreover, the way this matrix acts on the $2^5 = 32$ vertices of the hypercube $\{-1,1\}^5$ is interesting: 8 of the 32 vertices are sent to vectors with maximal entry 2 (in absolute value) and the remaining 24 vertices are sent to vertices with maximal coordinate $\sqrt{3}$ (in absolute value). We do not know whether this matrix is indeed optimal. Indeed, even for relatively small dimensions, like $n = 3$ or $n = 4$, rigorously proving that a certain candidate matrix is indeed extremal appears to be a highly nontrivial problem.

3.1.5 Known results.

Relatively little is known for the bad science matrix problem. Steinerberger proved that, as $n \to \infty$, the 'worst' matrix scales as

$$\beta(A) = (1 + o(1)) \cdot \sqrt{2 \log n}.$$

The upper bound follows relatively quickly from Hoeffding's inequality and the union bound. As for the lower bound, it is shown that a random matrix with entries $\pm 1/\sqrt{n}$ very nearly saturates the upper bound. The proof is completely non-constructive and does not yield a single example of a truly 'bad' bad science matrix. Moreover, some basic numerical experiments suggest that the extremal matrices (at least in low dimensions, say, $n \leq 8$) are quite different from random $\pm 1/\sqrt{n}$ matrices and have interesting behavior: in particular, their entries were noted to be exclusively algebraic numbers like $\sqrt{3}$ or $3/\sqrt{29}$. We also note that the bad science matrix problem appears to admit *many* extremal matrices: there is an obvious symmetry in permuting rows and columns but it appears that there is a great degree of non-uniqueness. Finally, nothing is known about the more general bad science matrix problem for rectangular matrices.

3.2 Main Results

3.2.1 Structure Theorem and Consequences.

Our first main result is a theorem about the structure of extremal matrices. The result proves a strong one-to-one correspondence between the i^{th} row of an extremal matrix and the vertices of the hypercube $\{-1, 1\}^n$ that are being mapped by the matrix to a point whose coordinates attain their maximum absolute value in the i^{th} coordinate.

Theorem 10 (Structure Theorem)**.** *Let* $A = \begin{bmatrix} a_1 \cdots a_m \end{bmatrix}^\top$ *be an* $m \times n$ *matrix maximizing the value of* $\beta(A)$ *among all such matrices with rows normalized in* ℓ^2*. Introducing the set* W_i *of vertices of the hypercube that are being mapped to a vector whose largest entry is in* *the i^{th} coordinate,*

$$W_i = \left\{ x \in \{-1,1\}^n \;\middle|\; \|Ax\|_\infty = a_i^\top x \right\},$$

then the i^{th} row of A is given by

$$a_i = \frac{\sum_{x \in W_i} x}{\left\| \sum_{x \in W_i} x \right\|}.$$

This structure theorem has a number of immediate consequences. The first consequence is something that was already observed (without proof) in [81]: entries of extremal entries appear to be either rational numbers or square roots of rational number (indeed, this was essentially in the construction of extremal candidates in [81]). This is indeed the case and follows easily from the Structure Theorem.

Corollary. *The entries of an optimal bad science matrix are square roots of rational numbers.*

We observe that Theorem 10 gives us something much stronger, it shows that extremal entries are really rescaled vectors which themselves are merely the sum of all vertices of a subset of the hypercube. We will use this structural restriction to solve the bad science matrix problem in dimensions $n \leq 4$.

Corollary (Cases $n \leq 4$)**.** *If $A \in \mathbb{R}^{n \times n}$ has normalized rows, then*

$$\frac{1}{2^n} \sum_{x \in \{1,1\}^n} \|Ax\|_\infty \leq \begin{cases} \sqrt{2} & \text{when } n = 2, \\ (\sqrt{2}+\sqrt{3})/2 & \text{when } n = 3, \\ \sqrt{3} & \text{when } n = 4. \end{cases}$$

These bounds are attained by the matrices

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}, \quad \text{and} \quad \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & -1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 \end{bmatrix}.$$

Our proof of Corollary 2 makes extensive use of the computer: the structure theorem can be used to deduce that extremal matrices have to be of a very specific form. One can then, for small values of n, explicitly analyze matrices of that size. Thus, in some sense, our proof of Corollary 2 is based on explicit enumeration. However, this scales rather badly in n and even the case $n = 5$ appears to be firmly out of reach. It is rather remarkable that we do not currently have a simple way of establishing extremality for small n. We conclude with a very simple observation.

Corollary. *The number of extremal $n \times n$ matrices is at most $2^{n \cdot 2^n}$.*

This can be seen very easily: the number of subsets of $\{-1, 1\}^n$ is $\leq 2^{2^n}$ which shows that there are at most this many possibilities for each row. There are n rows and the inequality follows. Since there is always at least one extremal matrix (by compactness), permutation of rows and columns shows that the number of extremal matrices is at least $\geq n!$ and probably much larger. It could be interesting if this number could be further narrowed down but this appears to be a difficult problem.

3.2.2 Construction 1: Lifting Construction.

A natural question is whether it is possible to explicitly construct examples of matrices A for which $\beta(A)$ is large. In what follows, we present a number of different approaches to this question. The first observation is that from one good example, one can obtain larger good examples: this is somewhat similar to Sylvester's construction of Hadamard matrices and uses a classic lifting trick.

Theorem 11. *For any $A \in \mathbb{R}^{n \times n}$ with rows normalized to length 1, the matrix*

$$B = \frac{1}{\sqrt{\beta(A)^2 + 1}} \begin{bmatrix} \beta(A) \cdot A & \mathrm{Id}_{n \times n} \\ \beta(A) \cdot A & -\mathrm{Id}_{n \times n} \end{bmatrix} \qquad \textit{satisfies} \qquad \beta(B) = \sqrt{\beta(A)^2 + 1}.$$

In particular, starting with the optimal 2×2 matrix A_2 and applying Theorem 11 repeatedly, we arrive at an explicit infinite family of matrices $(A_{2^k \times 2^k})_{k=1}^{\infty}$.

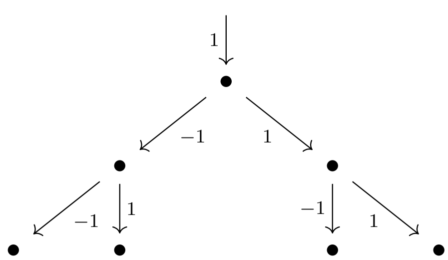
Corollary. *We have*

$$\beta(A_{2^k\times 2^k}) = \sqrt{k+1} = \sqrt{\log_2(n)+1}.$$

We note that, in particular,

$$\beta(A_{2\times 2}) = \sqrt{2}, \qquad \beta(A_{4\times 4}) = \sqrt{3}, \qquad \text{and} \qquad \beta(A_{8\times 8}) = 2$$

which matches the currently best known constructions in these dimensions (see [81]). Moreover, by Corollary 2 above, we know that the 2×2 and 4×4 constructions are optimal. It stands to reason that the so arising 8×8 matrix is also optimal but this remains open. Since $\sqrt{2\log n}/\sqrt{\log_2 n} = \sqrt{2\log 2} \approx 1.177$, this construction is a factor of approximately 1.177 away from optimal. The family of matrices matches a construction given by Steinerberger [81] which is itself based on a construction used by Kunisky [56] for the Komlós problem. We note that the construction is limited to $n\times n$ matrices where $n = 2^k$ is a power of 2. The next two constructions do not suffer from this drawback.



$$\begin{bmatrix} 1 & -1 & -1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Figure 3.1: An unsatisfiable tree and the corresponding matrix

3.2.3 Construction 2: Unsatisfiable Trees.

A more general construction can be given using highly balanced binary trees. We quickly recall their definition.

Definition 1 (Highly balanced binary trees)**.** *For a fixed integer $n \geq 1$, fill up a binary tree with vertices from left to right until one has n leaves, and finally add an edge that points into* *the root.*

We label the edges of such a highly balanced binary tree in the following way: edges that point left have label -1, edges that point right have label 1, and the edge that points to the root has label 1. From here, for a leaf v, walk along the unique path from the root to v. Then the edge labels of this path becomes a row of the matrix this method generates, where if the length of the path is less than n, we make the rest of the entries 0. The case $n = 4$ is illustrated in Fig. 1. One of our main results is a precise analysis of this construction: as it turns out, the construction performs exactly as well as the Hadamard-style tensor construction when n is a power of 2.

Theorem 12. *Let $A \in \mathbb{R}^{n\times n}$ be the matrix whose rows correspond to the n prefixes in the binary tree described above. This construction yields*

$$\begin{aligned}\beta(A) = &\left(2\sqrt{\lfloor \log_2 (n) + 1\rfloor} - \sqrt{\lfloor \log_2 n\rfloor + 2}\right) \\ &+ \frac{n}{2^{\lfloor \log_2 n\rfloor}}\left(\sqrt{\lfloor \log_2 n\rfloor + 2} - \sqrt{\lfloor \log_2 n\rfloor + 1}\right)\end{aligned}$$

When $n = 2^k$ is itself a power of 2, the matrices derived from this method satisfy $\beta(A_{2^k\times 2^k}) = \sqrt{k+1}$ which matches Corollary 3.2.2. However, the construction is more versatile and gives good examples of $n \times n$ matrices where n is not a power of 2.

3.2.4 Wide Matrices

So far, our focus has been on $n \times n$ matrices, however, there is no particular reason for doing so. We give two solutions for the $1 \times n$ case.

Theorem 13. *For any vector $a \in \mathbb{R}^n$, we have*

$$\frac{1}{2^n}\sum_{x\in\{-1,1\}^n} |\langle a, x\rangle| \leq \|a\|_2.$$

Equality is attained when $a = (1, 0, \ldots, 0)$.

Theorem 13 follows from the sharp Khintchine inequality [46]. There is also an elementary proof using only Jensen's inequality. After solving the $1 \times n$ case, one would try the $2 \times n$ case next. This case appears to already be remarkably challenging and gives rise to some amusing problems. We believe to have identified the correct answer but were unable to prove this rigorously.

Conjecture 1. *An optimal $2 \times n$ matrix is given by*

$$A = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \end{bmatrix}$$

The structure theorem immediately suggests the connection to set partitions. While investigating this connection, we came across a question that we were able to resolve and which appears to be of intrinsic interest.

Theorem 14. *For any $A \subset \{-1,1\}^n$ we have*

$$\left\| \sum_{x \in A} x \right\|_2 \le 2^{n-1}$$

with equality if we choose A to be the set of all vectors whose $i-th$ coordinate is fixed to be either 1 or -1.

3.2.5 ℓ^p variants.

The ℓ^∞ norm leads to fascinating extremizers but is otherwise difficult to deal with. A natural question is whether and in what way things change if we investigate the problem for other p. As it turns out, the natural other cases $p = 1$ and $p = 2$ are not quite as interesting and can be completely resolved.

Theorem 15 ($p = 1$ and $p = 2$)**.** *We have*

$$\max_{A \in \mathbb{R}^{n \times n}} \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \|Ax\|_1 = n$$

with equality being attained at the identity matrix. Moreover,

$$\max_{A\in\mathbb{R}^{n\times n}} \frac{1}{2^n} \sum_{x\in\{-1,1\}^n} \|Ax\|_2 = \sqrt{n}$$

with equality if A is an isometry. For $2 < p < \infty$, we have, for some universal constant $c > 0$, that

$$\max_{A\in\mathbb{R}^{n\times n}} \sum_{x\in\{-1,1\}^n} \|Ax\|_{\ell^p} \leq c\cdot\sqrt{p}\cdot n^{1/p}.$$

We note that the intermediate range $2 < p < \infty$ remains open but Theorem 15 gives an upper bound for the order of growth.

3.3 Proofs

3.3.1 Proof of the Structure Theorem (Theorem 10)

Proof. Define W_i to be the set of $x \in \{\pm1\}^n$ so that i is the smallest index with absolute largest coordinate in Ax. By construction, we see that the W_i are closed under $x \mapsto -x$, that they are disjoint, and hence partition $\{\pm1\}^n$. Thus we can write:

$$\beta(A) = \sum_{x\in\{\pm1\}^n} \|Ax\|_\infty = \frac{1}{2^n}\sum_{i=1}^{n}\sum_{x\in W_i} |a_i^\top x|$$

Now take $U_i = \left\{x \in W_i \;\middle|\; |a_i^T x| = a_i^T x\right\}$ to be the positive half of W_i. For each $x \in W_i$ with $a_i^T x = 0$, U_i will include both x and $-x$, but we will eventually want U_i to be antipodal. So in this case keep only one of x and $-x$. Then we have that:

$$\sum_{x\in W_i} |a_i^\top x| = 2\sum_{x\in U_i} a_i^\top x.$$

The idea is to now optimize each of the sums $\sum_{x\in U_i} a_i^\top x$ separately, since we have removed the independence among the rows from the objective function. One has

$$2\sum_{x\in U_i}\langle a_i, x\rangle = 2\left\langle a_i, \sum_{x\in U_i} x\right\rangle$$

By the Cauchy-Schwarz inequality,

$$\left\langle a_i, \sum_{x \in U_i} x \right\rangle \leq \|a_i\| \left\| \sum_{x \in U_i} x \right\| = \left\| \sum_{x \in U_i} x \right\|$$

since each of A's rows have norm 1. Now, if A does not have equality for all of the above inequalities, then we could replace A's rows with

$$\frac{\sum_{x \in U_i} x}{\| \sum_{x \in U_i} x\|}$$

to make the objective function strictly larger. But A is optimal, so we see that a_i is a positive multiple of the above vector, again by Cauchy-Schwarz. Since a has unit norm, we conclude that

$$a_i = \pm \frac{\sum_{x \in U_i} x}{\| \sum_{x \in U_i} x\|},$$

completing the proof. □

Notice that since each of the $x \in U_i$ are vectors on the hypercube, namely have every coordinate of ± 1, $\sum_{x \in U_i} x$ is a vector with integer coordinates. Thus we see that $\left\|\sum_{x \in U_i} x\right\|$ is a square root of an integer, and hence every entry of a_i is a square root of a rational number.

3.3.2 Proof of Corollary 2

Proof. This argument uses a computer search to check all possible candidate matrices. Theorem 10 implies each row a_i of an optimal matrix A is a normalized sum of vectors contained in some subset $V \subseteq \{-1, 1\}^n$, i.e.

$$a_i = \frac{\sum_{x \in V} x}{\left\|\sum_{x \in V} x\right\|}.$$

We can now check all possible matrices formed by combining rows of this form. Note, however, that this exhaustive search becomes prohibitive very quickly. Indeed, there are 2^{2^n} possible subsets of $\{-1, 1\}^n$, and must consider all possible $n-$tuples of rows of this form. Naïvely, this yields $(2^{2^n})^n = 2^{n \cdot 2^n}$ matrices to check. For $n = 3$ the search space is of size

$2^{24} \approx 1.7 \times 10^7$, and for $n = 4$ we have $2^{64} \approx 1.8 \times 10^{19}$ possibilities. We exploit the following symmetries.

(i) The β value of a matrix is invariant under permutations over rows.

(ii) An optimal matrix may not have duplicate rows.

(iii) An optimal matrix doesn't have zero rows.

(iv) The β value of a matrix is invariant under row-wise sign flips

Observations (i) and (ii) allow us to consider row combinations instead, yielding $\binom{2^{2^n}}{n}$ matrices to check. By observation (iii), we further reduce the search space by deleting the zero rows (for instance the full subset $V = \{-1, 1\}^n$ yields such a row) and by observation (iv) we eliminate duplicate rows equal up to a sign flip (for instance, we keep only one of the rows arising from the subsets $V_1 = \{1\} \times \{-1, 1\}^{n-1}$ and $V_2 = \{-1\} \times \{-1, 1\}^{n-1}$). For the $n = 3$ case, we start with $2^{2^3} = 256$ rows, 238 of which are non-zero. Furthermore, since there may be multiple subsets V leading to the same row, we eliminate redundant ones, yielding 50 unique rows. Finally, removing redundant rows up to a sign-change reduces the search space to 25 final rows. This leaves $\binom{25}{3} = 2300$ candidate matrices, which can be easily computed. The $n = 4$ case is approached in the same manner. Performing the same reductions as before, we go from $2^{2^4} = 65536$ rows to 680 candidate rows, yielding $\binom{680}{4} = 8830510430$ matrices to check. These can be checked in a rather long but attainable amount of time. For context, a naïve sequential search took 72 hours to complete but we suspect it could be vastly parallelized to reduce the runtime. However, the $n = 5$ case is prohibitively expensive, since the initial rows $2^{2^5} \approx 4.3 \times 10^9$ cannot even be loaded into memory. Similarly, the rectangular bad science problem of dimensions 2×3, 2×4 and 3×4 can be checked, with search spaces of size $\binom{25}{2} = 300$, $\binom{680}{2} = 230860$ and $\binom{680}{3} = 52174360$, respectively. The first two cases took less than a minute to run, but the 3×4 took one hour on Google Colab. The Jupyter Notebooks containing the row reductions and the exhaustive search, with step-by-step comments, can be found in the Github repository [6]. $\square$

3.3.3 Proof of Theorem 11

Proof. Fix

$$\gamma = \frac{\beta(A)}{\sqrt{\beta(A)^2+1}}$$

and let $x = [x_1\ x_2] \in \{-1,1\}^{2n_0}$, with $x_1 \in \{-1,1\}^{n_0}$ and $x_2 \in \{-1,1\}^{n_0}$. Then, Bx can be written as

$$Bx = \begin{bmatrix} \gamma A x_1 + \sqrt{1-\gamma^2}x_2 \\ \gamma A x_1 - \sqrt{1-\gamma^2}x_2 \end{bmatrix}$$

Note that coordinate j of Bx has the form

$$(Bx)_j = \begin{cases} \gamma a_j^\top x_1 + \sqrt{1-\gamma^2}(x_2)_j & \text{if} \quad 1 \le j \le n_0 \\ \gamma a_{j-n_0}^\top x_1 - \sqrt{1-\gamma^2}(x_2)_{j-n_0} & \text{if} \quad n_0 < j \le 2n_0 \end{cases}$$

We now note that each term of the form $\gamma \cdot a_j^\top x_1$ contributes to two different coordinates of Bx: it does once to coordinate j, with the term being added to $\sqrt{1-\gamma^2}(x_2)_j$, and again to coordinate $j+n_0$, with the term being subtracted $\sqrt{1-\gamma^2}(x_2)_j$. Now, we may bound the absolute value of the coordinates of Bx in a straightforward manner. Letting the index $1 \le i \le n_0$ be the one for which $\|Ax_1\|_\infty = |a_i^\top x_1|$ holds, for $1 \le j \le n_0$ we have

$$\begin{aligned} |(Bx)_j| &\le \max\{|\gamma a_j^\top x_1 + \sqrt{1-\gamma^2}(x_2)_j|, |\gamma a_j^\top x_1 - \sqrt{1-\gamma^2}(x_2)_j|\} \\ &\le \gamma |a_j^\top x_1| + \sqrt{1-\gamma^2} \\ &\le \gamma |a_i^\top x_1| + \sqrt{1-\gamma^2} = \gamma \|Ax_1\|_\infty + \sqrt{1-\gamma^2} \end{aligned}$$

and equally for the second half of the coordinates $|(Bx)_{j+n_0}|$. Thus

$$\|Bx\|_\infty \le \gamma \|Ax_1\|_\infty + \sqrt{1-\gamma^2}.$$

The important observation comes from noting that if $a_j^\top x_1$ and $\sqrt{1-\gamma^2}(x_2)_j$ agree in sign, then coordinate j will be larger in absolute value than coordinate $j+n_0$, and vice-versa if $a_j^\top x_1$ and $\sqrt{1-\gamma^2}(x_2)_j$ disagree in sign. With this in mind, we see that one of the choices $j = i$ or $j = i+n_0$ (the one in which the terms $\gamma a_i^\top x_1$ and $\sqrt{1-\gamma^2}(x_2)_i$ or $\gamma a_i^\top x_1$ and $-\sqrt{1-\gamma^2}(x_2)_i$

agree in sign) achieves equality in each step, yielding $\|Bx\|_\infty = \gamma \|Ax_1\|_\infty + \sqrt{1-\gamma^2}$. It is easy to see that summing over all possible $x = \{-1,1\}^{2n_0}$ translates this result to the β setting:

$$\beta(B) = \frac{1}{2^{2n_0}} \sum_{x\in\{-1,1\}^{2k}} \|Bx\|_\infty = \frac{1}{2^{n_0}\cdot 2^{n_0}} \sum_{x_1,x_2\in\{\pm1\}^{n_0}} \gamma \|Ax_1\|_\infty + \sqrt{1-\gamma^2}$$
$$= \gamma\beta(A) + \sqrt{1-\gamma^2}$$

Plugging in $\gamma = \beta(A)/\sqrt{\beta^2(A)+1}$, we obtain

$$\beta(B) = \frac{1}{\sqrt{\beta(A)^2+1}} + \frac{\beta^2(A)}{\sqrt{\beta(A)^2+1}} = \sqrt{\beta(A)^2+1}.$$

Finally, we note that if A has normalized rows then B does too, since $\gamma^2 + 1 - \gamma^2 = 1$. This completes the proof. $\square$

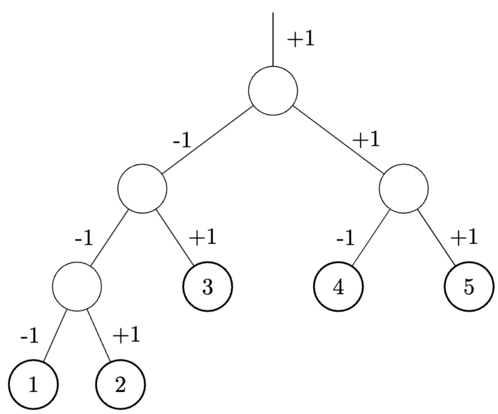


Figure 3.2: Example of binary tree with five leaves.

3.3.4 Proof of Theorem 12

Definition 2. *We define A as a set of square matrices which we construct from building a binary tree which is highly balanced as in Definition 1. Every row in the matrix corresponds to some leaf in the tree, and every column in the matrix corresponds to a level of the binary tree. Given this binary tree we yield the corresponding matrix in the following way:*

- *We let the first column be the all 1's column*
- *For every path from the root to the leaves: we correspond $A_{i,j}$ to the leaf with index $i \in [n]$ at the j'th vertex in the path from the root. If this vertex represents the left child of its parent vertex, we let this entry be -1, otherwise we let this entry be $+1$. Entries that do not correspond to vertices in this tree correspond to 0's in the matrix.*

We construct the corresponding matrix by mapping every leaf to the row with the same index and the index of every column corresponding to the level in the tree. For example, to construct row 3, we examine the root-to-leaf path of the leaf labeled with 3 which is $\{+1, -1, +1\}$. We set the first 3 entries of the row to match this prefix and set the rest of the entries of the row to 0. We repeat this process for the rest of the rows in this binary tree and before normalization, the corresponding matrix would look as follows:

$$A_5 = \begin{bmatrix} +1 & -1 & -1 & -1 & 0 \\ +1 & -1 & -1 & +1 & 0 \\ +1 & -1 & +1 & 0 & 0 \\ +1 & +1 & -1 & 0 & 0 \\ +1 & +1 & +1 & 0 & 0 \end{bmatrix}$$

We would then normalize the rows of this matrix to discuss its β value. We now use these matrices to prove Theorem 12.

Proof of Theorem 12. Suppose $2^k < n < 2^{k+1}$ (deferring the case when $n = 2^k$ to Corollary 3.2.2), i.e. $k = \lfloor \log_2(n) \rfloor$. Note that all rows of A are of the form $\{\pm 1\}^{k+1}$ followed by $n-k-1$ zeros or $\{\pm 1\}^{k+2}$ followed by $n-k-2$ zeros. We leverage the fact that if $|a_i^T x| = \|Ax\|_\infty$ for x in the cube, then $x = \pm a_i$ entry-wise except for the zero entries of a_i. This is because if we let $\tilde{a}$ be obtained by truncating the zeros off of a_i, Cauchy-Schwarz tells us that $\max_{\|\tilde{x}\|=1} |\tilde{a}^T \tilde{x}|$ is obtained at $\tilde{x} = \pm \tilde{a}$. Therefore, for any $x \in \{-1, 1\}^n$, $\|Ax\|_\infty$ is $\sqrt{k+1}$ or $\sqrt{k+2}$, corresponding to the length of the matched prefix. The number of possible vectors x which agree (in the aforementioned manner) with a_i is given by either

$2^{n-k-1} \cdot 2$ or $2^{n-k-2} \cdot 2$, (two choices for each zero entry, and one last doubling to account for negation). Then, we have that

$$\begin{aligned}
\beta(A) &= \frac{1}{2^n} \cdot \left[\sqrt{k+1} \cdot (2^{k+1} - n) \cdot 2^{n-k} + \sqrt{k+2} \cdot (2n - 2^{k+1}) \cdot 2^{n-k-1} \right] \\
&= \sqrt{k+1} \cdot \frac{2^{k+1} - n}{2^k} + \sqrt{k+2} \cdot \frac{2n - 2^{k+1}}{2^{k+1}} \\
&= \left(2\sqrt{k+1} - \sqrt{k+2} \right) + \frac{n}{2^k} \cdot \left(\sqrt{k+2} - \sqrt{k+1} \right)
\end{aligned}$$

Replacing k as $\lfloor \log_2(n) \rfloor$ recovers the desired expression. □

3.3.5 Proof of Theorem 13

Proof. We can and will assume without loss of generality that $\|a\|_{\ell^2} = 1$. First recall that

$$\beta(a) = \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} |a^\top x|.$$

Let x be a random vector whose entries are ± 1 with equal probability. Then by definition we have that $\beta(a) = \mathbb{E}[|a^\top x|]$. It remains to prove $\mathbb{E}[|a^\top x|] \leq 1$. We start by showing that $\mathbb{E}[xx^\top] = \mathrm{Id}_{n \times n}$. Notice that

$$\mathbb{E}[xx^\top]_{ij} = \mathbb{E}[x_i x_j] = \begin{cases} 1 & i = j \\ 0 & \text{otherwise.} \end{cases}$$

Now by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}[|a^\top x|] &= \mathbb{E}\left[\sqrt{(a^\top x)^2} \right] \leq \sqrt{\mathbb{E}[a^\top x x^\top a]} = \sqrt{a^\top \mathbb{E}[xx^\top] a} \\
&= \sqrt{a^\top a} = 1
\end{aligned}$$

since $\|a\|_2 = 1$. This completes the proof. □

3.3.6 Proof of Theorem 14

We first start with a simple Lemma.

Lemma 14. *The set $A \subset \{-1,1\}^n$ maximizing*

$$A \to \left\| \sum_{a \in A} a \right\|_2$$

has cardinality $\#A = 2^{n-1}$ and does not contain any antipodal points.

Proof. For the sake of contradiction, assume the optimal set A contains a set of antipodal points $a, -a \in A$. Then, $A = U \cup \{-a, a\}$ for some set $U \subset \{-1,1\}^n$. We now consider the term $\left\|\sum_{u \in U} u\right\|^2$. By switching a with $-a$ if necessary, we may without loss of generality assume that $\langle \sum_{u \in U} u, a \rangle \geq 0$. Then, by adding the strictly positive term $\|a\|^2$ and the nonnegative term $2\left\langle \sum_{u \in U} u, a \right\rangle$, we obtain the strict inequality

$$\left\| \sum_{u \in U} u \right\|^2 < \left\| \sum_{u \in U} u \right\|^2 + 2\left\langle \sum_{u \in U} u, a \right\rangle + \|a\|^2 = \left\| \sum_{u \in U \cup \{a\}} u \right\|^2.$$

Finally, we note that $\left\|\sum_{u \in A} u\right\|^2 = \left\|\sum_{u \in U} u\right\|^2$ since a and $-a$ cancel out in the sum. This shows the set A is not optimal, a contradiction. Having established this, it is easy to see that $\#A = 2^{n-1}$. Indeed, for an optimal set A, if $\#A > 2^{n-1}$ the pigeonhole principle ensures two antipodal points, a contradiction. If $\#A < 2^{n-1}$, we may apply the pigeonhole principle to the complement set $\{-1,1\}^n \backslash A$ of size $2^n - \#A > 2^n - 2^{n-1} > 2^{n-1}$, which shares the same objective value as A, to determine it is non-optimal. $\square$

Proof of Theorem 14. Lemma 7 shows that it suffices to consider sets A of size 2^{n-1} with no pair of antipodal points (if $a \in A$ then $-a \notin A$). Working under such assumptions, we let $u = \sum_{a \in A} a$, expand the square of the norm, and use the linearity of the inner product:

$$\left\| \sum_{a \in A} a \right\|^2 = \sum_{a,b \in A} a^\top b = \sum_{b \in A} \left(\sum_{a \in A} a \right)^\top b = \sum_{b \in A} u^\top b.$$

Next, we bound each term by its absolute value and add all the antipodal points into the sum, dividing the total expression by 2 to maintain equality:

$$\sum_{b \in A} u^\top b \leq \sum_{b \in A} \left| u^\top b \right| = \frac{1}{2} \sum_{b \in \{\pm 1\}^n} \left| u^\top b \right|.$$

We furthermore assume that $u \neq 0$ since this case is trivially suboptimal. Thus, we normalize u by multiplying and dividing by $\|u\|$,

$$\frac{1}{2} \sum_{b \in \{\pm 1\}^n} \left| u^\top b \right| = \frac{\|u\|}{2} \sum_{b \in \{\pm 1\}^n} \left| \frac{u}{\|u\|}^\top b \right|$$

Since we have a normalized vector and sum over all possible hypercube vectors, we may apply Theorem 13 to obtain

$$\frac{\|u\|}{2} \sum_{b \in \{\pm 1\}^n} \left| \frac{u}{\|u\|}^\top b \right| \leq \frac{\|u\|}{2} \sum_{b \in \{\pm 1\}^n} \left| e_i^\top b \right|$$

for all $1 \leq i \leq n$, where $e_i \in \mathbb{R}^n$ is the ith standard unit vector. Since $|e_i^\top b| = 1$ for all $b \in \{-1, 1\}^n$, the right hand side is simply

$$\frac{\|u\|}{2} \sum_{b \in \{\pm 1\}^n} \left| e_i^\top b \right| = \frac{\|u\|}{2} \cdot 2^n = \|u\| \, 2^{n-1}.$$

It now suffices to divide both sides by $\|u\| = \left\| \sum_{a \in A} a \right\|$ to obtain the upper bound

$$\left\| \sum_{a \in A} a \right\| \leq 2^{n-1}.$$

Finally, we determine the optimal sets by asking when equality is achieved. By Theorem 15, we note that the inequalities become equalities if and only if $u / \|u\| = e_i$. We see that $A = A_i = \{a \in \{-1, 1\}^n : a_i = 1\}$ satisfies

$$\frac{\sum_{a \in A} a}{\left\| \sum_{a \in A} a \right\|} = e_i,$$

and the proof is complete. □

3.3.7 Proof of Theorem 15

Proof. For $1 \leq p \leq \infty$ and $n \in \mathbb{N}$, we abbreviate

$$\beta_{n,p} = \max_{A \in \mathbb{R}^{n \times n}} \frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \|Ax\|_p = \max_{A \in \mathbb{R}^{n \times n}} \mathbb{E} \left[\|Ax\|_p \right].$$

We also note that $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{n}$ since A has normalized rows.

The case $p = 1$: Taking a_i to be the ith row of A, we have that,

$$\frac{1}{2^n} \sum_{x \in \{-1,1\}^n} \|Ax\|_1 = \frac{1}{2^n} \sum_{x \in \{\pm 1\}^n} \sum_{i=1}^{n} \left|a_i^\top x\right| = \sum_{i=1}^{n} \mathbb{E}\left[|a_i^\top x|\right] \leq n$$

since $\mathbb{E}[|a_i^T x|] \leq 1$ when x is a vector with ± 1 random entries with equal probability, and $|a_i| = 1$ by Theorem 13 (the $1 \times n$ case). The identity matrix achieves this, since $\|x\|_1 = n$ for every $x \in \{\pm 1\}^n$.

The case $p = 2$: Again by Jensen's inequality, we have that

$$\mathbb{E}\left[\|Ax\|_2\right] = \mathbb{E}\left[\sqrt{x^\top A^\top A x}\right] \leq \sqrt{\mathbb{E}[x^\top A^\top A x]}.$$

Then, since if $b \in \mathbb{R}$ then we can treat b as a 1×1 matrix and have $b = \mathrm{Tr}(b)$, yielding

$$\begin{aligned}
\sqrt{\mathbb{E}[x^\top A^\top A x]} &= \sqrt{\mathbb{E}[\mathrm{Tr}(x^\top A^\top A x)]} \\
&= \sqrt{\mathbb{E}[\mathrm{Tr}(A^\top A x x^\top)]} && \text{(since } \mathrm{Tr}(ABC) = \mathrm{Tr}(CAB)) \\
&= \sqrt{\mathrm{Tr}(A^\top A \mathbb{E}[x x^\top])} && \text{(linearity of expectation)} \\
&= \sqrt{\mathrm{Tr}(A^\top A)} = \sqrt{\mathrm{Tr}(A A^\top)} && \text{(since } \mathbb{E}[x x^\top] = I)
\end{aligned}$$

The entry of $AA^\top$ at position (i,i) is just $a_i^\top a_i = \sum_j A_{ij}^2$ if a_i is the ith row of A. The sum of all these is then seen to be $\sum_{i,j} A_{ij}^2 = n$ as we showed above. This shows that,

$$\mathbb{E}[\|Ax\|_2] \leq \sqrt{n}$$

Now we shall show that if A is orthogonal then the above becomes an equality. Recall that Jensen's inequality becomes an equality if the random variable is constant. When A is orthogonal, we have that $\|Ax\|_2^2 = x^\top A^\top A x = x^\top x = n$. Thus in this case $x^\top A^\top A x$ is constant and the above becomes an equality.

The case when $2 < p < \infty$. By Khintchine's inequality [46], if $p > 2$ we have that:

$$\mathbb{E}|a^\top x|^p \leq \frac{2^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \leq 2^{p/2}(p/2)!$$

Thus, by Jensen's inequality and Stirling's approximation,

$$\begin{aligned}
\mathbb{E}\left[\|Ax\|_p\right] &\leq \left(\sum_{i=1}^{n} \mathbb{E}_x|a_i^\top x|\right)^{1/p} \leq \left(n \cdot 2^{p/2}(p/2)!\right)^{1/p} \\
&\leq Cn^{1/p}(p/2)!^{1/p} \leq Cn^{1/p}\frac{(p/2)^{1/2+1/p}}{e^{1/2-1/p}} \leq Cn^{1/p}(e/2)^{1/p}p^{1/p}p^{1/2} \\
&\leq Cp^{1/2}n^{1/p}
\end{aligned}$$

For some constant C independent of p (The constant changed 3 times in 1 line). This shows that $\beta_{n,p} \leq C\sqrt{p} \cdot n^{1/p}$. However, due to the factor of $\sqrt{p}$, the bounds deteriorate for small n as p increases. $\square$

Chapter 4

PARERGA AND PARALIPOMENA

4.1 Fractals

4.1.1 Introduction

The Chaos Game is an algorithm that is commonly used to render fractals. One particular instance can be described as follows. Given a polygon Q and some initial point p, iteratively select a vertex of Q (chosen randomly) and move p to the midpoint of its current position and the selected vertex. The trajectory of p tends toward a particular set, which can be a fractal. Explicitly, let $f_i(x) = \frac{1}{2}(x + p_i)$, where $p_i \in \mathbb{R}^2$ and $i \in I$ for some finite index set I. Famously, when the p_i are 3 non-collinear points, the attractor of the Chaos Game is the famous Sierpinski Triangle.

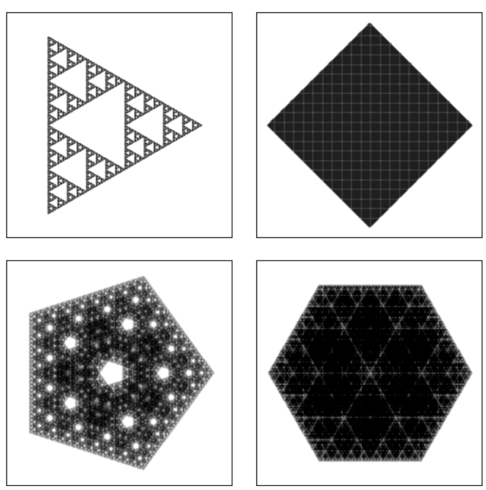


Figure 4.1: The Chaos Game on four polygons.

Practical implementations of the Chaos Game typically use the following algorithm.

- Fix some arbitrary point $q_0 \in \mathbb{R}^2$.
- Compute $q_{i+1} = f(q_i)$, where f is chosen uniformly at random from $\{f_i : i \in I\}$.

Note that if q_0 is chosen to be in the attractor of the IFS, then the entire sequence q_i is guaranteed to be in the attractor of the IFS. Alternatively, let $J \in I^k$, i.e. $J = (j_1, \ldots, j_{|J|}$, where each $j_i \in I$. Let

$$f_J = f_{j_1} \circ \ldots \circ f_{j_{|J|}}.$$

Then one can also implement the Chaos Game as computing the set of points $f_J(q_0)$ for all $J \in I^k$. The validity of theory behind these approaches is discussed in [13].

Various extensions of the Chaos Game have been explored. One major variant involves changing the jump factor, i.e. taking $f_i(x) = (1-\alpha)x + \alpha p_i$, for $\alpha \in (0,1)$ not necessarily equal to $\frac{1}{2}$. Depending on $|I|$, there exists a notion of optimal α, as computed by [1]. This optimality is related to iterated function systems and the Open Set Condition, which we discuss later. In particular, the optimal α computed in [1], is the largest value of α allowing the open set condition to apply when viewing the chaos game as an iterated function system.

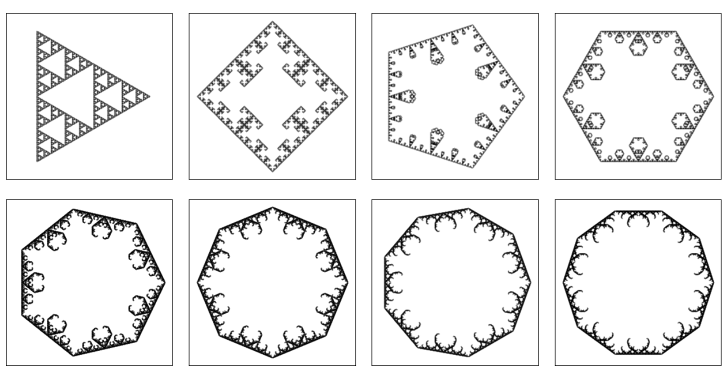


Figure 4.2: A restriction of the Chaos Game on eight polygons.

Another variant adjusts the above algorithm in the following way.

- Fix some arbitrary point $q_0 \in \mathbb{R}^2$.
- Compute $q_{i+1} = f(q_i)$, where f is chosen randomly from $\{f_i : i \in I\}$ according to some probability distribution D_i on $\{f_i : i \in I\}$.

For the rest of this section, we consider this variant where D_i puts equal mass on f_{i-1}, f_i, f_{i+1} (we take this addition modulo n), and zero mass elsewhere. We can alternatively describe this variant in similar language to our initial description of the Chaos Game. Whatever vertex was previously chosen, the next jump must be toward that same vertex or one of its neighbors. We will show that these fractals can be understood both as the attractors of graph-directed iterated function systems and the union of attractors of a particular iterated function system.

4.1.2 Iterated Function Systems

Recall that an *iterated function system* (IFS) is a set of contraction mappings on $\mathbb{R}^n$. This definition is of interest due to Hutchinson[50].

Theorem (Hutchinson)**.** *Let $\{f_i\}_{i \in I}$ be an IFS. There exists a unique non-empty compact set S, known as the attractor of the IFS, satisfying*

$$S = \bigcup_{i \in I} f_i(S).$$

Moreover, Moran provides a powerful tool known as the Open Set Condition (OSC) to compute the Hausdorff dimension of these attractors [67].

Theorem (Moran)**.** *Given an IFS $\{f_i\}_{i \in I}$ with attractor S, if there exists an open set V such that*

$$\bigcup_{i \in I} f_i(V) = V$$

and all $f_i(V)$ are disjoint, then the Hausdorff dimension of S is equal to its similarity dimension.

The Chaos Game is an IFS as defined in the introduction, and we show several examples of attractors (using $\alpha = 1/2$) of the Chaos Game in Figure 4.1.

A generalization of the IFS is the graph directed IFS, which is a directed graph $G = (V, E)$ (with loops permissible), where to each $e \in E$ we equip a contraction f_e. Similar to the relationship between the definition of the IFS and Hutchinson's theorem, this definition is interesting because of the following fact due to [33, 64].

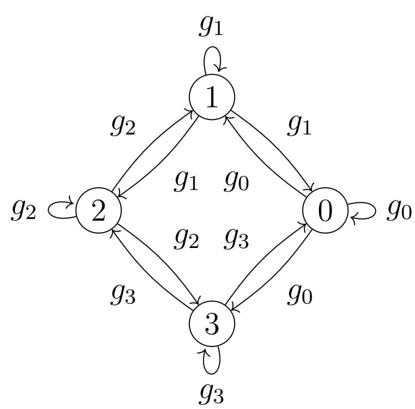


Figure 4.3: A graph G_n with $n = 4$.

Theorem. *Let $G = (V, E)$ with contractions $\{f_e\}_{e \in E}$ be a graph directed IFS, where $|V| = n$. Then there exists a unique collection of non-empty compact sets $\{S_v\}_{v \in V}$ such that*

$$S_v = \bigcup_{v \in v} \bigcup_{(v,w) \in E} f_{(v,w)}(S_w).$$

We call $S = \bigcup_{v \in v} S_v$ the attractor of the graph directed IFS.

Notice that the variant of the Chaos Game we discussed is an instance of a graph directed IFS, though this is less clear. Let G_n denote the graph on $\{0, \ldots, n-1\}$ with edges between any i, j such that $|i - j| \leq 1$, where this subtraction is taken modulo n. To edge (i, j), associate the map

$$g_i(x) = \frac{1}{2}(x + p_i).$$

This may feel slightly backwards; one might expect the edge (i, j) to be associated to the map g_j, since the edge is from i to j. It is necessary in this instance to define our graph directed IFS this way to be consistent with the definition of the attractor S_v. The alternative would yield a completely different fractal.

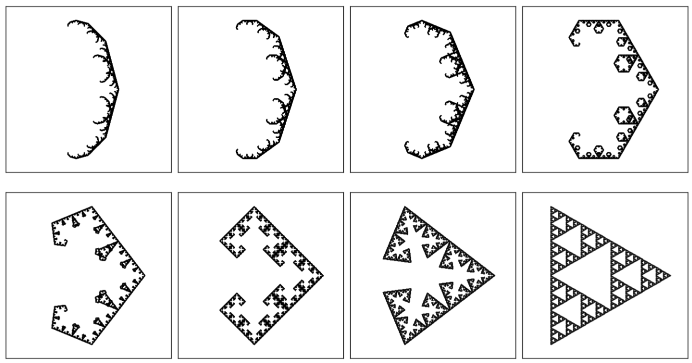


Figure 4.4: Attractors of the IFS with selected β values.

We show several examples of attractors of this variant of the Chaos Game in Figure 4.2. Visually, the attractors shown in Figure 4.2 exhibit some similarities, especially in the second row. They look so alike, in fact, that one might reasonably ask if there is some continuous deformation between these discrete instances. Consider the IFS given by

$$f_i(x) = \frac{1}{2}(T_\beta^i x + e_1),$$

for $i \in \{-1, 0, 1\}$, where $e_1 = (1, 0)$, and T_β denotes counterclockwise rotation about the origin by angle β. We claim that unions of the attractors of this IFS yield the attractors of this graph directed IFS (see Figure 4.4).

4.1.3 Hausdorff Dimension

Proposition 12. *For $\beta \in [\pi/2, 2\pi/3]$, the attractor S of the above IFS has Hausdorff dimension $log_2(3)$.*

The proof is essentially elementary geometry, and we provide a sketch below.

Proof Sketch. Consider the convex quadrilateral P with vertices given by

$$\{(0,0), f_1(e_1), e_1, f_{-1}(e_1)\},$$

and let V be its interior. Let V_i and P_i denote $f_i(V)$ and $f_i(P)$, respectively. Note that the V_i are disjoint if and only if $T_\beta^i V$ (for $i \in \{-1,0,1\}$) are disjoint. One can check that the angle between e_1 and $f_1(e_1)$ is $\frac{\beta}{2}$. Therefore, V is contained in a sector with angle β, in particular

$$V \subset S := \left\{(r\cos(\theta), r\sin(\theta)) : -\frac{\beta}{2} < \theta < \frac{\beta}{2}\right\}.$$

So long as $0 < \beta < \frac{\pi}{2}$, certainly $T_\beta^i S$ are disjoint, so the V_i must also be disjoint. Consider the convex hull of all V_i. One can check that this polygon is contained in V. Therefore, we see that V satisfies the Open Set Condition. We conclude that for $\beta \in [\pi/2, 2\pi/3]$, S has Hausdorff dimension $\log_2(3)$. $\square$

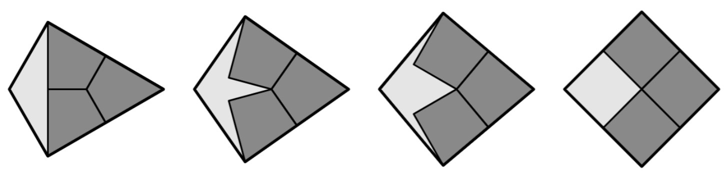


Figure 4.5: The set V as described in the proof of Proposition 12 for $\beta = \frac{2\pi}{3}, \frac{11\pi}{18}, \frac{5\pi}{9}, \frac{\pi}{2}$.

For $\beta < \pi/2$, this construction does not satisfy the Open Set Condition. We suspect that there does not exist an open set satisfying the OSC, but this is of course difficult to verify. Numerics and direct inspection of various f_J suggest that some values of β may yield IFSs that exhibit exact overlap. If these β correspond to contractions with algebraic coefficients, then Hochman showed that there must be dimension drop in these cases [49].

4.2 Evolutes

In 2018, Pat Devlin introduced[1] to us the notion of an evolute by asking the question of what happens in the limit to the intersection of the normals of a curve. At first thought, one might assume that these limits do not exist, as the normals of a line never intersect at all. This, however, is very clearly not the case for a circle, as indeed all normals of a circle intersect at its center. One can check then that the evolute is the locus of a curve's centers of curvature. This motivation of defining the evolute as the limits of intersections of normals (i.e. envelope of normals) leads to the natural question of what happens if we instead use rotated normals. Certainly a $\pi/2$ rotation should intuitively recover the original curve, while a 0 rotation of course yields the evolute itself. Would continuously varying this angle parameter then produce a homotopy from a curve to its evolute? This idea was explored several hundred years ago, and the resulting curves are known as evolutoids, which are well-defined so long as the original curve has a curvature with constant sign. Evolutoids seem to have been largely forgotten, as many resources defining the evolute do not reference the evolutoid.

Figure 4.6: Curve γ with osculating circle (left) and evolute (right).

[1] By introduced, we mean during a discussion over lunch in a Yale dining hall with some other undergraduate students.

4.2.1 Introduction

Evolutes.

The evolute E of a curve γ is the locus of the centers of curvature of γ. One may imagine (see Fig 1.) the 'best-fitting' circle, also known as the osculating circle, touching the curve in a point: the center of this circle is then part of the evolute of the curve γ and the union of all these points for form the evolute of the curve.

This definition requires that the osculating circle be well defined: we require that the curve is at least C^2 and that the curvature does not vanish. If these conditions are satisfied, then the evolute is easily computed as follows: given a curve $\gamma(t) = (x(t), y(t))$, the local normal direction can be found easily from rotating the tangent vector $\gamma'(t)$, and the scale of the osculating circle is given by the inverse curvature, which is explicitly given by

$$\kappa(t) = \frac{\dot{x}\ddot{y} - \ddot{x}\dot{y}}{(\dot{x}^2 + \dot{y}^2)^{3/2}}.$$

Evolutes are classical objects already discussed by Apollonius in 200 BC. We will not summarize all their properties here and instead refer to [2, 57, 63, 89].

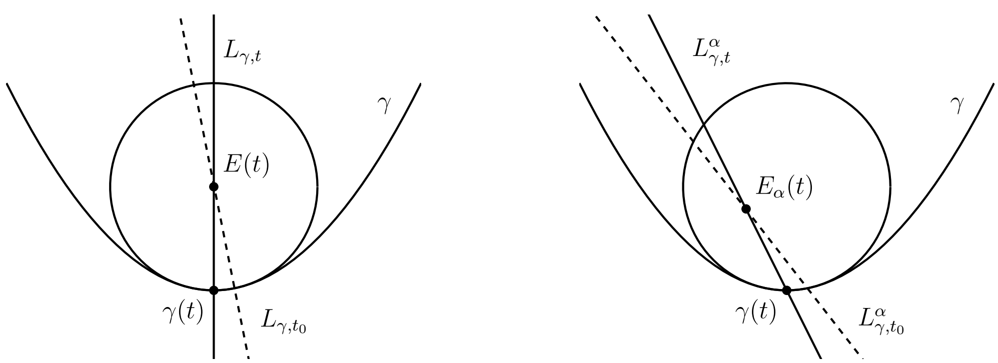


Figure 4.7: Construction of the evolute (left) and evolutoid (right).

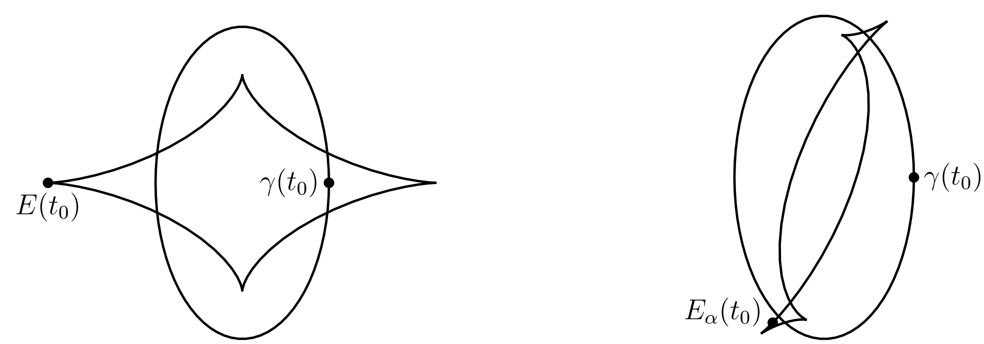
Evolutoids.

We were motivated by a very particular way of defining evolutes that is illustrated in Figure 2: take two nearby points, compute the normal lines and see where they meet. As the two points get closer and closer, in the limit, that meeting point is on the evolute.

One could now wonder what happens if we perform the exact same operation but tilt the two normal lines by α degrees counterclockwise. They will naturally meet elsewhere, and one could wonder whether the union of these points has any nice properties. When $\alpha = 0$, this is the original point on the evolute, and when $\alpha = \pi/2$, it turns out that we recover the original point on the curve.

We now provide the formal definition of this object: for a twice-differentiable curve γ, let $L_{\gamma,t}$ denote the line normal to γ passing through $\gamma(t)$ and $L^{\alpha}_{\gamma,t}$ denote the line through $\gamma(t)$ formed by rotating $L_{\gamma,t}$ counterclockwise by angle α. Define

$$E_\alpha(t) = \lim_{t_0 \to t} I(L^{\alpha}_{\gamma,t}, L^{\alpha}_{\gamma,t_0}),$$

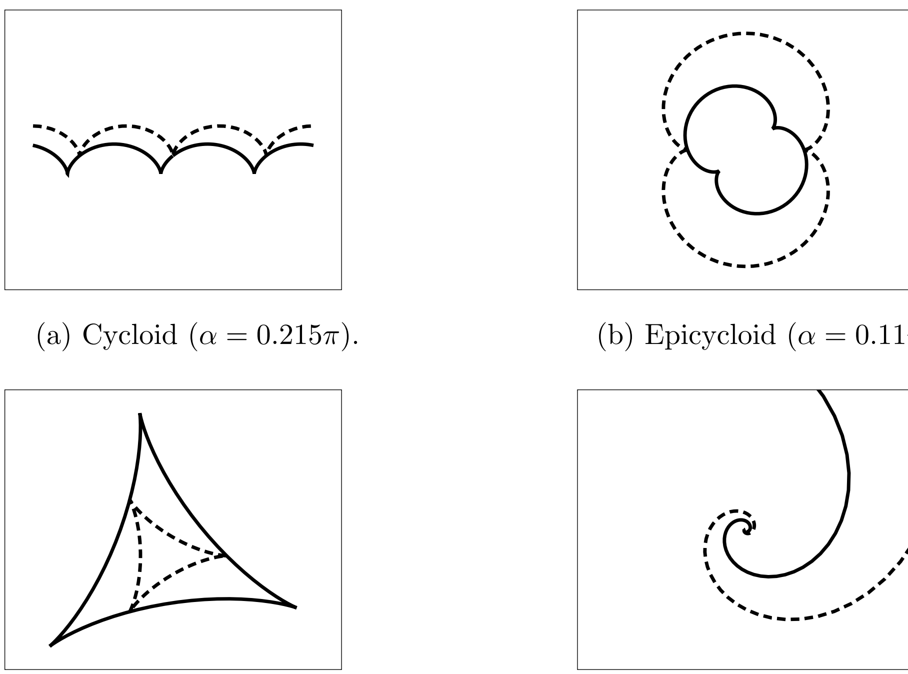


(a) An ellipse γ and its evolute E.

(b) An ellipse γ and an evolutoid E_α.

Figure 4.8: An ellipse, its evolute, and its evolutoid with $\alpha = 0.8$. The plotted points ($t_0 = 0$) show that despite a bijection between cusps and points of extremal curvature, the parametrization of the evolutoid induced by γ does not respect this bijection.

where $I(L_1, L_2)$ denotes the intersection point of lines L_1, L_2. We call $E_\alpha(t)$ the **evolutoid** of γ with parameter α. We remark that this definition of the evolutoid is essentially a more verbose way of simply stating that the evolutoid is the envelope of the α-rotated normals of γ (see [39]). This notion is very old and has a complicated history which we summarize in §1.3. There exist several known classical examples of curves whose evolutes are similar to themselves [57, 63], i.e. curves γ such that $\mathrm{Im}(\gamma) = T(\mathrm{Im}(E))$, where T is a map which sends $x \mapsto \rho U x + b$, for $\rho \in \mathbb{R}$, $U \in O(2)$, and $b \in \mathbb{R}^2$. More concisely, these curves are similar to their evolutes. The classical examples are the cycloid, the epicycloid, the hypocycloid, and the logarithmic spiral (and these are shown in Fig. 4). The circle could be included in this list for trivial reasons. Remarkably, these curves also have the property that they are similar to all of their evolutoids.



(a) Cycloid ($\alpha = 0.215\pi$).

(b) Epicycloid ($\alpha = 0.11\pi$).

(c) Hypocycloid ($\alpha = 0.25\pi$).

(d) Logarithmic spiral ($\alpha = 0.25\pi$).

Figure 4.9: Theorem illustrated: all evolutoids (one shown in solid) of curves (dashed) are similar to the original curve.

Our main result states that this property hold in general.

Theorem (Main Result)**.** *Suppose γ is a piecewise smooth periodic closed curve whose evolute E is similar to γ itself. Then each evolutoid E_α is also similar to γ.*

It has already been discussed in the 1920s that there exist infinitely many families of such curves [34]; however, explicit descriptions of these curves is tricky, as they require integrating the local inverse of a non-injective function. Since this task has become a lot easier with the advent of computers, we can characterize such curves, numerically approximate them and their evolutoids and show some examples, this is done in §3.

Related results.

The history of the evolutoid is complicated. The evolute was known since Apollonius, and it took almost two thousand years for the key idea of normals to be extended by Réaumur to rotated normals. Hamann [47] studied evolutes and evolutoids of ovals (C^2-smooth strictly convex curves). Yoshida and Saito [95] showed that the evolute of a log-aesthetic curve with parameter α is another log-aesthetic curve with parameter $-1/(\alpha-2)$. Note that this implies that the logarithmic spiral (which is log-aesthetic with parameter 1) is the unique log-aesthetic curve that is similar to its evolute. Izumiya and Takeuchi [51] studied evolutoids and their relations to pedaloids (a similar generalization of pedal curves) in the context of frontals. In particular, they establish a certain identity between evolutoids and pedaloids. Reznik, Garcia, and Stachel [72] investigated area (and signed area) invariances of a number of different curves including the evolute and evolutoid of the ellipse. Giblin and Warder [39] also studied evolutoids with a particular focus on cusps and wavefronts. Zwierzyński [97] studied sets of singularities/cusps of all evolutoids of a given curve. Similarly, Cambraia and Lemos considered the evolution and appearance of singularities as the angle parameter of the evolutoid changes. Apostol and Mnatsakanian [8] explored the analogous object of generalizing involutes, which they called tanvolutes. In fact, they showed that a curve is the α-evolutoid of any of its α-tanvolutes. We would have expected the word "involutoid" to appear somewhere. It turns out the word does seem to exist in the context of gear design [30] and as a generalization of tanvolutes to Legendrian curves [60]. Aguilar-Arteaga et l. [3] focus particularly on certain properties of convex curve evolutoids, including Steiner points.

4.2.2 Proofs of the Results

The circle, part I

We begin by introducing a parametrization of the evolutoid. This can be done in multiple ways, e.g. solving a system of Serret-Frenet equations, as in [39]. Our method is very elementary and uses the description in §1.2 as well as the fact that smooth curves with non-vanishing curvature are locally circular. Indeed, our approach will be to study the circle and understand its evolute entirely (Lemma 1 and Corollary 1), and then treat generic curves as approximately circular to explicitly describe their evolutoids.

Lemma 15. *Let $\gamma(t) = (\cos t, \sin t)$. Then $E_\alpha(0) = \frac{1}{2} \cdot (1 - \cos(2\alpha), -\sin(2\alpha))$.*

Proof. Now and henceforth, we will use the abbreviation

$$T_\alpha = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

to denote rotation matrices. Recall that the evolute of a circle is simply its center, so $E(t) = (0, 0)$. Then we can compute $E_\alpha(0)$. Parameterize the line which is an α counterclockwise rotation to the normal of γ at $t = 0$ by

$$L^\alpha_{\gamma,0}(s) = (1, 0) + T_\alpha \cdot (-1, 0) \cdot s,$$

and similarly for $t = \varepsilon$ by

$$L^\alpha_{\gamma,\varepsilon}(s) = (\cos\varepsilon, \sin\varepsilon) + T_\alpha \cdot (-\cos\varepsilon, -\sin\varepsilon) \cdot s,$$

where T_α denotes counterclockwise rotation by α (see Figure 4.10). One can compute that the intersection of these two lines is

$$\sin(\alpha)\cdot(\sin(\alpha)+\cos(\alpha)\tan(\varepsilon/2),-\cos(\alpha)+\sin(\alpha)\tan(\varepsilon/2))\,,$$

and therefore, sending $\varepsilon \to 0$, we have that

$$E_\alpha(0)=(\sin^2(\alpha),-\cos(\alpha)\sin(\alpha))=\frac{1}{2}\cdot(1-\cos(2\alpha),-\sin(2\alpha))\,.$$

$\square$

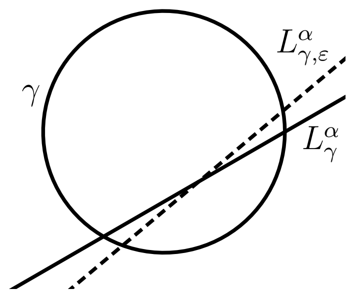


Figure 4.10: Lines $L^{\alpha}_{\gamma,0}$ and $L^{\alpha}_{\gamma,\varepsilon}$ in the proof of Lemma 15.

Note that as α varies from 0 to π, $E_\alpha(0)$ travels around a circle of radius 0.5 centered at $(0.5,0)$. In other words,

$$E_\alpha(0)=(1,0)-(0.5,0)-T_{2\alpha}(0.5,0)=\gamma(0)+\frac{1}{2}\vec{N}(0)+\frac{1}{2}T_{2\alpha}\vec{N}(0),$$

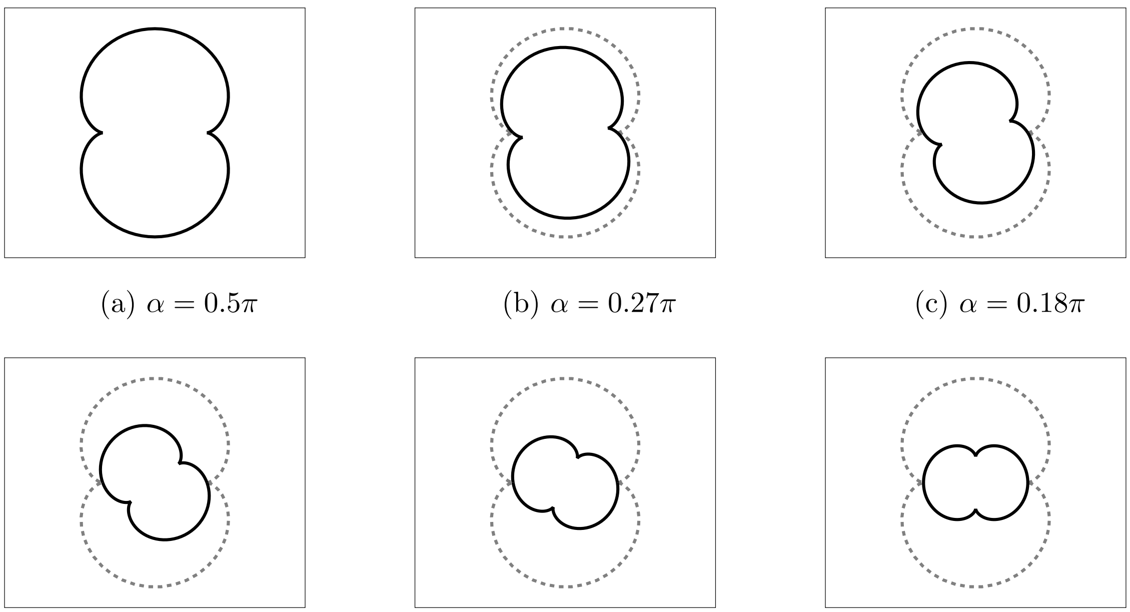
where $\vec{N}$ denotes the unit normal vector to γ pointing toward the center of curvature and κ denotes its curvature.

The circle, part II

Given the behavior of a particular point on the evolutoid of the unit circle, rotational symmetry gives us the full picture of the unit circle. In particular, with respect to α, any point of the evolutoid of the unit circle traces out a circle of radius $\frac{1}{2}$.

Corollary. *If $\gamma(t)=(\cos t,\sin t)$, let $\vec{N}$ denote the unit normal vector to γ pointing toward* *the center of curvature. Then*

$$\begin{aligned} E_\alpha(t) &= \gamma(t) + \frac{1}{2}\vec{N}(t) + \frac{1}{2}T_{2\alpha}\vec{N}(t) \\ &= \sin(\alpha)\cdot(\sin(\alpha+t), -\cos(\alpha+t)) \,. \end{aligned}$$



(a) $\alpha = 0.5\pi$ (b) $\alpha = 0.27\pi$ (c) $\alpha = 0.18\pi$

(d) $\alpha = 0.11\pi$ (e) $\alpha = 0.05\pi$ (f) $\alpha = 0.0\pi$

Figure 4.11: A nephroid (gray, dashed), i.e. an epicycloid with $R = 2, r = 1$, along with its evolutoids (black).

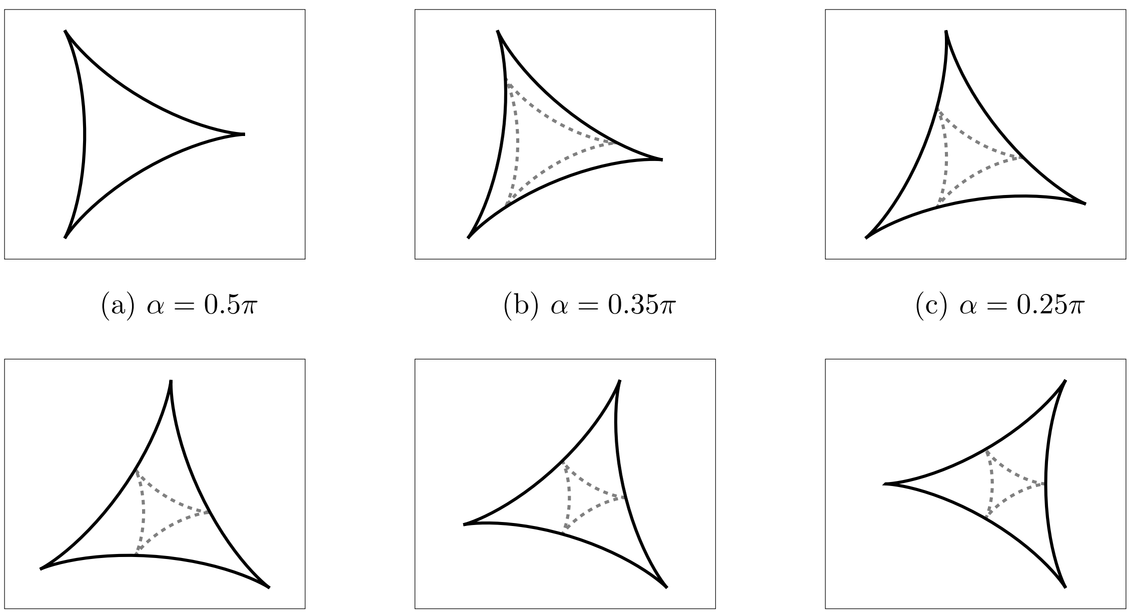
Proof. By rotational symmetry, the behavior of $E_\alpha(t)$ can be obtained by simply rotating $E_\alpha(0)$ by t, i.e.

$$E_\alpha(t) = \frac{1}{2}T_t\left(1-\cos(2\alpha), -\sin(2\alpha)\right).$$

Expanding the rotation gives

$$\begin{aligned} E_\alpha(t) &= \frac{1}{2}\left(\cos(t) - \cos(2\alpha - t), \sin(t) - \sin(2\alpha - t)\right) \\ &= (\cos(t), \sin(t) - \frac{1}{2}(\cos(t), \sin(t)) - \frac{1}{2}(\cos(2\alpha - t), \sin(2\alpha - t)) \\ &= \gamma(t) + \frac{1}{2}\vec{N}(t) + \frac{1}{2}T_{2\alpha}\vec{N}(t), \end{aligned}$$

This can also be written as $\sin(\alpha) \cdot (\sin(\alpha + t), -\cos(\alpha + t))$. $\square$



(a) $\alpha = 0.5\pi$ (b) $\alpha = 0.35\pi$ (c) $\alpha = 0.25\pi$

(d) $\alpha = 0.16\pi$ (e) $\alpha = 0.08\pi$ (f) $\alpha = 0.0\pi$

Figure 4.12: A deltoid (gray, dashed), i.e. a hypocycloid with $R = 3$, $r = 1$, along with its evolutoids (black).

While the final expression using only trigonometric functions, t, and α is perhaps most direct, the previous expression using γ and $\vec{N}$ turns out to be more useful. For a generic circle γ

with radius r, we would have

$$\begin{aligned} E_\alpha(t) &= \gamma(t) + r \cdot \frac{1}{2}\vec{N}(t) + r \cdot \frac{1}{2}T_{2\alpha}\vec{N}(t) \\ &= \gamma(t) + \frac{1}{2}\frac{\vec{N}(t)}{\kappa} + \frac{1}{2}T_{2\alpha}\frac{\vec{N}(t)}{\kappa}, \end{aligned}$$

where κ denotes the curvature of γ (which in the case of a circle, is of course constant). Notice that for $\alpha = 0$, this recovers one definition of the evolute, which can be written as

$$E(t) = \gamma(t) + \frac{\vec{N}(t)}{\kappa(t)}.$$

Evolutoids trace out circles.

We are now able to describe how the evolutoid behaves for any curve. The evolutoid of γ with $\alpha = 0$ is, by definition, the evolute of γ. Intuitively, the evolutoid of γ with $\alpha = \pi/2$ should recover the curve γ itself. Then one could conjecture that as α varies continuously, the evolutoid describes a continuous deformation from γ to its evolute. In fact, this is exactly what happens, so long as γ does not have vanishing second derivative. Moreover, the pointwise behavior of the evolutoid is remarkably simple.

Proposition 13. *Suppose γ is a curve with finite non-zero curvature. Then for any t in the domain of γ, the trajectory of E_α is the unique circle tangent to γ at t and passes through $E(t)$.*

$$E_\alpha(t) = \sin(\alpha) \cdot T_{a-\pi/2} \cdot \gamma(t) + \cos(\alpha) \cdot T_\alpha \cdot E(t). \tag{4.1}$$

Despite having a simple pointwise behavior, the global behavior of the evolutoid can be unintuitive. The evolute often produces cusps even when the original curve γ is smooth. These cusps arise directly from points of locally extremal curvature. In particular, if $\gamma(t_0)$ is a local minimum or maximum in curvature, then $E(t_0)$ is a cusp (CITE). The evolutoid also produces cusps, but while they might still correspond to points of extremal curvature, it is not true that $E_\alpha(t_0)$ is a cusp.

Proof of Proposition 13. Because the osculating circle of γ is a second order approximation, we treat γ as locally circular. Therefore,

$$E_\alpha(t) = \gamma(t) + \frac{1}{2} \cdot \frac{\vec{N}(t)}{\kappa(t)} + \frac{1}{2} T_{2\alpha} \frac{\vec{N}(t)}{\kappa(t)},$$

which can be rewritten and grouped as

$$\begin{aligned} E_\alpha(t) &= \frac{1}{2}\gamma(t) + \frac{1}{2}\gamma(t) + \frac{1}{2} \cdot \frac{\vec{N}(t)}{\kappa(t)} - \frac{1}{2} T_{2\alpha}\gamma(t) + \frac{1}{2} T_{2\alpha}\gamma(t) + \frac{1}{2} T_{2\alpha} \frac{\vec{N}(t)}{\kappa(t)} \\ &= \frac{1}{2}\gamma(t) + \left(\frac{1}{2}\gamma(t) + \frac{1}{2} \cdot \frac{\vec{N}(t)}{\kappa(t)}\right) - \frac{1}{2} T_{2\alpha}\gamma(t) + \left(\frac{1}{2} T_{2\alpha}\gamma(t) + \frac{1}{2} T_{2\alpha} \frac{\vec{N}(t)}{\kappa(t)}\right), \end{aligned}$$

which is just $\frac{1}{2}\left[\gamma(t) - T_{2\alpha}\gamma(t) + E(t) + T_{2\alpha}E(t)\right]$. One can check that for $\alpha \in [0, \pi]$, using I_2 denotes the identity map from $\mathbb{R}^2$ to itself,

$$I_2 - T_{2\alpha} = 2\sin(\alpha) \cdot T_{\alpha - \pi/2}, \quad \text{and} \quad I_2 + T_{2\alpha} = 2\cos(\alpha) \cdot T_\alpha.$$

Thus, we see that

$$E_\alpha(t) = \sin(\alpha) \cdot T_{\alpha-\pi/2} \cdot \gamma(t) + \cos(\alpha) \cdot T_\alpha \cdot E(t).$$

$\square$

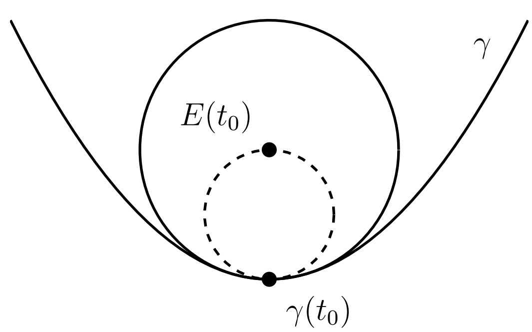


Figure 4.13: An example of Proposition 13: the curve γ, its osculating circle, and the α-trajectory of $E_\alpha(t_0)$ (dashed).

This expression of the evolutoid allow us to immediately see that for special values of α, E_α recovers the original curve and its evolute. In particular, for $\alpha = 0$ or $\alpha = \pi$, $E_\alpha(t) = E(t)$, and for $\alpha = \pi/2$, $E_\alpha(t) = \gamma(t)$. Moreover, so long as γ has finite non-zero curvature, E_α provides a continuous deformation from γ to its evolute. Moreover, if γ has finite non-zero curvature, then $H(t, \alpha_0) = E_{\pi\alpha_0/2}(t)$ is a homotopy between γ and E.

Intrinsic Equations

Up to this point, our study of evolutoids has dealt directly with coordinates. One could attempt to prove the theorem in coordinates, but this is, at best, impractical and tedious. Instead, we can look at intrinsic properties of a curve, i.e. quantities including curvature, arc length, and tangential angle. From this point on, for some curve γ, we use R to denote the radius of curvature, ψ for the tangential angle, s for arc length, κ for curvature, and k for signed curvature, all of which are parametrized by t, the parameter of γ. Equations which relate these properties of a curve are called *intrinsic* or *natural* and are "coordinate-free", so to speak. One such equation relates curvature and arc length, i.e. $\kappa = f(s)$. It is discussed in [7, 89] that this relation allows one to express

$$x = \operatorname{sgn}(k) \int_{\psi_0}^{\psi} R\cos(\varphi)d\varphi \quad \text{and} \quad y = \operatorname{sgn}(k) \int_{\psi_0}^{\psi} R\sin(\varphi)d\varphi, \tag{4.2}$$

$$\text{with} \quad \varphi = \int_{u_0}^{u} \kappa dv, \tag{4.3}$$

which [7] calls 'The Fundamental Theorem of Plane Curves', as this implies that any two curves with the same curvature with respect to arc length are the same up to a Euclidean transformation. This can be directly read from the above equations, changing the constants of integration in Equation 4.2 gives a translation, and changing the constant of integration in Equation 4.3 gives a rotation. Note that the expressions in Equation 4.2 differ slightly from those found in [89] as their use of the symbols κ, R differs from ours in that they allow κ, R to be negative. In particular, our symbol k is their κ. Incorporating a factor of $\operatorname{sgn}(k)$ accounts for this difference. Intuitively, the quantity $\operatorname{sgn}(k)$ simply captures whether a curve is turning counterclockwise (positive) or clockwise (negative).

One can verify that if γ_1 has an intrinsic equation given by $\kappa = f(s)$, and γ_2 has intrinsic equation $\kappa = af(s)$, then γ_1 and γ_2 are similar curves. In English, this means that a curve is uniquely determined up to Euclidean transformation once its curvature (with respect to some parametrization) is fixed. Similarly, the same statement holds if we instead use an intrinsic equation of the form $R = f(\varphi)$, as $R = \kappa^{-1}$, and we parametrize R (locally) with tangential angle. This fact allows us to check for similarity between two curves.

Rewriting the evolute

With the previously established notation, the definition of the evolute can be expressed as

$$E(t) = \gamma(t) + R(t) \cdot \vec{N}(t),$$

where $\vec{N}(t)$ denotes the unit vector pointing from $\gamma(t)$ to the center of the osculating circle to γ at t. If we use $\vec{T}(t)$ to denote the unit tangent vector to γ, then

$$\vec{T}(t) = \begin{pmatrix} \cos(\psi(t)) \\ \sin(\psi(t)) \end{pmatrix} \quad \text{and} \quad \vec{N}(t) = \pm \begin{pmatrix} -\sin(\psi(t)) \\ \cos(\psi(t)) \end{pmatrix} = \pm T_{\pi/2}\vec{T},$$

where the $\pm$ depends on the signed curvature of γ. Note that

$$\frac{d\vec{T}}{d\psi} = \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix} = T_{\pi/2}\vec{T},$$

and therefore,

$$\frac{d\vec{T}}{d\psi} = \operatorname{sgn}(k)\vec{N} \Rightarrow \vec{N} = \operatorname{sgn}(k) T_{\pi/2}\vec{T} \tag{4.4}$$

which in turn means that

$$\frac{d\vec{N}}{d\psi} = \operatorname{sgn}(k) T_{\pi/2}\frac{d\vec{T}}{d\psi} = \operatorname{sgn}(k) T_{\pi/2}\left(\operatorname{sgn}(k)\vec{N}\right) = T_{\pi/2}\vec{N} = -\operatorname{sgn}(k)\vec{T}. \tag{4.5}$$

This allows us to express the evolute entirely in terms of intrinsic parameters, i.e.

$$E = \gamma + R \cdot \operatorname{sgn}(k) \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix} \tag{4.6}$$

A Curvature Lemma.

The following lemma is, as far as we are aware, new, but is derived from the well known statement that the arc length of the evolute between two points equals the difference between the radii of curvature at those two points. Proofs of this fact can be found in [62, 77, 88]. One version of this lemma appears in [61] with no discussion or proof, so we provide one here.

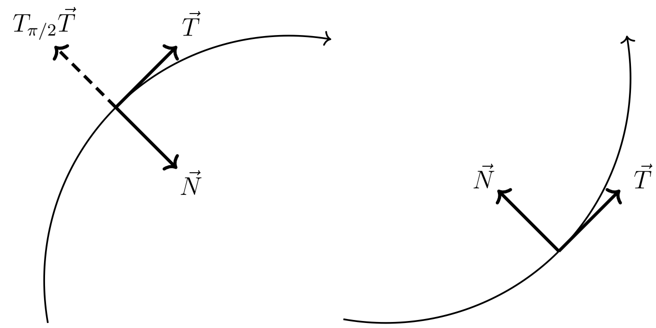


Figure 4.14: A visualization of Equation 4.1. $\vec{T}$ and $\vec{N}$ for a curve with negative signed curvature (left) and a curve with positive signed curvature (right).

Lemma 16. *Let γ be a piecewise smooth curve with evolute E. Then the radius of curvature R_E of the evolute can be written as*

$$R_E = \frac{dR}{d\psi} \cdot \operatorname{sgn}\left(\frac{dR}{ds}\right) \cdot \operatorname{sgn}(k).$$

Proof. Recall that $R = \left|\frac{ds}{d\psi}\right| = \frac{ds}{d\psi} \cdot \operatorname{sgn}(k)$. Then from Equation 4.6,

$$E = \gamma + R \cdot \operatorname{sgn}(k) \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix}.$$

Then so long as the curvature of γ does not vanish, we can compute

$$\begin{aligned} \frac{dE}{ds} &= \frac{d\gamma}{ds} + \frac{dR}{ds} \cdot \operatorname{sgn}(k) \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix} + R \cdot \operatorname{sgn}(k) \cdot \begin{pmatrix} -\cos(\psi) \\ -\sin(\psi) \end{pmatrix} \cdot \frac{d\psi}{ds} \\ &= \vec{T} + \frac{dR}{ds} \cdot \operatorname{sgn}(k) \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix} - \vec{T} \cdot R \cdot \operatorname{sgn}(k) \cdot \frac{d\psi}{ds} = \pm \frac{dR}{ds} \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix}, \end{aligned}$$

and in particular, $\left|\frac{dE}{ds}\right| = \left|\frac{dR}{ds}\right|$. Then we can write the arc length of E as

$$s_E = \int_{s_0}^{s_1} \left|\frac{dE}{ds}\right| ds = \int_{s_0}^{s_1} \left|\frac{dR}{ds}\right| ds,$$

and so long as R does not change sign from s_0 to s_1,

$$= \begin{cases} R_1 - R_0 & \frac{dR}{ds} > 0 \\ R_0 - R_1 & \frac{dR}{ds} < 0 \end{cases} = (R_1 - R_0) \cdot \operatorname{sgn}\left(\frac{dR}{ds}\right)$$

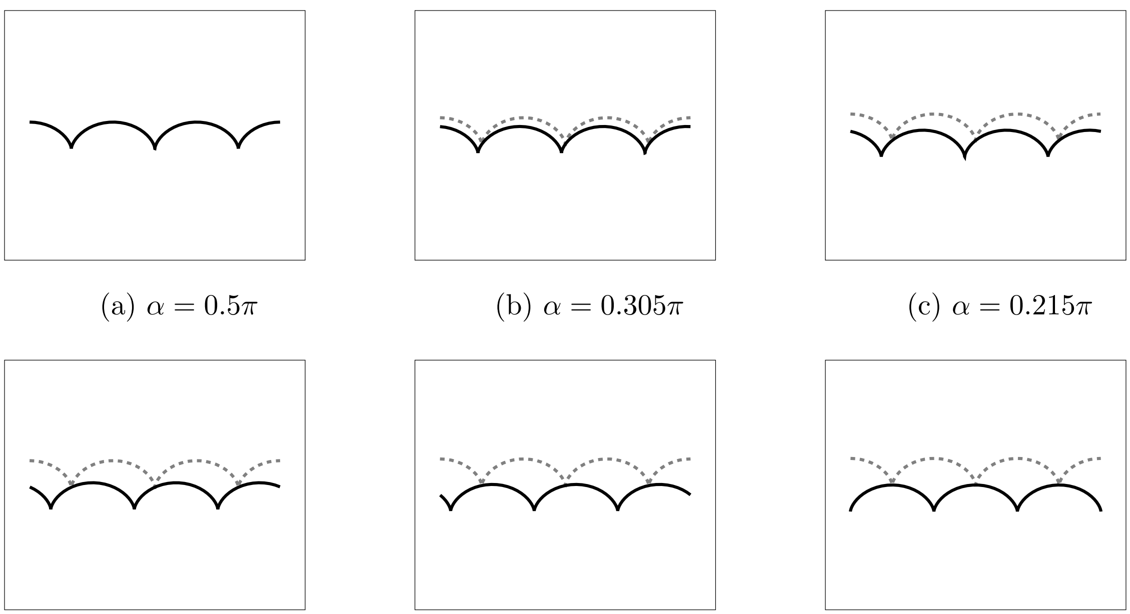
with respect to some fixed initial point where γ has curvature R_0. Therefore, we conclude that

$$R_E = \left|\frac{ds_E}{d\psi}\right| = \left|\frac{dR}{d\psi} \cdot \operatorname{sgn}\left(\frac{dR}{ds}\right)\right|,$$

where we note that $dR/d\psi$ and dR/ds have the same sign iff γ has positive signed curvature. Thus, we see that

$$R_E = \frac{dR}{d\psi} \cdot \operatorname{sgn}\left(\frac{dR}{ds}\right) \cdot \operatorname{sgn}(k),$$

□



(a) $\alpha = 0.5\pi$ (b) $\alpha = 0.305\pi$ (c) $\alpha = 0.215\pi$

(d) $\alpha = 0.135\pi$ (e) $\alpha = 0.065\pi$ (f) $\alpha = 0.0\pi$

Figure 4.15: A cycloid (gray, dashed) and its evolutoids (black).

Evolutoid angle shift

The following lemma can be viewed as immediately following the definition of the evolutoid as an envelope, but we provide a proof of this fact based on Proposition 13.

Lemma 17. *If the tangent line to γ at $\gamma(t)$ forms an angle of ψ with the x-axis, then the tangent line to E_α at $E_\alpha(t)$ forms an angle of $\psi + \alpha - \frac{\pi}{2}$ with the x-axis.*

Proof. Suppose that γ is parametrized by its arc length, i.e. we can replace t with s. Then as we saw in the proof of Lemma 16,

$$\frac{dE}{ds} = \operatorname{sgn}(k) \cdot \frac{dR}{ds} \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix}.$$

Then since

$$E_\alpha(t) = T_\alpha \left[\sin(\alpha) \cdot T_{-\pi/2}\gamma(t) + \cos(\alpha) \cdot E(t)\right],$$

we have

$$\frac{dE_\alpha}{ds} = T_\alpha \left[\sin(\alpha) \cdot T_{-\pi/2}\frac{d\gamma}{ds} + \cos(\alpha) \cdot \operatorname{sgn}(k) \cdot \frac{dR}{ds} \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix}\right].$$

Remembering that $\frac{d\gamma}{ds} = \binom{\cos(\varphi)}{\sin(\varphi)}$ give us that

$$\frac{dE_\alpha}{ds} = T_\alpha \left[\sin(\alpha) \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix} + \cos(\alpha) \cdot \operatorname{sgn}(k) \cdot \frac{dR}{ds} \cdot \begin{pmatrix} -\sin(\psi) \\ \cos(\psi) \end{pmatrix}\right],$$

and combining terms appropriately yields

$$\frac{dE_\alpha}{ds} = \left(\sin(\alpha) + \cos(\alpha) \cdot \operatorname{sgn}(k) \cdot \frac{dR}{ds}\right) T_{\alpha-\pi/2}\frac{d\gamma}{ds}.$$

Since $\frac{d\gamma}{ds}$ is the unit tangent vector of γ, and $\frac{dE_\alpha}{ds}$ is a tangent vector of E_α, we are done. $\square$

Proof of Theorem

Before we prove the Theorem, we establish one last fact. Equation 4.2 above implies

$$\frac{dx}{d\psi} = \operatorname{sgn}(k) R \cos(\psi), \quad \text{and} \quad \frac{dy}{d\psi} = \operatorname{sgn}(k) R \sin(\psi).$$

In other words,

$$\frac{d\gamma}{d\psi} = \operatorname{sgn}(k) R \cdot \vec{T}. \tag{4.7}$$

Note that this looks similar to the classical statement $R = |ds/d\psi|$, with a key difference of γ instead of s. There is, in fact, no contradiction here, as again we take all intrinsic quantities to be parametrized by some consistent parameter.

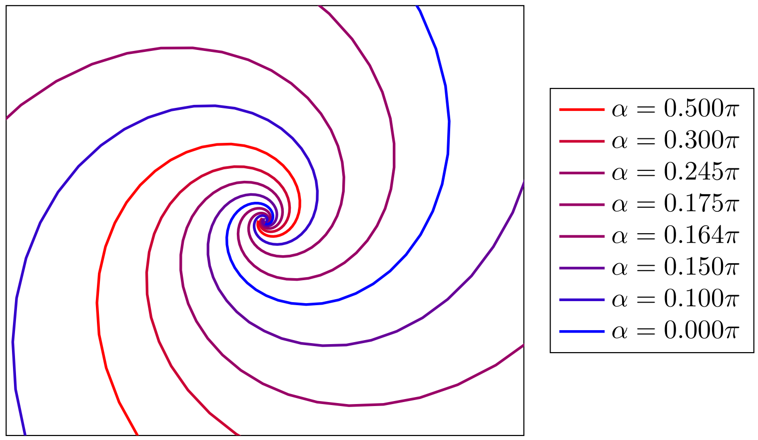


Figure 4.16: A logarithmic spiral (in red) along with its evolutoids (shades between red and blue).

Proof of Theorem. We would like to show that $R_\alpha(t)$, the radius of curvature of E_α, is a constant multiple of $R(t)$. To be precise, $R(t)$ denotes the radius of curvature of γ at t according to some parametrization of γ, and $R_\alpha(t)$ refers to the radius of curvature of E_α at t according to the parametrization induced by γ, i.e. Equation 4.1. Using Proposition 13, we write the evolutoid as

$$\begin{aligned} E_\alpha &= T_\alpha \left[\sin(\alpha) \cdot T_{-\pi/2}\gamma + \cos(\alpha) \cdot E\right]. \\ &= T_\alpha \left[\sin(\alpha) \cdot T_{-\pi/2}\gamma + \cos(\alpha) \cdot \left(\gamma + R \cdot \vec{N}\right)\right]. \end{aligned}$$

Because of Equation 4.7, we will take the derivative $\frac{dE_\alpha}{d\psi_\alpha}$ to extract R_α. Recall that $\psi_\alpha = \psi + \alpha - \frac{\pi}{2}$. Then taking a derivative with respect to ψ_α is the same as taking a derivative with respect to ψ, so

$$\frac{dE_\alpha}{d\psi_\alpha} = \frac{dE_\alpha}{d\psi} = T_\alpha \left[\sin(\alpha) \cdot T_{-\pi/2}\frac{d\gamma}{d\psi} + \cos(\alpha) \cdot \frac{dE}{d\psi}\right].$$

Isolating the evolute derivative, we have

$$\frac{dE}{d\psi} = \frac{d\gamma}{d\psi} + \frac{dR}{d\psi} \cdot \vec{N} + R \cdot \frac{d\vec{N}}{d\psi}$$

which, using Equations 4.5 and 4.7 for derivatives of $\gamma, \vec{N}$ with respect to ψ, gives

$$\frac{dE}{d\psi} = \operatorname{sgn}(k)R \cdot \vec{T} + \frac{dR}{d\psi} \cdot \vec{N} - \operatorname{sgn}(k)R \cdot \vec{T} = \frac{dR}{d\psi} \cdot \vec{N}.$$

Plugging this back in and using Equation 4.7 once more gives

$$\frac{dE_\alpha}{d\psi_\alpha} = T_\alpha \left[\operatorname{sgn}(k)R\sin(\alpha) \cdot T_{-\pi/2}\vec{T} + \cos(\alpha) \cdot \frac{dR}{d\psi} \cdot \vec{N} \right].$$

Applying Equation 4.4 and grouping everything together, we have

$$\begin{aligned} \frac{dE_\alpha}{d\psi_\alpha} &= T_\alpha \left[\operatorname{sgn}(k)R\sin(\alpha) \cdot T_{-\pi/2}\vec{T} + \cos(\alpha)\frac{dR}{d\psi} \cdot \vec{N} \right] \\ &= T_\alpha \left[\operatorname{sgn}(k)R\sin(\alpha) \cdot \vec{N} \cdot (-\operatorname{sgn}(k)) + \cos(\alpha)\frac{dR}{d\psi} \cdot \vec{N} \right] \\ &= \left(-R\sin(\alpha) + \frac{dR}{d\psi} \cdot \cos(\alpha) \right) \cdot T_\alpha\vec{N}. \end{aligned}$$

Therefore, by Equation 4.7, we have that

$$R_\alpha = \left| \frac{dE_\alpha}{d\psi_\alpha} \right| = \left| -R\sin(\alpha) + \frac{dR}{d\psi} \cdot \cos(\alpha) \right|.$$

Since γ and E are similar, if R can be written as a function of tangential angle ψ, with $R = f(\psi)$, then $R_E = af(\psi + \beta)$, where a (constant) is the ratio of similitude, and β (constant) is a shift. Franklin [34] and Light [61] already discussed this and suggested that f must be of the form

$$f(\psi) = \sum_j C_j e^{m_j \psi}.$$

with $m_j = ae^{m_j\beta}$ to ensure the differential equation in the statement of Lemma 16 holds. We formalize this line of reasoning as follows. If R is a function of ψ, we can use Lemma 16 to see that

$$\frac{dR}{d\psi} = f'(\psi) = R_E \cdot \operatorname{sgn}\left(\frac{dR}{ds}\right) \cdot \operatorname{sgn}(k) = \pm af(\psi + \beta).$$

Therefore, we are interested in solving the delayed differential equation

$$f'(\psi) = \pm a f(\psi + \beta), \tag{4.8}$$

where the sign depends on the sign of dR/ds. Since f is periodic, write

$$f(\psi) = \sum_{l \in \mathbb{Z}} a_l e^{il\psi} \Rightarrow f'(\psi) = \sum_{l \in \mathbb{Z}} il a_l e^{il\psi}.$$

Then

$$\sum_{l \in \mathbb{Z}} il a_l e^{il\psi} = \pm a \sum_{l \in \mathbb{Z}} a_l e^{il(\psi+\beta)} = \pm a \sum_{l \in \mathbb{Z}} a_l e^{il\beta} e^{il\psi}$$

implies that either $a_l = 0$ or $il = \pm a e^{il\beta}$. In the latter case, we see that $|a| = |l|$ and β is a rational multiple of π, which enforces that

$$f(\psi) = z e^{ia\psi} + \overline{z} e^{-ia\psi} = C \cos(a\psi),$$

for some constant C. We see that the presence of sign terms in the equation of Lemma 16 serves to correct the sign issue of f which we see above. Then we have

$$R_\alpha = |-C\cos(a\psi)\sin(\alpha) - aC\sin(a\psi)\cos(\alpha)|,$$

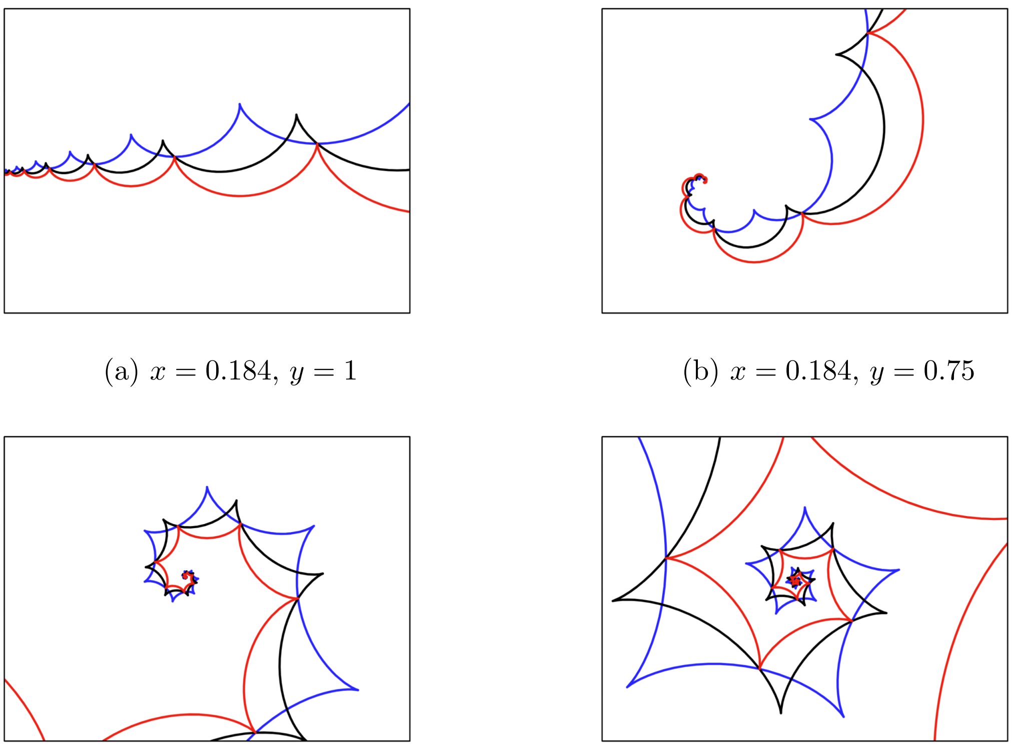
which by phasor addition is also of the form $D\cos(a\psi + \theta)$, for some constants D, θ. Thus, we see that R_α is, up to an argument shift, a constant multiple of R. □

4.2.3 New Examples

This leads naturally to the question of which curves have evolutes similar to themselves. One might be tempted to believe that this is a very rare property – which is, in some sense, true. However, already more than a hundred years ago Franklin [34] and Light [61] showed with an ansatz that one could use this machinery to produce new examples. Their investigation is purely theoretical as they neglected to actually produce examples. The procedure to generate new examples is straightforward. Observe that the function

$$f(\psi) = Ce^{x\psi}\cos(y\psi), \tag{4.9}$$

for real constants C, x, y, solves the delayed differential equation discussed in the proof of the theorem (with appropriate constants a, β). Integrate Equation 4.2 with $R = f$. We show in Figure 4.17 some examples of new curves which satisfy the condition that they are similar to their respective evolutes as well as their evolutoids.



(a) $x = 0.184$, $y = 1$

(b) $x = 0.184$, $y = 0.75$

(c) $x = 0.184$, $y = 1.44$

(d) $x = 0.184$, $y = 1.75$

Figure 4.17: New curves whose evolutes and evolutoids are similar. Original curves in red, evolute in blue, and evolutoids in black. Parameters provided for Equation 4.9

4.2.4 Concluding Remarks

We discovered recently that many of the ideas discussed in this chapter have been explored in [55], but our work was done independently. Their perspective, however, is slightly different. In particular, they do not seem to address the relationship between a curve being similar to its evolute and being similar to its evolutoids. Instead, they directly characterize the condition that a curve is similar to its evolute or evolutoids without relating the two. We would have liked to prove the general statement that a curve which is similar to its evolute is also similar to all its evolutoids. Doing so requires characterizing all solutions to the delayed differential equation 4.8 which is nontrivial. Instead, we restricted our attention to curves that had periodic curvature in order to be able to use the Fourier series expansion which is used by Franklin [34] as an ansatz. The ideas used [55] would likely be able to extend our result to other cases.

Chapter 5

JUNIPER GREEN

5.1 Introduction

5.1.1 History.

Juniper Green is a simple combinatorial game that was originally invented by Rob Porteous to teach children about multiplication and division (Juniper Green is the name of the school where Porteous taught). Porteous explained it to his father, the mathematician Ian Porteous, who in turn explained it to Ian Stewart, who then popularized it in a 1997 article [85] in *Scientific American* and further described in his 2010 book [87]. The history may be more complicated, Paul Blatz [86] recalls that the game was discussed in a number theory course given at Princeton University by Eugene Wigner in the late 1930s. The game has since also been discussed in the pedagogical literature [18, 79, 91].

5.1.2 Rules.

The game involves two players and is played on $[n] = \{1, 2, \ldots, n\}$ subject to the following rules.

1. Players take turns selecting numbers from $[n]$ such that each selected number must be a factor or multiple of the previously selected number.
2. Once a number has been used, it cannot be selected again.
3. A player loses if they have no legal moves.

One can check that the first player has a straightforward winning strategy by picking a prime p greater than $n/2$. This forces a response of 1, allowing player one to end the game by picking a second prime q also greater than $n/2$. This requires only the existence of at least 2 distinct primes $p, q \in (n/2, n)$. For sufficiently large n, this is guaranteed to be the case. Because of this "boring" strategy, it is standard to introduce the additional rule.

(4) The first move must be even[1].

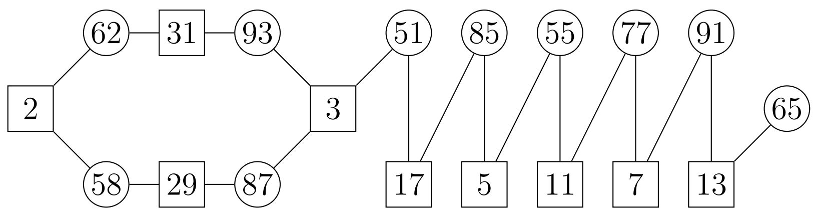


Figure 5.1: An induced subgraph of the divisibility graph for $n = 100$ on the vertices relevant to a particular winning strategy.

Stewart described this game in [85, 87] and for $n = 100$, describes how the first player can win with 58 or 62 as the opening move. A complete description of the possible sequence of moves in that case is shown in Figure 5.1. It is unclear whether this particular strategy for $n = 100$ can generalize. Stewart's article finishes with the following question.

> What about [...] completely general n? Can anyone find patterns? Or solve the whole thing? (Ian Stewart [85])

5.1.3 Lemoine's Theorem

Stewart's question was completely answered by Julien Lemoine in a 2022 paper.

Theorem (Lemoine [59])**.** *For any $n \geq 120$, the first player has a winning strategy.*

The argument relies on the fact that there always exist 3 primes in $[n/4, n/3]$ for n sufficiently large, a similar strategy for the first player can be generalized for arbitrarily large n, leaving

[1] One might reasonably ask why this modification is not to instead restrict the first move to a composite. We asked ourselves the same thing and have no satisfying answer.

only finitely many cases to directly analyze. Lemoine furthermore directly analyzes the finitely many remaining cases in $n \leq 120$ and tabulates which player can force a win by an exhaustive analysis.

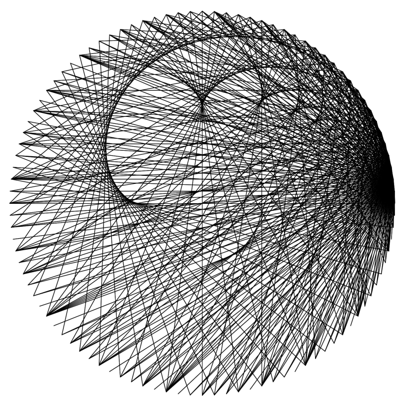


Figure 5.2: The divisibility graph on $[n]$ for $n = 100$, where the vertices are arranged in a circle. In particular, vertex i is placed at $(\cos(2\pi i/n), \sin(2\pi i/n)) \in \mathbb{R}^2$. The curves that appear as envelopes of these edges are epicycloids.

5.2 Result

We were motivated by Stewart's question whether more can be said: is it possible to understand the structure of winning moves for each n? In this sense, we refer to Juniper Green as the game subject to rules (1) – (3) and not (4) with the understanding that starting with a large prime number is a winning move for player one (which was the reason for rule (4) in the first place). Our main result can be informally summarized by saying that winning moves in Juniper Green correspond to a specific set in the Gallai-Edmonds decomposition of the divisibility graph on $[n]$ (see Figure 5.2). We recall that the divisibility graph, which we denote G_n, is the undirected graph with vertices $V = \{1, 2, \ldots, n\}$, where an edge exists between i, j when $i|j$.

5.2.1 The Gallai-Edmonds decomposition

We first describe the main ingredient which was discovered, independently, by Tibor Gallai [36, 37] and Jack Edmonds [32]. Given a graph $G = (V, E)$, the Gallai-Edmonds decomposition is a partition of the vertex set into three disjoint sets

$$V = D(G) \cup A(G) \cup C(G),$$

as follows: a vertex is said to either be *essential* (if it is covered by every maximum matching in G) or *inessential*. The decomposition is now as follows:

1. $D(G) \subseteq V$ consists of all *inessential* vertices.

2. $A(G) \subseteq V$ contains all essential vertices with an inessential neighbor.

3. $C(G) \subseteq V$ contains all remaining essential vertices.

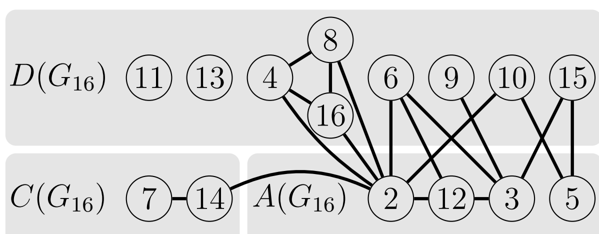


Figure 5.3: The Gallai-Edmonds Decomposition of G_{16} with 1 ($1 \in A(G_{16})$) omitted to reduce clutter. For n sufficiently large, 1 will always be a member of $A(G_n)$.

5.2.2 Main Result.

We can now state our main result.

Theorem (Main Result)**.** *The winning moves for player one in Juniper Green on* $[n]$ *are precisely the vertices in* $D(G_n)$ *of the Gallai-Edmonds decomposition of* G_n*.*

The result follows from Berge's Theorem [15, 16], see §3 for the details and a proof. As is often the case, it is difficult to answer a question without raising a new question; here, the new question evidently becomes

> What can be said about the Gallai-Edmonds decomposition of the divisibility graph on $[n]$?

As shown in Figure 5.4, it appears nontrivial to say very much at all. The result of Lemoine implies that $|D(G_n) \cap 2\mathbb{N}| \geq 1$ once $n \geq 120$. Computations using the blossom algorithm [53] to construct the decomposition suggest a most curious picture. It appears as if $|A(G_n)|$ is constantly small while $|C(G_n)|$ and $|D(G_n)|$ seem to dramatically alternate in size: these transitions seem to occur whenever n has many small prime factors (making it a particularly centrally vertex in the divisibility graph). This makes intuitive sense as illustrated with the following example. G_{101} does not differ much from G_{100}, as only one new vertex and one new edge are added; however, G_{100} differs very much from G_{99}, as from 99 to 100 adds many edges to the new vertex. That said, it remains unclear how one can make this precise. From Figure 5.5, there are a number of observations that can be made. It appears that the majority of the elements of $A(G_n)$ are less than $n/3$, and there are no members of $A(G_n)$ between $n/3$ and $n/2$. The interval $[n/2, 2n/3]$ seems to have the highest density of elements of $D(G_n)$. Vertical bands that seem to correspond to primes appear to belong either to $D(G_n)$ or $C(G_n)$ depending on n. Horizontal bands seem to indicate some kind of persistence a particular number has in terms of which part of the Gallai-Edmonds Decomposition it belongs to. It is easy to make further guesses, but it seems difficult to prove anything along these lines rigorously.

5.3 Proof of the Main Result

Juniper Green is a specific instance of the 'snake in the box' game on an undirected graph as described in [16]. Snake in the box is a two player combinatorial game subject to the following rules.

1. Players take turns selecting vertices of a graph G such that each selected vertex must be a neighbor of the previous vertex.

2. Once a vertex has been used, it cannot be selected again.

3. A player loses if they have no legal moves.

Notice that indeed, the above rules are entirely analogous to the rules described in §1. We can therefore understand Juniper Green as snake in the box played on the graph with vertex set $[n]$ and edge set $\{(i,j) : i|j\}$. In other words, Juniper Green is snake in the box played on the divisibility graph G_n.

Recall that a matching of a graph is a set of edges, no two of which share an edge. A maximum matching is simply a maximum cardinality matching. Note that this is not to be confused with a maximal matching, which is a matching M such that $M \cup \{e\}$ is not a matching for all $e \in E \setminus M$. Using maximum matchings, Berge provides a necessary and sufficient condition for the first player to have a winning strategy for snake in the box.

Theorem (Berge, [16])**.** *For snake in the box on a graph G, the first player has a winning strategy if and only if a maximum matching of G does not cover every vertex.*

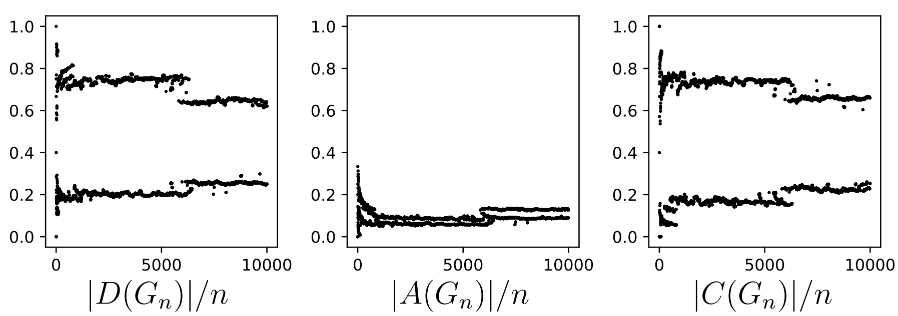


Figure 5.4: Proportions of $D(G_n)$, $A(G_n)$, and $C(G_n)$.

The proof of this theorem is quite elegant, using Berge's alternating chain lemma. The proof is important for our purposes as it provides as a corollary an exact set of moves the first player can make to force a win.

Lemma 18 (Alternating Chain Lemma)**.** *A matching M of a graph G is a maximum match-*

ing if and only if there are no two vertices uncovered by M that are connected by an alternating chain, i.e. a chain whose edges alternate between belonging to M and $E(G) \setminus M$.

Interestingly, the alternating chain lemma is sometimes called the alternating chain theorem or Berge's theorem (not to be confused with the above theorem). Its proof is slightly more involved, so for the interested reader, we refer to [16]. We now give Berge's proof characterizing which player wins snake in the box.

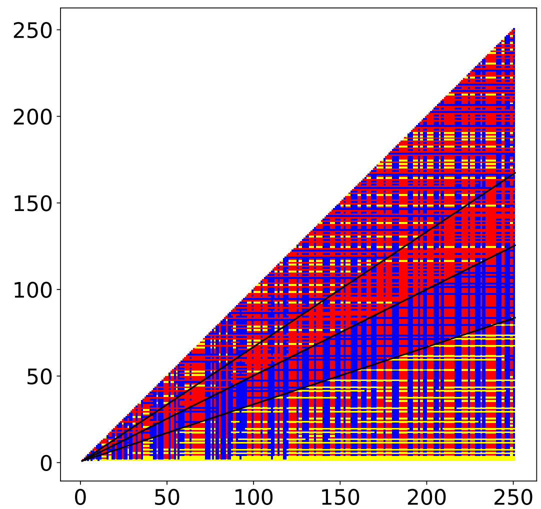


Figure 5.5: The elements of $[n]$ plotted against n colored by whether they belong to $D(G_n)$ (red), $A(G_n)$ (yellow), or $C(G_n)$ (blue). The black lines plotted are $n/3$, $n/2$, and $2n/3$.

Proof of Theorem. Let M be a maximum matching of G. If M does not cover a vertex v_1, the first player can select v_1. The second player's response must be covered by some edge $e = (v_2, v_3) \in M$, as otherwise, $M' = M \cup e$ would yield a larger matching. Suppose the second player chooses v_2. The first player now can choose the other vertex of e, namely v_3. From this point forward, the second player may not select a vertex uncovered by M, as this would violate the alternating chain lemma. Since the first player opened with a vertex not covered by M, the second player must always select a vertex u covered by $e' \in M$ such that if $e' = (u, u')$, u' has not yet been selected. This means that regardless of what the second player chooses, the first player will have an available move. Therefore, the first player cannot lose, i.e. player one wins.

Figure 5.6: If M is a maximum matching of G and $e \in M$, then any neighbor u of v_3 must also be covered by some edge of M, as otherwise, replacing e with (v_1, v_2) and (v_3, u) would yield a matching with higher cardinality.

If M covers every vertex of G, the second player has the following winning strategy. The first player chooses some vertex u, allowing the second player to respond with v, where $(u, v) \in M$. Since every vertex of G is covered by some edge of M, we see that the first player will always select some vertex u' such that if $e' = (u', v') \in M$ covers u', v' is available for the second player to choose.

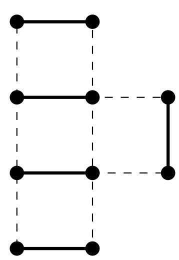


Figure 5.7: Whatever the first player selects, the second player can always respond by following the edges of a maximum matching M.

Therefore, the second player always has an available move by following M, so they cannot lose, i.e. player two wins. $\square$

Berge's proof gives us all the relevant tools to prove our result.

Proof of Main Result. Berge's proof tells us that the set of opening moves that the first player can take and win is precisely the set of inessential vertices, those elements $v \in V$ such that there exists a maximum matching M that does not cover v. Thus, since Juniper Green is snake in the box on G_n, we conclude that the set of winning openings for the first player is $D(G_n)$. $\square$

Chapter 6

RUPERT

6.1 Introduction

6.1.1 Introduction

According to John Wallis [94], Prince Rupert of the Rhine, whom Wallis calls *vir magno ingenio & sagacitate* [a man of great ingenuity and sagacity] posed the question (and wagered in favor of its answer being the affirmative) of whether a cube can "pass through itself", i.e. whether hole can be drilled into a cube such that an identical cube can be passed through that hole.

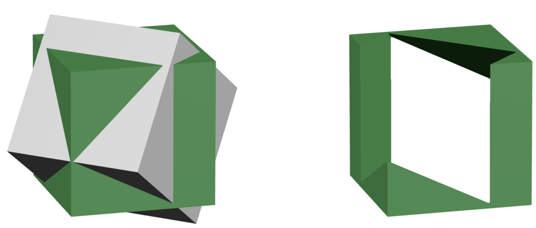


Figure 6.1: An illustration that the cube is Rupert.

Rupert's question was answered in the affirmative by Wallis himself, who showed that the flat projection of the cube into a square can fit inside the regular hexagonal projection of the cube along its space diagonal (see Figure 6.1). In fact, Wallis's construction allows a cube with side length $\sqrt{6} - \sqrt{2} \approx 1.03$ to pass through a hole drilled through a unit cube. Later, Peter Nieuwland (1764 – 1794) showed that an by using a slightly different projection, the same is true of a cube with side length $\sqrt{18}/4 \approx 1.06$ and that this improvement is optimal. Nieuwland himself never published this; it was published posthumously by Nieuwland's advisor in [93].

To make all of this precise, we introduce the following notation. We use $\mathcal{P}$ to denote a polyhedron which refers to the convex hull of a finite set of points in $\mathbb{R}^3$ in convex position. Moreover, let $\mathcal{P}^\circ$ denote the interior of $\mathcal{P}$. Let $P : \mathbb{R}^3 \to \mathbb{R}^2$ denote the projection which drops the x coordinate. Then Prince Rupert asked, do there exist $R_1, R_2 \in \mathrm{SO}(3)$ and $t \in \mathbb{R}^2$ such that

$$P \circ R_1(\mathcal{P}_C) + t \subset P \circ R_2(\mathcal{P}_C^\circ),$$

where $\mathcal{P}_C$ is the set of vertices of the cube. Note that the use of of the interior on the right side of the containment is important, as otherwise, we could trivially choose $R_1 = R_2$. We say that a polyhedron $\mathcal{P}$ has the *Rupert property* (or is *Rupert*) if it satisfies the above condition, and that the orientations R_1, R_2 are a *Rupert passage* for $\mathcal{P}$. Moreover, one can measure "how Rupert" a polyhedron $\mathcal{P}$ is by testing what the largest scaling $s\mathcal{P}$ is, that can still pass through $\mathcal{P}$. This notion is known as the *Nieuwland number* or *Nieuwland constant* (see [35, 84]) and is defined more precisely as the largest $s \in \mathbb{R}$ such that there exists $R_1, R_2 \in \mathrm{SO}(3)$ and $t \in \mathbb{R}^2$ for which

$$P \circ R_1(s\mathcal{P}) + t \subset P \circ R_2(\mathcal{P}^\circ)$$

holds. For instance, as discussed above, Nieuwland showed that the Nieuwland number of the cube is $\sqrt{18}/4$.

6.1.2 State of the Art

Currently, it is known that all 5 Platonic solids, 11 out of 13 Catalan solids, and 87 of the 92 Johnson solids have the Rupert property. Note that proving that a polyhedron is Rupert is a far easier task than showing that it is not. For the former, one need only exhibit two projections and verify containment, while for the latter, it is not immediately clear how one might even begin. The fact that so many nice solids (as well as other polyhedra) have the

Rupert property led to the following conjecture.

Conjecture (Jerrard-Wetzel-Yuan [52]). *All convex polyhedra are Rupert.*

Jerrard, Wetzel, and Yuan [52] state their conjecture "with a certain hesitancy". This caution seems to have been appropriate, as recently, Steininger and Yurkevich [84], constructed a convex polyhedron that cannot pass through itself.

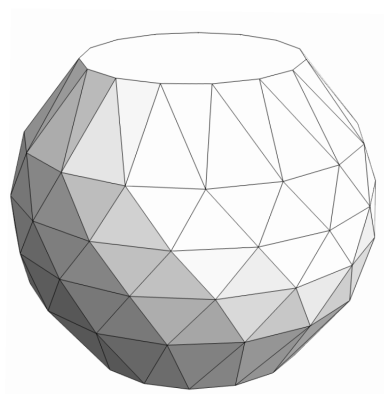


Figure 6.2: The Noperthedron of Steininger and Yurkevich.

Theorem (Steininger-Yurkevich [84]). *The Noperthedron is not Rupert.*

Their proof is extremely computational and relies on a cleverly constructed polyhedron (shown in Figure 6.2) with 90 vertices and $2\pi/15$ rotational symmetry, allowing them to reduce the size of their parameter space. In particular, their polyhedron is centrally symmetric, meaning that for every vertex v of the polyhedron, $-v$ is also a vertex. This condition effectively allows the removal of t in the above containment condition. Intuitively, the existence of the Noperthedron as a counterexample would suggest that there should exist many other potential non-Rupert polyhedra. Steininger and Yurkevich made the following conjecture [83] which, to the best of our knowledge, remains open.

Conjecture (Steininger-Yurkevich). *The rhombicosidodecahedron is not Rupert.*

The rhombicosidodecahedron is an Archimedean solid with 62 faces (20 triangles, 30 squares, and 12 pentagons), 60 vertices, and 120 edges. Our main motivation was that there should surely exist simpler counterexamples, i.e. polyhedra that are not Rupert and have many fewer vertices and a simpler face structure, leading us to ask the following.

Question. What is the 'simplest' polyhedron that is not Rupert?

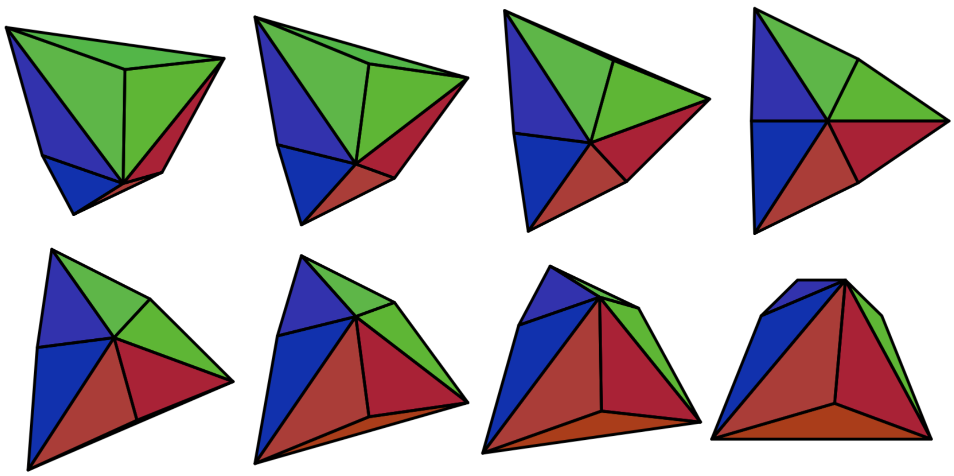


Figure 6.3: The stellated tetrahedron $\mathcal{P}_{11/20}$. Faces with similar colors share the same base tetrahedral face prior to stellation. We conjecture this convex polyhedron to not be Rupert.

6.1.3 The Stellated Tetrahedron

The Triakis Tetrahedron is a polyhedron with 8 vertices, 18 edges, and 12 faces. It can be obtained by starting with a regular tetrahedron and then adding an additional point for each of the 4 faces; this point has the property that its orthogonal projection onto face is exactly the centroid of the triangle (this process is known as 'stellation'). We stellate each face of a regular tetrahedron such that the resulting polyhedron has constant dihedral angle, i.e. the angle between any two adjacent faces is constant. Fredriksson showed in [35] that the Triakis Tetrahedron is Rupert with a lower bound for its Nieuwland number to be roughly $1 + 4 \times 10^{-6}$. The techniques used by Fredriksson likely ensure that this is not far off from the true Nieuwland number. This value is remarkably low which led us to consider similar polyhedra as potential non-Rupert candidates.

We consider the stellated tetrahedron $\mathcal{P}_a$ with vertices

$$\begin{aligned} p_0 &= (1,1,1), & q_0 &= (-a,-a,-a), \\ p_1 &= (1,-1,-1), & q_1 &= (-a,a,a), \\ p_2 &= (-1,1,-1), & q_2 &= (a,-a,a), \\ p_3 &= (-1,-1,1), & q_3 &= (a,a,-a). \end{aligned}$$

Edges are between all p_i, p_j and p_i, q_j for $i \neq j$, and $a \in (0.5, 0.57)$. Note that $a = 0.6$ corresponds to the Triakis tetrahedron. An algorithm implemented by Fredriksson [35] is capable of quickly verifying the Rupert property for many polyhedra, including several that were previously not known to be Rupert. For stellated tetrahedra with $a \in (0.5, 0.57)$, however, the code is unable to find a Rupert passage (see Figure 6.4). Scott, in [76], provides a sufficient condition for a polyhedron to be locally Rupert, which these stellated tetrahedra do not satisfy. We conjecture that these stellated tetrahedra are not Rupert.

Conjecture. *The stellated tetrahedron $\mathcal{P}_{11/20}$ is not Rupert.*

The choice of $a = 11/20$ is, admittedly, somewhat arbitrary. It is a nice enough rational number in the interval $(0.5, 0.57)$ which is convenient to type as input into a computer program. In support of this conjecture, we provide numerical evidence that the majority of pairs of orientation do not yield a Rupert passage.

The space of all pairs of orientations, that is $\mathrm{SO}(3) \times \mathrm{SO}(3)$, is a 6-dimensional manifold. Since we can encode an element of $\mathrm{SO}(3)$ using three angles α, θ, ϕ, we can define a parameter space Λ_0 as a 6-dimensional space, i.e. $[0, 2\pi]^6$ over which to search for Rupert passages. We discuss in §3 how Λ_0 can be reduced to $\Lambda = [0, 2\pi]^2 \times [0, \pi] \times [0, \pi/2]^2$. If we use the product measure on Λ, we can formulate the following statement.

Theorem. *There is no Rupert passage for $\mathcal{P}_{11/20}$ in 88% of Λ.*

Additional computational efforts may suffice to do increase the number and establish the result for $(100 - \varepsilon)\%$ for any $\varepsilon > 0$. However, we lack a satisfactory local theory ensuring that 'infinitesimal rotations' cannot create a valid Rupert passage, i.e. we cannot check that $\mathcal{P}_{11/20}$ is not locally Rupert. It is not clear to us whether $\mathcal{P}_{11/20}$, a polyhedron on 8 vertices, is the simplest example of a convex polyhedron that is not Rupert (if indeed it is not Rupert); however, it is a reasonable candidate.

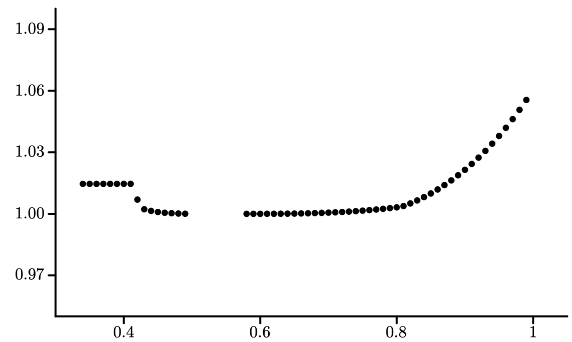


Figure 6.4: Numerical Nieuwland numbers for $\mathcal{P}_a$ plotted against a computed using the techniques of [35].

6.1.4 Main Result

6.2 Lemmas

Note that throughout this section, we refer to $\mathcal{P}_a$ instead of $\mathcal{P}_{11/20}$ as the statements and results are not specific to a particular value of a.

6.2.1 The Polygon Setting

Given two orientations R_1, R_2 of $\mathcal{P}_a$, we would like to check whether $R_1\mathcal{P}_a$ can "fit inside" $R_2\mathcal{P}_a$. Moreover, if $R_1\mathcal{P}_a$ "cannot fit inside" $R_2\mathcal{P}_a$, we would hope to say that $R_1'\mathcal{P}_a$ also "cannot fit inside" $R_2'\mathcal{P}_a$ for orientations R_1', R_2' that are "close" to R_1, R_2 respectively. We now establish the necessary machinery to make all of this precise. Let P, Q be convex polygons in $\mathbb{R}^2$ with vertices $\{p_i\}_{i\in[n]}$, $\{q_j\}_{j\in[m]}$ (ordered counterclockwise) respectively. Like in the polyhedron setting, when we say polygon, we refer to the convex hull of a finite set of points in convex position in the plane. Let $\{n_i\}$ denote the outward facing normals of Q, i.e.

$$n_i = R_{\pi/2}(q_i - q_{i+1}).$$

Then $x \in Q$ if and only if $\langle n_i, x\rangle \leq b_i$ for all i, where $b_i = \langle n_i, q_i\rangle$. Let $s \in \mathbb{R}$ and $t \in \mathbb{R}^2$, and consider the set of inequalities

$$\langle n_j, sp_i + t\rangle \leq b_j, \tag{6.1}$$

for all $i \in [n], j \in [m]$. Geometrically, values of s, t that satisfy this system of inequalities correspond to a scaling of P that can be translated to fit inside Q, i.e.

$$sP + t \subset Q^\circ. \tag{6.2}$$

We say that P *fits inside* Q or Q *contains* P if there exists $t \in \mathbb{R}^2$ and $s > 1$ satisfying 6.2. If s^* is the largest possible value of s, we say that P *fits inside* Q *up to scale* s^*. Note that this relation is transitive, i.e. for polygons A, B, C if A fits inside B and B fits inside C, then A fits inside C. We remark that others use similar language to allow rotating P at the price of a more complicated verification [26]. In our setting, we drop this rotation in favor of a simpler condition due to the relative simplicity of our polyhedron.

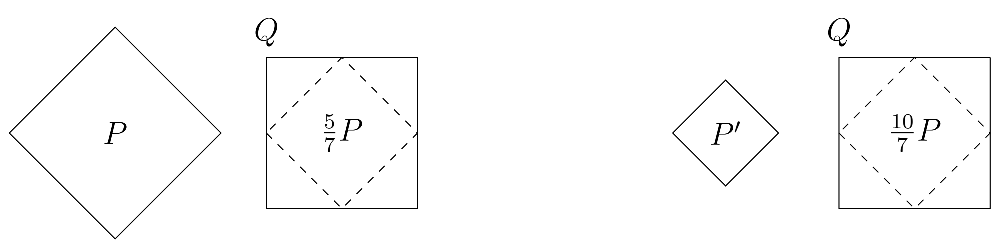


Figure 6.5: P does not fit inside Q, but P' does fit inside Q.

For a convex polygons P, we say that a δ-perturbation of P is any polygon P' such that $\|p_i - p'_i\| < \delta$ for all i. Then we have the following lemma.

Lemma 19. *For convex polygons P, Q, let P' be a polygon that is contained by all δ_1-perturbations of P and Q' be a polygon that contains all δ_2-perturbations of Q. Let s^* be the optimal solution to the linear program with objective function s and constraints*

$$\langle n_j, sp'_i + t\rangle \leq \langle n_j, q'_j\rangle,$$

for all $i \in [n], j \in [m]$. If $s^* < 1$, then no δ_1-perturbation of P can fit inside any δ_2-perturbation of Q.

Proof. If $s^* < 1$, then P' does not fit inside Q'. Let $\widetilde{P}$ be a δ_1-perturbation of P and $\widetilde{Q}$ be a δ_2-perturbation of Q. Then since P' fits inside $\widetilde{P}$ and $\widetilde{Q}$ fits inside Q', we immediately have that $\widetilde{P}$ cannot fit inside $\widetilde{Q}$. $\square$

6.2.2 The Polyhedron Setting

We encode our orientations in the following way. Let $M_{\theta,\phi}$ be the rotation matrix given by

$$\begin{aligned} M_{\theta,\phi} &= \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\cos\theta\cos\phi & -\sin\theta\cos\phi & \sin\phi \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sin\phi & 0 & \cos\phi \\ 0 & 1 & 0 \\ -\cos\phi & 0 & \sin\phi \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \end{aligned}$$

and let R_α denote counterclockwise rotation by α in the plane. Then a polyhedron $\mathcal{P}$ is Rupert if and only if there exist angles $\alpha, \theta, \phi, \alpha', \theta', \phi'$ such that $P = R_\alpha \circ M_{\theta,\phi}\mathcal{P}_a$ fits inside $Q = R_{\alpha'} \circ M_{\theta',\phi'}\mathcal{P}_a$. Notice that we can always assume $\alpha' = 0$, since we can rotate P by $-\alpha'$. This allows us to encode $\mathrm{SO}(3) \times \mathrm{SO}(3)$ as a 5-dimensional rectangular prism, i.e. $[0, 2\pi]^5$. If we have that $P = R_\alpha \circ M_{\theta,\phi}\mathcal{P}_a$ fits inside $Q = M_{\theta',\phi'}\mathcal{P}_a$, we say that $(\alpha, \theta, \phi, \theta', \phi')$ is a *Rupert passage*.

Lemma 20 (Steininger-Yurkevich)**.** *For $\varepsilon > 0$, if $|\alpha - \alpha'|, |\theta - \theta'|, |\phi - \phi'| \leq \varepsilon$, then*

$$\|M_{\theta,\phi} - M_{\theta',\phi'}\| \leq \sqrt{2}\varepsilon, \quad \textit{and} \quad \|R_\alpha M_{\theta,\phi} - R_{\alpha'} M_{\theta',\phi'}\| \leq \sqrt{5}\varepsilon.$$

For $P = R_\alpha \circ M_{\theta,\phi}\mathcal{P}_a$, we say that $P' = R_{\alpha'} \circ M_{\theta',\phi'}$ is an ε-angle perturbation of P if $|\alpha - \alpha'|, |\theta - \theta'|, |\phi - \phi'| \leq \varepsilon$. Then with these two lemmas, we obtain as a corollary the following.

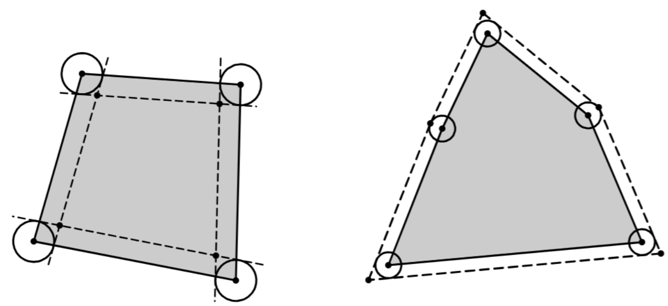


Figure 6.6: The construction of polygons P' (left) and Q' (right) as in the Corollary.

Corollary. *Let $P = R_\alpha \circ M_{\theta,\phi}\mathcal{P}_a$ and $Q = M_{\theta',\phi'}\mathcal{P}_a$. Let P' be a polygon that is contained by all ε-angle perturbations of P and Q' be a polygon that is contained by all ε-angle perturbations of Q and suppose P' fits inside Q' up to scale s. If $s < 1$, then there is no Rupert passage for any ε-angle perturbations of P, Q.*

So long as we have a way of constructing such P' and Q', this allows us a way of verifying that there is no Rupert passage within a full ε side-length cube in our 5-dimensional parameter space. One can do so using tangent lines to circles whose radii are the perturbation parameter, as in Figure 6.6.

6.2.3 Symmetries

In practice, $[0, 2\pi]^5$ is rather large, since the cubes we end up computing have small side length, in particular order 10^{-1} or smaller. Because the stellated tetrahedron exhibits tetrahedral symmetry, there are a number of symmetries that can reduce our parameter space and computation time.

Lemma 21. $M_{\theta+\pi/2,\phi}\mathcal{P}_a = -M_{\theta,\phi}\mathcal{P}_a = R_\pi \circ M_{\theta,\phi}\mathcal{P}_a$.

Proof. Compute

$$\begin{aligned} M_{\theta,\phi}(b) &= \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\cos\theta\cos\phi & -\sin\theta\cos\phi & \sin\phi \end{bmatrix} \begin{pmatrix} b_x \\ b_y \\ b_z \end{pmatrix} \\ &= \begin{pmatrix} -b_x\sin\theta + b_y\cos\theta \\ \cos\phi(-b_x\cos\theta - b_y\sin\theta) + b_z\sin\phi \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} M_{\theta+\pi/2,\phi}(c) &= \begin{bmatrix} -\cos\theta & -\sin\theta & 0 \\ \sin\theta\cos\phi & -\cos\theta\cos\phi & \sin\phi \end{bmatrix} \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix} \\ &= \begin{pmatrix} -c_x\cos\theta - c_y\sin\theta \\ \cos\phi(c_x\sin\theta - c_y\cos\theta) + c_z\sin\phi \end{pmatrix}. \end{aligned}$$

Then if $c_x = -b_y$, $c_y = b_x$, and $c_z = b_z$, we have that $M_{\theta,\phi}(b) = M_{\theta+\pi/2,\phi}(c)$. In particular, one can check that

$$\begin{aligned} M_{\theta,\phi}(p_0) &= -M_{\theta+\pi/2,\phi}(p_1), & M_{\theta,\phi}(q_0) &= -M_{\theta+\pi/2,\phi}(q_1), \\ M_{\theta,\phi}(p_1) &= -M_{\theta+\pi/2,\phi}(p_3), & M_{\theta,\phi}(q_1) &= -M_{\theta+\pi/2,\phi}(q_3), \\ M_{\theta,\phi}(p_2) &= -M_{\theta+\pi/2,\phi}(p_0), & M_{\theta,\phi}(q_2) &= -M_{\theta+\pi/2,\phi}(q_0), \\ M_{\theta,\phi}(p_3) &= -M_{\theta+\pi/2,\phi}(p_2), & M_{\theta,\phi}(q_3) &= -M_{\theta+\pi/2,\phi}(q_2). \end{aligned}$$

□

Let $\widehat{\theta} = \theta \pmod{\pi/2}$ and $\widehat{\theta'} = \theta' \pmod{\pi/2}$. Checking whether $P = R_\alpha \circ M_{\theta,\phi}\mathcal{P}_a$ fits in $Q = M_{\theta',\phi'}\mathcal{P}_a$ or vice versa is equivalent to checking whether P_2 fits in Q_2 or vice versa, where $P_2 = R_{\alpha+\beta} \circ M_{\widehat{\theta},\phi}\mathcal{P}_a$ and $Q_2 = M_{\widehat{\theta'},\phi'}$ for some angle $\beta \in \{0, \pi\}$. Consequently, we may restrict both θ parameters to the interval $[0, \pi/2]$, thereby reducing our parameter space to $[0, 2\pi]^3 \times [0, \pi/2]^2$.

Lemma 22. *$M_{\theta,\phi+\pi}\mathcal{P}_a = U \circ M_{\theta,\phi}\mathcal{P}_a$, where U denotes reflection across the x-axis in the plane.*

Proof. Compute

$$M_{\theta,\phi+\pi} = \begin{bmatrix} -\sin\theta & -\cos\theta & 0 \\ \cos\theta\cos\phi & \sin\theta\cos\phi & -\sin\phi \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \circ \begin{bmatrix} -\sin\theta & \cos\theta & 0 \\ -\cos\theta\cos\phi & -\sin\theta\cos\phi & \sin\phi \end{bmatrix} = U \circ M_{\theta,\phi}.$$

□

Let $\widehat{\phi} = \phi \pmod{\pi}$ and $\widehat{\phi}' = \phi' \pmod{\pi}$. If both $\phi, \phi' > \pi$, then checking whether $P = R_\alpha \circ M_{\theta,\phi}\mathcal{P}_a$ fits in $Q = M_{\theta',\phi'}$ or vice versa is equivalent to checking the same for P_2, Q_2, where $P_2 = R_\alpha \circ M_{\theta,\widehat{\phi}}\mathcal{P}_a$ and $Q_2 = M_{\theta',\widehat{\phi}'}$. Therefore, we need not have both ϕ, ϕ' range over the entirety of $[0, 2\pi]$. In particular, we can restrict ϕ to $[0, \pi]$, thereby reducing our parameter space again by an additional factor of 2, yielding a parameter space of $\Lambda = [0, 2\pi]^2 \times [0, \pi] \times [0, \pi/2]^2$.

6.3 Computational Approach

6.3.1 Implementation Details

There are a number of technical details that require discussion. First, computing s^* such that P fits in Q up to scale s^* is achieved by an LP solver. We use `fitScale` to denote a method that takes two polygons P and Q and returns s^*. This is a fundamental computation that is used many times when implementing the Corollary. To do so, we implement the subroutine `Check` (see Algorithm 1), which takes as input a 5-tuple $(\alpha, \theta, \phi, \theta', \phi')$ and a threshold b and returns a value ε such that no Rupert passage is possible within the ε side-length box centered on $(\alpha, \theta, \phi, \theta', \phi')$.

This is achieved by starting with a large initial perturbation value δ, which will result in P'

Algorithm 1 Check

```
P ← R_α ∘ M_{θ,φ} 𝒫_{11/20}
```

$P \leftarrow R_\alpha \circ M_{\theta,\phi}\mathcal{P}_{11/20}$

$Q \leftarrow M_{\theta',\phi'}\mathcal{P}_{11/20}$

$s_P \leftarrow \texttt{fitScale}(P, Q)$

$s_Q \leftarrow \texttt{fitScale}(Q, P)$

if $\max(s_P, s_Q) > 1$ **then**

  return INVALID

end if

$\delta \leftarrow 0.25$

while $\delta > b$ **do**

  $P_- \leftarrow \texttt{buffer}(P, -\sqrt{5}\delta)$

  $Q_+ \leftarrow \texttt{buffer}(Q, \sqrt{2}\delta)$

  $P_+ \leftarrow \texttt{buffer}(P, \sqrt{5}\delta)$

  $Q_- \leftarrow \texttt{buffer}(Q, -\sqrt{2}\delta)$

  $\delta \leftarrow 0.8\delta$

  $s_- \leftarrow \texttt{fitScale}(P_-, Q_+)$

  $s_+ \leftarrow \texttt{fitScale}(P_+, Q_-)$

  if $\max(s_-, s_+) < 1$ **then**

    return 2δ

  end if

end while

return 0

and Q' polygons that are too disparate in size allowing on to fit into the other. We then iteratively reduce δ until P' and Q' cannot fit into each other or until δ is below our threshold b. Note that the pairs (P_-, Q_+) and (P_+, Q_-) play the roles P' and Q' in the Corollary. We need two checks here to ensure that there is no containment in either direction. The `buffer` method computes buffer polygons as illustrated in Figure 6.6 using the `shapely` Python package. With `Check` described, we are now equipped to implement our main algorithm, whose aim is illustrated in Figure 6.7 and pseudocode in Algorithm 2.

Algorithm 2 Orthtree Search

$L \leftarrow \{C_1, \ldots, C_{32}\}$
$S \leftarrow \{\}$
while $|L| > 0$ **do**
  Pop cube C from L
  $w \leftarrow$ side length of C
  if C has side length less than ε **then**
    $S \leftarrow$ `Join`$(S, \{C\})$
  else
    $\varepsilon \leftarrow$ `Check`(C)
    if $\varepsilon < w$ **then**
      $\{C'_1, \ldots, C'_{32}\} \leftarrow$ `Split`(C)
      $L \leftarrow$ `Join`$(L, \{C'_1, \ldots, C'_{32}\})$
    end if
  end if
end while
return S

Our parameter space Λ can be decomposed into 32 cubes. Initialize a queue L with these 32 cubes $\{C_1, \ldots, C_{32}\}$. We maintain a list S of boxes with side length less than b. For each cube we apply `Check` to see if we can rule out any Rupert passages in the entire cube. If we can, we move on to the next cube. If we cannot, subdivide the cube into smaller cubes and append them to our queue. What this effectively does is partition our space in the manner of an orthtree, the arbitrary dimension generalization of quadtrees and octrees. This will rule out the regions of our parameter space that are sufficiently far from the 3-dimensional diagonal submanifold of $\mathrm{SO}(3) \times \mathrm{SO}(3)$. Here, `Join` appends queues or sets together, and `Split` takes as input a cube C and returns the 32 cubes obtained by splitting C in half along each coordinate direction. Running our algorithm yields our main result.

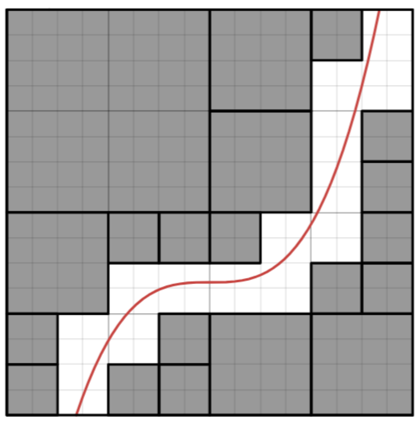


Figure 6.7: A 2-dimensional illustration of the algorithm.

6.3.2 Discussion

With some mild parallelization and a value of $w = 0.08$ (yielding cubes of side length $\pi/64 \approx 0.049$ in S), our algorithm ran in roughly two days. Reducing w would greatly increase increase this runtime, due to the exponential factor of 32 whenever `Split` is called. This would also increase the amount of Λ ruled to have no Rupert passages, though it is not clear by how much, as we have verified that some `Check` does struggle with some pairs of orientations even if they yield very different projections. This is likely due to the fact that the polygon constructions for the Corollary are too aggressive. However, even with unlimited computational power, our techniques would be insufficient to show that $\mathcal{P}_{11/20}$ is not Rupert as we lack a way to show that $\mathcal{P}_{11/20}$ is not locally Rupert. That is to say, we do not have a result of the form *for all* $R_1 \in SO(3)$*, there is no Rupert passage* R_1, R_2 *for any orientation* R_2 *that is* ε*-close to* R_1. Such a statement exists in [84] as their polyhedron has a particular nice structure, and perhaps most importantly, is centrally symmetric. This allows them to removing the translation term in the containment condition 6.2, which greatly simplifies local analysis. We do not have a local theorem that accounts for the translation term. Nevertheless, we conjecture that this stellated tetrahedron is not Rupert.

BIBLIOGRAPHY

[1] Abdulrahman Abdulaziz and Judy Said. On the contraction ratio of iterated function systems whose attractors are sierpinski n-gons. *Chaos, Solitons & Fractals*, 150:111140, 2021.

[2] Elsa Abenna, Simon Salamon, and Alfred Gray. *Modern Differential Geometry of Curves and Surfaces with Mathematica*. Chapmn and Hall/CRC, 2006.

[3] VA Aguilar-Arteaga, R Ayala-Figueroa, I González-García, and J Jerónimo-Castro. On evolutoids of planar convex curves ii. *Aequationes mathematicae*, 89:1433–1447, 2015.

[4] Arash Ahadi, Ali Dehghan, Mohammad-Reza Sadeghi, and Brett Stevens. On the algorithmic complexity of decomposing graphs into regular/irregular structures, arxiv:1801.08876.

[5] Saieed Akbari, Tommy R Jensen, and Mark Siggers. Decompositions of graphs into trees, forests, and regular subgraphs. *Discrete Mathematics*, 338(8):1322–1327, 2015.

[6] Alex Albors. Bad science matrix optimization github repository. https://github.com/alexalbors7/Bad-Science-Matrix-Optimization, 2024.

[7] Hilário Alencar, Walcy Santos, and Gregório Silva Neto. *Differential geometry of plane curves*, volume 96. American Mathematical Society, 2022.

[8] Tom M Apostol and Mamikon A Mnatsakanian. Tanvolutes: generalized involutes. *The American Mathematical Monthly*, 117(8):701–713, 2010.

[9] Peter Ashwin and Arek Goetz. Invariant curves and explosion of periodic islands in systems of piecewise rotations. *SIAM Journal on Applied Dynamical Systems*, 4(2):437–458, 2005.

[10] Peter Ashwin and Arek Goetz. Polygonal invariant curves for a planar piecewise isometry. *Transactions of the American Mathematical Society*, 358(1):373–390, 2006.

[11] Peter Ashwin, Arek Goetz, Pedro Peres, and Ana Rodrigues. Embeddings of interval exchange transformations into planar piecewise isometries. *Ergodic Theory and Dynamical Systems*, 40(5):1153–1179, 2020.

[12] Wojciech Banaszczyk. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998.

[13] Michael F Barnsley. *Fractals everywhere*. Academic press, 2014.

[14] Julien Bensmail, Martin Merker, and Carsten Thomassen. Decomposing graphs into a constant number of locally irregular subgraphs. *European Journal of Combinatorics*, 60:124–134, 2017.

[15] Claude Berge. Two theorems in graph theory. *Proceedings of the National Academy of Sciences*, 43(9):842–844, 1957.

[16] Claude Berge. Combinatorial games on a graph. *Discrete Mathematics*, 151(1-3):59–65, 1996.

[17] Sandro Bettin, Giuseppe Molteni, and Carlo Sanna. Greedy approximations by signed harmonic sums and the thue–morse sequence. *Advances in Mathematics*, 366:107068, 2020.

[18] Rick Billstein, Shlomo Libeskind, and Johnny W Lott. *A problem solving approach to mathematics for elementary school teachers*. Pearson London, 2013.

[19] Michael Boshernitzan and Arek Goetz. A dichotomy for a two-parameter piecewise rotation. *Ergodic Theory and Dynamical Systems*, 23(3):759–770, 2003.

[20] Henk Bruin and Jonathan Deane. Piecewise contractions are asymptotically periodic. *Proceedings of the American Mathematical Society*, 137(4):1389–1395, 2009.

[21] Darryn Bryant, Daniel Horsley, and Barbara Maenhaut. Decompositions into 2-regular subgraphs and equitable partial cycle decompositions. *Journal of Combinatorial Theory, Series B*, 93(1):67–72, 2005.

[22] Yann Bugeaud. *Approximation by algebraic numbers*, volume 160. Cambridge University Press, 2004.

[23] Horst Bunke. On a relation between graph edit distance and maximum common subgraph. *Pattern recognition letters*, 18(8):689–694, 1997.

[24] E Calabi and A Dvoretzky. Convergence-and sum-factors for series of complex numbers. *Transactions of the American Mathematical Society*, 70(1):177–194, 1951.

[25] Gary Chartrand, David Erwin, Garry L Johns, and Ping Zhang. Boundary vertices in graphs. *Discrete Mathematics*, 263(1-3):25–34, 2003.

[26] Bernard Chazelle. The polygon containment problem. *Advances in Computing Research*, 1:1–33, 1983.

[27] Xiaoyang Chen, Hongwei Huo, Jun Huan, and Jeffrey Scott Vitter. An efficient algorithm for graph edit distance computation. *Knowledge-Based Systems*, 163:762–775, 2019.

[28] Yitwah Cheung, Arek Goetz, and Anthony Quas. Piecewise isometries, uniform distribution and 3log 2- $\pi 2/8$. *Ergodic Theory and Dynamical Systems*, 32(6):1862–1888, 2012.

[29] Jonathan HB Deane. Piecewise isometries: Applications in engineering. *Meccanica*, 41(3):241–252, 2006.

[30] David B Dooner. On evolutoids and spatial involutoids to define hypoid flank geometry. *Mechanism and Machine Theory*, 156:104150, 2021.

[31] A. Dvoretzky and H. Hanani. Sur les changements des signes des termes d'une série á termes complexes. In *C. R. Acad. Sci. Paris*, 225, pages 516–518, 1947.

[32] Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of mathematics*, 17:449–467, 1965.

[33] Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2013.

[34] Philip Franklin. Discussions: On curves whose evolutes are similar curves. *The American Mathematical Monthly*, 28(4):168–170, 1921.

[35] Albin Fredriksson. Optimizing for the Rupert property. *The American Mathematical Monthly*, 131(3):255–261, 2023.

[36] Tibor Gallai. Kritische graphen, ii. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(3):373–395, 1963.

[37] Tibor Gallai. Maximale systeme unabhängiger kanten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 9(3):401–413, 1964.

[38] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13:113–129, 2010.

[39] Peter J Giblin and J Paul Warder. Evolving evolutoids. *The American Mathematical Monthly*, 121(10):871–889, 2014.

[40] Arek Goetz. Dynamics of a piecewise rotation. *Discrete and Continuous Dynamical Systems*, 4:593–608, 1998.

[41] Arek Goetz. Sofic subshifts and piecewise isometric systems. *Ergodic Theory and Dynamical Systems*, 19(6):1485–1501, 1999.

[42] Arek Goetz. Stability of piecewise rotations and affine maps. *Nonlinearity*, 14(2):205–219, 2001.

[43] Arek Goetz. Piecewise isometries—an emerging area of dynamical systems. *Fractals in Graz 2001: Analysis—Dynamics—Geometry—Stochastics*, pages 135–144, 2003.

[44] Arek Goetz and Anthony Quas. Global properties of a family of piecewise isometries. *Ergodic Theory and Dynamical Systems*, 29(2):545–568, 2009.

[45] Arkadiusz Goetz. *Dynamics of piecewise isometries.* University of Illinois at Chicago, 1996.

[46] Uffe Haagerup. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

[47] Marco Hamann. A note on ovals and their evolutoides. *Contributions to Algebra and Geometry*, 50(2):433–441, 2009.

[48] Theodore E Harris. The existence of stationary measures for certain markov processes. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 113–124, 1956.

[49] Michael Hochman. On self-similar sets with overlaps and inverse theorems for entropy. *Annals of Mathematics*, pages 773–822, 2014.

[50] John E Hutchinson. Fractals and self similarity. *Indiana University Mathematics Journal*, 30(5):713–747, 1981.

[51] Shyuichi Izumiya and Nobuko Takeuchi. Evolutoids and pedaloids of plane curves. *Note di Matematica*, 39(2):13–24, 2020.

[52] Richard P Jerrard, John E Wetzel, and Liping Yuan. Platonic passages. *Mathematics Magazine*, 90(2):87–98, 2017.

[53] Dieter Jungnickel. *Matchings*, pages 405–439. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[54] Vladimir M Kadets and Mikhail I Kadets. *Rearrangements of series in Banach spaces*, volume 86. American Mathematical Society Providence, RI, 1991.

[55] Sergiy Koshkin and Ivan Rocha. Caustics of light rays and euler's angle of inclination. *arXiv preprint arXiv:2206.09510*, 2022.

[56] Dmitriy Kunisky. The discrepancy of unsatisfiable matrices and a lower bound for the komlós conjecture constant. *SIAM Journal on Discrete Mathematics*, 37(2):586–603, 2023.

[57] J.D. Lawrence. *A Catalog of Special Plane Curves.* Dover Books on Mathematics. Dover Publications, 1972.

[58] Michel Ledoux. *The concentration of measure phenomenon.* Number 89. American Mathematical Soc., 2001.

[59] Julien Lemoine. R\'esolution du jeu de juniper green. *arXiv preprint arXiv:2202.09864*, 2022.

[60] Enze Li, Donghe Pei, and Masatomo Takahashi. Enveloids of legendre curves in the unit tangent bundle over the euclidean plane. *Bulletin of the Malaysian Mathematical Sciences Society*, 45(6):3011–3041, 2022.

[61] GH Light. Discussions: Note on curves whose evolutes are similar curves. *The American Mathematical Monthly*, 27(7/9):303–306, 1920.

[62] Ernst Lindelöf. Differentiali-ja integralilasku ja sen sovellutukset i. *Yhden muuttujan funktiot*, 2, 1950.

[63] E.H. Lockwood. *A Book of Curves*. University Press, 1967.

[64] R Daniel Mauldin and Stanley C Williams. Hausdorff dimension in graph directed constructions. *Transactions of the american mathematical society*, 309(2):811–829, 1988.

[65] Martin Merker. Graph decomposition. *Technical University of Denmark*, 2016.

[66] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[67] Pat AP Moran. Additive functions of intervals and hausdorff measure. In *Classics on Fractals*, pages 240–257. CRC Press, 2019.

[68] Michel Neuhaus, Kaspar Riesen, and Horst Bunke. Fast suboptimal algorithms for the computation of graph edit distance. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, SSPR 2006 and SPR 2006, Hong Kong, China, August 17-19, 2006. Proceedings*, pages 163–172. Springer, 2006.

[69] Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods*. SIAM, 1992.

[70] Paul P Park, Paul B Umbanhowar, Julio M Ottino, and Richard M Lueptow. Mixing with piecewise isometries on a hemispherical shell. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(7), 2016.

[71] Jakub Przybyło. On decomposing graphs of large minimum degree into locally irregular subgraphs. *The Electronic Journal of Combinatorics*, 23(2):P2–31, 2016.

[72] Dan Reznik, Ronaldo Garcia, and Hellmuth Stachel. Area-invariant pedal-like curves derived from the ellipse. *Beiträge zur Algebra und Geometrie/Contributions to Algebra and Geometry*, pages 1–19, 2021.

[73] Kaspar Riesen and Horst Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image and Vision computing*, 27(7):950–959, 2009.

[74] Kaspar Riesen and Horst Bunke. Improving approximate graph edit distance by means of a greedy swap strategy. In Abderrahim Elmoataz, Olivier Lezoray, Fathallah

Nouboud, and Driss Mammass, editors, *Image and Signal Processing*, pages 314–321, Cham, 2014. Springer International Publishing.

[75] Kaspar Riesen, Stefan Fankhauser, and Horst Bunke. Speeding up graph edit distance computation with a bipartite heuristic. In *MLG*, pages 21–24. Citeseer, 2007.

[76] Evan Scott. Two sufficient conditions for a polyhedron to be (locally) Rupert. *arXiv preprint arXiv:2208.12912*, 2022.

[77] George F Simmons. *Calculus gems: brief lives and memorable mathematics*, volume 55. MAA, 2007.

[78] Lauren D Smith, Paul B Umbanhowar, Richard M Lueptow, and Julio M Ottino. The geometry of cutting and shuffling: An outline of possibilities for piecewise isometries. *Physics Reports*, 802:1–22, 2019.

[79] L Lynn Stallings and Patricia L Bullock. Juniper green. *Mathematics Teaching in the Middle School*, 4(7):438–440, 1999.

[80] Stefan Steinerberger. The boundary of a graph and its isoperimetric inequality. *Discrete Applied Mathematics*, 338:125–134, 2023.

[81] Stefan Steinerberger. Bad science matrices. *arXiv preprint arXiv:2402.03205*, 2024.

[82] Stefan Steinerberger and Tony Zeng. A curious dynamical system in the plane. *arXiv preprint arXiv:2409.08961*, 2024.

[83] Jakob Steininger and Sergey Yurkevich. An algorithmic approach to rupert's problem. *Mathematics of Computation*, 92(342):1905–1929, 2023.

[84] Jakob Steininger and Sergey Yurkevich. A convex polyhedron without Rupert's property. *arXiv preprint arXiv:2508.18475*, 2025.

[85] Ian Stewart. Juniper green. *Scientific American*, 276(3):118–120, 1997.

[86] Ian Stewart. The lore and lure of dice. *Scientific American*, 277(5):110–113, 1997.

[87] Ian Stewart. *Professor Stewart's hoard of mathematical treasures.* Basic Books, 2010.

[88] James Johnston Stoker. *Differential geometry*, volume 20. John Wiley & Sons, 2011.

[89] Dirk J Struik. *Lectures on classical differential geometry.* Courier Corporation, 2012.

[90] Rob Sturman. The role of discontinuities in mixing. *Advances in applied mechanics*, 45:51–90, 2012.

[91] İşıkhan Uğurel and Sevgi Moralı. Matematik ve oyun etkileşimi. *Gazi Eğitim Fakültesi Dergisi*, 28(3):75–98, 2008.

[92] Johannes van der Corput. Verteilungsfunktionen ii. In *Nederl. Akad. Wetensch. Proc.*, volume 38, pages 1058–1066, 1935.

[93] Jean Henri van Swinden. *Grondbeginsels der meetkunde*. P. den Hengst en zoon, 1816.

[94] John Wallis. *De Algebra Tractatus; Historicus & Practicus*, volume 2. E Theatro, Sheldoniano, Oxoniae, 1685.

[95] Norimasa Yoshida and Takafumi Saito. The evolutes of log-aesthetic planar curves and the drawable boundaries of the curve segments. *Computer-Aided Design and Applications*, 9(5):721–731, 2012.

[96] Zhiping Zeng, Anthony K. H. Tung, Jianyong Wang, Jianhua Feng, and Lizhu Zhou. Comparing stars: On approximating graph edit distance. *Proc. VLDB Endow.*, 2(1):25–36, aug 2009.

[97] Michał Zwierzyński. The singular evolutoids set and the extended evolutoids front. *Aequationes mathematicae*, 96(4):849–866, 2022.